Muddling or modelling your way through normalization?

Professor Ernst Wit
University of Groningen

Joint work with Luigi Augugliaro, University of Palermo

e.c.wit@rug.nl

http://www.math.rug.nl/∼ernst

5 December 2008

Two philosophies

There are essentially two attitudes to "normalization":

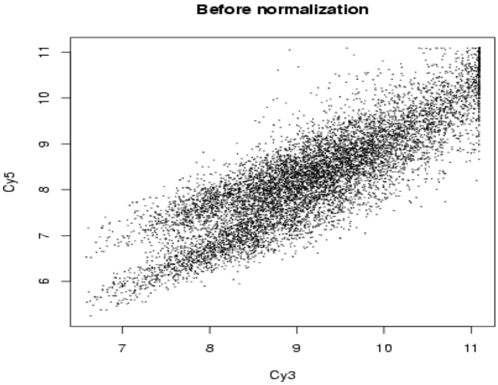
- **Computer Scientist's Attitude: Muddling**
  a **pre**processing activity, whereby data are cleaned before further analysis.
- **Statistician's Attitude: Modelling**
  a joint modelling activity, whereby analysis and accounting for nuisance effects are combined.

It is easy to see why the former is more prevalent:

- Computationally less intensive;
- Convenient to separate normalization and analysis;
- There are more computer scientists than statisticians.

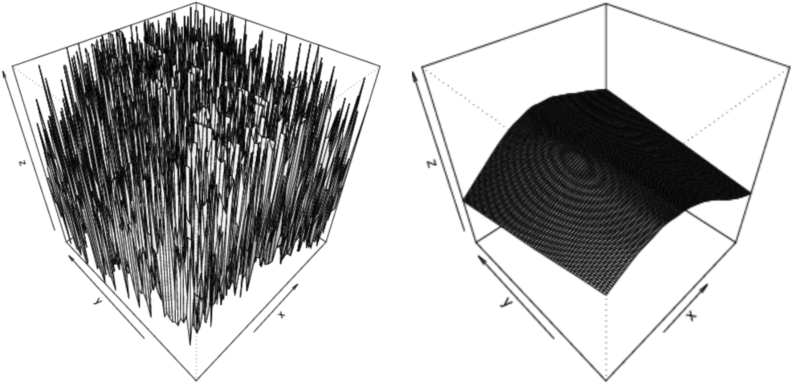
Example of the Computing Scientist Attitude

Rule: *Normalize all local features first; then progress to normalizations that involve several and, finally, all arrays.*



Spatial Normalization

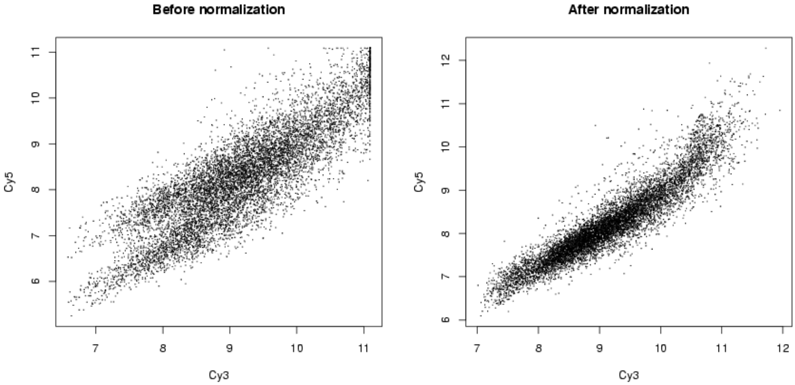
Location: Fit smooth surface to data and subtract it.
Scale: Fit smooth surface to residuals and divide by it.



Then rescale and relocate by the median of the two surfaces.

Example of Spatial Normalization

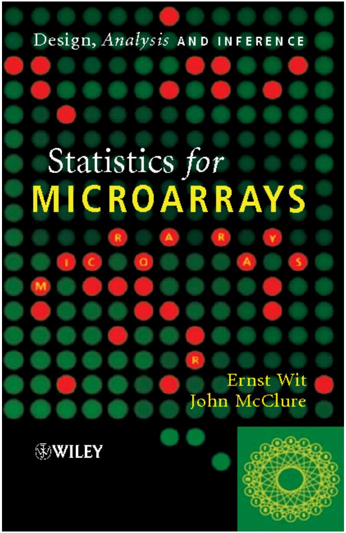
Spatial normalization before dye normalization is essential!



(a)

(b)

… etc.



And you can do it also for

- Background “subtraction”
- Dye normalization
- Between-slides normalization
- ….

As done, e.g., in this “computer scientist” book by

Ernst Wit & John McClure
John Wiley & Sons

What are the drawbacks of “muddling”?

- **False believe** that the normalized data are clean (and typically no way of checking whether this is true).
- The uncertainly inherent in the normalization is not carried forward to the analysis: results can be **too liberal**.
- Most pre-processing methods **can’t deal with additional structure** in the data.

As an alternative we proprose a statistical model, in order to

- check the validity of our normalization model.
- carry the uncertainty in the normalization over to inference.
- deal with the peculiar structure of the EMERALD dataset.

What are the essential features of the EMERALD data

- **Comparison of interest:** 2 tissue types: kidney and liver,
  - measured in 0/1, 0.25/0.75, 0.75/0.25, 1/0 mixtures,
  - each repeated 3 times (per rat, per platform)
  - plus some additional pools
- 3 different laboratories each with their own platform.
- 6 normal rats, repeatedly used in each lab.
- 96 arrays in each platform.

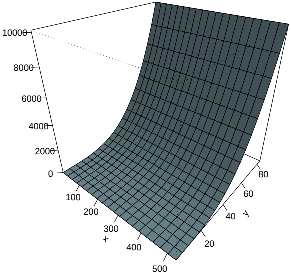
Therefore,

- Platform is confounded with laboratory.
- Low replication number: only 6 degrees of freedom for comparing kidney/liver across thousands of genes; deal with lots of technical replication.
- Mixtures are introduced, which need to be modelled.

What are the nuisance (but relevant) features of the EMERALD data?

- There might be spatial variation across the slides.
- Depending on the platform, there is information about
  - Fluidics station,
  - Fluidics Machine en
  - Scanner

  that was used in the experiment on each array.



Model Part 1: what we want to know

We want to learn which genes behave differently in the liver and the kidney, so our primary model should be:

$$E \log(y_{gti}) = \alpha_{gt} + \ldots, \qquad \text{for gene } g, \text{tissue } t \text{ and replicate } i$$

which is equivalent with

$$E \log(y_{gti}) = \mu_g + \delta_g \times p_t + \ldots,$$

where

- μ_g = expression of gene g for liver.
- δ_g = amount of differential expression of kidney w.r.t. liver.
- p_i = fraction of kidney tissue in the sample i $(0, \frac{1}{4}, \frac{3}{4}, 1)$.

Model Part 1: random effects model

We assume that

- $\mu_g \sim N(\mu_0, \sigma_0^2), \quad g = 1, \ldots.$
- $\delta_g \sim N(\mu_1, \sigma_1^2), \quad g = 1, \ldots.$

The advantages over a usual regression model

- We require only 4 parameters instead of 40,000!
- We can still do inference on the basis of the random effects;
- It allows a more subtle normalization model.

Model Part 2: Hybridization artifacts

For the Affy data: information about hybridization instruments
For Affy and Agilent: spot location information known.
This can be translated into a model for the structural nuisance effects in the data:

$$E \log y_{smcxy} = \ldots + FS_s + FM_m + S_c + L(x, y) + \ldots .$$

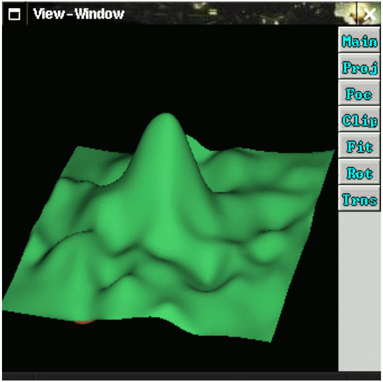
Where

- FS_s = fluidic station effect
- FM_m = fluidic machine effect
- S_c = scanner effect
- $L(x, y)$ = spatial effect at point (x,y) on the array.

B-splines

For the spatial function we use a smooth cubic B-spline,

$$L(x,y) = \sum_{i=1}^{m} P_i b_{i,3}(x) + \sum_{i=1}^{m} Q_i b_{i,3}(y)$$



Model Part 3: Technical replication

FACT: Multiple measurements of same individual are more similar than multiple measurement across different individuals.

Therefore, in the model we include a discriminating factor for measurements across two different individuals:

$$E \log y_{ab} = \ldots + \sum_{b=1}^{6} f_{ab} B_b + \ldots$$

where

- B_b = amount of biological variation away from the mean for indvidual b.
- f_{ab} = fraction of biological sample b on array a.

It common to take $B_b \sim N(\mu_2, \sigma_2^2)$, but here are only 6 individuals.

Scale and Variation differences between platforms

Maybe the most challenging aspect of this analysis: the combination of data from 3 platforms.

- Do the platforms have the same scale?
- Do the platforms have the same variability?

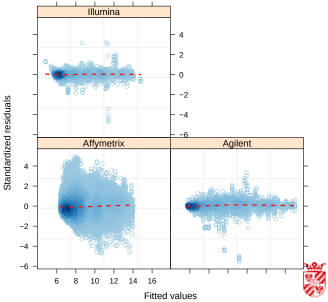
Scale?

	Average
Affy	5.67
Agilent	5.32
Illumina	5.67

$$\log(y_a i) = \ldots + M_a + \epsilon_{ai}$$

where $\epsilon_{ai} \sim N(0, \sigma_a^2)$

Variability?



Complete model

$$\begin{aligned}\log y_{gtmcxybai} &= \mu_g + \delta_g \times p_t + \sum_{i=1}^{3} P_i b_{i,3}(x) + \sum_{i=1}^{3} Q_i b_{i,3}(y) \\ &\quad + B_b + M_a + FS_s + FM_m + S_c + L(x,y) + \epsilon_{ai}\end{aligned}$$

consists of ±300 fixed effect parameters and a couple of random effect parameters.

	DF	denDF	F-value	p-value
Other fixed	12	30801	340.07	0.00
Spatial	288	30801	9.26	0.00

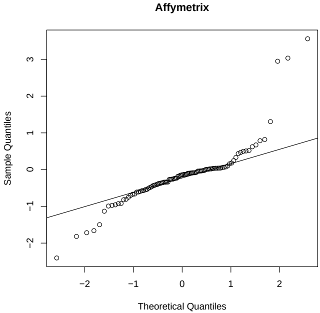
Fixed effects

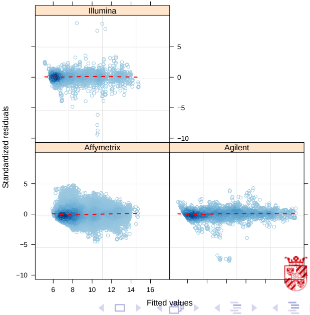
	Value	Std.Error	DF	t-value	p-value
(Intercept)	8.46	0.19	30801.00	43.68	0.00
Fluidics.station2	−0.09	0.10	30801.00	−0.92	0.36
Fluidics.station3	0.01	0.10	30801.00	0.09	0.93
Fluidics.station4	0.19	0.09	30801.00	2.24	0.03
Fluidics.station0	−0.18	0.17	30801.00	−1.08	0.28
Fluidics.machine2	−0.11	0.09	30801.00	−1.24	0.22
Fluidics.machine3	−0.08	0.11	30801.00	−0.70	0.48
Fluidics.machine7	−0.05	0.11	30801.00	−0.44	0.66
Fluidics.machine8	0.39	0.12	30801.00	3.33	0.00
Fluidics.machine9	0.20	0.14	30801.00	1.50	0.13
Scanner2	0.31	0.07	30801.00	4.13	0.00
Bio.Sample2	−0.03	0.01	30801.00	−2.73	0.01

Random effects

```
             StdDev     Corr
(Intercept)  1.7484842  (Intr)  prop   Agilent
prop         0.9380541  -0.153
Agilent      1.7295239  0.355   0.097
Illumina     1.4767537  -0.078  0.247  0.338

Residual     0.8560642
```





Results

	(Intercept)	prop	Agilent	Illumina
RGD1311100(predicted)	0.83	−2.46	−0.36	0.27
Bspry	0.68	−2.07	0.83	−0.95
RGD1565941(predicted)	0.89	−2.00	0.68	−1.14
Prss23	1.87	−1.97	1.30	1.05
LOC361596	4.16	−1.62	2.00	−5.65
...				
Reln	−2.17	1.79	0.01	2.20
LOC364773	1.67	2.49	−0.79	0.62
Fn1	1.64	3.10	1.56	1.15
Clu	1.51	3.39	1.71	1.60
Smp2a	−1.74	3.60	1.97	1.92

Computational efforts

The bad news:
It takes several hours to process the data (approximately 500,000 data points) and fit the model.

Computational efforts

The bad news:
It takes several hours to process the data (approximately 500,000 data points) and fit the model.

The good news:
The method can be run in any package with mixed model capabilities.

Conclusions

- The muddling approach to normalization has and will have a role to play in large datasets;
- Mixed effects models make it possible to replace the muddling approach by a modelling approach, which means that quality of the inference improves.
- Fantastic dataset for the development of intra-platform methods.