

# Latent class analysis for Genomic Micro-arrays

Hans C van Houwelingen

Dept. Medical Statistics  
Leiden University Medical Centre

[jcvanhouwelingen@lumc.nl](mailto:jcvanhouwelingen@lumc.nl)

Comparative Genomic Micro-arrays are used to measure genomic imbalances in DNA content (gains and losses of DNA along the genome) as they frequently occur in cancer.

This leads to a restricted number of possibly true states at the measured positions: deleted, normal, amplified and over-amplified.

Examples of such data from our local laboratory are discussed in Rosenberg C, Geelen E, IJszenga MJ, et al.

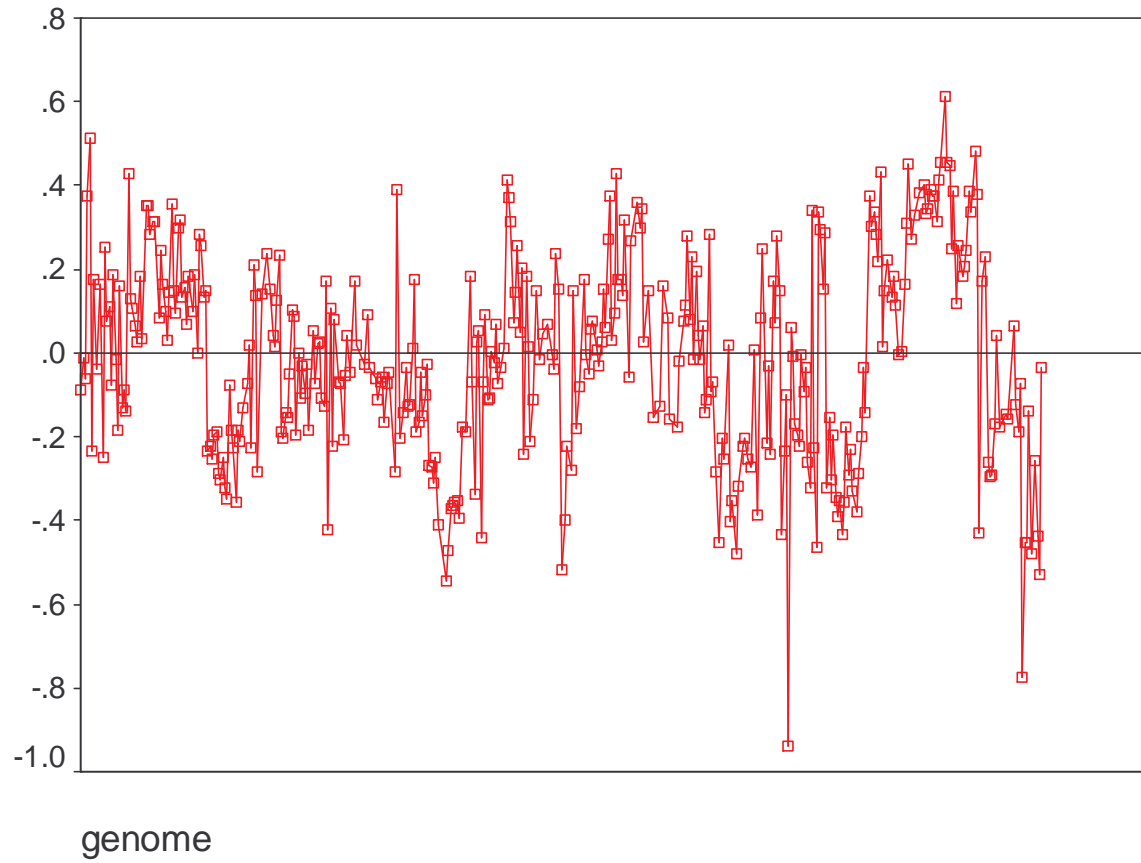
Spectrum of genetic changes in gastro-esophageal cancer cell lines determined by an integrated molecular cytogenetic approach  
CANCER GENET CYTOGEN 135 (1): 35-41 MAY 2002

Example: data on 448 genes for 21 individuals

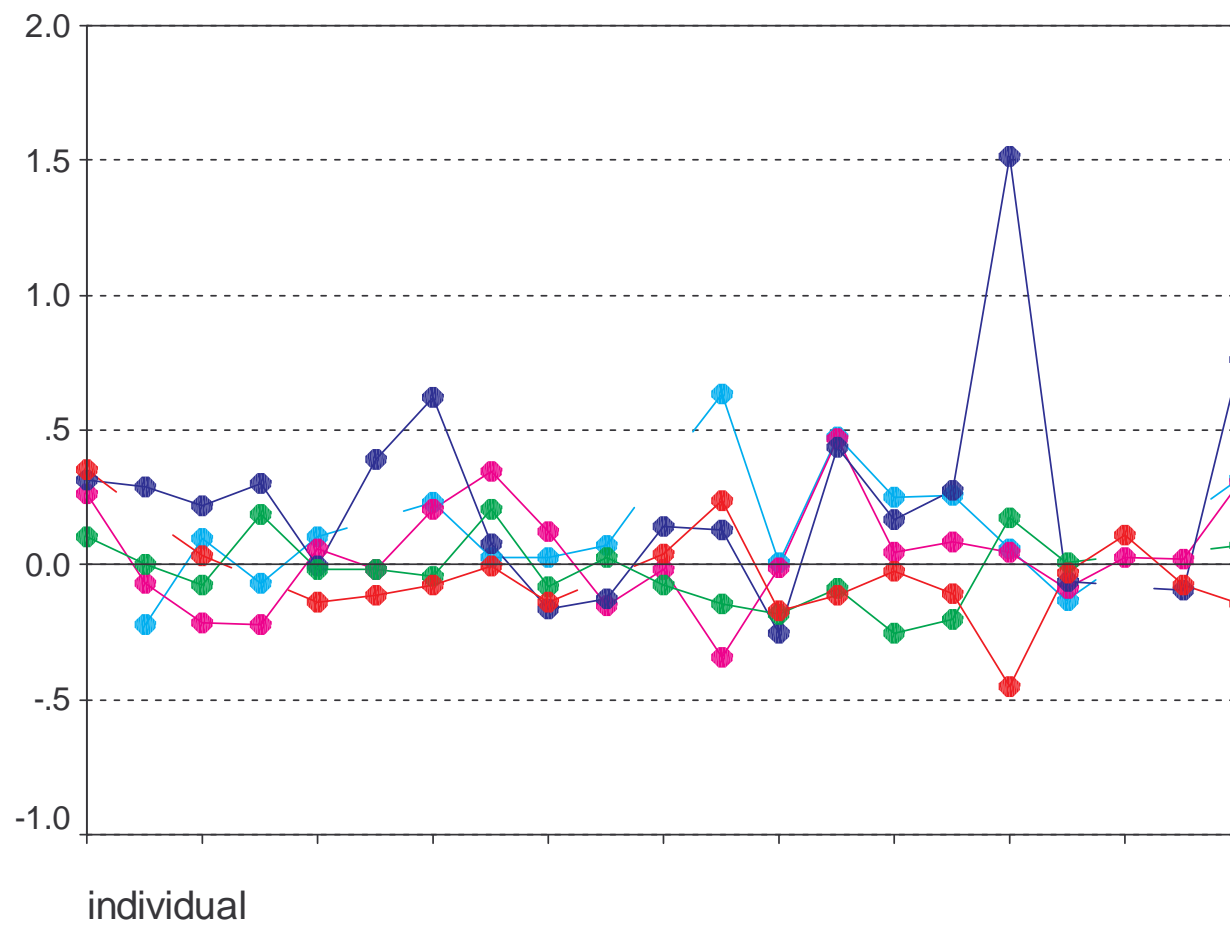
Data are expressed on  $\ln(\text{ratio})$  scale and normalized (mean/median/mode equal to zero).

Lots of missing data.

# data for one individual

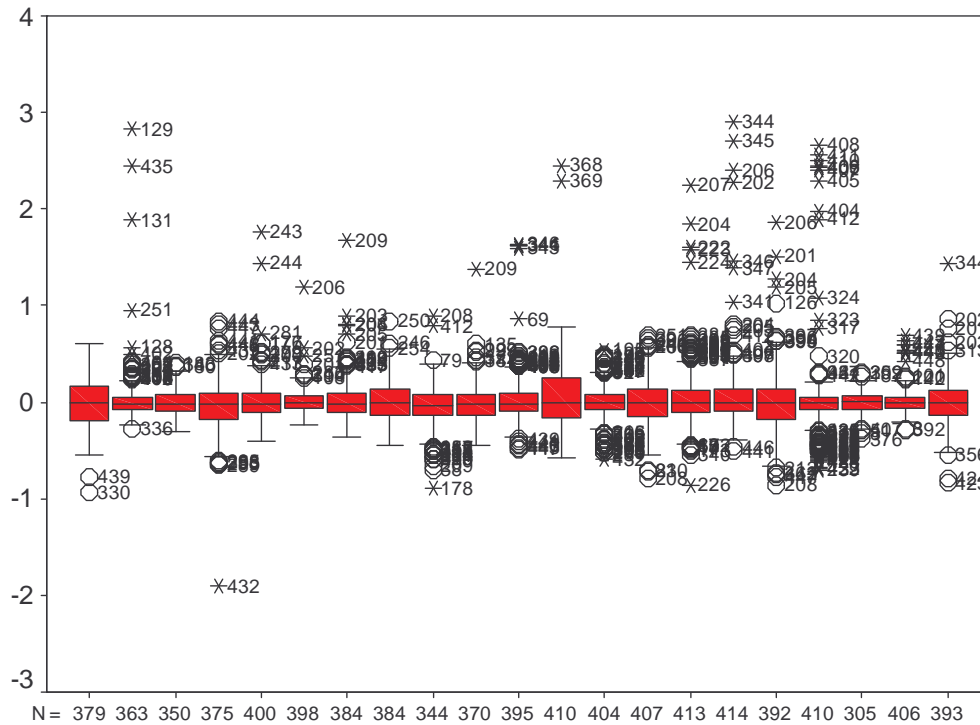


data for a few genes

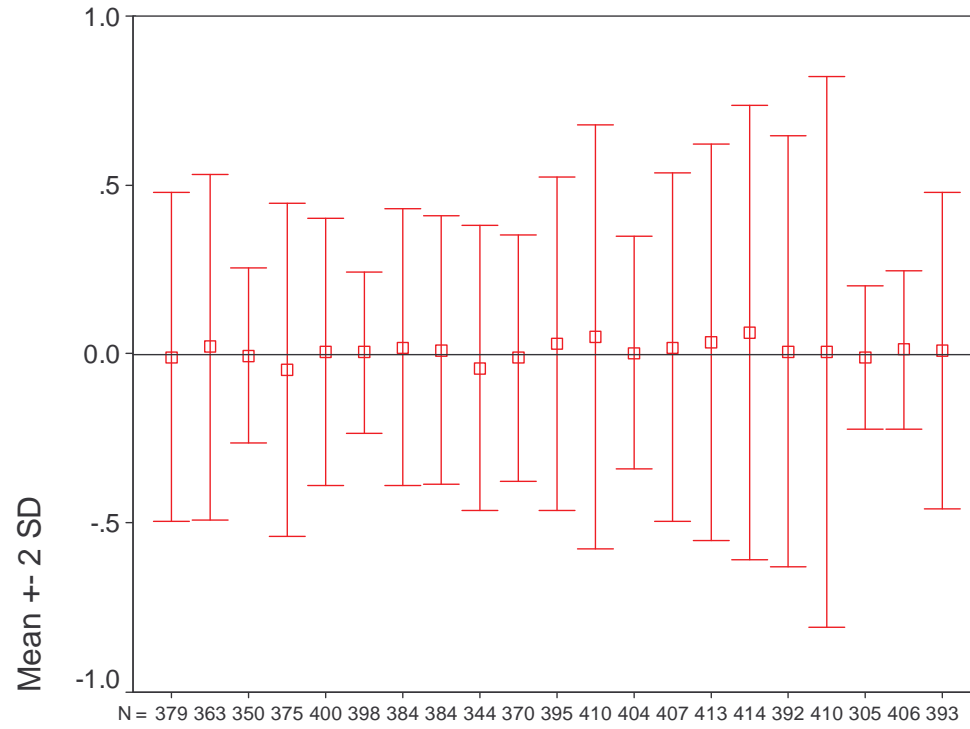


After normalization, there appear to be big differences in standard deviations between the arrays.

Box plot for 21 samples

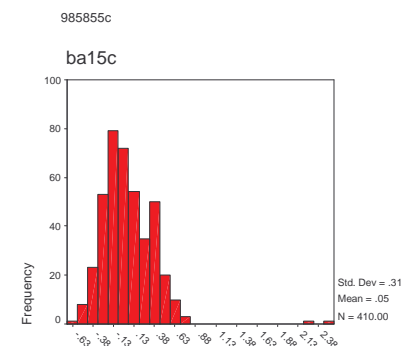
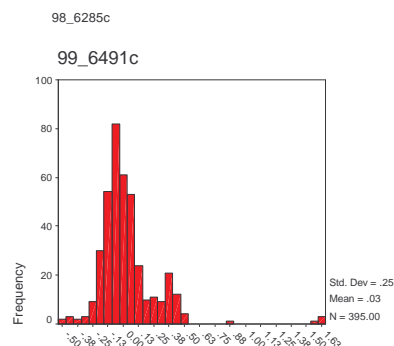
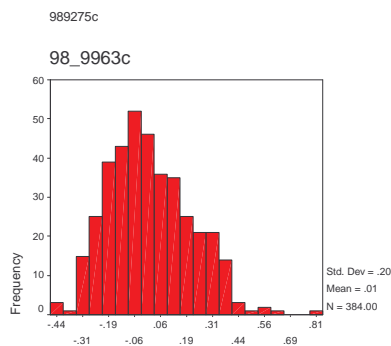
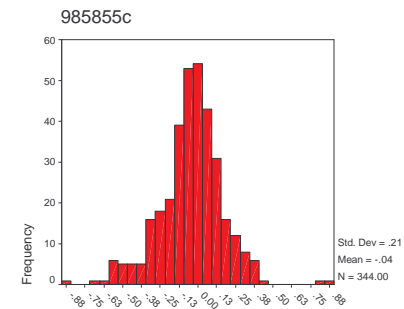
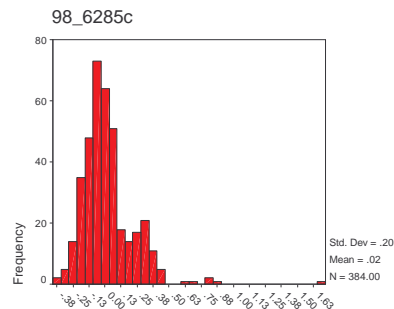
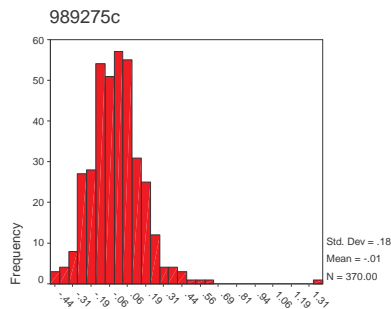


mean  $\pm$  2 sd for 21 samples



# Suggested analysis in the literature: Clusteranalysis per array á la Parmigiani.

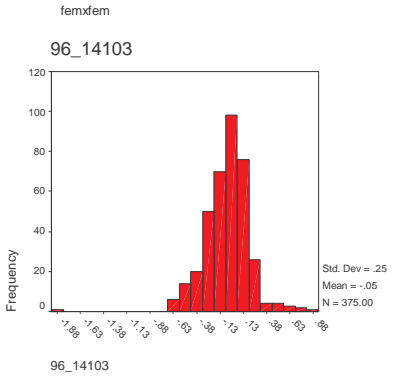
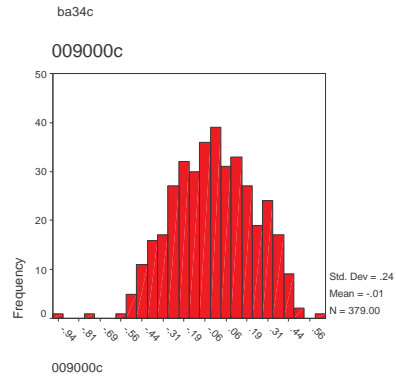
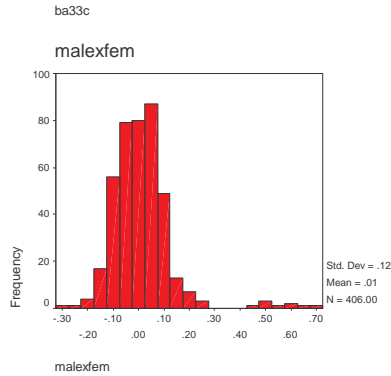
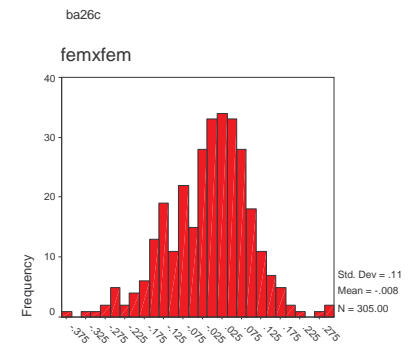
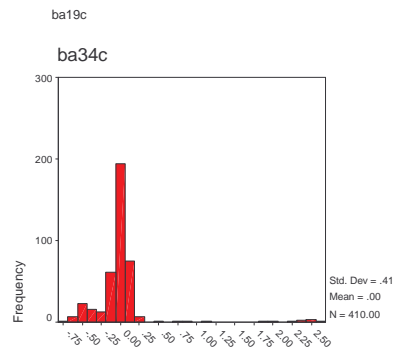
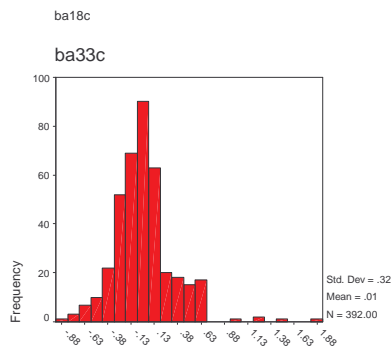
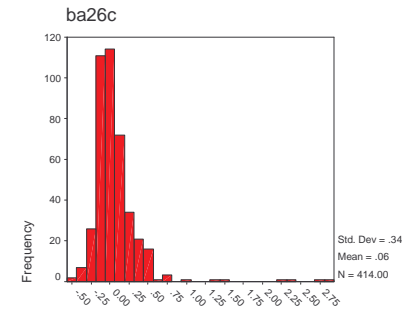
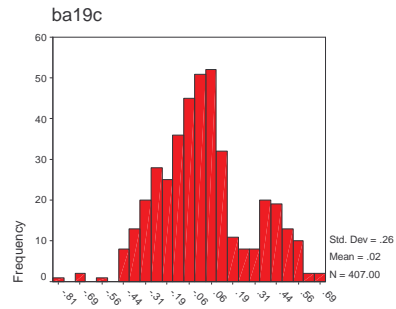
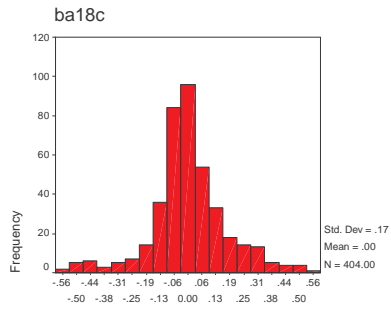
Not very easy to perform

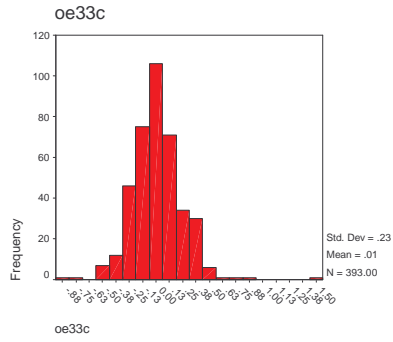


98\_9963c

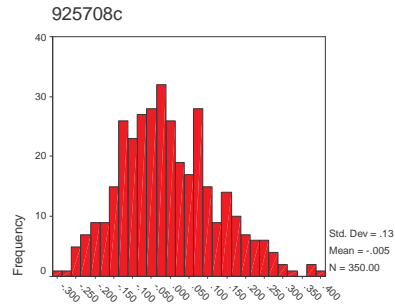
99\_6491c

ba15c

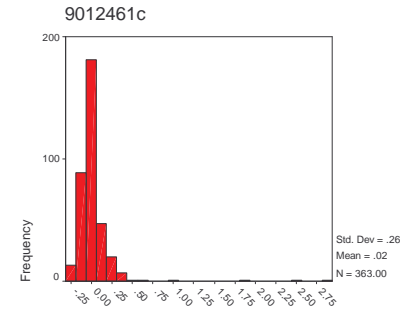




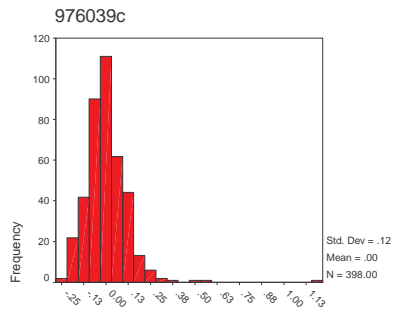
oe33c



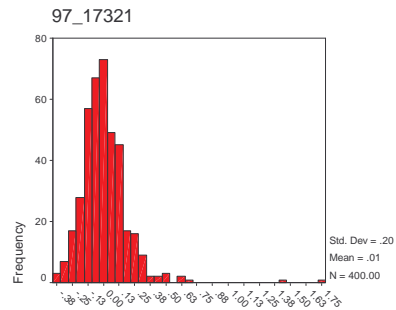
925708c



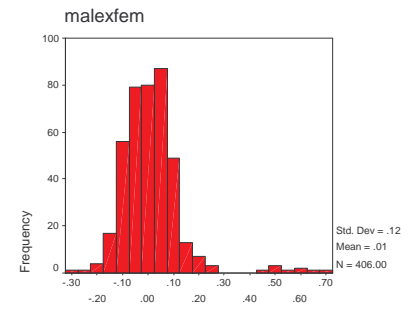
9012461c



976039c

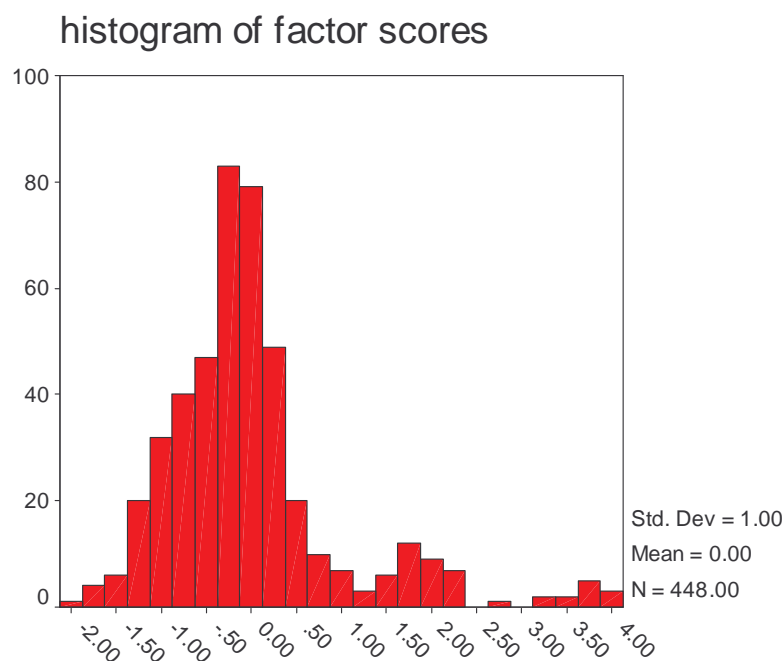


97\_17321



malexfem

Alternative. Combine data through factor analysis and cluster then.



This gives some indication of the four latent classes.

Better: put class into the model from the start: latent class analysis.

## Elements of data analysis (sketchy, not perfect yet)

- latent class model assuming independent genes
  - fit the model by EM
  - compute posterior probabilities
- latent class model combined with Markov Chain model for genes
  - fit the model by pseudo-likelihood
  - compute posteriors by using neighboring genes

Model:

- genes cluster in 4 different clusters with
  - probabilities  $\pi_1, \pi_2, \pi_3, \pi_4$
  - effect parameters  $\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4$ 
    - $\vartheta_2 = 0$
    - $\sum_{i=1}^4 \pi_i \vartheta_i^2 = 1$
- observation on gene  $j$  in cluster  $c$  for individual  $i$

$$X_{i,j(c)} \sim N(\beta_i \vartheta_c, \sigma_i^2)$$

- marginal model 
$$X_{i,j} \sim \sum_{c=1}^4 \pi_c N(\beta_i \vartheta_c, \sigma_i^2)$$

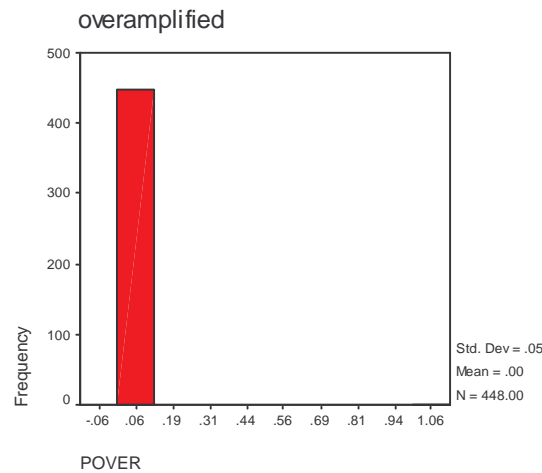
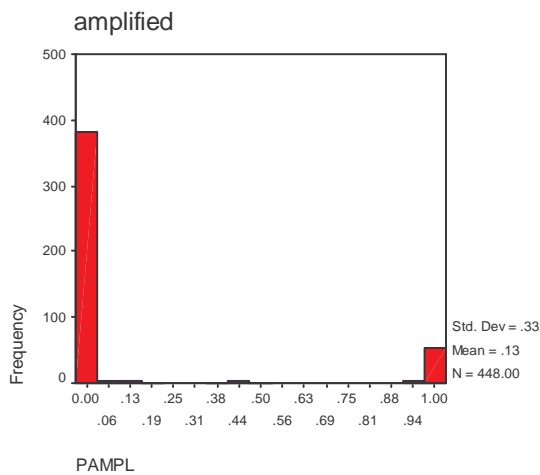
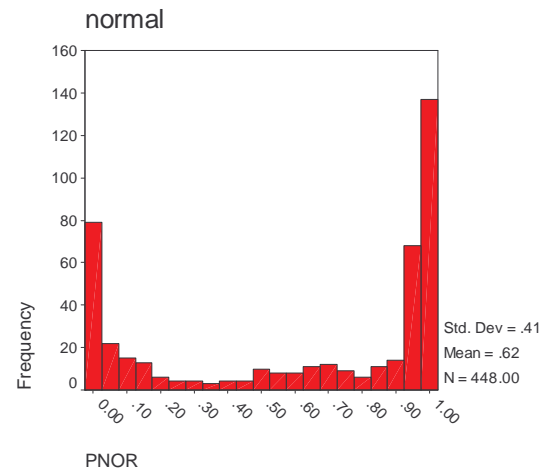
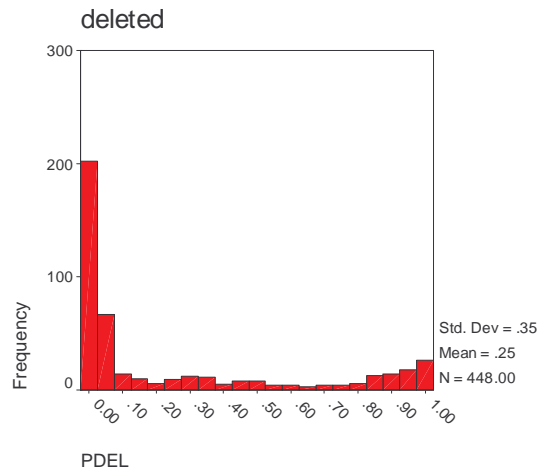
Ideal model to fit by EM

Result

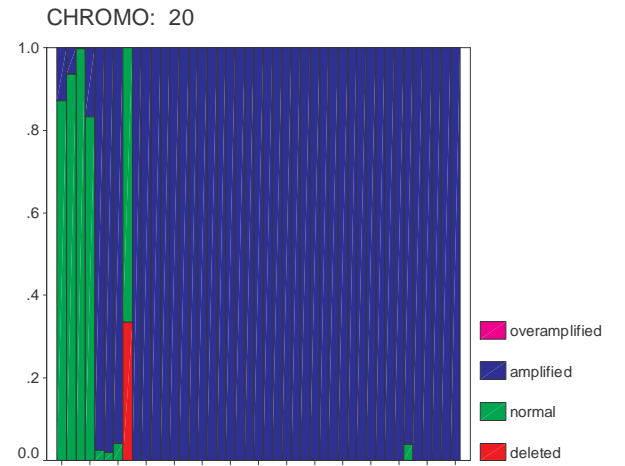
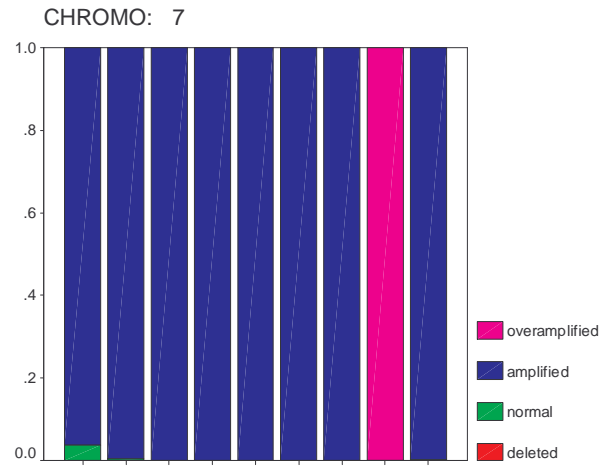
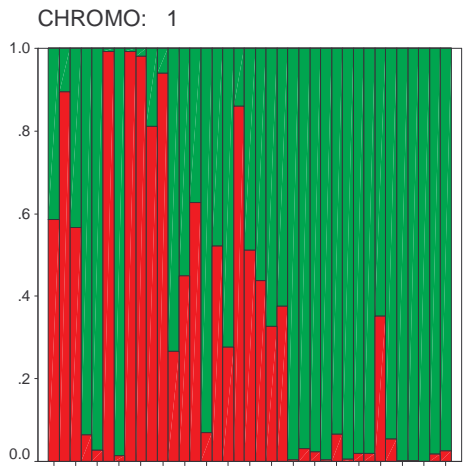
$\pi$	$\theta$
0.2477	-0.9319
0.6180	0.0000
0.1321	2.2988
0.0022	6.2428

individual	$\beta$	$\sigma$	explained variation
<b>1</b>	<b>0.1298</b>	<b>0.2056</b>	<b>0.2851</b>
2	0.0724	0.2446	0.0806
3	0.0206	0.1269	0.0258
4	0.0102	0.2502	0.0017
5	0.0495	0.1910	0.0631
<b>6</b>	<b>0.0724</b>	<b>0.0952</b>	<b>0.3663</b>
<b>7</b>	<b>0.1323</b>	<b>0.1533</b>	<b>0.4270</b>
8	0.0303	0.1956	0.0234
9	0.0672	0.2022	0.0996
10	0.0533	0.1738	0.0861
<b>11</b>	<b>0.1568</b>	<b>0.1991</b>	<b>0.3826</b>
12	0.1188	0.2951	0.1394
13	0.0079	0.1716	0.0021
<b>14</b>	<b>0.1714</b>	<b>0.1940</b>	<b>0.4385</b>
<b>15</b>	<b>0.1912</b>	<b>0.2235</b>	<b>0.4227</b>
<b>16</b>	<b>0.2314</b>	<b>0.2492</b>	<b>0.4630</b>
<b>17</b>	<b>0.2048</b>	<b>0.2387</b>	<b>0.4242</b>
18	0.1409	0.3814	0.1201
19	0.0063	0.1059	0.0036
20	-0.0045	0.1177	0.0015
<b>21</b>	<b>0.1790</b>	<b>0.1618</b>	<b>0.5505</b>

Results can be used to compute posterior cluster probabilities per gene.



These can be plotted along the chromosome



Model does not take into account the “spatial correlation” of clustering along the chromosome.

Simple stable Markov Chain model for “spatial” clustering

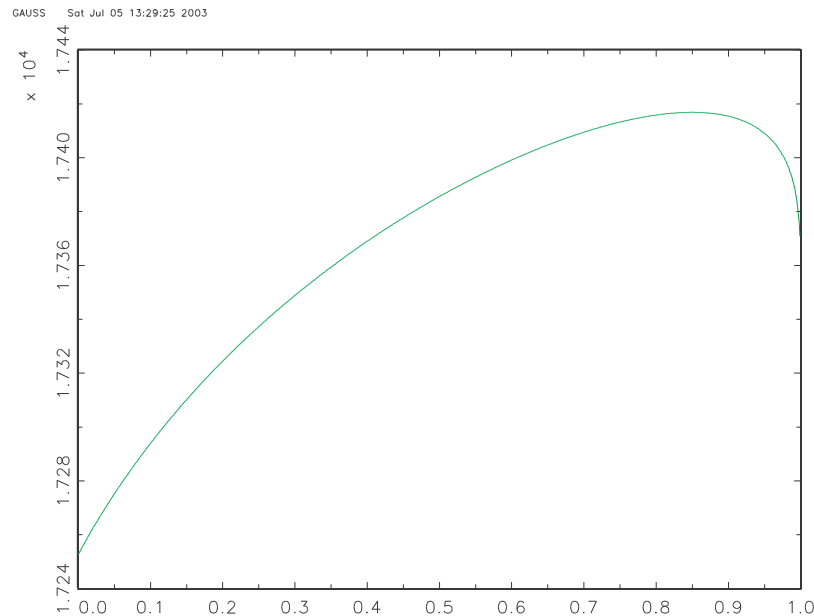
Let  $Z_j$  be the cluster status for gene  $j$

$$P(Z_j = k' | Z_{j-1} = k) = \alpha \delta_{k,k'} + (1 - \alpha) \pi_{k'}$$

Bivariate representation

$$P(Z_{j-1}, Z_j) = \alpha \text{diag}(\pi) + (1 - \alpha) \pi \pi'$$

Additional parameter  $\alpha$  is estimated by pseudo-likelihood using adjacent genes (on the same chromosome)

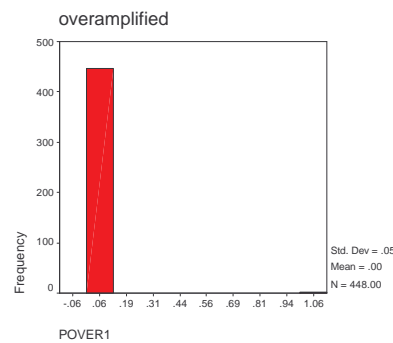
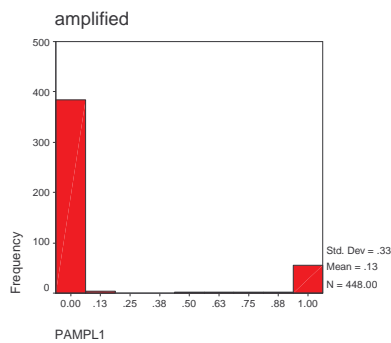
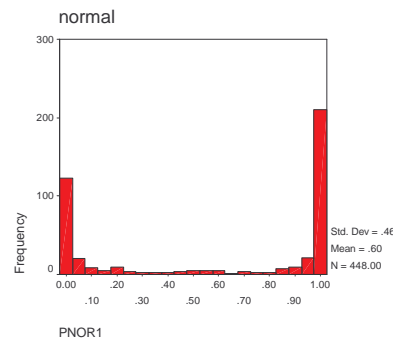
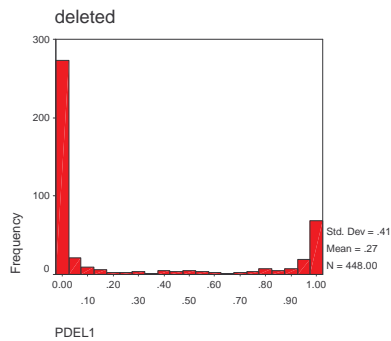


Estimate  $\hat{\alpha} = 0.85$

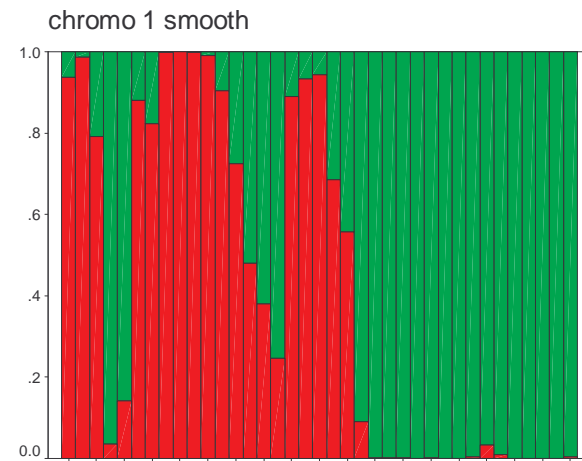
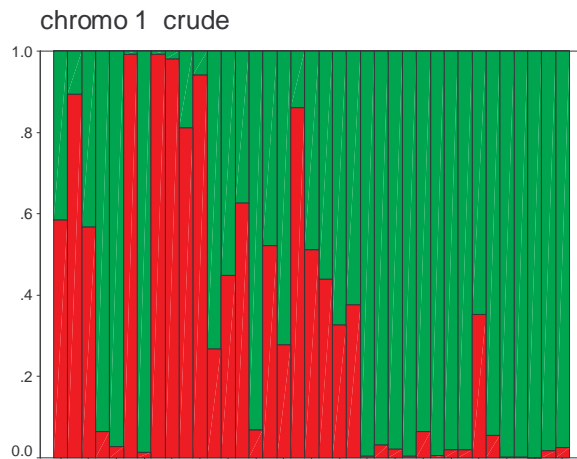
Full posterior probabilities are hard to compute.

Practical solution is to use the data on the gene itself and its two neighbouring genes.

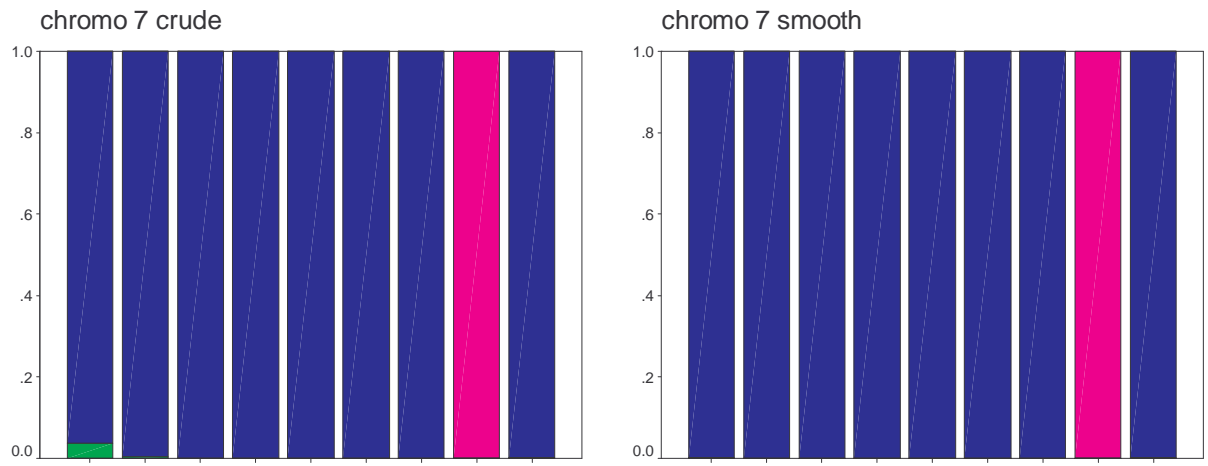
Histogram of posteriors much more outspoken.



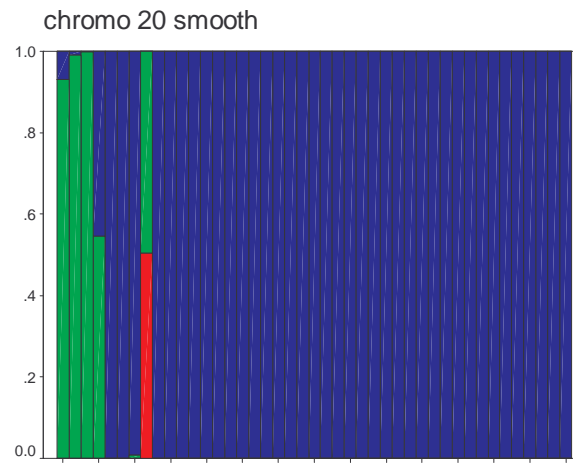
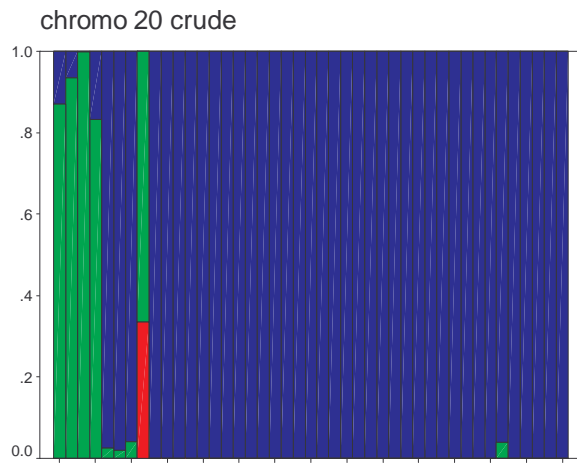
This also shows along the chromosomes



This shows clear smoothing effect



Outlying over-amplified gene is maintained



Again, clear effect of smoothing.

## Conclusion

- we can learn a lot from psychometry and spatial modelling
- data structure should be exploited as much as possible
- mixture models can be fitted without MCMC, using relevant pieces of information