NEW PROCEDURES FOR FALSE DISCOVERY CONTROL

Christopher R. Genovese[*]

Department of Statistics
Carnegie Mellon University

Elisha P. Merriam[†]

Center for the Neural Basis of Cognition
University of Pittsburgh

ABSTRACT

Following Benjamini and Hochberg [2], the False Discovery Rate has emerged as a viable alternative to strong control of Type I error in multiple testing problems such as those which arise in functional neuroimaging. This paper reports on new methods for false discovery control that can usefully be applied to functional neuroimaging data, especially for thresholding statistical maps. The methods are based on controlling the unobserved False Discovery Proportion (FDP) – the number of false rejections divided by the number of rejections – simultaneously over all thresholds. From this, one can design thresholding procedures to control the False Discovery Rate and other features of the false discovery process, including the probability that FDP exceeds a specified level or the expected proportion of false regions. We give two variants on these procedures, one for unsmoothed data and one for smoothed data based on random field models. We illustrate the methods using data from a visual remapping study.

1. INTRODUCTION

Among the many challenges raised by the analysis of functional neuroimaging data is the problem of accurately identifying voxels with nontrivial treatment effects – loosely, "active voxels." Many commonly used analysis methods generate a statistic at each voxel that is used to perform a statistical hypothesis test for the desired effects. Choosing a threshold for so many simultaneous tests sharpens the usual trade-off between power and rate of false positives (type I errors): a low threshold can reveal interesting regions of marginal effect size at the cost of many falsely active voxels; a high threshold can nearly eliminate falsely active voxels but also eliminates many truly active regions of interest.

The traditional approach to threshold selection is simultaneous control of type I error, so that under the null hypothesis, the probability of *any* false positives is no bigger than a prespecified bound. In fact, slightly more is typically demanded — strong control. Strong control requires that over any subset of voxels for which the null hypothesis is true the probability of one or more false positives from among that subset of voxels no greater than the target level. A method with strong control thus allows localization of active voxels. In contrast, with only weak control of type I error, it is only possible to detect whether *some* voxels are active; no localized assessment can be made.

One commonly used method that guarantees strong control of type I error is the Bonferroni correction (see for example [10]) Unfortunately, Bonferroni thresholds tend to have low power (sensitivity), which prevents their widespread use in functional neuroimaging studies.

Viable alternatives between Bonferroni and uncorrected testing do exist. For example, there are procedures that attain more power than Bonferroni while maintaining strong control of type I error [9,10], though in practice the difference appears small. An approach that has been successful in neuroimaging is to model the smoothed data as a realization of a smooth random field and derive thresholds via geometric properties of the sets that exceed various levels [17]. Other common approaches include permutation methods [12], restriction to specific regions of interest, and deriving thresholds from pilot data.

The False Discovery Rate (FDR), introduced by Benjamini and Hochberg ([2], BH), offers an alternative error-control criterion. The FDR is the expected proportion of false positives among all voxels declared active. Let TA denote the number of truly active voxels, FDA the number of voxels falsely declared active, and DA the total number of voxels declared active, correctly or incorrectly. Throughout m will denote the total number of voxels tested. Define the False Discovery Proportion (FDP; [7]) and the FDR by

$$\mathrm{FDP} = \begin{cases} \frac{FDA}{DA} & \text{if } DA > 0 \\ 0 & \text{if } DA = 0. \end{cases} \qquad \mathrm{FDR} = \mathsf{E}\,\mathrm{FDP}, \tag{1}$$

where E denotes expectation. The FDP is an unobserved random variable; the FDR gives the ensemble average of this quantity. Note that both FDP and FDR depend on the threshold used for each test.

Benjamini and Hochberg [2] also demonstrated a simple procedure that controls the FDR. Let $0 = P_{(0)} < P_{(1)} < \ldots < P_{(m)}$ be the p-values for the hypothesis tests sorted in increasing order. For any chosen $0 < \alpha < 1$, the BH method uses a threshold $T = \max\{P_{(j)} : P_{(j)} \le \alpha j/m\}$, so that any null hypothesis with p-value $\le T$ is rejected. This threshold guarantees that

$$\mathrm{FDR} \le (1 - TA/m)\alpha \le \alpha. \tag{2}$$

Note that TA/m is unobserved. The BH technique usually has more power than Bonferroni and related methods of strong con-

*Research supported by NIH Grant 1 R01 NS047493-01.

†Research supported by NSF Grant DGE-9987588.

trol while still providing a principled error control criterion. See [6] for more detail in a neuroimaging context.

The BH method is simple and fast, and the BH FDR-bound holds for certain forms of dependence among test statistics [4]. Yet several features of the BH result have motivated much current research. First, BH overcontrols FDR by the factor $1 - TA/m$, raising the possibility that power can be improved, for example, by estimating the proportion of truly active voxels [3, 7, 14]. Second, though the BH bound (2) holds for dependence structures that include Gaussian statistics with non-negative correlations — an often reasonable but not obviously true assumption in neuroimaging data – it is hard to "break" the bound. This suggests a more general result (see [15] for results in this direction.) Third, the BH method controls the mean of the unobserved FDP, but given that FDP distributions can be skewed, controlling the mean can leave a nontrivial probability of large FDP. Fourth, as with traditional approaches to multiple testing, the BH method imposes no relations among the different tests, but in spatial problems like neuroimaging, spatial clustering in active voxels is informative and can be used to improve power.

Spurred by this considerations, we describe two new methods for false discovery control. The first, which we call confidence thresholds, produces a threshold T that controls the probability that FDP exceeds a pre-specified level. That is, in place of the BH mean bound $\mathrm{FDR} \leq \alpha$, the procedure guarantees that $\mathsf{P}\{\mathrm{FDP} \leq \alpha\} \geq 1 - \gamma$ for chosen $0 < \alpha, \gamma < 1$, typically small. The second method, based on a random-field definition of FDP, constructs a threshold T that controls the proportion of falsely identified *regions* of active voxels.

2. CONFIDENCE THRESHOLDS

Our procedure for confidence thresholds is based on that in [7,8]. Confidence thresholds provide an interpretable and tunable family of procedures with progressively stronger control. As we will see, these can achieve nearly the power obtained by FDR control despite making a stronger guarantee.

The procedure is conceptually simple and computationally fast. We begin by viewing the FDP as a *as a function of threshold*. Note that FDP is unobserved. We construct a function $\overline{\mathrm{FDP}}$ bounds FDP above with high confidence. $\overline{\mathrm{FDP}}$ is obtained by inverting a test for uniformity on all subsets of the p-values. On *every subset* of the p-values from the m voxels, we perform a (level γ) statistical hypothesis test that the corresponding p-values were drawn from a uniform distribution (i.e., that the null is true). Those subsets of p-values for which uniformity cannot be rejected give configurations of true and false null hypotheses that are consistent with the data, and in turn a possible FDP function. Taking the maximum of FDP over all such configurations and all thresholds gives $\overline{FDP}$ which thus satisfies $\mathsf{P}\big\{\mathrm{FDP}(t) \leq \overline{\mathrm{FDP}}(t) \quad \text{for all } t\big\} \geq 1 - \gamma$. Now define the threshold T to be the biggest t such that $\overline{FDP}(t) \leq \alpha$. We call T a $(\alpha, 1 - \gamma)$ confidence threshold.

Of course, it is infeasible in practice to test all subsets of the p-values; fortunately, for a variety of good tests, this is not necessary. A good test for this purpose should allow fast computation and should have highest power for discriminating from uniformity in the smaller p-values, which have the biggest impact on the size of FDP. One family of tests that satisfies both desiderata use the kth order statistic of a subset as the test statistic. For any choice of k, the confidence threshold can be computed in linear time and give good power. A priori choice of k based on anticipated effect sizes tends to work well, but a simple data dependent method — estimating a parametric form of the test statistic distribution and using the optimal k for that distribution — improves power while maintaining nearly nominal coverage.

3. FALSE DISCOVERY CONTROL FOR RANDOM FIELDS

A successful approach to thresholding statistical maps from neuroimaging data is based on modeling the statistics as a realization of a random field [16,17]. False discovery control also carries over to a random field context [13] by defining the False Discovery Proportion as a proportion of *areas* on which the random field exceeds a threshold level. Using a strategy of inverting tests as in confidence envelopes and p-value approximations from random field theory, we can design thresholds to control FDR, tail probabilities, or proportions of false regions.

One interesting application is to define an error criterion on the *regions*, rather than voxels, declared active. Define a declared active region as false at tolerance $0 \leq \tau \leq 1$ if at least a proportion τ of the region's area corresponds to truly inactive locations. The random field method of [13] leads to a threshold T that guarantees, for fixed τ, α, γ, that with probability at least $1 - \gamma$ fewer than a proportion α of the declared active regions are false at tolerance τ.

4. EXAMPLE

To illustrate our methods, we use data from an fMRI study of visual remapping [11]. When the eyes move so that a neuron's receptive field lands on a previously stimulated location, that neuron fires even though no stimulus is present. This implies that eye movements are associated with a transformation — or *remapping* — in neural representation (Duhamel et al. 1992). Monkeys exhibit such visual remapping in parietal cortex, and [11] sought evidence for remapping in human cortex. The experiment uses an event-related design to isolate the visual and remapped responses. Imaging for the subject presented here was acquired on a GE Signa 3T scanner with EPI-RT acquisition, TR 2s, TE 30ms, 20 oblique slices, 3.125mm $\times$ 3.125mm $\times$ 3mm voxels, 1mm gap. See [11] for full details on background, paradigm, and methods.

Here, we present results from one subject to illustrate the methods described above. First, we use unsmoothed data and apply the confidence thresholds described above. Figure 1 shows a surface rendering of the subject's brain with active regions determined by a $(\alpha = 0.05, 1 - \gamma = 0.9)$ confidence threshold. A comparison of the confidence threshold results to the results obtained with Benjamini-Hochberg and Bonferroni thresholds shows that the confidence threshold essentially the same active areas as BH while Bonferroni misses several areas of interest (esp. superior parietal).

Second, we use smoothed data (10mm FWHM Gaussian filter) and apply our random field methods to control the proportion of false clusters. Figure 2 shows a similar surface rendering with thresholded values chosen so that the proportion of false regions is less than 10% with probability at least 90%, where "false regions" means an overlap with truly inactive voxels of at least 10% of volume.

5. DISCUSSION

Research on false discovery control has blossomed in recent years and is now gaining wider use in neuroimaging. The Benjamini-Hochberg method is a fast and effective technique for gaining power while maintaining principled control of errors. Yet there are several directions in which the BH method can be extended. We have presented two such methods.

Confidence thresholds have practical advantages for False Discovery Control. In particular, we gain a stronger inferential guarantee with little effective loss of power. Although it has not yet been proved conclusively, confidence envelopes appear to be valid under the same positive dependence conditions as work for FDR control.

For spatial applications, the shape of an active region can be highly informative, but multiple-testing methods like BH and Bonferroni that ignore possible spatial relations do not account for this. As a step toward region-based false discovery control, we show how to control the proportion of false regions, where false refers to the proportion of area overlapping inactive voxels.

References

[1] Adler, R.J. (1990). *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, Institute of Mathematical Statistics.

[2] 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.

[3] Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational Behavior, Statistics*, **25**, 60–83.

[4] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165-1188.

[5] Duhamel, J. R., Colby, C. L., and Goldberg, M. E. (1992a). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, **255**, 90–92.

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

[7] Genovese, C. R. and Wasserman L. (2004). A stochastic process approach to False discovery control. *The Annals of Statistics*, to appear.

[8] Genovese, C.R. and Wasserman L. (2004). Exact control of the false discovery process. Manuscript in progress.

[9] Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, **9**, 811–818.

[10] Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*, John Wiley & Sons, New York.

[11] Merriam, E.P., Genovese, C.R., and Colby, C.L. (2003). Spatial Updating in Human Parietal Cortex, *Neuron*, **39**, 361–373.

[12] Nichols, T.E. and Holmes, A.P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, **15**, 1–25.

[13] Pacifico Perone, M., Genovese, C.R., Verdinelli, I., and Wasserman, L. (2004). False Discovery Rates for Random Fields, *JASA*, to appear.

[14] Storey J.D. (2003). The positive False Discovery Rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, to appear.

[15] Storey J.D., Taylor J.E. and Siegmund D. (2003). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B*, to appear.

[16] Worsley, K. J., Liao, C.H., Aston, J., Petre, V., Duncan, G.H., Morales, F., and Evans, A.C. (2002). A General Statistical Analysis for fMRI data. *NeuroImage*, **15**, 1–15.

[17] 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.

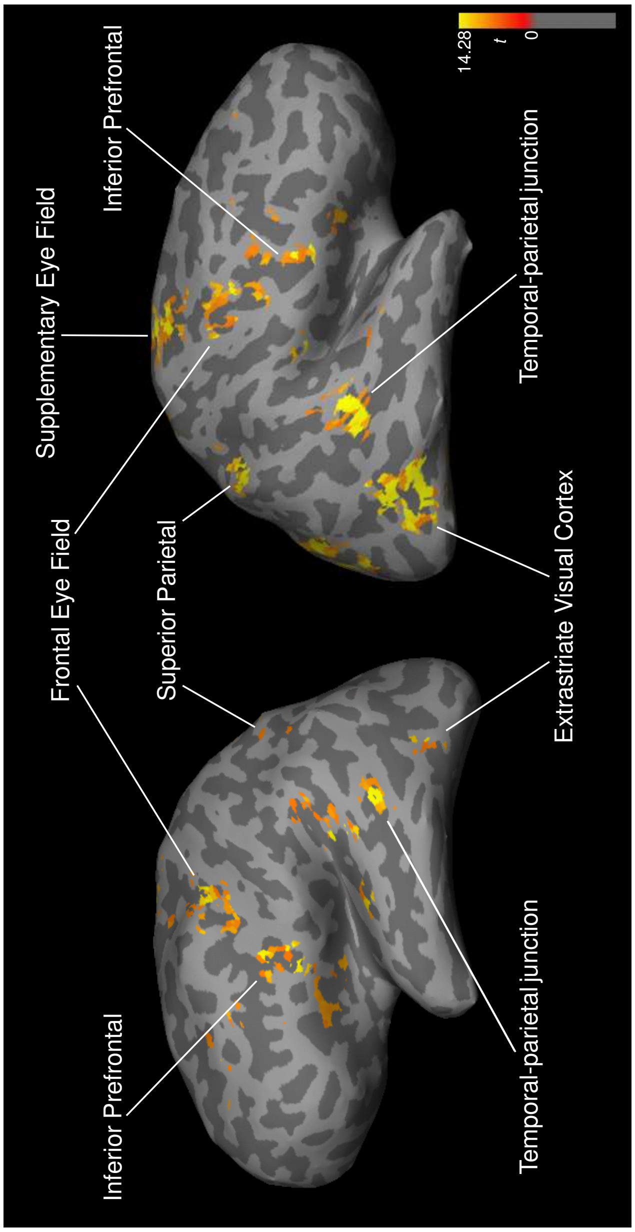

Figure 1. Cortical surface renderings with superimposed active regions determined by $(\alpha = 0.05, 1 - \gamma = 0.9)$ confidence threshold. Tests at each voxel are t-tests for detection of positive amplitude for visual or remapped response in the event-related design. The data were not smoothed.

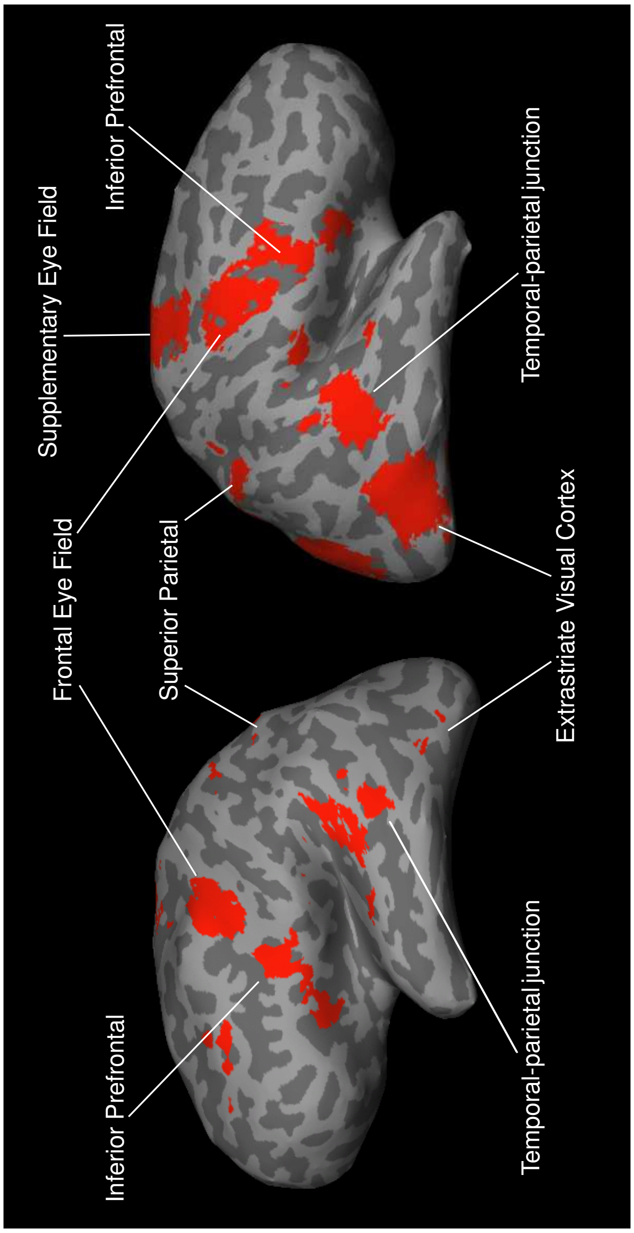

Figure 2. Cortical surface renderings with superimposed active region determined so that the proportion of false clusters is less than 10% with probability at least 90%, where the tolerance for identifying a false cluster is 10% of its area in the null set. Data were smoothed with a 10mm FWHM Gaussian filter.