Using the MiniMax Statistic

Gabriel Odom and Lily Wang

8/29/2021

# Introduction and Overview

This vignette will give a brief story about what motivated us to work on this problem, some mathematical details about the method, and walk through some components of an example of how to use it in practice throughout these main sections.

library(flextable)  
library(tidyverse)  
library(pathwayMultiomics)

# Motivating Example

Some time ago, our research team was attempting to use some multi-omics tools to answer a few questions concerning breast cancer. While the tools were often well-documented and had nice statistical properties, we consistently ran into one of the following two constraints:

* Multi-Omics tools must match across platforms *by samples*, or
* Multi-Omics tools must match across platforms *by features*.

## Joining Data Platforms

In many of the widely-available cancer data sets for a single cancer, we could usually find a nice overlapping set of samples between two platforms of data (for instance copy number variation and DNA methylation). However, as soon as we added in a third or fourth platform (such as gene and/or protein expression), the intersection of the samples recorded with these new platforms rapidly approached the empty set. We could start with 600+ samples, and we’d barely have 50 by the time we added a third platform, and only a handful of samples (or even none at all) if we added a fourth.

If, instead, we tried to match on features, we would run into a different set of issues. What is the best way to map from the “widest” data, site- or probe-level data (as in DNA methylation, SNP, or RNA editing levels), to the “thinnest” data, such as expressions of small sets of individual proteins? Using a feature-matching strategy would require us to discard many relevant biological features simply because they were not recorded in the most restricted data.

## A Pathway-Based Solution

Rather than using only the intersection of the samples, or only the intersection of the genomic features, we suggest to match across data sets by data summaries of biological pathway activity. For example, consider a set of three data platforms recorded for colorectal adenocarcinoma: copy number variation, DNA methylation, and protein expression. These three data sets may share samples, but that isn’t required. Instead of trying to create two different many-to-one mappings between regions of the genome and particular proteins, instead, statistically assess the activity of a set of biological pathways in each genomic data layer independently, return a vector of -values corresponding to the pathway collection for each data set, then match the pathway results on these pathways. The resulting summary data would have one row per pathway in the collection, and three columns of -values for each pathway.

In this package, we provide an example of SNP, DNA methylation, and gene expression results for 640 subjects with Alzheimer’s disease (from the ROSMAP study) evaluated over the Broad Institute’s C2 CP collection. We show the first 10 rows (of the 2833 pathways in the collection with fewer than 200 genes or more than 4 genes):

| pathway | snpPval | dnamPval | rnaseqPval |
| --- | --- | --- | --- |
| BIOCARTA\_41BB\_PATHWAY | 0.7792572 | 0.81637091 | 0.037130842 |
| BIOCARTA\_ACE2\_PATHWAY | 0.8778045 | 0.22534182 | 0.001187332 |
| BIOCARTA\_ACETAMINOPHEN\_PATHWAY | 0.2675038 | 0.29729642 | 0.613386613 |
| BIOCARTA\_ACH\_PATHWAY | 0.5626200 | 0.53475657 | 0.038961039 |
| BIOCARTA\_ACTINY\_PATHWAY | 0.4538971 | 0.13514098 | 0.035964036 |
| BIOCARTA\_AGPCR\_PATHWAY | 0.5998621 | 0.01189386 | 0.018774580 |
| BIOCARTA\_AGR\_PATHWAY | 0.4561920 | 0.98926144 | 0.266733267 |
| BIOCARTA\_AHSP\_PATHWAY | 0.7556747 | 0.46082890 | 0.294705295 |
| BIOCARTA\_AKAP13\_PATHWAY | 0.8725796 | 0.32379009 | 0.280719281 |
| BIOCARTA\_AKAP95\_PATHWAY | 0.7723686 | 0.25845492 | 0.036963037 |

# The MiniMax Statistic and its Distribution

## The MiniMax Statistic

From a systems biology perspective, when we perform a multi-omics analysis, we often care about biological processes which are disregulated across *multiple* layers of the protein transcription process. For example, a moderate cascading effect from DNA methylation upstream through gene expression to protein expression may be of more practical significance than a heavily disregulated effect in a single layer of the process. The latter scenario will be easily and quickly detected in single-omics analysis, but the former scenario will be missed.

Therefore, the pathway MiniMax statistic is defined to be the **min**imum -value among all pairwise **max**ima for that biological pathway. Biologically, it indicates if the pathway is disregulated in two or more layers of genomic data. Statistically and mathematically, such a measure is equivalent to requiring that the second smallest -value is significant. That is, we equivalently define the MiniMax to be the second order statistic of the -values.

## The Distribution of the MiniMax Statistic

Let us consider the scenario wherein we test that a single pathway is disregulated for a single data platform . Under the null hypothesis (that there is no signal within that pathway for genomics data set ), the -value of a well-defined statistical test for this question follows a uniform distribution; that is, . Further consider such tests for that pathway, and for the sake of argument, assume that they are independent (a horribly inaccurate assumption to be sure); then under ,

It is a known result that the order statistics for such a collection of uniform random variables follows a Beta distribution (denoted herein). Thus,

## Adjusting the Parameters for Dependence Across Data Platforms

We know that that this assumption of independence will not, and *should* not, hold under real data. We believe that the pathway activity for one layer of genomic data should be related to the activity for that pathway in another genomic data layer. However, we believe that the use of the Beta distribution itself will still be appropriate, but that the parameters need to be adjusted to account for the dependence between .

In order to estimate these Beta distribution parameters, we recommend that you run your single-platform analyses twice: once with the real data, and once with a random permutation of that data (to simulate pathway conditions under ). Once this has been completed, pass the MiniMax statistcs calculated from platform -values for each pathway under to the MiniMax\_estBetaParams() function. This function has options to use either the Method of Moments or Maximum Likelihood to estimate the parameters of this Beta distribution. If you do not have access to the original data (such as in the case of a meta-analysis), then the only option we offer right now is to use the closed-form definitions of these parameters ().

## Continuing the Alzheimer’s Disease Example

In the data set above, the MiniMax statistic and corresponding -value are given below. Note that we assume that ; we do not have access to the original data from the ROSMAP study with which to estimate the parameters as we would like, but showed *in silico* that these fixed values don’t yield horrible performance (more on that shortly). Therefore, we can add the statistic and it’s significance as follows:

### The results table ###  
alzheimersMultiOmics\_df %>%   
 print(width = 70)  
#> # A tibble: 2,833 × 8  
#> pathway size snpPval snpFDR dnamPval dnamFDR rnaseqPval  
#> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
#> 1 BIOCARTA\_GRANU… 14 9.67e- 1 9.99e- 1 0.00241 0.359 0.919   
#> 2 BIOCARTA\_LYM\_P… 11 6.79e- 1 9.99e- 1 0.0156 0.538 0.0749  
#> 3 BIOCARTA\_BLYMP… 12 3.31e-69 1.17e-66 0.138 0.858 0.999   
#> 4 BIOCARTA\_CARM\_… 24 3.30e- 1 9.99e- 1 0.0263 0.643 0.0287  
#> 5 BIOCARTA\_LAIR\_… 17 2.17e- 1 9.99e- 1 0.0463 0.718 0.630   
#> 6 BIOCARTA\_VDR\_P… 24 8.78e- 1 9.99e- 1 0.115 0.840 0.0350  
#> 7 BIOCARTA\_MTA3\_… 18 7.98e- 1 9.99e- 1 0.0494 0.721 0.105   
#> 8 BIOCARTA\_GABA\_… 9 3.71e- 1 9.99e- 1 1 1 0.0280  
#> 9 BIOCARTA\_EGFR\_… 11 6.77e- 1 9.99e- 1 0.0339 0.668 0.118   
#> 10 BIOCARTA\_MONOC… 10 5.26e- 1 9.99e- 1 0.657 1 0.167   
#> # … with 2,823 more rows, and 1 more variable: rnaseqFDR <dbl>  
  
  
### Pathway Multi-Omics Significance with the MiniMax ###  
# We accept the default values of the Beta Distribution  
adMiniMax\_df <-   
 alzheimersMultiOmics\_df %>%   
 # The MiniMax() function takes in the gene set name and the three p-value  
 # columns  
 select(pathway, ends\_with("Pval")) %>%   
 rename(  
 SNP = snpPval, DNAm = dnamPval, RNAseq = rnaseqPval  
 ) %>%   
 MiniMax()  
  
adMiniMax\_df %>%   
 print(width = 70)  
#> # A tibble: 2,833 × 7  
#> pathway SNP DNAm RNAseq MiniMax MiniMaxP drivers   
#> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>   
#> 1 PID\_PDGFRB\_PAT… 6.99e- 1 0.000145 1.67e-4 1.67e-4 8.33e-8 DNAm an…  
#> 2 WP\_CHEMOKINE\_S… 8.19e- 1 0.000317 1.94e-5 3.17e-4 3.02e-7 DNAm an…  
#> 3 KEGG\_HEMATOPOI… 3.67e-36 0.000340 7.61e-1 3.40e-4 3.48e-7 DNAm an…  
#> 4 PID\_TCR\_PATHWAY 4.48e- 4 0.0275 4.90e-4 4.90e-4 7.20e-7 RNAseq …  
#> 5 WP\_REGULATION\_… 3.76e- 5 0.0332 6.55e-4 6.55e-4 1.29e-6 RNAseq …  
#> 6 KEGG\_CHEMOKINE… 7.90e- 1 0.000672 2.98e-4 6.72e-4 1.35e-6 DNAm an…  
#> 7 PID\_KIT\_PATHWAY 2.55e- 1 0.000684 1.10e-4 6.84e-4 1.40e-6 DNAm an…  
#> 8 WP\_KIT\_RECEPTO… 3.05e- 2 0.000368 1.41e-3 1.41e-3 5.94e-6 DNAm an…  
#> 9 PID\_CXCR4\_PATH… 4.87e- 4 0.00150 7.29e-2 1.50e-3 6.76e-6 DNAm an…  
#> 10 REACTOME\_TCR\_S… 6.06e-52 0.207 2.16e-3 2.16e-3 1.40e-5 RNAseq …  
#> # … with 2,823 more rows

Finally, we can adjust these pathway -values for multiple comparisons, and filter to those most significant.

adRes\_df <-   
 adMiniMax\_df %>%   
 mutate(MiniMaxFDR = p.adjust(MiniMaxP, method = "fdr")) %>%   
 filter(MiniMaxFDR < 0.01)   
  
adRes\_df %>%   
 select(pathway, MiniMaxFDR, drivers) %>%   
 mutate(pathway = str\_trunc(pathway, width = 45)) %>%   
 qflextable()

| pathway | MiniMaxFDR | drivers |
| --- | --- | --- |
| PID\_PDGFRB\_PATHWAY | 0.0002359943 | DNAm and RNAseq |
| WP\_CHEMOKINE\_SIGNALING\_PATHWAY | 0.0003282175 | DNAm and RNAseq |
| KEGG\_HEMATOPOIETIC\_CELL\_LINEAGE | 0.0003282175 | DNAm and SNP |
| PID\_TCR\_PATHWAY | 0.0005099928 | RNAseq and SNP |
| WP\_REGULATION\_OF\_TOLLLIKE\_RECEPTOR\_SIGNALI... | 0.0005686064 | RNAseq and SNP |
| KEGG\_CHEMOKINE\_SIGNALING\_PATHWAY | 0.0005686064 | DNAm and RNAseq |
| PID\_KIT\_PATHWAY | 0.0005686064 | DNAm and RNAseq |
| WP\_KIT\_RECEPTOR\_SIGNALING\_PATHWAY | 0.0021048227 | DNAm and RNAseq |
| PID\_CXCR4\_PATHWAY | 0.0021267169 | DNAm and SNP |
| REACTOME\_TCR\_SIGNALING | 0.0039757422 | RNAseq and SNP |
| REACTOME\_HATS\_ACETYLATE\_HISTONES | 0.0071556422 | RNAseq and SNP |
| REACTOME\_MHC\_CLASS\_II\_ANTIGEN\_PRESENTATION | 0.0071556422 | RNAseq and SNP |
| WP\_INSULIN\_SIGNALING | 0.0071556422 | DNAm and RNAseq |
| REACTOME\_RAC1\_GTPASE\_CYCLE | 0.0076144591 | DNAm and RNAseq |
| REACTOME\_DDX58\_IFIH1\_MEDIATED\_INDUCTION\_OF... | 0.0076144591 | RNAseq and SNP |
| KEGG\_NEUROTROPHIN\_SIGNALING\_PATHWAY | 0.0076144591 | DNAm and RNAseq |
| WP\_TRANSCRIPTION\_FACTOR\_REGULATION\_IN\_ADIP... | 0.0077424430 | DNAm and RNAseq |
| REACTOME\_SUMOYLATION\_OF\_CHROMATIN\_ORGANIZA... | 0.0094064793 | RNAseq and SNP |
| REACTOME\_HCMV\_INFECTION | 0.0094064793 | RNAseq and SNP |
| REACTOME\_NEUROTRANSMITTER\_RECEPTORS\_AND\_PO... | 0.0097309051 | RNAseq and SNP |

Many of these pathways have been implicated in Alzheimer’s Disease or in neurodegenerative conditions at large.

# Session Information

sessionInfo()  
#> R version 4.1.0 (2021-05-18)  
#> Platform: x86\_64-apple-darwin17.0 (64-bit)  
#> Running under: macOS Big Sur 10.16  
#>   
#> Matrix products: default  
#> BLAS: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.dylib  
#> LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib  
#>   
#> locale:  
#> [1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
#>   
#> attached base packages:  
#> [1] stats graphics grDevices utils datasets methods base   
#>   
#> other attached packages:  
#> [1] pathwayMultiomics\_0.0.0.9003 forcats\_0.5.1   
#> [3] stringr\_1.4.0 dplyr\_1.0.7   
#> [5] purrr\_0.3.4 readr\_2.0.0   
#> [7] tidyr\_1.1.3 tibble\_3.1.3   
#> [9] ggplot2\_3.3.5 tidyverse\_1.3.1   
#> [11] flextable\_0.6.8   
#>   
#> loaded via a namespace (and not attached):  
#> [1] Rcpp\_1.0.7 lubridate\_1.7.10 assertthat\_0.2.1 digest\_0.6.27   
#> [5] utf8\_1.2.2 R6\_2.5.0 cellranger\_1.1.0 backports\_1.2.1   
#> [9] reprex\_2.0.1 RcppZiggurat\_0.1.6 evaluate\_0.14 httr\_1.4.2   
#> [13] pillar\_1.6.2 gdtools\_0.2.3 rlang\_0.4.11 uuid\_0.1-4   
#> [17] readxl\_1.3.1 rstudioapi\_0.13 data.table\_1.14.0 rmarkdown\_2.10   
#> [21] munsell\_0.5.0 broom\_0.7.9 compiler\_4.1.0 modelr\_0.1.8   
#> [25] xfun\_0.25 pkgconfig\_2.0.3 systemfonts\_1.0.2 base64enc\_0.1-3   
#> [29] htmltools\_0.5.1.1 tidyselect\_1.1.1 fansi\_0.5.0 crayon\_1.4.1   
#> [33] tzdb\_0.1.2 dbplyr\_2.1.1 withr\_2.4.2 grid\_4.1.0   
#> [37] jsonlite\_1.7.2 gtable\_0.3.0 lifecycle\_1.0.0 DBI\_1.1.1   
#> [41] magrittr\_2.0.1 scales\_1.1.1 Rfast\_2.0.3 zip\_2.2.0   
#> [45] cli\_3.0.1 stringi\_1.7.3 fs\_1.5.0 xml2\_1.3.2   
#> [49] ellipsis\_0.3.2 generics\_0.1.0 vctrs\_0.3.8 tools\_4.1.0   
#> [53] glue\_1.4.2 officer\_0.4.0 hms\_1.1.0 parallel\_4.1.0   
#> [57] yaml\_2.2.1 colorspace\_2.0-2 rvest\_1.0.1 knitr\_1.33   
#> [61] haven\_2.4.3