mvGST: TOOLS FOR MULTIVARIATE AND DIRECTIONAL GENE SET TESTING

9 April 2014
M.S. Statistics Defense

Dennis Mecham
Motivating Example

• An experiment was performed to understand how obatoclax mesylate treats leukemia.
• A 3x2 full factorial design was used with 2 replicates.
• The 3 levels of obatoclax mesylate were:
  – High dose (HIGH)
  – Low dose (LOW)
  – Control
• There were 2 blood cell lines treated:
  – RS4:11
  – SEM-K2
• Our objective is to determine how thousands of biological processes’ activity levels are affected by obatoclax dose level
  – after accounting for cell line differences [non-stratified]
  – in each cell line separately [stratified]
Idea of Gene Sets

• Gene Ontology groups genes into sets that perform the same biological processes.
• Only genes that contribute to the process are included in each group.
• Genes that inhibit a process are not included in the gene set for that process.
• If genes in set are active, then the process proceeds.
• If even a single gene is not active, then a process may be “disturbed”.
Statistical Model Used

• Expression level of genes in set is used as a proxy for activity level of corresponding biological process

• Per-gene model:

\[ Y_{jkl} = \mu + D_j + L_k + DL_{jk} + \varepsilon_{jkl} \]

- \( Y_{jkl} \) is the log (base 2) of the expression level in replicate l of dosage level (D) j, for leukemia cell line (L) k
- \( D_j \) is the dosage level: high, low, or control.
- \( L_k \) is the leukemia cell line: RS4:11 or SEM-K2
- \( \varepsilon_{jkl} \) follows a normal distribution with gene-specific variance \( \sigma^2 \)

• Model fit using limma package in R
• Suppose a gene is significantly more active for HIGH vs. CTL in the RS4:11 line, less active for LOW vs. CTL in the RS4:11 line, and not significantly different for either dose vs. CTL in the SEM-K2 line.
• The profile could be summarized as:
  1, -1, 0, 0
• or the profile could be stratified and summarized as:
  1, -1 for RS4:11
  0, 0 for SEM-K2
Multivariate and Directional Differential Expression

- The term “multivariate” is used because of simultaneous interest in multiple contrasts
- The term “directional” is used because of interest in one-sided alternatives
- Four contrasts were tested for each gene (HIGH/LOW vs CTL at each cell line) with:
  - $H_0$: HIGH/LOW = CTL
  - $H_a$: HIGH/LOW > CTL
- These 4 tests can be summarized in 1 profile with 4 dimensions, or 2 profiles with 2 dimensions.
# Summarizing Profiles: Non-Stratified

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>#</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

...
## Summarizing Profiles: Stratified

<table>
<thead>
<tr>
<th>HIGH</th>
<th>LOW</th>
<th>RS4:11</th>
<th>SEM-K2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>0</td>
<td>-1</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>#</td>
<td>#</td>
</tr>
</tbody>
</table>
P-value Combination

• For each contrast, p-values of individual genes are combined to obtain a single p-value for each gene set.
• Fisher’s Method
  – Alternative Hypothesis: at least one gene is significant
  – Not symmetric
• Stouffer’s Method
  – Alternative Hypothesis: consensus of significance
  – Symmetric
• Whitlock (2005) showed that Stouffer’s method is more powerful for the more meaningful alternative of consensus (see slide 3; since even 1 inactive gene may “disturb” a biological process, consensus is most appropriate)
• Symmetry preserves interest in directionality
  – i.e. if the result of combining $p_1$ and $p_2$ is $p_{12}$, then the result of combining $(1- p_1)$ and $(1- p_2)$ should be $(1- p_{12})$
Multiple Hypothesis Testing

• Multiple hypothesis tests are performed on each gene set
  – Each test is another chance to make an error
  – Motivating example has 12,260 gene sets (x 4 contrasts)
  – Some adjustment must be made

• What to control?
  – Methods that control family wise error rate are often too conservative (but more confirmatory)
  – Controlling False Discovery Rate (FDR) is more powerful (but more exploratory)

• Benjamini – Yekutieli adjustment is used to control FDR
  – P-values of gene sets are dependent because genes are in multiple sets (all genes in child sets are also in the parent sets)
  – Dependency structure is unknown
  – Benjamini – Yekutieli adjustment allows for any dependency structure
profileTable

• Takes matrix of p-values, vector of gene names, and vector of contrasts and produces desired profile summaries (slides 7 and 8)

• Necessary arguments:

  – gene.names: a vector of gene names where the $i^{th}$ gene name corresponds to the $i^{th}$ row of pvals

  – contrasts: a character vector of contrasts tested; must be in one of the forms: Var1 or Var1.Var2 (stratified). The $j^{th}$ contrast corresponds to the $j^{th}$ column of pvals

  – pvals: a matrix of p-values (row = gene, column = contrast)
### Non-Stratified Output

<table>
<thead>
<tr>
<th>RS4Low</th>
<th>RS4High</th>
<th>SEMK2Low</th>
<th>SEMK2High</th>
<th>BP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10714</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>340</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>191</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>159</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>149</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>114</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>106</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>92</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>68</td>
</tr>
</tbody>
</table>
# Stratified Output

<table>
<thead>
<tr>
<th>Low</th>
<th>High</th>
<th>RS4</th>
<th>SEMK2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>11398</td>
<td>10970</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>18</td>
<td>647</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>251</td>
<td>370</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>432</td>
<td>88</td>
</tr>
<tr>
<td>0</td>
<td>-1</td>
<td>53</td>
<td>130</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>75</td>
<td>48</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>33</td>
<td>7</td>
</tr>
</tbody>
</table>
Gene Name Translation

• Depending on naming system used, gene names may need to be translated to Entrez

• Differences in naming systems cause one-to-many and many-to-one problems

• Options in *profileTable*:
  – method 1: ignore problem
  – method 2: Stouffer combine p-values in many-to-one
  – method 3: arbitrarily ignore all but one of one-to-many
  – method 4: use method 2, then method 3
pickOut

• Returns a vector containing the ID’s of gene sets that fit a specified profile

• Necessary Arguments:
  – \textit{mvgst}: a mvGST object as returned by \textit{profileTable}
  – \textit{row}: the row number of the desired profile in the table returned by \textit{profileTable}
Non-Stratified Output

> gene.sets <- pickOut(example2a, 15, 1)

[1] "GO:0001510" "GO:0006400" "GO:0006417" "GO:0006431" "GO:0006437"
[6] "GO:0006564" "GO:0006839" "GO:0007005" "GO:0008614" "GO:0008615"
[16] "GO:0034975" "GO:0042819" "GO:0055129" "GO:0071301" "GO:0071494"

Stratified Output

> gene.sets <- pickOut(example2b, 7, 2)

[1] "GO:0006813" "GO:0043266" "GO:0043268" "GO:0048745" "GO:0051481"
[6] "GO:0071526" "GO:0090075"
go2Profile

• Given a specific gene set(s), find its profile
• Returns a table, or list of tables, that are similar to the table from `profileTable` except that only one gene set is included
• Necessary Arguments:
  – `names`: a character vector with the names, or ID's, of the gene sets of interest
  – `object`: a mvGST object as returned by `profileTable`
Non-stratified Output

```r
> profiles <- go2Profile(c("GO:0001510", "GO:0006171"), example2a)

`GO:0001510`
RS4Low   RS4High   SEMK2Low   SEMK2High   BP
-1       -1        0          0          1

`GO:0006171`
RS4Low   RS4High   SEMK2Low   SEMK2High   BP
1        1         0          0          1
```

Stratified Output

```r
> profiles <- go2Profile(c("GO:0006813"), example2b)

`GO:0006813`
Low   High   RS4   SEMK2
0     0       1     0
-1    0       0     1
```
graphCell

- Displays a GO graph of the gene sets that fit a specified profile and their parent sets

- Necessary Arguments:
  - `object`: a mvGST object as returned by `profileTable`
  - `row`: the row number of the desired profile in the table returned by `profileTable`
Non-Stratified Output

> graphCell(example2a, 15, 1, interact = FALSE, print.legend = FALSE)
Stratified Output

>`graphCell(example2b, 7, 2, interact = FALSE, print.legend = FALSE)`
library(mvGST)
data2 <- read.csv("C:/examples/second.csv")
gene.names2 <- as.character(data2[, 1])
contrasts2a <- c("RS4Low", "RS4High", "SEMK2Low", "SEMK2High")
contrasts2b <- c("Low.RS4", "High.RS4", "Low.SEMK2", "High.SEMK2")
pvals2 <- as.matrix(data2[, 2:5])
chip <- "hgu133plus2"

eample2a <- profileTable(gene.names2, contrasts2a, pvals2, gene.ID ="affy", organism = "hsapiens", ontology = "BP", affy.chip = "hgu133plus2")
gene.sets <- pickOut(example2a, 15, 1)
Demo continued

graphCell(example2a, 15, 1, interact = FALSE,
print.legend = FALSE)
profiles <- go2Profile(c("GO:0001510",
"GO:0006171"), example2a)

e.example2b <- profileTableTable(gene.names2,
contrasts2b, pvals2, gene.ID ="affy",
organism = "hsapiens", ontology = "BP",
affy.chip = "hgu133plus2")
gene.sets <- pickOut(example2b, 7, 2)
graphCell(example2b, 7, 2, interact = FALSE,
print.legend = FALSE)
profiles <- go2Profile(c("GO:0006813"),
example2b)
Hartung versus Stouffer

• When many gene names are translated to one gene name, the expression levels of the many genes may be biologically dependent.
  – If the “many” really are separate genes, they are expressed independently
  – If the “many” really are parts of the same gene, they may not be expressed independently

• Stouffer’s p-value combination method may overstate significance when p-values from biologically dependent “genes” are combined

• Hartung’s p-value combination method accounts for dependent p-values
  – Hartung’s assumes positive, constant covariance (between p-values), but is robust to non-constant covariance
• Covariance estimates come from the formula used in Hartung’s method
• Observed covariance estimates do not differ greatly from covariance estimates of simulated independent p-values
• No clear evidence that Hartung’s method is necessary
Recent Additions

• Add Short Focus Level adjustment as an option instead of Benjamini-Yekutieli adjustment
• Short Focus Level adjustment controls family-wise error rate while accounting for GO graph structure (Saunders 2014 dissertation)
• A few other minor modifications to improve usability (version 1.1)
Utility of mvGST Package

- Platform independent (Affymetrix, Next Gen Seq, ...)
- Design independent
- Controls for other factors
- Multivariate Summary
- Directional interest preserved
Acknowledgements

• Thank you to my committee for your time
• Thank you to Dr. Stevens for all of your help
• Thank you to Rachel for all you’ve done to let me work on this
• Thank you to the Agricultural Experiment Station for their work on the first motivating example (in the report)
• Thank you to the department for helping me get my degree