

Bayesian clustering of gene expression microarray data for subgroup identification

Metodi bayesiani per l'aggregazione di dati microarray di espressione genica per l'identificazione di gruppi

Leonardo Bottolo

Dip. di Ec. Pol. e Metodi Quantitativi
Università di Pavia

and

Department of Statistics
University of Oxford
bottolo@stat.ox.ac.uk

Guido Consonni

Dip. di Ec. Pol. e Metodi Quantitativi
Università di Pavia

guido.consonni@unipv.it

Riassunto: La tecnologia per la raccolta di dati genetici si sta sviluppando con grande rapidità generando grandi basi di dati, con particolare riferimento alle osservazioni *microarray* riferite all' "espressione " di geni. Sorge quindi l'esigenza di sviluppare analisi statistiche capaci di raccogliere questa sfida, anche se finora queste rimangono spesso confinate in un ambito descrittivo. Recentemente tuttavia sono stati proposti modelli ispirati all'impostazione bayesiana che hanno il pregio di fornire una base probabilistica coerente alle categorie concettuali impiegate (quali raggruppamento, profilo molecolare etc), oltre a consentire stime più efficienti, grazie al cosiddetto effetto di *borrowing strength*. In particolare sono stati proposti modelli gerarchici bayesiani nei quali le osservazioni sono generate da una distribuzione mistura a tre componenti (a seconda che il gene sia sottoespresso, normalmente espresso e sovraespresso); le distribuzioni iniziali dei parametri sono invece basate sull'ipotesi tradizionale di scambiabilità. In questo lavoro proponiamo un modello gerarchico nel quale i parametri che rappresentano le probabilità che i geni provengano dalla componente sotto- o sovraespressa non sono più contraddistinti dall'ipotesi di scambiabilità, ma piuttosto da una distribuzione mistura con un numero aleatorio di componenti. In questo modo il modello consente di cogliere la presenza di aspetti di eterogeneità conducendo di conseguenza ad un modello bayesiano di *clustering* sia per quanto attiene ai geni che alle unità. Il modello viene applicato ad un insieme di dati simulati e ad uno di osservazioni reali.

Keywords: Clustering; Gene-expression; Hierarchical mixture priors; Microarray data; Partition; Reversible Jump.

1. Introduction

Technologies for the collection of genetic data, such as those based on microarrays for gene expression, are developing at a very fast rate. This process generates huge data sets which call for timely statistical analysis.

A typical microarray data set for gene expression is represented by a matrix with rows representing genes while columns refer to samples: gene expressions are visually displayed using a color map, possibly after some suitable reordering of the rows and

⁰Partially funded by Università di Pavia (Fondi Ateneo Ricerca)

columns, based on standard clustering techniques, to better appreciate the structure of the output. Ordinarily the ratio between the number of genes and samples is around 100.

A crucial objective in these studies is clustering. In particular gene clustering based on the similarity of their level of expression is of paramount importance, in order to narrow down further investigations to a few genes, possibly with a view to prognosis and therapy.

A related clustering task concerns samples themselves: this goes under the name of molecular profiling, and aims at identifying similarities across samples based on a few genes identified from the previous stage of the analysis.

In this paper we employ a fully Bayesian hierarchical model. A key ingredient of our approach is represented by the use of mixture distributions. Specifically, mixtures are used at two levels in a very distinct fashion. First of all expression measurements are assumed to originate from a three-component mixture distribution, see [5] and [11], representing the underlying population classes (baseline, underexpression and overexpression) of samples relative to gene-expression. For each gene we assume a specific (random) probability of belonging to any of the three classes. In the second step, we assume that each gene-specific probability arises from a mixture of prior distributions with an unknown number of components: this represents the novelty of our approach, which was outlined for the first time in [3]. We shall detail the technicalities and advantages of such procedure, relative to current Bayesian implementations such as those described in [5], in the following pages. We validate our model using the simulated data set POEsimu taken from [5] which consists of a 1000x72 matrix of simulated data. The model is then applied to a real data set described by [1]. The original data set contains information on 1000 genes in 203 samples whereas the subset that we use is a 460x203 matrix. Of the 203 specimens 186 are snap-frozen lung tumors while the remaining 17 are normal.

2. Hierarchical modelling of gene expression measurements

Suppose data are arranged in a $G \times S$ expression matrix $A = (a_{gs})$ where a_{gs} represents the measured transcript abundance of gene g in sample s , or a transformation thereof. We follow [11] and [5], to which we refer for further details, in assuming that there exist three distinct categories of expression for each entry in the matrix A .

The observations a_{gs} are assumed to arise from a mixture of three components: $f_{0,g}$, $f_{-,g}$ and $f_{+,g}$ with weights π_g^- , π_g^0 and π_g^+ , say, where $\pi_g^- + \pi_g^0 + \pi_g^+ = 1$. In particular $f_{0,g}$ represents small biological variation across samples, as well as errors in measuring abundance in the hybridization experiment. On the other hand the densities $f_{-,g}$ and $f_{+,g}$, having disjoint supports, are meant to show the variation of expression for gene g in samples that exhibit underexpression, respectively overexpression, relative to the norm depicted by $f_{0,g}$.

It is expedient to introduce allocation variables e_{gs} , such that: $e_{gs} = -, 0, \text{ or } +$, if gene g has, respectively, abnormally low, normal, or high expression in sample s . Clearly π_g^- represents the probability that $e_{gs} = -$, and similarly for the remaining ones.

In the above formulation samples are implicitly regarded as experimental replications. In practice this cannot always be assumed, e.g. when they correspond to tumors, since heterogeneity across samples is to be expected. This suggests to incorporate a sample specific effect into the analysis, as we do in the following specification of the model which we adopt from [11]

$$\begin{aligned}
a_{gs} | \alpha_s, \mu_g, \zeta_g, \delta_g, \pi_g &\stackrel{ind}{\sim} \\
\pi_g^- U(-\delta_g^- + \alpha_s + \mu_g, \alpha_s + \mu_g) &+ \\
+\pi_g^0 N(\alpha_s + \mu_g, \zeta_g^2) &+ \\
+\pi_g^+ U(\alpha_s + \mu_g, \alpha_s + \mu_g + \delta_g^+) &,
\end{aligned} \tag{1}$$

where π_g stands for the collection $\pi_g^-, \pi_g^0, \pi_g^+$, $U(a, b)$ denotes a uniform density over (a, b) , while $N(\mu, \sigma^2)$ stands for the normal distribution with mean μ and variance σ^2 .

In practice normally expressed samples are centered around $\alpha_s + \mu_g$, while δ_g^- , and δ_g^+ , represent displacements for those underexpressed, respectively, overexpressed. Notice that the sample effects α_s only appear as additive components in locating the center of the distribution for normally expressed data.

The hierarchical model requires now a probability assignment to all conditioning random variables appearing in (1). Current Bayesian modeling adopt standard prior specifications, We remark that the parameters λ_g^- and λ_g^+ are modelled separately, merely on the ground of simplicity. Indeed it would be more appropriate, because of the constraint $\pi_g^- + \pi_g^0 + \pi_g^+ = 1$, to model $(\lambda_g^-, \lambda_g^+)$ jointly, see the Discussion in this paper. Save for the parameters α_s , all remaining parameters are assumed to be exchangeable, i.e. i.i.d. conditionally on some further hyperparameters. In practice all gene-specific parameters are assumed to arise from a single parent distribution. A possible justification is that measurement technology may affect in a similar way all data collected. Exchangeability is a simple yet powerful assumption: it allows to borrow strength across the whole genomic distribution for estimating each gene-specific parameter, and it avoids having to specify tedious identifiability constraints that would be needed in a standard frequentist setting (such as Maximum Likelihood Estimation). We shall return to the issue of exchangeability later in the paper.

A further prior specification involves the hyperparameters appearing in the exchangeable structure described above. Typically mildly informative proper priors will be used, while noninformative priors should be avoided as they may lead to identifiability problems.

As remarked in the Introduction, a first inferential goal concerns classifying genes in terms of differential expressions across samples. This would allow to identify a relatively small set of homogeneous genes, which can subsequently be used for molecular profiling (see below for details) in order to group samples based on similarities of gene expression patterns.

A current Bayesian procedure for classifying genes in terms of differential expression is described in [5]. We avoid details here but simply summarize some features of their work which motivated our approach. First the number, k say, of distinct gene categories need to be fixed. Next a *differential expression pattern* of interest has to selected. The idea here is to state *a priori* how many samples within a gene are expected to be underexpressed and how many overexpressed. For example the pattern $(0.05, 0.20)$ indicates that 5% of the samples are underexpressed, 20% of them are overexpressed, while the remaining 75% belong to the normally-expressed component of the mixture. An iterative algorithm is then applied in order to select, from each group, a representative gene. This is based on agreement scores as well as coherency scores. As the Authors remark, the identification of the distinct sets of genes is conditional on the number k as well as the chosen pattern. In other words both the clustering structure and the representative may

change depending on these inputs. This is clearly unfortunate, especially when no firm guidance can be given regarding the choice of such inputs. Moreover an obvious issue of robustness arises which may require running the procedure several times.

Having identified a representative gene for each group, samples can be clustered according to molecular profiling based on the joint probability

$$\Pr\{e_{g_1s} = i_1, e_{g_2s} = i_2, \dots, e_{g_ks} = i_k | \text{data}\}, \quad i_j \in \{-, 0, +\}.$$

3. Hierarchical mixture priors for the probability of abnormal expressions

In this paper we focus on the assignment of a prior distribution on λ_g^- , and λ_g^+ . As remarked earlier, conventional practice regards each set of quantities $\{\lambda_g^-\}$, $\{\lambda_g^+\}$ as exchangeable. While it is generally agreed that borrowing strength is a very useful aspect of hierarchical modelling, we feel that the assumption of (full) exchangeability of the parameters may be often too crude. In particular, gene-specific estimates are shrunk towards an *intermediate* value, thus influencing the corresponding clustering structure.

Of course the previous criticism is not peculiar to this setting, since it represents a well known drawback of any modelling based on exchangeability assumptions; alternative and more flexible approaches are now available, based on partition models ([8] and [4]), product partition models ([13]), or hierarchical mixture priors ([10] and [2]).

We here consider the approach based on hierarchical mixture priors. In this setting, gene-specific parameters are allowed to arise from a mixture of prior distributions in order to accommodate in a flexible way the underlying clustering structure. In this way heterogeneity is better accounted for, and borrowing strength takes place in a selective way, producing multiple shrinkage estimates.

Specifically, the prior distribution for the λ_g^- s is taken to be

$$\lambda_g^- \mid k^{\lambda^-}, \omega_j^{\lambda^-}, \theta_j^{\lambda^-}, \sigma_j^{\lambda^-} \stackrel{iid}{\sim} \sum_{j=1}^{k^{\lambda^-}} \omega_j^{\lambda^-} N\left(\theta_j^{\lambda^-}, \left(\sigma_j^{\lambda^-}\right)^2\right), \quad (2)$$

and similarly for that of the λ_g^+ s with obvious modifications.

We emphasize that our model allows for: i) a distinct random number of components in the mixture prior, one for λ_g^- and one for λ_g^+ ; ii) distinct random weights for the mixture components as well as distinct location-scale hyperparameters.

Again, it is expedient to introduce latent allocation variables $\{z_g^{\lambda^-}\}$ and $\{z_g^{\lambda^+}\}$ for computational and inferential purposes.

Consider for example $z_g^{\lambda^-}$. We have, independently for each $g = 1, \dots, G$,

$$\Pr\left\{z_g^{\lambda^-} = j \mid k^{\lambda^-}, \omega^{\lambda^-}\right\} = \omega_j^{\lambda^-}, \quad j = 1, \dots, k^{\lambda^-},$$

with similar expressions holding for the distribution of $z_g^{\lambda^+}$.

To complete the specification of the model, the distributions of all the remaining random quantities involved need to be assigned. Given the lack of space, we do not provide them here, as they follow closely previous work in the area, such as [10], to which we refer for further details; see also [2] for an application in a different context.

31 Prior assignments

A Bayesian hierarchical model is made up of two components: a structural part described in the previous section, and a quantitative part, which specifies numerical values for all fixed quantities appearing as boxes in the DAG of Fig. 2. The structural component is by far more important in the inferential process; nevertheless care must also be used in the quantitative assignment.

We adopt a pragmatic approach whereby substantial prior information should be used whenever available; otherwise we endorse the employment of weakly informative priors. The latter approach was adopted in this paper.

We now take into considerations two more specific aspects.

A first issue concerns the distribution of the observations a_{gs} described in (1). To account for very large or very low expressions we require that the uniform distributions extend sufficiently far in either direction. In particular we require $\Pr \{ \min\{\delta_g^-, \delta_g^+\} > c_0 \zeta_g \} = 1 - \epsilon$, $\epsilon > 0$, for some suitably large value of c_0 and small value of ϵ (in our case we set $c_0 = 3$ and $\epsilon = 0.01$). Clearly this imposes constraints on the hypercoefficients of the model which can be solved for numerically. Another strategy which was implemented in this paper is to use slice sampling techniques, see [9]. A second issue is related to the clustering structure. In other words when should λ_g^- , and λ_h^- (equivalently π_g^- , and π_h^-) be regarded as sufficiently similar to classify them in the same mixture component? This in turn implies fixing a maximal discrepancy between values of the parameters living in the same cluster, beyond which one should, with high probability, regard them as belonging to distinct clusters.

We concede that in our context it might be hard to make these assignments. We have therefore devised an automatic procedure for fixing a discrepancy value based on the data available. The idea is rather simple and was successfully implemented in [2]: preliminary estimates of λ_g^- , and λ_g^+ are obtained, perhaps based on MLEs. Next the distribution of all pairwise differences of such estimates are computed. Possible values for the discrepancy are then represented by selected quantiles of such distribution, e.g. the median or the third quartile. Clearly higher quantiles, corresponding to higher discrepancies, would induce coarser clustering structures (i.e. a smaller number of components), while the opposite occurs for lower quantiles. In this paper we chose the median.

32 Gene clustering and molecular profiling

Statistical inference with our model is performed using Markov Chain Monte Carlo (MCMC) methods. Specifically, since the number of mixture components k^{λ^-} and k^{λ^+} are random, a Reversible Jump MCMC (RJMCMC) algorithm was employed, see [6], that allows to handle models with varying-dimension parameters.

Defining a complete cycle of iterations over all random quantities involved as a sweep, we ran 200 sweeps for burn-in, followed by 1,000 further sweeps. Each sweep took an average of 20 seconds on a Pentium 4 computer for the first data set and 40 seconds on the second. To reduce serial correlation we sampled 1 every 4 such sweeps, thus effectively considering 250 sweeps for our results. Although the number of sweeps is not high, this does not cause problems as convergence to the stationary distribution occurs rapidly.

Routine informal diagnostic checks were performed on the MCMC output and they did not show signs of aberrant behavior apart for the parameters k^{λ^-} and k^{λ^+} : in this case the split/merge acceptance rate is rather low. We come back to this issue in the Discussion.

Clearly the RJMCMC output allows to estimate all parameters in the model. In particular, the output contains sample values from the posterior distribution of k^{λ^-} and k^{λ^+} which can be used to make inference, e.g. *via* a posterior mode, on the number of mixture components. This is in sharp contrast to [5], wherein it is needed to fix a value for the number of components in advance. Notice that the posterior distribution of the two quantities k^{λ^-} and k^{λ^+} may be quite different. This simply means that the clustering structure for underexpression is different from that for overexpression, and may represent potentially useful information. An allied profitable output concerns the sample values of the allocation variables $z_g^{\lambda^-}$ and $z_g^{\lambda^+}$. This is crucial to recover *partitions* of genes, as we now describe. At each iteration of the algorithm one obtains realizations for $z_g^{\lambda^-}$ and $z_g^{\lambda^+}$ which make up two distinct partitions of genes, namely one for each marginal allocation variable. In order to recover a combined partition, corresponding to a joint allocation, we suggest the following procedure: if two genes belong to the same subset in both partitions, so they will in the joint partition; otherwise each gene will make up a singleton subset. We use the superscript $(\lambda_g^-, \lambda_g^+)$ when referring to the joint partition.

We stress that the above procedure to obtain a joint partition is needed simply because we assigned two separate distributions to the parameters λ_g^- and λ_g^+ . Had we assigned a bivariate mixture prior to the pair $(\lambda_g^-, \lambda_g^+)$, we would have obtained directly a joint partition at each iteration; see the Discussion for further details.

Having identified a partition, perhaps that with the highest posterior probability, the next step involves choosing a representative gene from each subset in the partition. This is yet fairly unexplored territory. Currently we suggest to compute MCMC-based estimates for λ_g^- and λ_g^+ , plot them, fit a normal bivariate kernel, and then choose the gene within the subset that best resembles the modal value of the kernel estimate.

Having obtained the set of representative genes, one for each subset, one can proceed with molecular profiling along the lines described in [5], using the freely available R-based software POE (Probability of Expression) which can be accessed at <http://astor.som.jhmi.edu/poe>.

4. Model validation: POEsimu data set

As mentioned in the Introduction, we consider the data set POEsimu taken from [5] which consists of a 1000x72 matrix of simulated data. The first 50 genes are overexpressed in samples 1 to 24, underexpressed in samples 25 to 33, and normally expressed in samples 33 to 72. Genes 51 to 100 are underexpressed in every third sample, starting with sample 1, and overexpressed in every third sample between samples 1 and 26, beginning with sample 2. Genes 101 to 150 are overexpressed in samples 1 to 18 and underexpressed in samples 37 to 54. Underexpressed, normally expressed and overexpressed data were generated as follows: $a_{gs}|e_{gs} = - \sim N(-4, 2^2)$ $a_{gs}|e_{gs} = 0 \sim N(1, 1)$ $a_{gs}|e_{gs} = + \sim N(4, 2^2)$. It turns out that the probability of under- overexpression for the first 50 genes is 1/9 and 1/3 respectively, for the following 50 genes is 1/3 and 1/9, for genes from 101 to 150 is 1/4 and 1/4 and 0 otherwise. In particular we use a smaller part of this data set considering just the first 350 genes: while this reduction will not effect the underlying simulated structure, it reduces considerably the amount of computational time required to analyze it.

We start by considering the posterior distribution of the number of mixture components k^{λ^-} and k^{λ^+} . They are reported in the first two panels starting from the left of

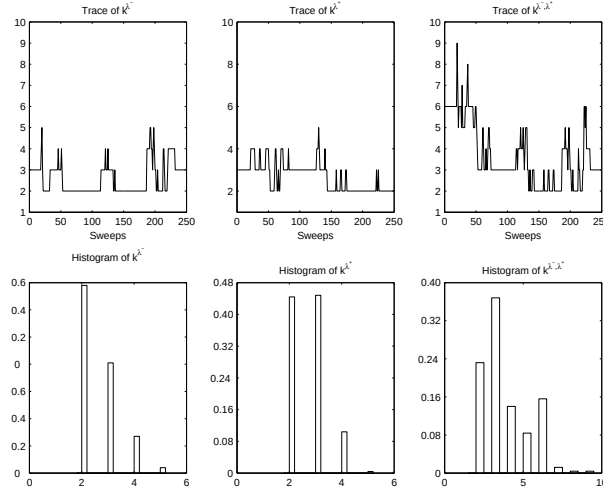


Figure 1: Top row from left to right: trace of k^{λ^-} , k^{λ^+} and k^{λ^-, λ^+} . Bottom row from left to right: histograms of k^{λ^-} , k^{λ^+} and k^{λ^-, λ^+} .

Figure 1: the upper panels describe the trace while the corresponding lower panels summarize the result through the posterior distribution. It appears that most of the posterior probability for k^{λ^-} is concentrated on the values 2 and 3, with 2 being clearly the mode; a similar situation occurs for k^{λ^+} : here however the values 2 and 3 are essentially both modal values. It therefore appears that there exist either two, or three, clusters as far as the probability of under or over expression is concerned. The rightmost panels of Figure 1 contain the same information with respect to the combined partition, which is constructed as described in Subsection 3.2. It appears that the modal value is now 3, with some probability on the value 2.

We now turn to the posterior estimates of the probabilities of under and over expression. The left panel of Figure 2 reports as dots the posterior expectation of π_g^- from our model (1) with superimposed the corresponding values obtained using the POE algorithm of [5] represented as plus signs. Solid bold lines are the true values used in the actual simulation. The right panel contains the same information relative to π_g^+ .

For genes 151 to 350, which are normally expressed i.e. simulated with $\pi_g^- = \pi_g^+ = 0$, it appears that our model estimates these probabilities much better than POE with posterior estimates around 0.01, while those of POE oscillate between 0.05 and 0.10. This is clearly due to the fact that our model separates out this collection of genes from the remaining one in a distinct cluster: therefore estimates are not pulled up, as it happens instead under the POE model because of the exchangeability assumption.

Consider now genes 1 to 150. Here both models show some difficulties in estimating the actual simulated values. Essentially POE performs better for “low” values, while our model does better for higher values. The explanation for this behavior is simple: since POE assumes exchangeability across genes, estimates for the probabilities related to genes 1 to 150 are pulled down by the corresponding estimates for genes 151-350 (which are very close to zero). On the other hand our model, which allows for heterogeneity and relegates genes 151-350 in a single group treats separately genes from 1 to 150 and performs multiple shrinkage. As for the group comprising the first 150 genes, our model essentially identifies either no further subset, or sometime an extra subset. Because of this, unique shrinkage occurs within the group containing genes 1-150 and does not allow

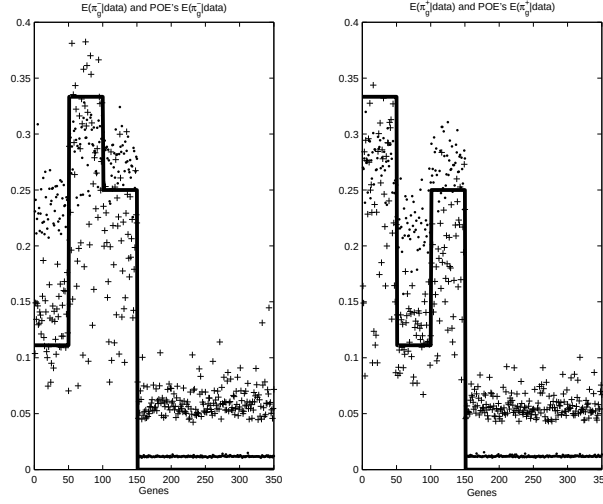


Figure 2: Left panel: posterior estimates of π_g^- . Right panel posterior estimates of π_g^+ . Dot signs represent $E(\pi_g^-|\text{data})$ and $E(\pi_g^+|\text{data})$ respectively obtained using model (1), while plus signs are the posterior estimates using POE algorithm. Bold line: simulated values of π_g^- and π_g^+ .

our model to track closely the true values. This is also evident from Figure 3 which reports the posterior probability of differential expression for each gene-sample combination. Let $p_{gs}^- = \Pr\{e_{gs} = -|\text{data}\}$, $p_{gs}^0 = \Pr\{e_{gs} = 0|\text{data}\}$ and $p_{gs}^+ = \Pr\{e_{gs} = +|\text{data}\}$. The top three panels display the above probabilities for our model, while those obtained under POE are contained in the bottom ones. On the whole it appears that the performance of our model is less noisy and better captures the underlying structure.

5. Application: human lung carcinomas

As recalled in the Introduction this data set is a 460x203 matrix. Of the 203 samples, 17 were normal lung specimens (“NL”); the 186 lung cancer specimens include histologically defined lung adenocarcinomas (127) (“adeno”), squamous cell lung carcinomas (21) (“SQ”), pulmonary carcinoids (20) (“PC”), small-cell lung carcinoma (6) (“SCLC”), while the remaining 12 adenocarcinomas were suspected to be extrapulmonary metastases based on clinical history (“adenoEM”), see [1] for a detailed discussion. Figure 4 reports the posterior expectations of π_g^+ and π_g^- (top panels) as well as the trace and histogram of $k_g^{\lambda^-}$, $k_g^{\lambda^+}$ and $k_g^{\lambda^-, \lambda^+}$ (intermediate and bottom panels). Concerning the trace of $k_g^{\lambda^-}$ and $k_g^{\lambda^+}$ it should be noticed that they do not appear to be satisfactory, as they remain for long spells in the same state. This suggests intervention in the Reversible Jump algorithm, see also the Discussion. On the other hand the trace of $k_g^{\lambda^-, \lambda^+}$ appears more satisfactory since it jumps more often across states: notice that the top mode for the number of classes is 3; however the next mode appears to be 5 which is the number of known biologically specified classes in the data set (including the normal tissues if “adeno” and “adenoEM” groups are considered together). When examining the actual distinct joint partitions we find a huge number of them, although the differences between many pairs of partitions may be quite trivial, so that picking a specific one becomes hard. This suggests an additional strat-

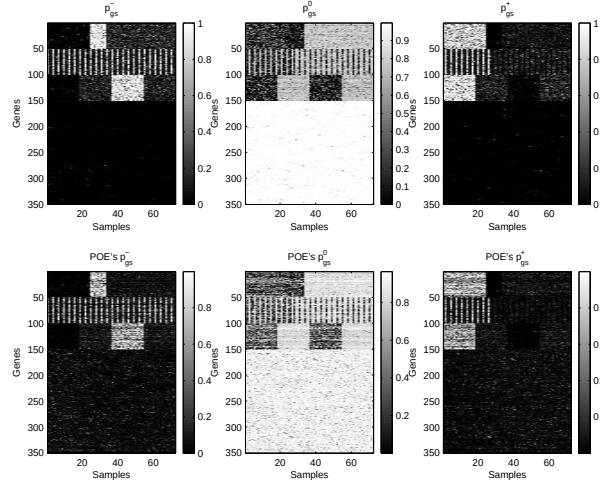


Figure 3: Top row from left to right: pictorial representation of the probability of differential expression p_{gs}^- , p_{gs}^0 and p_{gs}^+ respectively using model (1). Bottom row from left to right: pictorial representation of the probability of differential expression p_{gs}^- , p_{gs}^0 and p_{gs}^+ respectively using POE algorithm.

egy to identify a single partition. Consider for each gene, the mode of $z_g^{\lambda^-, \lambda^+}$, see the top panel of Figure 5. This specifies a unique partition of genes which, interestingly, is made up of 4 subsets, which comes closer to the number of biologically distinct tissues in the whole data set. The scatter plot of $(\text{logit}(E(\pi_g^- | \text{data})), \text{logit}(E(\pi_g^+ | \text{data})))$ is reported in the central panel of Figure 5, with distinct symbols for each of the four subsets: notice that they appear well separated. As a final step, a bivariate normal kernel density estimate is fitted to each cluster above and a representative gene for each cluster is identified as that with values closest to the modal value.

This constitutes the input for the molecular profiling exercise which is depicted in Figure 6. It is apparent from the picture that choosing $k = 4$ and the representative genes in the way described above gives rise to an increasing ability of POE profiling algorithm to discriminate across samples. In particular almost all normal lungs are clustered together and underexpressed in the first two genes, overexpressed in the third and normally expressed in the last $(-, -, +, 0)$, while pulmonary carcinoids are still well clustered together $(+, 0, +, +)$. Small-cell lung carcinomas show a less clear picture although a 4 out of 6 samples are in the same profile group $(0, 0, +, 0)$. Unfortunately for squamous cell lung carcinomas it seems there is not a unique classification pattern: most of them show a normal profile $(0, 0, 0, 0)$, while the remaining are clustered in different profiles (from left to right of Figure 6: $(0, 0, -, 0)$, $(0, 0, 0, -)$, $(-, 0, 0, 0)$, $(0, -, 0, 0)$, $(0, -, 0, -)$). Lung adenocarcinomas belong almost entirely to the normal profile $(0, 0, 0, 0)$ and therefore this group can be thought of as a reference group for the above classification.

6. Discussion

We have developed a Bayesian hierarchical model for gene expression data, extending current work in the area, notably by [5].

The main feature of our work is represented by the prior distribution on the collection

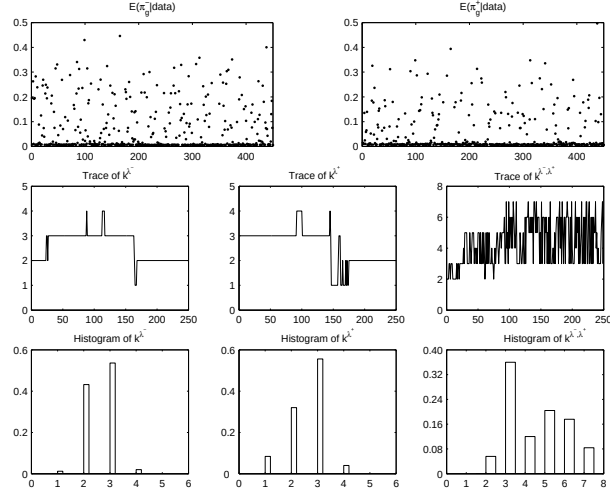


Figure 4: Top row from left to right: posterior estimates of π_g^- and π_g^+ . Central and bottom row from left to right: trace and histogram of k^{λ^-} , k^{λ^+} and k^{λ^-, λ^+} .

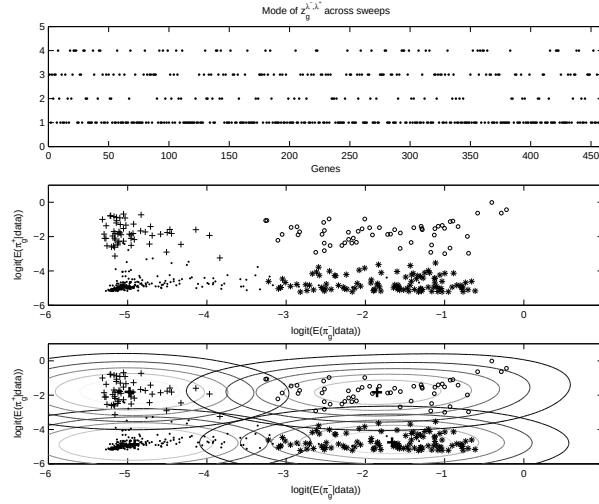


Figure 5: Top panel: mode of the joint allocation variable $z_g^{\lambda^-, \lambda^+}$ across sweeps. Central panel: scatterplot of $(\text{logit}(E(\pi_g^-|\text{data})), \text{logit}(E(\pi_g^+|\text{data})))$, where dot signs are the posterior estimates of (π_g^-, π_g^+) when $z_g^{\lambda^-, \lambda^+} = 1$, plus signs when $z_g^{\lambda^-, \lambda^+} = 2$, star signs when $z_g^{\lambda^-, \lambda^+} = 3$ and finally circle signs when $z_g^{\lambda^-, \lambda^+} = 4$. Bottom panel: contour plot of the bivariate normal kernel density apply to the posterior of π_g^- and π_g^+ conditionally on the joint allocation variable $z_g^{\lambda^-, \lambda^+}$. Bold plus sign corresponds to the mode of the bivariate normal kernel density.

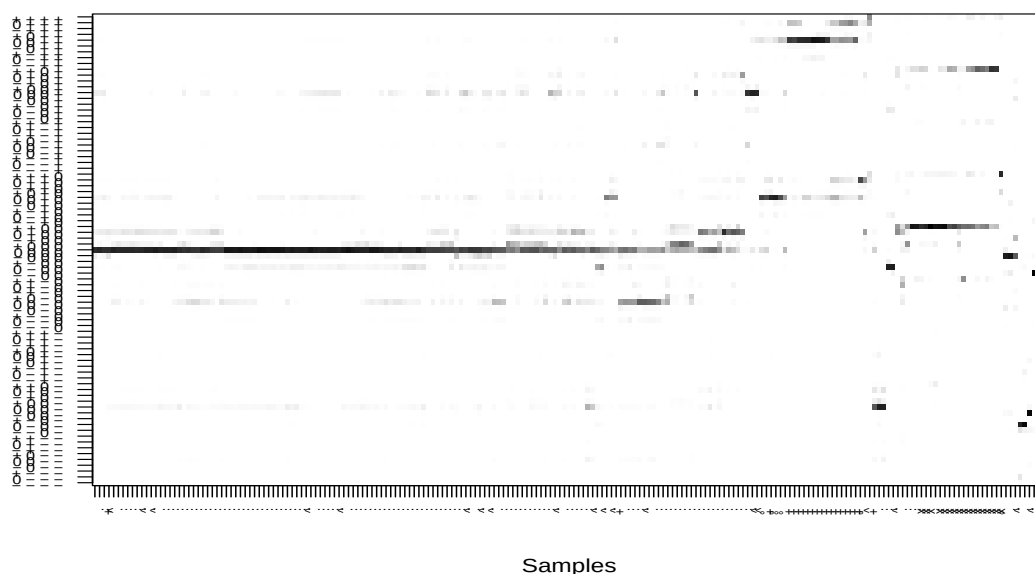


Figure 6: Image plot of molecular profiles by samples of lung cancer data using POE profiling algorithm based on representative genes (R91950, AA482169, AA872020, AA460115). Samples axis: dot signs = “adeno” and “adenoEM”, cross sign = “NL”, circle sign = “SCLC”, triangle sign = “SQ” and plus signs = “PC”.

of probabilities for under- and over expression, which is modelled as a mixture of prior distributions with a random number of components and weights. This allows to account for heterogeneity and induces a distribution on the partitions for genes which is later used for molecular profiling.

Validation of the model on a simulated data set is promising. In particular our model is able to capture very low or very high probabilities of gene expression.

Application of the model to the lung cancer data set described in [1] has identified a partition with 4 clusters, which comes close to the actual biologically known 5 clusters present in the data set (notice however that we only considered approximately 50% of the genes present in the original study). To some extent these clusters also emerge from the ensuing analysis of molecular profiling.

Areas which in our opinion require further work are: i) improving the Reversible Jump MCMC algorithm in order to explore alternative models more efficiently and increase the acceptance rate, possibly along the lines of [7] and [15]; ii) modelling simultaneously the probabilities of under and over expression; iii) summarizing the posterior distribution on the space of partitions in order to identify a representative partition.

References

- [1] Bhattacharjee *et al.* (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences, USA*, **98**, 13790-13795.

- [2] Bottolo, L., Consonni, G., Dellaportas, P. and Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes*, **6**, 25-47.
- [3] Bottolo, L. and Consonni, G. (2003). Hierarchical mixture priors for the analysis of gene expression data in cancer studies. In *Convegno della Società Italiana di Statistica* (Napoli, 9-11 giugno 2003). RCE Edizioni, Napoli, 103-110.
- [4] Consonni, G. and Veronese, P. (1995). A Bayesian method for combining results from several binomial experiments. *J. Am. Statist. Assoc.*, **90**, 935-944.
- [5] Garrett, E.S. and Parmigiani, G. (2003). POE: Statistical Methods for Qualitative Analysis of Gene Expression. In *The Analysis of Gene Expression Data: Methods and Software* (eds. G. Parmigiani, E.S. Garrett, R.A. Irizarry, S.L. Zeger), Chapter 16. Springer: New York.
- [6] Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.
- [7] Green, P.J. and Mira, A. (2001). Delay ed rejection in reversible jump Metropolis-Hastings. *Biometrika*, **88**, 1035-1053.
- [8] Malec, D. and Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, **79**, 593-601.
- [9] Neal, R.M. (2003). Slice sampling (with discussion). *Annals of Statistics*, **31**, 705-767.
- [10] Nobile, A. and Green, P.J. (2000). Bayesian analysis of factorial experiments by mixture modelling. *Biometrika*, **87**, 15-35.
- [11] Parmigiani, G., Garrett, E.S., Anbazhagan, R. and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer (with discussion). *J. Roy. Statistic Soc. B*, **64**, 717-736.
- [12] Parmigiani, G., Garrett, E.S., Irizarry, R.A. and Zeger, S.L. (2003). *The Analysis of Gene Expression Data: Methods and Software*. Springer: New York.
- [13] Quintana, F.A. and Iglesias, P.L. (2003). Bayesian clustering and product partition models. *J. Roy. Statistic Soc. B*, **65**, 557-574.
- [14] Richardson, S. and Green, P.J. (1997). On the Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statistic Soc. B*, **59**, 731-792.
- [15] Scaccia, L. and Green, P. J. (2003). Bayesian growth curves using normal mixtures with nonparametric weights, *J. Comp. Graph. Stat.*, **12**, 308-331.