Cogn Affect Behav Neurosci (2013) 13:703–713
DOI 10.3758/s13415-013-0185-3

Alternative-based thresholding with application to presurgical fMRI

Joke Durnez · Beatrijs Moerkerke · Andreas Bartsch · Thomas E. Nichols

Published online: 19 July 2013

Abstract Functional magnetic reasonance imaging (fMRI) plays an important role in pre-surgical planning for patients with resectable brain lesions such as tumors. With appropriately designed tasks, the results of fMRI studies can guide resection, thereby preserving vital brain tissue. The mass univariate approach to fMRI data analysis consists of performing a statistical test in each voxel, which is used to classify voxels as either active or inactive—that is, related, or not, to the task of interest. In cognitive neuroscience, the focus is on controlling the rate of false positives while accounting for the severe multiple testing problem of searching the brain for activations. However, stringent control of false positives is accompanied by a risk of false negatives, which can be detrimental, particularly in clinical settings where false negatives may lead to surgical resection of vital brain tissue. Consequently, for clinical applications, we argue for a testing procedure with a stronger focus on preventing false negatives. We present a thresholding procedure that incorporates information on false positives and false negatives. We combine two measures of significance for each voxel: a classical *p*-value, which reflects evidence against the null hypothesis of no activation, and an alternative *p*-value, which reflects evidence against activation of a prespecified size. This results in a layered statistical map for the brain. One layer marks voxels exhibiting strong evidence against the traditional null hypothesis, while a second layer marks voxels where activation cannot be confidently excluded. The third layer marks voxels where the presence of activation can be rejected.

Keywords fMRI · Power · False negative errors · Multiple testing · Pre-surgical fMRI

J. Durnez (✉) · B. Moerkerke
Department of Data Analysis, Ghent University, H. Dunantlaan 1, 9000 Ghent, Belgium
e-mail: Joke.Durnez@UGent.be

A. Bartsch
FMRIB Centre, Oxford University, Oxford, United Kingdom

A. Bartsch
Department of Neuroradiology, University of Heidelberg, Heidelberg, Germany

T. E. Nichols
Department of Statistics & Warwick Manufacturing Group, University of Warwick, Coventry, UK

Introduction

A common treatment for patients suffering from a brain tumor is surgical resection of the tumor. In order to minimize the risk of resecting brain tissue involved in essential brain functions, such as speech or language comprehension, these patients often undergo presurgical functional magnetic resonance imaging (fMRI). This is a technique that shows subject-specific neural activity changes in the brain. The resulting fMRI data can assist the surgeon in performing the tumor resection while preserving the brain tissue involved in important cognitive and sensorimotor functions (Bartsch, Homola, Biller, Solymosi, & Bendszus, 2006) and can even be used to predict the outcome of postoperative cognitive functioning (Richardson et al., 2004).

To analyze fMRI data, a huge number of statistical tests are performed simultaneously. In cognitive neuroscience, this technique is used to link neurological and neuropsychological functions with their respective location in the brain, supporting different theories of brain function. To be confident that a brain area is associated with a task, it is essential to account for the multiple testing problem. This can be done using corrections for either the familywise error rate (Friston, Frith, Liddle, & Frackowiak, 1991; Worsley et al., 1996) or the false discovery rate (Genovese, Lazar, & Nichols, 2002). These multiple testing corrections result in a more stringent control of the null hypothesis of no activation, and consequently, the probability of a false negative increases (Lieberman & Cunningham, 2009;

Logan & Rowe, 2004). In cognitive neuroscience, a false positive means fallacious support for a given cognitive theory. While false positives can often be discovered by unsuccessfully trying to replicate the study, much time, effort, and money can be expended. As a result, the scientific discipline generally deems stringent control of false positives necessary, accepting the concomitant sacrifices in sensitivity.

In a clinical setting such as presurgical fMRI, however, a loss in power means that true activation is not discovered, and this might result in the resection of vital brain tissue (Haller & Bartsch, 2009). Inversely, false positives have a less negative impact on the surgical result (Gorgolewski, Storkey, Bastin, & Pernet, 2012). The goal of classical hypothesis testing is to prevent the null hypothesis from being rejected by considering voxels as being active only when enough evidence against the null of no activation is found. This asymmetrical way of penalizing errors in statistical inference is undesirable in this context (Johnson, Liu, Bartsch, & Nichols, 2012), and instead, the focus should be on protecting the alternative hypothesis: one wants to exclude activation only when enough evidence against activation is found. We therefore present a new hypothesis thresholding procedure that incorporates information on both false positives and false negatives and, thus, is ideally suited for presurgical fMRI.

In classical hypothesis testing, the evidence against null hypothesis is measured with the *p*-value, the null hypothesis probability of data as or more extreme than that observed. Thresholding a *p*-value at α produces a statistical test that controls the false positive rate at α. To allow direct control of false negative risk, we present a symmetrical measure that quantifies evidence against the alternative hypothesis (Moerkerke, Goetghebeur, De Riek, & Roldan-Ruiz, 2006). Correspondingly, thresholding this probability measure at β ensures control of the false negative rate at β.

By combining thresholds on the classical and alternative *p*-values, we use information on the probability of false positives and false negatives. We show that thresholding both error measurements results in a layered statistical map for the brain, each layer marking voxels with evidence (or lack thereof) against the null and/or alternative hypothesis. One layer consists of voxels exhibiting strong evidence against the null of no activation, while a second layer is formed by voxels for which activation cannot be confidently excluded. The third level then consists of voxels for which the presence of activation can be rejected.

fMRI data can be analyzed in different ways. The most popular method is a confirmatory mass-univariate general linear model (GLM) analysis, where the measured time series in each voxel is regressed onto the design of the experiment, resulting in an estimate of the effect, for which a *T*-statistic with a corresponding classical *p*-value can be computed for each voxel. This method has been shown to be very effective and robust, but its downside is the mass-univariate character. While many attempts have been made to take into account the spatial character of the data with data smoothing and peak- and cluster-thresholding, the GLM fails to recognize patterns of activation or noise. In this light, statistical techniques for multivariate data have been successfully applied to fMRI data. Independent component analysis (ICA; Beckmann & Smith, 2004) is an exploratory method used to find hidden source signals, modeling the observed data as a (unobserved) linear mixture of (unobserved) sources. ICA therefore allows one to discover spatially and temporally structured noise. Given the popularity of the GLM and the upcoming interest for ICA, especially in a clinical context, we will introduce the thresholding procedure for both techniques. We show how the ideas can be translated to different statistical techniques.

In the Method section, we introduce and combine quantities to measure significance when testing for activation. To this end, we start with a simple setting in which test statistics are assumed to be Gaussian distributed and take the general form of the ratio of an observed effect and its standard error. These settings directly translate to the case of univariate linear modeling that makes use of *T*-distributions. We further demonstrate how to use the principle for ICA. In the Results section, we present the results of the procedure applied to presurgical fMRI data.

Method

Measures of evidence against the null and alternative

At each voxel i, $i = 1, \ldots, I$, we assume that a linear model is fit and produces $\widehat{\Delta}_i$, an unbiased estimate of the BOLD effect of interest Δ_i, and an estimate of the standard deviation of $\widehat{\Delta}_i$, its "standard error" $\mathrm{SE}(\widehat{\Delta}_i)$. We henceforth suppress the voxel subscript unless needed for clarity. We assume that the degrees of freedom are sufficiently large so that $\mathrm{SE}(\widehat{\Delta})$ has negligible variability, as is the case for fMRI time series. We further assume that the data, model, and contrast have been scaled appropriately so that $\widehat{\Delta}$ has units of percent BOLD change (or at least approximately, as when global brain intensity is scaled to 100[1]).

The null and the alternative hypotheses

The null hypothesis $H_0{:}\Delta = 0$ states that the true effect magnitude is zero and an underlying difference between conditions Δ is equal to 0. Classical statistical inference involves computing a test statistic, converted to a *p*-value, that measures the evidence against this null hypothesis. The decision procedure to reject H_0 is calibrated to maintain the type I error

[1] Note that, as of SPM8, the global brain intensity after intensity normalization is scaled to 200 or greater.

at α. However, failing to reject H_0 does not allow one to conclude that H_0 is true. The reason is that the probability calculation of the p-value is based on the assumption that the null hypothesis is true. It is a logical fallacy, "affirming the consequent" or "reasoning to a forgone conclusion," to begin by assuming something and then, eventually, conclude that the initial thing is true. More concretely, when we fail to reject H_0, it could simply be because there are only subtle deviations from H_0 that are not detected or because the precision on the observed effect is too small to reach statistical significance. Scientists frequently make this mistake, and there have been various guidelines for reporting study results (see, e.g., Meehl, 1978; Schmidt & Hunter, 2002), all of which stress the importance of complementing p-values with effect sizes.

Our procedure considers an "alternative hypothesis" p-value, p_1, that measures the evidence against H_a:$\Delta = \Delta_1$, the nonzero effect magnitude expected under activation. Often, fMRI studies are preceded by power analyses for sample size calculations, which also require the specification of Δ_1. In literature, different approaches to choosing a meaningful Δ_1 have been presented (Desmond & Glover, 2002; Hayasaka, Pfeiffer, Hugenschmidt, & Laurienti, 2007; Mumford & Nichols, 2008; Zarahn & Slifstein, 2001). Alternatively, in presurgical fMRI, one can estimate Δ_1 on the basis of data in previous patients.

Measures of significance

At a given voxel, we have a test statistic T with observed value

$$t = \frac{\widehat{\Delta}}{\mathrm{SE}(\widehat{\Delta})}. \tag{1}$$

We assume that T has a known distribution under H_0 (e.g., Student's t with given degrees of freedom or Gaussian), so that we can compute the classical p-value:

$$p_0 = P(T \geq t|H_0). \tag{2}$$

That is, p_0 quantifies the evidence against the null hypothesis H_0 of no task-related activation.

In a symmetrical fashion, the alternative p-value is defined as in Moerkerke et al., 2006):

$$p_1 = P(T \leq t|H_a). \tag{3}$$

Correspondingly, p_1 measures the evidence against H_a and corresponds to the classical p-value for testing a "null" H_1 versus an "alternative" H_0. In general, as the evidence in favor of H_1 grows, p_0 becomes smaller and p_1 becomes larger.

In order to compute p_1, we need the distribution of T under H_a, which requires specification of Δ_1. However, we expect not a single magnitude of true activation, but a distribution of different true values (Desmond & Glover, 2002). Therefore, in a Bayesian spirit, we specify a distribution of likely values of Δ_1 instead of a fixed value:

$$\Delta_1 \sim \mathcal{N}(\mu, \tau^2), \tag{4}$$

where μ is the expected magnitude of effect under true activation while acknowledging variation among voxels—specifically, Gaussian variation with standard deviation τ.

Assuming that T also follows a Gaussian distribution, it has the following distribution under H_a at voxel i:

$$T_i \sim \mathcal{N}\left(\frac{\mu}{\mathrm{SE}(\widehat{\Delta}_i)}, \frac{\mathrm{SE}(\widehat{\Delta}_i)^2 + \tau^2}{\mathrm{SE}(\widehat{\Delta}_i)^2}\right), \tag{5}$$

where voxel subscripts are used to emphasize that the values of μ and τ are *fixed* for the entire brain and based on prior knowledge or other experiments, while $\mathrm{SE}(\widehat{\Delta}_i)$ is from each individual voxel. With this distribution, we can compute p_1 at each voxel. An illustration of both measures of significance can be seen in Fig. 1. Since the alternative distribution depends on the voxel-specific standard error, the distance between the null and alternative distributions will be voxel specific. In particular, a large standard error results in a large overlap between H_0 and H_a, while small standard errors lead to a large distance and little overlap between H_0 and H_a.

Combining measures of significance

In classical null hypothesis significance testing, a threshold α on p_0 can be translated into a threshold t_α for the test statistic in Eq. 1. In parallel, a threshold β on p_1 can be translated into a test statistic threshold t_β. While t_α is determined by α (and degrees of freedom, if not using a Gaussian), t_β further depends on β, μ, τ, *and* $\mathrm{SE}(\widehat{\Delta}_i)$. Thus, t_β varies over the brain depending on the (estimated) standard error.

Figure 2 shows the possible results of this testing procedure, with α and β relatively small. In what is expected to be the typical scenario, with a standard error that is large relative to the true effect magnitude, $t_\beta < t_\alpha$ and three possible outcomes can be distinguished.

One outcome is when voxels exhibit evidence against H_0 and, at the same time, are consistent with H_a ($p_0 < \alpha$ and $p_1 > \beta$; red in Fig. 2). This is the most compelling case for the presence of true activation ($\Delta > 0$). The opposite outcome is a large p_0 and a small p_1 ($p_0 > \alpha$ and $p_1 < \beta$; gray in Fig. 2). Here, the data are consistent with the null, and there is evidence to reject the alternative; this is the most compelling case for true absence of activation ($\Delta = 0$). The third outcome is when the data are compatible with both the null and the

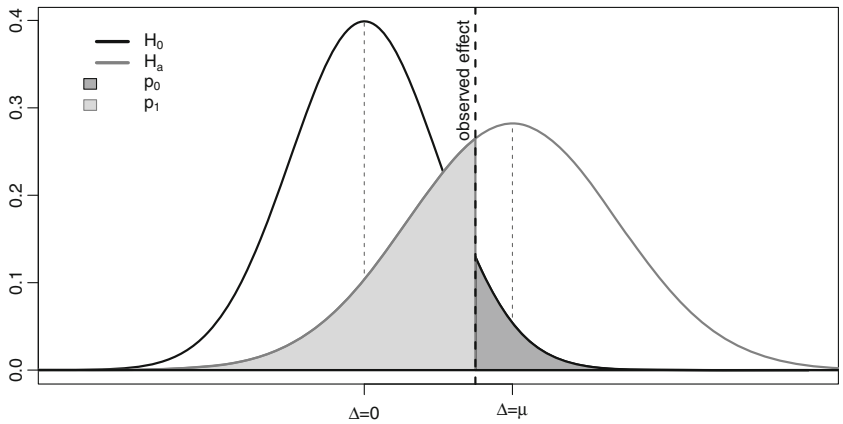

Fig. 1 The distributions of an effect under H_0 and H_a are displayed for an observed effect of $t = 1.5$, $\mathrm{SE}(\widehat{\Delta}) = 1$, $\Delta_1 = 2$, and $\tau = 1$. Note that H_a has a wider distribution than H_0 due to the uncertainty on Δ_1

alternative and neither can be excluded ($p_0 > \alpha$ and $p_1 > \beta$; yellow in Fig. 2).

A less frequent, albeit possible scenario appears when the standard error is small relative to the true effect magnitude, $t_\alpha < t_\beta$, and H_0 and H_a can be clearly distinguished. Voxels with no effect or strong effects will be identified as before ($p_0 > \alpha$ and $p_1 < \beta$, no activation; $p_0 < \alpha$ and $p_1 > \beta$, activation). However, for certain data, there is evidence against both H_0 and H_a ($p_0 < \alpha$ and $p_1 < \beta$; orange in Fig. 2). It indicates a case where the effect is so small as to lack *practical significance*.

For presurgical fMRI, this procedure provides information on which areas are confidently safe to be resected (gray areas), which areas should absolutely be avoided when resecting brain tissue (red areas), and in which areas the surgeon should take care because neither hypotheses can be rejected (yellow areas). When the fourth type of voxel is found, meaning both hypotheses can be rejected (orange areas), an abundance of caution suggests that again care be taken, since rejection of H_0 does suggest some association with the task, just at a possibly very small

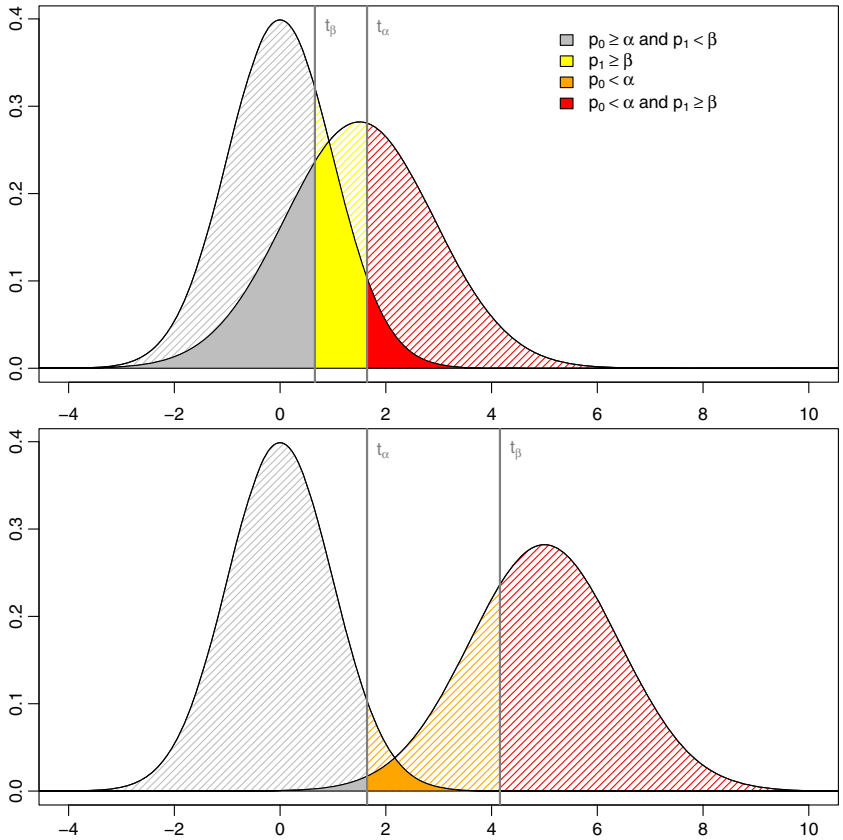

Fig. 2 When thresholding p_0 and p_1 at significance levels α and β, two possibilities arise: $t_\beta < t_\alpha$ (upper panel) or $t_\alpha < t_\beta$ (lower panel)

magnitude. The specific application to real data is shown in the Results section.

Alternative thresholding of independent component analysis

Above, we described the classical and alternative *p*-value for a traditional setting, where a test statistic is the ratio of an observed effect and its standard error, as is the case for *T*-statistics when the GLM is used. Here, we demonstrate that the technique is also applicable in more general settings—in particular, with maps from independent component analysis. Exact implementation details of ICA methods differ; our development here follows the FSL[2] software's implementation, MELODIC (Beckmann & Smith, 2004), but should be readily applicable to other ICA software.

ICA is a technique for multivariate data-driven analysis of fMRI data. It does not require the specification of the experimental design and produces spatiotemporal patterns that explain the variability in the data. ICA transforms the four-dimensional fMRI data into K pairs of spatial and temporal modes. Each spatial mode, or independent component (IC) image, is associated with one IC time series. The variation explained by each component is the IC time series scaled by the weights at each voxel in the IC image; equivalently, it is the spatial pattern in an IC image scaled by each value of the IC time series. Stated simply, the weights represent the association between the temporal activation pattern observed in the voxel and the temporal pattern in the K different components.

Let $\boldsymbol{Y}$ represent the $J \times I$ data matrix, where J is the number of time points and I is the number of voxels. We assume that the data at each voxel have been mean-centered; that is, the column means of $\boldsymbol{Y}$ are zero. ICA decomposes the data as per

$$\boldsymbol{Y} \approx \boldsymbol{M}\ \boldsymbol{S}_1\ \boldsymbol{C}\ \boldsymbol{S}_2\boldsymbol{S}_0, \tag{6}$$

where $\boldsymbol{M}$ is a $J \times K$ matrix with one temporal mode in each column and $\boldsymbol{C}$ is a $K \times I$ matrix with one spatial mode in each row; $\boldsymbol{S}_1$ ($K \times K$) is a diagonal scaling matrix that ensures that the temporal modes have unit variance, and $\boldsymbol{S}_0$ and $\boldsymbol{S}_2$ (both $I \times I$) are diagonal scaling matrices that ensure that background noise in the spatial modes have unit variance (see Appendix 1 for detailed definitions of these scaling factors).

In the presentation and interpretation of ICA results, each of the K spatial modes in $\boldsymbol{C}$ are visualized and explored. Since they have been noise-normalized, they are often treated as *z*-score images and thresholded to control a nominal false positive rate. The end result is an inference that quantifies the relation between the corresponding temporal mode in $\boldsymbol{M}$ and $\boldsymbol{Y}$. We seek to apply our alternative hypothesis thresholding procedure to these maps, but first we need to define a meaningful effect size in percentage of BOLD change and transform this to the scale of $\boldsymbol{C}$.

Meaningful BOLD effect sizes with ICA

Consider a particular IC of interest, $k \in \{1, \ldots, K\}$, and a particular voxel $i \in \{1, \ldots, I\}$ of interest in the spatial mode. Specifically, consider the contribution of the kth IC to the time series at voxel i:

$$\boldsymbol{m}_k\ \ s_{1,k}\ \ c_{ki}\ \ s_{2,i}\ \ s_{0,i}, \tag{7}$$

where $\boldsymbol{m}_k$ is the kth column of $\boldsymbol{M}$, $c_{ki} = (\boldsymbol{C})_{ki}$, and $s_{1,k}$, $s_{2,i}$, and $s_{0,i}$ are the indicated diagonal elements of the scaling matrices.

As was previously mentioned, the rows of $\boldsymbol{C}$ are normalized to have noise variance of 1, so c_{ki} has *z*-score (and not BOLD data) units. We need to compute a meaningful percentage of BOLD change effect. We will first compute this for a fixed Δ_1, in the units of c_{ki}, and will later impose a distribution on the effect size. Equation 7 shows that the temporal variation from IC k is determined not only by c_{ki} and the scaling factors, but also by $\boldsymbol{m}_k$. But $\boldsymbol{m}_k$ is scaled to unit variance and will not induce a unit BOLD change in the data. We propose scaling $\boldsymbol{m}_k$ so that it (roughly) expresses a unit BOLD effect and, as a result, preserves the units of the other terms. Specifically, we introduce h_k:

$$\boldsymbol{m}_k\ \ h_k h_k^{-1}\ \ s_{1,k}\ \ c_{ki}\ \ s_{2,i}\ \ s_{0,i}., \tag{8}$$

so that $\boldsymbol{m}_k h_k$ expresses a unit BOLD effect in the data. One way to set the factor h_k is so that $\boldsymbol{m}_k h_k$ has a baseline-to-peak range of 1. Another way is to regress $\boldsymbol{m}_k$ on a covariate $\boldsymbol{d}$ that expresses the anticipated (unit) experimental effect; setting h_k to the inverse of the regression coefficient will ensure that $\boldsymbol{m}_k h_k$ corresponds to an approximate unit BOLD effect.

Finally, we correct for the attenuation of the hypothesized effect based on the mismatch between $\boldsymbol{m}_k$ and $\boldsymbol{d}$. That is, even if we choose h_k well, $\boldsymbol{m}_k h_k$ may only be weakly correlated with $\boldsymbol{d}$. As a result, we scale the expected BOLD effect of IC k at voxel i by $\rho_{\boldsymbol{m}_k \boldsymbol{d}}$, the correlation between $\boldsymbol{m}_k$ and $\boldsymbol{d}$ (see Appendix 2 for details).

Now we can relate the expected (attenuated) percentage of BOLD change, $\rho_{\boldsymbol{m}_k \boldsymbol{d}} \Delta_1$, to the units of IC temporal mode. Let Δ_1^* be the expected alternative mean effect in the *z*-score statistic c_{ki}; then,

$$\rho_{\boldsymbol{m}_k \boldsymbol{d}}\ \ \Delta_1 \approx h_k^{-1}\ \ s_{1,k}\ \ \Delta_1^*\ \ s_{2,i}\ \ s_{0,i}, \tag{9}$$

and thus we can translate BOLD units into c_{ik} units with $\Delta_1^* \approx s_{ik}^* \Delta_1$, where

$$s_{ki}^* = \rho_{\boldsymbol{m}_k \boldsymbol{d}}\ \ h_k\ \ s_{1,k}^{-1}\ \ s_{2,i}^{-1}\ \ s_{0,i}^{-1}. \tag{10}$$

[2] http://www.fmrib.ox.ac.uk/fsl.

Finally, this implies that our distribution of alternative effects in c_{ki} units is

$$\Delta_1^* \sim \mathcal{N}\left(s_{ki}^* \ \mu, \ s_{ki}^{*2} \ \tau^2\right) \tag{11}$$

(cf Eq. 4).

Significance procedure

Since c_{ki} has unit noise variance, with an assumption of Gaussianity, the null distribution is given by

$$c_{ki}|H_0 \sim \mathcal{N}(0,1). \tag{12}$$

Under the alternative, we consider the addition of effect Δ_1^* to c_{ki} yielding the alternative distribution

$$c_{ki}|H_a \sim \mathcal{N}\left(s_{ki}^* \ \mu, \ 1 + s_{ki}^{*2} \ \tau^2\right). \tag{13}$$

Data

We consider data from a patient suffering from a left prefrontal brain tumor. The study design was a boxcar design, where the patient was asked to alternate between recitation of tongue-twisters and quiescence. Figure 3 shows a sagittal slice of the T2 image, with the tumor visible in the inferior prefrontal frontal cortex. For the application to mass univariate linear modeling, the data were analyzed with FEAT in FSL 4.1 (Smith et al., 2004). The application to independent component analysis was performed using MELODIC in FSL 4.1 (Beckmann & Smith, 2004).

Results

Univariate linear modeling

We applied these techniques to the data described in the Data section. We derived the expected effect magnitude for Δ_1 and the variability of that effect τ from 5 patients who underwent the same fMRI paradigm. We threshold the image of each individual using an FDR control at 0.05 and look at the average percent BOLD change units in each individual. The results are shown in Table 1. Therefore, we specify the expected effect magnitude for Δ_1 of $\mu = 0.73$ percent BOLD change units and variability of that effect as $\tau = \sqrt{\hat{\tau}^2} = 0.21$ percent BOLD change. These results are consistent with others in the literature (see, e.g., Desmond & Glover, 2002, Fig. 7A).

Results are shown in Fig. 4 with thresholds $\alpha = 0.001$ and $\beta = 0.20$. In other words, we specified a p_0 threshold for declaring an activation when there is none at 1-in-1,000; and we set the p_1 threshold for declaring the absence of activation when, in fact, the specified activation magnitude is present at 1-in-5. The red and the (scant) orange voxels show where H_0 can be confidently rejected, and, if presurgical planning was done only on the basis of classical null hypothesis testing, all

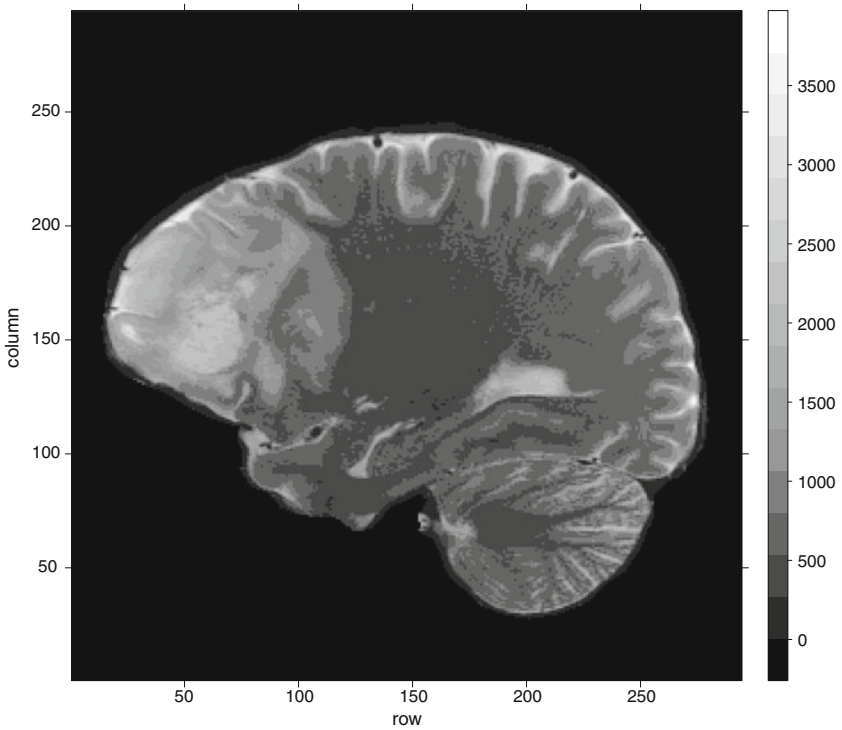

Fig. 3 Anatomical scan of the patient. The tumor can be clearly seen in the prefrontal cortex

Table 1 Average effect sizes in 5 previously tested patients in percent BOLD change units

	Average $\widehat{\mu}$	Average $\widehat{\tau}^2$
Patient 1	0.59	0.20
Patient 2	0.55	0.26
Patient 3	0.68	0.35
Patient 4	0.75	0.46
Patient 5	1.08	0.84
Average	0.73	0.43

other tissue would be regarded as "safe." Considering information on the alternative, we have the red voxels where, specifically, H_0 can be rejected and H_a cannot be rejected; that is, the red voxels are incompatible with the null and compatible with the alternative and, thus, are strong evidence for the effect. The yellow areas are areas where neither H_0 *nor* H_a can be rejected; here, the data are compatible with both the null and alternative and suggest a lack of confidence in ruling out activation. Finally, for voxels with no coloration, the H_0 cannot be rejected, but H_a can; the data are compatible with the null and incompatible with the alternative and, thus, have good evidence for a lack of activation and suggest that these brain regions can be safely resected. This shows the key strength of the procedure: Among voxels traditionally classified as "nonactive"—that is, those with insufficiently small p_0s, it distinguishes between voxels where there is compelling evidence for nonactivation (not colored) and those voxels where we cannot rule out the possibility of activation (yellow).

The orange voxels represent voxels for which the observed effect size is between the null hypothesis of no

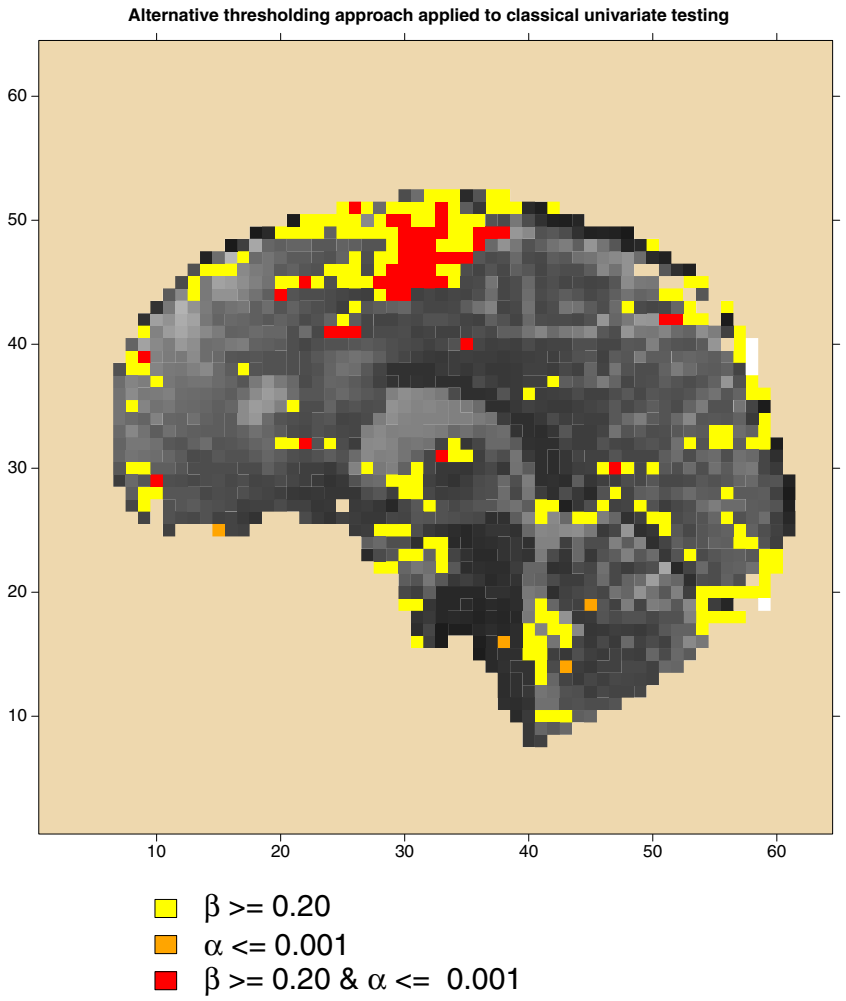

Fig. 4 Sagittal slice of "layered" activation inference overlaying gray-scale T2* reference image, threshold values of $\alpha = 0.001$ and $\beta = 0.20$. Red areas show areas of high confidence of activation (H_0 rejected, H_a not rejected), while yellow areas show areas where activation cannot be ruled out (neither H_0 nor H_a rejected); uncolored areas have high confidence of no activation (H_0 not rejected, H_a rejected), while the few orange voxels indicate voxels with significant but surprisingly small BOLD response magnitude (H_0 and H_a rejected)

activation and the expected effect size. In these voxels, both the null and the alternative hypotheses are rejected, which corresponds to very low residual noise in the GLM.

Independent components analysis results

We applied these techniques to the data described in the Data section. We used the same effect size and uncertainty as in the Univariate linear modeling section—that is, $\mu = 0.73$ and $\tau = 0.18$ percent BOLD change units.

MELODIC's automated dimensionality estimation method in MELODIC found 52 components. We chose one IC whose time series corresponded to the design matrix, shown in Fig. 5. Regressing this temporal mode on the design gives a coefficient of $\widehat{\beta} = 1.48$, and thus $h = \widehat{\beta}^{-1} = 0.677$ is the scaling factor used to have the temporal mode express a unit-BOLD effect (see Eq. 8). The pointwise correlation between the design and the chosen component is $\rho_{md} = 0.63$, which is used to attenuate the expected effect magnitude (see Eq. 9).

The layered thresholding procedure for this IC is shown in Fig. 6, for $\alpha = 0.001$ and $\beta = 0.20$. There is a set of voxels with strong evidence (red, H_0 rejected; H_a, accepted) but also additional voxels where both hypotheses are rejected (orange). As was mentioned above, in the setting of presurgical planning, these orange regions are best regarded as regions of possible activation and, thus, excluded from resection.

This result is quite different from the GLM results and is a reflection of the dramatically lower voxel-wise variance in the IC spatial mode relative to the GLM statistic image. The explanation is that the GLM result accounts for all noise variance, while the IC spatial map reflects only the noise in the subspace corresponding to the IC temporal mode (Beckmann & Smith, 2004).

Crucially, we stress that our thresholding procedure seeks only to improve the interpretability of the ICA result and does not produce confirmatory inferences; IC selection is intrincally post hoc and subsequent inferences circular, and all we attempt to do here is improve the thresholding of a selected IC spatial map.

Discussion

Statistical thresholding in the context of multiple tests is generally driven by the need to limit false positives. These stringent testing procedures in fMRI research lead to an abundance of false negatives (Lieberman & Cunningham, 2009) and are, therefore, less useful in the context of presurgical fMRI, where a false negative can have dire consequences. While many attempts have been made to propose more liberal testing criteria—for example, by controlling the FDR instead of the FWER (Genovese et al., 2002)—the focus is still on protecting the type I error rate. The unilateral focus on preventing false positives leads to a bias toward large obvious effects and against complex cognitive and affective effects (Lieberman & Cunningham, 2009). We therefore propose a measure that quantifies the evidence against the alternative hypothesis as introduced in Moerkerke et al. (2006). We use this quantity p_1 in addition to the classical p_0 value in a procedure that results in a thresholding procedure with multiple layers of significance. One layer consists of voxels exhibiting strong evidence of activation (red in Figs. 4 and 6), while another layer shows voxels with ambiguous evidence (yellow and orange), and a final layer then consists of voxels for which the presence of activation can be confidently rejected (an absence of overlaid statistic values). Thereby, we offer a more symmetrical interest toward both false positives and false negatives.

We have chosen to focus on voxel-wise inference instead of other topological features, such as peaks (Chumbley, Worsley, Flandin, & Friston, 2010) or clusters (Chumbley & Friston, 2009). These topological inference methods have reduced spatial specificity relative to voxel-wise inference and are, therefore, less suitable for presurgical fMRI, where maximal spatial precision is needed.

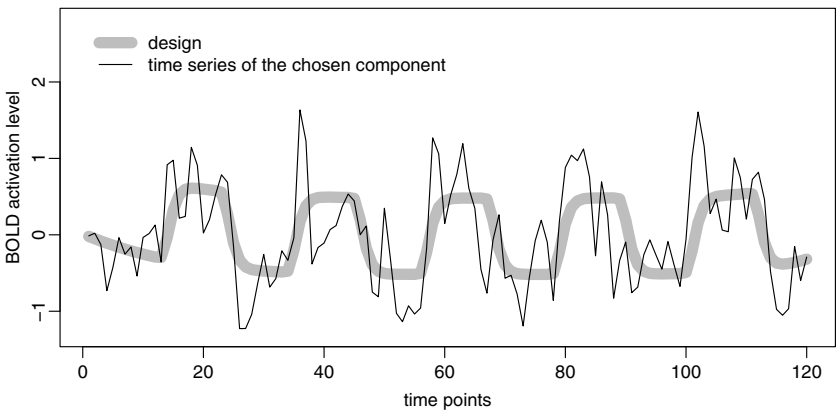

Fig. 5 The time series of a selected IC, with a least squares fit of regressing the series on the design shown in gray. The estimated response height is used to normalize the component to have unit BOLD effect, and the pointwise correlation between the design and the selected IC is used to attenuate the expected BOLD response magnitude

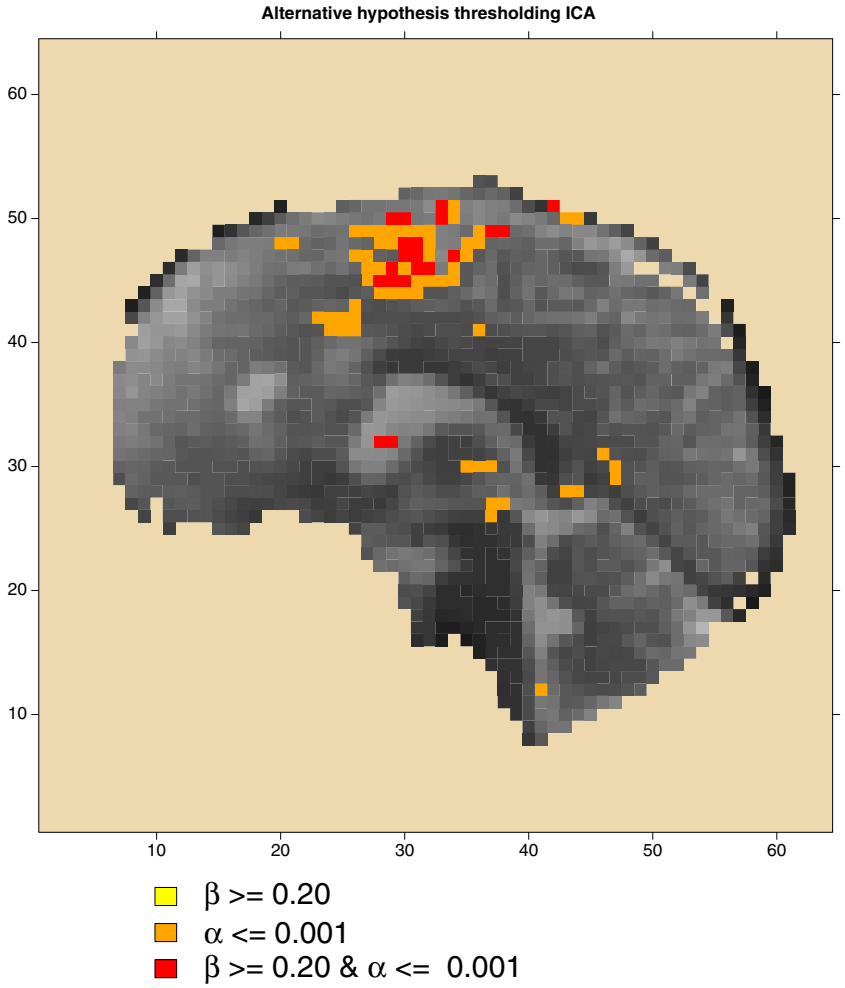

Fig. 6 Results of the alternative thresholding procedure when using ICA. Sagittal slice of "layered" activation inference overlaying gray-scale T2* reference image, threshold values of $\alpha = 0.001$ and $\beta = 0.20$. Red areas show areas of high confidence of activation (H_0 rejected, H_a not rejected), orange areas show voxels with significant but surprisingly small BOLD response magnitude (H_0 and H_a rejected); uncolored areas have high confidence of no activation (H_0 not rejected, H_a rejected)

To use the procedure described in this article, an expected effect size and its variance need to be defined on a BOLD scale. This is an arbitrary choice, however many possibilities are available. Desmond and Glover (2002) showed, for a specific experimental paradigm, the distribution of percentage of signal change with its distribution. They showed, on average, a BOLD effect size of 0.48 percent BOLD change. Another possibility for estimating the expected effect size can be based on previous research. Since, in presurgical fMRI, the same experiment is repeated over most patients, the effect size can be derived from patients who already underwent the experiment and surgery. The degree to which brain activation in patients is representative for the particular setting for which estimates are needed highly depends on the context and should be carefully judged. It can be expected that different methods will affect the estimates for μ and τ; however, we found that the estimates we obtained by averaging over voxels is close to the effect sizes that can be found in literature.

The two different analytical approaches we used, the GLM and ICA, showed somewhat different results. While both GLM and ICA analyses found similar sets of voxels that were confidently activated (H_0 rejected, H_a not), in the GLM analysis many voxels were found that did not show evidence against the null or against the alternative (yellow in Fig. 4). The explanation for this outcome is the high level of noise present in the data and, thus, confusion about the veracity of either H_0 or H_a. In contrast, in the ICA analysis, almost no voxels have this ambiguity, and instead, we find voxels that have evidence against both the null and the alternative. Since ICA is a good tool for identifying structured noise in a data-driven manner, it can be expected that the residual voxel-wise variance will be smaller. Low variances result in a large distance between the null and the alternative distribution functions. Whereas the difference between ICA and the GLM seem contradictory at first, we argue that the differences in our approach reflect real differences between the two analysis tools.

The quantity p_1 shows a relationship with the voxel-based statistical power defined by Van Horn, Ellmore, Esposito, and Berman (1998). The voxel-wise power in Van Horn et al. translates to the complement of the alternative p-value, p_1, in our study. However, the use of the quantity is fundamentally different. Whereas Van Horn et al. used the voxel-wise power to visualize and interpret the results of a certain study, we explicitly threshold the quantity. Moreover, the interpretation of both quantities is not so straightforward. When a high power is encountered in a certain voxel, with the method of Van Horn et al., it is interpreted as follows: "If the observed effect in the voxel is used as a cutoff when testing from H_0, we have a high probability of rejecting H_0 when H_0 is indeed false." However, a large voxel-wise power translates to a small p_1 and is, in our study, interpreted as follows: "When the alternative hypothesis is true, there is a small probability of observing this effect," and we will interpret this effect as evidence against the alternative hypothesis. This interpretation is much more straightforward and usable.

This procedure has been developed in light of presurgical fMRI, since false negatives can have harmful consequences for the patient. However, the lack of power is omnipresent in fMRI analyses (Lieberman & Cunningham, 2009), and therefore, this procedure is also very useful in all branches of cognitive neuroscience. For example, negative results (i.e., voxels that are not significantly related to the task) are sometimes regarded as evidence against activation. However, these conclusions are not provided by null hypothesis significance testing. The presented procedure, on the other hand, quantifies the evidence for no activation at each voxel and is, therefore, perfectly suited to interpreting negative results.

We would like to stress that this procedure does not abandon null hypothesis significance testing. The classical significance testing framework is still included in the procedure, represented by one layer of significance. The method is merely an extension of the thresholded statistical parametric map, thereby providing a new layer with information on type II error rate control. Mixture modeling is similar in spirit to this method, in that null and the alternative distribution are used; however, mixture model applications usually focus on only controlling type I errors. With a fitted mixture model, you could also apply our method and find p_0 and p_1 values; however, we take pains to estimate alternative effect magnitudes a priori, from separate data, to remove any circularity.

In this procedure, control of false positives remains possible, but our procedure also takes into account information on the false negative rate. We do not assert that our method alleviates all concerns with multiplicity, and one possible direction of future work is a multiplicity correction that adjusts both null and alternative hypothesis inferences for the number of tests.

Appendix 1. Scaling steps in ICA

Let $\boldsymbol{Y}$ be the $J \times I$ data matrix for time points $j = 1, \ldots, J$ and voxels $i = 1, \ldots, I$. Then the normalized data matrix $\boldsymbol{Y}^*$ can be expressed as

$$\boldsymbol{Y}^* = \boldsymbol{Y}\boldsymbol{S}_0^{-1}, \tag{14}$$

where $\boldsymbol{S}_0$ the $I \times I$ diagonal matrix with voxel-wise robust variance estimates on the diagonal. Then the independent component analysis results in the following decomposition

$$\boldsymbol{Y}^* \approx \boldsymbol{M}\boldsymbol{C}^*, \tag{15}$$

where $\boldsymbol{M}$ represents the $J \times K$ mixing matrix, where K is the number of components. When $K < J$, Eq. 15 is only an approximation. $\boldsymbol{C}^*$ is the $K \times I$ matrix with the original image component loadings.

In the probabilistic ICA framework (Beckmann & Smith, 2004), $\boldsymbol{C}^*$ is Gaussian distributed and can, therefore, be used as a test statistic. To normalize $\boldsymbol{C}^*$ to a standard Gaussian distribution,

$$\boldsymbol{C} = \boldsymbol{S}_1^{-1}\boldsymbol{C}^*\boldsymbol{S}_2^{-1}, \tag{16}$$

where the diagonal matrix $\boldsymbol{S}_1$ scales the components (over voxels) and the diagonal matrix $\boldsymbol{S}_2$ scales the voxels (over components):

$$\boldsymbol{S}_1^2 = diag\left\{\boldsymbol{M}^{-1}\left(\boldsymbol{M}^{-1}\right)'\right\} \tag{17}$$

$$\boldsymbol{S}_2 = diag\{\mathrm{SD}(\boldsymbol{Y}^* - \boldsymbol{M}\boldsymbol{C}^*)\}\frac{\sqrt{I-K}}{\sqrt{I-1}}, \tag{18}$$

where $\mathrm{SD}(\boldsymbol{Y}^* - \boldsymbol{M}\boldsymbol{C}^*)$ is the column-wise standard deviation of the residuals of the ICA approximation. Consequently, we can approximate $\boldsymbol{Y}$ as $\boldsymbol{M}\boldsymbol{S}_1\boldsymbol{C}\boldsymbol{S}_2\boldsymbol{S}_0$ in Eq. 6.

Appendix 2. Attenuation of anticipated BOLD effect for a given IC

A basic result in psychometrics (Spearman, 1904) holds that the correlation between unreliable measures is attenuated by the test–retest reliability of each measure. For example, if measure A imperfectly measures variable A^* and B imperfectly measures B^*, then the correlation of A and B is attenuated relative to the uncorrupted measures:

$$\mathrm{Corr}(A, B) = \mathrm{Corr}(A^*, B^*)\ \sqrt{\rho_{AA}}\ \sqrt{\rho_{BB}}, \tag{19}$$

where ρ_{AA} and ρ_{BB} are the test–retest correlations of A and B.

In our setting, let A be the BOLD response and B be the IC time course $\boldsymbol{m}$. The reproducibility of BOLD (ρ_{AA}) is respectable, with Vul, Harris, Winkielman, and Pashler (2009) reporting reliabilities between .66 and .94; however, most

procedures for fMRI data analysis do not take this into account, and hence we only consider $\rho_{AA} = 1$.

We are interested in the "reproducibility" of an IC time course *relative* to the experimental design $\boldsymbol{d}$. Of course, obtaining two replicates of an IC time course that both equally reflect $\boldsymbol{d}$ is not feasible, but we can indirectly estimate ρ_{BB} as follows.

Consider a general test–retest setting, where measurement B_1 is made at one time and, later, a "retest" gives measurement B_2. Assuming additive error, we can relate the "corrupted" measures to the "uncorrupted" measures as

$$B_1 = B^* + \varepsilon_1 \tag{20}$$

$$B_2 = B^* + \varepsilon_2, \tag{21}$$

where $\varepsilon_j, j = 1, 2$ are the measurement-specific errors, and we assume $\mathrm{Var}(B^*) = \sigma^2$ is the variance of the perfectly reproducible measure and $\mathrm{Var}(\varepsilon_j)$, is the variance of the corrupting noise. The test–retest correlation is then

$$\rho_{BB} = \mathrm{Corr}(B_1, B_2) = \frac{\sigma^2}{\sigma^2 + \tau^2}. \tag{22}$$

If, instead, one corrupted measure is correlated with the uncorrupted "true" measure, you find

$$\mathrm{Corr}(B_1, B^*) = \frac{\sigma}{\sqrt{\sigma^2 + \tau^2}} = \sqrt{\rho_{BB}}. \tag{23}$$

Or, equivalently, $\rho_{BB} = \mathrm{Corr}(B_1, B^*)^2$. In short, these results show that we can estimate the "reproducibility" of $\boldsymbol{m}_\mathrm{k}$ as a noisy sample of the true $\boldsymbol{d}$ as $\rho^2_{\boldsymbol{m}_k \boldsymbol{d}}$.

Finally, from Eq. 19, taking $\rho_{AA} = 1$ and $\rho_{BB} = \rho^2_{\boldsymbol{m}_k \boldsymbol{d}}$, we see that the attenuation factor needed to account for the mismatch between $\boldsymbol{m}_k$ and $\boldsymbol{d}$ is just $\rho_{\boldsymbol{m}_k \boldsymbol{d}}$.

References

Bartsch, A. J., Homola, G., Biller, A., Solymosi, L., & Bendszus, M. (2006). Diagnostic functional MRI: Illustrated clinical applications and decision-making. *Journal of Magnetic Resonance Imaging: JMRI, 23*(6), 921–932.

Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging, 23*(2), 137–152.

Chumbley, J., Worsley, K., Flandin, G., & Friston, K. (2010). Topological FDR for neuroimaging. *NeuroImage, 49*(4), 3057–3064.

Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage, 44*(1), 62–70.

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods, 118*(2), 115–128.

Friston, K. J., Frith, C. D., Liddle, P. F., & Frackowiak, R. S. (1991). Comparing functional (PET) images: The assessment of significant change. *Journal of Cerebral Blood Flow and Metabolism, 11*(4), 690–699.

Genovese, C. R., Lazar, N. A., & Nichols, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage, 15*(4), 870–878.

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., & Pernet, C. R. (2012). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in Human Neuroscience, 6*(245), 1–14.

Haller, S., & Bartsch, A. J. (2009). Pitfalls in FMRI. *European Radiology, 19*(11), 2689–2706.

Hayasaka, S., Peiffer, A. M., Hugenschmidt, C. E., & Laurienti, P. J. (2007). Power and sample size calculation for neuroimaging studies by non-central random field theory. *NeuroImage, 37*(3), 721–730.

Johnson, T. D., Liu, Z., Bartsch, A. J., & Nichols, T. E. (2012). A Bayesian non-parametric Potts model with application to presurgical FMRI data. *Statistical Methods in Medical Research*.

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience, 4*(4), 423–428.

Logan, B. R., & Rowe, D. B. (2004). An evaluation of thresholding techniques in fMRI analysis. *NeuroImage, 22*(1), 95–108.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Moerkerke, B., Goetghebeur, E., De Riek, J., & Roldan-Ruiz, I. (2006). Significance and impotence: Towards a balanced view of the null and the alternative hypotheses in marker selection for plant breeding. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(1), 61–79.

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage, 39*(1), 261–268.

Richardson, M. P., Strange, B. A., Thompson, P. J., Baxendale, S. A., Duncan, J. S., & Dolan, R. J. (2004). Pre-operative verbal memory fMRI predicts post-operative memory decline after left temporal lobe resection. *Brain: A Journal of Neurology, 127*(Pt 11), 2419–2426.

Schmidt, F., & Hunter, J. (2002). Are there benefits from NHST. *American Psychologist, 57*(1), 65–66.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., . . . Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage, 23 Suppl 1*, S208–19.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101.

Van Horn, J. D., Ellmore, T. M., Esposito, G., & Berman, K. F. (1998). Mapping voxel-based statistical power on parametric images. *NeuroImage, 7*(2), 97–107.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274–290.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping, 4*(1), 58–73.

Zarahn, E., & Slifstein, M. (2001). A reference effect approach for power analysis in fMRI. *NeuroImage, 14*(3), 768–779.