Supervised Learning: Methods for Analyzing Gene Expression Data

Kellie J. Archer, Ph.D.
Department of Biostatistics
Virginia Commonwealth University

kjarcher@vcu.edu
www.people.vcu.edu/~kjarcher
Taxonomy of Microarray Data Analysis Methods

- **Supervised Learning**: Class or group labels are known a priori and the goal of the statistical analysis pertains to
  - identifying differentially expressed genes (AKA feature selection) or
  - identifying combinations of genes that are predictive of class or group membership.
- **Unsupervised Learning**: The statistical analysis seeks to find structure in the data without knowledge of class labels.
Supervised Learning

• Class comparison/ Feature selection
  – T-test / Wilcoxon rank sum test (2 groups)
  – F-test / Kruskal Wallis test (>2 groups)
  – Adjustments for multiple comparisons

• Class Prediction
  – k-nearest neighbors
  – Compound Covariate Predictors
  – Classification trees
  – Support vector machines
  – etc.
Supervised Learning

- Class comparison/feature selection is an important step that should be conducted prior to predictive modeling in order to reduce the number of model parameters.
  - There are only a few predictive modeling methods (e.g., bootstrap aggregating, random forests) that do not require feature selection as a pre-filter to model fitting.
Outline

• Class comparisons
  – T-test
  – F-test
  – Multiple hypothesis testing issues
  – Permutation based p-values

• Class prediction
  – k-nearest neighbors
  – Compound Covariate Predictors

• Software implementation
Illustrative Dataset

• Suppose we assessed the expression of a single gene, peroxiredoxin 1 (PRDX1), in 16 samples using the following experimental design to determine whether day of test, module of the testing instrument, or whether the sample was fresh or frozen, had any impact on expression.
<table>
<thead>
<tr>
<th>Day</th>
<th>Fresh / Frozen</th>
<th>Module</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fresh</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Fresh</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>Fresh</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>Fresh</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>Frozen</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Frozen</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>Frozen</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>Frozen</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Frozen</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Frozen</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Frozen</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Frozen</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Frozen</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Frozen</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Frozen</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Frozen</td>
<td>4</td>
</tr>
</tbody>
</table>
Hypothesis Testing

• The hypothesis that two means $\mu_1$ and $\mu_2$ are equal is called a null hypothesis, commonly abbreviated $H_0$.
• This is typically written as $H_0: \mu_1 = \mu_2$
• Its antithesis is the alternative hypothesis, $H_A: \mu_1 \neq \mu_2$
Hypothesis Testing

• A statistical test of hypothesis is a procedure for assessing the compatibility of the data with the null hypothesis.
  – The data are considered compatible with $H_0$ if any discrepancy from $H_0$ could readily be due to chance (i.e., sampling error).
  – Data judged to be incompatible with $H_0$ are taken as evidence in favor of $H_A$. 
Hypothesis Testing

• If the sample means calculated are identical, we would suspect the null hypothesis is true.

• Even if the null hypothesis is true, we do not really expect the sample means to be identically equal because of sampling variability.

• We would feel comfortable concluding $H_0$ is true if the chance difference in the sample means should not exceed a couple of standard errors.
Hypothesis Testing

- The **mean** $\mu(X)$ of a random variable $X$ is a measure of central location of the density of $X$.
- The **variance** of a random variable is a measure of spread or dispersion of the density of $X$.
- $\text{Var}(X) = E[(X-\mu)^2] = \sum (X - \mu)^2/(n-1) = \sigma^2$
- Standard deviation $= \sqrt{\text{Var}(X)} = \sigma$
T-test

• In testing $H_0: \mu_1 = \mu_2$ against $H_A: \mu_1 \neq \mu_2$ note that we could have restated the null hypothesis as $H_0: \mu_1 - \mu_2 = 0$ and $H_A: \mu_1 - \mu_2 \neq 0$

• To carry out the t-test, the first step is to compute the test statistic and then compare the result to a t-distribution with the appropriate degrees of freedom (df)

$$t_g = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{SE(\bar{y}_1 - \bar{y}_2)} = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^4} + \frac{\sigma_2^4}{\sigma_2^4}\right)^2}{\frac{n_1 - 1}{\sigma_1^4} + \frac{n_2 - 1}{\sigma_2^4}}$$
T-test

• Data must be independent random samples from their respective populations
• Sample size should either be large or, in the case of small sample sizes, the population distributions must be approximately normally distributed.
• When assumptions are not met, non-parametric alternatives are available (Wilcoxon Rank Sum/Mann-Whitney Test)
## T-test: PRDX1

<table>
<thead>
<tr>
<th>Genechip</th>
<th>Day 1</th>
<th>Day 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2013.7</td>
<td>1974.6</td>
</tr>
<tr>
<td>2</td>
<td>2141.9</td>
<td>2027.6</td>
</tr>
<tr>
<td>3</td>
<td>2040.2</td>
<td>1914.8</td>
</tr>
<tr>
<td>4</td>
<td>1973.3</td>
<td>1955.8</td>
</tr>
<tr>
<td>5</td>
<td>2162.2</td>
<td>1963.0</td>
</tr>
<tr>
<td>6</td>
<td>1994.8</td>
<td>2025.5</td>
</tr>
<tr>
<td>7</td>
<td>1913.3</td>
<td>1865.1</td>
</tr>
<tr>
<td>8</td>
<td>2068.7</td>
<td>1922.4</td>
</tr>
<tr>
<td>$\bar{y}$</td>
<td>2038.5</td>
<td>1956.1</td>
</tr>
<tr>
<td>$\hat{\sigma}^2$</td>
<td>7051.284</td>
<td>3062.991</td>
</tr>
<tr>
<td>$n$</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>
T-test: \( PRDX1 \)

\[
t_g = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{SE(\bar{y}_1 - \bar{y}_2)}
\]

\[
t_g = \frac{(2038.5 - 1965.1) - 0}{\sqrt{\frac{7051.3}{8} + \frac{3062.99}{8}}} = 2.317
\]

\[
\text{df} = \frac{(7051.3 + 3062.99)^2}{\frac{7051.3^2}{8-1} + \frac{3062.99^2}{8-1}} = 12.116
\]

\[
P = 0.039
\]
Hypothesis Testing

• Type I error: The probability of rejecting a null hypothesis when it is true. (e.g., a gene is declared to be differentially expressed when it is not.)

• Type II error: The probability of accepting a null hypothesis when it is false. (e.g., a gene is declared to not be differentially expressed when it actually is.)
P-value

• The p-value for a hypothesis test is the probability, computed under the condition that the null hypothesis is true, of the test statistic being at least as extreme as the value of the test statistic that was actually obtained.
  – A large p-value (close to 1) indicates a value of t near the center of the t-distribution.
  – A small p-value indicates a value of t in the far tails of the t-distribution.
P-value

• The threshold value, on the p-value scale, is called the significance level of the test, and is usually denoted by the Greek letter $\alpha$.

• If the p-value for a given $H_0$ is less than or equal to $\alpha$, the data are judged incompatible with $H_0$.
  – In this case, we say that $H_0$ is rejected and that the data provide evidence in favor of $H_A$.
  – For $PRDX1$, the p-value was 0.039
F-test

• In testing $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ against $H_A$: The $\mu_j$’s are not all equal, we use an F-test.

• If we were to test whether the means for modules 1, 2, 3, and 4 were equal against the alternative that the mean for at least one module was different, the F-test would have been used.
F-test

\[
F_g = \frac{\text{MS(group)}}{\text{MS(error)}} = \frac{\frac{\text{SS(group)}}{\text{df(group)}}}{\frac{\text{SS(error)}}{\text{df(error)}}} = \frac{\sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y}_.)^2}{\frac{k - 1}{\sum_{i=1}^{k} \sum_{j=1}^{n_{ij}} (y_{ij} - \bar{y}_.)^2 - \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y}_.)^2}} \frac{1}{\sum_{i=1}^{k} (n_i - 1)}
\]

\[
\bar{y}_. = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_{ij}} y_{ij}}{\sum_{i=1}^{k} n_i}
\]

\[
\sim F\left(k - 1, \sum_{i=1}^{k} (n_i - 1)\right)
\]
### F-test: PRDX1

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Module 1</th>
<th>Module 2</th>
<th>Module 3</th>
<th>Module 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2013.7</td>
<td>2141.9</td>
<td>2040.2</td>
<td>1973.3</td>
</tr>
<tr>
<td>2</td>
<td>2162.2</td>
<td>1994.8</td>
<td>1913.3</td>
<td>2068.7</td>
</tr>
<tr>
<td>3</td>
<td>1974.6</td>
<td>2027.6</td>
<td>1914.8</td>
<td>1955.8</td>
</tr>
<tr>
<td>4</td>
<td>1963.0</td>
<td>2025.5</td>
<td>1865.1</td>
<td>1922.4</td>
</tr>
<tr>
<td>$\bar{y}_i$</td>
<td>2028.4</td>
<td>2047.5</td>
<td>1933.4</td>
<td>1980.1</td>
</tr>
<tr>
<td>$\left(\bar{y}<em>i - \bar{y}</em>-\right)^2$</td>
<td>967.2</td>
<td>2520.04</td>
<td>4083.2</td>
<td>295.84</td>
</tr>
<tr>
<td>$n_i$</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

\[
\bar{y}_- = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{k} n_i} = 1997.3
\]
F-test: \( PRDX1 \)

\[
F_g = \frac{\frac{4(967.2) + 4(2520.04) + 4(4083.2) + 4(295.84)}{4 - 1}}{\frac{(2013.7 - 1997.3)^2 + (2162.2 - 1997.3)^2 + \cdots + (1922.4 - 1997.3)^2 - [4(967.2) + 4(2520.04) + 4(4083.2) + 4(295.84)]}{(4 - 1) + (4 - 1) + (4 - 1) + (4 - 1)}}
\]

\[
F_g = \frac{31465.2}{\frac{3}{97967.2 - 31465.2}} = 1.89
\]

\[ P = 0.18 \]
Multiple Hypothesis Testing

- When many hypotheses are tested, each test with a specified Type I error probability, the probability that at least some Type I errors are committed increases sharply with the number of hypotheses tested.

- Investigators are often provided a list of “significant” genes ordered by ascending p-value.
  - Genes selected for further study may be impacted by recall bias.
Multiple Hypothesis Testing

• When testing $H_i, i=1,\ldots,22,283$ simultaneously, the family-wise error rate is the probability of at least one Type I error in the family (FWER)
  – Weak control: Provide control of Type I error when all null hypotheses are true (e.g., Fisher LSD, Newman-Keuls)
  – Strong control: Provide control of Type I error under any combination of true and false null hypotheses.
Multiple Hypothesis Testing

- Bonferroni procedure controls FWER
  - Testing $g$ null hypothesis
  - Reject any $H_i$ with $p_i \leq \frac{\alpha}{g}$
  - $0.05 / 22,283 = 0.0000022$
  - Controls the FWER to be $\leq \alpha$ and to be equal to $\alpha$ if all hypotheses are true.
  - As the number of hypotheses increases, the average power for an individual hypothesis decreases
  - Very conservative; no attempt to incorporate dependence between tests
Multiple Hypothesis Testing

• Šidák Single Step
  – Testing $g$ null hypothesis
  – Reject any $H_i$ with $p \leq 1 - \frac{g}{\sqrt{(1 - \alpha)}}$

$$1 - \frac{22,283}{\sqrt{(1 - 0.05)}} = 0.000002302$$
Multiple Hypothesis Testing

- Holm step-down procedure, FWE
  1) Order the p-values and hypotheses
     \[ P_{(1)} \geq \ldots \geq P_{(g)} \] corresponding to \( H_{(1)}, \ldots, H_{(g)} \)
  2) Let \( i = 1 \).
  3) If \( P_{(g-i+1)} > \alpha / (g - i + 1) \) then accept all remaining hypotheses \( H_{(g - i + 1)} \) and STOP.
  4) If \( P_{(g-i+1)} \leq \alpha / (g - i + 1) \) then reject \( H_{(g - i + 1)} \) and increment \( i \), then return to step 3.
Multiple Hypothesis Testing

• Šidák Step down

1) Order the p-values and hypotheses
   \[ P_1 \geq \ldots \geq P_g \] corresponding to \( H_1, \ldots, H_g \)

2) Let \( i = 1 \).

3) If \( P_{g-i+1} > 1 - \frac{(g-i+1)\sqrt{1-\alpha}}{1-\alpha} \) then accept all remaining hypotheses \( H_{g-i+1} \) and STOP.

4) If \( P_{g-i+1} \leq 1 - \frac{(g-i+1)\sqrt{1-\alpha}}{1-\alpha} \) then reject \( H_{g-i+1} \) and increment i, then return to step 3.
Multiple Hypothesis Testing

• Hochberg step-up

1) Order the p-values and hypotheses
   \[ P(1) \geq \ldots \geq P(g) \] corresponding to \( H(1), \ldots, H(g) \)

2) Let \( i = 1 \).

3) If \( P(i) \leq \alpha/i \) then reject all the remaining hypotheses \( H(i), \ldots, H(g) \) and STOP.

4) If \( P(i) > \alpha/i \) then accept \( H(i) \), increment \( i \), and Return to step 3.
## Multiple Hypothesis Testing

<table>
<thead>
<tr>
<th>Bonferroni</th>
<th>SidakSS</th>
<th>Holm</th>
<th>SidakSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002244</td>
<td>0.000002302</td>
</tr>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002244</td>
<td>0.000002302</td>
</tr>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002244</td>
<td>0.000002302</td>
</tr>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002244</td>
<td>0.000002302</td>
</tr>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002244</td>
<td>0.000002302</td>
</tr>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002244</td>
<td>0.000002302</td>
</tr>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002245</td>
<td>0.000002303</td>
</tr>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002245</td>
<td>0.000002303</td>
</tr>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002245</td>
<td>0.000002303</td>
</tr>
<tr>
<td>0.000002244</td>
<td>0.000002302</td>
<td>0.000002245</td>
<td>0.000002303</td>
</tr>
</tbody>
</table>
Multiple Hypothesis Testing

• The test statistics and hence the p-values are likely correlated due to co-regulation of the genes.

• Would like multiple testing procedures that take into account the dependence structure of the genes.

• This could be accomplished by estimating the joint null distribution of the unadjusted, unknown p-values.
Permutation based methods

• Permutation based adjusted p-values
  – Under the complete null, the joint distribution of the test statistics can be estimated by permuting the columns of the gene expression matrix
  – Permuting entire columns creates a situation in which membership to the Day 1 and Day 2 groups is independent of gene expression but preserves the dependence structure between genes
Permutation based methods

- Permutation algorithm for the $b^{th}$ permutation, $b=1,\ldots,B$
  1) Permute the $n$ columns of the data matrix $X$
  2) Compute test statistics $t_{1,b},\ldots,t_{g,b}$ for each hypothesis

- The permutation distribution of the test statistic $T_j$ for hypothesis $H_j$, $j=1,\ldots,g$ is given by the empirical distribution of $t_{j,1},\ldots,t_{j,B}$. 
## Permutation based methods

<table>
<thead>
<tr>
<th>Gene g on Chip i</th>
<th>Expression</th>
<th>Day</th>
<th>Example Permutations</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2013.7</td>
<td>1</td>
<td>1 2 1 1 ...</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2141.9</td>
<td>1</td>
<td>1 2 1 ...</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2040.2</td>
<td>1</td>
<td>1 1 2 ...</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1973.3</td>
<td>1</td>
<td>1 1 1 ...</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2162.2</td>
<td>1</td>
<td>1 1 1 ...</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1994.8</td>
<td>1</td>
<td>1 1 1 ...</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1913.3</td>
<td>1</td>
<td>1 1 1 ...</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>2068.7</td>
<td>1</td>
<td>1 1 1 ...</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1974.6</td>
<td>2</td>
<td>1 1 1 ...</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>2027.6</td>
<td>2</td>
<td>2 2 2 ...</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>1914.8</td>
<td>2</td>
<td>2 2 2 ...</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>1955.8</td>
<td>2</td>
<td>2 2 2 ...</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1963</td>
<td>2</td>
<td>2 2 2 ...</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>2025.5</td>
<td>2</td>
<td>2 2 2 ...</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1865.1</td>
<td>2</td>
<td>2 2 2 ...</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>1922.4</td>
<td>2</td>
<td>2 2 2 ...</td>
<td>1</td>
</tr>
</tbody>
</table>
Permutation based methods

• For two-sided alternative hypotheses, the permutation p-value for hypothesis $H_j$ is

$$ p^*_j = \frac{1}{b=1} \frac{\sum_{b=1}^{B} I\left(\left|t_{j,b}\right| \geq \left|t_j\right|\right)}{B} $$

where $I(*)$ is the indicator function, equaling 1 if the condition in parentheses is true and 0 otherwise.
Permutation based methods

• Permutation method permits estimation of the joint null distribution of the unadjusted unknown p-values.
• Dependency structure between the genes is preserved.
• May suffer from a granularity problem (when two groups, should have $\geq 6$ arrays in each group to use permutation based method).
Permutation based methods

- Compute test statistic for each gene
- Obtain adjusted p-values for a multiple testing procedure which strongly controls the family-wise Type I error rate.
Considerations for controlling the FWER

• Control over FWER is only appropriate in situations where the intent is to identify only a small number of genes that are truly different.

• Otherwise, the severe loss in power in controlling FWER is not justified.
Considerations for controlling the FWER

- Approaches that set out to control the FWER seek to control the probability of at least one false positive regardless of the number of hypotheses being tested.
  - When the number of hypotheses $N$ is very large, this may be too strict.
  - Controlling FWER with large $N$ may lead to many missed findings.
# Results of Hypothesis Tests

<table>
<thead>
<tr>
<th></th>
<th>Accept $H_0$</th>
<th>Reject $H_0$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>$N_{00}$</td>
<td>$N_{01}$</td>
<td>$N_0$</td>
</tr>
<tr>
<td>Null false</td>
<td>$N_{10}$</td>
<td>$N_{11}$</td>
<td>$N_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$N-N_r$</td>
<td>$N_r$</td>
<td>$N$</td>
</tr>
</tbody>
</table>
False discovery rates

• It may be more appropriate to emphasize the proportion of false positives among the differentially expressed genes.

• The expectation of this proportion is the false discovery rate (FDR) (Benjamini & Hochberg, 1995)
False Discovery Rates

- Controlling for the False Discovery Rate (FDR) allows one to identify as many genes with significant differences as possible, while incurring a relatively low proportion of false positives.

\[
FDR = E \left\{ \frac{N_{01}}{N_r \lor 1} \right\} \quad \text{where} \quad N_r \lor 1 = \max(N_r, 1)
\]
False Discovery Rates

Two procedures for controlling FDR:

- Fix the acceptable FDR level $\delta$ a priori, then find a data-dependent threshold so that the FDR $\leq \delta$. (Benjamini & Hochberg)
- Fix the threshold rule and then form an estimate of the FDR whose expectation is $\geq$ the FDR rule over the significance region. (Storey)
BH Procedure control of FDR

• Assuming the p-values from the null distribution are independent and uniformly distributed
  – Step 1. Let \( p_{(1)} \leq \ldots \leq p_{(N)} \) be the ordered p-values.
  – Step 2. Calculate \( \hat{k} = \max \left\{ k : p_{(k)} \leq \alpha k / N \right\} \)
  – Step 3. If \( \hat{k} \) exists, then reject the null hypotheses corresponding to \( p_{(1)} \leq \ldots \leq p_{(\hat{k})} \); otherwise, reject nothing.
Significance Analysis of Microarrays (SAM)

- SAM seeks to control the proportion of false rejections among the set of rejected hypotheses (FDR). (Tusher et al)
- Permutation method is used to calculate the null distribution of the modified t-statistics.
- The modified t-statistic is

\[ T_j = \frac{\left( \bar{y}_{1j} - \bar{y}_{2j} \right)}{s_j \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + a_0}} \]
SAM

- Because of the large number of genes in the microarray experiments, there will always be some genes with very small sum of squares across replicates, so the absolute t-statistics will be large whether or not the averages exhibit a large difference.

- The small constant $a_0$ is added to the denominator because of the problem in dividing by a very small error term.

- The constant $a_0$ was chosen to make the coefficient of variation of $T_j$ approximately constant as a function on $s_j$. 
SAM

• Plot the order statistics $t_{(1)}, \ldots, t_{(N)}$ are plotted against their null expectations assessed for $j=1, \ldots, N$ as

$$
\bar{t}_{0(j)} = \frac{1}{B} \sum_{b=1}^{B} t_{0(j)}^{(b)}
$$

• For a fixed threshold $\Delta$, start at the origin and move up to the right, finding the first $j=j_1$ such that $t_{(j)} - \bar{t}_{0(j)} \geq \Delta$

• All genes past $j_1$ are called “significant positive.”
SAM

- Second, start at the origin and move down to the left, finding the first $j=j_2$ such that
  \[ \bar{t}_{0(j)} - t_{(j)} \geq \Delta \]
- All genes past $j_2$ are called “significant negative.”
- Thus, the derived cutpoints defining the critical region are not required to be symmetric, and therefore may lead to more powerful tests in situations where more genes are over or under expressed.
SAM

• FDR can be estimated from the permuted null statistics.

• Tusher estimated FDR to be the proportion of the $t_{0,j}^{(b)}$ found to be significant.

• Use is typically to estimate FDR for a range of $\Delta$ values, then select that $\Delta$ which meets your FDR criterion.
False discovery rates

- Pre-selection of genes that pass FDR testing at a moderate level of significance may largely suppress noise introduced by non-differentially expressed genes which obscure the signal in the data (i.e., use for feature selection prior to performing class prediction/discriminant analysis)
- Statistical properties regarding the actual number of false rejections unclear.
Class Prediction

• K-nearest neighbors
• Compound Covariate Predictor
• Logistic Regression
• Classification and Regression Trees
• Etc.
Class Prediction

- Risk of over-fitting the data: may have a perfect discriminator for the data set at hand but the same model may perform poorly on independent data sets.
- Most prediction methods are intended for large ‘n’ (samples) small ‘p’ (covariates) datasets. **Filter the data first.**
- Fit model
- Check model adequacy
- Inference
K Nearest Neighbors

• For each chip, identify the k closest chips using a distance metric (i.e. Euclidean distance) to form a group.

• At each point, the class having the majority vote for the k cases in the neighborhood is assigned as the class label.

• Odd values of k are typically used to avoid ties.
Compound Covariate Predictor

1) Determine differentially expressed genes via the t-test.

2) The two-sample t-statistic of each differentially expressed gene serves as its weight in the compound covariate.

\[ c_j = \sum t_j x_{ij} \]

for sample j, where \( t_j \) is the t-statistic for the two group comparison of labels with respect to gene j, \( x_{ij} \) is the log-ratio measured in sample i for gene j and the sum is over all differentially expressed genes.
Compound Covariate Predictor

3) Establish a classification threshold such as $C_t = (C_1 + C_2)/2$ where $C_1$ and $C_2$ are the mean values of the CCP for samples in class labels 1 and 2 respectively.

4) A sample is predicted to be in class 1 if its CCP is closer to $C_1$ and to be in class 2 if its CCP is closer to $C_2$.

(note, equivalent to one-dimensional linear discriminant analysis on the CCP with equal prior probabilities for the two classes)
Class Prediction

• Regardless of algorithm used, it is essential that once the prediction rule has been defined, an unbiased estimate of the true error rate must be calculated.
  – Apply the prediction rule to an independent dataset and calculate misclassification error
  – For small datasets, k-fold cross-validation or ‘leave-one-out’ cross-validation should be used.
Class Prediction

• In a data rich situation,
  – randomly divide the dataset into two parts, representing a training and test dataset.
  – Build the prediction algorithm using the training dataset
  – Once a final model has been developed, the prediction rule is applied to the test dataset to estimate the generalization error
Class Prediction

• For small sample sizes, withholding a large portion of the data for validation purposes may limit the ability of developing a prediction rule. Therefore, use cross-validation techniques to assess the error.
Class Prediction

- K-fold cross-validation requires one to randomly split the dataset into K equally sized groups.
- Thereafter, the model is fit to K-1 parts of the data and the generalization error is calculated using the Kth remaining part of the data.
- This procedure is repeated so that the generalization error is estimated for each of the K parts of the data, providing an overall estimate of the generalization error and its associated standard error.
Class Prediction

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Leave out data in group 3
• Fit the model to the data in groups 1 – 2, 4 – 10 (learning dataset)
• Calculate the error using observations in group 3 as the test dataset
• Do this for each of the 10 partitions
Preprocessing microarrays

• Prior to applying statistical inferential methods to microarray data (such as class comparisons, class predictions, or clustering procedures) data from microarrays must be pre-processed.
  – Image analysis
  – Normalization/Background correction
  – Expression summaries
  – Quality assessments
Probe Set Expression
Summaries

• An Affymetrix Genechip is an oligonucleotide array consisting of a several perfect match (PM) and their corresponding mismatch (MM) probes that interrogate for a single gene.
  – PM is the exact complementary sequence of the target genetic sequence, composed of 25 base pairs
  – MM probe, which has the same sequence with exception that the middle base (13th) position has been reversed
  – There are roughly 11-20 PM/MM probe pairs that interrogate for each gene, called a probe set
PM and MM intensities are combined to form an expression measure for the probe set (gene)
Probe Set Expression Summaries

- There are various methods for estimating probe set expression summaries.
- The most commonly used methods are
  - Average Difference (old)
  - Microarray Suite 5.0 (MAS 5.0 method)
  - Model Based Expression Index (aka MBEI, Li & Wong, dChip)
  - Robust Multiarray Average (RMA)
    - Now, GC-RMA is available
- Once probe set expression summaries are estimated, tests of hypotheses are performed.
- Probe level data from Affymetrix GeneChip experiments are stored in *.CEL files.
R/Bioconductor example

affy package
multtest package
siggenes package
References


• Troendle, J. *JASA* (90)1995, 370-378.
Software


dChip  www.dchip.org

R  www.r-project.org

Bioconductor  www.bioconductor.org