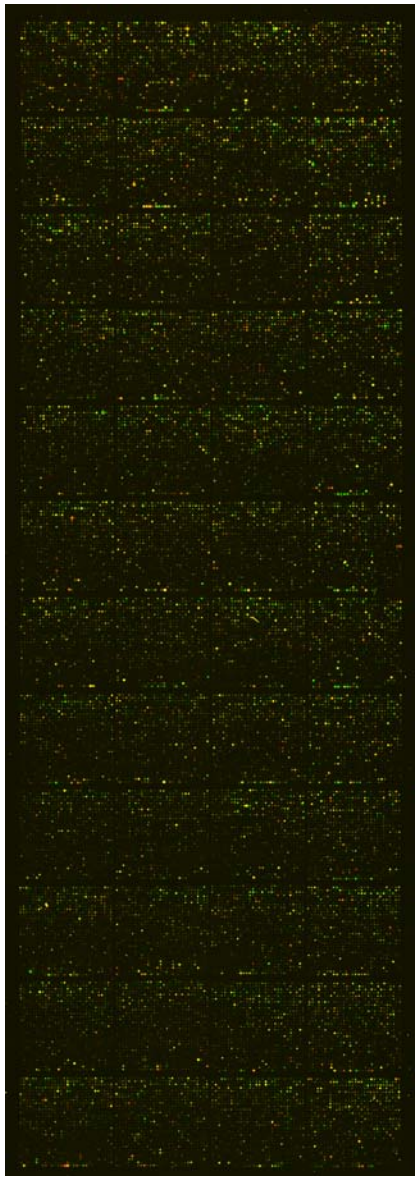
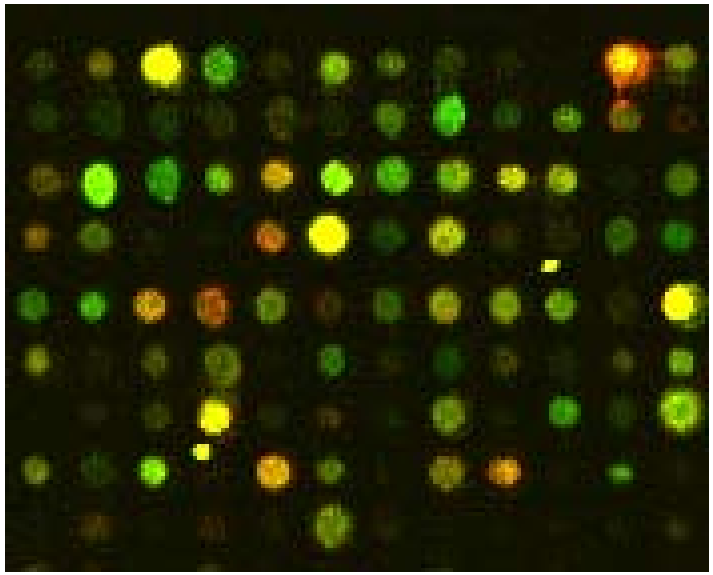
# Biweight Correlation as a Measure of Distance between Genes on a Microarray

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

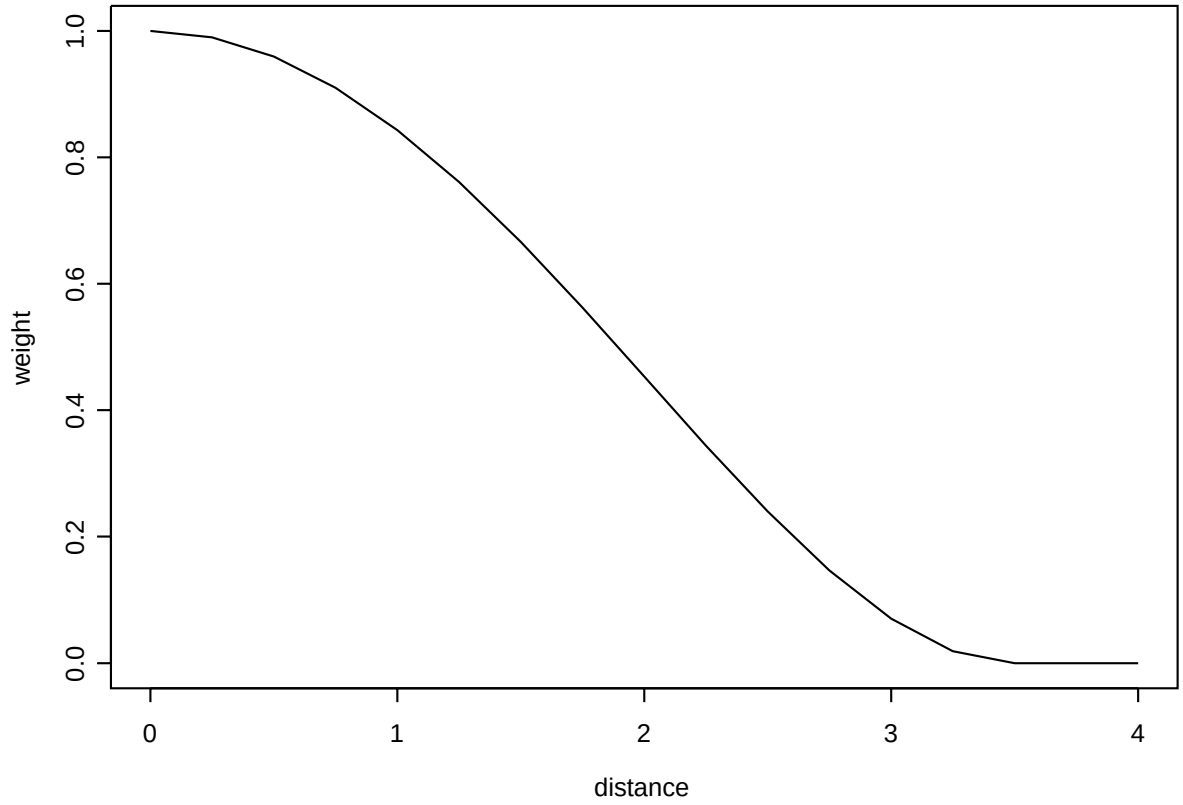
$$
\begin{aligned}
d_i &= \sqrt{(x_i - \tilde{\mu})' \tilde{\Sigma}^{-1} (x_i - \tilde{\mu})} && (1)\\
\tilde{\mu} &= \frac{\sum_i w(d_i) x_i}{\sum_i w(d_i)} && (2)\\
\tilde{\Sigma} &= \frac{\sum_i w(d_i)(x_i - \tilde{\mu})(x_i - \tilde{\mu})'}{\sum_i w(d_i)} && (3)
\end{aligned}
$$

Tukey's biweight

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

Use Minimum Covariance Determinant (MCD) for initial estimation of $\mu$ and $\Sigma$

# Plot of Biweight weight function ($w$)

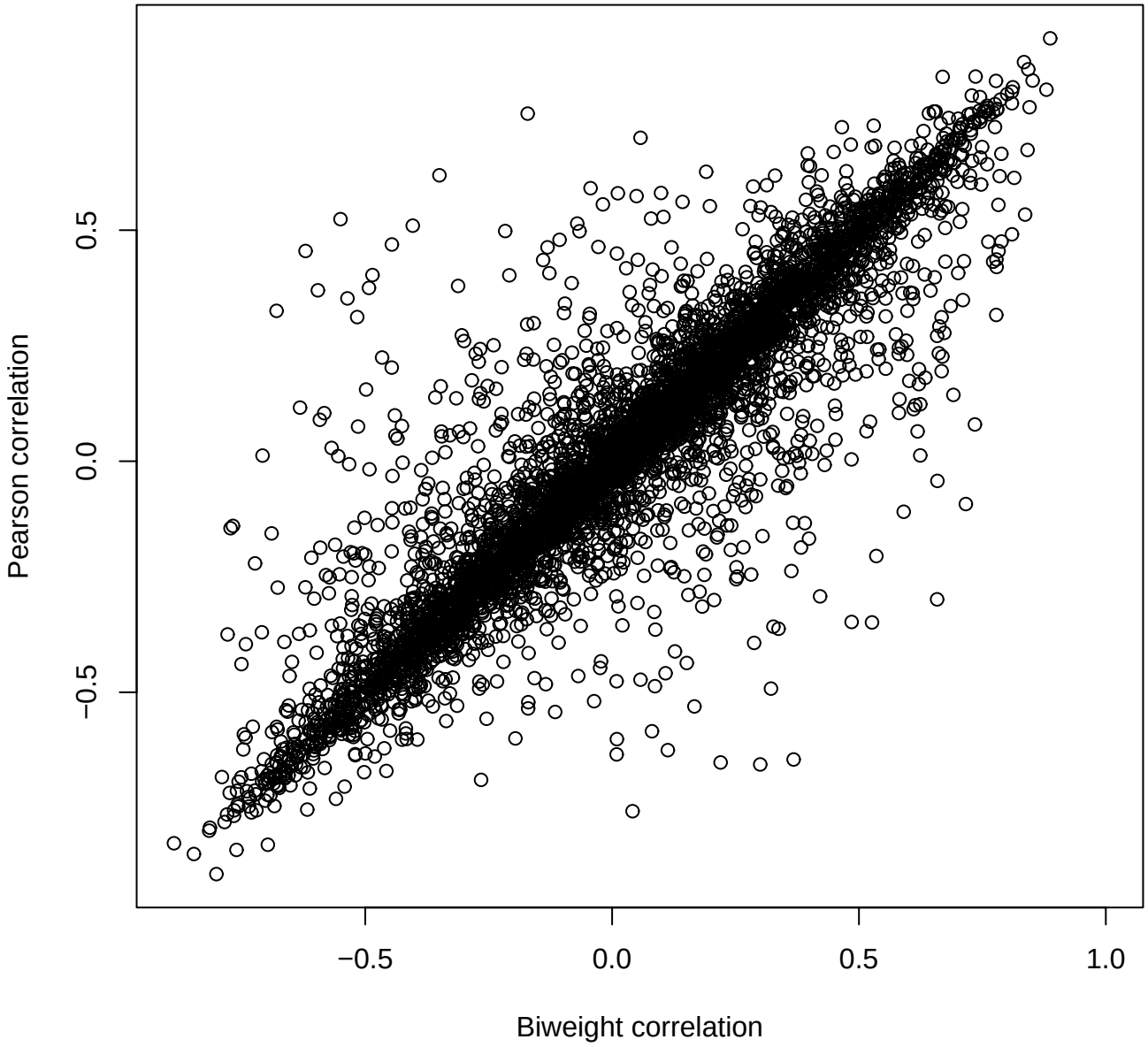


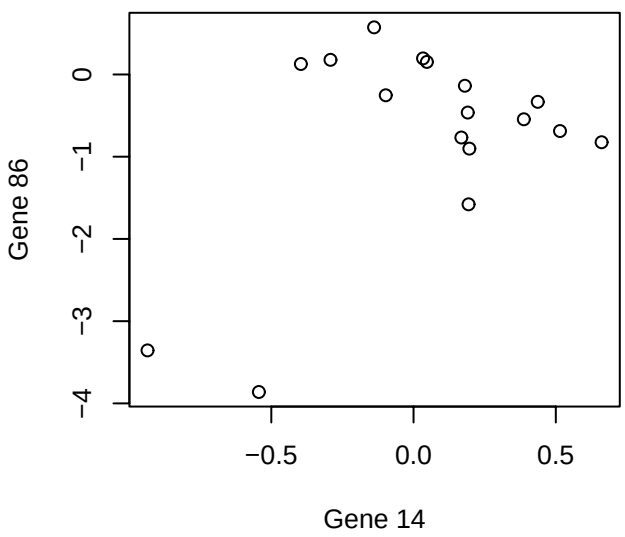
Biweight Correlation Coefficient

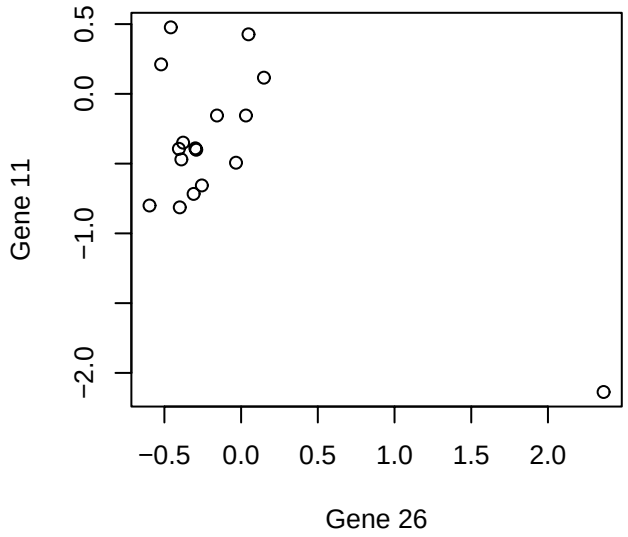
$$bwc_{jk} = \frac{\sigma_{jk}}{\sigma_{jj}\sigma_{kk}}$$

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


Pearson correlation
0.5
0.0
−0.5
−0.5
0.0
0.5
1.0
Biweight correlation


Gene 86
0
−1
−2
−3
−4
−0.5
0.0
0.5
Gene 14


Gene 11
0.5
0.0
−1.0
−2.0
−0.5
0.0
0.5
1.0
1.5
2.0
Gene 26

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

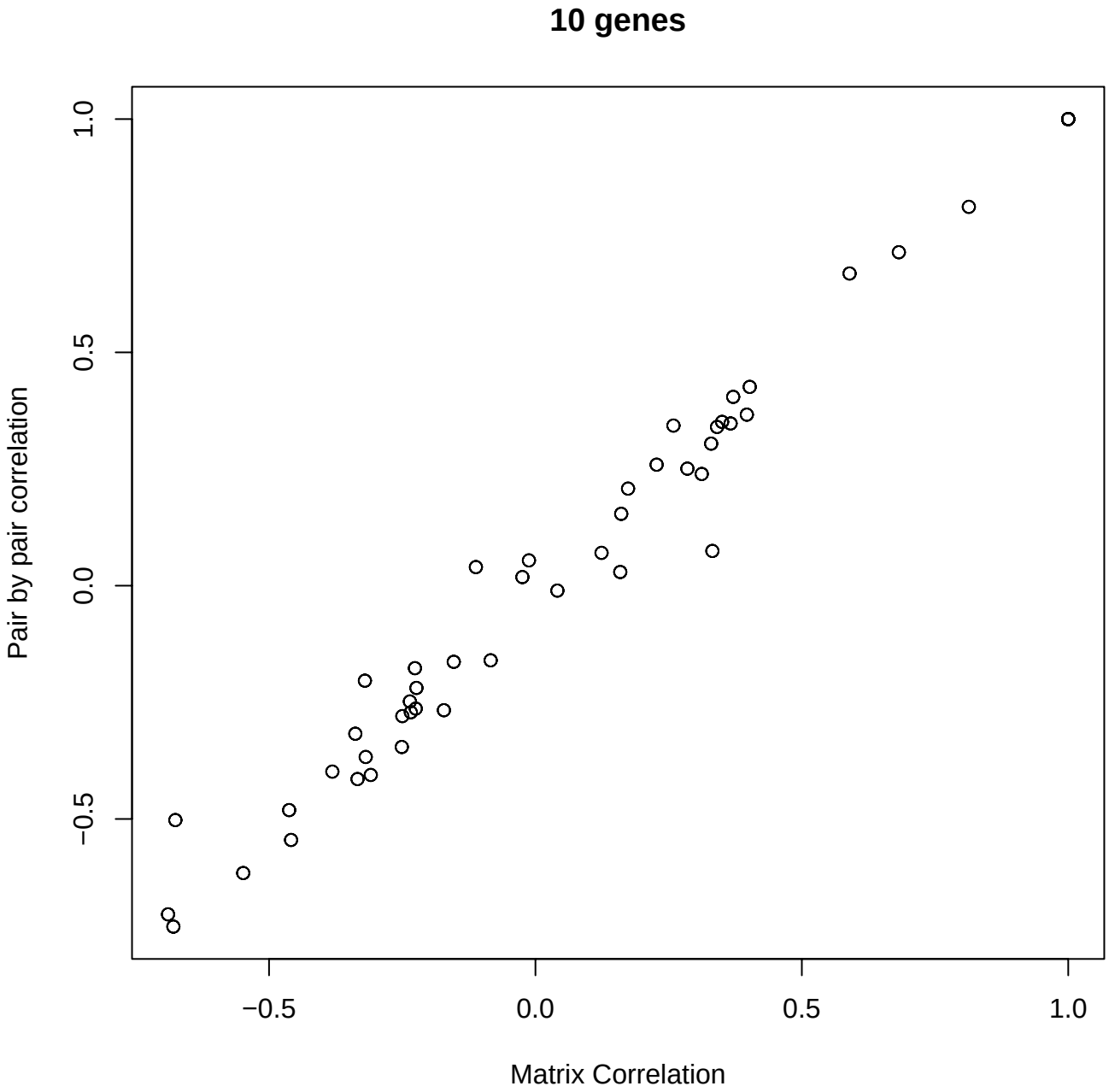
$$d_i = \sqrt{(x_i - \tilde{\mu})' \tilde{\Sigma}^{-1} (x_i - \tilde{\mu})} \quad (4)$$

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

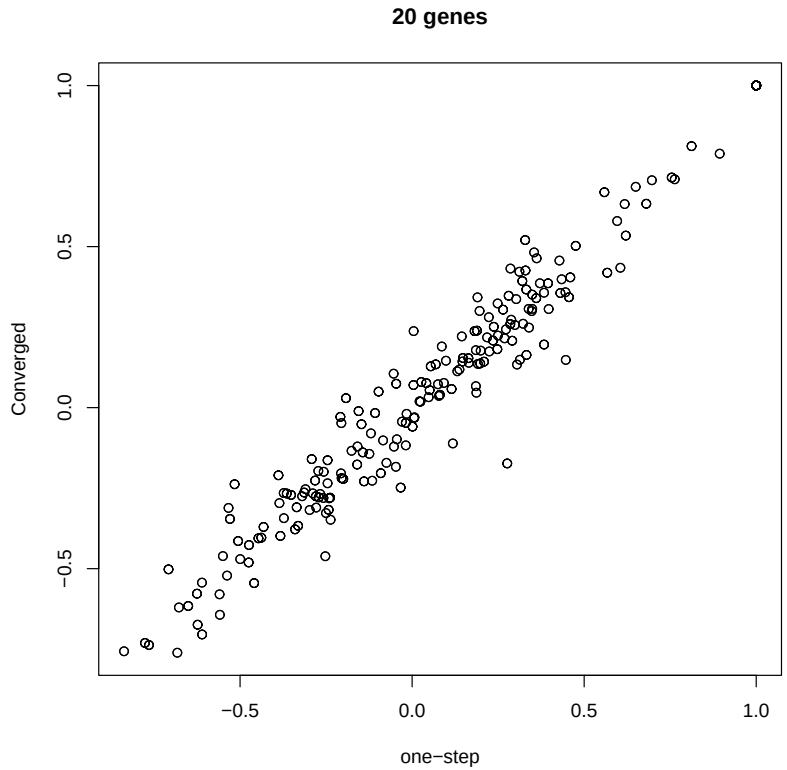
$$\tilde{\Sigma} = \frac{\sum_i w(d_i)(x_i - \tilde{\mu})(x_i - \tilde{\mu})'}{\sum_i w(d_i)} \quad (6)$$

$$\begin{pmatrix} mat.bwc_{11} & \dots & mat.bwc_{1n} \\ mat.bwc_{21} & \dots & mat.bwc_{2n} \\ \vdots & \ddots & \vdots \\ mat.bwc_{n1} & \dots & mat.bwc_{nn} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{nn} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \sigma_{21} & \dots & \sigma_{2n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \dots & \sigma_{nn} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{nn} \end{pmatrix}^{-1}$$

$$mat.bwc_{jk} = bwc_{jk}???$$

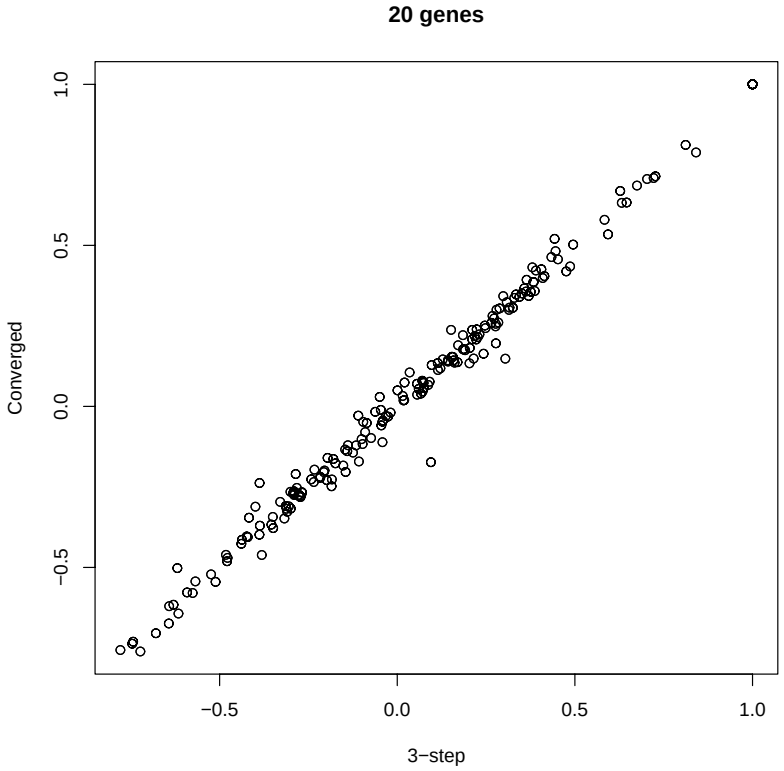

10 genes
Pair by pair correlation
Matrix Correlation
−0.5
0.0
0.5
1.0
−0.5
0.0
0.5
1.0

# One-step M-estimation

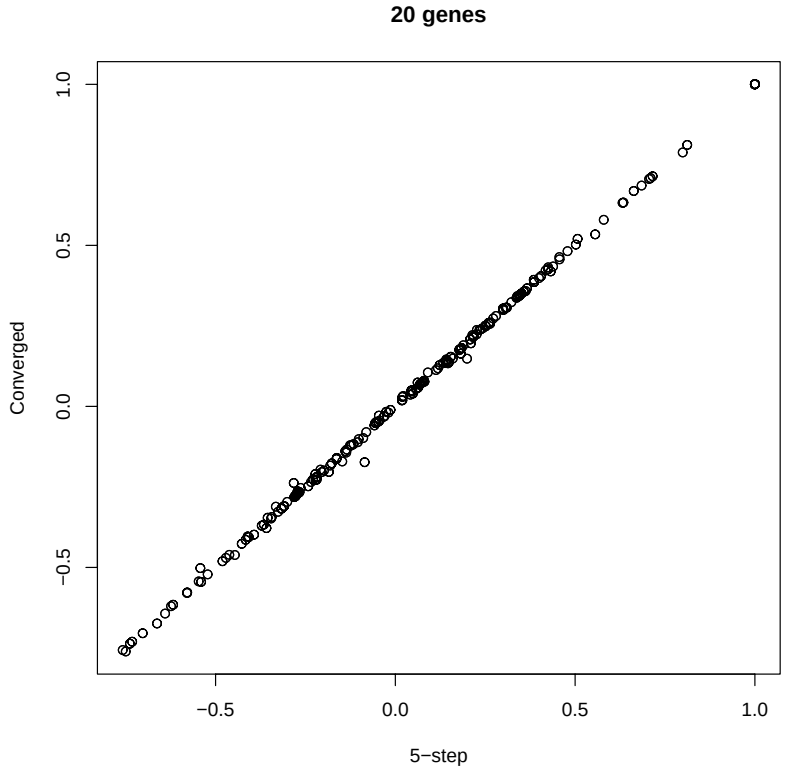



Converged M-estimation was doing 10-25 iterations on average
(Takes 11 seconds to compute 190 pairs of genes)

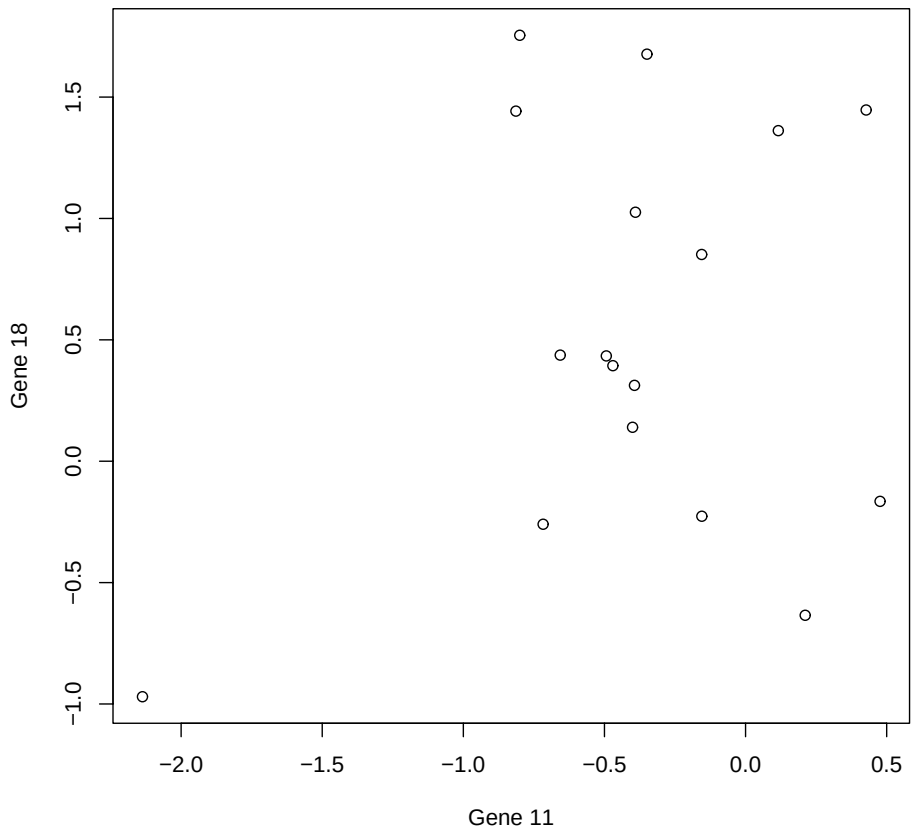
# Few-step



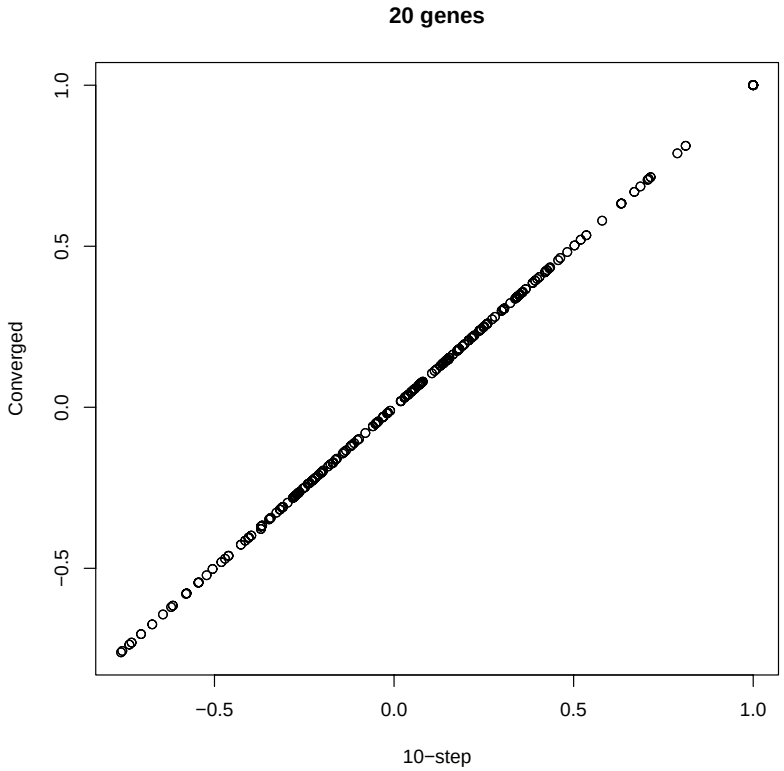

3.5 seconds




5.5 seconds


Gene 18
1.5
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
Gene 11

10-step

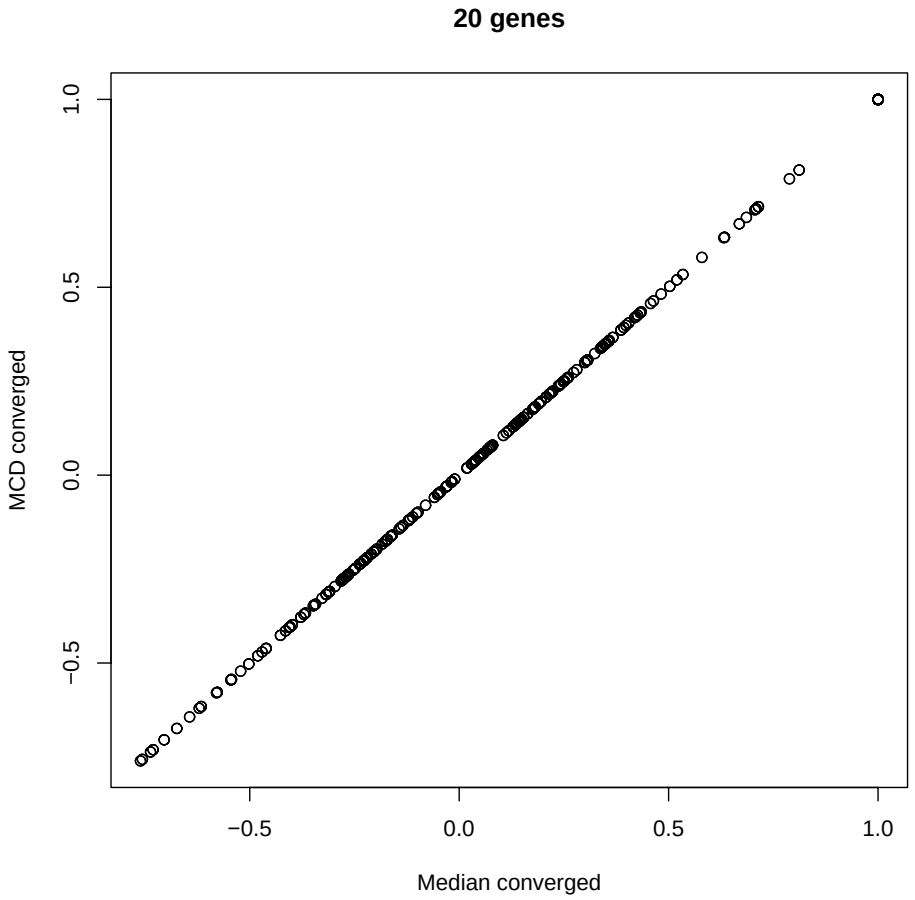



8 seconds

Median instead of MCD

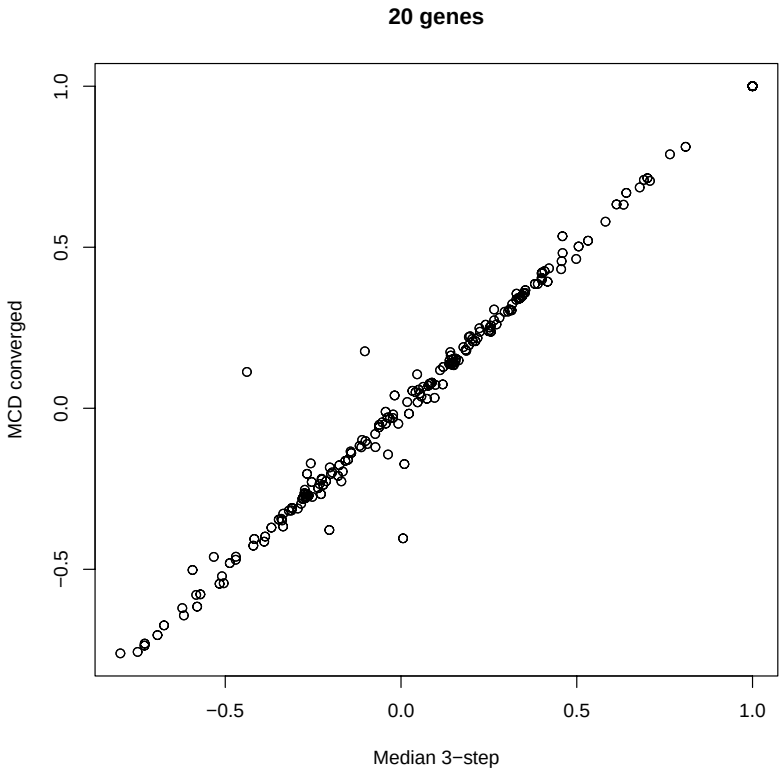
- Median for $\tilde{\mu}$
- Median absolute deviation (MAD) for $\tilde{\Sigma}$
  $\text{MAD}(X) = \text{median}|x_i - \text{median}(x_i)|$

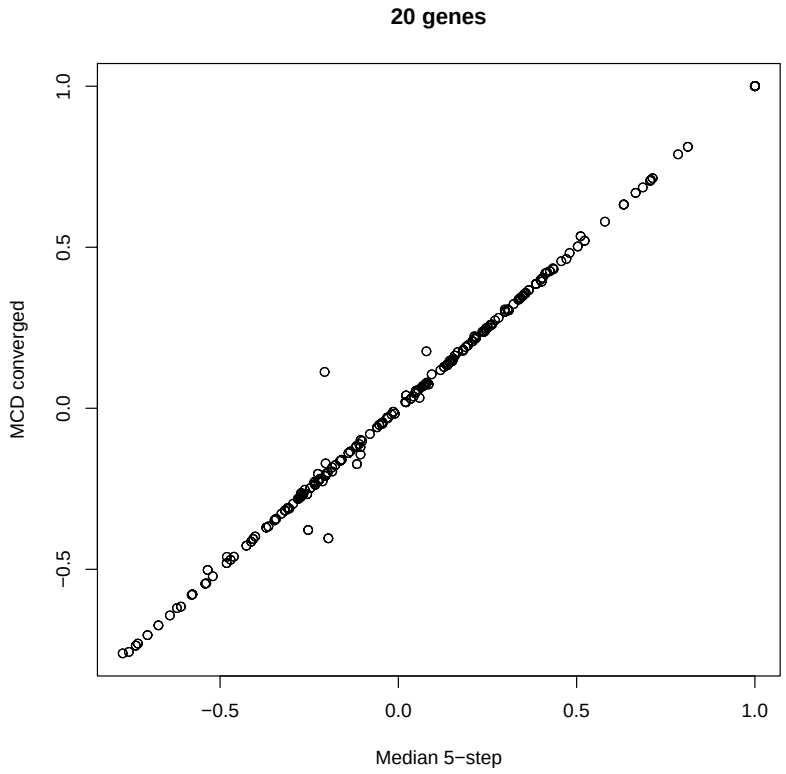
If converged $\rightarrow$ no difference


20 genes
MCD converged
Median converged
1.0
0.5
0.0
−0.5
−0.5
0.0
0.5
1.0

7 seconds

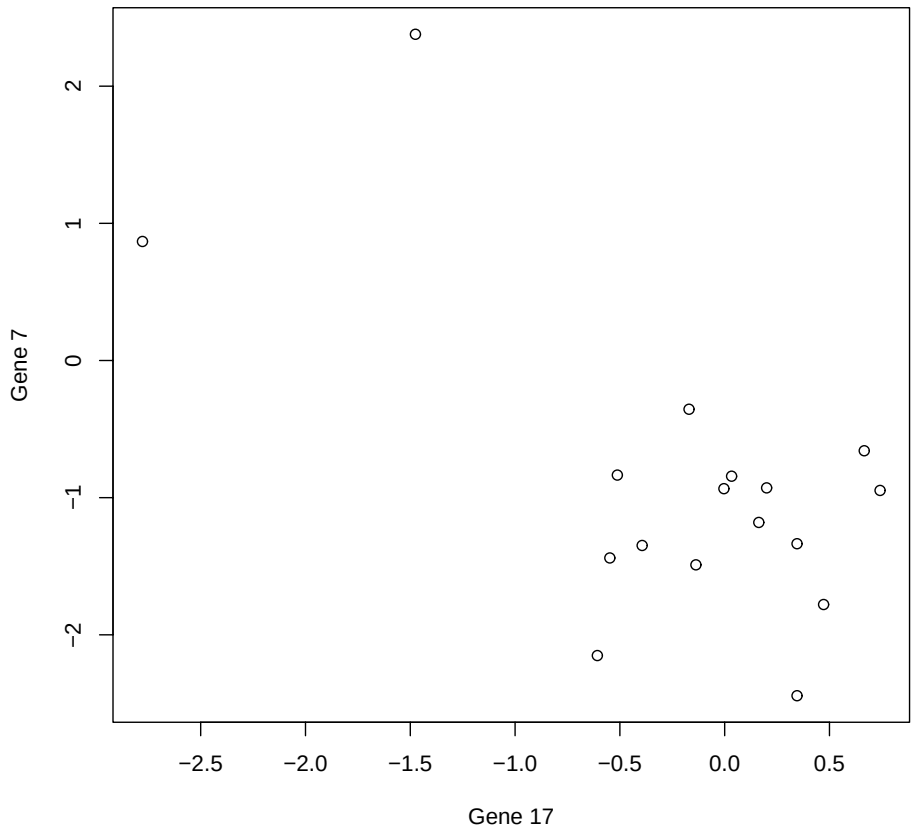
# Few-step median



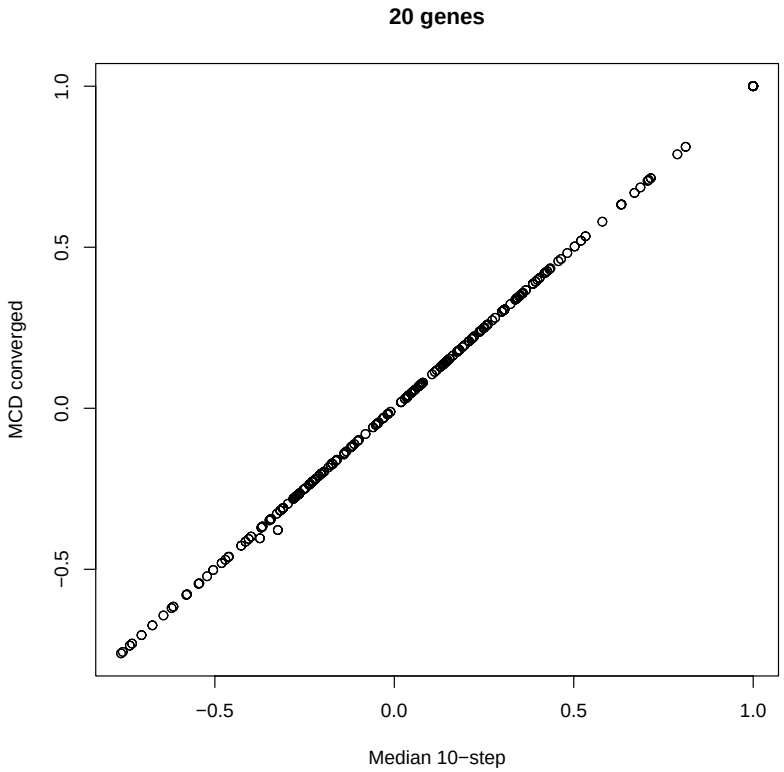




1.5 seconds

2.5 seconds

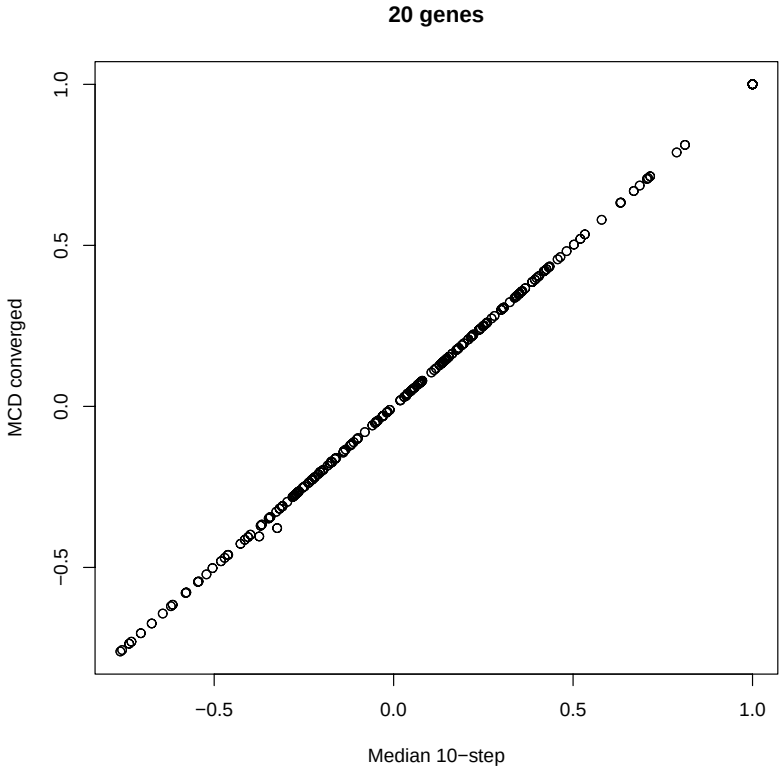

Gene 7
2
1
0
−1
−2
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
Gene 17

10-step median



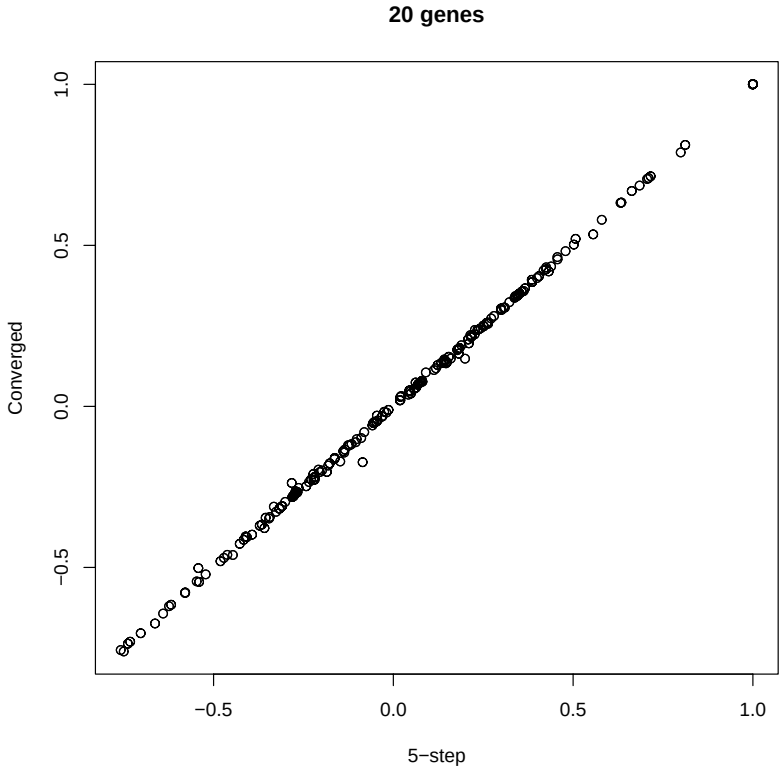

5 seconds

10-step median




5 seconds

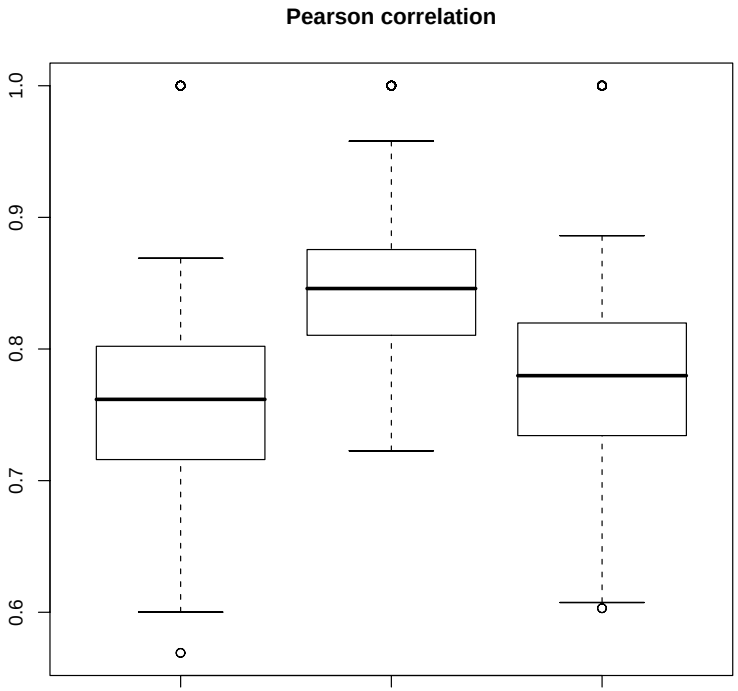
5-step MCD

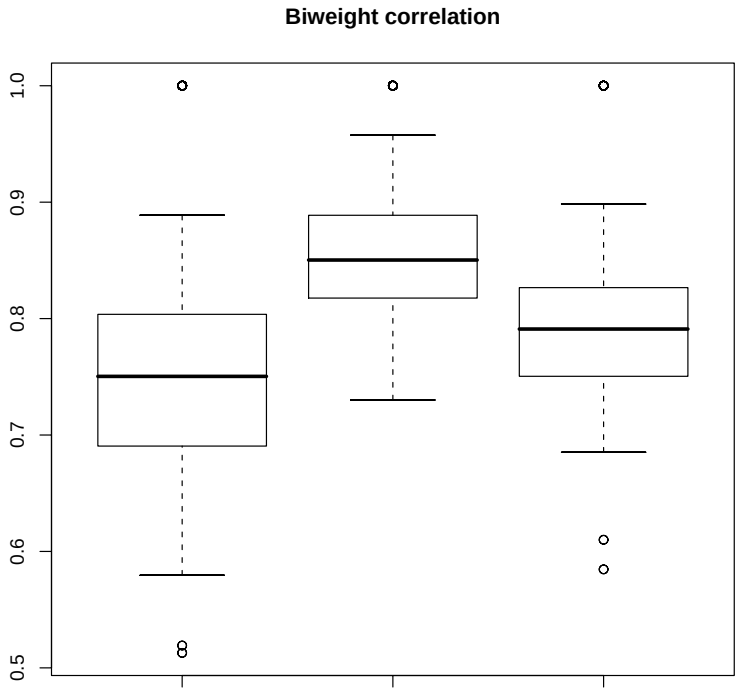



5.5 seconds

# Biweight correlation on clean data

How biased/variable compared to Pearson correlation?

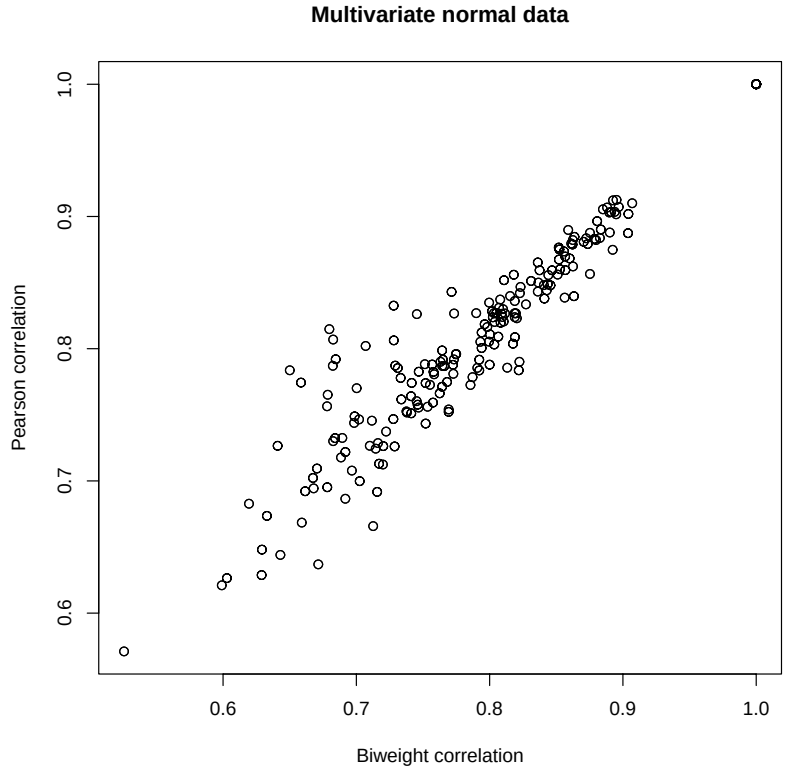


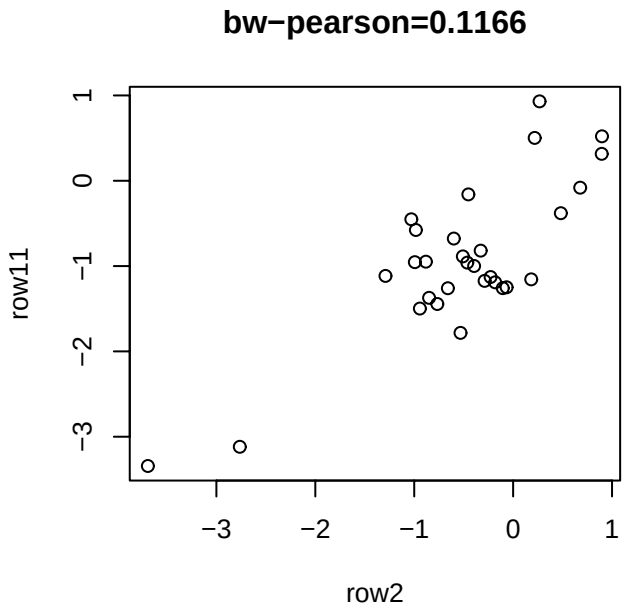



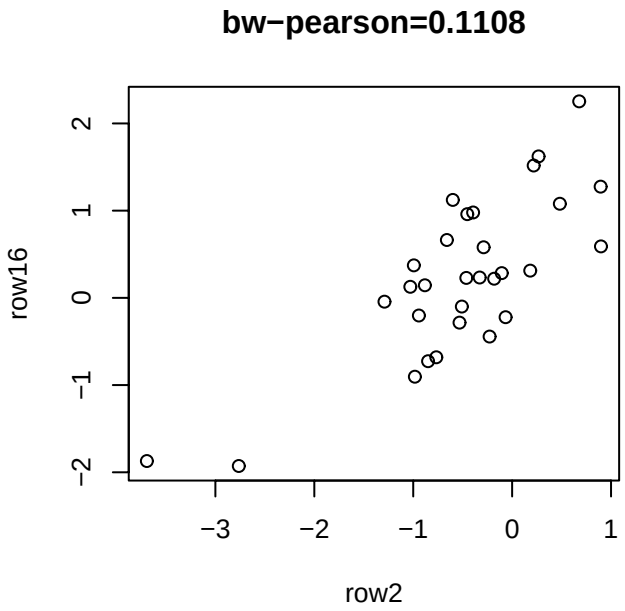

0.7636 0.8482 0.7850

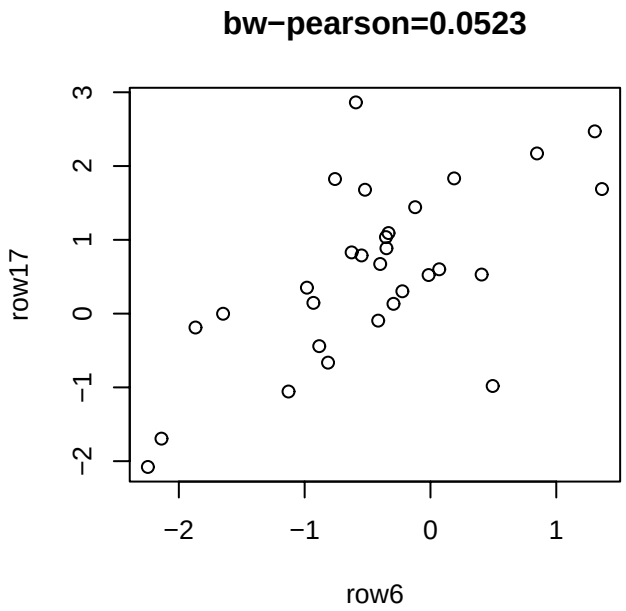
0.7523 0.8541 0.7945

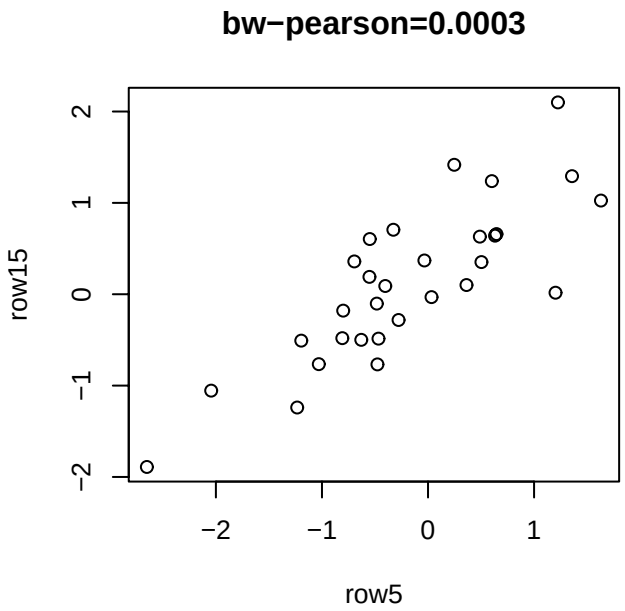
# What makes the difference?




bw−pearson=0.1166
row11
row2
−3
−2
−1
0
1


bw−pearson=0.1108
row16
row2
−3
−2
−1
0
1
2


bw−pearson=0.0523
row17
row6
−2
−1
0
1
2
3


bw−pearson=0.0003
row15
row5
−2
−1
0
1
2

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