Outline for Microarray Data Analysis

1. Introduction of Microarry
2 . Statistical Analyses and Data Visualization
   - Distance Measures in DNA Microarray Data Analysis
   - Cluster Analysis of Genomic Data
   - Analysis of Differential Gene Expression Studies
   - Multiple Testing Procedures and Applications to Genomics
   - Machine Learing Concepts and Tools for Statistical Genomics
3. Preprocessing
   - Preprocessing High-density Oligonucleotide and Two-Color Spotted Arrays
   - Preprocessing SELDI-TOF Mass Spectrometry Protein Data

Microarray Platforms

- Two main classes of platforms:
  - High-density oligonucleotide array (e.g. Affymetrix GeneChips): Contain one set of probe-level data per microarray; some probes for specific finding and others for nonspecific finding.
  - Two-color spotted array (e.g. cDNA): Two colors represent the two samples (experiment and reference) competitively hybridized.

Microarray Data

- Two-color spotted array (e.g. cDNA): Measure relative abundance of a probe sequence in experimental and reference samples
  - Relative expression measure
  - log ratios of intensities
- High-density oligonucleotide array (e.g. Affymetrix GeneChips): Measure overall abundance of a probe in the experimental samples
  - Absolute expression measure
  - log intensities

Comparison of cDNA Arrays

- Usually a common reference is used across multiple slides; it provides a *baseline* for direct comparison of expression measures between arrays.
  - Comparable: Gene X in patient i and patient j (between-sample, within-gene)
  - Incomparable: Gene X and Gene Y in patient i (within-sample, between-gene)

Comparison of Affymetrix Arrays

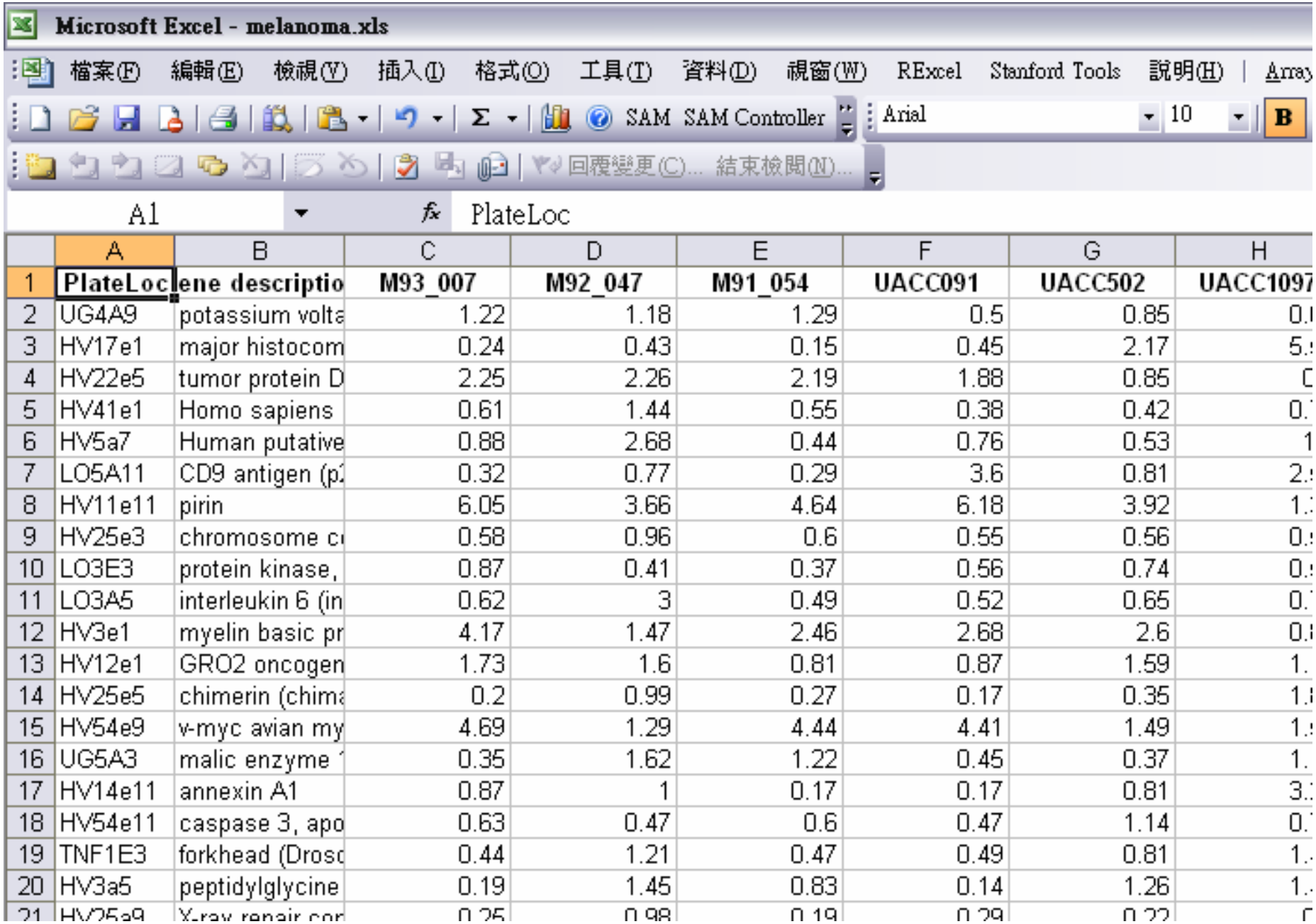
- No common reference across slides; certain normalization techniques have to be applied before comparison.
  - Comparable after normalization: Gene X in patient i and patient j (between-sample, within-gene)
  - Incomparable: Gene X and Gene Y in patient i (within-sample, between-gene)

Comparison of Affymetrix Arrays

- Mimic reference sample for Affymetrix (?!)

Select a particular array to use as a reference and take ratios of all expression measures to this reference?!

Not quite as successful as it is for cDNA arrays, with which both experimental and reference samples are co-hybridized to the same slide.

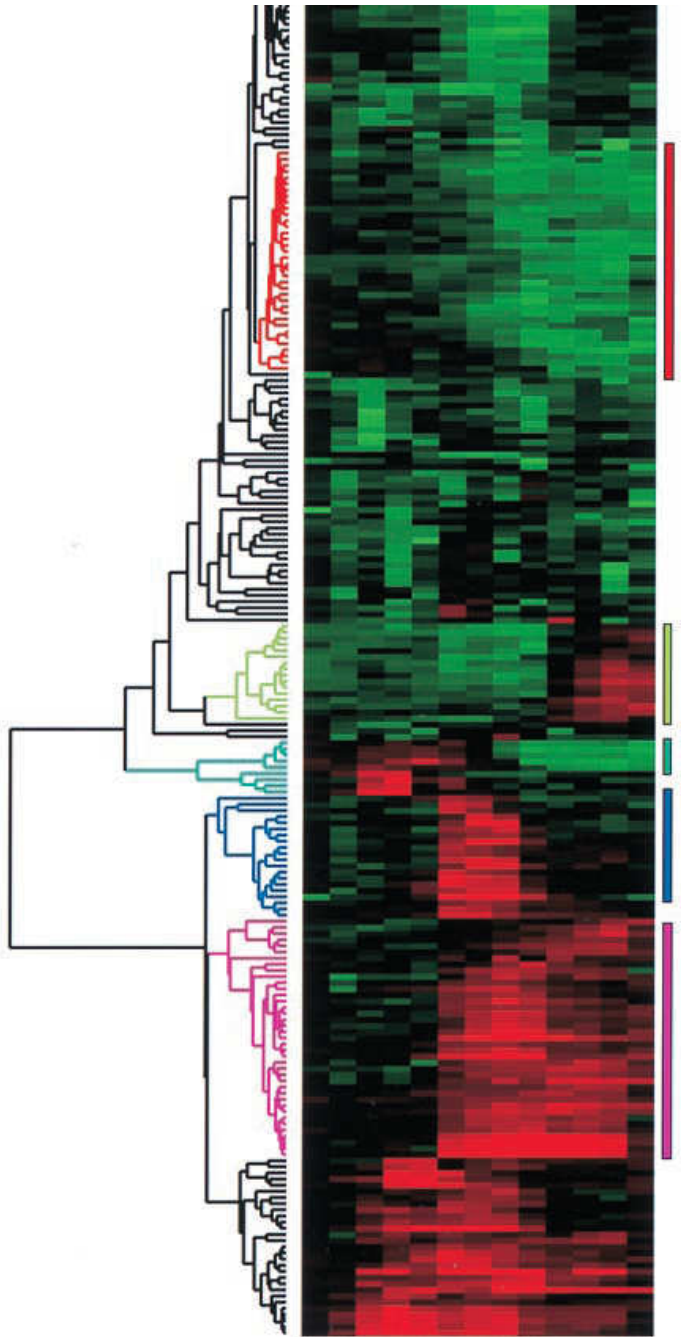


	A	B	C	D	E	F	G	H
1	PlateLoc	ene descriptio	M93_007	M92_047	M91_054	UACC091	UACC502	UACC1097
2	UG4A9	potassium volta	1.22	1.18	1.29	0.5	0.85	0.
3	HV17e1	major histocom	0.24	0.43	0.15	0.45	2.17	5.
4	HV22e5	tumor protein D	2.25	2.26	2.19	1.88	0.85	0
5	HV41e1	Homo sapiens	0.61	1.44	0.55	0.38	0.42	0.
6	HV5a7	Human putative	0.88	2.68	0.44	0.76	0.53	1
7	LO5A11	CD9 antigen (p	0.32	0.77	0.29	3.6	0.81	2.
8	HV11e11	pirin	6.05	3.66	4.64	6.18	3.92	1.
9	HV25e3	chromosome c	0.58	0.96	0.6	0.55	0.56	0.
10	LO3E3	protein kinase,	0.87	0.41	0.37	0.56	0.74	0.
11	LO3A5	interleukin 6 (in	0.62	3	0.49	0.52	0.65	0.
12	HV3e1	myelin basic pr	4.17	1.47	2.46	2.68	2.6	0.
13	HV12e1	GRO2 oncogen	1.73	1.6	0.81	0.87	1.59	1.
14	HV25e5	chimerin (chim	0.2	0.99	0.27	0.17	0.35	1.
15	HV54e9	v-myc avian my	4.69	1.29	4.44	4.41	1.49	1.
16	UG5A3	malic enzyme	0.35	1.62	1.22	0.45	0.37	1.
17	HV14e11	annexin A1	0.87	1	0.17	0.17	0.81	3.
18	HV54e11	caspase 3, apo	0.63	0.47	0.6	0.47	1.14	0.
19	TNF1E3	forkhead (Drosc	0.44	1.21	0.47	0.49	0.81	1.
20	HV3a5	peptidylglycine	0.19	1.45	0.83	0.14	1.26	1.
21	HV25a9	X-ray repair cor	0.25	0.98	0.19	0.29	0.22	0

Data Visualization

- Visualization is about to convey *important* information to the reader *accurately*.
- Color is an important aspect of visualization. The color schemes should be intuitive, consistent, and ergonomic.

Example for an unergonomic color scheme: red-green color.



Data Visualization

- Useful visualization tools:
  - Showing feature of the expression level of one particular gene (sample): histogram, box plot
  - Comparing expression levels of two genes (samples): side-by-side box plot, scatter plot, MA plot
  - Presenting the similarity among multiple genes (samples): side-by-side box plot, heatmap

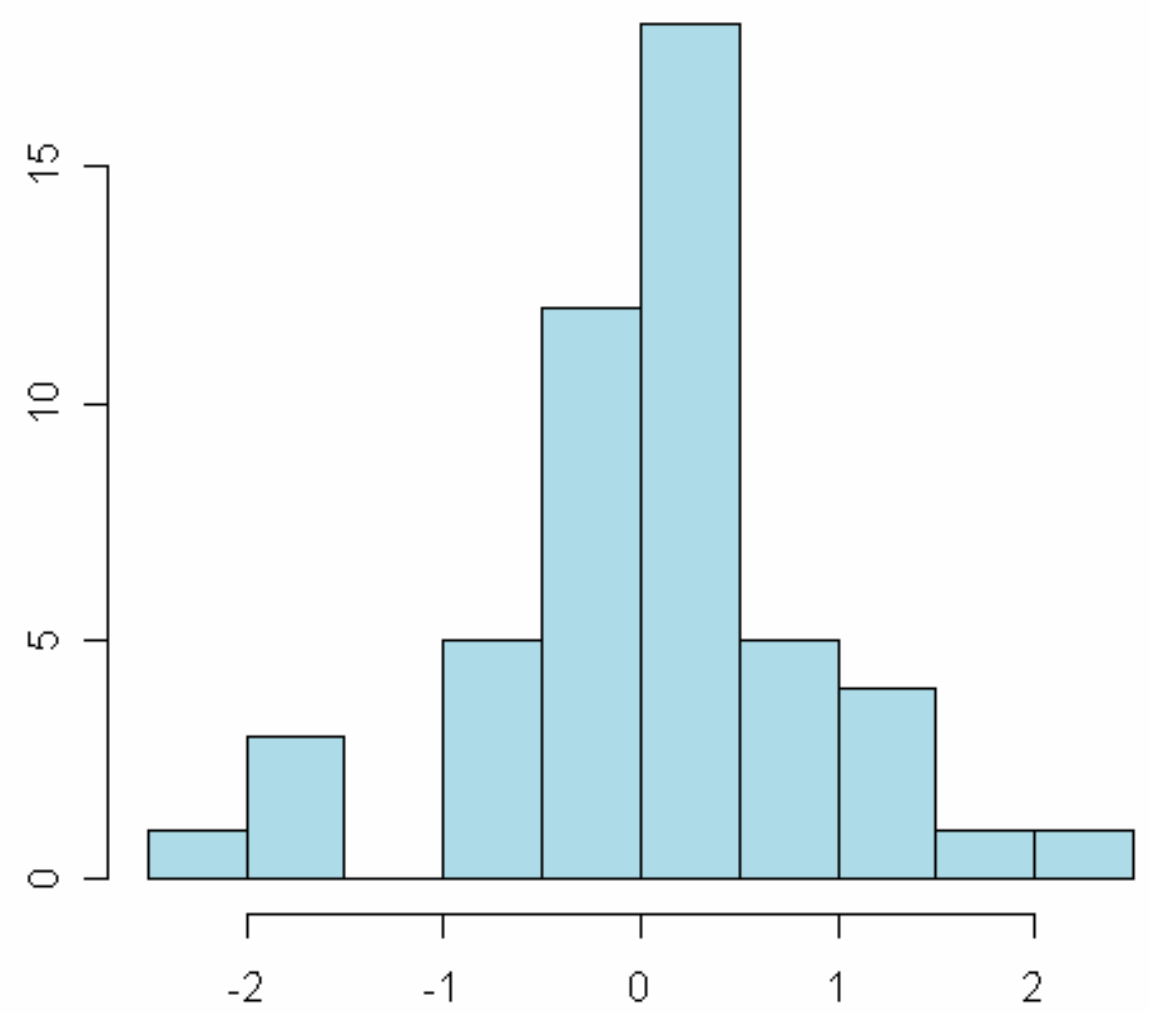
Histogram

- Histogram: the graph shows the frequency distribution of the values in a given data set.

  Step 1: Fractionate the entire range of values encountered in the data set into several intervals; these intervals are called bins in the histogram.

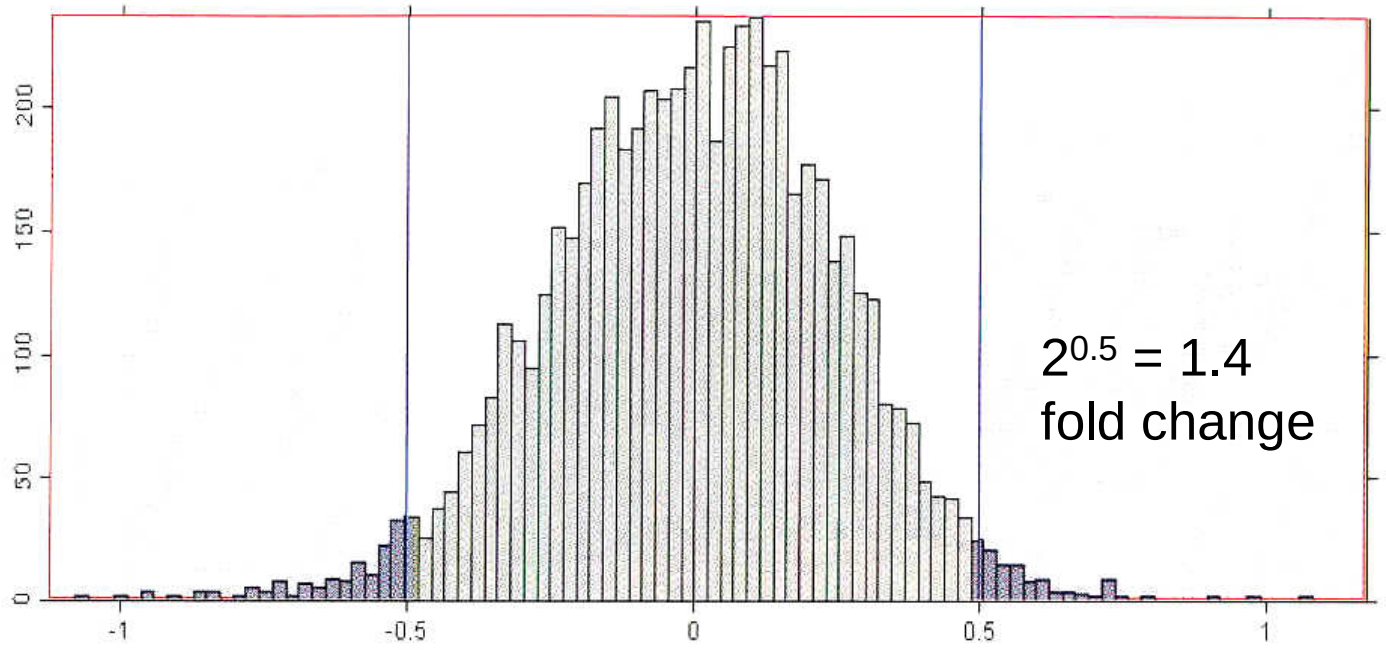
  Step2: Draw a bar for each bin and the height of the bar will be equal to the number of values falling in the interval represented by the bin.

Histogram – An Example



Histograms for log Ratios

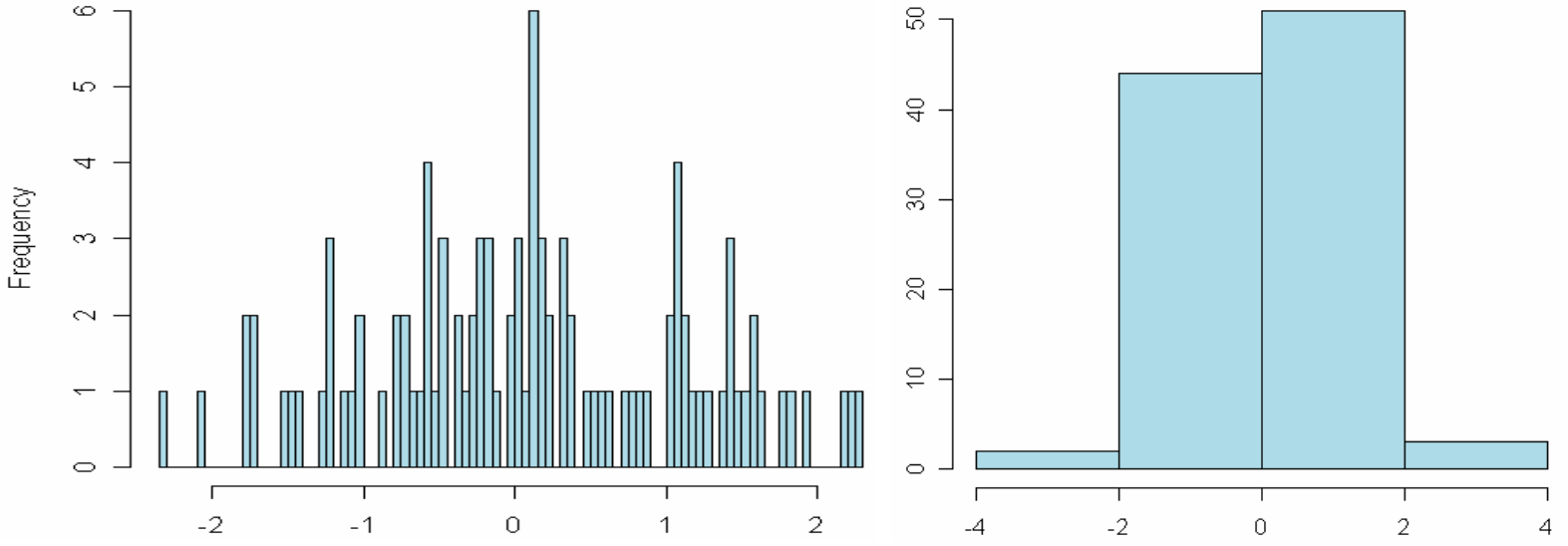
- If the log ratios have been well normalized, differentially expressed genes will be found in the tails of the histogram. Therefore, the histogram can be used to select the genes that have a minimum desired fold change.



Histogram – Determine Bin Size

- Too few or too many bins result in less informative histograms.

  How to determine the number of bins?



Histogram – Determine Bin Size

- To determine the number of bins:
  - Rule of thumb = $\sqrt{N}$
  - Sturges' rule = $1 + \log_2 N$
  - Scott's rule = $R\sqrt[3]{N}\Big/\left(3.5\sqrt{\mathrm{V}(X)}\right)$
  - Friedman-Diaconis = $R\sqrt[3]{N}\Big/\left(2\times \mathrm{IQD}\right)$

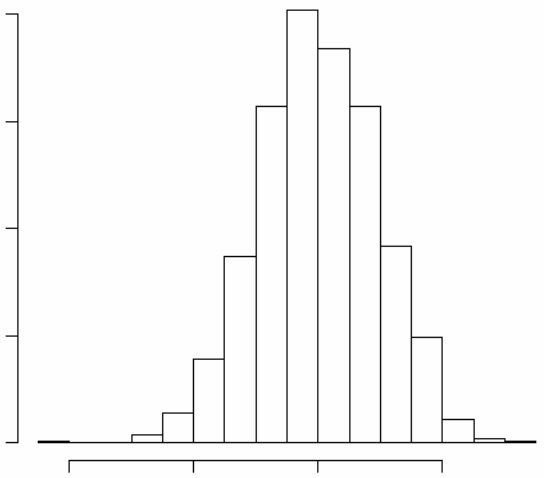
  Note: N = number of observations

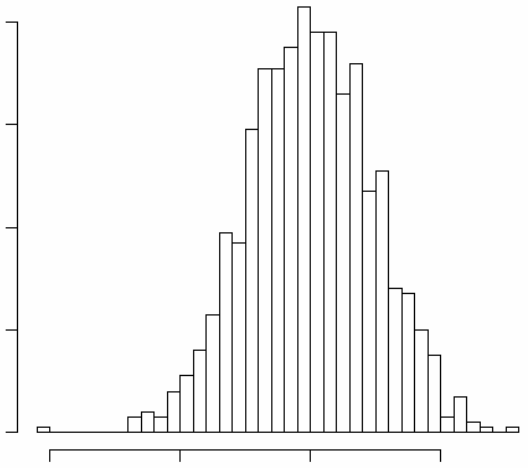
  R = range of observations = max – min

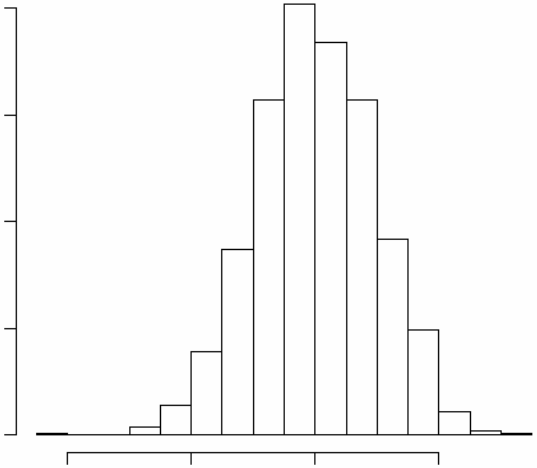
  IQD = inter-quantile distance

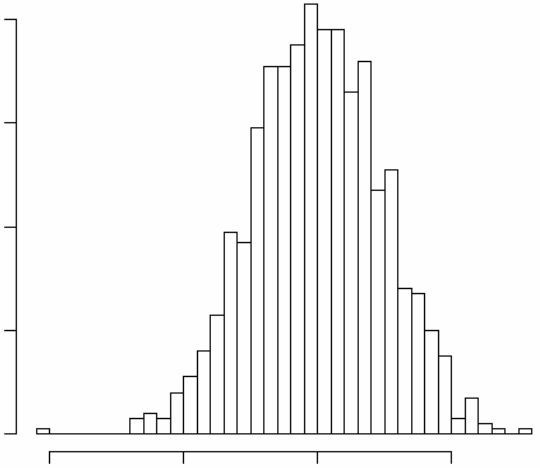
R: Histogram

```
> hist(x)      # Sturges' rule by default
> hist(x,sqrt(length(x)))    # rule of thumb
> hist(x,"scott")     # Scott's rule
> hist(x,"FD")     # Friedman-Diaconis
```



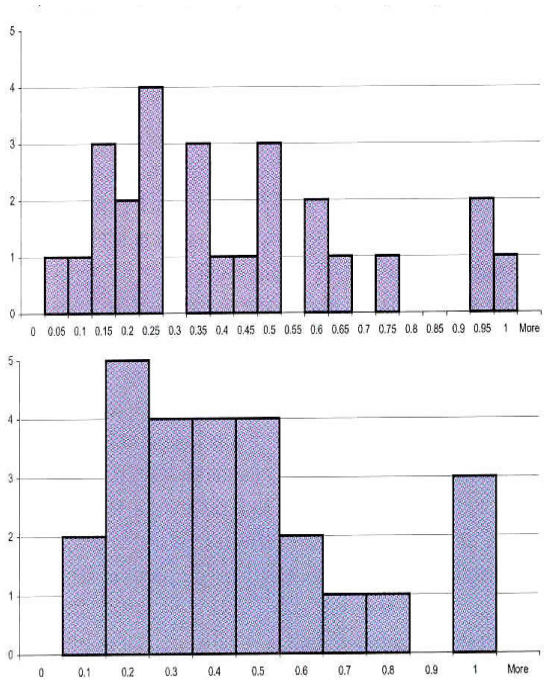




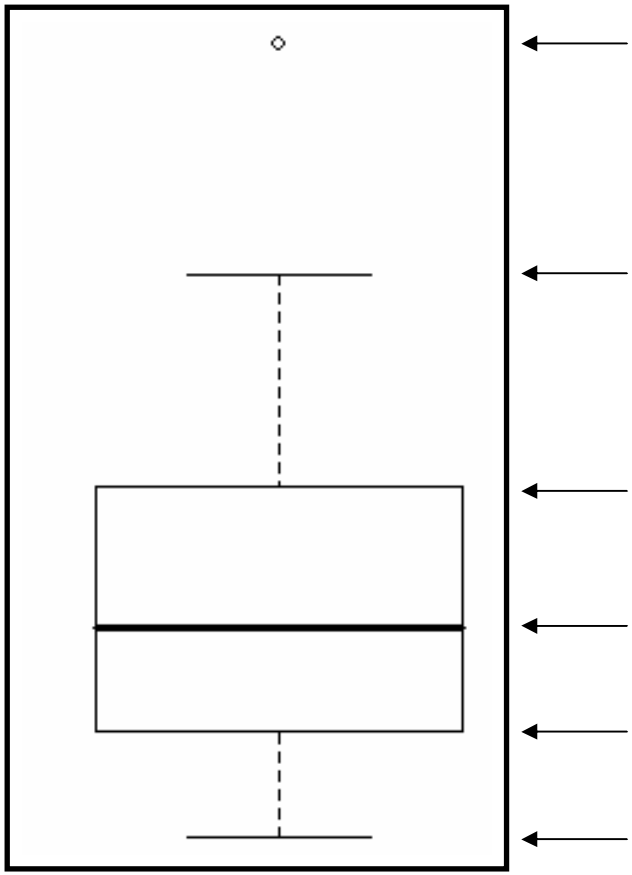


Histogram – Binning Artifact

- Inappropriate binning in histograms may cause information loss or false interpretation.
- Binning artifacts usually can be detected by plotting histograms given different bin sizes.



Box plots



outlier(s): observations that are greater than UQ+1.5*IQD or less than LQ-1.5*IQD

The largest observation that is smaller than UQ+1.5*IQD

Upper Quantile (UQ)

Median

Lower Quantile (LQ)

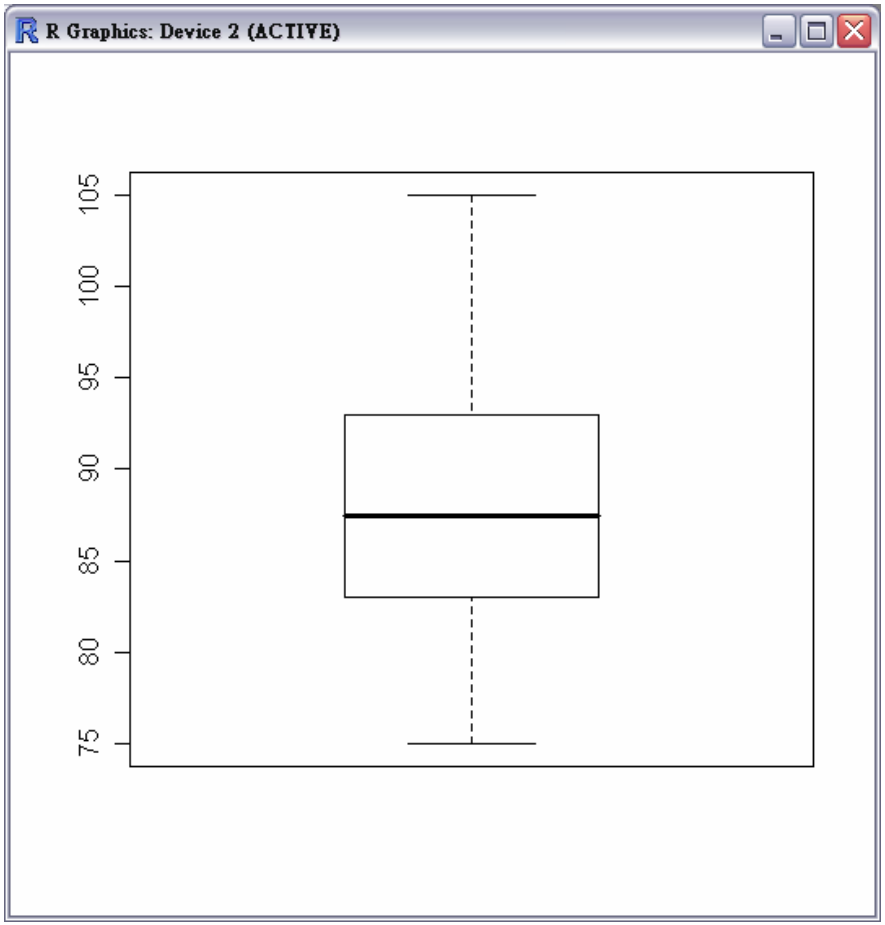
IQD = UQ-LQ

The smallest observation that is greater than LQ-1.5*IQD

R: Box plots

- Box plots for single variable:

  > boxplot(x)



Visualization: Single Gene

Data download:

http://homepage.ntu.edu.tw/~lyliu/IntroBioinfo/BreastCancer_ERp.xls

.xls → .csv

```
> bcdata = read.csv("BreastCancer_ERp.csv")
> y = bcdata[,3]

> hist(y,xlab=names(bcdata)[3],main="")
> boxplot(y)
```

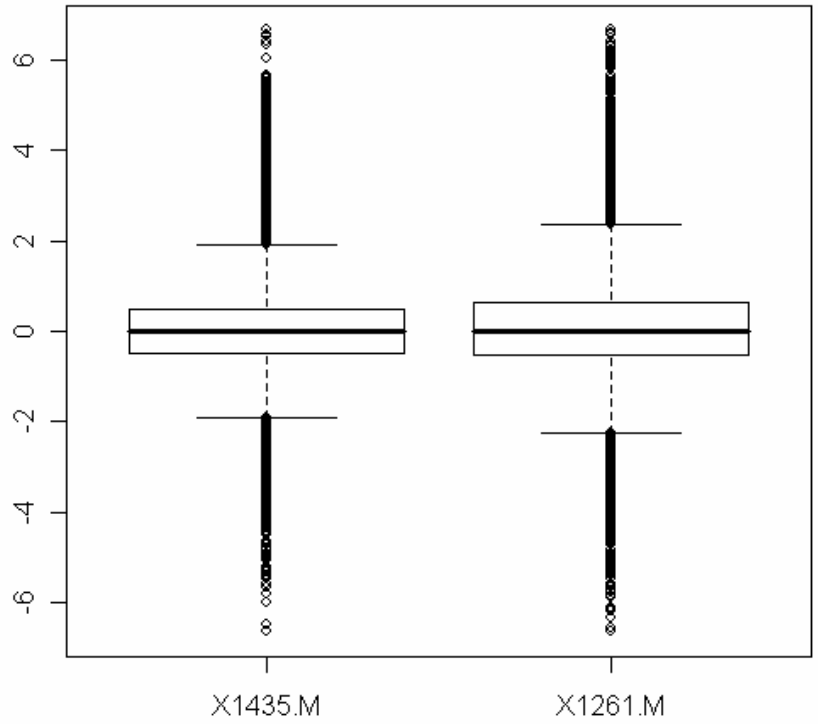
Data Visualization

- Useful visualization tools:
  - Showing feature of the expression level of one particular gene (sample): histogram, box plot
  - Comparing expression levels of two genes (samples): side-by-side box plot, scatter plot, MA plot
  - Presenting the similarity among multiple genes (samples): side-by-side box plot, heatmap

Side-by-side Box Plots

- Side-by-side box plots for two or more groups:

```
> bcdata = read.csv("BreastCancer_ERp.csv")
> x = bcdata[,3:4]
> boxplot(x)
```

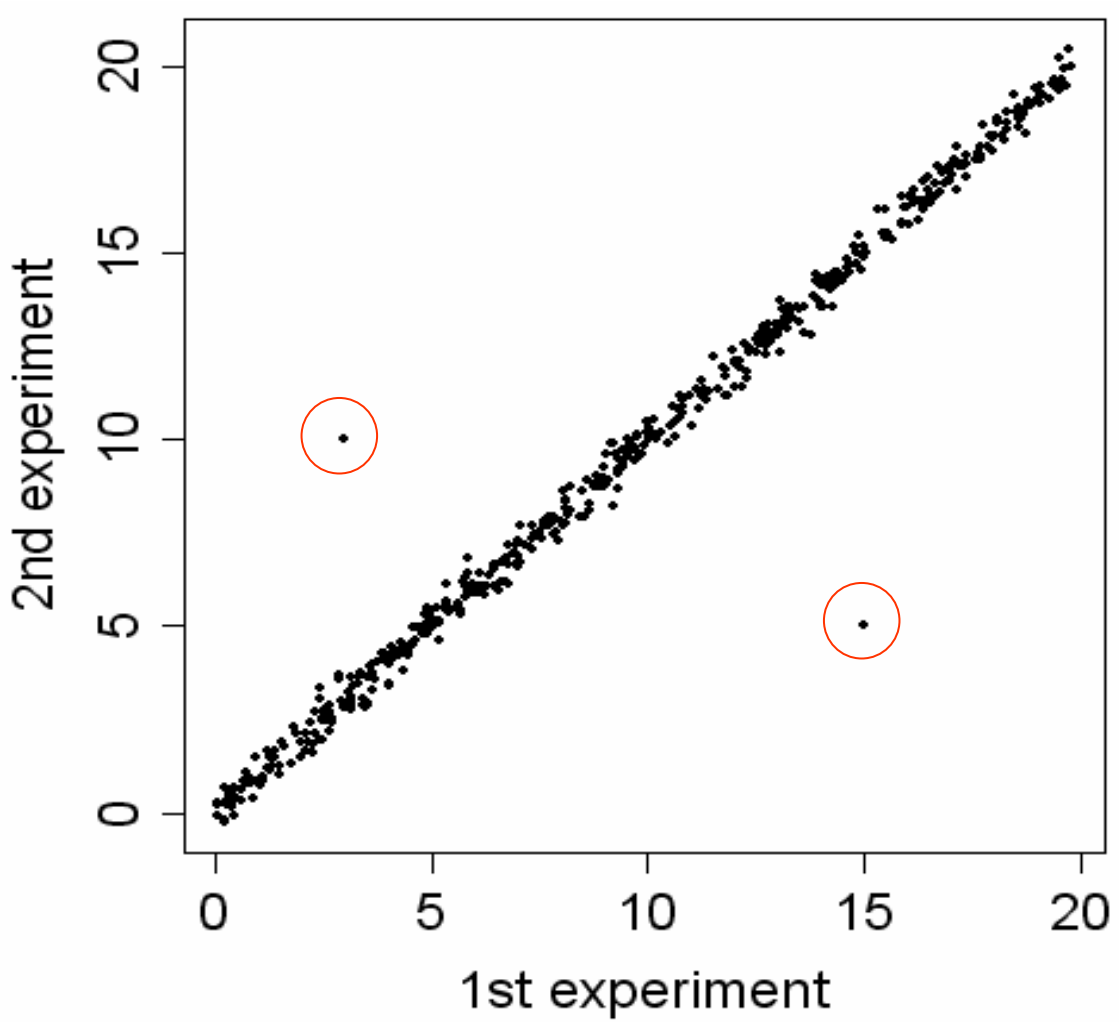


Scatter Plots

- Suppose a gene G has an expression level of e_1 in the 1st experiment and that of e_2 in the 2nd experiment, the point representing G will be plotted at coordinates (e_1, e_2) in the scatter plot.

Note: Genes with similar expression levels in two experiments will appear around the first diagonal of the coordinate system.

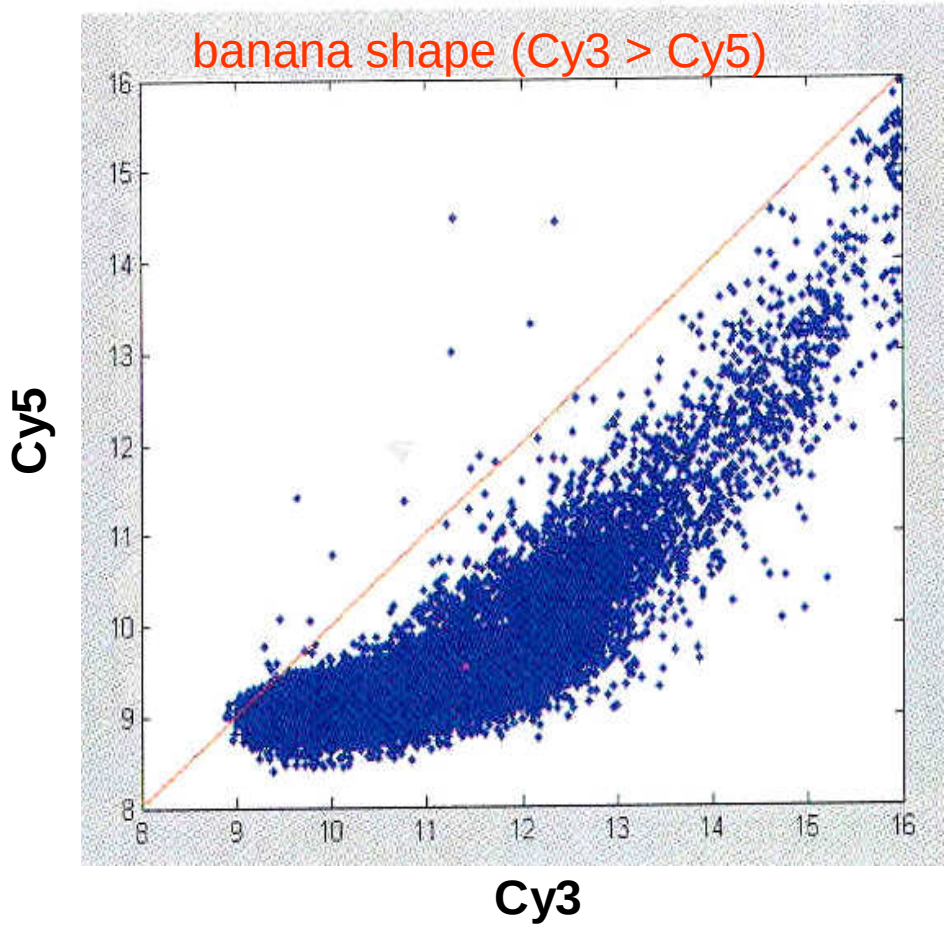
Scatter Plots



Scatter Plots

- Scatter plots allow us to observe certain important features of the data:

Example: Dye swap -- the banana shaped blob indicates nonlinear dye effect.



Scatter Plots v.s. MA Plots

- The MA plot (or ratio-intensity plot) is a variant of the scatter plot. It is commonly used for two-channel cDNA array data.
- Let

  $M = \log(y) - \log(x) = \log(y/x)$

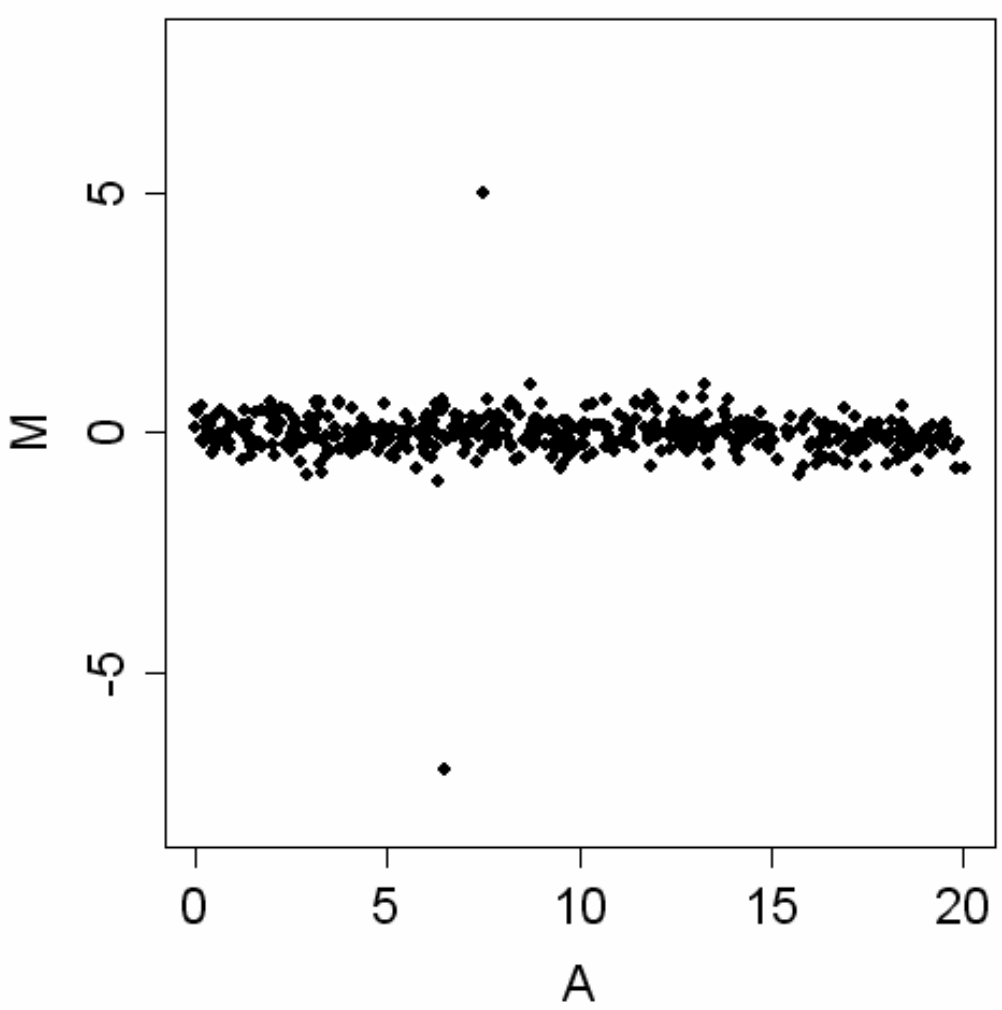
  $A = (\log(y) + \log(x))/2$

  The MA plot is the scatter plot of M against A.

Scatter Plots v.s. MA Plots

M = log(y) – log(x)

A = (log(y) + log(x))/2



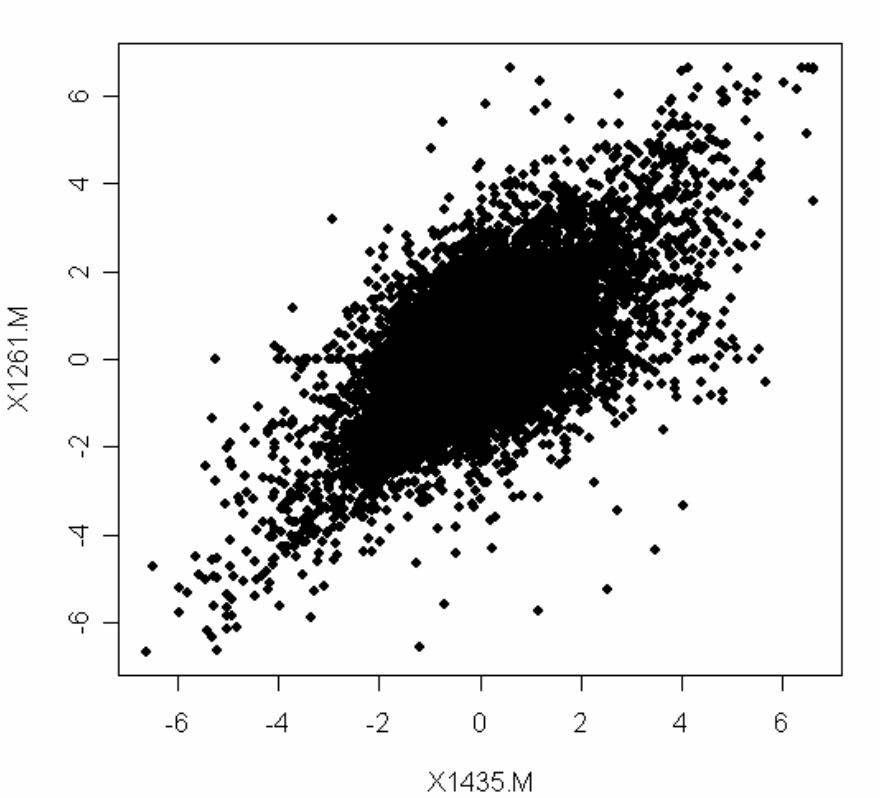
Scatter Plots v.s. MA Plots

- In MA plot, genes with similar expression levels in two experiments will appear around the horizontal line y = 0.
- Points off the horizontal line y = 0 indicate the values measured on one of the two channels to be higher than the values measured on the other channel.

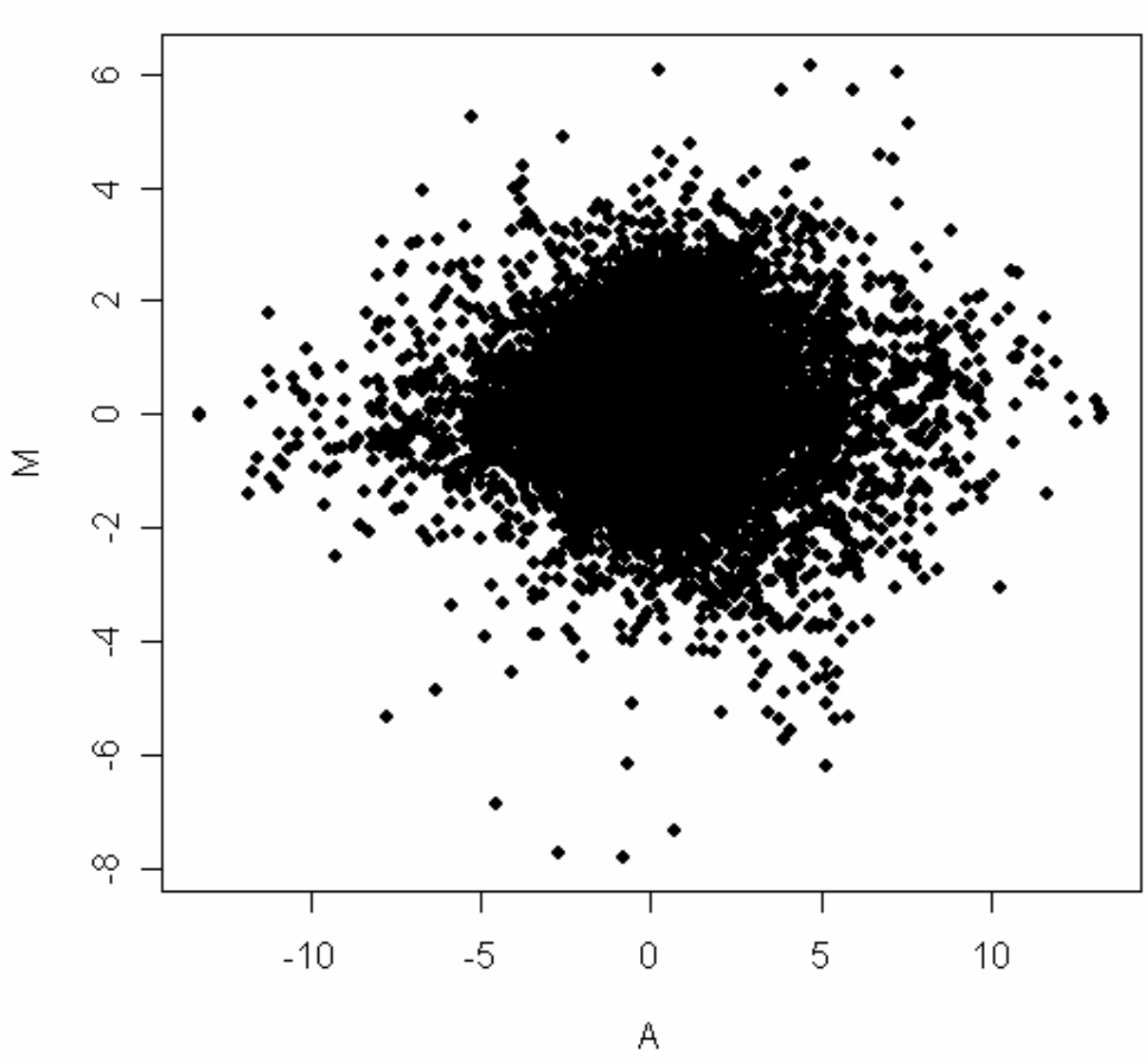
R: Scatter Plots

- Scatter plots of y against x:

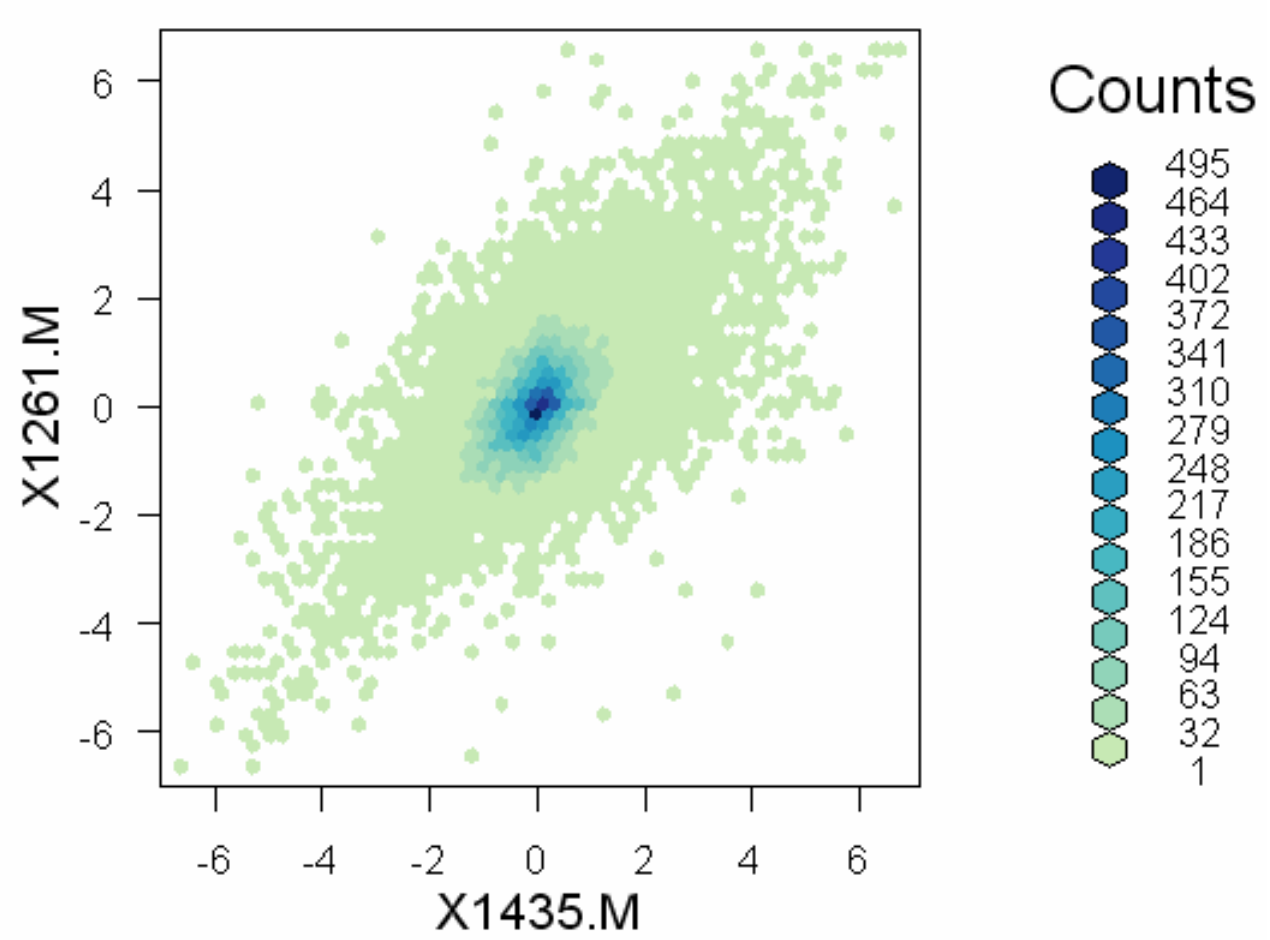
```
> bcdata = read.csv("BreastCancer_ERp.csv")
> x = bcdata[,3:4]
> plot(x,pch=16)
```



```
> y = as.matrix(x) %*% cbind(A=c(1,1), M=c(-1,1))
> plot(y,pch=16)
```



```
> library(RColorBrewer)
> hb=hexbin(x,xbins=50)
> plot(hb,colramp=colorRampPalette(brewer.pal(9,"YlGnBu")[-c(1:2)]))
```



RColorBrewer package:RColorBrewer R Documentation

ColorBrewer palettes

Description:

Creates nice looking color palettes especially for thematic maps

Usage:

```
brewer.pal(n, name)
display.brewer.pal(n, name)
display.brewer.all(n=NULL, type="all", select=NULL, exact.n=TRUE)
```

Arguments:

n	Number of different colors in the palette, minimum 3, maximum depending on palette
name	A palette name from the lists below

Palettes names:
Blues BuGn BuPu GnBu Greens Greys Oranges OrRd PuBu PuBuGn PuRd Purples RdPu Reds YlGn YlGnBu YlOrBr YlOrRd

colorRamp package:grDevices R Documentation

Color interpolation

Description:

These functions return functions that interpolate a set of given colors to create new color palettes (like 'topo.colors') and color ramps, functions that map the interval [0, 1] to colors (like 'grey').

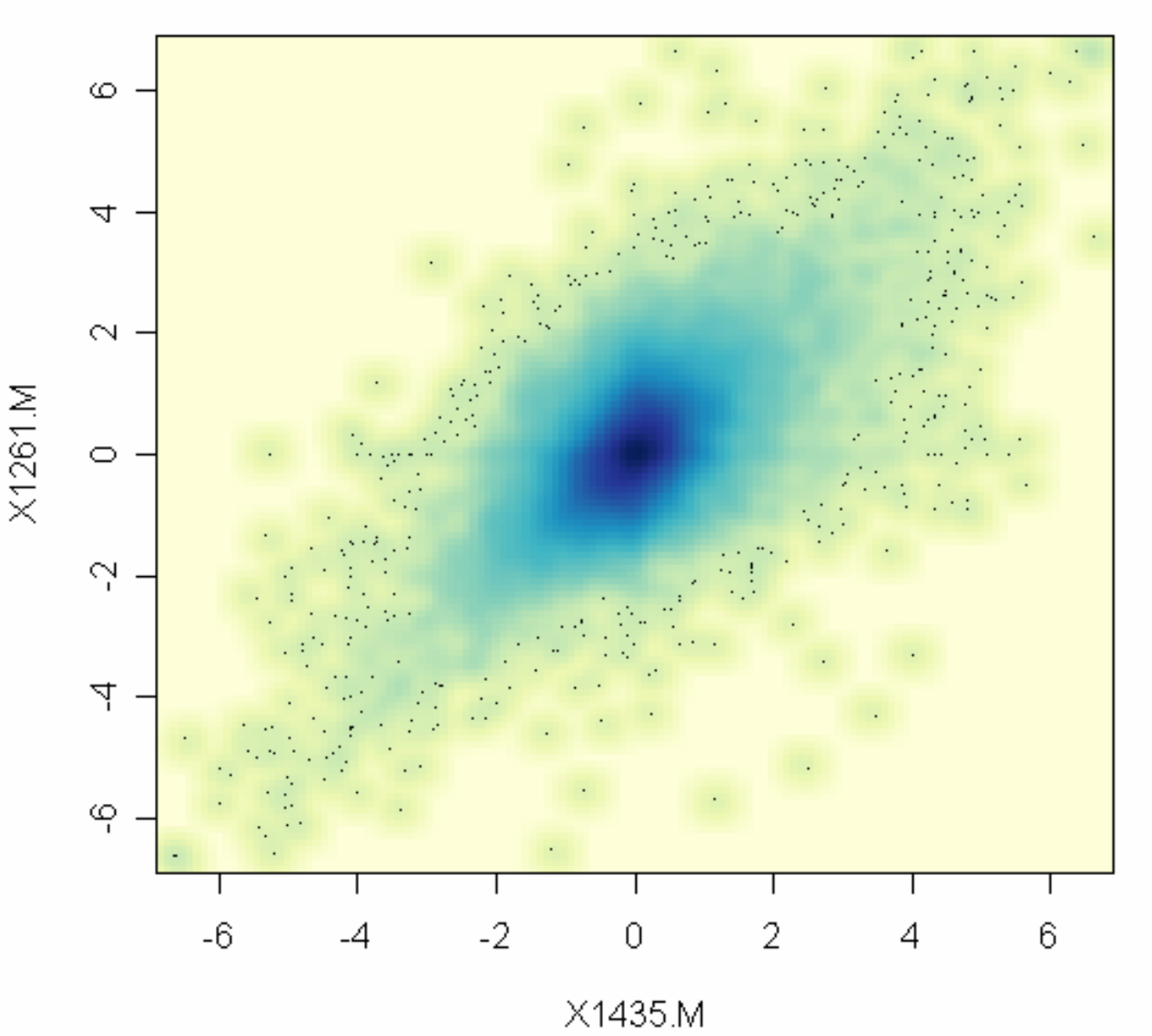
Usage:

```
colorRamp(colors, bias = 1, space = c("rgb", "Lab"),
          interpolate = c("linear", "spline"))
colorRampPalette(colors, ...)
```

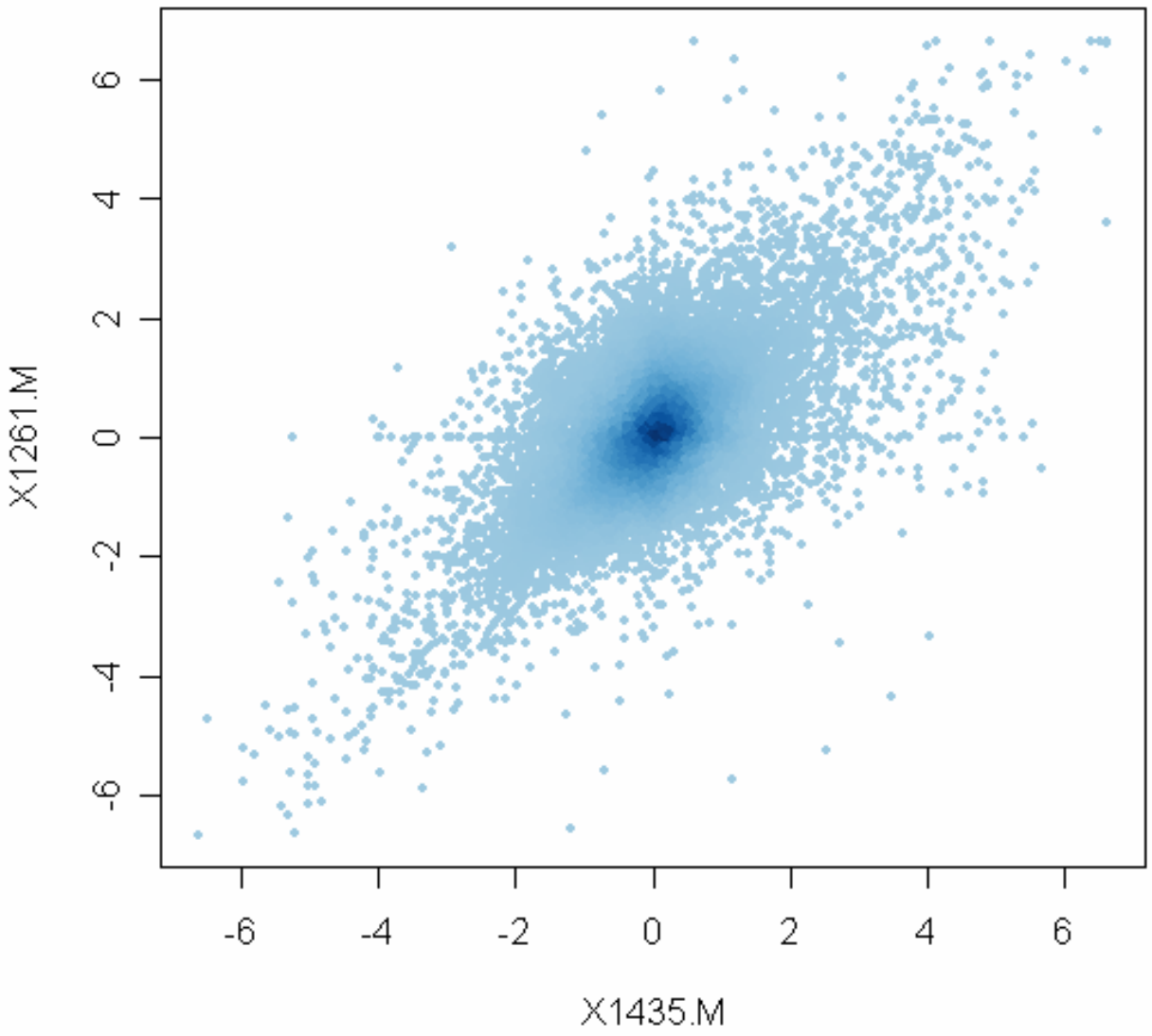
Arguments:

colors	Colors to interpolate

```
> library(geneplotter)
> library(prada)
> smoothScatter(x,nrpoints=500,colramp=colorRampPalette(brewer.pal(9,"YlGnBu")))
```



```
> plot(x,col=densCols(x),pch=20)
```

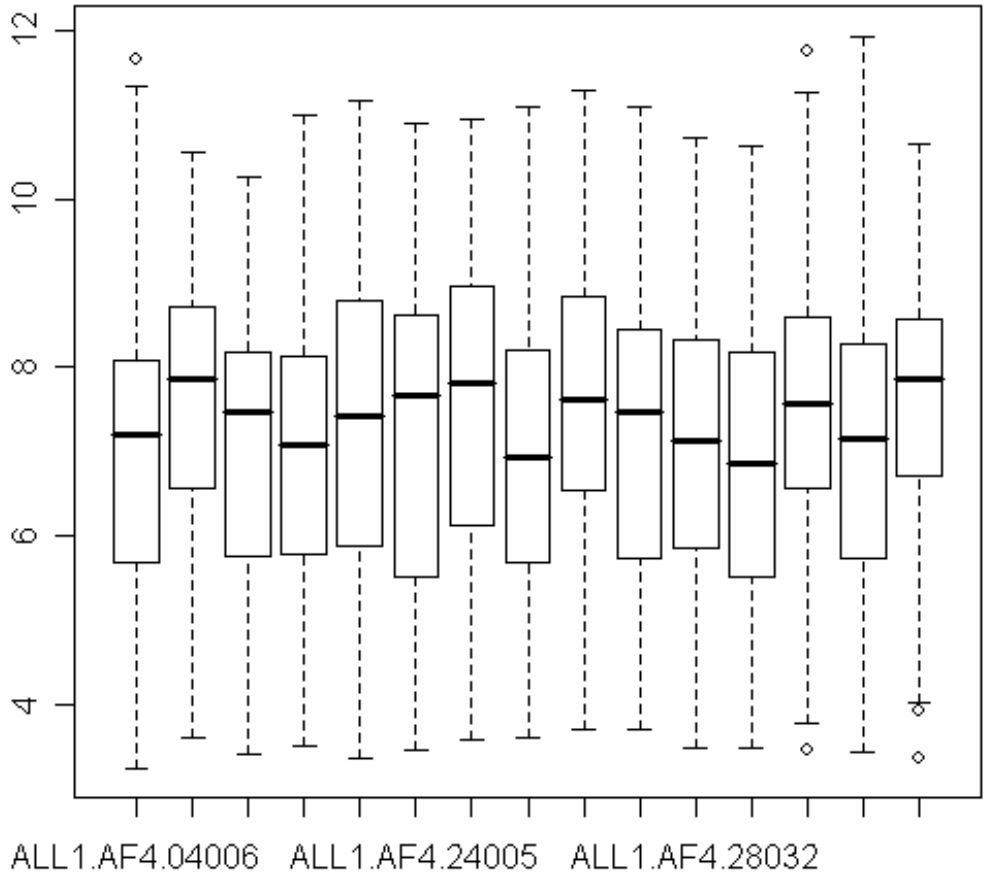


Data Visualization

- Useful visualization tools:
  - Showing feature of the expression level of one particular gene (sample): histogram, box plot
  - Comparing expression levels of two genes (samples): side-by-side box plot, scatter plot, MA plot
  - Presenting the similarity among multiple genes (samples): side-by-side box plot, heatmap

Side-by-side Box Plot

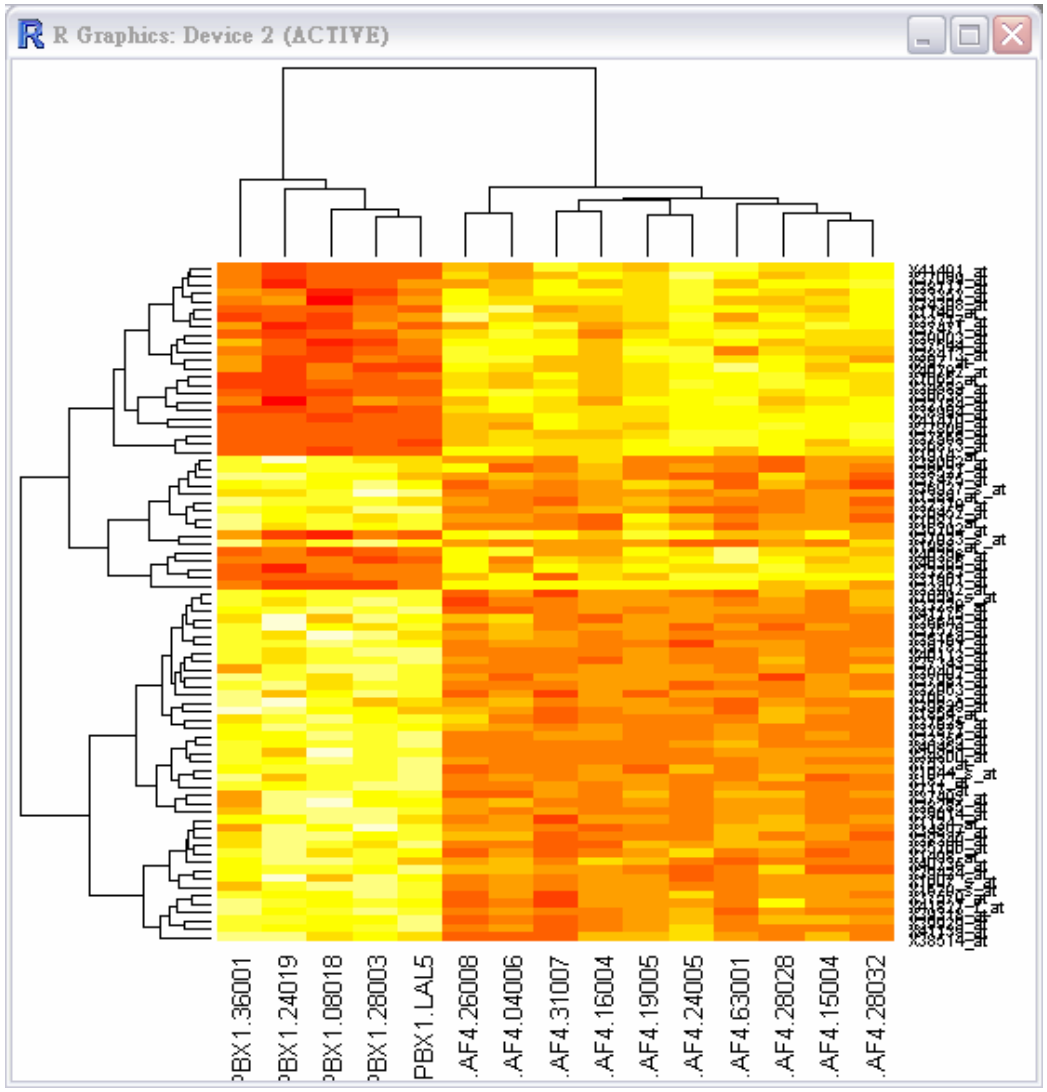
```
ALLhm = read.csv("ALL_hmdemo.csv",row.names="Genes")
boxplot(ALLhm)
```



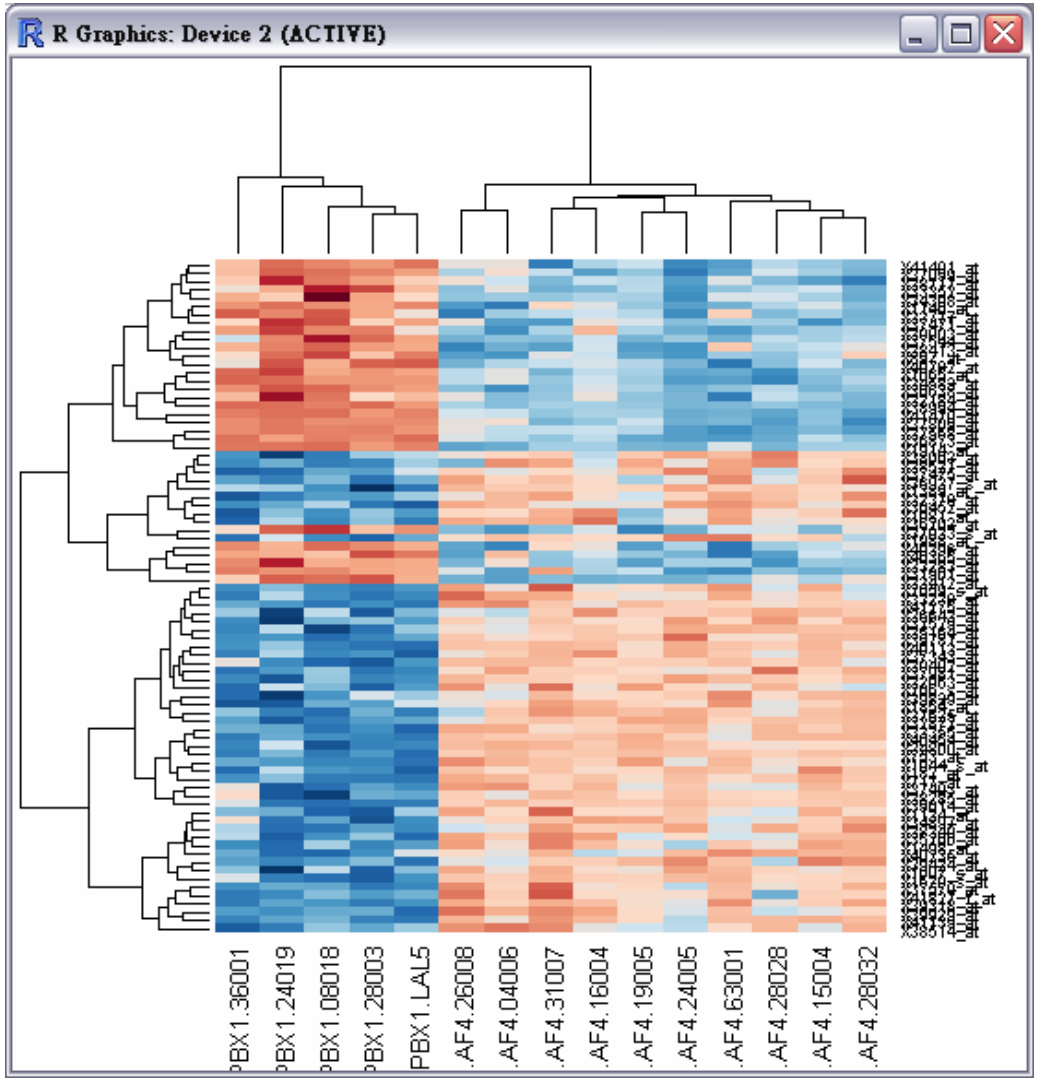
Heatmaps

- A heatmap is a two-dimensional, rectangular, colored grid. It displays data that themselves come in the form of a rectangular matrix:
  - The color of each rectangle is determined by the value of the corresponding entry in the matrix.
  - The rows and columns of the matrix are rearranged independently so that similar rows and columns are placed next to each other, respectively.

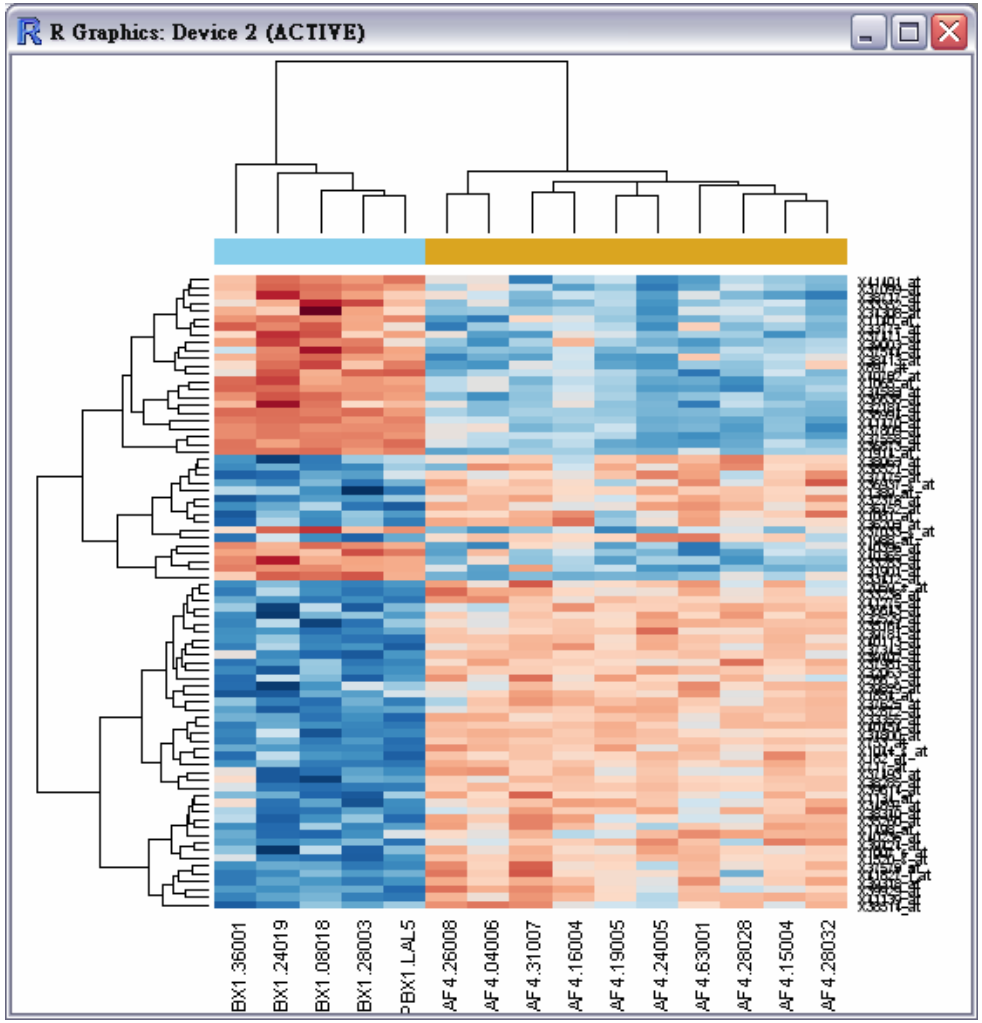
```
> ALLhm = read.csv("ALL_hmdemo.csv",row.names="Genes")
> heatmap(as.matrix(ALLhm))
```



```
> hmcol = colorRampPalette(brewer.pal(10,"RdBu"))(256)
> heatmap(as.matrix(ALLhm),col=hmcol)
```



```
> spcol = ifelse(substr(names(ALLhm),1,8) == "ALL1.AF4" ,
+                                                  "goldenrod","skyblue")
> hmcol = colorRampPalette(brewer.pal(10,"RdBu"))(256)
> heatmap(as.matrix(ALLhm),col=hmcol)
```



Measure of Distance

- To group entities that are similar, we need to define a measure of similarity, usually called distance metric.

- A distance measure must satisfy:
  - Symmetry: $d(x, y) = d(y, x)$
  - Positivity: $d(x, y) \geq 0$
  - Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$
  - Definiteness: $d(x, y) = 0$ if and only $x = y$

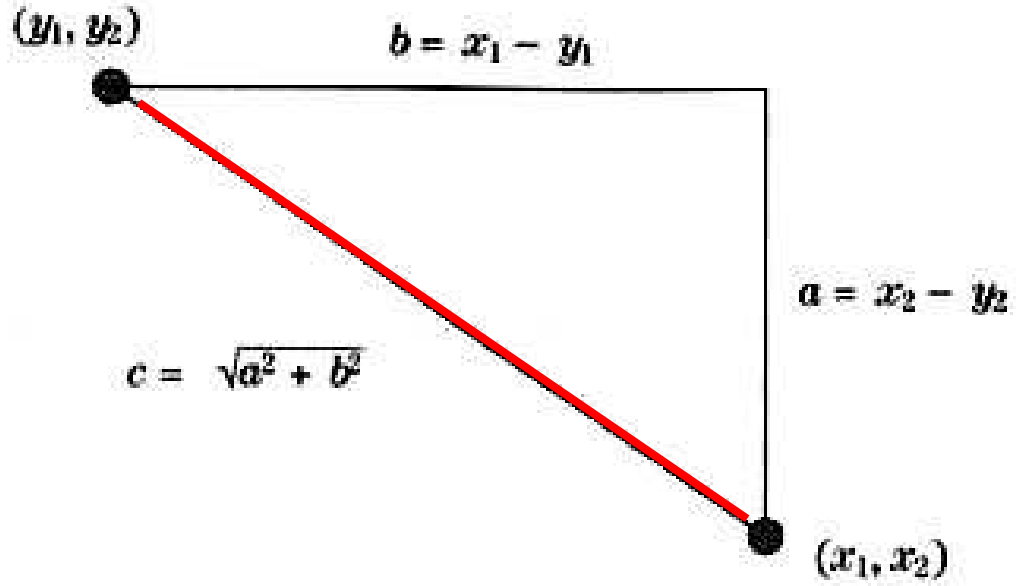
Measure of Distance

- We wish to define the distance between two objects
- Distance metric between points:
  - Euclidean distance (EUC)
  - Manhattan distance (MAN)
  - Pearson sample correlation (COR)
  - Angle distance (EISEN – considered by Eisen et al., 1998.)
  - Spearman sample correlation (SPEAR)
  - Kandall's τ sample correlation (TAU)
  - Mahalanobis distance
- Distance metric between distributions:
  - Kullback-Leibler information
  - Hamming's mutual information

Euclidean Distance

- The Euclidean distance:

$$d_E(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



Euclidean Distance

- Example: the distance from *O* (0,0) to *A* (3,4)

$$d_E(O,A) = \sqrt{3^2 + 4^2} = \sqrt{25} = 5$$

- A change of one unit in one of the coordinates determined a change of 13% respect to the truth.

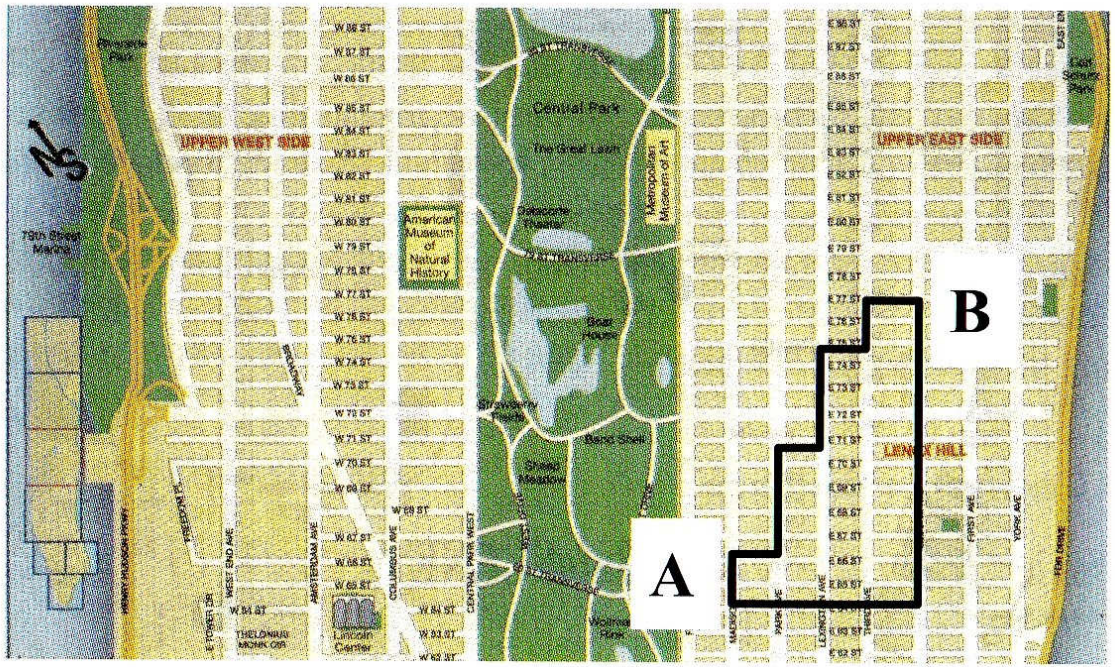
$$d_E(O,A') = \sqrt{4^2 + 4^2} = \sqrt{32} = 5.65$$

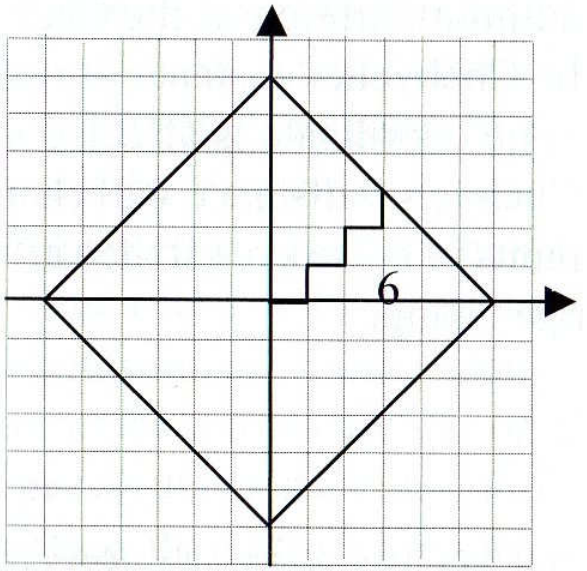
$$\frac{5.65}{5} = 1.13$$

Manhattan Distance

- The Manhattan distance:

$$d_M(\mathrm{x},\mathrm{y}) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n| = \sum_{i=1}^{n} |x_i - y_i|$$





Manhattan Distance

- Example: the distance from *O* (0,0) to *A* (3,4)

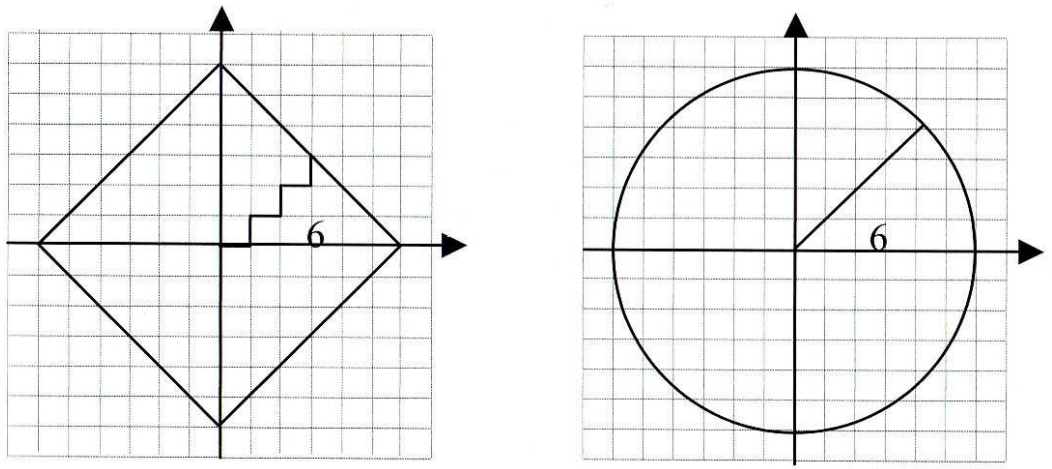
$$d_M(O, A) = 3 + 4 = 7$$

- A change of one unit in one of the coordinates determined a change of 14% respect to the truth.

$$d_M(O, A') = 4 + 4 = 8$$

$$\frac{8}{7} = 1.14$$

Euclidean vs Manhattan Distances



- Manhattan distance yield a larger numerical value for the same relative position of points.
- Manhattan distance slightly emphasizes the outlier of the dataset; a outlier will appear a bit further away.

Pearson Correlation

- The Pearson correlation focuses on whether the two points change in the same way:

$$d_R(\mathrm{X},\mathrm{y}) = 1 - r_{xy}$$

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x}\sqrt{s_y}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\because -1 \le r_{xy} \le 1 \quad \therefore 0 \le d_R(\mathbf{x},\mathbf{y}) \le 2$$

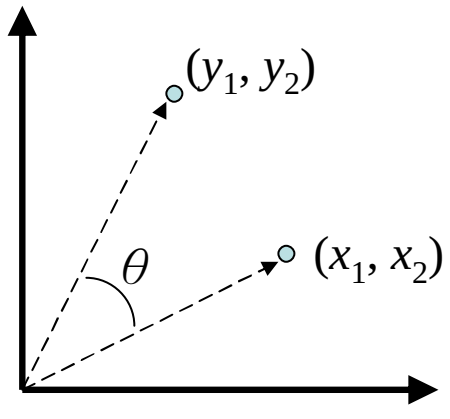
Note: the Pearson correlation is affected greatly if the measurement along a particular dimension are very different!

Angle Distance

- The angle distance:

$$d_\alpha(\mathbf{x},\mathbf{y}) = \cos^{-1}(\theta) = \cos^{-1}\left(\frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}\right)$$

where $\mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{n} x_i y_i$, $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$



Note: the angle distance will be the same if the point moves alone the line going through the original position and the origin (scaling).

Other Correlations

- The Spearman correlation is the correlation of rank statistics.
- The Kendall's τ :

$$d_{tau}(x,y) = 1 - \tau(x,y)$$

$$\tau(x,y) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} C_{x_{ij}} C_{y_{ij}}}{n(n-1)},$$

where $C_{x_{ij}} = sign(x_i - x_j)$ and $C_{y_{ij}} = sign(y_i - y_j)$

Mahalanobis Distance

- The Mahalanobis distance:

$$d_{Ml}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})}$$

- If the space warping matrix *S* is taken to be the identity matrix, the Mahalanobis distance reduces to the classical Euclidean distance.

$$d_{Ml}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T (\mathbf{x}-\mathbf{y})} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

When to Use What Distance?

- Normalization process such as location-scale normalization maybe necessary before calculating the distance, especially when different types of variables need to be mixed together.
- Surely, different distance measure has difference emphasis. Here we summarize all measurements introduced.

A Comparison of Various Distances

1. Euclidean distance:

   describes the geometric distance; the most commonly used measure.

2. Manhattan:

   slightly emphasizes the outlier of the dataset than Euclidean distance.

3. Angle between vectors:

   takes into consideration only the angles, not the magnitude. For example, (1,1) and (100,100) will have the distance equal to 0. $\Rightarrow$ same cluster!

4. Mahalanobis:

   can warp the high dimensional space in a convenient way

A Comparison of Various Distances

5. Correlation-based distance: Correlation-based distances are adversely affected by outliers and then the non-parametric versions (SPEAR or TAU) are preferred.