4-2022

# Empirically Adjusted Weighted Ordered P-values Method

Wimarsha Jayanetti
*Old Dominion University*

Sinjini Sikdar
*Old Dominion University*

N. Rao Chaganty
*Old Dominion University*

# Empirically Adjusted Weighted Ordered P-values Method

WIMARSHA T. JAYANETTI, SINJINI SIKDAR AND N. RAO CHAGANTY

Department of Mathematics and Statistics
Old Dominion University

## Introduction

- Meta analysis is used to integrate summary results from multiple studies targeting the same questions.
- One common practice is to use Fisher's [1] or Stouffer's [2] method for combining the p-values from multiple studies.
- The traditional methods aim at testing the alternative hypothesis that at least one of the studies is non-null.
- In recent years, researchers are more interested in identifying genes which are differentially expressed in majority of studies.
- WOP method [3] combines ordered p-values, weighting them based on their order, assuming p-values from individual studies are uniformly distributed under the null.
- In large-scale multiple testing, empirical distribution of p-values may not be uniform - so adjustments are needed.

## Simulation Study

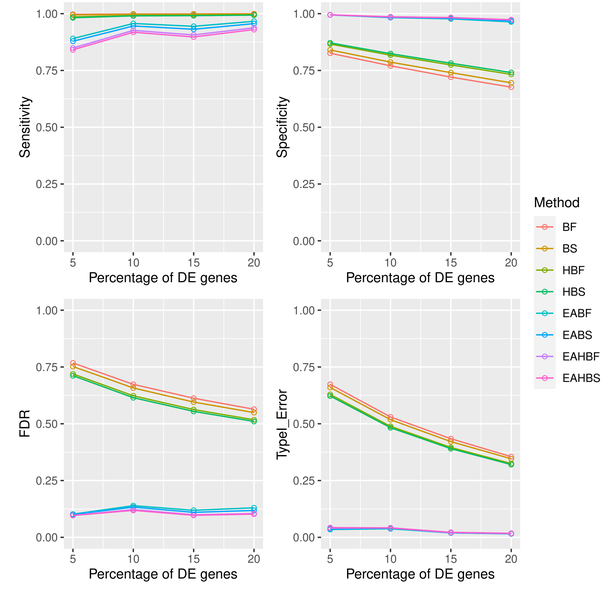
- Considered 10 studies each with 3000 genes.
- 50 genes considered to be differentially expressed (DE) in 1, 2,$\cdots$, 10 studies respectively.
- Considered 10% of the genes to be DE between two groups in at least 5 studies.
- Generated (log) gene expression values for the $i^{th}$ gene of the $k^{th}$ subject in the $j^{th}$ group for each experiment as:

$$y_{ijk} = \mu + G_i + V_j + GV_{ij} + W_{ijk} + e_{ijk}$$

where $\mu$: overall mean effect, $G_i$: effect of the gene, $V_j$: effect of the group, $GV_{ij}$: interaction effect between gene and group, $W_{ijk}$: hidden confounder effect, $e_{ijk}$: random error term

- Set $\mu$, $G_i$ and $V_j$ as zero for simplicity.
- Set the differences in magnitudes of DE genes between the two groups as 8 through the interaction term $GV_{ij}$.
- Generated hidden confounder as $W_{ijk} = u_{ijk}I(s_{ijk} = 1)$, where $s_{ijk} \sim Bernoulli(0.4)$ and $u_{ijk}$'s from Normal distribution such that effect of hidden confounder depends on gene ID, experiment ID and subject group.
- Compared performance of our Empirically Adjusted (EA) methods with original WOP methods. All results were averaged over 500 replicates.

| Method | Sensitivity | Specificity | FDR | Type I Error |
|---|---|---|---|---|
| BF | 0.998 | 0.766 | 0.677 | 0.530 |
| **EABF** | 0.954 | 0.982 | 0.138 | **0.037** |
| BS | 0.997 | 0.783 | 0.660 | 0.517 |
| **EABS** | 0.944 | 0.984 | 0.129 | **0.039** |
| HBF | 0.992 | 0.814 | 0.626 | 0.488 |
| **EAHBF** | 0.925 | 0.985 | 0.119 | **0.041** |
| HBS | 0.989 | 0.821 | 0.618 | 0.482 |
| **EAHBS** | 0.916 | 0.986 | 0.116 | **0.041** |




## Proposed Method

- Consider $K$ independent studies where each study consisting of $G$ genes.
- Let $\theta_{ij}$ denotes the underlying true effect size for the $i^{th}$ gene in the $j^{th}$ study, $i = 1, 2 \ldots, G$: $j = 1, 2, \ldots, K$.

**Hypothesis setting**

For the $i^{th}$ gene,

$$HS_m: \{H_0: \sum_{j=1}^{K} I(\theta_{ij} \neq 0) = 0 \; vs \; H_1^m: \sum_{j=1}^{K} I(\theta_{ij} \neq 0) \geq m\}$$

where $m = \lceil K/2 \rceil$, i.e., $m$ is the smallest integer that is not lower than $K/2$.

**Algorithm**

- **Step 1**: For gene $i$ in study $j$, obtain the p-value $p_{ij}$ for testing the hypothesis of interest.
- **Step 2**: Consider the inverse z-transformation to get the corresponding z-scores as $\Phi^{-1}(p_{ij})$.
- **Step 3**: Let $\hat{\delta}_0$ and $\hat{\sigma}_0$ be estimated mean and standard deviation of the null distribution using central matching method [4]. Modify the z-scores, obtained in step 2 as:

$$z_{ij}^{'} = \frac{z_{ij} - \hat{\delta}_0}{\hat{\sigma}_0}$$

- **Step 4**: Convert the empirically adjusted z-scores into corresponding p-values as:

$$p_{ij}^{'} = \Phi(z_{ij}^{'})$$

- **Step 5**: For a gene $i$ order the p-values over the $K$ independent studies. Let $p_{i(j)}^{'}$ denote the $j^{th}$ ordered p-value for gene $i$. Calculate the summary statistic as in [3] :

$$T_i = \sum_{j=1}^{K} w_j H(p_{i(j)}^{'})$$

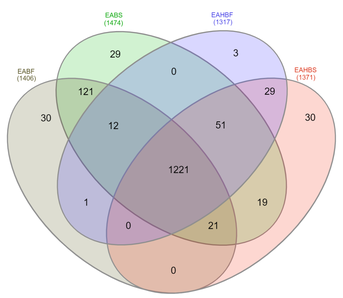
  - Fisher's: $H(p_{i(j)}) = -2log(p_{i(j)})$
  - Stouffer's: $H(p_{i(j)}) = \Phi^{-1}(1 - p_{i(j)})$
  - Binomial: $w_j^b = f(j - 1: K - 1, 0.5), j = 1, 2, ..., K$, where $f(x: n, p)$ denotes the pmf of $Bin(n, p)$ for $x = 0, 1, \ldots, n$
  - Half-binomial: $w_j^{hb} = w_j^b \; for \; m \leq j \leq K \; and \; 0 \; for \; j < m$

- **Step 6**: For gene $i$, obtain p-value $p^i$ by comparing the statistic, defined in step 5, to the numerical distribution by simulating $U(0, 1)$ random variables, $i = 1, 2 \ldots, G$.
- **Step 7**: Finally, apply the Benjamini-Hochberg method to account for multiple testing.

## Data Analysis

- Identified DE genes between two lung cancer types.
- Data includes 5 studies with 7200 genes each.
- Tested the alternative hypothesis that genes are DE in at least 3 experiments out of 5 experiments ($HS_3$).

| Method | Number of DE genes (Percentage) | |
|---|---|---|
| | WOP | EAWOP |
| BF | 4921 (68.3%) | 1406 (19.5%) |
| BS | 4672 (64.9%) | 1474 (20.4%) |
| HBF | 4286 (59.5%) | 1317 (18.3%) |
| HBS | 4208 (58.4%) | 1371 (19.0%) |




- Pathway analysis identified biologically relevant pathways such as cell cycle, p53 signaling pathway, etc.

## Conclusions

- The proposed method has significantly better performance than the original WOP method especially in presence of hidden confounder.
- Type I errors are controlled at 5% for our methods while they are extremely high for the original WOP methods. FDR values are also significantly lower for the proposed methods.
- Our methods have slightly lower sensitivity values but much higher specificity values compared to the original methods.

## References

[1] Fisher R.A. *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, Scotland, 1925.

[2] Stouffer S.A., Suchman E.A., DeVinney L.C., Star S.A., and Williams Jr R.M. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1. 1949.

[3] Li Y. and Ghosh D. Meta-analysis based on weighted ordered p-values for genomic data with heterogeneity. *BMC bioinformatics*, 15(1):1–12, 2014.

[4] Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.