# Supplemental Materials for "Hierarchical Testing in the High-Dimensional Setting with Correlated Variables"

Jacopo Mandozzi and Peter Bühlmann

Seminar for Statistics, ETH Zürich

January 9, 2015

## Supplemental material to Section 2

### An alternative bottom-up hierarchical adjustment

The procedure described in Section 2 is based on a top-down hierarchical adjustment of the p-values $P_h^C = \max_{D \in \mathcal{T}: C \subseteq D} P^C$. Another possibility is the following bottom-up approach.

We begin with clustering as in Section 2.1 and screening as in Section 2.2. Then we take the p-values $p^{C,(b)}$ as in (2) and define

$$\overline{p}_h^{C,(b)} = \min\{\, 2|\hat{S}^{(b)}| \min_{D \in \mathcal{T}: D \subseteq C} p^{C,(b)} \, , \, 1\}.$$

Finally we define for $\gamma \in (0,1)$ the aggregated p-values

$$\overline{Q}_h^C(\gamma) = \min\big\{\, 1 \ , \ q_\gamma\big(\{\overline{p}_h^{C,(b)}/\gamma;\, b = 1, \ldots, B\}\big)\big\}$$

and eliminate $\gamma$ taking

$$\overline{P}_h^C = \min\big\{\, 1 \ , \ (1 - \log \gamma_{\min}) \inf_{\gamma \in (\gamma_{\min}, 1)} \overline{Q}_h^C(\gamma)\big\}.$$

The price one has to pay for minimizing among p-values of children clusters instead of maximizing among p-values of parents clusters is a factor $|C \cap \hat{S}^{(b)}|$ in the multiplicity adjustment.

Although none of the two methods theoretically dominates the other, simulations with some scenarios as in Section 4 have shown that the top-down method exhibits substantially higher power than the bottom-up method. Hence we put our focus on the top-down method.

## Supplemental material to Section 4

### Variability of Performance 1 and Performance 2 in the simulation study

To give some idea about the variability among the different simulation runs, we show in Figures 4 and 5 the Performance 1 and Performance 2 measures, respectively for all 100 runs of some of the scenarios.

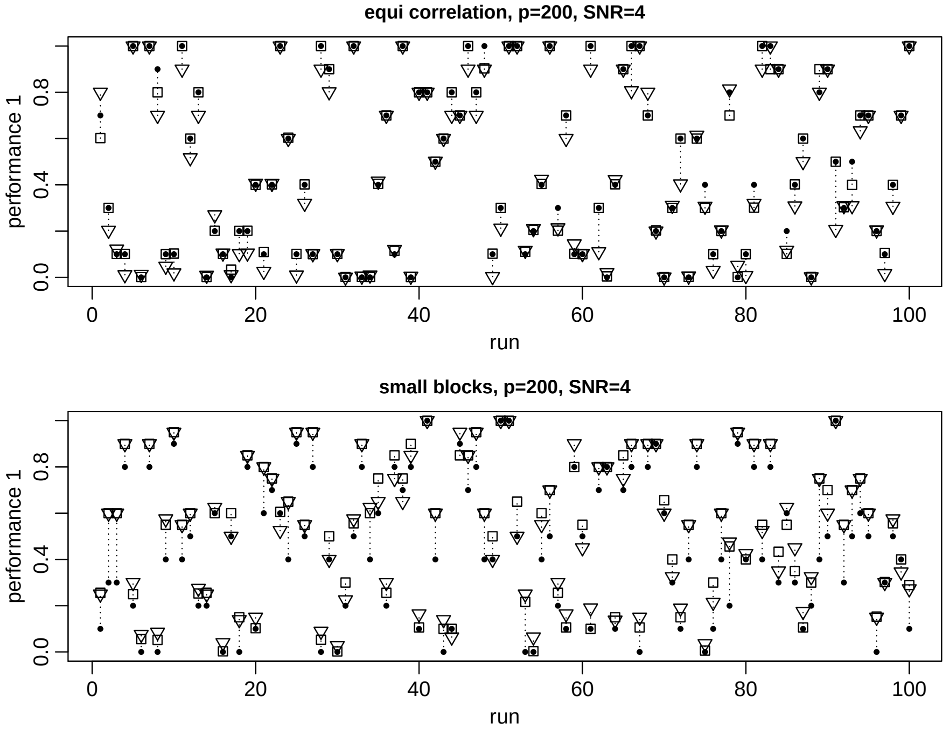

Figure 4: The Performance 1 measure for all 100 runs for 2 different scenarios described in the header of the plots. Single variable method (filled small circle), the hierarchical method with canonical correlation clustering (empty square) and `hclust` clustering (triangle).

In Figure 4 we consider Performance 1 for two synthetic scenarios, one where the single variable method is favored and another where the hierarchical method is better. In Figure 5 we adopt the same approach for Performance 2 considering two scenarios based on semi-real datasets.

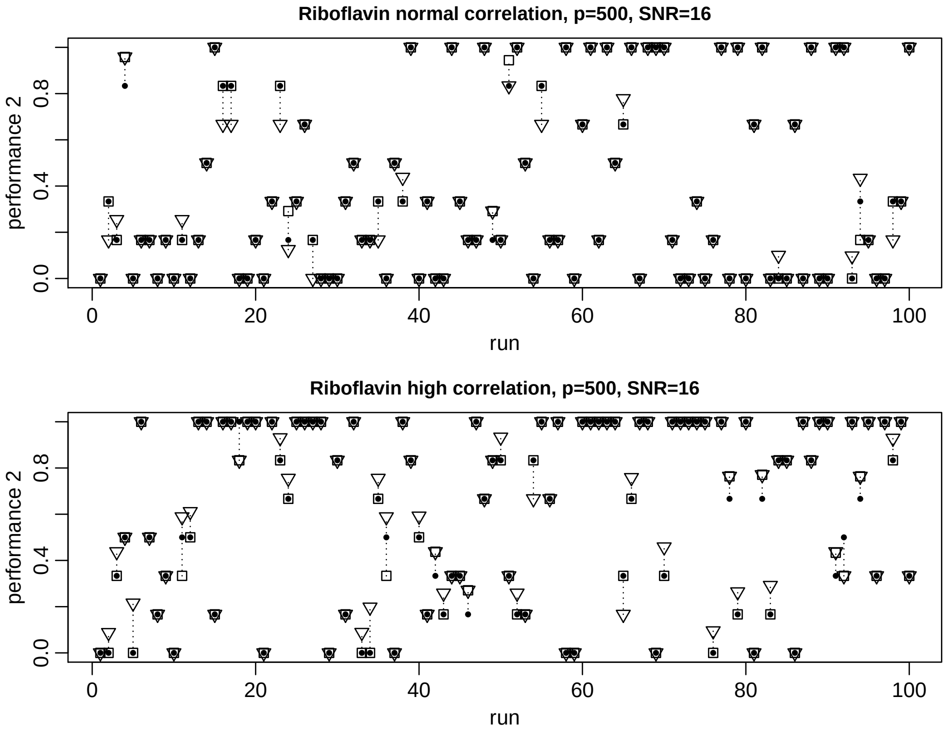

Figure 5: The Performance 2 measure for all 100 runs for 2 different scenarios described in the header of the plots. Single variable method (filled small circle), the hierarchical method with canonical correlation clustering (empty square) and `hclus` clustering (triangle).

## Variability of MTDs in Section 4.3

We show in Figures 6 and 7 the number of MTDs for all simulation runs of the "small blocks"-design with SNR = 8 and $\rho = 0.8$ and $\rho = 0.9$, respectively, and for the "large blocks"-design with SNR = 8 and $\rho = 0.7$ and $\rho = 0.95$, respectively. For each of the 100 simulation runs and cardinalities from 1 to 20, the number of MTDs for the hierarchical method with `hclus` clustering is depicted in black while the number of MTDs for the single variable method is depicted in gray, for graphical convenience at the bottom of the y-axis (since the cardinality of the MTDs of the single variable method is always equal to 1).

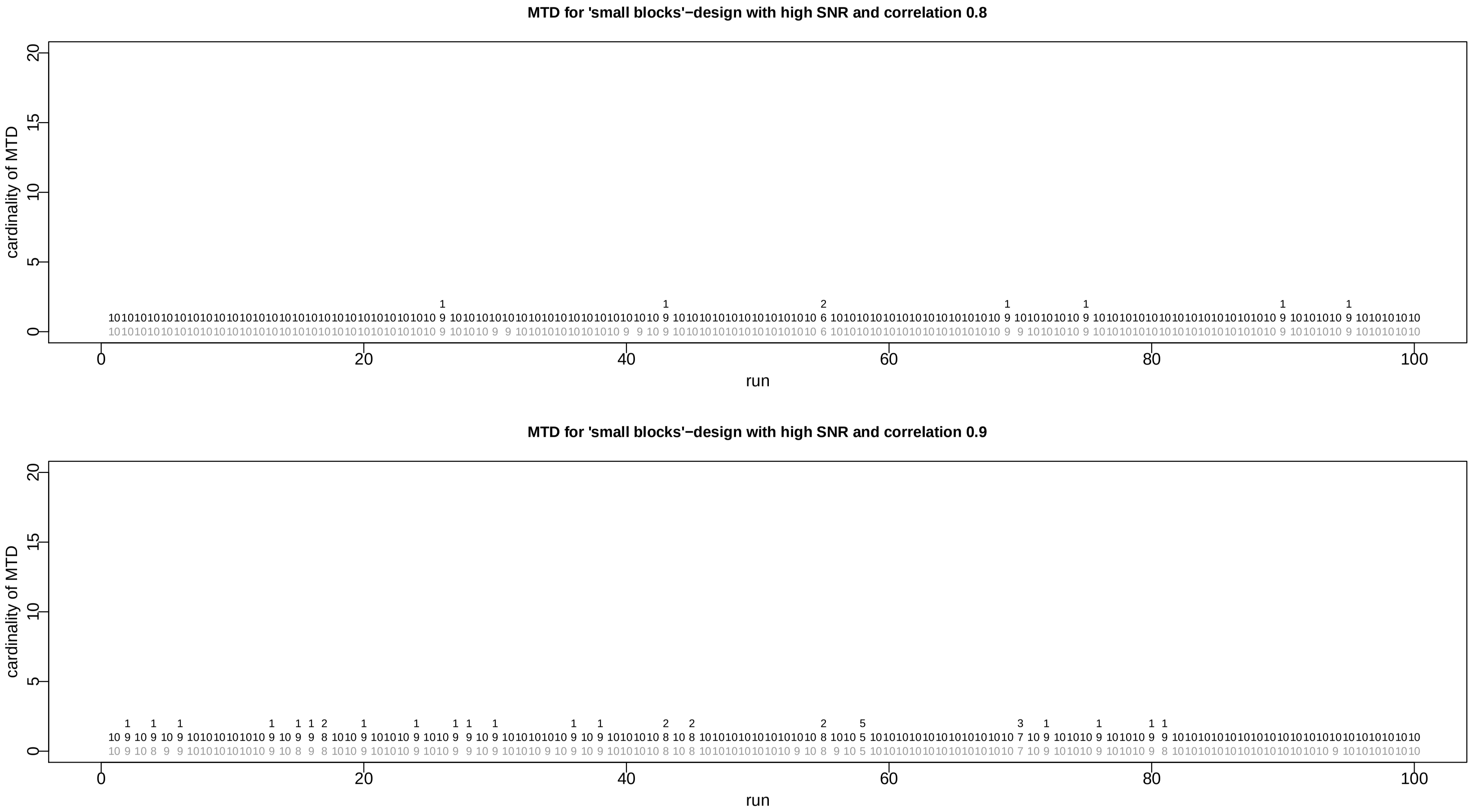

Figure 6: Number of MTDs for “small blocks”-design with high SNR (SNR=8) and $\rho = 0.8$ resp. $\rho = 0.9$. For each of the 100 simulation runs (x-axis) and every cardinality (y-axis), the number of MTDs for the hierarchical method with `hclus` clustering (in black) and for the single variable method (in gray, for graphical convenience at the bottom of the y-axis).

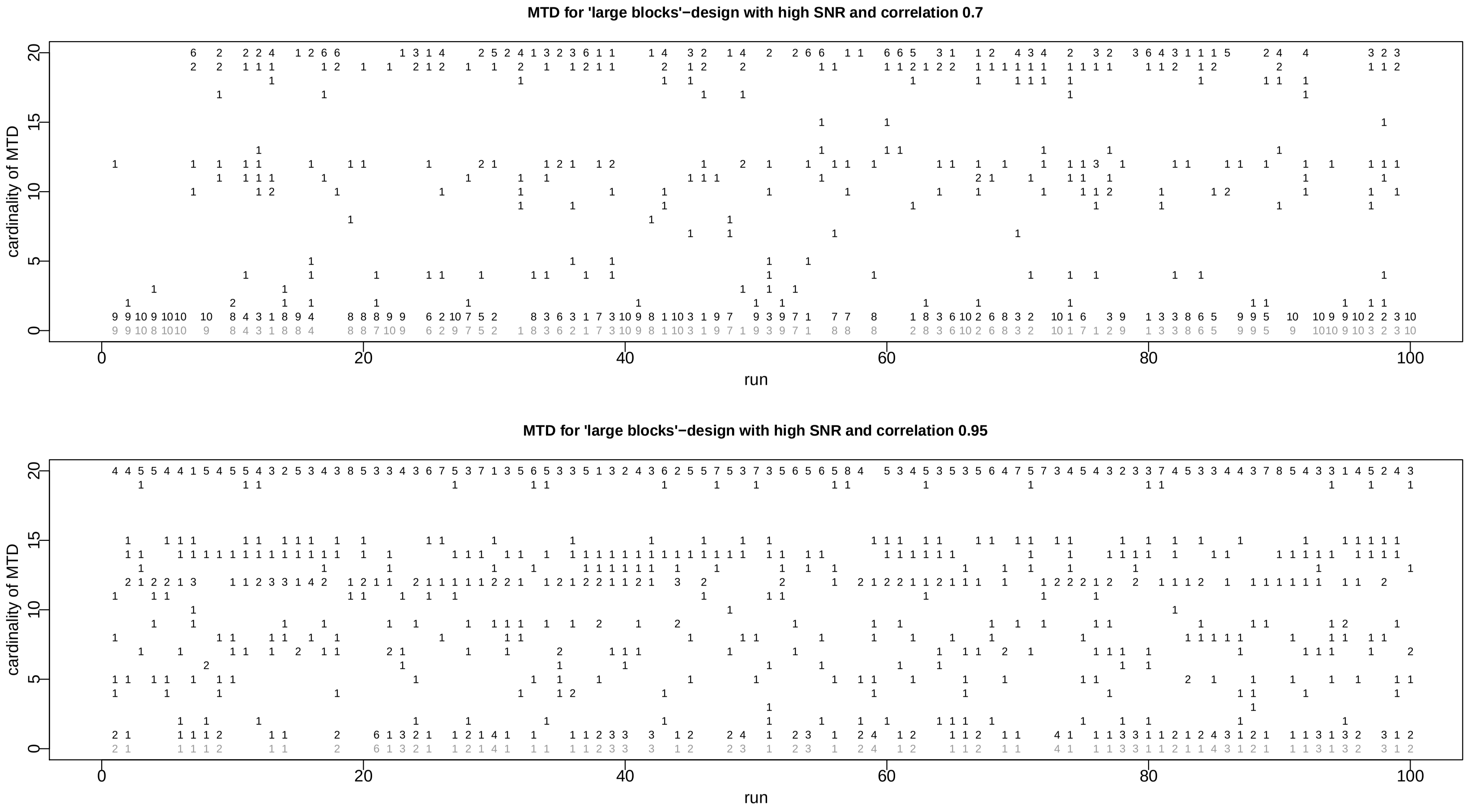

Figure 7: Number of MTDs for “large blocks”-design with high SNR (SNR=8) and $\rho = 0.7$ resp. $\rho = 0.95$. For each of the 100 simulation runs (x-axis) and every cardinality (y-axis), the number of MTDs for the hierarchical method with `hclus` clustering (in black) and for the single variable method (in gray, for graphical convenience at the bottom of the y-axis).

### Extension of the considerations of Section 4.3 for low SNR

We present here the same detailed analysis as in Section 4.3 for the signal to noise ratio SNR $= 4$. The empirical results presented below show that the power of all considered methods is significantly affected by the change of SNR (e.g. for the "large blocks"-design with $\rho \geq 0.7$ detecting at least one singleton is difficult when SNR $= 4$), but they also confirm the superiority of the hierarchical in comparison to the single variable methods reported in the main paper in Section 4.3.

Table 5 reports some average results over 100 simulation runs. As for the case in the main paper with high SNR, the number of singleton detections are again similar for all methods. The large number of MTDs with cardinality 2 in the "small blocks"-design emphasizes the powerful advantage of automatically going to the finer possible resolution with the hierarchical method.

To better illustrate what happens in a typical simulation run, we show in Figure 8 the dendrograms for a representative simulation run of the "large blocks"-design with $\rho = 0.9$ (here with SNR $= 4$), for the single variable method and the hierarchical method with `hclus` clustering. The active variables are labeled in black and the truly detected non-zero variables along the hierarchy are depicted in black. While the single variable method "only" detects one singleton, the hierarchical method detects the same singleton and achieves 6 more MTDs.

Figure 9 is the analogous of Figure 8 in the main paper for a typical run of the "small blocks"-design with $\rho = 0.7$. It shows that the hierarchical method improves the results of the single variable method (which detects 5 singletons) providing 2 more MTDs of cardinality 2.

In Figure 10 we show the number of MTDs for all 100 simulation runs of the "small blocks"-design with SNR $= 4$ and $\rho = 0.7$ and $\rho = 0.9$, respectively. In Figure 11 we do the same for the "large blocks"-design with SNR $= 4$. For each simulation run and cardinalities from 1 to 20, the number of MTDs for the hierarchical method with `hclus` clustering is depicted in black while the number of MTDs for the single variable method is depicted in gray, for graphical convenience at the bottom of the y-axis (since the cardinality of MTDs of the single variable method is always equal to 1).

Finally, we illustrate in Figure 12 the true positive (TPR) rates and false positive rates (FPR) of the Lasso, the single variable and the hierarchical method with `hclus` clustering as points in the

| $\rho$ | $\delta$ | FWER | | | # MTD | | | # MTD for given cardinality | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | $\lvert \cdot \rvert = 1$ | | | $\lvert \cdot \rvert = 2$ | | $3 \leq \lvert \cdot \rvert \leq 10$ | | $11 \leq \lvert \cdot \rvert \leq 20$ | |
| | | S | C | H | S | C | H | S | C | H | C | H | C | H | C | H |
| "small blocks"-design with low SNR | | | | | | | | | | | | | | | | |
| 0 | 0.15 | 0 | 0 | 0 | 9.57 | 9.60 | 9.53 | 9.57 | 9.52 | 9.42 | 0 | 0 | 0.03 | 0.06 | 0.05 | 0.01 |
| 0.4 | 0.20 | 0 | 0 | 0 | 8.84 | 8.91 | 8.65 | 8.84 | 8.82 | 8.36 | 0.04 | 0.05 | 0 | 0.08 | 0.05 | 0.03 |
| 0.7 | 0.33 | 0 | 0 | 0 | 5.87 | 7.18 | 7.28 | 5.87 | 6.01 | 5.65 | 1.00 | 0.98 | 0.01 | 0.21 | 0.14 | 0.14 |
| 0.8 | 0.47 | 0 | 0 | 0 | 5.53 | 6.64 | 6.79 | 5.53 | 5.56 | 5.33 | 0.89 | 0.88 | 0.19 | 0.55 | 0 | 0.02 |
| 0.85 | 0.70 | 0.03 | 0.03 | 0.03 | 2.97 | 4.72 | 5.21 | 2.97 | 2.99 | 2.82 | 1.34 | 1.11 | 0.33 | 0.95 | 0.03 | 0.17 |
| 0.9 | 0.63 | 0.01 | 0.01 | 0.01 | 3.35 | 5.26 | 5.49 | 3.35 | 3.37 | 3.22 | 1.38 | 1.27 | 0.47 | 0.95 | 0.04 | 0 |
| 0.95 | 0.97 | 0.46 | 0.46 | 0.46 | 1.02 | 2.85 | 4.04 | 1.02 | 1.02 | 0.90 | 1.49 | 1.34 | 0.34 | 1.66 | 0 | 0.07 |
| 0.99 | 0.97 | 0.55 | 0.55 | 0.54 | 3.62 | 5.81 | 6.01 | 3.62 | 3.62 | 3.42 | 1.94 | 1.89 | 0 | 0.66 | 0 | 0 |
| "large blocks"-design with low SNR | | | | | | | | | | | | | | | | |
| 0 | 0.27 | 0 | 0 | 0 | 8.42 | 8.48 | 8.38 | 8.42 | 8.35 | 7.98 | 0 | 0.01 | 0.02 | 0.08 | 0 | 0.08 |
| 0.4 | 0.21 | 0 | 0 | 0 | 7.61 | 7.89 | 8.98 | 7.61 | 7.60 | 7.44 | 0 | 0.13 | 0 | 0.42 | 0.17 | 0.91 |
| 0.7 | 0.66 | 0 | 0 | 0 | 0.67 | 3.54 | 5.90 | 0.67 | 0.67 | 0.59 | 0 | 0.01 | 0.19 | 0.30 | 2.28 | 3.99 |
| 0.8 | 0.80 | 0 | 0 | 0 | 0.27 | 4.58 | 6.02 | 0.27 | 0.27 | 0.24 | 0 | 0.02 | 0.27 | 0.47 | 3.56 | 3.64 |
| 0.85 | 0.97 | 0 | 0 | 0 | 0 | 1.58 | 3.38 | 0 | 0 | 0 | 0.01 | 0.01 | 0 | 0.04 | 0.79 | 0.75 |
| 0.9 | 0.94 | 0 | 0.07 | 0.06 | 0.38 | 7.07 | 7.59 | 0.38 | 0.38 | 0.38 | 0 | 0.01 | 0.53 | 0.63 | 6.01 | 5.91 |
| 0.95 | 1.00 | 0.03 | 0.40 | 0.16 | 0.45 | 8.34 | 8.67 | 0.45 | 0.45 | 0.44 | 0 | 0 | 1.19 | 1.45 | 6.61 | 6.68 |
| 0.99 | 1.00 | 0.97 | 1.00 | 1.00 | 1.47 | 5.83 | 5.28 | 1.47 | 1.47 | 1.47 | 0.01 | 0.17 | 1.73 | 2.24 | 2.62 | 1.40 |

Table 5: Results of the simulation with the "large blocks"- and "small blocks" -design with low SNR (SNR=4) for different correlations. $\rho$ is the correlation in the design, $\delta$ the relative frequency of screenings with $\hat{S} \not\supset S_0$, MTD denotes "minimal true detections", "$2 \leq \lvert \cdot \rvert \leq 5$" indicates that MTD of cardinality between 2 and 5 are considered, S, C and H represent the "single variable" resp. "canonical correlation clustering" and "hierarchical with `hclus` clustering" method.

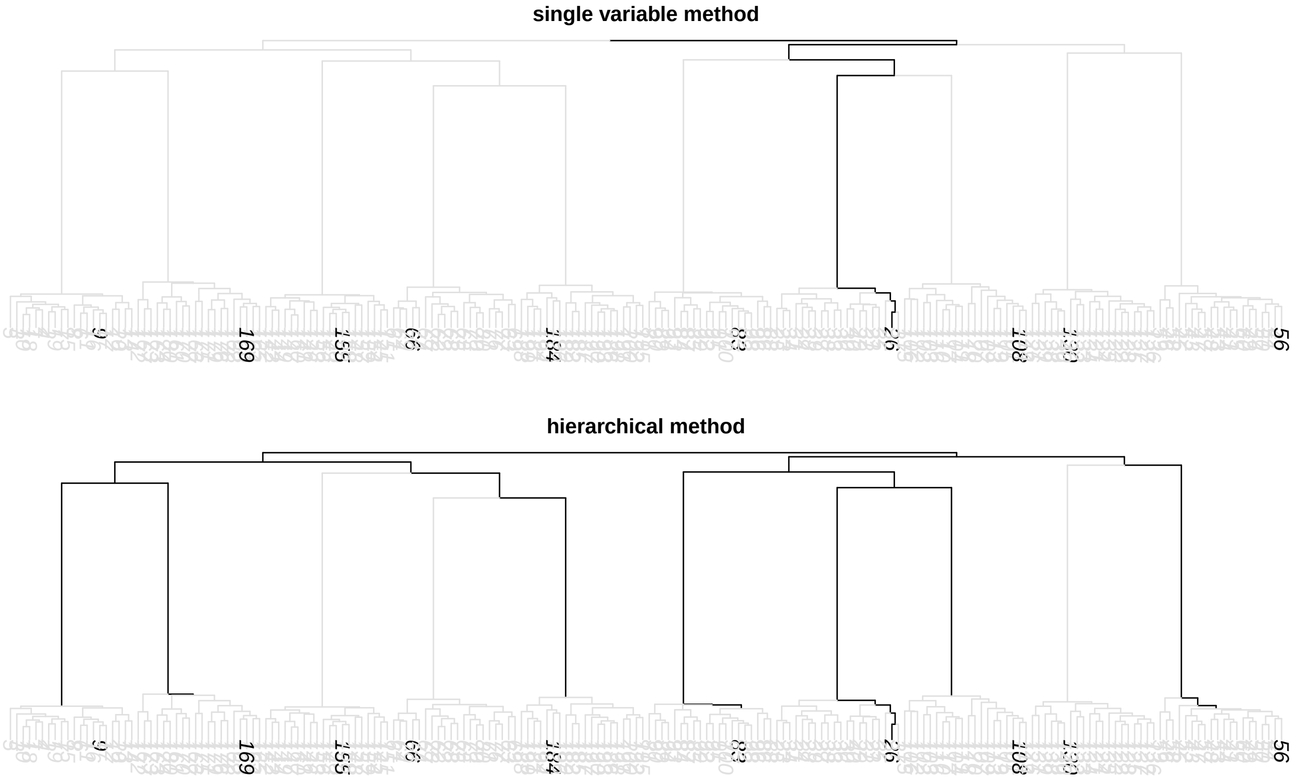

Figure 8: Dendrograms for a representative run of the “large blocks”-design with low SNR (SNR=4) and $\rho = 0.9$. The active variables are labeled in black and the truly detected non-zero variables along the hierarchy are depicted in black.

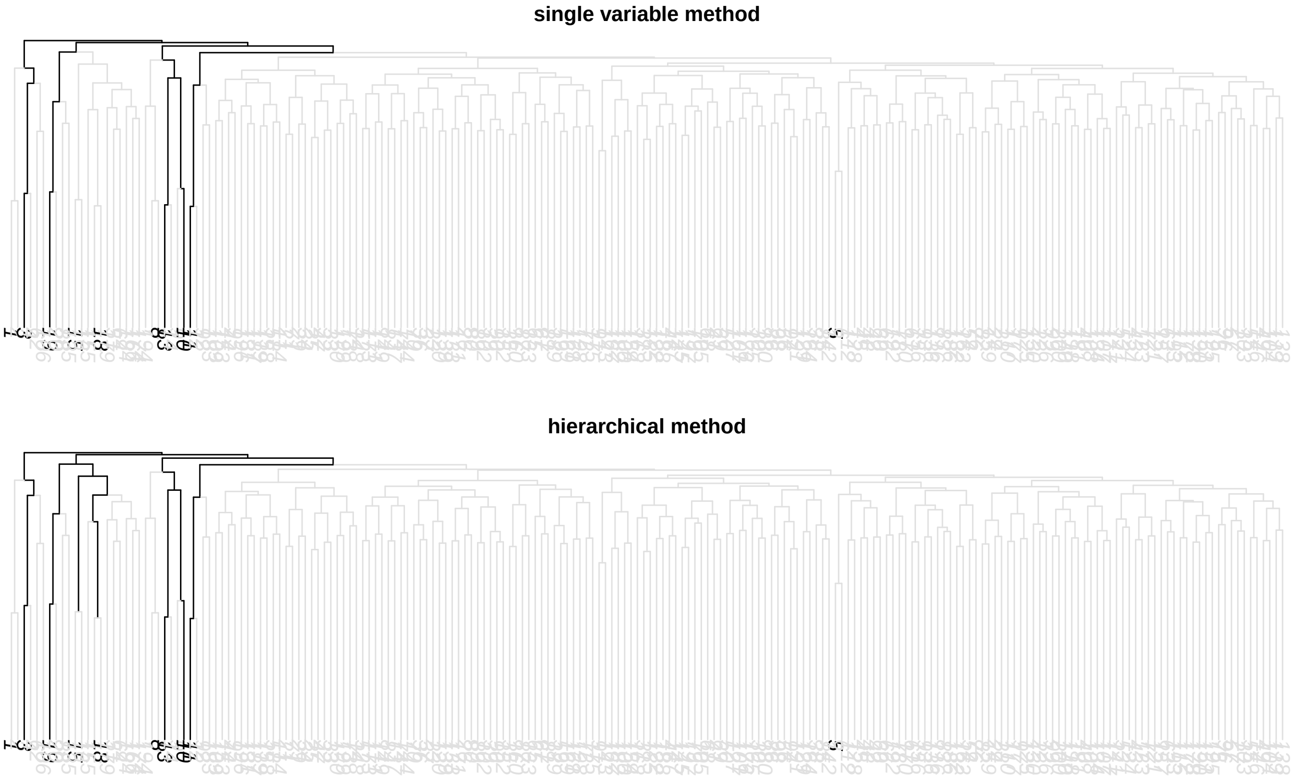

Figure 9: Dendrograms for a representative run of the "small blocks"-design with low SNR (SNR=4) and $\rho = 0.7$. The active variables are labeled in black and the truly detected non-zero variables along the hierarchy are depicted in black.

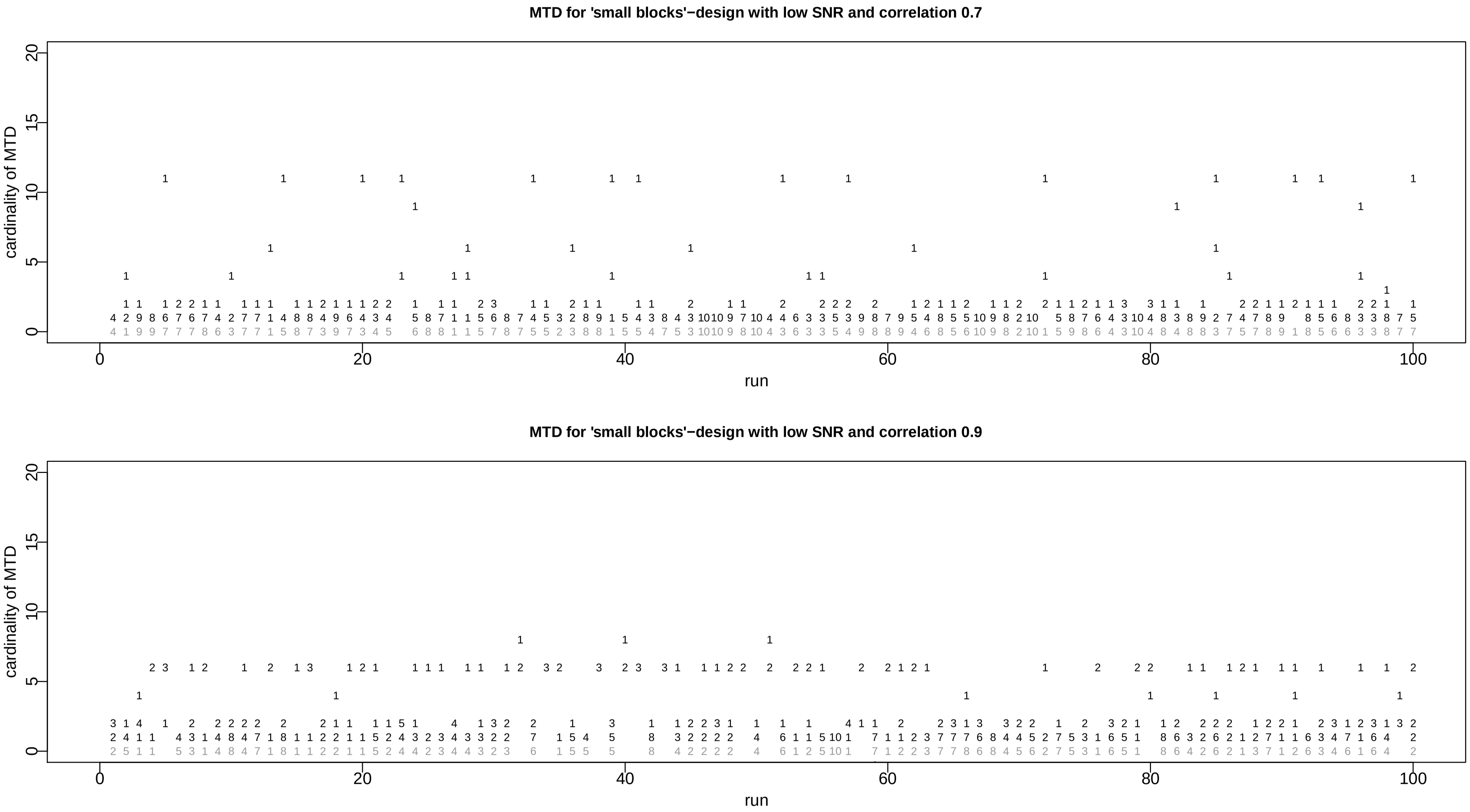

Figure 10: Number of MTDs for “small blocks”-design with high SNR (SNR=4) and $\rho = 0.7$ resp. $\rho = 0.9$. For each of the 100 simulation runs (x-axis) and every cardinality (y-axis), the number of MTDs for the hierarchical method with `hclus` clustering (in black) and for the single variable method (in gray, for graphical convenience at the bottom of the y-axis).

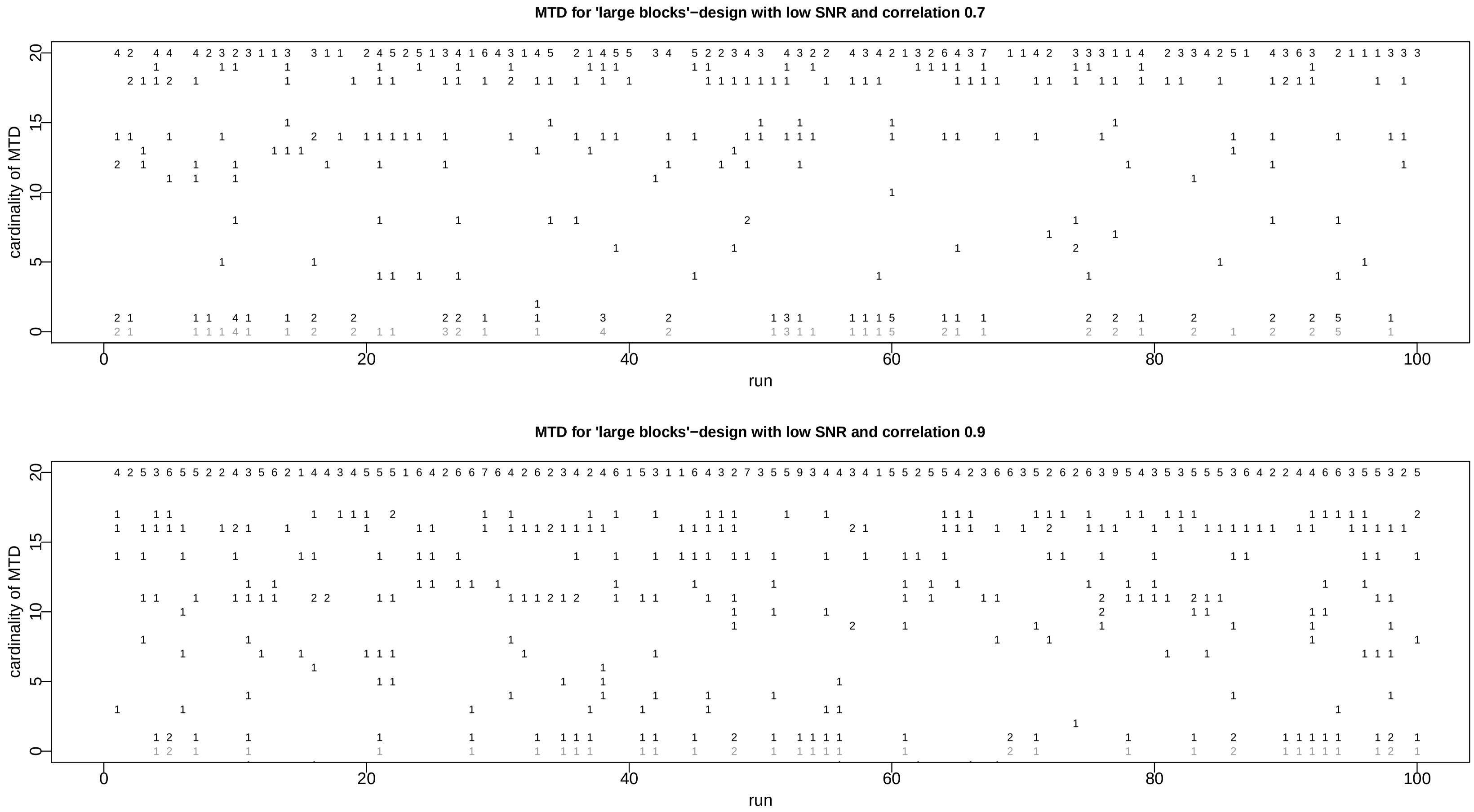

Figure 11: Number of MTDs for "large blocks"-design with high SNR (SNR=4) and $\rho = 0.7$ resp. $\rho = 0.9$. For each of the 100 simulation runs (x-axis) and every cardinality (y-axis), the number of MTDs for the hierarchical method with `hclus` clustering (in black) and for the single variable method (in gray, for graphical convenience at the bottom of the y-axis).

ROC space.

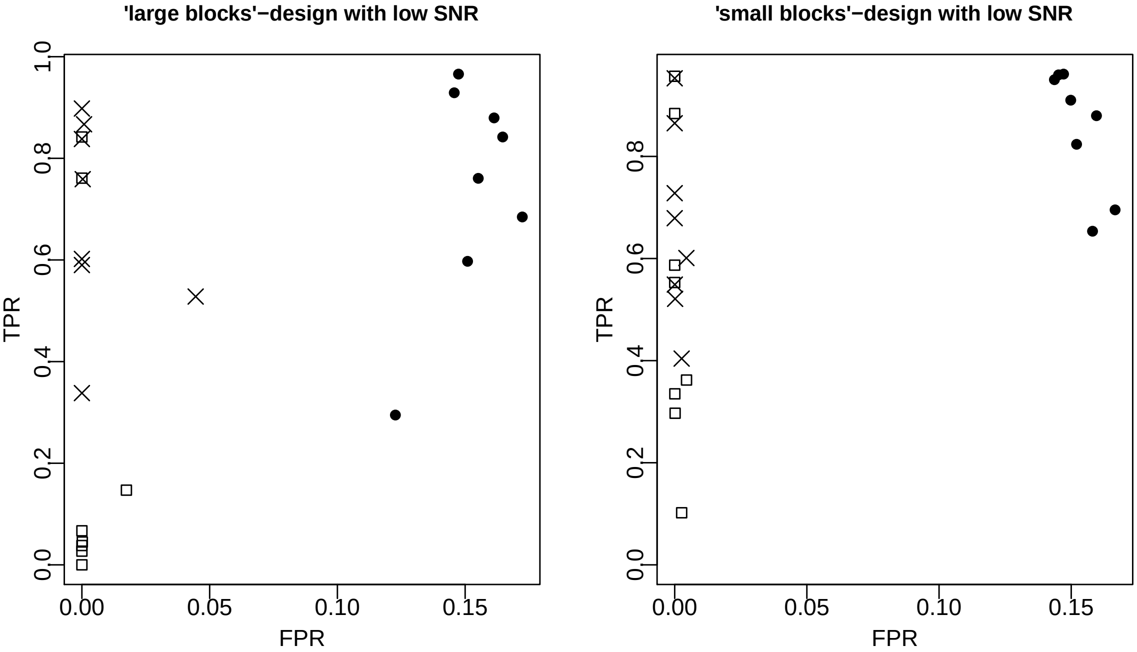

Figure 12: True positive rate (TPR) and false positive rate (FPR) for the Lasso (bullet), the single variable method (box), and the hierarchical method with `hclust` clustering (cross) for different scenarios as indicated in the header of the plots.

Comparing Figure 12 with Figure 3 in the main paper, we see that the negative impact of low SNR is more striking on the TPR then on the FPR which remains very similar. Regarding a comparison of the methods, the same conclusions as for high SNR = 8 can be drawn: the single variable and hierarchical method do much better than the Lasso in terms of FPR. The price one has to pay for the higher reliability is a lower TPR and the hierarchical method improves the TPR of the single variable method to the level of the Lasso (when considering MTDs).

# Proofs

## Proof of Theorem 1

Our proof is following ideas from the proofs of Theorem 3.1-3.2 in Meinshausen et al. (2009) and the proof Theorem 1 in Meinshausen (2008).

**Proof of first assertion of Theorem 1.**

First note that

$$\begin{aligned}\mathbb{P}(\mathcal{T}_{rej}^{\gamma} \cap \mathcal{T}_0 \neq \emptyset) &= \mathbb{P}(\exists C \in \mathcal{T}_0 \, : \, Q_h^C(\gamma) \leq \alpha) \\ &= \mathbb{P}(\exists C \in \tilde{\mathcal{T}}_0 \, : \, Q_h^C(\gamma) \leq \alpha)\end{aligned}$$

where $\tilde{\mathcal{T}}_0$ is the set of all clusters which fulfill the null hypothesis and are maximal in the sense that

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

This holds, since a direct consequence of the definition of the hierarchically adjusted p-values $Q_h^C(\cdot)$ is that $Q_h^{C'}(\gamma) \leq Q_h^C(\gamma)$ for $C \subset C'$ and hence an error committed on a cluster $C \in \mathcal{T}_0 \backslash \tilde{\mathcal{T}}_0$ implies an error in a set $C' \in \tilde{\mathcal{T}}_0$, with $C \subset C'$. Moreover, since $Q_h^C(\gamma) \geq Q^C(\gamma)$,

$$\mathbb{P}(\exists C \in \tilde{\mathcal{T}}_0 \, : \, Q_h^C(\gamma) \leq \alpha) \leq \mathbb{P}(\exists C \in \tilde{\mathcal{T}}_0 \, : \, Q^C(\gamma) \leq \alpha) = \mathbb{P}(\min_{C \in \tilde{\mathcal{T}}_0} Q^C(\gamma) \leq \alpha).$$

Hence it remains to show that

$$\mathbb{P}(\min_{C \in \tilde{\mathcal{T}}_0} Q^C(\gamma) \leq \alpha) \leq \alpha.$$

We consider the event

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

where all screenings are satisfied. Because of the $\delta$-screening assumption it holds

$$P(\mathcal{A}) \geq (1-\delta)^B.$$

In the following we omit the function $\min\{1, \cdot\}$ from the definition of $Q^C(\gamma)$ in order to simplify the notation (this is possible since the level $\alpha$ is smaller than 1). Define for $u \in (0,1)$ the function

$$\pi^C(u) := \frac{1}{B} \sum_{b=1}^{B} 1\{p_{adj}^{C,(b)} \leq u\}.$$

Then it holds

$$
\begin{aligned}
Q^C(\gamma) \le \alpha \quad &\Longleftrightarrow \quad q_\gamma(\{p_{adj}^{C,(b)}/\gamma;\, b = 1, \dots, B\}) \le \alpha \\
&\Longleftrightarrow \quad q_\gamma(\{p_{adj}^{C,(b)};\, b = 1, \dots, B\}) \le \alpha\gamma \\
&\Longleftrightarrow \quad \sum_{b=1}^{B} 1\{p_{adj}^{C,(b)} \le \alpha\gamma\} \ge B\gamma \\
&\Longleftrightarrow \quad \pi^C(\alpha\gamma) \ge \gamma.
\end{aligned}
$$

Thus

$$
\begin{aligned}
\mathbb{P}(\min_{C \in \tilde{\mathcal{T}}_0} Q^C(\gamma) \le \alpha) \quad &\le \quad \sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{E}(1\{Q^C(\gamma) \le \alpha\}) \\
&= \quad \sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{E}(1\{\pi^C(\alpha\gamma) \ge \gamma\}) \\
&\le \quad \frac{1}{\gamma} \sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{E}(\pi^C(\alpha\gamma)),
\end{aligned}
$$

where for the last inequality a Markov inequality was used. Now, using the definition of $\pi^C(\cdot)$,

$$
\begin{aligned}
\frac{1}{\gamma} \sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{E}(\pi^C(\alpha\gamma)) \quad &= \quad \frac{1}{\gamma} \sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{E}(\frac{1}{B} \sum_{b=1}^{B} 1\{p_{adj}^{C,(b)} \le \alpha\gamma\}) \\
&= \quad \frac{1}{\gamma} \frac{1}{B} \sum_{b=1}^{B} \sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{E}(1\{p_{adj}^{C,(b)} \le \alpha\gamma\}) \\
&= \quad \frac{1}{\gamma} \frac{1}{B} \sum_{b=1}^{B} \sum_{C \in \tilde{\mathcal{T}}_0,\, \hat{S}^{(b)} \cap C \ne \emptyset} \mathbb{E}(1\{p_{adj}^{C,(b)} \le \alpha\gamma\})
\end{aligned}
$$

where the last equality holds since $p_{adj}^{C,(b)} = 1$ if $\hat{S}^{(b)} \cap C = \emptyset$. Now, for $C$ such that $\hat{S}^{(b)} \cap C \ne \emptyset$

and on $\mathcal{A}$

$$\begin{aligned}
\mathbb{E}(1\{p_{adj}^{C,(b)} \le \alpha\gamma\}) &= \mathbb{P}(p_{adj}^{C,(b)} \le \alpha\gamma) \\
&= \mathbb{P}\Big(p^{C,(b)} \frac{|\hat{S}^{(b)}|}{|C \cap \hat{S}^{(b)}|} \le \alpha\gamma\Big) \\
&= \mathbb{P}\Big(p^{C,(b)} \le \alpha\gamma \frac{|C \cap \hat{S}^{(b)}|}{|\hat{S}^{(b)}|}\Big) \\
&\le \alpha\gamma \frac{|C \cap \hat{S}^{(b)}|}{|\hat{S}^{(b)}|}.
\end{aligned}$$

This is a consequence of the uniform distribution of the p-values $p^{C \cap \hat{S}^{(b)}}$ given $S \subseteq \hat{S}^{(b)}$ and the sample split $\{1, \ldots, n\} = N_{in}^{(b)} \sqcup N_{out}^{(b)}$. We can hence conclude that on $\mathcal{A}$

$$\begin{aligned}
\mathbb{P}(\min_{C \in \tilde{\mathcal{T}}_0} Q^C(\gamma) \le \alpha) &\le \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B} \sum_{C \in \tilde{\mathcal{T}}_0,\, \hat{S}^{(b)} \cap C \neq \emptyset} \alpha\gamma \frac{|C \cap \hat{S}^{(b)}|}{|\hat{S}^{(b)}|} \\
&= \alpha \frac{1}{B} \sum_{b=1}^{B} \frac{1}{|\hat{S}^{(b)}|} \sum_{C \in \tilde{\mathcal{T}}_0,\, \hat{S}^{(b)} \cap C \neq \emptyset} |C \cap \hat{S}^{(b)}| \\
&\le \alpha \frac{1}{B} \sum_{b=1}^{B} 1 \le \alpha,
\end{aligned}$$

since by definition the sets in $\tilde{\mathcal{T}}_0$ are disjoint and hence

$$\sum_{C \in \tilde{\mathcal{T}}_0,\, \hat{S}^{(b)} \cap C \neq \emptyset} |C \cap \hat{S}^{(b)}| \le |\hat{S}^{(b)}|.$$

Finally we have

$$\begin{aligned}
&\mathbb{P}(\min_{C \in \tilde{\mathcal{T}}_0} Q^C(\gamma) \le \alpha) = \\
= \;& \mathbb{P}(\min_{C \in \tilde{\mathcal{T}}_0} Q^C(\gamma) \le \alpha \,|\, \mathcal{A})\, P(\mathcal{A}) + \mathbb{P}(\min_{C \in \tilde{\mathcal{T}}_0} Q^C(\gamma) \le \alpha \,|\, \mathcal{A}^c)\, P(\mathcal{A}^c) \\
\le \;& \alpha + 1 - (1-\delta)^B
\end{aligned}$$

**Proof of second assertion of Theorem 1.**

We show that

$$\mathbb{P}(\exists C \in \mathcal{T}_0 \,:\, P_h^C \leq \alpha) \leq \alpha.$$

Defining $\tilde{\mathcal{T}}_0$ as in the proof of Theorem 1 and using similar arguments as there we obtain

$$\mathbb{P}(\exists C \in \mathcal{T}_0 \,:\, P_h^C \leq \alpha) = \mathbb{P}(\exists C \in \tilde{\mathcal{T}}_0 \,:\, P_h^C \leq \alpha) \leq \mathbb{P}(\exists C \in \tilde{\mathcal{T}}_0 \,:\, P^C \leq \alpha).$$

As in the proof of Theorem 1 we consider the event

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

with $P(\mathcal{A}) \geq (1-\delta)^B$. The uniform distribution of the p-values $p_{\mathrm{partial\,F-test}}^{C \cap \hat{S}^{(b)}}$ given $S \subseteq \hat{S}^{(b)}$ and the sample split $\{1, \ldots, n\} = N_{in}^{(b)} \sqcup N_{out}^{(b)}$, together with the fact that sets in $\hat{S}^{(b)}$ are disjoint, provides on $\mathcal{A}$

$$\mathbb{E}\Big(\frac{1\{p^{C,(b)} \leq \alpha\gamma\}}{\gamma}\Big) = \frac{1}{\gamma}\mathbb{P}(p^{C,(b)} \leq \alpha\gamma) \leq \alpha.$$

Moreover, on $\mathcal{A}$

$$\begin{aligned}
\mathbb{E}\Big(\max_{C \in \tilde{\mathcal{T}}_0} \frac{1\{p_{adj}^{C,(b)} \leq \alpha\gamma\}}{\gamma}\Big) &\leq \mathbb{E}\Big(\sum_{C \in \tilde{\mathcal{T}}_0} \frac{1\{p_{adj}^{C,(b)} \leq \alpha\gamma\}}{\gamma}\Big) \\
&\leq \mathbb{E}\Big(\sum_{C \in \tilde{\mathcal{T}}_0,\, \hat{S}^{(b)} \cap C \neq \emptyset} \frac{1\{p_{adj}^{C,(b)} \leq \alpha\gamma\}}{\gamma}\Big) \\
&= \frac{1}{\gamma} \sum_{C \in \tilde{\mathcal{T}}_0,\, \hat{S}^{(b)} \cap C \neq \emptyset} \mathbb{P}(1\{p_{adj}^{C,(b)} \leq \alpha\gamma\}) \\
&\leq \frac{1}{\gamma} \sum_{C \in \tilde{\mathcal{T}}_0,\, \hat{S}^{(b)} \cap C \neq \emptyset} \mathbb{P}\Big(p^{C,(b)} \frac{|\hat{S}^{(b)}|}{|C \cap \hat{S}^{(b)}|} \leq \alpha\gamma\Big) \\
&\leq \frac{1}{\gamma} \sum_{C \in \tilde{\mathcal{T}}_0,\, \hat{S}^{(b)} \cap C \neq \emptyset} \frac{|C \cap \hat{S}^{(b)}|}{|\hat{S}^{(b)}|}\alpha\gamma \leq \alpha
\end{aligned}$$

For a random variable $U$ taking values in $[0,1]$,

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

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

$$\begin{aligned} \mathbb{E}\Big(\sup_{\gamma \in (\gamma_{\min},1)} \frac{1\{U \le \alpha\gamma\}}{\gamma}\Big) &= \int_0^{\alpha\gamma_{\min}} \gamma_{\min}^{-1} dx + \int_{\alpha\gamma_{\min}}^{\alpha} \alpha x^{-1} dx \\ &= \gamma_{\min}^{-1} x\Big|_{x=0}^{x=\alpha\gamma_{\min}} + \alpha \log x\Big|_{x=\alpha\gamma_{\min}}^{x=\alpha} \\ &= \alpha + \alpha(\log\alpha - \log(\alpha\gamma_{\min})) \\ &= \alpha\big(1 - \log\frac{\alpha}{\alpha\gamma_{\min}}\big) \\ &= \alpha(1 - \log\gamma_{\min}). \end{aligned}$$

We apply this using as $U$ the uniform distributed $p^{C\cap\hat{S}^{(b)}}_{\mathrm{partial\,F-test}}$ and obtain that on $\mathcal{A}$

$$\mathbb{E}\Big(\sup_{\gamma \in (\gamma_{\min},1)} \frac{1\{p^{C,(b)} \le \alpha\gamma\}}{\gamma}\Big) \le \alpha(1 - \log\gamma_{\min}),$$

and similarly as above

$$\sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{E}\Big(\sup_{\gamma \in (\gamma_{\min},1)} \frac{1\{p^{C,(b)}_{adj} \le \alpha\gamma\}}{\gamma}\Big) \le \alpha(1 - \log\gamma_{\min}).$$

We can now consider the average over all random splits

$$\sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{E}\Big(\sup_{\gamma \in (\gamma_{\min},1)} \frac{(1/B)\sum_{b=1}^{B} 1\{p^{C,(b)}_{adj} \le \alpha\gamma\}}{\gamma}\Big) \le \alpha(1 - \log\gamma_{\min})$$

and defining $\pi^C(\cdot)$ as in the proof of Theorem 1 and using a Markov inequality

$$\sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{E}(\sup_{\gamma \in (\gamma_{\min},1)} 1\{\pi^C(\alpha\gamma) \ge \gamma\}) \le \alpha(1 - \log\gamma_{\min}).$$

We use now the fact, that the events $\{Q^C(\gamma) \leq \alpha\}$ and $\{\pi^C(\alpha\gamma) \geq \gamma\}$ are equivalent and deduce that on $\mathcal{A}$

$$\sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{P}( \inf_{\gamma \in (\gamma_{\min}, 1)} Q^C(\gamma) \leq \alpha) \quad \leq \quad \alpha(1 - \log \gamma_{\min}),$$

therefore on $\mathcal{A}$

$$\begin{aligned} \mathbb{P}(\exists C \in \tilde{\mathcal{T}}_0 \,:\, P^C \leq \alpha) &= \mathbb{P}(\min_{C \in \tilde{\mathcal{T}}_0} P^C \leq \alpha) \\ &\leq \sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{P}(P^C \leq \alpha) \\ &\leq \sum_{C \in \tilde{\mathcal{T}}_0} \mathbb{P}( \inf_{\gamma \in (\gamma_{\min}, 1)} Q^C(\gamma)(1 - \log \gamma_{\min}) \leq \alpha) \\ &\leq \alpha. \end{aligned}$$

Finally

$$\begin{aligned} & \mathbb{P}(\exists C \in \tilde{\mathcal{T}}_0 \,:\, P^C \leq \alpha) = \\ = \; & \mathbb{P}(\exists C \in \tilde{\mathcal{T}}_0 \,:\, P^C \leq \alpha \,|\, \mathcal{A}) \, P(\mathcal{A}) + \mathbb{P}(\exists C \in \tilde{\mathcal{T}}_0 \,:\, P^C \leq \alpha \,|\, \mathcal{A}^c) \, P(\mathcal{A}^c) \\ \leq \; & \alpha + 1 - (1 - \delta)^B \end{aligned}$$

and the proof is concluded.

## Proof of Theorem 2

As the only change to be considered with respect to Theorem 1 is the Shaffer multiplicity adjustment (6), it suffices to show that

$$\sum_{C \in \tilde{\mathcal{T}}_0, \, \hat{S}^{(b)} \cap C \neq \emptyset} |C|_{eff}^{\hat{S}^{(b)}} \leq |\hat{S}^{(b)}|.$$

It holds

$$\sum_{C\in\tilde{\mathcal{T}}_0,\,\hat{S}^{(b)}\cap C\neq\emptyset} |C|_{eff}^{\hat{S}^{(b)}} = \sum_{C\in\tilde{\mathcal{T}}_0,\,\hat{S}^{(b)}\cap C\neq\emptyset} \Big( |C\cap\hat{S}^{(b)}| + \\ +|\mathrm{si}(C)\cap\hat{S}^{(b)}|\,1\{\nexists E\in\mathrm{ch}(\mathrm{si}(C))\text{ s.t. } E\cap\hat{S}^{(b)}\neq\emptyset\}\Big).$$

As noted in the proof of Theorem 1 the sets in $\tilde{\mathcal{T}}_0$ are disjoint. Moreover for any cluster $D\in\tilde{\mathcal{T}}_0$ with $\mathrm{si}(D)\neq\emptyset$, $H_{0,\mathrm{si}(D)}$ is false, otherwise because of the assumption that $\mathcal{T}$ is binary $H_{0,\mathrm{pa}(D)}$ would also be true, leading to a contradiction to $D\in\tilde{\mathcal{T}}_0$. Consider now two sets $C,D\in\tilde{\mathcal{T}}_0$ with $\hat{S}^{(b)}\cap C\neq\emptyset$ and $\hat{S}^{(b)}\cap D\neq\emptyset$, since $H_{0,C}$ is true and $H_{0,\mathrm{si}(D)}$ is false it must be $\mathrm{si}(D)\not\subset C$. On the other hand if $C\subset\mathrm{si}(D)$ then the term $|\mathrm{si}(D)\cap\hat{S}^{(b)}|$ wouldn't be considered in the sum, hence only disjoint $C$ and $\mathrm{si}(D)$ are considered in the sum. Finally, suppose that for two sets $C,D\in\tilde{\mathcal{T}}_0$ with $\hat{S}^{(b)}\cap C\neq\emptyset$ and $\hat{S}^{(b)}\cap D\neq\emptyset$ it is $\mathrm{si}(C)\subset\mathrm{si}(D)$. Then if $|\mathrm{si}(D)\cap\hat{S}^{(b)}|$ is considered in the sum it must be $\mathrm{si}(C)\cap\hat{S}^{(b)}=\emptyset$. Putting all this together we conclude that all sets giving nontrivial contributions to the sum are disjoint.

## Proof of Theorem 3

In order to prove Theorem 3 we introduce four Lemmas.

**Lemma 1.**

$$\big(\hat{\beta}_{I_2}^{\hat{S}}-\beta_{\hat{S}}^0\big)\sim\mathcal{N}\Big(\big(\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}}\big)^{-1}\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}^c}\beta_{\hat{S}^c}^0,\sigma^2\big(\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}}\big)^{-1}\Big)$$

*Proof.* By definition

$$\begin{aligned}\hat{\beta}_{I_2}^{\hat{S}} &= \big(\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}}\big)^{-1}\mathbf{X}_{I_2}^{\hat{S},T}Y_{I_2} = \big(\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}}\big)^{-1}\mathbf{X}_{I_2}^{\hat{S},T}\big(\mathbf{X}_{I_2}^{\hat{S}}\beta_{\hat{S}}^0+\mathbf{X}_{I_2}^{\hat{S}^c}\beta_{\hat{S}^c}^0+\varepsilon_{I_2}\big)\\ &= \beta_{\hat{S}}^0+\big(\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}}\big)^{-1}\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}^c}\beta_{\hat{S}^c}^0+\big(\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}}\big)^{-1}\mathbf{X}_{I_2}^{\hat{S},T}\varepsilon_{I_2}\end{aligned}$$

and $\hat{\beta}_{I_2}^{\hat{S}}$ is as linear transformation of a normal distributed random variable also normal distributed. From the formula above its is easy to see that the expected value $\big(\hat{\beta}_{I_2}^{\hat{S}}-\beta_{\hat{S}}^0\big)$ is $\big(\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}}\big)^{-1}\mathbf{X}_{I_2}^{\hat{S},T}\mathbf{X}_{I_2}^{\hat{S}^c}\beta_{\hat{S}^c}^0$.

For the covariance we can calculate

$$\begin{aligned}
\mathrm{Cov}\big(\hat{\beta}^{\hat{S}}_{I_2} - \beta^0_{\hat{S}}\big) &= \mathrm{Cov}\big(\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}^c}_{I_2}\beta^0_{\hat{S}^c} + \big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}\mathbf{X}^{\hat{S},T}_{I_2}\varepsilon_{I_2}\big) \\
&= \big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}\mathbf{X}^{\hat{S},T}_{I_2}\mathrm{Cov}(\varepsilon_{I_2})\mathbf{X}^{\hat{S}}_{I_2}\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1} \\
&= \sigma^2\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}
\end{aligned}$$

□

**Lemma 2.** *$P^{\hat{S}}_{I_2}$ resp. $Q^{\hat{S}}_{I_2}$ is an orthogonal projection of $\mathbb{R}^{|I_2|}$ in $\mathbb{R}^{|\hat{S}|}$ resp. $\mathbb{R}^{|I_2|-|\hat{S}|}$.*

*Proof.* It follows from the definition of $P^{\hat{S}}_{I_2}$ and $Q^{\hat{S}}_{I_2}$, that they satisfy the equation $X^T = X = X^2$. Moreover

$$\begin{aligned}
\mathrm{tr}(P^{\hat{S}}_{I_2}) &= \mathrm{tr}(\mathbf{X}^{\hat{S}}_{I_2}\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}\mathbf{X}^{\hat{S},T}_{I_2}) = \mathrm{tr}(\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}) = \\
&= \mathrm{tr}(I_{|\hat{S}|}) = |\hat{S}| \\
\mathrm{tr}(Q^{\hat{S}}_{I_2}) &= \mathrm{tr}(I_{|I_2|} - P^{\hat{S}}_{I_2}) = \mathrm{tr}(I_{|I_2|}) - (P^{\hat{S}}_{I_2}) = |I_2| - |\hat{S}|
\end{aligned}$$

and this concludes the proof. □

**Lemma 3.** *$(\hat{\sigma}^{\hat{S}}_{I_2})^2$ and $\hat{\beta}^{\hat{S}}_{I_2}$ are independent.*

*Proof.* We show that $\hat{\varepsilon}^{\hat{S}}_{I_2}$ and $\hat{Y}^{\hat{S}}_{I_2}$ are uncorrelated, then the Lemma follows because of

$$\begin{aligned}
\hat{\beta}^{\hat{S}}_{I_2} &= \big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}\mathbf{X}^{\hat{S},T}_{I_2}Y_{I_2} = \big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}\mathbf{X}^{\hat{S},T}_{I_2}P^{\hat{S}}_{I_2}Y_{I_2} \\
&= \big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}\mathbf{X}^{\hat{S},T}_{I_2}\hat{Y}^{\hat{S}}_{I_2}
\end{aligned}$$

and the fact that the random variables involved are normally distributed.

$$\begin{aligned}
\mathrm{Cov}\big(\hat{\varepsilon}^{\hat{S}}_{I_2}, \hat{Y}^{\hat{S}}_{I_2}\big) &= \mathrm{Cov}\big(Q^{\hat{S}}_{I_2}Y_{I_2}, P^{\hat{S}}_{I_2}Y_{I_2}\big) = \mathrm{Cov}\big(Y_{I_2}\big)Q^{\hat{S}}_{I_2}P^{\hat{S},T}_{I_2} \\
&= \sigma^2\big(I_{I_2} - P^{\hat{S}}_{I_2}\big)P^{\hat{S}}_{I_2} = \sigma^2\big(P^{\hat{S}}_{I_2} - (P^{\hat{S}}_{I_2})^2\big) = 0
\end{aligned}$$

□

**Lemma 4.** *$(\hat{\sigma}_{I_2}^{\hat{S}})^2$ is an unbiased estimator of $\sigma^2$ and*

$$\big(|I_2| - |\hat{S}|\big)\frac{(\hat{\sigma}_{I_2}^{\hat{S}})^2}{\sigma^2} \sim \chi^2_{|I_2|-|\hat{S}|}$$

*Proof.* We calculate

$$\begin{aligned}
\mathbb{E}\big[(\hat{\sigma}_{I_2}^{\hat{S}})^2\big] &= \frac{1}{|I_2| - |\hat{S}|}\,\mathbb{E}\big[\hat{\varepsilon}_{I_2}^{\hat{S},T}\hat{\varepsilon}_{I_2}^{\hat{S}}\big] = \frac{1}{|I_2| - |\hat{S}|}\mathrm{tr}\Big(\mathbb{E}\big[\hat{\varepsilon}_{I_2}^{\hat{S}}\hat{\varepsilon}_{I_2}^{\hat{S},T}\big]\Big) \\
&= \frac{1}{|I_2| - |\hat{S}|}\mathrm{tr}\big(Q_{I_2}^{\hat{S}}\,\mathbb{E}\big[Y_{I_2}Y_{I_2}^T\big]Q_{I_2}^{\hat{S},T}\big) \\
&= \frac{\sigma^2}{|I_2| - |\hat{S}|}\mathrm{tr}\big(Q_{I_2}^{\hat{S}}Q_{I_2}^{\hat{S},T}\big) = \sigma^2
\end{aligned}$$

To see that the given random variable is chi-square distributed we use a geometrical approach. Consider a basis of $|I_2|$ orthogonal vectors, s.t. the first $|\hat{S}|$ vectors span the space given by the vectors of $\mathbf{X}_{I_2}^{\hat{S}}$ and call the corresponding transformation matrix $G$ (the columns of $G$ corresponds the coordinates of the new basis vectors in the old coordinate system). Then $G$ is orthogonal and using a star for the new coordinate system we have $Y_{I_2}^* = G^T Y_{I_2},\ \varepsilon_{I_2}^* = G^T\varepsilon_{I_2}$. By construction it is

$$\begin{aligned}
(\hat{Y}_{I_2}^{\hat{S}})^* &= (Y_1^*, \ldots, Y_{|\hat{S}|}^*, 0, \ldots, 0)^T \\
(\hat{\varepsilon}_{I_2}^{\hat{S}})^* &= (0, \ldots, 0, \varepsilon_{|\hat{S}|+1}^*, \ldots, \varepsilon_{|I_2|}^*)^T,
\end{aligned}$$

using the orthogonality of $G$ we get

$$\hat{\varepsilon}_{I_2}^{\hat{S},T}\hat{\varepsilon}_{I_2}^{\hat{S}} = (\hat{\varepsilon}_{I_2}^{\hat{S},T})^*(\hat{\varepsilon}_{I_2}^{\hat{S}})^* = \sum_{|\hat{S}|+1}^{I_2}\varepsilon_i^{*2}.$$

Again because of the orthogonality of $G$, it holds $\varepsilon_{I_2}^* = G^T\varepsilon_{I_2} \sim \mathcal{N}(0, \sigma^2 I_{I_2})$ and the proof is concluded. □

**Proof of Theorem 3.**

Theorem 3 follows from the Lemmas 1, 2, 3 and 4 and the following considerations. First rewrite

$$\begin{aligned} & \frac{(A\hat{\beta}^{\hat{S}}_{I_2} - A\beta^0_{\hat{S}})^T \big(A\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}A^T\big)^{-1}(A\hat{\beta}^{\hat{S}}_{I_2} - A\beta^0_{\hat{S}})}{q(\hat{\sigma}^{\hat{S}}_{I_2})^2} \\ = & \left(\frac{(A\hat{\beta}^{\hat{S}}_{I_2} - A\beta^0_{\hat{S}})^T \big(A\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}A^T\big)^{-1}(A\hat{\beta}^{\hat{S}}_{I_2} - A\beta^0_{\hat{S}})}{q\sigma^2}\right)\left(\frac{(\hat{\sigma}^{\hat{S}}_{I_2})^2}{\sigma^2}\right)^{-1}. \end{aligned}$$

Because of Lemma 3 the two terms in the big brackets are independent. Because of Lemma 4 the term in the second big bracket would be $\chi^2_{|I_2|-|\hat{S}|}$-distributed, if multiplied by $|I_2| - |\hat{S}|$. Let's consider the term in the first big bracket. Because of Lemma 1, the quadratic form given by the term in the first big bracket multiplied by $q$ corresponds to the quadratic form $Z^T Z$ where

$$Z = \frac{1}{\sigma}\big(A\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}A^T\big)^{-1/2}A(\hat{\beta}^{\hat{S}}_{I_2} - \beta^0_{\hat{S}}) \sim \mathcal{N}(\text{BIAS}, I_q)$$

with

$$\text{BIAS} = \frac{1}{\sigma}\big(A\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}A^T\big)^{-1/2}A\big(\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}}_{I_2}\big)^{-1}\mathbf{X}^{\hat{S},T}_{I_2}\mathbf{X}^{\hat{S}^c}_{I_2}\beta^0_{\hat{S}^c}$$

and this concludes the proof.

## References

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95:265–278.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681.