Model-based segmentation for improved activation detection in single-subject functional Magnetic Resonance Imaging studies

Wei-Chen Chen\textsuperscript{a,1}, Ranjan Maitra\textsuperscript{b,2},
\textsuperscript{a}Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring MD, USA.
\textsuperscript{b}Department of Statistics, Iowa State University, Ames, Iowa, 50011-1090, USA.

ABSTRACT

Functional Magnetic Resonance Imaging (fMRI) maps cerebral activation in response to stimuli but this activation is often difficult to detect, especially in low-signal contexts and single-subject studies. Accurate activation detection can be guided by the fact that very few voxels are, in reality, truly activated and that these voxels are spatially localized, but it is challenging to incorporate both these facts. We address these twin challenges to single-subject and low-signal fMRI by developing a computationally feasible and methodologically sound model-based approach, implemented in the R package MixfMRI, that bounds the \textit{a priori} expected proportion of activated voxels while also incorporating spatial context. An added benefit of our methodology is the ability to distinguish voxels and regions having different intensities of activation. Our suggested approach is evaluated in realistic two- and three-dimensional simulation experiments as well as on multiple datasets. Finally, the value of our suggested approach in low-signal and single-subject fMRI studies is illustrated on a sports imagination experiment that is often used to detect awareness and improve treatment in patients in persistent vegetative state (PVS). Our ability to reliably distinguish activation in this experiment potentially opens the door to the adoption of fMRI as a clinical tool for the improved treatment and therapy of PVS survivors and other patients.

© 2022 Elsevier B. V. All rights reserved.

1. Introduction

Functional Magnetic Resonance Imaging (fMRI) is an imaging tool for understanding the spatial characteristics of human cognitive and motor function (Belliveau et al., 1991; Kwong et al., 1992; Bandettini et al., 1993; Howseman and Bowtell, 1999; Lindquist, 2008; Lazar, 2008). It is predicated on the significant discovery over three decades ago that the firing of neurons in response to the application of a stimulus or the performance of a task is accompanied by changes in the blood oxygen levels in neighboring vessels, yielding the so-called Blood Oxygen Level Dependent (BOLD) contrast and that this BOLD effect is also present in human brains (Ogawa et al., 1990a). The BOLD effect causes a shift in the MR signal and is used in fMRI as a surrogate for neural activity. Images of BOLD measurements at each voxel, or volume element, are acquired at multiple times during or without the task- or stimulus-related activity and, before preprocessing to address the effects of motion and other artifacts, form the raw data from an fMRI experiment.

An important outcome of many fMRI experiments is the construction of voxel-wise maps to identify regions of neural acti-
eration. Such activation maps are often obtained by fitting a general linear model (Friston et al., 1995) to the aforementioned raw data, that is, the BOLD time series observations at each voxel. Statistical Parametric Mapping (Friston et al., 1990) then reduces the data at each voxel to a test statistic that summarizes the association between the BOLD response and the stimulus time-course (Bandettini et al., 1995). The test statistics (or their p-values) are thresholded to determine significantly activated voxels (Worsley et al., 1996; Genovese et al., 2002; Logan and Rowe, 2004; Monti, 2011). Many factors (Biswal et al., 1996; Hajnal et al., 1994) challenge detection accuracy, leading to much variation in identified activation from one study to another even when the same subject is scanned under the same paradigm (Maia et al., 2002; Giuliani et al., 2005; Maia, 2009a). In single subjects or high-level cognitive tasks, the signal- or contrast-to-noise ratios (SNR or CNR) are typically low, hampering accurate activation detection. Improving detection accuracy is important for clinical adoption of fMRI to understand and evaluate individual cognitive functions and to aid diagnosis by identifying pathologies.

There are many thresholding approaches to control the False Discovery Rate (FDR) (Genovese et al., 2002; Nichols and Hayasaka, 2003; Benjamin and Heller, 2007; 2008), with refinements (Benjamin and Heller, 2007) for weighted FDR to include spatial context and determine activation of clusters of voxels. Other approaches (Friston et al., 1991; Worsley et al., 1992; Worsley, 1994; Worsley and Taylor, 2003) use Random Field Theory and specified correlation structure on test statistics to calculate the threshold that produces an expected Euler characteristic (Adler, 1981) equal to the required p-value (Worsley, 1994; Worsley et al., 1996). Cluster-wise thresholding is perhaps the most popular approach and involves drawing clusters of a minimum number of connected voxels having p-values below a specified threshold (Friston et al., 1994; Hayasaka and Nichols, 2003). The method requires specification of neighborhood order, thresholding p-value and minimum cluster size. There has been little study on how neighborhood order impacts activation detection accuracy, but many researchers use a second-order neighborhood. Earlier p-value threshold recommendations (Forman et al., 1995; Lazar, 2008) of between 0.02 and 0.03 were deemed too liberal in Woo et al. (2014) who suggested a more conservative default of 0.001. Threshold-free cluster enhancement (TFCE) methods (Smith and Nichols, 2009; Heller et al., 2006; Spisak et al., 2019) obviate the need for threshold choice but can have low spatial specificity for larger regions or need precise a priori information (Woo et al., 2014). FDR in cluster-wise thresholding is controlled by choosing a minimum cluster size based on simulations under a null-activation model, but recent empirical studies have indicated cluster-wise thresholding implementations of popular fMRI software to have inflated false positive rates, both in single subject (Eklund et al., 2015) and group (Eklund et al., 2016) studies, though corrections and re-evaluation (Eklund et al., 2016; Kessler et al., 2017; Cox et al., 2017) have since indicated that the conclusion may have been unnecessarily alarmist.

The Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group Report to the Advisory Committee to the Director, NIH (2014) highlighted the need for holistic approaches that account for the spatial, temporal and behavioral properties of the experiment and to improve SNR to make reliable inference at the single-subject level while also moving the field to identify individual pathologies and help improve patient care in a clinical setting. In this context, Brown and Behrmann (2017) call for the linking of statistical methodology and the fMRI research paradigm to develop more accurate analysis methods, ensuring reliability and reproducibility.

One currently unexploited fact of fMRI activation studies is that only a very small (Heller et al., 2006; Lazar, 2008; Almodóvar-Rivera and Maitra, 2019) proportion (about 1-3%) of in-brain voxels are a priori expected to be truly activated. So, in this paper, we incorporate this information along with spatial context to improve detection accuracy even at the single-subject level in a computationally practical framework with the following objectives:

- Constrain a priori, the expected proportion of activated voxels. Importantly, this means we do not constrain the actual proportion of activated voxels in a particular study, but rather use this constraint to reflect our prior beliefs to guide inference.
- Promote spatial contiguity and context in the detected activation.

It is common to address the latter goal through the use of smoothing or through other devices such as a spatial mixture model (Hartvig and Jensen, 2000), or Markov Random Field priors such as the Ising model (Genovese, 2000; Smith and Fahrmeir, 2007); however, these approaches have not been successfully coupled with our first objective that places a limit on the a priori expected proportion of activated voxels. Our approach postulates a mixture model for the distribution of the p-value of the test statistic at each voxel – the first mixture component represents the inactive voxels and has mixing proportion set to be no less than the a priori expected proportion of inactive voxels. Spatial contiguity in the detected activation is encouraged through a practical penalty term added to the log-likelihood function. Our showcase application, introduced in Section 4, lays out the challenges in accurately detecting activation in a sports imagination experiment, illustrating both the need and value of our methodology because of its implications in moving fMRI to improve treatment and therapy of, for example, persistent vegetative state (PVS) patients.

The remainder of this paper is organized as follows. We provide practical methods for accurate activation detection in single-subject fMRI and discuss its implementation in Section 5. We next comprehensively evaluate performance in simulation experiments in Section 6 for a variety of levels and problem difficulty, in 2D, and different noise levels in 3D. We also evaluate performance on many standard datasets in the literature, Section 5 analyzes the sports imagination dataset introduced in Section 2. The paper concludes with some discussion in Section 6. We also have appendices providing additional details on the preprocessing of our dataset, statistical methodology, experimental illustrations, performance evaluations, and data analysis.
2. Improving treatment of patients in persistent vegetative state via sports imagination experiments

The Multi-Society Task Force on PVS (1994) reported that some PVS patients are known to retain some cognitive functions (Schiff et al., 2002). Indeed, the supplementary motor area (SMA) was found to be significantly affected in the fMRI study (Owen et al., 2006) of a 25-year-old female Traumatic Brain Injury (TBI) survivor in PVS while she was verbally instructed to alternately rest and imagine playing tennis. Similar activation was observed in 12 healthy volunteers, so it has been surmised (Owen et al., 2006) and demonstrated (Monti et al., 2010) that fMRI can be used to diagnose conscious awareness and to communicate with PVS patients. A more elaborate study (Bardin et al., 2011) found consistent activation (beyond the SMA, into the parietal cortex and other regions) in 14 normal subjects, but murky consistency in seven TBI patients. While healthy subjects necessarily have cognitive function that is mostly intact and regionally consistent in standardized (e.g., Talairach) space, TBI patients with injuries impacting brain structure and function may be missing some standardized regions, or show inconsistent disruptions across TBI-afflicted brains. It is therefore important that activation detection be reliable and accurate at the individual subject level, and achieving these goals is the rationale behind our proposed methodology.

We illustrate the challenges to accurate detection in single-subject fMRI experiments, by considering a healthy 40-year-old female volunteer (Bardin et al., 2011) who was instructed to alternately rest or imagine playing tennis, in blocks of 30 seconds each, and with the entire 210-second study starting and ending with a resting block. Bardin et al. (2011) do not mention analysis of this dataset further, but the raw data were publicly released by Tabelow and Polzehl (2011) to demonstrate features of their fmri package in R (R Core Team, 2018). We refer to Appendix A for details on data collection, preprocessing and analysis but note here that several methods, for example, after controlling for FDR at \( q = 0.05 \) (Genovese et al., 2002; Benjamini and Yekutieli, 2001), permutation-based testing (Winkler et al., 2014), thresholding by the value that produces an expected Euler characteristic (Adler, 1981; Worsley, 1994) equal to the desired p-value (taken to be 0.05 in this article), highlighting the challenges faced by global thresholding methods in accurately and reliably constructing activation maps in low-signal single-subject experiments. Such was also the result using spatial mixture modeling (Hartvig and Jensen, 2000). We also employed cluster-wise thresholding methods to detect activation, with the minimum number of contiguous voxels in each activated region determined using the 3dClustSim function of the Analysis of Functional Neuroimages (AFNI) software (Cox, 1996; Cox and Hyde, 1997; Cox, 2012) at a threshold of 0.001 (following the recommendations of Woo et al., 2014) and using first-, second- and third-order neighborhoods. These three orders of neighborhood correspond to the definition of neighboring voxels accordingly as when their (1) faces, (2) faces or edges, and (3) faces, edges or corners touch. All three types of neighborhood found similar activated regions so we only display results (Fig. 1a) using a second-order neighborhood. The activation detected is mostly unemphatic. For instance, there is a very tiny blip of activation in the subject’s somatosensory cortex (left hemisphere) but not in her motor cortex or the SMA. The precuneus and the primary visual cortex are identified, but the extent of spatial activation (and clustering) is low. The difference in the described activation map could have led to misinterpretations (if we were not, for instance, aware of what we were expecting to be activated based on our prior knowledge of the activity or on the fact that we had a normal subject). It is important to note that in a clinical setting, where inference is desired at an individual subject level it is often not known beforehand (e.g., a patient with a TBI) the particular regions of the brain that are expected to be activated.

Unclearer activation maps are obtained by probabilistic Threshold-Free Cluster Enhancement (pTFCE) (Spisak et al., 2019) (Fig. 1b) – a refinement of Threshold-Free Cluster Enhancement (TFCE) (Smith and Nichols, 2009) – or by the robust fast adaptive and smoothed thresholding (AR-FAST) method of Almodovar-Rivera and Maitra (2019) at \( \alpha = 0.05 \) (Fig. 1c). Activation is detected in barely in the precuneus, or (with AR-FAST) the primary motor cortex (PMC). Other FAST methods (Almodovar-Rivera and Maitra, 2019) were also unsatisfactory.

This example provides some indication of the challenges and inadequacies associated with the most common activation detection methods. For this normal subject, activation in the SMA...
is not easily detected by standard thresholding methods, despite the fact that several studies have found this region to be reliably activated in healthy subjects during sports imagination experiments. These deficits point to the need for improving activation detection methodology, and we explore doing so by incorporating the additional knowledge provided by both spatial context and prior belief on the expected proportion of activated voxels. These two ingredients are important and central to the application of activation detection in fMRI and have hitherto never been successfully incorporated into the statistical methodology for analysis. It is the goal of this paper to redevelop such statistical analysis for the purpose of improving activation detection in fMRI. We return to this dataset in Section 5 after developing and evaluating our needed methodology.

3. Methodological Development

3.1. Background and Preliminaries

Most current approaches to the statistical analysis of fMRI data are not specifically geared towards fMRI activation detection, but, rather, adapted during post-hoc analysis. For instance, a popular approach to incorporating spatial context in activation detection is to use cluster-wise thresholding, or its derivatives, a popular approach to incorporating spatial context in activation detection. These deficits point to the need for improving activation detection in fMRI. We return to this dataset in Section 5 after developing and evaluating our needed methodology.

From (ii), \( \delta \) is the \textit{a priori} expected minimum proportion of inactivated voxels while (iii) expresses that \( p \) is from activated components are expected to be declared significant.

The generative model for (1) stems from the distribution of the (observed) \( p \)-value at each voxel along with its (missing) class indicator. Let \( W_{ik} \) be 1 if the \( i \)th voxel is in the \( k \)th class \((k = 0, 1, \ldots, K) \) and 0 otherwise. Writing \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)^\top \), \( \beta = (\beta_1, \beta_2, \ldots, \beta_K)^\top \) and \( W = ((W_{ik}))_{1 \leq i \leq n, 0 \leq k \leq K} \), the complete log likelihood, assuming that the dependence in the \( p \) is through the generative \( W \), is

\[
\ell_c(\pi, \alpha, \beta; W, p) = \sum_{i=1}^{n} W_{i0} \log \pi_0 + \sum_{i=1}^{n} \sum_{k=1}^{K} W_{ik} \log(\pi_k + \log(b(p; \alpha_k, \beta_k))).
\]  

The Expectation-Maximization (EM) algorithm \cite{Dempster,McLachlan2008} iteratively maximizes \( \mathbb{E}[\ell_c(\pi, \alpha, \beta; W, p) | p] \) to yield maximum likelihood estimates (MLEs) of \((\pi, \alpha, \beta)\).

3.2. Incorporating spatial context

A shortcoming of (2), and thus of (1), is its disregard of spatial context. Modeling the \( W_{ik}s \) via a Markov Random Field (e.g. Potts’ model) is computationally impractical because the expectation step (E-step) expressions are intractable, computed at each iteration only through time-consuming Markov Chain Monte Carlo (MCMC) or other methods. Our solution is to penalize (2) by adding the term

\[
-\sum_{i=\ell, j=1}^{n} \sum_{k=0}^{K} W_{ik} W_{jk} \sum_{d=1}^{3} \frac{(v_{id} - v_{jd})^2}{h_{kd}^2} - \log h_{kd},
\]  

with \( v_{d} \) as the \( \ell \)th voxel’s coordinates in the 3D image volume (\( v_{d} \) is omitted for 2D slices). For given \( k, (3) \) has the least effect when \( W_{ik} W_{jk} = 1 \) for nearby voxels \( i \) and \( j \), and is like a regularization term in density estimation (but note that the \( v_{d} \)s are fixed here, while \( W_{ik} \)s are unobserved). The \( h_{kd}^2 \)s in the denominator modulate the influence of (3), varying by axes and with \( k \) to allow for localization. The additional log \( h_{kd} \) terms are for computational convenience (ensuring also that (3) is, but for a constant term of addition, a log density).

The terms for each axis \( d (4) \) in (3) can be recast as

\[
\sum_{i=\ell, j=1}^{n} \sum_{k=0}^{K} W_{ik} W_{jk} (v_{id} - v_{jd})^2 = \sum_{i=1}^{n} \sum_{k=0}^{K} W_{ik} (v_{id} - \mu_{kd})^2 / \sigma_{kd}^2.
\]

where we recast \( \mu_{kd} = \sum_{i=1}^{n} W_{ik} v_{id} / \sum_{i=1}^{n} W_{ik} \) and \( \sigma_{kd}^2 = h_{kd}^2 / 2 \) for \( d = 1, 2, 3 \). The penalized complete log likelihood ((2) + (3)) is

\[
\sum_{i=1}^{n} \sum_{k=0}^{K} W_{ik} \log \pi_k + \log b(p; \alpha_k, \beta_k)
\]

\[
-\log(|\Sigma_k|/2 - (v_i - \mu_k)^\top \Sigma_k^{-1}(v_i - \mu_k)/2).\]
where \( \alpha_0 = \beta_0 \equiv 1, \mu_k = (\mu_{k1}, \mu_{k2}, \mu_{k3})^T, v_i = (v_{i1}, v_{i2}, v_{i3})^T \) and \( \Sigma_k = diag(\sigma^2_{k1}, \sigma^2_{k2}, \sigma^2_{k3}) \). Maximizing (1) after incorporating the spatial penalty term is equivalent to maximizing
\[
\ell(\Theta; p, v) = \sum_{i=1}^n \log \left( \sum_{k=0}^{K} \pi_k f(p_i; v_i; \Gamma_k) \right) + f(p_i; v_i; \Gamma_k) = b(p; \alpha_k, \beta_k) \phi(v; \mu_k, \Sigma_k),
\]
where \( \Theta = \{\Theta_0, \Theta_1, \ldots, \Theta_K\} \) with \( \Theta_k = \{\pi_k, \Gamma_k\} \) and \( \phi(x; \mu, \Sigma) \) is the multivariate normal density with parameters \( \mu, \Sigma \).

### 3.3. Parameter estimation and model-fitting

Adding (2) and (3) yields the penalized complete data log-likelihood
\[
\sum_{i=1}^n \sum_{k=0}^{K} W_{ik} \left[ \log \pi_k + \log b(p; \alpha_k, \beta_k) - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (v_i - \mu_k)^T \Sigma_k^{-1} (v_i - \mu_k) \right]
\]
while the penalized observed-data log likelihood equation is
\[
\ell(\Theta; p, v) = \sum_{i=1}^n \log \left( \sum_{k=0}^{K} \pi_k f(p_i; v_i; \Gamma_k) \right)
\]
\[
\equiv \sum_{i=1}^n \log \left( \sum_{k=0}^{K} \pi_k b(p; \alpha_k, \beta_k) \phi(v; \mu_k, \Sigma_k) \right).
\]

Development of a classical EM algorithm for parameter estimation given \( K \) is almost immediate, but the inseparability of \( (\alpha_k, \beta_k) \)'s from the beta function, the large scale of the problem with hundreds of thousands of voxels (in 3D), the need for constrained optimization and repeated initialization call for many refinements (Lange, 2010; Wu, 1983; Boyd and Vandenberghe, 2004; Meng and van Dyk, 1997). We develop strategies for parameter estimation and provide custom initialization schemes and a novel parallel Expectation-Gathering-Maximization (EGM) method for implementing the Alternating Partial Expectation Conditional Maximization (APECM) algorithm (Chen and Maitra, 2011) through its APECMa variant (Chen et al., 2013).

#### 3.3.1. Parameter estimation using the EM algorithm

The classical EM algorithm has the E-step for calculating \( w_{ik}^{(s+1)} = \mathbb{E}[W_{ik} \mid p, v, \Theta^{(s)}] = \pi_k^{(s)} f(p_i; v_i; \Gamma_k^{(s)}) / \sum_{r=0}^{K} \pi_r^{(s)} f(p_i; v_i; \Gamma_r^{(s)}) \) at the \( (s+1) \)th iteration, with \( \Theta^{(s)} \) being the current value of \( \Theta \) obtained at the end of the \( s \)th iteration. At that iteration, the maximization step (M-step) maximizes
\[
Q(\Theta; \Theta^{(s)}) = \mathbb{E}[\ell(\Theta; p, v; \Theta^{(s)})]
\]
\[
= \sum_{i=1}^n \sum_{k=0}^{K} w_{ik}^{(s+1)} \log \pi_k + \log f(p_i; v_i; \Gamma_k^{(s)})
\]
within the constraints (i), (ii) and (iii) that govern (1). The maximization of (6) is separable in \( \mu_k, \Sigma_k, \pi \) and \( (\alpha, \beta) \) and can be handled one subset of parameters at a time. M-step updates for \( \mu_k \) and \( \Sigma_k \) are immediate, with \( \mu_k^{(s+1)} = \sum_{i=1}^n w_{ik}^{(s+1)} v_i / \sum_{i=1}^n w_{ik}^{(s+1)} \) and the (diagonal) matrix \( \Sigma_k^{(s+1)} \) containing the diagonal elements of \( \sum_{i=1}^n w_{ik}^{(s+1)} (v_i - \mu_k^{(s)}) (v_i - \mu_k^{(s)})^T / \sum_{i=1}^n w_{ik}^{(s+1)} \).

**Lemma 1.** Let \( g(\pi, W) = \sum_{i=1}^n \sum_{k=0}^{K} w_{ik} \log \pi_k \), with \( \pi = (\pi_0, \pi_1, \ldots, \pi_K)^T \). Suppose that
\[
(\alpha) \quad \pi^* \text{ maximizes } g(\pi, W) \text{ under the constraint that } h_i(\pi) = \pi_1 - 1 = 0 \text{ where } 1 = (1, 1, \ldots, 1)^T \text{ inside the convex set } \Pi = \{\pi : 0 \leq \pi_k \leq 1, k = 0, 1, \ldots, K\}.
\]
Then, under the additional inequality constraint
\[(\beta) \quad h_i(\pi) = \pi_0 - \delta \geq 0, \quad g(\pi, W) \text{ is maximized at } \pi_k^* = \max(\delta, \pi_0^*), \quad \pi_k^* = (1 - \pi_0^*)/1 - \pi_0^* \quad \text{for all } k = 1, 2, \ldots, K.
\]

**Proof.** See Appendix B.1.1

Lemma 1 means that we can first ignore the inequality constraint and estimate \( \pi \) only under the equality constraint (a) to get \( \hat{k}_k^{(s+1)} = \sum_{i=1}^n w_{ik}^{(s+1)}/\sum_{i=1}^n w_{ik}^{(s+1)} \). Then, if \( \pi_k^{(s+1)} \geq \delta \), then \( \hat{k}_k^{(s+1)} \equiv \hat{k}_k^{(s)} \) for all \( k = 0, 1, \ldots, K \). Otherwise, set \( \hat{k}_0^{(s+1)} \equiv \delta \) and \( \hat{k}_k^{(s+1)} = (1 - \delta) \sum_{i=1}^n w_{ik}^{(s+1)}/\sum_{i=1}^n w_{ik}^{(s+1)} \) for all \( k = 1, 2, \ldots, K \).

Since we do not have analytical expressions for M-step updates of \( \alpha \) and \( \beta \) within the constraints imposed by (iii), we use constrained optimization function subject to linear inequality constraints and using an adaptive barrier, as in the constrOptim() function (Lange, 2010) in R (R Core Team, 2018). Here also, \( (\alpha_k, \beta_k) \) can be addressed separately for each \( k = 1, 2, \ldots, K \). The constrained optimization for \( \alpha_k \) and \( \beta_k \) may not result in convergence at each M-step iteration owing to numerical issues. However, an increase in (6) is enough to guarantee convergence of the ML estimates for (5) as per the generalized EM algorithm (Wu, 1983). Therefore, for \( \alpha \) and \( \beta \), we ensure an increased (6) within each M-step. We use the relative increase in the value of (5) over successive iterations \(|(\ell(\Theta^{(s)}; p, v) - \ell(\Theta^{(s)}; p, v))| / \ell(\Theta^{(s)}; p, v) | < \epsilon\) to determine convergence of our EM algorithms. In this paper, we set \( \epsilon = 10^{-6} \).

#### 3.3.2. Initialization

The EM algorithm only guarantees convergence to a local maximum with the initializing parameter values greatly influencing the converged ML estimates (Maitra, 2009b). The need for effective initialization becomes more acute with increasing
dimensionality and $K$ (Maitra and Melnykov, 2010). Therefore, the choice of initial values is an important determinant in EM's performance. In order to increase the possibility of convergence to a global maximum, we modify the Rnd-EM approach of Maitra (2009b) to initialize $\Theta^{(0)}$ in the presence of constraints. Exact specifics on the choice of the random initializing seeds are in Appendix B.1.2.

### 3.3.3. Parallelization of computations

Despite its wide general appeal and ready applicability here, the EM algorithm of Section 3.3.1 is plagued by slow convergence and can be time-consuming for typical 3D fMRI datasets with potentially up to a million voxels. While many speedup methods ( McLachlan and Krishnan, 2008) have been proposed, we incorporate the Alternating Partial Expectation Conditional Maximization (APECM, APECMa) algorithms of Chen and Maitra (2011) and Chen et al. (2013) that utilize extra parallelism on multi-core multi-processor computing architectures. Specifically, APECM and APECMa modify the Alternating Expectation Conditional Maximization (AECM) algorithm of Meng and van Dyk (1997) for mixture models by introducing a partial calculation of the E-step from the previous iteration (denoted as PE-step). This sort of parsimonious computing yields faster convergence rates of the EM and shorter computing times. Further, data-distributed algorithms, such as the Expectation Gathering Maximization (EGM) algorithm Chen et al. (2013) for distributed computing architectures can be coupled with APECM and APECMa to parallelize the M-step calculation (really the conditional M- or CM-step) when local sufficient statistics are available. We refer to Appendix B.1.3 for the exact details of our implementation of EGM with APECMa for our problem.

#### 3.3.4. Choosing $K$

Many methods (see e.g. Melnykov and Maitra, 2010) exist for assessing $K$ in a mixture model. The most popular are Akaike’s An (AIC) (Akaike, 1973) and Bayes (BIC) (Schwarz, 1978) Information Criteria (respectively, selecting $K$ to be the one minimizing $B_K = -2\ell(\Theta_K; p, v) + 2dK$. Here $\Theta_K = (\hat{\theta}_k, \Gamma_k) : k = 0, 1, \ldots, K$) contains MLEs of $\Theta$ obtained for a given $K$, and $dK$ is the number of unrestricted parameters to be estimated in the model. Our approach to choosing $K$ uses BIC, but additionally incorporates Kass and Raftery’s (1995) suggestion that only reductions of BIC beyond 10 providing very strong evidence in favor of the more parameter-rich (larger-$K$) model. Formally, we choose the number of components as the first $K$ for which $B_K \leq B_{K+1} + 10$, for $K = 0, 1, \ldots, K_{\text{max}} - 1$. We set $K = K_{\text{max}}$ if no $K$ satisfies this condition.

#### 3.4. Merging of components and voxel classification

##### 3.4.1. Merging with inactive component

Having determined $K$, we obtain an initial maximum a posteriori (MAP) classification of each voxel using the mixture model and estimated MLEs. Constraint (iii) of (1) on the activated beta components ideally means that for $k > 0$, we get beta components with parameter estimates $(\hat{\alpha}_k, \hat{\beta}_k) \hat{\alpha}_k/(\hat{\alpha}_k + \hat{\beta}_k) \leq \eta$. This last constraint on the expectation of the (activated) beta components logically follows from our expectation that these $p$-values are low for the activated components. We use $\eta = 0.05$ to match the most common threshold used in declaring an alternative hypothesis significant. However, the spatial penalty can also yield beta components with true mean $\alpha_k/(\alpha_k + \beta_k) \geq \eta$, because the spatial context does not involve the constraint (iii) or the observed $p$-value. We merge such components with the inactivated (zeroeth) component via a hypothesis test of

$$\log \frac{\Lambda_k}{\sum_{k=1}^K \log \left( \prod_{i=1}^n \left[ b\left(p_i; \hat{\alpha}_k, \hat{\beta}_k\right) \right] \right)} \leq \frac{1}{2} K \log \frac{\chi^2_{K-1}}{\chi^2_{K-1}}$$

against $H_0 : \alpha_k/(\alpha_k + \beta_k) \leq \eta$. We use the likelihood ratio test statistic (LRTS) $\Lambda_k = -2 \sum_{i=1}^n \log \left[ b\left(p_i; \hat{\alpha}_k, \hat{\beta}_k\right) \right] - \log \left[ b\left(p_i; \alpha_k, \beta_k\right) \right]$ where $n_k$ is the number of voxels classified to the $k$th component, $p_{ik}$ are the $p$-values of those voxels, $\alpha_k$ and $\beta_k$ are MLEs under $H_0$ and $\hat{\alpha}_k$ and $\hat{\beta}_k$ are MLEs under $H_0 \cup H_a$. The $p$-value of each of these $K$ LRTSs is calculated using $\Lambda_k \sim \chi^2_{K-1}$, then converted to $q$-values to control for false discoveries (Benjamini and Hochberg, 1995). Components for which we fail to reject $H_0$ are merged with the inactivated component. We write $\hat{K}$ as the number of activated comments remaining after these mergers.

##### 3.4.2. Merging of active components

Similar to the case of inactivated components being possibly split into multiple groups because of spatial constraints, some of the remaining activated components are possibly identified as distinct only because they are spatially disjoint. We distinguish between spatially disjoint pairs of activated components with similar effect sizes from those that have distinct effect sizes. For each pair of activated components, we compared the parameters of the beta distributions representing the $p$-values of the components. Specifically, we tested the $K$th and $(K+1)$th active component means using $H_0 : (\hat{\alpha}_k, \hat{\beta}_k) = (\hat{\alpha}_l, \hat{\beta}_l)$ against the alternative $H_a : (\hat{\alpha}_k, \hat{\beta}_k) \neq (\hat{\alpha}_l, \hat{\beta}_l)$ using a LRTS which is given by $\Lambda_{kl} = \sum_{i=1}^n \log \left[ b\left(p_i; \hat{\alpha}_k, \hat{\beta}_k\right) \right] + \sum_{i=1}^n \log \left[ b\left(p_i; \hat{\alpha}_l, \hat{\beta}_l\right) \right] - \sum_{i=1}^n \log \left[ b\left(p_i; \hat{\alpha}_k, \hat{\beta}_k\right) \right] - \sum_{i=1}^n \log \left[ b\left(p_i; \hat{\alpha}_l, \hat{\beta}_l\right) \right]$ where $p_{ik}, i = 1, 2, \ldots, n$, is the set of $p$-values at the $n_i$ voxels classified to the $r$th component, $r \in (k, l)$, $(\hat{\alpha}_k, \hat{\beta}_k)$ is the MLE of $(\alpha_k, \beta_k)$ obtained from fitting a beta distribution to the $p$-values classified to the $r$th component, and $\alpha_k, \beta_k$ are the parameter MLEs obtained upon fitting a beta distribution to the combined sample of $p$-values. Under the null hypothesis, the asymptotic distribution of $-2\Lambda_{kl} \sim \chi^2_{2}$. To account for the $\binom{\hat{K}}{2}$ pairwise tests, we used $q$-values (Benjamini and Hochberg, 1995) larger than 0.05 to merge pairs of spatially separated active clusters with indistinguishable beta distribution parameters.

#### 3.4.3. Voxel classification

Voxels are initially classified into the initial $\hat{K}$ groups in terms of the highest MAP. After the potential mergers of Section 3.4, the voxels in the merged groups provide a final classification of each voxel as inactivated or activated (in one of the merged, activated groups). The (merged) activated groups represent activated voxels of different intensities and correspond to different strengths of activation.
3.5. Extension to a two-sided testing framework

Our methodology is easily extended to situations involving two-sided \( p \)-values. In this situation, we consider the two-sided \( p \)-values (for the left and right tail) individually, and run our methodology, with \( \delta \) replaced in each case by \( \delta_l \) and \( \delta_r \) depending on the practitioner’s minimum \textit{a priori} expected proportion of inactive voxels in the right and left tail. If the practitioner can only provide an overall minimum \textit{a priori} expected proportion of inactive voxels (\( \delta \)), then we suggest using \( \delta_l = \delta_r = 1 - (1 - \delta)/2 \). We illustrate and evaluate performance using the above methodology on 26 Flanker task datasets in Section 3.2.3.

3.6. Activation detection in group studies

Our methodology is really intended for single-subject studies, however, it can be easily adapted to apply to activation detection in multi-subject group studies. In this case, we simply apply our methodology to the map of the \( p \)-values of the SPM obtained from the group study.

In this section, we have developed and fine-tuned statistical methodology for application to detecting activation in fMRI studies. We now evaluate its performance in realistic simulation experiments before applying it to our sports imagination experiment dataset.

4. Performance Evaluations

We very comprehensively evaluated our activation detection methodology. Our evaluation setups included simulation experiments in realistic 2D and 3D settings, and on 81 real datasets from four different task paradigms. The performance of our activation detection methodology (denoted as MixfMRI or Mf in our figures) was compared with a host of commonly used alternatives, that either have software available or are easily coded, and, because our evaluations are over hundreds of simulated and real datasets, are computationally practical to use, implement and evaluate. Consequently, we do not include comparisons with Bayesian methods, such as those in (Smith and Fahrmeir 2007) that use computationally expensive MCMC methods and were demonstrated only in 2D, and for which code is unavailable. Our comparisons only use the enhanced versions of methods if the improvements have been shown to improve activation detection: for instance, we use pTFCE instead of TFCE (Spisák et al., 2019), and the fully-automated FAST methods (Almodovar-Rivera and Maitra, 2019) instead of structural adaptive segmentation (Polzehl et al., 2010). An added benefit here is the ready availability of software for both pTFCE and FAST (while the R package fmri implements structural adaptive segmentation, the software when applied to our datasets threw up errors that were not resolved even after communicating with the authors). We note also that (Almodovar-Rivera and Maitra, 2019) showed poorer performance of both TFCE (Smith and Nichols, 2009) and adaptive segmentation (Polzehl et al., 2010) in their extensive simulation and real data experiments. Therefore, for comparison, we use (a) thresholding, after controlling for false discoveries (Benjamini and Yekutieli, 2001) at 0.05 (abbreviated as FDR in our figures), (b) thresholding computed by using Random Field theory (abbreviated using RF) to get a specified \( p \)-value of 0.05 (Worsley, 1994), (c) cluster-based thresholding with a first-order (CT-1st), second-order (CT-2nd) or (and in the case of 3D) third-order (CT-3rd) neighborhoods, all with cluster sizes decided by AFNI and a thresholded \( p \)-value of 0.001, following (Woo et al., 2014), (d) the three sets of FAST methods (Almodovar-Rivera and Maitra, 2019) with parameter \( \alpha = 0.01 \) or 0.05: AM-FAST (AM, \( \alpha = 0.01 \) or AM, \( \alpha = 0.05 \) in our figures, AR-FAST (AR, \( \alpha = 0.01 \) or AR, \( \alpha = 0.05 \)), and ALL-FAST (ALL, \( \alpha = 0.01 \) or ALL, \( \alpha = 0.05 \)), (e) permutation testing at \( \alpha = 0.05 \) (“Permutation” in our figures), and (f) the spatial mixture model method (or SMM in our figures), of Hartvig and Jensen (2000). We also evaluated performance, in the case of our 2D experiments, using our methodology that included the proportion-of-inactives (\( \pi_0 \geq \delta \)) constraint, but not the spatial constraint (denoted as MfnoX in our figures). Therefore, including the methodology developed here, and with different specifications of \( \delta \), our comparisons are over a total of 19 methods. Our methodology, with or without spatial constraints, was applied using our R package MixfMRI (Chen and Maitra, 2017), while we used our coded R scripts for FDR and permutation testing. The R package AnalyzeFMRI (Bordier et al., 2011) was used for RF, cluster-wise thresholding and spatial mixture modeling, while we used pTFCE (Spisák et al., 2019) for pTFCE, and RFASTfMRI (Almodovar-Rivera and Maitra, 2019) for the FAST methods. For all methods, activation detection accuracy was numerically assessed by the Jaccard index, \( J \) (Jaccard, 1901; Maitra, 2010) that is the ratio of the number of voxels that are both identified as active and truly so and the number of voxels that are either truly active, so detected or both. The index \( J \) takes values in [0,1], with 0 indicating that no voxels are correctly identified as activated and 1 indicating perfect detection. Finally, we evaluated the unique ability of MixfMRI to detect the different kinds of activation by calculating the adjusted Rand index (Hubert and Arabi, 1985) between our estimated activation maps and the ground truth in our simulation experiments. This index, or ARI, takes values in (\(-\infty, 1\)] with the upper limit indicating perfect predicted classification, and values farther away from 1 signifying poorer agreement of the classification with the truth. The ARI is expected to be zero when the voxels have been randomly classified.

4.1. Simulation Experiments

4.1.1. Two-dimensional framework

Experimental setup. We evaluated our methodology on digitized 2D phantom (Yardi et al., 1985; Maitra and O’Sullivan, 1998) datasets commonly used in Positron Emission Tomography experiments and specifically redesigned by us to mimic \( p \)-values typically obtained in fMRI experiments. Our setup is in 2D to permit a very thorough and detailed large-scale analysis, while also being realistic enough to capture the challenges in 3D settings. Our phantom (Fig. 2) had different-sized and oriented digitized ellipses, each representing hypothesized structures in the brain with some of them presumed activated by stimulation. We considered three scenarios. The first case (Fig. 2a) had very little (0.9% of all “in-brain” pixels) true ac-
Fig. 2: The phantom with three areas of simulated activation that was used in our experiments. In case (a), the phantom has very few areas of true activation while in case (b), the phantom shows activation only on one side (and with smaller spatial extent) while in (c) the phantom shows activation that is bilateral (on both sides).

 Activation. The second case (Fig. 2b) had 2.3% truly activated in-brain pixels, with all activation on one side of the brain, as would happen, say, in a one-hand finger-tapping experiment. The third case (Fig. 2c) had 3.97% of truly activated in-brain pixels with activation that is symmetric on both sides (as expected, say, in a bilateral finger-tapping experiment.) In each case, two regions (shown in darker colors) of differing strength are truly activated. Thus, the true $\pi$ is (0.991, 0.005, 0.004), (0.977, 0.019, 0.004) or (0.96, 0.034, 0.006) for the scenarios in Fig. 2. For all cases, $K = 2$ is the number of active components when ignoring spatial context. However the activated components span multiple homogeneous spatial regions. Ignoring spatial context, each of the two activated components corresponds to one of the two non-uniformly distributed components of our generative mixture model whose choice we now discuss.

For simulating the $p$-values for our experiments, we forwent the beta distributions for the $p$-values of the activated regions, and instead chose the distribution of the $p$-value of a one-sided normal test statistic: $\psi(p; \nu_k) = \Phi(\Phi^{-1}(1 - p); \nu_k, 1)/\Phi(\Phi^{-1}(1 - p))$, where $\Phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal distribution and density functions, and $\Phi(\cdot; \nu, 1)$ is the normal density function with mean $\nu$ and variance 1. (For inactivated pixels, $v_0 = 0$ and $\psi(p; v_0)$ is the uniform density. We reiterate that $\psi(p; v_1)$ and $\psi(p; v_2)$ are the densities corresponding to the two activated components in our generative model.) We use $\psi(p; v_k)$ rather than $h(p; \alpha_k, \beta_k)$s in simulating our datasets in order to gauge performance when the $p$-value mixture distributions are only approximately modeled by (1). Therefore, our choice of simulation model does not match the model on which our estimation methodology and software are built, but as a result, provides us more real-world performance assessments. Note also that if the FAST methods, pTFCE and spatial mixture modeling are meant for use on Gaussian SPMs, so the generative model may be considered to be somewhat more favorable for their frameworks. For these methods which detect activation using SPMs, we converted each of our simulated $p$-values to the corresponding upper standard normal quantile before application.

We performed experiments in scenarios ranging from easy (typical in low-level, easily-differentiated motor task experiments) to difficult activation detection scenarios (as in higher-level tasks requiring finer levels of cerebral processing). We obtained these scenarios by relating the values of $v_0 \equiv 0$, $v_1$ and $v_2$ to the pairwise overlap measure $\omega$ (Maitra and Melnykov 2010) calculated between the unique pairs of densities characterized by $(v_0, v_1, v_2)$. Specifically, given $v_0 = 0$ and for each of the $\pi$s in Fig. 2, we chose $(v_1, v_2)$ such that the overlap between every pair of the three mixture components was set to be $\omega$ for $\omega \in \Omega = (0.01, 0.1, 0.25, 0.5, 0.75, 0.95)$. Thus, our simulated mixtures had components ranging from mildly well-separated ($\omega = 0.01$) to essentially indistinguishable ($\omega = 0.95$). We call $\omega$ the identification complexity parameter and regard it as a surrogate for the difficulty of the particular activation detection problem, with higher values reflecting greater difficulty in detecting activation. Figs. C.13, C.14 and C.15 display sample mixture distributions and simulated statistical parametric maps (SPMs) for three $\omega$s, and for each of the three phantoms. These SPMs were obtained from our simulated datasets by using the inverse standard Gaussian cumulative distribution function on the simulated $p$-values at each voxel.

We simulated 25 datasets at each $\omega \in \Omega$ for each phantom, and obtained activation maps using our APECMa-EGM algorithms and its competitors. Recognizing that a practitioner may not (be able to) specify $\delta$ with complete accuracy, we evaluated performance of our method, and its counterpart that ignored the spatial penalty, with $\delta \in (0.95, 0.975, 0.99)$ to evaluate sensitivity of our results to such potential $\delta$-misspecification. For all experiments, we set $K_{\max} = 8$.

Simulation results. We now detail and discuss our results.

Sample illustrations Fig. 3 illustrates activation detected using our method on sample realizations (see Figs. C.13, C.14 and C.15) of each of the phantoms in Fig. 2 and with identification complexity parameter $\omega$ ranging from the easy ($\omega = 0.01$) to the substantially ($\omega = 0.5$) to very difficult ($\omega = 0.75$). Fig. C.13 also has activation detected using (where applicable, the best-performer among) the other methods for the simulated dataset corresponding to Fig. 2h, while Figs. C.14 and C.15 have the corresponding results for Figs. 2b and 2c. (AM-FAST generally does poorly in many of these cases, so is excluded in these individual displays for compactness.) Our approach is generally a top performer in each of the examples, and this better performance is sustained even with increasing activation detection difficulty ($\omega$). However, in these examples, the SMM of Hartvig and Jensen (2000) is often the best performer. Cluster thresholding (CT-2nd or CT-1st, that is not shown), and FDR also do excellently, but under very low $\omega$. With very localized true activation (corresponding to the phantom in Fig. 2h), ALL-FAST does relatively well even when $\omega$ increases (Fig. C.13). For the other phantoms, pTFCE also does relatively well with low $\omega$, but only SMM, ALL-FAST and AR-FAST are competitive with higher identification complexity. Despite the good $\delta$s, we see that the FAST methods overestimate the spatial extent of activation. Permutation-testing is a poor performer across the board, in these examples. In general, many of the alternative methods degrade quicker in performance than our method. Further, setting only the $\delta$ constraint provides good performance,
but not always, and in any case, the improvement is not as much as that obtained when also including the spatial penalty. Finally, the unique ability of our methodology to potentially detect the different kinds of activation shows some promise. The results displayed here are on one candidate realization at the sample settings, so we now report performance of all the methods in our large, comprehensive simulation study.

**Comprehensive evaluations**  
Fig. 4 summarizes the ability of MixfMRI to identify the different kinds of activation in each phantom for each \( \delta \) and \( \omega \).

Fig. 3: Activation obtained using our method on sample realizations obtained using different \( \omega \) for the three phantoms in Fig. 2.

Fig. 4: The adjusted Rand index summarizing our method’s ability to identify the different kinds of activation in each phantom for each \( \delta \) and \( \omega \).

MixfMRI in recovering the different kinds of activation is encouraging. Most other methods do not have this ability, mostly providing only a binary classification of each voxel as activated or not, so we now evaluate all the methods in terms of \( \mathcal{J} \).

Fig. 5 summarizes the results of all the two-dimensional simulation experiments, with Fig. 5a displaying the performance of all methods on the three phantoms and under different \( \omega \)s through the \( \mathcal{J} \)s. Fig. 5b provides the \( q \)-value of the paired Wilcoxon signed rank test for whether the \( \mathcal{J} \)s obtained by MixfMRI with a given \( \delta \) is different from any of the competing methods. (The display differentiates between when MixfMRI is better or worse, and shows that we are significantly better (dark greens) than the competitor in most cases.) For easy problems (\( \omega = 0.01 \)), Fig. 5a repeats the trends shown in Fig. 3 and does not show much important distinction between our methodology and the best-performers such as CT-1st or CT-2nd (which report near-perfect \( \mathcal{J} \)s), or the spatial mixture model approach was the best performer in these 2D experiments in all but for the highest \( \omega \)s.

Interestingly, Almodóvar-Rivera and Maitra (2019) suggested using \( \alpha = 0.01 \) and \( \alpha = 0.05 \) for low- and high-noise situations, but in these experiments, the FAST methods with \( \alpha = 0.01 \) generally had higher \( \mathcal{J} \)s than those obtained using \( \alpha = 0.05 \). While all other methods degrade with increasing identification complexity, our method performs better relative to the alternatives. Even with \( \omega = 0.75 \), the median \( \mathcal{J} \) was very good for our methods, though there was greater variability in performance. The spatial mixture model approach was the best performer in these 2D experiments in all but for the highest \( \omega \) (and in many cases, significantly better than our methods). As in the case of the sin-
(a) The Jaccard index summarizing results of all methods on all replicates in our 2D simulation experiments. The row panel indicates the phantoms (indexed by the true \( \pi_0 \)). Each group of boxplots corresponds to a given identification complexity (\( \omega \)), with the methods in each group in the same order as the labels in the figure.

(b) The \( q \)-value obtained from the two-sided paired Wilcoxon signed rank test of the \( \chi^2 \)s obtained using our method (on the horizontal axis, for a given \( \delta \)) and each competitor (on the y-axis) for each phantom and experimental setting (here \( \omega_\alpha \) denotes identification complexity of \( \alpha \)). In all cases, darker values indicate significant differences between the competing methods. The yellow-to-green map displays the \( q \)-value for the cases when our method is better than a given competitor for a specific experimental setting, while the yellow-to-red map displays the \( q \)-value for the cases where our method is worse. (The \( q \)-value scale in our display is piecewise linear, with a steeper gradient in \([0,0.05]\) than in \([0.05,1]\)).

Fig. 5: Comprehensive summary of the results of our 2D simulation experiments. The methods are in the same order in both figures and use the abbreviations outlined at the beginning of Section 4.
ingle realization case earlier, permutation testing was the worst performer across all ωs and phantoms. Finally, including the spatial penalty in our method improves performance over our method implemented without the spatial penalty, so for concision and ease of presentation, we drop displays and discussion on the case that only involves the proportion-of-inactive-voxels constraint and not the spatial penalty.

Our comprehensive evaluations on 2D simulated single-subject scenarios show competitive performance of our methodology that incorporates both a priori expected proportion of activated voxels and spatial context. Our methodology is robust and a good performer even under high identification complexity, which portends well for its use in experiments involving high-level cognitive tasks, as in the case of our showcase application. Additionally, as seen in Figs. 3 and 4, we can identify the different intensities of activation fairly well. We now evaluate performance in realistic 3D simulation settings.

4.1.2. Three-dimensional framework

Our 3D simulation experiments used the neuRosim (Welvaert et al., 2011) R package that simulates time series fMRI data according to a desired experimental paradigm. We used the simulation setup presented in Welvaert et al. (2011) that mimics the real-life repetition priming experiment of Henson et al. (2002), with the exception that the radii of the five activated regions on Page 12 of Welvaert et al. (2011) were each decremented by one to match a more realistic proportion (2.74%) of in-brain activated voxels (and that in the real-life study in Henson et al. 2002). We used four different SNRs of 0.9925, 1.785, 3.87, and 7.74 to characterize varying levels of noise that can impact activation detection ability. Our evaluation was on the SPM formed by the faces-versus-baseline contrast, with the SPM converted to a p-value map where needed, such as for our methods. For each SNR, we simulated 25 four-dimensional fMRI time series datasets, and obtained the corresponding 25 SPMs and p-value maps. Activation was detected using our method (with \( K_{\text{max}} = 11 \)) and its competitors. Fig. 6 shows that the ARIs are modest under very low noise, and quite high in all other cases, indicating good ability of our method in determining the different kinds of activation. Fig. 7a displays the \( \mathcal{F} \)s obtained using our method and its competitors, with the significance of the differences between our method and its competitors.
summarized in Fig. 7. The ALL-FAST methods are the best performers among the competitors (and in a few cases even significantly better than our methods for the lowest SNR), but not significantly different (even if slightly worse) than our methods for moderately low SNR. Also, SMM is worse, but not significantly so, than our method for very low SNR. Further, while pTFCE is significantly better than our method at the highest SNR level, Fig. 7a indicates that this difference in performance may not be significant. Against all other methods, our method is a substantially better performer, and significantly so. AM-FAST and permutation testing are the worst performers at all SNRs. An interesting observation is the poorer performance of SMM than in the 2D simulation studies earlier.

Our experiments show good performance of our methodology for a large range of realistic 2D and 3D simulation experiments. In some cases, we are bettered by the competitors, with, and especially in 3D settings, the differences being sometimes significant but not always important. On the other hand, we are better in most cases, and in many of these cases the differences are both significant and substantial enough to be important. Our methodology is also seen to do fairly well in detecting the different kinds of activation. We now evaluate performance of the method and its competitors over some real data examples.

4.2. Performance on Real Data Examples

Our next tranche of evaluations is on 81 datasets from three sets of fMRI studies. We first evaluated the performance of our algorithm in null activation resting state scenarios for each of 31 typically developing children (Nebel et al. 2014). The second set of experiments was from the right- and left-hand finger-tapping study (Maitra et al., 2002; Maitra, 2009b, 2010; Almodovar-Rivera and Maitra, 2019) of a normal male volunteer over 12 sessions spanning two months. The final set of experiments were from a Flanker experiment (Kelly et al., 2008) on 26 right-handed adults. For all but the spatial mixture model and the FAST methods, we spatially smoothed the SPMs obtained in each of the experiments using the automated robust method of Garcia (2010) that uses the generalized cross-validation to adaptively estimate the smoothing parameter. These smoothed maps were used to obtain the p-value maps of activation or significance at each voxel. We do not apply smoothing to the SPM for the FAST methods where smoothing is an integral part of the adaptive smoothing and thresholding algorithm. Similarly, Hartvig and Jensen (2000) also express concerns about smoothing of the SPM before activation detection, and so we report on their results obtained without prior smoothing. (Each of these methods were also seen to perform better without prior smoothing of the SPM, while the other methods performed better with the prior smoothing of the SPM.) Further, there is no ground truth available for the finger-tapping and the Flanker task experiments, so for each experiment, we used the summarized Jaccard index (\( \hat{J} \)) of Maitra (2010) to measure consistency in the activation identified between the different replicates. The summarized Jaccard index, proposed by Maitra (2010), for evaluating consistency in several activation maps, first constructs a matrix \( \Omega \) of the Jaccard index between every pair of activation maps and then calculates \( \hat{J} \) over all the (M) activation maps in terms of its largest eigenvalue \( \omega_{(1)} \). Formally, \( \hat{J} = (\omega_{(1)} - 1)/(M - 1) \). It is easy to see (Maitra 2010) that \( 0 \leq \hat{J} \leq 1 \) with \( \hat{J} = 0 \) when there is no agreement at all between any of the maps, because then \( \Omega \) is an identity matrix. Further, \( \hat{J} = 1 \) when all the activation maps are in complete agreement with each other. We now describe and discuss the results in each experimental study.

4.2.1. Resting state datasets

The data used in this set of experiments are from the resting state fMRI scans of 31 typically developing adolescent children that were acquired using a single-shot, parallel (SENSE) gradient-recalled echo planar sequence (TR = 2500 ms, TE = 30 ms, and a flip angle of 70°) over 156 time points, with the image volumes at each time-point having axial slices of 3mm each and no slice gap. We used AFNI (Cox, 1996; Cox and Hyde, 1997; Cox, 2012) to pre-process the dataset and then fit a general linear model with autoregressive errors for a purported auditory experiment to the data. The SPMs obtained from each dataset, and their associated p-values were then submitted to our methodology (\( K_{\text{max}} = 11 \)) and its competitors to obtain activation maps. Fig. 8 provides a heatmap of the proportion of voxels identified as activated in-brain by each method. (For our method,
we only report performance with $\delta = 0.95$ since our results obtained using other values of $\delta$ were identical.) No activation was detected, correctly, using our methodology, RF, pTFCE, cluster thresholding, or FDR. The spatial mixture model approach correctly found no activation in 30 of the datasets, but also was unable to provide a solution for the remaining dataset. The FAST methods find varying amounts of activation in at least some of the SPMs, while permutation testing detected a very small proportion of active voxels in every dataset. Eklund et al. (2016) has stressed the importance of evaluating the false positive rate of activation detection algorithms on resting state datasets: we find it encouraging that our method is among those that passes this test here with each of the 31 SPMs.

4.2.2. Finger-tapping experiments

The next set of experiments involved the 12 replicated SPMs of Maitra et al. (2002) from the right-hand and left-hand finger tapping study of a right-hand-dominant male adult. Fig. 9 displays the $\tilde{J}$ obtained for the 12 replications on each hand.

![Image](image_url)

Fig. 9: The summarized Jaccard index ($\tilde{J}$) for assessing the similarity in the activation maps obtained by each method for the 12 right- and left-hand finger tapping experiments. Because of concerns (Maitra, 2010), that the right-hand dominant male may have been tapping his right-hand fingers in the eighth dataset, $\tilde{J}$ is also obtained for activation maps from the left-hand datasets sans the eighth replication.

For this experiment, we used $K_{\text{max}} = 5$ and higher values of $\delta = 0.975, 0.985, 0.995$ to reflect our view that this is a more experiment with few and very focused activation areas (and around 1-2% expected proportion of activated voxels). For the right-hand experiments, our method at $\delta = 0.975$ is marginally worse ($\tilde{J} = 0.248$) than that obtained using CT ($\tilde{J} = 0.252$) or pTFCE ($\tilde{J} = 0.259$), however, our results obtained using $\delta = 0.985$ ($\tilde{J} = 0.283$) and $\delta = 0.995$ ($\tilde{J} = 0.308$) are substantially better than all competitors. For the left hand, we see that our method outperforms all its competitors. Maitra (2010) has expressed some concern about the quality of the data in the eighth left-hand experiment replicate, therefore we also evaluated $\tilde{J}$ of the results after excluding this experiment. The $\tilde{J}$ for almost all methods improve, but even here, MixfMRI is by far the best performer.) For these datasets, cluster thresholding does quite well, and is the next-best performer (after MixfMRI) in the left-hand experiments. The spatial mixture model approach performs modestly on these experiments, and is mildly to substantially (for the right-hand experiments) worse than FDR. AM-FAST, AR-FAST, ALL-FAST and permutation testing are, in that order, among the worse performers. An encouraging aspect of the $\tilde{J}$s for our method is its consistency in the right hand, that indicates that a higher restriction on $\delta$ identifies more similar areas of activation in each map. For the left-hand experiments, $\tilde{J}$ goes up from $\delta = 0.975$ to $\delta = 0.985$, but dips slightly from there to $\delta = 0.995$, potentially attributed to the fact that this is a left-hand finger-tapping experiment performed by a right-hand dominant male. However, whether with the right or the left hand, our methodology is overall the better-performing method. We also comment that the generally low values of $\tilde{J}$ for all methods provide support to concerns Maitra (2010) Almodóvar-Rivera and Maitra (2019) on potential issues in quality and preprocessing of this dataset.

4.2.3. Flanker task experiments

The Flanker task data are from a study Kelly et al. (2008) on 26 right-handed adults who were vicenarians or tricenarians, and were imaged using echoplanar imaging (TR=2000 ms, TE = 30 ms, flip angle = 80°) over 146 time-points to get, 146 64×64×40 image volumes of voxel size 3×3×4 mm$^3$, while participating in 12 congruent and 12 incongruent Eriksen flanker task trials, that were presented to them at varying intervals of 8 to 14 seconds, in pseudo-random order. We used AFNI to first preprocess the dataset to a common MNI template for each subject, and then fit a general linear model with the appropriate design matrix and first-order autoregressive moving average errors. From each dataset, we obtained a SPM that captured the contrast between the congruent and the incongruent task, and obtained the corresponding $p$-values at each voxel. This is an example of an experiment with a two-sided alternative hypothesis, and our method is applied using the additional development of Section 3.5. Once again, there is no ground truth for comparison, but we note that although these SPMs are for different subjects, they are all of normal younger adults and in standardized space, and so we evaluate performance by calculating $\tilde{J}$ between the activation (or significance) maps obtained on each SPM by each method. Here, the activation map specifies the significance or otherwise of a contrast, as determined by the applied method. (Here, we evaluate MixfMRI with $K_{\text{max}} = 11.$
and using \( \delta = 0.9, 0.95, 0.975 \), with the higher values reflecting the circumstance that this is a two-sided hypothesis. Also, memory requirements in the AnalyzeFMRI implementation of cluster thresholding precluded us from evaluating any of the CT-1st, CT-2nd, or CT-3rd methods so we do not include these methods in our comparisons.) Fig. 10 summarizes the results

**Fig. 10:** The summarized Jaccard index for assessing the similarity in the significance maps obtained by each method for the 26 Flanker task experiments.

and shows that our methods yield \( \hat{J} \)-values that are almost twice as good as its closest competitor (pTFCE). The performance of FDR, spatial mixture model and the AR-FAST approaches is middling, while permutation testing and the other FAST methods perform poorly. It is unclear if the very marginal decrease in the consistency of the significance maps with increasing \( \delta \) are themselves significant, but if so, may point to the difficulty of consistency identifying significance in higher-level tasks. However, we see that we are by far the best performer in this two-sided hypothesis testing scenario.

### 4.3. Summary of Results

The results of our experiments on both simulation and real data examples show good performance of our MixfMRI methodology developed in this paper. In both 2D and 3D simulation setups, our method is always a top performer. The superlative performance of the spatial mixture model in the 2D simulation experiments (for all but the highest model complexity) does not quite translate to the 3D simulation setup, and even more so, to the real-data 3D examples, where we are almost always the best method. Our method is also among those that correctly identify no activation in all 31 resting state datasets. Among the other methods, the second-best performer is unclear in the 3D simulation experiments, but is pTFCE in the case of the finger-tapping and Flanker task real-life data experiments. Interestingly, and in general, the performance of permutation testing for detecting activation was anemic at best – a finding also reported in Almodovar-Rivera and Maitra (2019) and somewhat contrary to that of Ekland et al. (2015).

We close this section with a brief discussion on the specification of \( \delta \), the constraint on which is an integral part and major contribution of our work. Our simulation studies indicate that the choice of \( \delta \) slightly affects performance for all three 2D phantoms and identification complexities, with greater relative (albeit marginal) inaccuracy when \( \delta \) is misspecified downwards (i.e. \( \delta \) is less than the true \( \pi_0 \)) than with \( \delta \) is specified to be greater than the true \( \pi_0 \), where it still provides quite accurate identification. Given that it is impractical to expect a practitioner to precisely specify \( \delta \), it is encouraging to note the robustness of our method to the exact choice of \( \delta \).

### 5. Activation in sports imagination experiment with implications in treatment and therapy of PVS patients

We now re-analyze the sports imagination experiment dataset of Section 2. Recall that the value of this experiment is in providing for a way to communicate with, and provide therapy and treatment for PVS patients (Owen et al., 2006), who because of TBI or other reasons, may have individualized brain structure and function that needs to be accurately identified. We therefore evaluate whether we can, as a follow-up to our very encouraging results of Section 4, accurately recover activation in a known normal subject where the results can be explained and interpreted against known cerebral structure and function, and where Section 2 has suggested unclear results with current standard analysis methods. We revisit this dataset by applying our refined methodology to the voxel-wise \( p \)-values obtained, as described in Appendix A.2. Here the most plausible value of \( \delta \) is 0.99 but we also try \( \delta = 0.975 \) and 0.999 to check the effect of \( \delta \)-misspecification. With \( \hat{K} \max = 20 \), BIC (and ICL-BIC) chose \( \hat{K} \) = 11, 12 and 6 for solutions using \( \delta = 0.975, 0.99 \) and 0.995. In either case, the inactive component did not merge with any of the other groups in the procedure outlined in the first part of Section 3.4. Using the second part of Section 3.4, the activated components merged to form four, five and four clusters with \( \delta = 0.975, 0.99, 0.995 \), respectively. We now discuss and interpret the results.

#### 5.1. Analysis and interpretation of results

We focus our discussion on results obtained using \( \delta = 0.99 \). Fig. 11 summarizes the results and indicates the five distinct regions identified as activated. We discuss these regions numbered (i) through (v), in decreasing order, of their averaged \( t \)-statistic (highlighted in Fig. 11), recalling that the experimental paradigm is of a healthy female volunteer imaged alternately while at rest and while imagining playing tennis.

The region with the highest average (standardized) intensity of activation is among the smallest, and spans the calcarine sulcus and the inferior occipital cortex/lingual gyrus. The primary visual cortex – where dynamic and static information from the visual hemifields is organized coherently for later processing in
Fig. 11: (a) Activation detected using our method with $\delta = 0.99$ in a healthy female volunteer performing a sports imagination experiment. (b) The five activated regions, identified by their indicator color, the number of voxels ($\#v$), the average $t$ statistic and estimated beta distribution parameters for each region. In that order, the five regions span (i) the calcarine sulcus and the inferior occipital cortex/lingual gyrus, (ii) the PMC, (iii) the pre-cuneus, (iv) the cuneus, and (v) the SMA.

The cuneus – is primarily located in the calcarine sulcus, while the inferior occipital/lingual gyrus is important for visual memory, attention, and in orienting the brain to visual location of an existing or remembered scene. That these regions would be activated with high intensity makes sense given that the subject is tasked with imagining playing tennis, and needs to recall visual imagery of the activity and in processing information in pretend-planning and executing her moves. But for the PVC which is identified rather faintly, these regions are not identified using cluster-wise thresholding, pTFCE (see Fig. 11., b) or many other methods.

The next highest average $t$-statistic is a large region in the primary motor cortex (PMC) that executes functions planned by the pre-motor areas. This area also includes a somatotopic map comparable to the somatosensory cortex, and has been associated with mental motor imagery and sensorimotor planning in multiple fMRI studies involving imagined movements (Roth et al., 1996; Lotze et al., 1999; Stippich et al., 2002). Once again, the fact of these regions being activated makes sense in the context of the experimental paradigm, however, but for the tiny blip of activation in the somatosensory cortex, these regions were largely missed in cluster-wise thresholding or AR-FAST, and completely by methods other than pTFCE.

The region of activation with the third highest averaged $t$-statistic is in the pre-cuneus which is an area of the brain that is known to pre-process 3D space, color and other information while the next region includes the SMA that works in concert with the pre-SMA, the striatum, and other regions to collect prefrontally driven information about what the subject is doing and guides appropriate movement in space. This higher-order information is then sent to the PMC, which eventually directs motor movements. Once again, activation of these areas are in keeping with the experimental paradigm. Cluster-wise thresholding and pTFCE also identify this region, though the spatial extent of these regions is lower.

The fifth region is the smallest activated region and is mostly the Supplementary Motor Area (SMA). This region works in concert with pre-SMA, the striatum, and other regions to collect prefrontally driven information about what the organism is doing and where it should move in space to execute its goals. This higher order information is then sent to primary motor cortex, which eventually send movement commands to the limbs, face, and trunk. Once again, it is reasonable to expect activation in this region, considering the experimental paradigm, but AR-FAST, cluster thresholding and pTFCE miss this region entirely.

5.1.1. Activation detection in sports imagination experiment with $\delta = 0.975$ and $\delta = 0.995$

Fig. 12 displays activation regions detected by our method with $\delta = 0.975$ and $\delta = 0.995$. Compared to Fig. 11, both choices of $\delta$ result in one less region, but for different reasons. With $\delta = 0.975$, Fig. 12a shows one less region than Fig. 11 merging the cuneus with the calcarine sulcus and the inferior occipital cortex/lingual gyrus, while Fig. 12b has the SMA and PMC placed in the same activated group. So the results are not very dissimilar and exhibit robustness with regard to modest misspecification of $\delta$. 

| Region | Indicator color | $t_{\text{stat}}$ | $\#v$ | $\alpha$ | $\beta$ |
|--------|-----------------|-----------------|-------|---------|--------|
| (i) Purple | 3.736 | 178 | 0.524 | 404.09 |
| (ii) Sea green | 3.420 | 417 | 0.551 | 140.48 |
| (iii) Pink | 3.025 | 387 | 0.695 | 112.27 |
| (iv) Orange | 2.761 | 533 | 1.652 | 181.74 |
| (v) Green | 2.463 | 175 | 1.533 | 79.55 |
5.2. Significance of our findings

Our results with this imagination dataset indicate that our refinements that account for both spatial context and a priori constraints on the expected proportion of activated voxels can provide improved and interpretable activation detection in a single-subject sports imagination experiment. Specifically, when compared to the results from cluster-wise thresholding, pFTCE or AR-FAST (Fig. 1) that only identified the first three of our regions and did not identify the cuneus which preprocesses information on 3D space and color, or the SMA that gathers prefrontal information to send to the PMC, our identified activation regions are clear and interpretable in the context of the experimental paradigm, providing confidence in our findings even though they are derived from a single subject. This demonstrated ability of our approach to identify these regions, each with different kinds of activation, is a major strength relative to all existing methods. This subject was a normal healthy female and so while her activation maps are interpretable in the context of the study, we would not a priori know this, for example with a TBI survivor or a patient in PVS, or more generally in a clinical setting. The original goal of sports imagination experiments in patients in PVS was to facilitate therapy and provide ways of communication with them but, as explained in Section 2 this benefit has not been realized because of murky activation detection even in group PVS patient experiments (Bardin et al., 2011). It is essential to have reliable activation detection at the individual subject level in order to permit therapy and communication with PVS patients and the development and analysis of this paper show promise in making this goal possible. More generally also, our refinements and improvements can improve activation detection at the individual level, making it possible to adopt fMRI in a clinical setting.

6. Discussion

We provide an improved activation detection approach in task-based fMRI that incorporates spatial context and also the proportion of voxels a priori expected to be activated as per the experimental paradigm. Our statistically sound model-based methodology is computationally practical and implemented in our MixfMRI R package. Simulation experiments on settings of low to very high detection difficulty indicated our methods to have superior performance over existing ones. Our methodology also correctly predicted no activation in null activation settings and was a top performer, in terms of consistency, or real-world datasets. When applied to a single-subject sports imagination experiment study, with deep implications in the treatment of PVS patients, the method provided interpretable results and improved activation detection, identifying regions that had not been activated by current methods. An additional benefit of our approach is the ability to differentiate intensities and kinds of activation. This augurs well for applying our methodology to clinical settings, such as with TBI survivors or subjects with impacted or individualized brains where accurate activation detection is essential to identify pathologies, treatments and therapy.

Our refinements are easily extended to situations beyond voxel-wise \( p \)-values of the SPM. For instance, our methodology could be easily extended to the SPM itself. Further, even in the context of \( p \)-values, we have developed our methodology using an approximate mixtures-of-beta mode, but we could easily incorporate more precise models, as for instance provided in Maitra (2009a). Activation detection could potentially be further improved by semi-supervised approaches that include known areas of inactivation, voxel-wise a priori activation probability maps drawn from normal subjects, or better processing such as including phase in the fMRI time-series data analysis (Adrian et al., 2018). Further, it may we worth investigating if the performance of competing methods, e.g., the spatial mixture model, can be further improved by incorporating a priori constraints on the expected proportion of activated voxels. Thus, we see that while we have provided a substantial contribution to accurate activation detection in fMRI, a number of interesting extensions and investigations remain that may benefit from further attention.

Acknowledgements

The authors are very grateful to B. Klinedinst and A. Willette of the Program in Neuroscience and the Department of Food Sciences and Human Nutrition at Iowa State University for help with the interpretation of Figure 11.

Funding

R. Maitra was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) under Grant R21EB016212. The content of this paper is however solely the responsibility of the authors and does not represent the official views of the FDA, the NIBIB or the NIH.

Data Availability

The dataset used in the 2D simulations is publicly available within the authors’ R package MixfMRI implementing the methods. Software to generate the 3D simulations is also publicly available via the neuRosim package in R. The sports imagination experiment dataset is publicly available at https://doi.org/10.18637/jss.v044.i11

Availability of computer code

The methodology developed in this article is implemented in publicly available R package MixfMRI publicly available as a “Contributed package” at https://www.r-project.org

Appendix A. Sports imagination experiment dataset: Additional details

Appendix A.1. Background

The time-course sequence of images were acquired on a MR scanner using a gradient echo-echo planar imaging sequence
with an echo time (TE) of 40 milliseconds, relaxation time (TR) of 2 seconds and an excitation flip angle of 80 degrees. At each time-point, the scanner acquired 30 axial slices of 4 mm thickness each and a field-of-view of 24 cm with an imaging grid of size 64 × 64, yielding in-plane voxel dimensions of 3.75 mm. Bardin et al. (2011) do not provide detailed discussion and specific analysis of this dataset, but it was later used (without pre-processing) and publicly released by Tabelow and Polzehl (2011) to demonstrate features of their fmri software package in R (R Core Team 2018).

Appendix A.2. Data Preprocessing

We used the umbrella sub.dartject.py in the Analysis of Functional Neuroimaging (AFNI) software (Cox, 1996; Cox and Hyde, 1997; Cox and Jesmanowicz, 1999; Saad et al., 2009) to pre-process, register (Cox and Hyde, 1997; Cox, 2012) to pre-process, register and analyze the dataset. The script consists of two main steps. First, the functional imaging data cube at all time points is registered with the anatomical data cube. Our registered and aligned 3D dataset had 46×55×43 voxels, of which 33,753 voxels were inside the brain. There were 105 time-points with TR=2s, and the time series at each voxel was fit using the model

$$Y_t = \beta_0 + \beta_1 s_t + \beta_2 d_t + \beta_3 d_t^2 + \epsilon_t$$

where, at the $t^{th}$ time-point, $Y_t$ is the observed BOLD response, $s_t$ is the expected BOLD response obtained by convolving the input stimulus and the Hemodynamic Response Function (HRF), and $d_t$ is the time drift parameter (that is assumed to have a quadratic effect on the observed BOLD response), and $\epsilon_t$ is the noise or error in the observation with an assumed autoregressive moving average ARMA(1,1) dependence structure. (Here, we have temporarily dropped the subscript for the voxel for readability but it is present in all the parameters and the random quantities in (A.1).) Our voxel-specific parameters allow us to account for local inhomogeneities – the voxel-wise fitting of the model is also computationally practical. At this stage, we do not account for any spatial structure in the model. Although all parameters are fit, our interest in this experiment is only in knowing if the parameter $\beta_1$ is positive or not. Framed in terms of a hypothesis testing problem, we test $H_0 : \beta_1 = 0$ v.s. $H_a : \beta_1 > 0$ at each voxel which means that there is a positive association between the observed BOLD response at that voxel and the expected BOLD response after accounting for other factors. All other parameters are nuisance parameters and of no primary importance. At each voxel, we test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 > 0$ which means that there is a significant voxel response to stimulus when $H_0$ is rejected. Thus, the t-statistics for testing the hypothesis of no activation against the alternative of positive association with the expected BOLD response are obtained at each voxel.

Because our analysis outlined above does not account for spatial context, the t-statistics were spatially smoothed for consistency using the automated robust method of Garcia (2010) which uses generalized cross-validation to adaptively estimate the smoothing parameter. The $p$-value of activation at each voxel was computed at each voxel from these test statistics. These obtained $p$-values were used to construct activation maps.

Appendix A.3. Constructing Activation Maps

The $p$-values at each voxel were thresholded to construct activation maps using some common approaches. These maps and the test-statistics at those voxels were then mapped using the SUnFaced MAPping tool (SUMA) of Saad and Reynolds (2012). We used FDR with $q = 0.05$ (Genovese et al., 2002; Benjamin and Yekutieli, 2001) and found no activation, highlighting the challenges faced by a global thresholding method such as controlling FDR for constructing activation maps for experiments such as this one. We also used cluster-wise thresholding methods to detect activation with AFNI’s 3DClustSim package to determine the minimum number of contiguous voxels in each activation region at a threshold of 0.001 (following Woo et al., 2014) using first-, second- and third-order neighborhoods.

Appendix B. Statistical Methodology: Additional Details

Appendix B.1. Parameter Estimation via The EM Algorithm

Appendix B.1.1. Proof of Result

Define the Lagrangian $L(\pi, \lambda, \lambda_1) = g(\pi; W) - \lambda_1 h_1(\pi) - \lambda_1 h_2(\pi)$ with convex set $\Pi$. Then $\pi^*$ is a local maximum if and only if there exists exactly one $\lambda_1^*$ satisfying stationarity (i.e. having $\nabla_\pi L(\pi^*, \lambda_1^*, \lambda_2^*) = 0$) and the Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vandenberghe, 2004). Note that the constraint functions $h_1(\pi)$ and $h_2(\pi)$ are both affine, so all KKT conditions are satisfied.

The stationarity condition for maximizing $g(\pi; W)$ yields $\pi^*$ satisfying the equations:

$$-\frac{1}{\pi^*_k} \sum_{i=1}^{n} w_{ik} \lambda^*_i + \lambda^*_i 1_{\delta k} = 0, \quad \text{for } k = 0, 1, \ldots, K. \tag{B.1}$$

The inequality constraint is inactive if $\lambda^*_k = 0$. This means that the local minimizer $\pi^*$ of $g(\pi; W)$ under the equality constraint ($h(\pi) = 0$) is a feasible solution for our problem. We now consider the case of the active inequality constraint, that is, $\lambda^*_k > 0$, for which we use the KKT conditions. Primal feasibility is itself defined by the constraint in (b). The dual feasibility condition stipulates that $\lambda^*_k \geq 0$ so with an active inequality constraint it is enough to focus on $\lambda^*_k > 0$. Combining this condition ($\lambda^*_k > 0$) with complementary slackness ($\lambda^*_k h_i(\pi^*) > 0$) yields $\pi^* = \delta$. From (B.1) we get $\pi^*_k = \sum_{i=1}^{n} w_{ik}/\lambda_1$ for each $k = 1, 2, \ldots, K$. From the constraint (a), we have $\pi^*_0 + \sum_{i=1}^{K} \pi^*_i = 1$ so that $\sum_{k=0}^{K} \pi^*_k = 1 - \delta$ under the active equality constraint. Thus, we get $\lambda^*_k = \sum_{k=0}^{K} \pi^*_k w_{ik}/(1 - \delta)$ from where we get $\pi^*_i = (1 - \delta) \sum_{k=0}^{K} \pi^*_k w_{ik}/(1 - \delta) \sum_{k=0}^{K} \pi^*_k = (1 - \delta) \pi^*_k/(1 - \pi^*_0)$ for $k = 1, 2, \ldots, K$. The result follows upon combining the cases with active and inactive inequality constraints.

Appendix B.1.2. Initialization and Classification

The Rnd-EM approach (Maitra, 2009a) chooses – from a pre-set number $M$ ($M = 50$ in this paper) of “valid” randomly-realized initial values $\Theta^*$ – the configuration maximizing $\pi$ and uses that to initialize EM which is then run till convergence. We now discuss the issue of choosing the initializing random realizations. We choose each random initializer one component
and utilize iteration. The CM-steps are performed in turn on the $i$th voxel in the volume and the choice of 0.5 is from the mean of the uniform distribution. For the initializing parameters of the beta distributions of the activated components, we randomly select $K$ other values from among the $p$-values of voxels of potential interest, for example, from voxels whose $p$-values were below some $p_{\text{max}}$ (say 0.05). The coordinates from each of these $K$ voxels are our initializers for these $K \mu_k$'s.

We now group each voxel in terms of the closest Euclidean distance of the observed $(p_i, v_i^T)^T$ to the potential initializing candidates. (Note that the voxel coordinates $v$'s are scaled to lie in $[0, 1]$, so that all voxel coordinates and $p$-values are of comparable importance in the Euclidean distance for grouping. For each group $k$, we get the initially assigned $p$-values and voxel coordinates $(p_j, v_j^T)^T$: $j = 1, 2, \ldots, n_k$, set $n_k = n_i/n$ to be the proportion of voxels assigned to the $k$th group and the remaining components of $\Theta_k$ from the maximizer of the log-density $\sum_i^m \log[b(p_j; \alpha_k, \beta_k) b(v_j; \mu_k, \Sigma_k)]$ under constraints (ii) and (iii).

We declare a candidate initializer “invalid” when a random initial grouping yields a single observation in some group, resulting in a degenerate initial $\Sigma_k$ for that group. (Invalid candidates are discarded from further consideration.) We select the best initial value as $\Theta^{(0)} = \text{argmax}_\Theta \ell(\Theta; p, v)$ from which the EM algorithm of the constrained mixture model is applied. We consider a MLE valid if the EM algorithm of Appendix B.1 converges to non-degenerate parameter estimates and leads to non-degenerate groups (i.e., each cluster has at least a few observations, say 1 + $c_v$). The converged result from the best initial value does not always lead to a valid MLE, so we repeat the Rnd-EM approach until we get a valid MLE. From the converged valid MLE $\hat{\Theta}$, we estimate $\hat{\omega}_k$ the maximum a posteriori (MAP) evaluated at $\hat{\Theta}$ and classify the $i$th voxel to the $k$th component where $l = \text{argmax}_k \hat{\omega}_k$.

### Appendix B.1.3. Parallelizing EM for use with fMRI data

Let our APECMa algorithm for maximizing $\mathcal{L}$ consist of the $C = K + 1$ cycles given by $\mathcal{C} = \{[\pi, \Gamma_0], [\pi, \Gamma_1], \ldots, [\pi, \Gamma_K]\}$ with each cycle containing a CM-step followed by a partial E-step (PE). For example, PE-steps only update $u^{(s)+}[c|c_k]$ for $c = 0, 1, 2, \ldots, C - 1$. The extra space here is $(u_k(d))_{1 \leq d \leq n_k, 0 \leq s \leq K}$ corresponding to the $(w_k(d))_{1 \leq d \leq n_k, 0 \leq s \leq K}$. There is no need to update components other than the $k$th component because $u_k^{(s)}$ for $k < c$ have been updated in a previous cycle of the same iteration $s$, while $u_k^{(s+c)} = u_k^{(s)}$ for $k > c$ have not changed from the previous iterated value at $\Gamma_k^{(s+c)}$. Therefore, $w_k^{(s+c)} = \frac{\pi_k^{(s+c)} \mu_k^{(s+c)} \Sigma_k^{(s+c)}}{\sum_{k=0}^K \pi_k^{(s+c)} \mu_k^{(s+c)} \Sigma_k^{(s+c)}}$ for all $i$ and $k$, completing the E-step for the $c$th cycle of the $k$th iteration. The CM-steps are performed in turn on the $c$th cycle and utilize $w_k^{(s+c)}$ in the same manner as the M-step of the EM algorithm, but restricted to $[\pi, \Gamma_c]$ for the $c$th cycle.

The EMG algorithm requires no further changes to the PE-steps, since we may distribute the dataset $X_{\text{new}(1+c)} = ((p_i, v_i^T), 2)_{1 \leq i \leq n, 1 \leq j \leq (1+c)}$ in consecutive row blocks into $D$ cores for the EGM algorithm. For example, each core contains one row block of the dataset, say $X_{\text{new}(1+c)}$, with $n_g$ observations in the $d$th core and $\sum_{d=1}^D n_d = n$. ($n_d \approx n / D$ is ideally for gaining computation performance.) Similarly, we distribute $\omega_k$'s and $u_k$'s into the $D$ cores with dimensions $n_g \times (K + 1)$. Therefore, the PE-steps in the data-distributed computed environment are the same as in the serial environment and no communication across cores is needed. Moreover, each core needs less computation – about $1/D$ on the average – for the EGM algorithm.

The CM-steps, however, need adjustment for application of the EGM algorithm. For the mixing proportion, the unconstrained estimates are

$$\hat{\pi}_k^{(s+c)} = \frac{\sum_{d=1}^D \sum_{i=1}^{n_d} \omega_k^{(s+c)}(v_i; \mu_k^{(s+c)} \Sigma_k^{(s+c)})}{\sum_{k=0}^K \sum_{d=1}^D \sum_{i=1}^{n_d} \omega_k^{(s+c)}(v_i; \mu_k^{(s+c)} \Sigma_k^{(s+c)})}$$

where $\sum_{d=1}^D \omega_k^{(s+c)}(v_i; \mu_k^{(s+c)} \Sigma_k^{(s+c)})$ is the sufficient statistics calculated locally for each $d$ core and is gathered/reduced in the G-step from all the other $(D - 1)$ cores. The constrained estimates for $\hat{\pi}_k^{(s+c)}$ can be adjusted accordingly. Unlike $\Gamma_k$'s, $\pi$ is involved in every cycle so that all $\pi_k$'s need to be updated in each cycle in order to fulfill conditions that guarantee validity of the EM steps, such as monotonicity of the log likelihood and space-filling of the parameter space (Chen et al. 2013). For the beta distribution parameters with constraints, we numerically optimize the objective function (in terms of $\alpha_k$ and $\beta_k$ for $c > 0$) required by the constrOptim() function in R, such that $\sum_{d=1}^D \omega_k^{(s+c)}(v_i; \mu_k^{(s+c)} \Sigma_k^{(s+c)})$ for each $d$ core is maximized under the constraint (iii) (where $\sum_{d=1}^D \omega_k^{(s+c)}(v_i; \mu_k^{(s+c)} \Sigma_k^{(s+c)})$ is the sufficient statistics of the objective function calculated locally given each new numerical update of $(\alpha_k, \beta_k)$). For the normal components, we compute

$$\mu_c^{(s+c)} = \frac{\sum_{d=1}^D \sum_{i=1}^{n_d} \omega_k^{(s+c)}(v_i; \mu_c^{(s+c)} \Sigma_c^{(s+c)})}{\sum_{d=1}^D \sum_{i=1}^{n_d} \omega_k^{(s+c)}(v_i; \mu_c^{(s+c)} \Sigma_c^{(s+c)})}$$

and $\Sigma_c^{(s+c)}$ which is the diagonal matrix containing the diagonal elements of

$$\sum_{d=1}^D \sum_{i=1}^{n_d} \omega_k^{(s+c)}(v_i; \mu_c^{(s+c)} \Sigma_c^{(s+c)}) (v_i - \mu_c^{(s+c)} \Sigma_c^{(s+c)})^T$$

where $\sum_{d=1}^D \omega_k^{(s+c)}(v_i; \mu_c^{(s+c)} \Sigma_c^{(s+c)})$, $\sum_{d=1}^D \omega_k^{(s+c)}(v_i; \mu_c^{(s+c)} \Sigma_c^{(s+c)})$ and $\sum_{d=1}^D \omega_k^{(s+c)}(v_i; \mu_c^{(s+c)} \Sigma_c^{(s+c)})$ are the sufficient statistics for the G-step.

### Appendix C. Performance evaluations: Additional details

#### Appendix C.1. Illustrative examples to show the effect of the identification complexity parameter

We first provide sample illustrations (Figs. C.13, C.14, and C.15) of the generative mechanism used to simulate the $p$-values in our 2D simulation. The identification complexity
Fig. C.13: Sample realizations and activation detection for identification complexity $\omega = 0.01$ (top set), $\omega = 0.50$ (middle set) and $\omega = 0.75$ (bottom set) on simulations obtained using the phantom in Fig. 2a. For each set of values, we have, in the left panel, a density of $p$-values, including a realization (placed atop the density plots) of $p$-values using the generative model for $\pi_0 = 0.977$. The right panel displays a (i) sample SPM, obtained by calculating the upper quantile of the simulated $p$-values, as well as activation maps (along with their $J$) obtained using (ii) RF, (iii) FDR, (iv) CT-2nd, (v) pTFCE, (vi) permutation testing (Perm in the figure), (vii) AR-FAST, with $\alpha = 0.01$, (viii) ALL-FAST, with $\alpha = 0.01$, (ix) SMM, and (x) MfnoX, with the best-performing $\delta$ from among $\{0.95, 0.975, 0.99\}$, and the choice indicated by the first value in the legend.
Fig. C.14: Sample realizations and activation detection for identification complexity $\omega = 0.01$ (top set), $\omega = 0.50$ (middle set) and $\omega = 0.75$ (bottom set) on simulations obtained using the phantom in Fig. 2b. For each set of values, we have, in the left panel, a density of $p$-values, including a realization (placed atop the density plots) of $p$-values using the generative model for $\pi_0 = 0.977$. The right panel displays a (i) sample SPM, obtained by calculating the upper quantile of the simulated $p$-values, as well as activation maps (along with their $J$) obtained using (ii) RF, (iii) FDR, (iv) CT-2nd, (v) pTFCE, (vi) permutation testing (Perm in the figure), (vii) AR-FAST, with $\alpha = 0.01$, (viii) ALL-FAST, with $\alpha = 0.01$, (ix) SMM, and (x) MfnoX, with the best-performing $\delta$ from among $\{0.95, 0.975, 0.99\}$, and the choice indicated by the first value in the legend.
Fig. C.15: Sample realizations and activation detection for identification complexity \( \omega = 0.01 \) (top set), \( \omega = 0.50 \) (middle set) and \( \omega = 0.75 \) (bottom set) on simulations obtained using the phantom in Fig. 2. For each set of values, we have, in the left panel, a density of \( p \)-values, including a realization (placed atop the density plots) of \( p \)-values using the generative model for \( \pi_0 = 0.977 \). The right panel displays (i) sample SPM, obtained by calculating the upper quantile of the simulated \( p \)-values, as well as activation maps (along with their \( J \)) obtained using (ii) RF, (iii) FDR, (iv) CT-2nd, (v) pTFCE, (vi) permutation testing (Perm in the figure), (vii) AR-FAST, with \( \alpha = 0.01 \), (viii) ALL-FAST, with \( \alpha = 0.01 \), (ix) SMM, and (x) MfnoX, with the best-performing \( \delta \) from among \( \{0.95, 0.975, 0.99\} \), and the choice indicated by the first value in the legend.
parameter $\omega$ controls the difficulty and the specificity of the activation problem and we display sample densities, realizations and detected activation images for representative $\omega = 0.01, 0.5, 0.75$. Thus, the difficulty of our activation detection problem in our sample illustrations spans from the moderately easy to the very difficult. As explained in Section 4.1, the $p$-values are realizations from one of $k$ densities given by

$$\phi(p; v_k) = \phi(\Phi^{-1}(1 - p); v_k, 1) / \phi(\Phi^{-1}(1 - p)).$$

Note that $\nu_0 = 0$ and at that value, the density reduces to the standard uniform density (corresponding to the distribution of $p$-values from the inactivated component). $\nu_0, \nu_1$ and $\nu_2$ are set to be the proportion of pixels in each region for the given kind of phantom. With these $\nu_0, \nu_1$ and $\nu_2$, and $v_1$ and $v_2$ are chosen to satisfy $\omega k_l = \omega$ for all $k \neq l \in \{0, 1, 2\}$, where $\omega k_l$ is the overlap measure as defined in [Maitra and Melynkov (2010)]. The left panels of Figs. C.13 C.14 and C.15 show sample densities obtained for $\omega = 0.01, 0.5, 0.75$. Simulated $p$-values are then realized from $\phi(p; v_k')$. Sample SPMs obtained from realizations of $p$-values obtained from each of the realized densities in Figs. C.13 C.14 and C.15 are also displayed (in terms of values, atop each density figure) and as the first image in the right panels.

Appendix C.2. Illustrative example of performance in simulation settings

The other images in the right panels of Figs. C.13 C.14 and C.15 display, for each setting, and in order, activation images obtained using our competitors (using values, where appropriate, parameter values that yielded the better performance).

References

Adler, R., 1981. The geometry of random fields. Wiley, New York.
Adrian, D.W., Maitra, R., Rowe, D.B., 2018. Complex-valued time series modeling for improved activation detection in fMRI studies. Annals of Applied Statistics 12, 1451–1478. doi:https://doi.org/10.1214/17-AOAS1177
Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle, in: Selected Papers of Hirotugu Akaike. Springer, pp. 199–213.
Allison, D.B., Gadbury, G.L., Heo, M., Fernández, J.R., Les, C.K., Prolla, J.A., Weinindruch, R., 2002. A mixture model approach for the analysis of microarray gene expression data. Computational Statistics and Data Analysis 39, 1–20.
Almodovar-Rivera, I.A., Maitra, R., 2019. FAST adaptive smoothed thresholding for improved activation detection in low-signal fMRI. IEEE Transactions on Medical Imaging 38, 2021–2026. doi:10.1109/TMI.2019.2915052
Bandettini, P.A., Jesmanowicz, A., Wong, E.C., Hyde, J.S., 1993. Processing strategies for time-course data sets in functional MRI of the human brain. Magnetic Resonance in Medicine 30, 161–173.
Bardin, J.C., Fins, J.I., Katz, D.I., Hersh, J., Heier, L.A., Tabelow, K., Dyke, J.P., Ballon, D.J., Schif, N.D., Voss, H.U., 2011. Dissociations between behavioural and functional magnetic resonance imaging-based evaluations of cognitive function after brain injury. Brain 134, 769–782. doi:10.1093/brain/auro005
Belliveau, J.W., Kennedy, D.N., McKinstry, R.C., Buchbinder, B.R., Weisskoff, R.M., Cohen, M.S., Vevea, J.M., Brady, T.J., Rosen, B.R., 1991. Functional mapping of the human visual cortex by magnetic resonance imaging. Science 254, 716–719.
Benjamini, Y., Heller, R., 2007. False discovery rates for spatial signals. Journal of the American Statistical Association 102, 1272–1281. doi:10.1198/016214507000000941
Benjamini, Y., Heller, R., 2008. Screening for partial conjunction hypotheses. Biometrics 64, 1215–1222. doi:10.1111/j.1541-4420.2007.00983.x
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300. URL: http://dx.doi.org/10.2307/2346101
Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29, 1165–1188. URL: http://www.jstor.org/stable/26814075
Biswal, B., DeVeoe, A.E., Hyde, J.S., 1996. Reduction of physiological fluctuations in fMRI using digital filters. Magnetic Resonance in Medicine 35, 107–113. URL: http://view.ncbi.nlm.nih.gov/pubmed/8771028
Bordier, C., Dojat, M., de Miechoux, P.L., 2011. Temporal and spatial independent component analysis for fMRI data sets embedded in the AnalyzeFMRI R package. Journal of Statistical Software 44, 1–24. URL: http://www.jstatsoft.org/v44/i09/
Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press, Cambridge.
Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group Report to the Advisory Committee to the Director, NIH, 2014. BRAIN 2025: A SCIENTIFIC VISION.
Brown, E.N., Behrmann, M., 2017. Controversy in statistical analysis of functional magnetic resonance imaging data. Proceedings of the National Academy of Sciences 114, E3368–E3369. URL: http://www.pnas.org/content/114/17/E3368.doi:10.1073/pnas.1705513114
arXiv: http://www.pnas.org/content/114/17/E3368.full.pdf
Chen, W.C., Maitra, R., 2017. MixfMRI: fMRI clustering analysis. R Package, http://cran.r-project.org/package=MixfMRI
Chen, W.C., Ostrouchov, G., Pugmire, D., Prabhath, M., Wehner, M., 2013. A parallel em algorithm for model-based clustering with application to explore large spatio-temporal data. Technometrics 55, 513–523.
Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computers and biomedical research, an international journal 29, 162–173.
Cox, R.W., 2012. AFNI: What a long strange trip it has been. NeuroImage 62, 743–747.
Cox, R.W., Chen, G., Glen, D.R., Reynolds, R.C., Taylor, P.A., 2017. fMRI clustering and false-positive rates. Proceedings of the National Academy of Sciences 114, E3370–E3371. URL: http://www.pnas.org/content/114/17/E3370.doi:10.1073/pnas.1614981114
arXiv: http://www.pnas.org/content/114/17/E3370.full.pdf
Cox, R.W., Hyde, J.S., 1997. Software tools for analysis and visualization of fMRI data. NMR in Biomedicine 10, 171–178.
Cox, R.W., Jesmanowicz, A., 1999. Real-time 3d image registration for functional MRI. Magnetic Resonance in Medicine 42, 1014–1018.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood for incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society. Series B 39, 1–38.
Eklund, A., Nichols, T., Andersson, M., Knutsson, H., 2015. Empirically investigating the statistical validity of SPM, FSL and AFNI for single subject fMRI analysis, in: IEEE International symposium on biomedical imaging (ISBI), p. 1376–1380.
Eklund, A., Nichols, T.E., Knutsson, H., 2016. Correction for Eklund et al., cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences 113, E4029–E4029. URL: http://www.pnas.org/content/113/33/E4029.doi:10.1073/pnas.1612033113
arXiv: http://www.pnas.org/content/113/33/E4029.full.pdf
Forman, S., Cohen, J., Fitzgerald, M., Eddy, W., Mintun, M., Noll, D., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magnetic Resonance in Medicine 33, 636–647.
Friston, K.J., Frith, C., Liddle, P., Frackowiak, R., 1991. Comparing functional (pet) images: the assessment of significant change. Journal of Cerebral Blood Flow and Metabolism 11, 690.
Friston, K.J., Frith, C.D., Liddle, P.F., Dolan, R.J., Lammertsma, A.A., Frackowiak, R.S.J., 1990. The relationship between global and local changes in PET scans. Journal of Cerebral Blood Flow and Metabolism 10, 458–466.
Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: A general
linear approach. Human Brain Mapping 2, 189–210.
Friston, K.J., Jezzard, P., Turner, R., 1994. Analysis of functional MRI time-series. Human Brain Mapping 1, 153–171.
Garcia, D., 2010. Robust smoothing of gridded data in one and higher dimensions with missing values. Computational Statistics and Data Analysis 54, 1167–1178.
Genovese, C.R., 2000. A Bayesian time-course model for functional Magnetic Resonance Imaging data (with discussion). Journal of the American Statistical Association 95, 691–719.
Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15, 870–878.
Gullapalli, R.P., Maitra, R., Roys, S., Smith, G., Alon, G., Greenspan, J., 2005. Reliability estimation of grouped functional imaging data using penalized maximum likelihood. Magnetic Resonance in Medicine 53, 1126–1134.
Hajnal, J.V., Myers, R., Oatridge, A., Schweiso, J.E., Young, J.R., Bydder, G.M., 1994. Artifacts due to stimulus-correlated motion in functional imaging of the brain. Magnetic Resonance in Medicine 31, 283–291.
Hartvig, N.V., Jensen, J.L., 2000. Spatial variable modeling of fMRI data. Human Brain Mapping 11, 233–248. URL: http://dx.doi.org/10.1002/1097-0193(20000211)11:4<233::AID-HBM10>.3.0.CO;2-F doi: 10.1002/1097-0193(20000211)11:4<233::AID-HBM10>.3.0.CO;2-F
Hayasaka, S., Nichols, T.E., 2003. Validating cluster size inference: random field and permutation methods. Neuroimage 20, 2343–2356.
Heller, R., Stanley, D., Yuketilde, D., Rubin, N., Benjamini, Y., 2006. Cluster-based analysis of fMRI data. NeuroImage 33, 599–608. URL: http://dx.doi.org/10.1016/j.neuroimage.2006.04.233 doi:10.1016/j.neuroimage.2006.04.233
Henson, R., Shallice, T., Gorno-Tempini, M., Dolan, R., 2002. Face Repetition Effects in Implicit and Explicit Memory Tests as Measured by fMRI. Cerebral Cortex 12, 178–186. doi:10.1093/cercor/12.2.178
Howseman, A.M., Bowtell, R.W., 1999. Functional magnetic resonance imaging: imaging techniques and contrast mechanisms. Philosophical Transactions of the Royal Society, London 354, 1179–94.
Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of classification 2, 193–218.
Jaccard, P., 1901. Etude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Societ`e Vaudoise des Sciences Naturelles 37, 547–579.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795. doi:10.1080/01621459.1995.10476572
Kelly, A.C., Uddin, L.Q., Biswal, B.B., Castellanos, F.X., Milham, M.P., 2008. Competition between functional brain networks mediates behavioral variability. NeuroImage 39, 527–537. doi:10.1016/j.neuroimage.2007.08.008
Kessler, D., Angstadt, M., Sripada, C.S., 2017. Reevaluating “cluster failure” in fMRI analysis. NeuroImage 50, 124–135.
Maitra, R., Melnykov, V., 2010. Simulating data to study performance of finite mixture modeling and clustering algorithms. Journal of Computational and Graphical Statistics 19, 354–376. doi:10.1198/jcgs.2009.08054
Maitra, R., O’Sullivan, F., 1998. Variability assessment in Positron Emission Tomography and related generalized deconvolution models. Journal of the American Statistical Association 93, 1340–1355.
Maitra, R., Roys, S.R., Gullapalli, R.P., 2002. Test-retest reliability estimation of functional MRI data. Magnetic Resonance in Medicine 48, 62–70.
McLachlan, G.J., Krishnan, T., 2008. The EM Algorithm and Extensions. Wiley.
Melnykov, V., Maitra, R., 2010. Finite mixture models and model-based clustering. Statistics Surveys 4, 80–116.
Meng, X., van Dyk, D., 1997. The EM algorithm — an old folk-song sung to a fast new tune (with discussion). Journal of the Royal Statistical Society B 59, 51–567.
Monti, M.M., 2011. Statistical analysis of fMRI time-series: A critical review of the GLM approach. Frontiers in Human Neuroscience 5, 1–13. doi:10.3389/fnhum.2011.00028
Monti, M.M., Vanhoudenhuyse, A., Coleman, M.R., Boly, M., Pickard, J.D., Tshibanda, L., Owen, A.M., Laureys, S., 2010. Willful modulation of brain activity in disorders of consciousness. New England Journal of Medicine 362, 579–89.
Nebel, M.B., Joel, S., Muschelli, J., Barber, A., Caflø, B., Pekar, J., Mostofsky, S., 2014. Disruption of functional organization within the primary motor cortex in children with autism. Human Brain Mapping 35, 567–580.
Nichols, T.E., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research 12, 419–446.
Ogawa, S., Lee, T.M., Kay, A.R., Tank, D.W., 1990a. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proceedings of the National Academy of Sciences 87, 9868–9872.
Ogawa, S., Lee, T.M., Nayak, A.S., Glynn, P., 1990b. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. Magnetic Resonance in Medicine 14, 68–78.
Owen, A.M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S., Pickard, J.D., 2006. Detecting awareness in the vegetative state. Science 313, 1402. doi:10.1126/science.1130197
Parker, R.A., Rothenberg, R.B., 1988. Identifying important results from multiple statistical tests. Statistics in Medicine 7, 1031–1043.
Polzehl, J., Voss, H.U., Tabelow, K., 2010. Structural adaptive segmentation for statistical parametric mapping. NeuroImage 52, 515–523.
Pounds, S., Morris, S.W., 2003. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinformatics 19, 1236–1242.
R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/
Roth, M., Decety, J., Kaybaudi, M., Massarelli, R., Delon-Martin, C., Segabarth, C., Morand, S., Gemignani, A., Decors, M., Jeannerod, M., 1996. Possible involvement of primary motor cortex in mentally stimulated movement: a functional magnetic resonance imaging study. NeuroReport 7, 1280–1284.
Saad, Z.S., Glen, D.R., Chen, G., Beauchamp, M.S., Desai, R., Cox, R.W., 2009. A new method for improving functional-to-structural mri alignment using local pearson correlation. NeuroImage 44, 839–848.
Saad, Z.S., Reynolds, R.C., 2012. SUMA. NeuroImage 62, 767–772.
Schif, N.D., Ribary, U., Moreno, D.R., Beattie, B., Kronberg, E., Blasberg, R., Giacino, J., McCagg, C., Fins, J.J., Llinas, R., Plum, F., 2002. Residual cerebral activity and behavioural fragments can remain in the persistently vegetative brain. Brain 125, 1210–1234. doi:10.1093/brain/wsf131
Schwarz, G., 1978. Estimating the dimensions of a model. Annals of Statistics 6, 461–464.
Smith, M., Frazier, L., 2007. Spatial bayesian variable selection with application to functional magnetic resonance imaging. Journal of the American Statistical Association 102, 417–431. URL: http://pubs.amstat.org/content/102/476/417/ff
Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in fMRI activation maps. NeuroImage 50, 124–135.
cluster inference. Neuroimage 44, 83–98.
Spisák, T., Spisák, Z., Zunhammer, M., Bingel, U., Smith, S., Nichols, T., Kincses, T., 2019. Probabilistic TFCE: A generalized combination of cluster size and voxel intensity to increase statistical power. NeuroImage 185, 12 – 26. doi: https://doi.org/10.1016/j.neuroimage.2018.09.078
Stippich, C., Ochmann, H., Sartor, K., 2002. Somatotopic mapping of the human primary sensorimotor cortex during motor imagery and motor execution by functional magnetic resonance imaging. Neuroscience Letters 331, 50–54.
Tabelow, K., Polzehl, J., 2011. Statistical parametric maps for functional MRI experiments in R: The package fmri. Journal of Statistical Software 44, 1–21. URL: http://www.jstatsoft.org/v44/i11
The Multi-Society Task Force on PVS, 1994. Medical aspects of the persistent vegetative state. New England Journal of Medicine 330, 1499–1508. doi: 10.1056/NEJM199405263302107
Vardi, Y., Shepp, L.A., Kaufman, L.A., 1985. Statistical model for Positron Emission Tomography. Journal of the American Statistical Association 80, 8–37.
Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., Rosseel, Y., 2011. neuRosim: An R package for generating fMRI data. Journal of Statistical Software 44, 1–18. URL: http://www.jstatsoft.org/v44/i10/
Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–397.
Woo, C.W., Krishnan, A., Wager, T.D., 2014. Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. Neuroimage 91, 412–419.
Worsley, K.J., 1994. Local maxima and the expected Euler characteristic of excursion sets of $\chi^2$, $F$ and $t$ fields. Advances in Applied Probability 26, 13–42.
Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. Journal of Cerebral Blood Flow and Metabolism 12, 909–918.
Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C., 1996. A unified statistical approach for determining significant voxels in images of cerebral activation. Human Brain Mapping 4, 58–73.
Worsley, K.J., Taylor, J.E., 2005. An improved theoretical $p$ value for SPMs based on discrete local maxima. NeuroImage 28, 1056–1062.
Wu, C., 1983. On the convergence properties of the EM algorithm. Annals of Statistics 11, 95–103.