Electronic Journal of Statistics
Vol. 4 (2010) 461–471
ISSN: 1935-7524
DOI: 10.1214/09-EJS496

Multiple testing along a tree

Werner Ehm

Institute for Frontier Areas of Psychology and Mental Health, 79098 Freiburg, Germany
e-mail: ehm@igpp.de

Jürgen Kornmeier

Institute for Frontier Areas of Psychology and Mental Health, 79098 Freiburg, Germany and University Eye Clinic, Section Functional Vision Research, 79106 Freiburg, Germany
e-mail: juergen.kornmeier@uni-freiburg.de

Sven P. Heinrich

University Eye Clinic, Section Functional Vision Research, 79106 Freiburg, Germany
e-mail: sven.heinrich@uniklinik-freiburg.de

Abstract: Suitable sequentially rejective multiple test procedures allow to "zoom in" on clusters of relevant variables in high-dimensional regression (Meinshausen [7]), or on regions of interest in some search space (Heinrich et al. [3]; Meinshausen et al. [8]). As a common framework for these schemes we propose to consider multiple testing along a tree of hypotheses together with a "keep rejecting until first acceptance" rule. Particular topics addressed in this note are control of the familywise error, and some variants and basic properties of the procedure.

AMS 2000 subject classifications: 62G10, 62J15, 62L99.
Keywords and phrases: Conquer and Divide, event-related potentials, familywise error, (nested) multiple testing, tree of hypotheses.

Received October 2009.

Contents

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 test 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 test procedures and variants thereof (Marcus et al. [6], Holm [4], Goeman & Mansmann [2]). 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 [1]). 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 a few 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 corresponding multiple test procedure has recently been introduced by Meinshausen [7], in the context of testing for variable importance in a high-dimensional linear regression setting. Clusters of highly correlated regressors are tested for their joint influence on the dependent variable, and get subdivided until the test result indicates irrelevance. Meinshausen showed that the familywise error of the test procedure can be controlled by suitably adjusting the p-values of the single tests; moreover, that considerable gains in power are attainable with this scheme compared to other multiple test procedures.

Virtually the same procedure, though for a different purpose, has been proposed by Heinrich et al. [3] under the name Conquer and Divide [CaD]. It was devised for detecting regions of potential interest (in regard to time, frequency, and space [viz. electrode]) in neurophysiological recordings such as the EEG. CaD proceeds by successively subdividing the se arch 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. Another closely related scheme is the "blind search" algorithm of Meinshausen et al. [8]: regions of interest are successively narrowed down to the "needle(s) in the haystack," i.e., to a few instances of massively violated hypotheses amongst a huge number of true (no effect) null-hypotheses. The focus in that paper is on computational cost and its reduction by means of an optimal search strategy. All these procedures work from coarse to fine resolution levels within the search space. Treelets, introduced by Lee et al. [5] as a fully adaptive alternative to principal components analysis, follow the converse path. The single variables are grouped into clusters represented by newly constructed variables that together with their residuals form a sparse orthonormal basis via a kind of multi-resolution scheme.

A common feature of all these procedures is that they operate on an underlying tree structure. The purpose of this note is to present the scheme underlying Meinshausen's [7] and Heinrich et al.'s [3] procedures in the general setting of

multiple testing along a tree of hypotheses. We prove control of the familywise error of the procedure under a suitable local Bonferroni condition (Section 2), discuss a few of its extensions and basic properties (Section 3), present an illustrative example (Section 4), and end with some conclusions (Section 5).

2. Testing along a tree of hypotheses

Consider a rooted tree with vertex set $\mathcal{V}$. For definiteness, the tree is supposed to be hanging upside-down, with the *root* $v_0 \in \mathcal{V}$ on top. Thus following a branch "downward/upward the tree" ultimately leads to its leaves/its root, respectively. Each vertex v gives rise to its *children* vertices, imagined as lying one *layer* below v. Let $c(v) \subset \mathcal{V}$ denote the children of v, the number (≥ 1) of which may differ across vertices. Ramification 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 supposed to be *complete* in the sense that all branches end at the bottom layer. (The completeness assumption is made only for convenience of presentation and can be dropped; cf. Remark 2.2 below.)

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 $\leq \alpha(v)$. Any such multiple testing problem will be called a *tree testing problem.* 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 $\sum_{v' \in c(v)} \alpha(v') \leq \alpha(v)$.

The proposed multiple test procedure by successive subdivision may now be described as follows. As in Heinrich et al. [3] we dub it Conquer and Divide.

[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 so far.*

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. We will show that the familywise error of the CaD procedure does not exceed α. A closely related result pertaining to hierarchical variable selection is due to Meinshausen [7, Theorem 1].

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

Remark 2.2. The theorem immediately extends to the case where one has a collection of rooted trees, not necessarily with identical depths, provided the test levels at the respective roots are controlled by Bonferroni. Incomplete trees with branches ending in a leaf strictly above the bottom layer can be completed trivially: any branch ending in a leaf v at a layer $\ell < L$ is prolonged to a branch

ending at layer L by letting the local testing problem at any newly added vertex be an identical copy (regarding hypothesis, test procedure, and outcome) of the testing problem at vertex v.

Remark 2.3. The familywise error controls other common error criteria, too. Thus domination by the familywise error is guaranteed for any criterion representable as the expected value of an (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 [1, p. 291]). Note that no assumption is required here about the joint distribution of the test statistics.

Proof of Theorem 2.1. Let P denote the probability measure underlying the observations. Given P, the hypothesis $\mathcal{H}_0(v)$ at vertex v is either true or false. 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

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

whence it suffices to show that

$$\sum\nolimits_{v\in F} \alpha(v) \le \alpha. \tag{2}$$

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:

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

We argue by induction on the depth ℓ of $\mathcal{U}$ ($0 \le \ell \le L$). The case $\ell = 0$ is trivial (since $\mathcal{U}$ then consists of its root only), so let $1 \le \ell\,(\le L)$ 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

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

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

3. Extensions

3.1. Strict testing problems

Thus far null-hypotheses could be true or false without any restriction. Often, however, restrictions result from logical dependencies between hypotheses (Shaffer [9]). In such cases the conditions imposed on the individual test levels can be relaxed, and gains in test power achieved. This possibility was elaborated by Meinshausen [7] under the heading Shaffer improvement. Of particular interest here is *strict* logical dependency.

Definition 3.1. *A tree testing problem is called* strict *if for every false hypothesis $\mathcal{H}_0(v)$ above the bottom layer, at least one of the hypotheses $\mathcal{H}_0(v')$, $v' \in c(v)$ is false, too.*

Corollary 3.2. *Consider a strict tree testing problem such that for any parent-child pair v, $v' \in \mathcal{V}$ with v' a* leaf *of $\mathcal{V}$ (i.e. v' is a vertex at the bottom layer L, and $v' \in c(v)$) one has*

$$\alpha(v') \leq \tfrac{\alpha(v)}{|c(v)|-1} \quad \textit{(if } |c(v)| \geq 2\,;\ \alpha(v') = 1 \textit{ otherwise),}$$

whereas for all other vertices (i.e., children vertices at layers $< L$) condition (LB) *is satisfied. (This modification of condition* (LB) *is referred to as condition* (LBs)*.) Then $\pi_1 \leq \alpha$.*

Proof. The proof of Theorem 2.1 applies up to the following modification. Consider any vertex v' at the bottom layer, with parent vertex v. If $v' \notin F$ then it does not contribute to π_1, so suppose $v' \in F$. By the definition of F and strictness one has $t(v') = 1$, $t(v) = 0$, and $t(v'') = 0$ for at least one $v'' \in c(v)$, so that ($|c(v)| \geq 2$ and) at most $|c(v)| - 1$ vertices in $c(v)$ belong to F, i.e., can contribute to π_1. Thus by (LBs), the probability of a false rejection at any $v' \in c(v)$ is controlled by $\alpha(v)$, and the proof is complete.

Remark 3.3. The strictness condition generally is satisfied in the tree testing problems envisaged here. Suppose that $x(t)$, $t \in T$ is a stochastic process with mean $m(t) = Ex(t)$ and we wish to learn where $m(t)$ deviates (substantially) from zero. With CaD, such regions may be searched for by successively bisecting T into smaller intervals I down to a certain level, and test-

ing $\mathcal{H}_0(I) : m(t) = 0$ $(t \in I)$ along each subdivision branch until first acceptance. Evidently, the corresponding tree testing problem is strict. It remains so also if the null-hypothesis at vertex I is less rigid than above, say $\mathcal{H}_0(I) : \int_I m(t)\,dt = 0$. However, in the latter case a hypothesis can be true with both its children hypotheses being false.

Remark 3.4. (Optimality) In the common case of a binary tree $\mathcal{V}$ and uniform distribution of errors across the two children, condition (LB) means that test levels have to be halved from layer to layer (downwards $\mathcal{V}$), so that $\alpha(v) = \alpha 2^{-\ell}$ for any vertex v at the ℓ-th layer. Under condition (LBs) the last halving step can be omitted, and one may choose $\alpha(v) = \alpha 2^{-(L-1)}$ for every vertex v at the bottom layer (Meinshausen [7]). Can one further relax the conditions imposed on the test levels $\alpha(v)$? The following argument shows that at least for small α *sizable improvemens beyond condition (LBs) are not possible in general.*

Let us assume that the individual test levels are identical within each layer, and let α_ℓ denote their common value at the ℓ-th layer. It will be shown that, essentially, the α_ℓ cannot be chosen larger than indicated above. We consider two cases. First, suppose there is a branch $v_0, v_1, \ldots, v_L$ reaching from the root to the bottom layer of $\mathcal{V}$ such that all hypotheses $\mathcal{H}_0(v_\ell),\ 0 \le \ell \le L$ are false, while for the respective sibling v'_ℓ of v_ℓ (sharing the same parent) the hypothesis $\mathcal{H}_0(v'_\ell)$ is true $(1 \le \ell \le L)$. Suppose further that every $\mathcal{H}_0(v_\ell)$, $0 \le \ell < L$ is rejected with probability close to 1, say equal to 1. Then every sibling hypothesis $\mathcal{H}_0(v'_\ell)$, $1 \le \ell \le L$ gets tested, and the probability of any error occurring thereby can be estimated by, and assumed to be not much less than[1] $\sum_{\ell=1}^{L} \alpha_\ell =: \epsilon_1$. Thus, $\pi_1 \gtrsim \epsilon_1$. Secondly, fix a layer $\ell \ge 1$ and suppose that for each pair of siblings at this layer one of the null-hypotheses is false and the other one is true; furthermore, that all hypotheses at layers strictly above ℓ are false; and finally, that these false hypotheses are rejected with high probability, say 1. Then similarly as in the first case, the probability of any false rejection at layer ℓ can be assumed to be approximable by $2^{\ell-1}\alpha_\ell$. This holds for every ℓ, so $\pi_1 \gtrsim \max_{1\le\ell\le L} 2^{\ell-1}\alpha_\ell =: \epsilon_2$. Consequently, $\epsilon_1,\ \epsilon_2$ have to be $\le \alpha$ if the procedure is to control the familywise error.

Naturally, one would like to maximize the smallest test level, which under the constraint $\epsilon_2 \le \alpha$ amounts to setting $\alpha_L = \alpha 2^{1-L}$. If one then puts $\alpha_\ell = \alpha 2^{-\ell}$ for $0 \le \ell < L$ as in condition (LBs), then the approximate lower bound ϵ_1 to π_1 already attains the maximally allowed value α,

$$\epsilon_1 = \sum\nolimits_{\ell=1}^{L} \alpha_\ell = \alpha \left(\sum\nolimits_{\ell=1}^{L-1} 2^{-\ell} + 2^{1-L} \right) = \alpha,$$

showing that no room is left for substantial improvements beyond condition (LBs).

Similar considerations apply to q-nary trees where each vertex has q children $(q > 2)$. In this case condition (LBs) amounts to setting $\alpha(v) = \alpha q^{-\ell}$ or $\alpha(v) = \alpha q^{1-L}/(q-1)$, respectively, according as vertex v lies above or at the bottom

[1] Such is the case if the test statistics are independent and the test levels are small.

(L-th) layer. Arguments entirely analogous to the case $q = 2$ show that, again, sizable improvements beyond this choice are not possible.

3.2. Nested multiple testing

With CaD, the probability of any false rejection at the children $v' \in c(v)$ of a vertex $v \in \mathcal{V}$ is controlled by (local) Bonferroni type estimates. A natural idea is, then, to apply instead a more sophisticated multiple testing procedure to the children hypotheses $\mathcal{H}_0(v')$, $v' \in c(v)$. Any tree testing problem in the sense of Section 2 may likewise be conceived as a *nested* testing problem: with any vertex v (strictly) above the bottom layer L one associates the local problem $\mathcal{M}(v)$ consisting of the hypotheses $\mathcal{H}_0(v')$, $v' \in c(v)$[2] along with an associated multiple test procedure. The quantity $\alpha(v)$ then has to be interpreted as the familywise error of that local test procedure.

For definiteness, suppose the $m = |c(v)|$ children hypotheses $\mathcal{H}_0(v')$, $v' \in c(v)$ are tested at the level $\alpha(v)$ using Holm's [4] sequential test procedure. (Any multiple test procedure other than Holm's that controls the familywise error locally at every $\mathcal{M}(v)$ could be applied as well.) At the next layer, $\mathcal{M}(v)$ splits into m children problems $\mathcal{M}(v')$, $v' \in c(v)$, where each $\mathcal{M}(v')$ corresponds to a subdivision of the single hypothesis $\mathcal{H}_0(v')$ into $m' = |c(v')|$ further null-hypotheses which, again, are tested using Holm's procedure. The *nested CaD* procedure [nCaD] stops at vertex v if the local procedure associated with $\mathcal{M}(v)$ accepts *at least one* of the single hypotheses $\mathcal{H}_0(v')$, $v' \in c(v)$. Otherwise it continues at *all* children problems $\mathcal{M}(v')$, $v' \in c(v)$. The familywise error π_1 of nCaD is defined as the probability that any of the local test procedures $\mathcal{M}(v)$, $v \in \mathcal{V}$ produces a false rejection, which equals the probability that any of the single null-hypotheses $\mathcal{H}_0(v')$, $v' \in c(v)$, $v \in \mathcal{V}$ is falsely rejected.

Although above nCaD was developed starting out from a tree testing problem in the sense of Section 2, such an ascription is not necessary. The single hypotheses making up the local multiple testing problem $\mathcal{M}(v_1)$ generally need not have any relationship to those making up $\mathcal{M}(v_2)$ ($v_1 \neq v_2$).

Corollary 3.5. *Under condition* (LB) nCaD *satisfies* $\pi_1 \leq \alpha$.

Proof. It suffices to assign truth values as follows: $t(v) = 1$ if any of the single hypotheses $\mathcal{H}_0(v')$, $v' \in c(v)$ is true, and $t(v) = 0$ otherwise. The correspondingly defined set F then retains its original meaning: one readily verifies that if nCaD 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 Theorem 2.1.

Towards a comparison of CaD and nCaD in the case of a binary tree testing problem, suppose that nCaD is implemented using Holm's procedure. Let $\widehat{p}(v)$ denote the p-value of the test of hypothesis $\mathcal{H}_0(v)$. Then nCaD rejects none or at

[2]Note that even though the single hypotheses making up $\mathcal{M}(v)$ are located at layer $\ell + 1$ if v belongs to layer ℓ, $\mathcal{M}(v)$ is considered as being located at layer ℓ in the nested version.

least one of the two children hypotheses of vertex v according as $\min_{v' \in c(v)} \widehat{p}(v')$ is $> \alpha(v)/2$ or $\leq \alpha(v)/2$, respectively, and in the latter case it rejects both if in addition $\max_{v' \in c(v)} \widehat{p}(v') \leq \alpha(v)$. For comparison, CaD rejects hypothesis $\mathcal{H}_0(v')$ if $\widehat{p}(v') \leq \alpha(v)/2$ (separately for both $v' \in c(v)$). Therefore, individual test levels are more restrictive with CaD. This does not imply that CaD is less powerful, however, because in return nCaD has a more restrictive stopping rule. In fact, if nCaD stops at vertex v in some branch of $\mathcal{V}$, then CaD (i) does not stop earlier than at the children $v' \in c(v)$ (i.e., not earlier than nCaD; see footnote 2); (ii) may continue testing further downward the tree, namely if $\min_{v' \in c(v)} \widehat{p}(v') \leq \alpha(v)/2$—whereas nCaD will stop at v if $\max_{v' \in c(v)} \widehat{p}(v') > \alpha(v)$. Thus in general none of the two procedures dominates the other one.

4. Application

For illustration let us consider a signal plus noise model

$$x(t_i) = f(t_i) + \epsilon(t_i), \quad t_i = (i - 1/2)/n,\ 1 \leq i \leq n \tag{4}$$

with i.i.d. Gaussian noise rv's having mean 0 and (known) variance σ^2. One wants to know whether $f(t_i) = Ex(t_i) \neq 0$ for some t_i, and if so, at which t_i. When (n)CaD is applied with a bisection strategy the vertices of the search tree $\mathcal{V}$ correspond to dyadic intervals. Starting at the root $v_0 = [0, 1)$, continued bisection yields 2^ℓ vertices of the form $v = [k2^{-\ell}, (k + 1)2^{-\ell})$, $0 \leq k < 2^\ell$ at the ℓ-th layer. The testing problem at vertex v involves the null-hypothesis $\mathcal{H}_0(v) : f(t_i) = 0\ \forall\, t_i \in v$ along with some test that rejects $\mathcal{H}_0(v)$ if, e.g.,

$$T_1(v) = \#\{t_i \in v\}^{-1} \left| \sum\nolimits_{t_i \in v} x(t_i)/\sigma \right| \qquad \text{or} \qquad T_2(v) = \sum\nolimits_{t_i \in v} (x(t_i)/\sigma)^2$$

is too large. Since these test statistics have known distributions under $\mathcal{H}_0(v)$, all elements required for running the (n)CaD procedure are specified. Simulation results are presented below.

For the electroencephalographic (EEG) applications considered in Heinrich et al. [3], x may be thought of as representing the difference of two event-related potentials (ERP) recorded at time points t_i under different experimental conditions.[3] In this case the i.i.d. assumption on the noise variables is inappropriate, and the above tests are not applicable. Feasible valid tests can easily be obtained, however, using randomized assignment of the single trials to the two experimental conditions. Apart from that the model (4) captures the basic situation encountered when looking for ERP-components (significant deflections of the difference-ERP from the null-line) in EEG studies.

We demonstrate the performance of CaD and nCaD for the case indicated above, with a signal f mimicking a typical ERP shape, noise level $\sigma = 0.7$, $n =$

[3] An ERP is an average of short segments, called "trials," of a continuously recorded EEG trace that are temporally aligned to certain periodically recurring events such as the onset of a stimulus.

2^{10}, and $\alpha = .05$; bisection continues to layer $L = 5$. The signal f is a linear combination of shift-scale versions $g_{a,b}(t) = g((t-b)/a)$ $(a > 0, b \in \mathbf{R})$ of the smoothed indicator function

$$g(t) = 1\ (|t| \le 1/2), \quad = 0\ (|t| \ge 3/2), \quad = \cos^2((|t|-1/2)\pi/2)\ (1/2 < |t| < 3/2).$$

The blue and red lines in Fig. 1 represent *rejection profiles* for CaD and nCaD, respectively, which are obtained as follows. For every t_i the ordinate of the profile is computed as the relative frequency, among 10^4 Monte Carlo simulations of (4), of the event that hypothesis $\mathcal{H}_0(v)$ is rejected for each $v \in \mathcal{V}$ containing t_i. A rejection profile thus is a step function constant on the dyadic intervals at the finest resolution level.

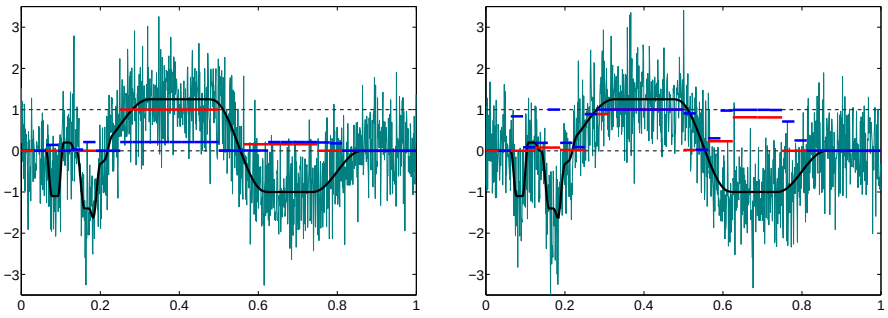

FIG 1. *Rejection profiles (see text) of CaD (blue) and nCaD (red) procedures, for test statistics T_1 (left) and T_2 (right). Black: signal f; greenish: sample noisy versions of f.*

The example shows that at least for the test statistics of type T_1 neither CaD nor nCaD completely dominates the other procedure in regard to power as measured by the rejection profile. When varying the parameters defining f we often found that CaD performed better than nCaD particularly where f oscillates heavily or deviates strongly from zero on small temporal regions, as for $.05 < t < .25$ in our example. By contrast, nCaD can be more sensitive against consistent, less localized deviations; compare the two rejection profiles for $\frac{1}{4} < t < \frac{1}{2}$ in the left-hand plot. The low performance of CaD in this case is due to the fact that the null-hypothesis at the root $v_0 = [0,1)$ is only sligthly violated because the normalized mean deviation, $\sqrt{n}/\sigma$ times the average of all $f(t_i)$s, has a rather moderate value of about 1.2. Therefore the test at the root v_0 accepts $\mathcal{H}_0(v_0)$ fairly often and the CaD procedure (applied with the T_1-type tests) stops already at v_0. In contrast, nCaD starts testing the two hypotheses $\mathcal{H}_0(v')$ (conceived as a multiple test problem) corresponding to the intervals $v' = [0, 1/2)$ and $v' = [1/2, 1)$, respectively, which are both strongly violated, hence let nCaD continue with high probability. Afterwards, however, nCaD tends to stop earlier than CaD, namely when there is no sizeable deflection in one of the two subdivision intervals.

Concerning familywise errors the procedures were strongly conservative, for both test types. With an overall level of $\alpha = .05$ one would expect that an error of the first kind (familywise) occurs in about 500 among the 10^4 simulated trials. However, with CaD such an error occurred only in 9 (T_1), respectively 82 (T_2) trials; with nCaD these figures were 160 (T_1) and 174 (T_2), respectively.

Subdivision into disjoint intervals can be disadvantageous if deviations of interest happen to occur in small vicinities of the subdivision points. It may then be meaningful to allow for some overlap, e.g., divide $[0, 1)$ into $[0, 1/2 + \delta)$ and $[1/2 - \delta, 1)$ for some $\delta \in (0, 1/2)$, and similarly at finer resolution levels. Note that there is no need for subdividing the search space into disjoint portions since the testing problems at the vertices of the tree may be entirely arbitrary in general. When analyzing images in 2 (or 3) dimensions rather than time courses, successive triangulation or subdivision into 4 (or 8) subregions, perhaps with overlap, may be appropriate, yielding more strongly ramified trees.

5. Conclusions

Multiple testing problems for hypotheses that can be arranged in a tree structure have important applications in statistics. As a particular feature of the procedures studied here, the significance levels of the tests are moderate initially, and become restrictive only downward the tree (toward finer resolution levels). This increases the chance of rejecting hyotheses at the upper layers of the tree (coarser levels), and thus to get some rough information where to look more closely. By contrast, with bottom-up sequential procedures such as Holm's [4], the most restrictive test is carried out first, implying an increased risk that the procedure stops immediately and yields no information at all.

Much of the present work is already contained in Meinshausen [7]. In particular, Meinshausen proposed a multiple test procedure equivalent to CaD (Conquer and Divide; Heinrich et al. [3]) and proved that it controls the familywise error; furthermore, that it is amenable to Shaffer improvement (cf. Section 3.1). However, this material was stated in a somewhat special context, hierarchical testing for variable importance. Our main purpose here was to present the basic problem and procedure in the general, context-free form of multiple testing along a tree of hypotheses, which covers a wide range of search and selection problems including those considered by Meinshausen [7] and Heinrich et al. [3].

Some further extensions and results, mostly straightforward but useful, include the following. First, test levels at the children vertices in the local Bonferroni conditions need not be balanced. This allows to allocate test power selectively, for instance when regions of interest are known. Secondly, arbitrary trees are covered: the number of children may vary from vertex to vertex, and branches need not all have identical lenghts. Third, the Shaffer-improved CaD procedure has a certain (approximate) optimality property: subject to familywise error control the individual test levels cannot in general be relaxed to a sizable degree (Remark 3.4). Finally, nested multiple testing along with a generalized CaD procedure, nCaD, was introduced, where the individual test

problem at a vertex v may be a multiple test problem itself. Depending on the type of tests used nCaD sometimes can dominate CaD on broad regions in search space showing consistent deviations from the null-hypothesis (Figure 1, left). Our simulation results and heuristic considerations indicate, however, that CaD may often be superior to nCaD, particularly for hypotheses corresponding to sparse, narrowly localized deflections.

Acknowledgements

Support by the Deutsche Forschungsgemeinschaft (grants HE 3504/2, HE 3504/4, BA 877/16, BA 877/18) is gratefully acknowledged.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300. MR1325392

[2] Goeman, J.J. and Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* **24**, 537–544.

[3] Heinrich, S.P., Bach, M. and 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.

[4] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70. MR0538597

[5] Lee, A.B., Nadler, B. and Wasserman, L. (2009). Treelets—An adaptive multi-scale basis for sparse unordered data. *Ann. Appl. Statist.* **2**, 435–471. MR2524336

[6] Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660. MR0468056

[7] Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika* **95**, 265–278. MR2521583

[8] Meinshausen, N., Bickel, P. and Rice, J. (2009). Efficient blind search: Optimal power of detection under computational cost constraints. *Ann. Appl. Statist.* **3**, 38–60.

[9] Shaffer, J.P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81**, 826–831.