Jacopo Mandozzi* and Peter Bühlmann

A Sequential Rejection Testing Method for High-Dimensional Regression with Correlated Variables

Abstract: We propose a general, modular method for significance testing of groups (or clusters) of variables in a high-dimensional linear model. In presence of high correlations among the covariables, due to serious problems of identifiability, it is indispensable to focus on detecting groups of variables rather than singletons. We propose an inference method which allows to build in hierarchical structures. It relies on repeated sample splitting and sequential rejection, and we prove that it asymptotically controls the familywise error rate. It can be implemented on any collection of clusters and leads to improved power in comparison to more standard non-sequential rejection methods. We complement the theoretical analysis with empirical results for simulated and real data.

Keywords: familywise error rate, hierarchical clustering, high-dimensional variable selection, inheritance procedure, lasso, linear model, minimal true detection, multiple testing, sample splitting, sequential rejection principle, singleton true detection

1 Introduction

Error control of false selection or false positive statements based on p-values is a primary goal of statistical inference and an established, broadly used tool in many areas of science. It relies on standard statistical hypothesis testing and procedures which give provable guarantees in presence of multiple, potentially very large scale multiple testing [1–3]. While being standard in the classical low-dimensional setup, statistical significance testing in the more challenging high-dimensional setting where the number of variables p might be much larger than the sample size n has only received attention recently.

We consider here a linear regression 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 $\mathbf{Y}$. We allow for high-dimensional scenarios where $p \gg n$. We assume that the regression coefficient vector is sparse with many coefficients of β^0 being equal to zero, that is, the active set of variables

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

is assumed to be a small subset of $\{1, \ldots, p\}$ corresponding to all variables.

A few methods for assigning p-values and constructing confidence intervals for individual parameters $\beta_j^0 (j = 1, \ldots, p)$ have been suggested [4–10], and some of them have been compared against each other in various settings [11, 12]. The inferential statements can easily be adjusted for multiplicity, thanks to the methodology and theory in multiple testing ([1], cf.). However, and important for practical applications, some major issues in presence of highly correlated variables still need further attention: typically, when $p \gg n$, none or only a few of the individual β_j^0's turn out to be significant which is a consequence of their near non-identifiability (even when some theoretical conditions on well-posedness on the design matrix $\mathbf{X}$ ([13], cf.). However, a *group* of (correlated) variables is often much better identifiable, but one can then not determine anymore the relevant variables within such a group [14–16].

*Corresponding author: Jacopo Mandozzi,** Libera AG, Zürich, Switzerland, E-mail: jacopo.mandozzi@libera.ch
Peter Bühlmann, Libera AG, Zürich, Switzerland

Thus, our main goal is testing of significance of groups of parameters: for a group or cluster $C \subseteq \{1, \ldots, p\}$ we consider the following null- and alternative hypothesis, respectively:

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

Given a collection $\mathscr{C}$ of clusters, we propose a general method for obtaining a collection $\mathscr{R} \subseteq \mathscr{C}$ of rejected clusters such that familywise error rate (FWER) is strongly controlled. That is, for a given nominal level $\alpha \in (0, 1)$:

$$\mathbb{P}[\mathscr{R} \subseteq \mathscr{F}] \geq 1 - \alpha,$$

where $\mathscr{F} = \{C \in \mathscr{C} \text{ s.t. } H_{0,C} \text{ is false}\}$ i.e., $\mathscr{F}$ is the collection of false null hypotheses. Our new method has the following main features:

- It can be implemented on any collection of clusters $\mathscr{C}$.
- It is modular in the sense that it requires four basic building blocks that have to satisfy certain assumptions.
- Its modular conception allows for a better insight of the procedure's power and improvements thereof.

We are particularly interested to use the procedure for *hierarchically ordered* clusters of (correlated) variables. Such a hierarchical structure can be obtained from the output of a hierarchical clustering algorithm: since it operates on the design matrix **X** only and does not involve the response **Y**, the inference for β^0 remains correct (for fixed design or by conditioning on **X**). With such a hierarchical cluster tree, our inference method (Sections 2.5 and 4.2) first tests the cluster $C = \{1, \ldots, p\}$ containing all the variables (the top node in the tree): if the corresponding null-hypothesis is rejected, we test some refined clusters, and we proceed down the cluster tree, in a *sequential manner*, until a cluster is not significant anymore. Figures 1 and 2, based on the results of the simulations in Section 5, provide some graphical illustrations. This procedure has the remarkable property that the resolution level of the significant clusters is automatically controlled by the sequential testing method: if the signal is strong
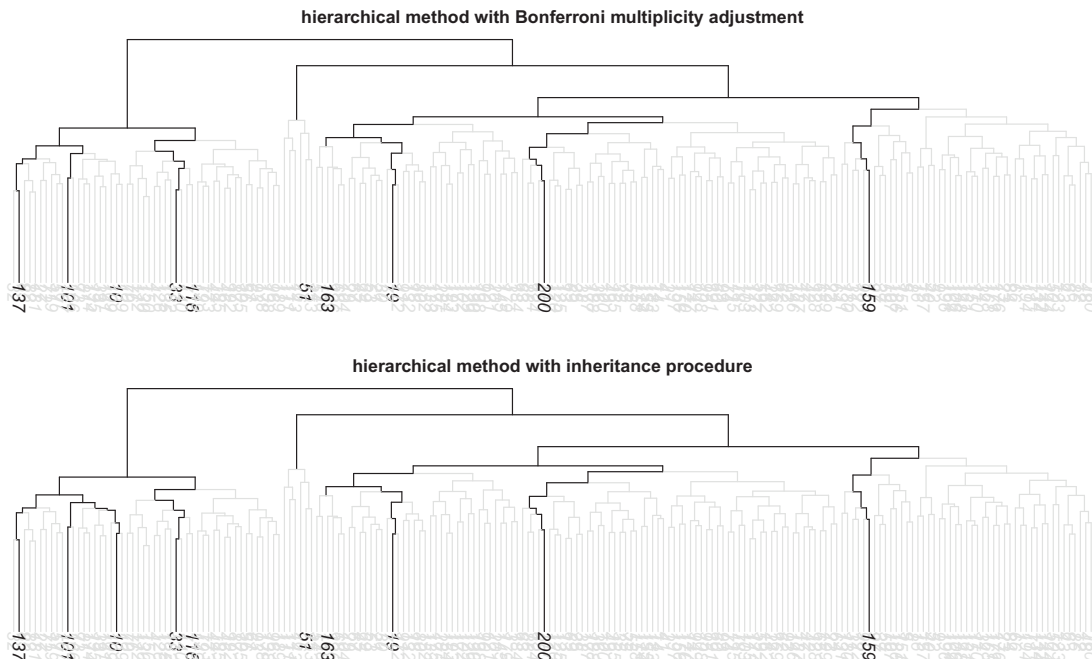

Figure 1: Dendrograms for a paradigmatic simulation run of the "equi correlation"-design with $p = 200$ and *SNR* $= 4$. The active variables are labeled in black and the truly detected non-zero variables along the hierarchy are depicted in black.

Figure 2: Dendrograms for a paradigmatic simulation run of the "high correlation within small blocks"-design with $p = 200$ and $SNR = 4$. The active variables are labeled in black and the truly detected non-zero variables along the hierarchy are depicted in black.
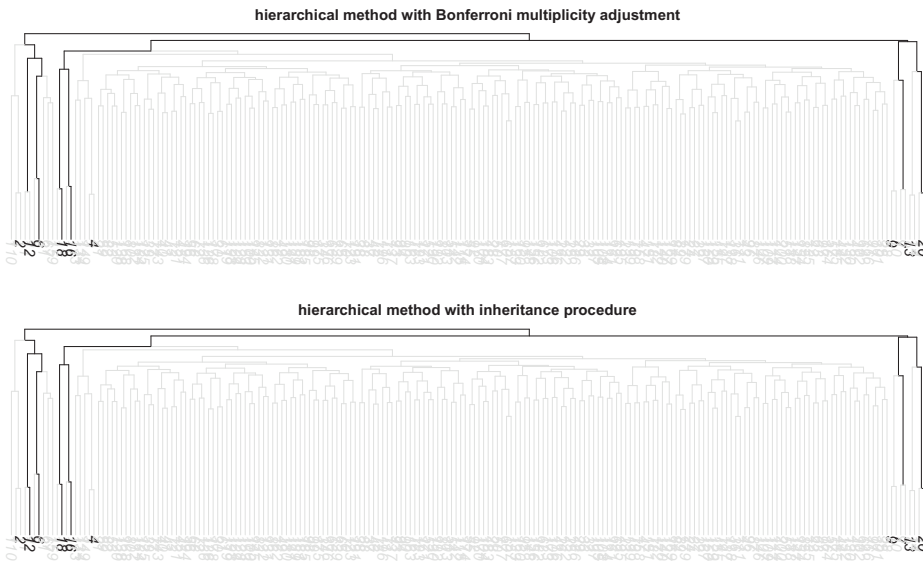
(e.g. large absolute values of components of β^0) and the variables are not too highly correlated, one can detect small clusters or even single variables and vice-versa, if the signal isn't very strong or the variables are highly correlated, only larger groups can be detected as significant.

1.1 Relation to other work

Our proposed method is based on the multi sample splitting method from Meinshausen et al. [7] and the sequential rejection principle of Goeman and Solari [17]. It is a generalization and power improvement over the multi sample splitting technique for inference of single variables [7] and for hierarchically ordered clusters of variables [15]. The improvement in power is strict, and in analogy to the gain of power of Holm's procedure [23] over the Bonferroni adjustment. Thus, even if the increased power might be only small for some datasets, one cannot do worse with the new procedure. The only price to pay is a slightly more complicated algorithm: we provide an implementation in the R-package hdi. We note that the mentioned improvement, based on the work by Goeman and Solari [17] and Goeman and Finos [18], is not entirely trivial to derive in the setting of the multiple sample splitting scheme. The modular set-up based on four building blocks presented here is addressing the issue how the improvement can be constructed.

1.2 Outline of the paper

In Section 2 we describe the four basic building blocks of the method and the assumptions that are sufficient to establish in Section 3 its strong FWER control. In Sections 4.1 and 4.2, respectively, we focus on the inference of two specific kinds of cluster collections: singletons and hierarchically ordered clusters. In Section 4.3 we show how logical relationships can be used to improve the power. Finally, we provide in Section 5 a comparison based on empirical results for error control and power, with a focus on minimal true detections, and we apply the new method to a real dataset.

2 A construction based on four building blocks

Our method is based on four basic building blocks that satisfy certain assumptions.

One main ingredient is multi sample splitting. For $b = 1, \ldots, B$ where B is the number of repeated sample splitting, the original data of sample size n is split into two disjoint groups, $N_{in}^{(b)}$ and $N_{out}^{(b)}$, i.e., a partition

$$\{1, \ldots, n\} = N_{in}^{(b)} \cup N_{out}^{(b)}$$

is randomly chosen. The groups are chosen of equal size if n is even or satisfy $|N_{out}^{(b)}| = |N_{in}^{(b)}| + 1$ if n is odd. The idea is to use data from $N_{in}^{(b)}$ to select a few variables and the other data from $N_{out}^{(b)}$ to perform the statistical hypothesis testing in the low-dimensional submodel with the selected variables from $N_{in}^{(b)}$. The details are described next.

2.1 Screening of variables

We consider variable screening where an estimator $\hat{S}^{(b)} \subseteq \{1, \ldots, p\}$, based on data corresponding to $N_{in}^{(b)}$, is aiming at including all active variables S_0. A prime example is the Lasso [19], while a detailed empirical comparison of five popular screening procedures can be found in Bühlmann and Mandozzi [20]. Assume that the screening procedure satisfies the following properties for any sample split b:

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

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

The *sparsity property* in (A1) implies that for each sample split b it holds that $|\hat{S}^{(b)}| < |N_{out}^{(b)}|$, a condition which is necessary for applying classical tests as described in Section 2.2 below. The δ*-screening property* in (A2) ensures that all the relevant variables are retained with high probability (where $\delta > 0$ is typically small). We indicate in Section 3.1 that under some assumptions, the Lasso satisfies (A1) and (A2).

2.2 Testing and p-values

The idea is to perform a classical statistical test on the other half sample from $N_{out}^{(b)}$ in a low-dimensional problem with variables from $\hat{S}^{(b)}$ only. For each sample split b, based on the second half of the sample corresponding to $N_{out}^{(b)}$, consider a testing procedure, e.g. the classical partial F-test (see also Section 3.1), that provides correct p-values $p^{C,(b)}$ for the null hypothesis $H_{0, C \cap \hat{S}^{(b)}}$ for each screened set $\hat{S}^{(b)}$, in the sense that for each nominal level $\alpha \in (0, 1)$

(A3) *Correct testing property* : Under the null hypothesis $H_{0, C \cap \hat{S}^{(b)}}$ it holds

$$\mathbb{P}[p^{C,(b)} \leq \alpha] \leq \alpha.$$

We note that the probability is with respect to the data generating random variables corresponding to the second half $N_{out}^{(b)}$, and the null-hypothesis is fixed with respect to $N_{out}^{(b)}$. Due to the screening property (A2), when $\delta \to 0$, the null-hypothesis $H_{0, C \cap \hat{S}^{(b)}}$ approximates the unconditional hypothesis $H_{0, C}$ which we aim to test for. If $C \cap \hat{S}^{(b)} = \emptyset$ define $p^{C,(b)} = 1$. This provides a (correct) p-value $p^{C,(b)}$ for each cluster $C \in \mathscr{C}$ and each sample split $b \in \{1 \ldots B\}$.

2.3 Multiplicity adjustment

Consider for each sample split b and each cluster $C \in \mathscr{C}$ a multiplicity adjustment procedure $m_C^{(b)} : 2^{\mathscr{C}} \to [1, \infty]$ that for each collection $\mathscr{R}$ of rejected clusters provides a multiplicity adjustment $m_C^{(b)}(\mathscr{R}) \geq 1$ and satisfies the following properties:

$$(A4)\ \textit{Monotonicity property}: \ \text{If } \mathscr{R} \subseteq \mathscr{S} \text{ then } m_C^{(b)}(\mathscr{R}) \geq m_C^{(b)}(\mathscr{S}).$$

$$(A5)\ \textit{Single-step property}: \ \sum_{C \in \mathscr{C} \setminus \mathscr{R}} \frac{1\{C \cap \hat{S}^{(b)} \neq \emptyset\}}{m_C^{(b)}(\mathscr{R})} \leq 1,$$

where we define $1/\infty = 0$. Such a family of multiplicity adjustments for $b = 1, \ldots, B$ are often naturally induced from a global multiplicity adjustment procedure m_C.

2.4 Aggregation of p-values

Consider a collection of screened sets of variables $\hat{S}^{(b)}$, a cluster $C \in \mathscr{C}$, a collection of p-values $p^{C,(b)}$ for the null-hypothesis $H_{0,C \cap S^{(b)}}$ (which approximates $H_{0,C}$, see comment after (A3)) and a collection of multiplicity adjustments $m_C^{(b)} \geq 1$ (we drop here the dependence on $\mathscr{R}$).

The goal is to aggregate the p-values $p^{C,(1)}, \ldots, p^{C,(B)}$ to a single p-value which is adjusted for multiplicity. An aggregation procedure is a monotone increasing function $aggr : [0,1]^B \to [0,1]$. Assume it satisfies the following property:

$$(A6)\ \textit{Aggregation property}: \ \text{If } \mathbb{P}[p^{C,(b)} \leq \alpha] \leq \alpha, \ \forall \alpha \in [0,1], \text{ then}$$

$$\mathbb{P}[aggr(p^{C,(1)} m_C^{(1)}, \ldots, p^{C,(B)} m_C^{(B)}) \leq \alpha]$$

$$\leq \frac{\alpha}{B} \sum_{b=1}^{B} \frac{1\{C \cap \hat{S}^{(b)} \neq \emptyset\}}{m_C^{(b)}}, \ \forall \alpha \in [0,1].$$

2.5 The procedure

Our procedure is based on the four building blocks above. First, we proceed with screening of the variables based on the first half sample from $N_{in}^{(b)}$ (Section 2.1), e.g., in Section 5.1 we use the Lasso with regularization parameter chosen by 10-fold cross-validation (see also Section 3.1). Then, we construct the p-values based on the second half sample from $N_{out}^{(b)}$ by using the partial F-test (Section 2.2 and see also Section 3.1). This leads to a (correct) p-value $p^{C,(b)}$ for each cluster $C \in \mathscr{C}$ and each sample split $b \in \{1 \ldots B\}$.

The multiplicity adjustment is done sequentially (Section 2.3). Based on a chosen significance level $\alpha \in (0,1)$ and for a collection of currently rejected sets $\mathscr{R}$, define the successor of $\mathscr{R}$ as

$$\mathscr{N}(\mathscr{R}) = \{C \in \mathscr{C} \setminus \mathscr{R} \text{ s.t. } aggr(p^{C,(1)} m_C^{(1)}(\mathscr{R}), \ldots, p^{C,(B)} m_C^{(B)}(\mathscr{R})) \leq \alpha\}$$

Start from "no rejections" $\mathscr{R}_0 = \emptyset$, define $\mathscr{R}_{i+1} = \mathscr{R}_i \cup \mathscr{N}(\mathscr{R}_i)$ and $\mathscr{R}_\infty = \lim_{i \to \infty} \mathscr{R}_i$ (although $\mathscr{R}_\infty$ is never constructed due to finite-ness of all possible subset of the variables). Concrete choices of $m_C^{(1)}(\mathscr{R}), \ldots, m_C^{(B)}(\mathscr{R})$ are discussed in Section 4.

Finally, we aggregate the p-values as indicated in Section 2.4. Concrete aggregation methods are described in Proposition 1 in Section 3.1.

3 Familywise error control

We show here that the method from Section 2.5 (strongly) controls the FWER at each step $i = 0, 1, 2, \dots$

Theorem 1 *Assume that (A1)-(A6) hold. Then for any $i \in \mathbb{N}_0 \cup \infty$*

$$\mathbb{P}[\mathscr{R}_i \subseteq \mathscr{F}] \geq (1-\delta)^B - \alpha,$$

where $\mathscr{F} = \{C \in \mathscr{C} \text{ s.t. } C \cap S_0 \neq \emptyset\}$ is the collection of false null hypotheses.

A proof is given in the Appendix.

3.1 Screening, testing and aggregation: their properties

We discuss here some choices for screening, testing and aggregation which we use in the implementation in the R-package hdi. The issue of sequential multiplicity adjustment is treated separately in Section 4.

For variable screening, we use the Lasso with regularization parameter chosen by 10-fold cross-validation. Theoretical justification of the sparsity and screening property (A1) and (A2) can be derived by assuming a compatibility or restricted eigenvalue condition on the fixed design matrix $\mathbf{X}$ and a beta-min assumption requiring that $\min_{j \in S_0} |\beta_j^0| \gg \sqrt{|S_0| \log(p)/n}$ is sufficiently large: we refer to Bühlmann and van de Geer ([13], Ch. 2.7 and Ch. 6) for the details.

For construction of the p-values (in the low-dimensional setting, due to variable screening in the first half of the sample) we use the partial F-test. Then, assuming fixed design $\mathbf{X}$ and Gaussian errors, condition (A3) holds.

For aggregation of the p-values, ensuring that (A6) holds, we have the following result for two slightly different methods.

Proposition 1 *Denote by $q_\gamma(u)$ the empirical γ-quantile of the values occurring in the components of a vector u. The monotone increasing functions $[0,1]^B \to [0,1]$*

$$(\tilde{p}^{(1)}, \dots, \tilde{p}^{(B)}) \mapsto Q(\gamma) = \min\{1, q_\gamma(\tilde{p}^{(1)}/\gamma, \dots, \tilde{p}^{(B)}/\gamma)\}$$

$$(\tilde{p}^{(1)}, \dots, \tilde{p}^{(B)}) \mapsto \min\left\{1, (1 - \log \gamma_{\min}) \inf_{\gamma \in (\gamma_{\min}, 1)} Q(\gamma)\right\}$$

satisfy the aggregation property (A6) for any $\gamma, \gamma_{\min} \in (0, 1)$.

A proof, which was basically given in Meinshausen et al. [7], can be found in the Appendix.

4 Some concrete methods for multiplicity adjustment

We discuss here the issue of multiplicity adjustment, and justify assumption (A4) and (A5) for different inference procedures.

4.1 Inference of single variables

This first example is paradigmatic for the advantages of the modular approach: a simple improvement of the multiplicity adjustment procedure allows for a better power, basically in the same way as in a low-dimensional setting in Goeman and Solari [17].

Concretely, we consider the problem of inferring single variables, i.e., the collection of clusters $\mathscr{C} = \{\{i\}; i = 1, \ldots, p\}$. The method proposed in Meinshausen et al. [7] corresponds to the method of Theorem 1 with the aggregation procedures of Proposition 1 and the following Bonferroni-based [21, 22] multiplicity adjustment procedure:

$$m_{\{i\}}^{(b)}(\mathscr{R}) = |\hat{S}^{(b)}|. \tag{2}$$

As the multiplicity adjustments are independent from the (previously) rejected collection of sets, the monotonicity property (A4) is trivially satisfied, while the single-step property (A5) follows from

$$\sum_{C \in \mathscr{C} \backslash \mathscr{R}} \frac{1\{C \cap \hat{S}^{(b)} \neq \emptyset\}}{m_C^{(b)}(\mathscr{R})} = \sum_{\{i\} \in \mathscr{C} \backslash \mathscr{R}} \frac{1\{\{i\} \cap \hat{S}^{(b)} \neq \emptyset\}}{|\hat{S}^{(b)}|} \leq 1.$$

The power of the method can be improved taking instead of (2) the following Bonferroni-Holm-based [23] multiplicity adjustment procedure:

$$m_{\{i\}}^{(b)}(\mathscr{R}) = |\hat{S}^{(b)} \cap (\mathscr{C} \backslash \mathscr{R})| = |\{j \in \hat{S}^{(b)} \text{ s.t. } \{j\} \notin \mathscr{R}\}|. \tag{3}$$

The monotonicity property (A4) is still satisfied since

$$|\hat{S}^{(b)} \cap (\mathscr{C} \backslash \mathscr{R})| \geq |\hat{S}^{(b)} \cap (\mathscr{C} \backslash \mathscr{R})|$$

for $\mathscr{R} \subseteq \mathscr{S}$, whereas

$$\sum_{C \in \mathscr{C} \backslash \mathscr{R}} \frac{1\{C \cap \hat{S}^{(b)} \neq \emptyset\}}{m_C^{(b)}(\mathscr{R})} = \sum_{\{i\} \in \mathscr{C} \backslash \mathscr{R}} \frac{1\{\{i\} \cap \hat{S}^{(b)} \neq \emptyset\}}{|\hat{S}^{(b)} \cap (\mathscr{C} \backslash \mathscr{R})|} = 1$$

proves the single step property (A5).

4.2 Inference of hierarchically ordered clusters of variables

When dealing with the challenge of inferring hierarchically ordered clusters of variables, e.g. from the tree-structured output of a hierarchical clustering algorithm, one considers a collection of clusters $\mathscr{C} = \{C_i\}_i$ where for any two clusters $C_i, C_{i'} \in \mathscr{C}$, either one cluster is a subset of the other, or they have an empty intersection. The method proposed in Mandozzi and Bühlmann ([15], Section 2), which is based on the procedure of Meinshausen [24], corresponds to the one as in Theorem 1 with the aggregation methods of Proposition 1 and the following multiplicity adjustment:

$$m_C^{(b)}(\mathscr{R}) = \begin{cases} \infty, & \text{if } \operatorname{anc}(C) \nsubseteq \mathscr{R} \\ 1, & \text{if } \operatorname{anc}(C) \subseteq \mathscr{R} \text{ and } \hat{S}^{(b)} \cap C = \emptyset \\ \dfrac{|\hat{S}^{(b)}|}{|\hat{S}^{(b)} \cap C|}, & \text{otherwise.} \end{cases} \tag{4}$$

Here, $\operatorname{anc}(C)$ denotes the ancestors in a hierarchically ordered cluster tree. To check the monotonicity property (A4), consider $\mathscr{R} \subseteq \mathscr{S}$. For $C \in \mathscr{C}$ with $\operatorname{anc}(C) \subseteq \mathscr{R}$ it holds $\operatorname{anc}(C) \subseteq \mathscr{S}$ and hence $m_C^{(b)}(\mathscr{R}) = m_C^{(b)}(\mathscr{S})$, while for $C \in \mathscr{C}$ with $\operatorname{anc}(C) \nsubseteq \mathscr{R}$ one has $m_C^{(b)}(\mathscr{R}) = \infty \geq m_C^{(b)}(\mathscr{S})$. The single step property (A5) follows from

$$\sum_{C \in \mathscr{C} \backslash \mathscr{R}} \frac{1\{C \cap \hat{S}^{(b)} \neq \emptyset\}}{m_C^{(b)}(\mathscr{R})} = \frac{1}{|\hat{S}^{(b)}|} \sum_{C \in \mathscr{C} \backslash \mathscr{R} \text{ s.t. } \operatorname{anc}(C) \subseteq \mathscr{R}} |\hat{S}^{(b)} \cap C| \leq 1,$$

where in the inequality we have used the fact that for two sets in the sum above, one cannot be a subset of the other and hence, by definition of the hierarchy $\mathscr{C}$, they are disjoint.

4.2.1 The inheritance procedure in the high-dimensional setting

In Goeman and Solari ([17], Section 6.3) and Goeman and Finos [18], the authors propose various possibilities on how the sequential rejection principle can be used to improve the power of the hierarchical procedure in Meinshausen [24].

We consider here the most powerful one, the inheritance procedure of Goeman and Finos [18] which we extend to the high-dimensional setting with hierarchical cluster trees. In order to do that, we apply the method of Theorem 1 with the aggregation procedures of Proposition 1 and the following multiplicity adjustment:

$$
m_C^{(b)}(\mathscr{R}) = \begin{cases} \infty, & \text{if } \operatorname{anc}(C) \not\subseteq \mathscr{R} \\[2mm] 1, & \text{if } \operatorname{anc}(C) \subseteq \mathscr{R} \text{ and } \hat{S}^{(b)} \cap C = \emptyset \\[2mm] \dfrac{|\hat{S}^{(b)}|}{|\hat{S}^{(b)} \cap C|} \displaystyle\prod_{D \in \operatorname{anc}(C)} n_D^{(b)}(\mathscr{R}), & \text{otherwise,} \end{cases} \tag{5}
$$

where

$$
n_D^{(b)}(\mathscr{R}) = \frac{1}{|\hat{S}^{(b)} \cap D|} \sum_{E \in \operatorname{ch}(D) \setminus \mathscr{E}(\mathscr{R})} |\hat{S}^{(b)} \cap E|
$$

and

$$
\mathscr{E}(\mathscr{R}) = \{ C \in \mathscr{C} \text{ s.t. } \operatorname{of}(C) \subseteq \mathscr{R} \}
$$

the set of extinct branches, i.e., the set of hypotheses which have been rejected together with all their offspring denoted by $\operatorname{of}(C)$ (as before, $\operatorname{anc}(C)$ denotes the ancestors of cluster C). Note that since $n_D^{(b)}(\mathscr{R}) \leq 1$ this procedure leads to an uniform improvement over the method of the previous section.

The monotonicity property (A4) follows from the same considerations as above and

$$
\mathscr{R} \subseteq \mathscr{S} \;\Rightarrow\; \mathscr{E}(\mathscr{R}) \subseteq \mathscr{E}(\mathscr{S}) \;\Rightarrow\; n_D^{(b)}(\mathscr{R}) \geq n_D^{(b)}(\mathscr{S}).
$$

To check that the single step property (A5) holds note that

$$
\sum_{C \in \mathscr{C} \setminus \mathscr{R}} \frac{1\{C \cap \hat{S}^{(b)} \neq \emptyset\}}{m_C^{(b)}(\mathscr{R})} = \sum_{C \in \mathscr{C} \setminus \mathscr{R} \text{ s.t. } \operatorname{anc}(C) \subseteq \mathscr{R}} \frac{|\hat{S}^{(b)} \cap C|}{|\hat{S}^{(b)}|} \prod_{D \in \operatorname{anc}(C)} \frac{|\hat{S}^{(b)} \cap D|}{\sum_{E \in \operatorname{ch}(D) \setminus \mathscr{E}(\mathscr{R})} |\hat{S}^{(b)} \cap E|}
$$

$$
= \sum_{C \in \mathscr{C} \setminus \mathscr{R}} \frac{\alpha_C^{(b)}(\mathscr{R})}{\alpha}
$$

where $\alpha_C^{(b)}$ is as in Goeman and Finos ([18], eq. (5)) with the weights $w_C^{(b)} = |\hat{S}^{(b)} \cap C|$; therefore the single step property follows directly from the considerations in Goeman and Finos [18].

4.3 Exploiting logical relationships: Shaffer improvements

Logical relationships between hypothesis can be exploited to improve the power of the sequential rejection procedure. A first example of such an improvement for hierarchically ordered clusters was given in Meinshausen [24], while in Goeman and Finos [18] the improvement is applied to the inheritance procedure. Since those improvements are based on the considerations of Shaffer [25] they are called "Shaffer

improvements". For the high-dimensional setting a possible Shaffer improvement consists of multiplying the multiplicity adjustment $m_C^{(b)}(\mathscr{R})$ with the Shaffer factor

$$s_C^{(b)}(\mathscr{R}) = \max\{m_C^{(b)}(\mathscr{U})/m_C^{(b)}(\mathscr{R}) \text{ s.t. } C \notin \mathscr{U} \supseteq \mathscr{R}, \ \mathscr{U} \text{ congruent}\}, \tag{6}$$

where a set $\mathscr{U} \subseteq \mathscr{C}$ is called congruent if, by the logical implications, it can be a complete set of false hypothesis (e.g. for a collection $\mathscr{C}$ of hierarchically ordered hypothesis $\mathscr{U} \subseteq \mathscr{C}$ is congruent if for each $C \in \mathscr{U}$ it holds $\mathrm{anc}(C) \subseteq \mathscr{U}$ and at least one offspring leaf node of C is in $\mathscr{U}$).

Note that multiplication with the Shaffer factor never decreases the power of the method since by the monotonicity property (A4), $s_C^{(b)}(\mathscr{R}) \le 1$. Moreover, if $\mathscr{R}$ is congruent, then $s_C^{(b)}(\mathscr{R}) = 1$ and since the collection $\mathscr{F}$ of all false hypothesis is congruent, the Shaffer improvement doesn't affect the validity of eq. (8). Finally, for $\mathscr{R} \subseteq \mathscr{S}$,

$$m_C^{(b)}(\mathscr{R})s_C^{(b)}(\mathscr{R}) = \max\{m_C^{(b)}(\mathscr{U}) \text{ s.t. } C \notin \mathscr{U} \supseteq \mathscr{R}, \ \mathscr{U} \text{ congruent}\}$$

$$\ge \max\{m_C^{(b)}(\mathscr{U}) \text{ s.t. } C \notin \mathscr{U} \supseteq \mathscr{S}, \ \mathscr{U} \text{ congruent}\}$$

$$= m_C^{(b)}(\mathscr{S})s_C^{(b)}(\mathscr{S})$$

and hence the Shaffer improvement doesn't affect the validity of eq. (7) neither.

We want to apply this Shaffer improvement to the inheritance procedure described in Section 4.2.1. Following the same reasoning as in Goeman and Finos ([18], Section 6), with the weights $w_C^{(b)} = |\hat{S}^{(b)} \cap C|$ we get the Shaffer factor

$$s_C^{(b)}(\mathscr{R}) = \begin{cases} \dfrac{w_C^{(b)} + u_C^{(b)} - v_C^{(b)}}{w_C^{(b)} + u_C^{(b)}}, & \text{if } C \notin \mathscr{R}, \mathrm{si}(C) \subseteq \mathscr{L} \backslash \mathscr{R} \\ 1, & \text{otherwise,} \end{cases}$$

where $\mathrm{si}(C) = \mathrm{ch}(\mathrm{pa}(C)) \backslash \{C\}$ denotes the siblings of C, $\mathscr{L} \subset \mathscr{C}$ denotes the collection of leaf nodes, $u_C^{(b)} = \sum_{D \in \mathrm{si}(C)} w_D^{(b)}$ and $v_C^{(b)} = \min_{D \in \mathrm{si}(C)} w_D^{(b)}$. If $\mathscr{C}$ is a binary tree the Shaffer factor becomes

$$s_C^{(b)}(\mathscr{R}) = \begin{cases} \dfrac{|\hat{S}^{(b)} \cap C|}{|\hat{S}^{(b)} \cap C| + |\hat{S}^{(b)} \cap \mathrm{si}(C)|}, & \text{if } C \notin \mathscr{R}, \mathrm{si}(C) \subseteq \mathscr{L} \backslash \mathscr{R} \\ 1, & \text{otherwise.} \end{cases}$$

Unlike as for the inheritance procedure in (5), the Shaffer factor (6) for the procedure in (4) is always 1. Nevertheless, a possibility how to exploit logical relationships to improve the power of the procedure (4) for binary trees, which provides a Shaffer improvement very similar to the one above, is illustrated in Mandozzi and Bühlmann [15].

5 Empirical results

5.1 Implementation of the methods and considered scenarios

In this section we compare the performance of the four methods illustrated in Sections 4.1 and 4.2 and refined in Section 4.3, i.e. single variable method with Bonferroni multiplicity adjustment (2), hierarchical method with Bonferroni-based adjustment (4) along with Shaffer improvement as in Mandozzi and Bühlmann [15], single variable method with Bonferroni-Holm multiplicity adjustment (3) and hierarchical method with inheritance procedure (5) along with Shaffer improvement (6). In the following we refer to the first two methods as the "non-sequential methods" (strictly seen, the hierarchical method with Bonferroni-

based adjustment is actually sequential, but there previous rejections are not used to improve subsequent multiplicity corrections) and the latter two methods as the "sequential methods".

We consider the same implementation of the methods and the same scenarios (with exactly the same sample splits) as in Mandozzi and Bühlmann [15], although here we use only standard hierarchical clustering for the hierarchical methods. Concretely, the following choices have been made for implementation:

- construction of the clusters with standard hierarchical clustering (using the R-function hclust) with distance between two covariables equal to 1 minus the absolute correlation between the covariables, and using complete linkage;
- screening with the Lasso [19] with regularization parameter chosen by 10-fold cross-validation;
- $B = 50$ sample splits (for each scenario exactly the same splits as in Mandozzi and Bühlmann [15]);
- for aggregation, the p-values P_h^C in Proposition 1 are computed over a grid of γ-values between $\gamma_{\min} = 0.05$ and 1 with grid-steps of size 0.025;
- nominal significance level $\alpha = 5\%$.

The following scenarios are considered (for the details we refer to Mandozzi and Bühlmann [15]):

- 42 scenarios based on 7 designs;
- for each design we consider 6 settings by varying the number of variables p in the model and the signal to noise ratio defined by

$$\text{SNR} = \sqrt{(\beta^0)^T \mathbf{X}^T \mathbf{X} \beta^0 n^{-1} \sigma^{-2}},$$

  namely for $p = 200$ we use $\text{SNR} = 4$ and $\text{SNR} = 8$, for $p = 500$ we use $\text{SNR} = 8$ and $\text{SNR} = 16$ and for $p = 1000$ we use $\text{SNR} = 16$ and $\text{SNR} = 32$;
- 3 designs based on synthetic data ("equi correlation", "high correlation within small blocks" and "high correlation within large blocks") and 4 designs based on semi-real data ("Riboflavin with normal correlation", "Breast with normal correlation", "Riboflavin with high correlation", "Breast with high correlation");
- sparsity $s_0 = 6$ for the two "Riboflavin"-designs and $s_0 = 10$ for the other five designs.

5.2 Familywise error rate control (FWER)

For each of the 42 scenarios described in Section 5.1 we consider exactly the same 100 independent simulation runs as in Mandozzi and Bühlmann ([15], section 4.2.2) by varying only the synthetic noise term ε and count the number where at least one false selection is made. According to Theorem 1, we expect this number to be at most $100\alpha = 5$ ($\alpha = 0.05$). The results for the Bonferroni-based methods can be seen in Mandozzi and Bühlmann ([15], Table 1): FWER control holds for 40 of the 42 scenarios and in 37 scenarios there is no false selection at all.

The results for the methods with sequential rejection are very similar, the only differences being that for the "high correlation within small blocks"-design with $p = 500$ and $\text{SNR} = 8$ the number of runs with at least a false selection increases (compared to Bonferroni-type methods) from 7 to 9 for the single variable method, and from 7 to 13 for the hierarchical method, respectively; for the same design with $p = 1000$ and $\text{SNR} = 16$ the number of runs with at least a false selection increases from 5 to 6 for both the single variable and hierarchical method. For all other scenarios, inclusively the "high correlation within large blocks"-design with $p = 200$ and $\text{SNR} = 4$, where the non-sequential hierarchical method slightly failed to control FWER (6 runs with at least a false detection), the sequential methods exhibit the same FWER control as their non-sequential counterparts.

Summarizing, FWER holds for all four methods in 39 out of 42 scenarios and the designs where it doesn't fully hold are "high correlation within small blocks" and "high correlation within large blocks",

Table 1: Empirical results, averaged over 100 simulation runs, for single variable method with Bonferroni (SB) and with Bonferroni-Holm (SH), resp. hierarchical method with Bonferroni (HB) and with sequential rejection induced by the inheritance procedure (HSR).

Design		p	# MTDs				# STDs		Perf 1 (in %)				Perf 2 (in %)	
			SB	SH	HB	HSR	HB	HSR	SB	SH	HB	HSR	HB	HSR
low SNR	equi corr	200	4.79	5.00	5.40	5.59	4.34	4.55	47.9	50.0	44.2	46.3	46.1	48.1
		500	3.97	4.13	4.74	4.84	3.73	3.84	39.7	41.3	37.7	38.7	38.3	39.3
		1,000	1.77	1.79	2.54	2.54	1.73	1.73	17.7	17.9	17.4	17.4	17.6	17.6
	small blocks	200	4.45	4.78	6.85	7.12	4.36	4.84	44.5	47.8	53.7	57.5	60.7	64.0
		500	3.15	3.33	5.18	5.27	3.15	3.42	31.5	33.3	38.3	40.1	44.2	45.5
		1,000	1.31	1.35	2.53	2.57	1.31	1.37	13.1	13.5	15.1	15.6	17.2	17.7
	large blocks	200	0.29	0.30	6.50	6.50	0.28	0.28	2.9	3.0	6.7	6.7	31.3	31.3
		500	0.06	0.06	2.76	2.76	0.06	0.06	0.6	0.6	1.1	1.1	1.1	1.1
		1,000	0.00	0.00	0.60	0.60	0.00	0.00	0.0	0.0	0.1	0.1	0.1	0.1
	Riboflavin normal corr	200	1.41	1.43	2.41	2.46	1.33	1.35	23.5	23.8	23.4	23.8	25.2	25.7
		500	0.90	0.90	1.84	1.85	0.77	0.79	15.0	15.0	13.5	13.8	14.2	14.5
		1,000	0.72	0.73	1.60	1.63	0.63	0.66	12.0	12.2	10.8	11.2	11.0	11.4
	Breast normal corr	200	4.05	4.16	5.00	5.11	3.84	3.94	40.5	41.6	39.5	40.6	41.6	42.9
		500	3.95	4.02	5.04	5.11	3.82	3.87	39.5	40.2	38.8	39.3	39.6	40.2
		1,000	3.30	3.34	4.25	4.27	3.10	3.13	33.0	33.4	31.2	31.5	31.7	31.9
	Riboflavin high corr	200	1.44	1.49	2.96	2.96	1.41	1.44	24.0	24.8	26.0	26.4	31.8	32.1
		500	1.72	1.79	2.95	2.98	1.69	1.72	28.7	29.8	29.9	30.4	32.8	33.3
		1,000	1.51	1.51	2.54	2.56	1.49	1.52	25.2	25.2	25.3	25.8	25.7	26.1
	Breast high corr	200	3.98	4.10	5.91	5.95	3.87	3.91	39.8	41.0	41.2	41.6	46.1	46.6
		500	5.13	5.22	6.51	6.56	4.87	4.93	51.3	52.2	49.9	50.4	51.7	52.3
		1,000	4.73	4.77	5.95	5.98	4.64	4.67	47.3	47.7	47.0	47.3	48.3	48.6
high SNR	equi corr	200	9.77	9.83	9.79	9.80	9.73	9.74	97.7	98.3	97.4	97.5	97.4	97.5
		500	7.28	7.38	7.63	7.67	7.18	7.24	72.8	73.8	72.0	72.5	72.1	72.6
		1,000	2.81	2.84	3.50	3.50	2.78	2.78	28.1	28.4	27.9	27.9	28.1	28.1
	small blocks	200	9.18	9.31	9.98	10.00	9.29	9.48	91.8	93.1	96.3	97.4	98.1	98.7
		500	6.99	7.03	8.05	8.14	7.02	7.15	69.9	70.3	73.5	74.7	76.4	77.4
		1,000	2.26	2.27	3.40	3.41	2.26	2.28	22.6	22.7	24.3	24.5	26.3	26.5
	large blocks	200	2.17	2.26	9.58	9.58	2.13	2.14	21.7	22.6	27.9	28.0	61.4	61.4
		500	1.17	1.20	5.38	5.38	1.15	1.15	11.7	12.0	12.6	12.6	13.2	13.2
		1,000	0.43	0.45	1.11	1.11	0.43	0.43	4.3	4.5	4.4	4.4	4.4	4.4
	Riboflavin normal corr	200	3.39	3.46	3.89	3.92	3.33	3.34	56.5	57.7	56.3	56.5	58.7	59.1
		500	2.24	2.25	2.90	2.90	2.15	2.15	37.3	37.5	36.4	36.4	36.9	36.9
		1,000	0.98	1.00	1.83	1.83	0.96	0.96	16.3	16.7	16.2	16.2	16.3	16.3
	Breast normal corr	200	8.65	8.70	8.89	8.93	8.60	8.65	86.5	87.0	86.4	86.8	87.2	87.7
		500	6.81	6.86	7.33	7.35	6.72	6.74	68.1	68.6	67.6	67.8	68.3	68.5
		1,000	3.95	3.97	4.81	4.84	3.79	3.82	39.5	39.7	38.1	38.4	38.4	38.7
	Riboflavin high corr	200	3.86	3.97	4.79	4.83	3.82	3.86	64.3	66.2	66.0	66.7	69.4	70.1
		500	3.69	3.72	4.40	4.43	3.65	3.68	61.5	62.0	61.8	62.3	63.7	64.2
		1,000	2.48	2.51	3.24	3.27	2.43	2.45	41.3	41.8	40.7	41.0	41.1	41.4
	Breast high corr	200	9.09	9.15	9.59	9.61	9.09	9.14	90.9	91.5	91.9	92.3	93.8	94.1
		500	7.75	7.82	8.38	8.40	7.71	7.72	77.5	78.2	77.8	78.0	78.9	79.1
		1,000	5.85	5.89	6.73	6.76	5.72	5.75	58.5	58.9	57.5	57.9	58.1	58.5
Average			3.65	3.72	4.98	5.02	3.58	3.64	40.4	41.1	41.0	41.6	43.9	44.4

which is not surprising since each active predictor is highly correlated with one or more inactive predictors from S_0^c and hence it is rather difficult for our screening method (the Lasso) to guarantee that $\hat{S} \supseteq S_0$.

5.3 Power

For measuring the power we consider four different aspects: the one-dimensional statistics defined in Mandozzi and Bühlmann ([15], section 4.2.1) as "Performance 1" and "Performance 2" (see below), the number of minimal true detections (MTDs, i.e., smallest significant groups of variables of any cardinality, containing at least one active variable, see below) and singleton true detections (STDs, i.e., MTDs with cardinality 1). Concretely, a cluster is said to be a MTD if it satisfies all of the following:

– C is a significant cluster, e.g., has p-value $< 5\%$ ("Detection");
–   There is no significant sub-cluster $D \subset C$ ("Minimal");
– $C \in \mathscr{F}$, i.e., there is at least one active variable in C ("True");

and we define:

$$\text{Performance 1} = \frac{1}{|S_0|} \sum_{MTD\ C} \frac{1}{|C|},$$

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

For each of the 42 scenarios outlined in Section 5.1, we consider exactly the same 100 independent simulation runs obtained in Mandozzi and Bühlmann ([15], section 4.2.3-4) by varying the synthetic noise term ε and the synthetic regression vector β^0. We then calculate the average Performance 1, Performance 2, number of MTDs and number STDs, over the 100 simulation runs. The results are shown in Table 1 (for the single variable methods each MTD is an STD and by definition Performance 1 and 2 are the same).

Considering both low and high SNR, the methods with sequential rejection improve the considered power measures in comparison to the analogous method without sequential rejection in 207 out of 252 cases, the absolute improvement being at least 0.05 for MTDs and STDs, and at least 0.5% for Performance 1 and Performance 2 in 133 cases out of 252 cases. For better interpretation of these results: an absolute improvement of 0.05 MTDs (resp. STDs) basically means that in one out of 20 runs one more MTD (resp. STD) could be detected. Averaging over all scenarios, the improvement given by the sequential rejection procedures lies between 0.04 and 0.06 for MTDs and STDs, and between 0.5% and 0.7% for Performance 1 and Performance 2. The biggest gain with sequential rejection can be found in the "high correlation within small blocks"-design with $p = 200$ and low SNR: it consists of 0.48 more STDs, 0.33 more MTDs and an absolute increase of 3.8% of Performance 1 and 3.3% of Performance 2, respectively. This basically means that in half of the runs the method with sequential rejection could find one STD more and in one third of the runs it could find one MTD more. Other particularly favorable scenarios for an improvement with sequential rejection are the "equi correlation"-design and the "breast normal correlation"-design, both with $p = 200$ and low SNR and the "high correlation within small blocks"-design with high SNR and $p = 200$, resp. $p = 500$.

In general, the improvement given by the sequential rejection procedures decreases with increasing number p of covariables and is substantial only when the power of the method without sequential rejection is intermediate. These empirical findings are not surprising, since looking at how the methods are defined and in particular at the eqs (2)–(5), we conclude that an improvement with the sequential rejection methods is only possible if the related non-sequential method provides at least an STD (and gets more likely the more STDs are provided by the non-sequential method). Moreover, an improvement with sequential rejection is more likely to happen when the number $|\hat{S}|$ of screened variables is small.

For a better illustration of what kind of an improvement is possible using sequential rejection, we show in Figures 1 and 2 the dendrograms (in gray) for a paradigmatic simulation run of the "equi correlation" – and the "high correlation within small blocks"-design, respectively, both with $p = 200$ and $SNR = 4$.

Figure 1 illustrates that sequential rejection allows the detection of a further singleton, increasing the number of STDs from 6 to 7 and the number of MTDs 8 to 9. In Figure 2 sequential rejection allows to detect a singleton that could previously only be detected together with another non-relevant variable in a cluster of cardinality 2, increasing the number of true STDs from 4 to 5 (while the number of MTDs remains to be 6).

Finally, we have performed a simulation with the same scenarios (and the same sample splits) as in Mandozzi and Bühlmann ([15], section 4.3), i.e. "small blocks"-designs and "large blocks"-designs with 8 different correlations

$$\rho \in \{0, 0.4, 0.7, 0.8, 0.85, 0.9, 0.95, 0.99\}.$$

The full results are shown in Table 2 in the Appendix. While the methods with sequential rejection control the FWER in exactly the same scenarios where it is also controlled by the non-sequential methods, they increase the average number of MTDs from 5.51 to 5.62 for the single variable method, and from 8.11 to 8.18 for the hierarchical method, and the number of STDs for the hierarchical method from 5.44 to 5.55, with improvements for a single scenario up to 0.48 MTDs and 0.56 STDs (averaged over 100 runs).

The empirical results can be summarized as follows. The methods with sequential rejection essentially controls the FWER in the same way as the non-sequential methods. Regarding power, sequential rejection allows for improvements, to a similar extent for the single variable and the hierarchical procedures. As already noted in Mandozzi and Bühlmann [15], for the non-sequential methods, the hierarchical methods have similar STDs as the single variable methods but allow for substantially more MTDs. Thus, our proposed hierarchical method with the inheritance procedure can be seen as the best of the considered methods.

5.4 Real data application: Motif Regression

We consider here a problem of motif regression [26] from computational biology. We apply the four methods described above, plus the two hierarchical methods (with and without sequential rejection) using the recently proposed canonical correlation clustering of Bühlmann et al. [14], to a real dataset with $n = 287$ and $p = 195$, used in Meinshausen ([24], section 4.3) and Mandozzi and Bühlmann ([15], section 4.4). The sequential rejection methods detects exactly the same significant structures as non-sequential methods, namely a single variable and a cluster containing 165 variables (the latter can be detected only with the hierarchical method with canonical correlation clustering). This can barely be considered as surprising, as with only one STD by the non-sequential methods, further improvements by the sequential methods are rather unlikely (see Section 5.3 for more explanation and empirical evidence).

6 Conclusions

We propose a general sequential rejection testing method for clusters and single variables in a high-dimensional linear model. In presence of high correlations among the covariables, due to serious problems of identifiability, it is essentially mandatory to focus on detecting significant groups of variables rather than single individual covariables. Our method asymptotically controls the familywise error rate (FWER), while, as a consequence of its modular structure, allowing for unburdened power optimization. We provide an implementation in the R-package hdi.

We use and study the procedure for inference of single variables but much more importantly, for hierarchically ordered clusters of variables. With the latter, we establish a powerful scheme for meaningful inference in a high-dimensional regression model, much beyond considering single variables only. Our

presented mathematical analysis on control of the FWER and power improvement is complemented by empirical results based on semi-real and simulated data confirming the theoretical results.

Appendix

Proof of Theorem 1

Proof. We show that the procedure satisfies monotonicity and single-step conditions as required by Goeman and Solari ([17], Theorem 1), i.e.

$$\mathscr{R} \subseteq \mathscr{S} \Rightarrow \mathscr{N}(\mathscr{R}) \subseteq \mathscr{N}(\mathscr{S}) \cup \mathscr{S} \tag{7}$$

$$\mathbb{P}[\mathscr{N}(\mathscr{F}) \subseteq \mathscr{F}] \geq (1-\delta)^B - \alpha. \tag{8}$$

Assume $\mathscr{R} \subseteq \mathscr{S}$ and $C \in \mathscr{N}(\mathscr{R})$. Then by definition

$$\mathrm{aggr}(p^{C,(1)} m_C^{(1)}(\mathscr{R}), \ldots, p^{C,(B)} m_C^{(B)}(\mathscr{R})) \leq \alpha.$$

The monotonicity property (A4) of the multiplicity adjustment and the fact that the aggregation procedure is monotone increasing imply

$$\mathrm{aggr}(p^{C,(1)} m_C^{(1)}(\mathscr{S}), \ldots, p^{C,(B)} m_C^{(B)}(\mathscr{S})) \leq \mathrm{aggr}(p^{C,(1)} m_C^{(1)}(\mathscr{R}), \ldots, p^{C,(B)} m_C^{(B)}(\mathscr{R}))$$

and hence either $C \in \mathscr{S}$ or $C \in \mathscr{N}(\mathscr{S})$ which proves (7). Consider the event

$$\mathscr{A} = \{\hat{S}^{(b)} \supseteq S_0, \forall b = 1 \ldots B\}$$

where all screenings are satisfied. Because of the δ-screening assumption (A2) it holds $\mathbb{P}(\mathscr{A}) \geq (1-\delta)^B$ and hence

$$\mathbb{P}[\mathscr{N}(\mathscr{F}) \not\subseteq \mathscr{F}] = \mathbb{P}[\mathscr{N}(\mathscr{F}) \not\subseteq \mathscr{F} \mid \mathscr{A}] \mathbb{P}(\mathscr{A}) + \mathbb{P}[\mathscr{N}(\mathscr{F}) \not\subseteq \mathscr{F} \mid \mathbb{P}\mathscr{A}^c](\mathscr{A}^c)$$
$$\leq \mathbb{P}[\mathscr{N}(\mathscr{F}) \not\subseteq \mathscr{F} \mid \mathscr{A}] + (1 - (1-\delta)^B).$$

Since

$$\mathbb{P}[\mathscr{N}(\mathscr{F}) \not\subseteq \mathscr{F} \mid \mathscr{A}] \leq \mathbb{P}[\bigcup_{\mathscr{C} \setminus \mathscr{F}} \{\mathrm{aggr}(p^{C,(1)} m_C^{(1)}(\mathscr{F}), \ldots, p^{C,(B)} m_C^{(B)}(\mathscr{F})) \leq \alpha\}]$$

$$\leq \sum_{\mathscr{C} \setminus \mathscr{F}} \mathbb{P}[\mathrm{aggr}(p^{C,(1)} m_C^{(1)}(\mathscr{F}), \ldots, p^{C,(B)} m_C^{(B)}(\mathscr{F})) \leq \alpha\}]$$

$$\overset{(A3)(A6)}{\leq} \sum_{\mathscr{C} \setminus \mathscr{F}} \frac{\alpha}{B} \sum_{b=1}^{B} \frac{1\{C \cap \hat{S}^{(b)} \neq \emptyset\}}{m_C^{(b)}(\mathscr{F})} \overset{(A5)}{\leq} \frac{\alpha}{B} \sum_{b=1}^{B} 1 \leq \alpha$$

we can prove (8) as follows:

$$\mathbb{P}[\mathscr{N}(\mathscr{F}) \subseteq \mathscr{F}] = 1 - \mathbb{P}[\mathscr{N}(\mathscr{F}) \not\subseteq \mathscr{F}]$$
$$\geq 1 - (\alpha + (1 - (1-\delta)^B)) = (1-\delta)^B - \alpha.$$

□

Proof of Proposition 1

Proof. The proof was basically given in the Appendix of Meinshausen et al. [7]. In the following we omit the function $\min\{1, \cdot\}$ from the definition of $Q(\gamma)$ in order to simplify the notation (this is possible since the level α is smaller than 1). Define for $u \in (0, 1)$ the function

$$\pi(u) := \frac{1}{B}\sum_{b=1}^{B} 1\{\tilde{p}^{(b)} \leq u\}.$$

Then it holds

$$Q(\gamma) \leq \alpha \Leftrightarrow q_\gamma(\tilde{p}^{(1)}/\gamma, \ldots, \tilde{p}^{(B)}/\gamma) \leq \alpha \Leftrightarrow q_\gamma(\tilde{p}^{(1)}, \ldots, \tilde{p}^{(B)}) \leq \alpha\gamma$$

$$\Leftrightarrow \sum_{b=1}^{B} 1\{\tilde{p}^{(b)} \leq \alpha\gamma\} \geq B\gamma \Leftrightarrow \pi(\alpha\gamma) \geq \gamma.$$

Thus,

$$\mathbb{P}(Q(\gamma) \leq \alpha) = \mathbb{E}(1\{Q(\gamma) \leq \alpha\}) = \mathbb{E}(1\{\pi(\alpha\gamma) \geq \gamma\}) \leq \frac{1}{\gamma}\mathbb{E}(\pi(\alpha\gamma))$$

$$= \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\mathbb{E}(1\{\tilde{p}^{(b)} \leq \alpha\gamma\}) = \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}(\tilde{p}^{(b)} \leq \alpha\gamma)$$

$$\leq \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\frac{\alpha\gamma}{m^{(b)}}1\{C \cap \hat{S}^{(b)} \neq \emptyset\} = \frac{\alpha}{B}\sum_{b=1}^{B}\frac{1\{C \cap \hat{S}^{(b)} \neq \emptyset\}}{m^{(b)}},$$

where the first inequality is a consequence of the Markov inequality and the last inequality is a consequence of the assumptions that $\mathbb{P}(\tilde{p}^{(b)} \leq \alpha) = \mathbb{P}(p^{(b)}m^{(b)} \leq \alpha) \leq \alpha/m^{(b)}$ and the definition $\tilde{p}^{(b)} = 1$ for $C \cap \hat{S}^{(b)} = \emptyset$. For a random variable U taking values in $[0, 1]$,

$$\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{1\{U \leq \alpha\gamma\}}{\gamma} = \begin{cases} 0, & U \geq \alpha \\ \alpha/U, & \alpha\gamma_{\min} \leq U < \alpha \\ 1/\gamma_{\min}, & U \leq \alpha\gamma_{\min}. \end{cases}$$

and if U has an uniform distribution on $[0, 1]$

$$\mathbb{E}\left(\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{1\{U \leq \alpha\gamma\}}{\gamma}\right) = \int_0^{\alpha\gamma_{\min}} \gamma_{\min}^{-1}dx + \int_{\alpha\gamma_{\min}}^{\alpha} \alpha x^{-1}dx$$

$$= \gamma_{\min}^{-1}x\big|_{x=0}^{x=\alpha\gamma_{\min}} + \alpha\log x\big|_{x=\alpha\gamma_{\min}}^{x=\alpha} = \alpha(1 - \log\gamma_{\min}).$$

We apply this using as U the uniform distributed $\tilde{p}^{(b)}/m^{(b)} = p^{(b)}$ for $C \cap S^{(b)} \neq \emptyset$ and obtain

$$\mathbb{E}\left(\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{1\{\tilde{p}^{(b)}/m^{(b)} \leq \alpha\gamma\}}{\gamma}\right) \leq \alpha(1 - \log\gamma_{\min}),$$

and similarly as above

$$\mathbb{P}\left(\inf_{\gamma \in (\gamma_{\min}, 1)} Q(\gamma) \leq \alpha\right) = \mathbb{E}\left(\sup_{\gamma \in (\gamma_{\min}, 1)} 1\{\pi(\alpha\gamma) \geq \gamma\}\right)$$

$$\leq \mathbb{E}\left(\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{1}{B}\sum_{b=1}^{B}\frac{1\{\tilde{p}^{(b)} \leq \alpha\gamma\}}{\gamma}\right)$$

$$= \mathbb{E}\left(\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{1}{B}\sum_{b=1}^{B}\frac{1\{\tilde{p}^{(b)} \leq \alpha\gamma\}1\{C \cap S^{(b)} \neq \emptyset\}}{\gamma}\right)$$

$$\leq \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\left(\sup_{\gamma \in (\gamma_{\min}, 1)} \frac{1\{\tilde{p}^{(b)}/m^{(b)} \leq \alpha\gamma/m^{(b)}\}1\{C \cap S^{(b)} \neq \emptyset\}}{\gamma}\right)$$

$$\leq (1 - \log\gamma_{\min})\frac{\alpha}{B}\sum_{b=1}^{B}\frac{1\{C \cap S^{(b)} \neq \emptyset\}}{m^{(b)}}$$

□

Additional empirical results

Table 2: Results of the simulation with the "small blocks"- and "large blocks"-design for 8 different correlations ρ, for single variable method with Bonferroni (SB) and with Bonferroni-Holm (SH), resp. hierarchical method with Bonferroni (HB) and with sequential rejection induced by the inheritance procedure (HSR).

ρ			FWER				# MTDs			# STDs
	SB	SH	HB	HSR	SB	SH	HB	HSR	HB	HSR
"small blocks"-design with low SNR (SNR = 4)										
0	0	0	0	0	9.57	9.69	9.53	9.63	9.42	9.53
0.4	0	0	0	0	8.84	9.06	8.65	8.81	8.36	8.51
0.7	0	0	0	0	5.87	6.26	7.28	7.60	5.65	6.13
0.8	0	0	0	0	5.53	5.76	6.79	7.22	5.33	5.89
0.85	0.03	0.04	0.03	0.04	2.97	3.08	5.21	5.56	2.82	3.14
0.9	0.01	0.01	0.01	0.01	3.35	3.55	5.49	5.86	3.22	3.60
0.95	0.46	0.47	0.46	0.48	1.02	1.11	4.04	4.07	0.9	0.99
0.99	0.55	0.56	0.54	0.58	3.62	3.78	6.01	6.27	3.42	3.71
"large blocks"-design with low SNR (SNR = 4)										
0	0	0	0	0	8.42	8.68	8.38	8.50	7.98	8.11
0.4	0	0	0	0	7.61	8.09	8.98	8.98	7.44	7.48
0.7	0	0	0	0	0.67	0.71	5.90	5.91	0.59	0.59
0.8	0	0	0	0	0.27	0.27	6.02	6.02	0.24	0.24
0.85	0	0	0	0	0	0	3.38	3.38	0	0
0.9	0	0	0.06	0.06	0.38	0.39	7.59	7.60	0.38	0.38
0.95	0.03	0.03	0.16	0.16	0.45	0.45	8.67	8.68	0.44	0.44
0.99	0.97	0.97	1.00	1.00	1.47	1.48	5.28	5.27	1.47	1.48
"small blocks"-design with high SNR (SNR = 8)										
0	0	0	0	0	9.87	9.89	9.90	9.90	9.86	9.86
0.4	0	0	0	0	10	10	10	10	10	10
0.7	0	0	0	0	10	10	10	10	10	10
0.8	0	0	0	0	9.85	9.89	9.98	9.98	9.90	9.91
0.85	0	0	0	0	9.26	9.38	9.89	9.92	9.39	9.53
0.9	0	0	0	0	9.59	9.65	10	10	9.67	9.79
0.95	0.21	0.23	0.21	0.28	8.36	8.46	9.82	9.78	8.36	8.61
0.99	0.92	0.93	0.92	0.95	6.72	6.85	8.06	8.04	6.73	6.99
"large blocks"-design with high SNR (SNR = 8)										
0	0	0	0	0	10	10	10	10	10	10
0.4	0	0	0	0	9.98	9.98	10	10	9.99	9.99
0.7	0	0	0	0	5.12	5.35	9.60	9.60	5.10	5.12
0.8	0	0	0	0	9.23	9.43	10	10	9.14	9.15
0.85	0	0	0	0	3.86	4.03	9.98	9.98	3.84	3.85
0.9	0	0	0	0	0.06	0.06	7.17	7.17	0.06	0.06
0.95	0	0	0	0	1.26	1.29	9.99	9.99	1.27	1.28
0.99	0.33	0.33	0.99	0.99	3.26	3.26	7.92	7.92	3.26	3.26

References

1. Dudoit S, van der Laan MJ. Multiple testing procedures with applications to genomics. Berlin: Springer Science & Business Media, 2007.
2. Efron B. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, volume 1. Cambridge: Cambridge University Press, 2010.

3. Westfall PH. Resampling-based multiple testing: Examples and methods for p-value adjustment, volume 279. New York: John Wiley & Sons, 1993.
4. Bühlmann P. Statistical significance in high-dimensional linear models. Bernoulli 2013;19:1212–42.
5. Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 2014;15:2869–909.
6. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. Ann Stat 2014;42:413–68.
7. Meinshausen N, Meier L, Bühlmann P. p-Values for High-Dimensional Regression. J Am Stat Assoc 2009;104:1671–81.
8. van de Geer S, Bühlmann P, Ritov Y, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Stat 2014;42:1166–202.
9. Wasserman L, Roeder K. High dimensional variable selection. Ann Stat 2009;37:2178–201.
10. Zhang C-H, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat SocSeries B 2014;76:217–42.
11. Bühlmann P, Kalisch M, Meier L. High-dimensional statistics with a view toward applications in biology. Ann Rev Stat Appl 2014;1:255–78.
12. Dezeure R, Bühlmann P, Meier L, Meinshausen N. High-dimensional Inference: Confidence intervals, p-values and R-software hdi. To appear in Statistical Science. Preprint arXiv:1408.4026v1.
13. Bühlmann P, van de Geer S. Statistics for High-Dimensional Data: Methods, Theory and Applications. New York, NY: Springer Verlag, 2011.
14. Bühlmann P, Rütimann P, van de Geer S, Zhang C-H. Correlated variables in regression: clustering and sparse estimation (with discussion). J Stat Plann Inference 2013;143:1835–71.
15. Mandozzi J, Bühlmann P. Hierarchical testing in the high-dimensional setting with correlated variables. J Am Stat Assoc 2015. Published online (DOI: 10.1080/01621459.2015.1007209).
16. Meinshausen N. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. J R Stat Soc Series B 2015;77:923–45.
17. Goeman JJ, Solari A. The sequential rejection principle of familywise error control. Ann Stat 2010;38:3782–810.
18. Goeman JJ, Finos L. The inheritance procedure: Multiple testing of tree-structured hypotheses. Stat Appl Genet Mol Biol 2012;11:1–18.
19. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B 1996;58:267–88.
20. Bühlmann P, Mandozzi J. High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. Comput Stat 2014;29:407–30.
21. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936;8:3–62.
22. Dunn OJ. Multiple Comparisons among Means. J Am Stat Assoc 1961;56:52–64.
23. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70.
24. Meinshausen N. Hierarchical testing of variable importance. Biometrika 2008;95:265–78.
25. Shaffer JP. Modified sequentially rejective multiple test procedures. J Am Stat Assoc 1986;81:826–31.