Gene expression

A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data

Andrew E. Teschendorff*, Yanzhong Wang, Nuno L. Barbosa-Morais, James D. Brenton and Carlos Caldas

Department of Oncology, Cancer Genomics Program, Hutchison-MRC Research Centre, University of Cambridge, Hills Road, Cambridge CB2 2XZ, UK

Received on October 11, 2004; revised and accepted on April 23, 2005
Advance Access publication April 28, 2005

ABSTRACT

Motivation: Accurate subcategorization of tumour types through gene-expression profiling requires analytical techniques that estimate the number of categories or clusters rigorously and reliably. Parametric mixture modelling provides a natural setting to address this problem.

Results: We compare a criterion for model selection that is derived from a variational Bayesian framework with a popular alternative based on the Bayesian information criterion. Using simulated data, we show that the variational Bayesian method is more accurate in finding the true number of clusters in situations that are relevant to current and future microarray studies. We also compare the two criteria using freely available tumour microarray datasets and show that the variational Bayesian method is more sensitive to capturing biologically relevant structure.

Availability: We have developed an R-package vabayelMix, available from www.cran.r-project.org, that implements the algorithm described in this paper.

Contact: aet21@cam.ac.uk

Supplementary information: http://bioinformatics.oxfordjournals.org

INTRODUCTION

The elucidation of the specific taxonomy of tumour types is an important preliminary step towards developing anti-cancer treatments that are specific to each patient. Such tumour- or patient-specific treatment is necessary given that all-too-often tumours with the same morphological and histopathological features have widely different clinical outcomes. Faced with this problem, large scale microarray experiments have been conducted (Golub et al., 1999; Alizadeh et al., 2000; Perou et al., 2000; Sorlie et al., 2003; Sotiriou et al., 2003) to profile the RNA expression levels over a wide range of specimens in an attempt to classify tumours or new tumour subtypes in terms of characteristic gene-expression profiles.

From a data analysis viewpoint, the subcategorization of a given tumour type in terms of the normalized and dimensionally reduced expression matrix can be tackled using unsupervised clustering algorithms (Hartigan, 1975) whereby specimens are clustered depending on how similar their gene-expression profiles are. A popular choice of unsupervised algorithm is hierarchical clustering (Eisen et al., 1998). Other unsupervised approaches that have been extensively applied to normalized microarray data include $k$-means (Tavazoie et al., 1999) and self-organizing maps (SOMs) (Tamayo et al., 1999).

These methods, however, have significant limitations that render them suboptimal for the subcategorization problem. Hierarchical clustering, by its very nature, does not allow rigorous identification of discrete clusters. However $k$-means and SOM allow distinct clusters to be inferred but the number of these have to be specified in advance. Such bias must be clearly avoided if the aim is to discover true sub-classes within a given class of tumours. Modified agglomerative and $k$-mean algorithms that attempt to infer the number of clusters in a dataset are described in Calinski and Harabasz (1974), Hartigan (1975) and Krzanowski and Lai (1985). More recently, a prediction-based resampling method (Dudoit and Fridlyand, 2002) was shown to perform better in comparison with other non-parametric methods. In spite of these continuous improvements, these methods are heuristic in nature and are not based on statistically rigorous principles. Consequently, they do not provide a rigorous framework to address important questions, e.g. to estimate the cluster membership probabilities of a specimen.

We support the view that the subcategorization problem is most naturally addressed within the context of a parametric mixture model. The advantage here is that we explicitly model the data as a mixture of an underlying set of components or clusters. The number of clusters is usually treated as an external parameter defining different models. The Bayesian information criterion (BIC) (Schwarz, 1978; Raftery, 1995) may then be used for model selection, that is, here for inferring the number of clusters. This is the approach taken in Fraley and Raftery (1999) where an Expectation Maximization (EM) based Gaussian mixture model was developed and subsequently applied to tumour microarray data in Yeung et al. (2001) and Ghosh and Chinnaiyan (2002). A comparison of the EM + BIC approach with leading non-parametric methods was provided in Yeung et al. (2001) and shown to be more accurate in cluster number prediction.

In spite of the clear advantages that a model-based approach to clustering provides, only a few subsequent model-based studies have been made within the microarray context (McLachlan et al., 2002; Muro et al., 2003; Monti et al., 2003). Muro et al. (2003) used a variational Bayesian mixture model (MacKay, 1995; Attias, 1999) for unsupervised clustering of genes in colorectal cancer. The variational Bayesian method is an attractive alternative to EM + BIC for unsupervised clustering and model selection problems. In particular, it may be used for inferring the number of clusters in a dataset.

*To whom correspondence should be addressed.
In this work, we compare the criterion for model selection derived from the variational Bayesian approach with the BIC approach (Fraley and Raftery, 1999) in a context that is relevant to existing and future gene-expression profiling studies. Specifically, the comparison is made within the setting of a Gaussian mixture model and uses both simulated data and freely available microarray data for evaluation purposes. We show that the variational Bayesian approach is more accurate in determining the true number of clusters for situations that are characteristic of current and future microarray studies. Furthermore, we provide a software package vabayelMix, written in R, that is freely available from www.cran.r-project.org

SYSTEMS AND METHODS

A variational Bayesian framework for inference and model selection

The variational Bayesian framework for inference and model selection has been described by MacKay (1995) and Attias (1999). The convergence properties and consistency of this approach has been established both empirically (Attias, 1999) and theoretically (Wang and Titterington, 2004). In common with other Bayesian methods it attempts to find a reasonably good approximation to a posterior density for which analytical marginalization is not possible. In this sense it provides an attractive alternative to well-established but computationally intensive methods, such as Markov chain Monte Carlo (MCMC), in that it uses a variational approach to optimize a proposal posterior.

The optimization algorithm can often be expressed in terms of a small number of iteration rules or equations that have fast convergence properties. Since the variational Bayesian method has not yet been extensively used in the microarray community we include a brief revision of the basic concepts below.

Let \( p(\theta|D, M) \) be the posterior density, where \( D \) is the data, \( \theta = \{\theta_i\}_{i=1,...} \) denotes the parameters to be inferred, and \( M \) denotes the model. Furthermore, let \( Q(\theta) \) denote the proposal posterior density.

We impose the constraint that \( Q(\theta) \) is separable, i.e. \( Q(\theta) = \prod_{i=1}^{n} Q_i(\theta_i) \) for an easy marginalization later. A suitable measure of how close \( Q(\theta) \) resembles \( p(\theta|D, M) \) is the Kullback-Leibler distance

\[
D(Q, P) = \int d\theta Q(\theta) \log \frac{Q(\theta)}{P(\theta)} \geq 0,
\]

which is positive definite for all \( Q \) and \( P \) and if and only if \( Q = P \). Using Bayes’ Theorem we can reformulate inequality (1) in terms of a ‘cost’ function \( C(Q, P) \),

\[
C(Q, P) = D(Q, P) - \log \int d\theta Q(\theta) \log \frac{Q(\theta)}{p(D|\theta, M)}
\]

\[
\geq -\log E(M),
\]

where \( p(D|\theta, M) \) is the likelihood, \( p(\theta|M) \) is the prior and \( E(M) \equiv p(D|M) \) is the evidence (or integrated likelihood) of the model. Thus, we seek to minimize \( C(Q, P) \) (or \( D(Q, P) \)) as a function of \( Q(\theta) \). Treating \( C(Q, P) \) as a functional of \( Q \) and assuming that the prior is also separable one can show after some calculation that the optimal \( Q_i \) have the form

\[
Q^*_i(\theta_i) = \frac{1}{Z_i} p(\theta_i|M)e^{-\log p(D|\theta_i, M)},
\]

where \( Z_i \) is the normalization factor

\[
Z_i = \int d\theta_i p(\theta_i|M)e^{-\log p(D|\theta_i, M)},
\]

and where \( <X>_{Q(\theta)} \) denotes the expectation of \( X \) under the combined posterior of all \( Q_j \), \( j \neq i \). Equation (3) can be solved iteratively until convergence. Note that even though the full posterior is separable the update of each individual parameter depends on the other posterior parameters.

After convergence of Equation (3) marginalization over the optimal separable approximation is easily accomplished.

A significant bonus of the variational approach is that it gives a criterion for model selection through a lower bound on the evidence as can be seen from Equation (2). As shown by Miskin (2000) for the case of simple mixture models, the bound provides in most circumstances a reasonably close approximation to the true evidence. Thus, given two models \( M_1 \) and \( M_2 \) we can compare their lower bounds \(-C(Q^*, P|M_1)\) and \(-C(Q^*, P|M_2)\) and select on the basis of the model which has the lowest cost function value.

Gaussian mixture modelling

In order to specify our mixture model within the Bayesian setting, we must make choices for the likelihood function as well as the prior densities. To keep things analytically tractable, we use the same choice of Gaussian components and conjugate priors as in Miskin (2000). Thus, writing the datamatrix \( D \) as \( \{x_s, s = 1, \ldots, N\} \) where each \( x_s \) is a \( d \)-dimensional vector and \( N \) denotes the number of observations (samples) to cluster, the likelihood is

\[
p(D|\theta, M) = \prod_{i=1}^{N} p(x_i|\theta, M) = \prod_{i=1}^{N} \sum_{c=1}^{K} \pi_c G(x_i|\mu_c, \Sigma_c),
\]

where \( c \) labels the component, \( \mu_c \) and \( \Sigma_c \) are the mean vector and inverse covariance matrix of Gaussian component \( c \) and \( \pi_c \) denotes the weight of component \( c \). It is important to note in the above that \( K \) is not the number of clusters that are inferred. On the contrary, it denotes the maximum possible number of clusters that could be inferred. Thus, without loss of generality it could be set equal to \( N \).

After convergence of the iteration rule (3), only a subset of weights \( \pi_c \) are found to be non-zero and it is the number in this subset that gives the estimate for the number of clusters.

The ensemble learning algorithm

To derive the learning rules for the parameters we are faced with the difficulty of evaluating \(-\log p(D|\theta, M)\) which cannot be done analytically for a mixture model. Nevertheless, Equation (2) may be brought into a tractable form using Jensen’s inequality and minimizing the cost functional also relative to an extra set of categorical weight parameters (Miskin, 2000).

For a Gaussian mixture model and the choice of conjugate priors as in Miskin (2000), the posterior densities take the same form as the priors leading to an iterative algorithm for learning the model parameters. The prior hyperparameters were chosen to yield complete uninformative priors. This was done either by fixing the prior means of the proposed clusters to be the same with broad prior variances or by using the data to choose random prior means and a variance estimate.⁴ For analytical tractability, we have only considered the optimization algorithm for a diagonal model in which the covariance matrices of the Gaussian components are assumed to be diagonal. Thus, this Gaussian mixture model corresponds to the case where the components may be elliptical clusters with unequal volume but that have been axis-aligned. Imposing the constraint that the components have equal orientation may sound too restrictive, yet this model has been shown to perform well in real microarray studies (Yeung et al., 2001). Moreover, for the applications and sample sizes (\( N \sim 100, d \sim 5 \)) we have in mind, the number of parameters that would need to be estimated for the complete unrestricted model would be of the order \( d^2 \times K \sim d^2 \sim 100 \), which would be of the order of the number of data points, thus making any inferences less reliable. A diagonal model is, therefore, in effect, the most complex model that we are considering given the constraints of the problem.

¹We found that the first method of implementing uninformative priors led to an instability of the learning algorithm for the univariate case. Thus, for the univariate case the second method was used.
limitations we do not give here the explicit iteration equations of the optimization learning algorithm. They can be found in Miskin (2000) and as R code in the software package vabayelMix.

**Number of clusters and cluster membership probabilities**

After the convergence of the optimization algorithm, the cluster membership probabilities of the samples are given by $p(c | x_i)$ where

$$p(c | x_i) = \int \pi(c) p(c | x_i, \theta) Q^*(\theta) d\theta$$

$$\propto \int d\theta_i p(x_i | c, \theta_i) Q^*(\theta_i)$$

$$\propto \pi^* \int d(\mu, \Sigma) G(x_i | \mu, \Sigma) Q^*(\mu) Q^*(\Sigma)$$

(6)

and a maximum probability criterion could be used to assign each sample to a specific cluster. Details of how the integral is computed can be found in vabayelMix. The estimated number of clusters is given by the number of non-zero weights $\pi^*$.

**Simulated data**

The use of Gaussian components to simulate microarray data is clearly not ideal, yet the Gaussian model has been shown to be a reasonably good approximation for suitably normalized real data (Yeung et al., 2001). Since the aim is to evaluate and compare the ability of two parametric clustering methods for capturing structure, without loss of generality, we may focus on the simplest case where the data are generated by two types of samples (Dasgupta, 1999). We consider both univariate and multivariate (two dimensions) Gaussian components. In all multivariate models we assume that the underlying clusters have the same covariance structure. Models A and B correspond to the case where the clusters are spherical, whereas in model C we allow the clusters to be elliptical. Thus, in models A and B the dimensions are symmetric, whereas in model C the degree of separation of the clusters depends on the specific dimension. Furthermore, we consider different relative cluster sizes to allow for unbalanced studies. Thus, in models A and C we assume the clusters to be of equal size, whereas in model B we allow for unequal sample sizes. All models are analysed for a range of different degrees of separation of the clusters, known as ‘c-separation’, as defined in Dasgupta (1999). Thus, we consider Gaussian components that are c-separated with $c > 2$ (linearly separable), two-separated (almost linearly separable) and c-separated with $c < 2$ (overlapping). We generated models in each of these three categories by fixing, without loss of generality, the centres to be at 0 and 1 for the univariate case, and (0,0) and (1,1) for the multivariate case, considering a range of different standard deviations (SD = 0.75, 0.5, 0.25). Thus, Model-A with (SD = 0.75) corresponds to $c = 4/3(<2)$, which corresponds to the case of overlapping clusters (Dasgupta, 1999). We did not consider c-separation with $c < 1$, since that corresponds to the extreme case of significant overlap between clusters and is not likely to be relevant for current classification studies. Finally, we consider a wide range of total sample sizes that are typical of current and future microarray studies.

**Microarray data**

**Dimensional reduction** If one wishes to apply parametric mixture models to unsupervised clustering of biological samples profiled using microarrays, a prior dimensional reduction step is necessary. PCA was used by Ghosh and Chinnaiyan (2002) for the dimensional reduction step and subsequent clustering was performed along the two principal components. In the present study, we use independent component analysis (ICA) (reviewed by Hyvaerinen et al., 2001) instead of PCA, since this method can find projections or modes that are biologically more meaningful (Marrogi et al., 2002; Liebermeister, 2002). Another advantage of ICA over PCA is that it only finds non-Gaussian projections. Thus, projections that are Gaussian and are likely to represent noise are filtered out. We have implemented the independent component analysis using a maximum-likelihood framework (Hyvaerinen et al., 2001) as an R-package, mllica, also available from www.cran.r-project.org.

**RESULTS AND DISCUSSION**

**Simulated data**

The aim of using artificial data is to provide a framework in which to objectively compare the prediction accuracy of two model-based clustering approaches. Our focus is on correct cluster number prediction, misclassification rates and on how the accuracy of prediction varies with the distributional parameters.

From a biologist’s perspective what matters most in the context of unsupervised clustering is whether the algorithm predicts structure or not, whether it distinguishes, in a statistically significant way, the case where there are distinct clusters from the case where there are not (McShane et al., 2002). Thus, we have defined an instance of an algorithm as giving an error if it predicts less than the true number of clusters. In what follows, we call this the error rate or cluster number error rate. Since the output of an algorithm may predict the correct number of clusters and yet misclassify many samples, we have also considered the total misclassification error (or accuracy) conditioned on the algorithm predicting the right number of clusters. To ensure the robustness of our results all simulation models were randomly generated 500 times and the resulting error rates recorded.

Figure 1 shows the mean error rate as a function of total sample size for the three models A, B and C and for the case of clustering over two dimensions. Different choices of cluster variances are shown and correspond to situations that differ in how distinct the two clusters are (see Supplementary Figure 1). Strikingly, we found that the VB algorithm is more sensitive to capturing structure than the combined EM and BIC approach (EM + BIC) (Fig. 1). In fact, the VB algorithm is more accurate for all cases where the clusters are less well differentiated. For example, for Model B and SD = 0.75 we found that the VB algorithm obtained error rates $<30\%$ over a wide range of different samples sizes, whereas the EM + BIC failed to capture structure in $>40\%$ of the runs. Similar difference in performance was observed for Models A and C. We verified that on increasing the sample sizes further to 400 (Model A) and 480 (Model B) we obtained zero error rates for the VB algorithm and very small error rates (0.09 and 0.04) for EM + BIC. Also, as expected, for well separated clusters (SD = 0.25) both algorithms achieved zero or close to zero error rates.

In spite of the better performance of the VB algorithm over EM + BIC in capturing structure, this in itself does not necessarily imply that the misclassification rates are better for the VB algorithm. To address this issue, we analysed the misclassification rate of the two algorithms conditioned to the case where the algorithm predicts the right number of clusters. The results for clustering over two dimensions and Models A, B and C are shown in Figure 2. Interestingly, we found that the VB algorithm outperformed the EM + BIC algorithm once again. In fact, for Models A and C the observed differences in misclassification rates were significant. Thus, while the VB algorithm (Model C) achieved $<20\%$ misclassifications for a total sample size of 200 and less well differentiated clusters [SD = (0.75, 0.5)], the misclassification rate of EM + BIC algorithm was $>30\%$. These results indicate, therefore, that the improved sensitivity of the VB algorithm to capturing structure in the data is due to a more accurate learning of the underlying clusters and their parameters.
Fig. 1. Error rates for predicted cluster numbers versus total sample size, for different 2D simulation models and for each parametric clustering algorithm. VB (black) denotes the variational Bayesian algorithm. BIC (grey) denotes the EM + BIC approach. Models A and C correspond to underlying clusters of the same size; Model B to relative cluster sizes in the proportion 3:1. The x-axis labels the total size of the sample set considered. For Models A and B the clusters are spherical, for Model C they are elliptical, the SD of the clusters as shown. The error rates given represent averages over 500 randomly generated simulations.
Fig. 2. Average conditional misclassification rates versus total sample size over 500 2D simulation models A, B and C and for each parametric clustering algorithm. VB (black) denotes the variational Bayesian algorithm. BIC (grey) denotes the EM + BIC approach.
Similar results were obtained for clustering over a single dimension (Supplementary Figure 2). Thus, for overlapping clusters \( SD = (0.75, 0.5) \) we found that the VB algorithm significantly outperformed EM + BIC over a wide range of sample sizes. Strikingly, even for large sample sizes (~400), EM + BIC failed to capture structure in >80% of the simulations, whereas the VB algorithm obtained cluster number error rates <20% with associated misclassification rates (data not shown) of ~40% (SD = 0.75) and 23% (SD = 0.5). In fact, the two algorithms showed widely different trends as sample sizes were increased. While the cluster number error rate of VB algorithm decreased consistently, the EM + BIC showed no significant improvement over the considered sample size range. It was only for very large sample sizes that we observed the convergence to zero error rates in the case of EM + BIC (e.g. for Model A and SD = 0.5, cluster number error rates were 58% and 1% for sample sizes of 1000 and 2000, respectively, with associated misclassification rates of ~16% for both sample sizes). For well differentiated clusters (SD = 0.25) we also found that the VB algorithm outperformed EM + BIC. Only for the special case of Model A and small sample sizes did the EM + BIC algorithm perform better.

Breast tumours

The complete molecular classification of breast tumours is still being elucidated. Many microarray studies (Perou et al., 2000; Sotiriou et al., 2003; Sorlie et al., 2003) have established that breast tumours fall into two main subtypes depending on whether the estrogen receptor gene, \( ESR1 \), is expressed or not. These studies have also provided strong evidence for other molecular subdivisions, such as the separation of breast tumours into basal and luminal subtypes as well as subtypes defined by differential expression of \( ERBB2 \). It is less clear whether the subdivision of normal basal and luminal tumours into further basal and luminal subtypes, as proposed in Sotiriou et al. (2003) and Sorlie et al. (2003), is statistically meaningful. For example, in Sotiriou et al. (2003) the clustering algorithm of McShane et al. (2002) was used to establish the existence of two clusters associated with the basal and luminal characteristics of the tumours, yet the algorithm failed to uncover any other statistically significant substructure. In spite of this, the authors (Sotiriou et al., 2003) found clinically interesting substructure within the main basal and luminal groups. In particular, the luminal group was shown to be subdivided into three main subtypes (luminal-1, -2 and -3) with the luminal-1 group having a statistically significant better outcome than the luminal-2 and luminal-3 subgroups.

The subcategorization of breast tumours provides a testground in which to apply and compare unsupervised clustering algorithms that aim to identify the number of subtypes. In particular, we aimed to compare the performance of the two parametric clustering methods. Since the methods are parametric and sample sizes are limited, we considered clustering only over a small number of dimensions. To compare the two methods objectively, we decided to apply the algorithms to cases where there is evidence for substructure and where this substructure is either related to a clinical parameter or is of biological significance.

We first devised a particular test to see whether the results obtained using simulated data occur within the realm of real expression data. To this end we asked whether the two algorithms could predict the clinically relevant substructure within the luminal types as claimed in Sotiriou et al. (2003). In their work, a luminal-1 subgroup was identified that was characterized by differential higher expression of \( KIT \), \( HGF \), \( ATF3 \) and \( IGFBP3 \), lower expression of cell-growth related genes, such as \( PCNA \), \( CDC2 \) and \( BUB1 \), and which compared favourably with the rest of the luminal samples with respect to clinical outcome. Hierarchical clustering over these 7 genes revealed two main clusters (17 ‘good-outcome’ and 42 ‘bad-outcome’ samples), with the smaller cluster mapping precisely to the luminal-1 subgroup as identified in Sotiriou et al. (2003). To evaluate and compare the two parametric methods, we considered clustering the luminal subtype samples over two of these seven discriminatory genes. This was done for all pairwise combinations (a total of 21), and for each such combination the predicted number of clusters was compared between the two methods (Fig. 3). From this figure we can see that the VB algorithm was able to predict structure in 95% (20/21) of clusterings, whereas for the EM + BIC algorithm structure was predicted only for 48% (10/21) of the runs. While the precise distribution of samples among the predicted clusters was highly dependent on the pair of genes used, we would expect the predicted clusters to be significantly correlated with outcome. We checked this with a \( k \times 2 \) contingency table test (Fisher-test), the entries of the table representing the number of samples from the ‘good-outcome’ and the ‘bad-outcome’ groups present in each predicted cluster (\( k \) clusters in total). Conditioning on the cases (20 in total) where the variational Bayesian algorithm was able to predict statistically significant structure (as opposed to no structure at all), we found that in 90% (18/20) of these cases the structure predicted was significantly associated (\( P < 0.05 \)) with the clinically relevant substructure identified in Sotiriou et al. (2003) (Fig. 3). Thus, we can conclude that the variational Bayesian approach is more sensitive than EM + BIC in capturing structure, and most importantly, we have shown that the clusters it predicts are not arbitrary partitions but relate to a partition that is significantly correlated with clinical outcome.

In view of the improved sensitivity of the variational approach to capturing structure, we decided to further compare the two parametric clustering methods within the semi-supervised setting in which we envisage using these model-based methods. This means either clustering over a small number of selected genes or a small number of specific 1D projections, such as the ones found using PCA or ICA. Here we decided to systematically apply the clustering algorithms to projections inferred using ICA. The significance of an independent component was determined mainly by two properties: (1) whether it separated the samples into two or more clusters and (2) whether it was enriched in terms of specific Gene Ontologies (GO) or whether it correlated significantly with a clinical parameter, such as \( ESR1 \) expression, Grade or Nodal Status. It is clear that to establish property (1), an algorithm that predicted the number of clusters objectively was required. The result of clustering with the two parametric methods on each IC is shown in Table 1 for two different datasets (Sotiriou et al., 2003; van Veer et al., 2002). We found that in both datasets the variational Bayesian method was able to capture significant structure along each inferred IC, whereas the EM + BIC did so only for a few modes.

Next, we describe specific modes in each dataset over which the VB algorithm predicted interesting structure. We emphasize that for some of these modes the EM + BIC failed to predict any clusters. Sotiriou

\(1\) IC-5 and IC-9. A Wilcoxon-test showed that these modes were significantly associated with ER status. The application of
A variational Bayesian mixture modelling framework

Fig. 3. A-1, A-2 and A-3: Scatterplots of normalized expression ratios of selected genes characterizing the luminal-1 subgroup (in blue) from the rest of luminal samples (red) as defined in Sotiriou et al. (2003). The different shapes specify the cluster membership of the samples as predicted by the VB algorithm. (B) Predicted cluster numbers obtained from clustering the luminal type samples along every possible pair of genes. Upper (Lower) triangular matrix represents results from the application of the EM + BIC (VB) algorithm. Out of the 21 clusterings using VB, 20 predicted significant structure, and 18 (90%) of these were significantly correlated (in red) with the clinical outcome (Fisher test \( P < 0.05 \)). Similarly, for the 10 clusterings in which EM + BIC predicted structure, only 5 (50%) out of these were significantly correlated (in red) with the clinical outcome.

the VB algorithm along these two modes predicted two main clusters (Supplementary Figure 3A) that were significantly associated with ESR1 expression (\( P < 10^{-13} \) Fisher test). Using EM + BIC over these two projections gave a very similar result (data not shown). Thus, in this case both algorithms were able to capture biologically relevant substructure.

(2) IC-4. The VB algorithm predicted two clusters of sizes 84 and 15 (Supplementary Figure 3B), while no structure was predicted using EM + BIC. This mode was found to be significantly associated with the degree of differentiation of the tumours (Kruskal–Wallis test \( P = 0.043 \)) and to be significantly enriched by genes relating to immune response, and membrane and MHC class II receptor activities [Hypergeometric test \( P < 10^{-10} \) using GOTM (Zhang et al., 2004) and Supplementary Table 1]. Furthermore, the subdivision predicted by the VB algorithm (Supplementary Figure 3B) was indeed significantly associated with Grade-1 versus Grade-3 (Table 2, Fisher test \( P = 0.014 \)). A similar subdivision was found for IC-2 in vant Veer (see below).

(3) IC-7. The VB algorithm predicted two clusters of sizes 78 and 21 (Supplementary Figure 4A and B), while no structure was predicted using EM + BIC. This mode was found to be significantly enriched by genes related to extracellular space and extracellular matrix functions, collagen binding, proteolysis, peptidolysis and bone formation (Hypergeometric test \( P < 10^{-7} \) and Supplementary Table 1). Even though the clusters predicted by VB were only weakly associated with Grade, this subdivision of breast tumours based on differential expression of the osteoblastic module (Supplementary Figure 4C) has been consistently observed across different tumour types (Segal et al., 2004) and across different breast cancer studies (A. Teschendorff et al., manuscript in preparation), and may, therefore, play an important role in elucidating breast cancer and general cancer taxonomy. A corresponding classification was found for IC-13 in vant Veer (see below).
We have compared two parametric approaches to unsupervised clustering of gene-expression data. Both methods fit a Gaussian mixture model to data but use different criteria for model selection. Thus, the EM method uses the BIC for predicting the optimal number of clusters, whereas the variational approach uses a lower bound on the evidence to select this number.

The two methods were first compared by applying them separately to the same simulated data. We considered a range of simulation models that differed in the degree of separation of the clusters. Also, since the methods were parametric, we focused on clustering only over a small number of dimensions. We also focused entirely on simulation models involving just two components. Clearly, both these restrictions are violated in practice, since typically we might want to cluster over a number of genes or 1D projections, and data are not often best described in terms of only two underlying clusters. In spite of these restrictions, we have found that the results obtained here generalize to both higher dimensions and to cases where there are more than two clusters (data not shown). Using the simulated data we have shown that the variational Bayesian approach is more sensitive at capturing structure than the EM + BIC algorithm. Furthermore, we showed that the misclassification rates were also smaller for the variational Bayesian approach. All this is true for a range of sample sizes that are typical of current and future microarray experiments. Our results are supported by similar observations made in more theoretical contexts (Beal and Ghahramani, 2003).

We reached similar conclusions by comparing both algorithms on real gene-expression data. By considering a set of genes that discriminated samples according to their clinical outcome and then clustering over all possible pairwise combinations of genes, we were able to show that the variational approach was more sensitive to predicting structure than EM + BIC, and that the predicted clusters were significantly associated with clinical outcome in breast cancer. We also applied the algorithms to the independent components inferred in two different datasets. Here, again, we found that the VB approach was able to predict biologically relevant structure for cases where the EM + BIC algorithm failed.

Clearly, the variational Bayesian framework advocated here also allows prior information to be easily incorporated. For example, if prior knowledge about the number of clusters is available, this can be incorporated by assigning correspondingly different weights to the parameters in the prior Dirichlet distribution. Models with different priors can then be objectively compared through the estimated lower bounds on the evidence.

Another advantage of the variational approach over BIC for model selection is that it treats the number of clusters as an internal parameter that is inferred together with the cluster centres and variances. Therefore, it explicitly avoids having to run separate models

### Table 1.

We give the number of clusters predicted by the VB and EM + BIC algorithms on each IC inferred in Sotiriou et al. (2003) and selected IC in van’t Veer et al. (2002)

<table>
<thead>
<tr>
<th>IC</th>
<th>IC1</th>
<th>IC2</th>
<th>IC3</th>
<th>IC4</th>
<th>IC5</th>
<th>IC6</th>
<th>IC7</th>
<th>IC8</th>
<th>IC9</th>
<th>IC10</th>
<th>IC12</th>
<th>IC13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sotiriou VB</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>EM + BIC</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>van’t Veer VB</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>EM + BIC</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 2.

Distribution of grades of breast tumours among the two main clusters obtained by the application of the VB algorithm along IC-4 (Sotiriou) and IC-2 (van’t Veer)

<table>
<thead>
<tr>
<th></th>
<th>Grade-1</th>
<th>Grade-2</th>
<th>Grade-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sotiriou Cluster-1</td>
<td>16</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>Cluster-2</td>
<td>0</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>van’t Veer Cluster-1</td>
<td>12</td>
<td>25</td>
<td>47</td>
</tr>
<tr>
<td>Cluster-2</td>
<td>0</td>
<td>1</td>
<td>31</td>
</tr>
</tbody>
</table>

(1) IC-5. The VB algorithm predicted two main clusters of sizes 14 and 103, while no structure was predicted using EM + BIC. This mode was found to be enriched with keratins (KRT1, 5, 10, 13, 15, 17, 23, 2A) (Supplementary Table 2) corresponding to filament cytoskeleton components (GOTM P = 10^-10). A similar subdivision of samples based on differential expression of the keratin gene family has been documented in Sotiriou et al. (2003) and Sorlie et al. (2003).

(2) IC-2. This mode was associated with grade (Kruskal–Wallis P < 10^-4). The VB algorithm predicted two clusters of sizes 32 and 84, which were also discriminative of Grades (Table 2; Fisher test p < 10^-6). Moreover, this mode was enriched with genes associated with immune response functions (GOTM P < 10^-3 and Supplementary Table 2), suggesting a correspondence with IC-4 described in Sotiriou subsection.

(3) IC-13. The VB algorithm predicted two clusters of sizes 57 and 55 (Supplementary Figure S5A and B), whereas EM + BIC predicted no structure. This mode was enriched with genes related to extracellular space and matrix functions (GOTM P < 10^-34 and Supplementary Table 2) and the osteoblastic module (Supplementary Figure S5C) (Segal et al., 2004), and corresponds to IC-7 described in Sotiriou subsection.

**CONCLUSIONS**

We have compared two parametric approaches to unsupervised clustering of gene-expression data. Both methods fit a Gaussian mixture model to data but use different criteria for model selection. Thus, the EM method uses the BIC for predicting the optimal number of clusters, whereas the variational approach uses a lower bound on the evidence to select this number.

The two methods were first compared by applying them separately to the same simulated data. We considered a range of simulation models that differed in the degree of separation of the clusters. Also, since the methods were parametric, we focused on clustering only over a small number of dimensions. We also focused entirely on simulation models involving just two components. Clearly, both these restrictions are violated in practice, since typically we might want to cluster over a number of genes or 1D projections, and data are not often best described in terms of only two underlying clusters. In spite of these restrictions, we have found that the results obtained here generalize to both higher dimensions and to cases where there are more than two clusters (data not shown). Using the simulated data we have shown that the variational Bayesian approach is more sensitive at capturing structure than the EM + BIC algorithm. Furthermore, we showed that the misclassification rates were also smaller for the variational Bayesian approach. All this is true for a range of sample sizes that are typical of current and future microarray experiments. Our results are supported by similar observations made in more theoretical contexts (Beal and Ghahramani, 2003).

We reached similar conclusions by comparing both algorithms on real gene-expression data. By considering a set of genes that discriminated samples according to their clinical outcome and then clustering over all possible pairwise combinations of genes, we were able to show that the variational approach was more sensitive to predicting structure than EM + BIC, and that the predicted clusters were significantly associated with clinical outcome in breast cancer. We also applied the algorithms to the independent components inferred in two different datasets. Here, again, we found that the VB approach was able to predict biologically relevant structure for cases where the EM + BIC algorithm failed.

Clearly, the variational Bayesian framework advocated here also allows prior information to be easily incorporated. For example, if prior knowledge about the number of clusters is available, this can be incorporated by assigning correspondingly different weights to the parameters in the prior Dirichlet distribution. Models with different priors can then be objectively compared through the estimated lower bounds on the evidence.

Another advantage of the variational approach over BIC for model selection is that it treats the number of clusters as an internal parameter that is inferred together with the cluster centres and variances. Therefore, it explicitly avoids having to run separate models.
with different numbers of clusters. Instead, each converged algorithm run predicts a specific number of clusters. Separate runs are, however, still required to check the robustness of this number and the associated sample distribution to the random initialization point of the algorithm. In general, we found that the number of clusters that the algorithm predicts is fairly robust to the initialization point, unless clustering is done over more dimensions (for our specific datasets and sample sizes this meant more than two dimensions). In these latter cases, the distinct solutions can be easily and objectively compared using their converged cost function values. Alternatively, the distinct clustering solutions may be combined using consensus clustering as proposed in Monti et al. (2003).

We found that convergence times depended on several factors including the structure inherent in the data. For the datasets used here, convergence times were fairly fast: 10 converged runs over two dimensions and 100 samples took \(\sim 35\) s on a single Pentium 4 3.1 GHz processor. In general, we found that the algorithm scaled reasonably well with an increase in dimensionality. Since most applications would involve clustering over at most tens of dimensions, this is not a limiting factor. More problematic could be the increase of convergence time with the number of samples. We found that as long as clustering is restricted to smaller sample sets (\(\sim 100–500\)) the ensemble learning algorithm implemented here is fairly fast. For future applications where one might want to cluster thousands of specimens longer running times (\(\sim 2–3\) h) may be expected.

In conclusion, the variational method provides a computationally tractable Bayesian framework to infer statistically significant clusters. It compares favourably against a popular alternative based on the BIC in terms of improved sensitivity and accuracy for cluster number prediction. We believe that the method will be valuable for the future large gene profiling studies that attempt to infer structure and find subcategories within poorly characterized tumour types.

ACKNOWLEDGEMENT

This research was supported by the Cambridge-MIT Institute (AET), Cancer Research UK (JDB), FCT Ministry of Science Portugal (NB-LB-M) and NTRAC (National Translational Cancer Research Network).

REFERENCES


