# Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate

Christopher R. Genovese* Nicole A. Lazar† Thomas Nichols‡

Address of Corresponding Author:

Christopher R. Genovese
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
Email: genovese@stat.cmu.edu
Phone: (412) 268-7836
Fax: (412) 268-7828

*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA. This work partially supported by NSF Grant SES 9866147.

†Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

‡Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.

Abstract

Finding objective and effective thresholds for voxelwise statistics derived from neuroimaging data has been a long-standing problem. With at least one test performed for every voxel in an image, some correction of the thresholds is needed to control the error rates, but standard procedures for multiple hypothesis testing (e.g., Bonferroni) tend to not be sensitive enough to be useful in this context. This paper introduces to the neuroscience literature statistical procedures for controlling the False Discovery Rate (FDR). Recent theoretical work in statistics suggests that FDR-controlling procedures will be effective for the analysis of neuroimaging data. These procedures operate simultaneously on all voxelwise test statistics to determine which tests should be considered statistically significant. The innovation of the procedures is that they control the expected proportion of the rejected hypotheses that are falsely rejected. We demonstrate this approach using both simulations and functional Magnetic Resonance Imaging data from two simple experiments.

**Keywords**: Functional Neuroimaging, False Discovery Rate, Multiple Testing, Bonferroni Correction

# 1 Introduction

A common approach to identifying active voxels in a neuroimaging data set is to perform voxelwise hypothesis tests (after suitable pre-processing of the data) and to threshold the resulting image of test statistics. At each voxel, a test statistic is computed from the data, usually related to the null hypothesis of no difference between specified experimental conditions. The voxels for which the test statistics exceed the threshold are then classified as active, relative to the particular comparison being made. While this approach has proved reasonably effective for a wide variety of testing methods, a basic problem remains: choosing the threshold.

When one uses theoretically-motivated thresholds for the individual tests, ignoring the fact that many tests are being performed, the probability that there will be false positives (voxels declared active when they are really inactive) among all the tests becomes very high. For example, for a one-sided $t$-test with a 0.05 significance level, the threshold would be 1.645, which would lead to approximately 1433 voxels declared active on average out of the 28672 voxels in a $64 \times 64 \times 7$ image when there is no real activity. The 5% error rate thus leads to a very large number of false positives in absolute terms, especially relative to the typical number of true positives.

The traditional way to deal with multiple testing is to adjust thresholds such that Type I error is controlled for all voxels in the brain, simultaneously. There are two types of error control, weak and strong. Weak control requires that, when the null hypothesis is true everywhere, the chance of rejecting one or more tests is less than or equal to a specified level

$\alpha$. Strong control requires that, for any subset of voxels where the null hypothesis is true, the chance of rejecting one or more of the subset's tests is less than or equal to $\alpha$. As concisely stated by Holmes *et al.* (1996), "A test with strong control declares nonactivated voxels as activated with probability at most $\alpha$ ...." A significant result from a test procedure with weak control only implies there is an activation *somewhere*; a procedure with strong control allows individual voxels to be declared active – it has localizing power.

There is a variety of methods available for controlling the false-positive rate when performing multiple tests. Among the methods, perhaps the most commonly used is the Bonferroni correction (see, for example, Miller, 1981). If there are $k$ tests being performed, the Bonferroni correction replaces the nominal significance level $\alpha$ (e.g., 0.05) with the level $\alpha/k$ for each test. It can be shown that the Bonferroni correction has strong control of Type I error. This is a conservative condition, and in practice with neuroimaging data, the Bonferroni correction has a tendency to wipe out both false and true positives when applied to the entire data set. To get useful results, it is necessary to use a more complicated method or to reduce the number of tests considered simultaneously. For instance, one could identify regions of interest (ROI) and apply the correction separately within each region. More involved methods include random field approaches (such as Worsley *et al.*, 1996) or permutation based methods (such as Holmes *et al.*, 1996). The random field methods are suitable only for smoothed data and may require assumptions that are very difficult to check; the permutation method makes few assumptions, but has an additional computational burden and does not account for temporal autocorrelation easily. ROIs are labor intensive to create, and further, they must be created prior to data analysis and left unchanged throughout, a

rigid condition of which researchers are understandably wary.

Variation across subjects has a critical impact on threshold selection in practice. It has frequently been observed that, even with the same scanner and experimental paradigm, subjects vary in the degree of activation they exhibit, in the sense of contrast-to-noise. Subjective selection of thresholds (set low enough that meaningful structure is observed, but high enough so that appreciable random structure is not evident) suggests that different thresholds are appropriate for different subjects. Without an objective method for selecting these thresholds, however, the meaning of the statistical tests can be subverted by the researcher by adjusting the thresholds, implicitly or explicitly, to give desirable results. Many researchers using neuroimaging therefore tend to choose a single threshold consistently for all data analyzed in an individual experiment. This choice is usually based on what has "worked well" in the past. For example, a $t$ threshold of 6 and a p value of less than 0.001 are commonly used, though completely arbitrary, values for thresholding maps. This practice avoids biases from *ad hoc* threshold adjustments, but its forced consistency can significantly reduce sensitivity (and waste data).

There have been a number of efforts to find an objective and effective method for threshold determination (Genovese, Noll, and Eddy, 1997; Worsley *et al.*, 1996; Holmes *et al.*, 1996). While these methods are promising, they all involve either extra computational effort or extra data collection that may deter researchers from using them. In this paper, we describe a recent development in statistics that can be adapted to *automatic and implicit* threshold selection in neuroimaging: procedures that control the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995; Benjamini and Liu, 1999; Benjamini and Yekutieli, 2000).

Whenever one performs multiple tests, the FDR is the proportion of false positives (incorrect rejections of the null hypothesis) among those tests for which the null hypothesis is rejected. We believe that this quantity gets at the essence of what one wants to control, in contrast to the Bonferroni correction, for instance, which controls the rate of false positives among *all* tests whether or not the null is actually rejected. A procedure that controls the FDR bounds the expected rate of false positives among those tests that show a significant result. The procedures we describe operate simultaneously on all voxels in a specified part of the data (e.g., the entire data set) and identify in which of those voxels the test is rejected. This implicitly corresponds to a threshold selection method that adapts to the properties of the given data set. These methods work for any statistical testing procedure for which one can generate a p-value. FDR methods also offer an objective way to select thresholds that is automatically adaptive across subjects.

An outline of the paper is as follows. In Section 2, we describe the FDR in more detail and present a family of FDR-controlling procedures that have been studied in the statistics literature. In Section 3, we present simple simulations that illustrate the performance of the FDR-controlling procedures. In Section 4, we apply the methods to two data sets, one describing a simple motor task (Kinahan and Noll, 1999), and the other from a study of auditory stimulation. Finally, in Section 5, we discuss some of the practical issues in the use of FDR.

## 2 The False Discovery Rate

In a typical functional Magnetic Resonance Imaging (fMRI) data analysis, one computes, for each voxel of interest, a test statistic that relates to the magnitude of a particular contrast among experimental conditions. A voxel is declared active if the corresponding test statistic is sufficiently extreme with respect to the statistic's distribution under the null hypothesis.

Let $V$ denote the total number of voxels being tested in such an analysis. Each voxel can be classified into one of four types, depending on whether or not the voxel is truly active and whether or not it is declared active, as shown in Table 1.

| | Declared Active | Declared Inactive | |
|---|---|---|---|
| Truly Active | $V_{aa}$ | $V_{ai}$ | $T_a$ |
| Truly Inactive | $V_{ia}$ | $V_{ii}$ | $T_i$ |
| | $D_a$ | $D_i$ | $V$ |

Table 1. Classifications of voxels in $V$ simultaneous tests.

For example, $V_{ia}$ denotes the number of false positives and $D_a = V_{aa} + V_{ia}$ denotes the number of voxels declared active. In any data analysis, we only observe $D_a$, $D_i$, and $V$; the remaining counts are unknown.

The False Discovery Rate (FDR) is given by the ratio

$$\text{FDR} = \frac{V_{ia}}{V_{ia} + V_{aa}} = \frac{V_{ia}}{D_a}; \tag{1}$$

that is, the proportion of declared-active voxels which are false positives. If none of the tests is rejected, the FDR is defined to be 0.

A procedure controlling the FDR specifies a rate $q$ between 0 and 1 and ensures that *on average* the FDR is no bigger than $q$. This works even though $V_{ia}$, the number of false positives, is unknown. The phrase "on average" here is important to the interpretation of the procedure. The guarantee is that if one were to replicate the experiment many times, then the average FDR over those replications would be no bigger than $q$. For any particular data analysis, the actual FDR might be larger than $q$.

In contrast to FDR, the Bonferroni procedure controls the probability of having any false positives: one specifies an error rate $\alpha$, and the procedure ensures that $\mathsf{P}\{V_{ia} > 0\} \leq \alpha$. While this does a good job of reducing false positives, it is conservative, meaning that $\mathsf{P}\{V_{ia} > 0\}$ is much less than $\alpha$, and in general the method has low power.

The FDR-controlling techniques introduced by Benjamini and Hochberg (1995) are easily implemented, even for very large data sets. These procedures guarantee control of the FDR in the sense that

$$\mathsf{E}(\text{FDR}) \leq \frac{T_i}{V} q \leq q, \tag{2}$$

where $\mathsf{E}$ denotes expected value and where the first inequality is an equality when the p-values are obtained from a continuous distribution. The unknown factor $T_i/V$, the proportion of truly inactive voxels, shows that the procedure somewhat overcontrols the expected FDR. In analyses of the entire data set, this factor will in practice be very close to 1 and can reasonably be ignored. For analyses of smaller ROIs, however, it might be useful to estimate $T_i/V$, and choose $q$ accordingly.

For the $V$ voxels being tested, the general procedure is as follows:

1. Select a desired FDR bound $q$ between 0 and 1. This is the maximum FDR that the researcher is willing to tolerate on average.

2. Order the p-values from smallest to largest:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(V)}.$$

   Let $v_{(i)}$ be the voxel corresponding to p-value $p_{(i)}$.

3. Let $r$ be the largest $i$ for which

$$p_{(i)} \leq \frac{i}{V}\frac{q}{c(V)},$$

   where $c(V)$ is a predetermined constant described below.

4. Reject the null hypothesis of inactivity for voxels $v_{(1)}, \ldots, v_{(r)}$.

The choice of the constant $c(V)$ depends on assumptions about the joint distribution of the p-values across voxels. The following choices control FDR under different conditions: (i) $c(V) = 1$ and (ii) $c(V) = \sum_{i=1}^{V} 1/i$. The size of the constant in (ii) is larger than that in (i). Hence, all else being equal, the corresponding cut-off for significance and number of voxels declared active are smaller. The first choice of $c(V)$ applies under the assumption of independence among the p-values at different voxels. The second choice of $c(V)$ applies for any joint distribution of the p-values across voxels. As voxels are rarely strictly independent, even in unsmoothed data, we generally favor (ii) for imaging data. Note that $c(V) = \sum_{i=1}^{V} 1/i = \ln(V) + \gamma + r(V)$ where $\gamma \approx 0.5772$ is Euler's constant and $r(V) \leq 1/V$. Hence, for large $V$, one can approximate the harmonic sum with $\ln(V) + \gamma$.

A graphical perspective can be helpful to understanding the procedure. One plots the ordered p-values $p_{(i)}$ and the line through the origin with slope $q/c(V)$ and finds the last undercrossing of the line by the p-values. All null hypotheses corresponding to p-values less than or equal to this undercrossing point are rejected. See Figure 1.

To implement the procedure, one must choose a value for the parameter $q$, but one strength of the method is that this is not an arbitrary choice. From equation (2), $q$ has a meaningful and rigorous interpretation that can be relied on in selecting its value and that makes it comparable across studies. While it is common to set $q$ to conventional levels for significance testing (e.g., 0.01–0.05), this is by no means required. For instance, values of $q$ in the range of 0.10–0.20 are reasonable in many problems (Benjamini, personal communication).

Another advantage of this method is that it is adaptive, in the sense that the chosen thresholds are automatically adjusted to the strength of the signal. The researcher chooses a tolerable rate of false discoveries, and the specific thresholds are determined from the data. This solves the threshold selection problem automatically, even for multiple subjects: there is no need to find an arbitrary and *ad hoc* threshold that works for all subjects simultaneously, or to use a complicated method of targeting the threshold to each subject.

# 3 Simulation Studies

To show how the FDR-controlling procedures perform, we give in this section the results of simulations in which some of the basic parameters ($V$, $T_a$, etc.) are systematically varied.

There are two important points about FDR to keep in mind. First, the procedures guarantee that the FDR will be below the specified bound *on average* over many replications of an experiment. For any given data set, the FDR need not be below the bound. Second, the FDR by itself does not tell us what proportion of the truly active voxels were detected. To find this we would need also the dual quantity, the False Nondiscovery Rate (FNR), which is the proportion of voxels declared inactive that are truly active. That is,

$$\text{FNR} = \frac{V_{ai}}{V_{ai} + V_{ii}} = \frac{V_{ai}}{D_i}, \tag{3}$$

with FNR = 0 if all voxels are declared active. The simulations enable us to find the underlying distribution of the FDR and to compute the FNR to assess power.

In our simulations, we generate random two-sample $t$-statistics (one-sided) that correspond to those computed from a time series of 98 images. We consider two image sizes, $64 \times 64$ and $128 \times 128$, which determines the number of tests performed. Within each image, we include four square blocks of active voxels. The effect size as measured by the shift in the $t$-distribution of the statistic is 0.5, 1, 2, and 3 across the blocks, providing a range of magnitudes for the task-related signal changes from barely detectable to easily detectable. We vary the block size (0, 10, 20, 30) across simulation runs, thus changing the proportion of truly active voxels. Within each run, we obtain 2500 samples using $q = 0.05$. Table 2 shows the simulation results. Figure 2 shows voxel-by-voxel proportions of rejections across one simulation run, with the truly active voxels delineated for comparison.

The expected FDRs follow the pattern predicted by equation 2 quite closely, in that they are all quite close to $T_i q/V = 0.05$. As the proportion of active voxels increases, the

| Image Size | Block Size | $\mathsf{E}$(FDR) | $\frac{T_1}{V}q$ | $\mathsf{P}\{\text{FDR} > q\}$ | $\mathsf{E}$(FNR) |
|---|---|---|---|---|---|
| $64 \times 64$ | 0 | 0.046 | 0.050 | 0.046 | 0.000 |
| | 10 | 0.047 | 0.045 | 0.430 | 0.057 |
| | 20 | 0.038 | 0.030 | 0.187 | 0.211 |
| | 32 | 0.000 | 0.006 | 0.000 | 1.000 |
| $128 \times 128$ | 0 | 0.054 | 0.050 | 0.054 | 0.000 |
| | 10 | 0.049 | 0.049 | 0.432 | 0.023 |
| | 20 | 0.048 | 0.045 | 0.441 | 0.057 |
| | 30 | 0.038 | 0.039 | 0.036 | 0.211 |

Table 2. Summaries of FDR and FNR over replications of simulated data, with $q = 0.05$. Each row in the table represents a different simulation run. In each run, the data set generated at each iteration consists of four blocks of the stated size with different degrees of activation and surrounding non-activating voxels.

distribution of the FDR becomes more concentrated, less skewed, and seems to approach a Gaussian. For the 32 block size of the $64 \times 64$ simulations, there are virtually no false discoveries ($\mathsf{E}(\text{FDR}) \approx 0$), because there are virtually no discoveries ($\mathsf{E}(\text{FNR}) \approx 1$); this suggests that FDR is most powerful with sparse signals. The probability that the FDR is larger than the tolerance $q$ drops precipitously as the number of active voxels increases. Figure 2 shows that the FNR decreases with the effect size. A shift of 0.5 in the $t$-statistics is barely detectable over the background, but a shift of 3 is almost completely recovered.

The FDR-controlling procedure indicates which voxels should be declared active. The largest p-value among these voxels corresponds to a threshold on the original test statistics.

Figure 3 shows the distribution of these equivalent $t$ thresholds across simulation runs for the $128 \times 128$ image with $10 \times 10$ active blocks. The distribution is centered on the value 4.16 with a standard deviation of approximately 0.21. This variation from data set to data set shows the FDR-based method adapting to local variations in the contrast-to-noise ratio.

# 4 Data Example

In this section, we consider the effectiveness of the FDR approach on real data examples. We demonstrate the methods on two datasets. One dataset was described by Kinahan and Noll (1999), where PET and fMRI studies of finger opposition were compared; we use the fMRI data from one subject. The other dataset is from a study of auditory stimulation; it is available on the web, at `http://www.fil.ion.ucl.ac.uk/spm/data`. Both datasets are used here with the kind permission of the respective authors.

For the finger opposition task, subjects sequentially touched their thumb to the fingers of the right hand, starting with the little finger. The movements were synchronized to a numeric visual cue presented at 2-Hz rate for 60 seconds. The control condition was the same visual cue for 60 seconds, though no movement was made. Data from 12 pairs of task-control blocks were collected. A GE 1.5T scanner was used, collecting T2*-weighted EPI images. The acquired volumes had dimensions $128 \times 64 \times 20$, with voxels of size 3.125mm $\times$ 3.125mm $\times$ 4.0mm (no skip); TR was 6 seconds, TE 45 ms. Images were trimmed to $64 \times 64 \times 20$. There were 10 images per block, 12 pairs of blocks and hence a total of 240 images. A $t$-test statistic image was created by comparing the rest to active blocks.

The auditory stimulation experiment consisted of 14 42-second blocks, the blocks alternating between silent rest and presentation of bi-syllabic words. Words were paced at 60 per minute. A modified 2T Siemens scanner was used to collect T2*-weighted EPI images. The acquired volumes had dimensions $64 \times 64 \times 64$, with voxel size 3.0mm $\times$ 3.0mm $\times$ 3.0mm (no skip); TR was 7 seconds. There were 6 images per block, 14 blocks and hence a total of 84 images. For this data we fit the authors' recommended model, a linear regression consisting of a boxcar function convolved with a canonical hemodynamic response, global image intensity and a 7 element discrete cosine basis effecting a high-pass filter with cutoff periodicity of 168 seconds. A $t$-statistic image was created based on the experimental covariate.

Three thresholding methods were applied – the arbitrary cutoff point of 4 in a $t$-map, the basic FDR procedure (with $c(V) = 1$) and the FDR procedure for arbitrary p-value distributions ($c(V) = \sum_{i=1}^{V} 1/i$). Both FDR procedures used $q = 0.05$. Prior to implementation of the FDR method, images were cropped to exclude air outside the head, where no activity should be observed.

As seen in Figures 4 and 5, there is a noticeable difference between the FDR results with and without the independence assumption, with the independence version of FDR leading to many more active voxels. The comparison with the $t$-maps thresholded at 4 in both figures shows that the distribution-free version of FDR highlights basically the same regions, although slightly fewer voxels. These relations are consistent across all the slices. The *ad hoc* threshold of 4 tends to resemble the results under the independence assumption.

# 5 Discussion

We have examined methods to control the False Discovery Rate as a solution to the threshold selection problem in neuroimaging data. These provide an interpretable and adaptive criterion with higher power than other methods for multiple comparisons, such as the Bonferroni correction. In contrast to purely subjective threshold selection, the threshold varies automatically across subjects with a consequent gain in sensitivity. In contrast to complicated threshold-selection schemes, the methods are simple to implement and computationally efficient even for large data sets.

Although the procedure for controlling the FDR was not developed for the case of many thousands of tests and has not often been used in that context, the method gives sensible results with both simulated and real data from two fMRI experiments. As seen in the reported studies, controlling the FDR offers no guarantee that the activation maps will reveal some new structure. What then is the advantage? We see three main strengths of FDR-based methods, all of which derive from the additional information provided about the proportion of voxels falsely declared active.

First, any single choice of threshold across data sets will give an error rate that is too high for some data and too low for others. The FDR method adapts its threshold to the features of the data, eliminating an unnecessary excess of errors. Second, the parameter $q$ has a definite and clear meaning that is comparable across studies. Researchers might obtain different thresholds through their personal choice of $q$, but because the criterion is clear, we can understand the differences that will result. Third, since the FDR-controlling procedure

works on the p-values, and not on the test statistics themselves, it can be applied with any valid statistical test.

Choosing $q$ is only one of the implementation issues that the researcher needs to consider. We have presented two slight variations on the basic procedure that differ in the assumptions they require about the joint distribution of the p-values across voxels. Which of these should be used in a given situation will in general be determined by the nature of the data and the willingness of the researcher to make assumptions about the p-values. When it applies, the independence procedure will have the highest power. However, strict independence is hard to verify and will often fail with neuroimaging data. Indeed, our analysis suggests that the independence FDR procedure is perhaps too liberal for fMRI data, including too many voxels as above threshold. The distribution-free procedure $c(V) = \sum_{i=1}^{V} 1/i$ therefore seems a reasonable default choice.

A second consideration relates to data smoothing. The FDR method becomes more conservative as correlations increase, and hence, it is most powerful for unsmoothed data. This is in contrast to random field methods which are typically more conservative for unsmoothed data.

A third issue is that because FDR procedures operate simultaneously on all voxels included in the analysis, it is important to remove those voxels (e.g., air, CSF in the ventricles) for which we already know the truth. While it is common practice to remove voxels outside the head, it is still a somewhat discretionary step when thresholding voxelwise statistics. For FDR-methods this is a necessary step. However, it is not necessary to be too exacting at boundaries; a few extra voxels here or there will likely have little impact on the results.

We have presented the FDR-controlling procedures here as part of the process of identifying active voxels. More generally, the procedures apply to any multiple testing situation. Many recent methods for the analysis of fMRI data rely on fitting sophisticated statistical models to the data (see, for example, Friston *et al.*, 1994; Genovese, 2000; Lange and Zeger, 1997). Part of such analyses inevitably involves examining the values of fitted parameters at each voxel to test hypotheses about the underlying value of those parameters. FDR-based methods can also be used to perform these voxelwise statistical tests.

# REFERENCES


Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Benjamini, Y. and Liu, W. (1999) A distribution-free multiple test procedure that controls the false discovery rate. Technical Report, Department of Statistics and Operations Research, Tel Aviv University.

Benjamini, Y. and Yekutieli, D. (2000) The control of the false discovery rate under dependence. Technical Report, Department of Statistics and Operations Research, Tel Aviv University.

Friston, K.J., Jezzard, P., and Turner, R. (1994). Analysis of Functional MRI time series, *Human Brain Mapping*, **1**, 153–171.

Genovese, C.R. (2000). A Bayesian time-course model for functional Magnetic Resonance Imaging data, *Journal of the American Statistical Association*, **95**, 691–719.

Genovese, C. R., Noll, D. C. and Eddy, W. F. (1997). Estimating test-retest reliability in fMRI I: Statistical methodology, *Magnetic Resonance in Medicine*, **38**, 497–507.

Holmes, A.P., Blair, R.C., Watson, J.D.G. and Ford, I. (1996) Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow Mapping*, **16**, 7–22.

Kinahan, P.E. and Noll, D.C. (1999) A direct comparison between whole-brain PET and BOLD fMRI measurements of single-subject activation response. *Neuroimage*, **9**, 430–438.

Lange, N. and Zeger, S.L. (1997). Nonlinear Fourier time series analysis for human brain mapping by functional Magnetic Resonance Imaging, *Applied Statistics*, **46**, 1–29.

Miller, R.G., Jr. (1981) *Simultaneous Statistical Inference* (second edition). Springer-Verlag, New York.

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

## Figure Captions

Figure 1. A graphical display of the FDR-controlling procedures. Sorted p-values are plotted in order (with index $i$ at $i/(V+1)$ for $i = 1, \ldots, V$). The point of last undercrossing of the line through the origin with slope $q$ determines that tests with smaller p-values reject their null hypotheses.

Figure 2. Proportions of tests rejected, by voxel, in the simulation runs with image size 128 and block size 20. See also Table 2. Boxes delineate voxels that are truly active. The true shift of the $t$ statistic increases from 0.5, 1, 2, 3, going counter-clockwise from the bottom left. False discoveries correspond to non-zero values outside the delineated boxes; false non-discoveries correspond to non-one values inside the delineated boxes.

Figure 3. Histogram of equivalent $t$ thresholds generated by the FDR-controlling procedure across simulation runs.

Figure 4. Coronal slice of suprathreshold pixels overlayed on mean T2* image. Colored pixels are $-\log_{10}$ of the p-value. Top, $t > 4$ threshold. Middle, threshold controling FDR at 5% based on independence assumption. Bottom, threshold controling FDR at 5% making no assumptions on p-value distribution.

Figure 5. Axial slice of suprathreshold pixels overlayed on T1 structural image. Colored pixels are $-\log_{10}$ of the p-value. Top, $t > 4$ threshold. Middle, threshold controling FDR at 5% based on independence assumption. Bottom, threshold controling FDR at 5% making no assumptions on p-value distribution.

Figure 1

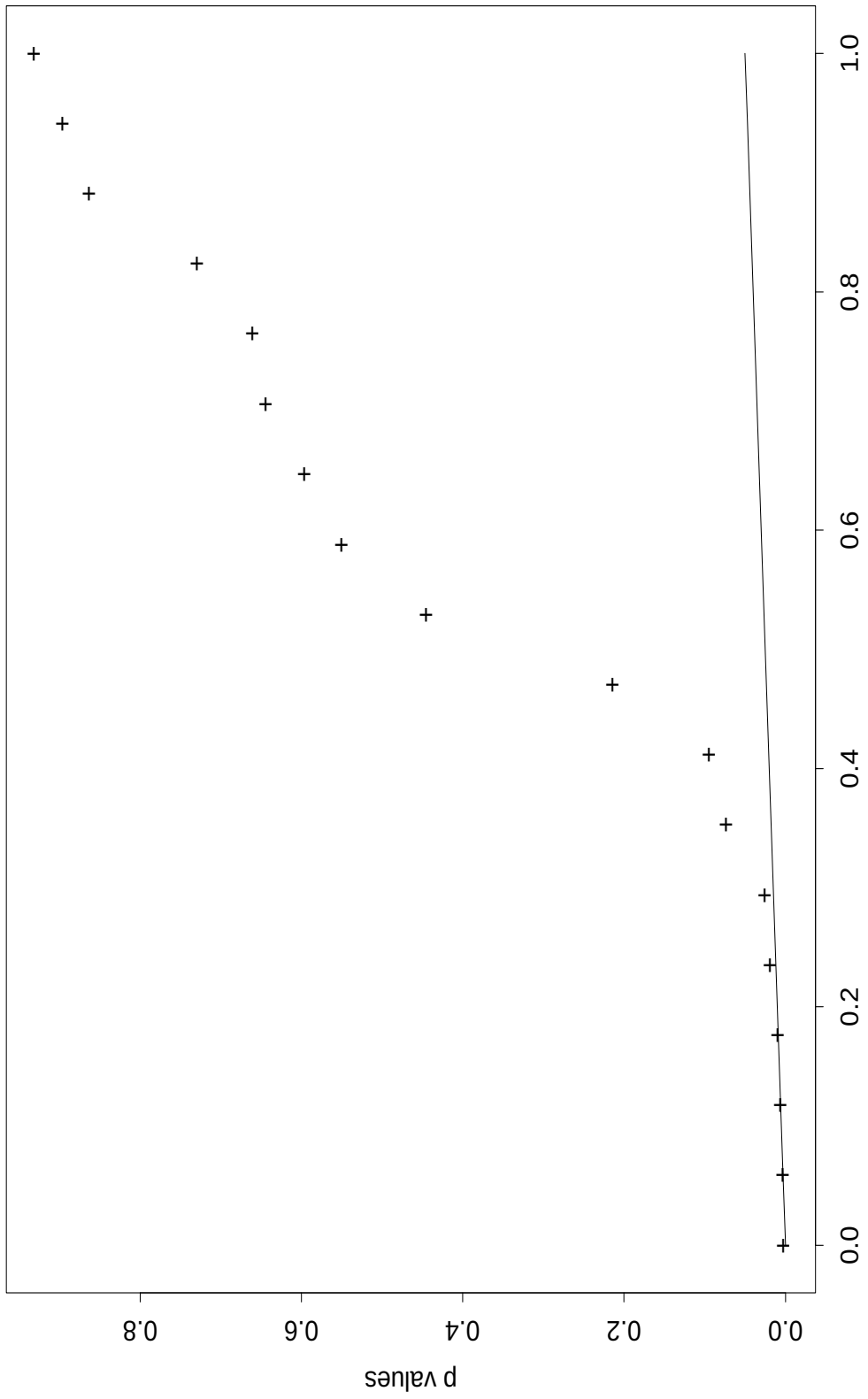

Figure 2

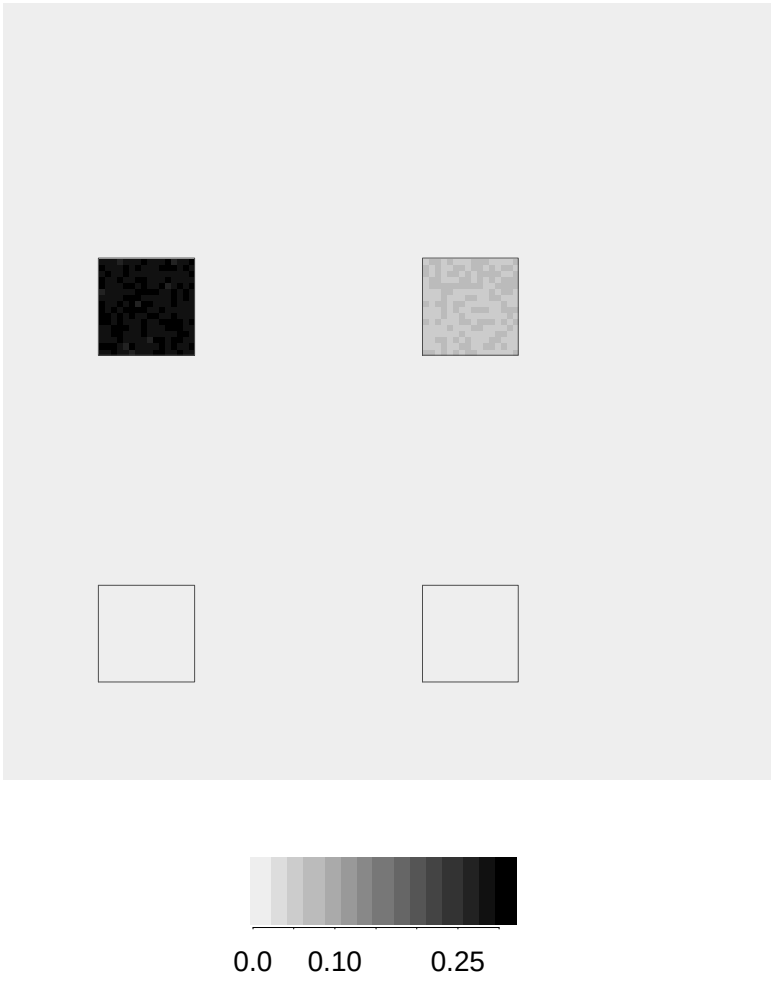

Figure 3

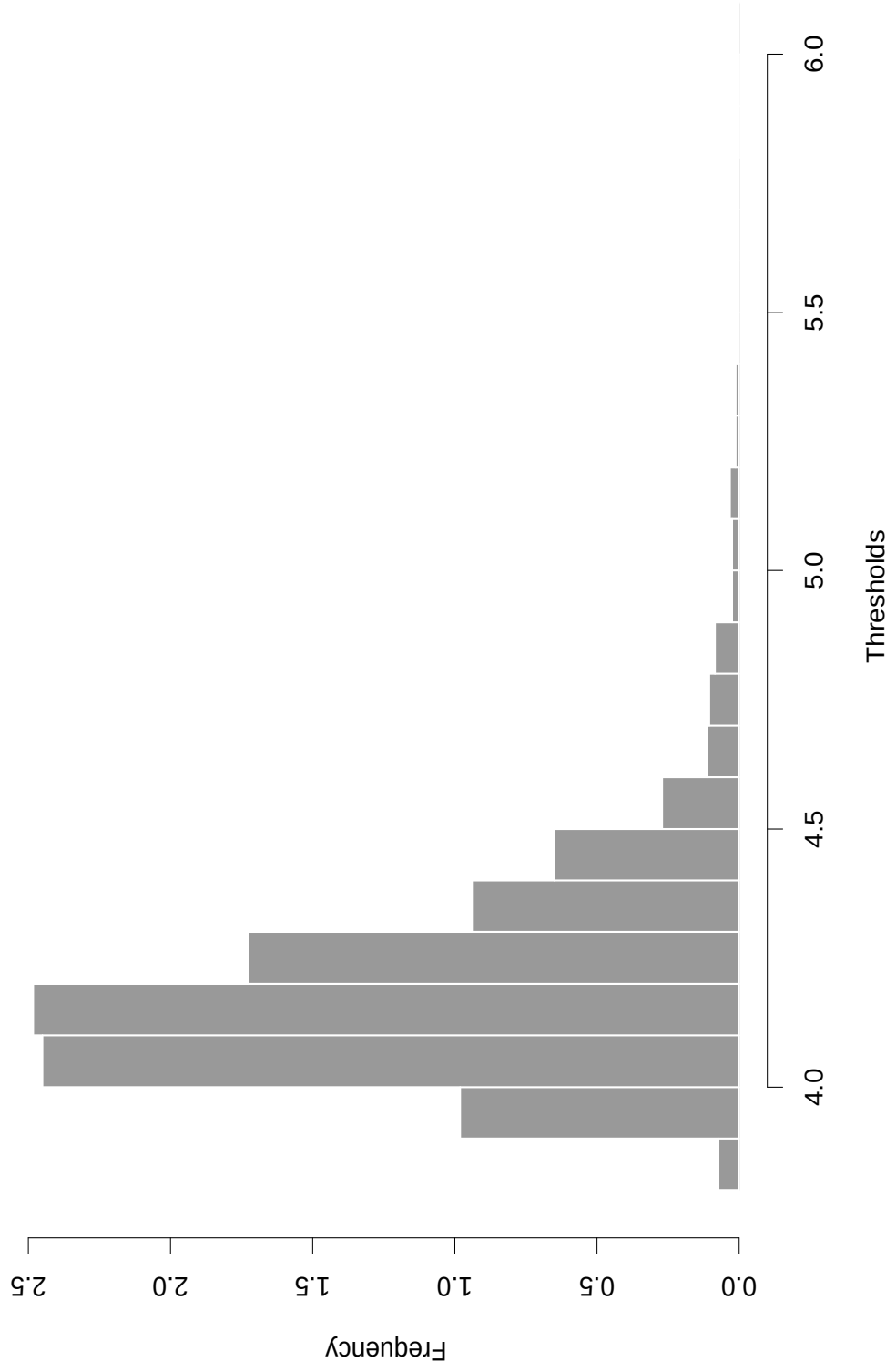

Figure 4

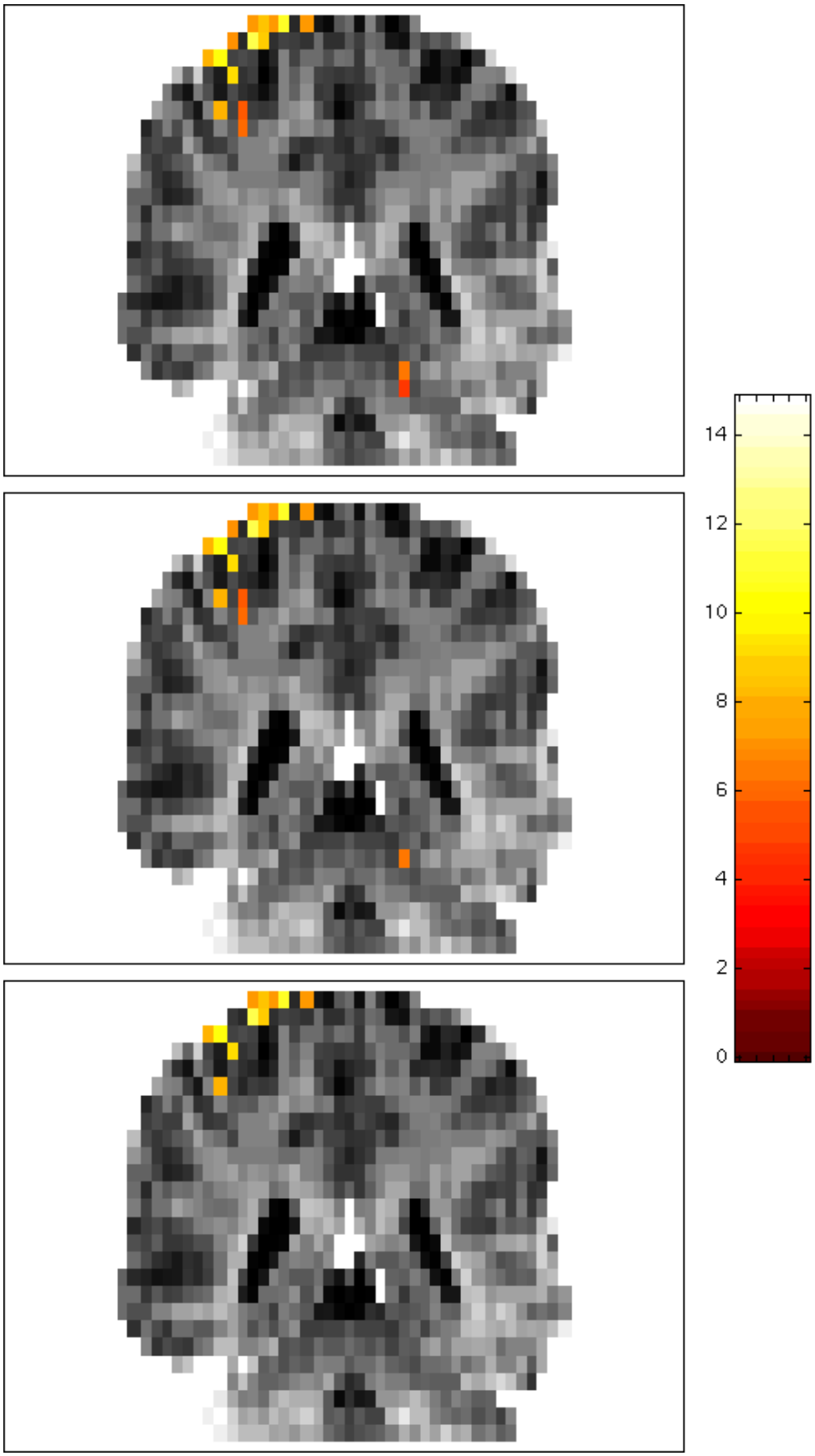

Figure 5