# BAYESIAN CHARACTERISATION OF NATURAL VARIATION IN GENE EXPRESSION

**Madhuchhanda Bhattacharjee**
**Mikko J. Sillanpaa**
**Elja Arjas**

**Rolf Nevanlinna Institute**
**University of Helsinki**
**Finland**

# Introduction

- We present a new latent variable based Bayesian clustering method for classifying genes into categories of interest.

- The approach is integrated in the sense that normalization and classification can be carried out jointly along with estimation of uncertainty.

- The observed expression is treated as a black box for the different effects which are considered jointly in a nested common structure.

- The residuals are then classified into different categories, which is of interest to us here.

- The approach is very general in the sense that it is easily customisable to different needs and can be modified with availability of additional information.

# Data

• A preliminary and an extended version of the model were applied to the expression data provided by Pritchard et al. (2001).

• The data contained median foreground and background intensities for about 5500 genes from experimental and reference samples taken from 3 organs of 6 mice each applied with 2 dyes and 2 replicates.

•This resulted in approximately 1.5 million data points.

• On several occasions the resulting intensities turned out to be negative. In absence of further clarification for such measured intensities, these were treated as missing data.

• We considered 5325 genes for each of which more than 50% of the log-ratio-of intensities were available.

# Model A

- We adjusted the observed expression log ratios by an effect for each organ and each of the 24 arrays.
- The adjusted data were then inspected for possible variation still remaining, if any, exhibited by the genes.
- It is anticipated that the genes may naturally behave differently in different organs from variation perspective.
- Accordingly each gene was classified independently for each organ with respect to its corresponding residual variance.
- We assume three latent variance classes with unknown ordered variances.
- Instead of variances, modelling was actually carried out using corresponding precision parameters.
- For each gene and for each organ, a latent variable indicates its variance-class membership in that organ, taking values in range (1,2,3).

# Model A

- Conditional distribution of the log-ratio of intensities $I_{ioj}$ is assumed to be given by

$$I_{ioj} = \mu_{oj} + e_{ioj} \text{ , where } e_{ioj} \sim N(0, 1/\tau(c_{io})),$$

$$i = 1, \ldots, 5325 \text{ (genes)},$$

$$o = K \text{ (Kidney)}, L \text{ (Liver) and } T \text{ (Testis)},$$

$$J = 1, \ldots, 24 \text{ (arrays)}.$$

- Posterior density $p(\mu,\tau,c,\lambda \mid I)$ is proportional to

$$p(I \mid \mu,\tau,c)\, p(c \mid \lambda)\, p(\tau)\, p(\mu)\, p(\lambda),$$

by assuming conditional independence between the parameters.

# Model A

- We assume vague priors for all model parameters.
- The array effects were assigned Normal priors .
- The precision parameters were assumed to have Gamma distributions *a priori*.
- The latent class-indicators were assigned Multinomial distributions with corresponding probabilities drawn from a Dirichlet distribution.
- In order to preserve compatibility the estimation of the model parameters for all three organs was carried out simultaneously.

# Model Implementation

- We implemented the model and performed parameter estimation using WinBUGS (Gilks et al. 1994).

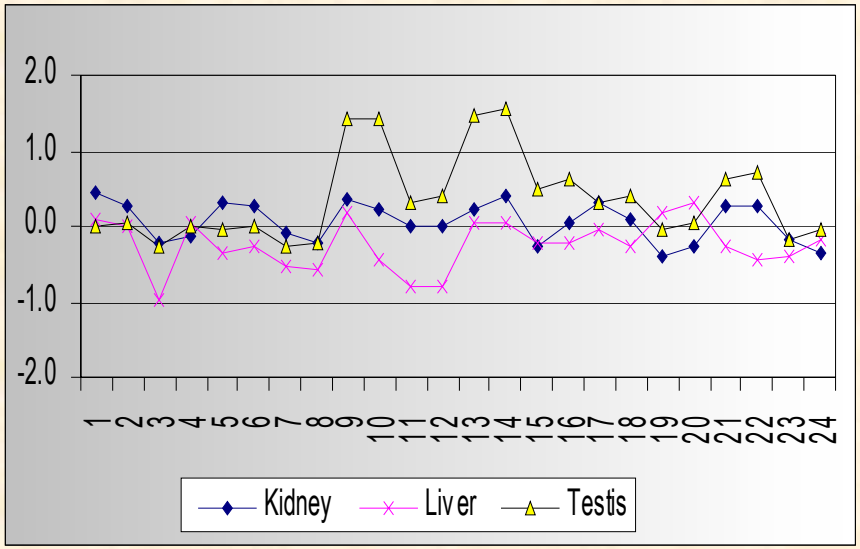
- Missing data points were treated as parameters in our model and were completed during estimation using data augmentation.

- 10,000 Markov chain Monte Carlo (MCMC) rounds were run (with additional burn-in rounds).

- The convergence of the chain was monitored by CODA and by inspecting the sample paths of the model parameters.

# Model A : Results

Figure 1: Plots of estimated posterior means for 24 arrays in three organs.




## Observations

- Array specific variations in the estimates.
- the estimates indicate an effect of dye on the observed log-ratio of intensities.
- No similar dye-pattern was observed from the Liver sample.
- Testis-samples indicated dye-effect and also possible mouse effect.

# Model A : Results

Table 1. Posterior estimates of precision parameters and proportions of genes in three precision groups (1,2,3).

| Parameter | Group | Kidney | Liver | Testis |
|---|---|---|---|---|
| Precision | 1 | 0.32 | 0.32 | 0.32 |
| | 2 | 2.95 | 2.95 | 2.95 |
| | 3 | 13.85 | 13.85 | 13.85 |
| Proportion of genes | 1 | *0.13* | *0.12* | *0.08* |
| | 2 | *0.41* | *0.39* | *0.43* |
| | 3 | *0.46* | *0.49* | *0.49* |

## Notes

- Posterior distributions of the three precision parameters were quite disjoint.
- Estimated distributions were highly concentrated around the posterior mean
- Genes were assigned to precision groups quite distinctly.

# Model A : Results

Table 2. Cross tabulation of genes (in %) according to estimated precision groups in the three organs.

| % of genes | | (T,1) | (T,2) | (T,3) | (T,1) | (T,2) | (T,3) | (T,1) | (T,2) | (T,3) | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (K,1) | (L,1) | 1.4 | 4.3 | 0.2 | | | | | | | 6.0 |
| | (L,2) | 0.7 | 3.1 | 1.3 | | | | | | | 5.0 |
| | (L,3) | 0.4 | 0.9 | 0.9 | | | | | | | 2.2 |
| (K,2) | (L,1) | | | | 0.7 | 2.4 | 0.5 | | | | 3.7 |
| | (L,2) | | | | 2.3 | 10.4 | 6.8 | | | | 19.5 |
| | (L,3) | | | | 0.7 | 7.6 | 8.9 | | | | 17.2 |
| (K,3) | (L,1) | | | | | | | 0.8 | 1.2 | 0.5 | 2.4 |
| | (L,2) | | | | | | | 0.8 | 7.0 | 6.1 | 14.0 |
| | (L,3) | | | | | | | 0.1 | 6.2 | 23.8 | 30.1 |
| Total | | 2.5 | 8.3 | 2.4 | 3.8 | 20.5 | 16.2 | 1.7 | 14.4 | 30.4 | 100 |

K : Kidney
L : Liver
T : Testis

1 : High variation
2 : Moderate variation
3 : Low variation

## Observations

- About 75% of genes were estimated to have moderate or low variation in all three organs.
- For some genes, estimated variance classes varied across organs.
- Only 1.4% genes were estimated to have high variation in all samples.

# Model B

- We noted that some genes can be expressed differently in one organ compared to its average expression in all three organs.

- We noted that for several genes, for a particular organ, the observed log-ratio-of-intensities could be far away from the expected zero value.

- This indicates that the expression levels of these genes are higher or lower in the experimental sample from that organ than in the reference sample.

- This also indicates that for the same genes in one or both of the remaining organs the log-ratio-of-intensities might behave in opposite way than the first organ.

# Model B

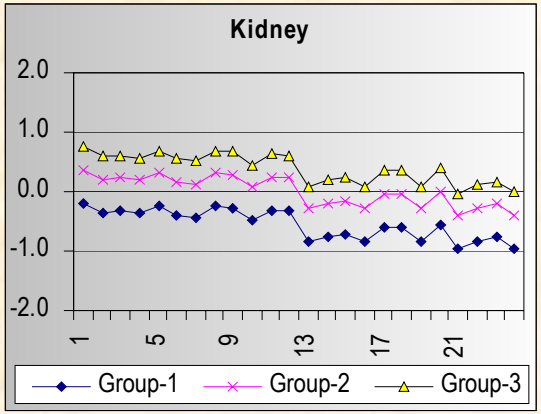
- Model continued to have array effects (as in Model A).
- Each gene was classified independently in each organ as having one of three possible expression groups ($d_{io}$).
- Accordingly each genes were assigned their group-effects ($\theta$).
- As before each gene was classified independently for each organ with respect to its corresponding residual variance ($c_{io}$).
- Conditional distribution of the log-ratio-of-intensities $I_{ioj}$ is assumed to be given by (with i, o, j as before),

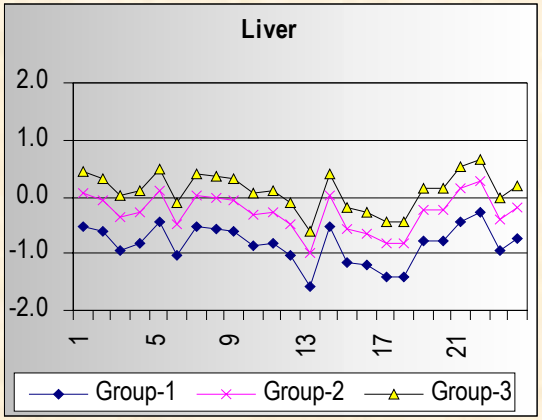
$$I_{ioj} = \mu_{oj} + \theta(d_{io}) + e_{ioj} \text{ , where } e_{ioj} \sim N(0, 1/\tau(c_{io})).$$

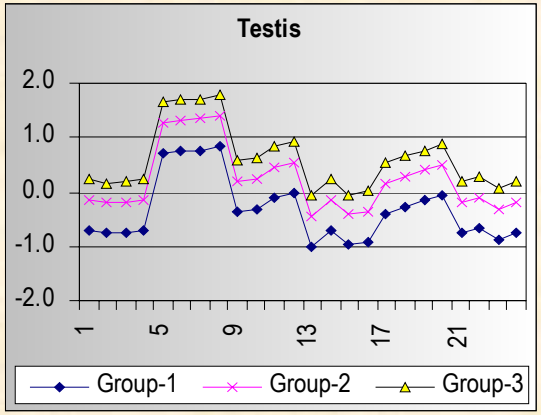
- Posterior density $p(\mu, \tau, c, \lambda_c, d, \lambda_d | I)$ is defined as before.

# Model B : Results

Figure 2: Plots of estimated posterior means for genes with three different group-effects (1-lower, 2-average, 3-higher) for 24 arrays.




**Note:** The posterior means for the group 2 were comparable to the average array effects obtained under Model-A. The other two groups, (group 1 and 3) correspond to a lower and a higher expression category respectively.

# Model B : Results

Table 3. Posterior estimates of precision parameters and proportions of genes in three precision groups (1,2,3) in the three organs (viz. Kidney, Liver and Testis).

| Parameter | Group | Kidney | Liver | Testis |
|---|---|---|---|---|
| Precision | 1 | 0.43 | 0.43 | 0.43 |
| | 2 | 4.24 | 4.24 | 4.24 |
| | 3 | 17.42 | 17.42 | 17.42 |
| Proportion of genes | 1 | *0.10* | *0.08* | *0.05* |
| | 2 | *0.33* | *0.35* | *0.36* |
| | 3 | *0.57* | *0.57* | *0.58* |

## Notes

- Each of the estimated precision parameters under Model B is higher than the respective ones under Model A.
- Additionally the estimated number of genes in the lower variance-class increased from Model A to Model B.
- Also the number of genes in higher variation class was reduced compared to Model A.

# Model B : Results

Table 4. Cross tabulation of genes ( in % ) according to their estimated precision groups (1,2,3) in the three organs.

| % of genes | | (T,1) | (T,2) | (T,3) | (T,1) | (T,2) | (T,3) | (T,1) | (T,2) | (T,3) | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (K,1) | (L,1) | 0.8 | 1.7 | 1.2 | | | | | | | 3.6 |
| | (L,2) | 0.5 | 2.7 | 1.0 | | | | | | | 4.2 |
| | (L,3) | 0.3 | 1.0 | 1.1 | | | | | | | 2.4 |
| (K,2) | (L,1) | | | | 0.6 | 1.2 | 0.5 | | | | 2.3 |
| | (L,2) | | | | 0.9 | 9.3 | 5.5 | | | | 15.7 |
| | (L,3) | | | | 0.6 | 6.0 | 7.6 | | | | 14.3 |
| (K,3) | (L,1) | | | | | | | 0.6 | 0.8 | 0.7 | 2.1 |
| | (L,2) | | | | | | | 0.8 | 6.0 | 8.0 | 14.8 |
| | (L,3) | | | | | | | 0.4 | 7.4 | 32.9 | 40.7 |
| Total | | 1.5 | 5.4 | 3.2 | 2.2 | 16.5 | 13.6 | 1.8 | 14.2 | 41.6 | 100 |

K : Kidney
L : Liver
T : Testis

1 : High variation
2 : Moderate variation
3 : Low variation

## Observations

- Under Model B, more genes were estimated to have moderate or low variation in all three organs, compared to A.
- For some genes, estimated variance classes still varied across organs .
- Even fewer number of genes (0.8%) were estimated to have high variation in all samples.

# Model B : Results

Table 5. Cross tabulation of genes according to their estimated group-effect classes (1-lower, 2-average, 3-higher) in the three organs (viz. Kidney, Liver and Testis).

| No. of genes | | (T,1) | (T,2) | (T,3) | (T,1) | (T,2) | (T,3) | (T,1) | (T,2) | (T,3) | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (K,1) | (L,1) | 7 | 26 | 631 | | | | | | | 664 |
| | (L,2) | 26 | 188 | 222 | | | | | | | 436 |
| | (L,3) | 106 | 87 | 11 | | | | | | | 204 |
| (K,2) | (L,1) | | | | 20 | 53 | 258 | | | | 331 |
| | (L,2) | | | | 38 | 982 | 653 | | | | 1673 |
| | (L,3) | | | | 159 | 433 | 50 | | | | 642 |
| (K,3) | (L,1) | | | | | | | 85 | 71 | 67 | 223 |
| | (L,2) | | | | | | | 119 | 481 | 93 | 693 |
| | (L,3) | | | | | | | 263 | 185 | 11 | 459 |
| Total | | 139 | 301 | 864 | 217 | 1468 | 961 | 467 | 737 | 171 | 5325 |

## Observations

- Some genes are estimated to have average expression in all three organs.
- Large numbers at the furthest off-diagonal entries of the three matrices support the hypothesis of differential expression across organs.

# Model B : Results

Table 6. Organ-wise cross tabulation of genes (in %) according to their estimated precision groups with the corresponding estimated group-effect classes.

Precision - v1 : low, v2 : moderate, v3 : high
Group effects - ge 1 : lower, ge 2 : average, ge 3 : higher

| Kidney | v 1 | v 2 | v 3 | Total |
|---|---|---|---|---|
| ge 1 | 7.1 | 9.3 | 8.1 | 24.5 |
| ge 2 | 2.0 | 13.9 | 33.8 | 49.7 |
| ge 3 | 1.0 | 9.1 | 15.7 | 25.8 |
| Total | 10.2 | 32.3 | 57.6 | 100.0 |

| Liver | v 1 | v 2 | v 3 | Total |
|---|---|---|---|---|
| ge 1 | 5.4 | 9.4 | 8.0 | 22.9 |
| ge 2 | 0.8 | 16.9 | 34.9 | 52.6 |
| ge 3 | 1.8 | 8.3 | 14.4 | 24.5 |
| Total | 8.0 | 34.7 | 57.3 | 100.0 |

| Testis | v 1 | v 2 | v 3 | Total |
|---|---|---|---|---|
| ge 1 | 3.7 | 7.5 | 4.2 | 15.5 |
| ge 2 | 1.6 | 20.2 | 25.3 | 47.1 |
| ge 3 | 0.1 | 8.4 | 29.0 | 37.5 |
| Total | 5.5 | 36.1 | 58.4 | 100.0 |

## Notes

- This model may be useful in additionally identifying genes with different expressions across organs.
- In all three organs, most of the genes with higher than average expression also have moderate or high precision.
- More than 70% of the genes with lower expression also have high or moderate precision.

# Model Comparison

Table 7. For Kidney data, cross tabulation of genes (in %) according to their estimated precision groups (1,2,3) under Model A with those under Model B.

| Precision group | | Model A | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | Tot |
| Model B | 1 | 8.6 | 0.7 | 0.0 | *9.3* |
| | 2 | 4.1 | 26.4 | 2.3 | *32.9* |
| | 3 | 0.0 | 15.3 | 42.4 | *57.8* |
| | Tot | *12.7* | *42.5* | *44.8* | *100.0* |

## Notes

- Recall that the precision groups have improved from Model A to Model B.
- 97% of the genes are estimated to be on the diagonal or in the sub-diagonal entry, implying improvement in precision from Model A to B.

# Model Comparison

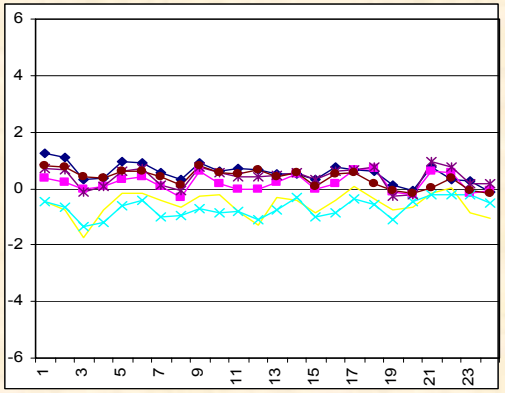
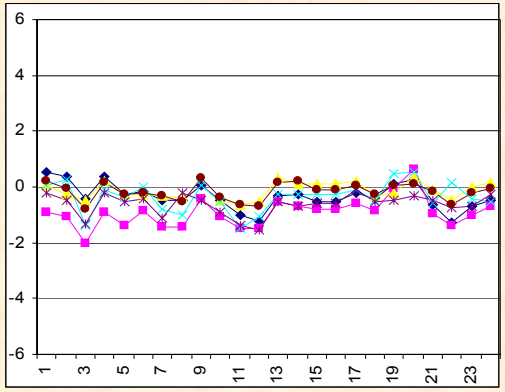
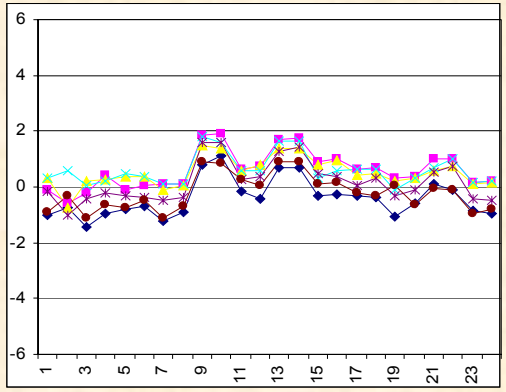
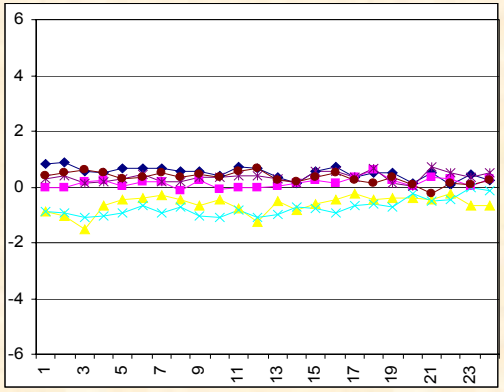
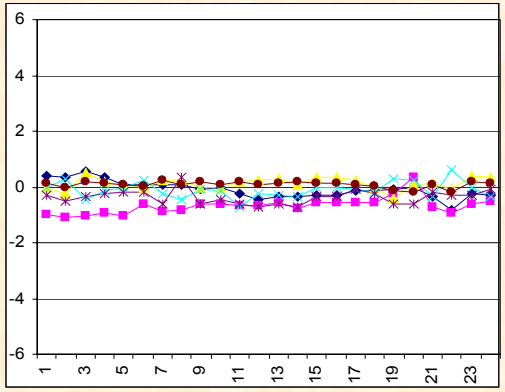
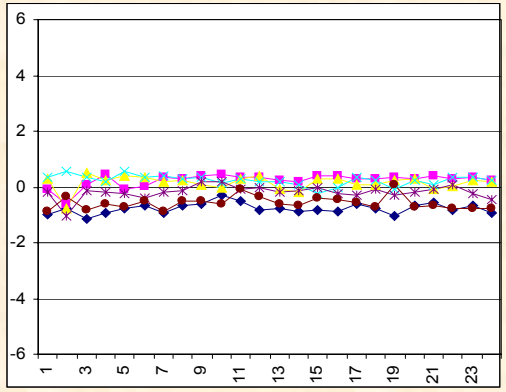
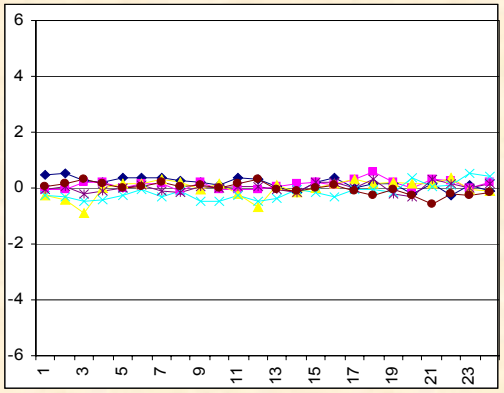
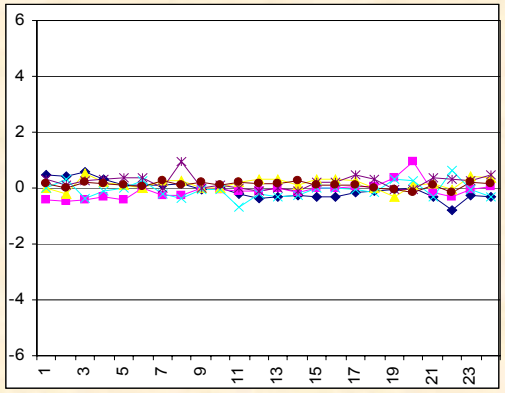
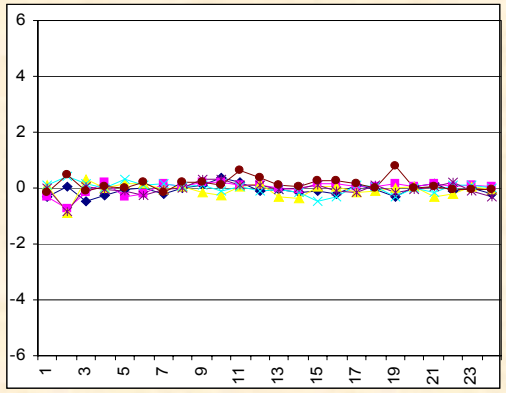
- In the following we give a brief example of how the two model works.

- Few genes were selected and log-ratio-of-intensities from each organ were plotted.

- Adjusted log-ratio-of-intensities under Model A show smoothening over arrays although model did not make any such assumption.

- Adjusted log-ratio-of-intensities under Model B show array-wise movement towards the origin resulting in narrowing of the plotted region.

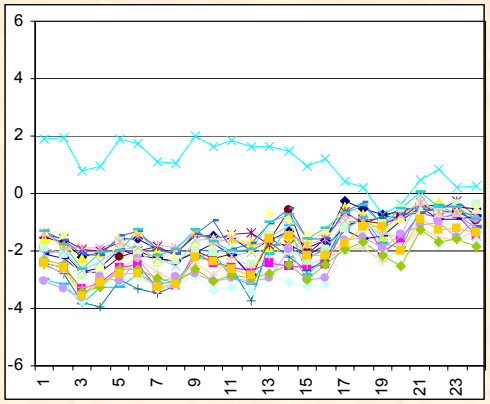
# Model Comparison

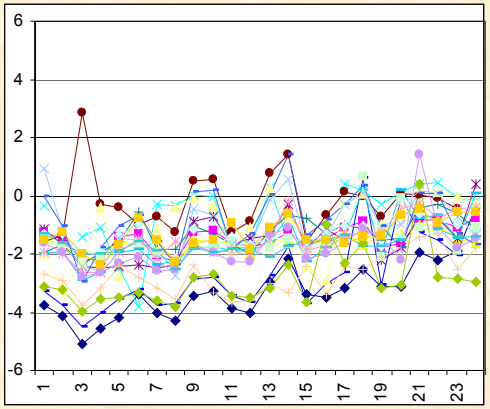
| Kidney data | Liver data | Testis data | |
|---|---|---|---|
|  |  |  | Original data |
|  |  |  | Model A |
|  |  |  | Model B |

# Model Comparison

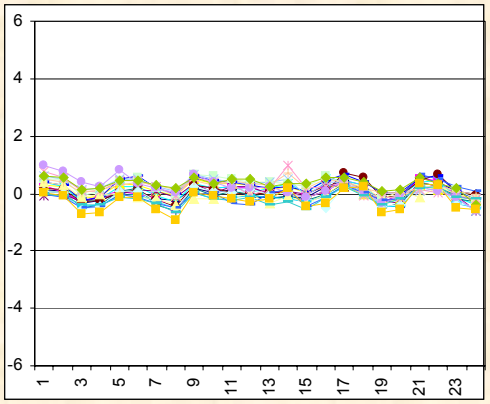
## Example from Kidney data

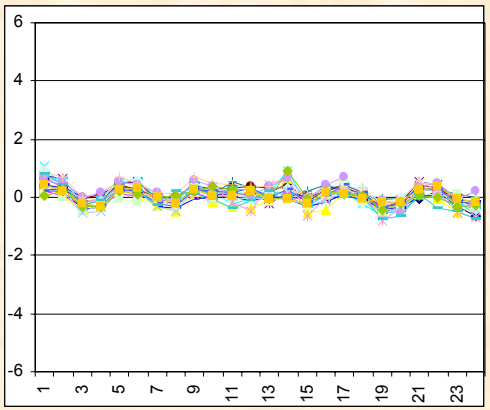
Genes estimated to have higher variance




Genes estimated to have moderate or low variance



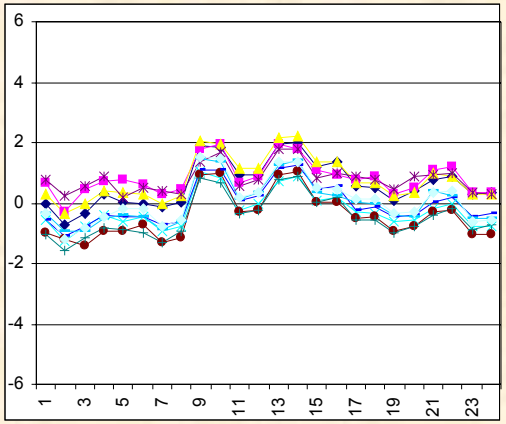




Mentioned by Pritchard et al. as highly varying

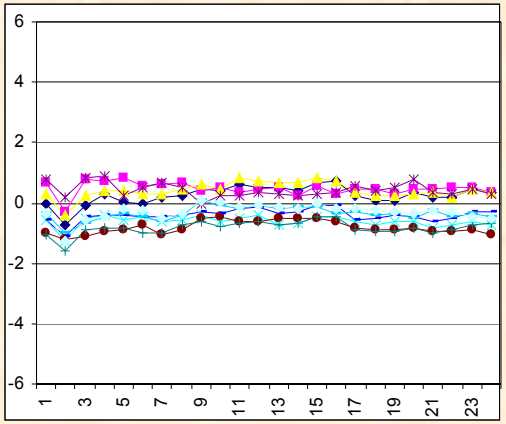
Not mentioned by Pritchard et al.

# Model Comparison

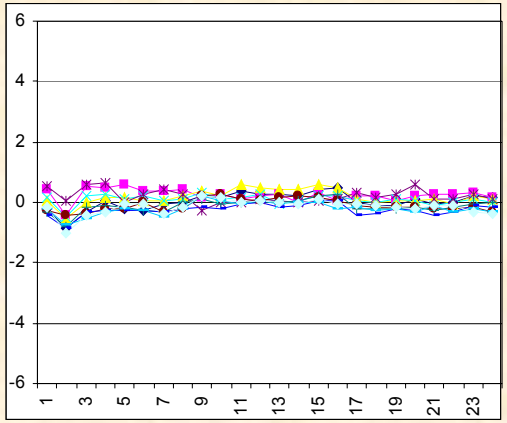
**Example from Testis data:** Plots for some genes mentioned by Pritchard et al. as highly varying




Plot of original data



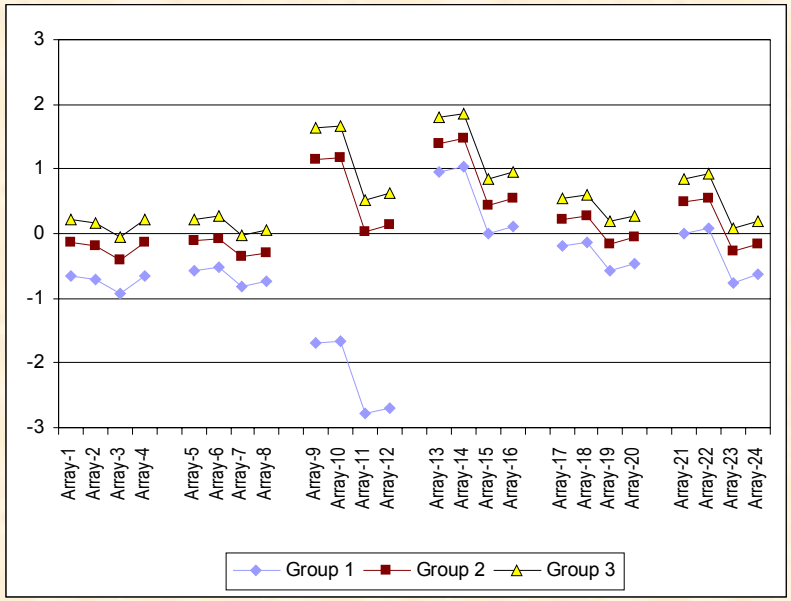

Plot of adjusted data under Model A



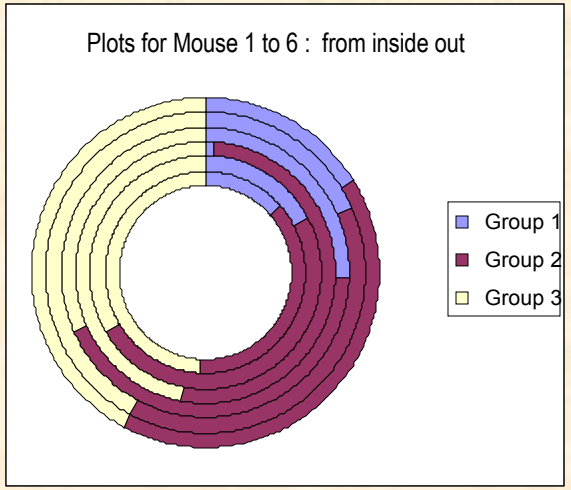

Plot of adjusted data under Model B

# Model Extension

## Mouse-specific model




Estimates from Testis data



# Concluding Remarks

- The approach is integrated in the sense that normalization and classification are being carried out simultaneously.

- Estimation of uncertainty is obtained along with classification, consequently it will be unnecessary to carry out a large number of testing of hypotheses.

- Model A takes into account normalisation for experimental factors distorting measurements of all genes on an array.

- Model B extends the previous model by incorporating available biological information.

- This approach of modelling can reduce dimension of the problem significantly.

- Extended models can be formed using information from one or both of biological and experimental factors influencing the observed data.