Hierarchical Testing of Variable Importance

Nicolai Meinshausen

367 Evans Hall
University of California, Berkeley
Berkeley, CA 94720
nicolai@stat.berkeley.edu

Abstract. A frequently encountered challenge in high-dimensional regression is the detection of relevant variables. Variable selection suffers from instability and the power to detect relevant variables is typically low if predictor variables are highly correlated. When taking the multiplicity of the testing problem into account, the power diminishes even further.
To gain power and insight, it can be advantageous to look for influence not at the level of individual variables but rather at the level of clusters of highly correlated variables. We propose a hierarchical approach. Variable importance is first tested at the coarsest level, corresponding to the global null hypothesis. If possible, the method tries then to attribute any effect to smaller sub-clusters or even individual variables. The smallest possible clusters which still exhibit a significant influence on the response variable are retained.
It is shown that the proposed testing procedure controls the family-wise error rate at a prespecified level, simultaneously over all resolution levels. The method has comparable power to Bonferroni-Holm on the level of individual variables and dramatically larger power for coarser resolution levels. The best resolution level is selected adaptively.

Keywords: Multiple linear regression, Multiple testing, Hierarchical clustering, Higher Criticism, High-dimensional alternatives

1. Introduction

It is increasingly common nowadays in many fields of the sciences that data are collected at a large scale, the number of measured variables often going into the hundreds or thousands. It is furthermore often believed that a small subset of the variables carries most or all of the interesting information. It is, however, usually not known in advance which those relevant variables are, which is the justification for gathering all the information in the first place. It is furthermore unknown whether there is indeed one unique set of important variables.

While one might only be interested in optimal predictive performance for some applications (see also Breiman, 2001), it can in other cases be interesting to find the variables that carry most or all of the relevant information.

We make matters more precise below for the simple case of a linear model. Considering the standard setting of a fixed design, the n-dimensional response vector $\boldsymbol{Y}$ is given by

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon}$ is i.i.d. standard normal distributed variable and X is the $n \times m$-dimensional matrix which columns consist of the m predictor variables $X_i, i = 1, \ldots, m$.

A sparsity assumption would entail that most entries of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ are actually zero or close to zero. These variables could be discarded as all the information about the response variable $\boldsymbol{Y}$ can be learned from the remaining variables.

In the following, we show some fallacies when testing for each variable whether $\boldsymbol{\beta}_k = 0$, as routinely done when fitting a standard or generalized linear model in R (R Development Core Team, 2005) and some other software packages. We show that the problem of testing for variable importance is in some sense ill-posed if there is a large number of correlated predictor variables in the model. Hierarchical testing of variable importance addresses this concern and leads to stable recovery of important clusters of variables.

Understanding variable importance can have many faces. While testing for variable importance is one approach, the issue is clearly related to the much broader "age old" problem of model selection in statistics. Abramovich et al. (2006) write: "The problem of model selection has attracted the attention of both applied and theoretical statistics for as long as anyone can remember". There is indeed by now an immense literature on model selection (Akaike, 1970; Schwarz, 1978; Mallows, 1973; Breiman and Freedman, 1983; Foster and George, 1994, to name a few). We only note that some of the known fallacies of model selection, most notably instability of the selected subset of variables (Breiman, 1995), are directly related to problems when testing the null hypothesis of a vanishing regression coefficient for each variable separately.

1.1. Hierarchical testing of variable importance.

For the linear model in (1), a test of variable importance is routinely done by many statistical software packages by testing for each variable $k = 1, \ldots, m$ the null hypothesis that the regression coefficient is zero

$$\begin{aligned} H_{0,k} &: \boldsymbol{\beta}_k = 0, \qquad \text{versus} \\ H_{A,k} &: \boldsymbol{\beta}_k \neq 0 \end{aligned} \tag{2}$$

Note that each test is multivariate and considers all other variables as nuisance parameters. This is very different from (and arguably more desirable than) testing only the marginal association between all predictor variables and the response variable.

If it is unrealistic to assume that regression coefficients are identically zero, one might want to use instead of (2) the null hypothesis that the absolute value of the regression coefficient is smaller than some constant. Nevertheless, we focus in the following on (2) for simplicity of exposition.

There are two reasons why testing at the level of individual variables is cursed with weak power.

- *Correlation.* The variance of $\hat{\boldsymbol{\beta}}_k$ is very high if variable k is strongly correlated with other variables.
- *Multiplicity.* The multiplicity of the testing problem has to be taken into account (Westfall and Young, 1993; Shaffer, 1995; Benjamini and Hochberg, 1995). If there are many predictor variables, this leads to a further reduction in the power to detect important variables.

To demonstrate the first point, a very simple example might be instructive. Consider $m = 2$ predictor variables with empirical correlation $0 \leq \rho \leq 1$. For an increasing correlation ρ, the variance of the estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ increases like $(1-\rho^2)^{-1}$. The sample size required to attain a certain power, when testing either the null hypothesis $H_{0,1}$ or $H_{0,2}$, diverges if $\rho \to 1$. For values of ρ close to 1, it might be possible to say that *at least one* of the variables 1 and 2 is relevant for prediction of the response variable, but it might be very difficult to establish *which* of the two variables has a non-vanishing regression coefficient (or, indeed, to establish whether both of them have a non-vanishing regression coefficient).

The question arises then if a cluster of, say, five highly correlated variables should be treated as one entity or five separate hypotheses. If treated as one entity, the null hypothesis for a cluster C of variables can be formulated as

$$H_{0,C} \quad : \quad \boldsymbol{\beta}_k = 0 \text{ for all } k \in C, \qquad \text{versus} \tag{3}$$

$$H_{A,C} \quad : \quad \boldsymbol{\beta}_k \neq 0 \text{ for at least one } k \in C \tag{4}$$

A null hypothesis for a cluster of variables is considered false if *at least one* variable in the cluster has a non-vanishing regression coefficient. A false null hypothesis at a coarse level indicates thus that some variables in this cluster are relevant. It is then of interest to find out *which* of these variables are the important ones. Conversely, if the null hypothesis is true for a cluster C, it is by definition also true for all subclusters $C' \subseteq C$ and the search for relevant variables in C can be terminated.

Here, we propose a hierarchical approach to the challenge of testing variable importance. If there is a significant effect in a cluster C and the null hypothesis $H_{0,C}$ can be rejected, an attempt is made to attribute the effect to individual variables in the cluster. The aim is to identify the smallest possible clusters of variables that exhibit

a significant influence on the response variable (an individual variable is considered as a cluster with cardinality 1).

1.2. Related work.

A related hierarchical approach to testing has been proposed by Blanchard and Geman (2005), where certain hierarchical testing designs are analyzed in terms of their cost-effectiveness, with applications in image analysis. Here we focus more on the statistical part of hierarchical testing procedures. While Pacifico et al. (2004) show FDR-control for random fields, a more closely related work is the technical report of Benjamini and Yekutieli (2003), where FDR-control for hierarchical tests is proposed. Their method relies, however, strongly on the independence assumption between test statistics. This assumption is, almost by definition, not fulfilled in the present context. For datasets with many correlated predictor variables, as they occur frequently in computational biology, Hastie et al. (2001) proposed *Tree Harvesting*. The main aim of *Tree Harvesting* is to improve prediction through use of a hierarchical clustering structure, whereas the contribution of the current work is a formal testing procedure for such hierarchical structures.

Some more recent work has focused on achieving model selection through ℓ_1-penalization of the coefficient vector (Frank and Friedman, 1993; Tibshirani, 1996; Knight and Fu, 2000; Fan and Li, 2001; Efron et al., 2004). It is often cited as an advantage of the Lasso over, say, Ridge regression that the Lasso leads to model selection as some regression coefficients are set to zero. The question turns up if the the Lasso (or any other model selection procedure for that matter) selects the "right" variables? From an asymptotic point of view, the answer seems to be often positive, even if the number of variables is very large (Meinshausen and Bühlmann, 2006). However, the consistency requires a condition on the design matrix (Meinshausen and Bühlmann, 2006; Zou, 2005; Zhao and Yu, 2006). Furthermore, Zou and Hastie (2005) point out that the Lasso has a tendency to pick just one variable in a cluster of highly correlated variables and disregard all other variables in this cluster. They propose the *elastic net* as an alternative to the Lasso; the elastic net tends to pick all variables in the cluster, thus reducing the variance of the selected set of coefficients. Incorporating a priori hierarchical knowledge into the penalization procedure has been proposed recently by (Yuan and Lin, 2006; Zhao et al., 2006; Meier et al., 2006). In contrast, the hierarchical structure is in this manuscript not imposed by prior knowledge, although the approach could be extended in this direction if one is dealing with factors as predictor variables for example. More fundamentally, the current approach is not so much about penalized estimation as it is about rigorous testing of variable importance. The price to pay for this is the current limitation to $m \leq n$ situations, as no test for the influence of a group of variables is available if the number of variables m exceeds the number n of samples.

2. Hierarchical Testing

Hierarchical testing steps through a given hierarchy of clusters. First, collective effects are measured at the coarsest level possible (the global null hypothesis that no variable exhibits an influence on the response variable.) If this global null hypothesis can be rejected, finer resolution levels are tested for an effect until the level of individual variables is reached. The motivation for hierarchical testing can be summarized as follows.

- *Any effect at all?* The influence of a group of variables on the response variable can be examined by testing whether the regression coefficients of *all* variables in the cluster can plausibly be zero.
- *Attribution of effects to sub-clusters.* If it is established that a cluster of variables does indeed have an influence on the response variable, it is desirable to attribute the effect to one or several subclusters.

If possible, the influence of a cluster of variables on the response variable is attributed to its subclusters. In each subcluster, it is again examined whether the collective effect can be attributed to even smaller subclusters of variables. The procedure retains the smallest possible clusters which exhibit a significant influence on the response variable.

The outlined idea assumes that there is a natural hierarchy of clusters of variables available. After discussing the choice of the hierarchy, the actual testing procedure is given in detail and error control is shown.

2.1. Hierarchies

The hierarchical testing procedure is based on a *given* hierarchy of clusters. This hierarchy can be derived from specific domain knowledge. In computational biology, it might for example be interesting to use the Gene Ontology (Ashburner et al., 2000) when testing for the influence of particular genes on survival times. The Gene Ontology does not posses the hierarchical nature of the hierarchies used below, but the approach could be made feasible (with some more cumbersome notation) for Gene Ontology and related hierarchies derived from domain knowledge.

If no specific domain knowledge exists and the hierarchies cannot be derived in another natural way (as possible for example for factor variables), one can employ hierarchical clustering (Hartigan, 1975; Ward, 1963) to select a suitable hierarchy of clusters. As discussed above, it is difficult to distinguish between effects of highly correlated variables. With hierarchical clustering, highly correlated variables tend to end up in a single small cluster. If such a cluster contains truly important variables, it can be easily identified with the chosen approach.

For the following, we assume that a hierarchy $\mathcal{T}$ is given, which is a set of clusters $C \subseteq \{1, \ldots, m\}$. The cardinality of clusters C is denoted by $|C|$. The root node $\{1, \ldots, m\}$ contains all variables and has cardinality m. The hierarchical structure implies that any two clusters $C, C' \in \mathcal{T}$ either have an empty intersection, $C \cap C' = \emptyset$, or that one cluster is a subset of the other.

2.2. Testing procedure

The null hypothesis (4) for a cluster $C \in \mathcal{T}$ is fulfilled if and only if *all* variables in this cluster have vanishing regression coefficients. A test for this hypothesis would typically be the partial F-test. However, other tests can be more powerful in the high-dimensional setting (Goeman et al., 2005). We assume for the moment that such a test is available. The p-value of a test of $H_{0,C}$ is denoted by p_C.

P-value adjustment. To take the multiplicity of the testing problem into account, p-values have to be adjusted. Define the adjusted p-value p^C_{adj} as

$$p^C_{adj} = p^C \frac{m}{|C|}. \tag{5}$$

The adjustment amounts to multiplying the p-value of each cluster C with the inverse of the fraction $|C|/m$ of variables it contains. The adjustment is thus resolution-dependent. At coarse resolutions, the penalty for multiplicity is weak, and it increases for finer resolution levels. The p-value of the root node is thus unadjusted, whereas individual variables receive a Bonferroni-type adjustment.

The hierarchical testing procedure rejects now all hypotheses $H_{0,C}$ with $C \in \mathcal{T}$ for which

(a) the adjusted p-value p^C_{adj} is below or equal to the specified level α and

(b) the parent node is rejected (this is always considered to be fulfilled for the root node).

Note that condition (b) is not a severe restriction. The null hypothesis $H_{0,C}$ of a node C is by definition always true if the null hypothesis $H_{0,\mathrm{pe}(C)}$ is true for the parent node $\mathrm{pe}(C)$. Hence it makes sense to stop testing in subtrees of nodes whose null hypothesis could not be rejected.

Hierarchical p-value adjustment. For notational simplicity, the procedure given above is expressed in a slightly different way. The hierarchically adjusted p-value is defined as

$$p^C_{h,adj} = \max_{D \in \mathcal{T}: C \subseteq D} p^D_{adj}. \tag{6}$$

The hierarchically adjusted p-value of a node is thus always smaller than the hierarchically adjusted p-value of the parent node. Using the definition of a hierarchically adjusted p-value, the set of rejected clusters in the hierarchy $\mathcal{T}$ is then given by

$$\mathcal{T}_{rejected} = \{C \in \mathcal{T} : p^C_{h,adj} \leq \alpha\}. \tag{7}$$

It is easy to see that the set of rejected nodes is completely equivalent to the set of rejected nodes with the previous procedure. In particular, no node can be rejected if its parent node has not been rejected.

Remarks. The global null hypothesis can be tested at level α and there is no adjustment for multiplicity at this coarsest resolution. The largest penalty is incurred at the finest resolution level, where individual variables are tested.

Even though the procedure offers simultaneous control over all resolution levels, the penalty for coarser resolutions (larger clusters) is not influenced by the fact that the procedure will turn to finer resolution levels later on. The global null hypothesis (corresponding to the root node) can for example be tested at level α, even though the procedure offers simultaneous control over all resolutions at level α.

It is shown below that the procedure offers indeed simultaneous control of the family-wise error rate at level α. It will be seen later that the method can be improved upon, so that the penalty at the level of individual hypothesis can be reduced by a factor 2 in general.

2.3. Control of the family-wise error rate

A node fulfills the respective null hypothesis $H_{0,C}$ according to (4) if and only if all regression coefficients of variables in this cluster are zero. The set of nodes (or clusters) that fulfill the null hypothesis is denoted by

$$\mathcal{T}_0 := \big\{C \in \mathcal{T} : H_{0,C} \text{ is fulfilled}\big\}. \tag{8}$$

Family-wise error rate control entails that the probability of rejecting any cluster in $\mathcal{T}_0$ is smaller than the pre-specified level.

Theorem 1. *For $\mathcal{T}_{rejected}$ defined as in (7), the family-wise error rate is controlled at level α,*

$$P\big(\mathcal{T}_{rejected} \cap \mathcal{T}_0 = \emptyset\big) \geq 1 - \alpha \tag{9}$$

Proof. The probability of an error can be rewritten as

$$P\big(\mathcal{T}_{rejected} \cap \mathcal{T}_0 \neq \emptyset\big) = P\big(\exists\, C \in \mathcal{T}_0 \;:\; p^C_{h,adj} \leq \alpha\big).$$

Define the set $\tilde{\mathcal{T}}_0$ as the set of all clusters which fulfill the null hypothesis and which are maximal in the following sense,

$$\tilde{\mathcal{T}}_0 := \big\{C \in \mathcal{T}_0 : \nexists D \in \mathcal{T}_0 \text{ with } C \subset D\big\}. \tag{10}$$

It holds that $\tilde{\mathcal{T}}_0 \subseteq \mathcal{T}_0$. It follows from the definition in (6) of a hierarchically adjusted p-value, that an error committed in a cluster $C \in \mathcal{T}_0 \setminus \tilde{\mathcal{T}}_0$ implies an error in a set $C' \in \tilde{\mathcal{T}}_0$, where $C \subset C'$. It follows that it is sufficient to focus on the probability of making an error in the set $\tilde{\mathcal{T}}_0 \subseteq \mathcal{T}_0$

$$\begin{aligned} P\big(\exists\, C \in \mathcal{T}_0 \;:\; p^C_{h,adj} \le \alpha\big) &= P\big(\exists\, C \in \tilde{\mathcal{T}}_0 \;:\; p^C_{h,adj} \le \alpha\big) \\ &\le P\big(\exists\, C \in \tilde{\mathcal{T}}_0 \;:\; p^C_{adj} \le \alpha\big). \end{aligned}$$

The right hand side is bounded by Bonferroni's inequality from above by

$$\sum_{C \in \tilde{\mathcal{T}}_0} P(p^C_{adj} \le \alpha) \le \sum_{C \in \tilde{\mathcal{T}}_0} (\alpha\, |C|/m), \tag{11}$$

where the last inequality follows from the definition (5) of the adjusted p-value. It is thus sufficient to show that $\sum_{C \in \tilde{\mathcal{T}}_0} |C| \le m$. Note that by definition of $\tilde{\mathcal{T}}_0$ in (10),

$$\forall C, C' \in \tilde{\mathcal{T}}_0 \;:\quad C \cap C' = \emptyset.$$

On the other hand it clearly holds that $\cup_{C \in \tilde{\mathcal{T}}_0} C \subseteq \{1, \ldots, m\}$. Hence it follows indeed that $\sum_{C \in \tilde{\mathcal{T}}_0} |C| \le m$, which completes the proof.

2.4. Shaffer-Improvement

In the current proposal (5) for hierarchical p-value adjustment, the adjustment for a cluster $C \in \mathcal{T}$ is achieved through multiplication of the relevant p-value with $m/|C|$, where $|C|$ is the number of variables that the cluster contains. The root node is thus not penalized at all, whereas individual variables receive a penalty proportional to the total number m of variables. The adjustment of p-values for individual variables is thus identical to the Bonferroni adjustment.

The Bonferroni adjustment assumes the worst case, namely that all hypotheses are true null hypotheses (similar for the Bonferroni-Holm adjustment). This possibility cannot be excluded a priori if one is just looking at the level of individual variables. In a hierarchical structure, not all combinations of null hypotheses are possible. Incorporating the constraints on the possible combinations can increase the power of the procedure, as already shown in a more general context by Shaffer (1986).

As will be seen, the penalty factor can in general be reduced by a factor 2. The improvement will be developed for binary trees only. Similar improvements are possible

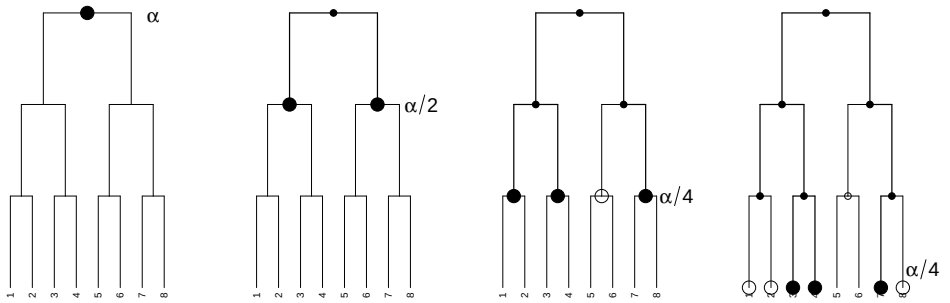

Fig. 1. *An example of the testing process for a binary tree. The global null hypothesis is tested first at level α (left panel), and the level of individual variables is reached last (right panel). Note that individual hypotheses can be tested at level $\alpha/4$ and not $\alpha/8$ as one might expect at first.*

for non-binary trees but are of marginal interest in this context and thus omitted for ease of exposition.

As a motivation, consider the case of a cluster $C = \{X_1, X_2\}$ of two variables. If both variables are true null hypotheses, then C fulfills the null hypothesis $H_{0,C}$. By definition of $\tilde{\mathcal{T}}_0$, only the cluster C will appear in the set $\tilde{\mathcal{T}}_0$ in this case (and the individual variables will not). If the cluster appears in $\tilde{\mathcal{T}}_0$, the adjustment at the level of the individual variables X_1, X_2 is irrelevant for the probability of making errors, as an error at the level of individual variables can only be made if the cluster C was already falsely rejected. Thus, one only needs to prepare for the case that *one* of the two individual variables corresponds to a true null hypothesis.

The *sibling* $\mathrm{si}(C)$ of a node C is defined as the children $\mathrm{ch}(D)$ of the parent node which are not identical to C, that is $\mathrm{si}(C) := \mathrm{ch}(\mathrm{pa}(C)) \setminus C$. With this definition, the adjusted p-value is now computed in analogy to (5) as

$$p^C_{adj} = p^C \frac{m}{|C|_{\mathit{eff}}}, \tag{12}$$

where the *effective* cluster size $|C|_{\mathit{eff}}$ is now given by

$$|C|_{\mathit{eff}} = \begin{cases} |C| & \text{if } \mathrm{si}(C) \text{ is not a leaf node} \\ |C| + |\mathrm{si}(C)| & \text{if } \mathrm{si}(C) \text{ is a leaf node} \end{cases}$$

Clusters which have a leaf node as a sibling have a larger *effective* cluster size and receive thus less penalty. If both children of a cluster are leaf nodes (and contain just one variable), this amounts to dividing the penalty for individual variables by a factor 2. The procedure remains unchanged otherwise.

Example. The improved procedure is demonstrated for a toy example with 8 hypotheses that form a binary tree in Figure 1. First, the global null hypothesis is tested at level α (leftmost panel). It is rejected. In the next step (second panel), it is examined if the effect can be attributed to one or both of the sub-clusters that follow in the hierarchy. Each of these two sub-clusters is tested at level $\alpha/2$. They are again both rejected in this example and the procedure turns to the next four clusters, which are tested at level $\alpha/4$ (third panel). Of these four hypotheses, one cluster made up of variables 5 and 6 is not rejected. Consequently, variables 5 and 6 are not tested anymore at the individual level in the last step (rightmost panel).

Note that the remaining 6 hypotheses can be tested at level $\alpha/4$ and not $\alpha/8$, as one might expect. At the level of individual variables, the procedure gains thus potentially additional power compared to a Bonferroni or Bonferroni-Holm testing.

Error control with the improved method. It is now shown that the level is still maintained after the improvement.

Theorem 2. *For $\mathcal{T}_{rejected}$ defined as in (7) with the adjusted p-values using definition (12) instead of (5), the family-wise error rate is still controlled at level α,*

$$P\big(\mathcal{T}_{rejected} \cap \mathcal{T}_0 = \emptyset\big) \geq 1 - \alpha \tag{13}$$

Proof. Starting from (11) in the proof of Theorem 1, it is sufficient to show that $\sum_{C \in \tilde{\mathcal{T}}_0} |C|_{\mathit{eff}} \leq m$. For simplicity of exposition, one can assume for the moment that there is only *one* cluster $D \in \tilde{\mathcal{T}}_0$ for which $|D|_{\mathit{eff}} > |D|$. Then

$$\sum_{C \in \tilde{\mathcal{T}}_0} |C|_{\mathit{eff}} = \sum_{C \in \tilde{\mathcal{T}}_0 \setminus D} |C| + |D|_{\mathit{eff}} = \sum_{C \in \tilde{\mathcal{T}}_0 \setminus D} |C| + |D| + |\mathrm{si}(D)|. \tag{14}$$

As $|D|_{\mathit{eff}} > |D|$ by assumption, it follows that the null hypothesis $H_{0,\mathrm{si}(D)}$ for the sibling $\mathrm{si}(D)$ is false. If this would not be the case, the null hypothesis $H_{0,\mathrm{pa}(D)}$ of the parent node $\mathrm{pa}(D)$ of node D would be true, which would imply that $D \notin \tilde{\mathcal{T}}_0$ by definition of $\tilde{\mathcal{T}}_0$ in (10). By contradiction, we can thus conclude that $H_{0,\mathrm{si}(D)}$ is not fulfilled. As $H_{0,\mathrm{si}(D)}$ is not fulfilled, it holds for all $C \in \tilde{\mathcal{T}}_0$ that $\mathrm{si}(D) \cap C = \emptyset$. Thus

$$\sum_{C \in \tilde{\mathcal{T}}_0 \setminus D} |C| + |D| \leq m - |\mathrm{si}(D)|,$$

which shows that the right hand side of (14) is indeed smaller than m, which completes the proof.

As the improved method is uniformly more powerful than the first proposal, it is always worthwhile to incorporate the improvement.

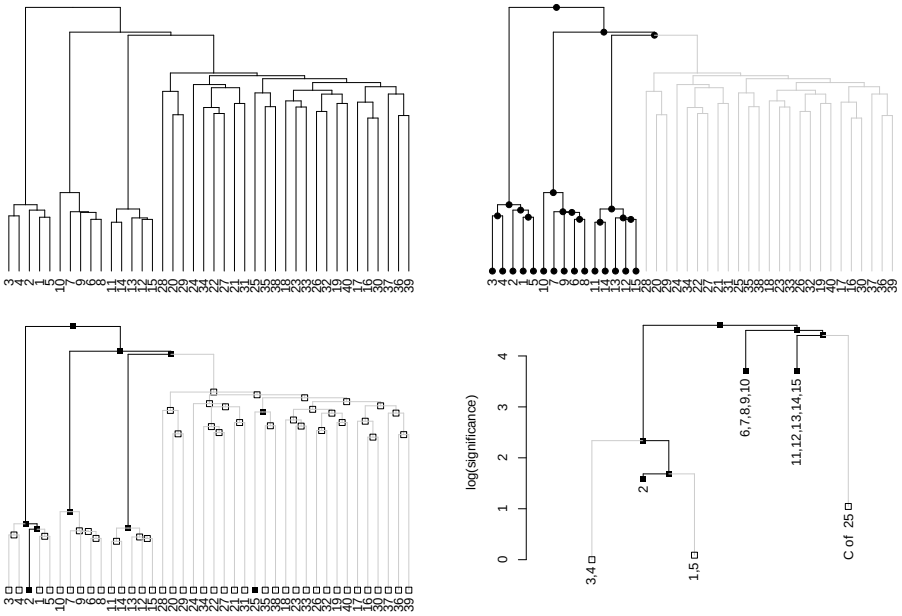

Fig. 2. *The hierarchical testing procedure illustrated for Example (d) of Zou and Hastie (2005). The hierarchical clustering structure that enters the testing procedure (top left); the true model structure (top right), where false null hypotheses are indicated by a dot and darker edges; the testing result (bottom left), where rejected nodes are indicated by darker edges; a visualization of the significance of the nodes (bottom right), where rejected nodes are again indicated by filled boxes and darker edges. In this latter plot, a larger height of a cluster of variables corresponds to a higher significance of the rejection and clusters that cannot be rejected are shown as one entity to give an uncluttered representation of the results. Hierarchical testing identifies the three relevant clusters of variables.*

2.5. Visualization

The procedure is illustrated graphically for Example (d) of Zou and Hastie (2005). On the top left of Figure 2, the hierarchical clustering structure is shown, as produced by complete linkage, with the distance taken as the Spearman correlation. Note that, at this stage, the response variable has not been used. A visualization of the "true model structure" is shown on the top right. All clusters which contain at least one variable with non-vanishing regression coefficient are indicated by a dot and darker edges.

The test result is shown on the bottom left. Clusters which show a significant effect (after multiplicity correction) are shown as dark boxes, whereas non-significant clusters are shown as unfilled boxes. Testing starts at the root node and progresses down the tree, stopping at the first non-significant result. There are two nodes that do have a significant adjusted p-value but whose hierarchically adjusted p-value is not significant; the procedure stops before it reaches those two nodes.

A compact visualization of the test result is shown in the bottom right panel. The height of a cluster node C is now taken to be $-\log_{10}(p_C)$, where p_C is the adjusted p-value of the cluster node. Every node above $-\log_{10}(0.05) \approx 1.3$ is hence significant at the 5% level. The higher the node, the more significant is the rejection of its respective null hypothesis. Due to the hierarchical p-value adjustment, nodes are always at the same height or below their parent nodes. Rejections are again shown by dark edges and filled boxes. To unclutter the visual display, non-significant clusters are only represented by their common node ("C of 25" stands for "Cluster of 25 variables").

The first cluster of five variables is broken apart and variable 2 is identified as individually significant. The other two relevant clusters are identified as significant entities of five variables each. The main principle of the hierarchical approach can be seen in this simple example. The method retains the smallest clusters possible that still exhibit a significant influence on the response variable.

2.6. Testing against high-dimensional alternatives

Above, we specified the hierarchical testing procedure under the condition that a test of $H_{0,C}$ is available for every cluster C in the hierarchy.

As pointed out in Goeman et al. (2005), the partial *F-test* has often rather weak power when testing against high-dimensional alternatives. Goeman et al. (2005) suggested as an alternative a *Score test*, which is a locally most powerful tests in some specific neighborhood of the distribution that corresponds to $H_{0,C}$.

As a third alternative, we consider *Higher Criticism*, which was introduced by Donoho and Jin (2004), based on a proposal by Tukey, and was shown to posses certain optimal properties for multiple testing when there are just very few false null hypotheses

(Donoho and Jin, 2004). Two modifications are necessary for the present purpose. First, Higher Criticism assumes independent test statistics. For a fixed design matrix, the covariance between parameter estimates is known, and we rotate the estimates of all variables in a considered cluster into a coordinate system such that the estimates are decorrelated. Second, the variance is estimated from the data. This additional source of variation is, however, easily incorporated in a simulation-based approach, as for example in Meinshausen (2006).

A drawback of all of these tests is that they require the number of samples n to be larger than the number of variables m. Developing a test for $n < m$ situations, possibly based on a regularized estimator, seems highly desirable yet challenging. For the moment, we are thus restricted to $m \leq n$ situations. This is not directly a restriction of the hierarchical approach. The method itself could easily be extended to $m > n$ if an appropriate test would be available.

3. Numerical Examples

To test the power of the proposed method, one needs artificial datasets with knowledge about the *true underlying model*. Here, we choose the four examples put forth in Zou and Hastie (2005). For sake of completeness, the four examples are described below. Complete linkage was chosen as the hierarchical clustering method, with the distance based on the Spearman correlation.

3.1. Datasets.

The response variable $\boldsymbol{Y}$ is sampled as $\boldsymbol{Y} = X\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}_i \overset{iid}{\sim} \mathcal{N}(0,1)$ for all $i = 1, \ldots, n$. The $n \times p$ matrix X of predictor variables is generated as follows.

(a) *Sparse Toeplitz*
In example (a), $\boldsymbol{\beta} = (3.1, 5, 0, 0, 2, 0, 0, 0)$, $\sigma = 3$, and $n = 20$. The pairwise correlation between X_k and X_l is $0.5^{|k-l|}$ for all $k, l \in \{1, \ldots, 8\}$.

(b) *Non-sparse Toeplitz*
Example (b) is the same as example (a), except that $\boldsymbol{\beta}_k = 0.85$ for all $k = 1, \ldots, 8$.

(c) *Uniform correlation*
For example (c), $n = 50$, $\sigma = 15$ and $\boldsymbol{\beta} = (0, \ldots, 0, 2, \ldots, 2, 0, \ldots, 0, 2, \ldots, 2)$, where each block of 0 or 2 entries respectively has length 10. The correlation between variables k and l is set 0.5 for all $k, l \in \{1, \ldots, 40\}$.

(d) *Factor model*
For example (d), $n = 50$, $\sigma = 15$ and $\boldsymbol{\beta} = (3, \ldots, 3, 0, \ldots, 0)$, where the first block has length 15 and the second block has length 25. The predictor variables are generated as follows:

$$X_k = \boldsymbol{Z}_{j(k)} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}_i \overset{iid}{\sim} \mathcal{N}(0,1)$ for all $i = 1, \ldots, 40$,

$$j(k) = \begin{cases} 1 & k \leq 5 \\ 2 & 6 \leq k \leq 10 \\ 3 & 11 \leq k \leq 15 \\ 4 & k > 15 \end{cases},$$

and, for $j = 1, \ldots, 4$, the vector $\boldsymbol{Z}_j$ has i.i.d. $\mathcal{N}(0, \sigma_j^2)$ distributed entries with $\sigma_1^2, \sigma_2^2, \sigma_3^2 = 1$ and $\sigma_4^2 = 0$.

It is maybe of interest that these datasets were used in Zou and Hastie (2005) in a similar context. It was in particular demonstrated that the *elastic net* performs well on these datasets, as all variables in a cluster tend to be selected. The *Lasso* in contrast tends to pick just one variable from a cluster of highly correlated variables and sets the coefficients of all other variables in the cluster to zero. The idea of hierarchical testing is similar to that of *elastic net* in this regard: if it is not possible to reliably single out the important variables, try at least to capture the important *clusters* of variables.

3.2. Power.

The power of the hierarchical testing procedure is clearly a function of the "resolution level" at which we are looking. The global null hypothesis is always tested at level α. The power at this coarsest resolution is rather large. For finer resolution levels, the power diminishes in general, as both the multiplicity of the testing problem increases and the effect of highly-correlated variables can be very difficult to separate.

Measuring power at different resolutions. To measure the power at a given resolution level, we fix a cluster size c with $1 \leq c \leq m$. Ideally, one would like to measure the power for clusters with cardinality equal to c. There are in general very few clusters with any given cardinality. To gain better estimates, we measure the power over all clusters that have *almost* cardinality c. First, select the set of all clusters $\mathcal{D}(c)$ who do not fulfill the null hypothesis and whose cardinality does not exceed c,

$$\mathcal{D}(c) = \{C \in \mathcal{T} : |C| \leq c \text{ and } H_{0,C} \text{ is false } \}.$$

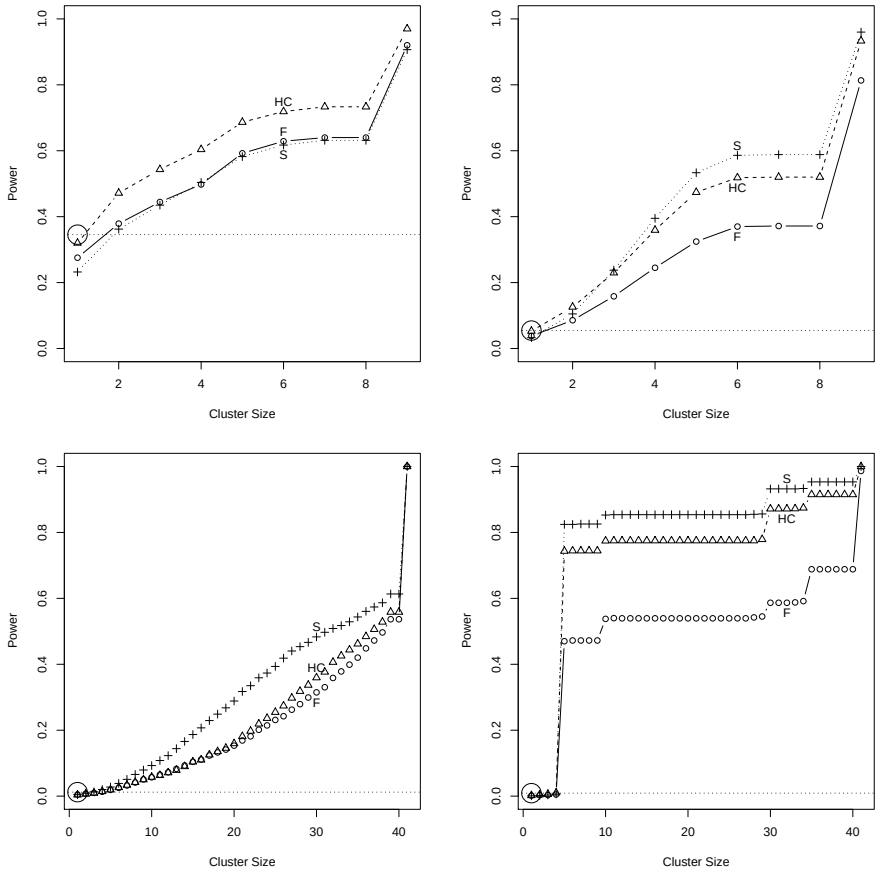

Fig. 3. *The power of the hierarchical testing procedure as a function of cluster size. Shown are the results for the F-test (∘, −), Higher Criticism (△, −−), and the Score test (+, ·), compared with the power of Bonferroni-Holm (large circle at cluster size 1 and horizontal dashed line) for Examples (a)-(d) (from top left to bottom right).*

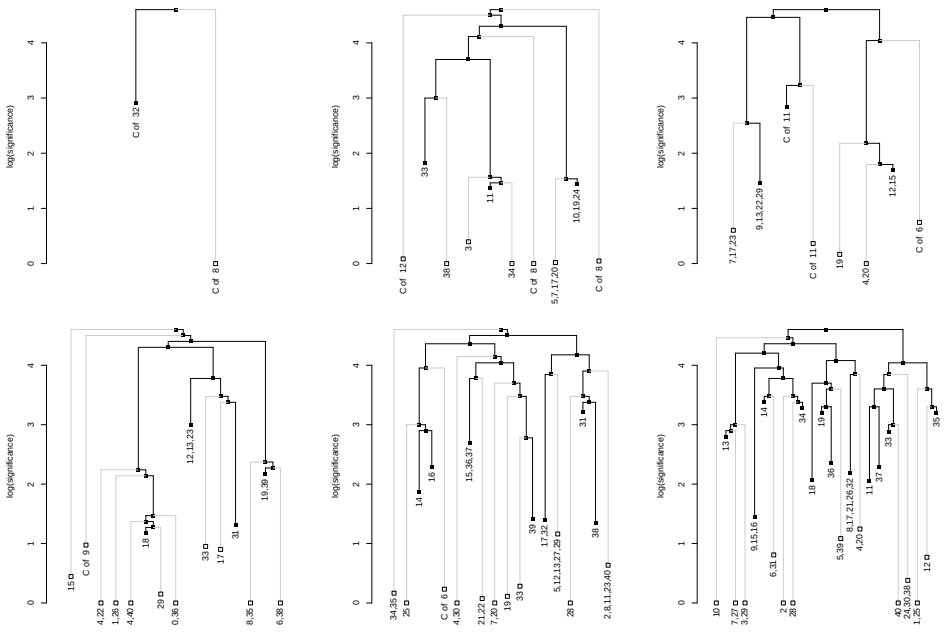

Fig. 4. *The effect of an increasing sample size on the hierarchical testing procedure. A plot of the rejected cluster nodes and rejected single variables, as in the bottom right panel of Figure 2 for Example (c) of Zou and Hastie (2005), where the height of each cluster is proportional to the logarithm of its significance. Sample size increases from top left to bottom right from 50 to 1600 (doubling at each step). The resolution of the test procedure increases as more samples become available.*

Of these cluster, we select the set $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ that contains the largest clusters in the sense that

$$C \in \tilde{\mathcal{D}} \quad \text{iff} \quad \{C \in \mathcal{D}\} \ \wedge \ \{\nexists\, C' \in \mathcal{D} : \ C \subset C'\},$$

where the dependence on the cluster size c has been omitted for notational simplicity. The power at cluster size c is then computed as the mean rejection rate of the clusters in $\tilde{\mathcal{D}}(c)$. In Figure 3, the power is shown as a function of cluster size for three mentioned tests of collective effects: the F-test, Higher Criticism of Donoho and Jin (2004) and the Score test proposed by Goeman et al. (2005).

Results. There are three main observations. First, at the finest resolution level (where individual variables are tested) the proposed hierarchical testing procedure has comparable power to the traditional Bonferroni-Holm testing. One could expect the hierarchical approach to have higher power at this level, as variables just have to be tested at level $2\alpha/m$ in general compared to α/m with the Bonferroni correction

or $\alpha/(m-q+1)$ with the Bonferroni-Holm procedure (where q is the rank of the p-value in question), see Holm (1979). However, due to the hierarchical nature, some variables are not even tested at this level as the procedure already stopped at a coarser resolution. These two effects seem to balance each other out.

Second, the power to detect important clusters of variables can be dramatically higher than the power to detect individual variables. Consider Example (d). The power to detect individual variables in the three important clusters is practically zero, as the correlation between the variables is so high that the effect of the cluster cannot be attributed to the individual variables. The power to detect the three important clusters of variables is in contrast around 0.8 with either the Score test or Higher Criticism and jumps to above 0.99 if the sample size is increased to $n = 100$ (not shown here).

Third, the choice of the test really does matter. The F-test seems to have the weakest power, as could already be expected from results in Goeman et al. (2005). Higher Criticism is the most powerful test for Example (a), while the same is true for the Score test for Examples (b)-(d). Both tests seem to be reasonable alternatives to the F-test when testing against high-dimensional alternatives.

3.3. Increasing sample size.

Figure 4 illustrates the effect of an increasing sample size on the "resolution" of the hierarchical testing procedure. Given only very few samples, the procedures returns just a few large clusters that exhibit significant influence on the response variable. If more samples become available, the resolution of the testing procedure increases as the effect can be attributed to smaller clusters of variables or individual variables.

4. Discussion

Model selection is in general very unstable with respect to small perturbations in the data, particularly if there are many highly correlated variables. Testing individual variables for relation with the response variable is thus cursed with weak power in such settings.

A hierarchical approach looks for the smallest possible clusters of variables that still have a significant relation with the response variable. If possible, the effect is attributed to individual variables.

The power of the hierarchical procedure is a function of the resolution. The power to detect variables at the individual level was shown to be comparable to the Bonferroni-Holm procedure. The detection rate of important small clusters is already dramatically higher, and increases further for coarser resolution levels.

Most importantly, the proposed method selects adaptively the right resolution. If evidence is weak, only a few large clusters are detected. For larger sample sizes, the effect can be attributed to increasingly smaller clusters. Even though the resolution level is chosen adaptively, the approach offers simultaneous control of making a single false rejection over all resolution levels.

References

Abramovich, F., Y. Benjamini, D. Donoho, and I. Johnstone (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics 34*.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics 22*, 203.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. *Nature Genetics 25*, 25 – 29.

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

Benjamini, Y. and D. Yekutieli (2003). Hierarchical FDR testing of trees of hypotheses. Technical report.

Blanchard, G. and D. Geman (2005). Hierarchical testing designs for pattern recognition. *Annals of Statistics 33*, 1155–1202.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics 37*, 373–384.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science 16*(3), 199–215.

Breiman, L. and D. Freedman (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association 78*, 131–136.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics 32*, 962–995.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–451.

Fan, J. and R. Li (2001). Variable selection via penalized likelihood. *JASA 96*, 1348–1360.

Foster, D. and E. George (1994). The risk inflation factor in multiple linear regression. *Annals of Statistics 22*, 1947–1975.

Frank, I. and J. Friedman (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics 35*, 109–148.

Goeman, J. J., S. A. van de Geer, and H. C. van Houwelingen (2005). Testing against a high-dimensional alternative. *JRSS B*.

Hartigan, J. (1975). *Clustering algorithms*. Wiley.

Hastie, T., R. Tibshirani, D. Botstein, and P. Brown (2001). Supervised harvesting of expression trees. *Genome Biology 2*, 1–12.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

Mallows, C. (1973). Some comments on C_p. *Technometrics 15*, 661–675.

Meier, L., S. van de Geer, and P. Bühlmann (2006). The group lasso for logistic regression. Technical report, Seminar für Statistik, ETH Zurich.

Meinshausen, N. (2006). False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics 33*, 227–237.

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics 34*, 1436–1462.

Pacifico, M., C. Genovese, I. Verdinelli, and L. Wasserman (2004). False discovery control for random fields. *JASA 99*, 1002–1014.

R Development Core Team (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Shaffer, J. (1986). Modified Sequentially Rejective Multiple Test Procedures. *Journal of the American Statistical Association 81*, 826–831.

Shaffer, J. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology 46*, 561–584.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B *58*, 267–288.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *JASA 58*, 236–244.

Westfall, P. and S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B 68*, 49–67.

Zhao, P., G. V. Rocha, and B. Yu (2006). Grouped and hierarchical model selection through composite absolute penalties. Technical report, Dep. Statistics, UC Berkeley.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. Technical Report 702, Department of Statistics, UC Berkeley, to appear in Journal of Machine Learning Research.

Zou, H. (2005). The adaptive lasso and its oracle properties. Technical Report 645, School of Statistics, University of Minnesota, to appear in Journal of the American Statistical Association.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *JRSS B 67*, 301–320.