NeuroImage 123 (2015) 102–113

Contents lists available at ScienceDirect

NeuroImage

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

Simultaneous control of error rates in fMRI data analysis

Hakmook Kang [a,b,*], Jeffrey Blume [a], Hernando Ombao [c], David Badre [d]

[a] Department of Biostatistics, Vanderbilt University, Nashville, TN 37203, USA
[b] Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37232, USA
[c] Department of Statistics, University of California at Irvine, Irvine, CA 92697, USA
[d] Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, RI 02912, USA

ARTICLE INFO

Article history:
Received 10 March 2015
Accepted 4 August 2015
Available online 10 August 2015

Keywords:
Multiple comparison
Likelihood paradigm
Likelihood ratio
Functional magnetic resonance imaging

ABSTRACT

The key idea of statistical hypothesis testing is to fix, and thereby control, the Type I error (false positive) rate across samples of any size. Multiple comparisons inflate the global (family-wise) Type I error rate and the traditional solution to maintaining control of the error rate is to increase the local (comparison-wise) Type II error (false negative) rates. However, in the analysis of human brain imaging data, the number of comparisons is so large that this solution breaks down: the local Type II error rate ends up being so large that scientifically meaningful analysis is precluded. Here we propose a novel solution to this problem: allow the Type I error rate to converge to zero along with the Type II error rate. It works because when the Type I error rate per comparison is very small, the accumulation (or *global*) Type I error rate is also small. This solution is achieved by employing the likelihood paradigm, which uses likelihood ratios to measure the strength of evidence on a voxel-by-voxel basis. In this paper, we provide theoretical and empirical justification for a likelihood approach to the analysis of human brain imaging data. In addition, we present extensive simulations that show the likelihood approach is viable, leading to "cleaner"-looking brain maps and operational superiority (lower average error rate). Finally, we include a case study on cognitive control related activation in the prefrontal cortex of the human brain.

Introduction

Functional magnetic resonance images (fMRIs) are typically used to identify activated regions of the brain. After image preprocessing (i.e., motion correction, co-registration, normalization, and spatial smoothing), the standard statistical approach is to use a voxel-level general linear regression model to characterize background neural activity and yield t-statistics for testing hypotheses of activation. Voxels are then classified as "activated" or "not activated" based on their t-statistics reaching some threshold. This results in a statistical map of brain activation. However, the voxel-wise t-threshold must be chosen carefully because the number of voxels is large and this results in family-wise Type I error (false positive) rate inflation. Simply increasing the threshold for significance/positivity across all voxels in a uniform manner controls the inflation, but it also increases Type II error (false negative) rate significantly.

Standard approaches to controlling this inflation are varied. Either they rely on finding an acceptable trade-off between Type II and Type I errors or they focus on controlling the false discovery rate (FDR) instead (Benjamini and Hochberg, 1995). Bonferroni corrections and the random field theory (RFT) (Worsley et al., 1992) are examples of the former: they control the family-wise error rate (FWER) at a predetermined level and are the most commonly used approaches in fMRI data analysis. The alternative approach of controlling FDR has been recently considered (Storey, 2002; Schwartzman et al., 2009). Even though controlling FDR is "less conservative" than controlling FWER, the classical FDR (Benjamini and Hochberg, 1995) and modified FDR proposed by (Schwartzman et al. (2009) methods still yield significant false negative findings. (Friston and Penny (2003) note that Bayesian approaches offer an alternative approach in which statistical inference relies on the Bayes factor (BF), proportional to posterior odds ratio between two competing hypotheses. The posterior odds of the alternative hypothesis is essentially a weighted average of the likelihood function over the parameter space weighted by the prior distribution. Consequently, the choice of the prior heavily influences the BF, one common criticism of the Bayes factor.

More importantly, these approaches do not address an obvious scientific challenge: as sample size approaches infinity, the Type I error rate remains fixed while the Type II error rate is indistinguishable from zero. As a result, even with myriad images collected, no brain activation map constructed from these approaches is ever completely free of false positive findings. It is tempting to dismiss this concern because no sample size is ever large enough to ensure perfect information (i.e. no sample size is equivalent to infinitely many images). But this reasoning is flawed: if the method is still wrong ($\alpha \times 100$) % of the time with infinite information (as are hypothesis testing methods under the null hypothesis), then they certainly can do no better with less information. In fact, they can do not worse either. The sample size is irrelevant when

* Corresponding author at: Vanderbilt University, Department of Biostatistics, 2525 West End Avenue. Suite 1100, Nashville, TN 37203, USA.
E-mail address: hakmook.kang@vanderbilt.edu (H. Kang).

http://dx.doi.org/10.1016/j.neuroimage.2015.08.009
1053-8119/

it comes to controlling Type I error rate. This is both the strength and the weakness of the traditional approach.

In this paper, we propose a novel alternative approach that is rooted in the likelihood paradigm. Instead of using the tail area probability (or p-value) as a *measure of the strength of evidence*, we use a likelihood ratio. We use the tail area probability as the *measure of how often misleading evidence* will be observed. Voxel-specific likelihood ratios are derived from an appropriately specified spatio-temporal model. They measure the strength of statistical evidence in the data about the level of activation in a voxel. We examine the long-run behavior of likelihood ratios in fMRI analyses and we show how to "control" it. For example, direct interpretation of likelihood ratios can be shown to minimize the average of the Type I and Type II error rates. In addition, the likelihood analog of the Type I error rate converges to zero as the sample size increases. Hence, the accumulation of (many) small Type I errors, i.e. the family-wise error rate, remains small and controllable even when the number of tests or comparisons is large.

The Likelihood paradigm offers distinct advantages over classical inferential tools and modern ones based on false discovery rates:

(1) The likelihood ratio serves only as the reportable measure of the strength of observed evidence; it is not adjusted for the number of simultaneous comparisons.
(2) Likelihood analogs of the false positive and false negative error rates are *both* controlled and they *both* converge to zero as the statistical information (e.g., the number of images) increases. This, in turn, drives both false discovery rates to zero.
(3) The *global* or *overall* error rates, analogous to family-wise error rates, also converges to zero despite inflation from simultaneous comparisons.

Points (2) and (3) above are distinct improvement over current applications of BF methodology. Also, the likelihood approach is free from the *a priori* information that is required in Bayesian approaches. More precisely, the interpretation of the data by the likelihood function holds for any prior and would be consistent in that sense with any Bayesian approach. An immediate consequence of these advantages is that likelihood depictions of brain activation are always correct in the limit and more often correct, on average, in finite samples.

For a concrete example, see Fig. 1, which is a representative realization from brain image simulation we describe later (Section 3). It is obvious from these images that the likelihood depiction (e) is the closest to the true activation image (a), while other approaches show moderate to severe deviation from the true image. This figure displays what we found to be true most often: the likelihood approach provides a visually appealing trade-off between Type I and II error rates while allowing both to virtually vanish as the statistical information increases.

To precisely detail the operational characteristics of the likelihood approach, some background material is required (Section 2). We derive the likelihood analogs of the Type I and II error rates in Section 2.1 and apply these methods in the context of simultaneous comparisons (Section 2.2). In Section 3, we empirically validate the claimed advantages and compare our method with other approaches, i.e., RFT, FDR, and BF. We also analyze real fMRI data in Section 4 with the likelihood approach. The two terms *voxel* and *pixel* are used interchangeably in this paper, particularly in Section 3 in which all the simulation studies were done in 2-dimensional images.

Likelihood paradigm

The law of likelihood (see Appendix A) explains how data represent statistical evidence for one hypothesis over another (Hacking, 1965; Royall, 1997). Specifically, data favor the hypothesis that better predicts the observed data or gives higher likelihood. The likelihood ratio measures the degree to which one hypothesis is better supported over another. Introductory material on the law of likelihood and recent applications and advances are available in the literature (e.g., Blume, 2002; Blume et al., 2007; Blume, 2011a; Choi et al., 2008; Royall and Tsou, 2003; Wang and Blume, 2011).

To fix ideas, consider independent samples $x_1, \ldots, x_n$ distributed as $f(X;\theta)$. The likelihood function is $L(\theta) \propto \prod_{i=1}^{n} f(x_i;\theta)$ where x_i represent observed data and θ the parameter of interest. Then, for any two competing hypotheses, say $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, the strength of the evidence for H_1 over H_0 is measured by the likelihood ratio $LR = L(\theta_1)/L(\theta_0)$ (Royall, 1997). A $LR = 1$ indicates neutral evidence. As the sample size grows, the likelihood ratio will converge to either 0 or ∞ in support of the true hypothesis (Royall, 1997; Blume, 2002). Upon collecting observations $x_1, \ldots, x_n$, the observed likelihood ratio will fall into one of three general regions for $k \geq 1$: small $LRs \in [0, 1/k]$ that indicate strong evidence for H_0 over H_1, midrange $LRs \in (1/k, k)$ that indicate weak (or inconclusive) evidence, and large $LRs \in [k, \infty)$ that indicate strong evidence for H_1 over H_0, which is illustrated in logarithmic scale in Fig. 2.

Conventional benchmarks of $k = 8, 32$ are points of reference along the gradual shift from weak ($k \leq 8$) to moderate ($8 < k < 32$) to strong evidence ($32 \leq k$) (see Appendix E). This is similar to established benchmarks for Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995). Notice that FWER or FDR methods also use benchmarks that depend on the sample size and initial specification of the alpha-level. Justifications for these benchmarks are well established elsewhere (Royall, 1997; Blume, 2002). The point is that likelihood ratios are fundamentally a descriptive tool. These benchmarks are not cuttoffs but rather guideposts. Sometimes, we are forced to collapse the continuous scale to a dichotomized one, such as when indicating when there is sufficient evidence that a voxel is activated or not. To do this, we will use a $k = 20$. It controls the probability of observing misleading evidence at $1/20 (= 0.05)$ under very general conditions, thus providing a familiar analog for strong control of the Type I error rate (Royall, 1997; Blume, 2002). Hence comparisons between methods that control the Type I error rate (or its analog) at the same rate are fairer. Importantly, the actual probability of observing misleading evidence is often much less than

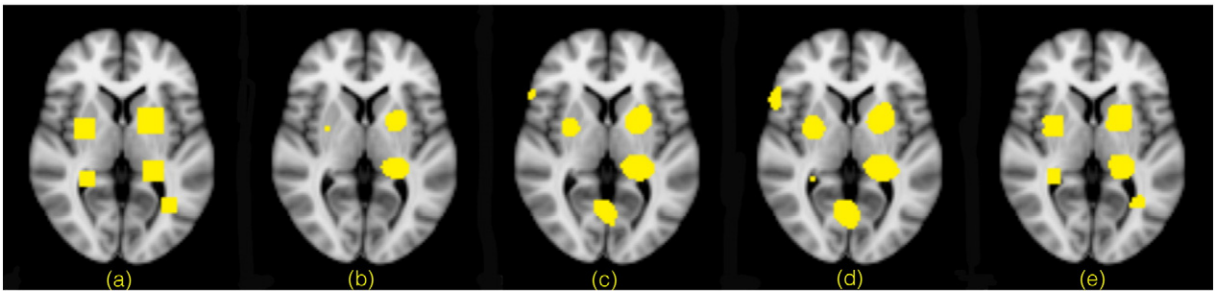

Fig. 1. Simulated voxel analysis for activation brain map. Data were simulated with two boxcar external stimuli, the spatial dimension of 91×109, and the temporal dimension of $T = 128$. The figures derived as follows: (a) truth (five truly active regions), (b) analysis using RFT, (c) analysis controlling FDR, (d) analysis using BF, (e) analysis using likelihood approach.

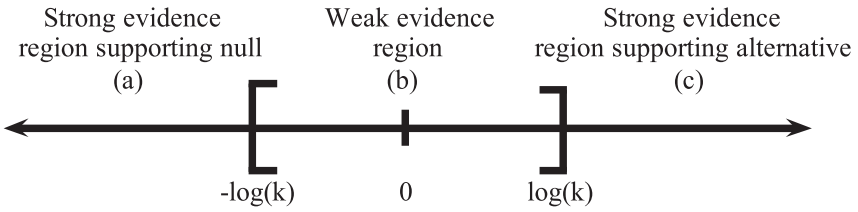

Fig. 2. Three evidence regions on logarithmic scale: (a) strong evidence for H_0, (b) weak evidence, and (c) strong evidence for H_1.

its universal bound, $1/k$, but the decrease is a function of the underlying model.

Evidential quantities and misleading evidence

The likelihood framework carefully distinguishes among three evidential quantities (Blume, 2011b):

(1) A measure of the strength of evidence: "How strong is the statistical evidence for a particular voxel that is claimed as activated?"
(2) The probability that a particular study design will generate misleading evidence: "What is the chance that the voxel-level evidence will incorrectly favor activation when it is truly inactivated?"
(3) The probability that observed data are misleading: "For the data supporting activation in some voxels, what is the chance that these observed data are misleading?"

Table 1, replicated from Blume (2011b), displays these evidential quantities and their mathematical form.

The first quantity (LR) communicates the strength with which the data support one hypothesis over another—it is the researcher's essential tool for understanding what the data say. The second quantities are the analog to the Type I (α) and Type II (β) error rates of hypothesis testing. They are used only for study design purposes; they play no role in the interpretation of data as evidence. In this paper, the two terms mis_0 and mis_1 refer to the "would-be" or analogous Type I and Type II error rates for the likelihood paradigm. It is important to note that the classical Type I and Type II error rates are conceptually similar to the probabilities of misleading evidence, but their differences are subtle, meaningful, and critical to our application. Both are used to assess the performance characteristics of the underlying approaches. But the approaches themselves have different goals (i.e., to make a decision vs. reporting the strength of evidence in favor of the hypotheses). The third quantity describes how likely it is that an observed result is misleading. It depends on prior information about the hypothesis of interest but is ultimately driven by the observed likelihood ratio (Blume, 2011b). The third quantity is closely related to the posterior probability of the hypothesis $P(H_i|data)$ for $i \in \{0, 1\}$, which is a type of false discovery rate. Failure to distinguish between these three evidential quantities leads to confusion about multiple comparisons (Blume and Peipert, 2004; Blume, 2011b).

Misleading evidence is defined as strong evidence in favor of the incorrect hypothesis, i.e., observing $LR > k$ when H_0 is true or $LR < 1/k$ when H_1 is true. The probabilities of observing misleading evidence, $mis_0 = P(LR > k|H_0)$ and $mis_1 = P(LR < 1/k|H_1)$, are both bounded by $1/k$ and their average is bounded by $1/(k+1)$ (Royall, 1997; Blume, 2002). The probability is often much less than the universal bound because $mis_i \to 0$ for $i \in \{0, 1\}$ as $n \to \infty$.

This likelihood approach also has a genuine frequentist justification. Instead of fixing one error rate and minimizing the other, the law of likelihood minimizes the average error rate at a given value k. Let $\gamma(k) = (\alpha(k) + \beta(k))/2$ be the average error rate with $\alpha(k) = P(LR > k|H_0)$ and $\beta(k) = P(LR < 1/k|H_1)$. Direct minimization of $\gamma(k)$ occurs at $k = 1$ for any sample size. A general weighted average of error rates $\omega\alpha(k) + (1-\omega)\beta(k)$ is minimized with $k = (\omega)/(1-\omega)$, where $0 \le \omega \le 1$. When we later use $k = 20$ as a sufficient level of evidence to indicate activation, we are implicitly setting weighting the Type I error rate 19-times more than the Type II error rate. Both error rates still shrink to zero, but this is slightly more conservative than using the natural $k = 1$ evidential marker.

Multiple comparisons

Multiple comparisons are handled differently in the likelihood context. Let θ denote the difference in the level of activation in a single voxel under two experimental conditions. Typically, $\theta = \beta_2 - \beta_1$ where the beta coefficients are corresponding to two distinct experimental conditions in an appropriately specified generalized linear model (see Section 3), but here it suffices to let $\hat{\theta} \sim N(\theta, \sigma^2/n)$ where σ^2 is known and n is the number of images collected (note that in conventional imaging analysis, n is often represented in time T). Replacing the unknown variance with a consistent estimate does not change the nature of these computations. Without loss of generality, let the hypotheses of no differential activation be $H_0 : \theta = 0$. Temporal correlation is ignored here, but its inclusion does not appreciably impact the example or the point we wish to illustrate.

For illustration, consider an alternative of interest of $H_1 : \theta = \sigma$, a differential level of activation equals to one standard deviation. For a single voxel, a standard hypothesis test with a one-sided Type I error rate of $\alpha = 0.05$ and a sample of 8 images has a Type II error rate of 0.118 (see Appendix C). With 4 independent voxels and a Bonferroni corrected Type I error rate of $\alpha = 0.0125 (= 0.05/4)$, the voxel-wise Type II error rate increases to 0.278. The family-wise Type I error rate

Table 1
Three evidential quantities, measure for each quantity, and its mathematical form.

Evidential quantity	What it measures	Name	Mathematical representation
1	Strength of the evidence	Likelihood ratio	LR
2	Propensity for study to yield misleading evidence	Probability of observing misleading evidence	$mis_0 = P(LR > k\|H_0)$ $mis_1 = P(LR < 1/k\|H_1)$
3	Propensity for observed results to be misleading	Probability that observed evidence is misleading	$P(H_0\|LR > k)$ $P(H_1\|LR < 1/k)$

is controlled at 0.05, but now the family-wise Type II error rate balloons to 0.728 ($=1-(1-0.278)^4$).

In comparison, consider a likelihood approach that uses $k = 8$ to indicate moderate evidence. The probability of observing moderate misleading evidence is (see Appendix D)

$$mis_0 = M(n,k) = \Phi(-\log k/\sqrt{n} - \sqrt{n}/2),$$

where Φ denotes the standard normal cumulative distribution function. By symmetry, $mis_0 = mis_1 = 0.016$. When $k = 20$, these probabilities drop to 0.007, illustrating that it is harder to observe stronger levels of misleading evidence. The probability of observing weak evidence $P(1/8 < LR < 8)$, under either hypothesis, is 0.233. The analogous family-wise likelihood error rates are both 0.062 ($=1-(1-0.016)^4$). The effective family-wise Type I error rate is inflated by 0.012, while the family-wise Type II error rate is reduced by 0.67.

Now, it could be argued that a direct comparison between likelihood and hypothesis testing is inherently biased in favor of likelihood because likelihood essentially ignores weak evidence while hypothesis testing treats weak evidence for the alternative as evidence for the null (not a typo; weak evidence is in the fail to reject region). Suppose then that we eliminate the weak evidence region by setting $k = 1$. Now both approaches have only two regions, making comparisons between them fairer. Of course, this means we would interpret any amount of evidence, no matter how fleeting, as strong enough evidence to be potentially misleading (we think this practice is scientifically questionable, but it provides a "fairer" computational benchmark despite philosophical inconsistencies). Not surprisingly, weak evidence is much less reliable than strong evidence and, as a consequence, the likelihood error rates increase from 0.016 to 0.078 and 0.062 to 0.280. The trade-off here is an increase of 0.23 in the family-wise Type I rate for a reduction of 0.45 in family-wise Type II rate, compared to the Bonferroni approach when the number of voxels is 4. This is displayed as the dichotomized likelihood paradigm, or DLP($k = 1$), in Table 2. Note that the average error, however, remains lower and the ability of our likelihood approach to distinguish between weak evidence and evidence in favor of the null hypothesis is a critical advance.

If instead we used $k = 8$ and disallowed weak evidence, then we would actually be labeling weak evidence for the alternative as evidence for the null hypothesis (i.e., this is DLP($k = 8$)). Problematic as this may sound, it is routinely done in hypothesis testing. This is what keep the Type I error rate from falling below α. Here the DLP($k = 8$) yields a family-wise Type I error rate of 0.016 for 1 voxel and 0.062 for 4 voxels; family-wise Type II error rate of 0.249 for 1 voxel and 0.682 for 4 voxels; and a family-wise average error rate of 0.133 for 1 voxel and 0.372 for 4 voxels. For completeness, we note that in the above illustration, the Bonferroni with 1 voxel approach is equivalent to DLP with $k = 1.92$, and with 4 voxels it is equivalent to $k = 10.37$.

Finally, the likelihood family-wise error rates can always be driven to zero with a sufficient large sample size:

$$\begin{aligned} \textit{global}\ \text{error rate} &= 1-(1-\text{ error rate per voxel })^m \\ &= 1-(1-mis_i)^m\ ,\ i=\{0,1\} \\ &\to 1-(1-0)^m \text{ as the number of images acquired} \to\infty \\ &= 0\ , \end{aligned}$$

where m is the number of voxels. Moreover, the probability of observing weak evidence also converges to zero as the sample size grows, so we are sure to be left with the strong evidence in the correct direction (Royall, 1997, 2000; Blume, 2002). With pre-determined *global* Type I error rate or false discovery rate, it would not be possible to achieve that the *global* Type I error rate converges to zero with a sufficient large sample size. In this paper, the amount of information (i.e., sample size) comes from either the number of subjects, the number of images collected, or both. In practice, it would be infeasible to have infinitely many subjects, but it would be feasible to collect a large number of images so that the asymptotic behavior of the LP approach can be warranted.

Spatio-temporal simulations

We validated our proposed approach with simulation studies based on spatially and temporally correlated time series with length T at each voxel. After fitting the appropriate voxel-specific general linear model, we aggregated results and estimated the Type I and II error rates for various approaches to controlling multiple comparisons: RFT, FDR, BF, LP, and DLP.

Data generation

To mimic the characteristics of functional magnetic resonance imaging data with two boxcar external stimuli, first we generated time series with different lengths, i.e., $T = 64$, 128, 256, and 320 using autoregressive model with order one (AR(1)), and the spatial dimension was 32×32 voxels. For the sake of simplicity and computational efficiency, we investigated only four different lengths of time series in a small 32×32 region even though $T = 64$ in such a small spatial dimension might not be common in practice. Let $Y(v,t)$ denote the response at a voxel v and time t. Then the model with $P = 2$, i.e., two conditions: active and rest, is

$$Y_v(t) = \sum_{p=1}^{P} X_p(t)\beta_v^p + \in_v(t), \tag{1}$$

where $X_p(t)$ is the convolution between the p^{th} stimulus impulse function (boxcar function for this simulation) and hemodynamic response function which is assumed to indirectly characterize the change of neural activity. We assume that the first stimulus is on during $[1, T/4 + 1]$ and $[2T/4, 3T/4 + 1]$ and the second is on otherwise, and $\in_v(t)$ follows AR(1) process with AR parameter of 0.4, $t = 1, \ldots, T$. For white noise component of AR(1) model, we assumed a Gaussian distribution with the mean of zero and the standard deviation of 1.5. The canonical hemodynamic response function from SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) was used to generate the design matrix $X_p(t)$ in (1). We assume that there are two square active blocks: one block consists of 64 voxels and the other of 36 voxels. The effect size as measured by $\beta^2 - \beta^1$ in (1) is one under the assumption that the hypothesis of interest is $H_0 : \beta^2 - \beta^1 = 0$ vs. $H_1 : \beta^2 - \beta^1 > 0$ at each voxel. Spatial correlation was imposed on the data via an exponential covariance function with the decaying parameter of 2 and variance of 2.5. For Fig. 1, the spatial dimension of 91×109 with $T = 128$ was used to create five active blocks consisting of a total of 300 voxels. For temporal correlation, a similar AR(1) process with the standard deviation of the white noise of 2.5 was employed for temporal correlation but Gaussian smoothing kernel with $\sigma = 1.5$ was applied to impose spatial

Table 2
Family-wise error rates for Bonferroni, likelihood, and dichotomized likelihood (DLP) approaches.

	Family-wise Type I error			Family-wise Type II error			Average family-wise error		
Voxels	Bonferroni	Likelihood	DLP(1)	Bonferroni	Likelihood	DLP(1)	Bonferroni	Likelihood	DLP(1)
1	0.050	0.016	0.078	0.278	0.016	0.078	0.164	0.016	0.078
4	0.050	0.062	0.280	0.728	0.062	0.280	0.340	0.062	0.280

correlation on the data instead of using the exponential covariance function. Contrast-to-noise ratio for the former simulation set is about 0.46 and that for the latter simulation related to Fig. 1 is about 0.34. All the simulation parameters used to generate Figs. 1 and 3 are summarized in Appendix E. To investigate the robustness of the Gaussian noise assumption, we generated data using a chi-square distribution with 3 degrees of freedom (something sufficient to be non-symmetric but not pathological), i.e., we set the white noise in the AR(1) process to be a chi-square distribution after removing the mean to have zero-mean random noise. Since a chi-square distribution with 3 degrees of freedom is known to be highly skewed, the robustness of each approach can be clearly assessed. Also, we examine the effect of different k values that indicate sufficient evidence for activation on the brain maps that are based on the likelihood paradigm when $T = 128$.

Results

Error rates with different T

To analyze the simulated data, we fit a common general linear model at each voxel while taking into account both spatial and temporal dependency in the data. For our approach (LP and DLP with $k = 20$), at each voxel, we also used the data from its immediately neighboring voxels to handle the underlying spatial correlation, while the temporal correlation was modeled as AR(1) process. For other approaches, we followed their standard protocol, i.e., employing spatial smoothing with FWHM = 8 and AR(1) process. We implemented the RFT procedure by precisely following Worsley (1994), FDR by following (Genovese et al. (2002), and BF by following Kass and Raftery (1995). For the sake of computational efficiency, Bayes factors were estimated via Laplace approximation instead of Markov chain Monte Carlo (MCMC) (Kass and Raftery, 1995). That is, our approach applies spatial smoothing on likelihood functions for each voxel and its immediately neighboring voxels, while the conventional approach applies spatial smoothing on the raw data and maximizes the likelihood function for the mean response at that voxel. For both cases, we used the iterative reweighted least squares (IRLS) to estimate the regression parameters and variance components. Then, we accounted for multiple comparisons by each method and computed the Type I and II error rates. We used $k = 20$ for the likelihood computations and we used the 95th percentile value (99th percentile value to generate Fig. 1) of the MLE of $\beta^2 - \beta^1$ as the value ($=\delta_1$) for alternative hypothesis: $H_0 : \beta^2 - \beta^1 = 0$ and $H_1 : \beta^2 - \beta^1 = \delta_1$. To approximate the BF at each voxel, a Gaussian distribution with the mean of zero and the variance of 100, $N(0,100)$, was used as a non-informative prior distribution for each beta coefficient. With this prior, we were able to avoid having an extremely conservative activation map, e.g., no positive findings with $N(0,10000)$. For the same reason, we classified each voxel as non-null if its log(BF) was greater than 0.5, which is generally less conservative than recommended, e.g., 3.0 ~ 5.0 (Kass and Raftery, 1995).

Comparing the two error rates across methods enables us to assess the performance of each method and the change in the performance as the sample size grows. Fig. 3 summarizes the results for 500 repetitions for each length of time series. In Fig. 3, the upper two plots (a) and (b) illustrate the behavior of the Type I and II error rates resulting from the five approaches, i.e., RFT (red), FDR (olive), BF (green), DLP (blue), and LP (purple) as the length of time series varies from $T = 64$ to $T = 320$. Controlling FWER maintains the *global* Type I error rate at or below the nominal error rate 0.05. However, because FDR and BF do not control the *global* Type I error rate, it can be very different from 0.05 as shown in Fig. 3(a). As T increases, it is evident that the Type I error rate corresponding to DLP decreases as explained in Section 2. By definition, the Type I error rate based on DLP should be

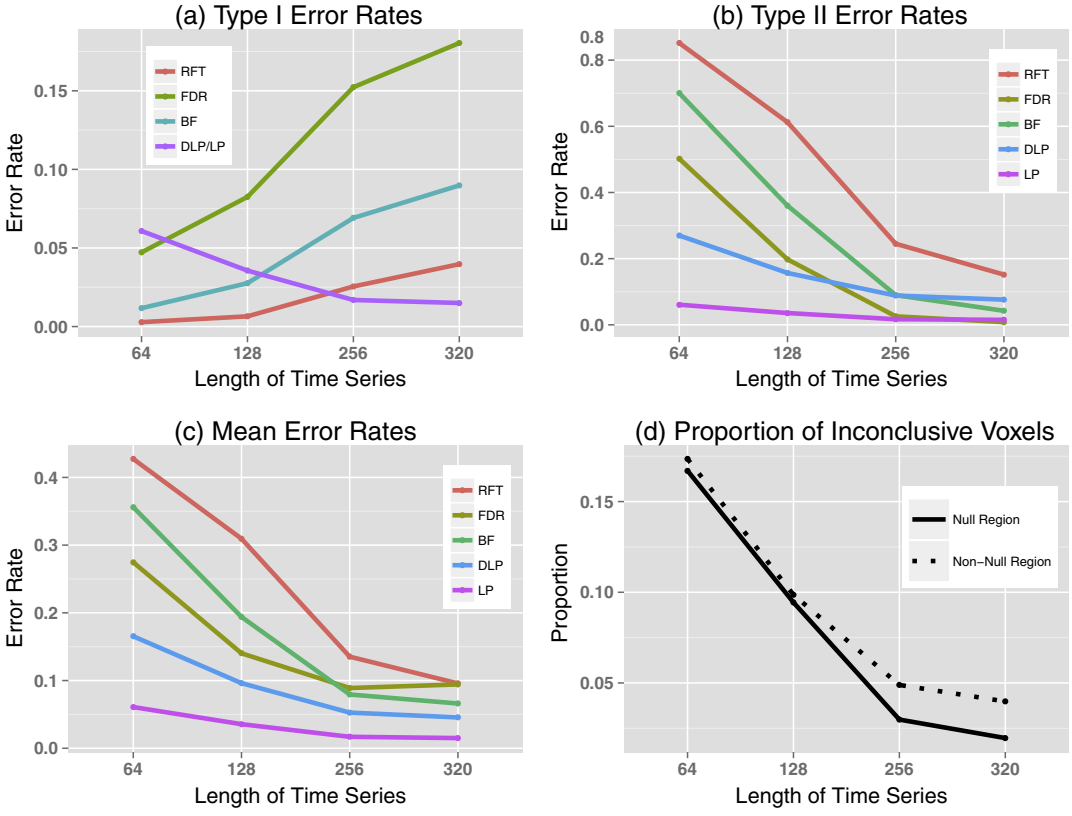

Fig. 3. Simulation results with different $T = 64, 128, 256, 320$ and $k = 20$ for LP and DLP. (a) Type I error rates (the average of the number of false positives divided by the total number of null voxels) (b) Type II error rates (the average of the number of false negatives divided by the total number of non-null voxels) (c) mean error rates for RFT (red), FDR (olive), BF(green), DLP (blue), and LP (purple) based on 500 iterations sampled from Gaussian distribution. (d) The proportion of inconclusive voxels in null (black solid) and non-null (black dotted) regions for LP at different length of time series.

the same as the one based on LP, illustrated as one purple line in Fig. 3(a), whereas the Type II and mean error rate based on DLP tends to be larger than those for LP, illustrated as two separate lines (blue and purple) in Fig. 3(b) and (c). It may seem counter-intuitive that the Type I error rate associated with RFT, FDR, and BF increase with the length of time series as shown in Fig. 3(a). However, this is a result of their extreme conservativeness in rejecting null hypotheses: when there is much less information, i.e., $T = 64$, the three methods are less likely to reject the null hypothesis, leading to very small Type I error rates (~0). With more information, i.e. $T = 128$, 256, and 320, the three methods tend to "find" more active voxels in error. The more they "find" active voxels, the more they can make mistakes.

Increases in sample size (or length of time series T) confer increased power, and a decrease in Type II error rates is shown in Fig. 3(b). Because the gain in power dominates the gain in Type I error rates, all mean error rates decline as T increases in Fig. 3(c). It is obvious that LP outperforms all the other methods and DLP outperforms RFT, FDR, and BF in terms of mean error rates. It is noteworthy that the difference in Type II error rates between DLP and LP decreases as T increases because DLP converges to LP as T grows, indicating that the number of voxels falling into the weak evidence region (or equivalent to inconclusive zone) in Fig. 2 sharply drops as T increases as shown in Fig. 3(d). The proportion of voxels classified as inconclusive in both null (solid black) and non-null (dotted black) regions diminishes from 16.7% and 17.35% when $T = 64$ to 2.0% and 4.0% when $T = 320$, respectively. The same estimation and inference tools, i.e., RFT, FDR, BF, and DLP, were used to analyze the data containing five active blocks illustrated in Fig. 1.

Error rates with different effect sizes and alternatives

The impact of the effect size $\beta^2 - \beta^1$ on error rates was investigated by assuming the effect was uniformly distributed [0.2, 1.5] with a mean set to 0.85. The error rates resulting from this simulation showed a pattern similar to that illustrated in Fig. 3, while varying the effect size tended to inflate both types of error rates in all the methods. We observed an increase of mean error rate of 0.048 for RFT, 0.037 for FDR, 0.036 for BF, 0.035 for DLP, and 0.002 for LP, on average across the four smoothing methods (described in Section 3.2.4) at $T = 128$. Therefore, our simulation results appear robust to the variation of the effect size across voxels as long as the effect sizes are not too small (not too near the null).

To illustrate the effect of choosing different alternative values on the performance of DLP, five percentile values, i.e., 65, 75, 85, 95, and 99, of MLE of $\beta^2 - \beta^1$ were chosen as alternative values. The corresponding mean error rates and proportions of voxels classified as inconclusive are summarized in Fig. 4. The mean error rates with respect to alternative values in DLP (purple line) are all smaller than the error rates resulted from three other approaches as shown in Fig. 4(a), although the performance of FDR is the same as DLP when the 99th percentile value of the MLE was used to choose an alternative value. We expected to achieve the lowest mean error rate when we set the true proportion of null voxels (about 90% in this simulation study) to be at the alternative value. In Fig. 4(b), as the percentile approaches the true proportion of the null voxels, the number of voxels classified as inconclusive in null (solid black) and non-null (dotted black) regions decreases. Although the proportion observed in non-null regions is slightly increased from 85th to 99th percentile, the overall proportion is deceasing due to a sharp decrease in the proportion in null regions. As seen in Fig. 4, it appears that a percentile between 90th and 99th of the MLE would work well for the alternative hypothesis compared to the other conventional methods if the true proportion of non-null voxels is not very high.

Error rates with chi-square distribution

Also, to assess the robustness of our approach, we followed the same estimating and inference procedures to analyze the second set of data generated using a chi-square distribution with 3 degrees of freedom. The overall pattern of Type I, II, and mean error rates illustrated in Fig. 5 looks similar to the corresponding plots in Fig. 3. However, at $T = 64$ and 128, RFT was only able to detect a small number of activated voxels, so the Type I error rates are close to zero while the Type II error rates are close to one. As shown in Fig. 5(c), even at $T = 320$, RFT

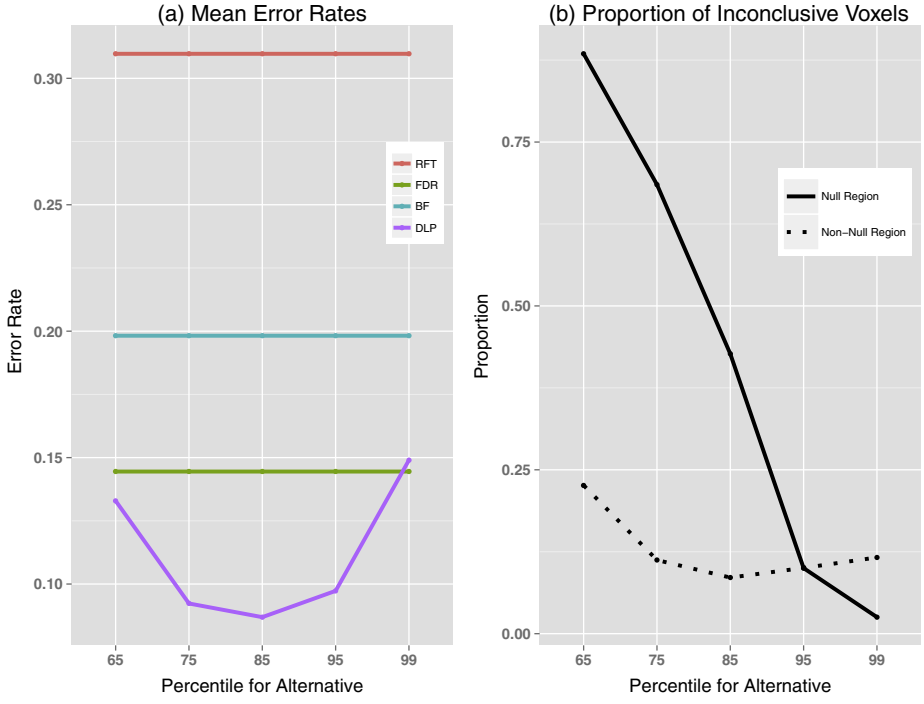

Fig. 4. Simulation results with 65th, 75th, 75th, 95th, and 99th percentiles in choosing an alternative value at $T = 128$ (a) mean error rates for RFT (red), FDR (green), BF (cyan), and DLP (purple) based on 500 iterations sampled from Gaussian distribution. (b) The proportion of inconclusive voxels in null (black solid) and non-null (black dotted) regions for LP at different alternative values.

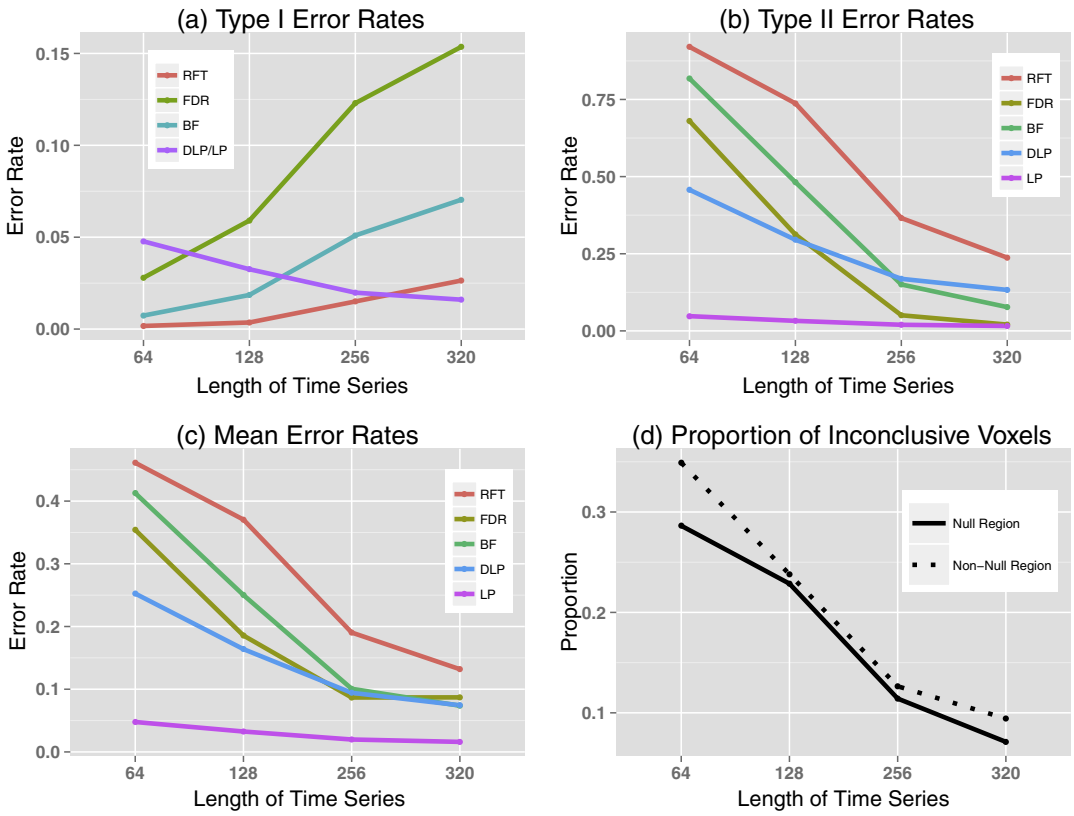

Fig. 5. Simulation results with different $T = 64, 128, 256, 320$. (a) Type I error rates; (b) Type II error rates; (c) mean error rates for RFT (red), FDR (olive), BF (green), DLP (blue), and LP (purple) based on 500 iterations sampled from chi-square distribution with 3 degrees of freedom; (d) the proportion of inconclusive voxels in null (black solid) and non-null (black dotted) regions for LP at different length of time series.

displays the worst performance compared to the other methods, indicating that RFT would be very conservative regardless of the true underlying error distribution in model (1). In Fig. 5(c), the LP and DLP outperform the conventional methods in terms of mean error rates. However, the similarity between Figs. 3 and 5 supports that all five methods are quite robust against violating the Gaussian assumption.

Error rates with different smoothing methods

To assess the effect of the size of the smoothing kernel for RFT, FDR, and BF, we generated the data from a Gaussian distribution with $T = 128$ combined with an exponential spatial covariance function with a decaying parameter of 2. Then, we estimated the parameters of interest and applied each method, i.e., RFT, FDR, BF, and DLP, to account for multiple comparisons. Note that we applied only DLP here as it is dominated by LP in any case. We evaluated the performance of each method at different smoothing kernel size: FWHM = 0 (no smoothing), FWHM = 4 (mild spatial smoothing), FWHM = 8 (high spatial smoothing), and using the information from immediately neighboring voxels (=UseNeighbors) in order to make a fair comparison of the methods after removing the effect of the size of the smoothing kernel. As shown in Fig. 6(a) and (b), RFT (red) attains the lowest Type I and highest Type II error rates regardless of smoothing kernel size and method, except for BF with UseNeighbors. Comparing to FDR (olive), DLP (purple) results in smaller Type I error rates and similar level of Type II error rates, but the mean error rates for both methods look alike in Fig. 6(c). The overall performance of each method is the worst when FWHM = 0, but it is much improved by taking into account the underlying spatial correlation, e.g., spatial smoothing or borrowing the information from neighboring voxels. It is surprising to observe a significant improvement in the mean error rate for RFT when combined with UseNeighbors instead of spatial smoothing. In contrast, the performance of BF when combined with UseNeighbors seems to be much worse than when the data were spatially smoothed, which is consistent across different length of time series. Note that the mean error rate for each method is highly dependent on the corresponding Type II error rate.

We further investigated the relationship between the two conventional approaches and DLP by characterizing the behavior of Type I error rates while matching their Type II errors. For each conventional approach, we found a value of k for DLP which resulted in the same Type II error rate at each of Monte Carlo iterations ($n = 500$). Searching such a value of k was performed using a simple exhaustive searching algorithm on a predefined set of values k at each iteration. The results are displayed in Fig. 6(d). There seems to be no difference in Type I error rates between RFT (red) and DLP (cyan) and between BF (green) and DLP (pink), whereas DLP (blue) slightly outperforms FDR (olive) both for FWHM = 8 and UseNeighbors.

These results indicate that the likelihood approach generally outperforms the other commonly used methods in the brain imaging analysis, regardless of the length of time series. This result is not due to the effect of the size of spatial smoothing kernel but due to the characteristics of the likelihood paradigm that are described in Section 2.

Effect of expanding weak evidence zone (benchmarking the magnitude of the LR, k)

Making fair comparisons among these approaches requires that we identify a comparison metric that is applicable across all of the approaches. The metric we have chosen is a simple binary map of brain activation, i.e., DLP map with $k = 20$. It is simple, intuitive, and visually appealing. However, the likelihood approach is not designed to yield a binary classification, so it must be modified slightly to permit such a comparison. To be clear about how we could construct a brain map, we present the following simple example. We will then further simplify this map so that it is comparable with standard approaches. We generated three representative evidence maps from the simulation at $T = 128$

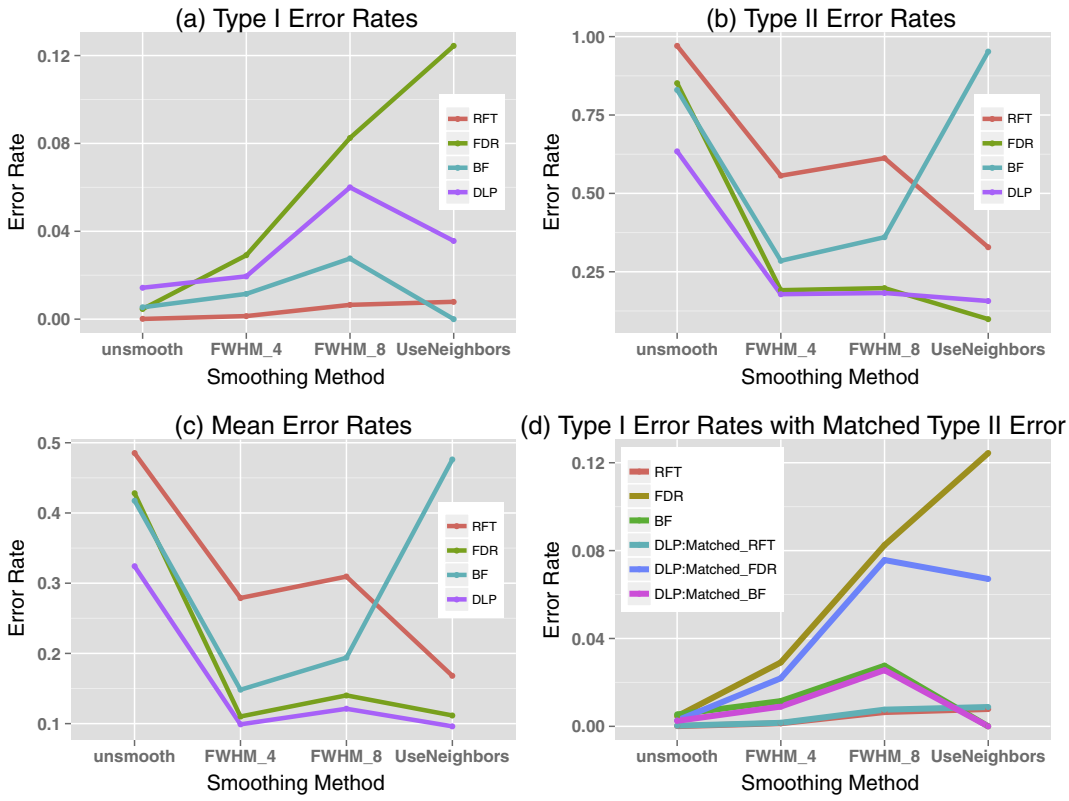

Fig. 6. Simulation results for RFT (red), FDR (olive), BF (blue), and DLP (purple) with $T = 128$ in terms of the Type I and II error rates, and the average of them at each method of smoothing, i.e., unsmoothing, FWHM = 4, FWHM = 8, and UseNeighbors: (a) Type I error rates, (b) Type II error rates, (c) mean error rates. Type I error rates are shown in (d) for DLP-RFT (cyan), DLP-FDR (blue), and DLP-BF (pink) pairs when each pair shares the same Type II error rates for each smoothing method.

corresponding to $k = 8$, 20, and 32 in Fig. 7(d), (e), and (f), where the "weak" evidence (i.e., LRs bewteen $1/k$ and k) are denoted as white. LRs greater than k that favor the activation hypothesis are denoted in red and LRs less than $1/k$ indicating no evidence for activation are denoted in blue. As k increases, the weak evidence region becomes wider, resulting in more voxels falling into the weak evidence region. Importantly, we see that the choice of k is largely inconsequential, except perhaps at the edges of the red regions. As illustrated in Fig. 3(d), however, the number of voxels classified as inconclusive decreases as the length of time series increases. We know from theoretical arguments that as the sample size increases, the LR will eventually support the correct hypothesis. Hence, the number of white regions will shrink with additional data. Even with $T = 128$, it is hard to distinguish between figures (d), (e), and (f) in Fig. 7, indicating the choice of the evidential level needed to show activation is somewhat robust even with time series as large as $T = 128$.

Notice that dark red areas are largely void of LRs that strongly favor the null (i.e., blue) and that the edges of the activation regions are almost uniformly white. This is to be expected as the edges are hard to identify and so the transition from red to blue voxels is buffered by a "weak evidence zone" which is in white. It is obvious that the likelihood depiction (b), (d), (e), or (f) possesses more information regarding the strength of evidence supporting the alternative (red color) and the null (blue color) than the binary map (c) in Fig. 7 at various k values. Generating different log-likelihood ratio maps with combinations of k and alternative values would be able to shed light on scientific findings that may be hindered by conservative conventional imaging analysis tools. A critical point is that standard brain maps based on p-values or hypothesis testings are unable to distinguish between the blue and white regions. This is because under the null hypothesis, the p-value is uniformly distributed, so all large p-values are inconclusive (i.e., weak evidence) and none can be interpreted as evidence supporting the null hypothesis.

fMRI data analysis

To identify activation patterns in a study designed to test cognitive control related activation in the prefrontal cortex (PFC) of the human brain, we used the proposed likelihood approach. Below, we describe the background, motivation, and description of these data that are relevant to the current analysis of detecting patterns of activation.

Based on prior work (e.g., Badre and D'Esposito, 2007; Badre et al., 2009; Long and Badre, 2009; Badre and Frank, 2012), we know that a fronto-parietal network is expected to be activated in an experimental situation where a subject selects one of two perceptual dimensions (i.e., shape or texture) of a stimulus in order to make a response. In this experiment, participants were expected to choose one of four keypress responses based on the relevant perceptual dimension that was cued by a color stimulus. More details about the experiment can be found in Long and Badre (2009) and Kang et al. (2012). For a given block of trials, either one dimension (e.g., only shape) would be cued throughout (D1) or two dimensions (both shape and texture) could be cued (D2). Cognitive control is required on D2 blocks in order to select the relevant dimension based on color, while minimal cognitive control is required on D1 blocks. Based on prior work (Badre and D'Esposito, 2007; Badre et al., 2009; Badre and Frank, 2012), the contrast of D2 > D1 should produce activation in the fronto-parietal network.

In the experiment, there were 288 trials, 144 of each dimension condition (D1, D2). Each trial lasted 2 s, with a variable inter-trial interval of 0–8 s. The trials were grouped into six scanning runs, with 48 trials per run. Each run consists of 4 blocks, which follow an ABBA format for

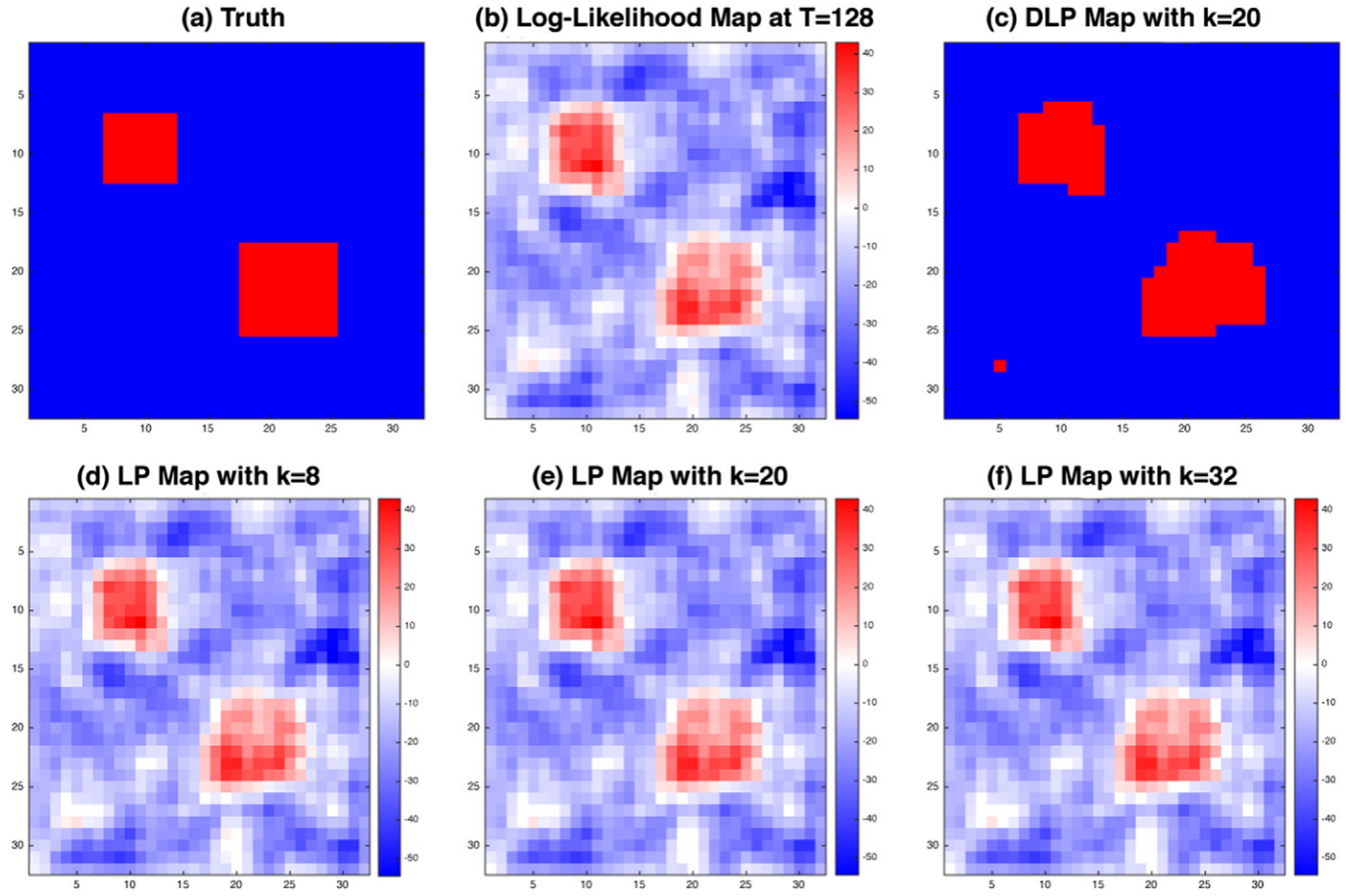

Fig. 7. Simulated voxel analysis for activation map. Data were simulated with two boxcar external stimuli as described in 3.1 with the temporal dimension of $T = 128$. The figures derived as follows: (a) truth (two truly active regions in red), (b) log-likelihood ratio map without applying k, (c) DLP map with $k = 20$, (d) LP map with $k = 8$, (e) LP map with $k = 20$, (f) LP map with $k = 32$. "LP map with $k = 8$ (20 or 32)" means that the weak evidence zone is from 1/8 (1/20 or 1/32) to 8 (20 or 32) and so any LRs in that set are set to white. The figures (d), (e), and (f) are not the same but very similar, indicating the minimal effect of cutoff value when T is long enough.

dimension type (e.g. D1, D2, D2, D1). The order of dimension condition is counterbalanced across subjects.

Whole-brain imaging was performed using the Siemens 3 T TIM Trio MRI system. Functional images were acquired using a gradient-echo echo-planar sequence (TR = 2 s; TE = 30 ms; flip angle = 90; 33 axial slices, 3 × 3 × 3.5 mm). After the five functional runs, high-resolution T1-weighted (MP-RAGE) anatomical images were collected for visualization (TR = 1900 ms; TE = 2.98 s; flip angle = 9; 160 sagittal slices, 1 × 1 × 1 mm).

Preprocessing were performed using FSL software package (FMRIB software Library, (Smith et al. (2004)), which includes slice timing correction, head motion correction across all runs, co-registration, and normalization on to the Montreal Neurological Institute (MNI) stereotaxic space. But the images were not spatially smoothed. After preprocessing steps, general linear models as shown in (1) were fitted to the data collected from a single subject as described in Section 3.2 for LP approach using Matlab (Mathworks, Natick, MA). The model contains only three covariates, which are for D1, D2, and the instruction period (IP). The three stimuli are convolved with the canonical HRF used in SPM8 and are denoted by X_{D1}, X_{D2}, and X_{IP}, respectively. The corresponding regression coefficients are β_v^{D1}, β_v^{D2}, and β_v^{IP} at voxel v, where $\beta_v^{D2} - \beta_v^{D1}$ is of main interest to test the hypothesis, $H_0 : \beta_v^{D2} - \beta_v^{D1} = 0$ and $H_1 : \beta_v^{D2} - \beta_v^{D1} = \delta_1$, where $v = 1, \ldots, V$. We used 95 percentile value of the distribution of $\hat{\beta}^{D2} - \hat{\beta}^{D1}$ as δ_1 and set $k = 20$ as in simulation studies. To illustrate the difference in resulting activation map, we also analyzed the data following typical protocol which includes slice timing correction, head motion correction across all runs, co-registration, normalization to the MNI space, spatial smoothing (FWHM = 6), fitting the first and second level GLM models, and then cluster thresholding using FSL. The cluster thresholding approach, consisting of a Z statistic thresholding at 2.3 and RFT (default method in FSL), is known to be much less conservative than controlling FDR at voxel-level.

Both the cluster thresholding approach and DLP($k = 20$) approach locate activation in lateral frontal and parietal cortex, consistent with the fronto-parietal network commonly observed in these response selection tasks as depicted in Fig. 8. Importantly, however, the RFT analysis locates primarily left lateralized activation in prefrontal cortex shown in Fig. 8(a). In Fig. 8(b), the DLP method locates activation bilaterally and as far rostral as frontal pole. In general, the activation located by RFT comprises a subset of those identified using DLP. This illustrates the reduced vulnerability to Type II errors when using the DLP approach, which is less likely to suppress important scientific findings that are often ignored by overly conservative classical Type I error correction tools.

Hence, using likelihood methods, we verify what is now known to be true: activation patterns are much broader than what standard methods in this setting typically discover (because they are too conservative). The LP clearly identifies a broader network associated with the contrast. Compared to the cluster thresholding, controlling voxel-level FWER using RFT or FDR at 0.05 instead of employing cluster thresholding approach resulted in activation maps with only a few blobs (not included here).

We need a method that can correctly identify the entire areas of activation in the fronto-parietal network that is engaged in sustained attention and working memory (Coull et al., 1996). The dysfunction of the network was found in people with neurobehavioral disorders, e.g., alcoholism and substance abuse (Boettiger et al., 2007), and bipolar

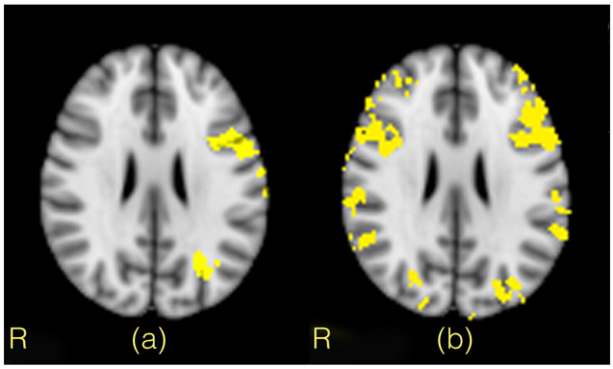

Fig. 8. Activation maps for the 50th axial slice (MNI Z-coordinate = 26), where yellow blobs indicate activated areas resulted from (a) controlling FWER at 0.05 using cluster thresholding in FSL, which consists of a Z statistic threshold at 2.3 and RFT (default method in FSL), after applying spatial smoothing (FWHM = 6), and (b) employing the DLP($k = 20$) approach without spatial smoothing.

disorder (Najt et al., 2013). By reliably identifying the cognitive control related activation patterns, we believe that our LP approach will shed new light on the neurobiological basis of such disorders associated with dysfunction in the fronto-parietal network.

Comments

The proposed likelihood approach appears to overcome the severe conservativeness of standard multiple comparisons methods. It results in significantly reduced Type II and *global* error rates compared to other conventional methods, enabling the discovery of more truly active voxels. A key advantage is the ability to describe evidence as weak. Moreover, the Type I error inflation, that can result from disallowing this advantage, appears to be generally low and quite acceptable. The simultaneous convergence of both likelihood version of *global* false positive and false negative error rates to zero is another notable property that present significant promise for statistical methods in neuroimaging. We note that likelihood methods can also be constructed in such a way as to impart robustness to model mis-specification (Blume et al., 2007).

One potential limitation is the dependence of evidential likelihood ratios on a specific alternative. Although we showed that a percentile between 90th and 99th of the MLE of linear contrast of regression coefficients, e.g., $\beta_1 - \beta_2$ would be a good alternative value, choosing a good alternative value still deserves further research. Evidence is like power in this way, and we view this as a welcome advance that eliminates the entrenched confusion between statistical significance and clinical significance. However, we note that power is never criticized for depending on a specific alternative and the concept of power averaged over all alternatives does not readily appear helpful. In our simulations and data-driven example, we choose for our alternative the 95th percentile of the distribution of the maximum likelihood parameter estimates. This data-driven approach appears to work well in practice and we are involved in ongoing research in this area. The caveat is that replacing the fixed simple alternative with a data-driven one (i.e., one that now changes with the sample size) results in a slight increase in the probability of observing misleading evidence. However, by using the 95th percentile of the MLE distribution, we drastically reduce the variability of the alternative and it behaves almost as if it were a fixed alternative.

Friston and Penny (2003) note that Bayesian or empirical Bayesian approaches offer an alternative approach with some advantages similar to that gained with a likelihood approach. However, based on our simulation results, it was shown that both LP and DLP would outperform the approach based on approximated Bayes factors proposed by Kass and Raftery (1995). Space does not permit full discussion regarding similarity and dissimilarity between Bayesian approach and the likelihood paradigm, but the likelihood approach can be thought of as prior-less Bayesian approach that retains excellent operational characteristics. In our estimation, the proposed likelihood approach offers a welcome compromise between classical frequentist approaches and Bayesian approaches. Importantly, likelihood retains the desirable properties of classical methods (e.g., control over error rates) while shedding the undesirable ones (e.g., ad hoc adjustments of error rates). Neuroimaging research could be well served by likelihood methods that promote gains in accuracy, efficiency, and flexibility.

Acknowledgments

The authors thank Dr. Ivor Cribben (University of Alberta) for helpful discussions. This research was supported in part by grants from NSF DMS (Hernando Ombao), NSF SES (Hernando Ombao), and NIH MINDS R01NS065046 (David Badre).

Appendix A. Law of likelihood

The law of likelihood was presented by Hacking (1965) and restated by Royall (1997):*The law of likelihood*: If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence.

When a random variable X follows a probability distribution with a parameter θ, observation x provides a likelihood function $L(\theta; x)$. Consider the simple null hypothesis $H_0 : \theta = \theta_0$ and simple alternative hypothesis $H_1 : \theta = \theta_1$. According to the law of likelihood, observation $X = x$ provides evidence supporting the alternative hypothesis if and only if $L(\theta_1; x) > L(\theta_0; x)$ and the ratio $L(\theta_1; x)/L(\theta_0; x)$ measures the strength of that evidence. Simply stated: the hypothesis with a higher likelihood is better supported.

Appendix B. Interpretation of k

A likelihood ratio of k means the same strength of evidence regardless of context. This is because the evidence from the data changes the prior odds into the posterior odds by exactly k. That is, the likelihood ratio represents the degree to which one's belief or knowledge is changed by the data (all priors, regardless of their magnitude, are changed by exactly the same amount on the odds scale). To illustrate for this context, we have the following thought experiment:Suppose that there are two hospitals: one hospital has only patients with mild cognitive impairment (MCI) and the other accepts equal numbers of MCI and Alzheimer's patients. You are masked to characteristics of both hospitals and are asked to identify which one accepts only MCI patients. With current MR technology, you believe that you can distinguish MCI from Alzheimer's patients by looking into the resting-state functional connectivity (FC) between two pre-determined regions of interests (ROIs), i.e., you are expected to see strong functional connectivity in an MCI patient while there will be no FC in an Alzheimer's patient. To this end, you decide to visit one hospital, randomly pick a patient with replacement, acquire that patient's resting-state fMRI, and investigate the FC between the two regions. Your hypothesis is, H_0: that hospital has the equal numbers of MCI and Alzheimer's patients, H_1: that hospital accepts only MCI patients. Suppose that you find two MCI patients in a row from the hospital. Then the likelihood ratio supporting H_1 is $LR = 1^2/0.5^2 = 4$, with three patients in a row, $LR = 8$, etc. Therefore, we can interpret that the strength of evidence associated with $LR = 4$ as the same evidence as that provided by two randomly sampled consecutive patients that turned out to have MCI. This would generally not be convincing, hence it is tagged as"weak" evidence. $LR = 20$ (or $LR = 32$) is equivalent to consecutively sampling 4.3 (or 5) patients in a row that all turn out to have MCI. If only half the patients at the hospital had MCI, it is unlikely for you to sample 5 MCI patients in a row. Hence LR of 32 marks the transition from moderate to strong evidence.

Appendix C. Computing Type II error rate

Let the null and alternative hypotheses of interest be: $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_0 + \sigma$ given $\hat{\theta} \sim N(\theta, \sigma^2/n)$. Suppose that we have 8 images ($n = 8$) and consider a single voxel. Because the Type I error rate is fixed at 0.05 (one-sided),

$$\begin{aligned}\text{Type I error rate } &= P(Z \geq 1.645|H_0)\\ &= P\left(\sqrt{n}\left(\hat{\theta}-\theta_0\right)/\sigma \geq 1.645\right)\\ &\quad \text{under the null : } \hat{\theta} \sim N(\theta_0, \sigma^2/n)\\ &= P\left(\hat{\theta} \geq 1.645\sigma/\sqrt{n}\right) \text{ if } \theta_0 = 0,\end{aligned}$$

where Z denotes a standard normal distribution, $Z \sim N(0, 1)$. Then, the Type II error rate is by definition,

$$\begin{aligned}\text{Type II error rate } &= P\left(\hat{\theta} < \theta_0 + 1.645\sigma/\sqrt{n}|H_1\right)\\ &= P(Z < 1.645 - \sqrt{n})\\ &\quad \text{under the alternative : } \hat{\theta} \sim N(\theta_0 + \sigma, \sigma^2/n)\\ &= 0.118.\end{aligned}$$

When we consider four voxels instead of a single voxel, the Type I error rate per voxel is 0.0125 (=0.05/4) with a Bonferroni correction and then it is straightforward to compute the Type II error rate of 0.278 per voxel following the steps described above.

Appendix D. Probability of observing misleading evidence

For the purpose of illustration, one example is used as follows. Suppose that $X_1, X_2, \ldots, X_n$ i.i.d. normal with unknown mean μ and known variance σ^2, and two hypotheses are $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$. Denote the likelihood function as $L_n(\mu) = \prod_{i=1}^{n} f(X_i; \mu)$. Then the probability of observing misleading evidence for μ_1 over μ_0 is

$$\mathbb{P}\left(\log\left\{\frac{L_n(\mu_1)}{L_n(\mu_0)}\right\} > \log k \middle| H_0\right) = M(n, k) \tag{D.1}$$

where n is the size of the data and $M(\cdot,\cdot)$ is a function of k and n. In the Eq. (D.1), $k > 1$ is any positive fixed number and does not depend on the error rates. The likelihood ratio in Eq. (D.1) can be further expanded as

$$\begin{aligned}\frac{L_n(\mu_1)}{L_n(\mu_0)} &= \frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right]}{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right]}\\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[n(\mu_1+\mu_0)(\mu_1-\mu_0) - 2(\mu_1-\mu_0)\sum_{i=1}^{n}x_i\right]\right\}\\ &= \exp\left\{\frac{n(\mu_1-\mu_0)}{\sigma^2}\left[\bar{X}_n - \frac{(\mu_1+\mu_0)}{2}\right]\right\}\end{aligned}$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Under the null, the probability that the data support the alternative hypothesis over the null hypothesis by at least a factor of k as shown in Blume (2002) is\kern1pt

$$\begin{aligned}M(n,k) &= \mathbb{P}\left\{\log\left\{\frac{L_n(\mu_1)}{L_n(\mu_0)}\right\} > \log k \middle| H_0\right\}\\ &= \mathbb{P}\left\{\bar{X}_n \geq \frac{\sigma^2}{n(\mu_1-\mu_0)}\log k + \frac{\mu_1+\mu_0}{2}\middle| H_0\right\}\\ &= \mathbb{P}\left\{Z \leq -\frac{\sigma}{\sqrt{n}(\mu_1-\mu_0)}\log k - \frac{\sqrt{n}(\mu_1-\mu_0)}{2\sigma}\right\}\\ &= \Phi\left(-\frac{\sigma}{\sqrt{n}(\mu_1-\mu_0)}\log k - \frac{\sqrt{n}(\mu_1-\mu_0)}{2\sigma}\right)\end{aligned} \tag{D.2}$$

where Φ denotes the standard normal cumulative distribution function. In the Eq. (D.2), it is obvious that for any fixed $k \geq 1$, as $n \to \infty$, then $\Phi(-\frac{\sigma}{\sqrt{n}(\mu_1-\mu_0)}\log k - \frac{\sqrt{n}(\mu_1-\mu_0)}{2\sigma}) \to \Phi(-\infty) = 0$ and so $M(n, k) \to 0$. Similarly, it is straightforward to show that the probability of observing weak evidence, $\mathbb{P}(-\log k < \log\{\frac{L_n(\mu_1)}{L_n(\mu_0)}\} < \log k | H_0) \to 0$ as $n \to \infty$. Note that this probability is the same as $\mathbb{P}(-\log k < \log\{\frac{L_n(\mu_1)}{L_n(\mu_0)}\} < \log k | H_1)$. Asymptotic behavior of the probability of observing misleading evidence under H_1, which is analogous to the Type II error rate, is the same as that of the probability under H_0.

Appendix E. Summary of parameters for simulation

We summarize all the parameters used for our main simulation studies resulting in Figs. 1 and 3 in the following table.

	Simulation parameters for Fig. 1	Simulation parameters for Fig. 3
Dimension	91×109	32×32
Number of "active" voxels	300	100
Number of images	$T = 128$	$T = 64, 128, 256, 320$
AR(1) parameter	0.4	0.4
Variance of random noise	2.5	1.5
Effect size $\beta_1 - \beta_2$	1	1
Spatial dependence	Spatial smoothing FWHM = 3.5	Exponential function Decaying parameter = 2 Variance = 2.5

References

Badre, D., D'Esposito, M., 2007. Fmri evidence for a hierarchical organization of the prefrontal cortex. J. Cogn. Neurosci. 19, 2082–2099.

Badre, D., Frank, M.J., 2012. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 2: evidence from fMRI. Cereb. Cortex 22, 527–536.

Badre, D., Hoffman, J., Cooney, J.W., D'Esposito, M., 2009. Hierarchical cognitive control deficits following damage to the human frontal lobe. Nat. Neurosci. 12, 515–522.

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

Blume, J.D., 2002. Likelihood methods for measuring statistical evidence. Stat. Med. 21, 2563–2599.

Blume, J., 2011a. Likelihood and its evidential framework. In: Gabbay, D., Woods, J. (Eds.), Handbook of The Philosophy of Science: Philosophy of Statistics. North Holland, San Diego.

Blume, J., 2011b. The likelihood paradigm in action: an application to fMRI data and evaluation of performance. Oral presentation at Eastern North American Region (ENAR) meeting, Miami, FL.

Blume, J., Peipert, J., 2004. Randomization in controlled clinical trials: why the flip of a coin is so important. J. Am. Assoc. Gynecol. Laparosc. 11, 320–325.

Blume, J., Su, L., Olveda, R.M., McGarvey, S.T., 2007. Statistical evidence for glm regression parameters: a robust likelihood approach. Stat. Med. 26, 2919–2936.

Boettiger, C.A., Mitchell, J.M., Tavares, V.C., Robertson, M., Joslyn, G., D'Esposito, M., Fields, H.L., 2007. Immediate reward bias in humans: fronto-parietal networks and a role for the catechol-o-methyltransferase 158 val/val genotype. J. Neurosci. 27, 14383–14391.

Choi, L., Caffo, B., Rohde, C., 2008. A survey of the likelihood approach to bioequivalence trials. Stat. Med. 27, 4874–4894.

Coull, J.T., Frith, C.D., Frackowiak, R.S.J., Grasby, P.M., 1996. A fronto-parietal network for rapid visual information processing: a pet study of sustained attention and working memory. Neuropsychologia 34, 1085–1095.

Friston, K., Penny, W., 2003. Posterior probability maps and spms. NeuroImage 19, 1240–1249.

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

Hacking, I., 1965. Logic of Statistical Inference. Cambridge University Press.

Jeffreys, H., 1961. Theory of Probability. 3rd ed. Oxford University Press, Oxford, UK.

Kang, H., Ombao, H., Linkletter, C., Long, N., Badre, D., 2012. Spatio-spectral mixed effects model for functional magnetic resonance imaging data. J. Am. Stat. Assoc. 107, 568–577.

Kass, R., Raftery, A., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.

Long, N.M., Badre, D., 2009. Testing hierarchical interactions in frontal cortex during cognitive control. Poster presented at the 16th Cognitive Neuroscience Society meeting.

Najt, P., Bayer, U., Hausmann, M., 2013. Right fronto-parietal dysfunction underlying spatial attention in bipolar disorder. Psychiatry Res. 210, 479–484.

Royall, R., 1997. Statistical Evidence: A Likelihood Paradigm. Chapman and Hall.

Royall, R., 2000. On the probability of observing misleading statistical evidence (with discussion). J. Am. Stat. Assoc. 95, 760–767.

Royall, R., Tsou, T.S., 2003. Interpreting statistical evidence using imperfect models: Robust adjusted likelihood function. J. R. Stat. Soc. Ser. B 65, 391–404.
Schwartzman, A., Dougherty, R.F., Lee, J., Ghahremani, D., Taylor, J.E., 2009. Empirical null and false discovery rate analysis in neuroimaging. NeuroImage 44, 71–82.
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., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N.D., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23, S208–S219.
Storey, J., 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B 64, 479–498.
Wang, S., Blume, J., 2011. An evidential approach to non-inferiority clinical trials. Pharm. Stat. 10, 440–447.
Worsley, K.J., 1994. Local maxima and the expected euler characteristic of excursion sets of $\chi 2$, f, and t fields. Adv. Appl. Probab. 26, 13–42.
Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for rCBF activation studies in the human brain. J. Cereb. Blood Flow Metab. 12, 900–918.