

The Latent Process Decomposition of cDNA Microarray Data



Simon Rogers and Mark Girolami

Bioinformatics Research Centre, Department of Computing Science, University of Glasgow

Luke Carrivick and Colin Campbell

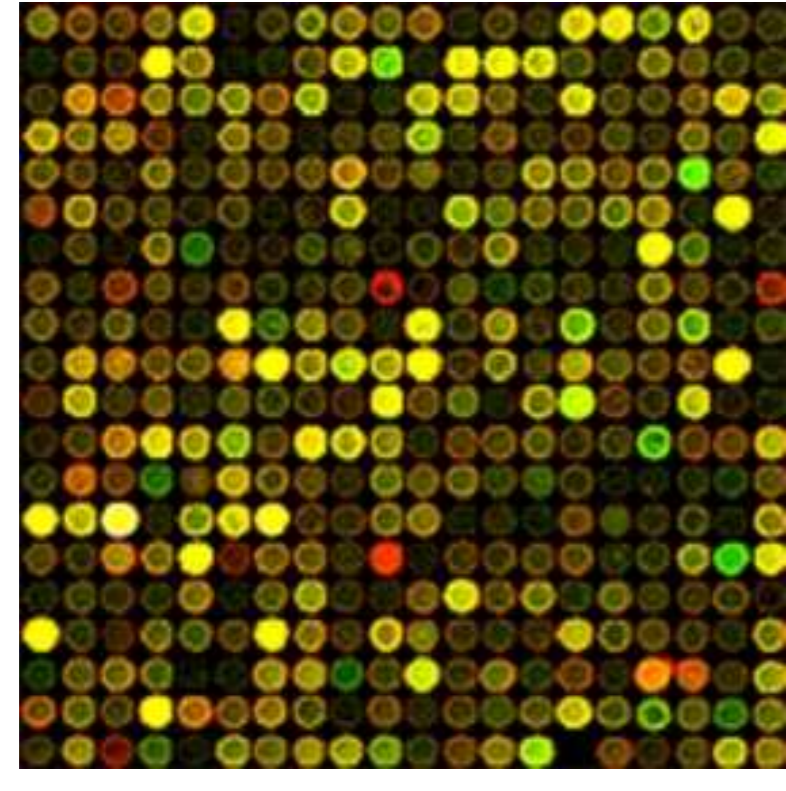
Department of Engineering Mathematics, University of Bristol

UNIVERSITY
of
GLASGOW

Introduction

We present a new computational technique which enables the probabilistic analysis of cDNA microarray data. An hierarchical Bayesian technique is introduced called Latent Process Decomposition (LPD) in which each sample (or gene) in the data set is represented as a combinatorial mixture over a finite set of latent processes that are expected to correspond to biological processes. Parameters in the model are estimated using efficient variational techniques. The technique has two main advantages over standard cluster analysis: Firstly, the ability to objectively assess the optimal number of processes, secondly, samples (or genes) may belong to several processes simultaneously - a more biologically relevant assumption than the standard hard clustering techniques. We demonstrate the power of the technique on two very different microarray datasets.

Aside - Microarrays

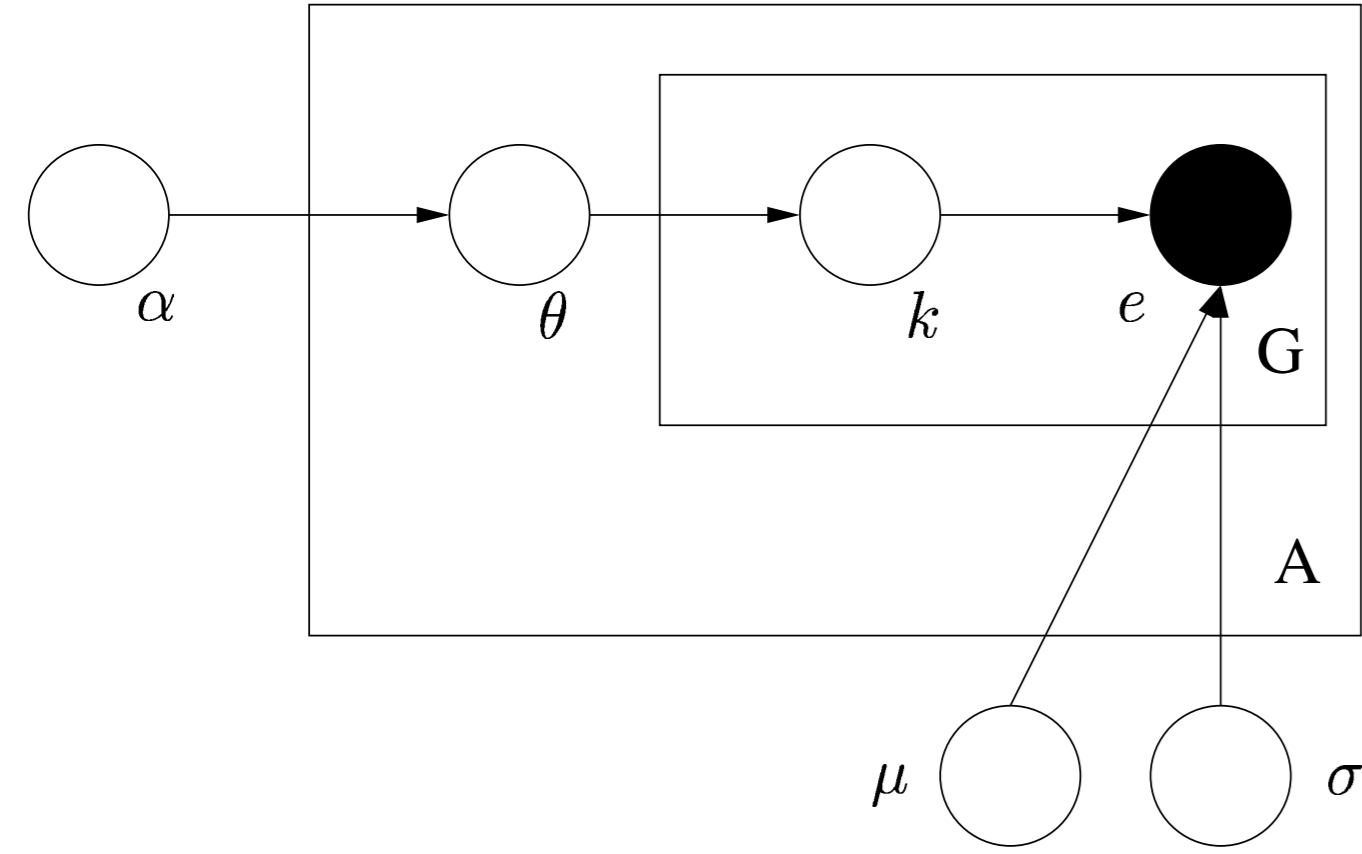


Microarrays facilitate the measurement of the *expression* of whole genomes simultaneously. Such measurements have the potential to provide crucial insights into the workings of organisms at a cellular level. Microarrays can be used to investigate the time-varying behaviour of genes (either due to cellular processes or response to external stimuli) or to measure how the expression of genes varies across tissue samples in different states (for example, healthy v disease or comparisons across different disease types or stages).

The results of a microarray experiment are usually provided as a $G(\text{genes}) \times A(\text{arrays})$ matrix of real values, e_{ga} . Obtaining useful biological information from such data has been the focus of much recent research.

Model Definition

The plates diagram for LPD can be seen below.



Given a set of arrays corresponding to an experiment, we can think of the following generative procedure:

- for each array a
 - Sample $\theta \sim \text{Dir}(\alpha)$
 - for each gene g
 - * Sample a process $k \sim M(\theta)$
 - * Sample the expression value $e_{ga} \sim \mathcal{N}(\mu_{gk}, \sigma_{gk})$

i.e., for each array, we sample a set of Multinomial parameters θ from a Dirichlet (parameterised by α). Then, for each gene in this array, we sample a process k according to $M(\theta)$ from which we sample the expression value e_{ga} . Note that the difference between this approach and a straightforward Gaussian mixture is in the repeated sampling of k for each gene rather than just sampling it once per array.

Inference

Marginalising over the hidden variables (k and θ) we obtain the following expression for the log-likelihood:

$$\log p(\mathcal{D}|\mu, \sigma, \alpha) = \sum_{a=1}^A \log \int_{\theta} \left\{ \prod_{g=1}^G \sum_{k=1}^K \mathcal{N}(e_{ga}|k, \mu_{gk}, \sigma_{gk}) \theta_k \right\} p(\theta|\alpha) d\theta \quad (1)$$

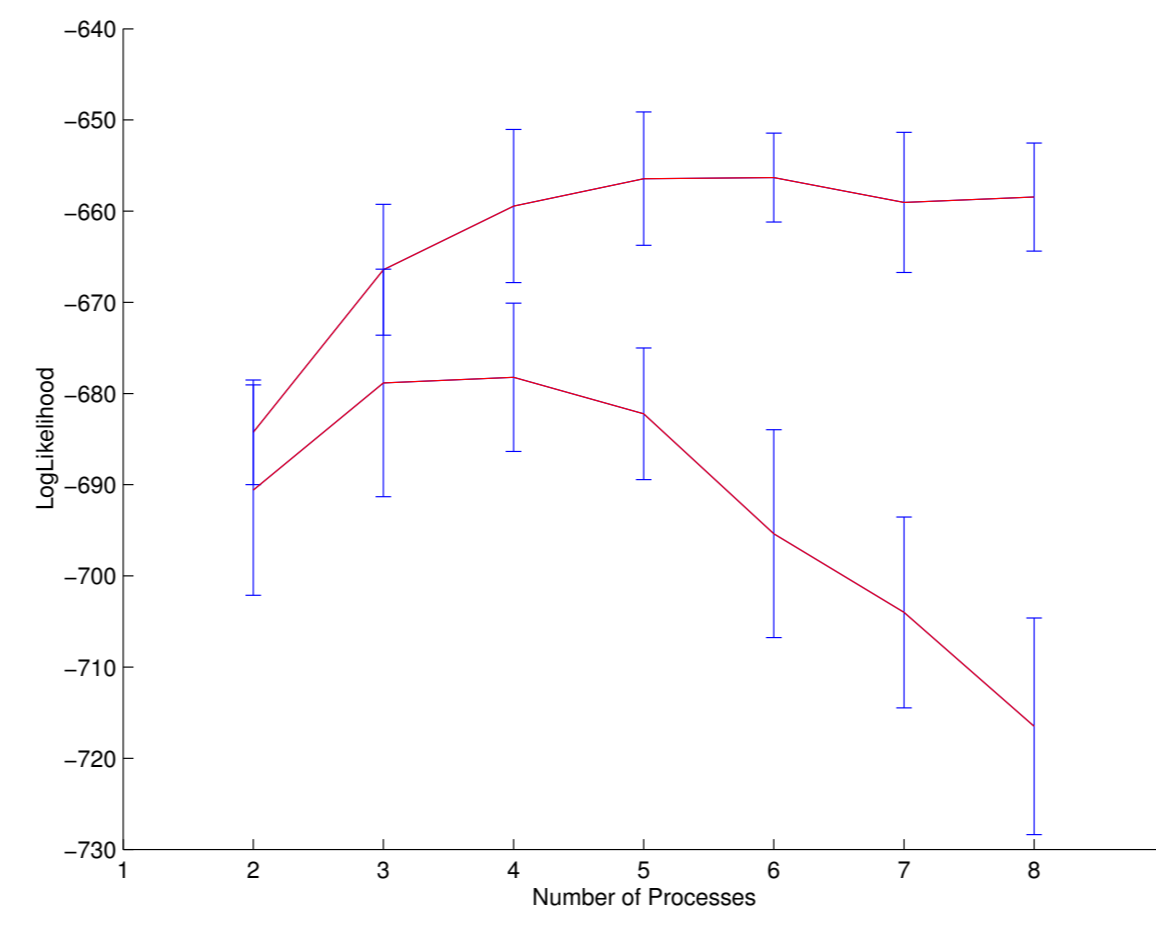
From which we can obtain the maximum-likelihood solution for μ_{gk} and σ_{gk} by twice lower bounding the expression via Jensen's inequality and the introduction of two sets of variational distributions:

- Q_{kga} - the probability that the expression for gene g in array a was produced by process k
- γ_a - parameters of an array specific Dirichlet

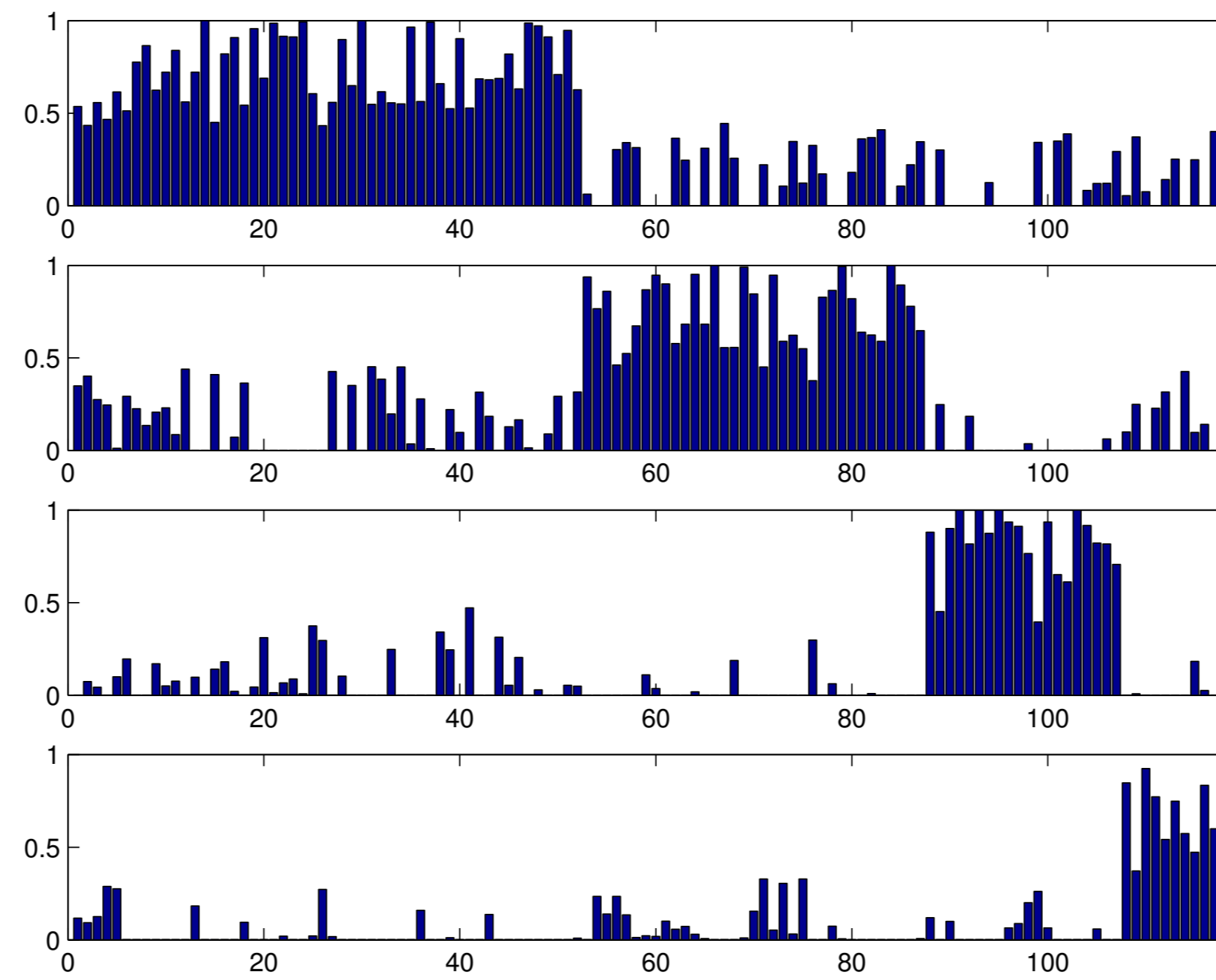
Note that it is trivial to obtain the MAP solution by placing suitable priors on μ_{gk} and σ_{gk} .

Breast Cancer Example

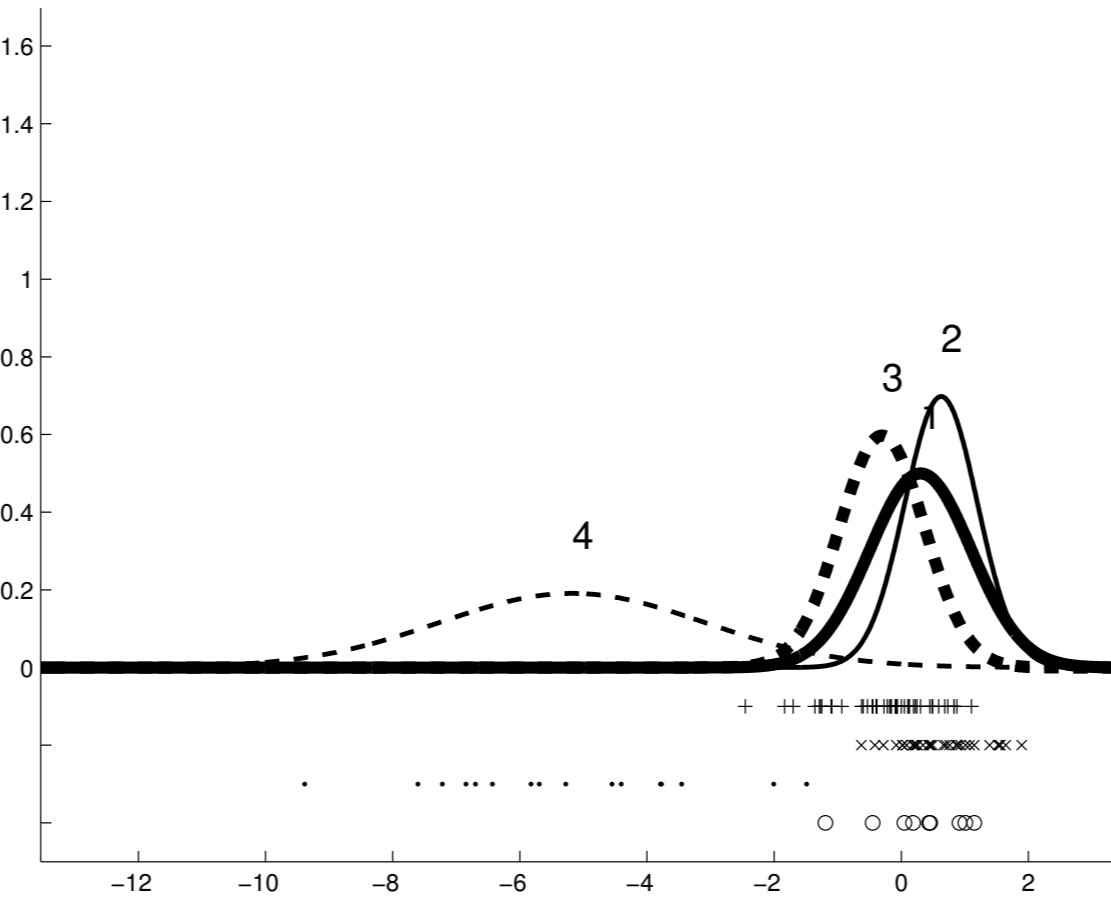
Using the data of Sorlie et al - consisting of expression of 534 genes across 115 breast carcinoma samples.



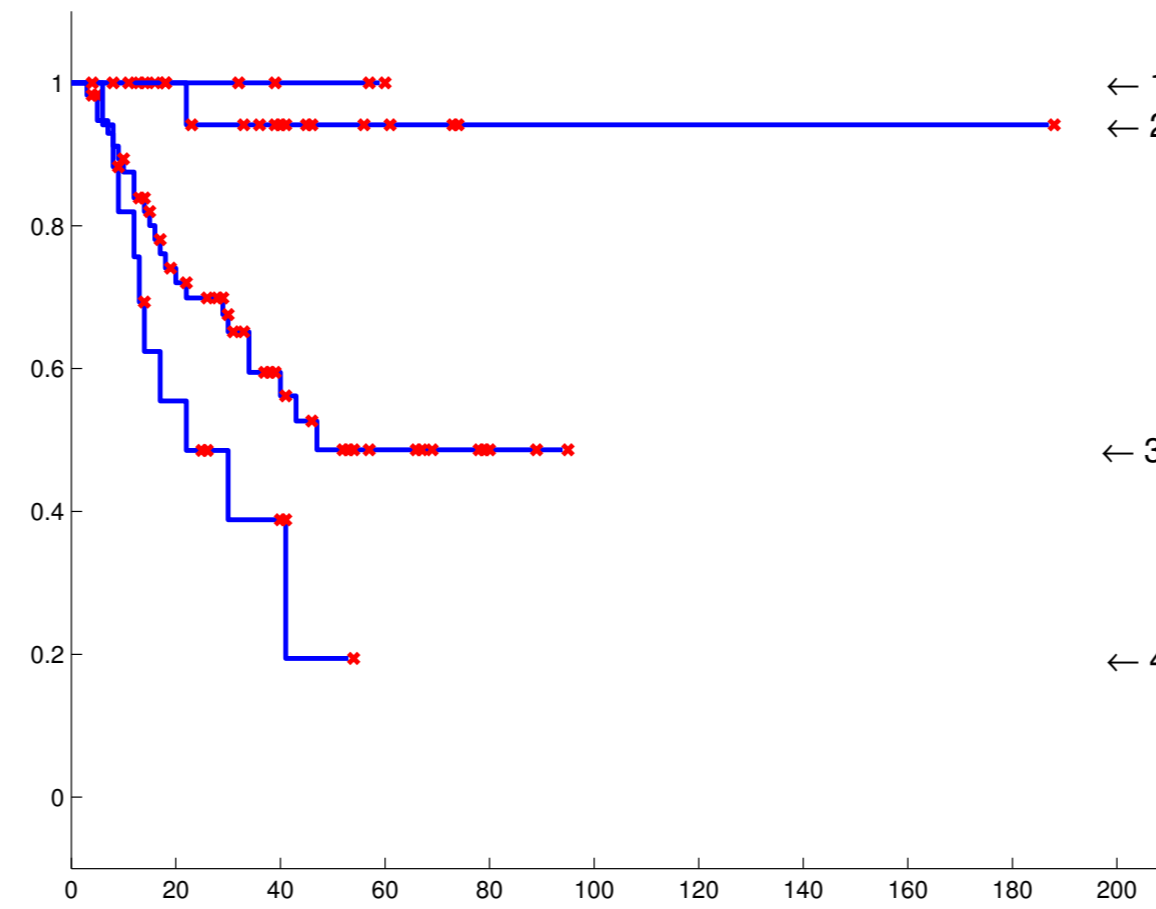
Held-out likelihood. Upper and lower curves are for the MAP and maximum likelihood solutions respectively. Both curves suggest in the region of $K = 4$ processes.



Normalised γ_a values for each of the 115 patients (columns) and 4 processes (rows). Each bar gives an indication of what proportion of expression values for a particular patient were due to a particular process.



Inferred densities for the FOXA1 gene for the 4 processes. Points plotted below show the actual values for patients assigned to each process. Note the consistent low expression for patients in process 4.



Kaplan-Meier plot showing the survival rates for patients in each process. Patients in process 1 have a far higher chance of survival than those in process 4. This suggests that genes similar to FOXA1 shown above would be excellent candidate for future investigation into this particular disease.

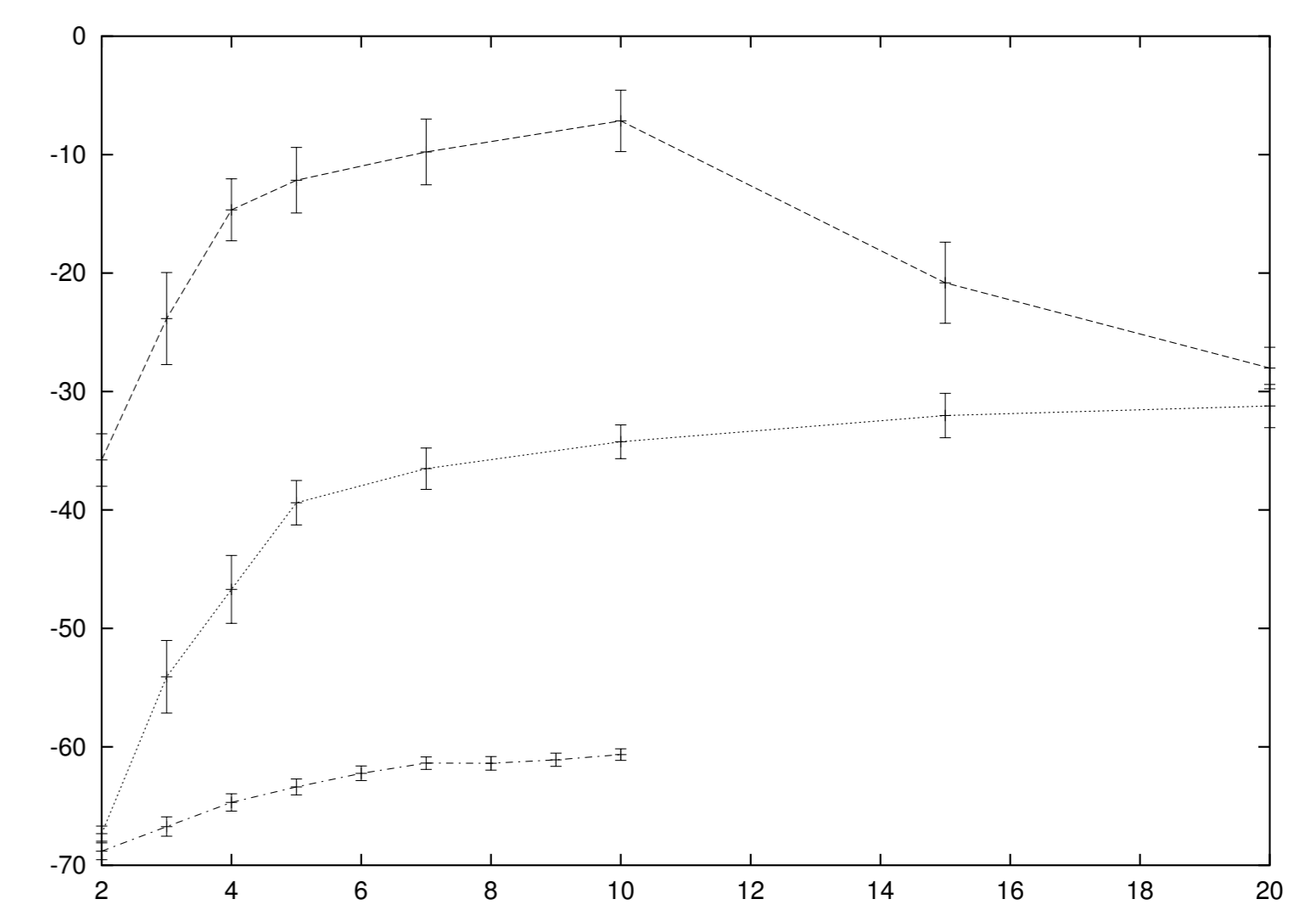
Similar results were found across several other Breast Cancer data-sets. Details given in Carrivick et al. (see references).

References

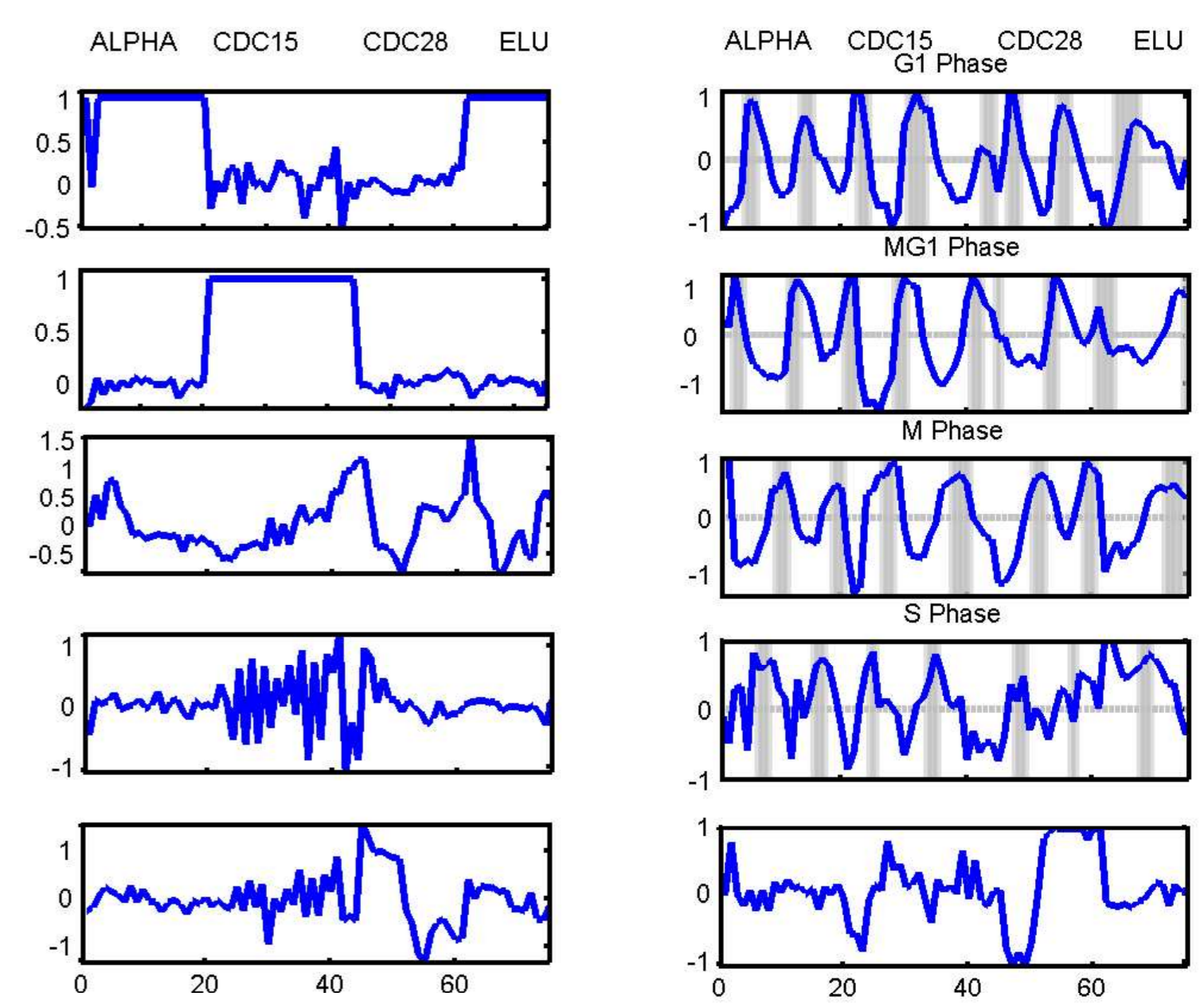
- The Latent Process Decomposition of cDNA Microarray Data. Rogers, S., Girolami, M., Campbell, C. and Breitling, R. (2005) *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2(2):143-156
- Identification of Prognostic Signatures in Breast Cancer Microarray Data using Bayesian Techniques. Carrivick, L., Rogers, S., Clark, J., Campbell, C., Girolami, M. and Cooper, C. (2005) *Journal of the Royal Society Interface* in press.
- Decomposing Gene Expression into Cellular Processes. Segal, E., Battle, A. and Koller, D. (2003) *Porc. Eighth Pacific Symp. Biocomputing (PSB)* pp. 89-100

Time Series Example

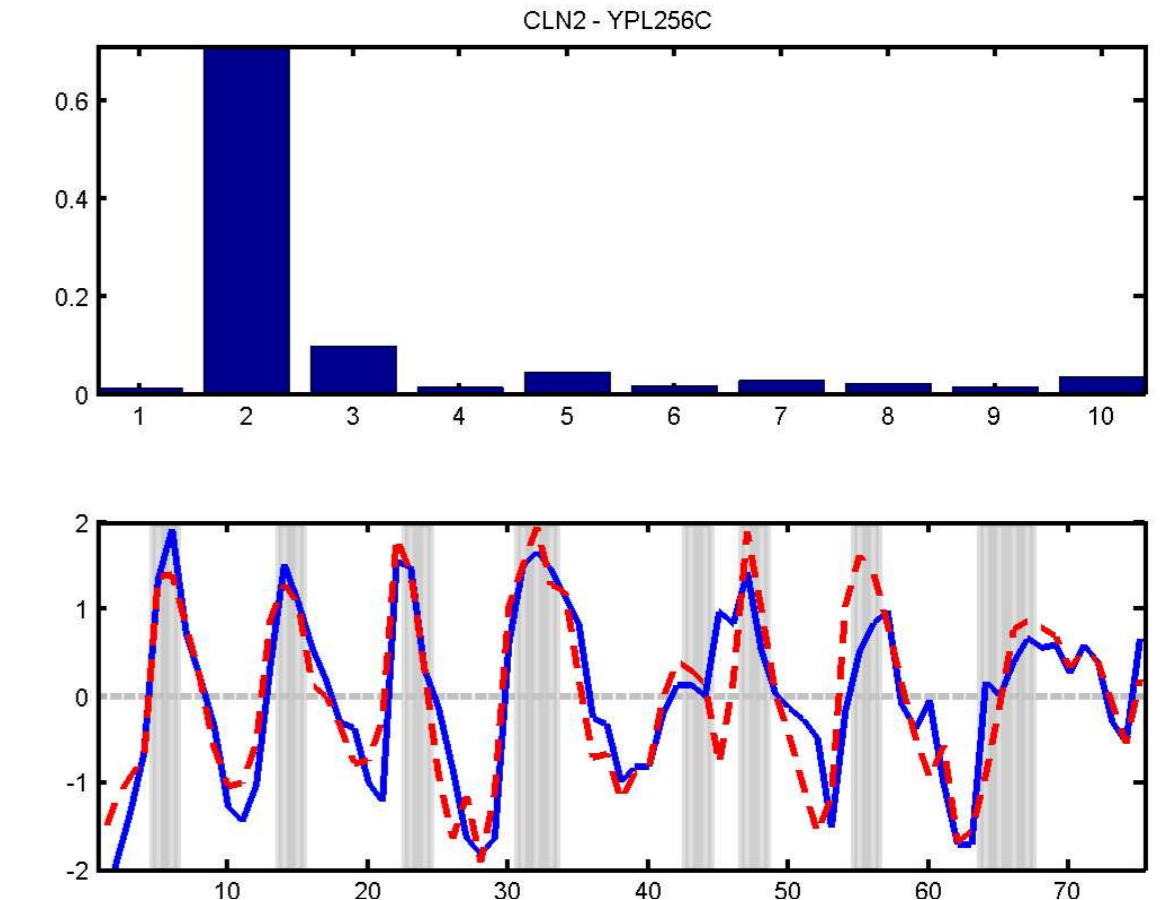
Using Spellman's yeast cell-cycle dataset. Now we decompose over genes rather than arrays - see Rogers et al for more information.



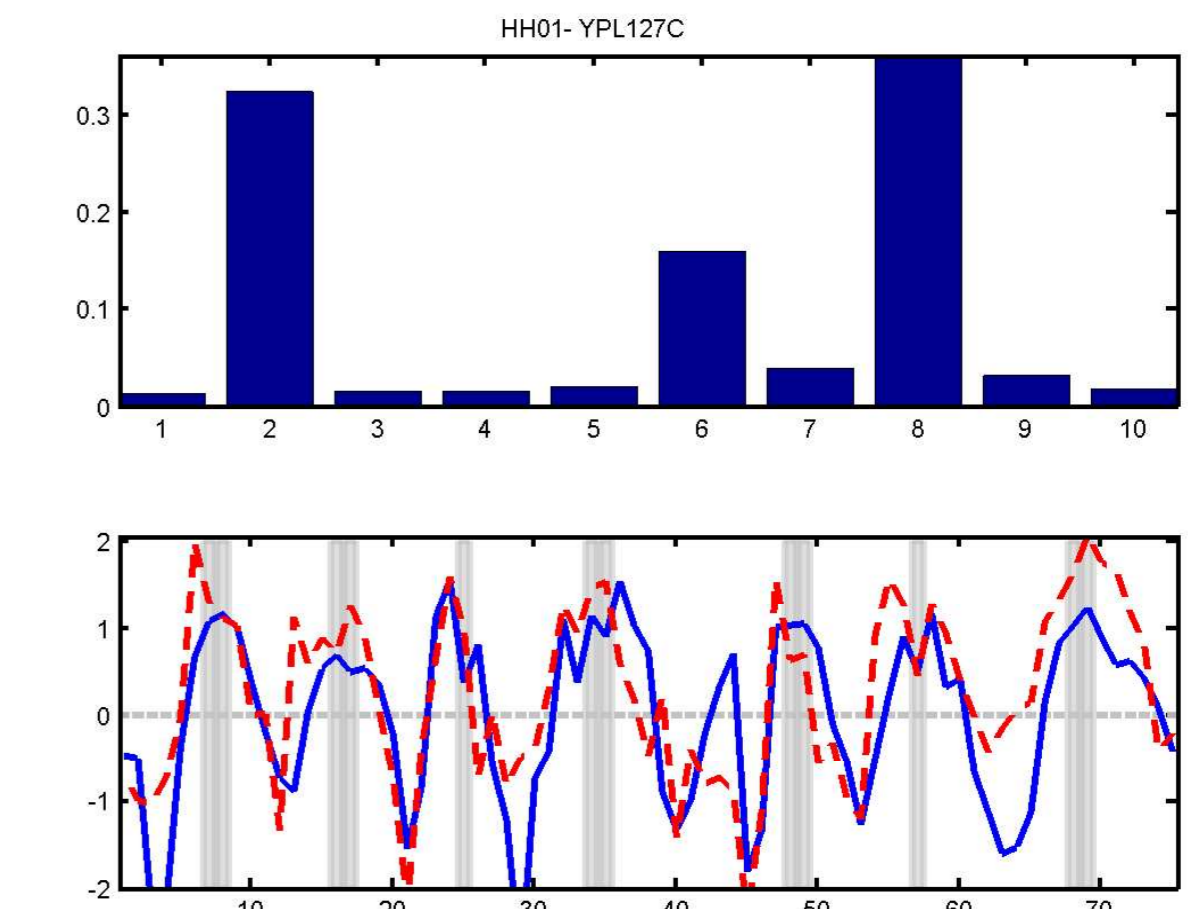
Held-out Likelihood versus the number of processes for LPD (top), Naive-Bayes Mixture (middle) and model of Segal et al. (bottom). LPD curve suggests a decomposition into $K = 10$ processes is sensible.



Process means for $K = 10$ process decomposition. 4 components corresponding to different phases of the cell-cycle are clearly visible.



Reconstructed (dotted line) and actual (solid line) expression of gene CLN2 - a gene known to express strongly in the G1 phase. As expected the gene is strongly dominated by process 2 which we have already seen corresponds to the G1 phase.



A more complicated example. This gene is known to peak in S phase and we see this in the large weight from process 8. However, we also have contributions from processes 2 (G1 phase) and 6 (M phase) reflecting the fact that the measured expression extends well beyond just the S phase.

Conclusions

Our experimental results strongly indicate that LPD is an effective technique for analysing microarray data. In the cancer example, LPD was able to decompose patients into groups that had strongly different survival characteristics and highlight the genes that were differentially expressed in these groups. In the cell-cycle example, LPD was able to extract several independent cycling components and identify genes associated with these components.

Future work includes integrating the model with a probabilistic model for binding site motifs and in-depth analysis of both public-domain and new datasets.