Statistical analysis of Megavariate Data

Javier Cabrera,
Rutgers University

Collaborators: Dhammika Amaratunga, Lan Yi, YS Lee,
Zhenya Cherkas, Volha Tryputsen,
Outline

1. Introduction: Megavariate Data from Genomics experiments
2. Enriched methods for supervised and unsupervised classification
3. Individual variable analysis
4. Group analysis
### (1) Megavariate Data

*Many variables with few observations. $X$ is a $G \times n$ matrix, $G \gg n$*

Examples: Microarray data, Chip-seq, SNP and other Sequence data, Proteomics data, Mass spec, Imaging, Google data, Census Data, ...

50000 Genes x 6 samples + Responses

You can also think of the data as:
- 50000 scatter-plots or boxplots vs same Y
- 50000 regression (linear or logistic) vs the same Y
Experiments that produce Megavariate Data in Genomics:

(i) Microarray data
(ii) HT-qPCR
(iii) Chip-seq, RNA-seq
(iv) SNP’s, Genome Wide Association Studies (GWAS)
(v) Protein arrays
(vi) LC-MS: Liquid chromatography Mass spectrometry
    Protein or metabolites or other small molecules
(vii) Copy number variation
(viii) Molecular imaging: CAT, PET, etc

Other sources: Clinical Data? Internet Data?
Some Properties

(i) The variables may all reflect one type of measurement
(ii) Mild to strong correlation structure among subgroups of variables.
(iii) Preprocessing may induce correlations among samples.
(iv) Data are not likely to be normally distributed. Sample outliers are difficult to detect.
(v) There is often subsidiary information available.
(vi) High throughput experiments that produce Megavariate data are exploratory in nature.

The central issue on modeling these data is how to avoid overfitting!
Signal Structure in Microarray Data

Microarray experiments:

\[
\text{Signal} = \text{Good signal} + \text{Spurious Signal}
\]

\[
\text{Good Signal} = \begin{array}{l}
\text{Specific} \\
(\text{Genes associated with Slc17A5})
\end{array} + \begin{array}{l}
\text{Non Specific} \\
(\text{Unknown Unbalance of the Samples})
\end{array}
\]

As dimension increases \(\Rightarrow\) Spurious signal may ultimately dominate all signals.

Spurious Signal \(\Rightarrow\) Likely to be made up of small incremental random signals

Solutions: Enriched methods (Enriched PCA …)
Thresholding methods (GLMNET …)
**Experiment:** Compare the gene expression profiles of 6 KO mice vs 6 WT mice using a microarray with 45101 genes.

**Questions:** Are there differences between gene expressions for WT and KO at 6h? 18 days? Are the differences happening in the same cellular processes?
Data: Expression measures for \( G \) genes in \( N \) samples:

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>83</td>
<td>94</td>
<td>82</td>
<td>111</td>
<td>130</td>
<td>122</td>
</tr>
<tr>
<td>G2</td>
<td>16</td>
<td>14</td>
<td>7</td>
<td>2</td>
<td>11</td>
<td>33</td>
</tr>
<tr>
<td>G3</td>
<td>490</td>
<td>879</td>
<td>193</td>
<td>604</td>
<td>1031</td>
<td>962</td>
</tr>
<tr>
<td>G4</td>
<td>46458</td>
<td>49268</td>
<td>74059</td>
<td>44849</td>
<td>42235</td>
<td>44611</td>
</tr>
<tr>
<td>G5</td>
<td>32</td>
<td>70</td>
<td>185</td>
<td>20</td>
<td>25</td>
<td>19</td>
</tr>
<tr>
<td>G6</td>
<td>1067</td>
<td>891</td>
<td>546</td>
<td>906</td>
<td>1038</td>
<td>1098</td>
</tr>
<tr>
<td>G7</td>
<td>118</td>
<td>111</td>
<td>95</td>
<td>896</td>
<td>536</td>
<td>695</td>
</tr>
<tr>
<td>G8</td>
<td>10</td>
<td>30</td>
<td>25</td>
<td>24</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>G9</td>
<td>166</td>
<td>132</td>
<td>162</td>
<td>27</td>
<td>109</td>
<td>213</td>
</tr>
<tr>
<td>G10</td>
<td>136</td>
<td>139</td>
<td>44</td>
<td>62</td>
<td>23</td>
<td>135</td>
</tr>
</tbody>
</table>

Preprocess: normalize and log transform

45101 rows (genes) x 12 columns (samples)
Enriched methods for classification

Classification in microarray experiments:

Curse of dimensionality (difficult) => Gene selection => classification

Gene selection is usually done with supervised hard thresholding.

Amaratunga Cabrera Kovtum (2006), Amaratunga, Cabrera, Li (2008) proposed an enriched weights scheme that is equivalent to soft thresholding in the context of unsupervised and supervised classification.
Enriched Methods
Assign a weight to each gene based on FDR

**FDR based Weights:**

Perform Statistical analysis of individual variables: Obtain p-values by testing statistically significant difference across the groups (t, F...)

Assume a null distribution for p-values, for example uniform p-value distribution $p_i \sim \text{Uniform}[0,1]$, then $p(i) \sim \text{Beta}[i, G-i]$ (order statistics)

FDR corrected p-values: $q(i) = p(i) / p_{(i), \alpha}$ and make $q(i)$ monotone on (i).
FDR corrected weights: $W_i = 1/q_i$ or $W_i = -\log(q_i)$

Suppose that $\alpha$ is a tuning constant and that $p_{(i), \alpha}$ is the $\alpha$-percentile of distribution of the $i^{th}$ order statistic.
We fix $\alpha = 0.05$ (Fisher) and calculate the weights.

**Simple version:** $q(i) = p(i) / (i/G)$ and $W_i = 1/q_i$ or $W_i = -\log(q_i)$
Enriched methods: Enriched Principal Components Analysis

PCA:  \( Z (n \times G) \)  
\[ Z = U D V' \Rightarrow S = 1/(n-1) VD U' \]  
\[ Z \text{ centered then } S = 1/(n-1) Z'Z \]
\[ \frac{1}{n-1}(n \times n) \]

\[ X (G \times n) \]  
\[ X = U D V' \Rightarrow S = 1/(n-1) UDV' \]  
\[ \frac{1}{G-1}(G \times G) \]

Weighted PCA:  
\[ W = \text{Diag}(W_i) \]
\[ X = W X = U*D^*V^* \]
\[ \frac{1}{n-1}(n \times n) \]

Covariance or Correlation  
\[ S^* = U^* D^*^2 U^*/(n-1) \]
\[ \frac{(n \times n)(n \times G)}{(G \times G)(G \times n) n \times n} \]
Enriched PCA Analysis
Biplot of day 0 gene expression data from Slc17A5 experiment
Enriched PCA Analysis
Biplot of day 0 gene expression data from Slc17A5 experiment
## Enriched PCA Analysis:
day 0 gene expression data from Slc17A5 experiment

### Important Genes with Annotation

<table>
<thead>
<tr>
<th>Vertical direction</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1429116_a</td>
<td>SLC gene that was knocked out in treatment group.</td>
</tr>
<tr>
<td>1435559_at</td>
<td>Myo6, Growth.</td>
</tr>
<tr>
<td>1437522_x_at</td>
<td>Gh growth hormone.</td>
</tr>
<tr>
<td>1454905_at</td>
<td>Inhibitor of Bruton agammaglobulinemia tyrosine kinase.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Horizontal direction</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1417210_at</td>
<td>Eukaryotic translation initiation factor 2, subunit 3, structural gene <strong>Y-linked</strong> Eif-2gy, Spy, Tfy</td>
</tr>
<tr>
<td>1426438_at</td>
<td>Box polypeptide 3, <strong>Y-linked</strong> 8030469F12Rik, D1Pas1-rs1, Dby</td>
</tr>
<tr>
<td>1427262_at</td>
<td>Inactive <strong>X specific</strong> transcripts A430022B11, AI314753. Exper Embryonic brain development, Expression profiling by array, count</td>
</tr>
<tr>
<td>1436936_s_at</td>
<td>Inactive <strong>X specific</strong> transcripts A430022B11, AI314753. Experiment Embryonic brain development, Expression profiling by array, count</td>
</tr>
</tbody>
</table>
Enriched PCA Analysis:
Gene expression data from Slc17A5 experiment
Day 0 predicts day 18
**Enriched PCA Analysis:**
Gene expression data from Slc17A5 experiment
Day 18 predicts day 0

1417210_at 1418243_at 1419483_at 1423310_at 1424754_at 1424903_at 1426438_at 1426439_at 1427076_at 1427262_at 1427263_at 1429116_at 1434411_at 1435477_s_at 1435559_at 1436650_at 1436936_s_at 1448148_at 1451941_a_at 1452077_at 1454905_at
## Enriched PCA Analysis:
Gene expression data from Slc17A5 experiment
Important Genes with Annotation

<table>
<thead>
<tr>
<th>Method</th>
<th>Gene Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrich PCA</td>
<td>Myo6</td>
</tr>
<tr>
<td>1437522_x_at</td>
<td>Gh growth hormone</td>
</tr>
<tr>
<td>1454905_at</td>
<td>Inhibitor of Bruton agammaglobulinemia tyrosine kinase</td>
</tr>
<tr>
<td>Horizontal direction</td>
<td>Anotationion</td>
</tr>
<tr>
<td>1417210_at</td>
<td>Eukaryotic translation initiation factor 2, subunit 3, structural gene Y-linked Eif-2gy, Spy, Tfy</td>
</tr>
<tr>
<td>1426438_at</td>
<td>Box polypeptide 3, Y-linked 8030469F12Rik, D1Pas1-rs1, Dby</td>
</tr>
<tr>
<td>1427262_at</td>
<td>Inactive X specific transcripts A430022B11, AI314753. Experiment embryonic brain development, Expression profiling by array, count</td>
</tr>
<tr>
<td>1436936_s_at</td>
<td>Inactive X specific transcripts A430022B11, AI314753. Experiment embryonic brain development, Expression profiling by array, count</td>
</tr>
</tbody>
</table>
Support Vector Machine:
The shaded area represents the separation region. The arrows indicate the location of the support vectors.
The Lasso

\[ \hat{\beta}^{Lasso} = \arg \min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 \]

subject to:
\[ \sum_{j=1}^{p} |\beta_j| \leq s \]

Quadratic programming algorithm needed to solve for the parameter estimates. Choose \( s \) via cross-validation.

\[ \tilde{\beta} = \arg \min_{\beta} \left( \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right) \quad q=0: \text{var. sel.} \]
\[ q=1: \text{lasso} \]
\[ q=2: \text{ridge} \]

Learn \( q \)?
### Comparing Enriched with LASSO and SVM methods for classification

<table>
<thead>
<tr>
<th>Method</th>
<th>Day 0 predicts Day 18</th>
<th>Day 18 predicts day 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched PCA</td>
<td>0%(0)</td>
<td>0%(0)</td>
</tr>
<tr>
<td>GLMnet</td>
<td>0%(0)</td>
<td>0%(0)</td>
</tr>
<tr>
<td>SVM(e1071)</td>
<td>50%(6)</td>
<td>50%(6)</td>
</tr>
<tr>
<td>Enriched (reciprocal) SVM(e1071)</td>
<td>0%(0)</td>
<td>0%(0)</td>
</tr>
<tr>
<td>Enriched(-log) SVM(e1071)</td>
<td>50%(6)</td>
<td>0%(0)</td>
</tr>
<tr>
<td>Proximal SVM</td>
<td>50%(6)</td>
<td>50%(6)</td>
</tr>
<tr>
<td>Penalized SVM (LASSO)</td>
<td>0%(0)</td>
<td>50%(6)</td>
</tr>
<tr>
<td>Penalized SVM (SCAD)</td>
<td>0%(0)</td>
<td>50%(6)</td>
</tr>
</tbody>
</table>
(2) Individual variable analysis

♦ Statistical analysis of individual variables: Seek variables that exhibit a statistically significant difference across the groups (via e.g., t, permutation test, Ct, SAM, limma, Bayes/EmpiricalBayes procedures).

♦ Adjust for multiplicity:

\[ \text{pFDR} = \text{Average} \left( \frac{\# \text{FalsePositives}}{\# \text{Positives}} \right). \]

Algorithm: Decision rule says “reject if \( T>c \)” \( \rightarrow h_0 \)

\( \rightarrow \) permute \( \rightarrow h_1 \) \( \rightarrow \) permute \( \rightarrow h_2 \) \( \rightarrow \ldots \) \( \rightarrow h_m \)

\( \rightarrow \) average=\( h^* \) \( \rightarrow \) pFDR=\( h^*/h_0 \) \( \rightarrow \) refine
The effect of small sample size

Often the sample size per group is small.

- unreliable variances (inferences)
- dependence between the test statistics ($t_g$) and the standard error estimates ($s_g$)

Borrow strength across genes

A model for borrowing strength

- Let $X_{gij}$ denote the preprocessed intensity measurement for gene $g$ in array $i$ of group $j$.

- Model: $X_{gij} = \mu_{gj} + \sigma_g \varepsilon_{gij}$

- Effect of interest: $\Delta_g = \mu_{g2} - \mu_{g1}$

- Error model: $\varepsilon_g \sim F(\text{loc}=0, \text{scale}=R), \sigma_g \sim F_\sigma$
Possible approaches

**Parametric:** Assume functional forms for $F_\varepsilon$ and $F_\sigma$ and apply either a Bayes or Empirical Bayes procedure $\rightarrow$ regularized test statistics.

\[
T_g = (\bar{X}_{g_1} - \bar{X}_{g_2}) / s_g
\]

\[
T_g (c) = (\bar{X}_{g_1} - \bar{X}_{g_2}) / (s_g + c)
\]

\[
T_g (d) = (\bar{X}_{g_1} - \bar{X}_{g_2}) / \sqrt{(d_g s_g^2 + d_0 s_0^2)}
\]

or

\[
T_g (c) = (\bar{X}_{g_1} - \bar{X}_{g_2}) / s_{\text{shrunken}}
\]
Compare C1–C3 vs T1–T3 using t tests

Test: t tests with $\alpha = 0.05$ (after preprocessing)

Result: Assume $\sigma_g^2 = ct$, $\varepsilon_{ij}$ Normally distributed

$RR: |T_g| > (z_{\alpha/2} \sigma)/s_g$
(3) Group Analysis

♦ Analysis of gene sets: Seek pre-defined gene sets that separate the groups.

Gene set scoring: \( \{p\} \rightarrow \text{MLP} = \text{mean} (-\log p) = 2.34^* \)

Over-representation analysis: 7/11 vs 167/12567 \( \rightarrow \text{odds ratio}=1.56^* \)

<table>
<thead>
<tr>
<th>Gene</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>11303</td>
<td>0.000651</td>
</tr>
<tr>
<td>14127</td>
<td>0.001703</td>
</tr>
<tr>
<td>14129</td>
<td>0.203787</td>
</tr>
<tr>
<td>14130</td>
<td>2.00E-05</td>
</tr>
<tr>
<td>14131</td>
<td>0.000292</td>
</tr>
<tr>
<td>16017</td>
<td>0.043791</td>
</tr>
<tr>
<td>17304</td>
<td>0.167931</td>
</tr>
<tr>
<td>19261</td>
<td>0.000415</td>
</tr>
<tr>
<td>56644</td>
<td>0.005529</td>
</tr>
<tr>
<td>70676</td>
<td>0.004842</td>
</tr>
<tr>
<td>380793</td>
<td>0.103618</td>
</tr>
</tbody>
</table>

Example:

*Phagocytosis engulfment* in Slc17A5-KO vs WT experiment

GO #3 is a “significant” GO term
Assessing MLP's significance

- Permute the $p$-values across the genes (thereby preserving correlations among gene sets).
- Recalculate MLP $\rightarrow$ MLP*.
- Repeat NS times for all gene sets.
- Estimate the $\alpha^{th}$ quantile of MLP* vs $m$ (gene set size) and interpolate the $p$-values.

- New improvements. Think of the problem as sampling from a finite known population with sample size $m$. Approximate the distribution of MLP with an Edgworth expansion.
♦ New improvements. Think of the problem as sampling from a finite known population with sample size m. Approximate the distribution of MLP with an Edgeworth or Saddle Point expansion.
Issues in gene set analysis

- Which test statistic to use?

MLP is a version of the Anderson-Darling statistic for comparing distributions. AD is known to be more powerful for picking up tail differences than Kolmogorov-Smirnov. Our simulations show the MLP approach has higher power than KS-based or Fisher-based row-permutation based approaches.

- To determine the significance of gene set $S$, what should one permute? rows? columns? both?

Row permutations: Are the genes in $S$ generally more significant than the genes in a random gene set that is the same size as $S$? [Comparing across gene sets]

Column permutations: Is $S$ significantly differentially expressed across classes? [Comparing across classes]
Importance of gene set analysis

- Interpretability: Better interpretation of microarray results is possible at the biological process level than at the individual gene level.

<table>
<thead>
<tr>
<th>GO</th>
<th>Description</th>
<th>Geneset.Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>16126</td>
<td>sterol biosynthesis</td>
<td>24</td>
</tr>
<tr>
<td>16125</td>
<td>sterol metabolism</td>
<td>54</td>
</tr>
<tr>
<td>8366</td>
<td>nerve ensheathment</td>
<td>15</td>
</tr>
<tr>
<td>6695</td>
<td>cholesterol biosynthesis</td>
<td>20</td>
</tr>
<tr>
<td>42551</td>
<td>neuron maturation</td>
<td>23</td>
</tr>
<tr>
<td>8203</td>
<td>cholesterol metabolism</td>
<td>50</td>
</tr>
<tr>
<td>48469</td>
<td>cell maturation</td>
<td>51</td>
</tr>
<tr>
<td>7272</td>
<td>ionic insulation of neurons by glial cells</td>
<td>12</td>
</tr>
<tr>
<td>42552</td>
<td>myelination</td>
<td>12</td>
</tr>
<tr>
<td>42553</td>
<td>cellular nerve ensheathment</td>
<td>12</td>
</tr>
<tr>
<td>6694</td>
<td>steroid biosynthesis</td>
<td>55</td>
</tr>
<tr>
<td>1508</td>
<td>regulation of action potential</td>
<td>14</td>
</tr>
<tr>
<td>6911</td>
<td>phagocytosis, engulfment</td>
<td>11</td>
</tr>
<tr>
<td>50764</td>
<td>regulation of phagocytosis</td>
<td>13</td>
</tr>
<tr>
<td>50766</td>
<td>positive regulation of phagocytosis</td>
<td>13</td>
</tr>
</tbody>
</table>
References:


Website:
www.rci.rutgers.edu/~cabrera/DNAMR