Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

More specific signal detection in functional magnetic resonance imaging by false discovery rate control for hierarchically structured systems of hypotheses

Konstantin Schildknecht[1], Karsten Tabelow[1], Thorsten Dickhaus[2]

submitted: June 23, 2015

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
E-Mail: Konstantin.Schildknecht@wias-berlin.de
Karsten.Tabelow@wias-berlin.de

[2] University of Bremen
Institute for Statistics
P. O. Box 330 440
28344 Bremen
E-Mail: dickhaus@uni-bremen.de

No. 2127
Berlin 2015

2010 *Mathematics Subject Classification.* 62J15, 62F03, 62P10.

Key words and phrases. Functional magnetic resonance imaging, grouped hypotheses, multiple hypotheses testing, voxels.

This research is partly supported by the Federal Ministry of Education and Research of Germany (BMBF) via grant No. 031A191 (EPILYZE project). We thank Henning U. Voss (Weill Medical College, New York, USA) for providing the sport imagination and the two fMRI finger tapping datasets.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

Signal detection in functional magnetic resonance imaging (fMRI) inherently involves the problem of testing a large number of hypotheses. A popular strategy to address this multiplicity is the control of the false discovery rate (FDR). In this work we consider the case where prior knowledge is available to partition the set of all hypotheses into disjoint subsets or families, e. g., by a-priori knowledge on the functionality of certain regions of interest. If the proportion of true null hypotheses differs between families, this structural information can be used to increase statistical power. We propose a two-stage multiple test procedure which first excludes those families from the analysis for which there is no strong evidence for containing true alternatives. We show control of the family-wise error rate at this first stage of testing. Then, at the second stage, we proceed to test the hypotheses within each non-excluded family and obtain asymptotic control of the FDR within each family in this second stage. Our main mathematical result is that this two-stage strategy implies asymptotic control of the FDR with respect to all hypotheses. In simulations we demonstrate the increased power of this new procedure in comparison with established procedures in situations with highly unbalanced families. Finally, we apply the proposed method to simulated and to real fMRI data.

1 Introduction

Modern research is increasingly concerned with large-scale experiments and complex experimental designs. From a statistical perspective the analysis of such experiments often involves the issue of multiple testing of a large number (say m) of individual hypotheses. The development of methods to deal with this issue is a very active field of research with many sophisticated procedures emerging, e. g., taking a specific structure in the set of hypotheses into account; see, for example, Sections 3.3 and 12.2 of Dickhaus [2014].

One example is the analysis of functional magnetic resonance imaging (fMRI) data; see Lazar [2008] for an overview. At each unit of measurement (voxel) on a regular grid a statistical test is to be performed for the null hypothesis of no activation versus the alternative hypothesis of activation of the voxel (a signal detection problem). In such an application, the number m is often of the order of magnitude of several hundreds of thousand hypotheses.

Two established notions for measuring the type I error of a multiple test are the family-wise error rate (FWER) and the false discovery rate (FDR). The FWER denotes the probability of at least one false rejection among the m individual tests, and a multiple test is said to control the FWER (in the strong sense), if the latter probability is bounded by a pre-defined significance level α over the whole parameter set of the statistical model. One simple way to control the FWER is to carry out every individual test at the adjusted level α / m, commonly referred to as the Bonferroni correction. However, this ignores the spatial correlations of the data (cf. Worsley [2003]), and can often be improved by multivariate methods. Another strategy for fMRI signal detection with FWER control incorporating the spatial dependencies of the hypotheses is based on the geometry of random fields, see Worsley et al. [1992] and Adler and Taylor [2007].

On the contrary, the FDR is defined as the expected proportion of type I errors among all rejections of the multiple test φ, and φ is said to control the FDR at level α if this expected proportion is smaller than a given level $\alpha \in (0,1)$ for all parameter values of the considered statistical model. Applying this criterion leads to more liberal multiple tests, meaning that on average more null hypotheses can be rejected. The Benjamini-Hochberg procedure (or linear step-up (LSU) test φ^{LSU}, see Benjamini and Hochberg [1995]) for FDR control has become very popular in fMRI research [Genovese et al., 2002]. Meanwhile, FDR control is an established criterion for the analysis of high-dimensional data, and is agreed upon to provide a suitable interpretation of the results.

When structural information regarding the hypotheses is at hand, it is often possible to incorporate this external knowledge into the statistical methodology in order to improve the test procedures with respect to power or specificity. In the fMRI context, weighted variants of φ^{LSU} considered in previous work incorporate different aspects of the spatial structure of the activation areas, which are typically organized as clusters of activation rather than as singular spots. Furthermore, the functional organization of the brain defines specific regions of interest related to specific functions that are accessible by suitable experimental paradigms [Huettel et al., 2014]. A very old example for such a functional atlas based on cytoarchitecture is the Brodmann atlas [Brodmann, 1909]. Heller et al. [2006] and Benjamini and Heller [2007] employed clustering techniques to define regions of interest, and incorporated the (in general) heterogeneous cluster sizes into φ^{LSU}. Relatedly, Hu et al. [2010] and Zhao and Zhang [2014] studied the case in which the set of hypotheses can be divided into disjoint groups with potentially different proportions of activated voxels by means of a-priori knowledge and showed higher power of their proposed weighted φ^{LSU} tests in comparison with the standard LSU procedure if the fraction of true null hypotheses differs between the groups.

Another class of weighted FDR-controlling multiple tests introduces a second layer of hypotheses which are added to the original set of the m individual hypotheses. Namely, each of the considered disjoint groups is associated with the group-specific null hypothesis of no activation of the whole group. This leads to a hierarchical hypotheses structure with two levels. One level consists of all the group hypotheses and the other of all the m individual hypotheses. In such a context, hierarchical multiple test procedures consist of two stages: First, the group hypotheses are tested, and families for which the group hypothesis cannot be rejected are excluded from the analysis. This strategy relaxes the (remaining) multiplicity for the second stage, where the individual hypotheses are tested. This situation was investigated, among others, by Yekutieli [2008]; Bogomolov [2011], and Benjamini and Bogomolov [2014], and is also widely applied in other application fields like genetic association studies [Yekutieli et al., 2006], gene expression analyses [Li and Ghosh, 2014], or in electroencephalography research [Singh and Phillips, 2010].

In this paper we develop a new two-stage method for FDR control in the fMRI context that takes into account an a-priori partition of the brain into disjoint families of voxels. The main innovation is that nonlinear critical values or rejection curves, respectively, are utilized in the second stage. To this end, we make use of the work of Finner et al. [2009] and Finner et al. [2012] who recently developed a theory for implicit adaptation of FDR-controlling multiple test procedures to the amount of signals. While the two latter papers only considered the individual hypotheses, we apply their reasoning within every group which is still under consideration in the second stage of the hierarchical two-stage test. This leads to high sensitivity regarding the voxels within such a group. This is combined with a Bonferroni-type multiplicity adjustment in the first stage, implying a good specificity during the detection of active regions (testing of the group hypotheses). We prove that this procedure controls the FWER on the set of the family hypotheses, as well as, asymptotically as $m \to \infty$, the FDR within each family and the global FDR (gFDR), which is the FDR with respect to all the individual hypotheses.

The remaining sections are structured as follows. In Section 2, the mathematical notation is set up,

some known results about FDR control are reported, the considered two-stage procedures are introduced and their statistical properties are analyzed. To evaluate the proposed new procedure we perform a number of simulations and an analysis of real fMRI data. To this end, the experimental setups are explained in Section 3, while the most important results are reported in Section 4. We conclude with a discussion in Section 5. Lenghty mathematical derivations are deferred to Appendix A. For sake of completeness, remaining experimental results are provided in Appendices B and C.

2 Statistical methodology

2.1 Notation and preliminaries

We denote the number of families of hypotheses by k and the families themselves by $\mathcal{H}_1, \ldots, \mathcal{H}_k$. Each set $\mathcal{H}_\ell$ is assumed to consist of $m_\ell > 0$ individual hypotheses $H_{\ell 1}, \ldots, H_{\ell m_\ell}$, $1 \leq \ell \leq k$. In addition, for each of the k groups we consider a screening (or family) hypothesis $H_\ell^f, 1 \leq \ell \leq k$ which we will formally define in Definition 4. The aims of the statistical analyses are (i) FDR control in each family $\mathcal{H}_\ell$ separately, (ii) FDR control with respect to all individual hypotheses pooled together, denoted by the global FDR, (iii) FWER control on the group level, i. e., with respect to $(H_\ell^f)_{1 \leq \ell \leq k}$. We assume that for each hypothesis a (marginal) p-value is available, which we identify by the same sub- and / or superscript as the corresponding hypothesis.

Definition 1 (Linear step-up test φ^{LSU})**.** *Denote by $p_{1:m} \leq p_{2:m} \leq \ldots \leq p_{m:m}$ the ordered p-values for a collection $\mathcal{H}_m = \{H_i, i \in I = \{1, \ldots, m\}\}$ of null hypotheses at hand. Furthermore, let $H_{1:m}, \ldots, H_{m:m}$ denote the re-ordered null hypotheses in $\mathcal{H}_m$, according to the ordering of the p-values. Then, the linear step-up test φ^{LSU} at* FDR *level $\alpha \in (0,1)$ rejects exactly the hypotheses $H_{1:m}, \ldots, H_{k:m}$, where*

$$k = \max\{i \in I : p_{i:m} \leq i\alpha/m\}. \tag{1}$$

If the maximum in (1) *does not exist, then no hypothesis is rejected.*

The linear step-up test belongs to the broad class of step-up-down (SUD) multiple tests, introduced by Tamhane et al. [1998].

Definition 2 (Step-up-down test of order λ in terms of p-values, cf. Finner et al., 2012)**.** *Let $p_{1:m} \leq p_{2:m} \leq \ldots \leq p_{m:m}$ and α be defined as in Definition 1. For a tuning parameter $\lambda \in \{1, \ldots, m\}$ a step-up-down test $\varphi^\lambda = (\varphi_1, \ldots, \varphi_m)$ (say) of order λ based on some critical values $\alpha_{1:m} \leq \cdots \leq \alpha_{m:m}$ is defined as follows: If $p_{\lambda:m} \leq \alpha_{\lambda:m}$, set $k = \max\{j \in \{\lambda, \ldots, m\} : p_{i:m} \leq \alpha_{i:m}$ for all $i \in \{\lambda, \ldots, j\}\}$, whereas for $p_{\lambda:m} > \alpha_{\lambda:m}$, put $k = \sup\{j \in \{1, \ldots, \lambda - 1\} : p_{j:m} \leq \alpha_{j:m}\}$ $(\sup \emptyset = -\infty)$. Define $\varphi_i = 1$ if $p_i \leq \alpha_{k:m}$ and $\varphi_i = 0$ otherwise $(\alpha_{-\infty:m} = -\infty)$.*

A step-up-down test of order $\lambda = 1$ or $\lambda = m$, respectively, is called step-down (SD) or step-up (SU) test, respectively. If all critical values are identical, we obtain a single-step test.

In case of φ^{LSU}, $\lambda = m$ and $\alpha_{i:m} = i\alpha/m$ for all $1 \leq i \leq m$. In general, the choice of the order λ and of the critical values employed in an SUD test for FDR control depends on model assumptions; cf. Table 5.1 of Dickhaus [2014].

Definition 3 (AORC-based critical values, cf. Finner et al. [2009] and Finner et al. [2012])**.** *Under the assumptions of Definitions 1 and 2, we denote by φ_λ^{AORC} the SUD test with critical values*

$$\alpha_{i:m} = \frac{i\alpha}{m - i(1-\alpha)}, \ 1 \leq i \leq m. \tag{2}$$

These critical values correspond to the so-called asymptotically optimal rejection curve (AORC) introduced by Finner et al. [2009]. For suitable choices of λ and under the assumption of stochastically independent p-values, φ_λ^{AORC} has been shown to exhaust the FDR level α asymptotically as $m \to \infty$, while φ^{LSU} is not exhausting α if the number of true null hypotheses is smaller than m.

In a two level situation with group hypotheses and individual hypotheses, a two-stage procedure can be employed. In our case we are interested in testing the hypotheses within a family $\mathcal{H}_\ell$ only if this family has been declared active, meaning that H_ℓ^f has been rejected in the first stage of analysis. To define activity of families we make use of the partial conjunction hypothesis as in Benjamini and Heller [2008].

Definition 4. *For a given integer $1 \leq u_\ell \leq m_\ell$, the u-partial conjunction hypothesis H^{u_ℓ/m_ℓ} for family $\mathcal{H}_\ell$ is defined as the set of parameters such that $\mathcal{H}_\ell$ contains less than u_ℓ false null hypotheses, with corresponding alternative given by the set of parameters such that the number of true alternatives in $\mathcal{H}_\ell$ is at least equal to u_ℓ. Based on this, we let $H_\ell^f = H^{u_\ell/m_\ell}$. According to Benjamini and Heller [2008] a valid p-value for testing H_ℓ^f can be defined as*

$$p^{u_\ell/m_\ell} = \min_{1 \leq i \leq m_\ell - u_\ell + 1} \left\{ \frac{m_\ell - u_\ell + 1}{i} p_{u_\ell - 1 + i : m_\ell} \right\}. \tag{3}$$

Another critical issue in connection with FDR control is the dependency structure among the p-values. The LSU test controls the FDR under the assumption of positive regression dependency on subsets (PRDS) regarding the joint distribution of the p-values, see Benjamini and Yekutieli [2001] and Sarkar [2002]. It was shown by Guo and Rao [2008] that φ^{LSU} cannot be improved uniformly if the dependency among the p-values is completely unknown. Other procedures as the one by Storey et al. [2004] assume weak dependency in the sense of Definition 5.

Definition 5 (Weak dependency)**.** *Let $p_1, \ldots, p_m$ denote (random) marginal p-values for a collection $\mathcal{H}_m = \{H_i, i \in I = \{1, \ldots, m\}\}$ of null hypotheses at hand. Let $I_N \subseteq I$ ($I_A \subseteq I$) with $|I_N| = m_N$ ($|I_A| = m_A$) denote the index set of true (false) null hypotheses in I. Then, $p_1, \ldots, p_m$ are called weakly dependent, if $q_N = \lim_{m\to\infty} m_N / m$ exists and*

$$\hat{F}_{N m_N}(t) = m_N^{-1} \sum_{i \in I_N} \mathbb{I}_{[0,t]}(p_i) \to F_N(t), \quad m \to \infty \tag{4}$$

$$\hat{F}_{A m_A}(t) = m_A^{-1} \sum_{j \in I_A} \mathbb{I}_{[0,t]}(p_j) \to F_A(t), \quad m \to \infty, \tag{5}$$

where $\mathbb{I}_S$ denotes the indicator function of the set S, convergence in (4) *and* (5) *is uniformly in $t \in [0,1]$ and almost surely, and F_N and F_A are continuous cumulative distribution functions with $0 < F_N(t) \leq t$ for all $t \in (0,1]$.*

Throughout this work, we assume that the p-values within each family are weakly dependent. While Logan et al. [2008] argued against this assumption in the fMRI context, the validity of weak dependency for p-values corresponding to voxel data has been discussed in Chen et al. [2009] on the basis of simulation studies for different magnitudes of positive correlation among the voxels. No situation militating against the assumption was found. The FDR behaviour of AORC-based multiple test procedures under the weak dependency assumption regarding the joint distribution of the p-values was investigated in Chapter 4 of Gontscharuk [2010].

2.2 Considered two-stage multiple tests

In Benjamini and Bogomolov [2014] a general method to design procedures coping with the selection of families has been provided. For a comparison with our proposed procedure φ^{HO} we make use of the so-called "simple selection adjusted procedure" from Bogomolov [2011], which is based on φ^{LSU} and denoted by φ^{Bog}. This procedure achieves global FDR control and FDR control within each family [Bogomolov, 2011].

Algorithm 1 (The procedure φ^{Bog})**.**

1. *Test the k families with the LSU procedure at level α applied to $(p^{1/m_\ell})_{1\le\ell\le k}$, see* (3)*. Obtain R rejections.*
2. *In the case of $R > 0$, apply in each of the R rejected families φ^{LSU} at level $R\alpha/m_\ell$, where ℓ denotes the index of a rejected family.*

We propose to apply the following procedure which harnesses the advantages of the AORC approach and exploits the structural information.

Algorithm 2 (The procedure φ^{HO})**.** *Let $\lfloor x \rfloor$ denote the largest integer smaller than or equal to x.*

1. *For a given tuning parameter $\kappa > k$, let $u_\ell = \lfloor \kappa^{-1} \cdot m_\ell \rfloor + 1$ for $1 \le \ell \le k$. Reject all families $\mathcal{H}_\ell$ for which*
$$p^{u_\ell/m_\ell} \le \frac{\alpha}{\kappa}.$$
Obtain R rejections.
2. *In the case of $R > 0$, apply in each of the R rejected families φ_λ^{AORC} at level α, with $\lambda = u_\ell$, where ℓ denotes the index of a rejected family.*

Under standard assumptions which are typically made in FDR theory, all three aims of the statistical analyses mentioned before are achieved by φ^{HO}, at least asymptotically as $\min_{1\le\ell\le k} m_\ell \to \infty$; see Appendix A.1 for details.

3 Experiments

We will compare the two hierarchical procedures φ^{HO} and φ^{Bog} with the multiple test φ_λ^{AORC} regarding the empirical power on the combined set of hypotheses in Section 3.1. In the simulations regarding fMRI data presented in Section 3.2, we will make the comparison of the LSU procedure φ^{LSU} with the hierarchical procedures on the combined set of voxels by means of their empirical FDRs. When evaluating real fMRI experiments, we compare the respective numbers of detections, i. e., rejections.

3.1 Computer simulations regarding the power of φ^{HO}

In this section we consider the performance of the procedures in terms of power of a multiple test. A standard notion of power of a multiple test procedure $\varphi_{(m)}$ for m hypotheses is given in Definition 1.4 of Dickhaus [2014] as
$$\text{power}_m(\varphi_{(m)}) = \mathbb{E}\left[\frac{S_m}{m_A \vee 1}\right],$$

Table 1: Parameter configurations in the one-sided normal means problem.

	$\pi = (\pi_1, \pi_2)$	$q_N = (q_{N1}, q_{N2})$
1	$(0.5, 0.5)$	$(0.5, 0.5)$
2	$(0.5, 0.5)$	$(0.8, 0.1)$
3	$(0.8, 0.2)$	$(0.8, 0.1)$
4	$(0.5, 0.5)$	$(0.99, 0.01)$
5	$(0.8, 0.2)$	$(0.99, 0.01)$

where S_m denotes the number of correct rejections and the expectation $\mathbb{E}$ refers to the true underlying measure. The global power of a multiple test procedure $\varphi_{(m)}$ that operates on a structured family of hypotheses as considered in Section 2 is given by

$$\text{gpower}_m(\varphi_{(m)}) = \mathbb{E}\left[\frac{S_m}{m_A \vee 1}\right] = \mathbb{E}\left[\frac{\sum_{\ell=1}^k S_\ell}{\sum_{\ell=1}^k m_{A\ell} \vee 1}\right],$$

where $m_{A\ell}$ and S_ℓ are the number of false null hypotheses and the number of correct rejections in family ℓ. For a given number B of Monte Carlo repetitions, the power of $\varphi_{(m)}$ is estimated by the average value

$$\widehat{\text{power}}_m(\varphi_{(m)}) = \frac{1}{B}\sum_{b=1}^{B}\frac{s_{m,b}}{m_A},$$

where $s_{m,b}$ denotes the realization of S_m in the b-th simulation run. In our simulations, we set $B = 10{,}000$ and $m = 2{,}500$.

The simulations refer to the one-sided normal means problem with $\Omega = \mathbb{R}^m$, an observable random vector $T = (T_1, \ldots, T_m)^\top$ with values in Ω such that $\mathcal{L}(T) = \mathcal{N}_m(\mu, I_m)$, where $\mu = (\mu_1, \ldots, \mu_m)^\top$, and hypotheses

$$H_j : \{\mu_j = 0\} \text{ vs. } K_j : \{\mu_j > 0\}, j \in \{1, \ldots, m\}.$$

The p-value for a hypothesis H_j is then given by

$$p_j(t_j) = \mathbb{P}_{H_j}(T_j > t_j) = 1 - \Phi(t_j),$$

where t_j denotes the observed value of T_j and Φ denotes the cumulative distribution function of the standard normal distribution.

For convenience, we set all μ_j, $j \in I_A$, to the same value $\mu_c > 0$. The power of the different procedures will be investigated for different effect sizes μ_c. The effect size μ_c will range from 0.5 up to 5 in steps of 0.5. Furthermore, we assume that the family $\mathcal{H}_m = (H_1, \ldots, H_m)$ is structured into two subfamilies $\mathcal{H}_{m_1}$ and $\mathcal{H}_{m_2}$. The parameter κ is set to 1000 and to 100, respectively, see Appendix A.2 for justification. We let $\pi_\ell = m_\ell/m$ and $q_{N\ell} = m_{N\ell}/m_\ell$, $\ell = 1, 2$. Table 1 lists the considered parameter configurations. The FDR level was set to $\alpha = 5\%$ in all simulations.

3.2 fMRI - data

Simulations and analysis of experimental data were all performed within the R language and environment for statistical computing and graphics R Development Core Team [2015]. The R-scripts for the creation of the simulated data and its analysis is available from the authors on request.

Simulated fMRI data We created simulated fMRI data using the R-package **neuRosim** Welvaert [2012] described in detail in Welvaert et al. [2011]. The simulated data consisted of 105 volumes of size 20 $\times$ 20 $\times$ 20 isotropic voxels. The simulated stimulus had onset times at the 16-th, 46-th and 76-th volume, a duration overlapping 15 volumes and a repetition time of two seconds. The expected hemodynamic response to this block design was created using a convolution of the task indicator function with the standard "double-gamma" hemodynamic response function Glover [1999]. The "activation" region in this data was set to a sphere of radius 3. The center of the sphere was set in voxel coordinates (5, 5, 5) for simulation A and in voxel (10, 10, 10) for simulation B. Noise was added using a Rician distribution including spatial and temporal correlations.

We then analyzed the data within a standard GLM approach using the R-package **fmri** Tabelow and Polzehl [2012, 2011] including corrections for temporal autocorrelations and quadratic signal trends. From the resulting statistical parametric map we determined local p-values.

We defined an arbitrary partition of the spatial domain into 8 families of voxels corresponding to the 8 "corners" of the data cube. For both simulation datasets we then applied the hierarchical testing procedure φ^{HO} described in this paper, the procedure φ^{Bog} as well the classical Benjamini-Hochberg procedure Benjamini and Hochberg [1995] using a level of 0.05.

SPM auditory fMRI test data For validation of our new inference method on experimental fMRI data we used a publicly available single subject fMRI dataset with an auditory stimulus design. The data can be downloaded at `http://www.fil.ion.ucl.ac.uk/spm/data/auditory/` together with details on its acquisition.

The number of volumes at an repetition time of 7 seconds was 6 with alternating blocks of rest and auditory stimulus, starting with rest, each lasting for 6 volumes. EPI data was acquired on a modified 2T Siemens MAGNETOM Vision system. The spatial dimension of the data was 64 $\times$ 64 $\times$ 64 isotropic voxels of length 3mm.

We analyzed the dataset using a standard GLM approach implemented in the R-package **fmri** Tabelow and Polzehl [2012, 2011] including corrections for temporal autocorrelations and quadratic signal trends. From the resulting statistical parametric map we determined local p-values.

To define suitable families of voxels we normalized AFNI's Cox [1996] EPI template (`TT_EPI-tlrc`) in Talairach space with Brodmann labels to the functional data using the normalization toolbox of SPM8. Thus each voxel in the functional data was assigned a label according to the Brodmann atlas. Any other suitable atlas or definition of families could have been used here.

We then applied the procedure φ^{HO}, φ^{Bog} and the classical Benjamini-Hochberg φ^{LSU} procedure on all voxel, that had been assigned any label by the atlas matching described above, restricting analysis to the labelled cortex areas only.

fMRI dataset using a sports imagination task We also re-used an fMRI dataset from Tabelow and Polzehl [2011] performed by one healthy adult female subject. The data is publicly available under `http://www.jstatsoft.org//v44/i11`. The alternating design of rest and task blocks, starting with rest, was identical to the one of the simulated fMRI data and resulted in 105 volumes. The rest and task blocks had a duration of 30 seconds, the repetition time was 2 seconds. The task was imagination of playing tennis. The spatial dimension of the data cube was 64 $\times$ 64 $\times$ 30 with an in-plane resolution of 3.75mm and a slice thickness of 4mm. The TE of the EPI sequence was 40ms and the flip angle was 80 degrees. Before the first rest block 6 dummy scans were discarded to allow for T_1 saturation.

We repeated the analysis described for the SPM auditory fMRI test data, i.e., normalizing the Brodmann labels to the functional data using SPM8 and performing a standard GLM analysis with the R-package **fmri** to determine local p-values.

Signal detection was performed using the procedure φ^{HO}, φ^{Bog} as well as with the Benjamini-Hochberg method φ^{LSU}.

Other fMRI datasets We also analyzed two more fMRI scan of another subject in a finger tapping task within the same task protocol as described for the sports imagination dataset. One of the datasets had a doubled in-plane resolution. The analysis yielded very similar results (with respect to the performance of the signal detection procedure) as the sports imagination dataset, such that we decided not show the results of the analysis here.

4 Results

4.1 Power simulations

The first five sub-figures in Fig. 1 refer to the five parameter configurations from Table 1 with the choice of $\kappa = 1{,}000$. The sixth sub-figure refers to the fifth parameter configuration from Table 1 with the choice of $\kappa = 100$.

In the second panel of Fig. 1 (comprising sub-figures 4 - 6), the ratios $q_{N\ell}$, $\ell = 1, 2$, are highly unbalanced. It can clearly be observed that this leads to an improvement in terms of power of the proposed procedure φ^{HO} over the existing multiple tests φ^{Bog} and $\varphi^{AORC}_{u_\ell}$, at least for $\mu_c \in [2, 3]$. In the first panel (comprising sub-figures 1 - 3), however, the empirical power of $\varphi^{AORC}_{u_\ell}$ is uniformly higher than that of φ^{Bog} and φ^{HO}, respectively.

We may remark that a more detailed analysis of the decision patterns of the three concurring multiple tests (not shown here) revealed that the higher power of $\varphi^{AORC}_{u_\ell}$ in sub-figures 2 and 3 is mainly due to the fact that φ^{Bog} and φ^{HO} discard the first family $\mathcal{H}_{m_1}$ already in the first stage of the analysis (with high probability). Often, such a behavior is wanted in practice, because few isolated signals are typically interpreted as artifacts, especially in the fMRI context.

4.2 fMRI - Simulations

We first show the results for simulation A, where the “activation area“ is fully located within one of the defined families compared with the known ground truth “activation“ in the simulation in Fig. 2, Fig. 3 and Fig. 4. Every procedure detects all true alternatives, but we can observe a different number of false discoveries. The hierarchical procedure φ^{HO} does not make any discoveries in families without activation.

Comparing the detected activation areas with the known ground truth, we estimated the global and within-family false discovery rates as well as the mean FDR over the families for 1000 Monte Carlo replications. We can observe differences regarding the detection of false positvives, see Table 2. The procedure φ^{LSU} has the most rejections, but violates the FDR in every family, except for the family in which the signal is located. The empirical level is below 5% for the other two procedures regarding all the FDRs of interest.

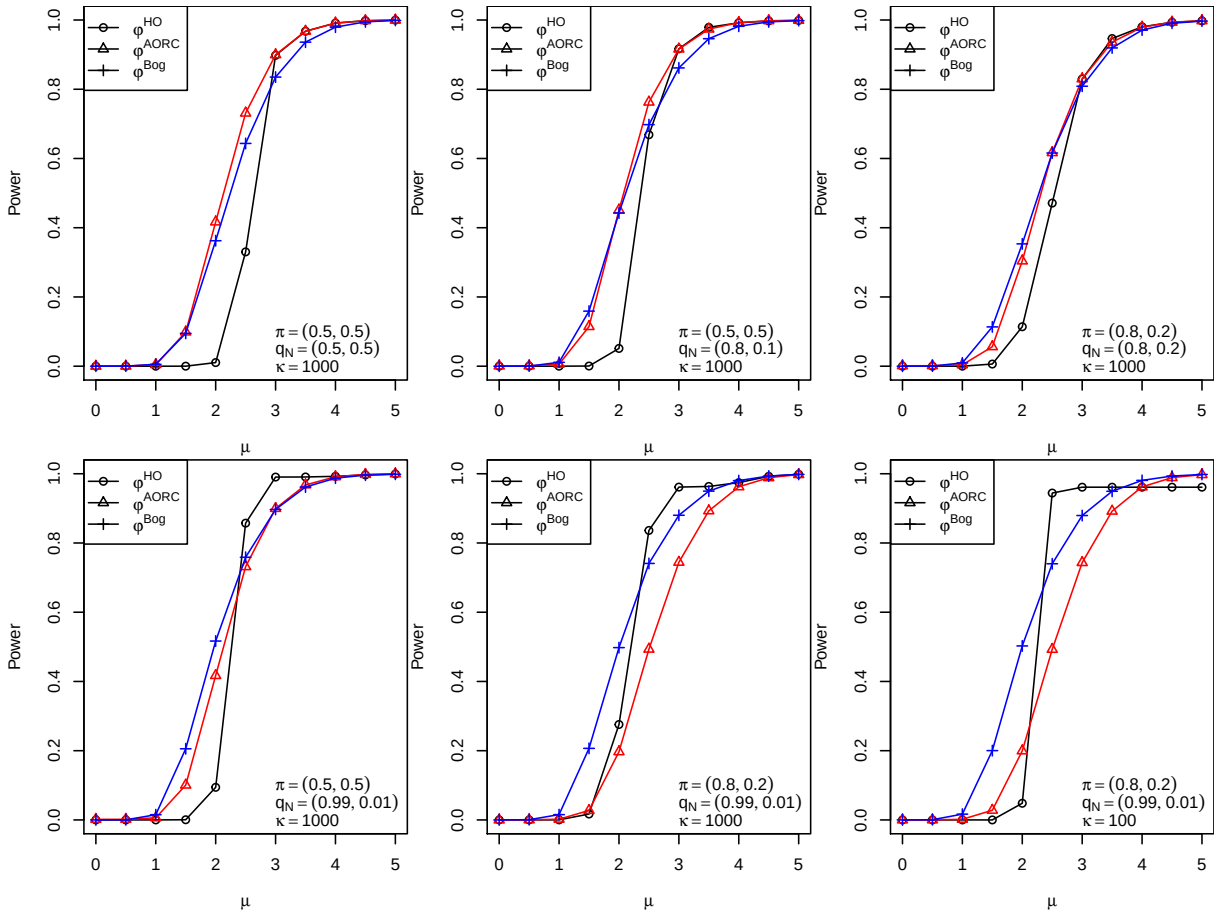

Figure 1: Empirical powers of the procedures φ^{HO} (black), φ^{Bog} (blue) and $\varphi^{AORC}_{u_\ell}$ (red) as a function of the effect size μ_c in the one-sided normal means problem. The total number of hypotheses equals $m = 2500$, and the number of groups equals $k = 2$. The parameter configurations $\pi = (\pi_1, \pi_2)$ and $q_N = (q_{N1}, q_{N2})$ are as in Table 1.

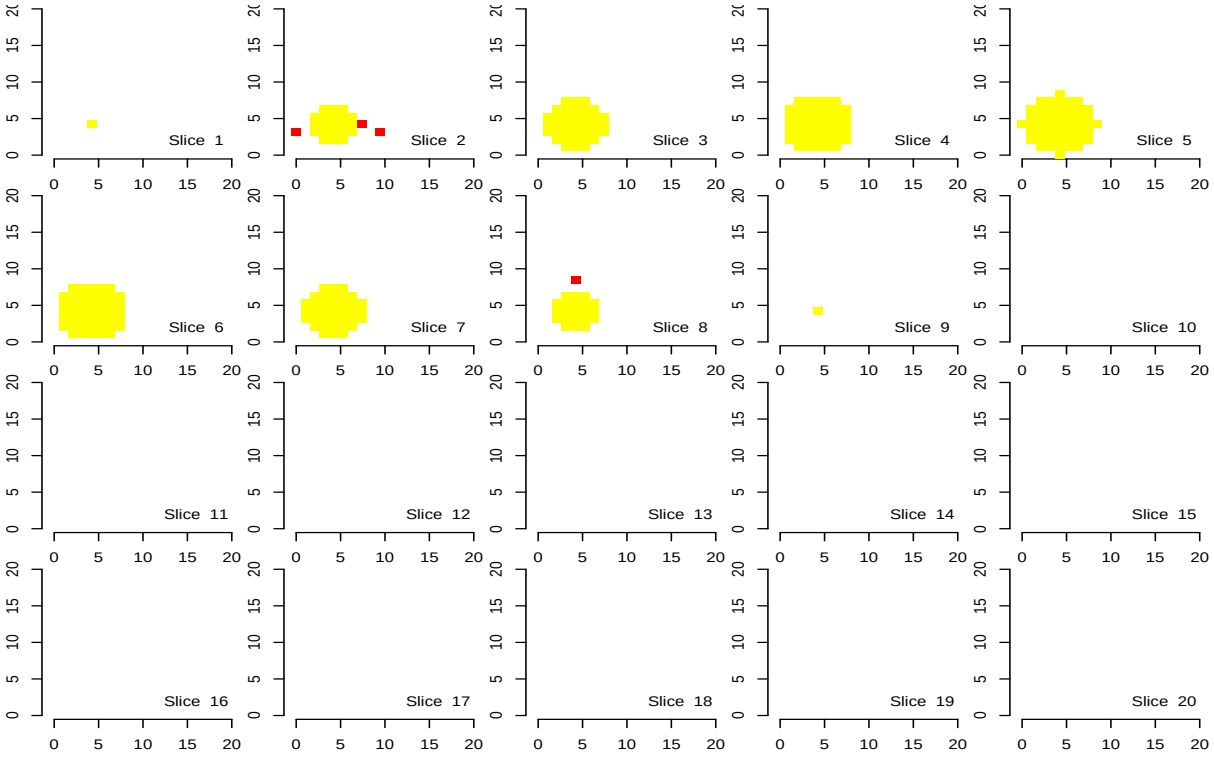

Figure 2: Discoveries of the procedure φ^{HO} in Simulation A on a cube with side length 20, there are 8 disjoint families consisting of cubes with side length 10 each with one edge in the corner of the original cube. Shown are 20 slices corresponding to the third dimension. Ground activation (yellow) and the false rejections of φ^{HO} (red) are shown.

We show the detection results for the simulation B, where true activations are located within all defined families of voxels, in Fig. 5,Fig. 7 and Fig. 6. First we show the slices of the data with the activated voxels determined by the three different procedures overlayed with the ground truth.

A visual inspection of the figures and the table shows the desired behaviour of the procedure. In the table 2 we can clearly observe that the families without activation are mostly excluded from the analysis by φ^{HO} and φ^{Bog}, as they do not show activation in many families, while they retain activation in the test via φ^{LSU}. It is not surprising that in families without signal the FDR in the family is not controlled for the Benjamini-Hochberg procedure. If the signal is found in every family (simulation B) one can not find an advantage in the use of the new procedure. The order of magnitude regarding the FDRs seems to be the same for the two procedures although the attained level of the procedure φ^{HO} is closer to 5%, suggesting higher power.

SPM auditory fMRI test data We show the detection results in the auditory cortex of the proposed procedure φ^{HO} overlayed on the functional division of the brain according to the Brodmann atlas and compare with the detections found by the procedure φ^{LSU} and φ^{Bog} in Figure 8. We can see that the hierarchical procedures detect voxels mainly located in the auditory areas, while the LSU procedure finds activations all over the brain. The full figures showing all slices can be found in the Appendix. The Table 3 shows the number of discoveries in the different Brodmann areas. It can be seen from the Table 3 more then from the Figure 8 that the proposed procedure leads to a far more concentrated signal detection in areas related to the auditory stimulus.

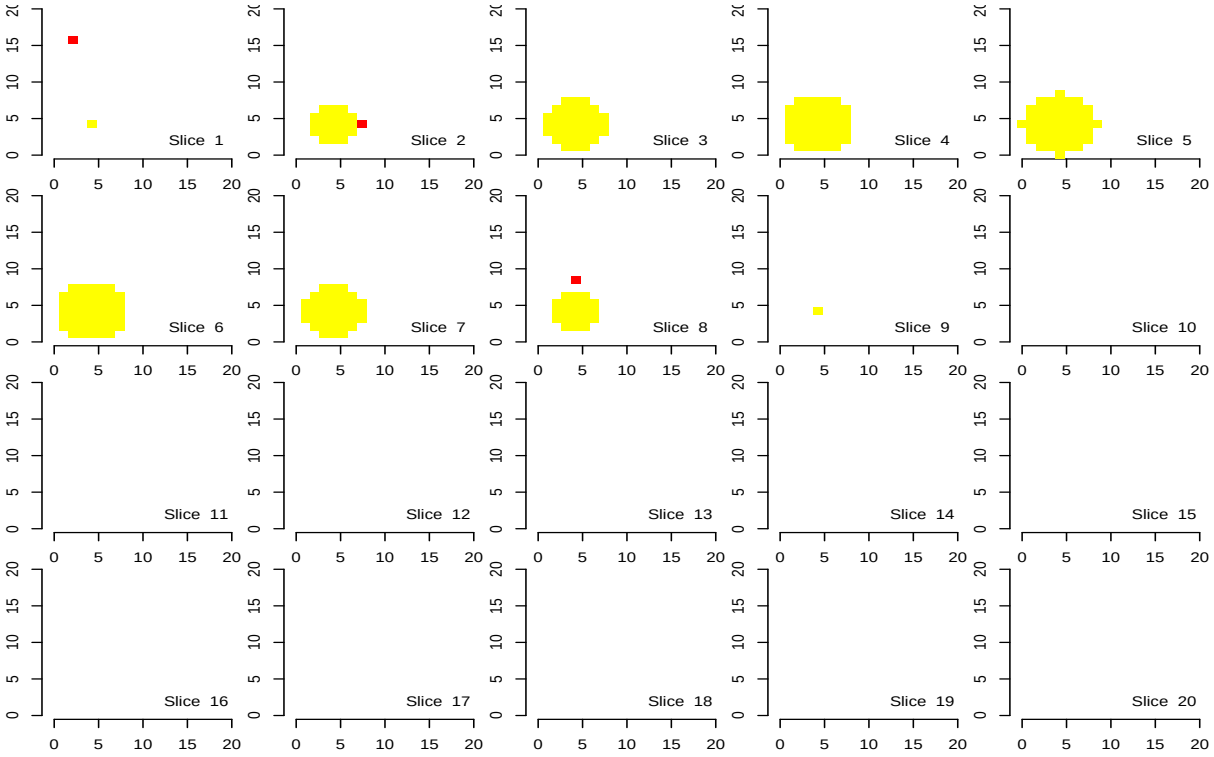

Figure 3: Discoveries of the procedure φ^{Bog} in Simulation A on a cube with side length 20, there are 8 disjoint families consisting of cubes with side length 10 each with one edge in the corner of the original cube. Shown are 20 slices corresponding to the third dimension. The Ground activation (yellow) and false rejections of φ^{Bog} (red) are shown.

fMRI dataset using a sports imagination task We show the detection results of the proposed procedures overlayed on the Brodmann atlas.

A visual inspection of the figures shows activation in the whole brain. As it can be seen in Table 4 in every area of the brain many activated voxels are detected by all procedures. We might hypothesize that stimulus of this experiment, which is an imagination task, is related to much less specific activation due to its complexity. Similar to the situation in fMRI Simulation B we do not observe that the hierarchical procedures perform more specific than φ^{LSU} regarding the Brodmann areas. The full figures with all slices can be found in the Appendix.

5 Discussion

This work focused on the use of structural information in a new procedure to control the FDR. We provided a rigorous mathematical analysis of this new procedure and proved asymptotic control of the FDR. In simulations we studied the performance of the proposed method in situation with finite m. Furthermore, we applied it to simulated and real fMRI data sets.

For fMRI analysis our procedure bears the unique advantage of being specific to the families/regions in which brain activity is located and is highly sensitive within each family. This conclusion can be clearly drawn from the Table 2 and is supported by the figures. Other FDR controlling procedures suffer from false positives in areas without signal. We filter first where strong signal can be found and continue to locate the voxels which are responsible for the strong signal, making use of the nonlinear critical values originating from the theory around the AORC. It was possible to present that when the activation is

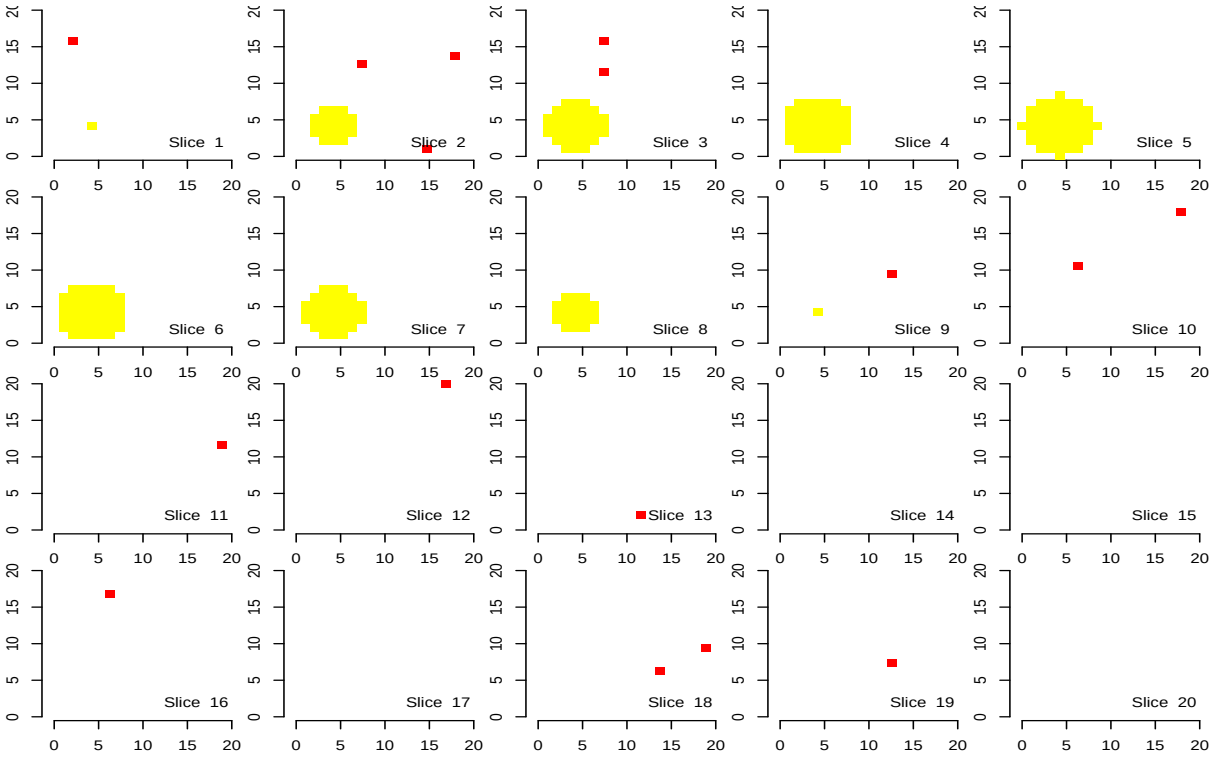

Figure 4: Discoveries of the procedure φ^{LSU} in Simulation A on a cube with side length 20, there are 8 disjoint families consisting of cubes with side length 10 each with one edge in the corner of the original cube. Shown are 20 slices corresponding to the third dimension. Ground activation (yellow) and the false rejections of φ^{LSU} (red) are shown.

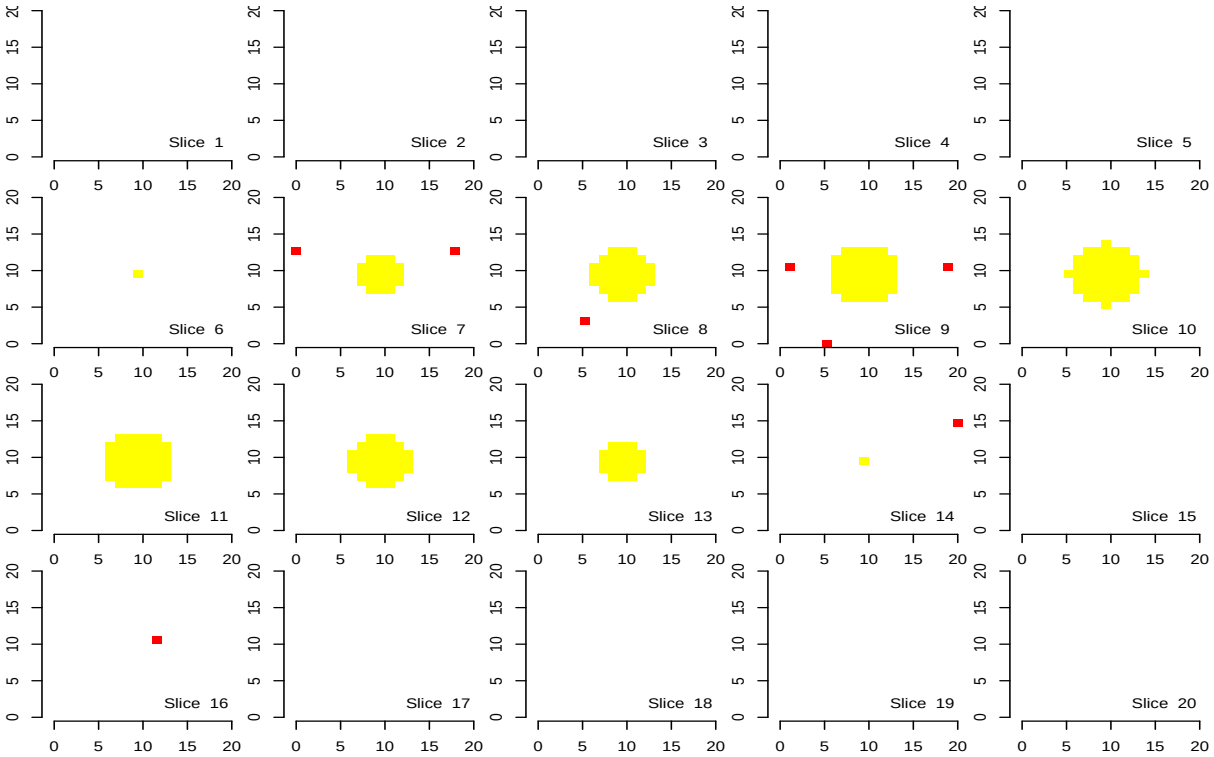

Figure 5: Discoveries of the procedure φ^{HO} in Simulation B on a cube with side length 20, there are 8 disjoint families consisting of cubes with side length 10 each with one edge in the corner of the original cube. The ground activation (yellow) and the false rejections of φ^{HO} (red) are shown.

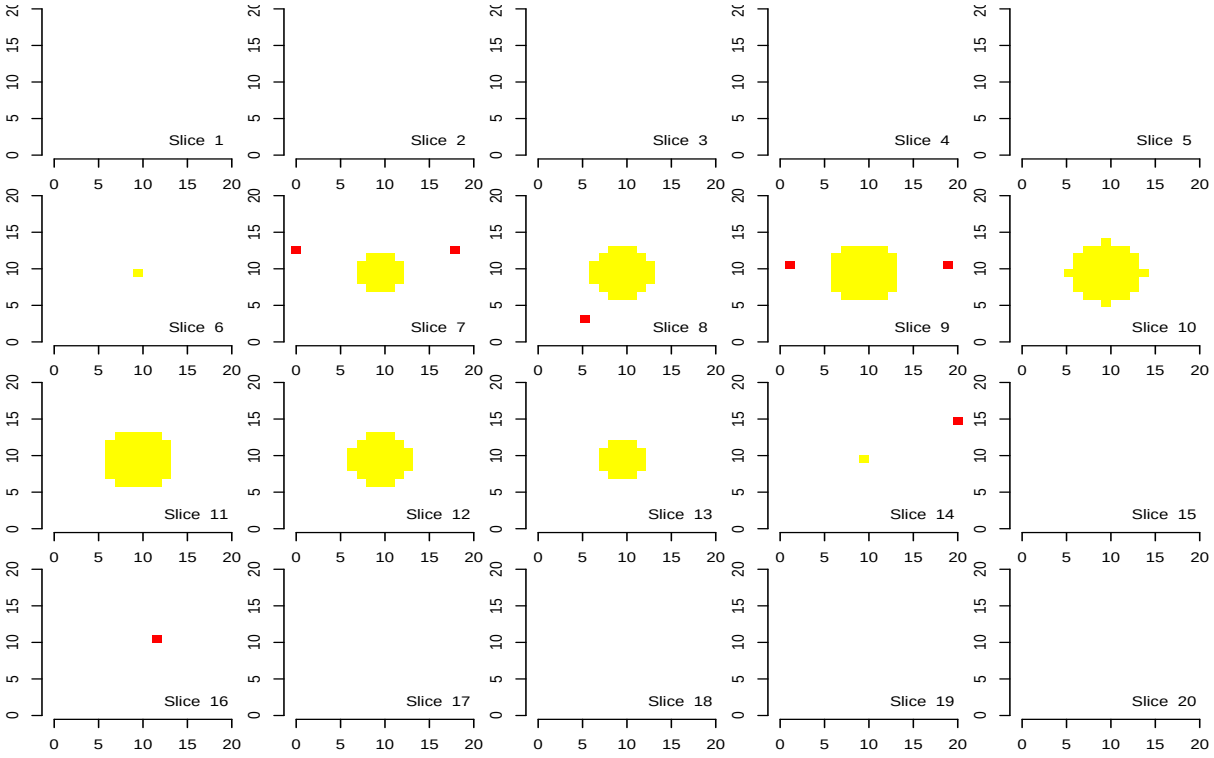

Figure 6: Discoveries of the procedure φ^{Bog} in Simulation B on a cube with side length 20, there are 8 disjoint families consisting of cubes with side length 10 each with one edge in the corner of the original cube. The ground activation (yellow) and the false rejections of φ^{Bog} (red) are shown.

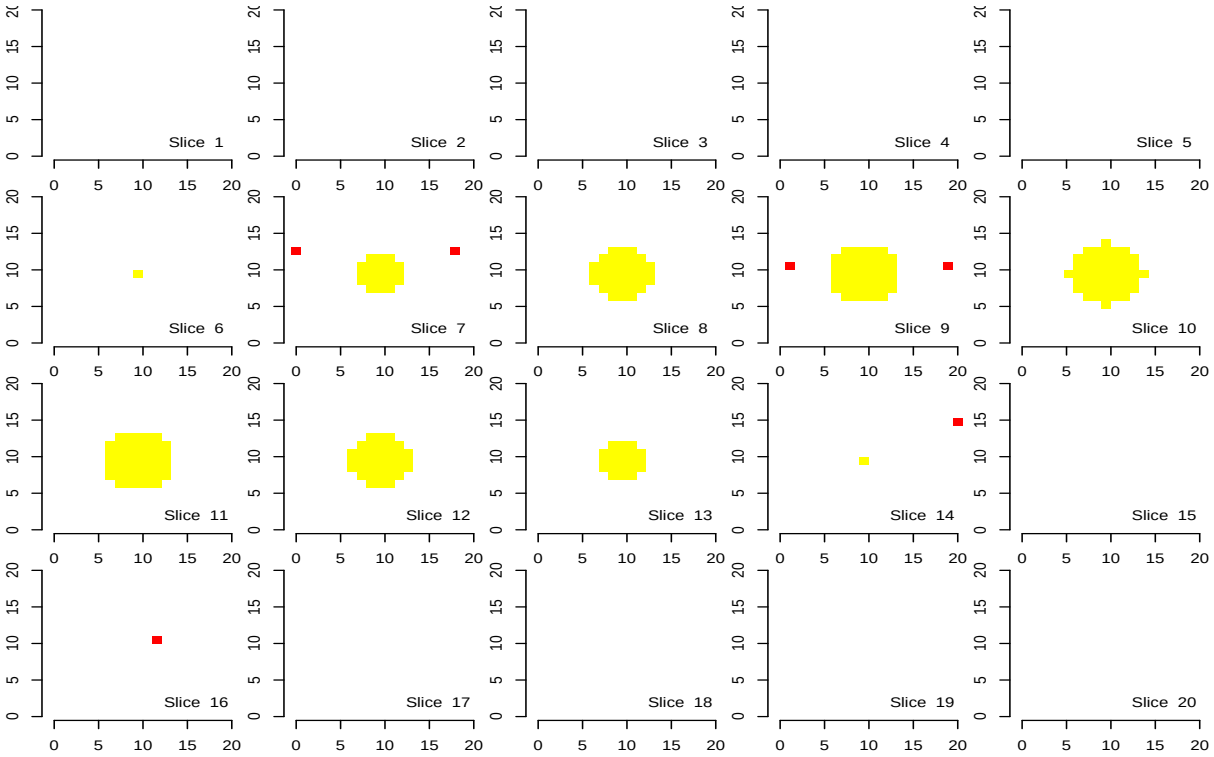

Figure 7: Discoveries of the procedure φ^{LSU} in Simulation B on a cube with side length 20, there are 8 disjoint families consisting of cubes with side length 10 each with one edge in the corner of the original cube. The ground activation (yellow) and the false rejections of φ^{LSU} (red) are shown.

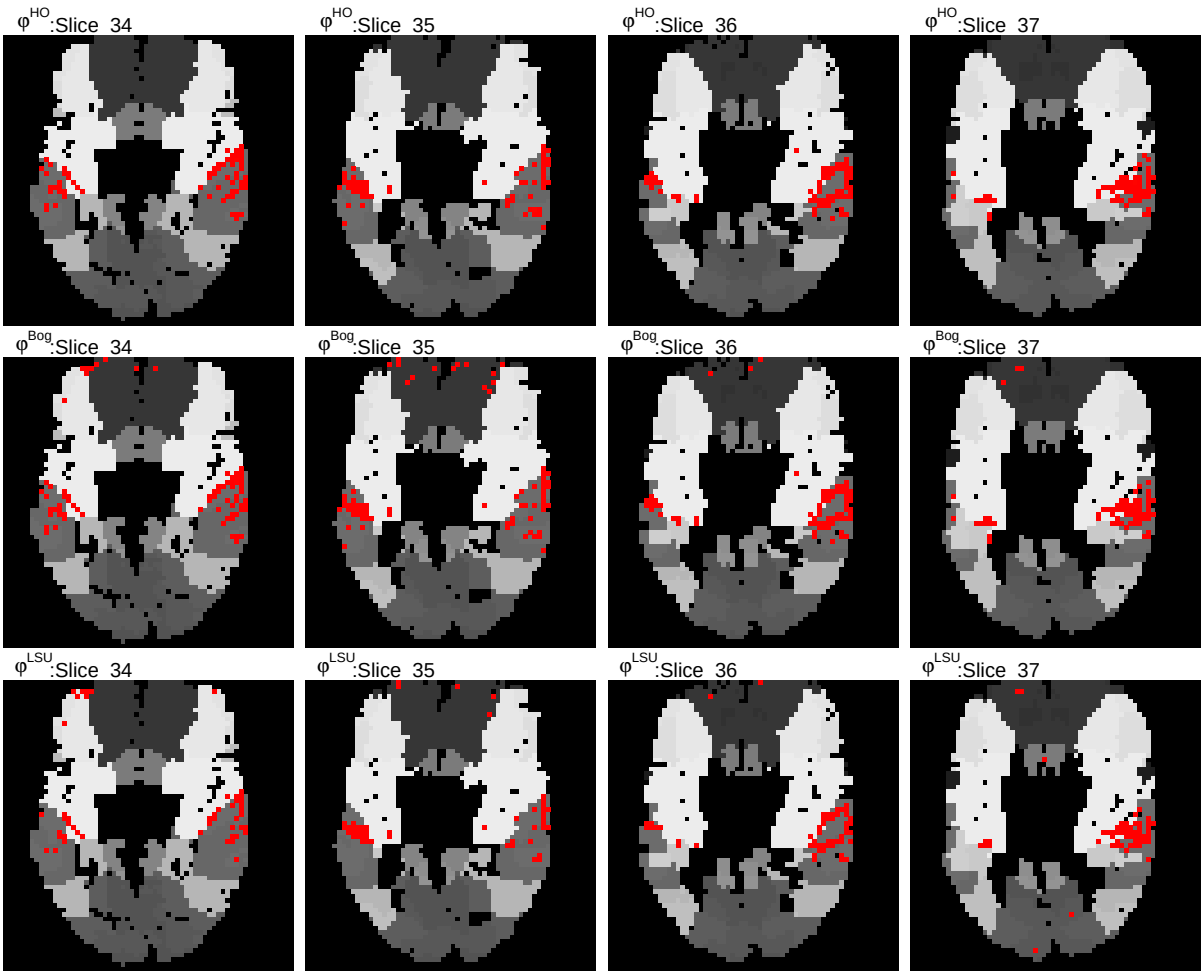

Figure 8: We present chosen slices (auditory cortex visible) of the brain for the SPM auditory fMRI dataset and the discoveries proposed procedure φ^{HO} in the upper row. The discoveries in the second row correspond to the procedure φ^{Bog} and in the third row the corresponding discoveries of φ^{LSU}.

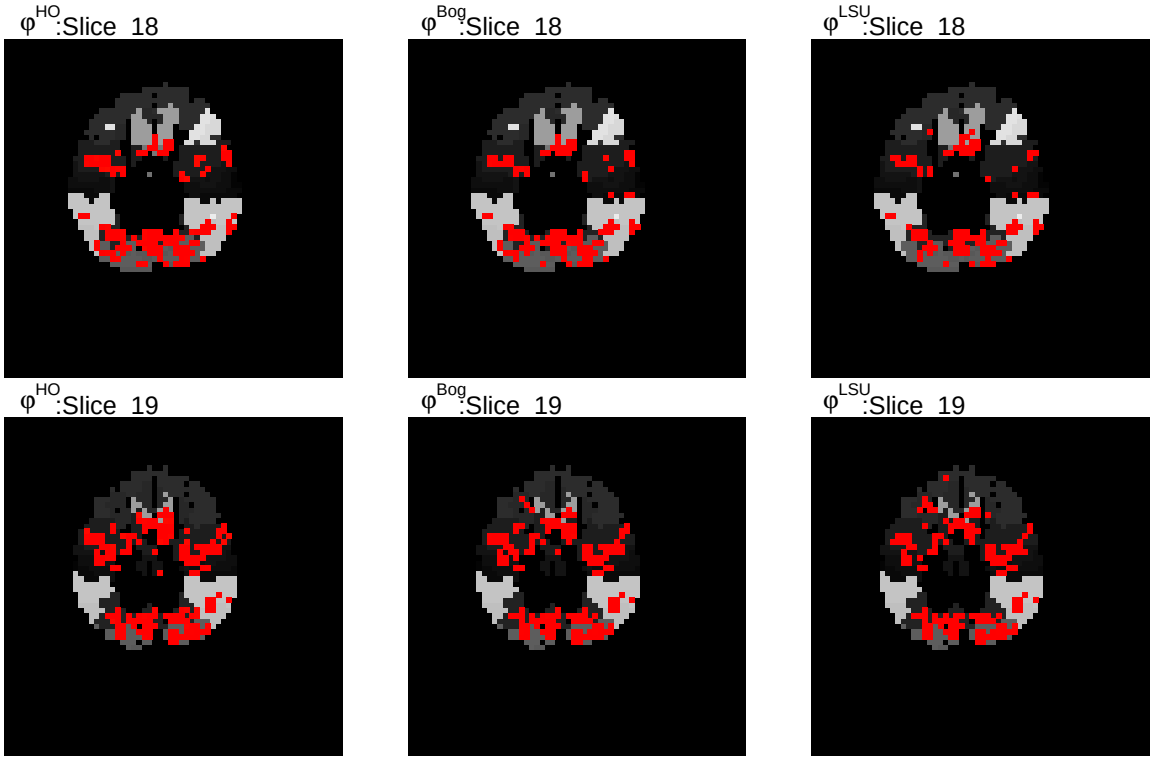

Figure 9: We present chosen slices (motor cortex visible) of the brain for the sports imagination task and highlight the discoveries of the procedure φ^{HO} in the first column. The discoveries in the second column correspond to the procedure φ^{Bog} and in the third column the corresponding discoveries of φ^{LSU}.

Table 2: Global FDR, mean FDR and within family FDR in the fMRI Simulation A and B for the different procedures.

	Simulation A			Simulation B		
	φ^{HO}	φ^{LSU}	φ^{Bog}	φ^{HO}	φ^{LSU}	φ^{Bog}
gFDR	0.0352	0.0333	0.026	0.0362	0.0341	0.0347
FDR Fam. 1	0.0352	0.003	0.0257	0.0354	0.0191	0.0331
FDR Fam. 2	0	0.685	0.006	0.0357	0.0288	0.006
FDR Fam. 3	0	0.696	0.015	0.0356	0.0295	0.014
FDR Fam. 4	0	0.678	0.006	0.0331	0.0419	0.006
FDR Fam. 5	0.001	0.676	0.005	0.0354	0.0285	0.005
FDR Fam. 6	0	0.683	0.01	0.0363	0.0443	0.01
FDR Fam. 7	0	0.693	0.012	0.0355	0.0428	0.012
FDR Fam. 8	0	0.691	0.005	0.0337	0.0603	0.005
mean FDR	0.0045	0.6006	0.0106	0.0351	0.0369	0.0114

concentrated in a-priori known regions the procedure can be used to increase the specificity on the level of the families while finding a similar number of discoveries as the standard approaches within the families of interest. The hierarchical approach was demonstrated to perform close to the non-hierarchical approach if families do not differ in the number of true alternatives. However, we forfeit sensitivity for weak signals if the pre-test is not passed. The use of the Brodmann atlas for the real fMRI data is just a simple example of a division of the brain into functionally different regions, which can (and should) be replaced by more suitable selections in specific applications. In summary our procedure shows superior specificity during the detection of active regions of interest in the brain while being highly sensitive regarding the voxels within a detected region, suggesting good applicability of the FDR in fMRI research.

From a more general perspective, the proposed procedure φ^{HO} is designed to discard families which contain only few scattered signals. This may result in sub-optimal global power, but leads to higher specificity on the group level, compared with non-hierarchical procedures which test all m hypotheses together. Often, as in the fMRI context discussed above, the groups are the experimental units of interest, and in such a situation the hierarchical approach is recommendable. The test φ^{HO} depends on a tuning parameter κ, which has to be chosen by the researcher before the start of the analysis. A value $\kappa \leq m_\ell$ for a family $\mathcal{H}_\ell$ has the interpretation, that a family is declared active if there is evidence that it contains at least κ^{-1} true alternatives. If $\kappa > m_\ell$ the partial conjunction hypothesis becomes the intersection hypothesis.

An interesting and challenging direction for future research is the consideration of additional layers of hierarchy in FDR-controlling multiple test procedures. For example, consider a hierarchical system $\mathcal{H}_m$ of m hypotheses which is closed under intersection. In the case that FWER control at level α is targeted, the closure principle (see Marcus et al. [1976]) allows one to test all m hypotheses in $\mathcal{H}_m$ at full level α, provided that the coherence rule is adhered to (a hypothesis can only be rejected if all its subsets have been rejected). How this principle can be transferred to the concept of (global) FDR control will be explored in future work.

References

Adler, R. J., Taylor, J. E., 2007. Random fields and geometry. New York, NY: Springer.

Benjamini, Y., Bogomolov, M., 2014. Selective inference on multiple families of hypotheses. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1), 297–318.

Benjamini, Y., Heller, R., 2007. False discovery rates for spatial signals. J. Am. Stat. Assoc. 102 (480), 1272–1281.

Benjamini, Y., Heller, R., 2008. Screening for partial conjunction hypotheses. Biometrics 64 (4), 1215–1222.

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

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29 (4), 1165–1188.

Blanchard, G., Dickhaus, T., Roquain, E., Villers, F., 2014. On least favorable configurations for step-up-down tests. Stat. Sin. 24 (1), 1–23.

Bodnar, T., Dickhaus, T., 2014. False discovery rate control under Archimedean copula. Electron. J. Stat. 8 (2), 2207–2241.

Bogomolov, M., 2011. Testing of Several Families of Hypotheses. Ph.D. thesis, Tel-Aviv University.

Brodmann, K., 1909. Vergleichende Lokalisationslehre der Großhirnrinde in ihren Prinzipien dargestellt auf Grund des Zellbaues. Leipzig: Barth.

Chen, S., Wang, C., Eberly, L. E., Caffo, B. S., Schwartz, B. S., 2009. Adaptive control of the false discovery rate in voxel-based morphometry. Hum. Brain Mapp. 30 (7), 2304–2311.

Cox, R. W., 1996. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput. and Biomed. Res. 29, 162–173.

Dickhaus, T., 2014. Simultaneous statistical inference. With applications in the life sciences. Berlin: Springer.

Finner, H., Dickhaus, T., Roters, M., 2009. On the false discovery rate and an asymptotically optimal rejection curve. Ann. Stat. 37 (2), 596–618.

Finner, H., Gontscharuk, V., Dickhaus, T., 2012. False Discovery Rate Control of Step-Up-Down Tests with Special Emphasis on the Asymptotically Optimal Rejection Curve. Scandinavian Journal of Statistics 39 (2), 382–397.

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

Glover, G. H., 1999. Deconvolution of impulse response in event-related BOLD fMRI. NeuroImage 9, 416–429.

Gontscharuk, V., 2010. Asymptotic and Exact Results on FWER and FDR in Multiple Hypotheses Testing. Ph.D. thesis, Heinrich-Heine-Universität Düsseldorf.

Guo, W., Rao, M. B., 2008. On control of the false discovery rate under no assumption of dependency. J. Stat. Plann. Inference 138 (10), 3176–3188.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., Benjamini, Y., 2006. Cluster-based analysis of FMRI data. NeuroImage 33 (2), 599–608.

Hu, J. X., Zhao, H., Zhou, H. H., 2010. False Discovery Rate Control With Groups. J. Am. Stat. Assoc. 105 (491), 1215–1227.

Huettel, S., Song, A., McCarthy, G., 2014. Functional Magnetic Resonance Imaging, 3rd Edition. Sinauer Associates, Inc.

Lazar, N. A., 2008. The Statistical Analysis of Functional MRI Data. Statistics for Biology and Health. Springer.

Li, Y., Ghosh, D., 2014. A two-step hierarchical hypothesis set testing framework, with applications to gene expression data on ordered categories. BMC Bioinformatics 15, Article 108.

Logan, B. R., Geliazkova, M. P., Rowe, D. B., Dec 2008. An evaluation of spatial thresholding techniques in fmri analysis. Hum. Brain Mapp. 29 (12), 1379–1389.

Marcus, R., Peritz, E., Gabriel, K. R., 1976. On closed test procedures with special reference to ordered analysis of variance. Biometrika 63 (3), 655–660.

R Development Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

Sarkar, S. K., 2002. Some results on false discovery rate in stepwise multiple testing procedures. Ann. Stat. 30 (1), 239–257.

Sen, P., 1999. Some remarks on simes-type multiple test of significance. J. Stat. Plan. Inference 82, 139 –145.

Singh, A. K., Phillips, S., 2010. Hierarchical control of false discovery rate for phase locking measures of EEG synchrony. NeuroImage 50 (1), 40–47.

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

Tabelow, K., Polzehl, J., 10 2011. Statistical parametric maps for functional MRI experiments in R: The package fmri. J. Stat. Softw. 44 (11), 1–21.

Tabelow, K., Polzehl, J., 2012. fmri: Analysis of fMRI Experiments. R package version 1.4.8.
URL http://CRAN.R-project.org/package=fmri

Tamhane, A. C., Liu, W., Dunnett, C. W., 1998. A generalized step-up-down multiple test procedure. Can. J. Stat. 26 (2), 353–363.

Welvaert, M., 2012. neuRosim: Functions to generate fMRI data including activated data, noise data and resting state data. R package version 0.2-10.
URL http://CRAN.R-project.org/package=neuRosim

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., Rosseel, Y., 2011. neuRosim: An R package for generating fMRI data. J. Stat. Softw. 44 (10), 1–18.

Worsley, K., 2003. Detecting activation in fMRI data. Stat. Methods in Med. Res. 12, 401–418.

Worsley, K. J., Evans, A. C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12 (6), 900–918.

Yekutieli, D., 2008. Hierarchical false discovery rate-controlling methodology. J. Am. Stat. Assoc. 103 (481), 309–316.

Yekutieli, D., Reiner-Benaim, A., Benjamini, Y., Elmer, G. I., Kafkafi, N., Letwin, N. E., Lee, N. H., 2006. Approaches to multiplicity issues in complex research in microarray analysis. Stat. Neerl. 60 (4), 414–437.

Zhao, H., Zhang, J., 2014. Weighted p-value procedures for controlling FDR of grouped hypotheses. J. Stat. Plann. Inference 151-152, 90–106.

A Theory

A.1 Mathematical proofs

First we introduce the basic setup and notation.

Model 1. *Let $(\Omega, \mathcal{F}, \{\mathbb{P}_\vartheta : \vartheta \in \Theta\})$ be a statistical experiment and let $\mathcal{H} = \{H_1, \ldots, H_m\}$ denote a set of null hypotheses of interest with $\emptyset \neq H_i \subset \Theta$ for all $i \in \{1, \ldots, m\}$. Let $p_i, i \in \{1, \ldots, m\}$, denote the marginal p-value for testing H_i versus $K_i : \Theta \setminus H_i$. A (non-randomized) multiple test procedure $\varphi_{(m)} = (\varphi_1, \ldots, \varphi_m)^\top$ for testing $\mathcal{H}_m$ is a vector of measurable mappings (individual tests) from the sample space into $\{0,1\}^m$. In this, the event $\{\varphi_i = 1\}$ means rejection of the i-th null hypothesis H_i. As convention, the index ℓ will be used to index families, while i is used to index individual hypotheses.*

Relevant quantities.

Definition 6. *Under the assumptions of Model 1, we let the total number of rejections, the number of erroneous rejections, the number of correct rejections, and the* FDR*, respectively, of $\varphi_{(m)}$ be defined as*

$$R_m(\varphi_{(m)}) = |\{i \in \{1, \ldots, m\} : \varphi_i = 1\}|, \tag{6}$$

$$V_m(\varphi_{(m)}) = |\{i \in \{1, \ldots, m\} : \varphi_i = 1 \text{ and } H_i \text{ is true}\}|, \tag{7}$$

$$S_m(\varphi_{(m)}) = |\{i \in \{1, \ldots, m\} : \varphi_i = 1 \text{ and } H_i \text{ is false}\}|, \tag{8}$$

$$\mathrm{FDR}_\vartheta(\varphi_{(m)}) = \mathbb{E}_\vartheta\left[\frac{V_m(\varphi_{(m)})}{R_m(\varphi_{(m)}) \vee 1}\right]. \tag{9}$$

The multiple test $\varphi_{(m)}$ is said to control the FDR *at level $\alpha \in (0,1)$ if*

$$\sup_{\vartheta \in \Theta} \mathrm{FDR}_\vartheta(\varphi_{(m)}) \leq \alpha.$$

It is said to control the FDR *asymptotically at level α as $m \to \infty$ if*

$$\limsup_{m \to \infty} \sup_{\vartheta \in \Theta} \mathrm{FDR}_\vartheta(\varphi_{(m)}) \leq \alpha.$$

If the m hypotheses are structured in disjoint families $\mathcal{H}_1, \ldots, \mathcal{H}_k$ with $|\mathcal{H}_\ell| = m_\ell$ for $1 \leq k \leq m$, a multiple test $\varphi_{(m_\ell)}$ is applied within each family, and we set $\varphi_{(m)} = (\varphi_{(m_1)}, \ldots, \varphi_{(m_k)})^\top$, we define the global FDR *of $\varphi_{(m)}$ by*

$$\mathrm{gFDR}_\vartheta(\varphi_{(m)}) = \mathbb{E}_\vartheta \left[\frac{\sum_{\ell=1}^k V_{m_\ell}(\varphi_{(m_\ell)})}{\left\{\sum_{\ell=1}^k R_{m_\ell}(\varphi_{(m_\ell)})\right\} \vee 1} \right].$$

In the sequel, all considered multiple test procedures are such that the quantities in (6) - (9) actually only depend on the joint distribution of the (random) p-values $p_1, \ldots, p_m$, and one may assume that $(\Omega, \mathcal{F}) = ([0,1]^m, \mathcal{B}([0,1]^m))$ without loss of generality.

Critical value functions and rejection curves. The critical values $\alpha_{i:m}$ from Definition 2 may be defined in terms of a critical value function $\rho : [0,1] \to [0,1]$, where ρ is non-decreasing and continuous, $\rho(0) = 0$ and $\alpha_{i:m} = \rho(i/m), i \in \{1, \ldots, m\}$. For a given critical value function ρ, the function r defined by $r(t) = \inf\{u : \rho(u) = t\}$ for $t \in [0,1]$ is called the rejection curve corresponding to ρ.

The AORC $r_\alpha : [0,1] \to [0,1]$ is defined by

$$r_\alpha(t) = \frac{t}{t(1-\alpha) + \alpha}, \quad t \in [0,1],$$

and the corresponding critical value function is given by $r_\alpha^{-1}(t) = 1 - r_\alpha(1-t)$, see Finner et al. [2009]. The critical values induced by this critical value function are the ones given in (2).

Lemma 1 (Sen [1999])**.** *Denote the empirical cumulative distribution function (ecdf) of the p-values $p_1, \ldots, p_m$ by $\hat{F}_m$, given by*

$$\hat{F}_m(t) = \sum_{i=1}^m \mathbb{I}_{[0,t]}(p_i).$$

Assume that $\alpha_{i:m} = \rho(i/m), i \in \{1, \ldots, m\}$ for a critical value function ρ with corresponding rejection curve r. Then it holds

$$p_{i:m} \leq \alpha_{i:m} \text{ if and only if } \hat{F}_m(p_{i:m}) \geq r(p_{i:m}).$$

Additional technical assumptions. Let $m_{N\ell}$ denote the number and $q_{N\ell}(m_\ell) = m_{N\ell}/m_\ell$ the proportion of true null hypotheses in family $\ell \in \{1, \ldots, k\}$. Define $\pi_\ell(m) = m_\ell/m$ as the proportion of hypotheses belonging to family ℓ. Consider an asymptotic setting such that $\forall \ell \in \{1, \ldots, k\}$: $m_\ell \to \infty$. For convenience, we assume $\pi_\ell(m) \to \pi_\ell \in (0,1)$ and $q_{N\ell}(m_\ell) \to q_{N\ell} \in [0,1]$.

Let $\vartheta^* = \vartheta^*(m_{N1}, \ldots, m_{Nk})$ denote a parameter value such that for every family $\mathcal{H}_\ell, 1 \leq \ell \leq k$, the $m_{N\ell}$ p-values corresponding to true null hypotheses are uniformly distributed on $[0,1]$ and jointly stochastically independent, and that the remaining $(m_\ell - m_{N\ell})$ p-values corresponding to false null hypotheses are almost surely equal to zero. Such a parameter value is commonly referred to as a Dirac-uniform configuration, see, e. g., Section 2.2.2 of Dickhaus [2014] and references therein. Notice that ϑ^* does not necessarily have to be contained in Θ. Under ϑ^*, the ecdf of the m_ℓ p-values in family $\mathcal{H}_\ell$, say $\hat{F}_{m_\ell,\ell}$, converges in the Glivenko-Cantelli sense to $\hat{F}_{\infty,\ell}$, given by $\hat{F}_{\infty,\ell}(t) = (1 - q_{N\ell}) + q_{N\ell} t$, $t \in [0,1]$. Furthermore, r_α and $\hat{F}_{\infty,\ell}$ possess a unique point of intersection on $[0,1)$, cf. Figure 5.2 of Dickhaus [2014]. We denote by $t_{q_{N\ell}}$ the abscissa of this point of intersection.

In general $t = \alpha_{i:m}$ is called a crossing point between $\hat{F}_m$ and r if it satisfies $\hat{F}_m(p_{i:m}) \geq r(p_{i:m})$ and $\hat{F}_m(p_{i+1:m}) < r(p_{i+1:m})$ for $i \in \{1, \ldots, m-1\}$ or $\hat{F}_m(p_{m:m}) \geq r(p_{m:m})$ for $i = m$.

Finally, we introduce the following assumption regarding the type I error behavior of φ^{HO} with respect to the parameter ϑ of the statistical model.

Assumption 1. *For given numbers $m_{N1}, \ldots, m_{Nk}$, the parameter value $\vartheta^* = \vartheta^*(m_{N1}, \ldots, m_{Nk})$ is a least favorable parameter configuration (LFC) for the FDR of $\varphi^{HO}_{(m_\ell)}$, $1 \leq \ell \leq k$, at least asymptotically as $\min_{1 \leq \ell \leq k} m_\ell \to \infty$, where $\varphi^{HO}_{(m_\ell)}$ denotes the proposed two-stage test applied in family $\mathcal{H}_\ell$. This means that $\mathrm{FDR}_\vartheta(\varphi^{HO}_{(m_\ell)}) \leq \mathrm{FDR}_{\vartheta^*}(\varphi^{HO}_{(m_\ell)})$ for all ϑ which are such that exactly $m_{N\ell}$ null hypotheses are true in family $\mathcal{H}_\ell$, $1 \leq \ell \leq k$.*

Assumption 1 is a standard assumption in FDR theory; see, among others, Blanchard et al. [2014] and Bodnar and Dickhaus [2014] and references therein.

Main results.

Theorem 1. *Let $\vartheta \in \Theta$ and assume that for $1 \leq \ell \leq k$ the multiple test $\varphi_{(m_\ell)}$ is an SUD test based on the critical value function $\rho \leq r_\alpha^{-1}$ (with corresponding rejection curve r). Furthermore, let the assumptions from above be fulfilled and let $\varphi_{(m)} = (\varphi_{(m_1)}, \ldots, \varphi_{(m_k)})^\top$. For notational convenience, let $R_{m_\ell} = R_{m_\ell}(\varphi_{(m_\ell)})$ and $V_{m_\ell} = V_{m_\ell}(\varphi_{(m_\ell)})$.*

If

$$\forall \ell \in \{1, \ldots, k\} : \lim_{m_\ell \to \infty} \mathbb{P}_\vartheta \left(\frac{R_{m_\ell}}{m_\ell} \in (0, r_\alpha(t_{q_{N\ell}(m_\ell)})] \right) = 1,$$

then it holds that

$$\limsup_{m \to \infty} \mathrm{gFDR}_\vartheta(\varphi_{(m)}) \leq \alpha.$$

Proof. The global FDR computes as

$$\mathrm{gFDR}_\vartheta(\varphi_{(m)}) = \mathbb{E}_\vartheta \left[\frac{\sum_{\ell=1}^k V_{m_\ell}}{\left\{ \sum_{\ell=1}^k R_{m_\ell} \right\} \vee 1} \right] = \mathbb{E}_\vartheta \left[\frac{m^{-1} \sum_{\ell=1}^k V_{m_\ell}}{m^{-1} \left(\left\{ \sum_{\ell=1}^k R_{m_\ell} \right\} \vee 1 \right)} \right]. \tag{10}$$

Let $t_{m_\ell} \in [0,1]$ denote the random crossing point between r and the ecdf of the p-values $\hat{F}_{m_\ell,\ell}$ characterizing the rejection rule of $\varphi_{(m)}$. This allows for the representation $R_{m_\ell}/m_\ell = r(t_{m_\ell}) = \hat{F}_{m_\ell,\ell}(t_{m_\ell})$ and $V_{m_\ell} = m_{N\ell} \hat{F}_{Nm_\ell,\ell}(t_{m_\ell})$. This means that the right-hand side of (10) equals

$$\mathbb{E}_\vartheta \left[\frac{\sum_{\ell=1}^k \pi_\ell(m) q_{N\ell} \hat{F}_{Nm_\ell,\ell}(t_{m_\ell})}{\sum_{\ell=1}^k \pi_\ell(m) r(t_{m_\ell})} \right] = \mathbb{E}_\vartheta \left[\frac{\sum_{\ell=1}^k \pi_\ell(m) q_{N\ell} \hat{F}_{Nm_\ell,\ell}(t_{m_\ell}) r(t_{m_\ell})/r(t_{m_\ell})}{\sum_{\ell=1}^k \pi_\ell(m) r(t_{m_\ell})} \right]. \tag{11}$$

An argumentation analogous to the one in the proof of Theorem 4.5 in Gontscharuk [2010] allows us to find an asymptotic non random upper bound for $q_{N\ell} \hat{F}_{Nm_\ell}(t_{m_\ell})/r(t_{m_\ell})$. According to (4), we can choose a $\delta > 0$ and m_ℓ large enough such that $\sup_{t \in [0,1]} |\hat{F}_{Nm_\ell}(t) - F_N(t)| \leq \delta$. Then it holds that

$$q_{N\ell} \hat{F}_{Nm_\ell}(t_{m_\ell})/r(t_{m_\ell}) \leq q_{N\ell} t_{m_\ell}/r(t_{m_\ell}) + \mathcal{O}(\delta) \leq q_{N\ell} t_{q_{N\ell}}/r_\alpha(t_{q_{N\ell}}) + \mathcal{O}(\delta).$$

By design of the function r_α, it holds that $q_{N\ell} t_{q_{N\ell}} / r_\alpha(t_{q_{N\ell}}) = \min\{\alpha, q_{N\ell}\}$. Thus, it holds that the right-hand side of (11) can for eventually all large m_ℓ be bounded from above by

$$\mathbb{E}_\vartheta \left[\frac{\sum_{\ell=1}^k \pi_\ell(m) r_\alpha(t_{m_\ell}) \min\{\alpha, q_{N\ell}\}}{\sum_{\ell=1}^k \pi_\ell(m) r_\alpha(t_{m_\ell})} \right] + \mathcal{O}(\delta).$$

Since δ can be chosen arbitrarily small, this entails

$$\limsup_{m\to\infty} \mathrm{gFDR}_\vartheta(\varphi_{(m)}) \leq \alpha.$$

■

Theorem 2 (Statistical properties of the procedure φ^{HO})**.** *Assume that the assumptions from above are fulfilled. Then, the proposed procedure φ^{HO} defined by Algorithm 2 controls the* FWER *at the stage of the families at level α. Furthermore, the global* FDR *of φ^{HO} and the* FDR *of φ^{HO} within each family are asymptotically bounded by α.*

Proof. Recall that the family $\mathcal{H}_\ell$ is selected at the first stage of analysis if and only if the corresponding conjunction p-value p^{u_ℓ/m_ℓ} does not exceed α/κ. Since $\kappa > k$, the Bonferroni inequality yields the first assertion.

In order to show asymptotic control of the global FDR, we notice that every hypothesis which is rejected by $\varphi^{HO}_{(m_\ell)}$ would also be rejected by $\varphi^{AORC}_{u_\ell,(m_\ell)}$ alone, where $\varphi^{AORC}_{u_\ell,(m_\ell)}$ denotes the SUD test which is applied in family $\mathcal{H}_\ell$ in the second stage of $\varphi^{HO}_{(m_\ell)}$, $1 \leq \ell \leq k$. This follows from the fact that κ and hence, u_ℓ, are fixed constants and the rejection rule of $\varphi^{HO}_{(m_\ell)}$ involves the additional condition regarding p^{u_ℓ/m_ℓ}. Hence, $R_{m_\ell}(\varphi^{HO}_{(m_\ell)}) \leq R_{m_\ell}(\varphi^{AORC}_{u_\ell,(m_\ell)})$. Under ϑ^* (cf. Assumption 1) and by construction of r_α, we have, by setting $t_{q_{N\ell}} = 1$ for $q_{N\ell} < \alpha$, that $R_{m_\ell}(\varphi^{AORC}_{u_\ell,(m_\ell)})/m_\ell \to r_\alpha(t_{q_{N\ell}})$ almost surely, cf. Corollary 5.1.(i) of Finner et al. [2009]. We conclude that $\limsup_{m_\ell\to\infty} R_{m_\ell}(\varphi^{HO}_{(m_\ell)})/m_\ell \leq r_\alpha(t_{q_{N\ell}})$ for all $\vartheta \in \Theta$. On the other hand, consider for each $1 \leq \ell \leq k$ such that $\mathcal{H}_\ell$ has been selected at the first stage of analysis the following chain of inequalities:

$$\begin{aligned}
p_{u_\ell:m_\ell} &\leq \min_{j=1,\ldots,(m_\ell-u_\ell+1)} \left\{ p_{(u_\ell-1+j):m_\ell} \right\} \\
&\leq p^{u_\ell/m_\ell} = \min_{j=1,\ldots,(m_\ell-u_\ell+1)} \left\{ \frac{(m_\ell-u_\ell+1)}{j} p_{(u_\ell-1+j):m_\ell} \right\} \\
&\leq \frac{\alpha}{\kappa} \leq r_\alpha^{-1}\left(\frac{m_\ell/\kappa}{m_\ell}\right) \leq r_\alpha^{-1}\left(\frac{\lfloor 1/\kappa \cdot m_\ell \rfloor + 1}{m_\ell}\right) = r_\alpha^{-1}\left(\frac{u_\ell}{m_\ell}\right).
\end{aligned}$$

Thus, if the family $\mathcal{H}_\ell$ is rejected, the SUD procedure $\varphi^{AORC}_{u_\ell,(m_\ell)}$ will reject at least u_ℓ hypotheses within $\mathcal{H}_\ell$. Notice that, by definition of u_ℓ, we have that $u_\ell/m_\ell \geq \kappa^{-1}$. We conclude that, in each selected family $\mathcal{H}_\ell$, $\liminf_{m_\ell\to\infty} R_{m_\ell}(\varphi^{HO}_{(m_\ell)})/m_\ell > 0$. Thus, Theorem 1 can be applied with k replaced by $|\{1 \leq \ell \leq k : \mathcal{H}_\ell \text{ has been rejected}\}|$.

Asymptotic FDR control within each family can be established as follows. If a family $\mathcal{H}_\ell$ is not rejected, we have $R_{m_\ell}(\varphi^{HO}_{(m_\ell)}) = V_{m_\ell}(\varphi^{HO}_{(m_\ell)}) = 0$. On the other hand, in each selected family $\mathcal{H}_\ell$, it holds $V_{m_\ell}(\varphi^{HO}_{(m_\ell)}) \leq V_{m_\ell}(\varphi^{AORC}_{u_\ell,(m_\ell)})$ by the same argumentation as for $R_{m_\ell}(\varphi^{HO}_{(m_\ell)})$. Under the LFC ϑ^*, this also entails that

$$\frac{V_{m_\ell}(\varphi^{HO}_{(m_\ell)})}{R_{m_\ell}(\varphi^{HO}_{(m_\ell)}) \vee 1} \leq \frac{V_{m_\ell}(\varphi^{AORC}_{u_\ell,(m_\ell)})}{R_{m_\ell}(\varphi^{AORC}_{u_\ell,(m_\ell)}) \vee 1}$$

almost surely, because the structure of an SUD test yields that, as soon as $V_{m_\ell}(\varphi^{HO}_{(m_\ell)}) \geq 1$, we have $R_{m_\ell}(\varphi^{HO}_{(m_\ell)}) = V_{m_\ell}(\varphi^{HO}_{(m_\ell)}) + (m_\ell - m_{N\ell})$, and the mapping $x \mapsto x/(x+a)$ is isotone in $x > 0$ for $a \geq 0$. Since $\varphi^{AORC}_{u_\ell,(m_\ell)}$ asymptotically controls the FDR under ϑ^*, this implies the assertion. ■

A.2 The choice of the tuning parameter κ

Here, we report results of a power study regarding the tuning parameter κ. The study was done in two setups for the normal means problem with effect size μ_c and variance 1, analogous to the simulations in Section 3.1. Our theoretical investigations indicate that we can expect the power of the procedure φ^{HO} within one selected family $\mathcal{H}_\ell$ (in our case of size $m_\ell = 2{,}000$) to depend on the ratio of true null hypotheses $q_{N\ell}$ within the family. To this end, we considered a balanced and a highly unbalanced case by setting $q_{N\ell} \in \{0.5, 0.99\}$. In both cases the power of φ^{HO} has been estimated as a function of $\mu_c \in [0, 5]$, and we let the parameter κ range from 1 to $10{,}000{,}000$ on a $\log_{10}$ scale.

The plots in Fig. 10 indicate that small values of κ lead to a good specificity in case of a large value of $q_{N\ell}$, while large values of κ lead to a good sensitivity in case of a moderate value of $q_{N\ell}$. This is line with the recommendation that κ should be chosen according to the amount of signals within a family which is considered relevant.

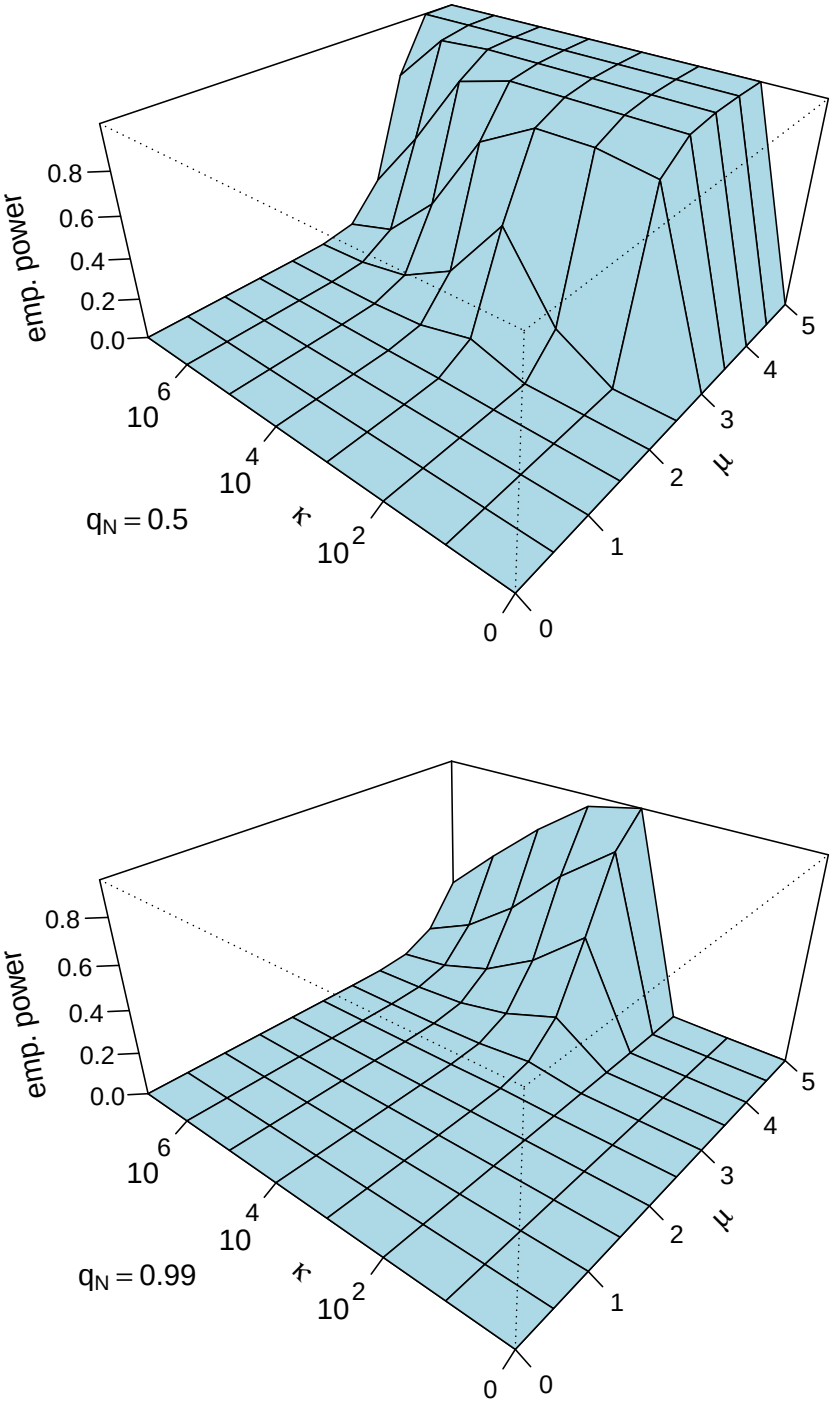

Figure 10: Empirical power of the procedure φ^{HO} for different choices of the fraction of true null hypotheses out of the hypotheses in the family q_N in dependence of the signal strength μ in the normal mean problem with variance 1 and the tuning parameter κ.

B Tables

For sake of completeness we present the full results regarding the rejections in the fMRI data sets.

Table 3: Number of discoveries in the SPM auditory experiment overall and in each Brodmann area of the procedure φ^{HO}, φ^{Bog} and φ^{LSU}.

Discoveries in	φ^{HO}	φ^{LSU}	φ^{Bog}
the whole brain	507	625	790
Brodmann area 1	0	0	0
Brodmann area 2	0	0	0
Brodmann area 3	0	0	0
Brodmann area 4	0	4	0
Brodmann area 5	0	1	0
Brodmann area 6	0	26	22
Brodmann area 7	0	0	0
Brodmann area 8	0	1	0
Brodmann area 9	0	3	2
Brodmann area 10	0	9	7
Brodmann area 11	0	58	102
Brodmann area 12	0	0	0
Brodmann area 13	0	0	0
Brodmann area 14	0	0	0
Brodmann area 15	0	0	0
Brodmann area 16	0	0	0
Brodmann area 17	0	6	0
Brodmann area 18	0	4	1
Brodmann area 19	0	6	0
Brodmann area 20	0	79	94
Brodmann area 21	157	108	155
Brodmann area 22	210	130	201
Brodmann area 23	0	0	0
Brodmann area 24	0	0	0
Brodmann area 25	0	1	0
Brodmann area 26	0	0	0
Brodmann area 27	0	0	0
Brodmann area 28	0	1	1
Brodmann area 29	0	0	0
Brodmann area 30	0	1	0
Brodmann area 31	0	0	0
Brodmann area 32	0	2	0
Brodmann area 33	0	0	0
Brodmann area 34	0	2	0
Brodmann area 35	0	1	1
Brodmann area 36	0	7	13
Brodmann area 37	0	18	18
Brodmann area 38	20	19	25
Brodmann area 39	0	0	0
Brodmann area 40	0	7	7
Brodmann area 41	18	11	17
Brodmann area 42	0	1	1
Brodmann area 43	0	0	0
Brodmann area 44	0	2	0
Brodmann area 45	0	3	1
Brodmann area 46	0	5	0
Brodmann area 47	0	21	21
Brodmann area 48	102	88	101

Table 4: Number of discoveries in the Imagination data set for each Brodmann area of the procedure φ^{HO}, φ^{Bog} and φ^{LSU}.

Discoveries in	φ^{HO}	φ^{LSU}	φ^{Bog}
the whole brain	1756	1661	1764
Brodmann area 1	0	5	7
Brodmann area 2	0	22	22
Brodmann area 3	41	39	39
Brodmann area 4	164	124	150
Brodmann area 5	28	27	27
Brodmann area 6	554	403	486
Brodmann area 7	215	150	180
Brodmann area 8	0	11	4
Brodmann area 9	0	10	2
Brodmann area 10	0	5	0
Brodmann area 11	0	10	0
Brodmann area 12	0	0	0
Brodmann area 13	0	0	0
Brodmann area 14	0	0	0
Brodmann area 15	0	0	0
Brodmann area 16	0	0	0
Brodmann area 17	0	36	39
Brodmann area 18	197	147	176
Brodmann area 19	148	116	135
Brodmann area 20	0	44	26
Brodmann area 21	50	53	48
Brodmann area 22	0	19	16
Brodmann area 23	0	2	0
Brodmann area 24	0	19	20
Brodmann area 25	0	3	0
Brodmann area 26	0	0	0
Brodmann area 27	0	0	0
Brodmann area 28	0	2	0
Brodmann area 29	0	0	0
Brodmann area 30	0	7	6
Brodmann area 31	0	0	0
Brodmann area 32	17	21	17
Brodmann area 33	0	0	0
Brodmann area 34	0	2	2
Brodmann area 35	0	2	1
Brodmann area 36	0	2	0
Brodmann area 37	97	89	94
Brodmann area 38	0	16	10
Brodmann area 39	35	35	35
Brodmann area 40	43	42	41
Brodmann area 41	0	8	7
Brodmann area 42	44	27	35
Brodmann area 43	0	1	1
Brodmann area 44	0	9	7
Brodmann area 45	0	6	2
Brodmann area 46	25	26	25
Brodmann area 47	0	13	9
Brodmann area 48	98	108	95

C Full figures

Here we present the full figures from the fMRI data analysis.

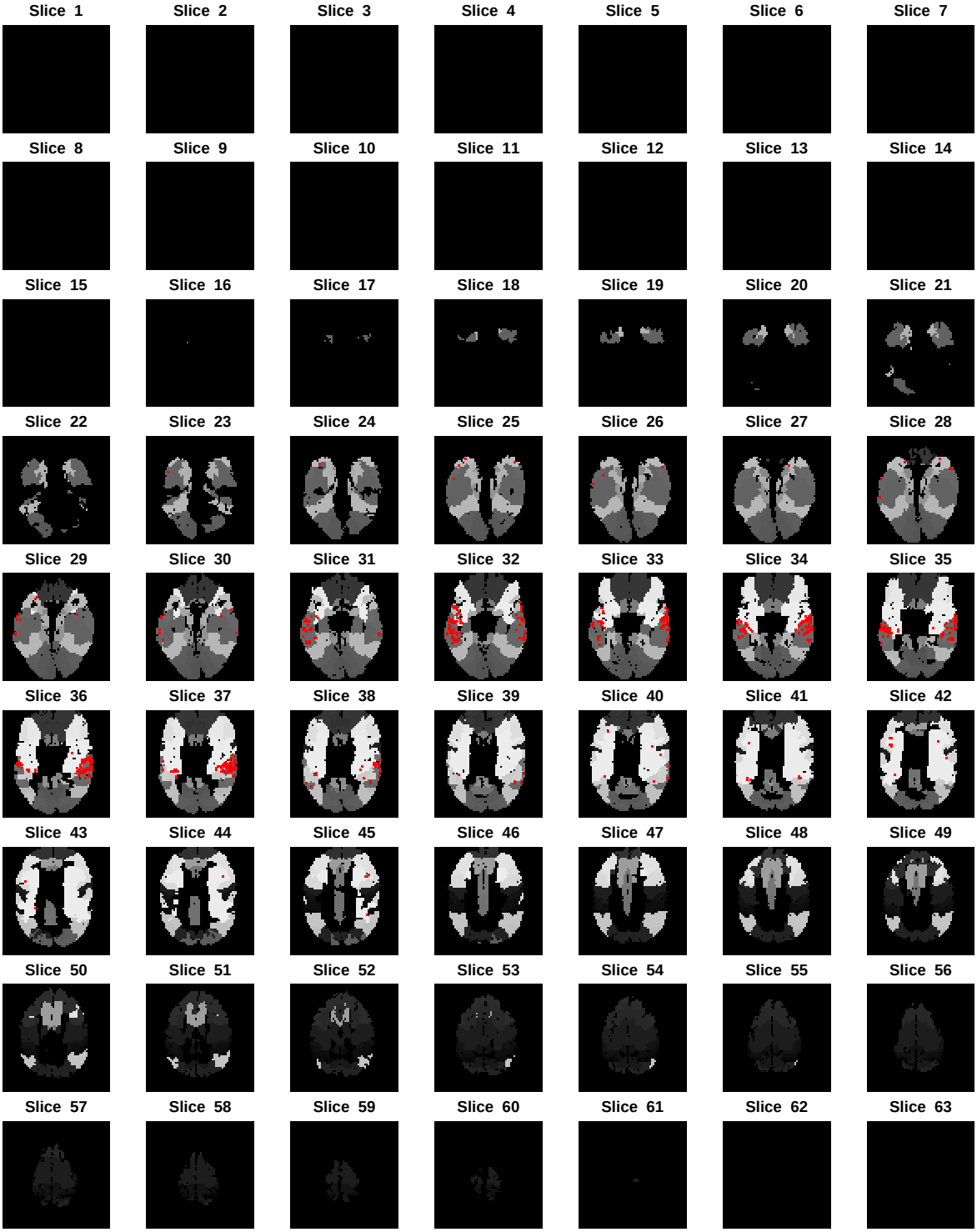

Figure 11: Discoveries of the proposed procedure φ^{HO} (red) for the SPM auditory fMRI dataset overlayed on the Brodmann areas of the brain.

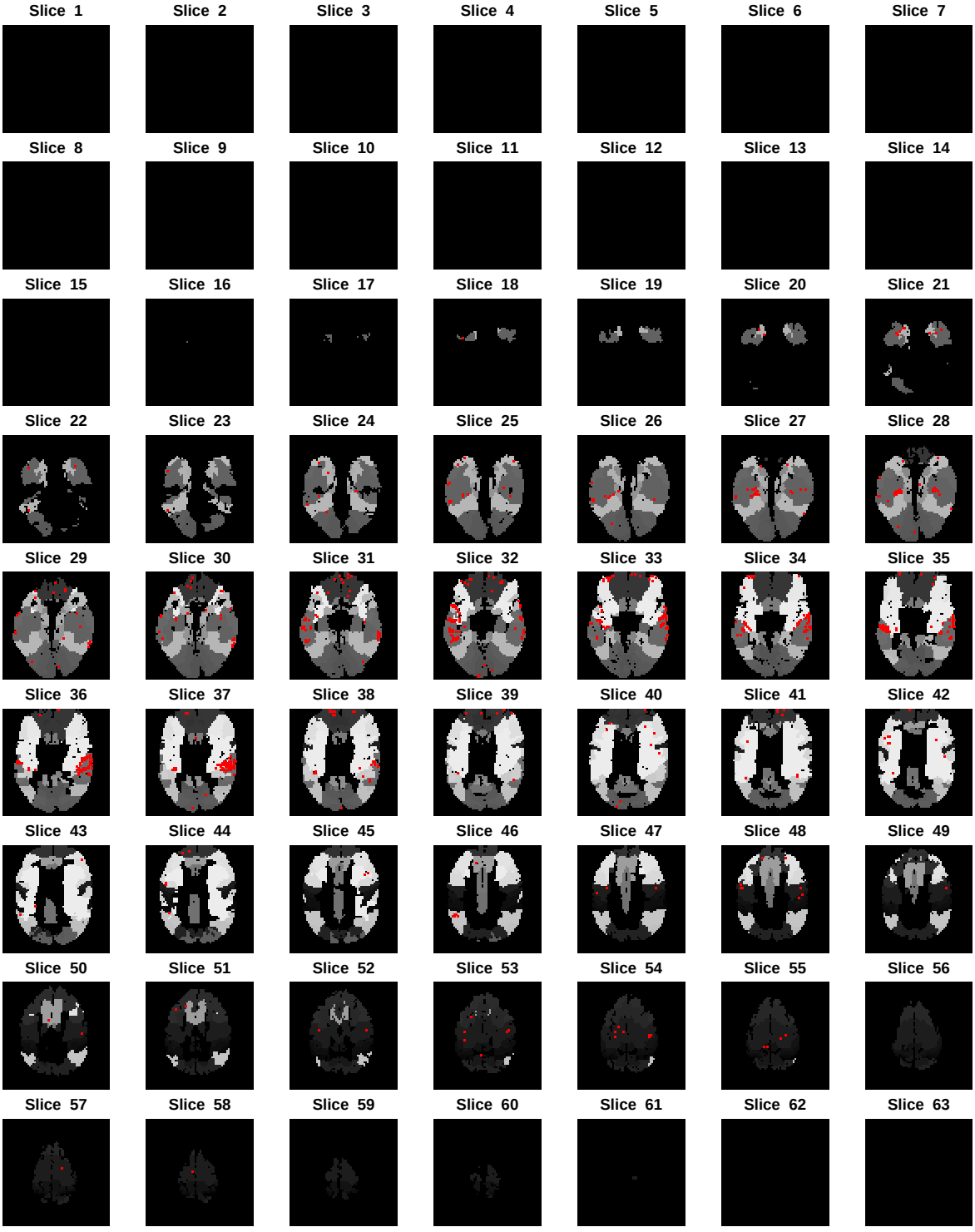

Figure 12: Discoveries of the procedure φ^{LSU} (red) for the SPM auditory fMRI dataset overlayed on the Brodmann areas of the brain.

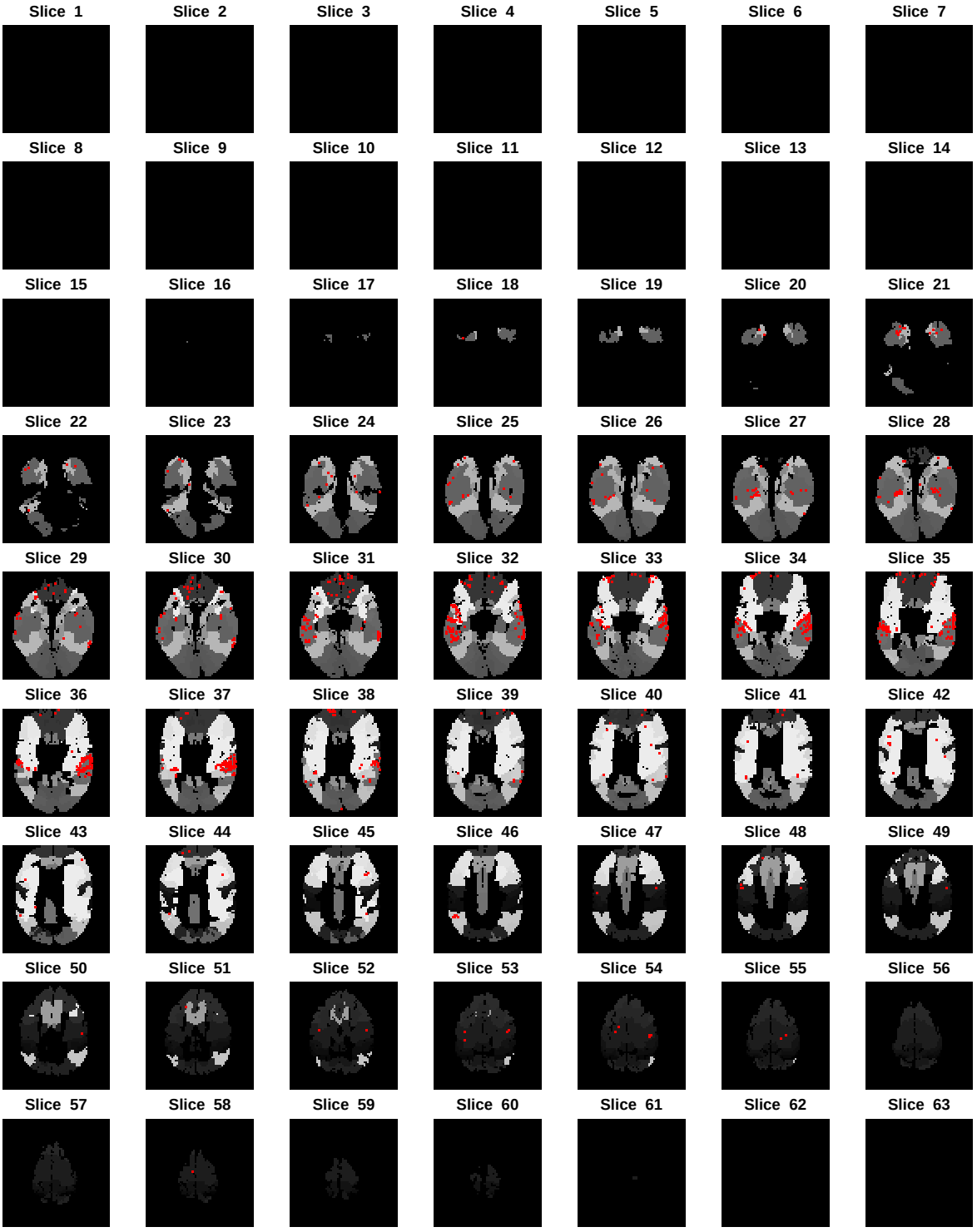

Figure 13: Discoveries of the procedure φ^{Bog} (red) for the SPM auditory fMRI dataset overlayed on the Brodmann areas of the brain.

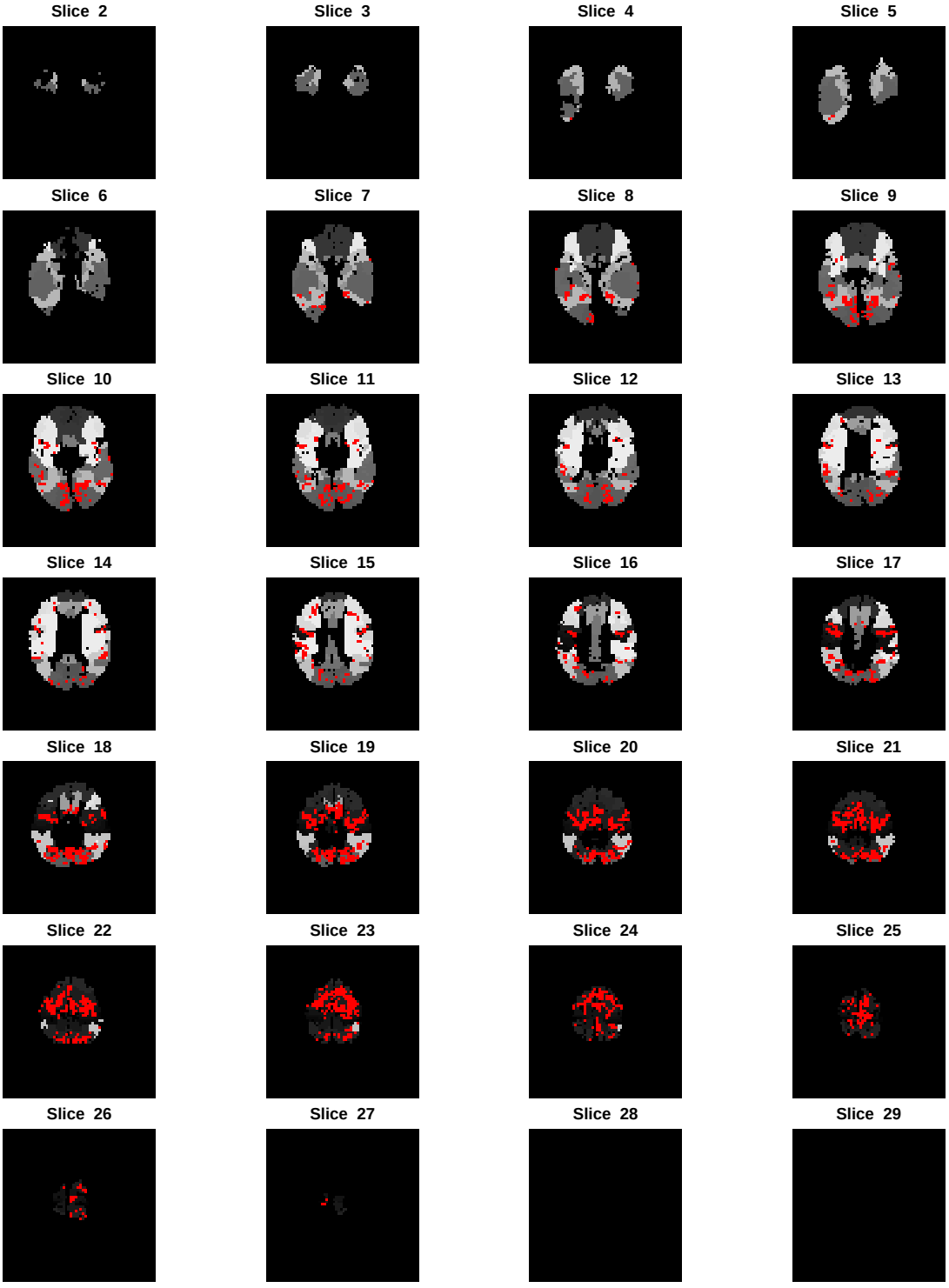

Figure 14: Discoveries of the procedure φ^{HO} for the Imagination dataset overlayed on the Brodmann areas of the brain.

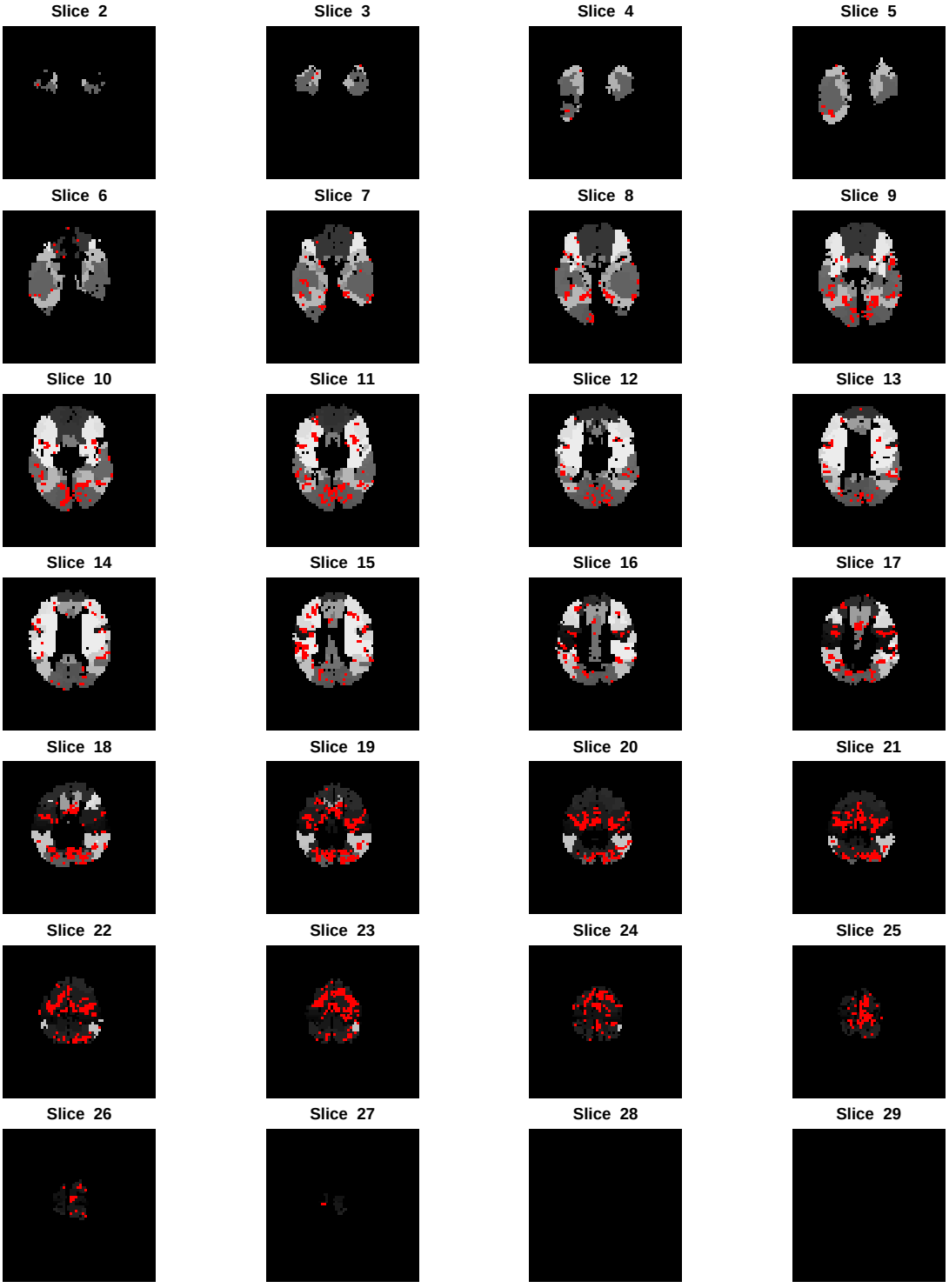

Figure 15: Discoveries of the procedure φ^{LSU} for the Imagination dataset overlayed on the Brodmann areas of the brain.

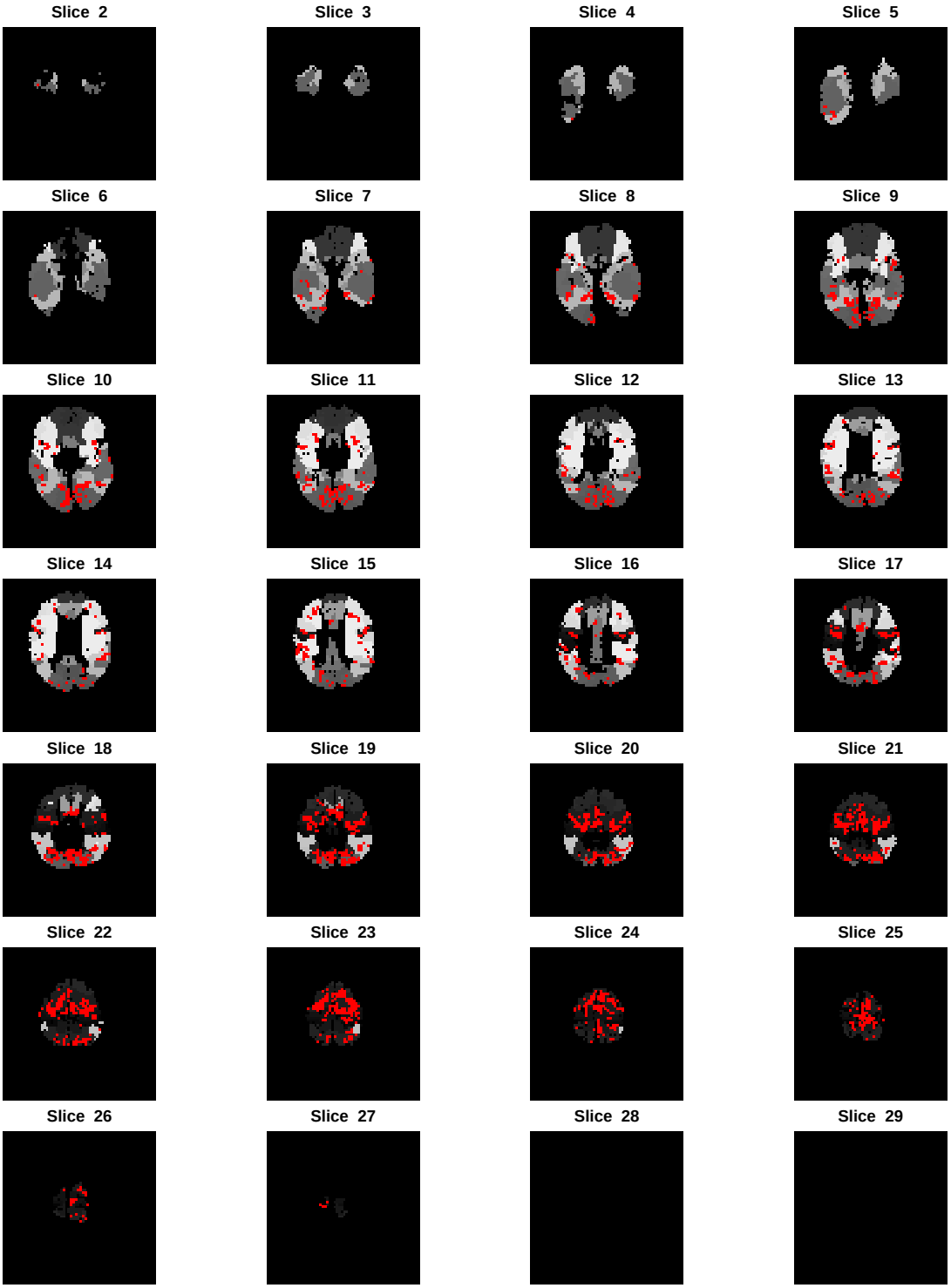

Figure 16: Discoveries of the procedure φ^{Bog} for the Imagination dataset overlayed on the Brodmann areas of the brain.