SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan
OF THE REPUBLIC OF BULGARIA

Benjamini-Hochberg procedure under positive dependency with block structure

19th International Summer Conference on Probability and Statistics

July 25, 2025
Sofia, Bulgaria

Nikolay I. Nikolov

Sofia University "St. Kliment Ohridski", SUMMIT, Project ΓAMMA
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

Collaboration

This talk is based on a joint work with Dean Palejev and Mladen Savov (Sofia University & Institute of Mathematics and Informatics, BAS).

This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No.BG-RRP-2.004-0008.

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Overview

1. Multiple testing problem
   - Benjamini-Hochberg procedure
   - Benjamini-Yekutieli procedure

2. Block structure dependency
   - Model description
   - Block based method to control the FDR
   - Structure estimation

3. Simulation study results

4. Summary

SUMMIT
SOFIA UNIVERSITY
MARKING MOMENTUM
FOR INNOVATION AND
TECHNOLOGICAL TRANSFER

Multiple testing problem

Multiple testing problem

Multiple comparisons problem:

- Several hypotheses are tested simultaneously, e.g., searching for differences in many features at the same time between two groups
- Application in biochemistry, genetics, plant sciences, ...
- In some studies thousands (or even millions) of tests are performed
- For a single test, each null hypothesis is rejected if the corresponding p-value is below a fixed significance level α
- For n tests, $n\alpha$ of them would be rejected **even if null is true** (even if the tested groups are identical)
- In gene studies, $n > 20000$ and at $\alpha = 0.05 \implies$
  > 1000 tests may be incorrectly rejected
  (> 1000 genes may be incorrectly identified as informative)
- Family-Wise Error Rate (FWER):
  the probability of at least one incorrect rejection

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Multiple testing problem

- False Discovery Rate (FDR):
  expected proportion of falsely rejected hypotheses among the set of rejected hypotheses (if there is at least one rejection, otherwise 0)

Benjamini–Hochberg (B–H) procedure:

- Based on ordered p-values: $p_{1:n} \leq p_{2:n} \leq \ldots \leq p_{n:n}$
- Rejects the null hypothesis for the tests corresponding to $p_{1:n}, p_{2:n}, \ldots, p_{k:n}$, where k is the largest integer such that
$$p_{k:n} \leq \frac{k}{n}\alpha$$
- If the raw p-values are iid with marginal distribution $\mathcal{U}(0,1)$, then the B–H procedure has an FDR equal to α
- Benjamini and Hochberg (1995) is among the Top 100 most cited scientific papers (Van Noorden et al. (2014))

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Multiple testing problem

Benjamini–Yekutieli (B–Y) procedure:

- Rejects the null hypothesis for the tests corresponding to $p_{1:n}, p_{2:n}, \ldots, p_{k:n}$, where k is the largest integer such that

$$p_{k:n} \leq \frac{k}{nc(n)}\alpha$$

- If the p-values are iid, the $c(n) = 1$, while in the case of arbitrary dependence Benjamini and Yekutieli (2001) suggest the n-th harmonic number, i.e.,

$$c(n) = \sum_{i=1}^{n} \frac{1}{i} \approx \log(n) + \frac{1}{2n} + \gamma,$$

where $\gamma \approx 0.5772$ is the Euler's constant

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Example

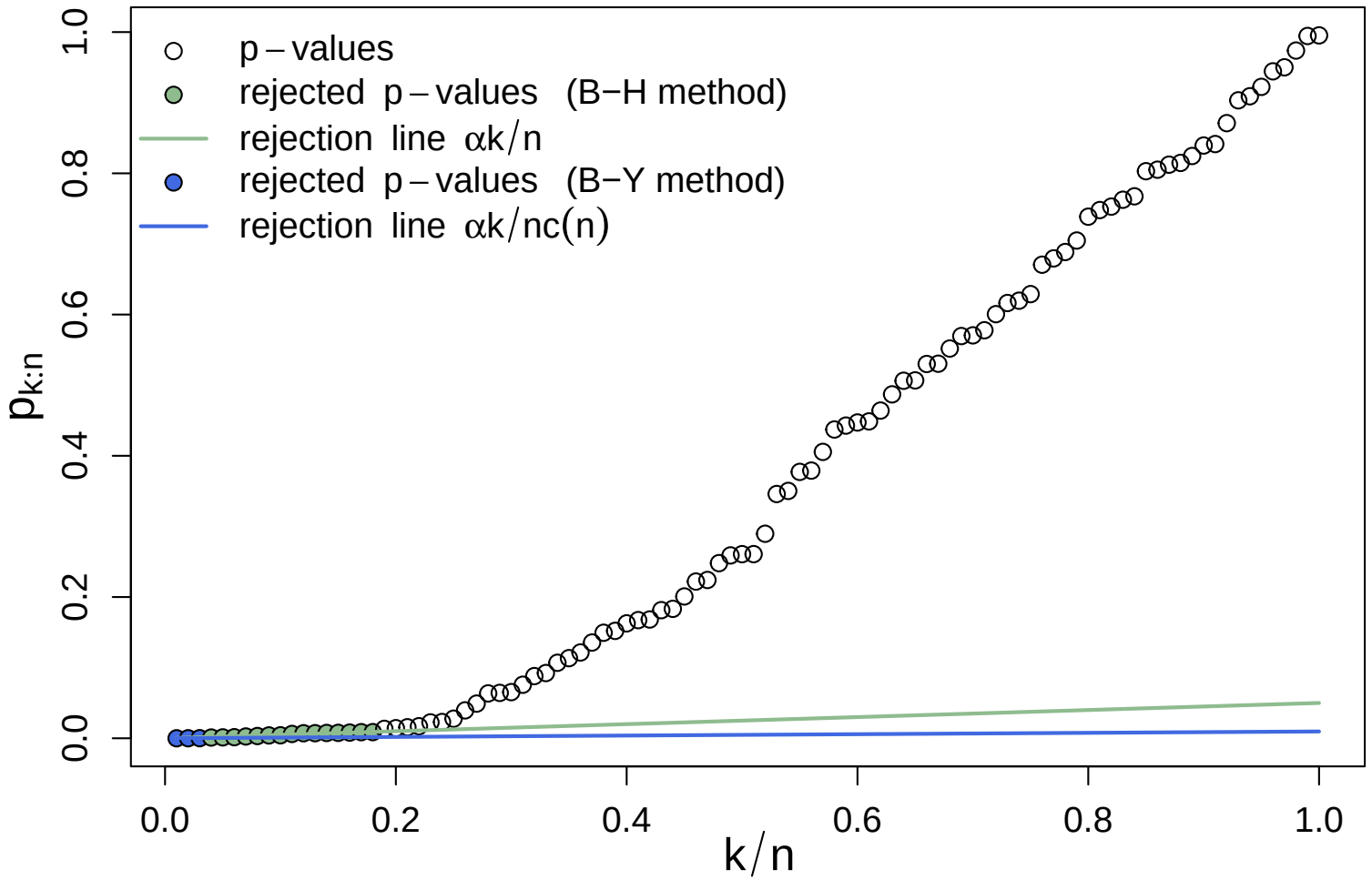


Figure 1: B–H and B–Y procedures applied to a set of $n = 100$ ordered p-values, using a FDR control level of $\alpha = 0.05$.

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Example (zoomed in)

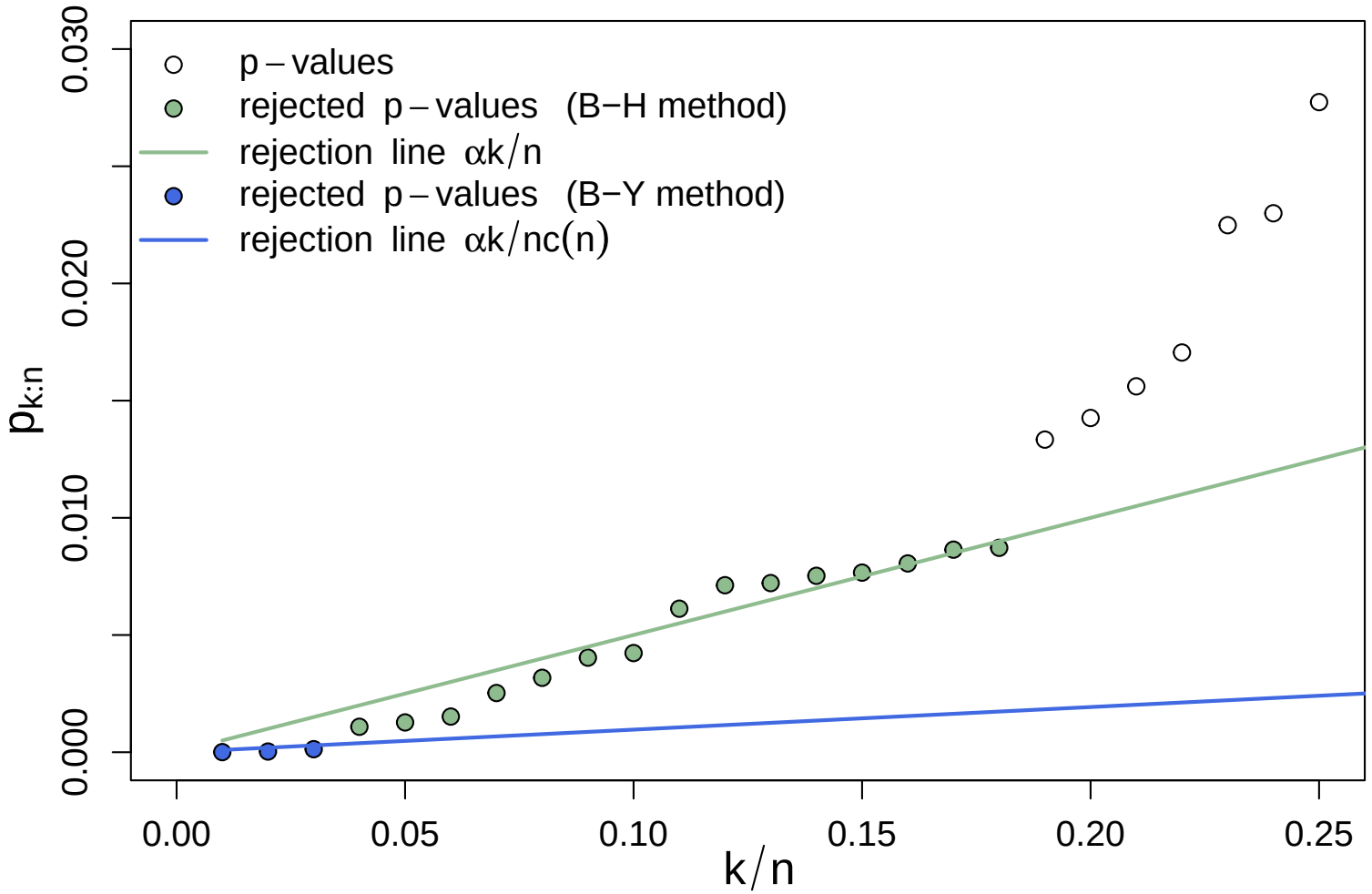


Figure 2: B–H and B–Y procedures applied to a set of $n = 100$ ordered p-values, using a FDR control level of $\alpha = 0.05$.

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Multiple testing problem

- False Discovery Rate:

$$\text{FDR} = \text{E}\left[\mathbb{1}\{R > 0\}\,\frac{V}{R}\right],$$

  where V is the number of true null hypotheses rejected and R is the total number of hypotheses rejected

- Average Power:

$$\text{Power} = \text{E}\left(\frac{R - V}{n - N}\right),$$

  with N being the number of true null hypotheses among all n tests

- In case of iid and true null p-values following $\mathcal{U}(0, 1)$, the FDR of B–H procedure converges to $\frac{N}{n}\alpha$, Benjamini and Hochberg (1995)

- The FDR of B–Y procedure is less than $\frac{N}{n}\alpha$, Benjamini and Yekutieli (2001)

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Block structure dependency

Notations

- Let $\vec{X}_i = \left(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(n)}\right)$ and $\vec{Y}_i = \left(Y_i^{(1)}, Y_i^{(2)}, \ldots, Y_i^{(n)}\right)$ be observations for the i-th individual, e.g. before and after treatment, or control and treatment groups of paired individuals, $i = 1, \ldots, m$, over n features, e.g. gene expression levels

- Then, $\left(X_1^{(j)}, Y_1^{(j)}\right), \left(X_2^{(j)}, Y_2^{(j)}\right), \ldots, \left(X_m^{(j)}, Y_m^{(j)}\right)$ is a paired sample for the j-th feature over m individuals, where $j = 1, 2, \ldots, n$

- $\mathbf{X} = \left\{X_i^{(j)}\right\}_{i=1,j=1}^{m,\,n}$ and $\mathbf{Y} = \left\{X_i^{(j)}\right\}_{i=1,j=1}^{m,\,n}$ are the observation matrices in the control and treatment groups, respectively

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Model description

Assumptions:

- $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_m$ and $\vec{Y}_1, \vec{Y}_2, \ldots, \vec{Y}_m$ are iid in each group
- $\vec{X}_i \sim \mathcal{N}(\vec{\mu}_X, \mathbf{\Sigma})$ and $\vec{Y}_i \sim \mathcal{N}(\vec{\mu}_Y, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is n by n covariance block matrix such that

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{A_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A_2} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{A_k} \end{pmatrix},$$

with $\mathbf{A_1} = \mathbf{A_2} = \cdots = \mathbf{A_k} = \mathbf{A}$ being s by s matrices ($s = [n/k]$)

Motivation:

- Kim et al. (2016):

  Different blocks correspond to different gene regulatory pathways and model the assumption that groups of genes in the same pathway are biologically or functionally correlated and interacting with each other, whereas genes in different pathways are uncorrelated.

Block structure

- Chi et al. (2025):

  One of the main challenges is that there are many definitions of what it means to be dependent. . . . The multitude of possibilities is daunting, perhaps explaining the lack of progress.

Parameterization of the block structure (examples):

$$\mathbf{A} = \mathbf{A}^*(\theta) = \begin{pmatrix} 1 & \theta & \dots & \theta^{s-1} \\ \theta & 1 & \dots & \theta^{s-2} \\ \dots & \dots & \dots & \dots \\ \theta^{s-1} & \theta^{s-2} & \dots & 1 \end{pmatrix} = \left\{ \theta^{|i-j|} \right\}_{i,j=1}^{s}$$

or

$$\mathbf{A} = \mathbf{A}'(\theta) = \begin{pmatrix} 1 & \theta & \frac{(2s-5)\theta}{2s-4} & \dots & \frac{\theta}{2} \\ \theta & 1 & \theta & \dots & \frac{(s-1)\theta}{2s-4} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{\theta}{2} & \frac{(s-1)\theta}{2s-4} & \frac{s\theta}{2s-4} & \dots & 1 \end{pmatrix}$$

$$= \left\{ \mathbb{1}\{i \neq j\} \frac{2s-3-|i-j|}{2s-4}\theta + \mathbb{1}\{i=j\} \right\}_{i,j=1}^{s}$$

with parameter $\theta \in (0,1)$

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Paired t-test

Student's t-tests for paired samples:

- $H_0 : \mu_X^{(j)} = \mu_Y^{(j)}$ vs $H_1 : \mu_X^{(j)} \neq \mu_Y^{(j)}$, for $j = 1, 2 \ldots, n$, where $\vec{\mu}_X = \left(\mu_X^{(1)}, \mu_X^{(2)}, \ldots, \mu_X^{(n)}\right)$ and $\vec{\mu}_Y = \left(\mu_Y^{(1)}, \mu_Y^{(2)}, \ldots, \mu_Y^{(n)}\right)$
- Matrix of differences, $\mathbf{D} = \left\{X_i^{(j)} - Y_i^{(j)}\right\}_{i=1,j=1}^{m,\,n} = \left\{D_i^{(j)}\right\}_{i=1,j=1}^{m,\,n}$
- Average difference for each feature (gene)

$$\overline{D}_j = \frac{1}{m}\sum_{i=1}^{m}\left(X_i^{(j)} - Y_i^{(j)}\right), \quad \text{for } j = 1, 2, \ldots, n,$$

- Student's t-statistic for each feature

$$T_m^{(j)} = \frac{\sqrt{m}\,\overline{D}_j}{\hat{\sigma}_m^{(j)}} \sim t(m-1), \quad \text{for } j = 1, 2, \ldots, n,$$

where $\hat{\sigma}_m^{(j)} = \sqrt{\dfrac{\sum_{i=1}^{m}\left(D_i^{(j)} - \overline{D}_j\right)^2}{m-1}}$

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Joint asymptotic distribution

Theorem (Nikolov et al. (2025))

Under H_0 and an arbitrary covariance (dependence) matrix $\mathbf{\Sigma}$, the joint distribution of $T_m^{(1)}, T_m^{(2)}, \ldots, T_m^{(n)}$ is asymptotically normal with

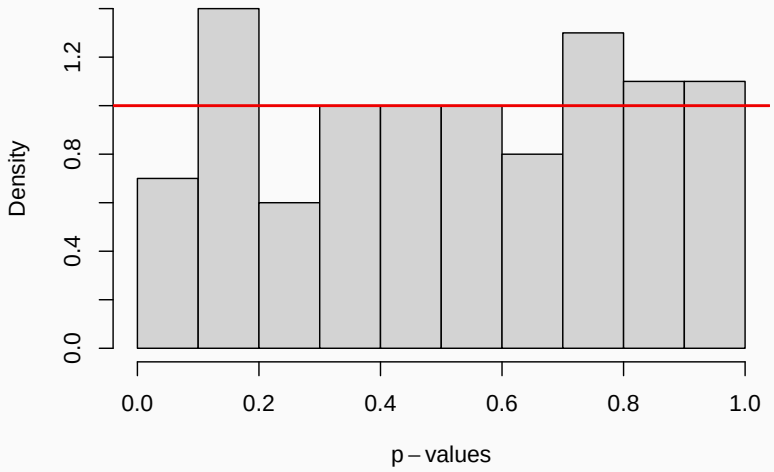
$$\begin{pmatrix} T_m^{(1)} \\ T_m^{(2)} \\ \ldots \\ T_m^{(n)} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} T^{(1)} \\ T^{(2)} \\ \ldots \\ T^{(n)} \end{pmatrix} \sim \mathcal{N}\left(\vec{0}, \mathbf{\Sigma}\right), \text{ as } m \to \infty.$$

- Use this dependence structure to adjust the p-values and improve the performance of the B–H method
- Block B–H method: reject all test rejected by B–H procedure together with the tests corresponding to the block with largest divergence from $\mathcal{U}(0,1)$

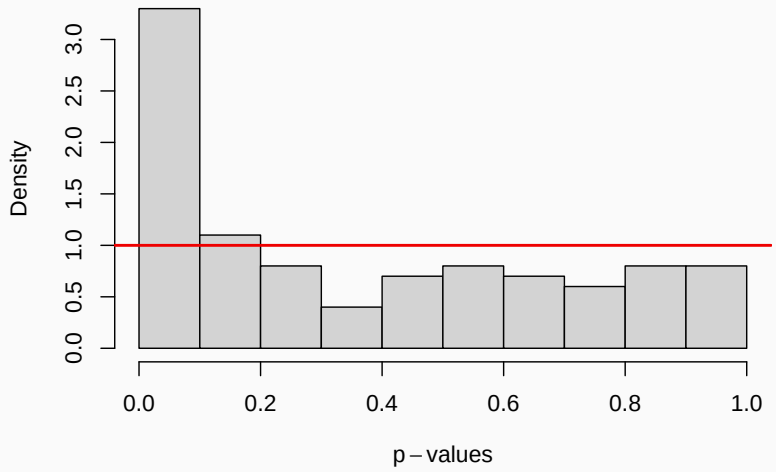
SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Kolmogorov-Smirnov measure

p-values in a block with true H_0



p-values in a block with true H_1



Kolmogorov-Smirnov distance:

$$D^+ = \max_{x \in [0,1]} \left(\hat{F}_p(x) - x \right),$$

where $\hat{F}_p(x)$ is the edf of the raw p-values in a block

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

Block structure estimation

Estimation of the parameter θ:

- Under true null hypotheses, $\vec{X}_i - \vec{Y}_i \sim \mathcal{N}(\vec{0}, 2\mathbf{\Sigma})$, for $i = 1, 2, \ldots, m$
- Let $\hat{\mathbf{C}}$ be the sample estimation of the covariance matrix $\mathbf{\Sigma}$
- Frobenius norm estimation of θ:

$$\hat{\theta} = \underset{\theta \in (0,1)}{\arg\min} \left\| \mathbf{A}(\theta) - \hat{\mathbf{C}} \right\|_F = \underset{\theta \in (0,1)}{\arg\min} \sqrt{\sum_{i=1}^{n} s_i^2 \left(\mathbf{A}(\theta) - \hat{\mathbf{C}} \right)},$$

  where $s_i(\mathbf{M})$ is i-th singular value of the matrix $\mathbf{M}$

- Max norm estimation of θ:

$$\tilde{\theta} = \underset{\theta \in (0,1)}{\arg\min} \left\| \mathbf{A}(\theta) - \hat{\mathbf{C}} \right\|_M = \underset{\theta \in (0,1)}{\arg\min} \left(\max_{i,j} \left\{ \mathbf{A}(\theta) - \hat{\mathbf{C}} \right\}_{i,j} \right)$$

- Other matrix norms $\rightarrow$ $L_{2,1}$ (sum of Euclidean norms), Trace norm (Schatten 1-norm), Spectral norm (Schatten ∞-norm)

Simulation study

Simulations setup

Parameter values:

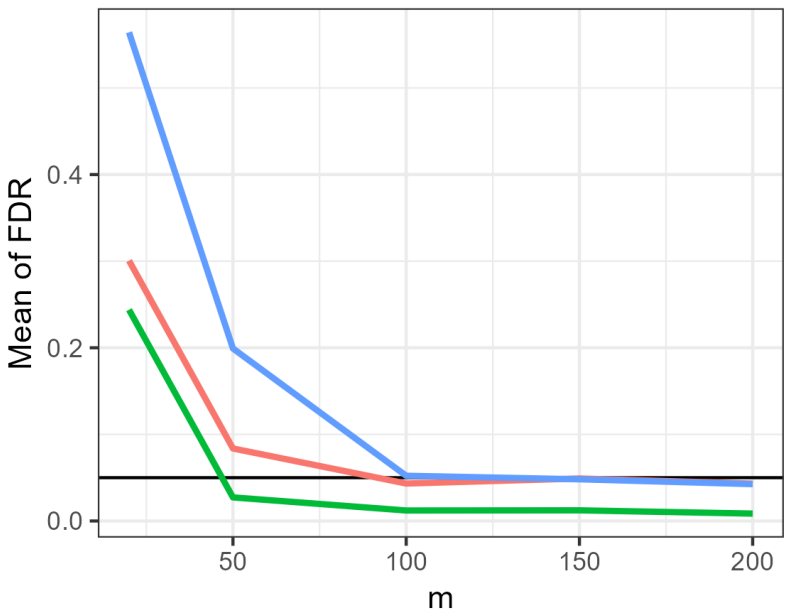
- $n \in \{100, 1000\} \rightarrow$ number of features (genes)
- $m \in \{20, 50, 100, 150, 200\} \rightarrow$ number of individuals (sample size)
- $k = 10 \rightarrow$ number of blocks (one block for the alternative, i.e. the proportion of true null hypotheses is 0.90)
- $\mu_X = 0$ and $\mu_Y = 0.5 \rightarrow$ location of null and alternative
- $\theta = 0.80 \rightarrow$ correlation coefficient in the block matrices
- $\mathbf{A}^*(\theta)$ and $\mathbf{A}'(\theta) \rightarrow$ block structure
- $\alpha = 0.05 \rightarrow$ significance level in the FDR procedures
- $1000 \rightarrow$ number of trials in each simulation

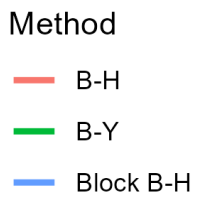
Results:

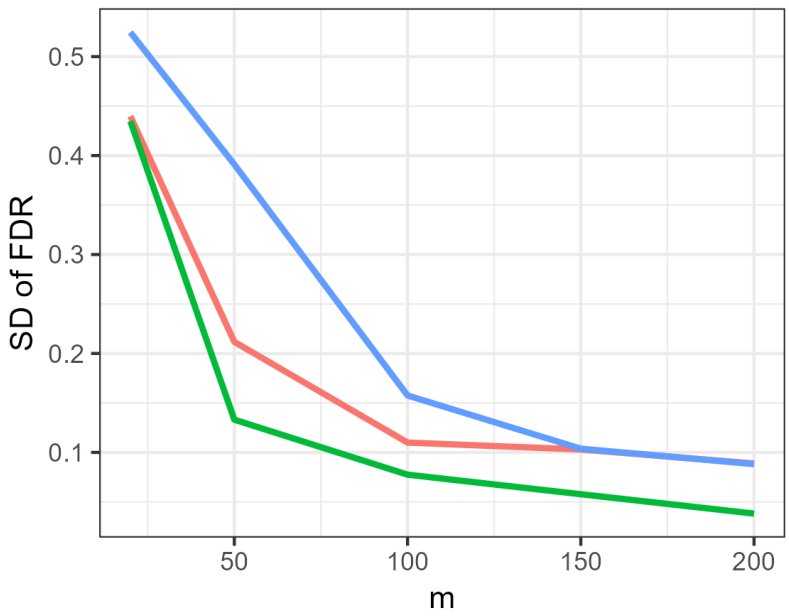
- FDR and Power of B–H, B–Y and Block B–H (Mean and SD of the results in each trial)
- Estimations $\hat{\theta}$ and $\tilde{\theta}$ Frobenius and Max norms, respectively (Mean, SD, MSE and Average Time of the results in each trial)

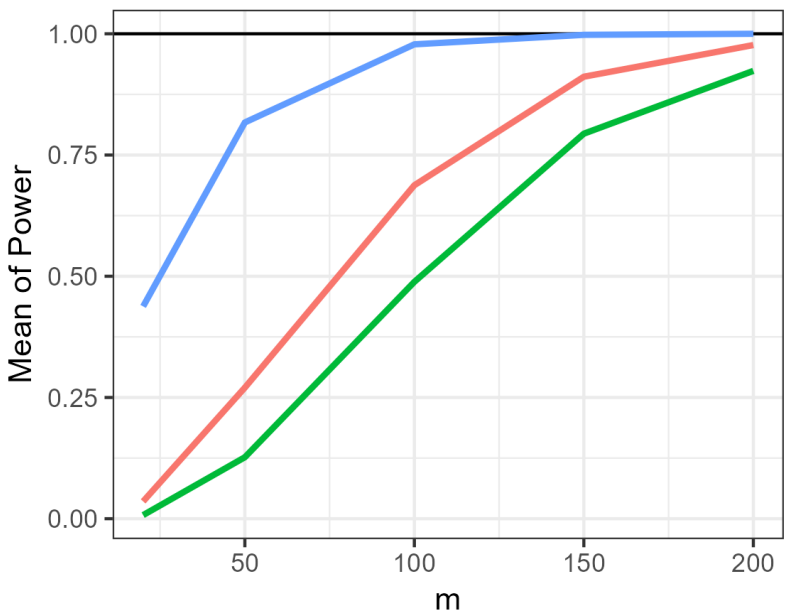
SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

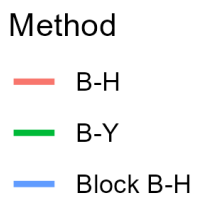
Simulation results ($n = 100, \mathbf{A}'(\theta)$)

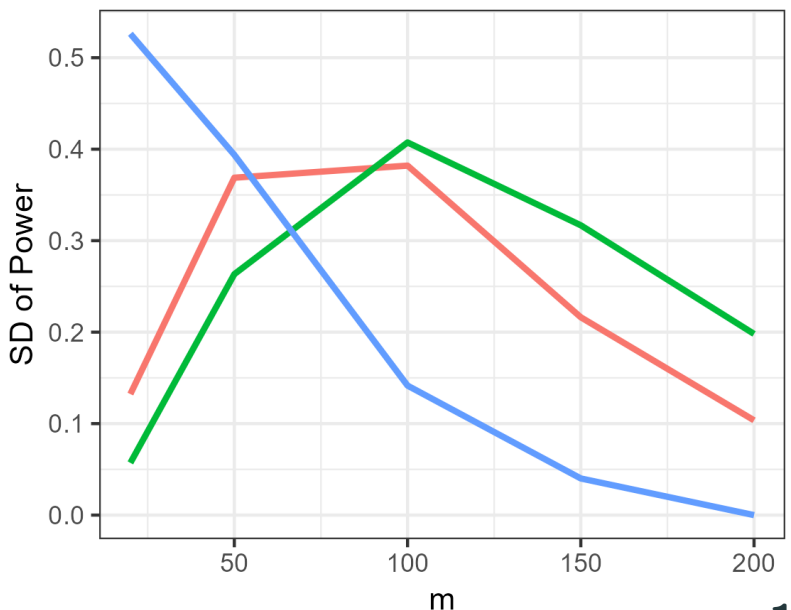




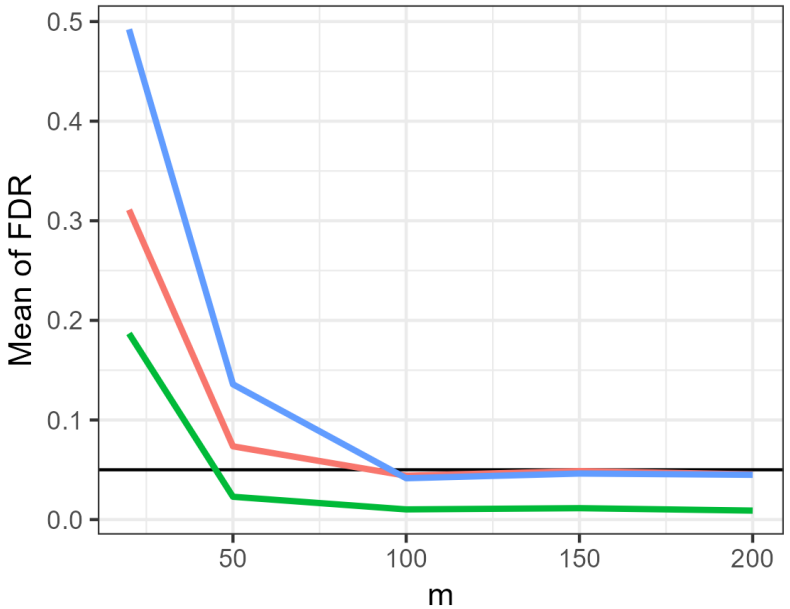


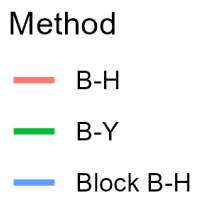


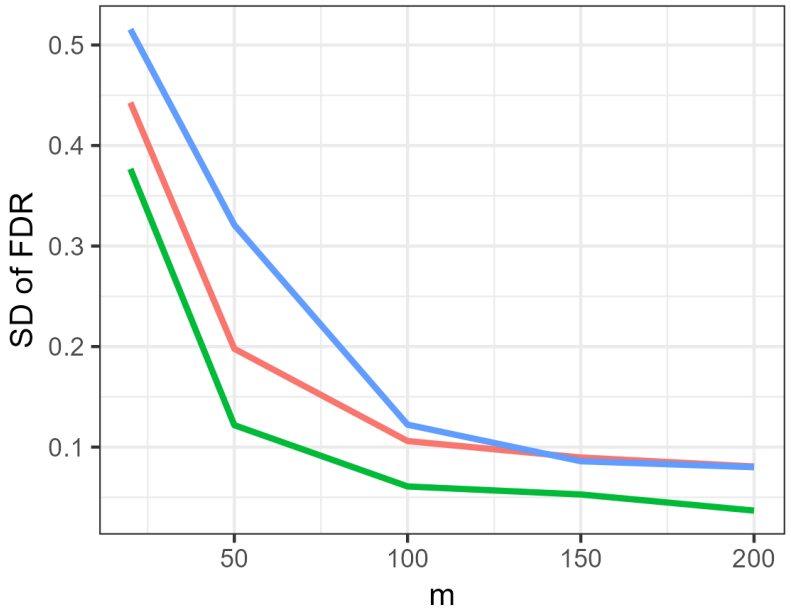


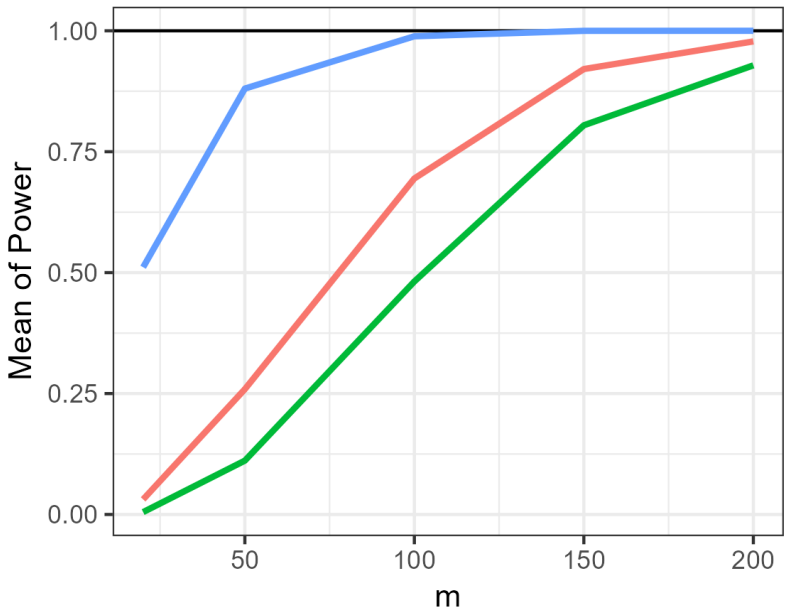


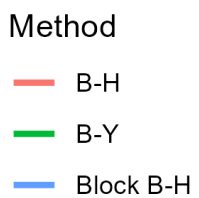
Simulation results ($n = 100, \mathbf{A}^*(\theta)$)

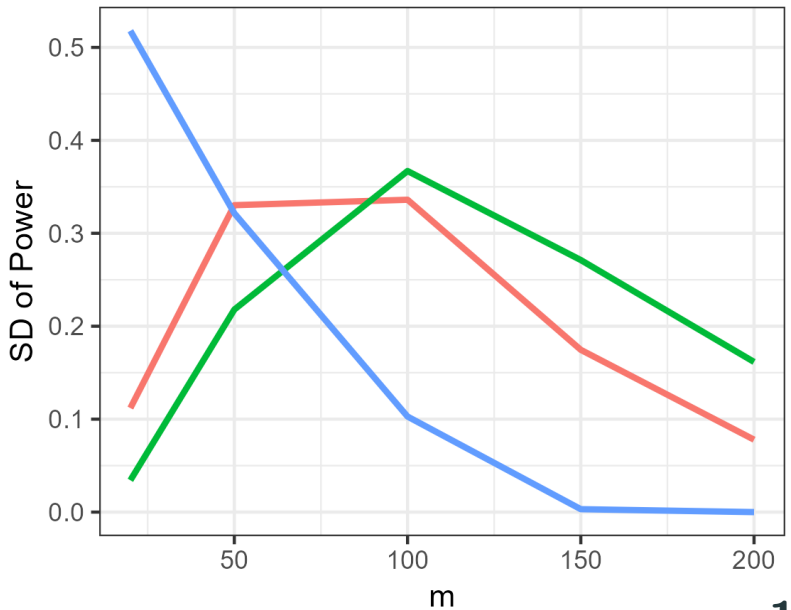




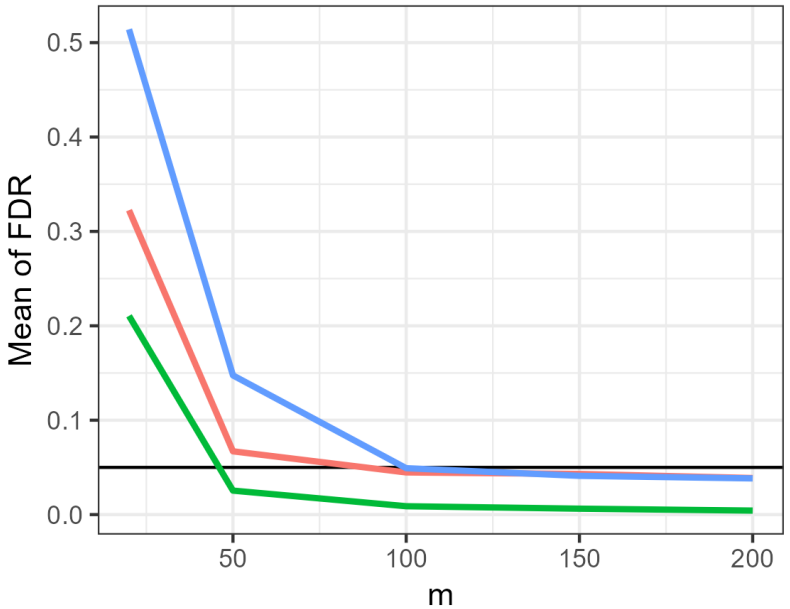


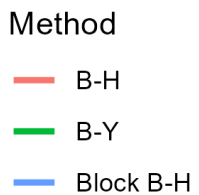


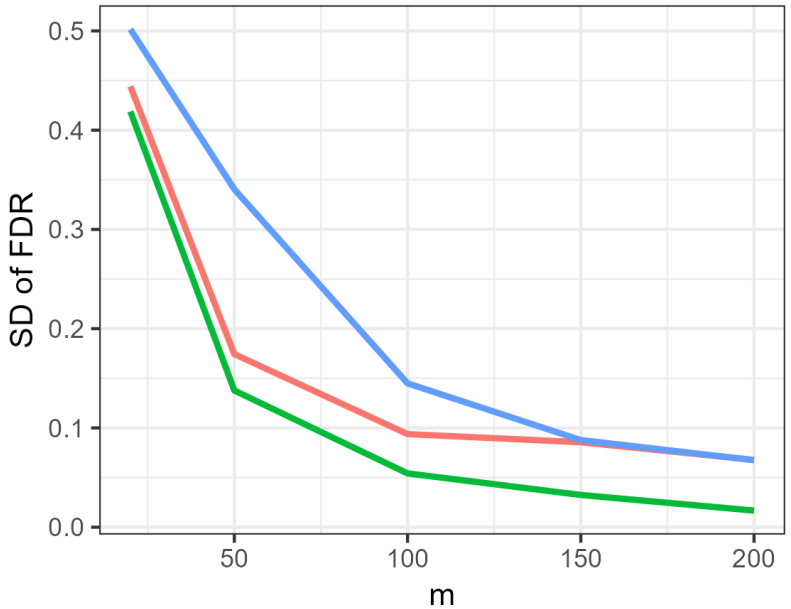


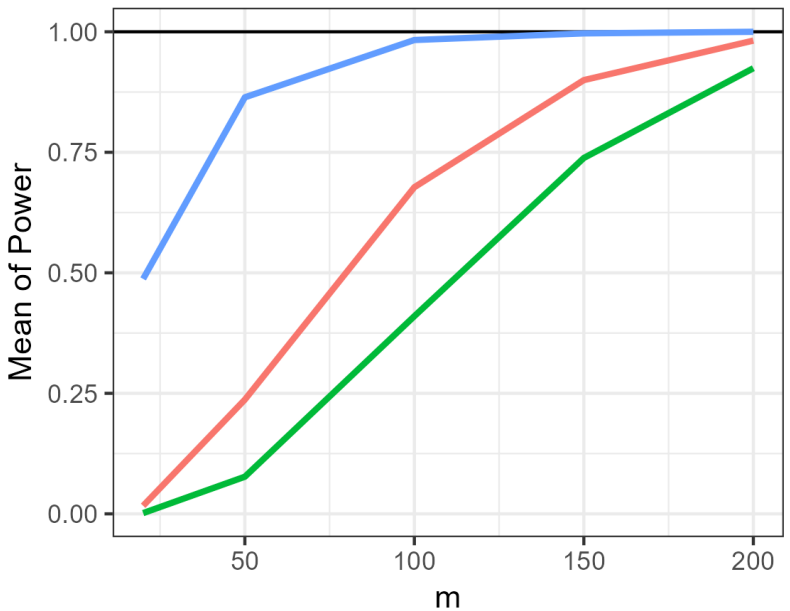


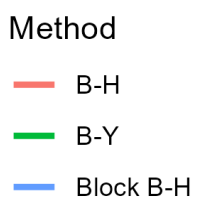
Simulation results ($n = 1000, \mathbf{A}'(\theta)$)

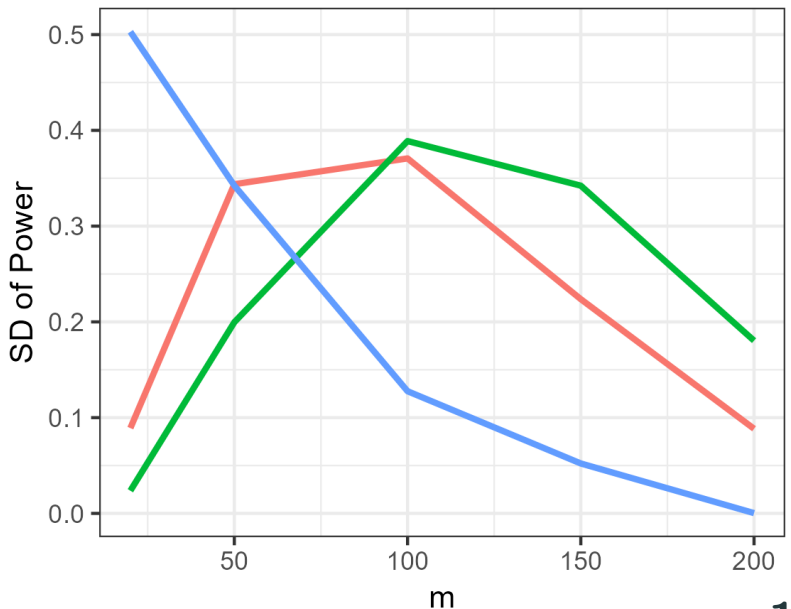




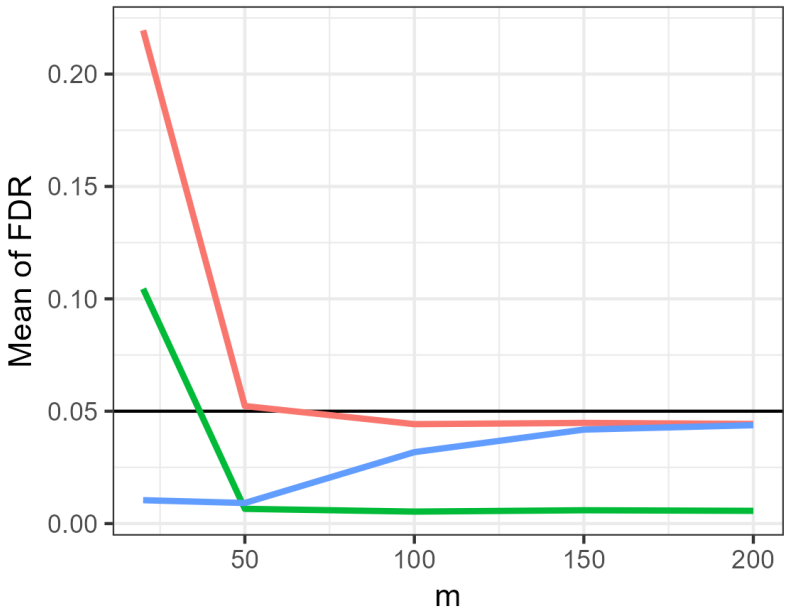


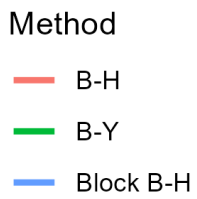


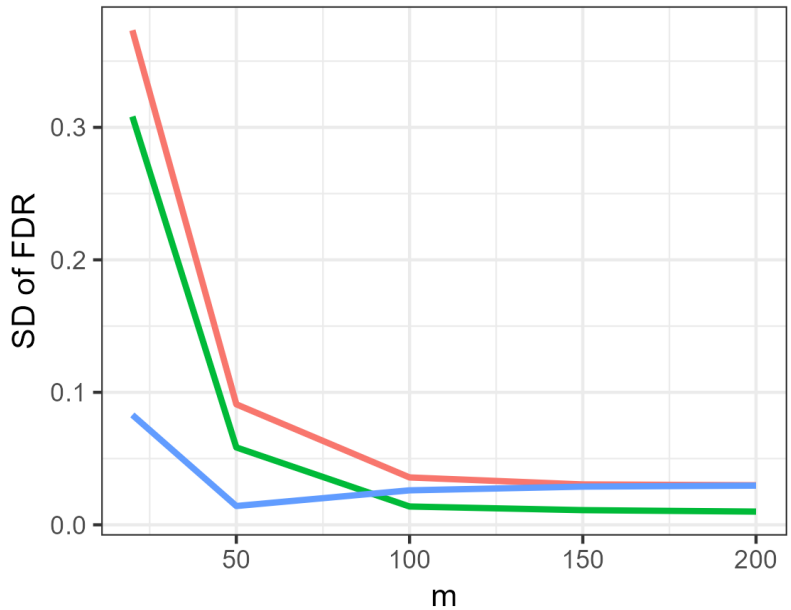


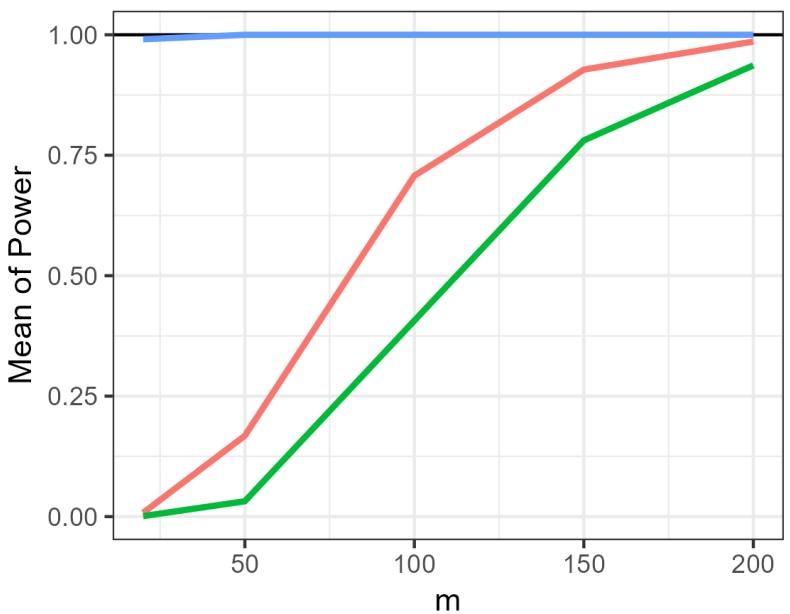


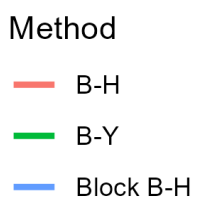
Simulation results ($n = 1000, \mathbf{A}^*(\theta)$)

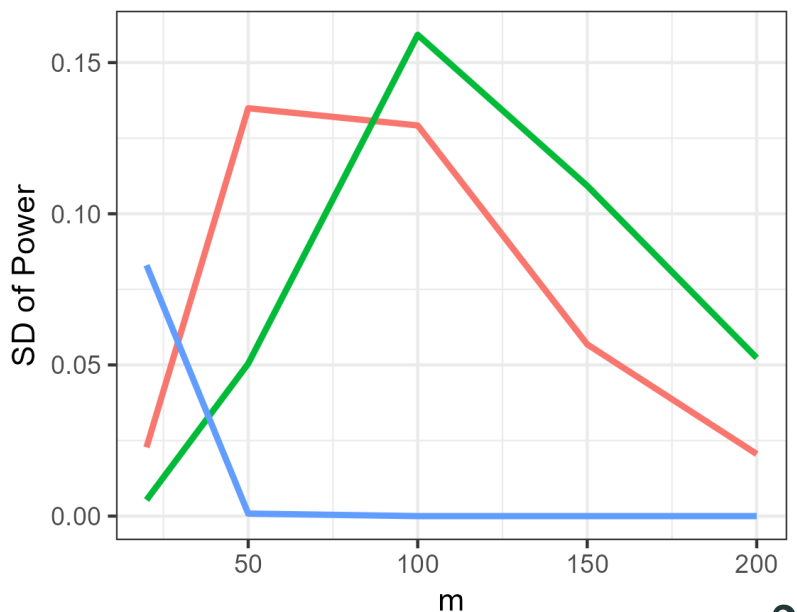




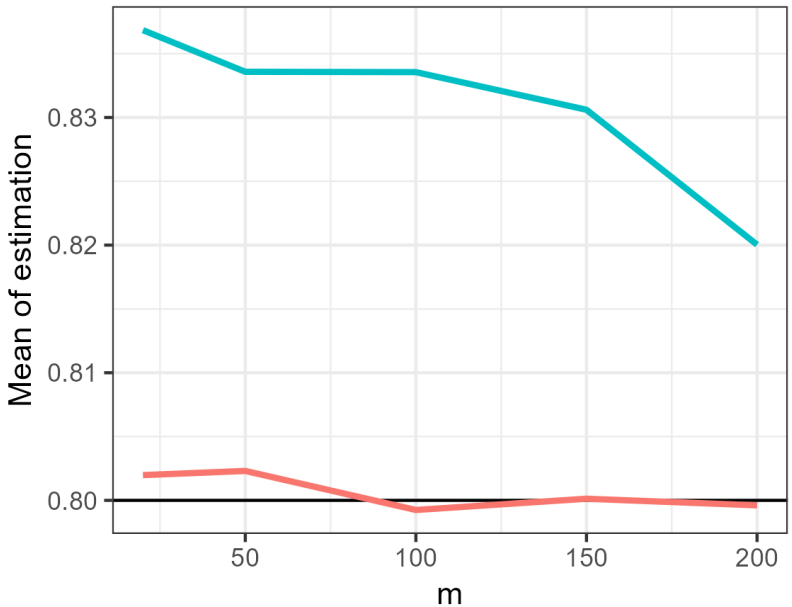


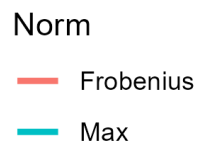


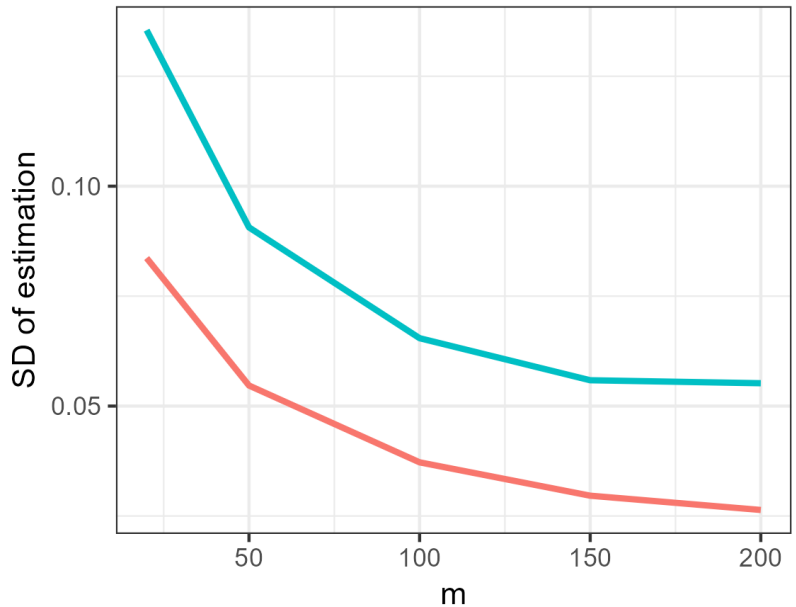


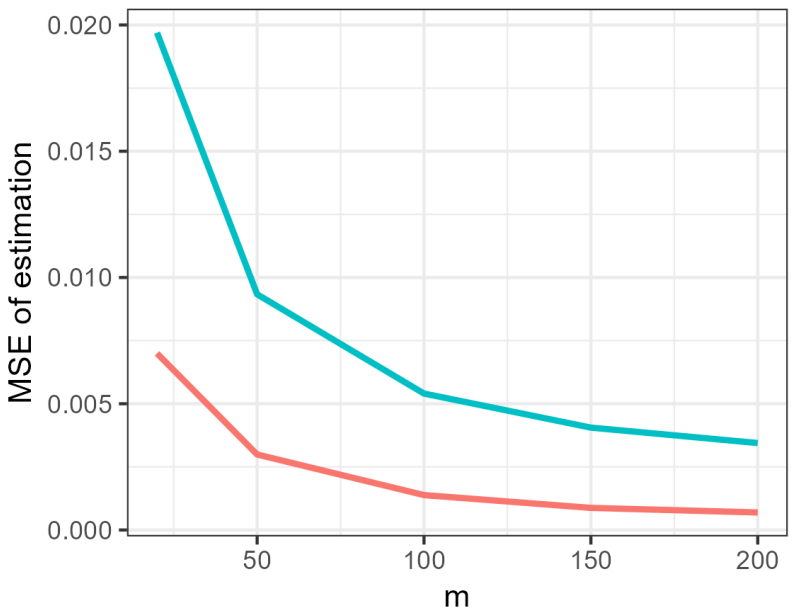


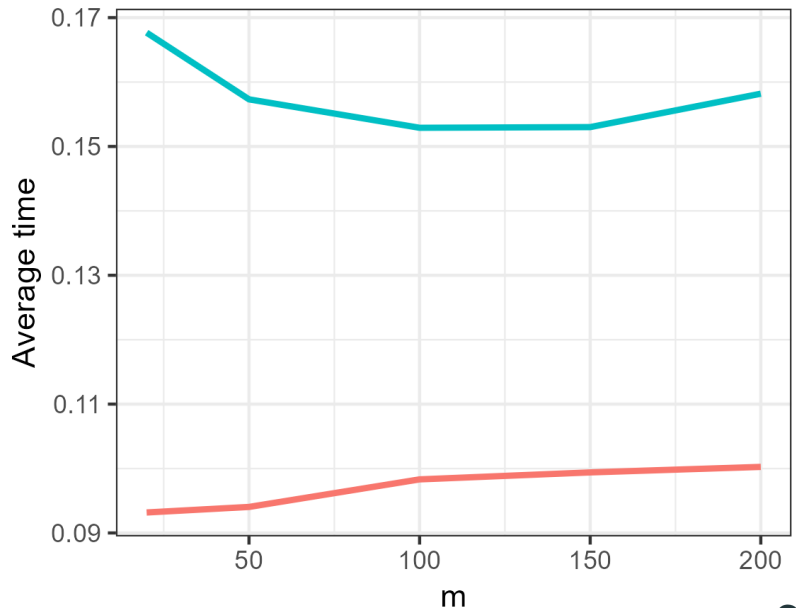
Simulation results ($n = 100, \mathbf{A}'(\theta)$)



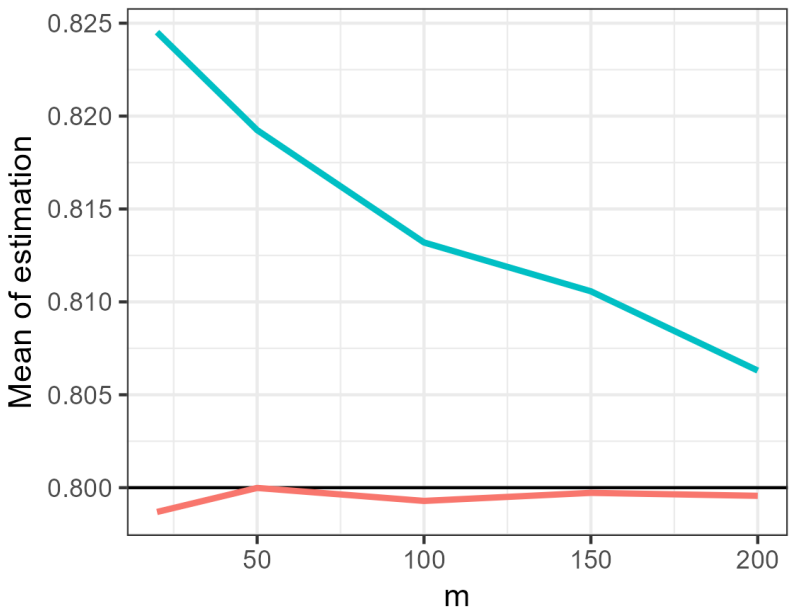


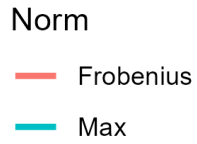


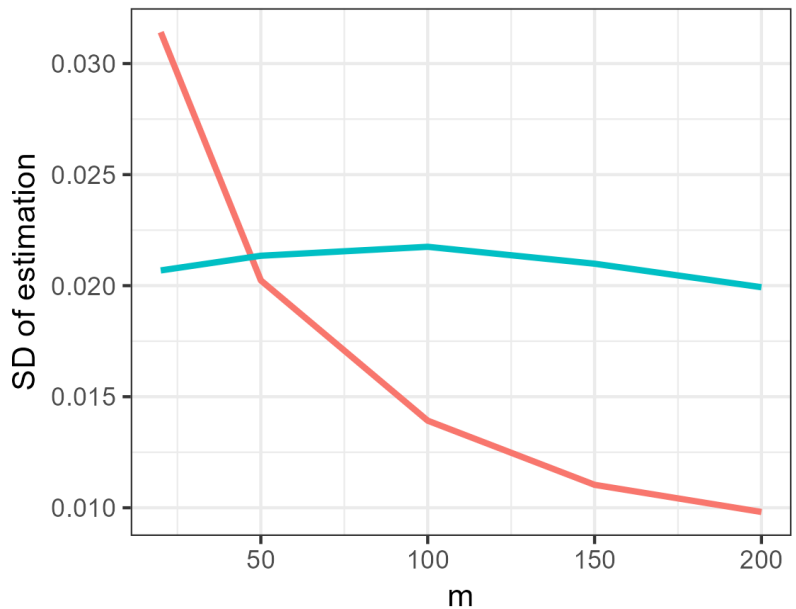


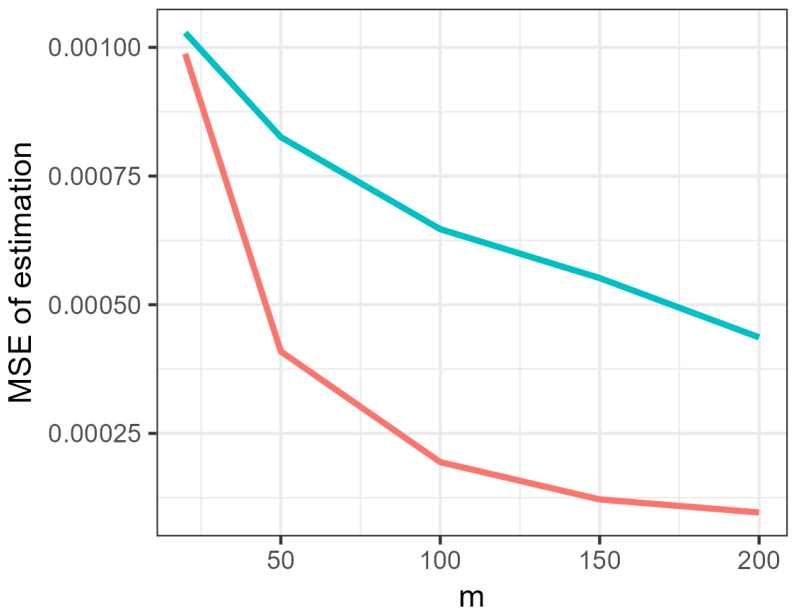


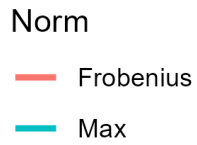
Simulation results ($n = 100, \mathbf{A}^*(\theta)$)

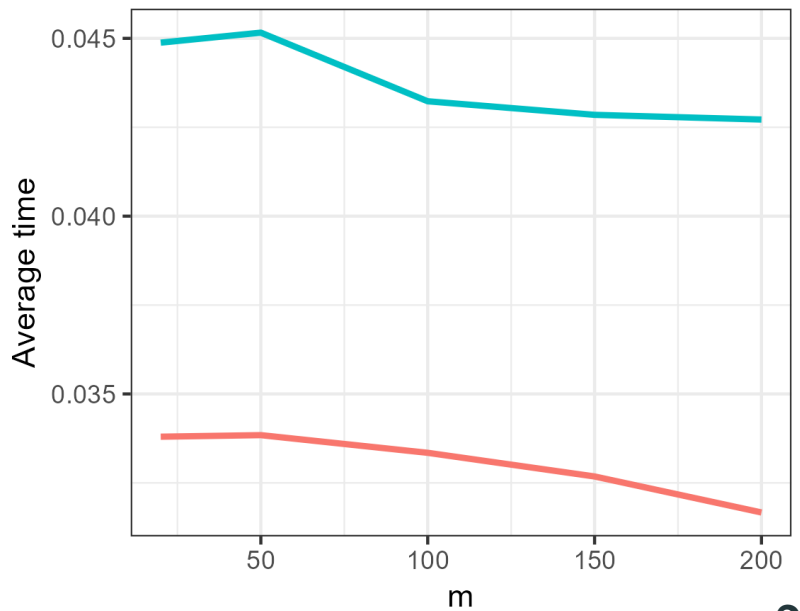




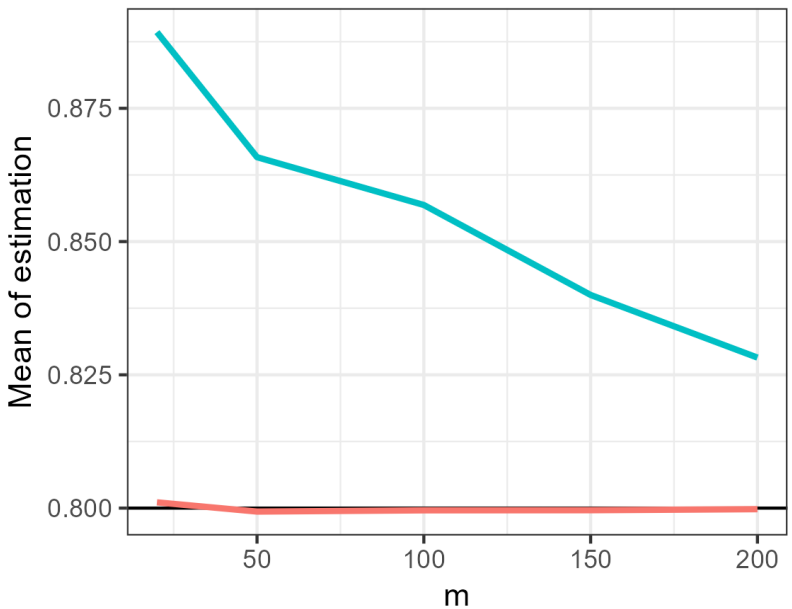


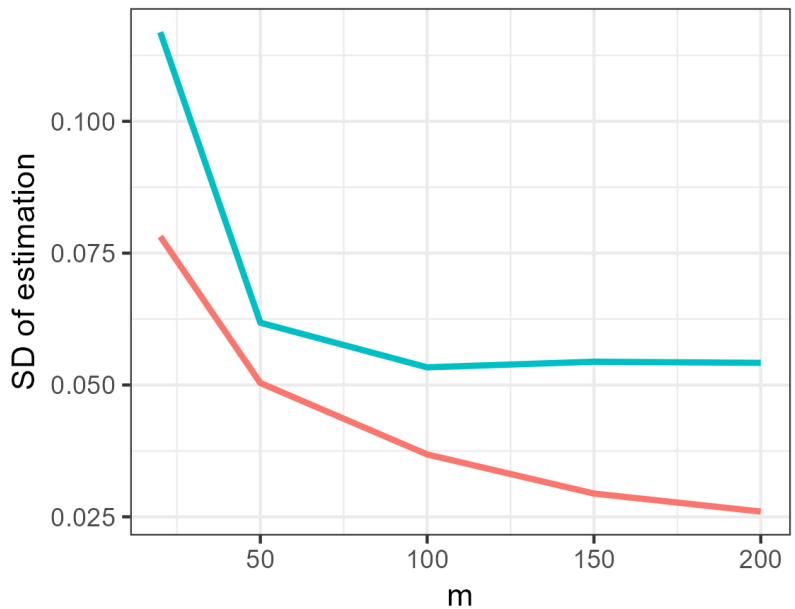


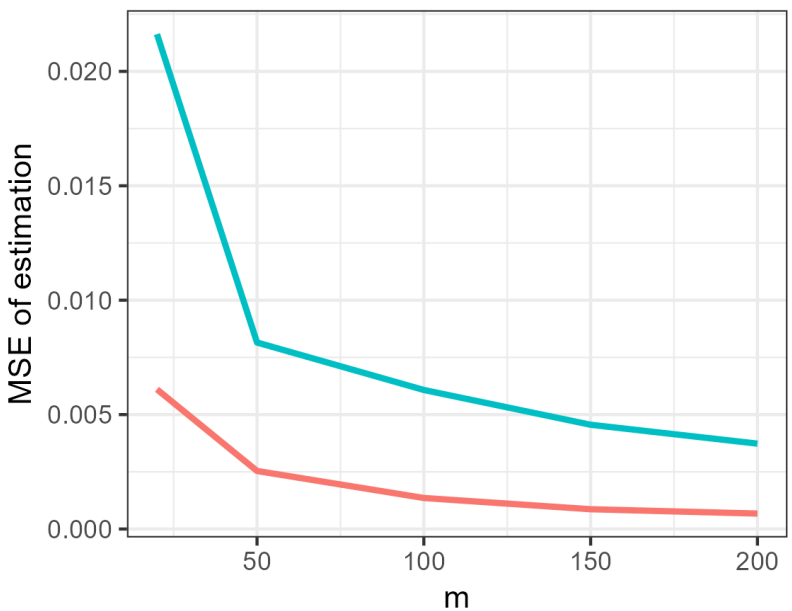


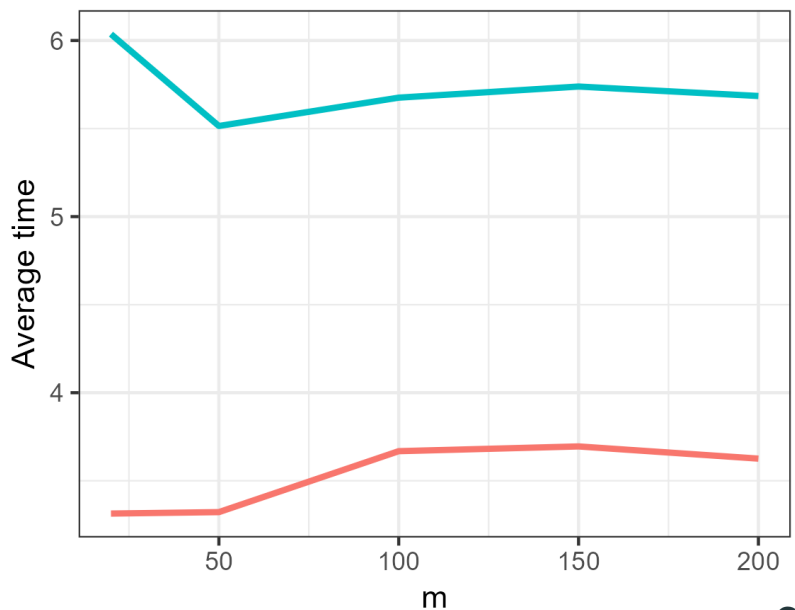


Simulation results ($n = 1000, \mathbf{A}'(\theta)$)

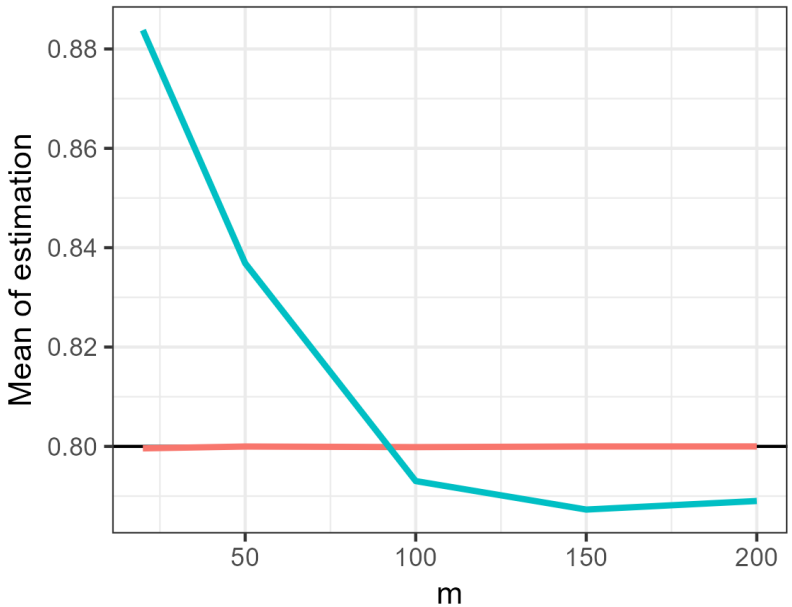


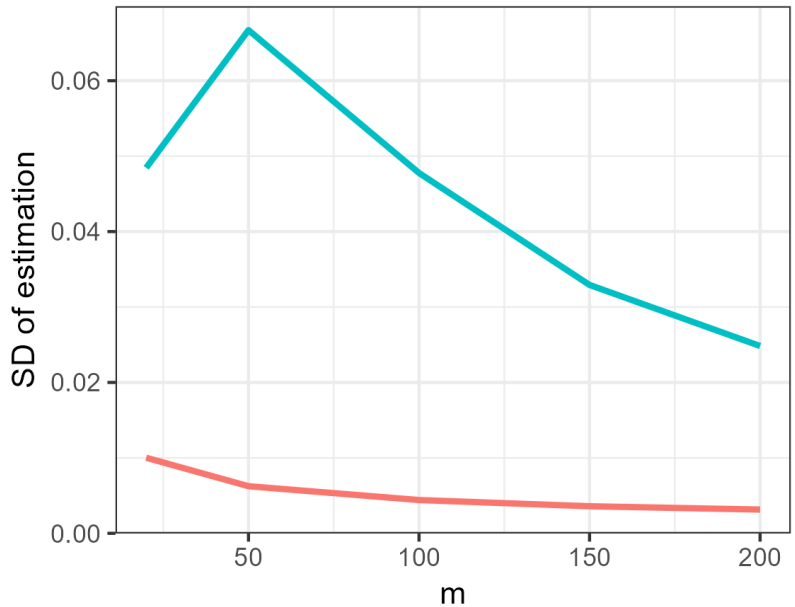


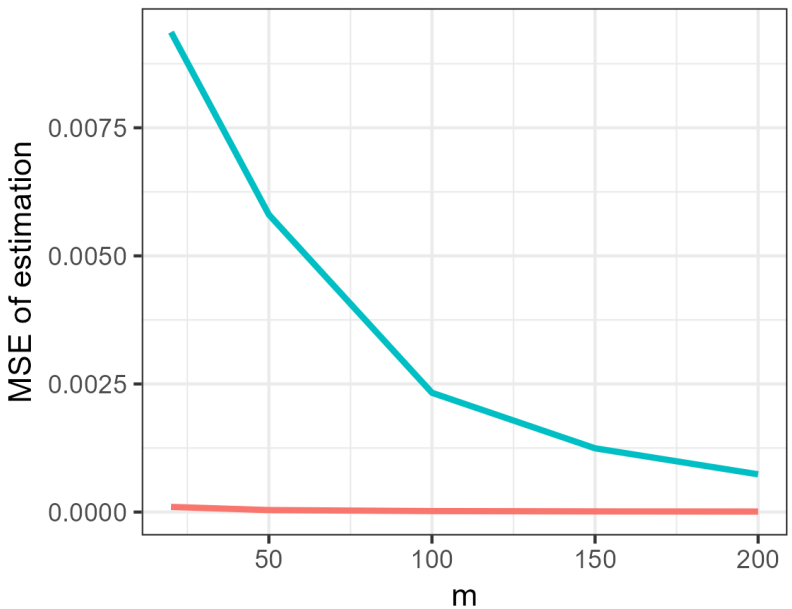


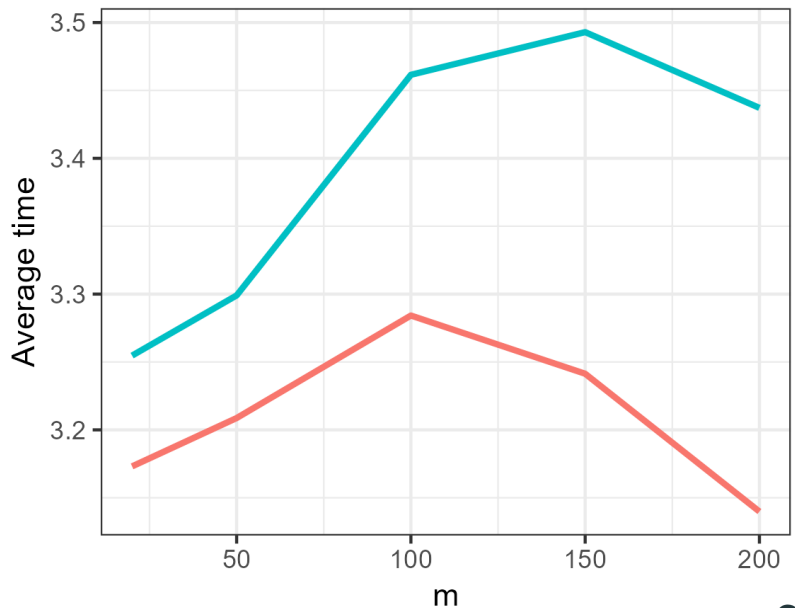


Simulation results ($n = 1000, \mathbf{A}^*(\theta)$)









Summary

Conclusion

Future work:

- Procedure for controlling the FDR by using the block structure in the joint distribution of the raw p-values
- Study the performance of the B–H procedures under block dependency (convergence rate)
- Improve the Block B–H procedure by introducing some critical value for the K–S distance D^+
- Adjust the block structure estimation in case the ordering in each block (pathway) is unknown
- Study the properties of the block estimation for θ (bias, consistency)

SUMMIT
SOFIA UNIVERSITY MARKING MOMENTUM FOR INNOVATION AND TECHNOLOGICAL TRANSFER

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 (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Chi, Z., Ramdas, A., and Wang, R. (2025). Multiple testing under negative dependence. *Bernoulli*, 31(2):1230–1255.

References ii

Kim, E., Ivanov, I., Hua, J., Chapkin, R. S., and Dougherty, E. R. (2016). Model-based study of the effectiveness of reporting lists of small feature sets using RNA-seq data. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 470–471.

Nikolov, N. I., Palejev, D., and Savov, M. (2025). Benjamini-Hochberg procedure performance under dependence structure. *Manuscript in preparation*.

Van Noorden, R., Maher, B., and Nuzzo, R. (2014). The top 100 papers. *Nature*, 514(7524):550–553.