A comparative evaluation of wavelet-based methods for hypothesis testing of brain activation maps

Jalal M. Fadili, E. T. Bullmore

A comparative evaluation of wavelet-based methods for hypothesis testing of brain activation maps

M.J. Fadili[a,*] and E.T. Bullmore[b]

[a]*Image Processing Group, GREYC CNRS UMR 6072 14050, Caen Cedex, France*
[b]*Brain Mapping Unit and Wolfson Brain Imaging Centre, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QQ, United Kingdom*

Wavelet-based methods for hypothesis testing are described and their potential for activation mapping of human functional magnetic resonance imaging (fMRI) data is investigated. In this approach, we emphasise convergence between methods of wavelet thresholding or shrinkage and the problem of hypothesis testing in both classical and Bayesian contexts. Specifically, our interest will be focused on the trade-off between type I probability error control and power dissipation, estimated by the area under the ROC curve. We describe a technique for controlling the false discovery rate at an arbitrary level of error in testing multiple wavelet coefficients generated by a 2D discrete wavelet transform (DWT) of spatial maps of fMRI time series statistics. We also describe and apply change-point detection with recursive hypothesis testing methods that can be used to define a threshold unique to each level and orientation of the 2D-DWT, and Bayesian methods, incorporating a formal model for the anticipated sparseness of wavelet coefficients representing the signal or true image. The sensitivity and type I error control of these algorithms are comparatively evaluated by analysis of "null" images (acquired with the subject at rest) and an experimental data set acquired from five normal volunteers during an event-related finger movement task. We show that all three wavelet-based algorithms have good type I error control (the FDR method being most conservative) and generate plausible brain activation maps (the Bayesian method being most powerful). We also generalise the formal connection between wavelet-based methods for simultaneous multiresolution denoising/hypothesis testing and methods based on monoresolution Gaussian smoothing followed by statistical testing of brain activation maps.

Keywords: Wavelet-based methods; Brain activation maps; False discovery rate

Introduction

Wavelet shrinkage

Nonparametric regression has been a fundamental tool in data analysis over the past two decades and is still an expanding area of research. The goal is to recover an unknown process, say g, based on sampled data that are contaminated with noise. It has been proven that wavelet "shrinkage" methods for nonparametric regression are optimal for this purpose, in the sense of closely approximating the minimax risk, assuming only that g belongs to a general class of functions with a prescribed regularity. It has also been shown empirically that wavelet-based denoising techniques can provide a very effective and simple way of finding structure in a variety of data sets without the imposition of a parametric regression model, see overview in Antoniadis et al. (2001). Wavelet shrinkage methods are a new subset of a prior class of nonparametric regression estimators, namely orthogonal series methods, and can be rapidly computed using fast algorithms (Mallat, 1989).

The seminal papers on wavelet shrinkage by Donoho and Johnstone defined it algorithmically as follows: the discrete wavelet transform of a noisy realisation of g is computed; the wavelet coefficients are thresholded by a universal threshold value λ applied identically to each coefficient according to a hard or soft thresholding rule η; the inverse wavelet transform of the coefficients that survive thresholding is used to estimate the denoised image g (Donoho and Johnstone, 1994, 1995; Donoho et al., 1995). Several variants or refinements of this basic scheme have since been proposed, including level-specific or data adaptive thresholding operators (see, e.g., Percival and Walden, 2000; Vidakovic, 1999 and references therein), and Bayesian approaches by which a prior distribution is specified to model the sparseness of the coefficients of the true image and the posterior probability is thresholded (Abramovich et al., 1998; Achim et al., 2001; Chang et al., 2000; Chipman et al., 1997; Clyde and George, 1999, 2000; Crouse et al., 1998; Johnstone and Silverman, 1998; Huang and Lu, 2000; Simoncelli and Adelson, 1996; Vannucci and Corradi, 1999; Vidakovic, 1998). In terms of mean squared error in finite

sample situations, it has been shown that Bayesian wavelet shrinkage rules often outperform "classical" thresholding algorithms, which operate on wavelet coefficients one at a time and without a prior distribution for coefficients of the true image. A detailed study involving recent classical and Bayesian methods in the development of high-performance wavelet shrinkage algorithms and their finite sample properties was reported by Antoniadis et al. (2001).

Brain activation mapping

Many difficulties must be addressed when processing fMRI data including the generally low signal-to-noise ratio and the multiple sources of artefacts. To detect activated brain regions, the most widely adopted data analytic procedures are based on linear (or, less often, nonlinear) regression theory. The experimental effects of interest are included in a design matrix that often also incorporates some other "nuisance" variables, e.g., low-frequency drift, head movements, etc. Then, the time course vector at each pixel (m, n) can be expressed as a linear combination of some covariates $\mathbf{x}_k$ via an additive linear model:

$$\mathbf{y}_{mn} = \mathbf{X}\beta_{mn} + \varepsilon_{mn}, \quad \varepsilon_{mn} \sim \mathcal{N}(0, \Sigma_{mn}) \tag{1}$$

where ε_{mn} is a zero mean random vector with covariance matrix $\mathbf{\Sigma}_{\mathrm{mn}}$. $\mathbf{X}$ is the design matrix whose columns are the $\mathbf{x}_k$'s. The classical assumption is that the errors ε_{mn} are independent and normally distributed, which is not generally realistic for fMRI time series with autocorrelated errors. There is also an extended literature on modelling ε_{mn} as a short-memory process (AR, ARMA, ARIMA; see Bullmore et al., 2001 and references therein). Recently, we have proposed a new wavelet-based estimator of such linear models that is robust to the presence of long-memory ($1/f$) error processes (Bullmore et al., 2001; Fadili and Bullmore, 2001).

Whatever method is used for regression model estimation, the output will be a 3D spatial map representing the linear model parameter of interest (often divided by its standard error) at each voxel in the image. Various statistical maps, including Student's t maps, Fisher–Snedecor F maps or Gaussian Z maps, have been reported in the neuroimaging literature. In the following, we will consider exclusively Gaussian statistical maps, although this is without loss of generality because any other standardised statistic map can be transformed to a Z map using appropriate integral transformations. Given such a spatial brain map of time series statistics, the problem at the heart of this paper is *how best to estimate the true activation map from its noisy realisation*? Almost universally in current practice, this problem is approached in two steps.

The first step is to reduce the noise in the observed map. This is most often done using a monoresolution Gaussian filter whose width must be arbitrarily specified. This will generally entail loss of resolution of spatially detailed features in the map and will cost sensitivity to detect any spatial features of the true image that do not conform in size or shape to the Gaussian kernel. Multiresolution Gaussian scale-space methods have been used by Poline and Mazoyer (1994) and revisited by Worsley et al. (1996a,b), but may be problematic due to the continuous nature of the scale parameter (which and how many scales should be used?), the overcompleteness or redundancy of Gaussian scale-space, and difficulties in knowing the distribution and covariance structure of non-Gaussian noise after Gaussian smoothing. For these reasons, it is easier to work with orthonormal transforms such as the discrete wavelet transform (DWT) that correspond to bases or tight frames.

The second step is to decide which voxels in the smooth map are activated, for a prespecified type I probability error α. This amounts to a binary decision problem where each voxel is compared to the α-level critical threshold. This decision step is currently implemented at the level of individual voxels or spatial clusters of contiguous suprathreshold voxels surviving a preliminary voxel-level test (Bullmore et al., 1999; Worsley et al., 1995). At voxel level, the search volume V, or total number of tests to be conducted, will typically be in the order of 1E4 and false-positive error may be controlled for multiple comparisons in terms of the family-wise error (FWE) rate or the false discovery rate (FDR); (see Friston et al., 1991; Hayasaka and Nichols, 2003; Nichols and Holmes, 2002; Worsley et al., 1996a,b for reviews).

Wavelet shrinkage and brain activation mapping

The overall aim of this paper is to advocate wavelets as a framework for combining the traditionally separate steps of denoising and hypothesis testing in a single operation. To put this another way, we will reformulate probabilistic wavelet shrinkage as a method for multiresolution denoising *and* hypothesis testing. The main advantages, in principle, of this approach are the following: (i) the discrete wavelet transform effects a multiresolution decomposition of spatial statistic maps, so denoising in the wavelet domain should be more adaptive to a range of spatial features in the true image than imposition of a monoresolution Gaussian kernel; and (ii) the wavelet transform often has whitening or decorrelating properties, meaning a set of hypothesis tests on wavelet coefficients may generally be regarded as independent of each other, which may confer some benefits in optimal control of type I error.

In the neuroimaging literature, some authors have already proposed using wavelet domain denoising to obtain an estimate of activation maps in fMRI (Brammer, 1998; Desco et al., 2001) and positron emission tomography (PET) (Ruttimann et al., 1994; Turkheimer et al., 1999; see also the review in Bullmore et al., 2003). However, these approaches have shared the drawback of applying a simple universal threshold to all wavelet coefficients without assigning any probability risk to the resulting activation maps, and the multiple comparisons problem has not always been explicitly addressed. To our knowledge, Ruttimann et al. (1995, 1996, 1998) were the first group to propose a wavelet-based hypothesis testing approach for brain activation mapping using Bonferroni-corrected thresholds unique to each level of the 2D-DWT of spatial statistic maps. In the same spirit, Feilner et al. (1999, 2000) advocated fractional spline wavelets for analysis of time series statistic maps, using the fractional order of the spline to control the smoothness of the reconstructed image. Raz and Turetsky (1999) combined ANOVA and wavelet-based false discovery rate thresholding for the analysis of single and group neuroimaging fMRI data, using a method very closely related to Ruttiman's approach. However, there has been no previous attempt to evaluate various wavelet shrinkage algorithms rigorously in terms of their relative sensitivity and type I error control, to apply such methods to mapping data derived from event-related experiments, or to formalise the conceptual links between wavelet

shrinkage and the prevailing alternative approach of Gaussian smoothing followed by a separate hypothesis testing step (with possible correction for multiple comparisons). Rather than being exhaustive, we concentrate our efforts on three algorithms because they are representative of the prior literature; one controlling the FDR, one the FWER and one Bayesian with no multiple comparison correction.

The rest of this paper is organised as follows: in the next section, we define wavelet shrinkage or nonparametric regression and introduce some notational aspects. In the third section, we specify three algorithms for probabilistic wavelet shrinkage: an algorithm using a universal threshold to control the false discovery rate; an algorithm using level-adaptive thresholds recursively to test coefficients; and a Bayesian algorithm for thresholding the posterior probability of the null hypothesis given the observed coefficients. We also briefly review some details of receiver operating characteristic (ROC) methodology that will be used to compare the performance of these algorithms. Then, we describe experimental procedures and preliminary data analysis and simulations. In "Results", we present some experimental results on null and experimental fMRI data sets, using ROC methods to compare performance of wavelet based hypothesis testing algorithms. Finally, some conclusions and directions of future work are briefly summarised.

Wavelet shrinkage or nonparametric regression

Let g_{mn}, $m, n = 0, \ldots, N-1$ be equally spaced samples of a real-valued image; without loss of generality, N is considered as a power of 2 ($N = 2^J$). Now consider the standard nonparametric regression model:

$$y_{mn} = g_{mn} + \epsilon_{mn} \tag{2}$$

where ϵ_{mn} are iid normal random variables with mean zero and variance σ^2 independent of g_{mn}. The goal is to recover the underlying function or true image g from the observed noisy data y_{mn}, without assuming any particular parametric structure for g. Let $\mathbf{y}$, $\mathbf{g}$ and $\boldsymbol{\epsilon}$ denote the matrix representations of the corresponding entities and let $\mathbf{D} = \mathcal{W}\mathbf{y}$, $\mathbf{S} = \mathcal{W}\mathbf{g}$ and $\mathbf{V} = \mathcal{W}\boldsymbol{\epsilon}$, where $\mathcal{W}$ is the two dimensional dyadic orthonormal wavelet transform (DWT) operator (Mallat, 1999); for background material on wavelets and fMRI, see reviews by Bullmore et al. (2003). In a two-dimensional setting, the subbands HH_j, HL_j and LH_j, $j = J_c, \ldots, J-1$ correspond to the detail coefficients in diagonal, horizontal and vertical orientations, and the subband LL_{J_c} is the approximation or the smooth component. J_c is the coarsest scale of the decomposition that will usually be specified as $J_c = \log_2 \log N + 1$ from asymptotic considerations. Let s^{oj}_{mn} be the detail coefficient of the true image $\mathbf{g}$ at location (m, n), scale j and orientation o, and similarly for d^{oj}_{mn} and v^{oj}_{mn}. Due to the orthogonality of the DWT, the wavelet coefficients of white noise will also be white noise with the same variance. It follows from Eq. (2) that:

$$d^{oj}_{mn} = s^{oj}_{mn} + v^{oj}_{mn}, \quad j = J_c,\ldots,J-1; \quad m,n = 0,\ldots,2^j - 1 \tag{3}$$

The sparseness of the wavelet expansion makes it reasonable to assume that essentially only a few large detail coefficients in $\mathbf{D}$ contain information about the underlying image $\mathbf{g}$, while small values can be attributed to the noise that uniformly contaminates all wavelet coefficients. It is also advisable to keep the approximation coefficients intact because they represent low-frequency terms that usually contain important features about the image $\mathbf{g}$. By thresholding or shrinking the detail coefficients and inverting the DWT, one can obtain an estimate of the underlying image $\mathbf{g}$. So the resulting three-step wavelet-based estimation procedure can be summarised by the following diagram:

$$\mathbf{y} \xrightarrow{\text{DWT}} \mathbf{D} \xrightarrow{\text{Nonlinear thresholding operator } \eta} \left\{\hat{\mathbf{S}} = \eta(\mathbf{D})\right\} \xrightarrow{\text{IDWT}} \hat{\mathbf{g}}$$

where η is a nonlinear (shrinkage or thresholding) operator. Examples of such an operator are the hard and soft thresholding rules with the universal threshold (Donoho and Johnstone, 1994).

These rules are given, respectively, by:

$$\eta^{\text{Hard}}_{\lambda}\left(d^{oj}_{mn}\right) = \begin{cases} 0 & \text{if } |d^{oj}_{mn}| \le \lambda \\ d^{oj}_{mn} & \text{otherwise} \end{cases}, \eta^{\text{Soft}}_{\lambda}\left(d^{oj}_{mn}\right)$$

$$= \begin{cases} 0 & \text{if } |d^{oj}_{mn}| \le \lambda \\ d^{oj}_{mn} - \text{sign}\left(d^{oj}_{mn}\right)\lambda & \text{otherwise.} \end{cases} \tag{4}$$

where λ is the threshold value.

Wavelet shrinkage as a hypothesis testing problem

The main idea here is simply to reformulate wavelet shrinkage as a hypothesis testing problem.

Classical approach

We mean by "classical" that no prior distribution is imposed on the true unknown wavelet coefficients, in contrast to the Bayesian approach where a prior density is specified to capture the sparseness of wavelet coefficients of the true image g. In both the classical hypothesis testing algorithms described below, the investigator has some control on the smoothness of the reconstructed image by prespecification of an arbitrary type I error probability threshold, e.g., $\alpha \le 0.05$.

Controlling false discovery rate

For each observed wavelet coefficient at each scale, orientation, and location, we test the following hypothesis:

$$H_0 : s^{oj}_{mn} = 0 \text{ vs. } H_1 : s^{oj}_{mn} \neq 0$$

The observed detail coefficient is distributed according to $d^{oj}_{mn} / s^{oj}_{mn} \sim \mathcal{N}(s^{oj}_{mn}, \sigma^2)$. This detail coefficient is retained in the reconstruction if H_0 is rejected with a risk α; otherwise, it is discarded. Classical approaches to multiple hypothesis testing face serious problems because of the large number of hypotheses being tested simultaneously. In other words, if the error is controlled at an individual level, the test is too permissive and the chance of erroneously retaining a coefficient over the search volume is extremely high; whereas if the family-wise error is controlled, the test is too conservative and the chance of falsely discarding a coefficient is extremely high. Abramovich and Benjamini (1995, 1996) and Shen et al. (2002) have proposed a way to control such dissipation of power based on type I error control in terms of the false discovery rate (FDR). From an estimation point of view, Abramo-

vich et al. (2000) showed recently that FDR thresholding provides a near-optimal way of adapting to unknown sparsity of the signal to be recovered; in other words, imposing an asymptotically negligible FDR level is asymptotically optimal in terms of the minimax risk.

Let T be the number of observed wavelet coefficients that are retained by the thresholding procedure. From these T coefficients, TP (true positives) are correctly retained and FP = T − TP (false positives) are erroneously retained. The error in such a procedure is expressed in terms of the random variable FPF = FP/T, i.e., the proportion of the retained wavelet coefficients that should properly have been rejected. Obviously, FPF is defined as zero when $T = 0$ because no error of this type can be made when no coefficient is retained. The FDR of empirical wavelet coefficients can now be defined as the expectation of FPF, i.e., the expected proportion of false positives among the total number of coefficients surviving the threshold.

Following Abramovich and Benjamini (1995, 1996), we propose maximizing the number of retained wavelet coefficients subject to the constraint FPF < α, which can be operationalised in terms of Algorithm 1 to calculate the global threshold of the map.

Algorithm 1 (*Classical wavelet shrinkage with FDR control*)

1: For each of the $nd = N \times N - 1$ observed wavelet coefficients $\{d_{mn}^{oj} : j = 0, \ldots, J-1; m, n = 0, \ldots, 2^j - 1; o = HH, HL, LH\}$, calculate the corresponding double-sided p value, p_{mn}^{oj} under H_0:

$$p_{mn}^{oj} = 2\left(1 - \Phi\left(\frac{|d_{mn}^{oj}|}{\hat{\sigma}}\right)\right) \tag{5}$$

where $\Phi(x)$ is the cumulative normal distribution of a standard normal variable. In general, the wavelet coefficient variance σ can be estimated from the coefficients in the horizontal HH orientation at the finest scale using the popular robust estimator (Donoho and Johnstone, 1994):

$$\hat{\sigma} = \frac{\mathrm{MAD}\left(d_{mn}^{HH_1}\right)}{0.6745} \tag{6}$$

where MAD is the median absolute deviation. In the context of this paper, we are dealing with normalised Gaussian scores that have unit variance by definition.

We can see that all the $N^2 - 1$ wavelet detail coefficients of the image are being tested in this step. In the context of fMRI, one is usually interested only in voxels representing brain tissue and not skull, scalp, or surrounding air. These noncerebral elements of the image are therefore discarded in a preliminary image preprocessing step (thresholding and then morphological closing); it follows that the number nd has to be reduced by taking into account only the number of intracranial voxels V.

2: Sort the p_{mn}^{oj} in an ascending order, $p_1 \le p_2 \le \ldots \le p_V$.

3: Find the p value p_i such that $i = \max_i (p_i \le (i/V)\alpha)$.

4: Calculate the critical threshold corresponding to this double-sided p value p_i:

$$\lambda_{\mathrm{FDR}} = \hat{\sigma}\Phi^{-1}\left(1 - \frac{p_i}{2}\right) \tag{7}$$

5: Use λ_{FDR} and apply classical hard (kill or keep) or soft (kill or shrink) thresholding rules to the observed wavelet coefficients up to coarsest level J_c.

6: Apply the inverse DWT to obtain an estimate of the image g.

Clearly this is a procedure in which the same threshold value λ_{FDR} is applied to coefficients at all scales and orientations of the decomposition. Note that Donoho's use of a universal threshold value can also be viewed as a (highly conservative) multiple hypothesis testing procedure rejecting *each* null hypothesis at a critical threshold λ_U. Using the well-known asymptotics $\Phi(x) \sim 1 - \phi(x)/x$ for x large, one can easily verify from Eq. (5) that the corresponding significance level for all tests will be $1/\left(N^2\sqrt{\pi \log N^2}\right)$ for N large. Thus, as mentioned by Abramovich and Benjamini (1995), Donoho's procedure is equivalent to the "panic" procedure of controlling the probability of even one erroneous inclusion of a wavelet coefficient at the level $1/\left(N^2\sqrt{\pi \log N^2}\right)$. The level at which this error probability is controlled goes to zero as N gets larger.

Thresholding as a change-point detection problem

Rather than seeking to include as many wavelet coefficients as possible (subject to constraint in terms of type I error), the recursive hypothesis testing procedure (Ogden and Parzen, 1996) includes a wavelet coefficient only when there is strong, affirmative evidence that it is needed in the reconstruction. Here we use prior work by Ogden and Parzen (1996) to develop a recursive procedure for multiple hypothesis testing that uses level-dependent thresholds λ_j.

For any orientation (and for sake of readability, we will omit the superscript o in the following notation), let $\tilde{d}_i^j = (d_i^j/\sigma)$ be independent random variables $\mathcal{N}(\tilde{s}_i^j, 1)$, $i = \{1, \ldots, n_j\}$ that represent the observed wavelet coefficients at any level j, where $n_j = 2^{2j}$ if all detail coefficients at scale j are tested. For fMRI data, and for the same reason as above, n_j is actually the number of intracranial voxels at scale j. Let I_{n_j} represent a non-empty subset of indices $\{1, \ldots, n_j\}$. Then, the multiple hypothesis testing problem can be expressed as:

$$H_0 : \tilde{s}_i^j = 0, i \in I_{n_j} \text{ vs. } H_1 : \tilde{s}_i^j \neq 0, i \in I_{n_j} \text{ and } \tilde{s}_i^j = 0, i \in I_{n_j} \tag{8}$$

To test this set of hypotheses, we can use the standard likelihood ratio test (LRT) (Ogden and Parzen, 1996). If the cardinality of the set I_{n_j} is not known, which is the case in practice, the LRT for the above hypotheses could be based on the test statistic $\sum_{i=1}^{n_j} |\tilde{d}_i^j|^2$, which is distributed as $\chi_{n_j}^2$ under H_0. However, this may not be the most appropriate test statistic as only a few of the $\tilde{s}_i^j$'s are non-zero, resulting in poor power of detection when I_{n_j} contains only a few coefficients. If one knows the cardinality c of I_{n_j}, then the LRT would more properly be based on the sum of squares of the c largest $\tilde{d}_i^j$'s. Because c is not known in practice, Ogden and Parzen (1996) suggested a recursive testing procedure for I_{n_j} containing one element each time. The LRT statistic would then be the largest $|\tilde{d}_i^j|^2$. It has been shown that the corresponding critical threshold at level α is equal to:

$$|\lambda_{\mathrm{LRT}}^i|^2 = \left\{\Phi^{-1}\left[\frac{(1-\alpha)^{1/n_j} + 1}{2}\right]\right\}^2 \tag{9}$$

Ogden and Parzen (1996) then proposed the recursive Algorithm 2 for choosing the threshold λ^j, which we extend to the 2D case for each level and orientation.

Algorithm 2 (*Wavelet shrinkage and recursive hypothesis testing*)

1: At each level and orientation, calculate $|\lambda_{\mathrm{LRT}}^{j}|^2$ using Eq. (9) and compare the largest $|\tilde{d}_i^j|^2$ with this critical value.
2: If the square value of $\tilde{d}_i^j$ is larger than threshold, this indicates that there is still significant signal among wavelet coefficients. Remove this $\tilde{d}_i^j$, set $n_j = n_j - 1$ and return to Step 1.
3: If $|\tilde{d}_i^j| < |\lambda_{\mathrm{LRT}}^{j}|$, then there is no strong evidence of signal in the remaining coefficients. Set the threshold λ_j of the current level and orientation to the largest remaining $|\tilde{d}_i^j|$.
4: Apply the soft thresholding scheme using λ_j so that small coefficients (indistinguishable from pure noise) are shrunk to zero and the significant coefficients included in the reconstruction are shrunk toward zero by the maximum absolute value of the small coefficients. This is accomplished in an adaptive data-driven way at each scale and orientation.
5: Apply the inverse DWT to obtain an estimate of the image g.

It is worth noting that Ogden's procedure is based on the square of the maximal statistic and could suffer from lack of power. Some alternatives that control the FWER could be used. We here cite the method of Hochberg (1979), also known as Bonferroni step-down (very similar to the Bonferroni but only a *little* less stringent), the Hochberg (1988) procedure or other related procedures (e.g., Hommel, 1988; Rom, 1990), which is the step-up analog of the Holm procedure. However, we stress the fact that our goal was not to use and compare all existing multiple testing procedures but rather to chose some that are representative of the wavelet/hypothesis testing literature.

A Bayesian approach

In contrast to these classical multiple hypothesis testing procedures, a Bayesian method for probabilistic wavelet shrinkage was considered by Vidakovic (1998). This method also involves testing the following hypothesis H_0: $d_{mn}^j = 0$ vs. H_1: $d_{mn}^j \neq 0$. However, the Bayesian framework here imposes a prior that describes the variability of the wavelet coefficients s_{mn}^j of the true image g.

In Bayesian testing of hypotheses, one usually specifies the prior probabilities $1 - p_j$ for H_0 being true and p_j for H_1 being true. This requires a prior distribution that has a point mass component; otherwise, the testing is impossible (Berger et al., 1996). Then, we will choose the prior mixture model according to Clyde et al. (1998) and Abramovich et al. (1998):

$$f_s(s) = p_j \tilde{f}(s) + (1 - p_j)\delta(s) \tag{10}$$

where p_j is the mixing proportion, $\delta(s)$ is a point mass at zero and $\tilde{f}(s)$ describes the behaviour of s_{mn}^j under H_1 (when s_{mn}^j is nonzero), which occurs with probability p_j. Considering $\tilde{f}(s)$ as a Gaussian pdf with zero mean and variance τ_j^2, Abramovich and Sapatinas (1999) proposed the ratio test (RT) statistic as the Bayes thresholding rule:

$$\eta_{mn}^j = \frac{P(H_1/d_{mn}^j)}{P(H_0/d_{mn}^j)} \underset{H_1}{\overset{H_0}{\lessgtr}} 1 \tag{11}$$

This quantity compares the posterior probabilities of the null hypothesis and its alternative (conditionally on the observation). If there is no evidence for presence of signal, i.e., only noise in the observed data, then the null hypothesis is more probable (conditionally) and the RT statistic will be correspondingly small (typically <1).

In contrast to previous authors, we will apply a threshold value α_{post} for the posterior probability of the null hypothesis conditional on the observation, namely $P(H_0/d_{mn}^j)$. This posterior conditional probability is closely connected to the p values in a conditional frequentist approach (Berger et al., 1996).

Using the Bayes rule, this probability can be expressed:

$$P(H_0/d_{mn}^j) = \frac{P(d_{mn}^j/H_0)P(H_0)}{f_d(d_{mn}^j)} = \frac{(1-p_j)\phi(d_{mn}^j;\sigma^2)}{f_d(d_{mn}^j)} \tag{12}$$

where $f_d(d_{mn}^j)$ is the marginal pdf of the observed wavelet coefficient:

$$f_d(d_{mn}^j) = p_j f_d/H_1(d_{mn}^j) + (1-p_j) f_d/H_0(d_{mn}^j)$$

$$= p_j \phi\left(d_{mn}^j;\tau_j^2+\sigma^2\right) + (1-p_j)\phi(d_{mn}^j;\sigma^2) \tag{13}$$

and $\phi(x;\sigma^2)$ is the centered Gaussian pdf with variance σ^2.

To test for the presence of signal, for arbitrary, prespecified posterior probability threshold value α_{post}, we then apply a hard thresholding rule to the estimated probability of the null hypothesis given the data:

$$\hat{s}_{mn}^j = d_{mn}^j 1\left(P(H_0/d_{mn}^j)\right) \leq \alpha_{\mathrm{post}} \tag{14}$$

where $1(x)$ is the indicator function. Arbitrary prespecification of $\alpha_{\mathrm{post}} \leq 0.05$ will provide some control over the smoothness of the reconstructed map.

Estimation of hyperparameters

To apply this Bayesian thresholding rule, the hyperparameters p_j, τ_j and σ must be appropriately estimated. Several solutions have been proposed in the literature. For example, one could use the robust estimate Eq. (6) for σ and an iterative expectation-minimisation (EM) algorithm to get maximum likelihood estimates (MLE) of p_j and τ_j (Clyde and George, 1999; Johnstone and Silverman, 1998). In our case, $\sigma = 1$ because normalised Gaussian maps are used, and only p_j and τ_j need to be estimated. This suggests Algorithm 3 for the Bayesian hypothesis testing approach.

Algorithm 3 (*Bayesian wavelet shrinkage*)

1: Calculate the DWT up to coarsest level J_c then for each orientation and level of the decomposition.
2: Use the EM algorithm to estimate the hyperparameters p_j and τ_j based on the intracranial wavelet coefficients (see Clyde and George, 1999; Johnstone and Silverman, 1998 for operational details).
3: Using these level-adapted hyperparameters, calculate the posterior probability of the null hypothesis conditional on each wavelet coefficient Eq. (12).
4: Apply the hard thresholding rule in Eq. (14) to the posterior probability of the null hypothesis.
5: Apply the inverse DWT to obtain an estimate of the image g.

Specification of prior distributions

It should be noted that other forms for $\tilde{f}$ in the mixture model of Eq. (10) have been considered. In their so-called BAMS method, Vidakovic and Ruggeri (2001) chose a standard exponential prior on the unknown σ^2, and obtained a double exponential pdf for $\tilde{f}$. Their results can be easily exploited to derive a closed-form expression for η_{mn}^{j}. For reasons of robustness, Vidakovic (1998) also suggested the use of central Student's t distributions (more heavily tailed than the normal) as a prior. However, no closed form expression is available for the corresponding Bayesian thresholding rule. In fact, many other prior distributions can be used provided that they are unimodal, centered and peaked at zero, and symmetric. Complicated prior pdfs can become useless in practice, although theoretically powerful, because closed-form expressions are not generally available for them, thus necessitating intensive numerical integration. It is also worth noting that the distribution of any Bayesian rule decision statistic quickly becomes a complicated function of d_{mn}^{j} as the complexity of the prior increases, so that its distribution under H_0 may not be theoretically tractable. In this case, one must resort to (computer-intensive) resampling techniques such as the bootstrap to estimate the distribution of the decision rule statistic.

ROC methods

The Receiver Operating Characteristic (ROC) is a well-known signal detection methodology for quantifying the detection accuracy of a test. A ROC curve is simply a plot of the sensitivity vs. the specificity of the test at different sizes of a probability threshold. The area under the ROC curve A_z is commonly regarded as a good single criterion for characterizing detection accuracy: the larger this area is (or the closer it is to unity), the better the detection accuracy of the test.

An ROC curve can be estimated from observed data using two main approaches. The first is fully parametric and has been well developed in the literature (Dorfman and Alf, 1969; Metz, 1986); the key assumption entailed is that the observed statistics are distributed according to a binormal mixture of distributions under the null and alternative hypotheses. Here, we will prefer a nonparametric method (Genovese et al., 1997), the main advantage of which is that it circumvents any assumptions about the distributional properties of the thresholded statistics.

To apply this method, we must begin by replicating the fMRI experiment K times and estimating the standardised GLM parameter vector β at each voxel in each replication. Then, we can use any hypothesis testing algorithm, applied over a range of sizes of test, to generate K thresholded maps in which each voxel has been labeled as either active or inactive. The probability that a voxel is labeled active $0 \leq k \leq K$ times is given by the following mixture model:

$$p(k) = \lambda \mathcal{B}_{(K,p_I)}(k) + (1-\lambda)\mathcal{B}_{(K,p_A)}(k) \tag{15}$$

where $\mathcal{B}_{(K,P)}$ is the binomial pdf with parameters K and P; p_I and p_A are, respectively, the probabilities that a truly inactive or active voxel is labeled active at the given threshold; that is, these probabilities are the observed FPF and TPF; and λ is the mixing proportion (here also the proportion of inactive voxels). Some assumptions underlying formulation of a binomial mixture model to estimate ROC curves based on thresholded statistic maps are summarised in Appendix A.

Experimental methods and preliminary analysis

fMRI data acquisition

(1) "Null" data sets: Five gradient echo echoplanar imaging (EPI) data sets were acquired from a single, healthy volunteer who was repeatedly scanned while lying quietly in the scanner with his eyes closed for 6 min. For each data set, 72 T_2*-weighted images were acquired at each of 26 contiguous oblique axial slices using a GE LX EchoSpeed system (General Electric, Milwaukee WI) operating at 1.5 T at CHU in Caen, France, with the following parameters: time to echo (TE) = 60 ms, time to repetition (TR) = 5 s, 64 × 64 voxel slices (N = 64 implying J = 6 scales in the 2D-DWT of these data), in-plane resolution 3.5 × 3.5 mm, slice thickness = 5 mm. In each time series, the first four volumes were eschewed to ensure magnetisation stabilisation.

(2) Event-related finger-movement datasets: We studied five male subjects during a discrete-trial or event-related (ER) experiment. The task was simply to oppose finger and thumb of the right hand repeatedly. The experiment was of 6-min duration and consisted of 11 trials, with a fixed interstimulus interval (ISI) of 30 s. Each trial was 5 s in duration. During this experiment, 26 slices of gradient echo echoplanar imaging data were acquired using the same scanner and acquisition parameters as for the null data acquisition. This simple experiment was designed to activate areas of the brain that are important in motor tasks.

Preliminary data analysis and simulation

Following correction for head movement, regression model parameters were estimated in each fMRI time series using wavelet-generalised least squares (Fadili and Bullmore, 2001). For the event-related experiment, the design matrix $\mathbf{X}$ comprised a vector simply coding the images acquired during finger movement. For the null data sets, three variants of $\mathbf{X}$ were specified, each comprising a periodic function of variable frequency [high (0.032 Hz), intermediate (0.016 Hz), or low (0.008 Hz)] and a unitary constant vector. Before parameter estimation, each column of all design matrices was convolved with a Poisson kernel (λ = 4s), to emulate the hemodynamic response function.

In evaluating the sensitivity of our methods, activation patterns with realistic spatial distribution were simulated in the null data sets. The spatial extent of the simulated clusters was defined by activation mapping of the event-related finger opposition task. A simulated periodic function (at high, intermediate or low frequency) was added to the "biological noise" of the null fMRI time series at voxel coordinates corresponding to locations of voxels activated by the finger-tapping experiment. The amplitude of simulated signal was expressed as a percentage of the mean value of the null time series.

Results

Type I error control

We assessed the relative performance of the three algorithms in terms of type I error control using the null fMRI data sets. For all three algorithms, the 2D-DWT of individual Z maps was implemented using the Daubechies wavelet with R = 4 vanishing moments and setting the coarsest level of the decomposition J_c = 2.

We tested the null hypothesis over a range of critical values corresponding to probabilities of type I error $0 < \alpha \leq 0.5$. For a valid hypothesis testing method, the number of positive tests generated by analysis of "null" data should be less than or equal to the number of positive tests expected under the null hypothesis, αV.

As shown in Fig. 1, all three algorithms demonstrated generally acceptable type I error control by this criterion. The FDR control algorithm was most conservative and the Bayesian algorithm was least conservative; indeed, the Bayesian method yielded slightly more than the predicted number of positive tests at the smallest size of posterior probability threshold.

Sensitivity

We assessed the relative performance of the three algorithms in terms of their power to detect simulated periodic signals of variable frequency and amplitude. In Fig. 2, the area under the ROC curve A_z is plotted as a function of signal amplitude. It can be seen that all three methods demonstrate improved sensitivity with increased signal amplitude but the Bayesian method is consistently most sensitive; the FDR controlling algorithm outperforms the recursive testing method for all but the smallest signal amplitudes.

Effects of DWT options

The 2D-DWT can be applied to spatial Z maps using wavelets of variable regularity (number of vanishing moments R) and the thresholding operations can be applied up to an arbitrary (coarsest) level of decomposition J_c. We systematically explored the impact of these options on both the specificity and sensitivity of probabilistic wavelet shrinkage.

(1) Effect of wavelet regularity: First we varied the regularity of the Daubechies wavelet, while maintaining the lowest level of the decomposition fixed at $J_c = 2$. As shown in Fig. 3, the false-positive fraction (FPF) observed by analysis of null data with $\alpha = 0.05$ tended to decrease as the regularity of the wavelet was increased in all three algorithms. In other words, increasing the number of vanishing moments tends to increase the conservativeness of all probabilistic wavelet shrinkage algorithms. There is also some evidence, in Fig. 4, for a monotonic decrease in sensitivity as R is increased, although this is a relatively minor trend compared to the regularity-related change in FPF. These findings agree with those reported by Desco et al. (2001).

An interpretation of this behaviour comes from the fact that more regular wavelets have better decorrelating properties. Hence, the assumption of independence for coefficients tested simultaneously in the wavelet domain is more valid. As far as the detection accuracy is concerned, it does not exhibit major change with increasing R, except for the FDR method applied to the low-frequency design matrix. Thus, as the wavelet becomes more regular, the method performs better in controlling the FPF but could suffer from a slight loss of detection power. However, a trade-off can be negotiated by choosing a Daubechies wavelet with 4 vanishing moments. This value keeps the FPF low without penalizing the detection power unduly (less than 2% loss).

(2) Effect of the coarsest level: In our notation, the resolution levels are numbered from 0 to $J - 1$, with 0 being the coarsest level (see "Wavelet shrinkage or nonparametric regression"). From Fig. 5, the observed FPF systematically decreases as the decomposition level increases for all the thresholding methods presented. This is

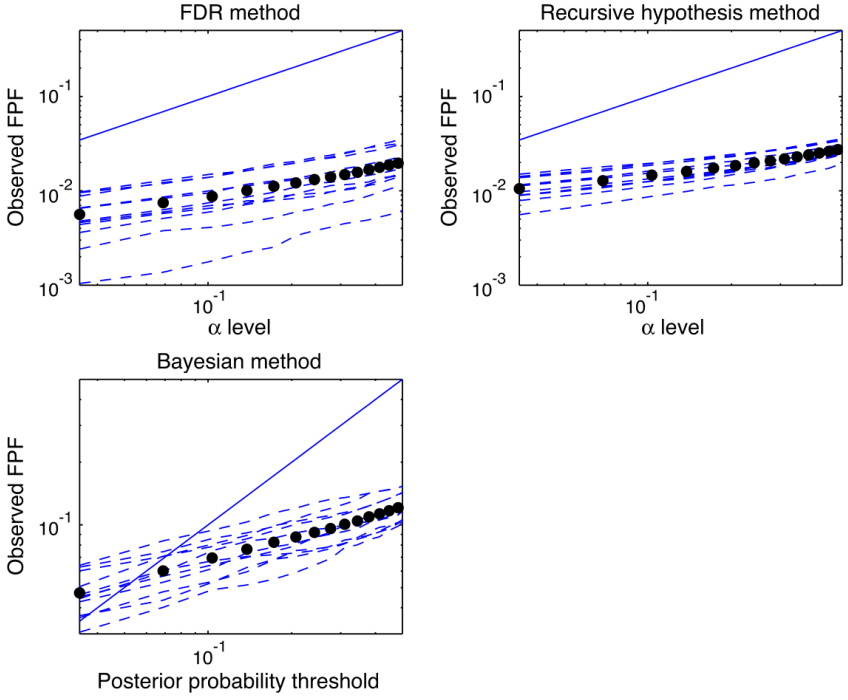

Fig. 1. Observed FPF (false-positive fraction) vs. prespecified risk α for the FDR and change-point detection (recursive hypothesis testing) approaches. For the Bayesian method, FPF is represented as a function of the posterior probability. Each dashed curve corresponds to results of thresholding a time series statistic map based on analysis of a single null fMRI data set with an arbitrary design matrix. The black filled circles correspond to the average curve. The solid line is the identity line. Number of vanishing moments $R = 4$ and coarsest level of decomposition $J_c = 2$.

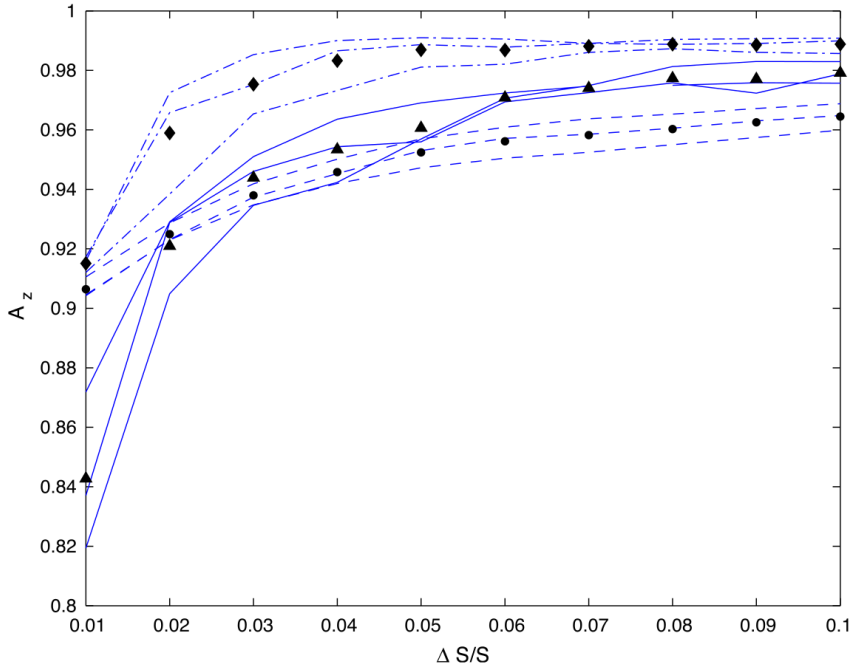

Fig. 2. Observed area under the ROC curve, A_z, vs. relative signal change, $\Delta S/S$. The FDR results are shown in solid lines; change-point detection (recursive hypothesis testing) is shown in dashed lines; and the Bayesian approach is shown in dashed-dotted lines. The mean curves, averaged over the three simulated frequencies, are represented by filled triangles, circles and diamonds, respectively. Number of vanishing moments $R = 4$ and coarsest scale of decomposition $J_c = 2$.

not surprising as the number of wavelet coefficients, and thus the number of hypotheses being tested simultaneously, increases as the decomposition level decreases. Meanwhile, as shown in Fig. 6, the area under the ROC curve shows a stable behaviour until level $J_c = 3$ and then decreases noticeably as the coarsest level increases further. Again, a trade-off between type I error control and power dissipation is suggested empirically by setting the coarsest level of decomposition $J_c = 3$. This choice keeps the FPF much less than the prespecified α level while ensuring good sensitivity. It is encouraging that this empirical choice of J_c is in agreement with

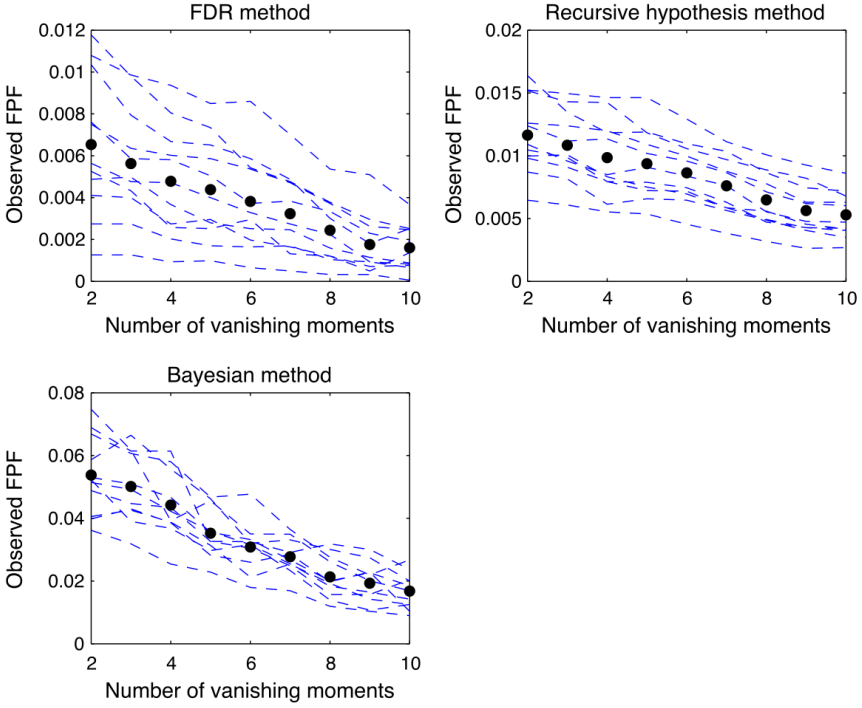

Fig. 3. Observed FPF (for a probability threshold of 0.05) as a function of the number of vanishing moments of the Daubechies wavelet. The coarsest decomposition level was 2. Each dashed line corresponds to a data set with an arbitrary design matrix (5 null data sets × 3 design matrices). The filled circles represent the average curve.

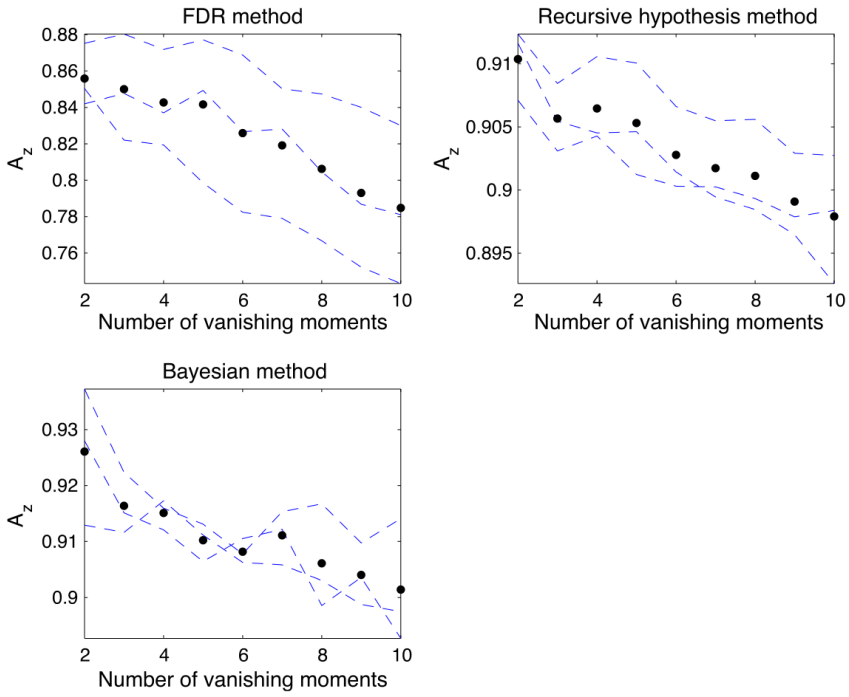

Fig. 4. Detection power A_z, measured by the area under the ROC curve, as a function of the number of vanishing moments of the Daubechies wavelet. The coarsest decomposition level was 2 and the relative signal change was 1%; probability threshold was set $\alpha = 0.05$.

the coarsest level asymptotically prescribed by the formula $J_c = \log_2 \log N + 1$ (Antoniadis et al., 2001).

Probabilistic wavelet shrinkage compared to Gaussian smoothing

These simulation results can also be interpreted by formal links to the theory of Gaussian smoothing, which has been central to inference using Statistical Parametric Mapping (SPM) software; see http://www.fil.ion.ucl.ac.uk/spm. Recently, VandeVille et al. (2003) have presented a first attempt to establish a link between SPM and the fractional splines wavelet transform, which involved investigating how well the scaling function (corresponding to the low-pass band) approaches a Gaussian. Here we generalise this approach to consideration of the compactly supported Daubechies wavelet with

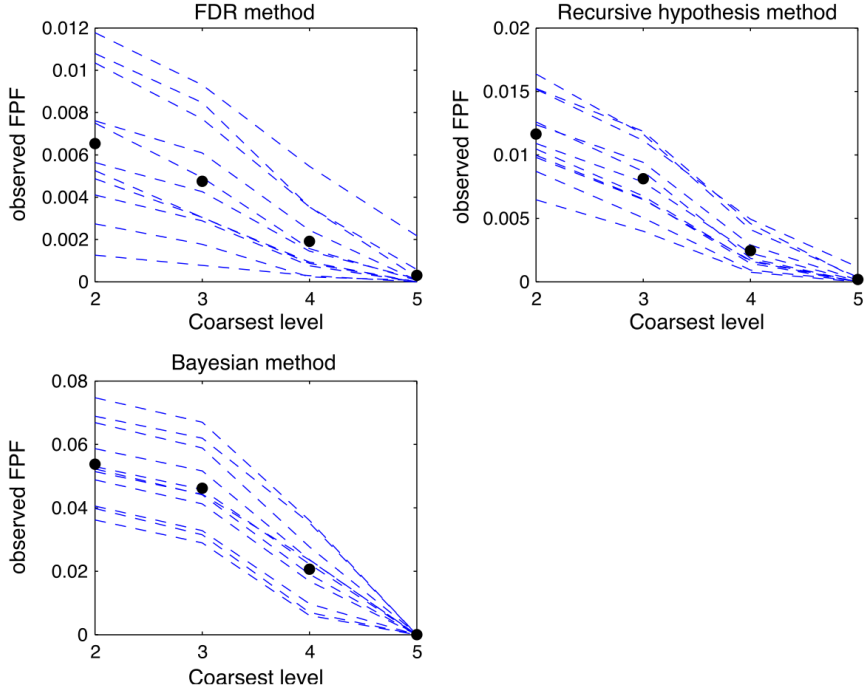

Fig. 5. Observed FPF (for a probability threshold of 0.05) as a function of coarsest thresholded scale. The Daubechies wavelet with 4 vanishing moments was used. Each dashed line corresponds to a data set with an arbitrary design matrix (5 null data sets × 3 design matrices).

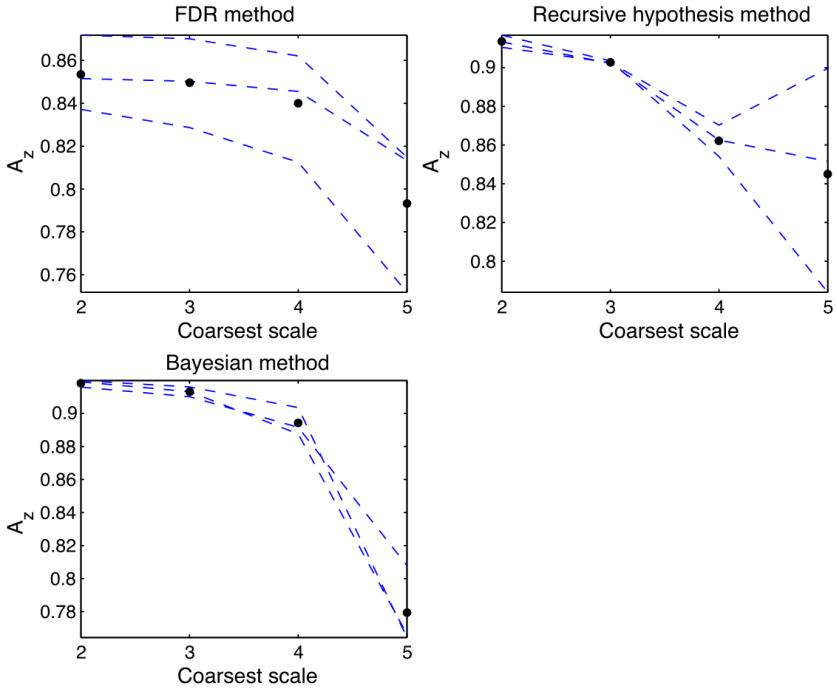

Fig. 6. Detection power A_z, measured by the area under the ROC curve, as a function of coarsest thresholded scale. The Daubechies wavelet with 4 vanishing moments was used and the relative signal change was 1%; probability threshold was set $\alpha = 0.05$.

R vanishing moments for which the Fourier transform of the corresponding low-pass QMF filter h can be written as:

$$H(\omega) = \text{const}\ e^{-i\frac{R\omega}{2}}\left(\cos\frac{\omega}{2}\right)^R \prod_{k=0}^{R-1}\left(1 - a_k e^{-i\omega}\right) \tag{16}$$

where a_k are roots inside the unit circle of a polynomial in $e^{i\omega}$ of degree $p-1$. This filter can be approximated as $\omega \to 0$ by:

$$H(\omega) \approx \text{const} \prod_{k=0}^{R-1}(1 - a_k)e^{-\frac{R\omega^2}{8}} \tag{17}$$

The Fourier transform of the scaling function up to a coarse scale $0 \le J_c < J$ can then be approximated by:

$$\varphi(\omega) = \prod_{p=0}^{J-1-J_c} H(2^p\omega)\alpha\ exp\left(-\frac{R\omega^2}{8}\frac{4^{J-J_c}-1}{3}\right) \tag{18}$$

By identifying this approximation to the Fourier transform of a Gaussian kernel, we can derive:

$$FWHM = 2\sigma\sqrt{2\ln 2} = \sqrt{2\ln 2}\sqrt{R}\sqrt{\frac{4^{J-J_c}-1}{3}} \tag{19}$$

where FWHM denotes the full width at half maximum of the Gaussian kernel. A closed form expression was derived for fractional splines in VandeVille et al. (2003). Actually, following these steps, similar expressions can be calculated for Battle–Lemarié, b-spline, symmlet or coiflet wavelets.

From Eq. (19), the FWHM depends on the coarsest decomposition level and the regularity of the wavelet. The equivalent FWHM increases very fast (exponentially) as we go coarser in scales $J_c \to 0$, and slowly (square root) as the number of vanishing moments increases. This means that for a given R and J_c setting, if the signal is reconstructed from *only* the low-pass subband, forcing the detail subbands to zero, this would be equivalent to applying a Gaussian smoothing with a FWHM as given by Eq. (19). Nonetheless, this approximation must be used very carefully as the approximation error, e.g., as measured by the L^2 norm of the residuals, also increases with J_c and R, as shown in Fig. 8. Consequently, the scaling function converges better to a Gaussian at coarser scales and smaller R. These findings provide a formal justification for previous results (Desco et al., 2001), which indicated that lower wavelet orders and resolution depths gave the optimal results in terms of sensitivity. However, these authors only used isotropic activation patterns for which Gaussian filters are optimal.

Fig. 7a shows the Fourier transform $\varphi(\omega)$ of the Daubechies scaling function and its Gaussian approximation for each number of vanishing moments. The coarsest scale was set $J_c = 3$. Fig. 7b depicts the same results when varying J_c while setting $R = 4$. In Fig. 8, we have plotted the L_2-norm residual error between the exact $\varphi(\omega)$ and its Gaussian approximation as a function of the coarsest decomposition scale and the number of vanishing moments. A Daubechies wavelet is compactly supported while a Gaussian kernel is not; this is one explanation for the lack of fit in Figs. 7a,b.

We can see that as we get coarser in scales ($J_c \searrow$), or as the wavelet gets more regular ($R \nearrow$), the equivalent FWHM increases. On the one hand, this "oversmoothing" naturally yields an increase in the observed FPF, which has been observed in Fig. 5. On the other hand, it is known that the sensitivity of Gaussian smoothing-based methods depends on the size of the Gaussian filter relative to the size of the signals to be detected (matched filter theorem). Our simulations included activation clusters whose isotropic equivalent size[1] ranged from 2 to 6 pixels with an average of 5 ($\sigma_{\text{size}} = 1.5$). The FWHMs calculated from Eq. (19) corresponding to $J_c \in [2,5]$ with $R = 4$ are, respectively, 18.5, 7.9, 3.3, 1.6 in pixel dimensions

[1] The isotropic equivalent size is the diameter of the disc with the same surface area as that of the simulated activation cluster.

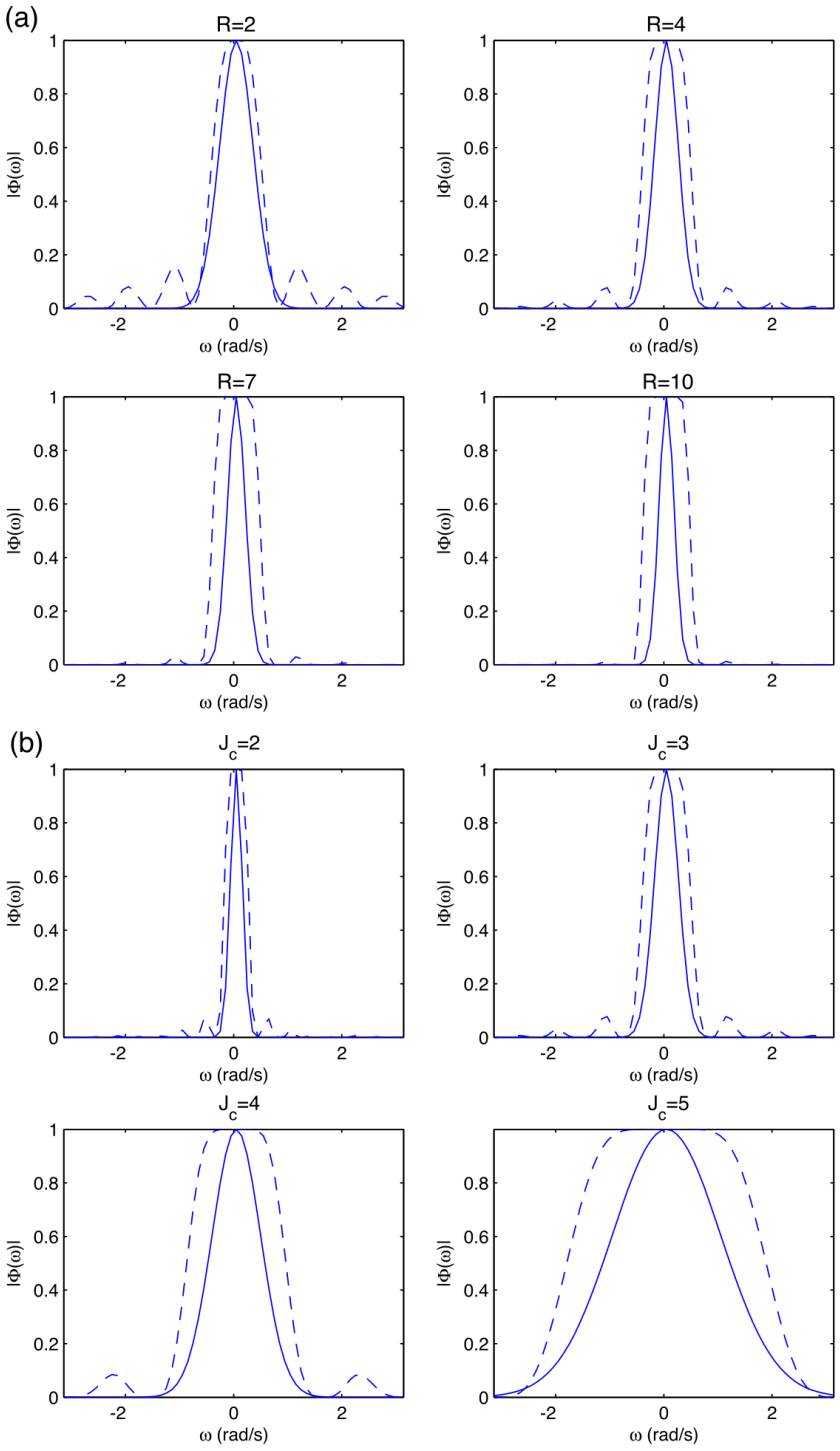

Fig. 7. (a) The Fourier transform φ (ω) of the Daubechies scaling function (solid line) and its Gaussian approximation (dashed line) for each number of vanishing moments. The coarsest scale J_c was set equal to 3. (b) φ (ω) and its Gaussian approximation when varying J_c while keeping $R = 4$. It can be seen that the Gaussian approximation to the Fourier transform of the scaling function is best when both J_c and R are smaller. The same point is illustrated by the approximation error surface in Fig. 8.

(the FWHMs were corrected to account for the error in the Gaussian approximation). Using a matched filter argument, one can then easily predict that the sensitivity will decrease as J_c increases, which is what we have observed in Fig. 6. Owing to its inherent multiscale nature, the wavelet transform with at least $J_c = 3$ is very efficient in detecting such activation clusters of different sizes. If we apply the same arguments when varying $R \in [2, 10]$ with $J_c = 3$, it turns out that the corrected equivalent FWHM increases slightly to around a value of 7.9 (in pixels). Again, this corroborates the loss of power as a function of the wavelet regularity observed in Fig. 4. Finally, if the corrected equivalent FWHM is calculated for each possible pair of parameters $(J_c, R) \in$

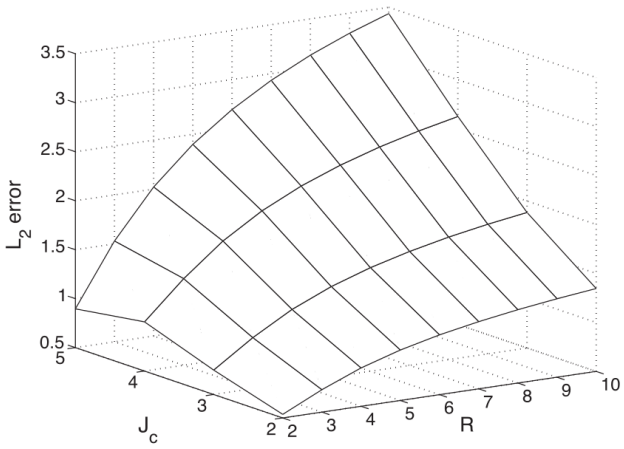

Fig. 8. L_2-error between the Daubechies scaling function and its Gaussian approximation in the Fourier domain. Error is smallest when both coarsest scale of decomposition J_c and regularity of wavelets R are small; and error increases rapidly as J_c increases.

[2, 5] × [2, 10], (J_c = 3, R = 4) gives a FWHM of 7.9, which corresponds to the size of most clusters in our simulations. Furthermore, for (J_c = 3, R = 4), the Gaussian approximation has proved good.

Functional MRI: activation mapping

All three wavelet-based methods gave broadly similar results for analysis of the event-related fMRI data. As shown in Fig. 9, areas of activation were located mainly in contralateral motor and somatosensory cortex and ipsilateral cerebellum ($\alpha \leq 0.01$ for all maps). Overall, the Bayesian method appears to provide a somewhat richer or more sensitive characterisation of the cerebral response, echoing the superior sensitivity of this method demonstrated by ROC curve analysis of simulated data. The activation maps generated by the FDR method are relatively conservative, which is also consistent with the ROC analysis results. The performance of the recursive hypothesis testing method is intermediate.

For comparative purposes, and to highlight the multiresolution adaptive nature of the wavelet methods, we also show in Fig. 10 activation maps for the same experimental data set analysed with SPM after three different Gaussian kernels have been applied (with FWHM = 6, 10 and 18 mm). A height threshold corrected for family-wise error $\alpha = 0.01$ was used for hypothesis testing. The strongest signal (in contralateral somatosensorimotor cortex) is significantly activated in all three monoresolution maps but the impact of arbitrary kernel size is demonstrated by the absence of cerebellar signal after smoothing with the largest kernel.

Discussion and conclusion

In this paper, we proposed a fully wavelet-based hypothesis testing framework for activation mapping based on functional magnetic resonance images of the human brain. Algorithms from the statistical theory of wavelets were adapted to the case of analysing 2D spatial maps of linear model parameters estimated by analysis of the fMRI time series observed at each voxel. Two classical methods and one Bayesian method for hypothesis testing were presented. Our results are promising and offer a naturally multiscale alternative to single scale Gaussian filtering as widely used in the neuroimaging community before hypothesis testing.

We have shown by ROC curve analysis of simulated data that a Bayesian thresholding algorithm (incorporating one of several possible priors for the sparse distribution of wavelet coefficients under the alternative hypothesis) has greater sensitivity than classical methods controlling the false discovery rate or recursively testing the coefficients for significance at each level of the wavelet decomposition. This result is consistent with prior work indicating greater sensitivity of wavelet thresholding by a Bayesian approach, and also with the results of applying all three methods to an experimental event-related data set. Our results are also broadly consistent with recent work indicating superior sensitivity of Bayesian compared to frequentist methods for multiple hypothesis testing of fMRI statistic maps in the spatial domain (Marchini and Presanis, in press).

We have systematically investigated the performance of our methods as a function of two critical factors—the regularity of the wavelet R and the coarsest level of the wavelet decomposition J_c. We have shown empirically that increasing J_c, i.e., decreasing the depth of decomposition, tends to improve the specificity of the methods at the expense of some sensitivity. Increasing the regularity of the wavelet was associated with the same effects on type I and type II error. A reasonable trade-off between these effects was found to exist with J_c = 3, as predicted asymptotically, and R = 4. However, we suggest that a more compelling argument in favor of this specification can be made by relating the low-pass filtering properties of the wavelet transform to the FWHM of an approximately equivalent Gaussian kernel. In this way, we showed that our empirically preferred choice of wavelet decomposition was approximately equivalent to smoothing by a Gaussian kernel with FWHM = 7 voxels, which was close to the mean size of activated clusters of voxels in our simulated data, predicting superior sensitivity of this algorithm by the matched filter theorem.

The comparison with monoresolution methods of smoothing, followed by hypothesis testing of regression coefficients in the spatial domain, was also pursued empirically. Analysis of experimental data following application of one of three different Gaussian kernels demonstrated again the familiar observation that results of activation mapping by this approach are conditional on the choice of kernel. A kernel much larger than the spatial extent of cerebellar signal obliterated evidence for significant activation in that region. All three multiresolution wavelet-based methods circumvent the need to specify a priori the size of signals expected, and therefore the optimal choice of smoothing kernel. Empirically, the wavelet-based methods seemed to provide a richer characterisation of distributed brain activation in the experimental data set.

However, several limitations are still to be addressed by further developments of this methodology. For example, due to critical sampling, the orthogonal bases used in the DWT are not translation invariant. Thus, “ringing” or Gibbs-like oscillation may occur at the boundaries of isolated activation clusters. To cope with this, a possible solution is to use overcomplete transforms, e.g., the undecimated DWT. Invariance to rotation could also be an important issue as the activation patterns in

fMRI will generally *not* be isotropic. However, decimated DWT-based analysis is sensitive to the orientation of the objects in the image under consideration, i.e., the wavelet coefficients of the original image and its rotated version are not the same ones rotated by the same angle, except for rotation angles that are multiple of (π/4) (horizontal, vertical and diagonal orientations). To address this issue may necessitate the use of transforms that adapt the orientation of their atoms to the geometry, such as curvelets or bandelets. Such new transforms (highly redundant) are the area of a very active current research in the image processing and harmonic analysis communities. However, the atoms involved in this type of analysis are frames but not bases, yielding a high redundancy and correlated coefficients under the null hypothesis. One must then adapt our methodologies to the case of non-independent tests.

Another important issue is anisotropic spatial resolution of fMRI data. Due to the separable nature of the DWT, only isotropic data sets can be processed efficiently. However, data acquired with fMRI are usually isotropic in-plane but the slice thickness is generally coarser. It is mainly for this reason that we did not use a 3D separable DWT transform immediately. To deal with anisotropic structures, one possible

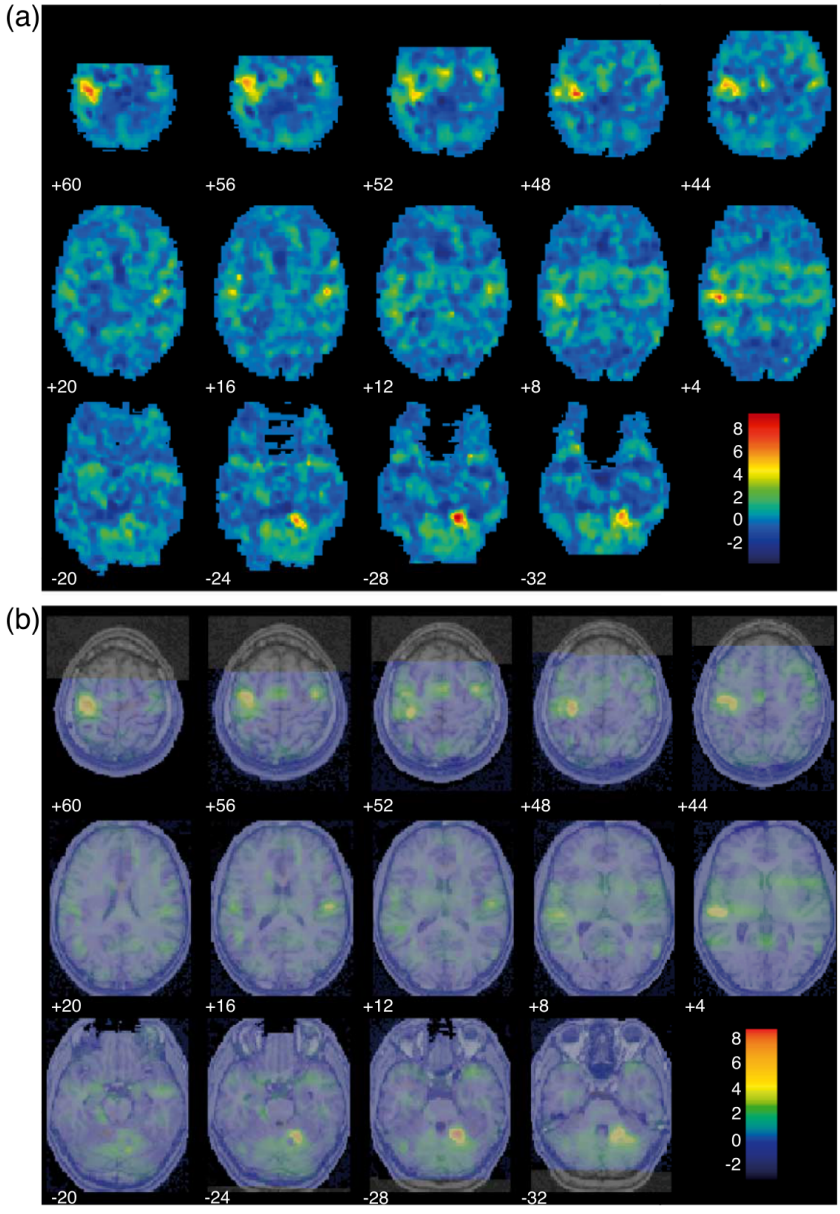

Fig. 9. Functional MRI data acquired during event-related right-handed finger-tapping experiment. (a) Example of an original statistical map representing linear model parameters estimated by wavelet-generalized least squares at each voxel of the 2D image. Estimated activation maps with $\alpha = 0.01$ using: (b) the FDR method, (c) the recursive hypothesis testing method, and (d) the Bayesian thresholding estimator. The right side of each map represents the right side of the brain; the z-coordinate of each slice in mm above or below the intercommissural plane in the space of Talairach and Tournoux is indicated by the number in bottom left corner of each panel. The activation task was repeated right index-thumb opposition that is expected to activate contralateral regions of motor and somatosensory cortex, supplementary motor area, and ipsilateral cerebellum. This pattern of activation is most clearly seen in the map obtained by the Bayesian thresholding operation. Number of vanishing moments $R = 4$ and coarsest scale of decomposition $J_c = 2$.

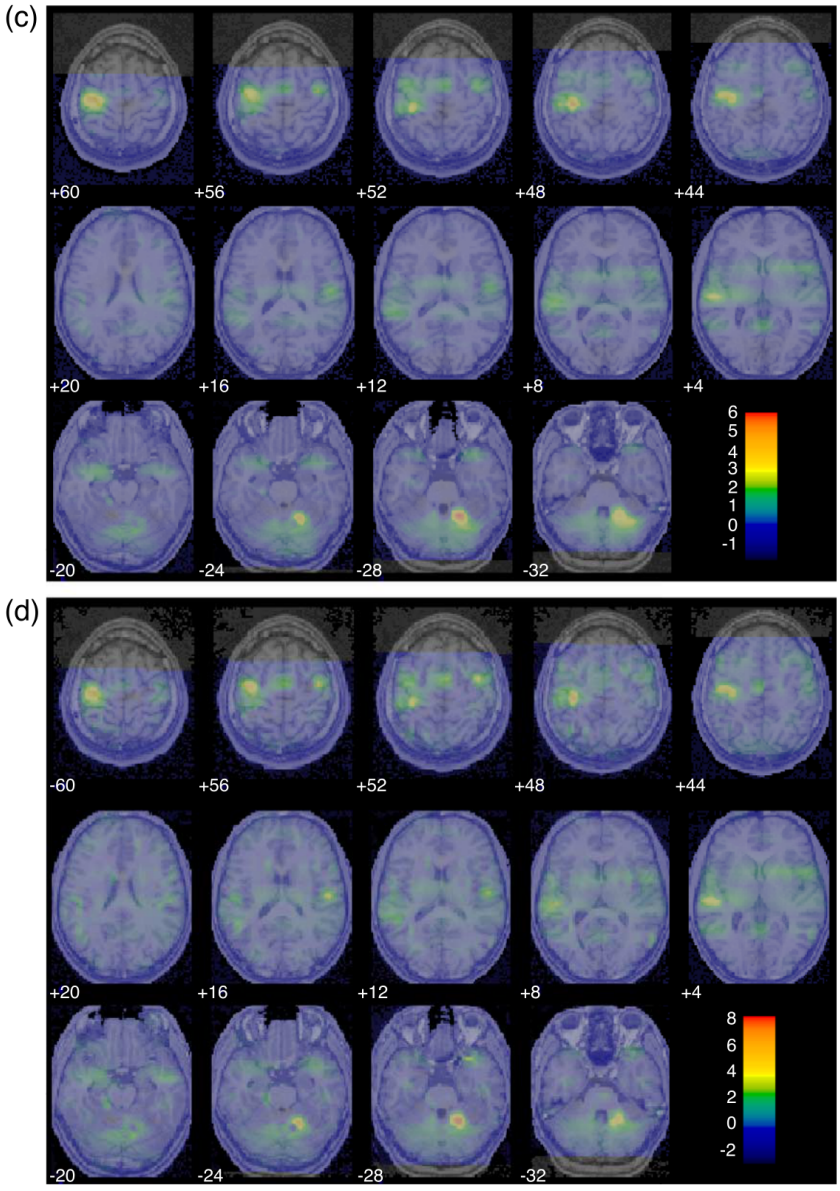

Fig. 9 (*continued*).

solution would consist in using different decomposition scales for X, Y and Z directions to compensate for the anisotropy. This can be clearly seen from Eq. (19). An alternative based on the $2D + Z$ quincunx wavelet transform has been proposed (VandeVille et al., 2003). Furthermore, because of the separable nature of the DWT, only isotropic activation patterns can be detected efficiently. A possible solution to this problem would consist in using one of the geometrical X-let transforms (X stands for any of these transforms), which adapt their shape to the pattern to be estimated. All these issues are the focus of ongoing work in our group.

Appendix A. ML estimation of the binomial mixture model

While the following algorithm can be generalised easily to any number of binomial mixture components, we only derive here the expressions for the two components case. Let $\Theta = (p_A, p_I, \lambda)$ denote the parameter vector defining the binomial mixture model. The ML estimate of Θ based on a set of N independent observations k_i is:

$$\hat{\Theta}_{\mathrm{ML}} = \arg\max_{\Theta} \sum_{i=1}^{N} \log\left(\lambda \mathcal{B}_{(K,p_I)}(k_i) + (1-\lambda)\mathcal{B}_{(K,p_A)}(k_i)\right) \quad (20)$$

This equation has no closed-form simple solution but it can be efficiently solved using the EM algorithm (McLachlan and Krishnan, 1996).

A.1. E-step

Compute the expected value of the complete likelihood, conditioned on the observed data and the current hyperparameter vector

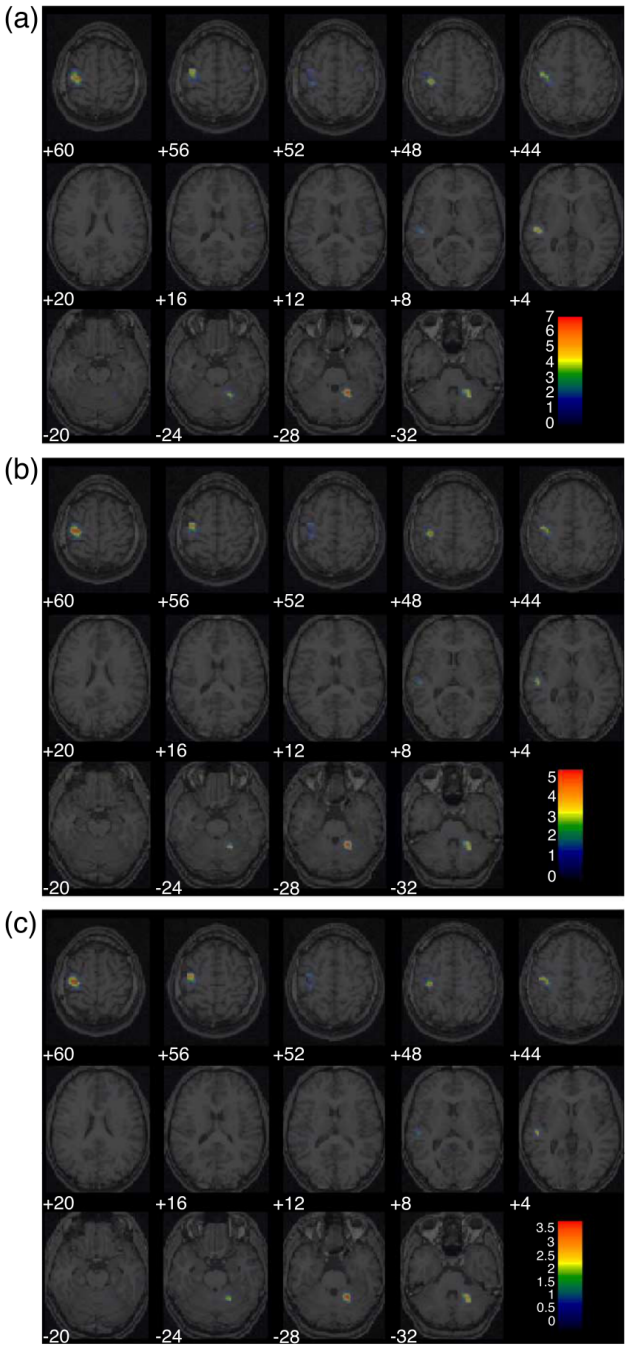

Fig. 10. Functional MRI data acquired during event-related right-handed finger-tapping experiment. Activation maps produced by SPM after different Gaussian kernels have been applied: (a) Gaussian kernel FWHM = 6 mm; (b) FWHM = 10 mm ; and (c) FWHM = 18 mm. A corrected critical height threshold at the significance level $\alpha = 0.01$ was applied. The right side of each map represents the right side of the brain; the z-coordinate of each slice in mm above or below the intercommissural plane in the space of Talairach and Tournoux is indicated by the number in bottom left corner of each panel. Note that significance of the ipsilateral cerebellar signal is conditional on kernel size and that plausible signals in medial premotor cortex and ipsilateral somatosensorimotor cortex recovered by multiresolution analysis/multiple hypothesis testing in the wavelet domain are not evident in these results of monoresolution smoothing followed by multiple hypothesis testing in the spatial domain.

estimate $\Theta^{(l)}$. For this mixture model, this reduces to calculating the posterior probabilities $P(I/k_i)$ and $P(A/k_i)$, i.e., the probability of having an inactive voxel or an active voxel, conditioned on the observed data and the current hyperparameter vector estimate $\Theta^{(l)}$:

$$P^{(l)}(I/k_i) = \frac{\lambda^{(l)}\mathcal{B}_{\left(K,p_1^{(l)}\right)}(k_i)}{\lambda^{(l)}\mathcal{B}_{\left(K,p_1^{(l)}\right)}(k_i) + \left(1-\lambda^{(l)}\right)\mathcal{B}_{\left(K,p_A^{(l)}\right)}(k_i)} \tag{21}$$

$$P^{(l)}(A/k_i) = 1 - P^{(l)}(I/k_i) \tag{22}$$

A.2. M-step

Update the hyperparameter vector estimate according to:

$$\lambda^{l+1} = \frac{1}{N}\sum_{i=1}^{N} P^{(l)}(I/k_i) \tag{23}$$

$$p_{\cdot}^{(l+1)} = \frac{\sum_{i=1}^{N} k_i P^{(l)}(./k_i)}{K\sum_{i=1}^{N} P^{(l)}(./k_i)} \tag{24}$$

Note that under the null hypothesis, i.e., in the absence of true activation (null fMRI data), only p_I can be estimated. The ML estimate in Eq. (24) then simply reduces to calculating the mean of the counts k_i, which coincides with calculating the observed FPF for each thresholded statistical map (replication) and then averaging these FPFs over replications.

A critical point of the EM algorithm is initialisation. Here we give some arguments to support our intuitive choice of the initialising set of hyperparameters. For instance, in our simulations, λ is known a priori as we know exactly the truly inactive and active voxels. Otherwise, one knows that, in the case of fMRI data, most voxels are inactive and only a few voxels are expected to be active. One should then expect a high value of λ, say 90%. As far as the probability p_I is concerned, it is should be of the same order as the FPF. One can then use as an initialisation a very conservative value obtained from the Bonferroni correction α/nd. For the parameter p_A, it is difficult to devise any automatic initialisation procedure as we do not have any prior knowledge about the sensitivity of the statistical estimation method in use. One can then use an arbitrary starting value of 0.5.

Another important issue is how to calculate the standard errors of the mixture hyperparameters (these provide some sense of how confident we should be about the parameters). One could use computer-intensive resampling methods such as the bootstrap. Alternatively, under regularity conditions, the standard errors can be estimated from the inverse of the expected information matrix. However, in the binomial mixture model, this involves nontrivial calculations. We here propose another way to proceed by approximating the information matrix within the EM framework (Basford et al., 1997). For independent observations, this approximation is given in terms of the gradient of the log-likelihood function:

$$I(\hat{\Theta}) \approx \sum_{i=1}^{N} \nabla LL_i(\hat{\Theta})\,\nabla^t LL_i(\hat{\Theta}) \tag{25}$$

where $\nabla LL_i\ (\hat{\Theta})$ is the gradient of the log-likelihood function based on the single observation k_i: $\nabla^t LL_i\ (\hat{\Theta}) = (\partial LL_i/\partial\lambda,\ \partial LL_i/\partial p_i,\ \partial LL_i,/\partial p_A)$

$$\frac{\partial LL_i}{\partial \lambda} = \frac{(1-p_1)^{K-k_i}p_1^{k_i} - (1-p_A)^{K-k_i}p_A^{k_i}}{\lambda(1-p_1)^{K-k_i}p_1^{k_i} + (1-\lambda)(1-p_A)^{K-k_i}p_A^{k_i}} \tag{26}$$

$$\frac{\partial LL_i}{\partial p_1} = \frac{-\lambda(1-p_1)^{K-1-k_i}p_1^{k_i-1}(Kp_1-k_i)((1-p_1)^{K-k_i}p_1^{k_i} - (1-p_A)^{K-k_i}p_A^{k_i})}{\left(\lambda(1-p_1)^{K-k_i}p_1^{k_i} + (1-\lambda)(1-p_A)^{K-k_i}p_A^{k_i}\right)^2} \tag{27}$$

$$\frac{\partial LL_i}{\partial_{pA}} = -\frac{(1-\lambda)(1-p_A)^{K-1-k_i}p_A^{k_i-1}(K_{pA}-k_i)\left((1-p_1)^{K-k_i}p_1^{k_i} - (1-p_A)^{K-k_i}p_1^{k_i}\right)}{\left(\lambda(1-p_1)^{K-k_i}p_1^{k_i} + (1-\lambda)(1-p_A)^{K-k_i}p_A^{k_i}\right)^2} \tag{28}$$

In our simulations, as λ is known and fixed, Eq. (26) can be ignored and the known λ can be substituted in Eqs. (27) and (28).

A.3. Assumptions of binomial model for ROC curve estimation

- Each voxel can only have two states (active or inactive).
- The mixing proportion (amount of inactive voxels) is supposed constant across replications and thresholds. In our experiments, we have not constrained this proportion to be constant across thresholds. This has been done for the sake of simplicity in the model and its solution by the EM algorithm. Nevertheless, a multinomial mixture including this supplementary constraint can be formulated and its EM solution derived. The convergence of the EM algorithm in this case is much slower (the number of parameters to be estimated jointly is $2 \times M - 1$ for M threshold levels). In addition, from our experiments, we have observed that the estimated mixing proportion is stable when varying the rating scale threshold.
- The replications (experiments) are supposed independent. Moreover, the number of replications K must be ≥ 3 for identifiability purposes (three parameters to estimate).
- When two or many estimation methods will be compared, their respective statistical maps will be assumed independent. This is the so-called independence model discussed by Genovese et al. (1997).

References

Abramovich, F., Benjamini, Y., 1995. Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In: Antoniadis, A., Oppenheim, G. (Eds.), Wavelets and Statistics. Springer-Verlag, New York, pp. 5–14.

Abramovich, F., Benjamini, Y., 1996. Adaptive thresholding of wavelet coefficients. Comput. Stat. Data Anal. 22, 351–361.

Abramovich, F., Sapatinas, T., 1999. Bayesian approach to wavelet decomposition and shrinkage. In: Muller, P., Vidakovic, B. (Eds.), Bayesian Inference in Wavelet Based Models. Springer-Verlag, New York, pp. 33–50.

Abramovich, F., Sapatinas, T., Silverman, B., 1998. Wavelet thresholding via a Bayesian approach. J. R. Stat. Soc., B 60, 725–749.

Abramovich, F., Benjamini, Y., Donoho, D., Johnstone, I., 2000. Adapting to unknown sparsity by controlling the false discovery rate (Tech Rep.). Department of Statistics, Stanford University.

Achim, A., Bezerianos, A., Tsakalides, P., 2001. Novel Bayesian multiscale method for speckle removal in medical ultrasound images. IEEE Trans. Med. Imag. 20, 772–783.

Antoniadis, A., Bigot, J., Sapatinas, T., 2001. Wavelet estimators in nonparametric regression: a comparative simulation study. J. Stat. Software 6 (6), 1–83.

Basford, K., Greenway, D., McLachlan, G., Peel, D., 1997. Standard errors of fitted means under normal mixture models. Comput. Stat. 12, 1–17.

Berger, J., Boukai, B., Wang, Y., 1996. Unified frequentist and Bayesian testing of a precise hypothesis. Stat. Sci. 12, 133–160.

Brammer, M., 1998. Multidimensional wavelet analysis of functional magnetic resonance images. Hum. Brain Mapp. 6, 378–382.

Bullmore, E., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M., 1999. Global, voxel and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans. Med. Imag. 18, 32–42.

Bullmore, E., Long, C., Suckling, J., Fadili, M., Calvert, G., Zelaya, F., Carpenter, A., Brammer, M., 2001. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. Hum. Brain Mapp. 12 (2), 61–78.

Bullmore, E., Fadili, M., Breakspear, M., Salvador, R., Suckling, J., Brammer, M., 2003. Wavelets and statistical analysis of functional magnetic resonance images of the human brain. Stat. Methods Med. Res. 12 (5), 375–399.

Chang, S., Yu, B., Vetterli, M., 2000. Adaptive wavelet thresholding for image denoising and compression. IEEE Trans. Image Process. 9 (9), 1522–1531.

Chipman, H., Kolaczyk, E., McCulloch, R., 1997. Adaptive Bayesian wavelet shrinkage. J. Am. Stat. Assoc. 92, 1413–1421.

Clyde, M.A., George, E.I., 1999. Empirical Bayes estimation in wavelet nonparametric regression. In: Muller, P., Vidakovic, B. (Eds.), Bayesian Inference in Wavelet Based Models. Springer-Verlag, New York, pp. 309–322.

Clyde, M., George, E., 2000. Flexible empirical Bayes estimation for wavelets. J. R. Stat. Soc., B 62, 681–698.

Clyde, M., Parmigiani, G., Vidakovic, B., 1998. Multiple shrinkage and subset selection in wavelets. Biometrika 85 (2), 391–401.

Crouse, M., Nowak, R., Baraniuk, R., 1998. Wavelet-based statistical signal processing using hidden Markov models. IEEE Trans. Signal Process. 46 (4), 886–902.

Desco, M., Hernandez, J., Santos, A., Brammer, M., 2001. Multiresolution analysis in fMRI: sensitivity and specificity in the detection of brain activation. Hum. Brain Mapp. 14, 16–27.

Donoho, D.L., Johnstone, I.M., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81 (3), 425–455.

Donoho, D.L., Johnstone, I.M., 1995. Adapting to unknown smoothness via wavelet shrinkage. J. Am. Stat. Assoc. 90 (432), 1200–1224.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D., 1995. Wavelet shrinkage: asymptopia? J. R. Stat. Soc., B 57 (2), 301–337.

Dorfman, A., Alf, E., 1969. Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating method data. J. Math. Psychol. 6, 487–496.

Fadili, M., Bullmore, E., 2001. Wavelet-generalised least squares: a new BLU estimator of linear regression models with 1/f errors. NeuroImage 15, 217–232.

Feilner, M., Blu, T., Unser, M., 1999. Statistical analysis of fMRI data using orthogonal filterbanks. Proceedings of the SPIE Conference on Mathematical Imaging: Wavelet Applications in Signal and Image Processing VII, vol. 3813. SPIE, Denver CO, USA, pp. 551–560.

Feilner, M., Blu, T., Unser, M., 2000. Analysis of fMRI data using spline wavelets. Proceedings of the Tenth European Signal Processing Conference (EUSIPCO'00), vol. IV. SPIE, Tampere, Finland, pp. 2013–2016.

Friston, K., Frith, P., Frackowiak, R., 1991. Comparing functional (PET) images: the assessment of significant change. J. Cereb. Blood Flow Metab. 11, 690–699.
Genovese, C.R., Noll, D.C., Eddy, W.F., 1997. Estimating test–retest reliability in functional MR imaging I: Statistical methodology. Magn. Res. Med. 38 (3), 497–507.
Hayasaka, S., Nichols, T., 2003. Validating cluster size inference: random field and permutation methods. NeuroImage 20 (4), 2343–2356.
Hochberg, Y., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6, 54–70.
Hochberg, Y., 1988. A sharper Bonferoni procedure for multiple tests of significance. Biometrika 75, 800–803.
Hommel, G., 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75, 383–386.
Huang, S., Lu, H., 2000. Bayesian wavelet shrinkage for nonparametric mixed effects models. Stat. Sin. 10, 1021–1040.
Johnstone, I., Silverman, B., 1998. Empirical Bayes approaches to mixture problems and wavelet regression (Tech. Rep.). Department of Mathematics, University of Bristol, UK.
Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. PAMI 11 (7), 674–693.
Mallat, S.G., 1999. A Wavelet Tour of Signal Processing. (second ed.) Academic Press, New York.
Marchini, J., Presanis, A., 2004. Comparing methods of analyzing fMRI statistical parametric maps. NeuroImage (in press).
McLachlan, G.J., Krishnan, T., 1996. The EM Algorithm and Extensions. Wiley, New York.
Metz, C.E., 1986. ROC methodology in radiological imaging. Invest. Radiol. 21, 722–733.
Nichols, T., Holmes, A., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.
Ogden, R.T., Parzen, E., 1996. Change-point approach to data analytic wavelet thresholding. Stat. Comput. 6 (2), 93–99.
Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge Press, Cambridge.
Poline, J.-B., Mazoyer, B., 1994. Analysis of individual brain activation maps using hierarchical description and multiscale detection. IEEE Trans. Med. Imag. 4, 702–710.
Raz, J., Turetsky, B., 1999. Wavelet ANOVA and fMRI. Proceedings of the SPIE Conference on Mathematical Imaging: Wavelet Applications in Signal and Image Processing VII, vol. 3813. SPIE, Denver CO, USA.
Rom, D., 1990. A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika 77, 663–665.
Ruttimann, U., Unser, M., Rio, D., 1994. Statistical analysis of image differences by wavelet decomposition. Proceedings of the Sixteenth Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Engineering Advances: New Opportunities for Biomedical Engineers (EMBS'94) vol. I. Baltimore MD, USA, pp. A28–A29.
Ruttimann, U., Ramsey, N., Hommer, D., Thévenaz, P., Lee, C., Unser, M., 1995. Analysis of functional magnetic resonance images by wavelet decomposition. Proceedings of the 1995 IEEE International Conference on Image Processing (ICIP'95) vol. I. Washington DC, USA, pp. 633–636.
Ruttimann, U., Unser, M., Thévenaz, P., Lee, C., Rio, D., Hommer, D., 1996. Statistical analysis of image differences by wavelet decomposition. In: Aldroubi, A., Unser, M. (Eds.), Wavelets in Medicine and Biology. CRC Press, Boca Raton FL, USA, pp. 115–144.
Ruttimann, U., Unser, M., Rawlings, R., Rio, D., Ramsey, N., Mattay, V., Hommer, D., Frank, J., Weinberger, D., 1998. Statistical analysis of functional MRI data in the wavelet domain. IEEE Trans. Med. Imag. 17 (2), 142–154.
Shen, X., Huang, H., Cressie, N., 2002. Nonparametric hypothesis testing for a spatial signal. J. Am. Stat. Assoc. 97, 1122–1140.
Simoncelli, E.P., Adelson, E.H., 1996. Noise removal via Bayesian wavelet coring. Third Int'l. Conf on Image Proc., vol. 1. IEEE Sig. Proc. Society, Laussanne, pp. 379–382.
Turkheimer, F., Brett, M., Visvikis, D., Cunningham, V., 1999. Multiresolution analysis of emission tomography images in the wavelet domain. J. Cereb Blood Flow Metab. 19, 1189–1208.
VandeVille, D., Blu, T., Unser, M., 2003. Wavelets versus resels in the context of fMRI: establishing the link with SPM. Proceedings of the SPIE Conference on Mathematical Imaging: Wavelet Applications in Signal and Image Processing X. SPIE, San Diego.
Vannucci, M., Corradi, F., 1999. Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. J. R. Stat. Soc., B 61, 971–986.
Vidakovic, B., 1998. Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. J. Am. Stat. Assoc. 93 (441), 173–179.
Vidakovic, B., 1999. Statistical Modeling by Wavelets. Wiley, New York.
Vidakovic, B., Ruggeri, F., 2001. BAMS Method: Theory and Simulations. Indian J. Stat. 63, 234–249.
Worsley, K., Poline, J.-B., Vandal, A., Friston, K., 1995. Tests for distributed, non-focal brain activations. NeuroImage 2, 183–194.
Worsley, K., Marrett, S., Neelin, P., Evans, A., 1996. Searching scale space for activation in PET images. Hum. Brain Mapp. 4, 74–90.
Worsley, K., Marrett, S., Neelin, P., Vandal, A., Friston, K., Evans, A., 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. 4, 58–73.