Multiple testing via successive subdivision

Werner Ehm[a] , *Jürgen Kornmeier*[a,b] , *Sven Heinrich*[b]

Abstract

A sequential multiple testing procedure recently introduced by Heinrich, Bach and Kornmeier allows to "zoom in" on, and thus identify regions with highly significant departures from null-hypotheses. The purpose of this note is to state a cognate of this procedure in general form and to prove that it controls the familywise error. Two possible applications are briefly indicated.

1 Introduction

Often in statistical applications heavy multiple testing is carried out leaving two major questions:
Q1: *Where* are significant departures from null-hypotheses?
Q2: What can be said about the overall error probability of the testing procedure?

In regard to Q2, the classical approach is to control the *familywise error,* i.e., to require that the probability of any false rejection is $\leq \alpha$, for some α fixed in advance. Such may be achieved using the Bonferroni inequality or, e.g., closed or sequential testing procedures (Marcus et al. 1976, Holm 1979). Particularly when the number of tested hypotheses is large, the desire to avoid any error of the first kind has to be paid by a low test power. Therefore, as an alternative it has been suggested to control instead the *false discovery rate* [FDR], i.e., to bound the expected proportion of false rejections among all rejections (Benjamini & Hochberg 1995). While test power generally is improved with this approach, it does not allow to pin down those tests for which the hypothesis can be safely rejected. Thus when using FDR control one only gets a vague answer to Q1.

There are cases, however, where some tests have very small p-values, suggesting a massive violation of the null-hypothesis. Naturally then, one would like to be able to reject precisely those null-hypotheses with guaranteed confidence. A sequential multiple testing procedure designed for such cases has recently been proposed by Heinrich, Bach & Kornmeier (2008) under the name "Conquer and Divide" [CaD]. CaD proceeds by successively subdividing the "search space" and continues testing along each "search path" until first acceptance of a null-hypothesis, thereby taking advantage of instances where some of the individual tests' p-values are very small.

The purpose of this short note is to develop a general, modified version of CaD (also called "CaD") and to prove that it controls the familywise error. This material appears in Section 2. Section 3 sketches two possible applications. An elaboration of this note is in progress.

2 The testing procedure

Consider a rooted tree with vertex set $\mathcal{V}$. For definiteness, the tree is supposed to be "hanging downward," with the *root* $v_0 \in \mathcal{V}$ on top. Each vertex v splits into its (immediate) *descendants,* imagined as lying one layer below v. Let $d(v) \subset \mathcal{V}$ denote the set of descendants of v, the number of which may differ across vertices. The splitting stops at

the L-th step ($L \geq 1$), such that the vertices of $\mathcal{V}$ come in L layers below the (0-th) root layer. In particular, the tree has *depth* L and is *complete* in the sense that all branches end at the bottom layer.

With each vertex (a "location in search space") is associated a testing problem: at every $v \in \mathcal{V}$ a test of a certain null-hypothesis $\mathcal{H}_0(v)$ is carried out whose probability of rejection under $\mathcal{H}_0(v)$ is $\leq \alpha(v)$. Let us write $\alpha = \alpha(v_0)$ for the test level at the root v_0. The test levels are assumed to satisfy the following *local Bonferroni* condition.

(LB) For every vertex $v \in \mathcal{V}$ above the L-th layer one has $\sum_{v' \in d(v)} \alpha(v') \leq \alpha(v)$.

The proposed multiple testing procedure by successive subdivision may now be described as follows.

[CaD] *Starting at the root v_0, keep testing downward each branch of the tree ("search path") as long as the respective null-hypothesis is rejected: stop testing upon first acceptance of a null-hypothesis, and reject all null-hypotheses that have been rejected thus far.*

We will show that the testing procedure is valid, in the sense that its familywise error does not exceed α. The familywise error, or probability of an error of the first kind of the *procedure* CaD, equals the probability π_1 that among the hypotheses rejected by CaD there is at least one true (hence falsely rejected) hypothesis.

Proposition 2.1 *Under condition* (LB) *one has* $\pi_1 \leq \alpha$.

Proof. Let P denote the probability measure underlying the observations. Given P, the hypothesis $\mathcal{H}_0(v)$ (about P) at vertex v is either true or false, independently of the experimental outcome. Thus given P, we get a valued tree by assigning vertex v the truth value $t(v) = 0$ if $\mathcal{H}_0(v)$ is false, and $t(v) = 1$ otherwise. For any vertex v let $U(v)$ denote the set of all vertices $v' \in \mathcal{V}$ that lie on the (unique) path leading from v up to v_0, except for v itself which is excluded. Let the set F consist of all vertices at which the null-hypothesis is true for the first time, 'first' in top-down direction. That is, F comprises all vertices $v \in \mathcal{V}$ with the following two properties: (i) $t(v') = 0$ for every $v' \in U(v)$; (ii) $t(v) = 1$. ($F = \{v_0\}$ if $t(v_0) = 1$.)

The significance of the set F is the following: (*) if (the application of) CaD happens to produce any error of the first kind (hereafter: "type I error"), then it also produces a type I error at some vertex $v \in F$. For suppose that CaD produces a type I error at vertex $v^* \in \mathcal{V}$, say. If $v^* \in F$, we are done. If $v^* \notin F$, then since $t(v^*) = 1$, there exists a first vertex v on the path from v_0 down to v^* with $t(v) = 1$, that is, there exists $v \in U(v^*) \cap F$. Moreover, the test at v rejects $\mathcal{H}_0(v)$ because otherwise the procedure would have stopped at v, leaving no occasion for a type I error to occur at v^*. Consequently, a type I error occurs at $v \in F$, and (*) is proven. But (*) implies

$$\begin{aligned} \pi_1 &= P\,[\mathcal{H}_0(v) \text{ is rejected for at least one } v \in F\,] \qquad (1) \\ &\leq \sum\nolimits_{v \in F} P\,[\mathcal{H}_0(v) \text{ is rejected}\,] \\ &\leq \sum\nolimits_{v \in F} \alpha(v), \end{aligned}$$

whence it suffices to show that

$$\sum\nolimits_{v \in F} \alpha(v) \leq \alpha. \qquad (2)$$

For any complete subtree $\mathcal{U}$ of $\mathcal{V}$ let $\rho_{\mathcal{U}}$ denote its root vertex. Then (2) is a consequence of the following more general claim:

$$\text{For every complete subtree } \mathcal{U} \text{ of } \mathcal{V}, \quad S_{\mathcal{U}} := \sum_{v \in F \cap \mathcal{U}} \alpha(v) \leq \alpha(\rho_{\mathcal{U}}). \tag{3}$$

We argue by induction on the depth ℓ of $\mathcal{U}$ $(0 \leq \ell \leq L)$. The case $\ell = 0$ is trivial (since $\mathcal{U}$ then consists of its root only), so let $1 \leq \ell\,(\leq L)$ and suppose that (3) holds for every complete subtree of depth $\ell - 1$. Let $\mathcal{U}$ be a complete subtree of depth ℓ. If $F \cap \mathcal{U}$ is empty or equals $\{\rho_{\mathcal{U}}\}$, there is nothing to prove. Otherwise let us decompose $\mathcal{U}$: each descendant v of $\rho_{\mathcal{U}}$ represents the root of a complete subtree $\mathcal{U}(v)$ of $\mathcal{U}$ of depth $\ell - 1$. Since the vertex sets of all these subtrees are pairwise disjoint, and $\rho_{\mathcal{U}} \notin F$ if $F \cap \mathcal{U} \neq \{\rho_{\mathcal{U}}\}$, the induction hypothesis and condition (LB) imply

$$S_{\mathcal{U}} = \sum\nolimits_{v \in d(\rho_{\mathcal{U}})} S_{\mathcal{U}(v)} \leq \sum\nolimits_{v \in d(\rho_{\mathcal{U}})} \alpha(v) \leq \alpha(\rho_{\mathcal{U}}).$$

Thus (3) holds for any complete subtree of depth ℓ, and the inductive proof is complete.

Remarks. The result immediately generalizes to the case where one has a collection of rooted trees, not necessarily with identical depths, provided the levels of the tests at the roots are controlled by Bonferroni. Note that the significance levels of the tests are moderate initially, and become restrictive only downward the tree. This is in contrast with other sequential procedures, e.g. Holm's (1979), where the most restrictive tests are carried out first. Note also that no assumption is required about the joint distribution of the test statistics. Finally, control of the familywise error implies that other common error criteria are controlled as well. In fact, domination by the familywise error is guaranteed for any criterion representable as the expected value of a (generally unobservable) random variable with values in $[0, 1]$ that assumes the value 0 whenever there is no false rejection. Examples include the false discovery rate and the per comparison error rate (Benjamini & Hochberg, 1995, p. 291).

A further generalization of the CaD procedure deals with the case where a vertex v may, itself, represent a "local" multiple testing problem along with an associated testing procedure, $\mathcal{M}(v)$, rather than just the test of a single hypothesis, $\mathcal{H}_0(v)$. The quantity $\alpha(v)$ then has to be interpreted as the familywise error of that testing procedure.

For example, $\mathcal{M}(v)$ may stand for the situation where $m = |d(v)|$ null-hypotheses $\mathcal{H}_0(v')$, $v' \in d(v)$ are tested using Holm's sequential testing procedure at the level $\alpha(v)$ (familywise). At the next layer, $\mathcal{M}(v)$ splits into m descendants $\mathcal{M}(v')$, $v' \in d(v)$, where $\mathcal{M}(v')$ corresponds to a subdivision of the single hypothesis $\mathcal{H}_0(v')$ into a number of further null-hypotheses which, again, are tested using Holm's procedure. Any multiple testing procedure other than Holm's that controls the familywise error can be applied as well. The CaD procedure stops at vertex v if the local procedure associated with $\mathcal{M}(v)$ accepts *at least one* of the single hypotheses $\mathcal{H}_0(v')$. Otherwise it continues at *all* descendants $\mathcal{M}(v')$, $v' \in d(v)$. The familywise error π_1 of the extended CaD procedure is defined as the probability that any of the local testing procedures $\mathcal{M}(v)$, $v \in \mathcal{V}$ produces a false rejection, which equals the probability that any of the single null-hypotheses $\mathcal{H}_0(v')$ is falsely rejected.

Corollary 2.2 *Under condition* (LB) *the extended* CaD *procedure described above satisfies* $\pi_1 \leq \alpha$.

Proof. It suffices to assign truth values as follows: $t(v) = 1$ if any of the single hypotheses $\mathcal{H}_0(v')$, $v' \in d(v)$ is true, and $t(v) = 0$ otherwise. The correspondingly defined set F then retains its original meaning: one readily verifies that if the extended CaD procedure produces a false rejection in the local testing problem $\mathcal{M}(v^*)$, then there is a vertex $v \in F \cap U(v^*)$ such that the procedure produces a false rejection in the local testing problem $\mathcal{M}(v)$. The remainder of the proof is analogous to that of the proposition.

The definition of the extended CaD procedure is chosen such that the original proof carries over easily. Other variants may also be of interest.

3 Two possible applications

Analysis of EEG data. This is the area CaD was developed for by Heinrich et al. (2008). In electroencephalographic studies [EEG] one often wants to know where in a time series $\{x(t),\, t \in T\}$ "something conspicuous" is happening, that is, locate one (or several) time region(s) $C_j \subset T$ showing distinct deviations from the behaviour to be expected under some null-hypothesis $\mathcal{H}_0$. E.g., $\mathcal{H}_0$ may mean "no systematic departure from zero", $E\,x(t) = 0$ for $t \in T$. With CaD, conspicuous regions are searched for by successively subdividing T into smaller intervals C_j down to a certain level, and testing $\mathcal{H}_0$ restricted to C_j along each subdivision path until first acceptance. Simulations carried out by Heinrich et al. (2008) suggested that CaD is conservative in the sense of Section 2, and revealed satisfactory power properties.

Thresholding of wavelet coefficients. Nonparametric curve estimation based on thresholding of wavelet coefficients was introduced by Donoho & Johnstone (1994). As emphasized by Abramovich & Benjamini (1995), thresholding may be regarded as a multiple testing problem, where an estimated wavelet coefficient $\widehat{w}_{j,k}$ is kept or set to zero, respectively, in accordance with the outcome of a test of the null-hypothesis that the "true" coefficient $w_{j,k} = 0$. In this context, the above testing procedure could be applied as follows. For $n = 2^{J+1}$ observations, the wavelet coefficients are grouped into resolution levels $j = 1, \ldots, J$ each comprising 2^j coefficients $w_{j,k}$, $k = 1, \ldots, 2^j$. They can thus be arranged as a binary tree in which coefficient $w_{j,k}$ "splits" into $w_{j+1,2k-1}$ and $w_{j+1,2k}$. This splitting corresponds to a halving of time intervals, as is most obvious for the Haar wavelet system. The CaD procedure applied with the tests of the hypotheses "$w_{j,k} = 0$" may then be regarded as a method of selecting thresholds for the estimated coefficients $\widehat{w}_{j,k}$. It differs from related proposals in the literature (e.g., Donoho & Johnstone (1994), Abramovich & Benjamini (1995), or, for a different setting, Donoho & Jin (2008)) in that the threshold is not the same for all coefficients (no matter how adaptive that global value is chosen), but increases with the resolution level. Useful implementations may require modifications of the tests at low resolution levels, in order to avoid too early stopping due to a possible "averaging out" of wavelet coefficients across longer intervals. The performance of the procedure can be studied along the lines of Abramovich & Benjamini's (1995) article.

References

Abramovich, F. & Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypothesis testing procedure. In *Wavelets and Statistics*, Ed. A. Antoniadis and G. Oppenheim, Lect. Notes Statist. Vol. 103, pp. 5-14. New York: Springer-Verlag.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289-300.

Donoho, D. & Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. (USA)* **105**, 14790-14795.

Donoho, D. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Heinrich, S. P., Bach, M. & Kornmeier, J. (2008). Conquer and Divide: A novel approach to spatiotemporal significance testing that accounts for alpha error inflation. *Neuroimage* **41** Suppl. 1, p. S159.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65-70.

Marcus, R., Peritz, E. & Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655-660.

Author addresses:

[a] Institute for Frontier Areas of Psychology and Mental Health
Wilhelmstr. 3a, 79098 Freiburg, Germany

[b] Department of Ophtalmology, University of Freiburg
Killianstr. 5, 79106 Freiburg, Germany

E-mail addresses:

ehm@igpp.de
kornmeier@igpp.de
sven.heinrich@uniklinik-freiburg.de