Optimal control of false discovery criteria in the general two-group model

Ruth Heller

www.math.tau.ac.il/∼ruheller

Joint work with Saharon Rosset

The two group model[1]

The observed test statistics $Z_1, \ldots, Z_K$ are assumed to be generated independently from the mixture model

$$Z \sim (1-\pi)g(\cdot \mid h=0) + \pi g(\cdot \mid h=1)$$

where:

- $h \sim Bernoulli(\pi)$;
  $\pi =$ Probability that the test statistic's null hypothesis is false.
- $g(\cdot \mid h=1) =$ The non-null density of Z (if $h=1$).
- $g(\cdot \mid h=0) =$ The null density of Z (if $h=0$).

[1] Efron, Tibshirani, Storey and Tusher (2001), *Empirical Bayes analysis of a microarray experiment.*

The two group model[1]

The observed test statistics $Z_1, \ldots, Z_K$ are assumed to be generated independently from the mixture model

$$Z \sim (1-\pi)g(\cdot \mid h=0) + \pi g(\cdot \mid h=1)$$

where:

- $h \sim Bernoulli(\pi)$;
  $\pi =$ Probability that the test statistic's null hypothesis is false.
- $g(\cdot \mid h=1) =$ The non-null density of Z (if $h=1$).
- $g(\cdot \mid h=0) =$ The null density of Z (if $h=0$).

Goal: Based on the observed $Z_1, \ldots, Z_K$, to discover as many non-null hypotheses ($h_k = 1$) as possible, while controlling for false discoveries.

[1] Efron, Tibshirani, Storey and Tusher (2001), *Empirical Bayes analysis of a microarray experiment*.

The general two group model[2]

- $\vec{h} = (h_1, \ldots, h_K)$ vector of hypotheses states with iid $Bernoulli(\pi)$ coordinates.
- $\vec{Z} = (Z_1, \ldots, Z_k)$ are sampled from the joint distribution given $\vec{h}$:

$$\vec{Z} \mid \vec{h} \sim g(\vec{z} \mid \vec{h})$$

[2] Xie, Cai, Mariz, Li (2011), *Optimal false discovery rate control for dependent data.*

The general two group model[2]

- $\vec{h} = (h_1, \ldots, h_K)$ vector of hypotheses states with iid $Bernoulli(\pi)$ coordinates.
- $\vec{Z} = (Z_1, \ldots, Z_k)$ are sampled from the joint distribution given $\vec{h}$:

$$\vec{Z} \mid \vec{h} \sim g(\vec{z} \mid \vec{h})$$

- For example, a reasonable model for the test statistics in GWAS studies is the mutlivariate mixture normal model:

$$\vec{Z} \mid \vec{h} \sim N\left(\beta\vec{h}, \Sigma + \tau^2 \times diag(\vec{h})\right).$$

[2] Xie, Cai, Mariz, Li (2011), *Optimal false discovery rate control for dependent data.*

The general two group model[2]

- $\vec{h} = (h_1, \ldots, h_K)$ vector of hypotheses states with iid $Bernoulli(\pi)$ coordinates.
- $\vec{Z} = (Z_1, \ldots, Z_k)$ are sampled from the joint distribution given $\vec{h}$:

$$\vec{Z} \mid \vec{h} \sim g(\vec{z} \mid \vec{h})$$

- For example, a reasonable model for the test statistics in GWAS studies is the mutlivariate mixture normal model:

$$\vec{Z} \mid \vec{h} \sim N\left(\beta\vec{h}, \Sigma + \tau^2 \times diag(\vec{h})\right).$$

Goal: Based on the observed $Z_1, \ldots, Z_K$, to discover as many non-null hypotheses ($h_k = 1$) as possible, while controlling for false discoveries.

[2] Xie, Cai, Mariz, Li (2011), *Optimal false discovery rate control for dependent data.*

Notation

We define the decision vector $\vec{D}(\vec{z}) = (D_1(\vec{z}), \ldots, D_K(\vec{z}))$, where :

$$D_k(\vec{z}) = \begin{cases} 1 & \text{if reject null hypothesis } k, \\ 0 & \text{otherwise .} \end{cases}$$

Notation

We define the decision vector $\vec{D}(\vec{z}) = (D_1(\vec{z}), \ldots, D_K(\vec{z}))$, where :

$$D_k(\vec{z}) = \begin{cases} 1 & \text{if reject null hypothesis } k, \\ 0 & \text{otherwise .} \end{cases}$$

The number of rejected and falsely rejected null hypotheses are:

$$R(\vec{D}(\vec{z})) = \sum_{k=1}^{K} D_k(\vec{z}), \quad V(\vec{D}(\vec{z})) = \sum_{k=1}^{K} D_k(\vec{z})(1 - h_k).$$

Notation

We define the decision vector $\vec{D}(\vec{z}) = (D_1(\vec{z}), \ldots, D_K(\vec{z}))$, where :

$$D_k(\vec{z}) = \begin{cases} 1 & \text{if reject null hypothesis } k, \\ 0 & \text{otherwise .} \end{cases}$$

The number of rejected and falsely rejected null hypotheses are:

$$R(\vec{D}(\vec{z})) = \sum_{k=1}^{K} D_k(\vec{z}), \quad V(\vec{D}(\vec{z})) = \sum_{k=1}^{K} D_k(\vec{z})(1 - h_k).$$

Popular error rates for the two-group model are[1].:

$$\text{pFDR}(\vec{D}) : \mathbb{E}\left(\frac{V}{R} \mid R > 0\right); \quad \text{mFDR}(\vec{D}) : \frac{\mathbb{E}V}{\mathbb{E}R}.$$

$$\text{FDR}(\vec{D}) : \mathbb{E}\left(\frac{V}{\max(R, 1)}\right) = \text{pFDR}(\vec{D})\Pr(R > 0).$$

[1] Storey, J. (2003), *The positive false discovery rate: A Bayesian interpretation and the q-value*

Goal: optimal policy with false discovery control

We seek to find the $\vec{D}$ that maximizes the expected number of true discoveries,

$$\max_{\vec{D}:\mathbb{R}^K \to \{0,1\}^K} \mathbb{E}(R - V) = \mathbb{E}(\vec{h}^t \vec{D}),$$

subject to

$$Err(\vec{D}) \leq \alpha,$$

where $Err(\vec{D}) \in \{\text{pFDR}(\vec{D}), \text{FDR}(\vec{D}), \text{mFDR}(\vec{D})\}$.

Goal: optimal policy with false discovery control

We seek to find the $\vec{D}$ that maximizes the expected number of true discoveries,

$$\max_{\vec{D}:\mathbb{R}^K \to \{0,1\}^K} \mathbb{E}(R - V) = \mathbb{E}(\vec{h}^t \vec{D}),$$

subject to

$$Err(\vec{D}) \leq \alpha,$$

where $Err(\vec{D}) \in \{\text{pFDR}(\vec{D}), \text{FDR}(\vec{D}), \text{mFDR}(\vec{D})\}$.

The optimal multiple testing (OMT) policy with *Err* control, OMT-Err, is denoted by $\vec{D}^*$.

Definition of the central statistic for the optimal policies

- The locFDR for the ith hypothesis is[1]

$$T_i(\vec{z}) = \Pr(h_i = 0 \mid \vec{z}) = \frac{(1-\pi)g(\vec{z} \mid h_i = 0)}{(1-\pi)g(\vec{z} \mid h_i = 0) + \pi g(\vec{z} \mid h_i = 1)},$$

where $g(\vec{z} \mid h_i)$ is the joint density of $\vec{z}$ given hypothesis state h_i only, rather than the entire vector $\vec{h}$.

[1] Xie, Cai, Mariz, Li (2011), *Optimal false discovery rate control for dependent data.*

[2] Efron, Tibshirani, Storey and Tusher (2001), *Empirical Bayes analysis of a microarray experiment.*

Definition of the central statistic for the optimal policies

- The locFDR for the ith hypothesis is[1]

$$T_i(\vec{z}) = \Pr(h_i = 0 \mid \vec{z}) = \frac{(1-\pi)g(\vec{z} \mid h_i = 0)}{(1-\pi)g(\vec{z} \mid h_i = 0) + \pi g(\vec{z} \mid h_i = 1)},$$

  where $g(\vec{z} \mid h_i)$ is the joint density of $\vec{z}$ given hypothesis state h_i only, rather than the entire vector $\vec{h}$.

- The marginal locFDR for the ith hypothesis is [2]

$$T_{marg}(z_i) = \Pr(h_i = 0 \mid z_i).$$

[1] Xie, Cai, Mariz, Li (2011), *Optimal false discovery rate control for dependent data.*

[2] Efron, Tibshirani, Storey and Tusher (2001), *Empirical Bayes analysis of a microarray experiment.*

Definition of the central statistic for the optimal policies

- The locFDR for the ith hypothesis is[1]

$$T_i(\vec{z}) = \Pr(h_i = 0 \mid \vec{z}) = \frac{(1-\pi)g(\vec{z} \mid h_i = 0)}{(1-\pi)g(\vec{z} \mid h_i = 0) + \pi g(\vec{z} \mid h_i = 1)},$$

  where $g(\vec{z} \mid h_i)$ is the joint density of $\vec{z}$ given hypothesis state h_i only, rather than the entire vector $\vec{h}$.

- The marginal locFDR for the ith hypothesis is [2]

$$T_{marg}(z_i) = \Pr(h_i = 0 \mid z_i).$$

- For the standard (i.i.d) two-group model,

$$T_i(\vec{z}) = \Pr(h_i = 0 \mid \vec{z}) = \Pr(h_i = 0 \mid z_i) = T_{marg}(z_i).$$

[1] Xie, Cai, Mariz, Li (2011), *Optimal false discovery rate control for dependent data.*

[2] Efron, Tibshirani, Storey and Tusher (2001), *Empirical Bayes analysis of a microarray experiment.*

The optimal policy with mFDR control

OMT-mFDR is a single step procedure:

threshold the locFDR with a fixed threshold[1],

$$D_i^*(\vec{z}) = \mathbb{I}\{T_i(\vec{z}) \leq C_{mFDR}\},$$

where C_{mFDR} is the largest value among all rejection policies of the form $T \leq t$, which guarantees mFDR $= \alpha$.

[1] Xie, Cai, Mariz, Li (2011), *Optimal false discovery rate control for dependent data.; Sun, W. and Cai, T. (2007), Oracle and adaptive compound decision rules for false discovery rate control.*

Main results

- For OMT-FDR and OMT-pFDR, the policy is also to reject the hypotheses with smallest locFDR values.

Main results

- For OMT-FDR and OMT-pFDR, the policy is also to reject the hypotheses with smallest locFDR values.
- The threshold is a function of the entire set of test statistics.

Main results

- For OMT-FDR and OMT-pFDR, the policy is also to reject the hypotheses with smallest locFDR values.
- The threshold is a function of the entire set of test statistics.
- We have an efficient algorithm for finding this threshold.

Outline for the remaining of the talk

1. Solving the optimization problem for $Err(\vec{D}) \in \{pFDR(\vec{D}), FDR(\vec{D})\}$
   - The mathematical solution
   - An efficient step-down algorithm
2. Numerical comparisons with thousands of hypotheses
   - Simulations that show: the power increase of the OMT procedures over their marginal counterparts can be very large; when power is low, OMT-pFDR has a more attractive policy than OMT-FDR and makes more discoveries than OMT-mFDR.
   - Gene expression data analysis
3. Summary and future work

The objective and constraint for the optimization problem

The joint density of $\vec{z}$ is

$$\mathbb{P}(\vec{z}) = \sum_{\vec{h}} g(\vec{z} \mid \vec{h}) \pi^{\vec{1}^t \vec{h}} (1-\pi)^{K - \vec{1}^t \vec{h}}.$$

The objective and constraint for the optimization problem

The joint density of $\vec{z}$ is

$$\mathbb{P}(\vec{z}) = \sum_{\vec{h}} g(\vec{z} \mid \vec{h}) \pi^{\vec{1}^t \vec{h}} (1-\pi)^{K - \vec{1}^t \vec{h}}.$$

- The objective is linear in $\vec{D}$:

$$\mathbb{E}(\vec{h}^t \vec{D}) = \int_{\mathbb{R}^K} \sum_{i=1}^{K} D_i(\vec{z})(1 - T_i(\vec{z}))\mathbb{P}(\vec{z}) d\vec{z},$$

The objective and constraint for the optimization problem

The joint density of $\vec{z}$ is

$$\mathbb{P}(\vec{z}) = \sum_{\vec{h}} g(\vec{z} \mid \vec{h}) \pi^{\vec{1}^t \vec{h}} (1-\pi)^{K - \vec{1}^t \vec{h}}.$$

- The objective is linear in $\vec{D}$:

$$\mathbb{E}(\vec{h}^t \vec{D}) = \int_{\mathbb{R}^K} \sum_{i=1}^{K} D_i(\vec{z})(1 - T_i(\vec{z}))\mathbb{P}(\vec{z}) d\vec{z},$$

- The constraint appears nonlinear in $\vec{D}$:

$$FDR(\vec{D}) = \int_{\mathbb{R}^K} \sum_{i=1}^{K} \frac{D_i(\vec{z})}{\vec{1}^t \vec{D}(\vec{z})} T_i(\vec{z}) \mathbb{P}(\vec{z}) d\vec{z} \leq \alpha,$$

$$pFDR(\vec{D}) = \frac{FDR(\vec{D})}{\int_{\mathbb{R}^K} \mathbb{I}\{\vec{1}^t \vec{D}(\vec{z}) > 0\} \mathbb{P}(\vec{z}) d\vec{z}} \leq \alpha,$$

Challenges in finding OMT-FDR and OMT-pFDR

The OMT solution is the decision function,

$$\vec{D} : \mathbb{R}^K \to \{0,1\}^K,$$

that maximizes the objective while controlling for the constraint.

Challenges in finding OMT-FDR and OMT-pFDR

The OMT solution is the decision function,

$$\vec{D} : \mathbb{R}^K \to \{0,1\}^K,$$

that maximizes the objective while controlling for the constraint.

The challenges seem great:

- The constraint appears to be not linear in $\vec{D}$.
- The optimization is over an infinite number of variables.
- This is a discrete optimization problem, which can be hard to solve even in finite dimensional cases.

Towards an exact solution: monotonicity and linearity

- Theorem: The optimal solution is *weakly monotone* in the locFDR values:

$$T_i(\vec{z}) \geq T_j(\vec{z}) \Leftrightarrow D_i^*(\vec{z}) \leq D_j^*(\vec{z}).$$

Towards an exact solution: monotonicity and linearity

- Theorem: The optimal solution is *weakly monotone* in the locFDR values:

$$T_i(\vec{z}) \geq T_j(\vec{z}) \Leftrightarrow D_i^*(\vec{z}) \leq D_j^*(\vec{z}).$$

- Given weak monotonicity , it turns out the constraints we consider are linear in $\vec{D}$.

Towards an exact solution: monotonicity and linearity

- Theorem: The optimal solution is *weakly monotone* in the locFDR values:

$$T_i(\vec{z}) \geq T_j(\vec{z}) \Leftrightarrow D_i^*(\vec{z}) \leq D_j^*(\vec{z}).$$

- Given weak monotonicity , it turns out the constraints we consider are linear in $\vec{D}$.

- We shall formalize the OMT problem for finding $\tilde{D}(\vec{z})$, where

$$\tilde{D}_k(\vec{z}) = D_{i_k}(\vec{z}), k = 1, \ldots, K,$$

for the sorting permutation $i_1, \ldots, i_K$ so $T_{i_1}(\vec{z}) \leq \ldots \leq T_{i_K}(\vec{z})$.

The objective and constraint for OMT-FDR and OMT-pFDR

Let $T_{(1)}(\vec{z}) \leq T_{(2)}(\vec{z}) \leq \ldots \leq T_{(K)}(\vec{z})$ and $\bar{T}_{k-1}(\vec{z}) = \frac{\sum_{l=1}^{k-1} T_{(l)}(\vec{z})}{k-1}$.

The objective and constraint for OMT-FDR and OMT-pFDR

Let $T_{(1)}(\vec{z}) \leq T_{(2)}(\vec{z}) \leq \ldots \leq T_{(K)}(\vec{z})$ and $\bar{T}_{k-1}(\vec{z}) = \frac{\sum_{l=1}^{k-1} T_{(l)}(\vec{z})}{k-1}$.

The objective is

$$\int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \sum_{i=1}^{K} \tilde{D}_i(\vec{z})(1 - T_{(i)}(\vec{z}))d\vec{z}$$

The objective and constraint for OMT-FDR and OMT-pFDR

Let $T_{(1)}(\vec{z}) \leq T_{(2)}(\vec{z}) \leq \ldots \leq T_{(K)}(\vec{z})$ and $\bar{T}_{k-1}(\vec{z}) = \frac{\sum_{l=1}^{k-1} T_{(l)}(\vec{z})}{k-1}$.

The objective is

$$\int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \sum_{i=1}^{K} \tilde{D}_i(\vec{z})(1 - T_{(i)}(\vec{z}))d\vec{z}$$

The FDR constraint is

$$\int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \left[\tilde{D}_1(\vec{z}) T_{(1)}(\vec{z}) + \sum_{k=2}^{K} \tilde{D}_i(\vec{z}) \frac{1}{k} \left(T_{(k)}(\vec{z}) - \bar{T}_{k-1}(\vec{z}) \right) \right] d\vec{z} \leq \alpha$$

The objective and constraint for OMT-FDR and OMT-pFDR

Let $T_{(1)}(\vec{z}) \leq T_{(2)}(\vec{z}) \leq \ldots \leq T_{(K)}(\vec{z})$ and $\bar{T}_{k-1}(\vec{z}) = \frac{\sum_{l=1}^{k-1} T_{(l)}(\vec{z})}{k-1}$.

The objective is

$$\int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \sum_{i=1}^{K} \tilde{D}_i(\vec{z})(1 - T_{(i)}(\vec{z}))d\vec{z}$$

The FDR constraint is

$$\int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \left[\tilde{D}_1(\vec{z}) T_{(1)}(\vec{z}) + \sum_{k=2}^{K} \tilde{D}_i(\vec{z}) \frac{1}{k} \left(T_{(k)}(\vec{z}) - \bar{T}_{k-1}(\vec{z}) \right) \right] d\vec{z} \leq \alpha$$

The pFDR constraint is

$$\int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \left[\tilde{D}_1(\vec{z})(T_{(1)}(\vec{z}) - \alpha) + \sum_{k=2}^{K} \tilde{D}_i(\vec{z}) \frac{1}{k} \left(T_{(k)}(\vec{z}) - \bar{T}_{k-1}(\vec{z}) \right) \right] d\vec{z} \leq 0$$

The linear program

We can put all our OMT problems in generic form:

$$\max_{\vec{D}:\mathbb{R}^K \to \{0,1\}^K} \quad \int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \sum_{i=1}^{K} \tilde{D}_i(\vec{z}) a_i(\vec{z}) d\vec{z}$$

$$\text{s.t.} \quad \int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \sum_{i=1}^{K} \tilde{D}_i(\vec{z}) b_i(\vec{z}) d\vec{z} \leq c_{Err},$$

$$\tilde{D}_1(\vec{z}) \geq \tilde{D}_2(\vec{z}) \geq \ldots \geq \tilde{D}_K(\vec{z}), \ \forall \vec{z} \in \mathbb{R}^K,$$

The linear program

We can put all our OMT problems in generic form:

$$\max_{\vec{D}:\mathbb{R}^K \to \{0,1\}^K} \quad \int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \sum_{i=1}^{K} \tilde{D}_i(\vec{z}) a_i(\vec{z}) d\vec{z}$$

$$\text{s.t.} \quad \int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \sum_{i=1}^{K} \tilde{D}_i(\vec{z}) b_i(\vec{z}) d\vec{z} \leq c_{Err},$$

$$\tilde{D}_1(\vec{z}) \geq \tilde{D}_2(\vec{z}) \geq \ldots \geq \tilde{D}_K(\vec{z}), \ \forall \vec{z} \in \mathbb{R}^K,$$

We relax the integer requirement, and end up with an infinite linear program to find the optimal $\tilde{D} : \mathbb{R}^K \to [0,1]^K$, which we prove has to be integer almost everywhere.

The step-down OMT procedure

1. For $\mu > 0$:

$$
\begin{aligned}
R_k(\vec{z}) &= a_k(\vec{z}) - \mu b_k(\vec{z}), k = 1, \ldots, K. \\
\tilde{D}_1^{\mu}(\vec{z}) &= \mathbb{I}\left\{ \cup_{l=1}^{K} \left(\sum_{k=1}^{l} R_k(\vec{z}) > 0 \right) \right\} \\
\tilde{D}_i^{\mu}(\vec{z}) &= \tilde{D}_{i-1}^{\mu}(\vec{z}) \times \mathbb{I}\left\{ \cup_{l=i}^{K} \left(\sum_{k=i}^{l} R_k(\vec{z}) > 0 \right) \right\}, \; i = 2, \ldots, K,
\end{aligned}
$$

2. We seek μ^* that satisfies

$$
\int_{\mathbb{R}^K} \mathbb{P}(\vec{z}) \left(\sum_{i=1}^{K} b_i(\vec{z}) \tilde{D}_i^{\mu^*}(\vec{z}) \right) d\vec{z} = c_{Err}.
$$

The optimal solution is $\tilde{D}^*(\vec{z}) = \tilde{D}^{\mu^*}(\vec{z})$.

Numerical comparisons

We compare the performance of the following procedures:

- OMT-FDR, OMT-pFDR, OMT-mFDR: the OMT procedure with FDR, pFDR and mFDR control, respectively.

Numerical comparisons

We compare the performance of the following procedures:

- OMT-FDR, OMT-pFDR, OMT-mFDR: the OMT procedure with FDR, pFDR and mFDR control, respectively.
- marg-FDR, marg-pFDR, marg-mFDR: the sub-optimal counterparts based on the marginal locFDRs.

Numerical comparisons

We compare the performance of the following procedures:

- OMT-FDR, OMT-pFDR, OMT-mFDR: the OMT procedure with FDR, pFDR and mFDR control, respectively.
- marg-FDR, marg-pFDR, marg-mFDR: the sub-optimal counterparts based on the marginal locFDRs.
- ind-FDR, ind-pFDR, ind-mFDR: the misspecified counterparts based on the iid assumption in the two-group model.

Competitors

- BH and adaptive BH[1], for which the threshold for significance of the ith largest p-value is $\frac{i\alpha}{K(1-\pi)}$ instead of the BH threshold $\frac{i\alpha}{K}$.
- est-mFDR[2], which first orders the marginal locFDRs,

$$T_{marg,(1)} \leq \ldots \leq T_{marg,(K)},$$

and then rejects the k hypotheses with smallest marginal locFDRs, where

$$k = \max\{i : \frac{1}{i}\sum_{j=1}^{i} T_{marg,(j)} \leq \alpha\}.$$

[1] Benjamini, Y., Krieger, A., and Yekutieli, D. (2006), *Adaptive linear step-up procedures that control the false discovery rate.*

[2] Sun, W. and Cai, T. (2007), *Oracle and adaptive compound decision rules for false discovery rate control*

The general two-group model

A $K = 5000$ dimensional multivariate mixture normal model:

- $h_1, \ldots, h_K$ is an iid sample from $Bernoulli(0.3)$.
- Given $\vec{h}$, the distribution of the test statistics is

$$\vec{Z} \mid \vec{h} \sim N\left(-1.5\vec{h}, \Sigma + 0.01 \times diag(\vec{h})\right).$$

where Σ is a block diagonal matrix with blocks

$$\begin{pmatrix} 1 & \rho_b & \rho_b & \rho_b & \rho_b \\ \rho_b & 1 & \rho_b & \rho_b & \rho_b \\ \rho_b & \rho_b & 1 & \rho_b & \rho_b \\ \rho_b & \rho_b & \rho_b & 1 & \rho_b \\ \rho_b & \rho_b & \rho_b & \rho_b & 1 \end{pmatrix}$$

and $\rho_b \in \{0.1, 0.5\}$ for block $b \in \{1, \ldots, 1000\}$.

The locFDRs computational complexity

- $\vec{T}_{marg}$ requires $O(K)$ calculations.
- $T_i(\vec{z})$ requires $O(2^K)$ calculations with a very naive implementation that considers all possible allocations of the vector $\vec{h}$.
  - e.g., $g(\vec{z} \mid h_i = 0) = \sum_{\vec{h} \in \{0,1\}^K : h_i = 0} \pi^{\vec{1}^t \vec{h}} (1-\pi)^{K - \vec{1}^t \vec{h} - 1} g(\vec{z} \mid \vec{h})$.
- In our setting, $\vec{T}(\vec{z})$ requires $O(K \times B \times 2^B)$ calculations for $K = 5000$ consisting of 1000 independent blocks of size $B = 5$

Results for $K = 5000$ z-scores generated from the multivariate mixture normal model.

	$\rho_b = 0.1$				$\rho_b \in \{0.1, 0.5\}$			
	FDR	pFDR	mFDR	TP	FDR	pFDR	mFDR	TP
OMT-FDR	.049	.159	.162	169	.050	.055	.059	**263**
marg-FDR	.050	.176	.179	167	.051	.178	.181	169
ind-FDR	.052	.177	.180	173	.056	.179	.183	185
OMT-pFDR	.051	.051	.147	166	.050	.050	.058	**263**
marg-pFDR	.050	.050	.163	158	.049	.049	.164	154
ind-pFDR	.052	.052	.163	163	.053	.053	.166	168
OMT-mFDR	.050	.050	.050	130	.050	.050	.050	**261**
marg-mFDR	.050	.050	.050	121	.050	.050	.050	121
ind-mFDR	.050	.050	.050	120	.050	.050	.050	121
est-mFDR	.050	.050	.050	120	.050	.050	.050	120
adaptive BH	.050	.050	.051	122	.050	.050	.052	122
BH	.035	.035	.037	73	.035	.035	.037	72

Results for $K = 5000$ z-scores generated from the multivariate mixture normal model.

	$\rho_b = 0.1$				$\rho_b \in \{0.1, 0.5\}$			
	FDR	pFDR	mFDR	TP	FDR	pFDR	mFDR	TP
OMT-FDR	.049	.159	.162	169	.050	.055	.059	**263**
marg-FDR	.050	.176	.179	167	.051	.178	.181	169
ind-FDR	.052	.177	.180	173	.056	.179	.183	185
OMT-pFDR	.051	.051	.147	166	.050	.050	.058	**263**
marg-pFDR	.050	.050	.163	158	.049	.049	.164	154
ind-pFDR	.052	.052	.163	163	.053	.053	.166	168
OMT-mFDR	.050	.050	.050	130	.050	.050	.050	**261**
marg-mFDR	.050	.050	.050	121	.050	.050	.050	121
ind-mFDR	.050	.050	.050	120	.050	.050	.050	121
est-mFDR	.050	.050	.050	120	.050	.050	.050	120
adaptive BH	.050	.050	.051	122	.050	.050	.052	122
BH	.035	.035	.037	73	.035	.035	.037	72

Results for $K = 5000$ z-scores generated from the multivariate mixture normal model.

	$\rho_b = 0.1$				$\rho_b \in \{0.1, 0.5\}$			
	FDR	pFDR	mFDR	TP	FDR	pFDR	mFDR	TP
OMT-FDR	.049	.159	.162	169	.050	.055	.059	**263**
marg-FDR	.050	.176	.179	167	.051	.178	.181	169
ind-FDR	.052	.177	.180	173	.056	.179	.183	185
OMT-pFDR	.051	.051	.147	166	.050	.050	.058	**263**
marg-pFDR	.050	.050	.163	158	.049	.049	.164	154
ind-pFDR	.052	.052	.163	163	.053	.053	.166	168
OMT-mFDR	.050	.050	.050	130	.050	.050	.050	**261**
marg-mFDR	.050	.050	.050	121	.050	.050	.050	121
ind-mFDR	.050	.050	.050	120	.050	.050	.050	121
est-mFDR	.050	.050	.050	120	.050	.050	.050	120
adaptive BH	.050	.050	.051	122	.050	.050	.052	122
BH	.035	.035	.037	73	.035	.035	.037	72

Results for $K = 5000$ z-scores generated from the multivariate mixture normal model.

	$\rho_b = 0.1$				$\rho_b \in \{0.1, 0.5\}$			
	FDR	pFDR	mFDR	TP	FDR	pFDR	mFDR	TP
OMT-FDR	.049	.159	.162	169	.050	.055	.059	**263**
marg-FDR	.050	.176	.179	167	.051	.178	.181	169
ind-FDR	.052	.177	.180	173	.056	.179	.183	185
OMT-pFDR	.051	.051	.147	166	.050	.050	.058	**263**
marg-pFDR	.050	.050	.163	158	.049	.049	.164	154
ind-pFDR	.052	.052	.163	163	.053	.053	.166	168
OMT-mFDR	.050	.050	.050	130	.050	.050	.050	**261**
marg-mFDR	.050	.050	.050	121	.050	.050	.050	121
ind-mFDR	.050	.050	.050	120	.050	.050	.050	121
est-mFDR	.050	.050	.050	120	.050	.050	.050	120
adaptive BH	.050	.050	.051	122	.050	.050	.052	122
BH	.035	.035	.037	73	.035	.035	.037	72

Conclusions from the numerical comparisons

- The power advantage of the OMT procedures over their marginal counterparts can be very large, and is increasing as the dependency increases.
- The policies that incorrectly assumes $\vec{z}$ comes from the two-group model for FDR and pFDR control can have levels above nominal, but for mFDR control the nominal level is maintained. The inflation increases as the dependency increases.
- The power gain of FDR and pFDR policies over the respective mFDR policy is large when the overall power is low, and it is due to high variation in $\frac{V}{\max(R,1)}$ which is manifest in the high mFDR levels. The variation in $\frac{V}{\max(R,1)}$ is greater with FDR control than with pFDR control policies.

Application to gene expression studies

For $K = 15270$ genes, we have the meta-analysis p-values of four studies of ulcerative colitis for up-regulation, and separately for down-regulation, of the genes [1].

	ID	pval.DOWNregulated	pval.UPregulated
1	A1BG	0.99545	0.36632
2	A1CF	0.00000	1.00000
3	A2M	0.99925	0.01869
...	...	...	...
15270	ZZZ3	0.64801	0.91332

[1] Shah, Guo, Wendelsdorf, Lu, Sparks, Tsang (2016), *A crowdsourcing approach for reusing and meta analyzing gene expression data*

Re-analysis methods

- Assuming the p-values are generated from the two group model, we want to compare OMT-FDR and OMT-pFDR with the competitors est-mFDR, adaptive BH and BH.

[1] Muralidharan, O. (2010), *An empirical Bayes mixture method for effect size and false discovery rate estimation*

Re-analysis methods

- Assuming the p-values are generated from the two group model, we want to compare OMT-FDR and OMT-pFDR with the competitors est-mFDR, adaptive BH and BH.
- We need to estimate the mixture components of the two group model for this purpose, and we do this using the R package *mixfdr* available from CRAN [1].

[1] Muralidharan, O. (2010), *An empirical Bayes mixture method for effect size and false discovery rate estimation*

Re-analysis methods

- Assuming the p-values are generated from the two group model, we want to compare OMT-FDR and OMT-pFDR with the competitors est-mFDR, adaptive BH and BH.
- We need to estimate the mixture components of the two group model for this purpose, and we do this using the R package *mixfdr* available from CRAN [1].
- The marginal locFDRs and optimal policy are computed assuming the observed test statistics are generated from the estimated two group model.

[1] Muralidharan, O. (2010), *An empirical Bayes mixture method for effect size and false discovery rate estimation*

Re-analysis methods

- Assuming the p-values are generated from the two group model, we want to compare OMT-FDR and OMT-pFDR with the competitors est-mFDR, adaptive BH and BH.
- We need to estimate the mixture components of the two group model for this purpose, and we do this using the R package *mixfdr* available from CRAN [1].
- The marginal locFDRs and optimal policy are computed assuming the observed test statistics are generated from the estimated two group model.
- OMT-FDR and OMT-pFDR coincide for both up-regulation and down-regulation.

[1] Muralidharan, O. (2010), *An empirical Bayes mixture method for effect size and false discovery rate estimation*

Results

	est-FDR	est-mFDR	adapt-BH	BH
# up regulated	**2409**	2305	2264	2211
# up regulated among confirmed discoveries	**2276**	2219	2189	2145
# down regulated	**2023**	1897	1837	1775
# down regulated among confirmed discoveries	**1815**	1731	1699	1671

Summary

- We have a complete mathematical treatment of OMT procedures for pFDR or FDR control in the general two-group model.

Summary

- We have a complete mathematical treatment of OMT procedures for pFDR or FDR control in the general two-group model.
- Other error measures that fit into the mathematical framework include $FDX = \mathbb{P}(FDP > \gamma)$, $FWER = \mathbb{P}(V > 0)$, $\mathbb{E}(V)$.

Summary

- We have a complete mathematical treatment of OMT procedures for pFDR or FDR control in the general two-group model.
- Other error measures that fit into the mathematical framework include $FDX = \mathbb{P}(FDP > \gamma)$, $FWER = \mathbb{P}(V > 0)$, $\mathbb{E}(V)$.
- For linear objective functions (not just the expected number of true discoveries!), we offer an efficient algorithm for computing the optimal rejection region:
  - for independent test statistics.
  - for the multivariate mixture model when the covariance structure has a block dependence structure.
- We showed the large potential gain from incorporating dependence.
- Paper available at `https://arxiv.org/abs/1902.00892`.

Future work

- We expect the OMT policies to be useful in genomic applications where the dependence is known. Specifically, for GWAS, the covariance is a known banded matrix. We plan to provide efficient computational tools for the general two-group model with this type of local dependence.

Future work

- We expect the OMT policies to be useful in genomic applications where the dependence is known. Specifically, for GWAS, the covariance is a known banded matrix. We plan to provide efficient computational tools for the general two-group model with this type of local dependence. .
- Extend the formulation to control more than one error rate, e.g., seek the OMT policy which controls the FDR as well as $\mathbb{E}(V)$, thus potentially creating a powerful policy with meaningful control over the false discovery proportion in expectation without allowing an unattractive policy which tends to reject many or very few hypotheses.