Something about robust regression

Patty Solomon

Department of Applied Mathematics and
Centre for the Molecular Genetics of Development
University of Adelaide

12 July 2002

http://www.maths.adelaide.edu.au/MAG

patty.solomon@adelaide.edu.au

about which I know almost nothing ...

about which I know almost nothing ...

- People *are* using robust regression to analyse data from multiple-slide factorial cDNA microarray experiments.
  - ★ *Because of the heavy-tailed distributions involved, this is a perfectly reasonable approach.*

about which I know almost nothing ...

- People *are* using robust regression to analyse data from multiple-slide factorial cDNA microarray experiments.
  - ⋆ *Because of the heavy-tailed distributions involved, this is a perfectly reasonable approach.*

- The purpose of this talk is to warn plu's not to trust, as a matter of course, the results of reputable, off-the-shelf robust regression methods for analysing cDNA microarray data.

Microarray Analysis Group

- The University of Adelaide
  - Patty Solomon
  - Gary Glonek*
  - Jonas Lloyd (Ph.D.)*
  - Michael Calvert (Grad. Dip.)
  - Adam Kister (Grad. Dip.)

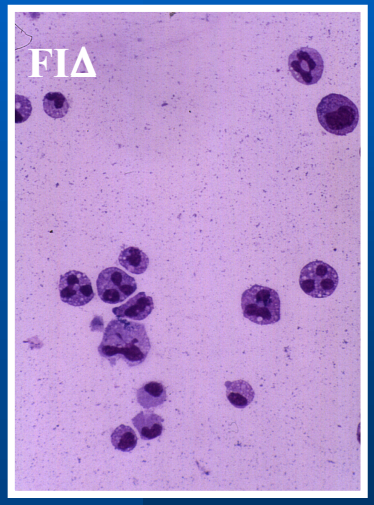
- Hanson Institute/IMVS
  - Anna Tsykin*

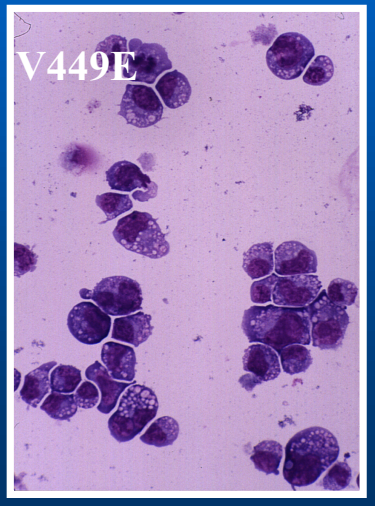
Microarray Analysis Group

- The University of Adelaide
  - Patty Solomon
  - Gary Glonek*
  - Jonas Lloyd (Ph.D.)*
  - Michael Calvert (Grad. Dip.)
  - Adam Kister (Grad. Dip.)
- Hanson Institute/IMVS
  - Anna Tsykin*
- *Acknowledgement: Terry Speed, UC Berkeley and WEHI*

Leukaemic mice project

Biological function of the activated mutants in FDB cells





⇒ Different signals generated by the two mutants

- *Aim: to identify genes that play an important role in receptor signalling and leukaemogenesis in mutant mice.*

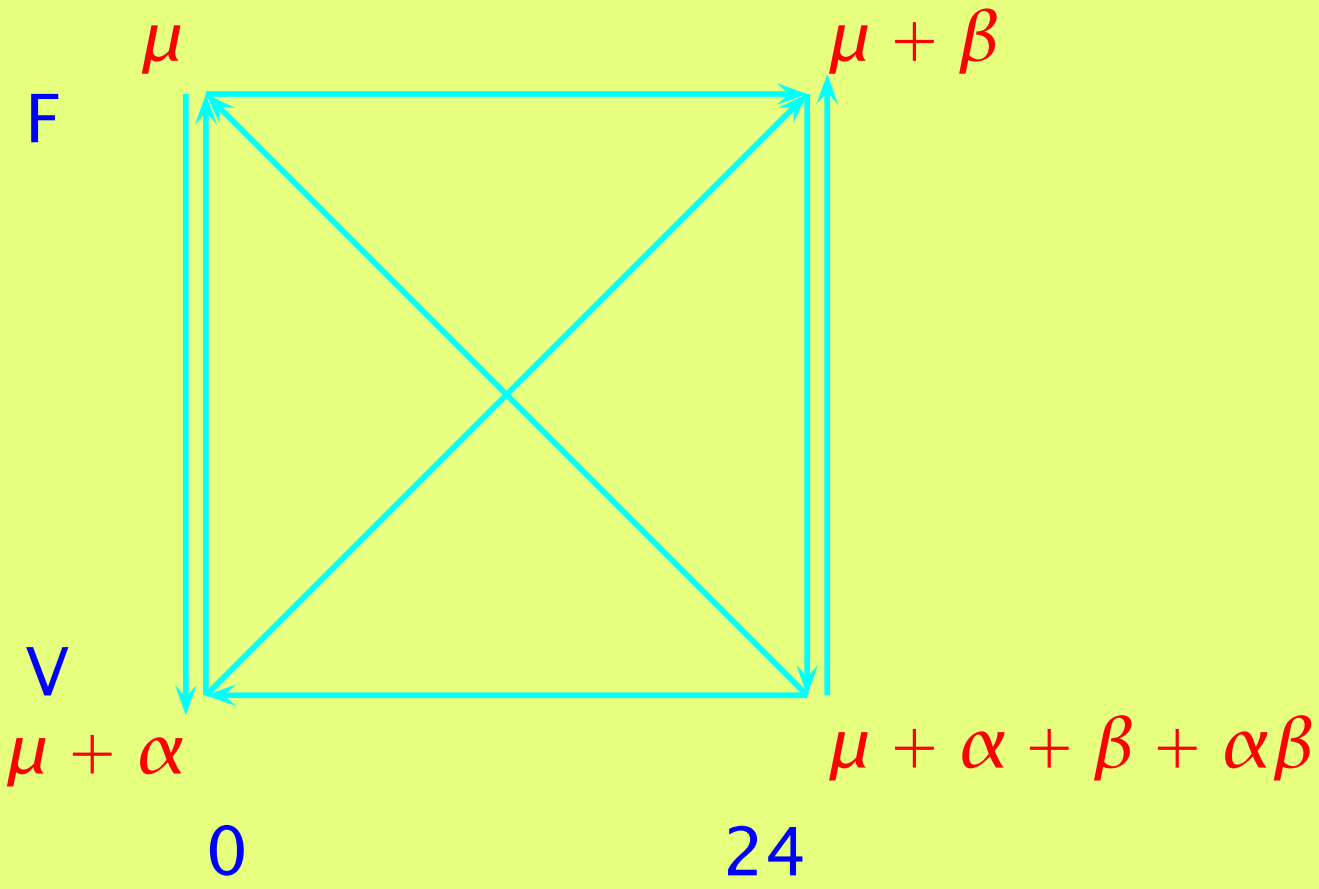
A 2x2 factorial design

- *Hypothesis:* that there is a set of genes induced specifically in response to expression of V449E that results in its leukaemic effects.
  - ⋆ It is anticipated that measuring changes over 24 hours will distinguish genes involved in promoting or blocking differentiation, or that suppress or enhance growth, as genes potentially involved in leukaemia.
- Two cell lines: FIΔ and V449E
  at two times 0 hours and 24 hours
  - ⋆ $\longrightarrow$ 2×2 factorial design of block size 2.

A 2x2 factorial design

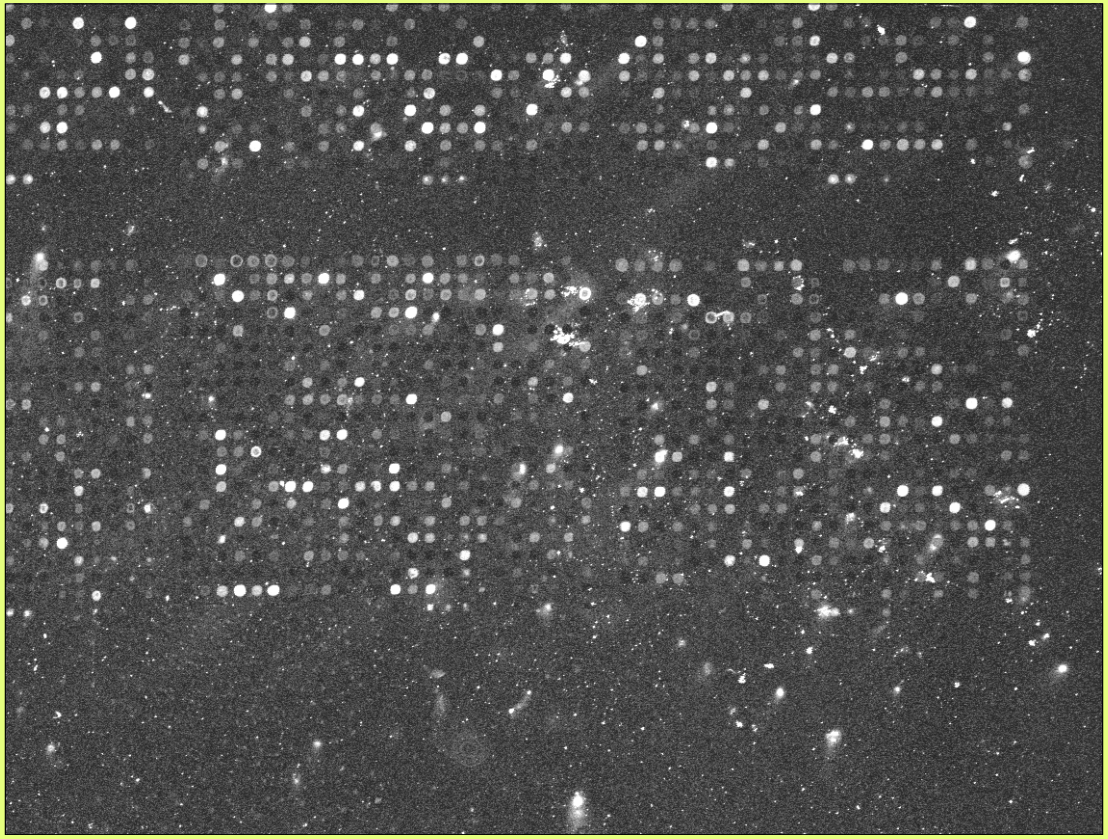
- *Hypothesis:* that there is a set of genes induced specifically in response to expression of V449E that results in its leukaemic effects.
  - ⋆ It is anticipated that measuring changes over 24 hours will distinguish genes involved in promoting or blocking differentiation, or that suppress or enhance growth, as genes potentially involved in leukaemia.
- Two cell lines: FIΔ and V449E
  at two times 0 hours and 24 hours
  - ⋆ $\longrightarrow$ 2×2 factorial design of block size 2.
- *Interaction of primary interest.*
- *Only a few genes are expected to change.*

Design for each gene

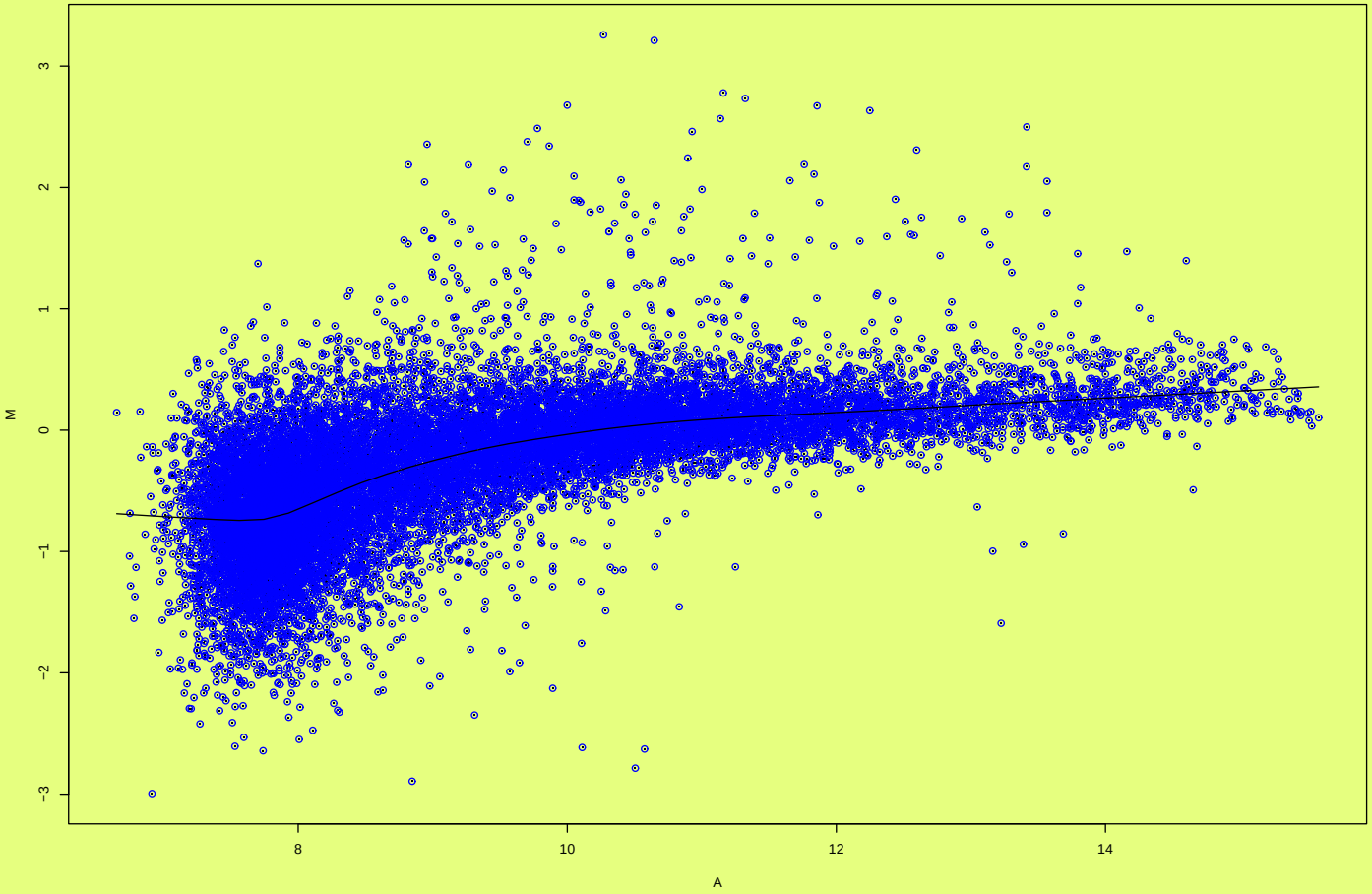


- 6 pairwise comparisons with dye-swaps on cell line comparisons at times 0 and 24 hours.

Precipitation, dust, high background, comet tails, ...

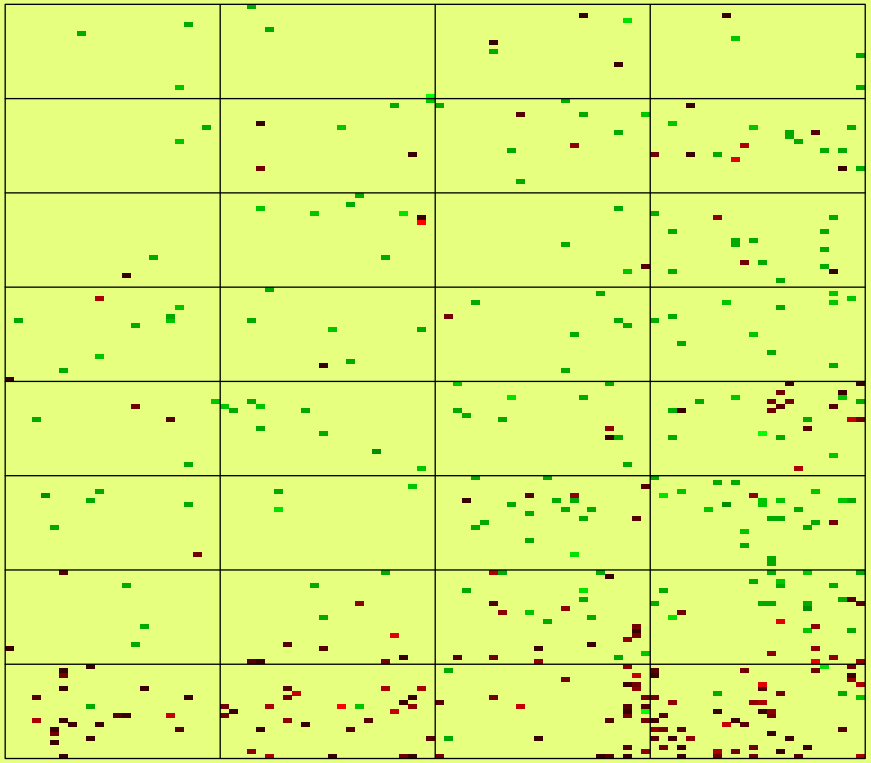


FIΔ versus V449E at 24 hours



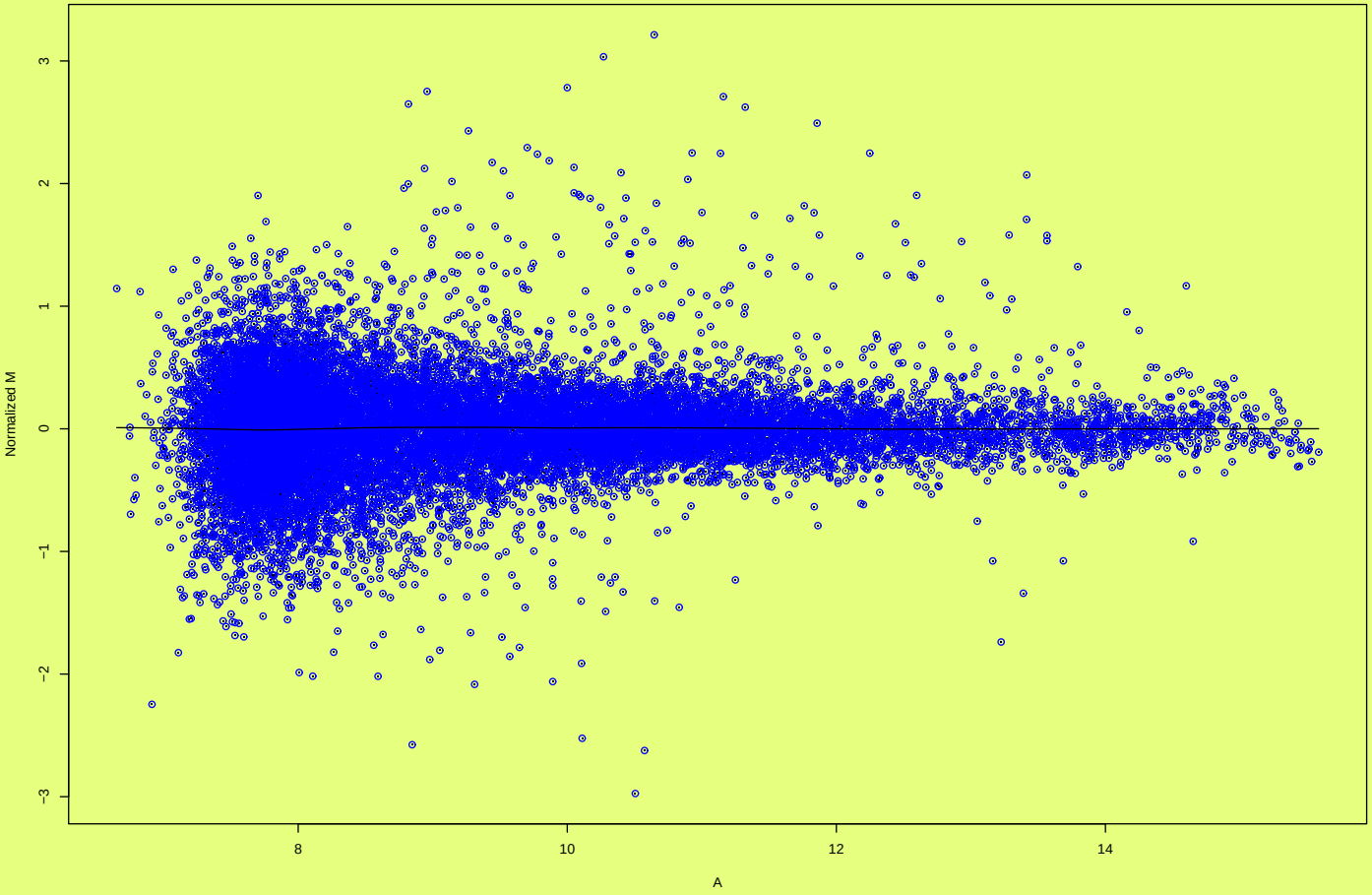
M is intensity dependent

M is spatially dependent



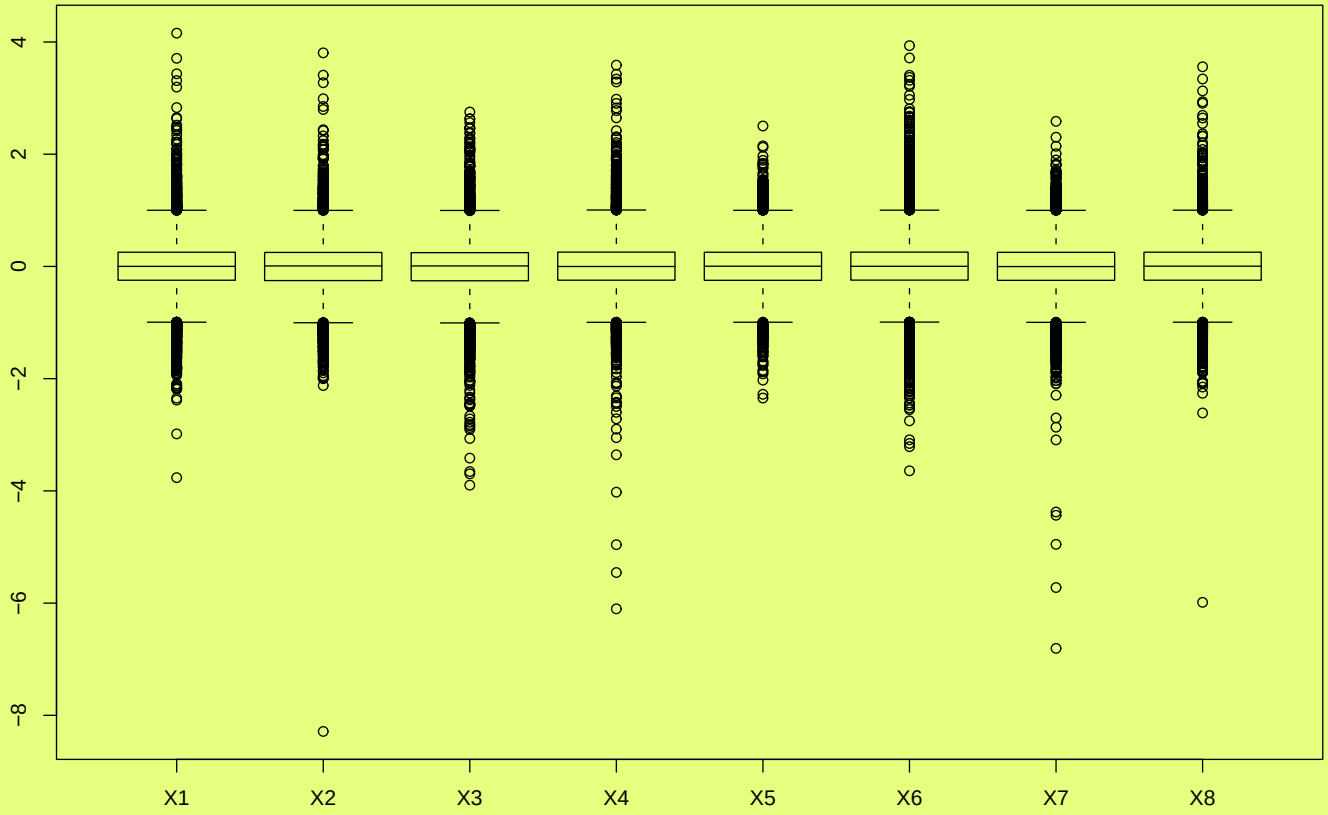
Print-tip effects

After print-tip normalization

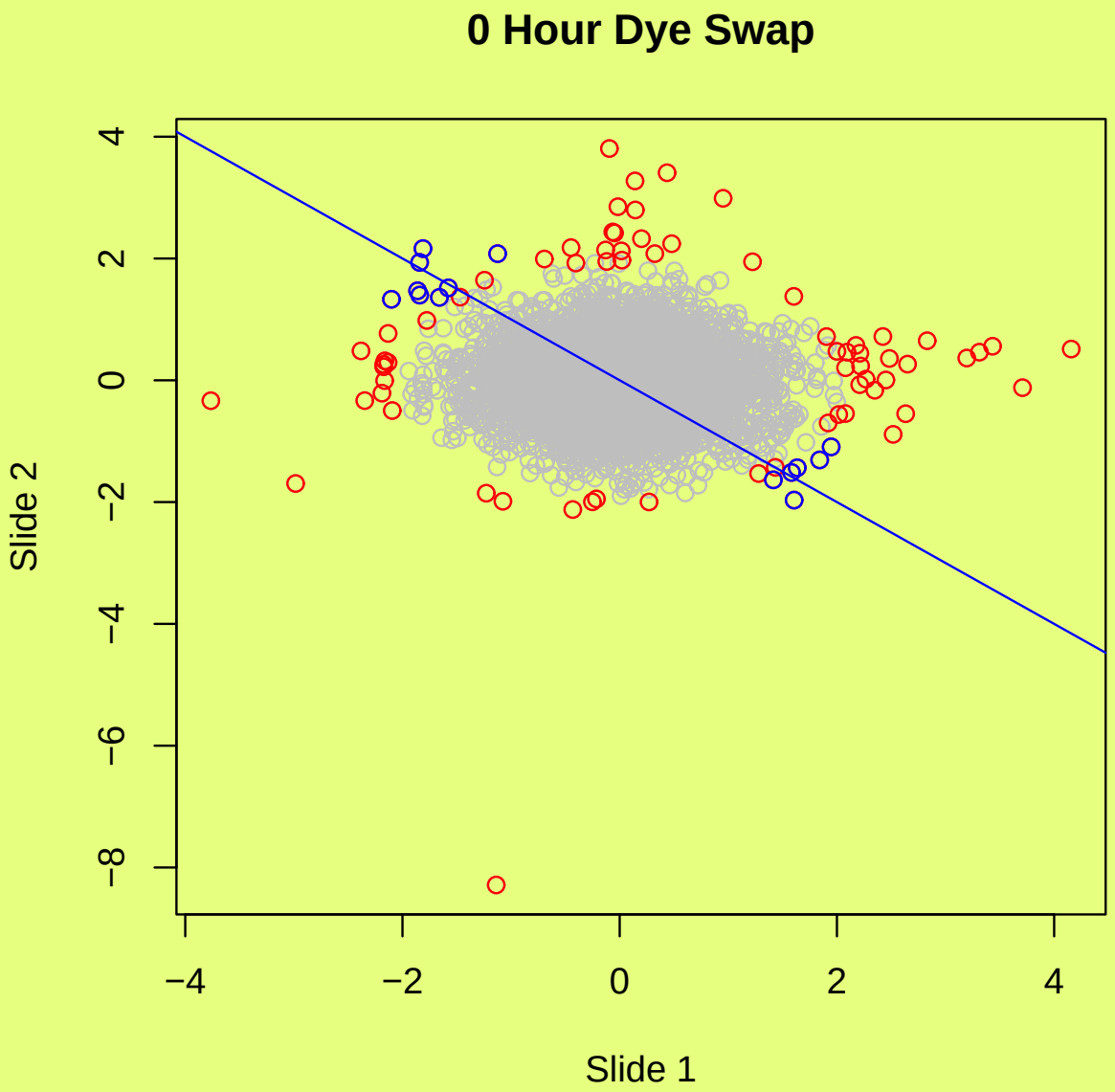


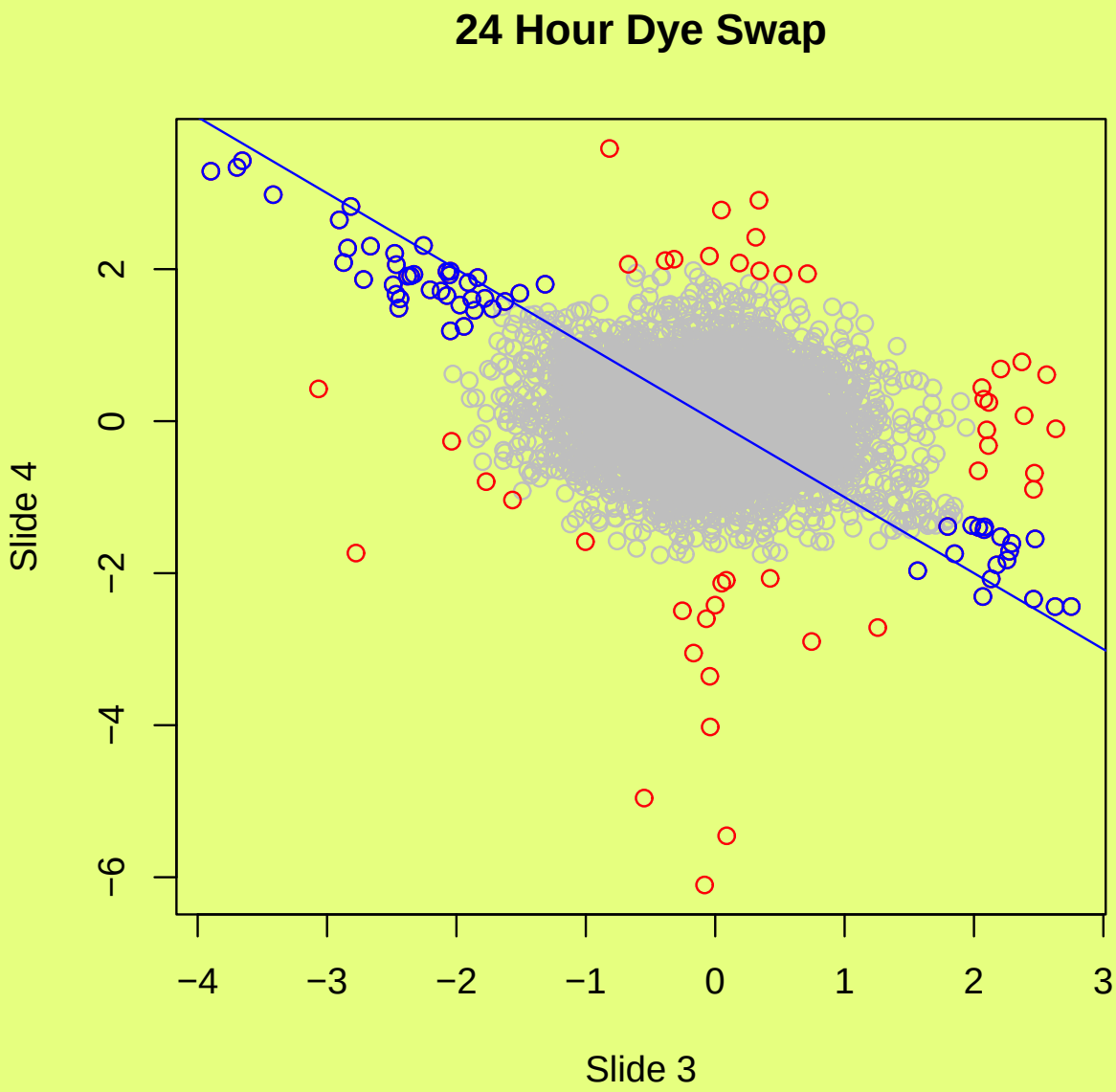
Changes are roughly symmetric about zero.

The 8 arrays after across-slide scale normalization



(Forward)


0 Hour Dye Swap
Slide 2
Slide 1
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4

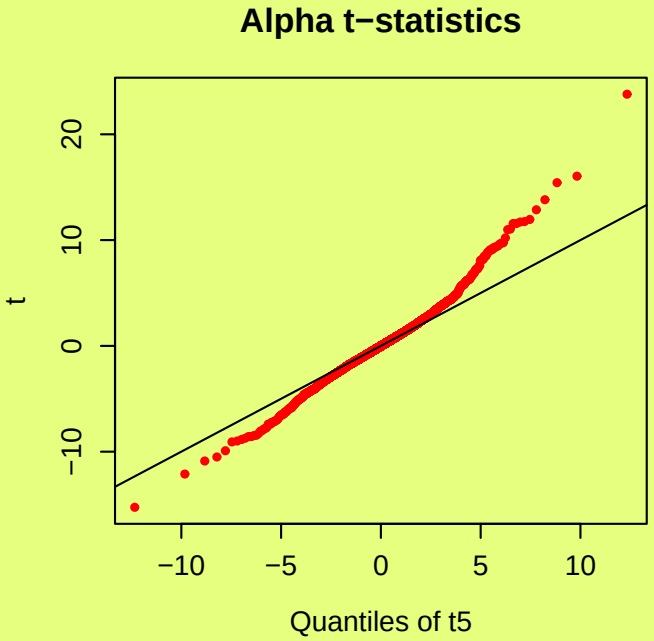

24 Hour Dye Swap
Slide 4
Slide 3
2
0
−2
−4
−6
−4
−3
−2
−1
0
1
2
3

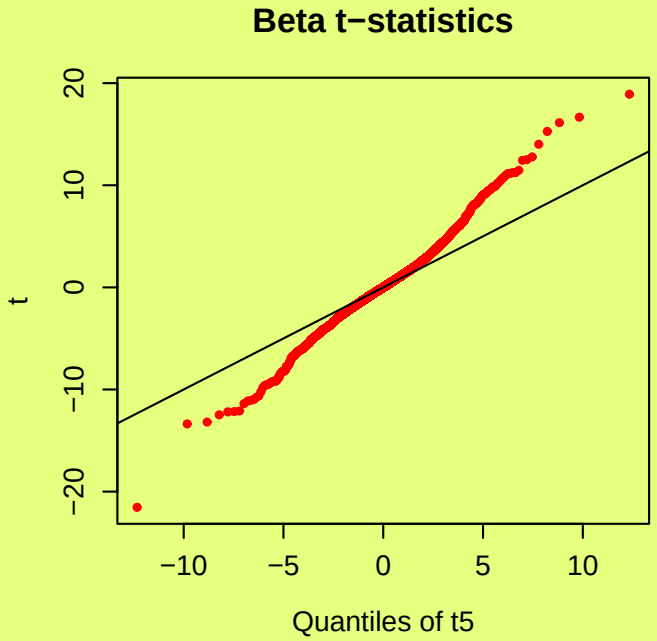
Combine 8 slides using regression

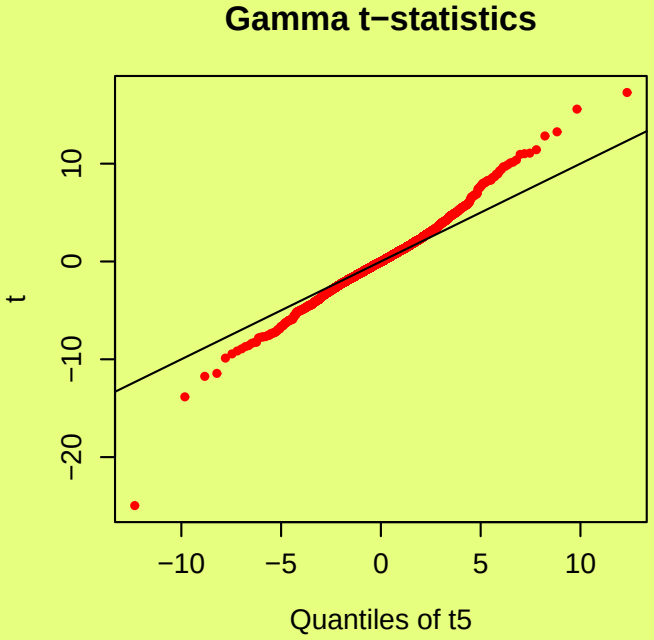
- Define a design matrix X such that $E(M) = X\theta$ where $\theta^T = (\alpha, \beta, \gamma)$. Find the least squares estimator of θ *for each gene*:

$$E\begin{pmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \\ M_8 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 0 \\ -1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$

$$\longrightarrow \hat{\theta} = (X^T X)^{-1} X^T m$$


Alpha t−statistics
t
20
10
0
−10
−10
−5
0
5
10
Quantiles of t5


Beta t−statistics
t
20
10
0
−10
−20
−10
−5
0
5
10
Quantiles of t5


Gamma t−statistics
t
10
0
−10
−20
−10
−5
0
5
10
Quantiles of t5

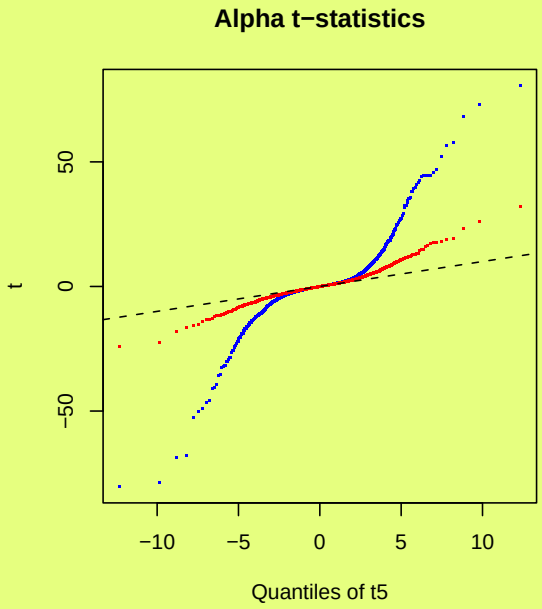
Robust regression

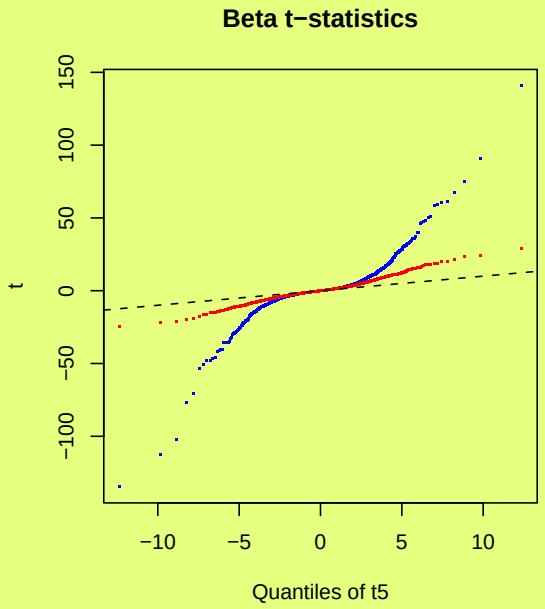
- A widely used ‘off-the-shelf’ method is *MM-estimation*[1]

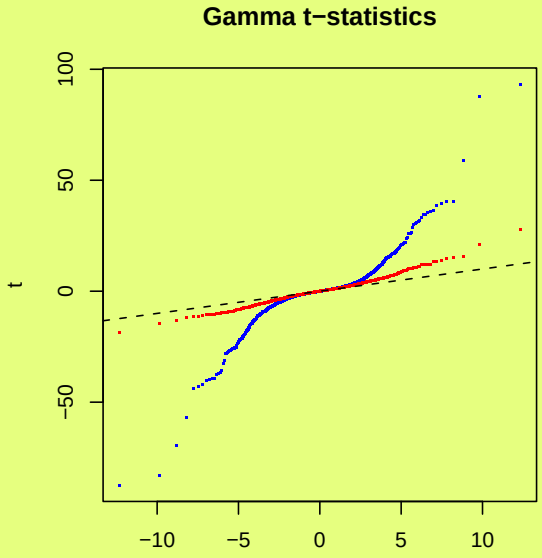
Robust regression

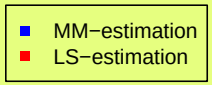
- A widely used 'off-the-shelf' method is *MM-estimation*[1]
  - ⋆ The two MM's stand for 'robust method within robust method'.
  - ⋆ We use `rlm()` from the VR library in R.

[1] Yohai 1987; Rousseeuw & Leroy 1987; Marazzi 1993


Alpha t−statistics
t
50
0
−50
−10
−5
0
5
10
Quantiles of t5

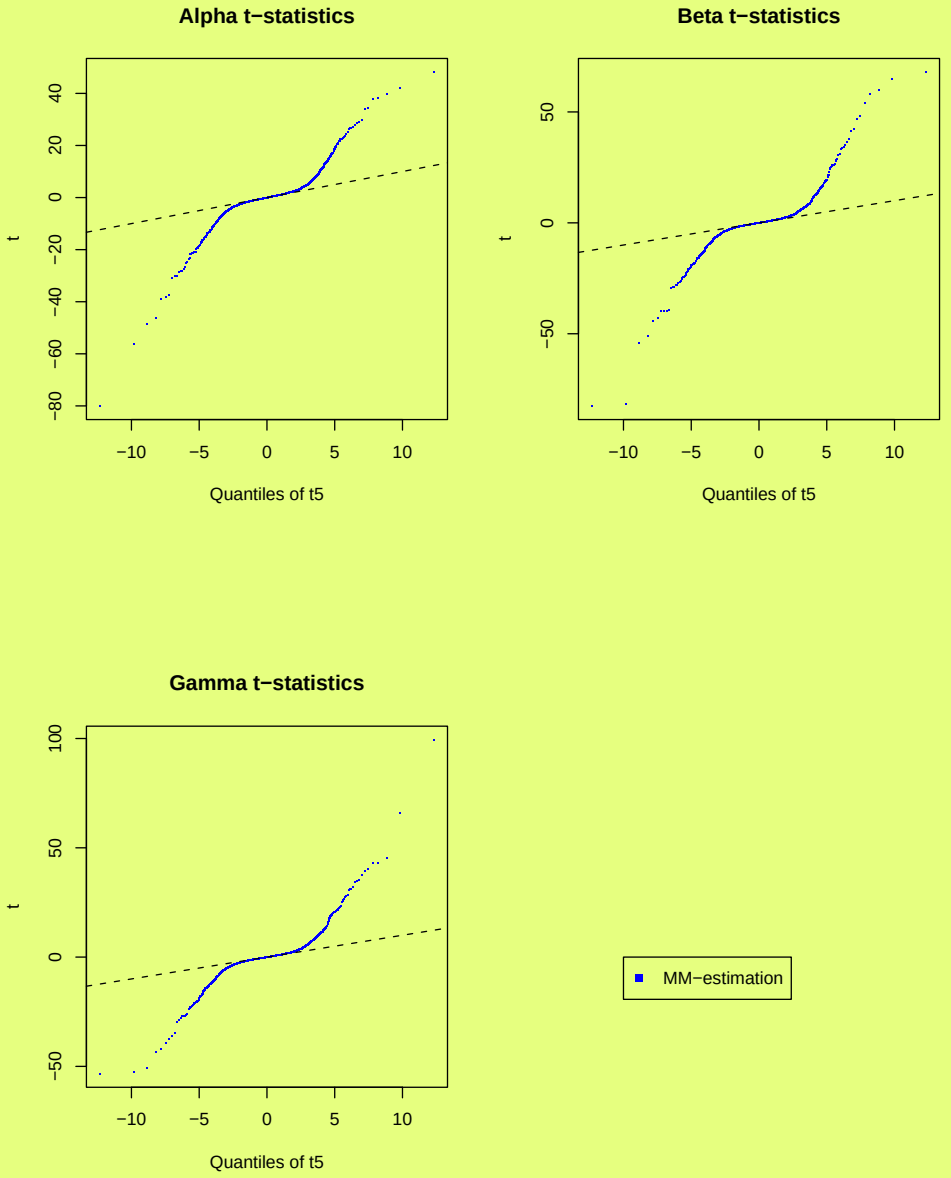

Beta t−statistics
t
150
100
50
0
−50
−100
−10
−5
0
5
10
Quantiles of t5

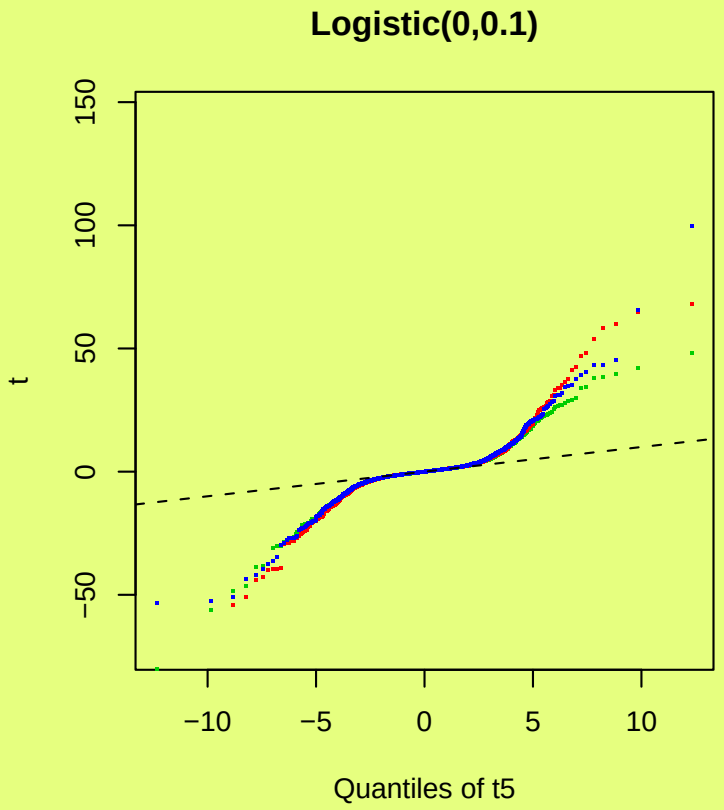

Gamma t−statistics
t
100
50
0
−50
−10
−5
0
5
10

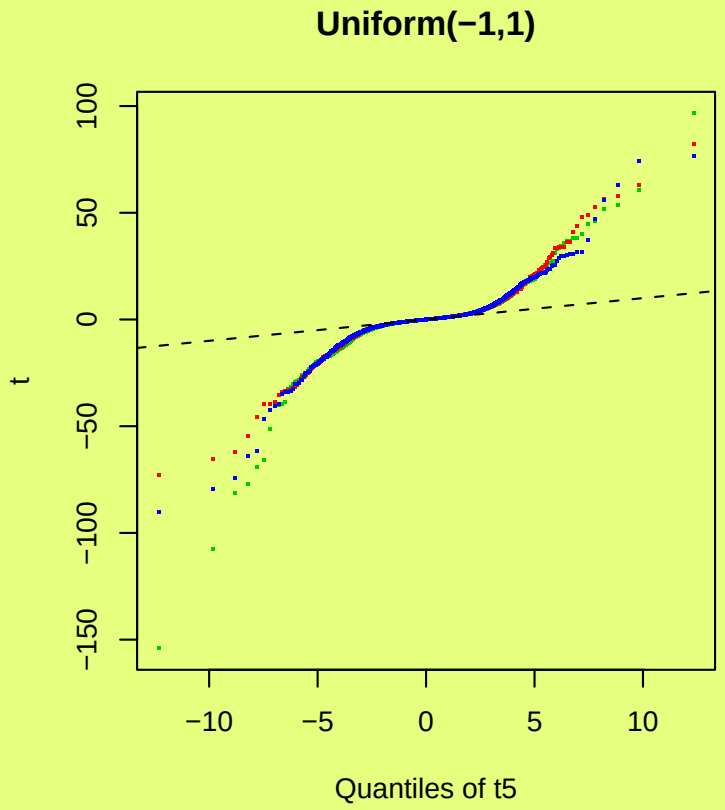

MM−estimation
LS−estimation

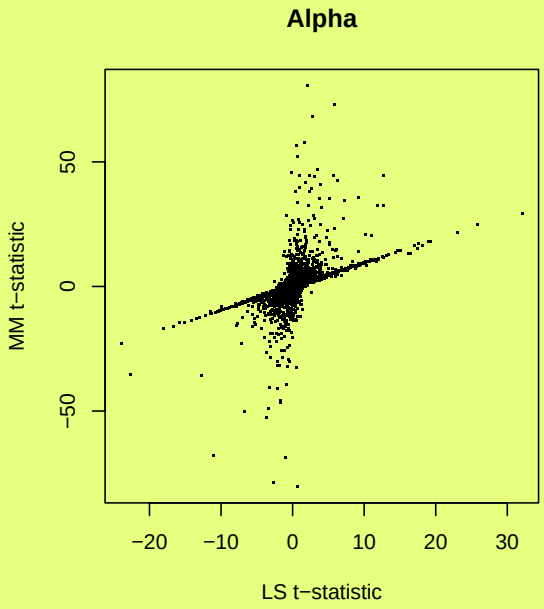
Quantile plots for MM t−statistics from 8−slide N(0,1) simulations

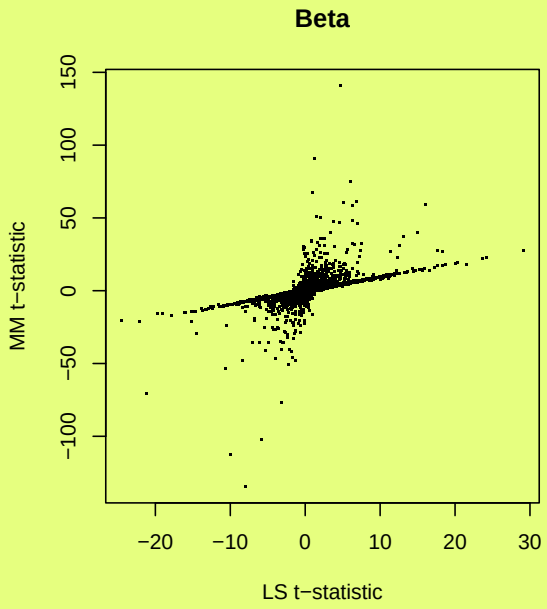
i.i.d. $N(0,1)$

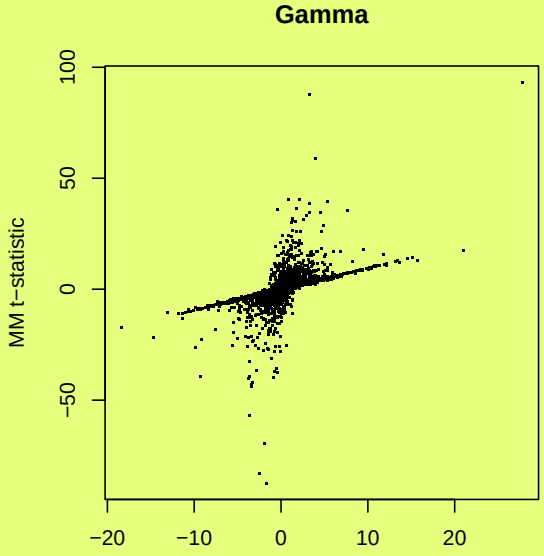


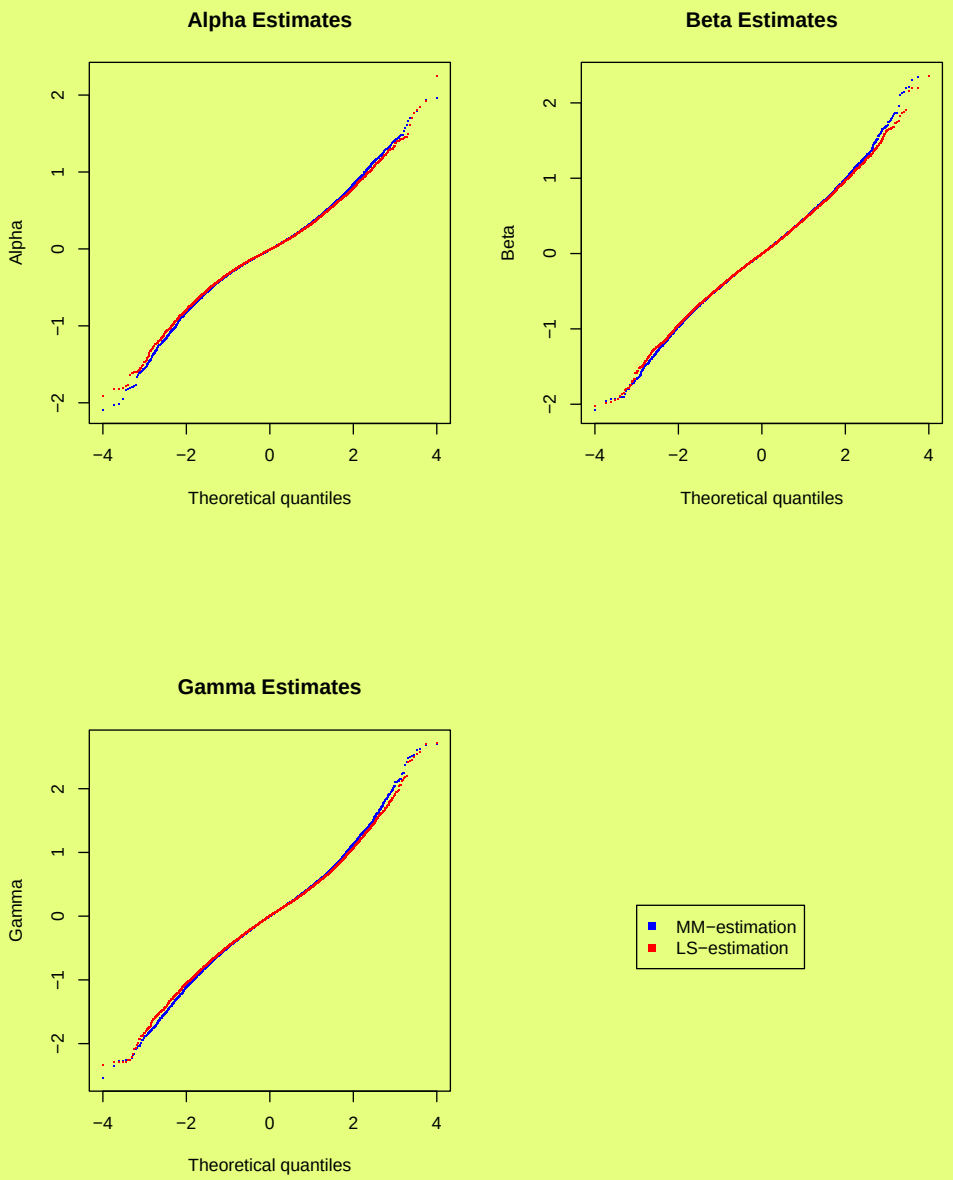

Logistic(0,0.1)
t
150
100
50
0
−50
−10
−5
0
5
10
Quantiles of t5

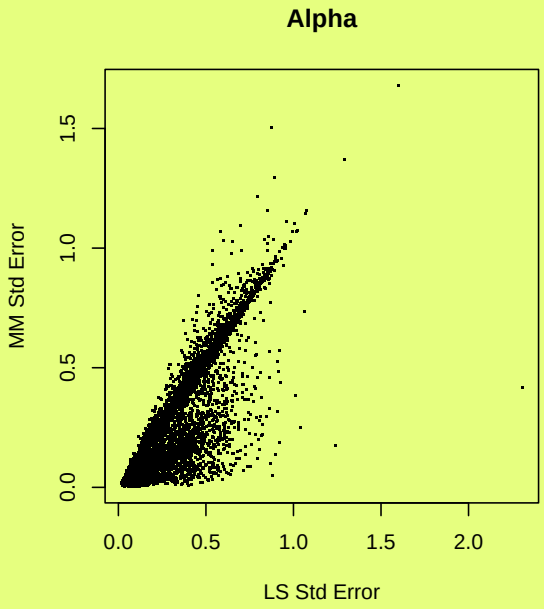

Uniform(−1,1)
t
100
50
0
−50
−100
−150
−10
−5
0
5
10
Quantiles of t5

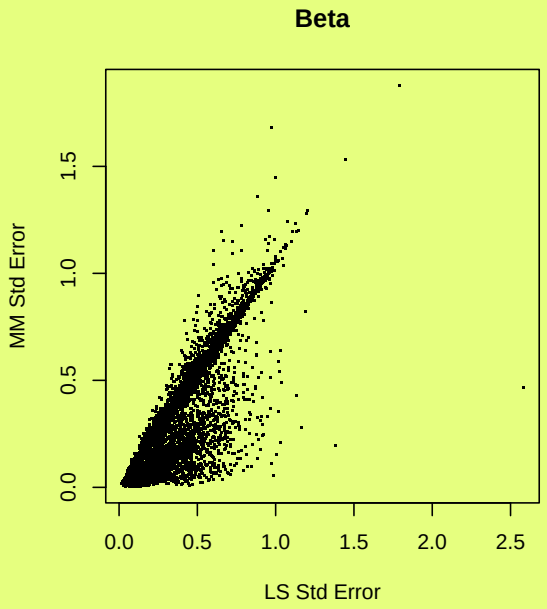

Alpha
MM t−statistic
LS t−statistic
50
0
−50
−20
−10
0
10
20
30

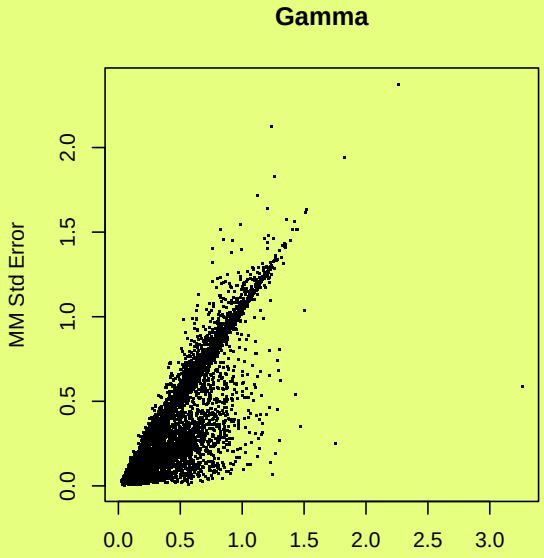

Beta
MM t−statistic
LS t−statistic
150
100
50
0
−50
−100
−20
−10
0
10
20
30

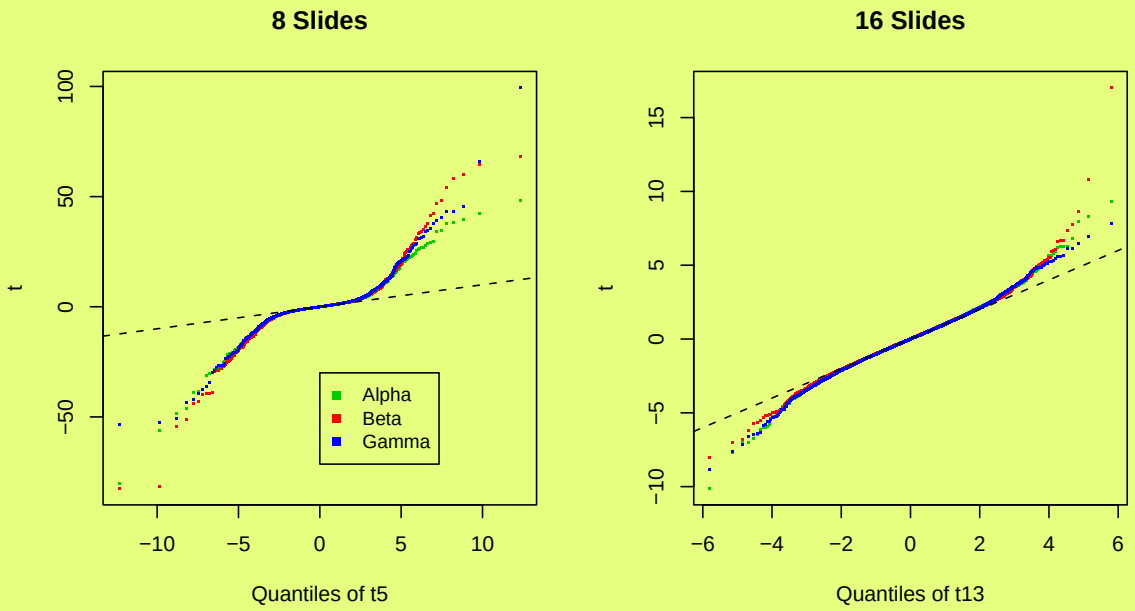

Gamma
MM t−statistic
100
50
0
−50
−20
−10
0
10
20


Alpha Estimates
Alpha
Theoretical quantiles
Beta Estimates
Beta
Theoretical quantiles
Gamma Estimates
Gamma
Theoretical quantiles
MM−estimation
LS−estimation


Alpha
MM Std Error
LS Std Error
0.0
0.5
1.0
1.5
2.0


Beta
MM Std Error
LS Std Error
0.0
0.5
1.0
1.5
2.0
2.5


Gamma
MM Std Error
0.0
0.5
1.0
1.5
2.0
2.5
3.0


8 Slides
16 Slides
24 Slides
32 Slides
t
Quantiles of t5
Quantiles of t13
Alpha
Beta
Gamma
100
50
0
−50
−10
−5
0
5
10
15
−6
−4
−2
2
4
6

Standard errors biased downwards

- *The parameter estimates themselves are well-behaved.*
- Biased standard errors a problem in small samples.

Standard errors biased downwards

- *The parameter estimates themselves are well-behaved.*
- Biased standard errors a problem in small samples.
  - ◇ Salibian-Barrera & Zamar (2002) use a fast bootstrap method to produce confidence intervals with better coverage.
  - ◇ DiCiccio & Monti (2002)
    - ⋆ derive asymptotic formulae for the bias and skewness of the t-statistic and
    - ⋆ *construct second-order accurate confidence intervals with improved coverage accuracy.*

Some concluding remarks

- When applying MM-estimation to cDNA microarray data
  - ◇ we believe it is an open question whether you can fine-tune `rlm()`, or whether you need to look more seriously at the accuracy of the estimation procedure.

Some concluding remarks

- When applying MM-estimation to cDNA microarray data
  - ◇ we believe it is an open question whether you can fine-tune `rlm()`, or whether you need to look more seriously at the accuracy of the estimation procedure.
- *Take-home-message: don't take the results of MM-estimation at face value.*

Acknowledgements

- Adelaide Women's and Children's Hospital
  - Brenton Reynolds
  - Richard D'Andrea
- Hanson Institute/IMVS
  - Greg Goodall
- Bionomics
  - Tom Gonda

Further reading

- http://www.statsci.org/micrarra/index.html
- Bioconductor Project: Open source bioinformatics using R. http://www.bioconductor.org
- MJ Buckley. The Spot user's guide. CSIRO Mathematical and Information Sciences, August 2000. http://www.cmis.csiro.au/IAP/Spot.htm
- S Dudoit, YH Yang, B Bolstad. Using R for the analysis of DNA microarray data. R News **2**, 24–32, 2002.
- R. Ihaka & R. Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314, 1996.
- I Lönnstedt, TP Speed. Replicated microarray data. *Statistica Sinica*, in press.

- Microarray Analysis Group. The University of Adelaide http://maths.adelaide.edu.au/MAG
- sma v. 0.5.6 (November 2001), B Bolstad, S Dudoit, YH Yang. http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html.
- GK Smyth, YH Yang, TP Speed (2002) Statistical issues in cDNA microarray data analysis. Research Report, WEHI.
- The Chipping Forecast. *Supplement to Nature Genetics*, **21**, 1999.
- TP Speed. Preprints, information and software (sma) www.stat.berkeley.edu/users/terry/zarray
- YH Yang, MJ Buckley, S Dudoit, TP Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical*

Statistics **11**, 2002.

- YH Yang, S Dudoit, P Luu, DM Lin, V Peng, J Ngai, TP Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, 2002.

THE END