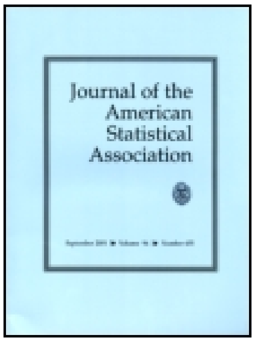

Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Hierarchical Testing in the High-Dimensional Setting With Correlated Variables

Jacopo Mandozzi & Peter Bühlmann

To cite this article: Jacopo Mandozzi & Peter Bühlmann (2016) Hierarchical Testing in the High-Dimensional Setting With Correlated Variables, Journal of the American Statistical Association, 111:513, 331-343, DOI: 10.1080/01621459.2015.1007209

To link to this article: http://dx.doi.org/10.1080/01621459.2015.1007209

Hierarchical Testing in the High-Dimensional Setting With Correlated Variables

Jacopo MANDOZZI and Peter BÜHLMANN

We propose a method for testing whether hierarchically ordered groups of potentially correlated variables are significant for explaining a response in a high-dimensional linear model. In presence of highly correlated variables, as is very common in high-dimensional data, it seems indispensable to go beyond an approach of inferring individual regression coefficients, and we show that detecting smallest groups of variables (MTDs: minimal true detections) is realistic. Thanks to the hierarchy among the groups of variables, powerful multiple testing adjustment is possible which leads to a data-driven choice of the resolution level for the groups. Our procedure, based on repeated sample splitting, is shown to asymptotically control the familywise error rate and we provide empirical results for simulated and real data which complement the theoretical analysis. Supplementary materials for this article are available online.

KEY WORDS: Familywise error rate; Hierarchical clustering; High-dimensional variable selection; Lassol; Linear model; Minimal true detection; Multiple testing; Sample splitting.

1. INTRODUCTION

High-dimensional statistical inference where the number p of (co-)variables might be much larger than the sample size n has become a key issue in many areas of applications. We focus here on the linear model

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon, \ \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I) \tag{1}$$

with $n \times p$ design matrix $\mathbf{X}$, $p \times 1$ regression vector β^0, and $n \times 1$ response $\mathbf{Y}$, allowing for high-dimensionality with $p \gg n$. Often, the active set of variables carrying the relevant information

$$S_0 = \{j; \beta_j^0 \neq 0\}$$

is assumed to be a small subset of all variables, that is, the model is sparse with many β_j^0 being equal to zero. Our main goal is testing of significance of groups of parameters: for a group or cluster $C \subseteq \{1, \dots, p\}$,

$$H_{0,C}: \ \beta_j^0 = 0 \text{ for all } j \in C,$$
$$H_{A,C}: \ \beta_j^0 \neq 0 \text{ for at least one } j \in C.$$

Significance testing in the high-dimensional framework is essential when looking beyond point estimation. Wasserman and Roeder (2009) proposed an approach based on single sample splitting, and Meinshausen, Meier, and Bühlmann (2009) improved the reliability and power of the method based on multiple sample splitting. Minnier, Tian, and Cai (2011) considered a perturbation technique, and (modified) bootstrap-type schemes were analyzed by Chatterjee and Lahiri (2013) and Liu and Yu (2013). Another line of methods have been proposed using low-dimensional regularized projections (e.g., on single variables for individual hypotheses $H_{0,j}$) which have some optimality properties (Bühlmann 2013; Javanmard and Montanari 2014a, 2014b; van de Geer et al. 2014; Zhang and Zhang 2014). However, in presence of highly correlated variables, all these methods are likely to fail for testing individual hypotheses $H_{0,j}$.

An interesting way to address the fundamental limitation of identifiability in presence of high correlation or near linear dependence is given by a hierarchical testing scheme proposed by Meinshausen (2008). First, the variables are grouped in a hierarchical way, for example, by hierarchical clustering. At the top of the hierarchy, the global hypothesis $H_{0,\{1,\dots,p\}}$ is tested. If it can be rejected, a finer partition with clusters $\{C_k\}_k$ is considered, and for the ones where H_{0,C_k} can be rejected, one proceeds down the hierarchy to finer partitions. The method has the powerful advantage that it *automatically* goes (from top to bottom in the hierarchy) to finer resolution with smaller clusters, depending on signal-strength and the correlation structure among the variables. At the end, significant clusters can be typically found, and if the signal for an individual variable is sufficiently strong, even significance of a single variable can be detected. Meinshausen (2008) worked out a simple yet powerful way for controlling the familywise error rate when performing multiple tests in the hierarchy, assuming that there is a method which leads to valid p-values for the various hypotheses tests; for example, when $p < n$ and with Gaussian errors, one can use partial F-tests.

1.1 Our Contribution

As one of our main contributions, we deal here with the problem to obtain valid p-values for hypotheses $H_{0,C}$ where C is an arbitrary group of (typically highly correlated) variables, for the high-dimensional scenario where $p \gg n$. We address this important and open issue; note that testing the global null-hypothesis $H_{0,\{1,\dots,p\}}$, in contrast to the “partial” hypothesis $H_{0,C}$ for some C with cardinality $1 < |C| < p$, is a rather different issue and has been addressed before; see Goeman, van De Geer, and Van Houwelingen 2006. Once we have valid p-values for $H_{0,C}$ for an arbitrary groups C, we make use of the method from Meinshausen (2008) leading to nonasymptotic bounds for strong control of the familywise error rate in a hierarchical structure. For construction of the p-values, we rely on

Jacopo Mandozzi is Pension Actuary, Libera AG, Stockerstrasse 34, CH-8022 Zürich, Switzerland (E-mail: *jacopo.mandozzi@libera.ch*). Peter Bühlmann is Professor in Statistics, Department of Mathematics, ETH Zürich, CH-8092 Zürich, Switzerland (E-mail: *buhlmann@stat.math.ethz.ch*). This work is based on part of the first author's PhD dissertation. The authors thank some anonymous reviewers for constructive and insightful comments.

Journal of the American Statistical Association
March 2016, Vol. 111, No. 513, Theory and Methods
DOI: 10.1080/01621459.2015.1007209

multiple sample splitting (Meinshausen, Meier, and Bühlmann 2009). While this might be suboptimal from a theoretical perspective, especially with respect to power, the method seems to perform well in a larger empirical study for individual hypotheses $H_{0,j}$ ($j = 1, \dots, p$) in terms of reliable control of the familywise error rate in multiple testing (Dezeure et al. 2014). We also extend the Shaffer improvement in Meinshausen (2008, sec. 3.6) to the high-dimensional scenario, increasing the power of the hierarchical method such that detection of more singletons than with the method from Meinshausen, Meier, and Bühlmann (2009) becomes possible.

Our second main contribution is the development of new methodology and theory for hierarchical inference and testing of hypotheses, using multiple sample splitting techniques (and *multiple* sample splitting is important for reproducibility (Meinshausen, Meier, and Bühlmann 2009)). Regarding methodology, the hierarchical approach allows for a substantially higher number of so-called minimal true detections (MTDs: significant smallest groups of variables) than the single variable analog and has the remarkable property of adaptively selecting a best resolution level (MTDs with the smallest possible cardinality). We prove strong control of the familywise error rate of the hierarchical method under a "zonal assumption" which is weaker than the standard beta-min condition (used in Meinshausen, Meier, and Bühlmann 2009), that is, we do not require that all nonzero coefficients in the regression vector β^0 are sufficiently large. We demonstrate the finite sample behavior of the method with various empirical results.

We note that recently, Meinshausen (2014) described another procedure for dealing with highly correlated variables and hierarchical testing of groups or clusters of variables. His method is an interesting alternative with the remarkable property that it does not require (major) regularity assumptions on the design matrix. The procedure is taking advantage of the special structure of a linear model while our approach is: (i) more generic and conceptually applicable to other (e.g., generalized linear) models, and (ii) computationally much more efficient due to variable screening in a first stage.

1.2 Outline of the Article

In Section 2 we describe our method for obtaining p-values for groups of variables and its use for hierarchical testing in high-dimensional settings. We show in Section 3 that the familywise error rate (FWER) is strongly controlled, and we describe a Shaffer improvement to increase the method's power while keeping control over the FWER. Section 4 is devoted to empirical results: we show that our procedure improves the single variable testing method of Meinshausen, Meier, and Bühlmann (2009) in settings with strong correlation among certain variables, particularly with respect to minimal true detections (MTDs). In Section 5 we provide theoretical evidence that the FWER is controlled even if a "screening assumption" required in Section 3 is not satisfied.

2. DESCRIPTION OF METHOD

Our method is based on four main steps: (i) hierarchical clustering of the variables, (ii) variable screening in a linear model, (iii) significance testing (with multiplicity adjustment) based on sample splitting, and (iv) aggregation over multiple sample splits and hierarchical multiplicity adjustment. See also Section 2.5 for a schematic summary.

2.1 Clustering

In a first step, we construct a hierarchy of clusters. A hierarchy, which can be represented as a tree-graph, $\mathcal{T}$ is a set of clusters $\{C_k\}_k$ with $C_k \subseteq \{1, \dots, p\}$: the root node of the tree $\{1, \dots, p\}$ contains all variables and for any two clusters $C_k, C_{k'} \in \mathcal{T}$, either one cluster is a subset of the other, or they have an empty intersection. We use the notation pa(C) for the parent of a cluster C (the smallest superset of C), ch(C) for the children of a cluster C (all clusters that have C as parent). Cluster C is called an ancestor of cluster D if $D \subset C$.

As noted by Meinshausen (2008), the hierarchy can be derived from specific domain knowledge or in some other natural way. The philosophy of the method is that highly correlated variables (or variables which are nearly linearly dependent) should end up in a single small cluster: it will then be relatively easy to identify the cluster as relevant, if it contains at least some variables from the active set S_0. For our empirical results, we consider standard hierarchical clustering based on correlation between variables, or a novel hierarchical scheme using canonical correlations between clusters (Bühlmann et al. 2013).

Once the hierarchical structure is given, the method goes on with a hierarchical version of the multi-sample-splitting procedure from Meinshausen, Meier, and Bühlmann (2009). The following two steps, described in Sections 2.2 and 2.3, have to be repeated for each sample split, indexed by $b = 1, \dots, B$ where B is the number of sample splits (since $B > 1$, we use the terminology multi-sample-splitting).

2.2 Screening

The original data of sample size n are split into two disjoint groups, $N_{\text{in}}^{(b)}$ and $N_{\text{out}}^{(b)}$, that is, a split $\{1, \dots, n\} = N_{\text{in}}^{(b)} \cup N_{\text{out}}^{(b)}$ is chosen. The groups are chosen of equal size if n is even or satisfy $|N_{\text{out}}^{(b)}| = |N_{\text{in}}^{(b)}| + 1$ if n is odd.

Then, using only $N_{\text{in}}^{(b)}$, estimate with a screening procedure the set of active predictors $\hat{S}^{(b)}$. A prime example is the Lasso (Tibshirani 1996).

2.3 Testing and Multiplicity Adjustment

By considering for each cluster C in the hierarchy $\mathcal{T}$ its intersection with $\hat{S}^{(b)}$, an induced hierarchy with root $\hat{S}^{(b)}$ is given. Due to this construction, assuming that the cardinality $|\hat{S}^{(b)}| < n/2$, the situation is not high-dimensional anymore. Therefore, on this induced hierarchy, we can apply testing procedure similar as in Meinshausen (2008), the difference being that the hierarchical adjustment is not performed at this stage but in Section 2.4 after the aggregation over many sample splits.

Based on the other half of the sample $N_{\text{out}}^{(b)}$, we use the classical partial F-test with the full model $\hat{S}^{(b)}$ and submodel $C \cap \hat{S}^{(b)}$ for the null hypothesis $H_{0,C\cap\hat{S}^{(b)}}$, where $C \in \mathcal{T}$ is a given cluster. Thereby, we implicitly assume that the submatrix $\mathbf{X}_{\hat{S}^{(b)}}$ with columns corresponding to $\hat{S}^{(b)}$ is of full rank (since $|\hat{S}^{(b)}| < n/2$). We then assign the p-value from the partial F-test to the entire cluster C, although we have only used the variables in $C \cap \hat{S}^{(b)}$. If a cluster C does not contain selected variables from

$\hat S^{(b)}$, we set the p-value to 1. In summary, we define:

$$p^{C,(b)} = \begin{cases} p^{C\cap S^{(b)}}_{\text{partial } F\text{-test}} \text{ based on } \mathbf{Y}_{N^{(b)}_{\text{out}}}, \mathbf{X}_{N^{(b)}_{\text{out}},\hat S^{(b)}}, & \text{if } C\cap \hat S^{(b)} \neq \emptyset,\\ 1, & \text{if } C\cap \hat S^{(b)} = \emptyset. \end{cases} \qquad (2)$$

Then, for $C \in \mathcal{T}$ define the multiplicity adjusted (nonaggregated) p-value as

$$p^{C,(b)}_{\text{adj}} = \min\left(p^{C,(b)} \frac{|\hat S^{(b)}|}{|C \cap \hat S^{(b)}|},\ 1\right) \qquad (3)$$

if $C\cap \hat S^{(b)} \neq \emptyset$ and $p^{C,(b)}_{\text{adj}} = 1$ otherwise.

2.4 Aggregation and Hierarchical Adjustment

By repeating the steps in Section 2.2 and 2.3 for $b = 1, \dots, B$, we obtain for each cluster C of the hierarchy $\mathcal{T}$ a set of B p-values $p^{C,(1)}_{\text{adj}}, \dots, p^{C,(B)}_{\text{adj}}$. We aggregate these p-values by considering their empirical quantile.

For $\gamma \in (0, 1)$ define the aggregated p-values

$$Q^C(\gamma) = \min\left\{1,\ q_\gamma\left(\{p^{C,(b)}_{\text{adj}}/\gamma;\ b = 1, \dots, B\}\right)\right\},$$

where $q_\gamma(\cdot)$ is the (empirical) γ-quantile function. Finally, define the hierarchically adjusted (aggregated) p-values as

$$Q^C_h(\gamma) = \max_{D\in\mathcal{T}:C\subseteq D} Q^D(\gamma)$$

such that the hierarchically adjusted (aggregated) p-value of a cluster C is always bigger than the hierarchically adjusted (aggregated) p-value of an ancestor cluster. In Section 3 we show that for any fixed $\gamma \in (0, 1)$ the $Q^C_h(\gamma)$ are correct p-values. At this stage, γ should be considered as a prespecified parameter of the method.

Similarly as in Meinshausen, Meier, and Bühlmann (2009), error control is not guaranteed if we optimize over γ, that is, for each C we would choose the minimal $Q^C_h(\gamma)$. Nevertheless, it is possible to eliminate parameter γ by proceeding as follows. Define

$$P^C = \min\left\{1,\ (1 - \log \gamma_{\min}) \inf_{\gamma\in(\gamma_{\min},1)} Q^C(\gamma)\right\}, \qquad (4)$$

for a lower bound $\gamma_{\min} \in (0, 1)$ for γ, typically $\gamma_{\min} = 0.05$. Then proceed with the hierarchical adjustment of P^C by defining

$$P^C_h = \max_{D\in\mathcal{T}:C\subseteq D} P^D.$$

These values P^C_h are the final output of our method: we will show again in Section 3 that P^C_h are a valid p-value controlling the familywise error rate when testing over all $C \in \mathcal{T}$.

Our proposed "top-down": method is schematically summarized in Section 2.5. In the supplementary material we illustrate a subideal alternative "bottom-up" approach which is empirically found to exhibit substantially less power.

2.5 Schematic Summary of the Method

We summarize our proposed method with the following schematic description.

Step 1: Clustering

$$\{\mathbf{X}_1, \dots, \mathbf{X}_p\} \xrightarrow{\text{clustering}} \mathcal{T}$$

Repeat for $b = 1, \dots, B$.

Step 2: Screening

$$\{1, \dots, n\} = N \xrightarrow{\text{sample split}} N^{(b)}_{in} \cup N^{(b)}_{\text{out}} \xrightarrow{\text{screening}} \hat S^{(b)}.$$

Step 3: Testing and multiplicity adjustment

$$|\hat S^{(b)}| < |N^{(b)}_{\text{out}}| \xrightarrow{\text{testing}} p^{C,(b)} \xrightarrow{\text{multiplicity adjustment}} p^{C,(b)}_{\text{adj}}$$

End of repeating for $b = 1, \dots, B$.

Step 4: Aggregation and hierarchical adjustment

$$p^{C,(b)}_{\text{adj}} \xrightarrow{\text{aggregation}} Q^C(\gamma) \xrightarrow{\text{hierarchical adjustment}} Q^C_h(\gamma)$$

$$p^{C,(b)}_{\text{adj}} \xrightarrow{\text{aggregation}} Q^C(\gamma) \xrightarrow{\text{elimination of } \gamma} P^C \xrightarrow{\text{hierarchical adjustment}} P^C_h.$$

3. FAMILYWISE ERROR RATE CONTROL

We show in this section, that if the variable selection procedure $\hat S$ satisfies two assumptions, then the p-values $Q^C_h(\gamma)$ and P^C_h defined in Section 2 control the familywise error rate. The assumptions are:

(A1) *Sparsity property*: $|\hat S| < n/2$.

(A2) δ-*Screening property*: $\mathbb{P}[\hat S \supseteq S_0] \geq 1 - \delta$, where $0 < \delta < 1$.

The *sparsity property* in (A1) implies that for each sample split b it holds that $|\hat S^{(b)}| < |N^{(b)}_{\text{out}}|$, a condition which is necessary to apply classical tests. The δ-*screening property* in (A2) ensures that all the relevant variables are retained with high probability (δ is typically small). While (A1) is the same condition as in Meinshausen, Meier, and Bühlmann (2009, sec. 3.1), we consider with (A2) a slight modification of the assumption in Meinshausen, Meier, and Bühlmann (2009, sec. 3.1) to obtain nonasymptotic bounds for familywise error rate control. We provide a relaxation of the screening property (A2) in Section 5.

Example. Consider the Lasso as a variable selection method $\hat S$. Assumption (A1) holds for any value of the regularization parameter. Assumption (A2) is ensured when requiring the following conditions:

1. The design matrix $\mathbf{X}$ satisfies the compatibility condition with compatibility constant ϕ_0^2 (Bühlmann and van de Geer 2011, see Equation (6.4)). Furthermore, it is normalized such that each column $\mathbf{X}^{(j)}$ satisfies $\|\mathbf{X}^{(j)}\|_2^2/n = 1$ for all $j = 1, \dots, p$.
2. A beta-min condition holds (we use here the notation $s_0 = |S_0|$):

$$\min_{j\in S_0} |\beta^0_j| > 16\sigma \sqrt{\frac{t^2 + 2\log(p)}{n}} s_0/\phi_0^2.$$

Then, the Lasso with regularization parameter $\lambda = 4\sigma_\varepsilon\sqrt{\frac{t^2+2\log(p)}{n}}$ satisfies (A2) with $\delta = 2\exp(-t^2/2)$ (Bühlmann and van de Geer 2011, Lem. 6.2, Thm. 6.1, (2.13)).

Especially when the correlation among the variables is high (violating the compatibility condition in the example above), one can hardly expect the Lasso or any other variable selection method to satisfy (A2) for very small δ. In Section 4 we present empirical results showing that the hierarchical p-value method still works well even when the screening property is not satisfied for a small value δ, and we provide some supporting theoretical results for this fact in Section 5.

For a given hierarchy $\mathcal{T}$, denote the set of clusters that fulfill the null hypothesis by

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

Furthermore, for some fixed parameter $\gamma \in (0,1)$ and some fixed significance level $\alpha \in (0,1)$,

$$\mathcal{T}^{\gamma}_{\text{rej}} = \{C \in \mathcal{T} \,:\, Q^C_h(\gamma) \le \alpha\}$$

is the set of rejected clusters based on the p-values $Q^C_h(\gamma)$ and analogously,

$$\mathcal{T}_{\text{rej}} = \{C \in \mathcal{T} \,:\, P^C_h \le \alpha\}$$

is the set of the rejected clusters when considering the p-values P^C_h. The latter does not require to choose or prespecify a parameter like γ.

Theorem 1. Assume that (A1) and (A2) hold. Then for any significance level $\alpha \in (0,1)$ and B denoting the number of sample splits:

1. For any fixed $\gamma \in (0,1)$, the p-values $Q^C_h(\gamma)$ control the familywise error rate in the sense that
$$\mathbb{P}(\mathcal{T}^{\gamma}_{\text{rej}} \cap \mathcal{T}_0 \neq \emptyset) \le \alpha + 1 - (1-\delta)^B \le \alpha + B\delta.$$
2. The p-values P^C_h control the familywise error rate in the sense that
$$\mathbb{P}(\mathcal{T}_{\text{rej}} \cap \mathcal{T}_0 \neq \emptyset) \le \alpha + 1 - (1-\delta)^B \le \alpha + B\delta.$$

A proof is given in the supplementary material. From Theorem 1, providing nonasymptotic bounds for familywise error rate control, one can easily derive asymptotic familywise error control using the assumption

(A2′) *Screening property*: $\lim_{n\to\infty} \mathbb{P}\left[\hat{S} \supseteq S_0\right] = 1.$

Example (continued). For the Lasso, under Assumption 1, described in the example above, and replacing Assumption 2 by an asymptotic beta-min condition

$$\min_{j\in S_0} |\beta^0_j| \gg \sqrt{\frac{\log(p)}{n}} s_0/\phi^2_0, \qquad (5)$$

we have that (A2′) holds (as $n \to \infty$, $p = p_n$ and $s_0 = s_{0;n}$ and $\phi^2_0 = \phi^2_{0;n}$ are allowed to change with n).

We then have the following result.

Corollary 1. Assume that (A1) and (A2′) hold. Then for any fixed $\gamma \in (0,1)$ and significance level $\alpha \in (0,1)$:

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathcal{T}^{\gamma}_{\text{rej}} \cap \mathcal{T}_0 \neq \emptyset\right) \le \alpha$$
$$\limsup_{n\to\infty} \mathbb{P}(\mathcal{T}_{\text{rej}} \cap \mathcal{T}_0 \neq \emptyset) \le \alpha.$$

3.1 Shaffer Improvement in High-Dimensional Setting

A similar version of the Shaffer improvement as described in Meinshausen (2008, Section 2.4) can be applied to our method. The main idea, shown by Shaffer (1986), is that in a hierarchical structure, some combinations of null hypothesis can be excluded a priori, and incorporating constraints on the possible combinations of null hypotheses can increase the power of the method.

Consider a binary hierarchy $\mathcal{T}$ and a screened set $\hat{S} \subset \{1, \ldots, p\}$. The siblings of a cluster C are the children of the parent of C which are not identical to C, $\text{si}(C) = \text{ch}(\text{pa}(C))\backslash C$. Define the effective cluster size $|C|^{\hat{S}}_{\text{eff}}$ of the cluster $C \in \mathcal{T}$ restricted to the screened set $\hat{S}$ as

$$|C|^{\hat{S}}_{\text{eff}} = \begin{cases} |C \cap \hat{S}|, & \text{if } \exists\, E \in \text{ch}(\text{si}(C)) \text{ s.t. } E \cap \hat{S} \neq \emptyset \\ |C \cap \hat{S}| + |\text{si}(C) \cap \hat{S}|, & \text{otherwise.} \end{cases}$$

Note that when no screening is performed ($\hat{S} = \{1, \ldots, p\}$) this definition coincides with the definition of the effective cluster size in Meinshausen (2008). Moreover, the condition "$\exists\, E \in \text{ch}(\text{si}(C))$ s.t. $E \cap \hat{S} \neq \emptyset$" is stronger than the condition "si(C) is not a leaf node" of Meinshausen (2008) and hence, the improvement given by our definition of restricted cluster size is bigger than the one given by a straightforward adaption of Meinshausen (2008).

The Shaffer improvement in the high-dimensional setting is then given by considering the multiplicity adjustment

$$p^{C,(b)}_{\text{adj}} = \min\left(p^{C,(b)} \frac{|\hat{S}^{(b)}|}{|C|^{\hat{S}^{(b)}}_{\text{eff}}},\ 1 \right), \qquad (6)$$

instead of using the multiplicity adjustment in (3).

Obviously, since the effective cluster size is always at least as big as the cluster size, the Shaffer improvement produces smaller p-values and hence, increases the power of the method while the familywise error rate control is still guaranteed, as described next.

Theorem 2. Assume the hierarchy $\mathcal{T}$ is binary. Then, Theorem 1 and Corollary 1 still hold when using the Shaffer improvement (6) as multiplicity adjustment, assuming the conditions of Theorem 1 or Corollary 1, respectively.

A proof is given in the supplementary material. We note that an extension of the results in Theorems 1 and 2 to control the false discovery rate (Benjamini and Hochberg 1995), instead of the FWER, for *hierarchically* ordered hypotheses (with corresponding dependent p-values) seems very challenging.

4. EMPIRICAL RESULTS

In this section we study the performance of our hierarchical method and compare it with the single variable testing method of Meinshausen, Meier, and Bühlmann (2009). Section 4.3 provides most informative results about our new method, in particular for understanding the differences in comparison to the single variable approach.

In a simulation study, we consider both synthetic and semireal data. The former are used to study special designs where we expect one of the two methods to perform clearly better. The semireal data are used to obtain insights of what happens when the design matrix comes from real high-dimensional datasets. In our simulation study, all the data are generated from a linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon,$$

where $\mathbf{X}$ is a $n \times p$ matrix from synthetic (designs 1–3) or real (designs 4–7) data, β^0 is a $p \times 1$ synthetic regression vector, and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ is a synthetic noise term. The data are always standardized such that $\mathbf{X}$ has columns with empirical mean zero and variance one.

We also apply the two different methods to a real dataset in Section 4.4.

4.1 Implementation of the Method

The implementation of our hierarchical method, and also of the single variable procedure from Meinshausen, Meier, and Bühlmann (2009), requires to make some choices. We mostly consider fairly standard and "easy-to-use" methods; unless there is some deeper methodological difference, as in our choice of additionally considering a less standard clustering procedure.

For clustering we consider the recently proposed canonical correlation clustering of Bühlmann et al. (2013) and the standard hierarchical clustering (using the R function `hclust`) with distance between two covariables set as 1 less the absolute correlation between the covariables, using complete linkage (other linkages lead to similar results). For variable screening, we use the Lasso (i.e., $\hat{S}$ from the nonzero estimated coefficients from Lasso) with regularization parameter chosen by 10-fold cross-validation.

As in Meinshausen, Meier, and Bühlmann (2009), we choose $B = 50$ as the number of sample splits. For aggregation, the p-values P_h^C in (4) are computed over a grid of γ-values between $\gamma_{\min} = 0.05$ and 1 with grid-steps of size 0.025. For both hierarchical methods we use the Shaffer improvement described in Section 3.1. As nominal significance level we always consider $\alpha = 5\%$.

4.2 Simulation Study With Synthetic and SemiReal Data

We consider 42 scenarios based on 7 designs. For each design we consider six settings by varying the number of variables p in the model ($p = 200$, $p = 500$, and $p = 1000$) and the signal-to-noise ratio (SNR, for each design and choice of p we consider a low and a high SNR, namely, for $p = 200$ we use SNR $= 4$ and SNR $= 8$, for $p = 500$ we use SNR $= 8$ and SNR $= 16$, for $p = 1000$ we use SNR $= 16$ and SNR $= 32$). The signal-to-noise ratio is defined by

$$\mathrm{SNR} = \sqrt{\frac{(\beta^0)^T \mathbf{X}^T \mathbf{X} \beta^0}{n\sigma^2}}$$

and our choices of signal-to-noise ratios are avoiding scenarios where the methods have degenerate performance of 0% or 100%, respectively. In designs 1–5 the sparsity s_0 is set to be 10, while in designs 6 and 7 it is set to be 6. The nonzero components of β^0 are randomly set as $\beta^0_j = 1$ or $\beta^0_j = -1$ for $j \in S_0$. The choice of S_0 is design-specific and hence, explained with the descriptions of the designs as follows.

Design 1: equicorrelation. We set $n = 100$ and generate $\mathbf{X}$ from a centered multivariate normal distribution with equal variances $\rho_{jj} = 1$ and covariances equal to $\rho_{jk} = 0.3$ between variables j and k for $j \neq k \in \{1, \dots, p\}$. The 10 active variables are chosen randomly among the p covariables.

Design 2: high correlation within small blocks. We set $n = 100$ and generate $\mathbf{X}$ from a centered multivariate normal distribution with covariance ρ_{jk} between variables j and k set as $\rho_{jj} = 1$ for all j, $\rho_{j,j+1} = \rho_{j+1,j} = 0.9$ for $j \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$ and $\rho_{jk} = 0$ otherwise. We choose in each of the 10 two-dimensional blocks with high correlation one active variable, that is, for $j \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$ we choose randomly either $j \in S_0$ or $j + 1 \in S_0$.

Design 3: high correlation within large blocks. We set $n = 100$ and generate $\mathbf{X}$ from a centered multivariate normal distribution with a block diagonal covariance matrix with 10 $p/10$-dimensional blocks $B_{p/10}(0.9)$ defined by $(B_{p/10}(0.9))_{jj} = 1$ and $(B_{p/10}(0.9))_{jk} = 0.9$ for $j \neq k$. We randomly choose in each of these $p/10$-dimensional blocks with high correlation one active variable.

Design 4: Riboflavin dataset with normal correlation. We consider the Riboflavin dataset (Bühlmann, Kalisch, and Meier 2014) with $n = 71$ and choose randomly p (i.e., 200, 500, or 1000 depending on the setting) among 4088 covariables in the whole dataset. The six active variables are chosen randomly among the p covariables.

Design 5: Breast dataset with normal correlation. We consider the Breast dataset (van't Veer et al. 2002) with $n = 117$ and choose randomly p (i.e., 200, 500, or 1000 depending on the setting) among 24,481 covariables in the whole dataset. The 10 active variables are chosen randomly among the p covariables.

Design 6: Riboflavin dataset with high correlation. We consider again the Riboflavin dataset as in design 4, but choose p covariables as follows: a covariable is randomly chosen among all 4088 covariables in the whole dataset. Then the nine covariables with the highest absolute correlation with the first one are chosen to build an "high correlated" 10-dimensional block. Then another covariable is chosen among the remaining 4078 covariables of the whole dataset and another "high correlated" 10-dimensional block is analogously built. We repeat this procedure until we have p covariables. The six active variables are chosen randomly among the set $\{j; j = 10k + 1, 0 \le k \le p/10 - 1\}$.

Design 7: Breast dataset with high correlation. We consider again the Breast dataset as in design 5, but choose p covariables as follows: a covariable is randomly chosen among all 24,481 covariables in the whole dataset. Then the nine covariables with the highest absolute correlation with the first one are chosen to build an "high correlated" 10-dimensional block. Then another covariable is chosen among the remaining 24,471 covariables of the whole dataset and another "high correlated" 10-dimensional block is analogously built. We repeat this procedure until we have p covariables. The 10 active variables are chosen randomly among the set $\{j; j = 10k + 1, 0 \leq k \leq p/10 - 1\}$.

4.2.1 Performance Measures for Simulation Study. Besides the familywise error rate, we consider, among other aspects, the following one-dimensional statistics measuring power (while Section 4.3 provides a more informative picture by avoiding to compress to one-dimensional performance measures).

We use two different performance functions. The first one is defined as

$$\text{Performance 1} = \frac{1}{|S_0|} \sum_{\text{MTD } C} \frac{1}{|C|}, \qquad (7)$$

where the sum is over all minimal true detections (which we denote by "MTD"). Thereby,

A cluster is said to be a MTD if it satisfies all of the following:

- C is a significant cluster, for example, has p-value $<5\%$. ("Detection")
- There is no significant subcluster $D \subset C$. ("Minimal")
- $C \notin \mathcal{T}_0$, that is, there is at least one active variable in C. ("True")

The Performance 1 is always between 0 and 1, and it is exactly 1 when each active variable is selected as a singleton. Moreover, the contribution to the Performance 1 of MTD C is independent of the number of active variables that are in C. Although this penalizes our new method, it reflects the fact that from $P_h^C < 5\%$ one can only conclude that there is at least one active variable in C without having further information whether there are additional active variables in C and which of the variables in C are active.

As second performance function we consider a slightly modified version of the Performance 1, where only MTDs with cardinality $|C| \leq 20$ are considered, and a "bonus" is given for each MTD, independently from its cardinality (if the latter is at most 20):

$$\text{Performance 2} = \frac{1}{|S_0|} \sum_{\text{MTD C with } |C| \leq 20} \frac{1}{2}\left(\frac{1}{|C|} + 1\right). \qquad (8)$$

The Performance 2 is also always between 0 and 1, and it is again exactly equal to 1 if each active variable is selected as a singleton.

Moreover, for the single variable method, both performance measures are the same as only singletons can be selected. Correct selection of a cluster with more than one variable is less valuable than a singleton, with both performance measures: Performance 2, however, is putting less emphasis on the size of a selected cluster. The choice of the bound for the cluster being at most 20 in Performance 2 is motivated by the idea that too large clusters are "uninteresting" in many practical applications (e.g., a genetic pathway consists of about up to 20 genes, and a cluster would represent a pathway).

4.2.2 Familywise Error Rate Control (FWER). For each of the 42 scenarios described in Section 4.2 we make 100 independent simulation runs varying only the synthetic noise term ε and count the number where at least one false selection is made (i.e., there exists a cluster $C \in \mathcal{T}_0 \cap \mathcal{T}_{\text{rej}}$). According to Theorem 2 we expect this number to be at most $100\alpha = 5$ ($\alpha = 0.05$).

Table 1. Familywise error rate in %: Number of cases with at least one false selection, out of 100 simulation runs.

		Familywise error rate (in %)					
		Low SNR			High SNR		
Design	p	Single	Cancorr	Hclus	Single	Cancorr	Hclus
	200	0	0	0	0	0	0
Equi	500	0	0	0	0	0	0
Corr	1000	0	0	0	0	0	0
High corr	200	1	1	1	0	0	0
Within small	500	7	7	7	0	0	0
Blocks	1000	5	5	5	3	3	3
High corr	200	0	7	6	0	0	0
Within large	500	0	0	0	0	0	0
Blocks	1000	0	0	0	0	0	0
Riboflavin	200	0	0	0	0	0	0
Normal	500	0	0	0	0	0	0
Corr	1000	0	0	0	0	0	0
Breast	200	0	0	0	0	0	0
Normal	500	0	0	0	0	0	0
Corr	1000	0	0	0	0	0	0
Riboflavin	200	0	0	0	0	0	0
High	500	0	0	0	0	0	0
Corr	1000	0	0	0	0	0	0
Breast	200	0	0	0	0	0	0
High	500	0	0	0	0	0	0
Corr	1000	0	0	0	0	0	0

NOTE: The scenarios where the critical value of 5 is overtaken are marked in gray

The results illustrated in Table 1 show that for 40 of the 42 scenarios, FWER control holds for all methods, while in one scenario it doesn't hold for any method and in one scenario it doesn't hold for the hierarchical methods. In 37 out of the 42 scenarios there is no false selection at all. It is not surprising that the most problematic designs with respect to FWER are the "high correlation within small blocks"—and "high correlation within large blocks"—designs, since there each active predictor is highly correlated with false variables from S_0^c and hence, it is rather difficult for our screening method (the Lasso) to guarantee $\hat{S} \supseteq S_0$.

4.2.3 Power: Performance 1. For each of the 42 scenarios described in Section 4.2 we make 100 simulation runs varying the synthetic noise term ε and the synthetic regression vector β^0. We then calculate the average Performance 1 in (7), that is, Performance 1 is averaged over 100 simulation runs.

The results are reported in Table 2. They show that, as expected, the hierarchical methods provide better results in the designs where the correlation among the variables is rather high (designs 2, 3, 6, and 7), while in the other designs the nonhierarchical method has in general a slightly better performance.

Table 2. Performance 1, averaged over 100 simulation runs, for the methods “single variable,” “hierarchical with canonical correlation clustering” and “hierarchical with `hclust` clustering.”

		Performance 1 in %					
		Low SNR			High SNR		
Design	p	Single	Cancorr	Hclus	Single	Cancorr	Hclus
	200	47.9	47.1	44.2	97.7	97.6	97.4
Equi	500	39.7	39.1	37.7	72.8	72.8	72.0
Corr	1000	17.7	17.6	17.4	28.1	27.8	27.9
High corr	200	44.5	54.0	53.7	91.8	96.3	96.3
Within small	500	31.5	37.1	38.3	69.9	73.3	73.5
Blocks	1000	13.1	15.0	15.1	22.6	24.2	24.3
High corr	200	2.9	6.5	6.7	21.7	27.5	27.9
Within large	500	0.6	0.7	1.1	11.7	11.9	12.6
Blocks	1000	0.0	0.0	0.1	4.3	4.3	4.4
Riboflavin	200	23.5	23.3	23.4	56.5	56.3	56.3
Normal	500	15.0	14.7	13.5	37.3	37.5	36.4
Corr	1000	12.0	11.5	10.8	16.3	16.4	16.2
Breast	200	40.5	40.5	39.5	86.5	86.5	86.4
Normal	500	39.5	39.6	38.8	68.1	68.0	67.6
Corr	1000	33.0	32.8	31.2	39.5	39.3	38.1
Riboflavin	200	24.0	24.3	26.0	64.3	65.0	66.0
High	500	28.7	29.2	29.9	61.5	61.1	61.8
Corr	1000	25.2	25.3	25.3	41.3	41.1	40.7
Breast	200	39.8	39.8	41.2	90.9	91.0	91.9
High	500	51.3	51.2	49.9	77.5	77.3	77.8
Corr	1000	47.3	47.3	47.0	58.5	58.4	57.5
Avg. normal corr.		26.0	25.8	25.4	52.9	52.9	52.7
Avg. high corr.		28.7	30.4	30.2	53.6	54.8	54.9
Average		27.5	28.4	28.1	53.3	54.0	54.0

The best and second best methods are marked in dark-gray and light-gray. The average performances in the bottom rows are averages over the corresponding or all scenarios, respectively

The method based on the `hclust` clustering is more sensitive with respect to high correlation among the variables than the analog based on canonical correlation clustering. In particular, the method based on the `hclust` clustering is best in 19 of the 24 scenarios which use designs 2, 3, 6, and 7 while it is the worst method in 15 of the 18 scenarios where the correlation is not particularly high. We note that the differences among the methods are rather small: this is mainly a consequence of our definition (7) of the Performance 1 and p being large. The biggest (absolute) difference in the Performance 1 can be found in design 2 where the hierarchical methods have a performance up to 9.5% higher than the single variable method, while the biggest deficit of a hierarchical method with respect to the single variable method can be found in design 1 and amounts to 3.7%. As expected, our results show that in general the Performance 1 (and also the differences between them when considering the different methods) lowers when p increases. Finally, it is interesting to note that in the scenarios that favor the single variable method, the Performance 1 of the method with canonical correlation clustering is very close to the Performance 1 of the single variable method (the difference is at most 0.8%), while in the other scenarios it might perform much better (differences of up to 9.5%).

4.2.4 Power: Performance 2. In Table 3 we show the average Performance 2 of the three considered methods for the 42 different scenarios, that is, for each scenario, Performance 2 is averaged over 100 simulation runs. While by definition, Performances 1 and 2 are the same for the single variable method, we find for both hierarchical methods that the Performance 2 is generally higher than the Performance 1 (only in one out of 84 cases it is lower and the difference is just 0.1%). This was expected as the idea of Performance 2 is to give a little extra reward to each correct selection, independently of the cardinality of the selected cluster (given the latter is at most 20). In particular, the method that benefits most from Performance 2 is the hierarchical method with `hclust` clustering which has an average Performance 2 of 43.9% while its average Performance 1 is 41.0% (average is meant over all scenarios). We also note that for Performance 2, the difference between the single variable and the hierarchical methods in the settings with high correlation is more evident.

Table 3. Performance 2, averaged over 100 simulation runs, for the methods “single variable,” “hierarchical with canonical correlation clustering” and “hierarchical with `hclust` clustering.”

		Performance 2 in %					
		Low SNR			High SNR		
Design	p	Single	Cancorr	Hclus	Single	Cancorr	Hclus
	200	47.9	47.3	46.1	97.7	97.7	97.4
Equi	500	39.7	39.2	38.3	72.8	73.0	72.1
Corr	1000	17.7	17.7	17.6	28.1	27.8	28.1
High corr	200	44.5	59.3	60.7	91.8	98.1	98.1
Within small	500	31.5	39.6	44.2	69.9	75.0	76.4
Blocks	1000	13.1	15.7	17.2	22.6	25.2	26.3
High corr	200	2.9	31.1	31.3	21.7	61.3	61.4
Within large	500	0.6	0.6	1.1	11.7	12.0	13.2
Blocks	1000	0.0	0.0	0.1	4.3	4.3	4.4
Riboflavin	200	23.5	23.3	25.2	56.5	56.7	58.7
Normal	500	15.0	14.7	14.2	37.3	37.7	36.9
Corr	1000	12.0	11.5	11.0	16.3	16.5	16.3
Breast	200	40.5	40.5	41.6	86.5	86.6	87.2
Normal	500	39.5	39.6	39.6	68.1	68.0	68.3
Corr	1000	33.0	32.8	31.7	39.5	39.3	38.4
Riboflavin	200	24.0	24.8	31.8	64.3	66.1	69.4
High	500	28.7	29.9	32.8	61.5	61.4	63.7
Corr	1000	25.2	25.3	25.7	41.3	41.1	41.1
Breast	200	39.8	40.0	46.1	90.9	91.4	93.8
High	500	51.3	51.2	51.7	77.5	77.3	78.9
Corr	1000	47.3	47.7	48.3	58.5	58.5	58.1
Avg. normal corr.		26.0	26.0	27.0	52.9	53.1	53.7
Avg. high corr.		28.7	33.2	34.5	53.6	58.1	58.7
Average		27.5	30.1	31.3	53.3	55.9	56.6

The best and second best methods are marked in dark-gray and light-gray. The average performances in the bottom rows are averages over the corresponding or all scenarios, respectively

Some additional results regarding the variability of both Performance 1 and Performance 2 measures among the 100 different simulation runs are given in the supplementary material.

4.3 A More Detailed Consideration

The power results in the previous section are given in terms of one-dimensional performance functions. Here, we provide more information on what our new method actually does and how it performs when looking beyond one-dimensional summary statistics. On the other hand, to keep the exposition at reasonable length, we focus on fewer simulation scenarios only.

We consider the “high correlation within small blocks”- and “high correlation within large blocks”-designs (designs 2 and 3 of Section 4.2) with $p = 200$, $\text{SNR} = 8$, $s_0 = 10$ with the nonzero components of β^0 randomly set as $\beta^0_j = \pm 1$ and various values for the nontrivial covariances: $\rho \in \{0, 0.4, 0.7, 0.8, 0.85, 0.9, 0.95, 0.99\}$. For each of the 16 scenarios we make 100 simulation runs varying the synthetic noise term ε. As results we consider the FWER (portion of runs with at least a false detection over all 100 runs) and, averaged over the 100 runs, the number of MTDs and the number of MTDs of some given cardinality. The results are as shown in Table 4.

FWER control (with nominal level $\alpha = 5\%$) holds for most settings, even for relatively high values of the fraction of failed screenings with $\hat{S} \not\supseteq S_0$ (represented by δ in Theorem 1). This indicates a robustness property of the methods in controlling FWER, beyond the results of Theorem 1 which requires that δ is very small.

Looking at the number of MTDs in Table 4, we see that the hierarchical method dominates the single variable method, with its superiority increasing with increasing correlation among the variables. Considering only the singleton detections (MTDs with cardinality 1), there are 5 scenarios out of 16 where at least one of the two hierarchical methods is worse than the single variable method, while in all other scenarios both hierarchical methods detect at least the same number of singletons as the single variable method. In general, the difference in the number of singleton detections of the considered methods is rather small.

Table 4. Results of the simulation with the "high correlation within small blocks"- and "high correlation within large blocks"-design with high SNR (SNR = 8) for different correlations. ρ is the correlation in the design, δ the relative frequency of screenings with $\hat{S} \not\supseteq S_0$, MTD denotes "minimal true detections," "$3 \leq |\cdot| \leq 10$" indicates that MTD of cardinality between 3 and 10 are considered, S, C, and H represent the "single variable," respectively, "canonical correlation clustering" and "hierarchical with hclust clustering" method

								# MTD for given cardinality								
		FWER			# MTD			$\lvert\cdot\rvert = 1$			$\lvert\cdot\rvert = 2$		$3 \leq \lvert\cdot\rvert \leq 10$		$11 \leq \lvert\cdot\rvert \leq 20$	
ρ	δ	S	C	H	S	C	H	S	C	H	C	H	C	H	C	H
							"High correlation within small blocks"-design with high SNR									
0	0.45	0	0	0	9.87	9.87	9.90	9.87	9.86	9.86	0	0.01	0	0.01	0	0
0.4	0.30	0	0	0	10	10	10	10	10	10	0	0	0	0	0	0
0.7	0.06	0	0	0	10	10	10	10	10	10	0	0	0	0	0	0
0.8	0.21	0	0	0	9.85	9.98	9.98	9.85	9.90	9.90	0.08	0.08	0	0	0	0
0.85	0.26	0	0	0	9.26	9.89	9.89	9.26	9.39	9.39	0.48	0.47	0.01	0.03	0.01	0
0.9	0.17	0	0	0	9.59	10	10	9.59	9.67	9.67	0.33	0.33	0	0	0	0
0.95	0.42	0.21	0.21	0.21	8.36	9.81	9.82	8.36	8.36	8.36	1.45	1.45	0	0.01	0	0
0.99	0.89	0.92	0.92	0.92	6.72	8.06	8.06	6.72	6.73	6.73	1.33	1.33	0	0	0	0
							"High correlation within large blocks"-design with high SNR									
0	0.20	0	0	0	10	10	10	10	10	10	0	0	0	0	0	0
0.4	0.01	0	0	0	9.98	9.98	10	9.98	9.98	9.99	0	0	0	0.01	0	0
0.7	0.22	0	0	0	5.12	6.85	9.60	5.12	5.12	5.10	0.15	0.18	0.67	0.68	0.87	3.54
0.8	0.04	0	0	0	9.23	9.93	10	9.23	9.17	9.14	0	0.03	0.51	0.48	0.25	0.35
0.85	0.21	0	0	0	3.86	9.98	9.98	3.86	3.82	3.84	0.04	0.10	0.85	1.16	5.27	4.88
0.9	0.75	0	0	0	0.06	6.62	7.17	0.06	0.06	0.06	0	0.02	0.20	0.23	6.10	6.05
0.95	0.63	0	0.05	0	1.26	9.94	9.99	1.26	1.25	1.27	0	0.20	0.77	1.61	7.92	6.91
0.99	0.99	0.33	0.88	0.99	3.26	8.55	7.92	3.26	3.26	3.26	0.14	0.18	2.80	2.70	2.35	1.78

We note that the hierarchical method can detect more singletons than the single variable method because the Shaffer improvement of Section 3.1 allows for better multiplicity adjustment. For the "high correlation within small blocks" scenarios with $\rho = 0.4$ and $\rho = 0.7$ and the "high correlation within large blocks" scenario with $\rho = 0$ all three methods exhibit a perfect accuracy.

It is interesting to note that for the "high correlation within small blocks"-designs, where the improvement given by the hierarchical over the single variable method is smaller than for the "high correlation within large blocks"-designs, the quality of the improvement should be considered as very high since almost all additional discoveries by the hierarchical method have cardinality 2 only and often sum up to essentially all possible discoveries. Further results regarding the MTDs for 4 (out of the 16 considered) scenarios are given in the supplementary material.

For additional illustration, we show in Figure 1 the dendrograms (in gray) for a representative simulation run of the "high correlation within large blocks"-design with $\rho = 0.85$, for the single variable method and the hierarchical method with hclust clustering. The active variables are labeled in black and truly detected nonzero variables along the hierarchy are depicted in black. While the single variable method "only" detects five singletons, the hierarchical method detects the same five singletons and achieves five more MTDs (three of which have cardinalities 5 or less and hence, are particularly informative). Figure 2 is analogous to Figure 1 for a simulation run of the "high correlation within small blocks"-design with $\rho = 0.8$. It shows that the hierarchical method improves the results of the single variable method (nine detected singletons) by additionally providing one MTD of cardinality 2 besides the same nine singletons of the single variable method. Thus, we provide evidence of the fact that the hierarchical method has the powerful advantage of automatically going to the finer possible resolution, depending on signal-strength and correlation structure among the variables.

Finally, we illustrate in Figure 3 the true positive (TPR) rates and false positive rates (FPR) of the Lasso, the single variable method and the hierarchical method with hclust clustering as points in the ROC space.

We note that, as expected from the philosophy of the single variable and hierarchical methods to control the FWER, there is a substantial difference between the FPR of the Lasso (0.15–0.18) and the FPR of the other two methods which is always less than 0.03 and equals 0 in most of the cases. For the "high correlation within small blocks"-design, this improvement of the FPR has no negative impact on the TPR, while for the more difficult "high correlation within large blocks"-design, a TPR comparable to that of the Lasso can only be achieved by the hierarchical method which significantly improves the TPR of the single variable method. It has to be remarked that the TPR and FPR are based on MTDs (regardless from their cardinality), hence, some care is needed when comparing the TPR and FPR of the hierarchical with those of the other methods where only singleton detections are possible. For a detailed analysis, we refer to Table 4.

In the supplementary material we present the same detailed analysis as in this section considering the same designs but with low SNR $= 4$ signal-to-noise ratio.

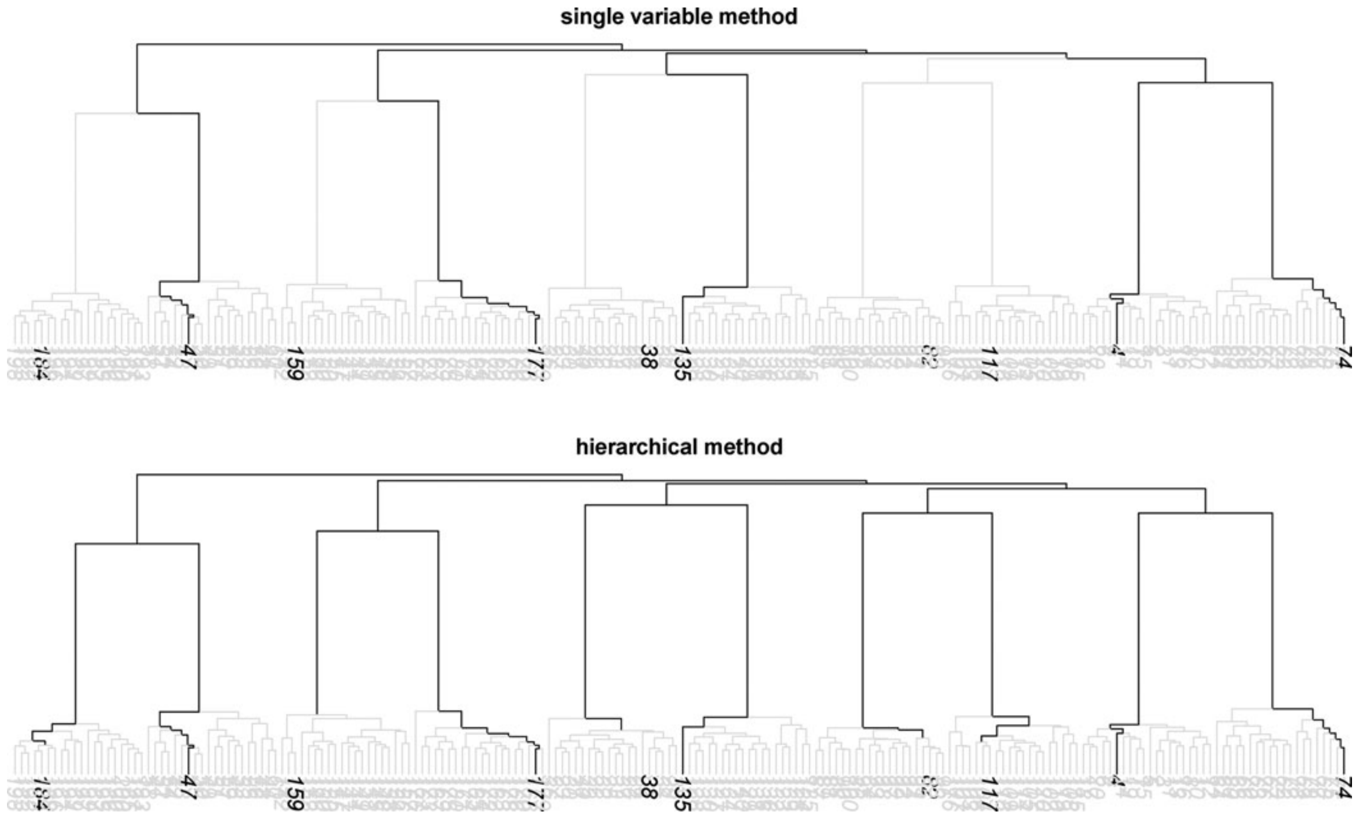

Figure 1. Dendrograms for a representative run of the "large blocks"-design with high SNR (SNR = 8) and $\rho = 0.85$. The active variables are labeled in black and the truly detected nonzero variables along the hierarchy are depicted in black.

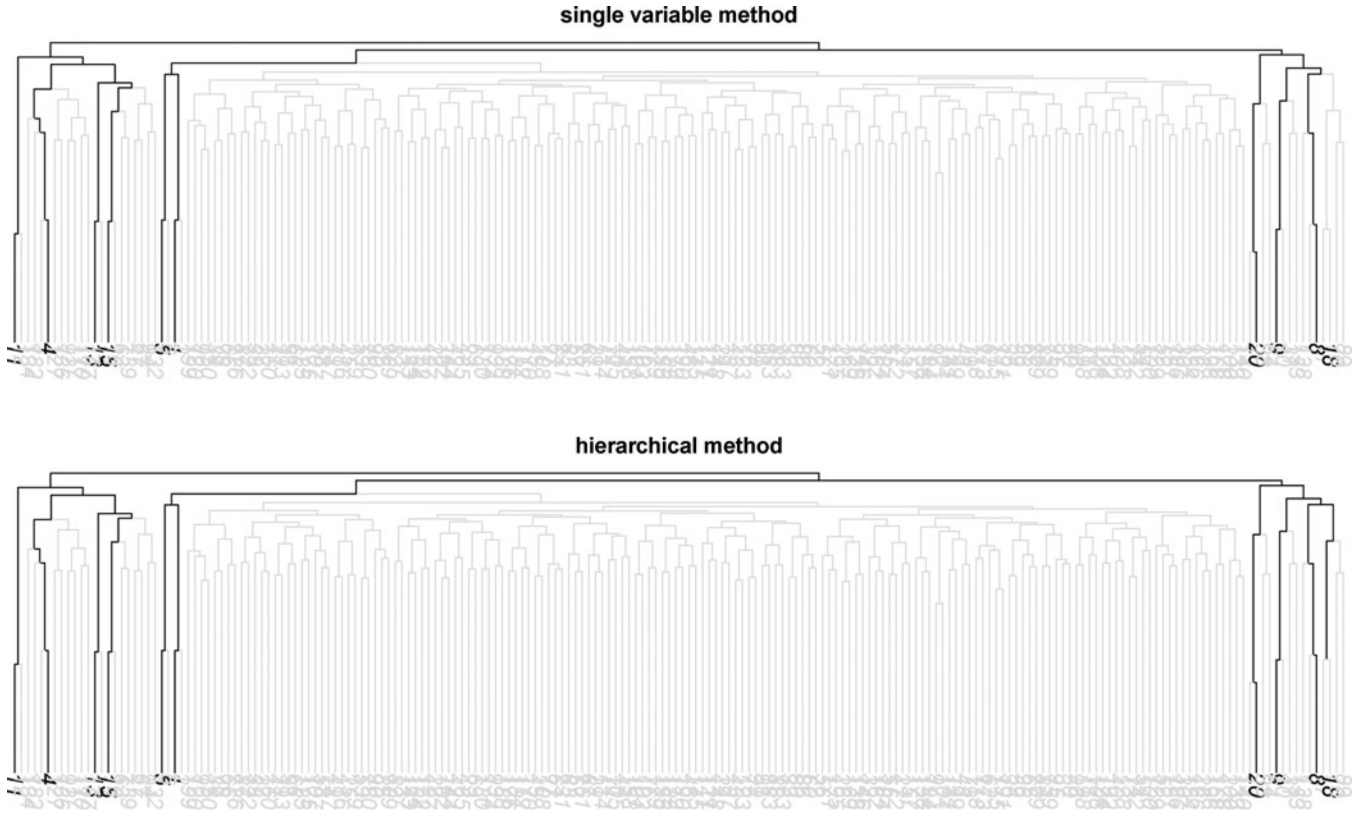

Figure 2. Dendrograms for a representative run of the "high correlation within small blocks"-design with high SNR (SNR = 8) and $\rho = 0.8$. The active variables are labeled in black and the truly detected nonzero variables along the hierarchy are depicted in black.

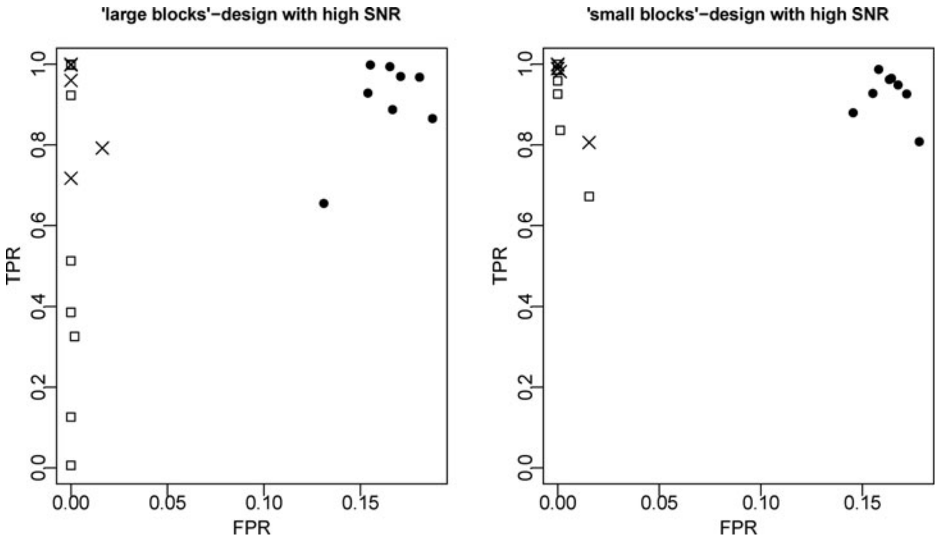

Figure 3. True positive rate (TPR) and false positive rate (FPR) for the Lasso (bullet), the single variable method (box), and the hierarchical method with `hclust` clustering (cross) for different scenarios as indicated in the header of the plots.

4.4 Real Data Application: Motif Regression

We apply the three methods described in Section 4.1 to a real dataset about motif regression (Conlon et al. 2003) with $n = 287$ and $p = 195$, used in Meinshausen, Meier, and Bühlmann (2009, sec. 4.3). The single variable method identifies one single predictor variable as significant (controlling the familywise error rate at 5%). The same variable is found to be significant with the hierarchical method with `hclust` clustering, while the hierarchical method with canonical correlation clustering identifies as significant clusters, in the sense of Section 4.2.1, the singleton, which is the same single predictor as found by the other two methods, and a very big cluster of 165 variables. This is an interesting finding saying that besides the single predictor variable, there are presumably other motifs, in the large cluster, which play a relevant role. However, there is not enough information to determine which of the variables in the large cluster are significant as a single motif.

4.5 Conclusions From the Empirical Results

We have studied error rate control and performance of the three methods over 42 scenarios. The familywise error rate control was respected for all methods in 40 out of 42 scenarios, for 2 scenarios it is slightly nonrespected (7 or less runs with at least a false selection out of 100). Considering Performance 1, we can see that the single variable method performs slightly better for settings where the correlation is not particularly high and the hierarchical methods perform better for settings with high correlation. If one looks at Performance 2, the disadvantage of the hierarchical methods in the "normal correlation" settings gets smaller, while their advantage in the "high correlation" settings gets more substantial, with an average (over all "high correlation" scenarios) improvement of 4.5% when considering canonical correlation clustering and 5.5% when considering `hclust` clustering.

Taking a more detailed and informative viewpoint in Section 4.3, the hierarchical method dominates the single variable method in terms of minimal true detections (MTDs), while both methods detect a similar number of singletons (the hierarchical method being slightly preferable in this aspect, too). While both methods exhibit a good performance for the scenario generated with $\rho = 0$, the clear superiority of the hierarchical method becomes apparent for increasing values of the correlations among the variables. The empirical findings supporting this statement are supported with additional results presented in the supplementary material.

Applying the hierarchical methods to a real dataset about motif regression (Conlon et al. 2003), we obtained an indication that there might be other potential motifs in a large cluster of size 165 which could play a significant role.

5. ROBUSTNESS OF THE METHOD WITH RESPECT TO FAILURE OF VARIABLE SCREENING

The variable screening assumption (A2) seems far from necessary for controlling the FWER as described in Theorem 1. Table 4 provides empirical support for this fact.

5.1 A Heuristic Explanation

The following argument yields some explanation why the screening property is a too restrictive assumption. Let us assume that the screening property fails because the beta-min condition (5) fails to hold. We then expect rather different selected sets $\hat{S}^{(1)}, \ldots, \hat{S}^{(B)}$, and the resulting p-values $p_{\mathrm{adj}}^{C,(1)}, \ldots, p_{\mathrm{adj}}^{C,(B)}$ based on these selected sets are likely to be rather different as well (since $\hat{S}^{(b)} \not\supseteq S_0$ for most of the b's): many of them would not exhibit a small value and thus, when aggregating these p-values, the resulting aggregated p-value is likely to be nonsmall. For example, when aggregating with the sample median ($\gamma = 1/2$ in Section 2.4), more than 50% of the p-values would need to be small such that the aggregated value would be small as well; and thus, the method only makes rejections if the single p-values $p_{\mathrm{adj}}^{C,(1)}, \ldots, p_{\mathrm{adj}}^{C,(B)}$ are stable and a substantial fraction of them are small (and hence, we expect conservative behavior with respect to FWER control). We note that failure of (A2) due to a different reason than failure of the beta-min condition (5), such as ill-posed correlations among the variables, might lead to stable p-values where a large fraction of them are spuriously small: and in such a circumstance, the method might perform poorly with respect to controlling the FWER.

5.2 A Mathematical Argument Based on Zonal Assumptions

We rigorously argue here that failure of the beta-min condition (5) still leads to control of the FWER, assuming alternative and weaker zonal assumptions (Bühlmann and Mandozzi 2014).

We partition the active set S_0 into sets with corresponding large and small regression coefficients, respectively,

$$\begin{aligned} S_0 &= S_{0,\text{large}}(a) \cup S_{0,\text{small}}(u), \\ S_{0,\text{large}}(a) &= \{j; |\beta_j^0| > a\}, \quad S_{0,\text{small}}(u) = \{j; |\beta_j^0| \le u\}, \end{aligned}$$

where $0 < u < a$.

Consider the model (1) with noise vector $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. It can be rewritten as

$$\mathbf{Y} = \mathbf{X}^{\hat{S}} \beta^0_{\hat{S}} + \mathbf{X}^{\hat{S}^c} \beta^0_{\hat{S}^c} + \varepsilon,$$

where $\hat{S} = \hat{S}(I_1) \subseteq \{1 \dots p\}$, $|\hat{S}| \le |I_2|$ and $\mathbf{X}^{\hat{S}}$ the design submatrix of $\mathbf{X}$ with columns corresponding to $\hat{S}$, and I_1, I_2 denote the two subsamples such that $I_1 \cup I_2 = \{1, \dots n\}$. Assume for the $|I_2| \times |\hat{S}|$ design submatrix $\mathbf{X}^{\hat{S}}_{I_2}$ of $\mathbf{X}$ with rows corresponding to I_2 and columns corresponding to $\hat{S}$:

$$\text{rank}((\mathbf{X}^{\hat{S}}_{I_2})^T \mathbf{X}^{\hat{S}}_{I_2}) = |\hat{S}|. \tag{9}$$

Then define the following least-square estimates based on the subsample I_2 and using only the variables from $\hat{S}$:

$$\begin{aligned} \hat{\beta}^{\hat{S}}_{I_2} &= \big((\mathbf{X}^{\hat{S}}_{I_2})^T \mathbf{X}^{\hat{S}}_{I_2}\big)^{-1} (\mathbf{X}^{\hat{S}}_{I_2})^T Y_{I_2}, \\ P^{\hat{S}}_{I_2} &= \mathbf{X}^{\hat{S}}_{I_2} \big((\mathbf{X}^{\hat{S}}_{I_2})^T \mathbf{X}^{\hat{S}}_{I_2}\big)^{-1} (\mathbf{X}^{\hat{S}}_{I_2})^T, \quad Q^{\hat{S}}_{I_2} = I_{I_2} - P^{\hat{S}}_{I_2}, \\ \hat{Y}^{\hat{S}}_{I_2} &= P^{\hat{S}}_{I_2} Y_{I_2} = \mathbf{X}^{\hat{S}}_{I_2} \hat{\beta}^{\hat{S}}_{I_2}, \quad \hat{\varepsilon}^{\hat{S}}_{I_2} = Q^{\hat{S}}_{I_2} Y_{I_2} = Y_{I_2} - \hat{Y}^{\hat{S}}_{I_2}, \\ (\hat{\sigma}^{\hat{S}}_{I_2})^2 &= \frac{\|\hat{\varepsilon}^{\hat{S}}_{I_2}\|_2^2}{|I_2| - |\hat{S}|}. \end{aligned}$$

Theorem 3. Consider any selector $\hat{S}$ which is based on the subsample I_1 and satisfies (9). Then, for a $q \times |\hat{S}|$-matrix A,

$$\frac{\left(A\hat{\beta}^{\hat{S}}_{I_2} - A\beta^0_{\hat{S}}\right)^T \left(A \left(\mathbf{X}^{\hat{S},T}_{I_2} \mathbf{X}^{\hat{S}}_{I_2}\right)^{-1} A^T\right)^{-1} (A\hat{\beta}^{\hat{S}}_{I_2} - A\beta^0_{\hat{S}})}{q(\hat{\sigma}^{\hat{S}}_{I_2})^2} \sim F_{q,|I_2|-|\hat{S}|}\,(\lambda_{\text{noncentral}})$$

is noncentral F-distributed with noncentrality parameter

$$\begin{aligned} \lambda_{\text{noncentral}} &= \sum_{i=1}^{q} (\text{BIAS})_i^2, \\ \text{BIAS} &= \frac{1}{\sigma} \left(A \left(\left(\mathbf{X}^{\hat{S}}_{I_2}\right)^T \mathbf{X}^{\hat{S}}_{I_2}\right)^{-1} A^T\right)^{-1/2} \\ &\quad \times A \left(\left(\mathbf{X}^{\hat{S}}_{I_2}\right)^T \mathbf{X}^{\hat{S}}_{I_2}\right)^{-1} \left(\mathbf{X}^{\hat{S}}_{I_2}\right)^T \mathbf{X}^{\hat{S}^c}_{I_2} \beta^0_{\hat{S}^c}. \end{aligned}$$

A proof is given in the supplementary material. Theorem 3 gives the distribution of the partial F-test statistic in the general case where a failure of screening is possible. The noncentrality parameter $\lambda_{\text{noncentral}}$, however, is unknown in practice. Clearly, if $\hat{S} \supseteq S_0$, then $\beta^0_{\hat{S}^c} = 0$ and the noncentrality parameter $\lambda_{\text{noncentral}} = 0$. Thus, if $\hat{S}$ is approximately correct for screening S_0, then $\lambda_{\text{noncentral}} \approx 0$.

In the following example we show that considering the Lasso as screening procedure and assuming zonal assumptions on the active variables, Theorem 3 implies asymptotically valid p-values when taking a partial F-test with central F-distribution (i.e., the noncentrality parameter is asymptotically negligible).

5.3 The Lasso as Selector $\hat{S}$ and Zonal Assumptions for β^0

For the Lasso, assuming that the compatibility condition holds with compatibility constant $\phi_0^2 > 0$ (Bühlmann and van de Geer 2011, see Equation (6.4)), with probability tending to one:

$$\begin{aligned} \|\hat{\beta} - \beta^0\|_\infty &\le \|\hat{\beta} - \beta^0\|_1 \le a\,(n, p, s_0, \mathbf{X}, \sigma) \\ &:= C\sigma s_0 \sqrt{\log(p)/n}/\phi_0^2 \end{aligned}$$

for some $C = C(\lambda) > 0$ when choosing the regularization parameter $\lambda \asymp \sigma\sqrt{\log(p)/n}$ (Bühlmann and van de Geer 2011, Thm. 6.1). Hence, on an event with high probability, we have for this $a = a(n, p, s_0, \mathbf{X}, \sigma)$,

$$\hat{S} \supseteq S_{0,\text{large}}(a)$$

(Bühlmann and Mandozzi 2014) and using the partitioning of S_0 it follows that

$$\|\beta^0_{\hat{S}^c}\|_\infty \le u \text{ and } \|\beta^0_{\hat{S}^c}\|_0 \le s_{0,\text{small}}(u).$$

Assuming constants C_1, C_2, and C_3 such that

$$\begin{aligned} \max_{j=1,\dots,p} \left(\mathbf{X}^T_{I_2} \mathbf{X}_{I_2}\right)_{jj} &\le C_1 |I_2| \\ \max_{j,k \in \hat{S}} | \left(\mathbf{X}^{\hat{S},T}_{I_2} \mathbf{X}^{\hat{S}}_{I_2}\right)^{-1} |_{jk} &\le C_2 |I_2|^{-1} \\ \max_{j,k=1,\dots,q} | \left(A \left(\mathbf{X}^{\hat{S},T}_{I_2} \mathbf{X}^{\hat{S}}_{I_2}\right)^{-1} A^T\right)^{-1/2} |_{jk} &\le C_3 |I_2|^{1/2} \\ \text{for each } q \times |\hat{S}|\text{-matrix } A \text{ with } q < |\hat{S}|,\ A_{jj} &\in \{0, 1\} \text{ and} \\ A_{jk} = 0 \text{ for } j \ne k&. \end{aligned} \tag{10}$$

It follows for the noncentrality parameter

$$\begin{aligned} \lambda_{\text{noncentral}} &\le q \max_{i=1,\dots,q} (\text{BIAS})_i^2 \\ &\le q \left(\frac{1}{\sigma} C_3 |I_2|^{1/2} |\hat{S}| C_2 |I_2|^{-1} |\hat{S}| C_1 |I_2| s_{0,\text{small}}(u) u\right)^2 \\ &\le \left(\frac{C_1 C_2 C_3}{\sigma} |\hat{S}|^{5/2} |I_2|^{1/2} s_{0,\text{small}}(u) u\right)^2. \end{aligned}$$

Now, assuming a more restrictive sparse eigenvalue condition on the design $\mathbf{X}$ we have $|\hat{S}| \le C_4 s_0$ for some constant $0 < C_4 < \infty$ (Zhang and Huang 2008; van de Geer, Bühlmann, and Zhou 2011) and hence, for some constant $D = D(C_1, C_2, C_3, C_4)$

$$\lambda_{\text{noncentral}} \le \left(\frac{D}{\sigma} s_0^{5/2} s_{0,\text{small}}(u) \sqrt{n} u\right)^2,$$

that is, the noncentrality parameter is negligible for u being at most of small order $o(n^{-1/2})$. Note that the inequality above is implicit in the value u since it involves $s_{0,\text{small}}(u)$: of course, we can give the upper bound

$$\lambda_{\text{noncentral}} \le \left(\frac{D}{\sigma} s_0^{7/2} \sqrt{n} u\right)^2,$$

implying that $u = o(s_0^{-7/2} n^{-1/2})$ suffices to obtain asymptotic negligibility of the noncentrality parameter.

We conclude as follows. Assume that (9) and (10) hold and that the design matrix satisfies a sparse eigenvalue condition with sparse eigenvalue bounded away from zero. Furthermore, replace the screening property in (A2) by zonal assumptions for the regression coefficients:

$$\begin{aligned} S_0 &= S_{0,\mathrm{large}}(a) \cup S_{0,\mathrm{small}}(u), \text{ with} \\ a &= C\sigma s_0 \sqrt{\log(p)/n}/\phi_0^2 \text{ for } C > 0 \text{ sufficiently large,} \\ u &= \tilde{C}\sigma s_0^{-5/2} s_{0,\mathrm{small}}^{-1}(u) n^{-1/2} \text{ for } \tilde{C} > 0 \text{ sufficiently small.} \end{aligned}$$

Then, when using the Lasso as selector $\hat{S}$, our hierarchical p-value method provides asymptotic strong error control of the familywise error rate.

6. CONCLUSIONS

We propose a method for testing whether (mainly) groups of correlated variables are significant for explaining a response in a high-dimensional linear model. In presence of highly correlated variables (or nearly collinear smaller groups of variables), as is very common in high-dimensional data, it seems indispensable to adopt such a kind of an approach going beyond multiple testing of individual regression coefficients. The groups of variables are ordered within a given hierarchy, for example, a cluster tree, which allows for powerful multiple testing adjustment. It automatically determines a good resolution level distinguishing between small and large groups of variables: the former are significant if the signal of one or few individual variables in such a small group is strong and/or the variables are not too highly correlated; and a large group can be significant even if the signals of (many) individual variables in the group are weak and the variables exhibit high correlation among themselves. The minimal true detections (MTDs) measure the power to detect significant smallest groups of variables, and our method performs well in terms of MTDs and substantially better than the analog of a single variable method.

Our procedure is based on repeated sample splitting which was empirically found to be "robust" and reliable for controlling Type I errors. We present some theory proving strong control of the familywise error rate, and our assumptions allow for scenarios beyond the beta-min condition saying that all nonzero regression coefficients should be sufficiently large. We also provide empirical results for simulated and real data which complement the theoretical analysis.

SUPPLEMENTARY MATERIALS

Supplementary material for "Hierarchical Testing in the High-Dimensional Setting with Correlated Variables": An alternative bottom-up hierarchical adjustment. Variability of Performance 1 and Performance 2 in the simulations study. Variability of MTDs in Section 4.3. Extension of the considerations of Section 4.3 for low SNR. Proofs.

[Received January 2014. Revised September 2014.]

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [334]

Bühlmann, P. (2013), "Statistical Significance in High-Dimensional Linear Models," *Bernoulli*, 19, 1212–1242. [331]

Bühlmann, P., Kalisch, M., and Meier, L. (2014), "High-Dimensional Statistics With a View Toward Applications in Biology," *Annual Review of Statistics and Its Application*, 1, 255–278. [335]

Bühlmann, P., and Mandozzi, J. (2014), "High-Dimensional Variable Screening and Bias in Subsequent Inference, With an Empirical Comparison," *Computational Statistics*, 29, 407–430. [342]

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. (2013), "Correlated Variables in Regression: Clustering and Sparse Estimation" (with discussion), *Journal of Statistical Planning and Inference*, 143, 1835–1871. [332,335]

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, New York: Springer Verlag. [333,342]

Chatterjee, A., and Lahiri, S. N. (2013), "Rates of Convergence of the Adaptive LASSO Estimators to the Oracle Distribution and Higher Order Refinements by the Bootstrap," *Annals of Statistics*, 41, 1232–1259. [331]

Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S. (2003), "Integrating Regulatory Motif Discovery and Genome-Wide Expression Analysis," *Proceedings of the National Academy of Sciences*, 100, 3339–3344. [341]

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2014), "High-Dimensional Inference: Confidence Intervals, *p*-Values and R-Software hdi," arXiv:1408.4026v1. [332]

Goeman, J. J., van De Geer, S. A., and Van Houwelingen, H. C. (2006), "Testing Against a High Dimensional Alternative," *Journal of the Royal Statistical Society*, Series B, 68, 477–493. [331]

Javanmard, A., and Montanari, A. (2014a), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [331]

Javanmard, A., and Montanari, A. (2014b), "Hypothesis Testing in High-Dimensional Regression Under the Gaussian Random Design Model: Asymptotic Theory," *IEEE Transactions on Information Theory*, 60, 6522–6554. [331]

Liu, H., and Yu, B. (2013), "Asymptotic Properties of Lasso+mLS and Lasso+Ridge in Sparse High-Dimensional Linear Regression," *Electronic Journal of Statistics*, 7, 3124–3169. [331]

Meinshausen, N. (2008), "Hierarchical Testing of Variable Importance," *Biometrika*, 95, 265–278. [331,332,334]

——— (2014), "Group Bound: Confidence Intervals for Groups of Variables in Sparse High Dimensional Regression Without Assumptions on the Design," *Journal of the Royal Statistical Society*, Series B, arXiv:1309.3489v2,. [332]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "P-Values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [331,332,333,335,341]

Minnier, J., Tian, L., and Cai, T. (2011), "A Perturbation Method for Inference on Regularized Regression Estimates," *Journal of the American Statistical Association*, 106, 1371–1382. [331]

Shaffer, J. P. (1986), "Modified Sequentially Rejective Multiple Test Procedures," *Journal of the American Statistical Association*, 81, 826–831. [334]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [332]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *Annals of Statistics*, 42, 1166–1202. [331]

van de Geer, S., Bühlmann, P., and Zhou, S. (2011), "The Adaptive and the Thresholded Lasso for Potentially Misspecified Models (and a Lower Bound for the Lasso)," *Electronic Journal of Statistics*, 5, 688–749. [342]

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002), "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature*, 415, 530–536. [335]

Wasserman, L., and Roeder, K. (2009), "High Dimensional Variable Selection," *Annals of Statistics*, 37, 2178–2201. [331]

Zhang, C.-H., and Huang, J. (2008), "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *Annals of Statistics*, 36, 1567–1594. [342]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [331]