The Latent Process Decomposition of cDNA Microarray Data Sets

Simon Rogers, Mark Girolami, Colin Campbell, and Rainer Breitling

Abstract—We present a new computational technique (a software implementation, data sets, and supplementary information are available at http://www.enm.bris.ac.uk/lpd/) which enables the probabilistic analysis of cDNA microarray data and we demonstrate its effectiveness in identifying features of biomedical importance. A hierarchical Bayesian model, called Latent Process Decomposition (LPD), is introduced in which each sample in the data set is represented as a combinatorial mixture over a finite set of latent processes, which are expected to correspond to biological processes. Parameters in the model are estimated using efficient variational methods. This type of probabilistic model is most appropriate for the interpretation of measurement data generated by cDNA microarray technology. For determining informative substructure in such data sets, the proposed model has several important advantages over the standard use of dendrograms. First, the ability to objectively assess the optimal number of sample clusters. Second, the ability to represent samples and gene expression levels using a common set of latent variables (dendrograms cluster samples and gene expression values separately which amounts to two distinct reduced space representations). Third, in constrast to standard cluster models, observations are not assigned to a single cluster and, thus, for example, gene expression levels are modeled via combinations of the latent processes identified by the algorithm. We show this new method compares favorably with alternative cluster analysis methods. To illustrate its potential, we apply the proposed technique to several microarray data sets for cancer. For these data sets it successfully decomposes the data into known subtypes and indicates possible further taxonomic subdivision in addition to highlighting, in a wholly unsupervised manner, the importance of certain genes which are known to be medically significant. To illustrate its wider applicability, we also illustrate its performance on a microarray data set for yeast.

Index Terms—cDNA microarray, latent variable modeling, cluster analysis.

1 Introduction

In recent years, there has been a large growth in the volume of microarray data. Machine learning techniques have been very useful in extracting information from these data sets. For example, unsupervised techniques such as hierarchical cluster analysis have been used to indicate tentative new subtypes for some cancers [1], [6] and to identify biologically relevant structure in large data sets. Using class labels, supervised learning methods will have an increasingly important role to play in indicating the detailed subtype of a disease, its expected progression, and the best treatment strategy.

Cluster analysis is the most common data analysis method used throughout the microarray literature. Generation of a dendrogram allows the inference of possible group structure within the data and, thus, the identification of related samples or functionally similar genes. A weakness of clustering by dendrogram is that there is no natural way to objectively assess the most probable number of structures which underlie the data. Probabilistic model-based clustering can overcome this problem and clustering based on mixtures of multivariate Gaussians has been proposed in [9], for example. However, as the number of features (gene probes) typically far exceeds the number of samples, the estimated within-class covariances will be singular. Therefore, extensive gene selection is a prerequisite to any probabilistic model-based clustering and in [9] a further level of feature extraction is required to obtain a sufficiently low-dimensional subspace within which the clustering is performed. For the purposes of biological interpretation, it would be better if group structure could be identified using the original set of genes probed in the microarray experiment with the model used to identify those genes responsible for the group structure. We will show that this is achieved naturally by the model proposed in this paper.

The main assumption underlying most clustering methods, whether dendrogram-based or probabilistic, is that a sample or gene is assigned exclusively to one class only. While this may be appropriate if mutually exclusive classes are inherent in the data, this assumption would not be valid if a number of biological processes interact to influence a given gene expression level, for example. Alternatively, a sample might validly share some characteristics with two or more sample clusters. It is therefore best to obtain models which are not restricted by this mutual exclusion of classes assumption.

Models overcoming this mutual exclusion assumption have already been proposed in the literature. An example is the Plaid Model of Lazzeroni and Owen [8] which allows for overlaps between clusters. Another example is a model

- *S. Rogers and M. Giralomi are with the Bioinformatics Research Centre, Department of Computing Science, A416, Fourth Floor, Davidson Building, University of Glasgow, Glasgow G12 8QQ, Scotland, United Kingdom. E-mail: {srogers, girolami}@dcs.gla.ac.uk.*
- *C. Campbell is with the Department of Engineering Mathematics, Queen's Building, Bristol University, Bristol BS9 1TR, United Kingdom. E-mail: C.Campbell@bris.ac.uk.*
- *R. Breitling is with the Molecular Plant Science Group & Bioinformatics Research Centre, Institute for Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom. E-mail: r.breitling@bio.gla.ac.uk.*

Manuscript received 12 Sept. 2004; revised 11 Nov. 2004; accepted 7 Dec. 2004; published online 2 June 2005.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0141-0904.

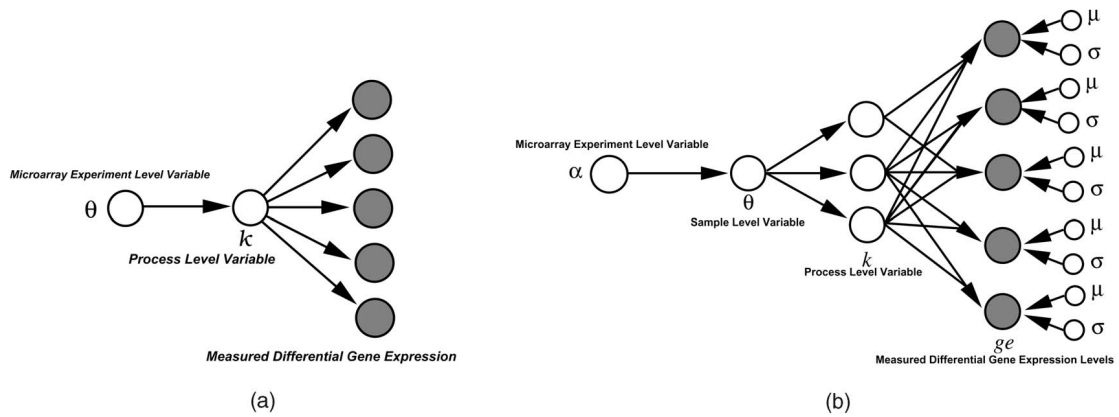


Fig. 1. Graphical model representations of LPD. (a) Cluster model structure. (b) LPD model structure.

recently proposed by Segal et al. [10], in which a probabilistic relational model (PRM) for gene expression is developed. PRM is shown to successfully identify known biological processes and provides improved interpretability over comparable methods [10]. However, PRM requires an estimation algorithm which, in its exact form, is NP-hard, and so various heuristics are required to make the model estimation tractable.

In this paper, we will present a new approach to the analysis of microarray data which is motivated by the above considerations and adopts the Latent Dirichlet Allocation (LDA) approach to data modeling [2]. We have called the method Latent Process Decomposition with a *process* defined as a set of functionally related datapoints (samples or genes). We use the term process, rather than cluster, because a sample or gene can have partial membership of several processes simultaneously, in contrast to many cluster analysis approaches, as discussed above.

Latent Process Decomposition has linear computational scaling in the number of genes, arrays and processes. It is able to provide a representation of microarray data sets which simultaneously reflects group structure across samples and genes as well as the possible interplay between multiple processes responsible for measured gene expression levels. Furthermore, it has advantages over alternative approaches such as the use of dendrograms, mixture models [9], Naive Bayes, and other approaches. In Section 2, we will outline the method, with full derivation and pseudocode listings given in the Appendix, while in Section 3, we will outline performance on medical and biological data sets. We have implemented LPD in Matlab and C^{++} and this code is freely available at http://www.enm.bris.ac.uk/lpd.

2 Latent Process Decomposition

2.1 Model Specification

In probabilistic terms, the data generation process for a tissue sample can be described as follows (Fig. 1b): For a complete microarray data set, a Dirichlet prior probability distribution for the distribution of possible processes is defined by the $\mathcal{K}$-dimensional parameter α. For a tissue sample a, a distribution θ over a set of mixture components indexed by the discrete variable k is drawn from a single prior Dirichlet distribution. Then, for each of the $\mathcal{G}$ features g (generally uniquely mapping to genes), we draw a process index k from the distribution θ with probability θ_k which selects a Gaussian defined by the parameters μ_{gk} and σ_{gk}. The level of expression e_{ga} for gene g from sample a is then drawn from the kth Gaussian denoted as $\mathcal{N}(e_{ga}|k, \mu_{gk}, \sigma_{gk})$. This is then repeated for each of $\mathcal{A}$ tissue samples. It can be easily seen that this generative process ensures that each sample has its own specific distribution of processes and that each measured level of gene expression in the sample is obtained by repeatedly sampling the process index from the sample-specific Dirichlet random variable θ, thus allowing the possibility of a number of processes having an effect on the measured levels of gene expression.

This is a biologically more realistic representation than the mixture model which underlies all forms of clustering, see Fig. 1a, where only one process or class is responsible for all the measured gene expression, thus disregarding gene-specific differences in effect (μ_{gk}) and noise (σ_{gk}) or the sample-specific interaction of various processes.

2.2 Learning with Uniform Priors: The Maximum Likelihood Solution

For a model with a set of parameters, $\mathcal{H}$, and data $\mathcal{D}$, Bayes's rule gives

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

where $p(\mathcal{D}|\mathcal{H})$ is the *likelihood* and $p(\mathcal{H})$ is the *prior* on our parameters $\mathcal{H}$. We are interested in finding the set of parameters $\mathcal{H}$ that maximizes $p(\mathcal{H}|\mathcal{D})$ (the *maximum posterior* or MAP solution). 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 this to a nonuniform (or informative) prior in Section 2.3. The log-likelihood of a set of $\mathcal{A}$ training samples is:

$$\log p(\mathcal{D}|\mu, \sigma, \alpha), \tag{2}$$

which can be factorized over individual samples as follows:

$$\log p(\mathcal{D}|\mu, \sigma, \alpha) = \sum_{a=1}^{\mathcal{A}} \log p(a|\mu, \sigma, \alpha). \tag{3}$$

Marginalizing over the latent variable θ allows us to expand this expression as follows:

$$\log p(\mathcal{D}|\mu,\sigma,\alpha) = \sum_{a=1}^{\mathcal{A}} \log \int_{\theta} p(a|\mu,\sigma,\theta)p(\theta|\alpha)d\theta, \quad (4)$$

where the probability of the sample conditioned on the means, variances, and process weightings can be expressed in terms of its individual components giving the following:

$$\log p(a|\mu,\sigma,\alpha) = \log \int_{\theta} \left\{ \prod_{g=1}^{\mathcal{G}} \sum_{k=1}^{\mathcal{K}} \mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk})\theta_k \right\} p(\theta|\alpha)d\theta, \quad (5)$$

where $\mathcal{N}(a|k,\mu,\sigma)$ denotes a normal distribution in process k with mean μ and standard deviation σ.

Variational inference can be employed to estimate the parameters of this model and full details are given in the Appendices. A lower bound on (5) can be inferred by the introduction of variational parameters and the following iterative update equations provide estimates of the two variational parameters Q_{kga} and γ_{ak}

$$Q_{kga} = \frac{\mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk})\exp[\psi(\gamma_{ak})]}{\sum_{k'=1}^{\mathcal{K}} \mathcal{N}(e_{ga}|k,\mu_{gk'},\sigma_{gk'})\exp[\psi(\gamma_{ak'})]}, \quad (6)$$

$$\gamma_{ak} = \alpha_k + \sum_{g=1}^{\mathcal{G}} Q_{kga}, \quad (7)$$

for given α_k and where $\psi(z)$ is the digamma function. The model parameters are obtained from the following update equations:

$$\mu_{gk} = \frac{\sum_{a=1}^{\mathcal{A}} Q_{kga} e_{ga}}{\sum_{a'=1}^{\mathcal{A}} Q_{kga'}}, \quad (8)$$

$$\sigma_{gk}^2 = \frac{\sum_{a=1}^{\mathcal{A}} Q_{kga}(e_{ga}-\mu_{gk})^2}{\sum_{a'=1}^{\mathcal{A}} Q_{kga'}}. \quad (9)$$

The update rule for the Dirichlet model parameter α_k is found from the derivatives of the α 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 the Appendix of this paper for further details).

The generative process described above for samples can equally be used over genes (the features). In probabilistic terms: For each gene g, a Dirichlet random variable is sampled, then for each experimental condition a, an index is drawn with probability θ_k, and then the expression level e_{ga} is sampled from the appropriate Gaussian with mean and variance μ_{ak}, σ_{ak}, respectively. Note that representing genes in this manner only requires a change of indices in the parameters which simply amounts to a transpose of the expression matrix. We will give one example of a decomposition over genes (the yeast cell cycle data set) later in Section 3.2.1.

2.3 Learning with Nonuniform Priors: The MAP Solution

In the previous section, we derived a set of update equations, guaranteed to converge to a local maximum of the lower bound on the likelihood. We now extend this argument to the MAP solution with informative priors. A suitable prior on the means could be a Gaussian distribution with zero mean. This would reflect a prior belief that most genes will be uninformative and will have log-ratio expression values around zero (i.e., they are unchanged compared to a reference sample). For the variance, we may wish to define a prior that penalizes overcomplex models and avoids overfitting. Overfitting may occur when Gaussian functions contract onto a single data point causing poor generalisation and numerical instability. With a suitable choice for the prior an extension of our previous derivation to the full MAP solution is straightforward. Our combined likelihood and prior expression is (assuming a uniform prior on α):

$$p(\mu,\sigma,\alpha|\mathcal{D}) \propto p(\mathcal{D}|\mu,\sigma,\alpha)p(\mu)p(\sigma). \quad (10)$$

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

$$\alpha,\sigma,\mu = \arg\max_{\alpha,\sigma,\mu} \quad \log p(\mathcal{G}|\mu,\sigma,\alpha) + \log p(\mu) + \log p(\sigma). \quad (11)$$

Thus, we can simply append these terms onto our bound on (1). Noting that they are functions of μ and σ only (and any associated hyperparameters), we conclude that these terms only change the update equations for μ_{ak} and σ_{ak}. Let us assume the following priors:

$$p(\mu_{gk}) \propto \mathcal{N}(0,\sigma_\mu), \quad (12)$$

$$p(\sigma_{gk}^2) \propto \exp\left\{-\frac{s}{\sigma_{gk}^2}\right\}, \quad (13)$$

i.e., a zero-mean Gaussian for the means and an improper prior for σ_{gk} (i.e., not integrating to unity over its domain) which gives zero weight to $\sigma_{gk}=0$ and asymptotically approaches 1 as σ_{gk} approaches infinity. Plots of these priors can be seen in Fig. 2 where $\sigma_\mu = 0.1$ and $s = 0.1$.

Adding these to our original bound, we now have to differentiate the following expression to obtain updates for μ_{gk} and σ_{gk}

$$\sum_{a=1}^{\mathcal{A}} Q_{gka} \log p(e_{ga}|\mu_{gk},\sigma_{gk}) + \log p(\mu_{gk}) + \log p(\sigma_{gk}). \quad (14)$$

Performing the necessary differentiations, we are left with the following new updates

$$\mu_{gk} = \frac{\sigma_\mu^2 \sum_{a=1}^{\mathcal{A}} Q_{gka} e_{ga}}{\sigma_{gk}^2 + \sigma_\mu^2 \sum_{a=1}^{\mathcal{A}} Q_{gka}}, \quad (15)$$

$$\sigma_{gk}^2 = \frac{\sum_{a=1}^{\mathcal{A}} Q_{gka}(e_{ga}-\mu_{gk})^2 + 2s}{\sum_{a=1}^{\mathcal{A}} Q_{gka}}. \quad (16)$$

These expressions clearly show the effect of the priors. In (15), we see that we now have a dependancy on both σ_{gk}^2 and σ_μ^2. In the limit of $\sigma_\mu^2 \to \infty$, we effectively have a uniform prior

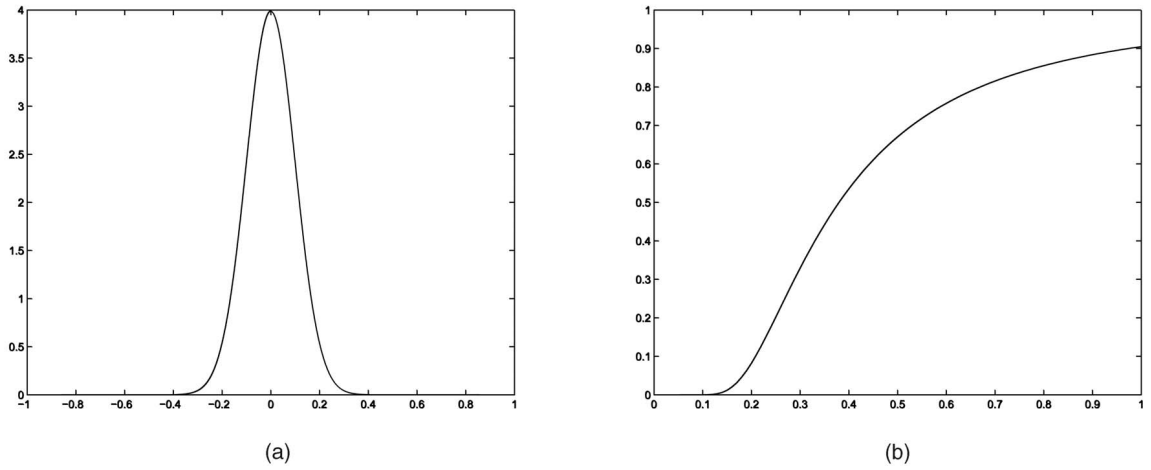


Fig. 2. Example priors for LPD μ_{gk} and σ_{gk} parameters. (a) Prior for μ_{gk}. (b) Prior for σ_{gk}.

and we recover our original update equations for the maximum likelihood solution. In the opposite limit of $\sigma_\mu^2 \to 0$, we find $\mu_{gk} \to 0$ as we would expect. In (16), the prior is effectively putting a lower bound on the value of σ_{gk}^2. As $s \to 0$, the prior approaches a uniform distribution over all positive reals and we recover our original σ_{gk}^2 update.

2.4 Computing the Likelihood and Parameter Initialization

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

$$\mathcal{L} = \prod_{a=1}^{\mathcal{A}'} \int_\theta \left\{ \prod_{g=1}^{\mathcal{G}} \sum_{k=1}^{\mathcal{K}} \mathcal{N}(e_{ga}|k, \mu_{gk}, \sigma_{gk})\theta_k \right\} p(\theta|\alpha) d\theta, \qquad (17)$$

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

$$\mathcal{L} \approx \prod_{a=1}^{\mathcal{A}'} \frac{1}{N} \sum_{n=1}^{N} \left\{ \prod_{g=1}^{\mathcal{G}} \sum_{k=1}^{\mathcal{K}} \mathcal{N}(e_{ga}|k, \mu_{gk}, \sigma_{gk})\theta_{kn} \right\}. \qquad (18)$$

Before iteratively updating Q_{kga}, γ_{ak}, μ_{gk}, σ_{gk}^2, and α_k various parameter values need to be initialized. In the numerical experiments reported below, the process mean values μ_{gk} were initialized to the mean expression value across the data set for each gene. Similarly, the variances σ_{gk}^2 were set to the variance of their respective genes. The α_k values were initialized to 1 for all k and the variational Dirichlet parameters γ_a were initialized to a positive random number.

3 Numerical Experiments

To evaluate the performance of Latent Process Decomposition (LPD) we applied it to four microarray data sets. For the first three data sets (prostate, lung, and lymphoma cancer), we show that LPD compares favorably with other cluster analysis methods, including dendrograms and Naive-Bayes, in its ability to classify samples. In the fourth example for the yeast cell cycle [11], we show that LPD can generate a rich, biologically relevant description of complex experiments. In these numerical studies, the only preprocessing was to take the logarithm of the data values (base 2) if this had not already been performed by the originators of the data set. Any likelihood calculations were approximated by (18) where samples from $p(\theta|\alpha)$ were taken by using k independent samples from gamma distributions with parameters α_k and then normalizing the result to sum to 1. The algorithm performs EM-like updates of the variational and model parameters. As such, convergence is guaranteed, although only to a local maximum of the training likelihood and not to the global maximum. Therefore, any analysis given below corresponds to the best solution (in terms of the likelihood) achieved over several starts with different initial γ_a values.

When estimating the best value of $\mathcal{K}$ for a particular data set, an N-fold cross-validation procedure was used. This consisted of randomly partitioning the data into N subsets, one of which would be removed before training and then used to calculate the likelihood. Each of the N subsets would be left out in turn and the N likelihood values would be averaged to give the estimate of the true value.

3.1 Prostate Cancer Data Set

Our first cancer data set comes from a study of prostate cancer by Dhanasekaran et al. [4] consisting of 54 samples with 9,984 features (we used their complete data set corresponding to the supplementary information of [4, Fig. 8]). This data set consists of 14 samples for benign prostatic hyperplasia (labeled BPH), three normal adjacent prostate (NAP), one normal adjacent tumour (NAT), 14 localized prostate cancer (PCA), one prostatitis (PRO), and 20 metastatic tumours (MET). We will use this data set to compare performance with uniform priors and nonuniform priors and to compare Latent Process Decomposition with hierarchical cluster analysis.

Fig. 3a shows the result of hierarchical cluster analysis using a dendrogram with average linkage and a Euclidean distance measure. Although some local structure can be identified, we see that the dendrogram gives a poor separation between known classes (in line with results presented in Dhanasekaran et al. [4]). However, the number of differentially expressed genes can be expected to be small relative to the total number of genes. As an unsupervised method we cannot use prior class knowledge but we can use a prior belief that genes varying little across samples are less likely to be interesting. Hence, to investigate the impact

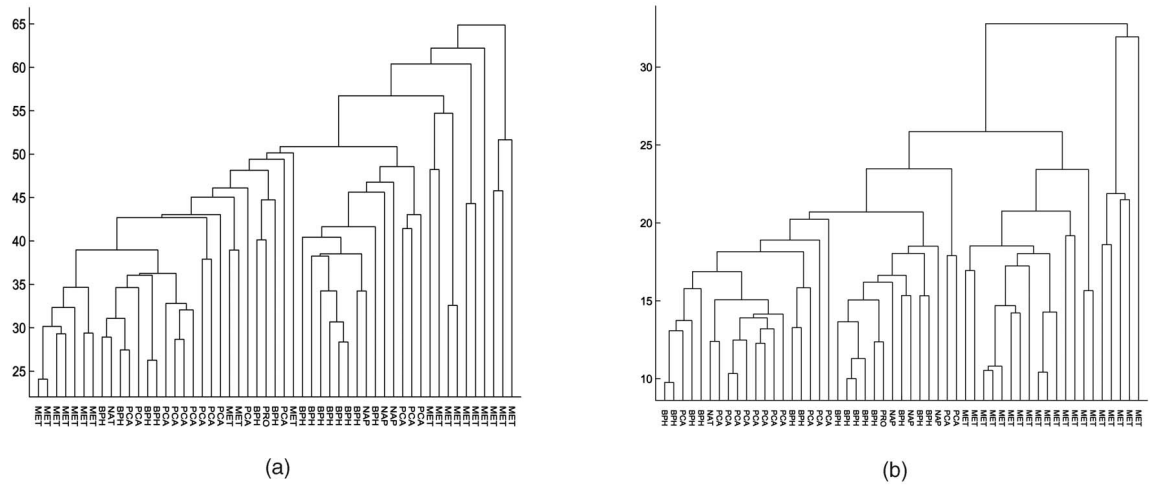


Fig. 3. Result of hierachichal cluster analysis on the prostate data set, using average linkage and a Euclidean distance measure. (a) Dendrogram with all features. (b) Dendrogram with top-ranked 500 features.

of feature selection we ranked genes based on variance across samples and used the subset of features with highest variance. Fig. 3b shows the corresponding result for performing hierarchichal clustering on this reduced data set. Although the dendrogram manages to split metastatic from the rest successfully, the difference between PCA and BPH is still unclear and it is very difficult to infer from this dendrogram just how many clusters are present.

In Fig. 4 (lower and descending curve) we give the hold-out log-likelihood versus $\mathcal{K}$, the number of processes, with a *uniform* prior and using the top-ranked 500 features. The log-likelihood peaks at about $\mathcal{K} = 4$ processes, after which a model with uniform prior overfits the training data as we increase the number of processes. This is to be expected with a maximum likelihood approach as we are allowing the model too much freedom by not penalising complexity. To solve this problem, we can incorporate the informative priors discussed in Section 2.3. We will discuss the determination of the parameters in these priors shortly. However, to illustrate performance we give the log-likelihood versus $\mathcal{K}$ plot in Fig. 4

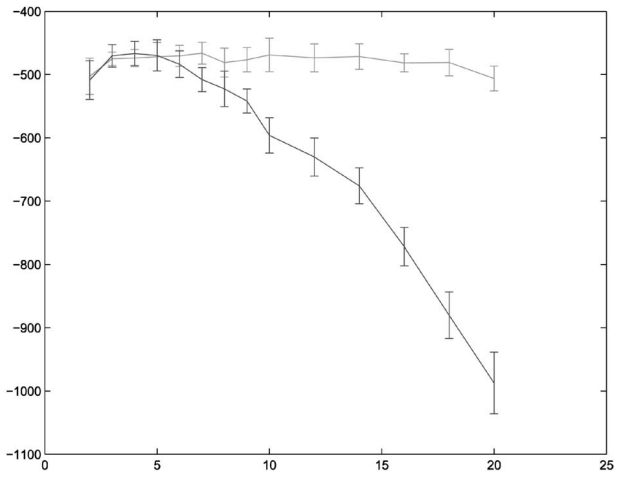


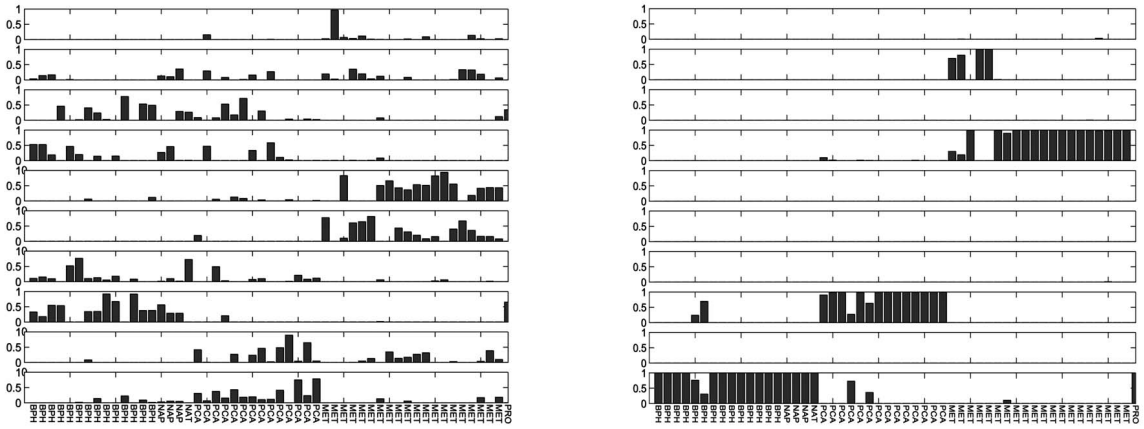
Fig. 4. Hold-out log-likelihood (y-axis) versus the $\mathcal{K}$ (x-axis) using the top ranked 500 genes in the prostate cancer data set using a nonuniform prior as given in Section 2.3 (upper, level curve) and a uniform prior (lower, descending curve) as described in Section 2.2.

(top, level curve) using values of $\sigma_\mu^2 = 0.1$ and $s = 0.1$. We see that the log-likelihood remains approximately constant as the number of processes is altered, rather than descending as $\mathcal{K}$ increases.

Fig. 5a provides an explanation for this behaviour: this is a decomposition of the prostate data into $\mathcal{K} = 10$ processes with a uniform prior (and using the top 500 genes ranked by variance). Given that the estimated optimum number of processes is 4, we might expect the model to severely overfit, as indeed it does. However, if we introduce priors with $\sigma_\mu^2 = 0.1$ and $s = 0.5$, we obtain the decomposition seen in Fig. 5b. We observe that the decomposition is over just four processes, with the remaining six processes empty. This is very useful since, given suitable values for the prior parameters, we do not need to decide the exact number of processes. We also notice that the decomposition is much cleaner. For example, the metastatic samples (MET) have no overlap with localized prostate cancer (PCA) and the benign and normal samples indicating they are a distinct state.

Although we can avoid overfitting with correct parameters for the prior, this still leaves the question of the correct choice of these prior parameters. Our numerical investigations across different data sets indicated that the mean parameter, σ_μ, had negligible effect and we typically set this parameter to 0.1. However, performance did depend on s, the variance parameter. In Fig. 6, we plot from a cross validation study of the log-likelihood versus s for the 500 top-ranked genes in the prostate data set. We notice a peak at $s = 0.1$, though any choice for s on the range 0.1 and 0.5 appears satisfactory.

It may appear that we have gained very little by moving from cross-validation over processes to one over s, but arguably this is not the case for two reasons. First, we are dealing with a small number of samples and, hence, by removing some samples for cross-validation, we could easily remove a substantial proportion of one particular class, suggesting a bias toward underestimating the number of classes present. The variance parameter on the other hand is a measure of noise that should be class independent and, hence, should be the same regardless of whether or not the full data set is being used. Also, the number of processes present will be dependent on the particular data set and could vary greatly depending on the given study. On the other



(a) (b)

Fig. 5. $\mathcal{K} = 10$ process decomposition of the prostate cancer data set using the top 500 genes. The peaks represent the confidence in the assignment of the ath sample to the kth process using γ_{ak}. (a) With a uniform prior. (b) Using a nonuniform prior with $\sigma_\mu^2 = 0.1$ and $s = 0.5$.

hand, the contribution made by noise is a consequence of the experimental procedure and we may expect the optimum value of s to vary very little across data sets. Indeed, in the following sections, we shall see that this is the case.

One advantage of LPD over standard clustering techniques is the ease with which information can be extracted from the derived model. In this prostate cancer example, we could be interested in finding those genes most differentially expressed between localized prostate cancer (PCA) and the metastatic samples (MET), for example. The following score can be used to rank genes:

$$Z = \frac{|\mu_{gi} - \mu_{gj}|}{\sqrt{\sigma_{gi}^2 + \sigma_{gj}^2}}, \tag{19}$$

where i and j correspond to the two processes being compared. Genes ranking highly will be those with large separation between their means and small standard deviations—corresponding to genes whose values fall consistently into two distinct clusters. In Table 1, we give a list of the 10 top-ranked genes based on this score. Plots of the mixture densities defined by the latent processes and the actual data values for three of these genes can be seen in Fig. 7. These three plots show that the extracted genes appear to distinguish well between localized cancer and metastasis and, importantly, have been highlighted as significant without the use of the class labels. An investigation of these genes (Dr. Jeremy Clark, personal communication) indicates that some of them are potentially significant for cancer biology. For example, CTGF is an immediate-early gene induced by growth factors or certain oncogenes. The expressed protein binds to integrins and modulates the invasive behavior of certain human cancer cells [3]. Similarly, it has been reported [5] that tumor angiogenesis and tumor growth are critically dependent on the activity of EGR1.

We have used the likelihood as our measure of model performance. One issue that we have not yet explored is how these likelihood values compare with those for alternative, and possibly simpler, cluster analysis methods. The most obvious comparison is with a standard Gaussian mixture model. The main difference between LPD and a

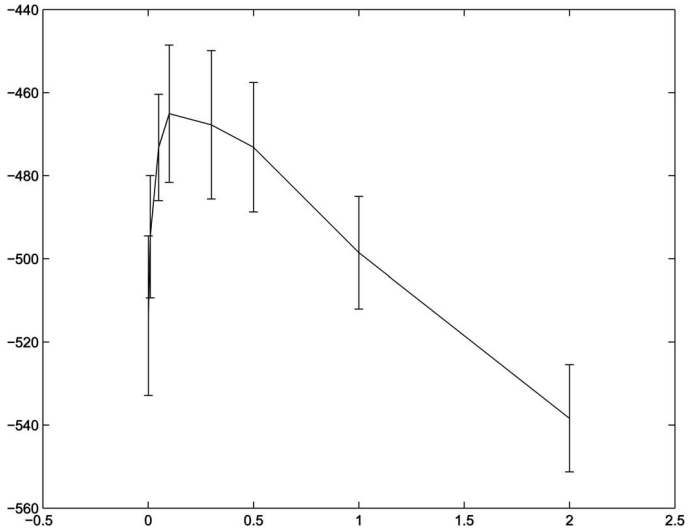
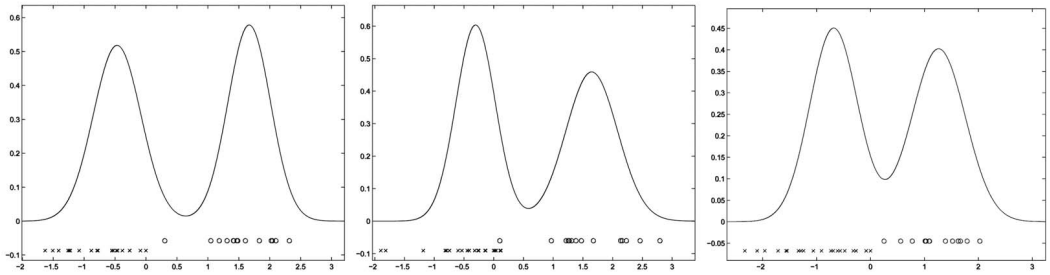


Fig. 6. Hold-out log-likelihood (y-axis) as a function of the variance prior parameter s (x-axis) using the top 500 features in the prostate data set.

TABLE 1
Top 10 Genes Separating Localized Prostate from Metastatic Prostate Cancer

Rank	Name	Z
1	translocating chain-associating membrane protein (TRAM)	4.243369
2	early growth response 1 (EGR1)	4.131861
3	connective tissue growth factor (CTGF)	3.579336
4	osteoglycin (osteoinductive factor, mimecan) (OGN)	3.339221
5	golgi membrane protein (GP73)	3.202468
6	butyrate response factor 1 (EGF-response factor 1) (BRF1)	3.074948
7	ribosomal protein L5 (RPL5)	2.985735
8	dopachrome delta-isomerase, tyrosine-related protein 2 (DCT)	2.929469
9	pericentriolar material 1 (PCM1)	2.877109
10	sprouty (Drosophila) homolog 4	2.755093

Genes are ranked by Z.



(a) (b) (c)

Fig. 7. Inferred density plots for three genes separating localized prostate cancer (PCA, process 2) from the metastatic samples (MET, process 4) in Fig. 5b. Actual data values are shown below the mixtures, separated by class with circles representing expression values for localized prostate cancer and crosses representing values for metastatic prostate cancer samples. (a) Early growth response (EGR1). (b) Connective tissue growth factor (CTGF). (c) Dopachrome delta-isomerase, tyrosine-related protein 2 (DCT).

Gaussian mixture model is the repeated sampling for each feature in LPD allowing each sample to be represented by a combination of processes.

Due to the large number of features involved, it is computationally infeasible to use a Gaussian mixture model with full covariance matrices and thus we shall constrain the matrices to be diagonal (equivalent to Gaussians aligned with the coordinate axis). Fig. 8 shows the log-likelihood curves using maximum likelihood (i.e., no priors) for LPD and Naive-Bayes. We see that Naive-Bayes begins overfitting immediately, suggesting only two clusters in the data. LPD rises above Naive-Bayes to a peak at $\mathcal{K} = 4$ processes before dropping toward the Naive-Bayes curve as $\mathcal{K}$ increases. Although the difference in the log-likelihood is not huge, it is statistically significant, given the error bars, and the suggestion of only two clusters from Naive-Bayes is misleading. Note that we have used the maximum likelihood approach for training LPD to give the fairest comparison with Naive-Bayes which is also trained by maximum likelihood. It is also worth noting the relative sizes of the error bars. For Naive-Bayes, it is possible to calculate the likelihood *exactly*, whereas with LPD we must take samples from the Dirichlet. The error bars for the two methods are certainly of the same order, suggesting that most variation is due to the cross-validation procedure thus allowing us to be confident in our LPD likelihood approximation method.

3.2 Lung Cancer Data Set

The lung cancer data set [6] consists of 73 gene expression profiles from normal and tumour samples with the tumors labelled as squamous, large cell, small cell, and adenocarcinoma. Ten-fold cross validation was performed for values of $\mathcal{K}$ between 2 and 20 with and without parameter priors and the results are shown in Fig. 9. As for the prostate cancer example, we notice a nonuniform prior avoids overfitting and causes the log-likelihood to plateau. In [6], the authors identified seven clusters in the data—with the adenocarcinoma samples falling into three separate clusters with strong correlation with clinical outcomes. For their ordering (which

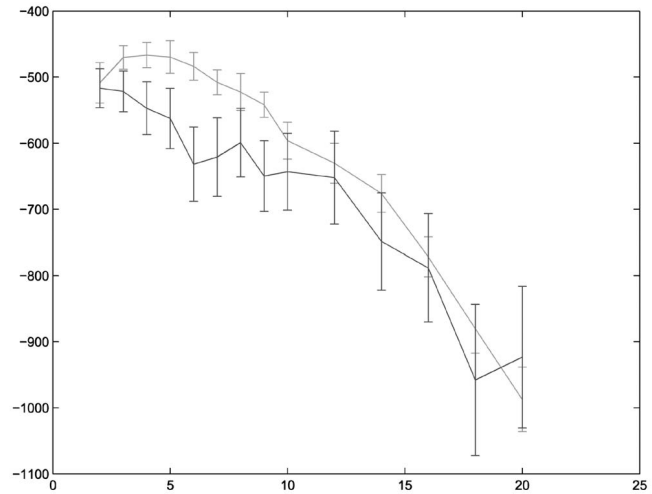


Fig. 8. Hold-out log-likelihood (y-axis) for maximum likelihood LPD (upper curve) and Naive-Bayes (lower curve) for various value of $\mathcal{K}$ (x-axis). The top 500 features are used for both models.

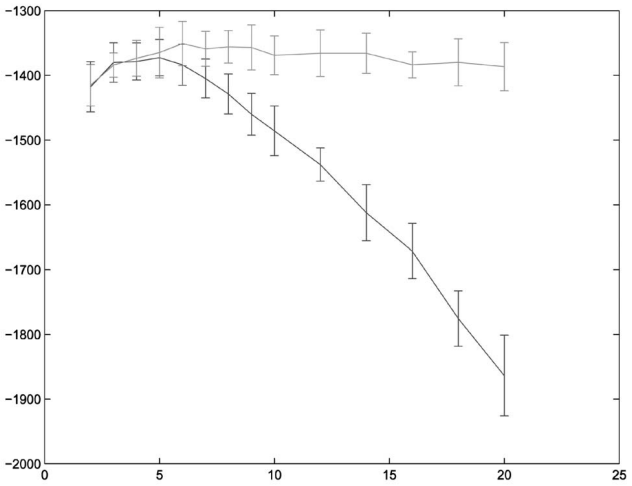


Fig. 9. Hold-out log-likelihood (y-axis) versus $\mathcal{K}$ (x-axis) for the lung cancer data set *with* (upper, level curve) and *without* parameter priors (lower, descending curve).

we will follow), samples 1-19 belong to adenocarcinoma cluster 1, samples 20-26 belong to adenocarcinoma cluster 2, samples 27-32 are normal tissue samples, samples 33-43 are adenocarcinoma cluster 3, samples 44-60 are squamous cell carcinomas, samples 61-67 are small cell carcinomas, and samples 68-73 are from large cell tumors. For this example and the following data set (for lymphoma), features selection was found to have little effect so we give the analysis without feature selection.

As in the previous section, the plateauing of the log-likelihood with a nonuniform prior suggests that we are using more processes than are necessary and additional processes are left empty. The onset of a plateau in the log-likelihood indicates the minimum number of processes to use and $\mathcal{K} = 10$ would seem reasonable. The cross-validation study over s is given in Fig. 10 and we can see a peak between 0.3 and 0.5. This is approximately the same peak value found for the prostate cancer data set.

In Fig. 11, we use LPD to decompose the data into 10 processes with the bars representing the normalized

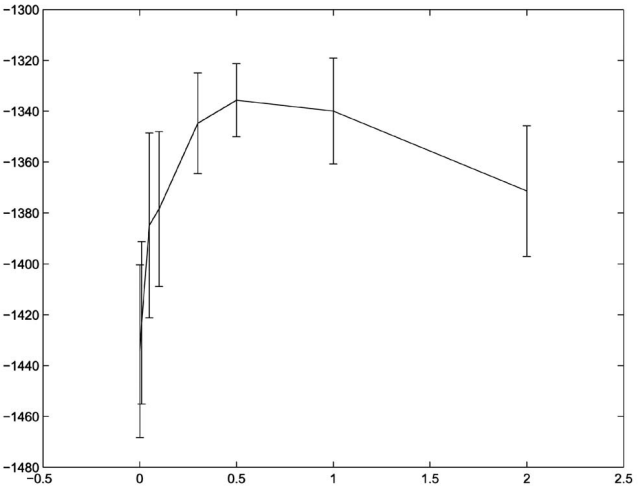


Fig. 10. Hold-out log-likelihood (y-axis) as a function of s (x-axis) for the lung cancer data set.

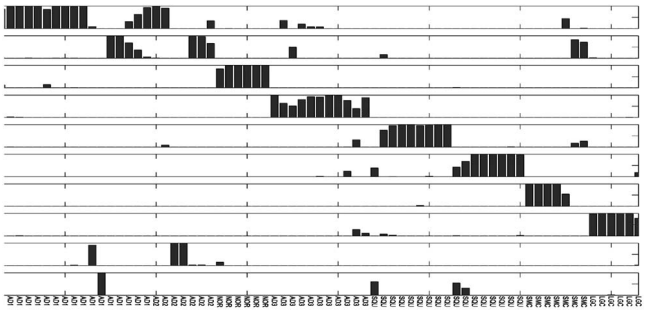


Fig. 11. Normalized γ_{ak} values for a 10 process decomposition of the lung cancer data.

γ_{ak} values (equivalent to $\mathbf{E}_{p(\theta|\gamma_a)}[\theta]$) and expressing the confidence in the allocation of the $a\text{th}$ sample to the $k\text{th}$ process. The samples are in the order in which they are presented in the original paper [6].

First, we observe that the clustering by dendrogram is accurately reproduced by latent process decomposition. One notable difference is that the last two entries for the small cell carcinoma grouping are placed with the adenocarcinoma cluster 2 by latent process decomposition with high values of probability (close to one). For the original clustering by dendrogram, these two samples are two adenocarcinomas inaccurately placed by the dendrogram in a cluster with small cell carcinomas. This difference is therefore in favor of latent process decomposition. For latent process decomposition, we find that the adenocarcinoma clusters 1 and 2 decompose across processes 1 and 2 and adenocarcinoma cluster 3 is almost exclusively represented by process 4. There are rare instances of adenocarcinoma placed in processes 9 and 10. The normal samples (process 3) and large cell tumors (process 8) are very distinct. The squamous cell carcinomas are decomposabled into processes 5 and 6, both of which have little commonality with any other tumor types in the data set. This might suggest possible further taxonomic decomposition into subtypes. However, this possibility should be treated with caution: decomposition into subprocesses could also mirror the different genetic stages of a single subtype, for example.

In [6], the class labels were used to extract genes which distinguished well between the different adenocarcinoma clusters. Using the $Z\text{-score}$ described in the previous section, we give the top 10 ranked genes distinguishing processes 1 and 4 (adenocarcinoma 1 and adenocarcinoma 3) in Table 2. Of these top genes, numbers 1, 2, and 10 were discussed in [6] as being significant for cancer biology. Also, the biological relevance of the detected genes is highlighted by the internal consistency of the list: two antagonistic proteins of sugar metabolism and a variety of transmembrane proteins (including two solute carriers) dominate the picture. Importantly, and in constrast to Garber et al. [6], these features have been extracted with no prior knowledge of class membership.[1]

Finally, as for the prostate cancer study, we can use the hold-out log-likelihood to compare LPD with Naive-Bayes. Again, we will use the maximum likelihood method to train

1. For a more complete list and comparisons between other processes, see http://www.enm.bris.ac.uk/lpd.

TABLE 2
Top Lung Genes, Ranked by Z

Rank	Name	Z
1	ESTs Hs.11607	2.5462
2	Solute carrier family 7, member 5	2.1824
3	Transmembrane, prostate androgen induced RNA	1.8716
4	Trophinin associated protein (tastin)	1.8706
5	Fructose-1,6-bisphosphatase 1	1.7742
6	Epididymal secretory protein (19.5kD)	1.7589
7	Phosphofructokinase, platelet	1.7040
8	ESTs Hs.76728 H00660	1.6998
9	Solute carrier family 2, member 1	1.6893
10	Protein tyrosine kinase 7	1.6774

both models to ensure a fair comparison. The results can be seen in Fig. 12. Although the difference between the two methods is small, LPD tends to overfit less quickly as $\mathcal{K}$ increases.

3.3 Lymphoma Cancer Data Set

Next, we consider the Lymphoma data set of Alizadeh et al. [1], which consists of 96 samples of normal and malignant lymphocytes. In this study, the authors used samples from diffuse large B-cell lymphoma (DLBCL), the most common form of non-Hodgkin's lymphoma in addition to including data for follicular lymphoma (FL), chronic lymphocytic leukaemia (CLL), and other cell types.

Determining the hold-out log-likelihood versus $\mathcal{K}$ with 10-fold cross-validation, we find overfitting with a maximum likelihood solution while the log-likelihood plateaus after $\mathcal{K} = 10$, suggesting this is the smallest number of processes to use for a good fit to the data. Using $\mathcal{K} = 10$ we can cross-validate over different values of the variance prior parameter s and the results are shown in Fig. 13.

A decomposition into $\mathcal{K} = 10$ processes (using the value of s that gave the maximum in Fig. 13) can be seen in Fig. 14. The samples are presented in the same order as Alizadeh et al. [1] with samples 1 to 16 largely consisting of DLBCL with samples 1 and 2 consisting of two DLBCL cell lines (OCI Ly3 and Ly10), while 10 and 11 are tonsil germinal cell samples.

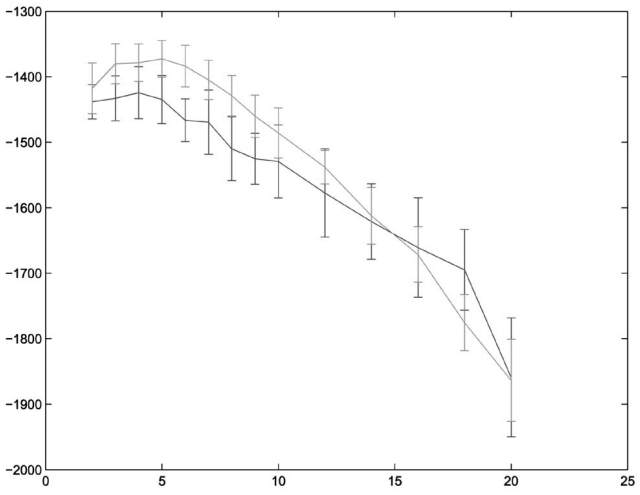


Fig. 12. Hold-out log-likelihood (y-axis) for maximum likelihood LPD (top curve for $\mathcal{K} = 5$) and Naive-Bayes (bottom curve for $\mathcal{K} = 5$) for various value of $\mathcal{K}$ (x-axis).

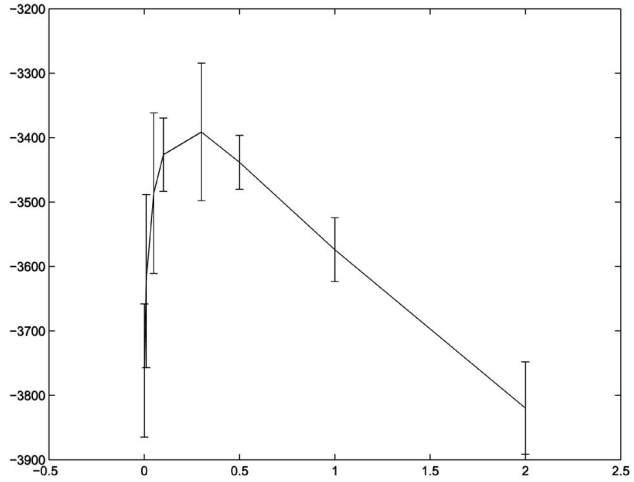


Fig. 13. Hold-out log-likelihood (y-axis) versus the variance prior parameter, s (x-axis), for the lymphoma data set.

Samples, 17 to 47 consist of DLBCL, samples 48 to 57, 58 to 63, and 65 to 70 consist of blood B and other samples, sample 64 is the OCI Ly1 cell line, sample 71 is a single DLBCL sample placed in with this general cluster of blood B and other tissues, samples 72-80 are follicular lymphoma (FL), samples 85 to 95 are CLL, and 96 is a lone DLBCL sample. The decomposition agrees well with the structure obtained using a dendrogram [1]. Decomposition of the DLCL samples into several processes suggests substructure supporting the discussion in [1]. LPD and the dendrogram do differ, e.g., sample 71 (DLCL) is now partly grouped with the other DLCL samples and not the FL samples in contrast to the dendrogram. Also, process 6 brings together the OCI samples which was not achieved by the dendrogram.

Finally, in Fig. 15, we compare the hold-out log-likelihood for LPD with that for the Naive-Bayes model. As in the previous examples, we see that LPD does not overfit as quickly and suggests a greater number of processes than the Naive-Bayes model. It also gives a higher maximum value, suggesting it is better suited to this particular data set.

3.4 Yeast Cell Cycle Data Set

For the next example, we have used the yeast cell cycle data set of Spellman et al. [11]. In this study, the authors investigate which yeast genes show periodically varying expression levels during cell cycle progression, using whole-genome microarrays covering approximately 6,200 genes. Using a cross-validation study of log-likelihood versus s with a

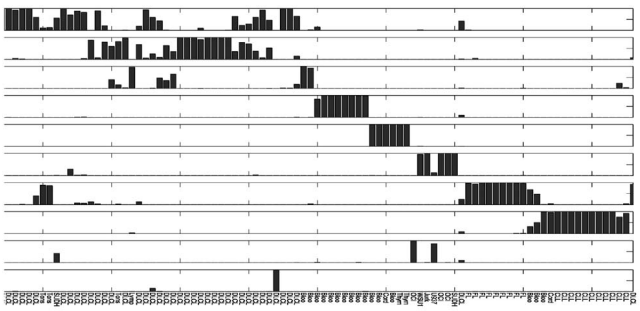


Fig. 14. $\mathcal{K} = 10$ process decomposition of the lymphoma data.

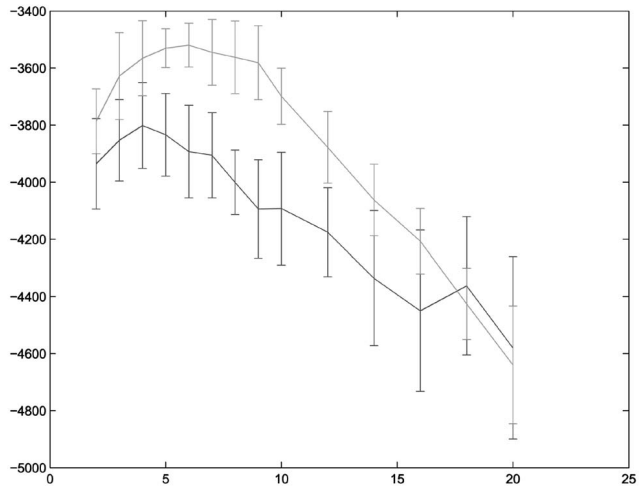


Fig. 15. Hold-out log-likelihood (y-axis) as a function of $\mathcal{K}$ (x-axis) for maximum likelihood LPD and the Naive-Bayes Gaussian mixture model.

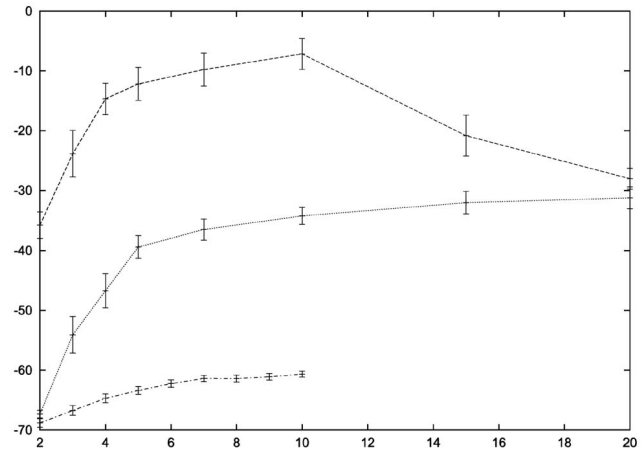


Fig. 16. Comparison of the log-likelihood (y-axis) versus $\mathcal{K}$ (x-axis) for LPD (top curve) and a Naive-Bayes mixture model (middle curve). We have additionally made a comparison with a Bayesian Network model introduced by Segal et al. [10] (bottom curve) which represents process membership over genes rather than samples.

nonuniform prior, we found the curve achieved its peak value at $s = 0$ suggesting that a nonuniform prior was not of benefit. Thus, the results in this section are presented using the maximum likelihood solution of Section 2.2. In contrast to the previous three examples, we will now decompose over genes rather than samples (see the end of Section 2.2). In Fig. 16, we perform a comparison of the log-likelihood curves versus $\mathcal{K}$ for three methods and in all cases LPD gives the best solution for these data sets.

Fig. 16 compares the predictive likelihood for LPD, Naive-Bayes and a Bayesian network model introduced by Segal [10]. Segal's model decomposed the data into a set of processes where each gene's membership of each process is defined by a separate binary variable. Each process has an activity value defined for each array. The expression value for a particular gene in a particular array is assumed to be Gaussian where the mean is the sum of the activities for each process to which this gene belongs. This method is considered to be state-of-the-art and as such we have included it in our comparisons. One problem with Segal's method is that estimating the binary membership parameters exactly is NP-hard and so for values of $\mathcal{K}$ greater than 10, it becomes very time consuming. It is for this reason that the likelihood curves for Segal's method stops at $\mathcal{K} = 10$ processes in Fig. 16. Various heuristics exist to estimate these variables, but given the relative difference in log-likelihood between this and the other two methods, it seemed unnecessary to implement them.

While the previous data sets for cancer mainly demonstrate the performance of LPD in categorising data, this cell cycle data can reveal the biological usefulness of the identified latent processes and their combinations in any given sample. In this respect, LPD is able to provide a far richer picture than is available from a simple clustering approach. Fig. 16 shows that for this data set between 5 and 10 processes give the highest predictive likelihood. From the well-characterized biology of the experiment, we expect one process corresponding to each of the known cell cycle phases (G1, S, G2, M, MG1, as defined in the original paper [7]) plus a small number of "classifying" processes that distinguish between the four different synchronization methods that were used in the experiment. Fig. 17 shows the 10 processes defined by LPD. It can easily be seen that processes 1 and 3 represent "classifying" processes that characterize alpha factor/elutriation synchronization and CDC15 synchronization, respectively. Processes 2, 4, 6, and 8 correspond to cycling processes and each of them can be uniquely mapped to one of the major cell cycle phases by reference to the biological information given in [7] (highlighted in Fig. 17). The fifth cell cycle phase, G2, which is very short in exponentially growing yeast cultures, is not recovered as a distinct process but can easily be obtained by the linear combination of the adjacent S and M phases (not shown). The remaining processes (5, 7, 9, and 10) seem to be dominated by noise and experimental artefacts (particularly in process 7). This emphasizes that identification of the most parsimonious number of latent processes is an extremely useful feature of the LPD approach. This would be even more important in experiments without previous knowledge of the underlying process structure, where the identification of the really informative processes would not be as easy as in the present test case.

From a biological point of view, it is particularly noteworthy that the cycling processes were identified without any prior reference to the time-series structure of the experiment. The original publication used a Fourier transformation to determine the expression peaks of cyclically expressed genes and Pearson correlation with known cell cycle-regulated genes to assign them to cell cycle phases. In the case of LPD, the same can be achieved by reference to the gene-specific variational parameter γ_{gk}. This is shown in Fig. 18, where the expression of a classical G1 phase marker gene (cyclin CLN2) is indeed represented most strongly by process 2, which according to Fig. 17, corresponds to peak expression in G1.

Fig. 19 shows a more complex example of the power of this approach. The histone genes, which code for the main protein component of chromosomes, are known to be most strongly expressed in S phase, when new chromosomes need to be assembled. Indeed, Fig. 19 shows that expression of one

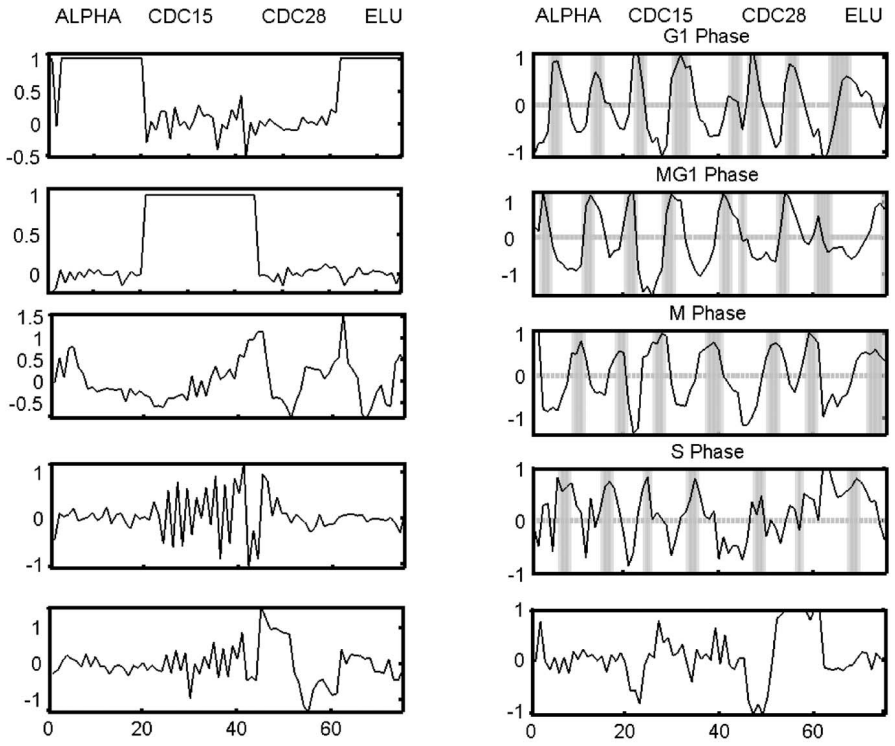


Fig. 17. The 10 processes identified by LPD for the yeast cell cycle. The cycling processes peak in the various phases of the cell cycle as highlighted on subplots 2, 4, 6, and 8. The four different synchronization methods are indicated at the top. Our numbering convention is odd numbers starting from the top subfigure for the left-hand figures and even numbers starting from the top for the right-hand figures.

representative histone (H1) is dominated by the S phase-specific process 8 (compare Fig. 17). However, there are also significant contributions by processes 2 and 6 (G1 and M phase) and the measured expression profile (solid line in Fig. 19, lower panel) shows clearly that this specific histone has a much broader expression range that extends well beyond S phase. This is not entirely unexpected as the production of histone proteins should precede the replication of DNA that defines S phase.

4 Conclusion

Our experimental results strongly indicate that latent process decomposition is an effective technique for discovering structure within microarray data. It has several advantages over the use of dendrograms or simple clustering: unambiguous determination of the number of processes via cross-validation using a hold-out (predictive)

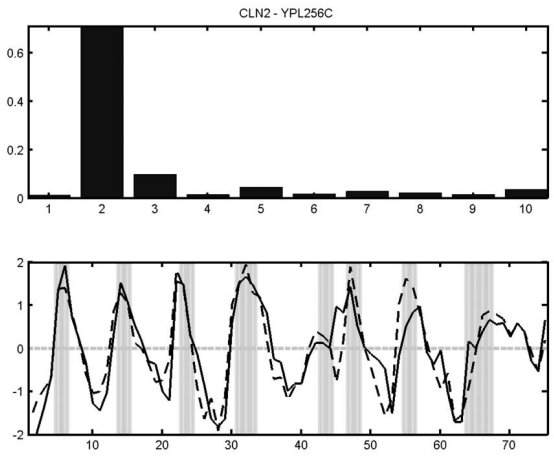


Fig. 18. The expression levels of the G1-specific cyclin CLN2 YPL256C peaks in G1 phase (lower panel, solid line). The dotted line is the sum of the mean for each process weighted by the value of the gene-specific variational parameter for each process (top bar chart). In this case, it can be seen that the second process profile (G1 phase) is indeed dominant for this gene.

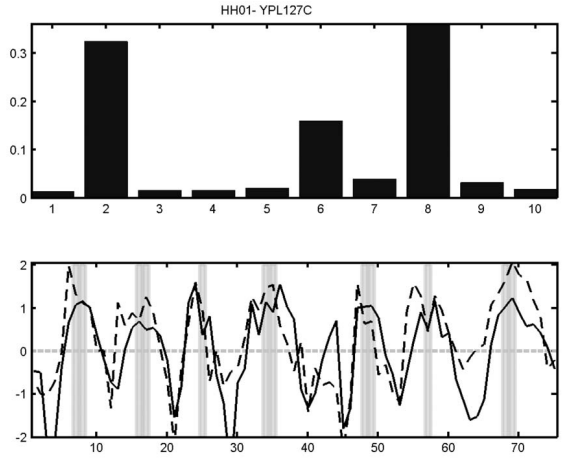


Fig. 19. The expression levels of the histone H1 gene (YPL127C) peaks in S phase (lower panel, solid curve). The dotted line is the sum of the mean for each process weighted by the value of the gene-specific variational parameter for each process (top bar chart). In this case, it can be seen that a combination of the second, sixth, and eighth process profiles is required for this histone. This combination corresponds to the G1, M, and S phase processes identified by LPD (Fig. 17).

likelihood. In addition, the model uses a shared latent space with the same explanatory variables for both samples and genes. Overfitting can be effectively controlled using a nonuniform prior, in contrast to many cluster analysis methods. Also, several processes can be simultaneously combined to determine gene expression levels so that the relationship between processes is more transparent. LPD has significant advantages over other cluster analysis methods, for example, models based on a mixture of Gaussians [9]. With the latter models, genes must be preselected followed by another level of feature extraction due to the rank deficiency of the estimated covariance matrix in the original space. This is not required in the approach advocated here. Given the large number of features used by microarray data sets, there is a redundancy of information. For this reason, we found that feature selection can sometimes improve the performance of the model. For lung and lymphoma data sets, we used no feature selection, but for the prostate cancer and cell cycle data sets, we used the top 500 and 1,000 genes as selected by variance spread. Feature selection in this context is an avenue for further research.

APPENDIX A

In this appendix, we will give a full derivation of the LPD model. We will outline the derivation of the maximum likelihood solution given in Section 2.2: the generalization to a nonuniform prior is straightforward and outlined in Section 2.3. The likelihood of a set of $\mathcal{A}$ training samples is as follows:

$$\prod_{a=1}^{\mathcal{A}} p(a|\mu,\sigma,\alpha) = \prod_{a=1}^{\mathcal{A}} \int_{\theta} p(a|\mu,\sigma,\theta)p(\theta|\alpha)d\theta, \tag{20}$$

where the probability of the sample conditioned on the means, variances, and process weightings can be expressed in terms of its individual components giving the following:

$$\prod_{a=1}^{\mathcal{A}} p(a|\mu,\sigma,\alpha) = \prod_{a=1}^{\mathcal{A}} \int_{\theta} \left\{ \prod_{g=1}^{\mathcal{G}} \sum_{k=1}^{\mathcal{K}} \mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk})\theta_k \right\} p(\theta|\alpha)d\theta. \tag{21}$$

For a concave function $f(x)$, Jensen's inequality states that $f\left(\mathbf{E}_{p(z)}[z]\right) \geq \mathbf{E}_{p(z)}[f(z)]$, where $\mathbf{E}_p(z) = \int zp(z)dz$ for continuous z. Thus, introducing a set of sample-specific variational Dirichlets $p(\theta|\gamma_a)$, we can derive a lower bound on the log-likelihood

$$\sum_{a=1}^{\mathcal{A}} \log p(a|\mu,\sigma,\alpha) \geq \sum_{a=1}^{\mathcal{A}} \mathbf{E}_{p(\theta|\gamma_a)} \left[\log \left\{ p(a|\mu,\sigma,\theta) \frac{p(\theta|\alpha)}{p(\theta|\gamma_a)} \right\} \right], \tag{22}$$

where

$$\mathbf{E}_{p(\theta|\gamma_a)}[f(\theta)] = \int_{\theta} f(\theta)p(\theta|\gamma_a)d\theta. \tag{23}$$

A set of variational parameters Q_{kga} are now introduced such that $\sum_k Q_{kga} = 1$ for bounding $p(a|\mu,\sigma,\theta)$ in the above inequality. These parameters can be regarded as the posterior distribution over the different processes for each g and a. Jensen's inequality can then be used again to give the following additional bound

$$\sum_{a=1}^{\mathcal{A}} \log p(a|\mu,\sigma,\theta) \geq \sum_{k=1}^{\mathcal{K}} \sum_{a=1}^{\mathcal{A}} \sum_{g=1}^{\mathcal{G}} Q_{kga} \log \left\{ \frac{\mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk})\theta_k}{Q_{kga}} \right\}. \tag{24}$$

Combining these bounds into one gives the following

$$\begin{aligned} \sum_a \log p(a|b,\mu,\sigma) \geq & \sum_a \mathbf{E}_{p(\theta|\gamma_a)}[\log p(\theta|\alpha)] \\ & - \sum_a \mathbf{E}_{p(\theta|\gamma_a)}[\log p(\theta|\gamma_a)] \\ & + \sum_a \sum_g \sum_k Q_{kga} \log \mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk}) \\ & - \sum_a \sum_g \sum_k Q_{kga} \log Q_{kga} \\ & + \sum_a \sum_g \sum_k Q_{kga} \mathbf{E}_{p(\theta|\gamma_a)}[\log \theta_k]. \end{aligned} \tag{25}$$

A.1 Model Parameters

The process mean and variance parameters μ_{gk} and σ^2_{gk} only appear in the third term of (25). This term can be expanded as follows:

$$\begin{aligned} & \sum_a \sum_g \sum_k Q_{kga} \log \mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk}) = \\ & \sum_a \sum_g \sum_k Q_{kga} \log \left[\frac{1}{\sqrt{2\pi\sigma^2_{gk}}} \right] - \sum_a \sum_g \sum_k Q_{kga} \frac{(e_{ga}-\mu_{gk})^2}{2\sigma^2_{gk}}. \end{aligned} \tag{26}$$

Taking partial derivatives of this expression with respect to μ_{gk} and σ_{gk}, setting the results to zero and solving yields the following updates

$$\mu_{gk} = \frac{\sum_a Q_{kga} e_{ga}}{\sum_a Q_{kga}}, \tag{27}$$

$$\sigma^2_{gk} = \frac{\sum_a Q_{kga} (e_{ga}-\mu_{gk})^2}{\sum_a Q_{kga}}. \tag{28}$$

The α parameter only appears in term 1 of (25), but partial differentiation of this term with respect to α_k reveals that a simple iterative procedure is not possible and a second order technique is required. Term 1 can be written

$$\begin{aligned} & \sum_a \mathbf{E}_{p(\theta|\gamma_a)}[\log(p(\theta|\alpha)] = \\ & \sum_a \mathbf{E}_{p(\theta|\gamma_a)} \left[\log \left\{ \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \right\} + \log \prod_k \theta_k^{(\alpha_k-1)} \right]. \end{aligned} \tag{29}$$

Thus, differentiating this expression we get

$$g_i = \mathcal{A}[\psi(\overline{\alpha}) - \psi(\alpha_i)] + \sum_{a=1}^{\mathcal{A}} [\psi(\gamma_{ia}) - \psi(\overline{\gamma_a})], \tag{30}$$

from which we can derive the corresponding Hessian

$$H_{ij} = \psi'(\overline{\alpha}) - \delta_{ij}\mathcal{A}\psi'(\alpha_i), \tag{31}$$

$$\alpha_{new} = \alpha_{old} - \mathbf{H}(\alpha_{old})^{-1}\mathbf{g}(\alpha_{old}), \tag{32}$$

where $\mathbf{H}(\alpha)$ is the Hessian matrix and $\mathbf{g}(\alpha)$ is the gradient. For $i, j = 1, \ldots, \mathcal{K}$, their components are given by:

$$g_i = \mathcal{A}[\psi(\overline{\alpha}) - \psi(\alpha_i)] + \sum_{a=1}^{\mathcal{A}}[\psi(\gamma_{ia}) - \psi(\overline{\gamma_a})], \tag{33}$$

$$H_{ij} = \psi'(\overline{\alpha}) - \delta_{ij}\mathcal{A}\psi'(\alpha_i), \tag{34}$$

where $\overline{\gamma_a} = \sum_{j=1}^{\mathcal{K}} \gamma_{jc}$, $\overline{\alpha} = \sum_{k=1}^{\mathcal{K}} \alpha_k$, and $\psi'(\theta)$ is the trigamma function. Since the Hessian matrix is of the form

$$\mathbf{H} = \mathrm{diag}(\mathbf{h}) + \mathbf{1}z\mathbf{1}^T, \tag{35}$$

matrix inversion can be avoided [2] and the ith component of the correction in the iterative update rule can be found from

$$(\mathbf{H}^{-1}\mathbf{g})_i = (g_i - r)/h_i, \tag{36}$$

where

$$r = \left(\sum_{j=1}^{k}(g_j/h_j)\right) / \left(z^{-1} + \sum_{j=1}^{k} h_j^{-1}\right). \tag{37}$$

A.2 Variational Parameters

First, Q_{kga} appears in terms 3, 4, and 5 of (25). Taking these terms, removing summations over k, a, and g and adding a Lagrangian to satisfy the summation constraint, we have

$$Q_{kga}\log\mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk}) - Q_{kga}\log Q_{kga} + Q_{kga}\mathbf{E}_{p(\theta|\gamma_a)} [\log\theta_k] - \lambda\left(\sum_a Q_{kga} - 1\right), \tag{38}$$

taking partial derivatives with respect to Q_{kga} and setting these to zero gives

$$0 = \log\mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk}) - (\log[Q_{kga}] + 1) + \mathbf{E}_{p(\theta|\gamma_a)}[\log\theta_k] - \lambda, \tag{39}$$

solving for λ and rearranging gives the following update for Q_{kga}

$$Q_{kga} = \frac{\mathcal{N}(e_{ga}|k,\mu_{gk},\sigma_{gk})\exp[\mathbf{E}_{p(\theta|\gamma_a)}[\log\theta_k]]}{\sum_{a'=1}^{\mathcal{A}}\mathcal{N}(e_{ga'}|k,\mu_{gk},\sigma_{gk})\exp[\mathbf{E}_{p(\theta|\gamma_{a'})}[\log\theta_k]]}, \tag{40}$$

where

$$\mathbf{E}_{p(\theta|\gamma_a)}[\log\theta_k] = \Psi(\gamma_{ak}) - \Psi\left(\sum_j \gamma_{aj}\right). \tag{41}$$

Now, the γ_{ak} variable appears in terms 1, 2, and 5 of the above bound. Extracting these terms and removing summations over k and a leaves

$$-\mathbf{E}_{p(\theta|\gamma_a)}[\log p(\theta|\gamma_a)] + \mathbf{E}_{p(\theta|\gamma_a)}[\log p(\theta|\alpha)] + \sum_g Q_{kga}\mathbf{E}_{p(\theta|\gamma_a)}[\log\theta_k]. \tag{42}$$

Now,

$$\begin{aligned}\mathbf{E}_{p(\theta|\gamma_a)}[\log p(\theta|\gamma_a)] &= \int_\Delta \log\left[\frac{\Gamma(\sum_k\gamma_{ak})}{\prod_k\Gamma(\gamma_{ak})} + \sum_k(\gamma_{ak}-1)\log[\theta_k]\right] p(\theta|\gamma_a)d\theta \\ &= \log\left[\frac{\Gamma(\sum_k\gamma_{ak})}{\prod_k\Gamma(\gamma_{ak})}\right] + \sum_k(\gamma_{ak}-1)\mathbf{E}_{p(\theta|\gamma_a)}[\log\theta_k] \\ &= \log\left[\frac{\Gamma(\sum_k\gamma_{ak})}{\prod_k\Gamma(\gamma_{ak})}\right] + \sum_k(\gamma_{ak}-1)[\Psi(\gamma_{ak}) - \Psi\left(\sum_j\gamma_{aj}\right)],\end{aligned} \tag{43}$$

and similarly for the $p(\theta|\alpha)$ term. Therefore, the relevant terms simplify to the following (again, omitting unnecessary summations and non-γ_{ak} terms)

$$\left(\alpha_k - \gamma_{ak} + \sum_g Q_{kga}\right)\left[\Psi(\gamma_{ak}) - \Psi\left(\sum_j\gamma_{aj}\right)\right] - \log\left[\frac{\Gamma(\sum_k\gamma_{ak})}{\prod_k\Gamma(\gamma_{ak})}\right]. \tag{44}$$

Taking partial derivatives with respect to γ_{ak} and setting these to zero leaves

$$\left(\alpha_k - \gamma_{ak} + \sum_g Q_{kga}\right)\left[\Psi'(\gamma_{ak}) - \Psi'\left(\sum_j\gamma_{aj}\right)\right] = 0, \tag{45}$$

giving the following update for γ_{ak}

$$\gamma_{ak} = \alpha_k + \sum_g Q_{kga}. \tag{46}$$

A.3 Pseudocode Listings

Algorithm 1: Latent Process Decomposition algorithm. We used a tolerance $tol_1 = 10^{-5}$ to avoid Gaussians collapsing onto a single point, and $tol_2 = 10^{-10}$ for the α updating.

```
Data: Read the data e
α ⟵ 1
Initialise μ and σ (see below)
γ ⟵ rand()
while iterations ≤ maxIterations do
    Update: Q as in equation (6)
    Update: γ as in equation (7)
    Update: μ as in equation (8) or (15)
    Update: σ² as in equation (9) or (16)
    if σ² < tol₁ then
        σ² = tol₁
    end
    while |α_new − α_old| > tol₂ do
        α_new ⟵ α_old − H(α_old)⁻¹ g(α_old)
        where g and H are given in equations (33) and (34) respectively
    end
    Determine the log-likelihood (see equation (18))
end
```

Parameter Initialization. From (6), we can see that in order to update Q_{kga}, we need values for μ_{gk}, σ_{gk}, and γ_{ak}. The first two are initialized to the array means and standard deviations, respectively. The γ_{ak} values were initialized to positive random numbers since, with the means and standard deviations uniform over the $\mathcal{K}$ processes, uniform

γ_{ak} values would place the algorithm in a local minima of the likelihood and training could not proceed.

Numerical Stability. There are several points where the algorithm could potentially lose numerical stability. If a Gaussian contracts onto a single point, its standard deviation will shrink to zero and its probability density function will approach ∞. To avoid this, we apply a lower bound to the value of σ^2_{gk}. The choice of the value of this bound is important since, if the value is too low, numerical problems will persist, while if it is too high it would overconstrain the model and prevent it from fitting the data. For the update equation for Q_{kga}, we find a second possible source of instability. From (6), we occasionally find that we reach the limits of machine precision and both the numerator and the denominator approach zero. To circumvent this, we added a small constant (10^{-100}) to each value in $\mathbf{Q}$ before normalization (i.e., before performing the division in (6)). Finally, when calculating the likelihood, we have to compute a product over the g. For a large number of genes, this product soon becomes smaller than the machine precision. However, we are ultimately interested in the logarithm of this product (the log-likelihood) and so we can perform the following numerical trick to avoid a problem. Namely, we observe that the expression we wish to evaluate is of the form $\log \sum_i \prod_j R_{ij}$. This can be rewritten

$$\log\left[\sum_i \exp \xi_i\right] = \log\left[\sum_i \exp\{\xi_i + K\}\exp(-K)\right]$$
$$= \log\left[\sum_i \exp(\xi_i + K)\right] - K,$$

if $\xi_i = \sum_j \log R_{ij}$, which can be readily evaluated given a suitable choice for K. Given that the ξ_i values will all be negative, a suitable choice is $-\max[\xi_i]$ so that we are effectively incrementing all ξ_i values so that the maximum value is 0. Following the implementation of these precautions, the algorithm is robust.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Jeremy Clark, Institute of Cancer Research, London, for comments on the genes mentioned in Section 3.1.1. The work of Rainer Breitling was supported by a BBSRC grant (17/GG17989) and Colin Campbell and Simon Rogers were supported by EPSRC grant GR/R96255 and an EPSRC DTA award.

REFERENCES

[1] A.A. Alizadeh et al., "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," *Nature,* vol. 403, no. 3, pp. 503-511, Feb. 2000.
[2] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.
[3] C.C. Chang et al., "Connective Tissue Growth Factor and Its Role in Lung Adenocarcinoma Invasion and Metastasis," *J. Nat'l Cancer Inst.,* vol. 96, pp. 344-345, 2004.
[4] S.M. Dhanasekaran et al., "Delineation of Prognostic Biomarkers in Prostate Cancer," *Nature,* vol. 412, pp. 822-826, 2001.
[5] R.G. Fahmy et al., "Transcription Factor EGR-1 Supports FGF-Dependent Angiogenesis during Neovascularization and Tumor Growth," *Nature Medicine,* vol. 9, pp. 1026-1032, 2003.
[6] E. Garber et al., "Diversity of Gene Expression in Adenocarcinoma of the Lung," *Proc. Nat'l Academy of Sciences of the USA,* vol. 98, no. 24, pp. 12784-12789, 2001.
[7] A.P. Gasch et al., "Genomic Expression Program in the Response of Yeast Cells to Environmental Changes," *Molecular Biology of the Cell,* vol. 11, pp. 4241-4257, 2000.
[8] L.C. Lazzeroni and A. Owen, "Plaid Models for Expression Data," *Statistica Sinica,* vol. 12, pp. 61-86, 2002.
[9] G.J. McLachlan, R.W. Bean, and D. Peel, "A Mixture Model-Based Approach to the Clustering of Microarray Expression Data," *Bioinformatics,* vol. 18, no. 3, pp. 413-422, 2002.
[10] E. Segal, A. Battle, and D. Koller, "Decomposing Gene Expression into Cellular Processes," *Proc. Eighth Pacific Symp. Biocomputing (PSB),* pp. 89-100, 2003.
[11] P. Spellman et al., "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell,* vol. 9, pp. 3273-3297, 1998.

Simon Rogers received a degree in electrical and electronic engineering from the University of Bristol (2001) during which he was awarded the Sander prize for the best final year examination results, and the PhD degree in engineering mathematics from the University of Bristol (2004). During June 2004, he was a visiting researcher in the Bioinformatics Research Centre at the University of Glasgow where he now holds a postdoctoral research position.

Mark Girolami received a degree in mechanical engineering from the University of Glasgow (1985), and the PhD in computer science from the University of Paisley (1998). He was a development engineer with IBM from 1985 until 1995 when he left to pursue an academic career. From May to December 2000, Dr. Girolami was the TEKES visiting professor at the Laboratory of Computing and Information Science in Helsinki University of Technology. In 1998 and 1999, he was a research fellow at the Laboratory for Advanced Brain Signal Processing in the Brain Science Institute, RIKEN, Wako-Shi, Japan.

Colin Campbell received a degree in physics from Imperial College, London (1981), and the PhD degree in applied mathematics from King's College, London (1984). Subsequently, he held postdoctoral positions at the University of Stockholm and Newcastle University. He joined the Faculty of Engineering at Bristol University in 1990. His main interests are learning theory, algorithm design, and decision theory. Current topics of interest include kernel-based methods, Bayesian generative models, and the application of these techniques to real world problems, mainly in the area of medical decision support, bioinformatics, and onco-informatics.

Rainer Breitling received a degree in biochemistry from the University of Hanover, Germany (1997), and the PhD degree in biochemistry from Technische Universitat Munchen (2001). From June 2001 to April 2003, he worked as a postdoctoral fellow in bioinformatics in the Department of Biology at San Diego State University. Since May 2003, he has been a research assistant at the University of Glasgow, Scotland, where he works on the development of automated interpretation techniques for biomedical microarray data sets.