ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 44 (2009) 71 – 82

Empirical null and false discovery rate analysis in neuroimaging

Armin Schwartzman,[a,*] Robert F. Dougherty,[b] Jongho Lee,[c]
Dara Ghahremani,[d] and Jonathan E. Taylor[e]

[a]*Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA*
[b]*Department of Psychology, Stanford University, Stanford, CA 94305, USA*
[c]*Advanced MRI Section, Laboratory of Functional and Molecular Imaging (LFMI), National Institute of Neurological Disorders and Stroke (NINDS), National Institutes of Health, Bethesda, Maryland 20824, USA*
[d]*Department of Psychology, University of California, Los Angeles, California 90095, USA*
[e]*Department of Statistics, Stanford University, Stanford, California 94305, USA*

Received 22 May 2007; revised 1 April 2008; accepted 5 April 2008
Available online 24 April 2008

Current strategies for thresholding statistical parametric maps in neuroimaging include control of the family-wise error rate, control of the false discovery rate (FDR) and thresholding of the posterior probability of a voxel being active given the data, the latter derived from a mixture model of active and inactive voxels. Correct inference using any of these criteria depends crucially on the specification of the null distribution of the test statistics. In this article we show examples from fMRI and DTI data where the theoretical null distribution does not match well the observed distribution of the test statistics. As a solution, we introduce the use of an empirical null, a null distribution empirically estimated from the data itself, allowing for global corrections of theoretical null assumptions. The theoretical null distributions considered are normal, t, χ^2 and F, all commonly encountered in neuroimaging. The empirical null estimate is accompanied by an estimate of the proportion of non-active voxels in the data. Based on the two-class mixture model, we present the equivalence between the strategies of controlling FDR and thresholding posterior probabilities in the context of neuroimaging and show that the FDR estimates derived from the empirical null can be seen as empirical Bayes estimates.

Introduction

A common objective in neuroimaging studies is to identify spatial locations where an effect is statistically significant. In voxel-based analyses, after image preprocessing (e.g., spatial registration and normalization) the statistical analysis typically consists of model fitting and computation of a test statistic at every voxel in the search region, followed by thresholding of the test statistic map, also called statistical parametric map (SPM). In the neuroimaging literature, at least three different thresholding criteria are currently in use: control of the family-wise error rate (FWER), control of the false discovery rate (FDR) and thresholding of posterior probabilities of activation. Correct inference using any of these criteria depends crucially on the appropriate assessment of the null distribution of the test statistics. We show in this paper several examples where the observed distribution of the test statistics does not match well the distribution specified by the theoretical model. As a solution to this problem we present a method for estimating the null distribution from the data itself and show that it may fundamentally change the results of the analysis and their interpretation. Based on a two-class mixture model, this empirical null methodology also provides an estimate for the proportion of null voxels. The known Bayesian interpretation of FDR then allows us to draw an equivalence between thresholding of SPMs based on FDR and thresholding based on posterior probabilities, and to see the estimates of FDR based on the empirical null as empirical Bayes estimates.

Traditional thresholding of SPMs has focused on control of the FWER, for which the most common methods are Bonferroni and random field corrections (Worsley et al., 1996, 2004). FDR controlling methods have also proven useful in neuroimaging studies (Genovese et al., 2002; Logan and Rowe, 2004; Marchini and Presanis, 2004; Schwartzman et al., 2005). The difference between these two criteria is that FWER corrections control the rate of any false positives in the entire search volume as a whole, whereas FDR corrections control the rate of false positives among all positives. By definition, FDR is more permissive of false positives and thus tends to yield lower thresholds. FDR might be preferable in exploratory studies where some false positive voxels can be tolerated as long as many more true positive voxels are found. FWER might be preferable in confirmatory studies where false

* Corresponding author. Department of Biostatistics, Dana-Farber Cancer Institute, CLS-11007, 44 Binney Street, Boston, MA 02115, USA. Fax: +1 617 632 2444.
E-mail address: armins@hsph.harvard.edu (A. Schwartzman).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.04.182

positives are not tolerated. For a comparison of the two methods, see Nichols and Hayasaka (2003).

In addition to the frequentist approach of controlling error rates, Bayesian thresholding rules have been proposed in the neuroimaging literature (Everitt and Bullmore, 1999; Hartvig and Jensen, 2000; Friston et al., 2002; Woolrich et al., 2005). These rules threshold the posterior probability of a voxel being a true positive given the data. In particular, Everitt and Bullmore (1999) proposed a mixture model in which the voxels were assumed to belong to either an active or an inactive class. By specifying a distribution on each of the two classes, they derived a posterior probability of a voxel being active given the data and proposed to threshold it at 0.5. Spatial regularization has been subsequently incorporated into this framework (Hartvig and Jensen, 2000; Woolrich et al., 2005).

Regardless of the chosen criterion, the inference depends crucially on the assumed null distribution of the test statistics. It has been observed in functional MRI (fMRI) (Ghahremani and Taylor, 2005) and Diffusion Tensor Imaging (DTI) (Schwartzman et al., 2008) that the observed distribution of the test statistics may not match the null distribution specified by the theoretical model. Inference based on an inappropriate theoretical null may produce misleading results. In this paper we present four examples of this phenomenon, of which more details will be given in Data examples. Briefly, the first example is an fMRI data set involving a standard visual (flashing checkerboard), auditory (monotonely increasing tones) and motor (self-paced finger tapping) stimulation in a single subject. Fig. 3a shows a histogram of z-scores from a standard linear model for $N=15{,}611$ voxels in the entire brain. The observed distribution of the data is substantially wider than the theoretical density $N(0, 1)$. The patterns of activation dictated by either null are also substantially different, as shown in Figs. 3c and d. The second example is an fMRI data set from a single subject involving a comparison between two repetitions of a visual word/nonword classification task. Fig. 4a shows a histogram of z-scores for $N=232{,}731$ voxels in the entire brain. The observed distribution of the data is shifted towards the left and narrower than the theoretical density $N(0, 1)$. In this case, activation is only detected using the empirical null (Fig. 4d). The third and fourth examples are respectively a steady-state free precession (SSFP) fMRI data set and a DTI data set. In both cases the voxel-wise analysis leads to test statistics whose theoretical null distributions are χ^2. Similar mismatches between the data and the theoretical null are observed (Figs. 5 and 6).

In this article we adopt an empirical Bayes approach to thresholding SPMs where: 1) the null distribution and the proportion of null voxels are estimated from the data itself; 2) the FDR is interpreted as a posterior probability. This approach was originally motivated by the analysis of genetic microarray data (Efron, 2004, 2007). As opposed to the usual FDR control seen in the neuroimaging literature based on ordering of p-values (Genovese et al., 2002; Logan and Rowe, 2004), the approach we advocate involves first estimating the FDR for every threshold and then picking a threshold that satisfies a given FDR. These two forms of FDR control are exactly equivalent when the null distribution is known (Benjamini and Hochberg, 1995; Yekutieli and Benjamini, 1999; Benjamini and Yekutieli, 2001; Storey et al., 2004) but have different interpretations. In the context of a two-class mixture model, the estimated FDR may be interpreted as the posterior probability of a voxel belonging to the active class given the data (Storey, 2003). Conversely, a threshold chosen for the posterior probability map provides control of the FDR at that level. Furthermore, the FDR and posterior probability afford graphical representation in terms of the distribution of the test statistics, aiding the interpretation of the results.

Viewed in this light, the empirical estimation of the null distribution can be seen as an empirical Bayes estimate. The empirical null was originally formulated for z and t-statistics (Efron, 2004, 2007). Motivated by a DTI analysis problem, it has been recently extended to χ^2 and F-statistics (Schwartzman et al., 2008) and general exponential families (Schwartzman, 2008). This paper demonstrates the relevance and implementation of the empirical Bayes approach in the analysis of standard neuroimaging data such as single-contrast fMRI and less standard neuroimaging data such as complex fMRI and DTI.

For simplicity, in what follows we borrow the language of fMRI and use the words "active"/"inactive" in a generic way to refer to voxels with positive/negative test results for both fMRI and DTI data sets.

Theory and methods

Empirical null

Let S be a search region containing N voxels, so that N is the number of tests. The search region can be divided into two sets: a null set $S_0 \subset S$ containing all the N_0 truly inactive (null) voxels and an active set $S_A = S \setminus S_0$ containing all the $N - N_0$ truly active (non-null) voxels. Let T_i be the test statistic at voxel i and assume that its null density $f_0(t)$ is the same in all voxels. The marginal density of the test statistics can be written as the mixture (Everitt and Bullmore, 1999; Hartvig and Jensen, 2000; Storey, 2003; Efron, 2004)

$$f(t) = p_0 f_0(t) + (1 - p_0) f_A(t), \tag{1}$$

where $p_0 = N_0/N$ is the fraction of truly inactive voxels in S and $f_A(t)$ is the marginal density within S_A, which may itself be a mixture.

In traditional parametric hypothesis testing, the null distribution is usually the result of theoretical modeling. In large scale multiple testing situations it is possible to estimate the null distribution from the data itself. The empirical null (Efron et al., 2001; Efron, 2004, 2007) relies on the following assumptions:

1. The number of tests (voxels) N is large, say in the thousands.
2. Most tests are null (i.e. most voxels are inactive), say at least 90%.

Consider a histogram of the N test statistics in S with bin width Δ chosen so that the largest bins contain a relatively large number of voxels, say in the hundreds. Under Assumption 1, the histogram is a reasonable empirical estimate of the scaled mixture density $N\Delta f(t)$, where $f(t)$ is given by Expression (1) and the scaling factor $N\Delta$ accounts for the fact that the histogram is made up of counts and grouped in bins. If p_0 is close to 1 (Assumption 2), we expect the bulk of the data histogram to follow the scaled null

$$\tilde{f}_0(t) = (N\Delta) p_0 f_0(t) \tag{2}$$

with the non-null voxels falling mostly in the tails. The idea is to fit a functional form of the scaled density (Expression (2)) to the bulk of the data histogram. Because the histogram is made up of counts, Poisson regression is the most appropriate fitting method. Here we consider four statistics whose null distributions are common in neuroimaging: z, t, χ^2 and F.

1. If the theoretical null is $N(0,1)$ (z-scores), the empirical null is a general normal $N(\mu, \sigma^2)$ with mean and variance possibly different from 0 and 1. This empirical null model respects the

shape of the theoretical null but can adjust for shift or scaling of the data (Efron et al., 2001; Efron, 2004).
2. If the theoretical null is χ^2 with ν_0 degrees of freedom (*d.f.*), the empirical null is a scaled χ^2 (i.e. gamma) with ν d.f. possibly different from ν_0 and a scaling parameter a possibly different from 1 (Schwartzman et al., 2008).
3. If the theoretical null is t with ν_0 *d.f.*, the t-scores are converted to z-scores by a one-to-one quantile transformation (Efron et al., 2001; Efron, 2004; Smith et al., 2004). Analysis then proceeds as in the z-score case.
4. If the theoretical null is F with ν_1 and ν_2 *d.f.*, the F-scores are converted to χ^2 scores with ν_1 *d.f.* by a one-to-one quantile transformation (Schwartzman et al., 2008). Analysis then proceeds as in the χ^2 case.

Details on how to fit the normal and χ^2 empirical nulls are given by Schwartzman (2008) and are summarized in the Appendix A. Fitting of parametric t and F distributions directly is more difficult because the normal and χ^2 distributions are exponential families while t and F are not (Schwartzman, 2008). For this reason we opt to do a transformation that eliminates the denominator *d.f.* Quantile transformations are standard in software packages such as FSL (Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), Oxford University, Oxford, U.K.) (Smith et al., 2004). While their use has been discouraged when the denominator number of *d.f.* is small (Nichols and Hayasaka, 2003), their effect is negligible when the denominator number of *d.f.* is large, say larger than 30.

Fitting the empirical null to the scaled density (Expression (2)) provides simultaneous estimates of p_0 and $f_0(t)$. However, if the analyst believes that the theoretical null $f_0(t)$ is correct, then the Poisson regression setup can be changed so that only p_0 is estimated (see Appendix A).

FDR estimation

The empirical null is particularly suited for FDR inference. Next we give a brief review of frequentist FDR when the null distribution is known and then show how the empirical null makes the FDR estimate an empirical Bayes estimate.

Assume, for simplicity, that the tests T_i are one-sided on the right: given a threshold u, voxel i is declared active if $T_i > u$. For any fixed threshold u, the right-tail FDR is the expected proportion of false discoveries, defined as (Benjamini and Hochberg, 1995)

$$\mathrm{FDR_r}(u) = E\left[\frac{V(u)}{\max\{R(u),1\}}\right] \tag{3}$$

where $R(u)$ is the number of voxels declared active (discoveries) and $V(u)$ is the number declared active that are truly inactive (false discoveries). The effect of the maximum operator in the denominator is to set the FDR to zero if $R(u)=0$. Heuristically, Expression (3) can be approximated by applying the expectation to the numerator only:

$$\mathrm{FDR_r}(u) \approx \frac{E[V(u)]}{\max\{R(u),1\}} = \frac{E[V(u)/N]}{\max\{R(u),1\}/N} = \frac{p_0 E[V(u)/N_0]}{\max\{R(u),1\}/N}.$$

Note that $E[V(u)/N_0]$ is exactly the probability $P_0[T_i > u] = P[T_i > u | H_0]$ of declaring a voxel active given that it is truly inactive, while $\max\{R(u),1\}/N$ is an empirical estimate of the unconditional probability $P[T_i > u]$ that a voxel is declared active. Based on this, when the null density $f_0(t)$ is known, the right-tail FDR can be estimated by the ratio (Storey et al., 2004)

$$\widehat{\mathrm{FDR}}_\mathrm{r}(u) = \frac{\hat{p}_0 P_0[T_i > u]}{\hat{P}[T_i > u]} \tag{4}$$

where the "hat" denotes estimation. While there are many ways to estimate the null proportion p_0 (Benjamini and Yekutieli, 2001; Genovese et al., 2002; Storey et al., 2004; Heller et al., 2007), here we use the "empirical null" estimate obtained when $f_0(t)$ is known. If $f_0(t)$ is also estimated by the empirical null, then the estimator (4) becomes

$$\widehat{\mathrm{FDR}}_\mathrm{r}(u) = \frac{\hat{p}_0 \hat{P}_0[T_i > u]}{\hat{P}[T_i > u]}. \tag{5}$$

Compare to Expression (4) and notice the hat on P_0 in Expression (5) provided by the empirical null.

Expressions (4) and (5) have a nice graphical interpretation as the ratio of the tail areas under the null and marginal densities respectively. In Figs. 1a and 3a–6a, this corresponds to the ratio between the right-tail areas under the null (dashed or solid line) and under the histogram for any fixed value u on the abscissa axis. Notice in Figs. 1b and 3b–6b that the right-tail FDR tends to decrease as u increases, while it tends to the estimate of p_0 as u decreases.

If a threshold u is fixed arbitrarily or is available from a previous analysis (e.g. FWER), then Expressions (4) and (5) give estimates of the FDR for that analysis. Otherwise, a threshold may be found so that the FDR is controlled at a tolerance level α by means of the following algorithm (Storey et al., 2004):

Algorithm 1.

1. Compute the estimate $\widehat{\mathrm{FDR}}_\mathrm{r}(u)$ for a range of thresholds u.
2. Set the final threshold as the smallest u for which $\widehat{\mathrm{FDR}}_\mathrm{r}(u) \le \alpha$.

If the curve does not cross the level α then no null hypotheses are rejected. When $f_0(t)$ is assumed known, Algorithm 1 with $\hat{p}_0 = 1$ is equivalent to the Benjamini and Hochberg (1995) procedure, so control of the FDR at level α is guaranteed asymptotically if the number of voxels is large and the test statistics are independent or weakly dependent. The estimates of p_0 and $f_0(t)$ provided by the empirical null are consistent if the test statistics are independent or weakly dependent (Schwartzman, 2008), so FDR control is also guaranteed asymptotically.

Left-tail and double-tail FDR can be treated in a similar way. For the left-tail FDR, the test rejects when $T_i < u$, while for the double-tail FDR, the test rejects when $|T_i| > u$. The FDR estimates are given by

$$\widehat{\mathrm{FDR}}_\mathrm{l}(u) = \frac{\hat{p}_0 P_0[T_i < u]}{\hat{P}[T_i < u]}, \quad \widehat{\mathrm{FDR}}_\mathrm{d}(u) = \frac{\hat{p}_0 P_0[|T_i| > u]}{\hat{P}[|T_i| > u]} \tag{6}$$

or with $\hat{P}_0$ instead of P_0 if the empirical null is used. For a required level α, the corresponding threshold is found using an algorithm similar to Algorithm 1. For the double-tail FDR there is no change, while for the left-tail FDR the final threshold is set as the *largest* u such that $\widehat{\mathrm{FDR}}_\mathrm{l}(u) \le \alpha$.

Bayesian interpretation of FDR

The FDR can be interpreted as a posterior probability under the mixture model (1). Before observing the data, a voxel i chosen at

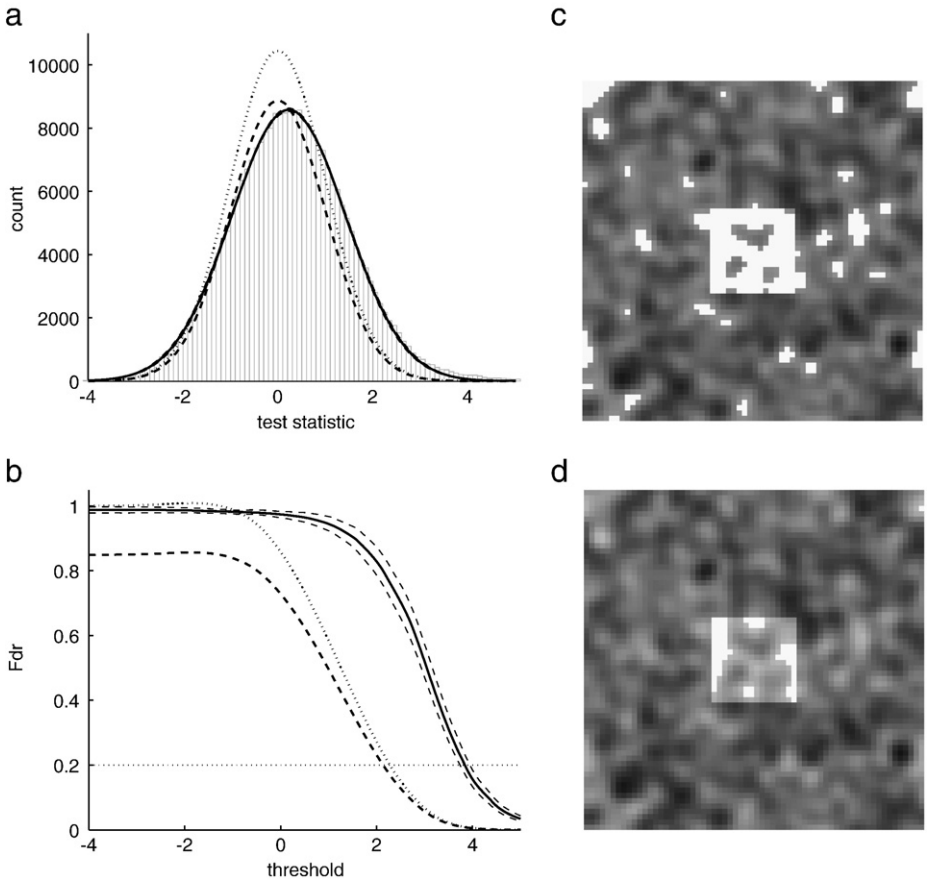

Fig. 1. Simulated example: (a) Histogram of the $N=262{,}144$ z-statistics (gray); theoretical null $N(0, 1)$ (dotted); scaled theoretical null (thick dashed); empirical null (thick solid). Confidence bands omitted for clarity. (b) Tail FDR curves: theoretical null (thick dotted); scaled theoretical null (thick dashed); empirical null (thick solid). Confidence bands are 95% point-wise (thin dashed). Horizontal dashed line is the chosen FDR level 0.1. (c) Significant voxels (white) at FDR level 0.1 according to the scaled theoretical null. The slice is No. 36 out of 64. (d) Significant voxels (white) at FDR level 0.1 according to the empirical null; same slice as in (c).

random uniformly over S has a prior probability $P[i\in S_0]=p_0$ of being inactive. After observing the data, if $T_i>u$ then the posterior probability of voxel i being inactive is

$$P[i\in S_0|T_i>u]=\frac{p_0P_0[T_i>u]}{P[T_i>u]}, \tag{7}$$

where $P[T_i>u]=p_0P_0[T_i>u]+(1-p_0)P_A[T_i>u]$, and the probabilities $P_0[T_i>u]$ and $P_A[T_i>u]$ are computed under f_0 and f_A respectively. The estimates (4) and (5) can now be seen as estimates of the posterior probability (Expression (7)).

When the test statistics are independent, Expression (7) is exactly equal to the positive FDR $\mathrm{pFDR_r}(u)=E[V(u)/R(u)|R(u)>0]$ (Storey, 2003), related to the original FDR (Expression (3)) via

$$\mathrm{FDR_r}(u)=\mathrm{pFDR_r}(u)\cdot P[R(u)>0]. \tag{8}$$

Expression (8) shows how FDR plays an intermediate role between the frequentist FWER and the Bayesian approach. When N is large and p_0 is away from 1, $P[R(u)>0]\approx 1$ and $\mathrm{FDR_r}(u)\approx \mathrm{pFDR_r}(u)=P[i\in S_0|T_i>u]$ has a Bayesian interpretation. On the other hand if we observe pure noise, i.e. $p_0=1$, then $\mathrm{pFDR_r}(u)=1$ and cannot be controlled. In this case the FDR takes the FWER interpretation: $\mathrm{FDR_r}(u)=P[R(u)>0]=P[V(u)>0]=\mathrm{FWER}(u)$.

Instead of thresholding Expression (7), Everitt and Bullmore (1999), Hartvig and Jensen (2000) and Woolrich et al. (2005) propose thresholding the posterior probability

$$P[i\in S_A|T_i=t]=\frac{(1-p_0)f_A(t)}{f(t)}. \tag{9}$$

This quantity directly relates to the FDR theory because it is equal to one minus the *local FDR* (Efron, 2004, 2007)

$$\mathrm{fdr}(t)=P[i\in S_0|T_i=t]=\frac{p_0f_0(t)}{f(t)}, \tag{10}$$

a version of Expression (7) based on densities and related to the right-tail FDR by the formula

$$\mathrm{pFDR_r}(u)=\frac{\int_u^\infty \mathrm{fdr}(t)f(t)dt}{\int_u^\infty f(t)dt} \tag{11}$$

(changing the integration limits to $(-\infty, t)$ gives a formula for the left-tail pFDR). Consequently, the posterior probability (Expression (9)) may be called the local true discovery rate (TDR). The corresponding (right-tail) cumulative version $\mathrm{pTDR_r}(u)=P[i\in S_A|T_i>u]$ measures the proportion of true positives among positives

and has been proposed as a measure of statistical power in multiple testing problems (Dudoit et al., 2003).

Simulations

To illustrate the behavior of the empirical null under local correlation, we simulated the following artificial example. A random Gaussian field occupying $64\times64\times64$ voxels was first generated by convolving a field of independent Gaussian samples with a spherical Gaussian kernel. The standard deviation of the Gaussian kernel was chosen as $s=1.5$ in each dimension, equivalent to a full-width-half-max (FWHM) of $2\sqrt{2\log 2}s = 2.5$. The field was then shifted and scaled in order to have marginal expectation $\mu=0.2$ and variance $\sigma^2=1.2^2$. A signal of intensity 3 occupying $16\times16\times16$ voxels was placed in the center of the $64\times64\times64$ region. The true null fraction is $p_0=1-16^3/64^3=0.9844$.

Fig. 1 a shows a histogram of the $N=64^3=262{,}144$ voxels using $\Delta=0.1$. By design, it is shifted with respect to the theoretical null and wider. The empirical null was computed using the same method as in Example 1 below, yielding estimates $p_0=0.988$, $\hat{\mu}=0.225$ and $\hat{\sigma}=1.204$, close to the target values. For comparison, estimation of p_0 using the theoretical null only gives $\hat{p}_0=0.849$ and does not fit the data well. As discussed in Schwartzman et al. (2008), the empirical null estimates may be slightly biased by the presence of the signal.

The FDR analysis in Fig. 1b shows that the misspecified theoretical null gives a misleadingly low threshold 2.14. The effect is seen in Fig. 1c, where many spurious "discoveries" are made, 18,200 voxels at FDR level 0.2. In contrast, the empirical null gives a conservatively higher threshold 3.84 and yields only 1732 active voxels, but these are mostly contained in the correct region, as shown in Fig. 1d. The empirical null also gives an indication of the error incurred in the activation map of Fig. 1c. The empirical null curve in Fig. 1b evaluated at the threshold 2.14 estimates the FDR at 0.8 ± 0.05, far higher than the nominal 0.2.

In order to assess the FDR properties of the empirical null under local correlation, the above simulation was repeated with the following changes. For speed, the field was reduced to $32\times32\times32$ with signal occupying an $8\times8\times8$ cube in the center of the region. The null proportion is $p_0=1-8^3/32^3=0.9844$. The smoothing parameter s was varied from 0 (no smoothing) to 3. Fig. 2 summarizes the results for 100 repetitions of the simulation. Panel a shows that the parameter estimates for the empirical null remain close to the target values while the null proportion obtained from the theoretical null is always off substantially. In all cases the standard error increases with the smoothing parameter because of the effective reduction in sample size. Panel b shows the achieved FDR. The empirical null achieves the desired FDR approximately, while the theoretical null does not.

Data examples

Example 1: BOLD fMRI

As a first example we consider a data set from a single subject acquired by BOLD fMRI during simultaneous exposure to three standard stimuli: visual (8 Hz flashing checkerboard), auditory (10 single frequency tones monotonely increasing from 1 kHz to 10 kHz, duration of 1/3 s each), and motor (self-paced finger tapping). The stimuli were presented for 3 min in blocks of 30 s off and 30 s on. Images were acquired 1/s for a total of 60 scans. The voxel size is $3.75\times3.75\times4$ mm. The search region comprises a whole-brain mask containing $N=15{,}611$ voxels.

The data were analyzed using FEAT in FSL (Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), Oxford University, Oxford, U.K.) (Smith et al., 2004). The analysis at each voxel involved fitting a standard linear model (Friston et al., 1995) and automatic conversion of the appropriate contrast from t-scale to z-scale. Since the number of $d.f.$ $60-2=58$ is relatively large, there is little difference between the converted z-scores and the unconverted t-scores.

Fig. 3 a shows a histogram of the 15,611 z-scores using a bin width $\Delta=0.05$. The histogram is substantially wider than the theoretical null $N(0, 1)$ and slightly shifted to the right. This is confirmed by the empirical null estimated parameters $\hat{p}_0=0.87$, $\hat{\mu}=0.14$ and $\hat{\sigma}=1.57$. These estimates were obtained using the Poisson regression method described in Appendix A.1 twice, once in the interval $[-1,1]$ to find a preliminary estimate of the mode location $\hat{\mu}^*$ and width $\hat{\sigma}^*$, and a second time in the interval $[\hat{\mu}^*-\hat{\sigma}^*, \hat{\mu}^*-\hat{\sigma}^*]$ to estimate the empirical

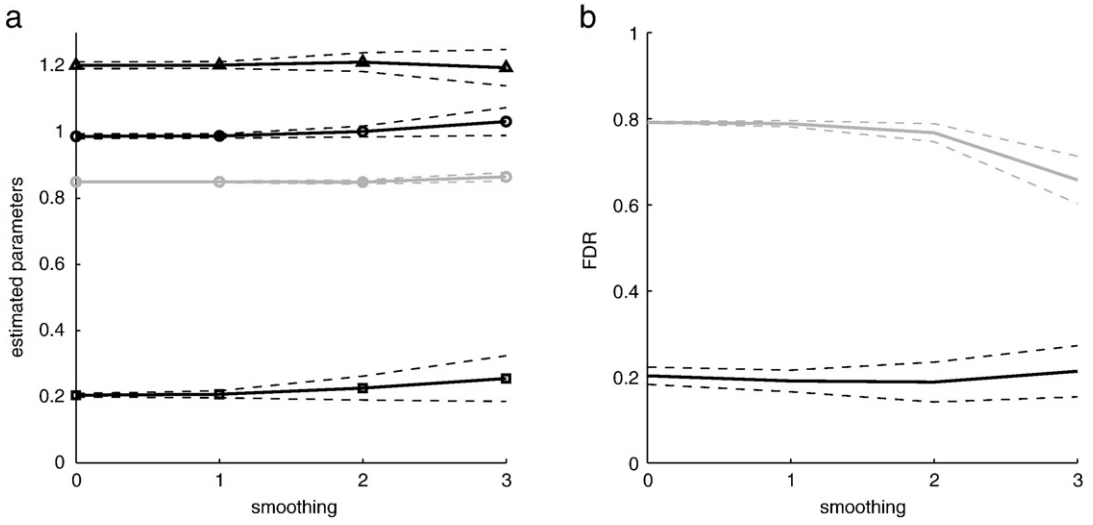

Fig. 2. (a) Estimated parameters. Theoretical null: $\hat{p}_0$ (gray, ○); empirical null: $\hat{p}_0$ (black, ○), $\hat{\mu}$ (black, □), $\hat{\sigma}$ (black, △). Confidence bands are 95% point-wise (thin dashed). (b) Realized FDR: theoretical null (gray), empirical null (black). Confidence bands are 95% point-wise (thin dashed).

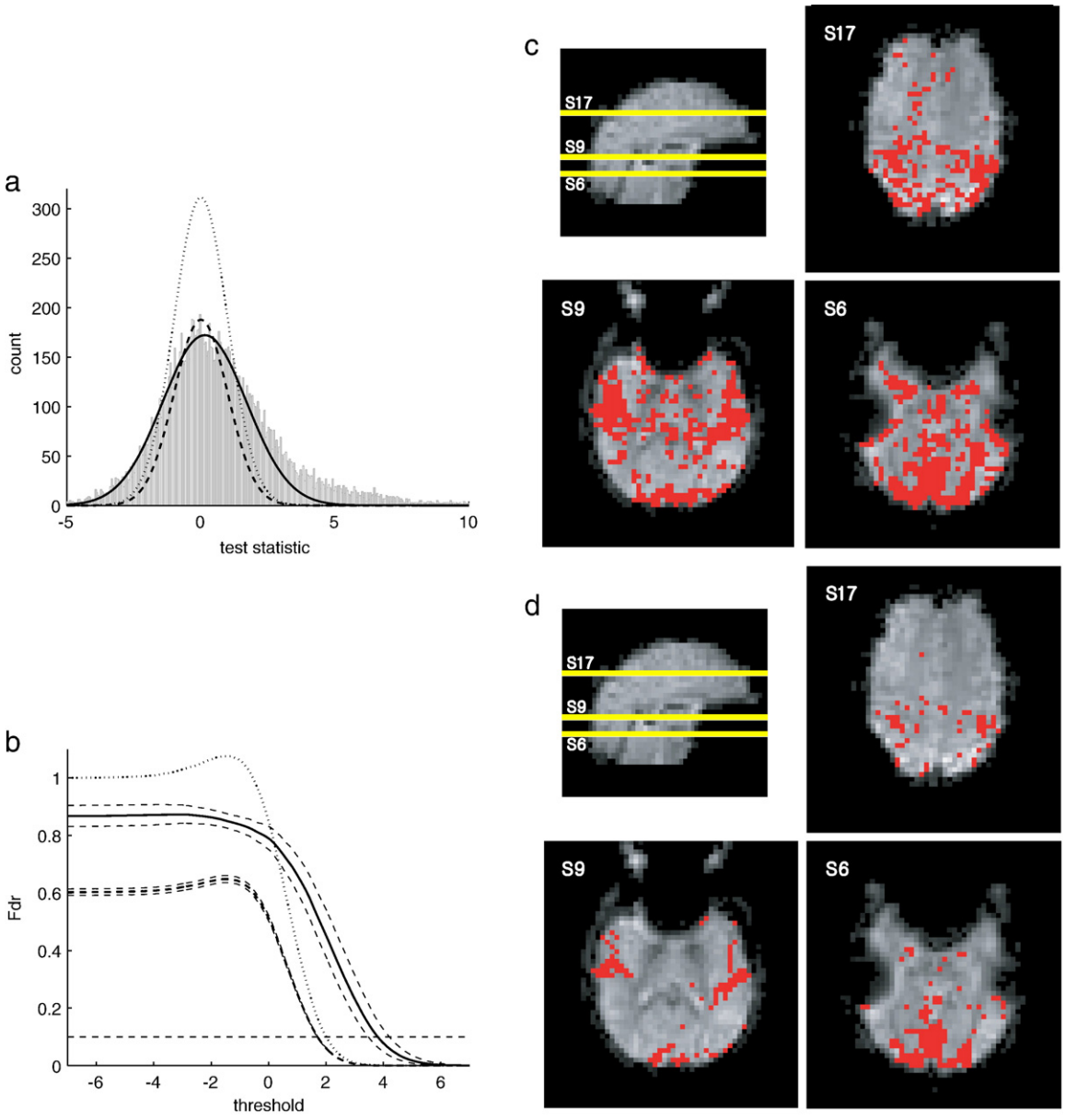

Fig. 3. fMRI example No. 1: (a) Histogram of the $N=15{,}611$ z-scores (gray); scaled theoretical null (thick dashed) 0.60 $N(0, 1)$; empirical null 0.87 $N(0.14, 1.57^2)$ (thick solid). Confidence bands omitted for clarity. (b) Right-tail FDR curves: theoretical null (thick dotted); scaled theoretical null (thick dashed); empirical null (thick solid). Confidence bands are 95% point-wise (thin dashed). Horizontal dashed line is the chosen FDR level 0.1. (c) Relevant significant voxels at FDR level 0.1 according to the scaled theoretical null. (d) Significant voxels at FDR level 0.1 according to the empirical null; same slices as in (c).

null parameters reported above. For comparison, the same method was applied in the interval $[-1,1]$ with $\mu=0$ and $\sigma=1$ fixed as prescribed by the theoretical null, yielding the estimate $\hat{p}_0=0.60$. Here the theoretical null estimates the total fraction of active voxels at 40% of the voxels, while the empirical null estimates it at 13%.

The FDR analysis for the right-tail is shown in Fig. 3b. The theoretical null curve corresponds to the estimator (4) with $\hat{p}_0$ set to 1, while the scaled theoretical null curve corresponds to the same estimator with $\hat{p}_0=0.60$. The empirical null curve corresponds to the estimator (5). Notice that all curves tend to decrease as the threshold u increases and tend to the corresponding estimate of p_0 as u decreases.

To obtain specific FDR thresholds, the FDR level was set at 0.1 and Algorithm 1 was applied. The threshold 2.01 obtained from the theoretical null is the same as the Benjamini and Hochberg (1995) threshold and is not too different from the threshold 1.70 obtained from the scaled theoretical null. The scaled theoretical null threshold 1.70 yielded 3501 active voxels. In contrast, the empirical null threshold 3.80 is substantially higher and yielded 1328 active voxels. The most relevant of these active voxels are shown in Figs. 3c and d, respectively. Notice that the empirical null better captures the visual, auditory and motor regions that are expected to be active, while the theoretical null produces many spurious active voxels throughout the brain.

Another way to interpret this result is the following. If the threshold was set arbitrarily to the scaled theoretical null threshold 1.70, yielding the activation pattern in Fig. 3c, then the empirical null curve in Fig. 3b indicates that the estimated FDR for that

activation pattern is 0.52±0.09, much higher than the nominal 0.1. To obtain the nominal 0.1, the empirical null prescribes the threshold 3.80, or to be on the safe side, the threshold corresponding to the right edge of the confidence interval, 4.21.

Example 2: BOLD fMRI

As a second example we consider a data set where the empirical null has the opposite effect to Example 1. In this event-related BOLD fMRI data set, a single subject was presented low-frequency words and "nonwords" (pronounceable nonwords) and was required to make a word/nonword classification via a button press. Each word was presented twice, and the comparison of interest was the difference between the first and second repetitions.

Each 3 s trial began with a 1.5 s word presentation and response period followed by a 1.5 s delay. Trials in which participants fixated their eyes on a cross in the center of the screen were pseudo-randomly inserted to allow the hemodynamic response to return to baseline. Forty-nine trials were presented per condition. As this was a rapid event-related design with trials appearing close together in time, we used a two-back trial-history counterbalancing procedure to produce a pseudo-randomized presentation order. This counterbalancing method ensured that each possible sequence of two trial types preceding a given trial was equally represented throughout the experiment (Hazeltine et al., 2000). An average of 20 intervening trials appeared between initial and repeated stimulus presentations.

Data analysis was performed using the SPM analysis package (SPM2, Wellcome Department of Cognitive Neurology). Slices of functional images were first temporally aligned (slice timing correction); images were then co-registered to account for head motion. The T_2 anatomical inplane was co-registered to the mean motion-corrected functional image and spatially normalized to the (MontrealNeurological Institute) MNI canonical brain. These spatial normalization parameters were applied to all functional images, which were later smoothed with a Gaussian, 6 mm FWHM isotropic kernel. Normalization resulted in a 2×2×2 mm voxel size. Differences between stimulus conditions were examined by using the general linear model (Friston et al., 1995), modeling activation at each voxel with a synthetic hemodynamic response function. In a fixed-effect analysis, contrast images were computed for each comparison of interest along with a corresponding t-statistic image. The t-statistics were converted to z-scores, but there is almost no difference between them as the number of $d.f.$ 320−2=318 is very large.

Fig. 4 a shows a histogram of the 232,731 z-scores using a bin width Δ=0.05. The histogram is substantially shifted to the left and

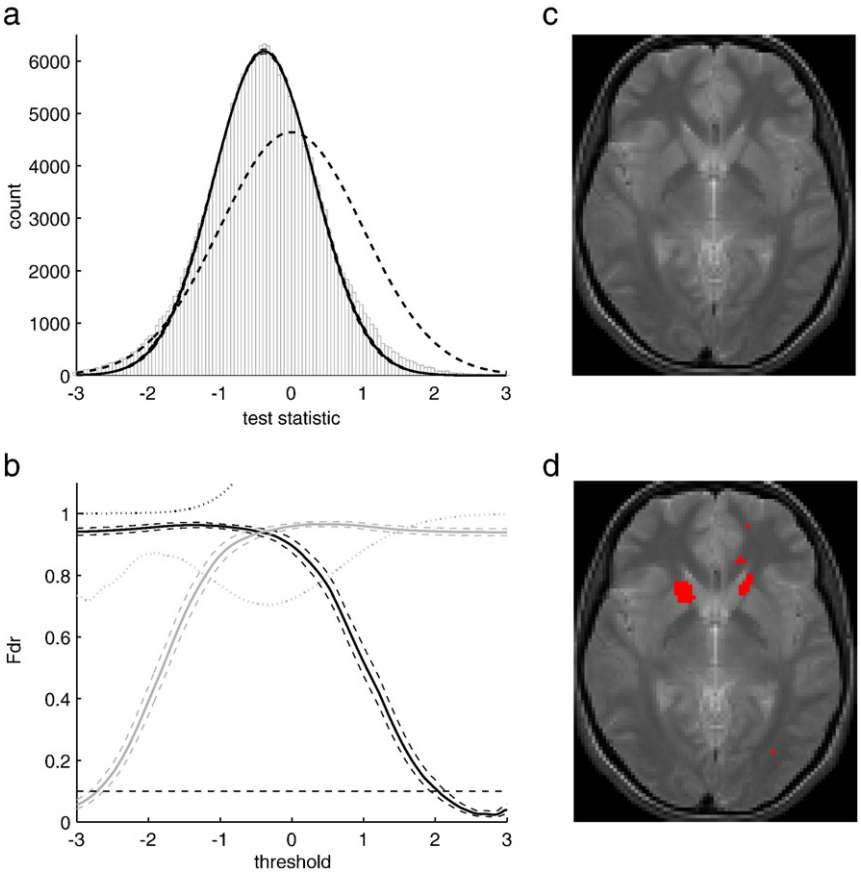

Fig. 4. fMRI example No. 2: (a) Histogram of the N=232,731 z-scores (gray); theoretical null $N(0, 1)$ (dotted); empirical null 0.94 $N(-0.38, 0.71^2)$ (thick solid). Confidence bands are 95% point-wise (thin dashed). (b) Right-tail (black) and left-tail (gray) FDR curves: theoretical null (thick dotted); empirical null (thick solid). Confidence bands omitted for clarity. Horizontal dashed line is the chosen FDR level 0.1. (c–d) Right-tail significant voxels at FDR level 0.1 according to the theoretical null (c) and empirical null (d). The slice corresponds to z=−2 mm in MNI coordinates, with the striatum shown active by the empirical null (d) but inactive by the theoretical null (c).

slightly narrower than the theoretical null $N(0, 1)$. This is confirmed by the empirical null estimated parameters $\hat{p}_0=0.94$, $\hat{\mu}=-0.38$ and $\hat{\sigma}=0.71$, which were obtained using the same method as in Example 1. For comparison, a scaled theoretical null was also fitted. The meaningless estimate $\hat{p}_0=1.10$ is an indication that the theoretical model is untenable for this data.

The FDR analysis is shown in Fig. 4b. The theoretical null FDR curves are meaningless because of the poor fit of the theoretical model to the data. In contrast, the empirical null FDR curves decrease nicely in both directions. According to the empirical null, setting the FDR level at 0.1 yielded the right-tail threshold 2.04 and a total of 655 significant voxels. The most relevant slice is shown in Fig. 4d. In contrast, the theoretical null yielded no significant voxels at the FDR=0.1 level, as emphasized by Fig. 4c. Another way to interpret this result is to say that if the right-tail threshold was arbitrarily set to 2.04, obtaining the active voxels of Fig. 4d, then the estimated FDR according to the empirical null would be 0.1 ± 0.02, while according to the theoretical null it would be meaningless.

Example 3: SSFP fMRI

Our third and fourth examples illustrate how the empirical null can be applied in cases where the test statistics are χ^2 or F. The third example is a data set from a single subject acquired by complex SSFP fMRI (Lee et al., 2007). The stimulus was a flashing checkerboard (15 s on/off for 2 min 18 s), for a total of $n=45$ observations. The data were analyzed using FEAT in FSL (Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), Oxford University, Oxford, U.K.) (Smith et al., 2004). The complex time-series data were decomposed into real and imaginary parts and modeled independently using the standard linear model. A Hotelling T^2 statistic was then computed for detecting differences between the real and imaginary components with theoretical null distribution $F(2, 43)$ (Lee et al., 2007).

Analogous to the transformation from t-scores to z-scores in the previous examples, the F-statistics were converted to $\chi^2(2)$-scores by a one-to-one quantile transformation. Fig. 5a shows a histogram of the $N=74{,}381$ $\chi^2(2)$-scores in the entire brain. The bin width is again $\Delta=0.05$. The method described in Appendix A.2 was applied within the interval [0,3], where 3 is the 78% quantile of the $\chi^2(2)$ distribution. The empirical null obtained was a scaled χ^2 with $\hat{\nu}=1.98$ *d.f.*, scaling parameter $\hat{a}=1.16$ and corresponding null proportion $\hat{p}_0=0.95$. For comparison, a scaled theoretical null was also fitted within the same interval, yielding a null proportion $\hat{p}_0=0.89$. The empirical null fits the data better than the scaled theoretical null and hints at less active voxels, 5% vs. 11%.

The effect of the empirical null is better seen in the FDR plot in Fig. 5b. Setting the FDR level at 0.1 results in the significantly higher threshold 13.66 for the empirical null vs. 10.79 for the scaled theoretical null, and therefore substantially less active

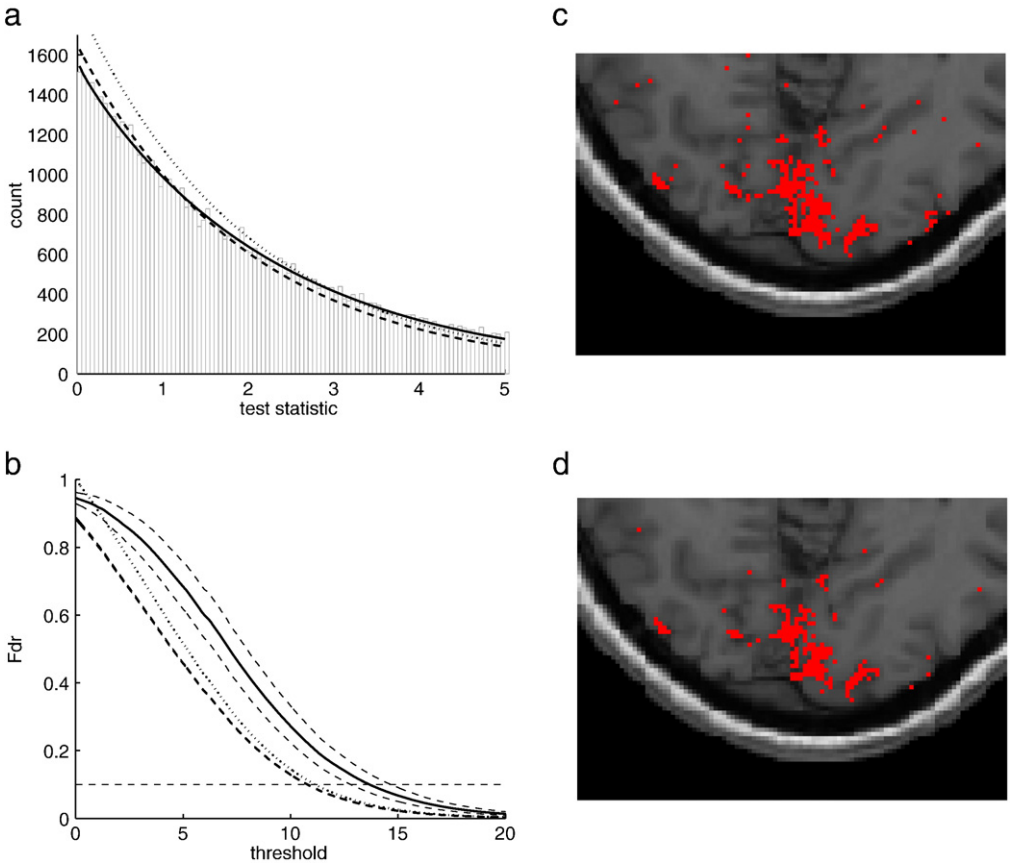

Fig. 5. Complex SSFP fMRI example: (a) Histogram of the $N=74{,}381$ $\chi^2(2)$-statistics (gray); theoretical null $\chi^2(2)$ (dotted); scaled theoretical null (thick dashed); empirical null (thick solid). Confidence bands omitted for clarity. (b) Tail FDR curves: theoretical null (thick dotted); scaled theoretical null (thick dashed); empirical null (thick solid). Confidence bands are 95% point-wise (thin dashed). Horizontal dashed line is the chosen FDR level 0.1. (c) Relevant significant voxels at FDR level 0.1 according to the scaled theoretical null. (d) Significant voxels at FDR level 0.1 according to the empirical null; same slice as in (c).

voxels, 1934 vs. 2994. The activation maps in Figs. 5c and d show that, in contrast to the theoretical null, the empirical null avoids spurious active voxels in regions far from the visual cortex. Another way to see this is that if the threshold was set arbitrarily to 13.66 corresponding to the activation pattern in Fig. 5c, the empirical null curve in Fig. 5b would indicate that the estimated FDR for that activation pattern is not 0.1 but 0.22±0.02.

Example 4: DTI

The fourth example shows an effect opposite to that of Example 3. The data for this example is a DTI subset from a study that investigated anatomical characteristics related to reading development in children aged 7–13 (Deutsch et al., 2005). The study was aimed at finding anatomical differences between poor readers with a previous diagnosis of dyslexia and normal readers. A previous analysis of this dataset found differences in the orientation of the principal diffusion direction in certain regions of the white matter (Schwartzman et al., 2005, 2008). For the current analysis, the images were registered to the MNI template. The two groups of images, 6 in each group, were then compared voxel-wise at each of $N=20{,}931$ voxels corresponding to white matter regions common to all subjects (Schwartzman, 2006).

In this analysis, the test is a multivariate test of differences in eigenvectors between group means of diffusion tensors (Schwartzman, 2006). The theoretical model in this case specifies an $F(3, 60)$ null distribution for the test statistics. The F-statistics were converted to $\chi^2(3)$-scores by a one-to-one quantile transformation. Fig. 6a shows a histogram of the 20,931 $\chi^2(3)$-scores. The bin width is again $\Delta=0.05$. The empirical null was obtained using the method described in Appendix A.2 within the interval [0,4.5], where 4.5 is the 79% quantile of the $\chi^2(3)$ distribution, yielding a scaled χ^2 with $\hat{\nu}=2.94$ *d.f.*, scaling parameter $\hat{a}=0.87$ and corresponding null proportion $\hat{p}_0=0.924$. For comparison, a scaled theoretical null was also fitted within the same interval, yielding a null proportion $\hat{p}_0=0.992$. The empirical null fits the data better than the scaled theoretical null and suggests more significant voxels, 7.6% vs. 0.7%.

The effect of the empirical null is better seen in the FDR plot in Fig. 6b. Setting the FDR level at 0.1 results in the significantly lower threshold 12.24 for the empirical null vs. 16.39 for the scaled theoretical null, and therefore substantially more active voxels, 511 vs. 196. The thresholded maps are shown in Figs. 6c and d.

Discussion

In this article, we have presented an empirical approach to FDR inference of neuroimaging data. The most important aspect of this approach is the estimation of the null density from the data itself. We have shown using four real examples that the use of an

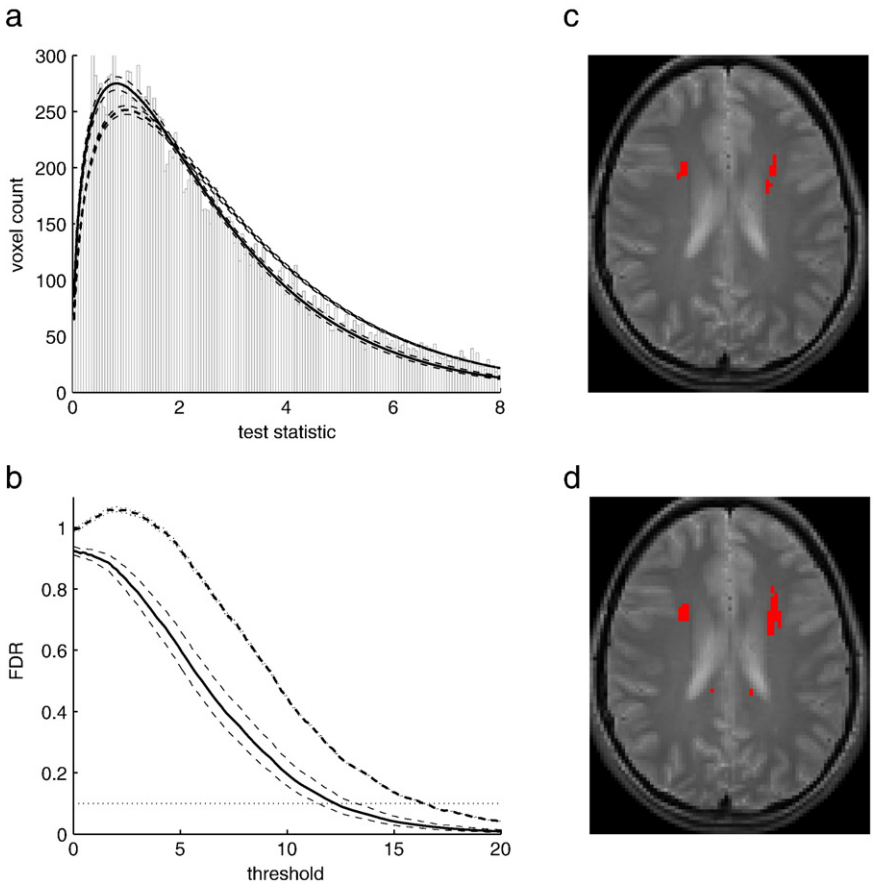

Fig. 6. DTI eigenvector frame example: (a) Histogram of the $N=20{,}931$ $\chi^2(3)$-statistics (gray); scaled theoretical null (thick dashed); empirical null (thick solid). Confidence bands are 95% point-wise (thin dashed). (b) Tail FDR curves: scaled theoretical null (thick dashed); empirical null (thick solid). Confidence bands are 95% point-wise (thin dashed). Horizontal dashed line is the chosen FDR level 0.1. (c) Significant voxels at FDR level 0.1 according to the scaled theoretical null. The slice corresponds to $z=26$ mm in MNI coordinates. (d) Significant voxels at FDR level 0.1 according to the empirical null; same slice as in (c).

empirical null is relevant and necessary for correct inference in neuroimaging studies. Depending on the case, the empirical null can produce more or less active voxels than the theoretical null at the same FDR level. Yet, the empirical null methodology is set up so that if the investigator still wishes to believe the theoretical null, the null proportion p_0 may still be estimated by itself using a special case of the same methodology. While being frequentist in nature, the empirical null also corresponds to an empirical Bayes interpretation of FDR.

The empirical null has been presented for z, t, χ^2 and F-statistics, all of which commonly occur in neuroimaging. In this paper we have shown applications of z and t-statistics in BOLD fMRI, and χ^2 and F-statistics in SSFP fMRI and DTI, but the methodology can be used in other applications as well. For example, t-tests may appear in the comparison of two conditions using scalar summaries of the diffusion tensor in DTI. F-tests may appear in the analysis of complex BOLD fMRI or in ANOVA analyses of several conditions in standard BOLD fMRI. In these cases, once an SPM is produced whose theoretical null distribution is one of the above four, then the analysis continues as described above in Empirical null.

The reasons for the mismatch between the theoretical null and the empirical null in the examples presented in this paper are unclear. Efron (2005, 2007) suggests one possible reason may be the presence of unobserved covariates, as these are all observational studies. The main message is that the theoretical null produces biased results, sometimes towards spurious findings (simulated example and examples 3 and 5) and sometimes towards hindered discovery (examples 4 and 6). Either bias would be invisible in a single test situation, but their existence is revealed by multiple testing. This is less striking in the χ^2 examples but certainly present there. The empirical null provides a way of accounting for that bias.

Our preference of FDR over FWER as a measure of error control is supported by Logan and Rowe (2004) and Marchini and Presanis (2004), who found that FDR controlling methods generally have higher power than FWER-controlling methods to detect active voxels. An important conclusion of Logan and Rowe (2004) is that incorporating voxel-wise correlation information does not appear to be crucially important to voxel-wise thresholding rules. We have shown by simulation that this is also true when the empirical null is used in conjunction with FDR. This claim also supports the interpretation of the FDR as a posterior probability in practice, even though this equivalence has only been shown mathematically to hold when the test statistics are independent.

It should be emphasized that, although the empirical null is well suited for FDR inference, it is really a statement about the nature of the null distribution and does not depend on the inference method that is used later for thresholding the SPM. In fact, the empirical null can also be used with FWER inference. If the empirical null is $N(\mu, \sigma^2)$, then any FWER method can be applied to the SPM of normalized z-scores $T^* = (T-\mu)/\sigma$. Similarly, if the empirical null is $\chi^2(\nu)$ with scaling factor a, then FWER methods for $\chi^2(\nu)$ distributions may be applied to the SPM with rescaled scores $T^* = T/a$.

An advantage of the empirical null methodology over the mixture fitting (Everitt and Bullmore, 1999; Hartvig and Jensen, 2000; Woolrich et al., 2005) is that the empirical null does not require the specification of an alternative density, as long as it may be assumed that the number of active voxels is relatively small and that test statistics from active voxels tend to have values far from zero. This is particularly useful when the active voxels are themselves governed by a mixture of distributions and are hard to describe by a single density. These assumptions allow the identification of the bulk of the test statistic histogram with the null. If these assumptions do not hold then a more specific mixture model may be necessary.

In single testing situations, it is customary to threshold at standard values such as 0.05 or 0.01. In multiple testing situations, the same thresholding values are common in FWER control. There are no such standards yet for FDR control. Efron (2004) has proposed 0.2 as a reasonable tolerable FDR level. In the examples analyzed in this paper, we have used 0.1. Given the Bayesian interpretation of FDR, thresholding the posterior probability of activation $P[i \in S_A | T_i > u]$ at 0.95 (Friston et al., 2002) would correspond to inference at a right-tail FDR level of $P[i \in S_0 | T_i > u] = 1 - 0.95 = 0.05$, which may be seen as unnecessarily conservative. On the other hand, the criterion of thresholding the posterior probability of activation $P[i \in S_A | T_i = t]$ at 0.5 (Everitt and Bullmore, 1999; Hartvig and Jensen, 2000; Woolrich et al., 2005) would be equivalent to inference at a local FDR level of $P[i \in S_0 | T_i = t] = 1 - 0.5 = 0.5$, which is excessively permissive.

The Bayesian interpretation of FDR provides a new meeting point between Bayesian and frequentist statistics. Friston et al. (2002) has described FWER inference as having uniform specificity regardless of effect size. In contrast, he has interpreted Bayesian inference as adjusting the FWER threshold according to the effect size, while keeping a uniform confidence in the results. This adaptability is similar to that exhibited by FDR, which controls the relative error rate as opposed to the absolute error rate. Since specificity is a frequentist concept, Friston et al. (2002) has suggested that reporting of the posterior probability of activation at each voxel should suffice and no thresholding should be necessary. This view would correspond to reporting the estimated FDR at each voxel as opposed to controlling it at a particular level. Ultimately, the appropriate inference practice in neuroimaging may become clearer as more experience is gained with experimental data and the answers they provide about the brain.

Matlab software for implementing the tools described in this article are available online at http://biowww.dfci.harvard.edu/~armin/software.html.

Acknowledgments

The authors thank Bradley Efron for his valuable guidance. This research was supported in part by a William R. and Sara Hart Kimball Stanford Graduate Fellowship, the Schwab Foundation for Learning, NIH grant EY-015000, NIMH Training Grant MH15157-20 (DGG) and NSF grant DMS-0405970.

Appendix A. Empirical null calculations

Appendix A.1. Normal case

When the theoretical null is $N(0, 1)$, an appropriate empirical null is $N(\mu, \sigma^2)$ given by

$$f_0(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

where we have used z for z-scores for clarity of notation. Replacing $f_0(z)$ in Expression (2), we have

$$\log\tilde{f}_0(z) = b_0 + b_1 z + b_2 z^2 \tag{12}$$

where

$$b_0 = \log(N\Delta) + \log p_0 - \log\sqrt{2\pi\sigma^2} - \frac{\mu^2}{2\sigma^2}, \tag{13}$$

$$b_1 = \frac{\mu}{\sigma^2}, \quad b_2 = -\frac{1}{2\sigma^2}.$$

The coefficients b_0, b_1 and b_2 are found by Poisson (or OLS) regression using the histogram log-counts as the response and 1, z and z^2 as predictors (with values given by the center of the histogram bins). To ensure capturing the null, the fit is restricted to an interval around the central peak of the histogram. The interval may be fixed a priori or data-dependent, say between the 20th and 80th percentiles of the data. The parameters p_0, μ and σ are then recovered from $\widehat{b}_0$, $\widehat{b}_1$ and $\widehat{b}_2$ by inverting Expression (13):

$$\widehat{\sigma}^2 = -\frac{1}{2\widehat{b}_2}, \quad \widehat{\mu} = \widehat{b}_1\widehat{\sigma}^2,$$

$$\log\widehat{p}_0 = \widehat{b}_0 - \log(N\Delta) + \log\sqrt{2\pi\widehat{\sigma}^2} + \frac{\widehat{\mu}^2}{2\widehat{\sigma}^2}$$

Notice that the estimate $\hat{p}_0$ is not restricted to the interval [0,1].

In the case that we believe the theoretical null $N(0, 1)$, model (12) becomes

$$\log\tilde{f}_0(z) = b_0 - \frac{z^2}{2}, \quad b_0 = \log(N\Delta) + \log p_0 - \log\sqrt{2\pi}.$$

The regression this time fits only the constant b_0, using $-z^2/2$ as an offset. The parameter p_0 is recovered by

$$\log\widehat{p}_0 = \widehat{b}_0 - \log(N\Delta) + \log\sqrt{2\pi}.$$

Appendix A.2. χ^2 case

When the theoretical null is χ^2 with ν_0 *d.f.*, an appropriate empirical null is a scaled χ^2 (i.e. gamma) with ν *d.f.* possibly different from ν_0 and a scaling parameter a possibly different from 1. The empirical null density is

$$f_0(t) = \frac{1}{(2a)^{\nu/2}\Gamma(\nu/2)} e^{-t/(2a)} t^{\nu/2-1}.$$

Replacing in Expression (2), we get

$$\log\tilde{f}_0(t) = b_0 + b_1 t + b_2 \log t \tag{14}$$

where

$$b_0 = \log(N\Delta) + \log p_0 - \log\left((2a)^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)\right) \tag{15}$$

$$b_1 = -\frac{1}{2a}, \quad b_2 = \frac{\nu}{2} - 1.$$

The coefficients b_0, b_1 and b_2 are found by Poisson (or OLS) regression using the histogram log-counts as the response and 1, t and $\log t$ as predictors (with values given by the center of the histogram bins). To ensure capturing the null, the fit is restricted to an interval around the central peak of the histogram from 0 up to a fixed number or a data-dependent number, e.g. the 80th percentile of the data. The parameters p_0, ν and a are then recovered from $\widehat{b}_0$, $\widehat{b}_1$ and $\widehat{b}_2$ by inverting Expression (15):

$$\widehat{\nu} = 2\left(\widehat{b}_2 + 1\right), \quad \widehat{a} = -\frac{1}{2b_1},$$

$$\log\widehat{p}_0 = \widehat{b}_0 - \log(N\Delta) + \log\left((2\widehat{a})^{\widehat{\nu}/2}\Gamma\left(\frac{\widehat{\nu}}{2}\right)\right).$$

Notice that the estimate $\widehat{p}_0$ is not restricted to the interval [0,1].

In the case that we believe the theoretical null $\chi^2(\nu_0)$, model (14) becomes

$$\log\tilde{f}_0(t) = b_0 - \frac{t}{2} - \left(\frac{\nu_0}{2} - 1\right)\log t,$$

$$b_0 = \log(N\Delta) + \log p_0 - \log\left(2^{\nu_0/2}\Gamma(\nu_0/2)\right).$$

The regression this time fits only the constant b_0, using $-t/2-(\nu_0/2-1)\log t$ as an offset. The parameter p_0 is recovered by

$$\log\widehat{p}_0 = \widehat{b}_0 - \log(N\Delta) + \log\left(2^{\nu_0/2}\Gamma\left(\frac{\nu_0}{2}\right)\right).$$

References

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

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188.

Deutsch, G.K., Dougherty, R.F., Bammer, R., Siok, W.T., Gabrieli, J.D., Wandell, B., 2005. Correlations between white matter microstructure and reading performance in children. Cortex 41 (3), 354–363.

Dudoit, S., Shaffer, J.P., Boldrick, J.C., 2003. Multiple hypothesis testing in microarray experiments. Stat. Sci. 18 (1), 71–103.

Efron, B., 2004. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J. Am. Stat. Assoc. 99 (465), 96–104.

Efron, B., 2005. Bayesians, frequentists and scientists. J. Am. Stat. Assoc. 100 (469), 1–5.

Efron, B., 2007. Size, power and false discovery rates. Ann. Stat. 35 (4), 1351–1377.

Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc. 96 (456), 1151–1160.

Everitt, B.S., Bullmore, E.T., 1999. Mixture model mapping of brain activation in functional magnetic resonance images. Human Brain Mapping 7, 1–14.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2 (4), 189–210.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging. NeuroImage 16, 465–483.

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

Ghahremani, D., Taylor, J.E., 2005. Empirical and theoretical false discovery rate analyses for fMRI data. Poster, Organization for Human Brain Mapping. (June).

Hartvig, N.V., Jensen, J.L., 2000. Spatial mixture modeling of fMRI data. Hum. Brain Mapp. 11, 233–248.

Hazeltine, E., Poldrack, R., Gabrieli, J.D., 2000. Neural activation during response competition. J. Cogn. Neurosci. 12, 118–129.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., Benjamini, Y., 2007. Cluster-based analysis of fMRI data. NeuroImage 33, 599–608.

Lee, J., Shahram, M., Schwartzman, A., Pauly, J.M., 2007. A complex data analysis in high-resolution SSFP fMRI. Magn. Reson. Med. 57, 905–917.

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

Marchini, J., Presanis, A., 2004. Comparing methods of analyzing fMRI statistical parametric maps. NeuroImage 22, 1203–1213.

Nichols, T.E., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res. 12, 419–446.

Schwartzman, A., 2006. Random ellipsoids and false discovery rates: statistics for diffusion tensor imaging data. Ph.D. thesis, Stanford University.

Schwartzman, A., 2008. Empirical null and false discovery rate inference for exponential families. Harvard University Biostatistics Working Paper Series. http://www.bepress.com/harvardbiostat/paper77.

Schwartzman, A., Dougherty, R.F., Taylor, J.E., 2005. Cross-subject comparison of principal diffusion direction maps. Magn. Reson. Med. 53, 1423–1431.

Schwartzman, A., Dougherty, R.F., Taylor, J.E., 2008. False discovery rate analysis of brain diffusion direction maps. Ann. Appl. Statist. 2, 153–175.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., Luca, M.D., Drobnjak, I., Flitney, D.E., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23 (Suppl. 1), S208–S219.

Storey, J.D., 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Statist. 31 (6), 2013–2035.

Storey, J.D., Taylor, J.E., Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J. R. Stat. Soc., B 66 (1), 187–205.

Woolrich, M.W., Behrens, T.E., Beckmann, C.F., Smith, S.M., 2005. Mixture models with adaptive spatial regularisation for segmentation with an application to fMRI data. IEEE Trans. Med. Imag. 24 (1), 1–11.

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. Hum. Brain Mapp. 4, 58–73.

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

Yekutieli, D., Benjamini, Y., 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J. Stat. Plan. Inference 82, 171–196.