

# Approximate Local FDR: Integrating Correlations Into Multiple Testing Methodology

Rajesh Karmakar, Ruth Heller, Saharon Rosset

Tel Aviv University



## Introduction

- **Large scale multiple testing** Controlling measure of false discovery in the framework of two group model to draw valid statistical inference.
- **False discovery measure** We are interested in  $mFDR = \frac{\mathbb{E}V}{\mathbb{E}R}$ .

- **IID Two group model** For z-score  $Z_i$  and its state  $h_i$ ,

$$h_i \stackrel{i.i.d.}{\sim} Ber(\pi)$$

$$Z_i | h_i \stackrel{i.i.d.}{\sim} f(z_i | h_i)$$

- **General Two group model** For z-score  $Z_i$  and its state  $h_i$ ,

$$h_i \stackrel{i.i.d.}{\sim} Ber(\pi)$$

$$\mathbf{Z} | \mathbf{h} \sim g(\mathbf{z} | \mathbf{h})$$

## OMT Procedure

- Solution to the problem:

$$\max_{\mathbf{D}: \mathbb{R}^K \rightarrow \{0,1\}^K} TP(\mathbf{D}) \quad \text{s.t.} \quad mFDR(\mathbf{D}) \leq \alpha$$

where  $TP(\mathbf{D}) = \mathbb{E}(\mathbf{h}^T \mathbf{D})$ .

- Optimal policy under iid two group model [2] is thresholding the **marginal locFDR**

$$D_i(Z_i) = \mathbb{I}(\mathbb{P}(h_i = 0 | Z_i) < c)$$

- Optimal policy under general two group model [3] is thresholding the **oracle locFDR**

$$D_i(\mathbf{Z}) = \mathbb{I}(\mathbb{P}(h_i = 0 | \mathbf{Z}) < c)$$

- Calculating  $\mathbb{P}(h_i = 0 | \mathbf{Z})$  for large number of hypothesis  $K$  is computationally infeasible.

- Using  $\mathbb{P}(h_i = 0 | Z_i)$  instead of  $\mathbb{P}(h_i = 0 | \mathbf{Z})$  may lead to significant loss of power.

## Challenge

Devise a  $mFDR$  controlling method which is both

- Computationally feasible
- Statistically efficient

## Practical Problem

- In GWAS

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \mathbf{S})$  with  $\mathbf{S} = \text{diag}((\mathbf{X}^T \mathbf{X})^{-1})$ .
- Define  $\mathbf{Z} = \frac{1}{\sigma} \mathbf{S}^{-1/2} \hat{\boldsymbol{\beta}}$  and  $\Sigma = \mathbf{S}^{-1/2} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{S}^{-1/2}$ .
- A reasonable model is

$$h_i \stackrel{i.i.d.}{\sim} Ber(\pi), \quad i = 1, \dots, K$$

$$\mathbf{Z} | \mathbf{h} \sim \mathcal{N}_K(\mathbf{b}\mathbf{h}, \Sigma + \tau^2 \text{diag}(\mathbf{h})) \quad (1)$$

## Approach

- One possible way forward is to use **neighborhood** based locFDR statistics of the form  $T_{i,N} = \mathbb{P}(h_i = 0 | Z_{i-N}, \dots, Z_i, \dots, Z_{i+N})$  or  $\mathbb{P}(h_i = 0 | \mathbf{Z}_{i,N})$  where  $\mathbf{z}_{i,N} = (z_{\max(i-N,1)}, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_{\min(i+N,K)})$ .
- This makes sense for short ranged dependence structure.
- Are these statistics easy to compute?

$$T_{i,N} = \frac{\sum_{\mathbf{h}_{i,N}, h_i=0} \mathbb{P}(\mathbf{Z}_{i,N} | \mathbf{h}_{i,N}) \mathbb{P}(\mathbf{h}_{i,N})}{\sum_{\mathbf{h}_{i,N}} \mathbb{P}(\mathbf{Z}_{i,N} | \mathbf{h}_{i,N}) \mathbb{P}(\mathbf{h}_{i,N})}$$

- Are these statistics statistically efficient?
- Let's observe the behaviour of  $\mathbb{P}(h_i = 0 | \mathbf{Z}_{i,N})$  for different  $N$ .

## Procedure

**Algorithm 1** The Data-driven procedure for level  $\alpha$  mFDR control

- 1: **Input:** z-scores  $\mathbf{z}$ , covariance  $\Sigma$  of  $\mathbf{z}$ , realistic  $N \geq 0$ .
- 2: Choose  $\mathbf{z}_s$  of approximately independent z-scores from  $\mathbf{z}$ .
- 3: Given  $\mathbf{z}_s$  and  $\mathbf{z}$ , estimate  $b, \pi, \tau$  of model (1) by EM-Algorithm.
- 4: Calculate  $\hat{\mathbb{P}}(h_i = 0 | \mathbf{Z}_{i,N})$  using the estimated parameters.
- 5: Calculate the estimated cutoff  $\hat{c}_{\alpha,N}$  by bootstrap.
- 6: **Output:** The set  $\mathcal{R}_\alpha = \{i : \hat{\mathbb{P}}(h_i = 0 | \mathbf{Z}_{i,N}) \leq \hat{c}_{\alpha,N}\}$  of rejected hypothesis.

## Property of the Procedure

- Define

$$\mathcal{C}_N = \{\mathbf{D} : \mathbb{R}^K \rightarrow \{0,1\}^K \mid D_i \equiv D_i(\mathbf{Z}_{i,N}) \quad \forall i\}$$

- **Theorem ( $T_N$ -rule)**

The solution to the optimization problem

$$\max_{\mathbf{D} \in \mathcal{C}_N} TP(\mathbf{D}) \quad \text{s.t.} \quad mFDR(\mathbf{D}) \leq \alpha.$$

is  $\mathbf{D}_N^* = (D_{1,N}, \dots, D_{K,N})$  where

$$D_{i,N} = \mathbb{I}(\mathbb{P}(h_i = 0 | \mathbf{Z}_{i,N}) < c)$$

- **Corollary**

Power  $TP_N$  is increasing in  $N$ .

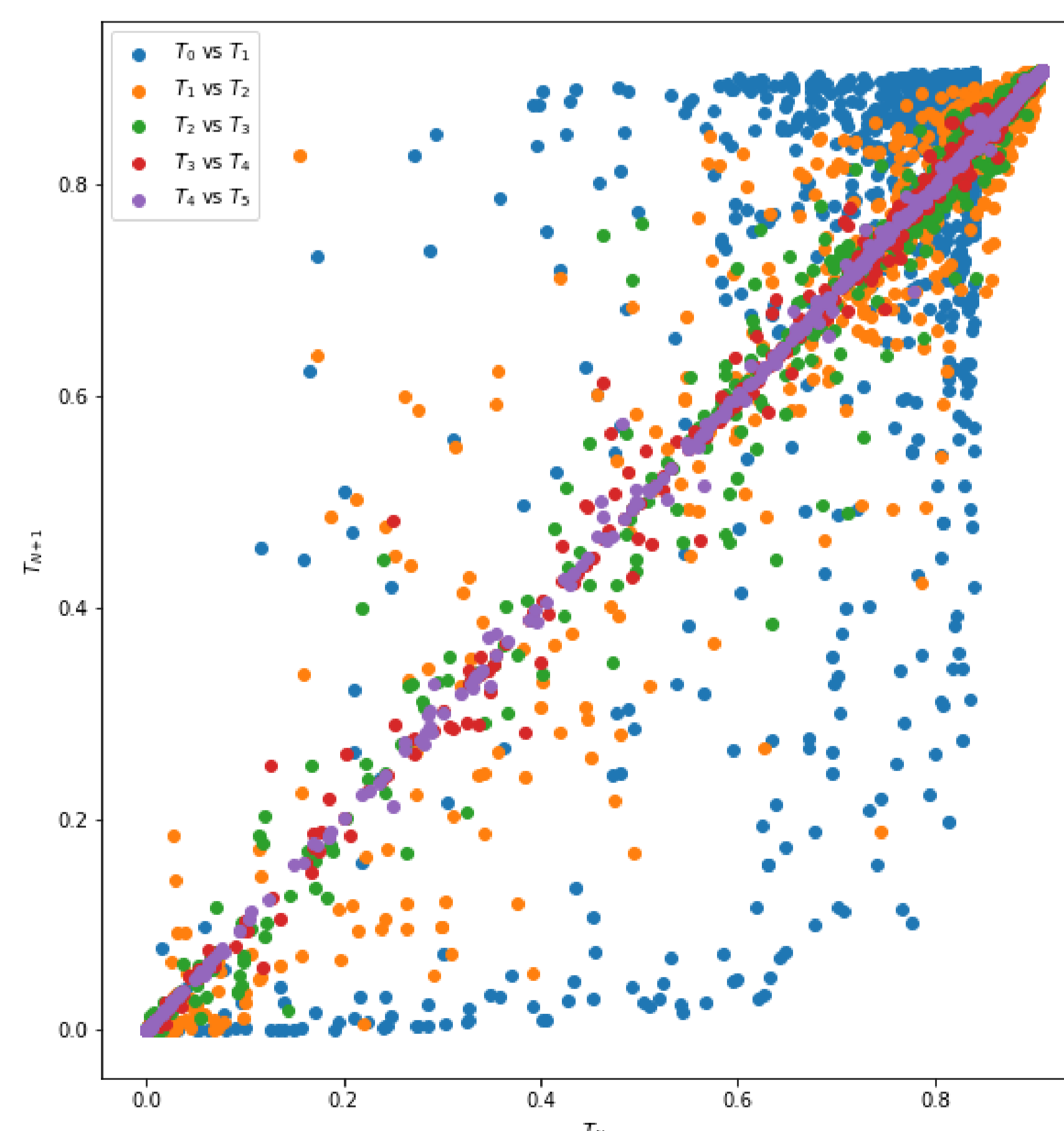
## Numerical Results

		Oracle			
Correlation	Error	$T_0$ -rule	$T_1$ -rule	$T_2$ -rule	
AR(1)	mFDR	0.0501	0.0499	0.0508	
	TP	61.19	124.08	139.36	
Long Range	mFDR	0.0489	0.0506	0.0497	
	TP	61.95	84.25	88.37	
Equi	mFDR	<b>0.0731</b>	0.0515	<b>0.0501</b>	
	TP	62.02	121.37	146.29	

**Table 1.** Power Gain with different covariance structure (Known Parameters). Fixed Parameters in all Simulations,  $K = 1000$ ,  $\pi = 0.3$ ,  $b = 0$ ,  $\tau = 2$

		Data Driven				
Correlation	Error	$T_0$ -rule	$T_1$ -rule	$T_2$ -rule	Sun&Cai	
AR(1)	mFDR	0.051	0.052	0.052	0.044	
	TP	248	336	<b>344</b>	237	
AR(1)	mFDR	0.0522	0.052	0.052	0.050	
	TP	136	202	<b>206</b>	134	

**Table 2.** Power Gain with different covariance structure. Fixed Parameters in all Simulations  $b = 0$ ,  $\tau = 2$ ,  $K = 4000$ ,  $\rho = 0.5$



**Figure 1.** AR(1) covariance with  $\rho = 0.8$ ,  $K=1000$ ,  $\tau=2$ ,  $\pi=0.3$ ,  $b=0.2$

## UK Biobank Height Data [1] Analysis

- We have height data of around 500K individuals.
- About 20K SNPs in chromosome 20.
- After pre-processing we selected 3.5K SNPs to analyze simultaneously.

Method	Number of Common Identified SNPs						
	S&C	BH	ABH	$T_0$ -rule	$T_1$ -rule	$T_2$ -rule	$T_3$ -rule
S&C	<b>36</b>	20	34	36	33	35	32
BH		<b>20</b>	20	20	19	20	20
ABH			<b>36</b>	36	33	35	33
$T_0$ -rule				<b>49</b>	43	45	40
$T_1$ -rule					<b>53</b>	52	49
$T_2$ -rule						<b>66</b>	60
$T_3$ -rule							<b>64</b>

**Table 3.** S&C indicates the Sun and Cai procedure  
BH indicates Benjamini-Hochberg procedure  
ABH indicates Adaptive Benjamini-Hochberg procedure

## Future Work

- Assessing replicability of findings in GWAS using dependence.
- Extending the approach to other models such as taking HMM structure of hypothesis states.
- Robust estimation method under any dependence.

## References

- [1] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.
- [2] Wenguang Sun and T Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901-912, 2007.
- [3] Jichun Xie, T Tony Cai, John Maris, and Hongzhe Li. Optimal false discovery rate control for dependent data. *Statistics and its interface*, 4(4):417, 2011.

## Acknowledgments

This study was supported by Israeli Science Foundation grant 2180/20. UK Biobank research has been conducted using the UK Biobank Resource under Application Number 56885.

