Identification of Prognostic Signatures in Breast Cancer Microarray Data using Bayesian Techniques

L. Carrivick [a], S. Rogers [b], J. Clark [c], C. Campbell [a]*,
M. Girolami [b] and C. Cooper [c]
[a]Advanced Computing Research Centre, Queen's Building,
University of Bristol, Bristol BS8 1TR, United Kingdom
[b]Bioinformatics Research Centre, Dept. of Computing Science,
University of Glasgow, Glasgow G12 8QQ, United Kingdom
[c]Section of Molecular Carcinogenesis,
The Institute of Cancer Research,
Sutton, SM2 5NG, United Kingdom

Abstract

We apply a new Bayesian data analysis technique (Latent Process Decomposition) to four recent microarray datasets for breast cancer. Compared to hierarchical cluster analysis, for example, this technique has advantages such as objective assessment of the optimal number of sample or gene clusters in the data, penalisation of overcomplex models fitting to noise in the data and a common latent space of explanatory variables for samples and genes. Our analysis provides a clearer insight into these datasets, enabling assignment of patients to one of four principal processes, each with a distinct clinical outcome. One process is indolent and associated with under-expression across a number of genes associated with tumour growth. One process is associated with over-expression of *GRB7* and *ERBB2*. The most aggressive process is associated with abnormal expression of transcription factor genes, including members of the *FOX* family of transcription factor genes.

Keywords : breast cancer, microarray data, cluster analysis

1 Introduction

Evidence from epidemiological studies, analysis of tumour progression and variability in response to treatment all indicate considerable diversity among human breast cancers. This view is supported by various independent microarray studies [6, 11, 12, 13, 21, 24, 26, 27]. For example, with one recent study [25], hierarchical cluster analysis suggested the existence of five major categories of breast cancer. Two groups of predominantly estrogen receptor positive (ER+) cancers had expression patterns similar to breast luminal cells (called luminal A and B). For the ER− cancers, three additional categories were identified that overexpressed genes associated with the *ERBB2* amplicon at 17q22, had a basal cell expression pattern or resembled normal breast tissue. The significantly different clinical outcomes of 4 of these groups (luminal A, luminal B, basal and *ERBB2*) highlighted the potential biological importance of this classification. Although these groups could be broadly defined, the fine structure of dendrograms varied between individual cluster analysis methods and the authors concluded that the observed high level branching was not always a reflection of biologically meaningful relationships.

*Address for correspondence: C.Campbell@bris.ac.uk

In this paper we will use a new Bayesian approach for finding informative structure in such datasets. This approach is called Latent Process Decomposition (LPD) [23] and it is modelled on the Latent Dirichlet Allocation method of Blei *et al* [2]. In the derived model each sample (or gene expression measurement) is represented as a combinatorial mixture over a finite set of latent processes (a *process* is an assumed functionally related set of samples or genes). Observations are not necessarily assigned to a single cluster. This reflects a prior belief that a number of processes could contribute to a given gene expression level or that a tumour could have a heterogeneous structure because it overlaps several defined states. By contrast, most cluster analysis methods use an implicit mutual exclusion of classes assumption, though several algorithms which avoid this assumption have been proposed recently [10, 19, 3]. The proposed approach has other advantages. For example, the optimal number of sample or gene clusters can be objectively assessed. Also samples and gene expression levels are modelled using a common space of explanatory variables. This is in contrast to the use of dendrograms where samples and gene expression values are typically clustered separately, amounting to two distinct reduced space representations which are not easily related. LPD can also readily handle missing values. Finally LPD has the advantage that we can incorporate a prior belief that experimental noise exists and thus use a Bayes prior penalising overcomplex models which would fit the noise. LPD also compares favourably to various cluster analysis methods [23]. To illustrate its potential we apply this approach to breast cancer datasets from Sorlie *et al* [25], West *et al* [27], van 't Veer *et al* [26] and de Vijver *et al* [6]. The method appears to give clearer insights into these datasets suggesting at least 4 principal processes, each associated with a different clinical outcome. The results presented in the next section derive from a variational approach to Latent Process Decomposition described in Appendix 2 (the reader is referred to Rogers *et al* [23] for a full description). To support these results we have additionally used a Markov Chain Monte Carlo (MCMC) approach to LPD, described in Appendix 2. The latter proved more computationally demanding than the variational approach, but gives a very similar picture.

2 The Application of Latent Process Decomposition to four Microarray Datasets for Breast Cancer

2.1 Sorlie *et al* dataset.

From the study of Sorlie *et al* [24] we used data from 115 primary breast carcinoma samples (labelled Norway/Stanford and very predominantly of invasive ductal type) and we used the same set of 534 genes selected in their study. In Figure 1 we give the log-likelihood curves for both a maximum likelihood and MAP (maximum a posterior) model using variational LPD [23]. For the maximum likelihood model the log-likelihood has an approximate peak at about 4 processes indicating this is a suitable number of processes to use. For the MAP model (Figure 1, upper curve) a Bayesian prior has been used to penalise construction of an over-complex model. The log-likelihood rises to a plateau after which no further gain is to be made by introducing further processes since the model will not exploit this extra freedom. In contrast, for the maximum likelihood solution, the log-likelihood falls as further processes are introduced since the algorithm will use these and construct an over-complex model.

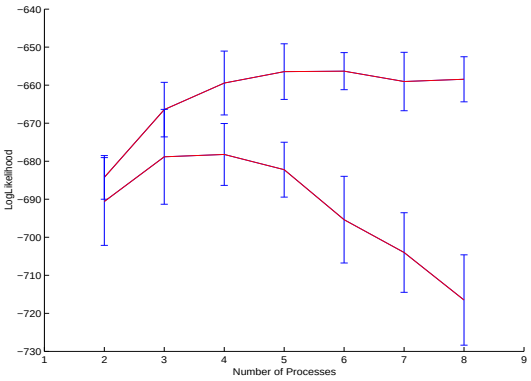


Figure 1: The log-likelihood (y-axis) versus number of processes (x-axis) using the MAP solution (upper curve) and maximum likelihood solution (lower curve) for the Sorlie *et al* dataset Stanford/Norway dataset [24].

Using a 4 process model we can derive the decomposition diagram in Figure 2 where the peaks represent the confidence that sample a is assigned to process k (these peaks are given by normalised γ_{ak} parameters, see Appendix 2, equation (4) for further details). Unlike most cluster analysis methods, samples can belong to several processes simultaneously.

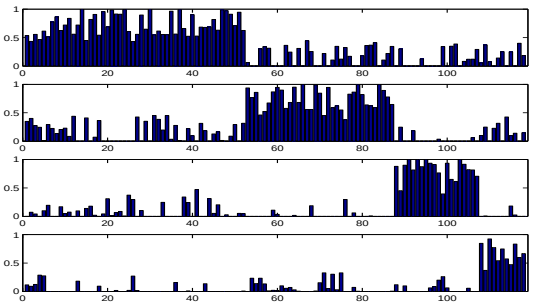
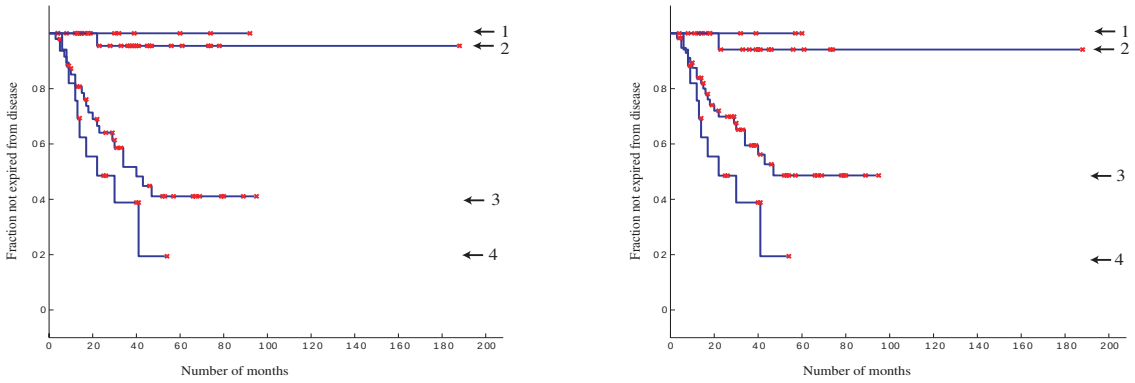


Figure 2: Decomposition diagram derived from LPD for the dataset of Sorlie *et al*. The top process is identified with the trend curve 3 in Figure 3(a), the second process is identified with 2, the third with 4 and the lowest is identified with the indolent process 1 in Figure 3(a).

We have used a threshold of 0.5 for assignment of sample a to process k and determined the corresponding Kaplan-Meier plot in Figure 3(a). The separation is more distinct than that made by the original authors [25] with one indolent subtype and three aggressive subtypes indicated.



(a) KM1

(b) KM2

Figure 3: Kaplan-Meier plots for the Sorlie *et al* dataset. The graphs show fraction not expired from the disease (y-axis) versus number of months (x-axis). For KM1 (left) there are 9 patients in process 1, 32 in 2, 48 in 3 and 18 in 4 (the remaining 8 samples are insufficiently identified with a process). A vertical drop indicates expiry from the disease and a star indicates the patient is not recorded as expired from the disease (this includes the point at which some patients exited the survey). KM2 corresponds to a different initialisation of the algorithm (see text) with 7, 23, 58 and 18 patients assigned to processes 1 to 4 respectively. With different initialisations there is some variability in the assignment of patients to processes 1 to 3, though process 4 remains quite distinct with 18 patients usually assigned, both using the variational LPD used here and the alternative MCMC approach described in Appendix 2 (see Figure 18(a)).

The likelihood function is not concave (local maxima can exist). Local maxima correspond to models with good fits to the data with the intervening regions in model space corresponding to poorer fits. Nevertheless, it is likely that models with good fits are sharply concentrated in model space. However, this does mean different initialisations of the algorithm can give different solutions. In fact, since many peaks in Figure 2 are near 0.5, the Kaplan-Meier plot is the most sensitive result dependent on this effect. Figure 3(b) is a typical result from a different initialisation in which some patients have moved between the outcome trends. To investigate this issue we restarted the algorithm with 50 randomly constructed initialisations and found that 32 of these gave a Kaplan-Meier plot in which no patient had expired from the disease in process 1. Furthermore, these 32 solutions had a distinctly higher average log-likelihood than those solutions with at least one patient expiring from the disease in process 1, indicating they are more appropriate models (Figure 4).

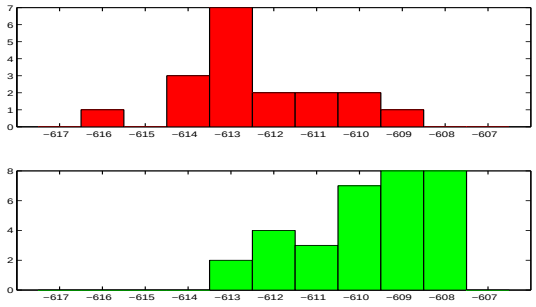
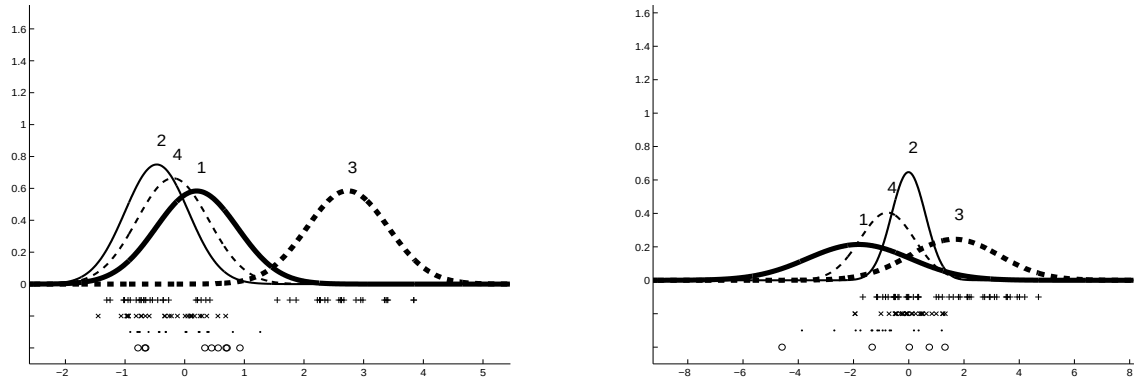


Figure 4: With 50 random initialisations, 32 instances gave Kaplan Meier plots with a purely indolent process 1 (lower histogram) and 18 cases had at least one patient expiring from the disease (upper histogram). The x-axis gives the value of the log-likelihood and the y-axis the frequency of occurence. Solutions with a purely indolent process 1 gave a higher average log-likelihood indicating they give a better fit to the data.

Apart from identifying samples with processes, LPD can be used to identify those genes which are most prominent in distinguishing processes. From the algorithm (equations (5,6,11,12) in Appendix 2), we can determine a mean μ_k and standard deviation σ_k for each process k and hence inferred density curves (estimating amount of data in a region). An example of two density curves is given in Figures 5(a) and 5(b). These density curves are derived from the dataset taken as a whole and are not one-dimensional fits to the expresssion values for that gene. We can thus use a score $Z_1 = |\mu_1 - \mu_2|/\sqrt{\sigma_1^2 + \sigma_2^2}$ to rank genes distinguishing processes 1 and 2, for example, and this score follows a normal probability distribution with $\mathcal{N}(0, 1)$. Apart from comparing two processes we could also compare one process with the rest e.g. by using the lowest pairwise Z_1-score. Unfortunately this score can be adversely influenced by large variances. Thus the gene depicted in Figure 7(a) does not score well because it has a large variance in the denominator of Z_1. Consequently we will also use a second, rank-based, score (based on the Mann-Whitney test [22]) to highlight such cases. This score will be denoted Z_2 and quantifies the probability of observing a sequence of ranked and labelled datapoints (ranked by expression level and labelled 1 (process of interest) or 2 (other processes)).
No single gene is a particularly distinct marker for process 1. However, of the top 20 ranked genes distinguishing process 1 from the rest, all but one exhibit relative under-expression in process 1. For the three aggressive processes (2-4), process 4 has the most distinctive genes and process 2 the least distinctive (the highest ranked gene is *LIV-1*). Using the Z_1-score the most distinctive gene in process 3 is *GRB7*, depicted in Figure 5(a). It has a score $Z_1 = 3.84$ ($p = 0.00006$) with only $Z_1 = 1.59$ ($p = 0.06$) for the next highest ranked gene (*PAPSS2*). *GRB7* is an adaptor-type signaling protein which is recruited via its SH2 domain to a variety of receptor tyrosine kinases (RTKs), including *ERBB2* and *ERBB3*. It is overexpressed in breast, esophagal and gastric cancers, and may contribute to invasiveness potential [20]. It is frequently co-amplified with *ERBB2* (*HER2*) in breast cancer and from Figure 5(b) we see that *ERBB2* is, indeed, only overexpressed in process 3.



(a) *GRB7* (b) *ERBB2*

Figure 5: Inferred densities for *GRB7* and *ERBB2* for the Sorlie *et al* dataset, with + the expression values for samples identified with process 3. Though only over-expressing in process 3 a subset of samples do not over-express *GRB7* suggesting a possible subprocess within this process. In this and subsequent figures individual expression values are marked ∘ if the samples are associated with process 1, × with 2, + with 3 and · if associated with process 4.

Process 4 has the most distinctive set of genes. In agreement with previous observations [25], this process has basal cell characteristics e.g. cytokeratin 5 appears up-regulated. Using the Z_1 score the top ranked gene distinguishing process 4 is *FLT1* (*VEGFR1*) (Figure 6). *VEGFR1* (especially its soluble isoform) is a negative regulator of vascular endothelial growth factor availibility. Indeed, *VEGFR1* overexpression is associated with improved survival in breast cancer [28]. Estrogen mediated decrease in *VEGFR1* expression can cause increased angiogenesis leading to enhanced breast tumour progression [9].

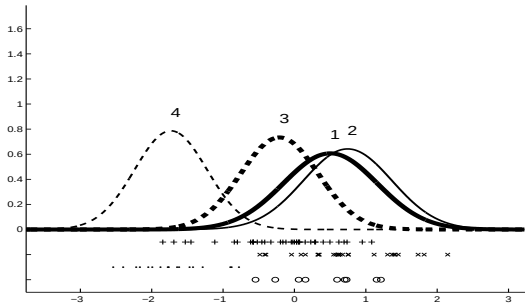


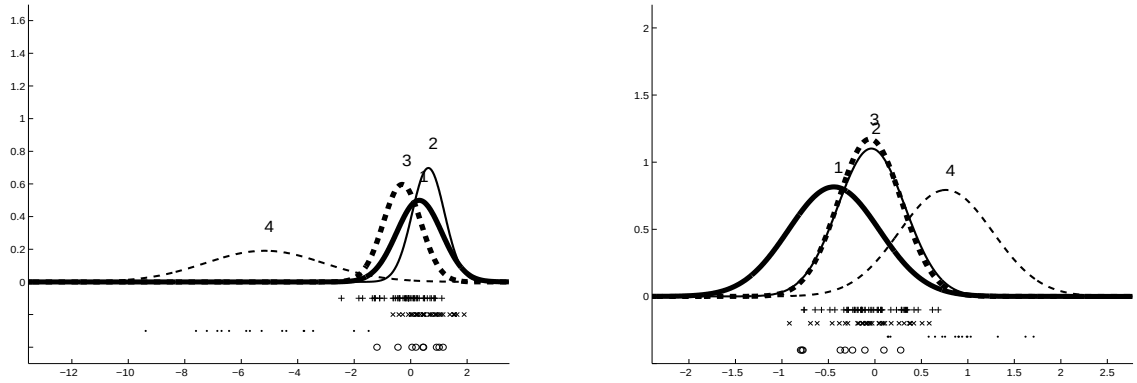
Figure 6: Inferred densities for *FLT1* (*VEGFR1*) in process 4 with · denoting the corresponding expression values.

The second ranked gene by Z_1-score is *MAFG* which is associated with upregulation of protective anti-oxidant enzymes under cellular conditions of oxidative stress [14]. Third ranked is *FOXC1*, a gene which expresses a forkhead transcription factor. The fourth ranked gene is *XBP1* expressing an X box binding protein and the fifth ranked gene expresses *AD021* protein. In the table below we list the top 12 probes ranked by the Z_2 score for process 4.

Rank	Gene	Z_2-score	Expression
1.	*TFF3*	6.35	Under
2.	*FOXC1*	6.32	Over
3.	*FOXA1*	6.30	Under
4.	*XBP1*	6.25	Under
5.	*GATA3*	6.11	Under
6.	*B3GNT5*	6.08	Over
7.	*FLJ14525*	6.05	Over
8.	*FLT1*	6.04	Under
9.	*GALNT10*	5.95	Under
10.	*FOXC1*	5.88	Over
11.	*FBP1*	5.76	Under
12.	*GATA3*	5.68	Under

Table 1: The top ranked genes distinguishing process 4 by Z_2-score for the dataset of Sorlie *et al*. Z_2 follows a normal distribution with $\mathcal{N}(0,1)$ thus the associated probabilities of occurrence are upper bounded by 10^{-8} reflecting the fact that the ordering of expression values for process 4 against the set of expression values for the other processes is highly improbable according to a null hypothesis. In the original data the *FOXC1* clone is annotated as *FLJ11796* and *FOXA1* as *HNF3A*.

FOXA1 and *FOXC1* are members of the forkhead family of transcription factor genes (Figure 7).



(a) *FOXA1* (b) *FOXC1*

Figure 7: *FOXA1* (*HNF3A*) underexpresses while *FOXC1* overexpresses in process 4 ($\cdot$ denotes the expression values in process 4).

FOXA1, *GATA3* and *XBP1* encode transcription factors and their roles and association with the estrogen receptor-α gene (*ESR1*) and trefoil factors (*TFF3* and *TFF1*) are reviewed by Lacroix and Leclerq [15].
In Appendix 1 we give the original dendrogram decomposition reported in Sorlie *et al* [25] along with the assignment to processes given in Figure 2. Sorlie *et al* [25] labelled a subset of the tumour samples as Luminal A and B, ERBB2+ and Basal. Their 18 Basal tumours match the 18 Process 4 samples. Indeed, we shall later see that this process is very distinctive. Elsewhere LPD labels a wider range of samples than labelled by Sorlie *et al* (though this would depend on the threshold

chosen for the significance of the peaks in Figure 2). Their 11 Luminal B and 11 ERBB2+ are exclusively subsets of process 3, while their 28 Luminal A are exclusively associated with processes 1 and 2. Indolent process 1 is exclusively sampled from some Luminal A samples and other samples which were left unlabelled in their study. If we use the MCMC-based approach to LPD we obtain a very similar picture (see Figure 18).

2.2 West *et al* dataset.

For the Affymetrix breast cancer dataset of West *et al* [27] we used data from 49 samples (exclusively derived from tumours of invasive ductal type) with 500 probes ordered using the p-values derived by the authors (though LPD can readily handle the full dataset, some feature selection is advisable since redundant information injects noise into the analysis). No survival data was available for this dataset, though time-to-metastasis was available. Nevertheless we can derive the corresponding MAP solution (Figure 8).

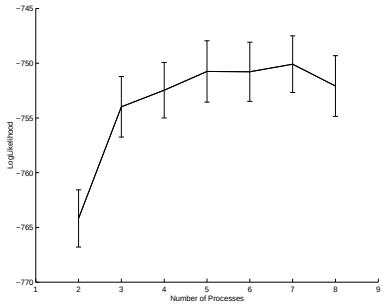


Figure 8: The log-likelihood (y-axis) versus number of processes (x-axis) using a MAP approach (right) for the West *et al* dataset.

The onset of the plateau is more ambiguous in this case and could indicate up to 5 processes. However, to conform with the analysis elsewhere we will use 4. We then get the following decomposition diagram:

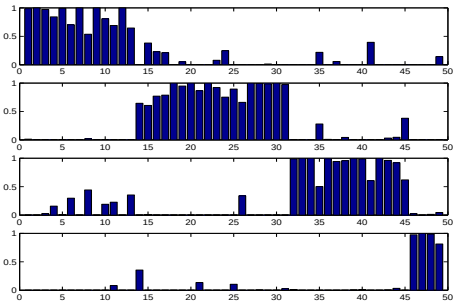


Figure 9: Decomposition diagram derived from LPD for the dataset of West *et al*.

As observed previously, process 4 has the most distinctive genetic signature which, from time-to-metastasis data, appears identified with the second row in Figure 9. The top-ranked genes distinguishing this process are given in the Table below:

Rank	Gene	Z_2-score	Expression
1.	*hCRHP*	5.51	Under
2.	*XBP1*	5.50	Under
3.	*FOXA1*	5.26	Under
4.	*FPB1*	4.98	Under
5.	*FLJ13710*	4.94	Under
6.	*GATA3*	4.94	Under
7.	*GATA3*	4.92	Under
8.	*CNAP1*	4.90	Over
9.	*NFIB2*	4.83	Over
10.	*Human complement factor B*	4.83	Under
11.	*TFF3*	4.79	Under
12.	*FLJ13710*	4.78	Under

Table 2: Top ranked genes using the Z_2-score distinguishing a tentative process 4. Using the Z_1 score *GATA3* is ranked 2nd, *FOXA1* is 3rd, *XPB1* is 4th and *TFF3* is 6th. The probabilities of occurrence are upper bounded by 2×10^{-6} (for $Z_2 = 4.78$).

Interestingly, *GATA3*, *FOXA1*, *XPB1*, *TFF3* and *FPB1* are in common between this Table and Table 1. Though *GRB7* and *ERBB2* were highlighted previously [25] the associated p-values and sample sizes indicate they do not have a statistically significant elevated expression here, though this fact most likely stems from the smaller dataset size.

2.3 van 't Veer *et al* dataset.

For the dataset of van 't Veer *et al* [26] we used samples from 78 patients with primary breast carcinomas, a further 18 samples from patients with *BRCA1* germline mutations and 2 samples with *BRCA2* mutations. We used 500 genes selected using the p-values derived by the authors [26], using those genes with a p-value of less than 0.01 in more than 30 tumours. Survival data is not available though we can still compute the log-likelihood curves (Figure 10) and this suggests a peak at 4 processes.

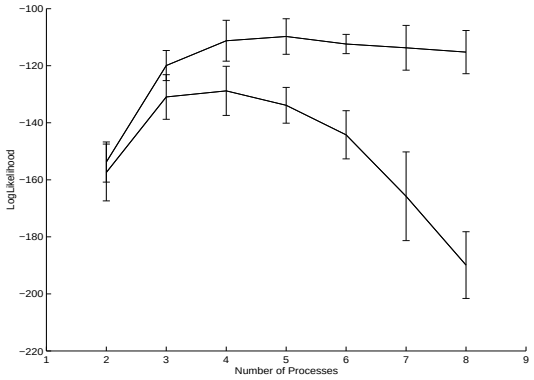


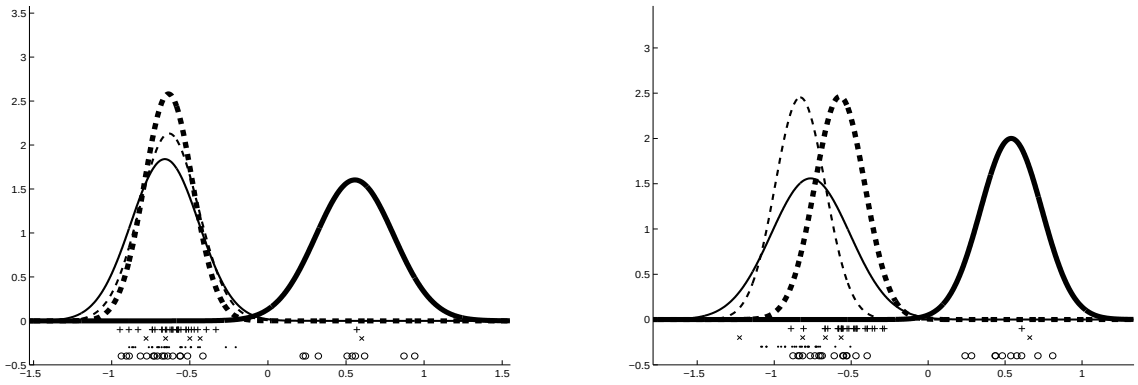
Figure 10: The log-likelihood (y-axis) versus number of processes (x-axis) using the MAP solution (upper, plateauing curve) and maximum likelihood (lower curve) solution for the Van 't Veer *et al* dataset [26].

The spectrum of peaks corresponding to Figure 2 indicated that 16 of the 18 *BRCA1* mutation

Rank	Gene	Z_2-score	Expression
1.	*TFF3*	7.02	Under
2.	*AGR2*	6.89	Under
3.	*FOXC1*	6.79	Over
4.	*GABA*	6.75	Over
5.	*VGLL1*	6.68	Over

Table 3: *TFF3* and *FOXC1* are first and third ranked for the most distinctive process in the dataset of van 't veer *et al*. Similarly they are first and second ranked for the most distinctive and aggressive process (4) in the data of Sorlie *et al* (Table 1).

carriers belonged in one process (which, from the time to metastasis data, appeared to be process 4 in Figure 3). The other 2 *BRCA1* samples were spread between processes and, interestingly, were the only 2 patients not to proceed to metastasis. The two *BRCA2* samples belonged together in the same process, distinct from the process associated with the *BRCA1* samples. This picture agreed with the interpretation by dendrogram of Sorlie *et al* [25].
Using the Z_1-score, one process has *ERRB2* (Figure 11(a)) and *GRB7* (Figure 11(b)) in second and third ranked position with the distribution of expression values having a similar bimodal distribution to that in Figures 5(a) and 5(b).



(a) *GRB7* (b) *ERBB2*

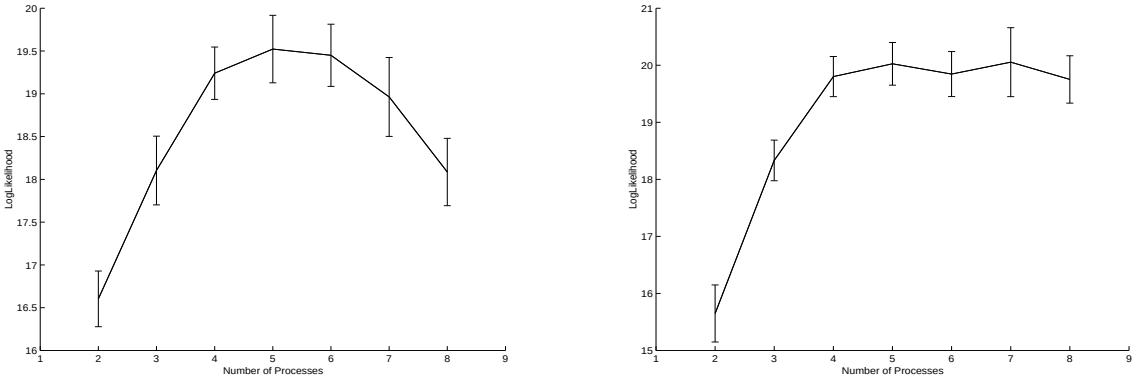
Figure 11: Inferred densities for *GRB7* and *ERBB2* for the dataset of van 't Veer *et al*.

The highest ranked Z_2-scores for genes in the four processes are 7.02, 5.85, 5.61 and 2.87. Interestingly, the most distinctive process (with $Z_2 = 7.02$) is associated with genes described previously for process 4, such as *TFF3* and *FOXC1* (Table 3). *TFF3*, and the *GATA3*, *FOXA1* and *XPB1* genes mentioned previously, all feature in a small gene expression graph derived from a sparse graphical model [7, 8] indicating genes closely linked with the estrogen receptor gene.

2.4 de Vijver *et al* dataset.

The study of van 't veer *et al* preceded a larger study by de Vijver *et al* [6] which used 295 samples from patients with primary breast carcinomas. The authors of this study discovered tentative signatures for poor and good prognosis using a reduced 70 gene set selected from 24,479. In

Figure 14 we present a Kaplan-Meier plot with the lower dashed curve corresponding to patients in the poor signature cohort and the upper dashed curve corresponding to the good signature cohort. In Figure 12(a) we have re-analysed the same dataset (295 samples, 70 features) using variational LPD and a maximum likelihood approach. The curve shows a peak in the range 4 to 6 processes, implying that the 2-process model proposed by the original authors [6] is a sub-optimal interpretation of the data. In Figure 12(b) we see that the likelihood curve for the MAP solution plateaus after using 4 processes.



(a) Maximum likelihood solution

(b) MAP solution

Figure 12: The log-likelihood (y-axis) versus number of processes (x-axis) using a maximum likelihood and MAP approach for the de Vijver *et al* dataset.

If we plot the corresponding Kaplan-Meier curves for Figure 13 we get the curves in Figure 14 in which the top process in Figure 13 is identified with curve 3 in Figure 14, the second process is identified with curve 4, the third process with 2 and the fourth (lowest) with 1. Compared to the original analysis of de Vijver *et al* (dashed curves in Figure 14), all patients in processes 3 and 4 derive from their lower (poor prognosis) group while 10 patients in process 1 are derived from their upper (good prognosis) group and 2 are derived from their poor prognosis group. All patients in process 2 derive from their good prognosis group. Thus our analysis is compatible with their description while enhancing the distinction between clinical outcomes (the solution presented here corresponds to the highest likelihood solution found in numerical experiments). With the MCMC-based algorithm we obtain a very similar Kaplan-Meier plot (Figure 19).

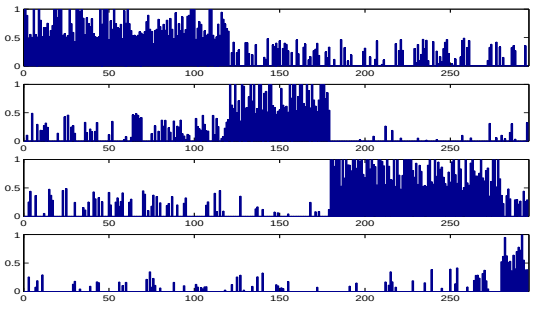


Figure 13: A 4 process decomposition of the dataset of de Vijver *et al* using variational LPD.

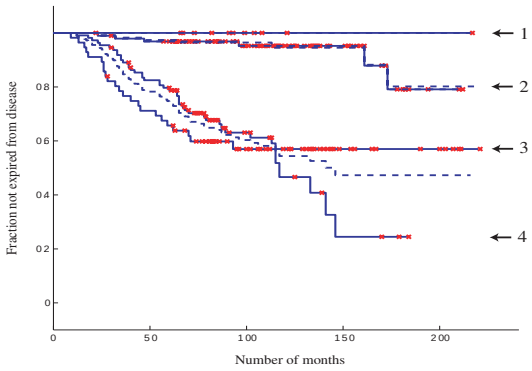
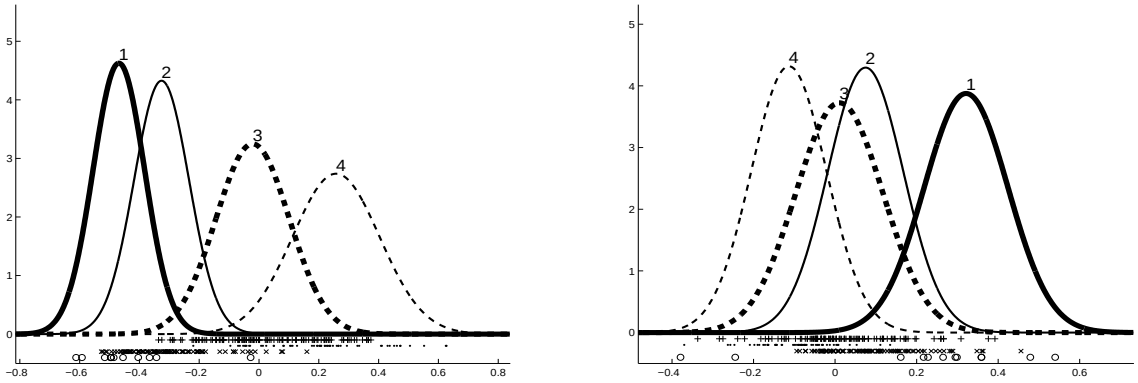


Figure 14: Kaplan-Meier plot for the processes identified in Figure 13: fraction not expired from the disease (y-axis), versus number of months (x-axis). The curves labelled 3 and 4 meet at the midpoint *but do not cross over*. The number of patients identified with each curve is 12 (process 1), 97 (2), 110 (3) and 56 (4) (these numbers do not sum to 295 because some samples are ambiguously identified). The original split of de Vijver *et al* [6] are given as dashed curves for comparison.

The inferred densities for two top-ranked genes separating processes 1 and 4 are given in Figures 15(a) and 15(b). In fact, of the 26 top-ranked genes separating processes 1 and 4, 21 genes move from under-expression to over-expression as we progress from indolent to the most aggressive subtype, following the trend in Figure 15(a), while 4 genes follow the reverse trend illustrated in Figure 15(b).



(a) *ORC6L* (b) *STK32B*

Figure 15: Inferred densities for *ORC6L* and *STK32B*. The individual expression values are given below the inferred density curves, with ∘ associated with process 1, × with 2, + with 3 and · with process 4.

The observation that most of the listed genes under-express in process 1 agrees with an observation for the dataset of Sorlie *et al* in which we found that 19 from the top ranked 20 genes distinguishing process 1 from the others under-expressed on the average in process 1. The gene names, their mean expression values per process and this trend are discussed in further detail in Appendix 3 to this paper.

3 Conclusion

The results are broadly consistent and indicate at least four principal processes for primary breast carcinoma. Our analysis suggests the existence of an indolent subtype distinguished by under-expression across a number of genes associated with tumour growth. Since some patients in this process do develop metastatic tumours this process is not wholly benign, nor does it consist of misidentified normal samples. There is a subtype closely related to the Luminal A subtype proposed by Sorlie *et al* [25]. In line with previous observations there is also a subtype marked by up-regulation of *ERBB2* (*HER2*) and *GRB7*. As noted in Figures 5 and 11 there is an apparent bimodal distribution and *ERBB2* and *GRB7* do not uniformly over-express in this process. Given the split observed in the dendrogram (Appendix 1) this may indicate two subprocesses, one with elevated expression levels for these genes. However, we did not find a statistically significant difference in clinical outcome for patients belonging to these two possible subclasses. The most aggressive subtype is also the most well defined: it is clearly and consistently identified by both variants of LPD (Figures 3 and 18(a)) and matches the basal subtype described by Sorlie *et al* (Figure 16). This subtype is marked by abnormal expression of the genes *FOXA1*, *FOXC1*, *GATA3*, *TFF3* and *XBP1*, for example, and it is associated with loss of regulation of the vascular growth factor *VEGF*. As already remarked, using a sparse graphical model [7, 8], we find that the transcription factor genes *FOXA1*, *GATA3*, *TFF3* and *XBP1* are closely linked with the estrogen receptor-alpha gene, which with the estrogen pathway, plays a crucial role in the development of many breast tumours. One target of ERα is the *TFF1* gene and *FOXA1* has a direct influence on transcription by this gene since there are binding sites for FOXA1 in its promoter region [1]. A number of other ERα-bound promoters have FOXA1 binding sites [16]. The role of *FOXA1* has been highlighted in a contemporary study by Laganiere *et al* [16]: expression by *FOXA1* correlates with the presense of ERα and it has been suggested that that this gene plays a crucial role in a transcriptional domain governing estrogen response. Reinforcing this result, a contemporary study by Carroll *et al* [4] has shown that forkhead factor binding sites are present in 54% of 57 ER binding regions. This strongly supports the significance of abnormal expression of *FOXA1* and *FOXC1* indicated by our analysis. Finally, in agreement with the analysis using a sparse graphical model [7, 8], there appears to be an important role played by *TFF3*, a close relative of *TFF1*.
The decomposition proposed here is at most a basic model since one would expect further sub-division as more data becomes available, thus enabling a higher resolution picture. As remarked previously, the effects of noise are averaged out as the dataset size increases. Thus for the dataset of Sorlie *et al* the peak in the likelihood curve is at 3-4 processes but, for the largest dataset of de Vijver *et al*, it is approximately 4-5. Certainly, our analysis suggests that the 2 process split of de Vijver *et al* [6] is too simple a model and at least 4 main processes are justified by the datasets used. The dataset for West *et al* was exclusively based on invasive ductal tumours and the Sorlie *et al* dataset had samples very predominantly of this type. However, use of samples consistently of the same histological type would also help reduce noise and improve definition. The indolent subtype 1 was not presented in the original analysis of Sorlie *et al* and the ability of the method to find this feature highlights the importance of using Bayesian methods in this context.

References

[1] S Beck, P Sommer, E Do Santos Silva, N Blin, and P Gott. Hepatocyte Nuclear Factor 3 (winged helix domain) activates trefoil factor gene TFF1 through a binding motif adjacent to the TATA box. *Cell Biology*, 18:157–164, 1999.

[2] D Blei, A Ng, and M Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] J Brunet, P Tamayo, T Golub, and J Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings National Academy Sciences*, 101:4164–4169, 2004.

[4] J Carroll et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FOXA1. *Cell*, 122:33–43, 2005.

[5] C Collesi, M Santoro, G Gaudino, and P Comoglio. A splicing variant of the RON transcript induces constitutive tyrosine kinase activity and an invasive phenotype. *Molecular Cellular Biology*, 16:5518–5526, 1996.

[6] M de Vijver et al. A gene expression signature as a predictor of survival in breast cencer. *New England Journal of Medicine*, 347:1999–2009, 2002.

[7] A Dobra, B Jones, C Hans, J Nevins, and M West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212, 2004.

[8] A Dobra and M West. Graphical model-based gene clustering and metagene expression analysis. Technical report, 2004.

[9] M Elkin, A Orgel, and H Kleinman. An angiogenic switch in breast cancer involves estrogen and soluble vascular endothelial growth factor receptor 1. *Journal of the National Cancer Institute*, 96:875–978, 2004.

[10] P Flaherty, G Giaever, J Kumm, M I Jordan, and A P Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics*, 21:3286–3293, 2005.

[11] S Gruvberger et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Research*, 61:5979–5984, 2001.

[12] I Hedenfalk et al. Gene-expression profiles in hereditory breast cancer. *New England Journal of Medicine*, 344:539–548, 2001.

[13] L Ben-Tovin Jones, S Ng, C Ambroise, K Monico, N Khan, and G J McLachlan. Use of microarray data via model-based classification in the study and prediction of survival from lung cancer. In J S Shoemaker and S M Lin, editors, *Methods of Microarray Data Analysis IV*, pages 163–173. New York:Springer, 2005.

[14] F Katsuoka, H Motohashi, J Engel, and M Yamamoto. NRF2 transcriptionally activates the MAFG gene through an antioxidant response element. *J Biol Chem*, 280:4483–4490, 2005.

[15] M Lacroix and G Leclerq. About GATA3, HNF3A and XBP1, three genes co-expressed with the oestrogen receptor-alpha gene (ESR1) in breast cancer. *Molecular and Cellular Endocrinology*, 219:1–7, 2004.

[16] J Laganiere et al. Location analysis of estrogen receptor α target promoters reveals that FOXA1 defines a domain of the estrogen response. *Proceedings National Academy Sciences*, 102:11651–11656, 2005.

[17] D Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[18] G McLachlan, R Bean, and D Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422, 2002.

[19] T Moloshok, R Klevecz, J Grant, F Manion, W Speier, and M Ochs. Application of bayesian decomposition for analysing microarray data. *Bioinformatics*, 18:566–575, 2002.

[20] S Pero, R Daly, and D Krag. GRB7-based molecular therapeutics in cancer. *Expert Reviews in Molecular Medicine*, 5:1–11, 2003.

[21] C Perou et al. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.

[22] D Rees. *Essential Statistics*. Chapman and Hall, 2001.

[23] S Rogers, M Girolami, C Campbell, and R Breitling. The latent process decomposition of cDNA microarray datasets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2:143–156, 2005.

[24] T Sorlie et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings National Academy Sciences*, 98:10869–10874, 2001.

[25] T Sorlie et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings National Academy Sciences*, 100:8418–8423, 2003.

[26] L van 't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–535, 2002.

[27] M West et al. Predicting the clinical status of human breast cancer using gene expression profiles. *Proceedings National Academy Sciences*, 98:11462–11467, 2001.

[28] L Zhukova, N Zhukov, and M Lichinitser. Expression of FLT-1 and FLK-1 receptors for vascular endothelial growth factor on tumor cells as a new prognostic criterion for locally advanced breast cancer. *Bull Exp Biol Med*, 135:478–481, 2003.

4 Appendix 1: Comparison with dendrogram of Sorlie *et al*

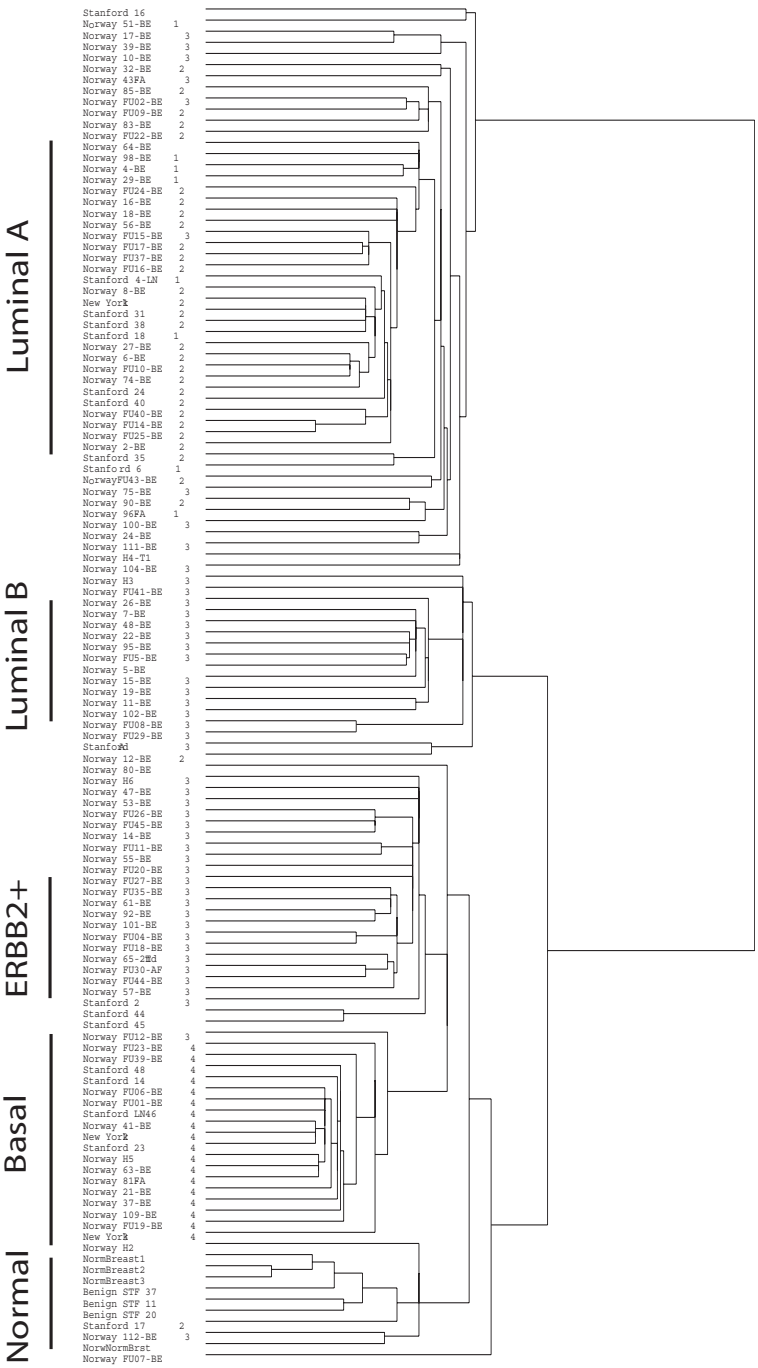


Figure 16: A comparison between the dendrogram reported in Sorlie *et al* [25], Figure 1B, and the decomposition by variational LPD given in Figure 2. To the left of the tree, the variational LPD assignment to process is designated by the numbers 1 to 4. Beside these numbers are the sample titles for identification with Sorlie *et al* [25], Figure 1B. Process assignment numbers are missing in a few cases because the peak in Figure 2 (normalised γ_{ak}, see equation 4, Appendix 2) is ambiguous in its assignment of sample to process.

5 Appendix 2: Latent Process Decomposition.

5.1 Variational Approach to LPD

We will briefly outline Latent Process Decomposition: for a more detailed description of the method the reader is referred to Rogers *et al.* [23]. As remarked in the text, a sample can be represented as a combinatorial mixture over multiple processes, in contrast to the implicit mutual exclusion of classes assumption of most cluster analysis methods. Thus we have used *process* rather than *cluster* to emphasis this difference with standard cluster analysis methods.
We are interested in constructing a model for the microarray data and this model will have parameters which we alter during the training process. We will suppose these parameters are $r_1, r_2, \ldots$ or, as a set, $\mathcal{R}$. Similarly the dataset will be denoted by $\mathcal{D}$. Thus we wish to maximise the probability of a model given the data, $p(\mathcal{R}|\mathcal{D})$, which from Bayes's rule can also be written:

$$p(\mathcal{R}|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{R})p(\mathcal{R}) \tag{1}$$

where $p(\mathcal{D}|\mathcal{R})$ is the *likelihood* and $p(\mathcal{R})$ is the *prior* on our parameters $\mathcal{R}$.
The approach we now outline is described in more detail elsewhere [23] and it adopts the Latent Dirichlet Allocation (LDA) approach to data modelling [2], comparing favourably with alternatives such as mixture models [18], Naive Bayes and other approaches (see [23]). In this approach we incorporate prior beliefs in the form of reasonable distributional assumptions e.g. the (logged) gene expression levels from a microarray experiment are assumed approximately normally distributed (for Affymetrix data we use a prior affine translation to bring expression data into an approximate $\mathcal{N}(0,1)$ distribution). Unfortunately, we cannot estimate the above posterior probability directly but we can lower bound this expression using Jensen's inequality. Thus our approach parallels the Latent Dirichlet Allocation method of Blei et al [2] which derives a similar lower bound for discrete data. This lower bound is found using an efficient algorithmic technique, described below.
We are interested in finding the set of paramaters $\mathcal{R}$ that maximises $p(\mathcal{R}|\mathcal{D})$. In the case of a uniform (or uninformative) prior, this is the maximum likelihood solution. We will begin by deriving the maximum likelihood solution and then extend the method to a non-uniform prior. The *log-likelihood* of a set of $\mathcal{A}$ training samples is $\log p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha})$, where $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha}$ are the model parameters, the process means, standard deviations and Dirichlet parameter respectively. Marginalising over the latent variable $\boldsymbol{\theta}$ allows us to expand this expression as follows

$$\log p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha}) = \sum_{a=1}^{\mathcal{A}} \log \int_{\theta} p(a|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta}. \tag{2}$$

A lower bound on this expression can be inferred by the introduction of two variational parameters Q_{kga} and γ_{ak} and the following iterative update equations provide estimates for these parameters

$$Q_{kga} = \frac{\mathcal{N}(e_{ga}|k, \mu_{gk}, \sigma_{gk}) \exp\left[\psi(\gamma_{ak})\right]}{\sum_{k'=1}^{\mathcal{K}} \mathcal{N}(e_{ga}|k, \mu_{gk'}, \sigma_{gk'}) \exp\left[\psi(\gamma_{ak'})\right]} \tag{3}$$

$$\gamma_{ak} = \alpha_k + \sum_{g=1}^{\mathcal{G}} Q_{kga} \tag{4}$$

for given α_k, with process index $k = 1, \ldots, \mathcal{K}$, and where $\mathcal{N}(\ldots)$ is a normal distribution and $\psi(z)$ is the digamma function. For gene g and process k, μ_{gk} and σ_{gk} are the means and standard deviations (for example, in Figure 5 these give the means and spreads for the 4 processes illustrated). γ_{ak}, normalised over the number of processes, gives the confidence of membership of sample a in process k. Let e_{ga} denote the expression level for gene g in sample a, then the model parameters are obtained from the following update equations:

$$\mu_{gk} = \frac{\sum_{a=1}^{\mathcal{A}} Q_{kga} e_{ga}}{\sum_{a'=1}^{\mathcal{A}} Q_{kga'}} \tag{5}$$

$$\sigma_{gk}^2 = \frac{\sum_{a=1}^{\mathcal{A}} Q_{kga} (e_{ga} - \mu_{gk})^2}{\sum_{a'=1}^{\mathcal{A}} Q_{kga'}} \tag{6}$$

The update rule for the Dirichlet model parameter α_k is found from the derivatives of the $\boldsymbol{\alpha}$ dependent terms in the likelihood [2]. Thus the α_k are modified after each iteration of the above updatings using a standard Newton-Raphson technique (see [2] Appendix A.4.2 and [23]).
The above argument can be extended to a maximum posterior (MAP) solution with non-uniform priors. Thus, a suitable prior on the means could be a Gaussian distribution with zero mean. This would reflect a prior belief that for cDNA microarrays most genes will be uninformative and will have logged expression ratios around zero (i.e. they are unchanged compared to a reference sample). For the variance, we may wish to define a prior that penalises over-complex models and avoids overfitting. Overfitting may occur when Gaussian functions contract onto a single data point causing poor generalisation. With a suitable choice for the prior an extension of our model to a full MAP solution is straightforward. Our combined likelihood and prior expression is (assuming a uniform prior on $\boldsymbol{\alpha}$):

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha} | \mathcal{D}) \propto p(\mathcal{D} | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha}) p(\boldsymbol{\mu}) p(\boldsymbol{\sigma}). \tag{7}$$

Taking the logarithm of both sides we see that the maximisation task is given by:

$$\boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\mu} = \arg \max_{\boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\mu}} \quad \log p(\mathcal{G} | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha}) + \log p(\boldsymbol{\mu}) + \log p(\boldsymbol{\sigma}). \tag{8}$$

Thus we can simply append these terms onto our bound on the log-likelihood. Noting that they are functions of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ only (and any associated hyper-parameters), we conclude that these extra terms only change the update equations for μ_{ak} and σ_{ak}. Let us assume the following priors:

$$p(\mu_{gk}) \propto \mathcal{N}(0, \sigma_\mu) \tag{9}$$

$$p(\sigma_{gk}^2) \propto \exp\left\{ -\frac{s}{\sigma_{ak}^2} \right\} \tag{10}$$

then we obtain the following new update equations instead:

$$\mu_{gk} = \frac{\sigma_\mu^2 \sum_{a=1}^{\mathcal{A}} Q_{gka} e_{ga}}{\sigma_{gk}^2 + \sigma_\mu^2 \sum_{a=1}^{\mathcal{A}} Q_{gka}} \tag{11}$$

$$\sigma_{gk}^2 = \frac{\sum_{a=1}^{\mathcal{A}} Q_{gka} (e_{ga} - \mu_{gk})^2 + 2s}{\sum_{a=1}^{\mathcal{A}} Q_{gka}}. \tag{12}$$

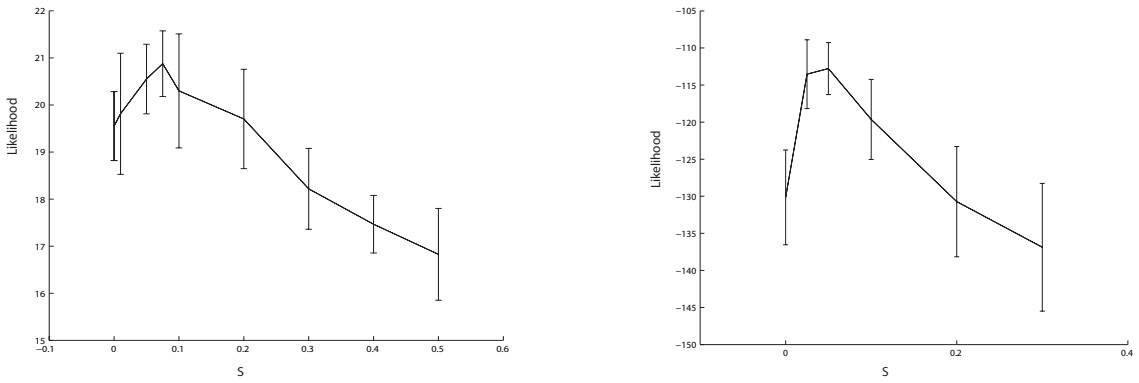
Once the model parameters have been estimated, we can calculate the likelihood for a collection of $\mathcal{A}'$ samples using:

$$\mathcal{L} = \prod_{a=1}^{\mathcal{A}'} \int_\theta \left\{ \prod_{g=1}^{\mathcal{G}} \sum_{k=1}^{\mathcal{K}} \mathcal{N}(e_{ga} | k, \mu_{gk}, \sigma_{gk}) \theta_k \right\} p(\boldsymbol{\theta} | \boldsymbol{\alpha}) d\boldsymbol{\theta} \tag{13}$$

where we estimate the expectation over the Dirichlet distribution by averaging over N samples drawn from the estimated Dirichlet prior $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$

$$\mathcal{L} \approx \prod_{a=1}^{\mathcal{A}'} \frac{1}{N} \sum_{n=1}^{N} \left\{ \prod_{g=1}^{\mathcal{G}} \sum_{k=1}^{\mathcal{K}} \mathcal{N}(e_{ga}|k, \mu_{gk}, \sigma_{gk}) \theta_{kn} \right\}. \tag{14}$$

Apart from using the likelihood to determine the best number of processes to use, it can be used to determine the parameters used in the prior. In Figure 17 we plot likelihood curves as a function of s, the prior parameter in equation (10). The peaks in these plots model the extent of noise in the data and enables the algorithm to avoid constructing an over-complex model which would fit to this noise. As reported elsewhere [23] the model is little affected by choice of the prior parameter σ_μ in equation (9) and we have set this value to 0.1.



(a) Sorlie *et al*

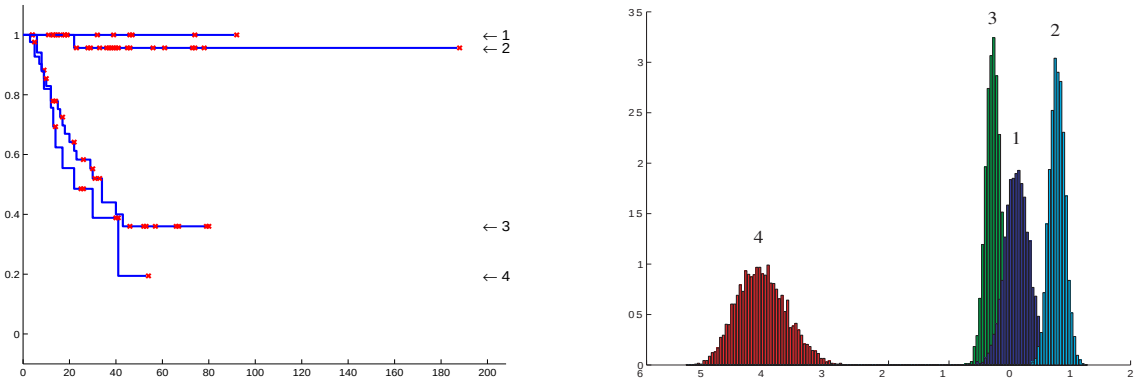
(b) van 't Veer *et al*

Figure 17: Hold-out log-likelihood as a function of s for the datasets of Sorlie *et al* (left) and van 't Veer *et al* (right).

5.2 A Markov Chain Monte Carlo approach to LPD

To validate the above variational method we re-derived the results using a Gibbs sampler-based approach for the datasets of Sorlie *et al* and de Vijver *et al*. The starting point, equation (2), is the same but otherwise the method is distinct. The approach we now describe is slow to execute (the cross-validation study of the number of processes proved prohibitive). However, it supports the results presented in the main text. Also, by using a Gibbs sampler we can obtain a full posterior distribution for the model parameters and hence investigate the accuracy of the point estimate approximations derived by the variational algorithm described above.

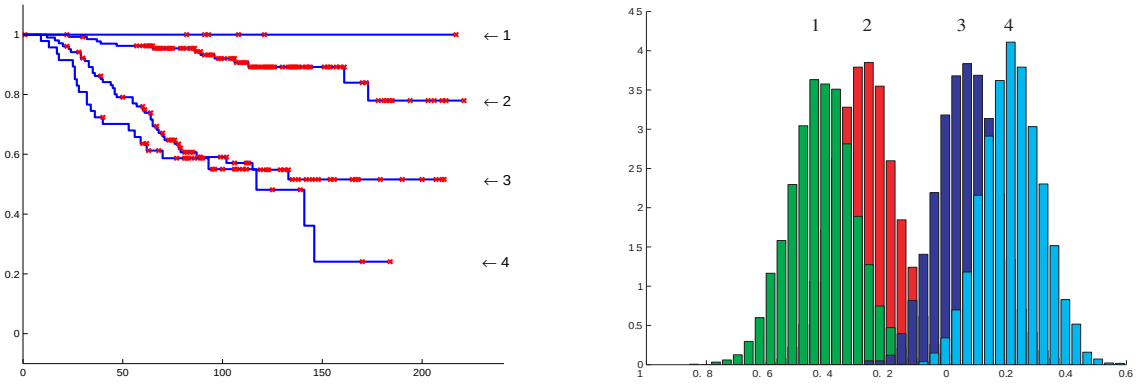
We implemented a standard Gibbs sampler [17] using conjugate priors for all model parameters. Each variable in the algorithm was initialised randomly. We used a *burn-in* period to allow the Monte Carlo algorithm to stabilise (100000 iterations for the Sorlie *et al* dataset and 40000 for de Vijver *et al*). The next 10000 samplings were used to form the posterior distribution. To compare with variational LPD we chose 4 processes. For process membership there is no γ parameter so instead we determined membership from the normalised mode of the posterior distribution of θ. For the Sorlie *et al* dataset we give the resulting Kaplan Meier plot in Figure 18(a), which can be compared to Figure 3(a) from the variational approach. The posterior distribution over model parameters supported the significance of genes already discussed. For example, in Figure 18(b) we give the distribution over means for *FOXA1* which can be compared to Figure 7(a) with point estimates of the means from the variational approach.



(a) KM Plot (b) *FOXA1*

Figure 18: Kaplan Meier plot (left) and distribution of means (right) for *FOXA1* from the MCMC algorithm applied to the Sorlie *et al* dataset. For the Kaplan Meier plot there are 11 patients in process 1, 30 in procss 2, 42 in process 3 and 18 in process 4. The right hand Figure shows the distribution of means for a selected gene (*FOXA1*) indicating the reliability of the point estimates of the means found using LPD (see Figure 7(a) for comparison).

For the dataset of de Vijver *et al* and using the MCMC approach, we give the Kaplan Meier plot in Figure 19(a). As for the variational approach we find one indolent process and further processes of increasing aggressiveness. For comparison with Figure 15(a) we give the distribution of means for *ORC6L* in Figure 19(b).



(a) KM Plot (b) *ORC6L*

Figure 19: Kaplan Meier plot (left) using the MCMC approach for the de Vijver *et al* dataset. For the Kaplan Meier plot there are 6 patients in process 1, 136 in procss 2, 103 in process 3 and 47 in process 4. The curves labelled 3 and 4 meet but do not cross. The right hand Figure shows the distribution of means for *ORC6L* giving a similar progression to that observed in Figure 15(a).

6 Appendix 3: Supplementary Material on the dataset of De Vijver *et al*.

In the original publication of de Vijver *et al* [6] 21 cDNA sequences had no gene name or information associated with them. Given this fact and the monotonic trends in mean expression values mentioned in the main text we have updated and examined ontology information for the 70 genes and their encoded proteins to examine their significance. A full description of all 70 entries and further information is available as supplementary data at www.enm.bris.ac.uk/lpd/bc.htm. In the table below we list the top ranked genes distinguishing process 1 vs process 4 (with $Z_1 > 2$) for the dataset of de Vijver *et al*. The 4 columns headed *Process* are the mean logged expression values (using log base 10). The processes are ranked in order of most indolent (1) to most aggressive (4) outcome. The end column highlights the progression trend across the 4 processes. Genes marked *BCSS1* and *BCSS2* correspond to hypothetical genes: *BCSS1* is 'moderately similar to T50635 hypothetical protein' and *BCSS2* is 'weakly similar to ISHUSS disulfide-isomerase'. The Z_1 values follow a normal probability distribution $\mathcal{N}(0, 1)$.

Of these genes, *ORC6L* is involved in DNA replication and serves as a platform for the assembly of additional initiation factors such as *CDC6* and *MCM*. siRNA gene silencing studies indicate that *ORC6L* plays an essential role in coordinating chromosome replication and segregation with cytokinesis. *STK32B* is a serine/threonine kinase. *KIAA1442* encodes a transcription factor with an IPT/TIG motif. These motifs are found in cell surface receptors such as Met and Ron as well as in intracellular transcription factors where it is involved in DNA binding. Intriguingly the Ron tyrosine kinase receptor shares with the members of its subfamily (Met and Sea) the control of cell dissociation, motility, and invasion of extracellular matrices (scattering) [5]. Two genes have no known function though Contig38288RC is weakly similar to ISHUSS protein disulfide-isomerase, an enzyme that participates in the folding of proteins containing disulfide bonds. In the Table we have labelled Contig55725RC as *BCSS1* and Contig38288RC as *BCSS2* (breast cancer survival signature 1 and 2). Many genes are involved in processes associated with tumour growth such as DNA replication (*MCM6*), cell cycle control (*CCNE2*), spindle associated factors (*NUSAP1*, *PRC1*), chromosome organisation (*CENPA*), actin filament assembly (*DIAPH3*) and vascular remodelling (*ITS*). All these genes are up-regulated for the most aggressive process versus the least aggressive. *DIAPH3*, which was unidentified in the original paper, appears three times in the 70 gene set.

Gene ID	Gene name	Process 1	Process 2	Process 3	Process 4	Z_1	Trend
NM_014321	*ORC6L*	−0.47	−0.32	−0.02	0.26	4.29	Up
Contig55725_RC	*BCSS1*	−0.80	−0.54	−0.22	0.39	4.15	Up
NM_018401	*STK32B*	0.32	0.07	0.01	−0.11	3.14	Down
AB037863	*KIAA1442*	0.28	0.05	−0.01	−0.29	3.07	Down
Contig38288_RC	*BCSS2*	−0.34	−0.16	−0.02	0.26	3.06	Up
NM_003981	*PRC1*	−0.45	−0.30	0.02	0.24	2.98	Up
NM_016359	*NUSAP1*	−0.50	−0.28	0.039	0.22	2.93	Up
NM_004702	*CCNE2*	−0.55	−0.32	−0.02	0.22	2.93	Up
NM_001809	*CENPA*	−0.52	−0.41	−0.06	0.29	2.80	Up
AL137718	*DIAPH3*	−0.30	−0.10	0.03	0.22	2.78	Up
NM_014791	*MELK*	−0.46	−0.21	0.01	0.26	2.71	Up
NM_016448	*RAMP*	−0.36	−0.17	0.05	0.15	2.65	Up
Contig40831_RC	*AI224578*	−0.39	−0.11	−0.05	0.19	2.57	Up
AL080059	*TSPYL5*	−0.53	−0.24	−0.15	0.25	2.50	Up
Contig46218_RC	*DIAPH3*	−0.35	−0.22	0.04	0.27	2.50	Up
NM_003875	*GMPS*	−0.34	−0.17	−0.05	0.21	2.45	Up
NM_020974	*SCUBE2*	0.24	0.19	−0.24	−0.99	2.39	Down
NM_000436	*OXCT1*	−0.29	−0.06	−0.10	0.15	2.37	Mixed
NM_005915	*MCM6*	−0.37	−0.14	0.00	0.23	2.31	Up
AA555029_RC	*AA555029*	−0.31	−0.09	−0.06	0.15	2.27	Up
NM_002916	*RFC4*	−0.29	−0.133	−0.01	0.20	2.27	Up
AL080079	*GPR126*	−0.59	−0.25	−0.12	0.17	2.22	Up
NM_015984	*UCHL5*	−0.21	−0.08	−0.01	0.15	2.13	Up
Contig20217_RC	*TGS*	−0.33	−0.17	−0.02	0.17	2.08	Up
NM_006117	*PECI*	0.21	0.05	0.01	−0.25	2.07	Down
Contig32185_RC	*ITS*	−0.33	−0.14	−0.08	0.15	2.02	Up