Biweight Correlation as a Measure of Distance between Genes on a Microarray

Aya Mitani
Pitzer College ’06
Advisor: Professor Johanna Hardin
Pomona College
April 29, 2006
About microarray

- Small chip

- Contains thousands of probes

- Measures mRNA activity in a particular cell type

- Contains control and treatment sample

- Expression level is measured from light intensity
Problem with microarray

- Noisy data
- Needs robust estimation of correlation
- Pearson correlation is often used
  - One outlier can greatly affect correlation
Last summer

M-estimation
weighed average with points farther from the center given less weight

\[
d_i = \sqrt{(x_i - \tilde{\mu})' \tilde{\Sigma}^{-1} (x_i - \tilde{\mu})}
\]

\[
\tilde{\mu} = \frac{\sum_i w(d_i) x_i}{\sum_i w(d_i)}
\]

\[
\tilde{\Sigma} = \frac{\sum_i w(d_i) (x_i - \tilde{\mu})(x_i - \tilde{\mu})'}{\sum_i w(d_i)}
\]

Tukey’s biweight

\[
w(d_i) = \begin{cases} 
  d_i \left(1 - \left(\frac{d_i}{c}\right)^2\right)^2 & d_i \leq c \\
  0 & d_i > c
\end{cases}
\]

Use Minimum Covariance Determinant (MCD) for initial estimation of \( \mu \) and \( \Sigma \)
Plot of Biweight weight function ($w$)
Biweight Correlation Coefficient

\[
bwc_{jk} = \frac{\sigma_{jk}}{\sigma_{jj}\sigma_{kk}}
\]

where \(\sigma_{jk}\) is biweight estimate of covariance of gene \(j\) and gene \(k\)
and \(\sigma_{jj}\) is biweight estimate of variance of gene \(j\)

Want to find out the correlation(similarities/differences) of two genes
Further work to be done

- Computational time

- Biweight correlation on clean data
This Spring

• Matrix correlation vs Pair by pair correlation

• One-step M-estimation

• Median vs MCD

• Biweight correlation good for clean data?
Instead of computing pair by pair correlation, compute correlation matrix from biweight covariance matrix simultaneously

\[ d_i = \sqrt{(x_i - \tilde{\mu})' \tilde{\Sigma}^{-1} (x_i - \tilde{\mu})} \quad (4) \]

\[ \tilde{\mu} = \frac{\sum_i w(d_i) x_i}{\sum_i w(d_i)} \quad (5) \]

\[ \tilde{\Sigma} = \frac{\sum_i w(d_i) (x_i - \tilde{\mu}) (x_i - \tilde{\mu})'}{\sum_i w(d_i)} \quad (6) \]

\[
\begin{pmatrix}
\text{mat.bwc}_{11} & \cdots & \text{mat.bwc}_{1n} \\
\text{mat.bwc}_{21} & \cdots & \text{mat.bwc}_{2n} \\
\vdots & \ddots & \vdots \\
\text{mat.bwc}_{n1} & \cdots & \text{mat.bwc}_{nn}
\end{pmatrix} = \begin{pmatrix}
\sigma_{11} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \sigma_{nn}
\end{pmatrix}^{-1} \begin{pmatrix}
\sigma_{11} & \cdots & \sigma_{1n} \\
\sigma_{21} & \cdots & \sigma_{2n} \\
\vdots & \ddots & \vdots \\
\sigma_{n1} & \cdots & \sigma_{nn}
\end{pmatrix} \begin{pmatrix}
\sigma_{11} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \sigma_{nn}
\end{pmatrix}^{-1}
\]

\[
\text{mat.bwc}_{jk} = \text{bwc}_{jk}??
\]
One-step M-estimation

Converged M-estimation was doing 10-25 iterations on average (Takes 11 seconds to compute 190 pairs of genes)
Few-step

3.5 seconds

5.5 seconds
10-step

8 seconds
Median instead of MCD

- Median for $\tilde{\mu}$

- Median absolute deviation (MAD) for $\tilde{\Sigma}$
  $$\text{MAD}(X) = \text{median}|x_i - \text{median}(x_i)|$$

If converged $\rightarrow$ no difference
20 genes

Median converged

MCD converged

7 seconds
Few-step median

20 genes

Median 3-step
MCD converged

1.5 seconds

Median 5-step
MCD converged

2.5 seconds
10-step median

![Plot showing 20 genes and MCD converged in 5 seconds](image)

5 seconds
10-step median

5-step MCD

20 genes

Median 10−step

MCD converged

−0.5 0.0 0.5 1.0

−0.5 0.0 0.5 1.0

20 genes

Converged

5 seconds

5.5 seconds
Biweight correlation on clean data

How biased/variable compared to Pearson correlation?

Pearson correlation

<table>
<thead>
<tr>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7636</td>
<td>0.8482</td>
<td>0.7850</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Biweight correlation

<table>
<thead>
<tr>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7523</td>
<td>0.8541</td>
<td>0.7945</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What makes the difference?

Multivariate normal data

Pearson correlation vs. Biweight correlation
Concluding remarks

- Biweight correlation is unbiased and similarly variable with Pearson correlation

- Median and median absolute deviation for initiation of $\tilde{\mu}$ and $\tilde{\Sigma}$ is as robust as MCD estimators

- Median and median absolute deviation for initiation of $\tilde{\mu}$ and $\tilde{\Sigma}$ is faster than MCD estimators

- Depending on how robust we want the result to be, computational time can be shortened by number of iterations for speed efficiency
  - Generally, 5 iterations or more is recommended