A Bayesian Approach to the Analysis of Microarray Datasets using Variational Inference

Luke Carrivick and **Colin Campbell**
Dept. of Engineering Mathematics, Queen's Building,
University of Bristol, Bristol BS8 1TR, United Kingdom
{Luke.Carrivick,C.Campbell}@bris.ac.uk

Abstract

Background: We present a variational Bayesian approach to inference in a probabilistic model for microarray gene expression data. The algorithmic approach efficiently maximises the probability of the model given the data and provides an unbiased indication of the most probable number of processes or soft clusters in the data. Compared to hierarchical cluster analysis, the method has a number of practical advantages such as an objective assessment of the number of soft clusters in the data, the ability to handle missing values and the ability to provide a confidence measure for process membership.

Results: We describe the method and its implementation. We compare the method to a previous variational graphical model, proposed by the authors, and argue that model selection is improved. As examples, we apply the algorithm to microarray datasets for breast cancer, prostate cancer and leukemia. The most detailed application is to breast cancer with a comparative study across 7 microarray datasets. We particularly focus on one subtype indicated by the method (the basaloid subtype) where it delineates a common genetic signature across all these datasets and it suggests a therapeutic target.

1 Background

Unsupervised learning methods have been extensively used for finding informative structure in microarray data and they have lead to the discovery of putative subtypes for a variety of cancers [2, 7, 28]. Hierarchical cluster analysis is commonly used for this purpose. However, probabilistic methods are a very attractive alternative. For example, whereas hierarchical cluster analysis is often performed separately on samples and genes, amounting to two distinct reduced space representations of the data, samples and genes can be modelled using a single explanatory space using probabilistic techniques. As we will see, probabilistic techniques can provide an objective measure of the number of clusters present and they can readily handle missing values. As Bayesian approaches favour simpler models, this approach can avoid overfitting due to noise in the data. For many cluster analysis methods there is an implicit mutual exclusion of clusters assumption. For example, for dendrograms, a sample is presumed identified with one sub-tree group. With probabilistic methods it is possible to relax this assumption and allow membership of several clusters simultaneously. Thus, in many biological contexts, it may be unreasonable to assume that samples belong to mutually exclusive clusters. For this reason, probabilistic models which relax this mutual exclusion of classes assumption have been introduced into the bioinformatics literature recently [17, 23]. A further advantage of relaxing the mutual exclusion of classes assumption is that we can derive probabilities of membership of specific clusters. For cancer applications, for example, it would be important to gear future treatments to specific subtypes

and thus a confidence measure for subtype membership carries important diagnostic information. Since cluster has a connotation of exclusive assignment to one group we will use the words *process* or *soft cluster* in this paper.
Given these motivations, there has been significant recent interest in the use of probabilistic methods for clustering microarray data [14, 15, 22]. Typically these methods use a large number of variables and so graphical models are often used to visually indicate dependencies between variables - the probabilistic graphical model in Figure 1 illustrates the method proposed here, for example. For probabilistic approaches involving Bayesian inference a critical task is the computation of the *marginal likelihood* which plays an important role since it enables model selection. Unfortunately, the marginal likelihood is difficult to compute and a range of techniques have been proposed to handle this problem such as annealed importance sampling [24], path sampling [18] and a number of MCMC based methods [10]. In this paper we will use variational methods [4, 5, 20] in which a lower bound on the marginal likelihood is computed using an efficient algorithmic technique. Thus, in a previous comparison of a variational method versus MCMC [12], the variational method appeared both computationally efficient and able to generate a good model with some consistency across microarray datasets. Variational Bayesian methods have been used with microarray data before. For example, a variational Bayesian mixture modelling approach has been considered by Teschendorff et al [30]. However, the model we propose here is more flexible than a mixture model since each sample is identified with a unique mixture over processes whereas, for a mixture model, each sample is identified with one individual mixture component.
The approach we propose is motivated by Latent Dirichlet Allocation [8, 9] and we refer to it as Latent Process Decomposition or LPD since each sample is represented as a combinatorial mixture over a set of latent processes. In Rogers et al [27] we used a variational Expectation Maximisation (EM) approach in which the likelihood is lower bounded using Jensen's inequality. There are two variants to this former approach: one based on a maximum likelihood approach (ML LPD) and the other a maximum a posteriori approach (MAP LPD). In this paper we introduce a variational Bayesian approach to inference, which is as fast and efficient as these earlier methods but which is fully Bayesian enabling determination of the full posterior distribution. An important practical advantage of the proposed method over ML and MAP LPD is an improved strategy for model selection, i.e. the determination of the number of clusters or processes underlying the data. Specifically, the new method has an inbuilt mechanism for model selection and there is no need to do a computationally intensive cross-validation study.
Aside from improved model selection, an important further motivation for the proposed algorithm has been to corroborate earlier discoveries. For example, in section 3.2, we show that the algorithm can isolate a subtype of primary breast carcinoma (the *basal-like* or *basaloid* subtype). Furthermore, it achieves a good alignment across 7 microarray breast cancer studies in the determination of the top-ranked genes distinguishing this subtype from other subtypes. The highlighted genes are biologically significant and in a parallel paper with cancer researchers [11], expression knockdown of an indicated target gene, using short interferring RNAs, has lead to induced loss of viability of more than 50% of tumour cells.
The paper is organised as follows. The method we propose is outlined in the next Section and in Section 3 we will illustrate the use of the method with three application studies to breast cancer, prostate cancer and leukemia. In Appendix A we outline the general variational Bayes's approach to provide further background to the method. Since the approach is mathematically defined in the next Section, a reader with a biological interest may wish to proceed to the Section 3 for examples of data analysis with this method.

2 A Variational Bayes Approch to LPD

With variational methods [32] a bound on probabilities is constructed via the introduction of variational distributions. There are two distinct approaches to variational inference. The first is most similar in motivation to the standard Expectation Maximisation (EM) algorithm and provides an iterative procedure to generate maximum

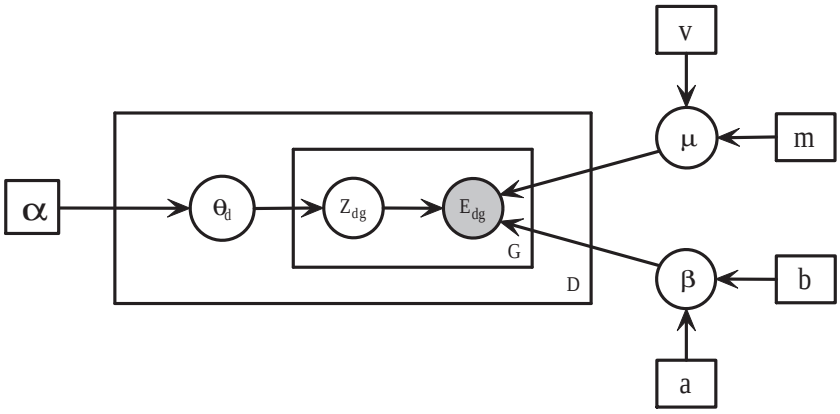


Figure 1: A Graphical Model Representation for the Variational Bayesian Latent Process Decomposition model proposed in this paper. E_{dg} denotes the expression value for gene g in sample d, μ will give the average gene expression for a subgroup and β is an inverse variance (these determe distributions as illustrated in Figure 5) v, m, b and a are hyperparameters on these variables, Z_{dg} is a hidden variable giving the label of gene g in sample d, θ_d gives the mixing over subgroups for sample d and α is a dirichlet parameter: the probability of θ_d given α is given by a Dirichlet distribution defined by this hyper-parameter. All square boxed variables are hyper-parameters for which we estimate a point value. We have not given the indices of variables outside the central plate.

likelihood (ML) or maximum a posterior (MAP) point estimates. The bound is introduced using Jensen's inequality. This bounded likelihood can be maximised by iteratively updating the maximum likelihood solution of the model parameters and the maximum likelihood solution of the variational distribution until convergence. As is the case with the EM algorithm, variational inference will give a point estimate of the posterior distribution.
In keeping with the Bayesian methodology, the second approach to variational inference attempts to *explain away* any latent uncertainty in the model by integrating out all hidden nodes. This is formulated by constructing a bound based on the negative free energy of the system and maximising this bound. Among the first authors to adopt this approach was Attias [3]. One advantage of a variational Bayesian approach over the ML and MAP algorithms is that model comparison can be performed more easily. Specifically there is an inbuilt mechanism for penalising over-complex models. For the ML and MAP methods a computationally wasteful cross validation study is required. This involves setting aside a certain percentage of the data and then estimating the parameters on the remaining data. A model accuracy score is then found from the likelihood of the left-out data.

2.1 Variational Bayes LPD

In Appendix A we outline the general methodology of the variational Bayes approach. The probabilistic graphical model we consider is illustrated in Figure 1. Experimental observations are E_{dg} denoting the expression value for gene g in sample d. μ and β (an inverse variance) are model parameters and they will determine a posterior distribution modelling gene expression values (see Figure 5 for an illustration). The posterior distribution for these

model parameters are governed by hyper-parameters v, m, b and a. There are two hidden variables: Z_{dg} is a hidden variable giving the label of gene g in sample d, while θ_d gives the mixing over subgroups for sample d. α_k is a dirichlet parameter, governing this distribution, with k the process index. If the α_k are uniformly valued there is uniform mixing but a small value relative to the others indicates fewer members in that group. We will use $\mathcal{K}$ to denotes the number of processes, and $\mathcal{G}$ and $\mathcal{D}$ similarly denote number of genes and samples. Given the graphical model in Figure 1, the joint likelihood of the observed data $\boldsymbol{E}$ and the latent variables $\boldsymbol{\theta}$, $\boldsymbol{Z}$, is then:

$$p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta}) = \prod_d p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_g p(Z_{dg}|\boldsymbol{\theta}_d) p(E_{dg}|\mu_g, \beta_g, Z_{dg}|\boldsymbol{\theta}_d) \tag{1}$$

where d is the sample index, g the attribute index and k labels the process. Furthermore, we make the following distributional assumptions in keeping with earlier models [27]:

$$\begin{aligned}
p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) &= \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_j \theta_j^{\alpha_j - 1} && (2)\\
p(Z_{dg} = k|\boldsymbol{\theta}_d) &= \theta_{dk} && (3)\\
p(E_{dg}|\mu_g, \beta_g) &= \mathcal{N}(E_{dg}; \mu_g, \beta_g) = \sqrt{\frac{\beta_g}{2\pi}} \exp\left(-0.5\beta_g (E_{dg} - \mu_g)^2\right) && (4)
\end{aligned}$$

where β is the inverse variance. Thus we assume an approximate normal distribution for the microarray data presented later, a Dirichlet distribution for $\boldsymbol{\theta}_d$, while θ_{dk} denotes the probability of a label k for sample d. We can extend Z_{dg} to a k dimensional vector of zeros except for a 1 in the location of the original Z_{dg} (hence Z_{dg} will be represented by $Z_{dg,k}$). The joint likelihood in equation (1) can then be re-expressed as:

$$p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta}) = \prod_d p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_{g,k} \left[p(Z_{dg,k}|\boldsymbol{\theta}_d) p(E_{dg}|\mu_g, \beta_g, Z_{dg,k})\right]^{Z_{dg,k}}$$

and the log joint likelihood is:

$$\begin{aligned}
\log p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta}) &= \textstyle\sum_{d,k} (\alpha_k - 1) \log \theta_{dk} + \sum_d \log\left(\sum_k \alpha_k\right) - \sum_k \log \Gamma(\alpha_k) \\
&\quad + \textstyle\sum_{d,g,k} Z_{dg,k} \left[\log \theta_{dk} - 0.5\beta_{gk}(E_{dg} - \mu_{gk})^2 + 0.5 \log \beta_{gk}\right]
\end{aligned} \tag{5}$$

We endow the model parameters with prior distributions, and give the form of the distributions for the latent variables as

$$\begin{aligned}
p(\boldsymbol{\mu}) &\sim \prod_{gk} \mathcal{N}(\mu_{gk}; m_0, v_0) && (6)\\
p(\boldsymbol{\beta}) &\sim \prod_{gk} \Gamma(\beta_{gk}; a_0, b_0) && (7)\\
p(\boldsymbol{Z}|\boldsymbol{\theta}) &\sim \prod_{dgk} \theta_{d,k}^{Z_{dg,k}} && (8)\\
p(\boldsymbol{\theta}|\boldsymbol{\alpha}) &= \prod_d p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \sim Dirichlet(\boldsymbol{\alpha}) && (9)
\end{aligned}$$

where v_0, like our previous β, is an inverse variance and where:

$$\Gamma(\beta_{gk}; a_0, b_0) = \beta_{gk}^{a_0-1} \exp(-\beta_{gk}/b_0)/(b_0^{a_0}\Gamma(a_0)) \tag{10}$$

We need to take expectations of the log likelihood given in equation (5) with respect to the approximate posterior distributions $q(\boldsymbol{\theta})$, $q(\boldsymbol{\alpha})$, $q(\boldsymbol{\mu})$ and $q(\boldsymbol{\beta})$. The approximate posteriors are assumed to factorise and have the form

$$\begin{aligned}
q(\boldsymbol{\theta}) &= \prod_d q(\boldsymbol{\theta}_d) \sim \prod_d Dirichlet(\tilde{\boldsymbol{\alpha}}_d) && (11)\\
q(\boldsymbol{\mu}) &\sim \prod_{gk} \mathcal{N}(\tilde{m}_{gk}, \tilde{v}_{gk}) && (12)\\
q(\boldsymbol{\beta}) &\sim \prod_{gk} \Gamma(\tilde{a}_{gk}, \tilde{b}_{gk}) && (13)\\
q(\boldsymbol{Z}) &\sim \prod_{dgk} r_{dg,k}^{Z_{dg,k}} && (14)
\end{aligned}$$

Note that $q(\boldsymbol{\alpha})$ does not have a definite form. This is because the Dirichlet distribution does not have a simple conjugate prior. Leaving out the parameter of interest and taking expectations with respect to the posterior distributions $q(\ldots)$ of all the remaining parameters we have equations (15) to (19).

$$\begin{aligned}
\langle \log p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta})\rangle_{\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\alpha}} = \textstyle\sum_{d,g,k} Z_{dg,k} \quad & [\langle \log \theta_{dk}\rangle + 0.5\langle \log \beta_{gk}\rangle \\
& -0.5\langle \beta_{gk}\rangle (E_{dg}^2 - 2E_{dg}\langle \mu_{gk}\rangle + \langle \mu_{gk}^2\rangle)\ ] \qquad (15)
\end{aligned}$$

$$\begin{aligned}
\langle \log p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta})\rangle_{\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\alpha}} &= \textstyle\sum_{d,k} (\langle \alpha_k\rangle - 1) \log \theta_{dk} + \sum_{d,g,k} \langle Z_{dg,k}\rangle \log \theta_{dk} \\
&= \textstyle\sum_{d,k} (\langle \alpha_k\rangle - 1 + \sum_g \langle Z_{dg,k}\rangle) \log \theta_{dk} \qquad (16)
\end{aligned}$$

$$\langle \log p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\mu})\rangle_{\boldsymbol{\theta},\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\alpha}} = -0.5 \sum_{d,g,k} \langle Z_{dg,k}\rangle \langle \beta_{gk}\rangle (\mu_{gk}^2 - 2E_{dg}\mu_{gk}) \tag{17}$$

$$\begin{aligned}
\langle \log p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\beta})\rangle_{\boldsymbol{\theta},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\alpha}} = \textstyle\sum_{d,g,k} \langle Z_{dg,k}\rangle \quad & \Big[-0.5\beta_{gk}(\langle \mu_{gk}^2\rangle - 2E_{dg}\langle \mu_{gk}\rangle + E_{dg}^2) \\
& +0.5 \log \beta_{gk}] \qquad (18)
\end{aligned}$$

$$\langle \log p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\alpha})\rangle_{\boldsymbol{\theta},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\beta}} = \sum_{d,k} (\alpha_k - 1)\langle \log \theta_{dk}\rangle \tag{19}$$

With the exception of α, these expectations take the form of simply the mean $E(X))$, second moments $E(X^2)$ or $E(\log(X))$ with expectation over a well formed posterior, and can be evaluated analytically in a standard way. Below we give their values with no working (see [26] for more details):

$$\langle \log \theta_{dk} \rangle_{q(\theta)} = \Psi(\tilde{\alpha}_{dk}) - \Psi(\sum_{k'} \tilde{\alpha}_{dk'}) = \log \tilde{\theta}_{dk} \tag{20}$$

$$\langle Z_{dg,k} \rangle_{q(Z)} = r_{dg,k} \tag{21}$$

$$\langle \beta_{gk} \rangle_{q(\beta)} = \tilde{a}_{gk} \tilde{b}_{gk} \tag{22}$$

$$\langle \log \beta_{gk} \rangle_{q(\beta)} = \Psi(\tilde{a}_{gk}) + \log \tilde{b}_{gk} \tag{23}$$

$$\langle \mu_{gk}^2 \rangle_{q(\mu)} = \tilde{m}_{gk}^2 + 1/\tilde{v}_{gk} \tag{24}$$

$$\langle \mu_{gk} \rangle_{q(\mu)} = \tilde{m}_{gk} \tag{25}$$

For the latent variable $\boldsymbol{Z}$, combining equations(15) and (A-6) we have

$$q(\boldsymbol{Z}) = \prod_{dgk} r_{d,k}^{Z_{dg,k}} = Mult(\boldsymbol{Z}; r_{dg,k}) \tag{26}$$

$$r_{dg,k} \propto \tilde{\theta}_{d,k} \exp\left[-0.5 \tilde{a}_{gk} \tilde{b}_{gk} (E_{dg}^2 - 2 E_{dg} \tilde{m}_{gk} + \tilde{m}_{gk}^2 + 1/\tilde{v}_{gk}) + 0.5(\Psi(\tilde{a}_{gk}) + \log \tilde{b}_{gk})\right] \tag{27}$$

For the latent variable $\boldsymbol{\theta}$, combining equations (16) and (A-5) we have

$$q(\boldsymbol{\theta}) \propto Dirichlet\left(\theta \,\middle|\, \langle \alpha_k \rangle + \sum_g r_{dg,k}\right) \propto \prod_{dk} \theta_{d,k}^{\langle \alpha_k \rangle + \sum_g r_{dg,k}} = \prod_{dk} \theta_{d,c}^{\tilde{\alpha}_{dk}}$$

For model parameters, from equation (A-7) $q(\boldsymbol{\Theta}) \propto \exp(\langle p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta}) \rangle) p(\boldsymbol{\Theta})$. Thus for the posterior distribution of the means $q(\boldsymbol{\mu})$:

$$q(\boldsymbol{\mu}) \propto \prod_{gk} \mathcal{N}\left(\mu_{gk}; \frac{\sum_d r_{dg,k} E_{dg}}{\sum_d r_{dg,k}}, \tilde{a}_{gk} \tilde{b}_{gk} \sum_d r_{dg,k}\right) \times \mathcal{N}(\mu_{gk}; m_0, v_0)$$

For a product of Gaussians the inverse variances are additive, hence:

$$q(\boldsymbol{\mu}) \propto \prod_{gk} \mathcal{N}(\mu_{gk}; \tilde{m}_{gk}, \tilde{v}_{gk})$$

where

$$\tilde{v}_{gk} = v_0 + \tilde{a}_{gk} \tilde{b}_{gk} \sum_d r_{dg,k}$$

$$\tilde{m}_{gk} = \frac{1}{v_{gk}} \left[v_0 m_0 + \tilde{a}_{gk} \tilde{b}_{gk} \sum_d r_{dg,k} E_{dg} \right]$$

For the posterior distribution of the Dirichlet parameter $q(\boldsymbol{\beta})$

$$p(\boldsymbol{\beta}) \propto \prod_{gk} \beta_{gk}^{a_0 - 1} \exp\left(-\frac{\beta_{gk}}{b_0}\right)$$

$$q(\boldsymbol{\beta}) \propto \prod_{gk} \Gamma\left(\boldsymbol{\beta}; 0.5\sum_d r_{gd,k}, \left[0.5\sum_d r_{dg,k}((E_{dg}-\tilde{m}_{gk})^2 + 1/\tilde{v}_{gk})\right]^{-1}\right) \times \Gamma(\boldsymbol{\beta}; a_0, b_0)$$

$$q(\boldsymbol{\beta}) \sim \prod_{gk} \Gamma(\beta_{gk}; \tilde{a}_{gk}, \tilde{b}_{gk})$$

where

$$\tilde{a}_{gk} = a_0 + 0.5\sum_d r_{dg,k}$$

$$\frac{1}{\tilde{b}_{gk}} = \frac{1}{b_0} + 0.5\sum_d r_{dg,k}\left((E_{dg}-\tilde{m}_{gk})^2 + 1/\tilde{v}_{gk}\right)$$

The iterative equations of interest for the latent variable parameters are therefore

$$\tilde{\alpha}_{dk} = \langle\alpha_k\rangle + \sum_g r_{dg,k} \tag{28}$$

$$r_{dg,k} \propto \tilde{\theta}_{d,k} \exp\left[-0.5\tilde{a}_{gk}\tilde{b}_{gk}(E_{dg}^2 - 2E_{dg}\tilde{m}_{gk} + \tilde{m}_{gk}^2 + 1/\tilde{v}_{gk}) + 0.5(\Psi(\tilde{a}_{gk}) + \log\tilde{b}_{gk})\right] \tag{29}$$

where $\tilde{\theta}_{d,k}$ is found from (20). $r_{dg,k}$ can be interpreted as the probability that gene g in sample d is generated by process k and hence should be normalised over $\mathcal{K}$. For the hyper-parameters:

$$\tilde{v}_{gk} = v_0 + \tilde{a}_{gk}\tilde{b}_{gk}\sum_d r_{dg,k} \tag{30}$$

$$\tilde{m}_{gk} = \frac{1}{\tilde{v}_{gk}}\left[v_0 m_0 + \tilde{a}_{gk}\tilde{b}_{gk}\sum_d r_{dg,k}E_{dg}\right] \tag{31}$$

$$\tilde{a}_{gk} = a_0 + 0.5\sum_d r_{dg,k} \tag{32}$$

$$\frac{1}{\tilde{b}_{gk}} = \frac{1}{b_0} + 0.5\sum_d r_{dg,k}\left((E_{dg}-\tilde{m}_{gk})^2 + 1/\tilde{v}_{gk}\right) \tag{33}$$

These update equations are in-line with expectations. The Dirichlet parameter $\tilde{\alpha}_{dk}$ is made up of a *prior mean* count and a number of observations. Similarly the parameters $\tilde{v}_{gk}$, $\tilde{m}_{gk}$, $\tilde{a}_{gk}$, $\tilde{b}_{gk}$ all decompose into the form $\xi_{new} = \xi_{prior} + \xi_{data}$ for a general parameter ξ.

2.2 Implementation

We can now outline the method in full (demonstration software is available [1]). First we need to fix $\langle\alpha_k\rangle$ in (28) and v_0, m_0, a_0 and b_0 in (30-33). For $\langle\alpha_k\rangle = \alpha_k/\sum_j \alpha_j$, we can assume the same prior on all the processes. Distributed on a simplex we therefore assume $\alpha_k = 1$, for all k and hence $\langle\alpha_k\rangle = 1/\mathcal{K}$, where $\mathcal{K}$ is the number of processes. Thus, in the numerical experiments below, the α_k are fixed throughout, though we can consider a

more general class of models where there are also hyperparameters on α_k. In the experiments below we used a linear translation of the data to give zero mean and unit variance and thus priors with hyperparameters $m_0 = 0.0$ and $v_0 = 1.0$ are suitable choices. For the priors on β (the inverse variance or precision, $\beta = 1/\sigma^2$), we note from equation (7) that the standard conjugate prior is a Gamma distribution. Since the mean of a Gamma distribution ($\Gamma(\beta; a, b)$ as defined in (10)) is ab and the variance is ab^2 we used $a_0 = 20.0$, $b_0 = 0.05$ giving a mean of 1 and a variance of 0.05. This gives a fairly peaked distribution near 1. We experimented with alternative choices for these hyperparameters but found the results were quite robust. From equations (32,33) we see that the choice of a_0 and b_0 gives a lower bound of $\tilde{a}_{gk}$ and $\tilde{b}_{gk}$ and hence they implicitly act as smoothing terms. For example, by avoiding $\tilde{a}_{gk} = 0$ we avoid a singularity in the digamma function in the update (29).

Having defined starting values and supplied the data E_{dg} (denoting the expression value of gene g in sample d), we then iteratively update $\tilde{\alpha}_{dk}$ and $r_{dg,k}$ and the hyperparameters $\{\tilde{v}_{gk}, \tilde{m}_{gk}, \tilde{a}_{gk}, \tilde{b}_{gk}\}$ until convergence. For many microarray experiments there may be missing values or poor quality readings which should be discarded. If the expression value for gene g in sample d is absent then we discard the corresponding d contributions in the summations over d in the update equations (28-33): for example, the $r_{dg,k}E_{dg}$ product term is discarded for the given sample d in the update for $\tilde{m}_{gk}$ in (31).

Having iterated the algorithm until the stopping criterion is satisfied, we derive the resulting model from the final values of latent variable parameters and hyper-parameters. $\tilde{\alpha}_{d,k}$ quantifies the extent to which sample d came from process k. From the final values of the hyper-parameters we can also estimate posterior distributions: an example of a posterior distribution of means is given in Figure 5 for two genes *FOXA1* and *TFF3* and the interpretation of these distributions and relation to the point estimate density estimates of ML and MAP LPD is described in Section 3.2. To determine these posterior distributions we use $\tilde{m}_{gk}$ which gives the mean of the distribution for gene g in process k. From equation (22) we can also estimate the inverse variances in the posterior distribtion from the product of $\tilde{a}_{gk}$ and $\tilde{b}_{gk}$. Using these means and variances we can also derive statistical scores to rank differentially expressed genes. For example, there is significant under-expression in process 4 in Figure 5(a) and hence this gene is distinguished by a high Fisher score in comparison to processes 1-3.

In Appendix A we give a general description of the variational Bayes method. Specifically, the idea behind the method is to maximise the evidence for the model ($p(Data)$) by maximising an expression called the free energy $F(\boldsymbol{\Theta})$, which is a lower bound on the log of the evidence, see (A-3). In Appendix B we give an expression for this lower bound. Since the free energy should increase with each iteration this provides a useful check on correct implementation and a stopping criterion for convergence (when the incremental change in the free energy is below a tolerance).

From equation (A-7) we also note that VB LPD method presented here has an inbuilt model selection mechanism. Specifically, the second term in $F(\boldsymbol{\Theta})$ is a Kullback-Leibler divergence between the approximate posterior and prior over parameters. As more processes are added the KL-divergence term will increase, causing the free energy to fall. The KL-divergence thus penalises complexity: the free energy increases until it passes through a peak as the KL-divergence penalises overcomplexity. Figures 4(a) and 6(a) are two examples.

The original ML LPD and MAP LPD [27] has a similar graphical representation to Figure 1 but without the hyper-parameters on the model variables. These two methods have no inbuilt method for penalising model over-complexity. ML LPD can simply overfit (Figure 2(a) is an example), while for MAP LPD there is a prior to avoid over-complex models. Either way, a cross-validation study using held-out data is necessary to determine the correct model complexity to use. As an approach to model selection, this has disadvantages. Firstly, we are setting aside some data to perform the cross-validation study: for VB LPD no such data is set aside. Secondly, the cross-validation study is computationally wasteful. Thirdly, for MAP LPD there is the question of what type of prior should we use to avoid over-complex models. Finally, there is the potential for bias. If we calculate the probability of the model given the data it will be much lower for a simple model fitting complex data, as opposed

to a complex model fitting simple data. Hence, there is the possibility that the likelihood doesn't start to fall rapidly enough after passing through the correct model complexity to use.

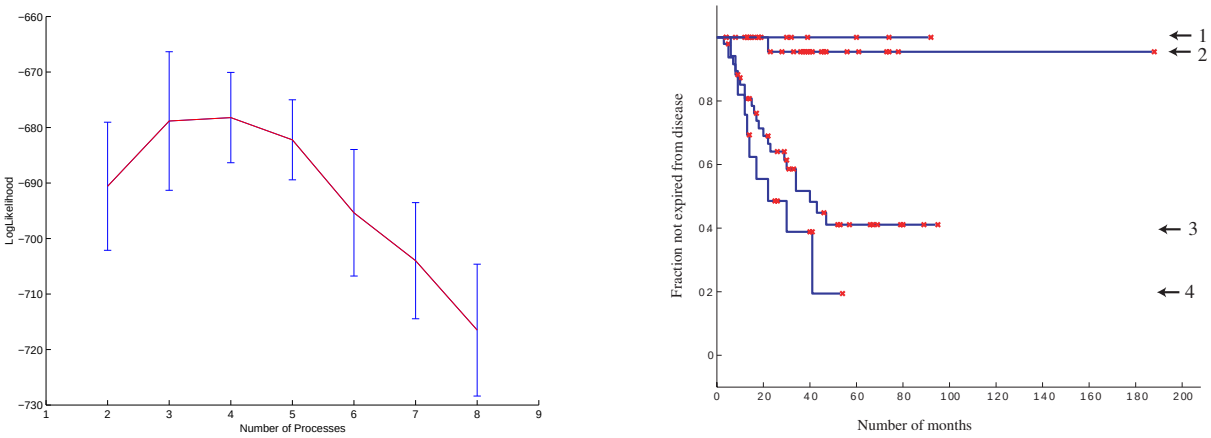
3 Results

3.1 Introduction

In this section we will demonstrate the above method on three applications in cancer informatics, specifically, identification of the genetic signature of the basaloid (basal-like) subtype of primary breast carcinoma, the identification of possible subtypes of prostate cancer and the identification of subtypes of leukemia. We will compare with hierarchical cluster analysis, generally the method of choice in the original data analyses. We will also compare and contrast with the original ML and MAP LPD algorithms.

3.2 Example 1: identifying the genetic signature of the basaloid subtype of primary breast cancer.

In a previous study [12] we investigated possible subtypes of primary breast carcinoma using ML and MAP LPD across 3 microarray datasets for breast cancer (principally of invasive ductal type). To differ from our earlier investigation, we will extend this study to a further 4 microarray datasets and focus on one subtype, demonstrating that the proposed method can successfully delineate the genetic signature of a subtype across a number of microarray studies. This example will also serve to illustrate the advantages of VB LPD over ML or MAP LPD. For our earlier study using ML and MAP LPD, we considered the dataset of Sorlie *et al* [28] consisting of 115 primary breast carcinoma samples (we used the 534 genes selected in their study). The corresponding ML LPD maximum likelihood curve is established using hold-out cross validation and it is given in Figure 2(a). As the number of processes increases, the likelihood curve passes through a peak. Prior to this peak underfitting occurs, whereas after the peak overfitting occurs: the algorithm would construct an over-complex model given the sample size and extent of noise in the data. The MAP solution is very similar except that the likelihood plateaus after 4 processes since further model complexity is not required. Thus the peak indicates that a 4 process model is most appropriate and, with a 4 process decomposition, samples can be identified with particular processes. The patients so identified have very different clinical outcomes: in Figure 2(b) we give the survival curves for patients identified with these 4 processes (for more details see Carrivick *et al* [12]).

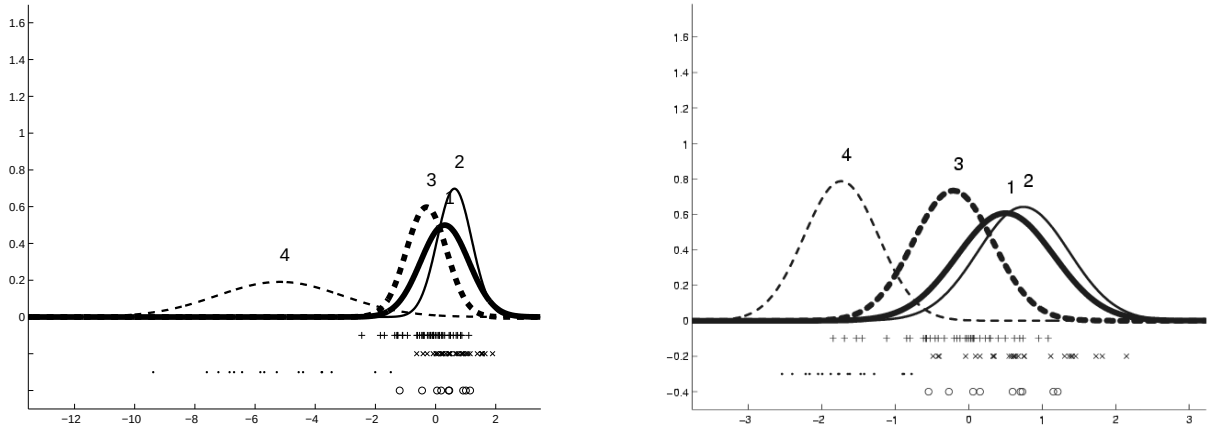


(a) Log-likelihood curve

(b) Kaplan Meier plot

Figure 2: Log-likelihood curve (left) and Kaplan Meier plot (right) using ML LPD for the dataset of Sorlie *et al* [28]. Figure 2(a) gives the log-likelihood (y-axis) versus number of processes (x-axis), while Figure 2(b) gives the the fraction not expired from the disease (y-axis) versus number of months (x-axis) for a 4 process decomposition.

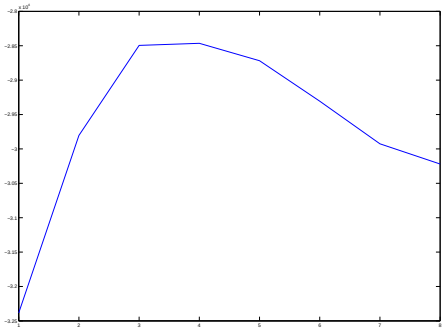
This analysis suggested a minimum one indolent and three aggressive subtypes. The most aggressive process 4 has the most distinctive profile and, with one exception, the patients belonging to this process are identified with the basaloid subtype of breast cancer found by Sorlie *et al* using hierarchical cluster analysis [28]. The density curves for two top-ranked genes distinguishing this process, *TFF3* and *FOXA1*, are given in Figure 3. Expression values associated with samples belonging to particular processes are given below the plot and the given density curves model the distribution of this data.



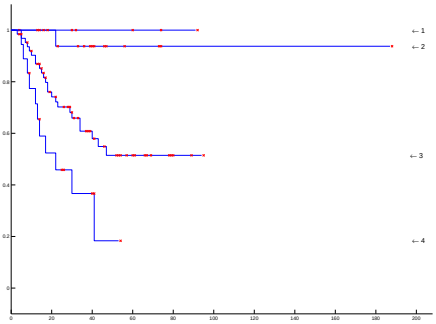
(a) FOXA1 (b) TFF3

Figure 3: Density curves derived using MAP LPD and expression values for two genes, *FOXA1* and *TFF3*, for the Sorlie *et al* [28] dataset. Expression values are below the curves with · indicating expression values for samples belonging to process 4 and the other symbols indicating expression values for samples belonging to other processes. The density curves model extent of data present on a given range. Thus *FOXA1* and *TFF3* underexpress in the most aggressive process 4 identified in Figure 2(b).

In Figure 4(a) we give the free energy curve using the VB LPD method proposed in this paper. The curve peaks at 4, suggesting this is the correct number of processes to use, in agreement with the ML LPD likelihood curve. Later, though, we will present examples where the peaks differ and VB LPD may give a more unbiased estimate for model selection. In Figure 4(b) we give the corresponding Kaplan-Meier plot for VB LPD using a 4 process decomposition. The top ranked genes distinguishing the most aggressive process 4 are given in Table 1 for VB LPD and, as for MAP LPD, *TFF3* and genes expressing forkhead box transcription factors *FOXA1* and *FOXC1* are prominant. The 19 process 4 samples are identified with the 19 basaloid samples of breast cancer described by Sorlie *et al* [28]. In Figure 5(a) we plot the posterior distributions for *FOXA1* and *TFF3*, derived using VB LPD. The proposed variational Bayes algorithm is more informative than ML and MAP LPD. Thus in Figure 3 MAP LPD uses point estimates for the density estimator means and variances whereas VB LPD gives the full posterior distribution: a wide spread in the peaks in Figure 5 would indicate that a range of models fit the data well and the density estimations in Figure 3 would be unreliable.

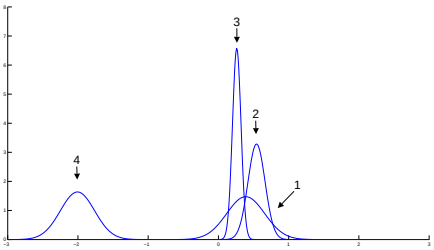


(a) Free Energy plot

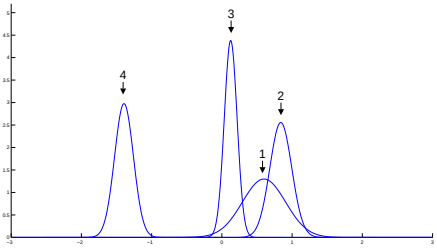


(b) Kaplan Meier plot

Figure 4: Free energy plot (left) and Kaplan Meier plot (right) for the dataset of Sorlie *et al* [28] using the variational Bayes method. These plots may be compared with Figure 2(a) and 2(b). The peak in the free energy is more pronounced than the result in Figure 2(a). The decomposition leads to a similar Kaplan Meier plot to Figure 2(b).



(a) FOXA1

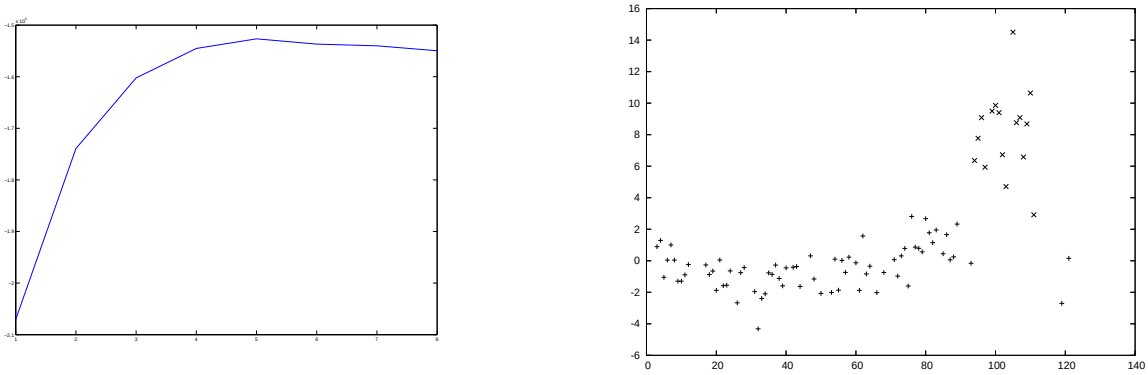


(b) TFF3

Figure 5: The distribution of means for two selected genes: these distributions indicate the reliability of the point estimates of the means found using MAP LPD and given in Figures 3(a) and 3(b) for comparison.

In a similar fashion to our earlier paper, Carrivick *et al* [12], we also used VB LPD with the 49 sample breast cancer dataset of West *et al* [34] and the dataset of van t'veer *et al* [31] with 78 samples. In Table 1 we show the genetic signatures matching the Sorlie *et al* basal signature (for West *et al* we have used time-to-metastasis to match processes, but for van't veer *et al* there was no survival data and the match is by correlated signature). So far, we find that VB LPD confirms the results of Carrivick *et al* [12]. However, to extend the study and to show that the method can successfully identify the genetic signature of a subtype across a large number of datasets, we will consider 4 further microarray datasets for breast cancer. These recent studies all use the Affymetrix Hu133A GeneChip and hence we will use them as a composite dataset of 614 breast carcinoma samples rather than use

them individually (the 4 component datasets are Yang *et al* with 28 samples [36], Farmer *et al* with 49 samples [16], Pawitan *et al* with 251 samples [25] and Wang *et al* with 286 samples [33]). The free energy plot is given in Figure 6(a) and the peak is now at 5. Though Figures 2(a) and 4(a) suggested 4 subtypes this result is in line with expectations: as we increase the sample set size the effects of noise are averaged out, model parameters are better estimated and a more detailed partitioning is achieved. Though this further partitioning affects the other subtypes (principally process 2), the basaloid subtype is distinct and unchanged if we use a 4 process or 5 process model (we will discuss this shortly with the results presented in Table 1). In Table 1 we give the top 20 ranked genes distinguishing the basaloid subtype using a 5-process split for this composite dataset. We observe that the large majority of these genes overlap with the top 20 genes listed for the previous 3 studies for Sorlie *et al*, West *et al* and van t' veer *et al*. In line with our remark that increasing dataset size reduces the effect of noise, we observe that the composite dataset, as the largest dataset, has the greatest alignment with the Sorlie *et al*, West *et al* and van t'veer *et al* signatures, whereas the smallest dataset, West *et al* with 49 samples, has least commonality with the other datasets. Of course, if we had decided to use these 4 datasets individually, rather than as a composite dataset, the observed gene ranking alignment is weaker because of the enhanced effects of noise. In our earlier comment on 4 versus 5 subtypes we mentioned that the genetic signature of this basaloid subtype is very robust. In Table 1 we also give the matching signature if we had used an 8 process split: a process with the same signature is apparent. This specific signature and failure to resolve into components with an increasing number of processes may indicate a single underlying cause for the genesis of this subtype.



(a) Free Energy plot for the composite dataset

(b) Fox-ratios (Sorlie *et al* dataset)

Figure 6: The free energy plot (left) for the composite dataset of 614 samples suggests 5 subtypes for breast cancer. The most aggressive subtype is the basaloid subtype and it is characterised by a high *FOX-ratio* (right). Figure 6(b) gives the ratio of *FOXC1* over *FOXA1* following linear rescaling of the data to zero mean and unit variance. The *FOX-ratios* are so high for the basaloid subtype that we have logged the ratios so they can be easily visualised on the same plot: $+$ give non-basaloid samples and $\times$ the basaloid samples as identified by MAP LPD.

It is therefore possible to claim that the basaloid signature is apparent across 7 microarray studies and it appears quite specific. Among the genes in the basaloid signature, *FOXA1* appears to play a pivotal role. It features in the top three positions for Sorlie *et al*, West *et al* and the composite dataset (it was absent from the van t' veer *et al* dataset). Indeed its importance is apparent from some of the other genes in the list: the X box-binding

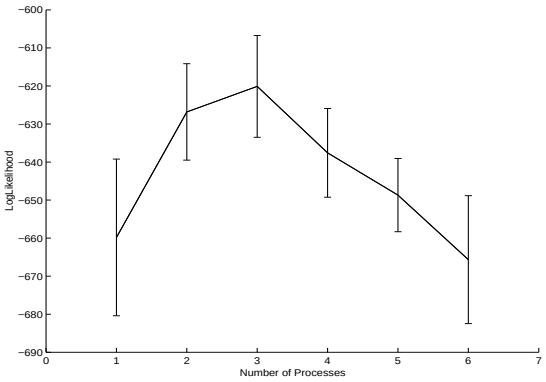
Sorlie et al	West et al	Van t' Veer et al	Composite (5-split)	Composite (8-split)
TFF3	*CRIP1*	*VGLL1*	**FOXA1**	**FOXA1**
XBP1	**XBP1**	**AGR2**	**AGR2**	*MLPH*
FOXA1	**FOXA1**	**TFF3**	**XBP1**	*FLJ20174*
GATA3	*CEBPD*	*ESR1*	*MLPH*	**AGR2**
B3GNT5	*HSPA8*	**CA12**	*FLJ20174*	**CA12**
GALNT10	**GATA3**	**DSC2**	**CA12**	*AK127020*
FBP1	*RARA*	**NAT1**	**GATA3**	**AR**
DSC2	*CRYAB*	*EST*	*AK127020*	**DSC2**
FOXC1	**GATA3**	**CDH3**	**CA12**	**XPB1**
FOXC1	**FBP1**	**FOXC1**	**CA12**	**CA12**
FLT1	*KRT18*	*SCUBE2*	**GATA3**	**GABRP**
FOXC1	*MSN*	**AR**	**AR**	**GATA3**
GATA3	*TCEAL1*	*Corf7*	**TFF3**	**CA12**
SLC11A3	*SCNN1A*	*SLC7A2*	*ABAT*	**GATA3**
SLC11A3	*NSEP1*	**GABRP**	**FBP1**	**TFF3**
MGC27171	**CDH3**	*EST*	**DSC2**	*ANP32E*
NAT1	*BF*	**XPB1**	**GATA3**	*ELF5*
MRPS14	**TFF3**	*BCMP11*	**CA12**	*ABAT*
LOC51313	*Hu. clone 23948*	*VAV3*	*TFF1*	**GATA3**
MGC10710	*FSCN1*	*EST*	**GABRP**	**CA12**

Table 1: The top-ranked genes distinguishing the basaloid subtype of breast cancer. Genes given in bold are common in the top 20 genes across more than one study. The composite dataset of 614 samples is taken as one study (see text). The composite dataset is derived from 4 datasets [16, 25, 33, 36], using the Affymetrix U133A chip. If these datasets are used individually poorer gene rank alignment is achieved due to the enhanced effects of noise. The genetic signature of the basaloid subtype is very robust: the peak in the free energy suggest 5 subtypes. However, if we choose 8 subtypes instead (right hand column) the same signature is retrieved with one process (see text). Note: (a) multiple entries for a gene in a column (e.g. *FOXC1* and *GATA3* under Sorlie et al) come from different probes for the same gene, (b) absense of a gene in a column can stem from the fact it is absent from the dataset e.g. *FOXC1* is listed under Sorlie *et al* and van t' veer *et al* but was absent from West *et al* and the composite dataset, (c) For the composite dataset (5-split) the genes ranked below position 20 are, in order: *DACH1*, *ESR1*, *ANP32E*, *MCCC2*, *KRT18*, *ABAT*, *GALNT6*, *INPP4B*, *SCUBE2*, *NAT1*, (d) *EST* is an expressed sequence tag.

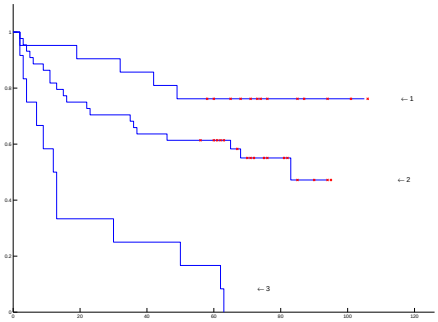
protein 1, *XBP1*, is believed to be regulated by *FOXA1* [13] as is the trefoil factor *TFF1* [6], a close relative of *TFF3*. The biological importance of *FOXA1* is also apparent from some recent results reported in the literature: a substantial number of estrogen response elements (EREs) have associated binding sites for *FOXA1* [13, 21]. In our earlier study [12] we found tumour samples from *BRCA1* mutation carriers were exclusively associated with the basaloid subtype and FOXA1 and BRCA1 proteins coregulate cell cycle inhibition [35]. *FOXA1* is a member of the forkhead box family of transcription factors, as is a second highly ranked gene: the developmental gene *FOXC1*. The latter regulates *DACH1* and the transforming growth factor TGFβ [29, 38]. *FOXA1* underexpresses in the basaloid subtype whereas *FOXC1* overexpresses. Indeed, if we evaluate the Pearson correlation coefficient for all possible gene pairings in the Sorlie *et al* dataset, the pairing *FOXA1* and *FOXC1* has the highest anticorrelation (for this reason the ratio of *FOXC1* over *FOXA1* appears a useful marker of the basaloid subtype, see Figure 6(b)). In a parallel paper [11] we report results on knockdown of expression by *FOXC1* using small inferring RNAs (siRNA) for a breast cancer cell line (BT549) which has a similar high *FOXC1:FOXA1* ratio as discussed here. We report loss of viability of more than 50% of cancer cells within 72 hours as a result. Thus the proposed method appears to have correctly highlighted a significant target.

3.3 Example 2: identifying subtypes of prostate cancer.

For our second example we consider the prostate cancer dataset of Glinsky *et al* [19]. These authors used a selected set of recurrence predictor genes and a training set of 21 tumours to evaluate prediction of recurrence versus non-recurrence on a further set of 79 tumours. In our case we will use all these samples, starting from the original dataset (12,625 probes), to determine any subtypes of prostate cancer. Specifically we are interested in determining if there are subtypes with differing frequency of recurrence or non-recurrence. This is a very important problem since currently there are no reliable methods for separating indolent from aggressive prostate cancers with the recognition that most prostate cancer patients have indolent subtypes which are overtreated. Using ML LPD we give the likelihood curve in Figure 7(a) and the corresponding clinical outcomes in Figure 7(b). The likelihood curve suggests 3 processes with quite different clinical outcomes: two subtypes appear to be heading toward a plateau with no disease recurrence, while the third subtype appears very aggressive with inevitable recurrence.



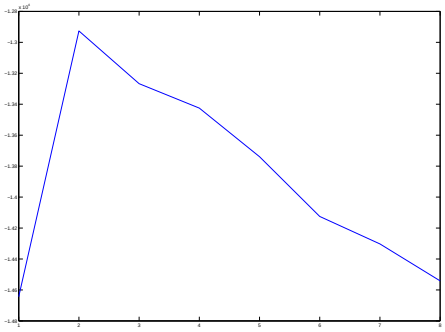
(a) Log-likelihood curve



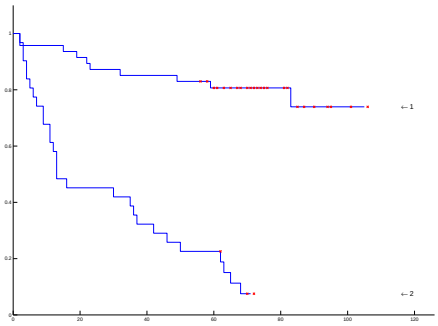
(b) Fraction without recurrence (y-axis) versus number of months (x-axis), $\star$ represents a patient remining in the survey without recurrence

Figure 7: For ML LPD the log-likelihood curve (left) suggests 3 subtypes. The clinical outcomes are distinct with one processes having disease recurrence for all patients within 60 months, whereas the other two processes appear to head toward a plateau with no recurrence. There are 21, 44 and 12 patients in processes 1, 2 and 3 respectively.

In Figure 8(a) we show the free energy curve for the dataset of Glinsky *et al* for VB LPD, with the corresponding clinical outcomes depicted in Figure 8(b).



(a) Free energy plot



(b) Recurrence curves

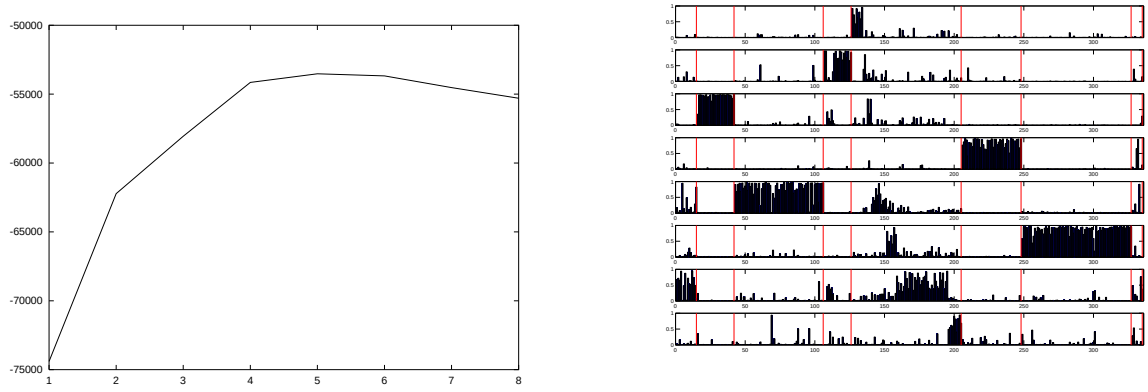
Figure 8: Free energy curve and disease recurrence plot for the prostate cancer dataset of Glinsky *et al* [19]. As for Figure 7(b) a drop in Figure 8(b) indicates disease recurrence and a star indicates the patient remains in the survey without recurrence. There are 47 patients in process 1 and 31 in process 2.

Interestingly, the peak is now at 2 with a clear split into two groupings, one with a very high probability of

recurrence and a second grouping with a milder form of the disease. Earlier we remarked that the ML and MAP LPD model could over-estimate the number of processes and this may be an instance where this occurs.

3.4 Example 3: identifying subtypes of leukemia.

As a last example, we applied the variational Bayes method to an oligonucleotide microarray dataset from 360 patients with acute lymphoblastic leukemia (ALL) [37]. ALL is known to have a number of subtypes with variable responses to chemotherapy. In many cases fusion genes are implicated in the genesis of the disease. For the Yeoh *et al* [37] dataset samples were drawn from leukemias with rearrangements involving *BCR-ABL*, *E2A-PBX1*, *TEL-AML1*, rearrangements of MLL gene, hyperdiploid karyotope (more than 50 chromosomes) and T lineage leukemias (*T-ALL*). The free energy is plotted in Figure 9(a) with a peak suggesting 5 subtypes.



(a) Free energy plot

(b) Decomposition diagram

Figure 9: The free energy curve (left) for the Leukemia dataset of Yeoh *et al* gives a peak at 5 processes. For the decomposition diagram (right) samples 1-15 are *BCR-ABL*, 16-42 are *E2A-PBX1*, 43-106 *Hyperdiploid*> 50, 107-126 *MLL*, 206-248 *T-ALL*, 249-327 *TEL-AML1*, 328-335 *Group23* and 127-205 are labelled as *Others*. *E2A-PBX1*, *T-ALL*, *TEL-AML1* and the hyperdiploid samples are very distinct groupings. The lack of distinction with the other groups probably explains the free energy peak at 5 rather than a higher level of partitioning.

In section 2.2 we mentioned that $\tilde{\alpha}_{d,k}$ quantifies the probability that sample d is generated by process k, and in Figure 9(b) we give a decompostion diagram using $\tilde{\alpha}_{d,k}$ (the peaks give the probability that sample d is in process k). We see that the subtypes for *E2A-PBX1*, *T-ALL*, *TEL-AML1* and the hyperdiploid samples are very distinct groupings. *BCR-ABL* and *MLL* are less distinct groupings which may explain why the peak is at 5 and not higher (with the decomposition diagram we allowed for 8 processes, with more processes the remainder are left empty). For the middle group (which the original authors marked as *Others*) we find some peaks suggesting a connection with known groupings with some evidence for two new groupings. A dendrogram was presented by Yeoh *et al* (their Figure 1), however, it only uses the top 40 genes most highly correlated with the 7 proposed class distinctions, with these genes being selected by a chi-squared statistic. This effectively creates a supervised learning problem. With no such use of class label information the corresponding dendrogram has a more difficult

interpretation (see Supplementary Information [1]).

4 Discussion

In this paper we have used the established variational Bayes approach to develop a method applicable to Latent Process Decomposition [27]. The method has advantages over hierarchical cluster analysis, such as a common explanatory space for samples and genes and the ability to handle missing values, for example. Compared to our earlier ML and MAP LPD methods [27], the method proposed here has advantages such as improved model selection and the fact that we obtain a full distribution over model parameters rather than point estimates of the density estimation (see captions to Figures 3 and 5, for example). More generally, as a Bayesian method, any assumptions or prior beliefs about the data are explicit. This is an important advantage over other data analysis approaches where implicit assumptions could be wholly inappropriate, hence degrading performance. As an example, hierarchical cluster analysis uses an implicit mutual exclusion of classes assumption (a sample is presumed identified with a unique grouping in a dendrogram), whereas no such assumption is made with LPD (a sample can be represented as a combinatorial mixture over processes). The use of inappropriate implicit assumptions could explain reported discrepancies between different data analysis methods and different studies whereas we can find reasonable agreement across datasets (e.g. for the basaloid example in Section 3.2).
Though these advantages justify the proposed approach, probabilistic methods have possible future advantages which may further strengthen this approach. Thus, microarray technology is intrinsically noisy, disrupting rank scoring of genes between different studies. Apart from increasing the size of datasets, another way to reduce the effects of noise is to incorporate more information from disparate types of data. Approaches such as hierarchical cluster analysis are ill-suited for incorporating other types of information, such as sequence data or pathway information. On the other hand, different types of data can usually be encoded into probabilistic constructs and consequently these types of techniques open an interesting avenue to information integration in the future. In general, medical data can be expected to be noisy and inexact thus justifying probabilistic approaches. For hierarchical cluster analysis, for example, samples are identified with groupings which reflect underlying subtypes, but no probability of membership of a subtype is indicated. On the other hand, probabilistic methods can give a confidence measure for class membership (e.g. in Figure 9(b) the peaks indicate degree of confidence in the class assignment). For these and other reasons, they will have important advantages in the future interpretation of medical data.

Appendix A: The Variational Bayes Method

In this Appendix we briefly summarise the general methodology for Variational Bayesian (VB) inference. VB seeks to find a lower bound on the evidence $p(Data)$, in a tractable form to be maximised. Approximations are made to the posterior distributions of all hidden and model variables so that they can be marginalised (integrated out). At each iteration of VB it is the hyperparameters, rather than parameters, that are updated. Thus, compared to the ML or MAP LPD we presented previously [27], the emphasis is shifted a step upwards. The probabilistic graphical model is presented in Figure 1. From this Figure there are two hidden nodes $\boldsymbol{Z}$ and $\boldsymbol{\theta}$ and a set of parameters $\boldsymbol{\Theta}$. In the applications we consider microarray expression data and hence $\boldsymbol{E}$ is used to denote the data. The evidence of some data $p(\boldsymbol{E})$ can be written as a ratio of the joint distribution (with respect to some variables) $p(\boldsymbol{E}, \boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z})$ and the posterior distribution of these variables given the data $p(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E})$.

$$p(\boldsymbol{E}) = \frac{p(\boldsymbol{E}, \boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z})}{p(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E})} \tag{A-1}$$

The log of this is written as:

$$\log p(\boldsymbol{E}) = \log p(\boldsymbol{E}, \boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}) - \log p(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E}) \tag{A-2}$$

Let us introduce an approximation to the posterior distributions of all model and hidden variables $q(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E})$. If we take expectations of expression (A-2) with respect to this approximate posterior $q(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E})$, the left hand side remains unchanged as this is independent of $\boldsymbol{\Theta}$, $\boldsymbol{\theta}$ and $\boldsymbol{Z}$.

$$\log p(\boldsymbol{E}) = \int q(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E}) \log p(\boldsymbol{E}, \boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}) d\boldsymbol{\Theta} d\boldsymbol{\theta} d\boldsymbol{Z} - \int q(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E}) \log p(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E}) d\boldsymbol{\Theta} d\boldsymbol{\theta} d\boldsymbol{Z}$$

Multiplying $p(\boldsymbol{E}, \boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z})$ top and bottom by $q(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E})$ and separating the terms we can now write

$$\log p(\boldsymbol{E}) = F(\boldsymbol{\Theta}) + KL(q(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E})||p(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E}))$$

where

$$F(\boldsymbol{\Theta}) = \int q(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E}) \log \frac{p(\boldsymbol{E}, \boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z})}{q(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E})} d\boldsymbol{\Theta} d\boldsymbol{\theta} d\boldsymbol{Z}$$

As the KL divergence is strictly greater than zero, we can now say that

$$\log p(\boldsymbol{E}) \geq F(\boldsymbol{\Theta}) \tag{A-3}$$

Equality holds when $KL = 0$, i.e. the approximate posterior q and true posterior p coincide. This is the case when our approximation becomes exact. The idea behind a variational Bayes approach is to maximise the evidence by maximising $F(\boldsymbol{\Theta})$. We shall now make an important assumption about the posterior. We assume that it factorises into separate terms, such that $q(\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{E}) = q(\boldsymbol{\Theta})q(\boldsymbol{\theta})q(\boldsymbol{Z})$ where the dependence on $\boldsymbol{E}$ is implied. By writing $p(\boldsymbol{E}, \boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{Z}) = p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})$ we can now expand $F(\boldsymbol{\Theta})$ as

$$F(\boldsymbol{\Theta}) = \int q(\boldsymbol{\Theta})q(\boldsymbol{\theta})q(\boldsymbol{Z}) \log \frac{p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{q(\boldsymbol{\Theta})q(\boldsymbol{\theta})q(\boldsymbol{Z})} d\boldsymbol{\Theta} d\boldsymbol{\theta} d\boldsymbol{Z}$$

Thus by expanding and integrating out $q(\boldsymbol{\theta})$ and $q(\boldsymbol{Z})$

$$F(\boldsymbol{\Theta}) = \int q(\boldsymbol{\Theta})q(\boldsymbol{\theta})q(\boldsymbol{Z}) \log \frac{p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta})}{q(\boldsymbol{\theta})q(\boldsymbol{Z})} d\boldsymbol{\Theta} d\boldsymbol{\theta} d\boldsymbol{Z} - KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta})) \tag{A-4}$$

In equation (A-4) the first term is an averaged likelihood and the second term $-KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}))$ is a measure of the *distance* between approximate posterior and prior over parameters, since this term increases with the number of parameters it can be seen as a penalising term for over complex models. Indeed it has been shown that in certain situations this reduces to the Bayesian information criteria (BIC) and the Minimum Description Length (MDL) (see [3] for further details).
To maximise $F(\boldsymbol{\Theta})$ in equation (A-4) we take zeroed gradients (functional derivatives in this case) with respect to the approximate posteriors $q(\boldsymbol{\Theta})$, $q(\boldsymbol{\theta})$ and $q(\boldsymbol{Z})$.

$$\frac{\delta F(\boldsymbol{\Theta})}{\delta q(\boldsymbol{\theta})} = \int q(\boldsymbol{\Theta})q(\boldsymbol{Z}) \log \frac{p(\boldsymbol{E}, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{\Theta})}{q(\boldsymbol{\theta})q(\boldsymbol{Z})} d\boldsymbol{\Theta} d\boldsymbol{Z} - \int \frac{q(\boldsymbol{\Theta})q(\boldsymbol{\theta})q(\boldsymbol{Z})}{q(\boldsymbol{\theta})} d\boldsymbol{\Theta} d\boldsymbol{Z} = 0$$

$$\int q(\boldsymbol{\Theta})q(\boldsymbol{Z})\log p(\boldsymbol{E},\boldsymbol{\theta},\boldsymbol{Z}|\boldsymbol{\Theta})d\boldsymbol{\Theta}d\boldsymbol{Z} - 1 - \log q(\boldsymbol{\theta})\int q(\boldsymbol{\Theta})q(\boldsymbol{Z})d\boldsymbol{\Theta}d\boldsymbol{Z} - \int q(\boldsymbol{\Theta})q(\boldsymbol{Z})\log q(\boldsymbol{Z})d\boldsymbol{\Theta}d\boldsymbol{Z} = 0$$

As the densities $q(\dots)$ integrate to one we can write

$$q(\boldsymbol{\theta}) \propto \exp\left[\int q(\boldsymbol{\Theta})q(\boldsymbol{Z})\log p(\boldsymbol{E},\boldsymbol{\theta},\boldsymbol{Z}|\boldsymbol{\Theta})d\boldsymbol{\Theta}d\boldsymbol{Z}\right] \tag{A-5}$$

Analogously

$$q(\boldsymbol{Z}) \propto \exp\left[\int q(\boldsymbol{\Theta})q(\boldsymbol{\theta})\log p(\boldsymbol{E},\boldsymbol{\theta},\boldsymbol{Z}|\boldsymbol{\Theta})d\boldsymbol{\Theta}d\boldsymbol{\theta}\right] \tag{A-6}$$

For any of the model parameters,

$$\frac{\delta F(\boldsymbol{\Theta})}{\delta q(\boldsymbol{\Theta})} = \int q(\boldsymbol{\theta})q(\boldsymbol{Z})\log\frac{p(\boldsymbol{E},\boldsymbol{\theta},\boldsymbol{Z}|\boldsymbol{\Theta})}{q(\boldsymbol{\theta})q(\boldsymbol{Z})}d\boldsymbol{\theta}d\boldsymbol{Z} - \log\frac{p(\boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} - 1 = 0$$

so

$$q(\boldsymbol{\Theta}) \propto \exp\left[\int q(\boldsymbol{\theta})q(\boldsymbol{Z})\log p(\boldsymbol{E},\boldsymbol{\theta},\boldsymbol{Z}|\boldsymbol{\Theta})d\boldsymbol{\theta}d\boldsymbol{Z}\right] p(\boldsymbol{\Theta}) \tag{A-7}$$

Equations (A-5) to (A-7) give the approximate posterior distributions for the latent variables and model parameters. They can be interpreted as the posterior taking the form of the exponential of the averaged log likelihood over all remaining variables. Thus all uncertainty is integrated away. The posterior forms of $q(\boldsymbol{\Theta})$, $q(\boldsymbol{\theta})$ and $q(\boldsymbol{Z})$ are determined directly from the optimisation via equations (A-5) to (A-7). In the case of model parameters, the prior distributions in (A-7) are chosen as conjugate to the derived exponentials so that the parametric form for $q(\boldsymbol{\Theta})$ remains the same. Having derived the general form of the posterior distributions we apply it to the LPD model represented in Figure 1 and this gives the approximate posteriors given in Section 2.1.

Appendix B: Evaluation of the Lower Bound

It is useful to be able to evaluate the free energy term $F(\boldsymbol{\Theta})$ given in equation (A-4). Firstly this acts as a test of correct implementation as is should increase with each iteration of the algorithm until convergence. Secondly it can be used as a comparative measure to determine the optimal number of components in a mixture distribution.

$$\begin{aligned} F(\boldsymbol{\Theta}) &= \textstyle\int q(\boldsymbol{\Theta})q(\boldsymbol{\theta})q(\boldsymbol{Z})\log\frac{p(\boldsymbol{E},\boldsymbol{\theta},\boldsymbol{Z}|\boldsymbol{\Theta})}{q(\boldsymbol{\theta})q(\boldsymbol{Z})}d\boldsymbol{\Theta}d\boldsymbol{\theta}d\boldsymbol{Z} - KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta})) \\ &= \langle\log p(\boldsymbol{E},\boldsymbol{\theta},\boldsymbol{Z}|\boldsymbol{\Theta})\rangle_{\boldsymbol{\theta},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\alpha}} - \langle\log(q(\boldsymbol{\theta}))\rangle_{\boldsymbol{\theta}} - \langle\log(q(\boldsymbol{Z}))\rangle_{\boldsymbol{Z}} \\ &\quad - \textstyle\int q(\boldsymbol{\Theta})\log\left[\frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta})}\right]d\boldsymbol{\Theta} \end{aligned} \tag{A-8}$$

Evaluating the elements of the bound given in equation (A-8):

$$\begin{aligned} \langle\log p(\boldsymbol{E},\boldsymbol{\theta},\boldsymbol{Z},|\Theta)\rangle_{\boldsymbol{\theta},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\alpha}} &= \textstyle\sum_{d,k}(\langle\alpha_k\rangle - 1)\langle\log\theta_{dk}\rangle \\ &\quad +\langle\log\Gamma(\textstyle\sum_k\alpha_k)\rangle \\ &\quad -\textstyle\sum_k\log\Gamma(\alpha_k) \\ &\quad +\textstyle\sum_{d,g,k} r_{dg,k}\left[\langle\log\theta_{dk}\rangle\right. \\ &\quad -0.5a_{gk}b_{gk}(E_{dg}^2 - 2E_{dg}m_{gk} + m_{gk}^2 + 1/v_{gk}) \\ &\quad \left.+0.5(\Psi(a_{gk}) + \log b_{gk})\right] \end{aligned} \tag{A-9}$$

$$\langle \log(q(\boldsymbol{\theta})) \rangle_{\boldsymbol{\theta}} = \sum_{dk} (\tilde{\alpha}_{dk} - 1) \langle \log \theta_{dk} \rangle$$

$$\langle \log(q(\boldsymbol{Z})) \rangle_{\boldsymbol{Z}} = \sum_{dgk} r_{dg,k} \log r_{dg,k}$$

The $KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}))$ term decomposes into three terms for the parameter set $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\alpha}\}$, these can be analytically evaluated making use of the same identities that were needed in evaluating the expectations earlier. Here, we shall quote the standard results for KL divergences as given in [26].
For the parameter $\boldsymbol{\mu}$, $p(\boldsymbol{\mu}) \sim \prod_{gk} \mathcal{N}(\mu_{gk}; m_0, v_0)$ and $q(\boldsymbol{\mu}) \sim \prod_{gk} \mathcal{N}(\tilde{m}_{gk}, \tilde{v}_{gk})$.

$$\begin{aligned} KL(q(\boldsymbol{\mu})||p(\boldsymbol{\mu})) &= \textstyle\sum_{gk} 0.5 \log \frac{v_{gk}}{v_0} + 0.5 v_0 \left[m_{gk}^2 + m_0^2 + 1/v_{gk} - 2 m_{gk} m_0 \right] - 0.5 \\ &= \textstyle\sum_{gk} \left[0.5 \log \frac{v_{gk}}{v_0} + 0.5 v_0 \left[m_{gk} - m_0 \right]^2 + 0.5 \left[\frac{v_0}{v_{gk}} - 1 \right] \right] \end{aligned} \tag{A-10}$$

where we have grouped the corresponding terms to show $KL = 0$ when the parameters from the two distributions are equal. For the parameter $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}) \sim \prod_{gk} \Gamma(\beta_{gk}; a_0, b_0)$ and $q(\boldsymbol{\beta}) \sim \prod_{gk} \Gamma(\tilde{a}_{gk}, \tilde{b}_{gk})$

$$\begin{aligned} KL(q(\boldsymbol{\beta})||p(\boldsymbol{\beta})) = \textstyle\sum_{gk} [\quad & (\tilde{a}_{gk} - 1) \Psi(\tilde{a}_{gk}) - \log \tilde{b}_{gk} - \tilde{a}_{gk} - \log \Gamma(\tilde{a}_{gk}) \\ & + \log \Gamma(a_0) + a_0 \log b_0 - (a_0 - 1)(\Psi(\tilde{a}_{gk}) + \log \tilde{b}_{gk}) + \tfrac{\tilde{a}_{gk} \tilde{b}_{gk}}{b_0}] \end{aligned} \tag{A-11}$$

Again we shall group the terms to show that $KL = 0$ when the parameter from the two distributions are equal.

$$\begin{aligned} KL(q(\boldsymbol{\beta})||p(\boldsymbol{\beta})) = \textstyle\sum_{gk} \quad & [\, (\tilde{a}_{gk} - a_0) \Psi(\tilde{a}_{gk}) + a_0 (\log b_0 - \log \tilde{b}_{gk}) \\ & + \log \Gamma(a_0) - \log \Gamma(\tilde{a}_{gk}) + \tilde{a}_{gk} \left[\tfrac{\tilde{b}_{gk}}{b_0} - 1 \right]] \end{aligned} \tag{A-12}$$

Availability and Requirements

Project name: VBLPD: demonstration software
Project home page: http://www.enm.bris.ac.uk/cig/pubs/code/vbcode.htm
Operating system: Windows
Programming language: Windows executable and C++

Authors Contributions: Luke Carrivick implemented the method. Both authors wrote the paper.

Acknowledgements: Funding was provided by EPSRC grant EP/E027296/1 and EU grant MCRTN504231 to Dr. Colin Campbell

References

[1] http://www.enm.bris.ac.uk/cig/pubs/code/vbcode.htm.

[2] A.A. Alizadeh et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(3):503–511, 2000.

[3] H. Attias. A variational bayesian framework for graphical models, 2000.

[4] M Beal. *Variational algorithms for approximate bayesian inference*. Doctoral disseration, University College, London, 2003.

[5] M J Beal. Variational algorithms for approximate bayesian inference. PhD thesis, gatsby computational neuroscience unit, university college london, 2003.

[6] 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.

[7] A Bhattacharjee et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci.*, 98:13790–13795, 2001.

[8] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

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

[10] S P Brooks. Markov chain monte carlo method and it application. *The Statistician*, 47:69–100, 1998.

[11] C Campbell et al. A bayesian analysis of breast cancer expression microarray datasets highlights the importance of FOXC1 in breast cancer development. *Journal submission*, 2006.

[12] L. Carrivick, S. Rogers, J. Clark, C. Campbell, M. Girolami, and C. Cooper. Identification of prognostic signatures in breast cancer microarray data using bayesian techniques. *Journal of the Royal Society: Interface*, 3:367–381, 2006.

[13] 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.

[14] W Chu, Z Ghahramani, F Falciani, and Wild D L. Biomarker discovery in microarray gene expression data with gaussian processes. *Bioinformatics*, 21:3385–3393, 2005.

[15] A Dubey, S Hwang, C Rangel, C E Rasmussen, Z Ghahramani, and D L Wild. Clustering protein sequence and structure space with infinite gaussian mixture models. In *Pacific Symposium on Biocomputing*, pages 399–410, 2004.

[16] P Farmer et al. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene*, 24:4660–4671, 2005.

[17] 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.

[18] A Gelman and X Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.

[19] G Glinsky et al. Gene expression profiling predicts clinical outcome of prostate cancer. *J. Clin. Invest.*, 113:913–923, 2004.

[20] M Jordan, Z Ghahramani, T Jaakola, and L Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[21] 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.

[22] M Medvedovic and S Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18:1194–1206, 2002.

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

[24] R M Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

[25] Y Pawitan et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, 7:R953–R964, 2005.

[26] W.D. Penny and S.J. Roberts. Variational bayes for 1-dimensional mixture models. Technical report, Department of Engineering Science, Oxford University, 2000.

[27] 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.

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

[29] Y Tamimi et al. Identification of target genes regulated by FOXC1 using nickel agarose-based chromatin enrichment. *Invest. Ophthalmol. Vis. Sci.*, 45:3904–3913, 2004.

[30] A Teschendorff et al. A variational bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, 21:3025–3033, 2005.

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

[32] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley, 2003., 2003.

[33] Y Wang et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365:671–679, 2005.

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

[35] E A Williamson, I Wolf, J O'Kelly, S Bose, S Tanosaki, and H P Koeffler. BRCA1 and FOXA1 proteins coregulate the expression of the cell cycle inhibitor p27(kip1). *Oncogene*, 25:1391–1399, 2006.

[36] F Yang et al. Laser microdissection and microarray analysis of breast tumors reveal er-α related genes and pathways. *Oncogene*, 25:1413–1419, 2006.

[37] E-J Yeoh et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.

[38] Y Zhou et al. Identification of FOXC1 as a TGF-β1 responsive gene and its involvement in negative regulation of cell growth. *Genomics*, 80:465–472, 2002.