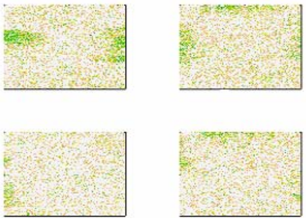
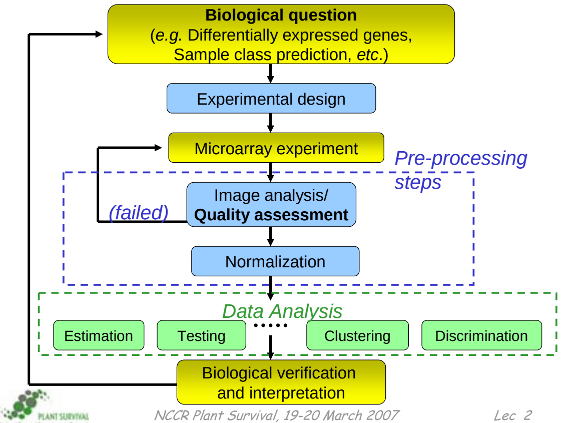
Statistics for Affymetrix GeneChips

Quality assessment and exploratory data analysis for Affymetrix experiments



http://www.isrec.isb-sib.ch/~darlene/NCCR-PS/



Biological question
(*e.g.* Differentially expressed genes, Sample class prediction, *etc.*)

Experimental design

Microarray experiment

Pre-processing steps

(failed)

Image analysis/ **Quality assessment**

Normalization

Data Analysis

Estimation

Testing

....

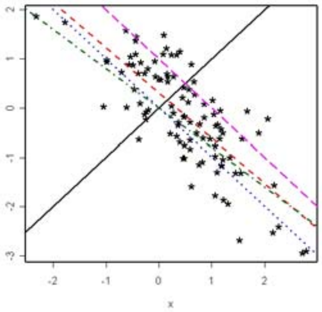
Clustering

Discrimination

Biological verification and interpretation

Simple linear modeling: which line?

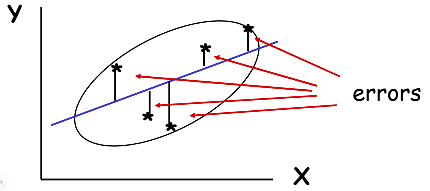
- There are *many possible lines* that could be drawn through the cloud of points in the scatterplot …
- How to choose?



Least Squares

- *Q*: Where does the regression equation come from?

  A: It is the line that is 'best' in the sense that it *minimizes* the sum of the *squared* errors (*residuals*) in the vertical (*Y*) direction



What is robustness?

- The term *robustness* is used to mean several possible things:
  - Lack of sensitivity to *distributional assumptions* (especially normality)
  - Lack of sensitivity to *outliers*
  - Small sets of the data *don't have a strong influence*

Do we need robust methods?

- Tukey (1962):

  "A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians."

Measures of center

- Mean – not robust
- Trimmed mean (‘Olympic judging’) - robust
  - obtained by discarding some percentage of the lowest and the highest values, then computing the mean of the remaining values
  - For example, for an 80% trimmed mean, you would throw out the highest 10% and lowest 10%
- Median – highly robust

Measures of spread

- SD (or variance) – not robust
- Range – not robust
- IQR (Interquartile range) – robust
- MAD (Median Absolute Deviation from the median) – robust

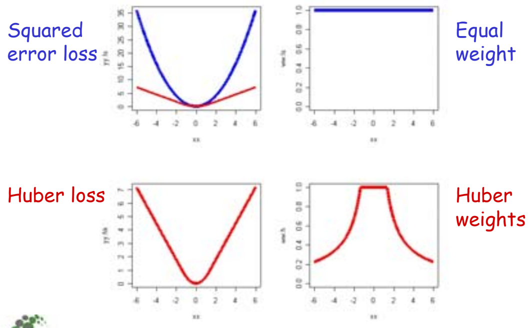
Robust Regression

- Idea: *downweight* observations that produce large residuals
- RMA carries out the robust fit with *median polish*
- There are also other ways of carrying out the fitting procedure, the technical details differ

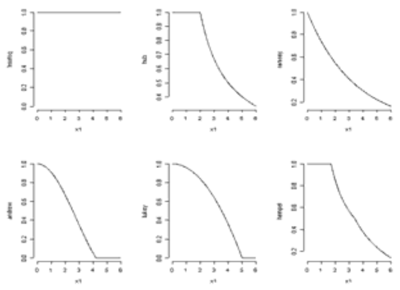
Loss, weight functions

- Least squares: ‘lose’ square of vertical error
- Here, squared error = *loss function*
- Each observation has *equal weight*
- Problem: *outliers* can have strong effect on estimates (slope, intercept of line; model parameters more generally)
- Solution: could use *other loss/weight functions*

Examples of Loss, Weight Functions



More weight functions



Robust regression in microarray analysis

- There are many ways that robust regression is used in analysis of microarray data
- We use it in two ways:
  - for *quantifying gene expression* measured with Affymetrix GeneChips (RMA)
  - for *assessing quality* of Affymetrix GeneChip gene expression measures

Using residuals from the fitting

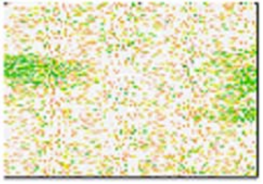
- The difference between the observed signal for a probe and its fitted value based on a model is called the *residual:*
  - *Residual = observed - predicted*
- Many types of chip problems will be reflected by *inflated residuals* from the fits to the probe + chip effect models
- *Summarizing the residuals* on a chip can provide good discrimination among chips producing data of varying quality

NUSE

- *NUSE* = 'Normalized Unscaled SE' – estimate SE(expression estimates) and summarize at the chip level
- Each chip will have a NUSE for each probe, which can be summarized by the *median*
- This provides one useful summary of the residuals, and can be used to judge quality relative to other chips
- Median NUSE is a number that fluctuates around the value 1 – 'high' values (> 1.05) indicate 'worse' (unusual) chips

Chip weight pseudo-images

- Image indicates the (robust regression) *weight* associated with the probe
- Areas of low weight are greener, high weights are light gray
- 'More color' ⇔ 'worse chip'

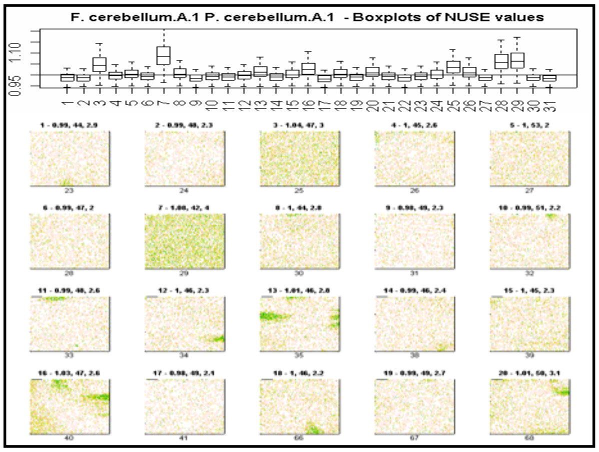


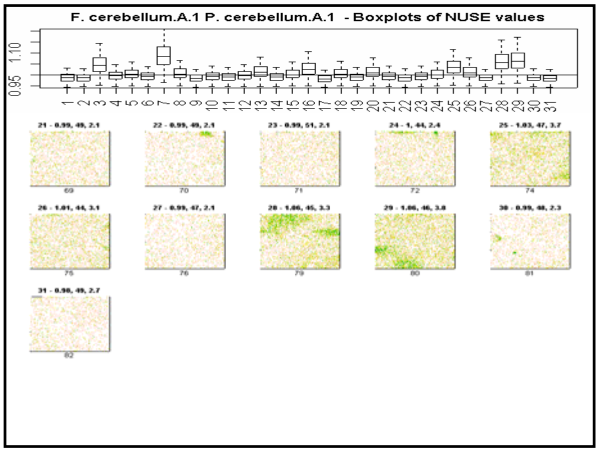
Example: HD

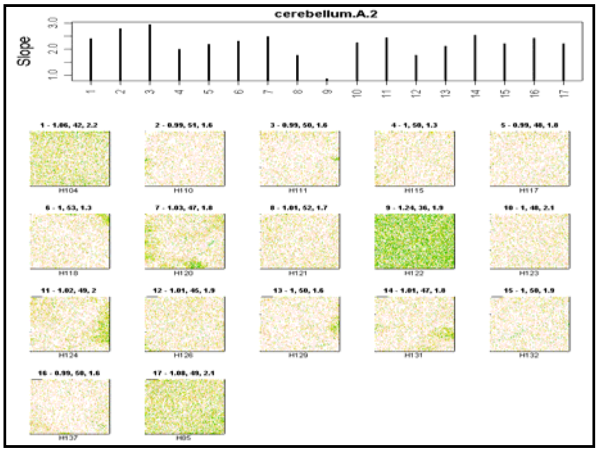
- About 70 individuals, U133A,B chips on each of 3 tissues
- Fitted RMA models
- Displays: NUSE plot, chip pseudo-image of residual weights

Title=ChipNo – Median NUSE, %P, SF

Subtitle=ChipId







Conclusions

- Model-based quality assessment appears to show good sensitivity to chip problems
- Provides useful basis for chip quality, inclusion/exclusion decisions

Software for Microarray Analysis

- Again, we will be using R packages from the BioConductor project
- http://www.BioConductor.org/
  - `affy`
  - `affyPLM`
  - `limma`
  - `affylmGUI`

Exploratory data analysis

- Signal intensity:
  - Pseudo-images
  - histograms
  - Boxplots
  - Pairwise scatterplots (MA version)
- Pseudo-images of *weights* (and/or residuals)
- Boxplots of *NUSE values*
- Boxplots of normalized signal values (RMA)

EDA with **`affylmGUI`**

- *demo…*