# Bayesian methods for microarray data

**Alex Lewin and Sylvia Richardson**

Department of Epidemiology and Public Health, Imperial College, Norfolk Place, London W2 1PG, UK

January 11, 2007

Summary

We review the use of Bayesian methods for analyzing gene expression data. We focus on methods which select groups of genes on the basis of their expression in RNA samples derived under different experimental conditions. We first describe Bayesian methods for estimating gene expression level from the intensity measurements obtained from analysis of microarray images. We next discuss the issues involved in assessing differential gene expression between two conditions at a time, including models for classifying the genes as differentially expressed or not. In the last two sections, we present models for grouping gene expression profiles over different experimental conditions, in order to find co-expressed genes, and multivariate models for finding gene signatures, *i.e.* for selecting a parsimonious group of genes that discriminate between entities such as subtypes of disease.

**Keywords:** Bayesian hierarchical models, differential expression, profile clustering, mixture models, gene selection, shrinkage priors, models for cDNA arrays, models for oligonucleotide arrays, gene expression.

**Related Chapters:** hsg006, hsg007

## 1. Introduction

High throughput technologies such as DNA microarrays have emerged over the last 5-10 years as one of the key source of information for functional genomics. Microarrays permit researchers to capture one of the fundamental process in molecular biology, the transcription process from genes into mRNA (messenger RNA), that will be subsequently translated to form proteins. This process is called gene expression. By quantifying the amount of transcription, microarrays allow the identification of the genes that are expressed in different types of cells, different tissues and to understand the cellular processes in which they intervene, thus giving a unique insight into the function of genes. However, transforming the huge quantity of data which is currently produced

*email:* a.m.lewin@imperial.ac.uk

in experiments which involve microarrays into useful knowledge for system biology is not trivial and research into ways of interpreting this rich body of data has become an active area, involving statisticians, machine learning and computer scientists.

Microarrays generally contain thousands of spots (or probes) at each of which a particular gene or sequence is represented. In effect a microarray experiment represents data comparable to that obtained by performing tens of thousands of 'experiments' of a similar type in parallel. The 'experiments' on a given array will share certain characteristics related to the manufacturing process of the particular array used and the extraction and handling of the biological sample hybridized to the array. The interest is in comparing expression levels between arrays with samples from different biological conditions of interest (e.g. cancerous against non-cancerous cells) and the challenge is identifying differences that are related to the biology of the samples rather than to technical experimental variation.

Many of the characteristic features of experiments involving microarrays render them particularly well suited to a flexible modeling strategy within the Bayesian framework. The aim of this chapter is to focus on the unique contribution that Bayesian methods offer and highlight this by discussing in detail the steps taken for modeling the variability in gene expression data at several levels.

The framework of Bayesian hierarchical modeling refers to a generic model building strategy in which unobserved quantities are organized into a small number of discrete levels with logically distinct and scientifically interpretable functions and probabilistic relationships between them that capture inherent features of the data. It is of course important to perform some basic exploration and visualization of the data before formulating complex models; *see hsg006 for examples*. The hierarchy of levels makes it particularly suitable for modeling gene expression data, which arises from a number of processes and is affected by many sources of variability. We shall see in the next sections an approach to modeling these different sources of variability using fixed effects, random effects and distributional assumptions.

One of the most important aspects of Bayesian hierarchical modeling as regards microarray data is the sharing of information across parallel units. For example, gene expression experiments used by biologists to study fundamental processes of activation /suppression frequently involve genetically modified animals or specific cell lines, and such experiments are typically carried out only with a small number of biological samples. It is clear that this amount of replication makes standard estimates of gene variability unstable. By assuming exchangeability across the genes, inference is strengthened by borrowing information from comparable units.

Another strength of the Bayesian framework is the propagation of uncertainty through the model. Due to the many sources of systematic variation between arrays and samples, gene expression data is often processed through a series of steps, each

time estimating and subtracting effects in order to make the arrays comparable. The end result of this process can be over-confident inference, as the uncertainty associated with each step is ignored. In a Bayesian model it is straightforward to include each of these effects simultaneously, thus retaining the correct level of uncertainty on the final estimates. Further, when including in the model structured priors that are associated with classification, e.g. mixture priors, estimates of uncertainty of the classification are obtained along with the fit of the model.

The field of microarray data analysis is very large, and it is not possible to cover all aspects in this chapter. In particular, we do not discuss methods for estimating graphical models and Bayesian networks that are aimed at understanding regulatory networks or metabolic pathways. There is a large literature on this subject, see for example a number of chapters in Do et al. (2006), and references therein. *See also* chapter hsg009, and hsg006 and hsg007 for non-Bayesian approaches.

We focus on methods which select groups of genes on the basis of their expression in RNA samples derived under different experimental conditions. We start, in the next section, by looking at two Bayesian methods for estimating gene expression level from the intensity measurements obtained from analysis of microarray images. This section includes some discussion of the steps involved in a microarray experiment. Section 3 discusses the issues involved in assessing differential gene expression between two conditions at a time. There is an extensive literature on this topic. Our presentation is divided into sections on normalization, gene variability and models for classifying the genes as differentially expressed or not. We include a brief explanation of mixture models, and some discussion of different decision rules used to choose the lists of genes considered to be differentially expressed. In Section 4 we present a range of models for grouping gene expression profiles, which are vectors of gene expression over several different experimental conditions, for ordered samples (e.g. time-course data) and for samples with no ordering. Finally Section 5 reviews the current work on multivariate methods for finding subsets of genes that can predict and classify phenotypes. We focus on discussing variable selection methods that have been used to find, for example, so-called 'gene signature' of different subtypes of disease, as well as Bayesian shrinkage methods. In both cases, the emphasis is on parsimony of the multivariate model in order to enhance interpretation. Inference in Bayesian models is made in either an empirical or fully Bayesian framework. In the case of fully Bayesian models, Markov Chain Monte Carlo (MCMC) is usually used to estimate the posterior distribution of the model. We do not go into details of these procedures, except in the case of non-standard algorithms. Table 1 gives URLs for software for the models discussed in this chapter.

A note on notation: we use $x_{g...}$ for gene expression measures used as data in Sections 3 and onward, rather than the more usual $y_{g...}$. This is to allow the standard formulation

for the variable selection models in Section 5, where $y$ stands for the outcome and $x$ stands for the variables (here genes). Throughout the chapter, $p$ stands for the number of genes and $n$ stands for the number of samples or experimental conditions.

## 2. Extracting Signal from Observed Intensities

The output from a microarray experiment starts with the image of an array (*see* hsg006). This image must be gridded and segmented into spots, which are the basis for inference about the sequences present in the sample of interest. There has been some work on Bayesian methods for image analysis (see for example Ceccarelli & Antoniol (2006); Gottardo et al. (2006a) and references therein). This work is beyond the scope of this chapter. The methods we describe in this section start with an intensity measurement for each pixel on an array, found from the image analysis, and use these to construct a summary measure of the amount of RNA present in the sample for each gene of interest.

There are two main types of microarrays in use, spotted or cDNA arrays, which are usually two-color, and oligonucleotide arrays, which are one-color. Spotted arrays are microscopic slides onto which long strands of cDNA are fixed in a regular grid layout. Each "spot" on the array will then contain millions of copies of the same (known) sequence of cDNA (called probes). One sequence corresponds to one gene, or EST (expressed sequence tag). In order to find out what sequences are present in a sample of mRNA, a sample of cDNA (the target) is produced from the mRNA by reverse transcription, and fluorescently labeled. This sample is introduced onto the array, where hybridization reactions take place between sequences of cDNA which match in the sample and on the array. The array is then washed to remove target cDNA which has not hybridized to the array, and scanned to detect the fluorescent labels of the cDNA strands which have hybridized. For two-color arrays, two samples of cDNA, labeled with dyes of two different frequencies (Cy3 and Cy5), are put on the array. The two samples are usually from different mRNA samples, enabling the concentrations of particular sequences to be compared between the two samples.

There are two particularly important statistical issues arising from the process of the microarray experiment. Firstly the Cy3 dye tends to appear brighter than the Cy5, due to differences in the reaction with the cDNA and different responses to the laser used in the scanning process. This leads to the so-called dye effect. In addition, the known cDNA sequences are printed onto the array using a number of spotting pins. The different pins may deliver slightly different amounts of cDNA to the array, thus there can be a systematic effect between spots printed with different pins. This is known as the print-tip effect.

Oligonucleotide arrays work in a similar way. There are three main differences from spotted arrays (from a data analysis point of view), the first being that just one sample is hybridized to each array, and thus only one dye is used, so there is no dye effect.

The second is that the same printing head is used for all spots, thus there is no print-tip effect. The third difference is that the probe sequences fixed to oligonucleotide arrays are shorter than those used in spotted arrays. For this reason, several probes (of different sequences) are used to detect one gene or EST. Spots on the array come in pairs: one containing the "perfect match" probe (PM) and an adjacent spot containing the "mis-match" probe (MM). The perfect match probe is a strand of cDNA which has the sequence of interest. The mis-match probe has the same sequence except for the central nucleotide, which is different.

The reason for this is to provide a measure of cross-hybridization: target cDNA having a similar but not identical sequence to the PM probe may hybridize to the PM, contaminating the signal. The idea with the MM probe is that these mis-matched target cDNAs would also hybridize to the MM probe, but the true matches to the PM would only hybridize to the PM and not to the MM. Thus the amount of cDNA with the exact same sequence as the PM could be estimated by subtracting the MM signal from the PM signal. In reality, target cDNA with exact match to PM also hybridizes partially to the MM, so estimating the correct amount is a complicated process (see Section 2.2).

### 2.1 *Spotted cDNA Arrays*

Most work with data from spotted arrays takes the ratio of the Cy3 and Cy5 intensities as a measure of the *relative* expression of each gene in the two RNA samples. These can be used in a fairly straightforward way to compare gene expression under different experimental conditions. Care must be taken to account for the dye and print-tip effects. These effects are often included as part of the model for differential expression, as will be seen in Section 3.1.

Here we discuss a Bayesian model developed by Frigessi et al. (2005) for obtaining estimates of concentrations of RNA from two-color arrays. This model includes the dye and print-tip effects, along with other aspects of the experimental process, usually absorbed into empirical normalization methods. The idea is to follow through the process which the RNA molecules undergo in order to be detected as hybridized to the array. The steps in this process are modeled with a hierarchical model.

The principal data used in the model are the intensity measurements in each pixel $j$ on each array $a$. These are denoted $L_{j,s}^{i,a}$, where $s$ labels spots to which the pixel belongs and $i$ labels the particular RNA sample hybridized to the array. The background intensity is assumed to have been subtracted as part of the image analysis. The quantity of interest to estimate is the concentration of RNA (in molecules per unit weight) for gene $g$ in sample $i$, denoted by $K_g^i$. The main steps relating this concentration to the observed intensities are hybridization, washing and scanning.

In order to model the scanning process, consider the number of molecules $J_s^{i,a}$ from sample $i$ left on spot $s$ after hybridization and washing. The observations which con-

tribute directly to $J_s^{i,a}$ are the intensities for the pixels in the spots corresponding to that gene; for a given sample and array $L_{j,s}^{i,a} \propto J_s^{i,a}/n_s^a$ where $n_s^a$ is the number of pixels in spot $s$. The constant of proportionality is $2^{f_{dye}.PMT^{i,a}}\alpha_{dye}$, where $PMT^{i,a}$ is the voltage used in scanning and $f_{dye}$ is the scanner amplification factor, both known. The factor $\alpha_{dye}$ accounts for the dye effect (this is estimated as part of the model). With this expected relation between intensity and number of molecules on a spot in place, the intensity measurements are modeled as coming from a Normal distribution,

$$L_{j,s}^{i,a} \sim N(2^{f_{dye}.PMT^{i,a}}\alpha_{dye}J_s^{i,a}/n_s^a, (\sigma_s^{i,a})^2). \tag{1}$$

The variance $(\sigma_s^{i,a})^2$ is estimated from the sample variance of the intensities, and treated as fixed in the analysis.

In the second level of the hierarchical model, the prior for $J_s^{i,a}$ depends on the concentrations $K_{g(s)}^i$, where $g(s)$ is the gene corresponding to spot $s$, and also on other parameters for various effects encountered in the hybridization step. This step is treated as a selection process where each molecule of gene $g(s)$ has an equal chance of hybridizing to and remaining on spot $s$. Thus the number of molecules has a Binomial distribution:

$$J_s^{i,a} \sim \text{Bin}(cn_s^a q^{i,a} K_{g(s)}^i, p_s^{i,a}) \tag{2}$$

where $q^{i,a}$ is the total weight of sample $i$ and $c$ is a hybridization factor (estimated in a calibration experiment). The probability of hybridization $p_s^{i,a}$ has several contributions:

$$p_s^{i,a} = L^{-1}(\gamma_0 + \gamma_{g(s,a)} + \gamma P^i + \beta X_s^a) \tag{3}$$

$P^i$ is a measure of the purity of sample $i$, and $X_s^a$ contains the covariates for spot $s$ on array $a$: probe length, probe quality, print-tip and array. Various link functions are used for $L$. To ensure identifiability, several arrays must be analyzed together. The main object of inference is $K_g^i$ and a purposely designed MCMC algorithm is used to get posterior samples.

### 2.2 *Oligonucleotide Arrays*

In contrast to cDNA arrays, the intensity measurements for spots on oligonucleotide arrays cannot be combined in a simple manner to form gene expression measurements. The simplest way to use the PM and MM measurements would be to use PM-MM as a measure of expression. However there is a problem with this, as very often the MM intensity is larger than the PM intensity (see Figure 1, top row). There has been much work in the microarray literature on methods to deal with this phenomenon. Here we present Bayesian models developed to model the PM and MM intensities in order to produce measures of gene expression.

Hein et al. (2005) present a fully Bayesian hierarchical model, estimated by MCMC, for obtaining gene expression measures for each gene in each experimental condition. If there are replicate samples for a condition (including biological replicates) the model produces an estimate for that condition, rather than separate estimates for each replicate. On the other hand, by using the variability of the probe sets for each gene, the model can be used with a single array for each condition and meaningful comparison between conditions without any replicates can be achieved (Hein & Richardson, 2006).

The data used for the model in Hein et al. (2005) are the perfect and mis-match intensities for probe-pair $j$ of gene $g$ in replicate $r$ of condition $c$, denoted by $PM_{gjcr}$ and $MM_{gjcr}$. Each $c, r$ pair corresponds to one physical array. The intensity observed at a PM probe is assumed to be the result of hybridization partly of fragments that perfectly match the probe (specific hybridization, signal: $S_{gjcr}$) and partly by fragments that do not perfectly match the probe (non-specific hybridization: $H_{gjcr}$). A similar pattern is assumed for the MM probe, with only a fraction $\phi$ of the signal $S_{gjcr}$ binding. Both specific and non-specific hybridization are estimated separately for each gene and probe. To account for the possibility of the MM being bigger than the PM the model includes an additive error on the normal scale.

$$\begin{aligned} \mathrm{PM}_{gjcr} &\sim N(S_{gjcr} + H_{gjcr}, \tau^2_{cr}) \\ \mathrm{MM}_{gjcr} &\sim N(\phi S_{gjcr} + H_{gjcr}, \tau^2_{cr}) \end{aligned} \tag{4}$$

At the next level of the model estimates for the specific hybridization for each gene in each condition $\mu_{gc}$ are obtained, averaging across probes and replicates. These are the final measures of interest. The non-specific hybridization is modeled with an array-wide distribution (indexed by $c$ and $r$).

$$\begin{aligned} \log(S_{gjcr} + 1) &\sim TN(\mu_{gc}, \sigma^2_{gc}) \\ \log(H_{gjcr} + 1) &\sim TN(\lambda_{cr}, \eta^2_{cr}) \end{aligned} \tag{5}$$

Here $TN$ stands for the Normal distribution truncated at zero on the left. This and the shifted log function allow the hybridization signals to be zero.

Array-specific parameters (those indexed by $c, r$) are given independent priors. The gene-specific variances $\sigma^2_{gc}$ are modeled exchangeably, to share information across the genes and stabilize the variance estimates.

Hein et al. (2005) fit this model to the GeneLogic spike-in data set at http://www.genelogic.com/media/studies/index.cfm. This is a widely-used data set consisting of gene expression measurements for replicate samples of cRNA from an acute myeloid leukemia (AML) tumor cell line, with eleven exogenous cRNAs spiked into each sample at a different known concentration in each sample. Each sample was hybridized on one array, thus all measurements for spike in genes on a particular array

correspond to the same cRNA concentration. The top row of Figure 1 shows the PM and MM measurements for four of the spiked in genes (all from the same array). It can be seen that the measurements of different probes within a gene vary widely, due to the different sequences being detected.

The lower panel shows results for the same four genes, when the model is fit to the single array. The plots show the posterior estimates of $S_{gjcr}$ compared with $log(PM_{gjcr} - MM_{gjcr})$. It can be seen that the posterior estimates get more precise for larger PM-MM, i.e. for probes with high specific hybridization, and consequently that the posterior credibility intervals for the $\mu_c$ are reduced. For probes with large MM, the estimates of $S_{gjcr}$ are drawn towards those for the rest of the probes for that gene.

## 3. Differential Expression

One of the most widely-studied problems in microarray analysis is that of differential gene expression between two experimental conditions, for example between knock-out and wildtype animals, or between cases and controls. Most work in this area starts with the gene expression measures for each gene on each array, or the log ratios of expression under two conditions. Expression measures have been observed to have increasing variability with increasing value (Schadt et al., 2000), so they are often modeled on the log scale. Sometimes a shifted log transform is used, as for example in Gottardo et al. (2006b). *Chapter hsg006* discusses many transformation used in the literature.

Many models that have been developed for differential expression can be written as a linear model for the log expression level $x_{gcr}$ for gene $g$, condition $c = 1, 2$ and replicate array $r$:

$$x_{gcr} = \mu_{gc} + \gamma_{cr} + \epsilon_{gcr} \tag{6}$$

where $\mu_{gc}$ represents the level of expression of gene $g$ for condition $c$, $\gamma_{cr}$ is a normalization term for the array containing the replicate $r$ sample of condition $c$, and $\epsilon_{gcr}$ is the residual.

Not all models we discuss can be fitted exactly into this format, for example Newton et al. (2004) and Kendziorski et al. (2003) use the Gamma distribution to model gene variability and so their models do not quite fit into the linear framework. However, they still involve parameters corresponding to the same biological quantities. In addition, the vast majority of models can be fitted into the linear framework, and thus it is useful to give these equations, in an attempt to clarify where models differ or otherwise. We will indicate in the text where models do not use the linear formulation (6).

The parameters of interest are the $\mu_{gc}$. Before we discuss how these are modeled, we look at the normalization and error terms.

### 3.1 *Normalization*

Microarray data show systematic differences between expression levels found on different arrays (e.g. Schadt et al., 2000). Some of these differences are due to dye and print-tip effects discussed in Section 2.1. This may be taken into account when analysing data at lower levels, but generally empirical differences between arrays are still found for the gene expression values. Often the systematic effect is such that there is non-linear relationship between the expression levels on different arrays.

Much work has been done in the classical statistical literature on different methods of accounting for these systematic non-linear differences (normalizing). These usually involve a transformation of the data before it is analysed with another method. Most work on Bayesian models for gene expression has also assumed that this process has been done beforehand. Bayesian models incorporating normalization include those proposed by Parmigiani et al. (2002) and Gottardo et al. (2006b). Both of these include a constant term in a linear model, estimated in a fully Bayesian manner.

Bhattacharjee et al. (2004) and Lewin et al. (2006) model normalization as a non-linear function of expression level. Bhattacharjee et al. (2004) use a normalization term $\gamma_{gcr}$ which is modeled as a piece-wise linear function of gene expression level. Due to marginalization over the joint posterior, posterior estimates of $\gamma_{gcr}$ will be reasonably smooth functions of expression level.

Lewin et al. (2006) propose a model starting with that given in Equation 6, but for which the normalization term has an additional gene index: $\gamma_{gcr} = f(\mu_{gc})$ where the function $f$ is a quadratic spline. They show that transforming the data first rather than modelling the normalization simultaneously with the other unknown quantities can introduce bias, as the gene expression levels $\mu_{gc}$ have to be estimated and thus have variability, as in measurement error problems (Carroll et al., 1995). Figure 2 shows the posterior mean array effects $\gamma_{gcr}$ as a function of expression level for a group of three arrays hybridized to cDNA from wildtype mice, as presented in Lewin et al. (2006).

### 3.2 *Gene variability*

There are many sources of variation in gene expression data. It is possible to put replicate RNA samples from the same individual on different microarrays, but this is usually considered unnecessary as it has been observed that these so-called technical replicates show very high correlation. More usually different arrays are hybridized with samples taken from different individuals. Thus the variability incorporated in the error term $\epsilon_{gcr}$ in Equation 6 represents the biological variability. It is generally accepted that different genes show different levels of biological variability, thus parameters in the distributions for the errors will depend on the gene index.

Several Bayesian models in the literature assume Normal errors (Lönnstedt & Speed, 2003; Baldi & Long, 2001; Bhattacharjee et al., 2004; Lewin et al., 2006). Gottardo et al. (2006c) use a t-distribution (bi-variate for cDNA data) to accommodate more

outlying data points. Newton et al. (2001) and Newton et al. (2004) give the data a Gamma likelihood rather than the lognormal implied by Equation 6. Simple model-checking techniques suggest the Gamma and lognormal families are equally suitable for gene expression data.

Since the numbers of individuals for each experimental condition is often small, independent estimates of gene variance parameters would be unstable. Therefore gene variances $\sigma^2_{gc}$ are usually shrunk, by assuming exchangeability across genes (and sometimes conditions). Both empirical Bayes (Lönnstedt & Speed, 2003) and fully Bayesian methods (Lewin et al., 2006; Gottardo et al., 2006c) relying on MCMC algorithms for inference have been used. Rather than allowing a separate variance for each gene, Bhattacharjee et al. (2004) allow gene variances to take one of three values, estimated as part of the model, as an alternative way of sharing information across genes. Baldi & Long (2001) allow gene variances to depend on expression level, by making the variances exchangeable amongst genes with similar expression levels (defined by a window on the expression level) and estimating these using empirical Bayes methods.

### 3.3 *Expression levels*

It is useful to write the expression levels in two experimental conditions as

$$\begin{aligned} \mu_{g1} &= \alpha_g - \delta_g/2 \\ \mu_{g2} &= \alpha_g + \delta_g/2 \end{aligned} \tag{7}$$

where $\alpha_g$ represents the overall expression level for gene $g$ and $\delta_g$ represents the log differential expression. For two-color arrays the data can be given as log fold changes between the conditions (the data is paired) and in that case there is no $\alpha_g$ parameter. When the data is given separately for the two conditions, $\alpha_g$ must be modeled. It is usually treated as a fixed effect, so no information is shared between genes for this parameter.

The fold change parameter $\delta_g$ can also be given an unstructured prior (e.g Baldi & Long, 2001; Bhattacharjee et al., 2004; Lewin et al., 2006), however many people choose to use mixture models to classify genes as differentially expressed or not. Usually this means putting a mixture prior on some measure of the difference between expression levels in the two experimental conditions. The mixture models can be classified into two groups: those which put a mixture prior on the model parameter $\delta_g$, and those which model the data directly as a mixture. These are not intrinsically different, as the parameters $\delta_g$ could be integrated out to give a mixture model on the data, but it is convenient to describe the models separately.

A finite mixture distribution for a quantity $\Delta_g$ is a weighted sum of probability

distributions ,

$$\Delta_g \sim \sum_{k=0}^{K-1} w_k f_k(\phi_k) \tag{8}$$

where the weights sum to one ($\sum_{k=0}^{K-1} w_k = 1$). Each mixture component has a certain distribution $f_k$, with parameters $\phi_k$. The weight $w_k$ represents the probability of $\Delta_g$ being assigned to mixture component $k$. In the context of differential expression, most mixture models used consist of two components ($K = 2$), one of which ($f_0$) can be thought of as representing the "null hypothesis" of there being no differential expression. The second component corresponds to the alternative hypothesis that there is differential expression. Of course it is not necessary to see the model in terms of hypothesis testing; in the Bayesian framework it is a straightforward procedure to classify each gene into one or other of the mixture components. This is usually done using the posterior probability of component membership (see Section 3.4 for details).

One of the earliest mixture models used in gene expression analysis was that of Efron et al. (2001). In this model $\Delta_g$ in Equation 8 is a regularized t-statistic $t_g$, one for each gene.

$$t_g \sim w_0 f_0 + w_1 f_1 \tag{9}$$

The densities of the mixture components are estimated non-parametrically using standard kernel density procedures. Regularized t-statistics are calculated using expression data from the same experimental condition, to provide an estimate of the null component $f_0$. An estimate of $w_0$ (which represents the proportion of genes in the null, or not differentially expressed) is obtained using Empirical Bayes methods. The whole mixture distribution ($w_0 f_0 + w_1 f_1$) can be estimated using all the $t_g$. Thus the second component $f_1$ can be inferred.

A fully Bayesian version of this model has been discussed by Do et al. (2005). In this work, the framework of Dirichlet Process Mixtures (DPM) is used to formulate a prior probability model for the distributions $f_0$ and $f_1$. A DPM model, characterized by a base measure $G^*$, a scalar parameter $\alpha$, and a mixing kernel to be specified, is one of the most popular nonparametric Bayesian models in reason of the simplicity of its representation and MCMC implementation (Escobar & West, 1995; Walker et al., 1999). Do et al. (2005) choose base measures $G_0^* \sim N(0, \tau^2)$ and $G_1^* \sim \frac{1}{2}N(-b, \tau^2) + \frac{1}{2}N(b, \tau^2)$ for $f_0$ and $f_1$ respectively and Gaussian mixing kernels with common variance parameter $\sigma^2$. The specification of $G_1^*$ reflects the prior belief that DE in either direction is equally likely, in the absence of more specific prior information. Using the stick-breaking construction of DP (Sethurman, 1994), leads to a useful representation of $f_k, k = 0, 1$ as an infinite

mixture of normals:

$$f_k = \sum_{h=1}^{\infty} p_{hk} N(\mu_{hk}, \sigma^2)$$

with $\mu_{hk} \overset{i.i.d}{\sim} G_k^*, k = 0, 1$ and the weights following the stick-breaking structure: $p_{hk} = U_h \prod_{j<h}(1-U_j)$ with $U_h \overset{i.i.d}{\sim} Beta(1, \alpha)$. In Do et al. (2005) all the model parameters are given hyperprior distributions, conjugate inverse Gamma for $\tau^2$ and $\sigma^2$ and conjugate normal for $b$, $\alpha$ is fixed at 1 and $w_0$ is given either a Beta prior or a Uniform prior away from 0. As in Efron et al. (2001), within-condition data differences are used to estimate $f_0$, while between-condition differences are modeled as arising from the mixture defined in (9). Figure 3 shows posterior estimates of $f_0$, $f_1$ and $f \equiv w_0 f_0 + w_1 f_1$ for the Alon colon cancer data set (Alon et al., 1999) as analyzed in Do et al. (2005). This is a data set of gene expression measurements for 2000 genes in 62 tissue samples (40 tumours and 22 normal samples). The density $f_1$ is bimodal, showing that there are genes which are expressed more in tumours than normal samples, and genes expressed more in normal samples. The estimate of the proportion of differentially expressed genes was around 1% in this data set. The performance of this model will depend on the number of within-replicate differences that are used to calibrate $f_0$ and on the information introduced in the hyperprior specification. When there are only a few replicates, the mixture might be close to non-identifiability.

Broët et al. (2002) suggest another model using a mixture at the data level to classify genes. Here the data is first transformed with a linear model to produce normalized log fold changes $d_g$. The $d_g$ are modeled using a fully Bayesian mixture of Normals which includes estimation of the proportion of differentially expressed genes (the weights in the mixture). The number of components in the mixture $K$ is not restricted to 2. There is still just one component representing the null, but several representing differentially expressed genes. This allows grouping of genes into different levels of differential expression. In fact $K$ is not fixed in this model, but estimated, in a fully Bayesian way, using the split and merge algorithm for mixtures with an unknown number of components introduced in Richardson & Green (1997).

When mixture distributions are put on parameters of the model (prior) rather than on the data (likelihood), care must be taken to ensure identifiability of the parameters of the mixture components. A common choice is to make the null component a point mass. This corresponds to testing the null hypothesis $\delta_g = 0$ versus the two-sided alternative.

Lönnstedt & Speed (2003), Lin et al. (2003) and Smyth (2004) use mixture priors on the parameter $\delta_g$ representing difference between conditions. Lönnstedt & Speed (2003) use a mixture of a point mass at zero and a conjugate Normal prior on the $\delta_g$. Smyth (2004) uses the same mixture model, but on data which has first been

transformed using a linear model similar to that in Equation 6 but using a robust estimation method, to obtain log fold changes. These two models are estimated using Empirical Bayes methods. The proportion of true nulls is not estimated, thus these methods produce a ranking of the genes rather than an actual estimate of how many genes are differentially expressed.

Rather than putting the mixture directly on the $\delta_g$, Newton et al. (2004) propose a mixture prior on the pair of parameters $\mu_{g1}, \mu_{g2}$. Their likelihood is Gamma, but the $\mu_{gc}$ still represent mean expression in the two conditions. One component of the mixture has $\mu_{g1} = \mu_{g2}$ drawn from one distribution, the other has $\mu_{g1}, \mu_{g2}$ drawn from two separate distributions. These distributions are estimated non-parametrically, estimated using an EM algorithm. Gottardo et al. (2006c) has a similar mixture on the pair $\mu_{g1}, \mu_{g2}$, this time using Normal priors, and a fully Bayesian estimation method, including estimating the proportion of differentially expressed genes. Reilly et al. (2003) has a model with a similar structure, but in addition incorporates prior information about certain genes being controls (and therefore not differentially expressed).

An early model for differential expression which does not employ a mixture model on the difference between the two conditions is that of Ibrahim et al. (2002). They model the data *in each condition* as coming from a mixture of a point mass and a log Normal distribution, the point mass representing the threshold for genes to be un-expressed. A measure for differential expression is formed from the ratio of expectation of expression in the two conditions.

As a final comment, note that finding differentially expressed genes can also be cast in a multivariate framework. This approach was adopted by Ishwaran & Rao (2003) who use multivariate shrinkage effected via a continuous version of the spike and slab variable selection model (see Section 5.1 for a discussion of variable selection approaches). They propose to detect differentially expressed genes by formulating the problem as a linear regression. They then use a multivariate shrinkage approach to find posterior means of the differentially expressed parameters and finally they compare these values to percentiles of a standard normal distribution (with a scaling coefficient) in order to select differentially expressed genes.

### 3.4 *Classifying genes as differentially expressed*

In differential expression problems, the aim is to produce a list of genes which are considered to be differentially expressed between the different experimental conditions. A decision rule is used to classify genes as either differentially expressed (DE) or not (non-DE). In Bayesian models this will either be based on the value of some model parameter (usually the posterior mean), or else on posterior probabilities of some criterion in the model, for example of being classified into a certain mixture component or of some parameter being above a certain threshold.

*Models with mixture priors for fold changes*

In the fully Bayesian mixture models described above, decisions are usually made using the posterior probabilities of a gene being allocated to the different mixture components. The mixture example given in Equation 8 can also be written as

$$\begin{aligned} \Delta_g|z_g &\sim w_{z_g} f_{z_g}(\phi_{z_g}) \\ \mathbb{P}(z_g = k) &= w_k \end{aligned} \tag{10}$$

where the $z_g$ are allocation parameters which label the mixture component to which gene $g$ is assigned. The posterior probability of gene $g$ being in component $k$ is $\mathbb{P}(z_g = k|\mathbf{x})$.

Defining a loss function enables one to form the decision rule. First, denote the set of genes declared to be DE by $S_1$ and the set of genes called non-DE by $S_0$. In the two-component mixture models, since there are two possible classifications for each gene, there are two possible penalties for mis-classification, one for false positives, one for false negatives. If the ratio of these two penalties is $\lambda$, the same for all genes, the loss function is proportional to

$$L \propto \sum_{g \in S_0} \mathbb{P}(z_g \neq 0|\mathbf{x}) + \lambda \sum_{g \in S_1} \mathbb{P}(z_g = 0|\mathbf{x}) \tag{11}$$

This is minimized by defining $S_0$ as the set of genes for which $\mathbb{P}(z_g = 0|\mathbf{x}) \geq 1/(1+\lambda)$, i.e. genes are classified using a threshold on the posterior probabilities of classification in the mixture. Müller et al. (2007) discuss different possible loss functions and the decision rules they lead to.

The posterior probabilities can also be used to obtain an estimate of the false discovery rate, which is the ratio of false positives to total declared positives:

$$\hat{FDR} = \frac{1}{|S_1|} \sum_{g \in S_1} \mathbb{P}(z_g = 0|\mathbf{x}) \tag{12}$$

(see Newton et al., 2004; Broët et al., 2004; Müller et al., 2007). An estimate of the false non-discovery rate (ratio of false negatives to total negatives) can be defined similarly. The false discovery rate is widely used in classical statistical analysis of gene expression data (*see hsg006 and hsg007*). It is useful to be able to give this estimate when comparing with different analysis methods and it has generally be found in simulation studies that (12) gives quite accurate estimates of the true FDR.

For mixtures of more than two components, one may consider different rules. The most obvious would be to assign genes to the component with highest probability, i.e. gene $g$ is assigned to component $k = \max_{k'} \mathbb{P}(z_g = k'|\mathbf{x})$. However when there are more than two components, this can lead to genes being declared DE (in a particular component) when their posterior probability of being classified into that component is

low. For example with 4 components, a gene which has almost equal probability of being classified in all components can be declared DE (into the best component for that gene) with posterior probability of 0.26 of being in the that component. An alternative, more conservative, suggestion would be to classify into one of the components representing DE only those genes for which the corresponding posterior probability of belonging to that component is above a set threshold, e.g. 50% or higher, and otherwise the genes are classified into the null. Again, evaluating the associated FDR of such rules will guide the choice of appropriate thresholds. Such a rule was used in a related context, that of modelling DNA copy number changes (gains or losses) in comparative genomic hybridization experiments by a spatially structured mixture model with 3 components (gain, loss, normal) in Broët & Richardson (2006). For typical noise to signal ratio, the authors found that classifying into the gain or loss components DNA sequences with a posterior probability above 0.8 gave good operational characteristics in this context, whereas the Bayes rule had poorer performance.

When mixture models are estimated using Empirical Bayes methods, without an estimate of the number of genes in the null, the posterior probability of being allocated to the null can only be estimated up to a constant. In this situation the posterior odds ratio can be used to rank genes:

$$Odds_g = \frac{\mathbb{P}(z_g = 0|\mathbf{x})}{\mathbb{P}(z_g \neq 0|\mathbf{x})} \tag{13}$$

(Lönnstedt & Speed, 2003; Smyth, 2004).

*Models with non-structured priors on fold change parameters*

In the non-mixture methods in the previous section, a variety of measures of differential expression are used to classify genes. Baldi & Long (2001) and Smyth (2004) propose so-called regularized or moderated t-statistics. These consist of the Bayesian posterior mean estimate of the log fold change parameter, divided by a shrunken estimate of standard deviation. This shrunken estimate is the square root of the posterior mean of the variance parameter, shrinkage being provided by the exchangeable prior on the variances estimated in an EB framework.

The model used by Bhattacharjee et al. (2004) allows gene variances to take one of three values (a mixture on gene variability). These can be used to classify genes into groups based on their variability within and between tissues.

With a non-informative prior on the $\delta_g$, Lewin et al. (2006) proposed a decision rule based on a threshold $\delta_{cut}$ set according to a biologically interesting level of differential expression. Differential expression is defined as $\delta_g$ being greater than $\delta_{cut}$, corresponding to an interval null hypothesis with the interval fixed a priori. The decision rule is that genes are declared to be differentially expressed if the posterior probability

$\mathbb{P}(|\delta_g| > \delta_{cut}|\mathbf{x})$ is greater than some threshold probability (e.g. 0.5). This rule combines statistical and biological significance.

When the interval for an interval null hypothesis is required not to be fixed a priori, Bochkina & Richardson (2006) suggest two types of decision rule based on tail posterior probabilities. They define a loss function

$$L \propto \sum_{g \in S_0} I[|\delta_g| > \theta(\sigma_g)|\mathbf{x}] + \lambda \sum_{g \in S_1} I[|\delta_g| \leq \theta(\sigma_g)|\mathbf{x}] \tag{14}$$

They consider two possibilities for $\theta$: firstly $\theta \propto \sigma_g$, which leads to a decision rule where $S_0$ is defined as the group of genes with $\mathbb{P}(|\delta_g/\sigma_g| \leq T^\alpha|\mathbf{x}) \geq 1/(1+\lambda)$, which is an analogue of a t-statistic procedure. The second choice is with constant $\theta$, in which case the decision rule defines $S_0$ as genes with $\mathbb{P}(|\delta_g| \leq \delta_g^\alpha|\mathbf{x}) \geq 1/(1+\lambda)$. A heuristic argument is used to choose the thresholds $T^\alpha$ and $\delta_g^\alpha$; these are defined as the percentiles of the distribution of $\delta_g/\sigma_g$ or $\delta_g$ found by hypothetically conditioning on $\bar{x}_{g2.} - \bar{x}_{g1.} = 0$ in the model, e.g. $f(\delta_g|\bar{x}_{g2.} - \bar{x}_{g1.} = 0, s_g^2)$. Bochkina & Richardson (2006) also consider a one-sided rule with threshold zero. This is shown to be equivalent to the moderated t-statistic of Smyth (2004) (when the same variance model is used).

### 3.5 *Multi-class data*

A number of models have been proposed which extend the methods used for differential expression in two conditions to compare expression in several conditions or classes. These might be used, for example, to compare the actions of several drugs and a control sample simultaneously, or to compare different tumor samples. As with the mixture models described previously, it can be useful to describe the classification of genes in terms of null and alternative hypotheses. There are a number of different choices of alternative hypothesis for multi-class data. Here we discuss models which use hypotheses of the type "the gene is differentially expressed (or not) in at least one condition", without distinguishing which condition it is. Section 4 deals with models which classify genes by clustering them according to the pattern of expression across the experimental conditions, thus distinguishing between being differentially expressed in condition 4 only and being differentially expressed in condition 2 only, for example. An intermediate approach is taken by Ishwaran & Rao (2005b) who, similarly to their work on differential expression, formulate the multi-class analysis as a multivariate regression problem, use variable selection and shrinkage to output lists of significant genes between any two conditions and then use these lists to highlight patterns of interest between the conditions.

A common formulation is the classical ANOVA model, which tests the null hypothesis "the gene has the same expression in all conditions" versus the alternative "the gene has differential expression in at least one condition". This is used in a Bayesian framework by Broët et al. (2004), who start with an F-statistic for each gene. These

F-statistics are transformed to the Normal scale and modeled with a two-component mixture to classify genes as differentially expressed or not. The null component is a standard Normal, while the alternative component is modeled semi-parametrically with a mixture of Normals. This type of formulation is also suggested by Smyth (2004), who uses moderated F-statistics, using shrunken estimates of variances as with the moderated t-statistics (see Section 3.4).

A slightly different formulation is considered by Bochkina & Richardson (2006), who use alternative hypotheses such as "the gene is DE in a set of pair-wise comparisons of interest" versus a compound null which is the opposite of this, i.e. genes are only selected to be of interest if they show changes in a predefined set of comparisons of interest.

## 4. Clustering Gene Expression Profiles

When gene expression is measured in several conditions simultaneously, the data is in the form of a matrix $x_{gc}$ with the index $c$ taking more than 2 values, $c = 1, ..., n$. In this case, the interest focuses on finding groups of genes which have the same pattern of co-expression across the conditions. These patterns are called expression profiles.

### 4.1 *Un-ordered samples*

The models in this section are designed for experiments in which there is no special ordering of the different conditions, for example several tumor samples. Most commonly, this sort of data has no replicate measurements, or at any rate samples from different individuals are not treated as replicates. When replicates under different conditions have been measured, the modeling of the profiles can take this into account to improve the classification, (Medvedovic et al., 2004; Dahl, 2006). Even with no replicates, the different samples can however be used together to estimate gene variances, even though the samples have different expression levels.

Most of these models start with a similar linear model to that in Equation 6. It is convenient to write the gene profile as a vector:

$$\vec{x}_g = \vec{\mu}_g + \epsilon_g \tag{15}$$

where the vectors $\vec{x}_g$ and $\vec{\mu}_g$ are of length $n$. The normalization term is omitted here, as most published Bayesian models do not include this term (instead requiring the data to have been transformed in a suitable way beforehand).

Choices for the distribution used to model gene variability are similar to those discussed in Section 3.2. Here our focus is on modeling the mean expression levels in the different groups. As with the differential expression models, a mixture model can be put on the $\vec{\mu}_g$ parameters, or on the data directly. The correspondence with null and alternative hypotheses can also be carried forward to this type of model, though now there may be several alternatives. The null is "no difference in expression

across conditions". The alternatives are usually all possible patterns showing some difference, though Kendziorski et al. (2003) allow the number of alternative patterns to be restricted, which is a useful feature for large numbers of experimental conditions.

The probability of expression (POE) model of Parmigiani et al. (2002) and Garrett-Mayer & Scharpf (2006) is a simple 3-components mixture model on the data. Its aim is to estimate allocation probabilities for each gene and condition to one of 3 groups: reference, under- and over-expressed. In its simplest form, biological information is available to classify some indices $c$ as giving 'normal' *i.e.* reference values. Reference values are modeled as arising from a normal distribution with additive gene plus condition global effects, whereas the under and over components are assumed to be uniformly left or right shifted from the reference mean with a range to be estimated. All unknown parameters are given prior distribution and the mixture is estimated in a fully Bayesian way via MCMC algorithms, (Parmigiani et al., 2002). The output of this mixture model is a simple transformation of the data matrix into a probability scale based on an underlying assumption that the information in the expression values is essentially categorical. In contrast to other work described below, clustering of the probabilities to define interesting subgroups of gene is not attempted within the model, but the authors suggest to use data mining tools at a second stage.

Kendziorski et al. (2003), Gottardo et al. (2006c) and House et al. (2006) suggest models for clustering gene expression profiles using a mixture model on the parameters $\vec{\mu}_g$. These models are extensions of those used in differential expression, and are formulated via combinations of point masses and continuous distributions for the various hypotheses/clusters. Kendziorski et al. (2003) extend the model of Newton et al. (2004), described in the differential expression section, to a mixture model on gene expression profiles. As an example, when there are three experimental conditions, the gene expression parameters are $\mu_{g1}, \mu_{g2}, \mu_{g3}$. There are 5 possible patterns for a profile over 3 conditions: one pattern with $\mu_{g1} = \mu_{g2} = \mu_{g3}$, three patterns with two of the $\mu_{gc}$ equal to each other and the third drawn from a separate distribution, and one pattern where all three $\mu_{gc}$ are drawn from different distributions. Their implementation allows the restriction to a few interesting patterns, as the number of possible patterns increases rapidly with the number of conditions. One version of the model of Kendziorski et al. (2003) is an extension of that used in Newton et al. (2004) where the likelihood is a Gamma and the mean expression parameters have Gamma priors. They also look at a version with log Normal likelihood and Normal priors. Gottardo et al. (2006c) propose a similar model for profiles, implemented for three experimental conditions. This model is an extension of their differential expression model mentioned in Section 3.3, using log Normal likelihood and Normal priors. They automatically consider all possible patterns. House et al. (2006) give another similar model, this time implemented in five conditions.

Vogl et al. (2005) gives an example of a mixture model on the data (see Section 3.3 for more explanation on mixture models). In this case $\vec{\mu}_g$ is replaced by $\vec{\mu}_k$ in Equation 15, where $k$ labels the mixture component and the mixture allocation parameter $z_g$ is equal to $k$. This model assumes a Normal distribution for each mixture component, with variance $\sigma_k^2$, i.e. equal variance for all genes in the same component. Note that gene variances integrated over different allocations will not be equal for all genes, as different genes will be allocated to different combinations of mixture components. The prior used in Vogl et al. (2005) for the $\vec{\mu}_k$ is a conjugate prior, with independence between different experimental conditions. The number of clusters $k = 1, ..., K$ is estimated in the model along the lines of Richardson & Green (1997). This model does not in fact automatically include the null cluster of equal expression in all conditions. Figure 4 shows some of the gene profiles found by Vogl et al. (2005) for the Spellman cell-cycle data (Spellman et al., 1998). Different clusters of genes peak at different phases of the cell cycle.

Rather than using finite mixture models for clustering profiles, a number of authors, (Medvedovic et al., 2004; Dahl, 2006; Lau & Green, 2006a,b), have recently developed fully Bayesian profile clustering, based on Dirichlet process mixtures (DPM). In this set-up, it is assumed that gene profiles characterized by parameters $\vec{\mu}_g, g = 1, \ldots, p$ are clustered according to a tractable distribution on partitions corresponding to the Dirichlet process and that within each cluster, the profiles follow the same distribution. DPM is a popular formulation for implementing clustering and partitions models. Indeed, besides their representation as infinite mixtures (see Section 3.3), DP models with baseline distribution $G^*$ and scalar $\alpha$ can be equivalently defined via a prior structure on the space $\mathbf{C}$ of partitions of $p$ items (here the genes) into $K$ clusters, $\{1, \ldots, p\} = \bigcup_{k=1}^{K} C_k$, with $p_k$ items in cluster $C_k$, a joint distribution on the partition given by:

$$p(C_1, C_2, \ldots, C_K) = \frac{\alpha^K \Gamma(\alpha) \prod_{k=1}^{K} (p_k - 1)!}{\Gamma(\alpha + p)},$$

associated i.i.d. draws of $\vec{\mu}_k^*$ from $G^*$, $k = 1, \ldots, K$, and setting $\vec{\mu}_g = \vec{\mu}_k^*$ if $g \in C_k$.

Authors differ in their choice of specification of the base measure distribution, and whether they choose a fully conjugate specification between the mixture kernels for the data part of the model and the base measure. In Lau & Green (2006b), the standard DPM approach is extended by replacing the DP model by a variant in which there is a background cluster not exchangeable with the others and in which there is a different prior distribution of the cluster-specific parameters. Fully conjugate specification (Dahl, 2006; Lau & Green, 2006a,b) is computationally advantageous as all the cluster parameters can be integrated out and efficient MCMC algorithms can be used that update solely the partition. Within cluster posterior distributions for the parameters are then sampled, conditional on the partition.

In general, drawing inference from the complex posterior clustering distribution is

not straightforward and as the number of partitions grows rapidly with increasing $p$, recording the partition with the highest posterior probability does not guarantee that it is close to the posterior mode. Medvedovic et al. (2004) suggest to compute the pairwise posterior probabilities for two genes to be in the same cluster and then to postprocess this output by traditional hierarchical clustering algorithms. Dahl (2006) proposes to choose among all the observed clustering, the clustering that minimizes, in the least square sense, the distance between its 0-1 association matrix and the estimated pairwise posterior probabilities. Lau & Green (2006a) formulate a Bayesian solution to define the optimal clustering that optimizes a posterior expected loss function. This loss function penalizes pairs that are clustered together when they should not be and vice-versa. They derive an efficient approximation to define the optimal clustering. Heard et al. (2006a) propose an hierarchical algorithm that approximates the posterior mode in a Bayesian clustering model without requiring MCMC computations (see next section for details).

It is possible to apply the models presented above to time-course or dose-response data, where there is an ordering of the samples, for example in Medvedovic et al. (2004) and Vogl et al. (2005). In those examples the un-ordered samples models work well. However, for data with less pronounced patterns it is better to use models which take into account the ordering information, such as those presented in the next section.

### 4.2 *Ordered samples*

Models for ordered samples also usually start with a linear model,

$$x_{gt} = \mu_{gt} + \epsilon_{gt} \tag{16}$$

where again we omit the normalization term, as this is usually assumed to be already taken care of. We use the index $t$ for the ordered data-points. For convenience we will refer to these as time-points, though they could be any ordered data. This type of data tends not to have repeated measurements for the same time point. Because of this, it is not possible to estimate separate $\mu_{gt}$ for each gene and perform clustering on these. The mixture prior for clustering must be at the data level, i.e. $\mu_{gt}$ is replaced by $\mu_{kt}$, where $k$ labels the mixture component to which gene $g$ is allocated.

Two broad classes of models for dependence between time-points have been proposed in the literature. One class models the parameter at any given time $t$ in terms of the previous time-point or several time-points. The other uses parametric basis functions to give a shape to the parameters across the time-points. Note also that the formulation presented in (Lau & Green, 2006a,b) allows the structuring of $\vec{\mu}_g$ as a linear function of a fixed set of covariates, in particular time (or function thereof), thus can be used effectively for both ordered and un-ordered samples.

Ramoni et al. (2002) implement the first class of model, using an auto-regressive (AR) prior on the $\mu_{kt}$ (thus $\mu_{kt}$ is regressed on $\mu_{k,t-1}, ..., \mu_{k,t-q}$, where $q$ is the order of the

AR prior). The AR prior assumes a stationary time-series, so will not be appropriate for many types of microarray data, especially as data is often measured at irregular time-points. The clustering used in this method is hierarchical and agglomerative, that is the posterior space of clusters is not explored fully, but a path is taken through the space in a similar way to classical hierarchical clustering methods. The scoring function used to decide which clusters to merge is based on ratios of posterior probabilities of the original and merged partitions.

A more flexible model in this class is given in Wakefield et al. (2003) and Zhou & Wakefield (2006), who use a random walk model on the $\mu_{kt}$:

$$\mu_{kt} = \mu_{k,t-1} + u_t \tag{17}$$

where $u_t \sim N(0, |X_t - X_{t-1}|\tau^2)$ and $X_t$ is the value of time at the $t$th time-point. Thus the closer adjacent time-points are, the more dependent they are. This model is fitted in a fully Bayesian way, with the number of clusters estimated as part of the model. Inference is made on the basis of posterior probability of cluster membership, with particular attention given to finding pairs of genes with high probability of being allocated to the same cluster. This is to find genes which are co-expressed.

A model in the second class is proposed by Heard et al. (2006a). This uses splines (with degree to be chosen) as basis functions for the trajectories of the genes over time,

$$\mu_{kt} = \sum_h X_{th}\beta_{hk} \tag{18}$$

where $X_{th}$ is the (fixed) value of the $h$-th basis function at time $t$ and $\beta_{hk}$ is the coefficient of the $h$-th basis function for cluster $k$. By using a fully conjugate specification, the joint distribution of the data conditional on any partition can be computed in closed form, hence the posterior probability of any partition can also be evaluated. The clustering method exploits this and proceeds by an agglomerative algorithm similar to that used by Ramoni et al. (2002), in order to find the partition that approximates the posterior mode. Heard et al. (2006b) extend this model to time-series taken in a number of different conditions, and estimate covariance between time-series in different conditions.

Wakefield et al. (2003) also fit a basis function model, for periodic data.

$$\mu_{gt} = A_g \sin(wX_t) + B_g \cos(wX_t) \tag{19}$$

applied to a cell-cycle data set.

## 5. Multivariate gene selection models

In the previous sections, we have discussed *gene expression association studies* where the aim is to find gene expression changes that relate to biological outcomes by comparing, for each gene, their differential expression under different conditions, and *profile*

*clustering* where the interest is to find patterns of co-expressed genes across different experimental conditions, in order to understand pathways. In this section we are concerned with a different but related problem, that of using gene expression for phenotype prediction. Our aim here is to build multivariate molecular profiles based on combination of the expression of a subset of genes which can characterize different phenotypes (e.g. clinical outcomes). We are thus in the framework of multivariate regression and classification models. The specific difficulty of genomic applications is that there are typically many more covariates than samples: the so-called "large $p$ (thousands of genes), small $n$ (50 to 100 samples) regression paradigm", and consequently standard regression/discrimination techniques do not apply. Further, the interest is in finding parsimonious regression models that include only small subsets of genes so that biological interpretation and validation can be attempted.

Bayesian approaches to multivariate gene selection have broadly followed two related lines of development: (i) regression methods with covariate selection, (ii) multivariate regression with shrinkage priors that favor sparsity. We shall review these in turn. Mostly, we shall discuss so-called supervised classification situations where the characteristic of the samples that one wants to predict are known. Variable selection can also be performed simultaneously with the task of uncovering clustering patterns of the samples, in an unsupervised manner.

### 5.1 *Variable selection approach*

Suppose that we have potentially $p$ predictor variables, each measured on a set of $n$ samples: $x_{gc}$ with $g = 1, \ldots, p$ and $c = 1, \ldots, n$. Thus for each predictor variable $g$, we have a vector of $n$ measurements $\vec{x}_g$. For the present the outcome variable, $y_c, c = 1, \ldots, n$ can be continuous (e.g. measuring a biomarker) or categorical (e.g. encoding a cancer subtype) and we denote by $\boldsymbol{\beta}$, the vector of regression parameters linking $\mathbf{X}$ and $\mathbf{Y}$, (these being the matrices for predictors and outcomes respectively). Thus $\beta_g$ is the regression parameter corresponding to the covariate $\vec{x}_g$.

Bayesian variable selection (BVS) is usually implemented through a hierarchical model, where all possible $2^p$ models are represented by a $p$-dimensional indicator variable $\gamma$:

$$\gamma_g = \begin{cases} 1 & \text{variable (gene) g is included} \\ 0 & \text{variable (gene) g is excluded} \end{cases}$$

A prior on the model space can be specified via a prior $p(\gamma)$. A common choice is $p(\gamma) = \prod_{g=1}^{p} \pi^{\gamma_g}(1-\pi)^{1-\gamma_g}$, and by choosing $\pi$ small, the number of variables selected can be controlled. Alternatively, a Beta prior distribution can be assumed for $\pi$ and the sparsity of the regression only controlled by the choice of prior for the $\beta$s.

This generic approach to variable selection, often referred to as the *spike and slab* approach, was taken in Mitchell & Beauchamp (1988), George & McCulloch (1993), George & McCulloch (1997) and in many subsequent papers (Clyde, 1999; Brown et al.,

1998, 2002). Much of the work on BVS was developed for linear models where $y_c$ is continous. In this context, authors differ in the choice of the prior distribution for $\boldsymbol{\beta}$, in particular whether the components of $\boldsymbol{\beta}$ are treated as independent or not, and whether a conjugate formulation is chosen so that the prior on $\boldsymbol{\beta}$ includes the noise parameter of the linear model. Typically, the prior for $\boldsymbol{\beta}$ is formulated via a mixture. Most models define a point mass at zero for $\beta_g$ when $\gamma_g = 0$, while when $\gamma_g = 1$, large variances are favored with a distribution to be specified. Ishwaran & Rao (2003, 2005a,b) use a modified spike and slab model that assumes a continuous bimodal prior for $\beta_g$, a scale mixture of two centered normals, one having a small variance. They show that this prior is useful in gene expression, see Sections 3.3 and 3.5. In the common case of a prior for $\beta_g$ with a point mass at zero, for ease of notation, we shall denote by $\boldsymbol{\beta}_\gamma$ all nonzero elements of $\boldsymbol{\beta}$ and correspondingly, we denote by $\mathbf{X}_\gamma$, the columns of $\mathbf{X}$ corresponding to those elements of $\gamma$ equal to 1.

A standard choice of prior for $\boldsymbol{\beta}_\gamma$ is the so-called g-prior, where $\boldsymbol{\beta}_\gamma \sim N(0, c({\mathbf{X}_\gamma}^T\mathbf{X}_\gamma)^{-1})$, where $c$ is a positive scale factor to be chosen. Similarly, if independent normals are specified for the components of $\boldsymbol{\beta}$, again a scalar has to be chosen. These choices influence the sparsity of the final regression model (Chipman et al. (2001)) and a full understanding of this aspect is the object of current research.

In the microarray context, because we are in a "large $p$, small $n$" situation, the posterior distribution over the model space of variable dimensions is multi-modal. Moreover, full posterior inference for the entire model space of size $2^p$ is not feasible if $p$ is larger than about 20. Hence, Markov chain Monte Carlo methods are rather used as stochastic search algorithms with the aim to quickly find many regions of high posterior probability. The Markov chain needs to move quickly around the support of the posterior distribution and as usual, it is useful to integrate out as many parameters as possible. For this reason, conjugate settings have been favored in the linear model. When proposing changes to $\gamma$, a key question is how to propose sensible changes to the regression vector $\boldsymbol{\beta}$. The current parameter values may be of little relevance in this case and joint moves which update simultaneously $\gamma$ and $\boldsymbol{\beta}$ produce an improvement (Holmes & Held, 2006).

Much of the application of variable selection in microarrays has concerned binary or categorical variables rather than the linear model. Typically, samples are classified as good or poor prognosis or linked to different clinical sub entities, like subtypes of cancer. There is no immediate conjugate formulation of Bayesian categorical regression, but following the approach of Albert & Chib (1993), probit regression can be efficiently implemented through the use of latent auxiliary variables which allows integration of the regression coefficients in the full conditional distribution of the indicator variables $\gamma$. This approach was taken by Lee et al. (2003) and Sha et al. (2004). These authors have implemented different MCMC schemes for updating $\gamma$ (Gibbs sampling for Lee et al.

(2003) which will tend to be slow mixing, Metropolis with add/delete/swap moves for Sha et al. (2004)). In a recent paper, Holmes & Held (2006) have proposed an auxiliary variable formulation of the logistic model which allows, in a similar way to the probit model, integration of the regression coefficients when updating the indicator variables, thus improving mixing. For both probit and logistic regression models, the calibration of the prior distribution of the regression coefficients again influences the outcome of the variable selection process. In this respect, the logistic regression, which is more commonly used for binary regression, is easier to calibrate than the probit model as it has heavier tails and so exhibits less sensitivity.

In general, MCMC variable selection algorithms in high dimension are difficult to implement due to slow convergence. Recent developments in stochastic simulation algorithms, such as population-based reversible jump MCMC and the use of parallel tempered chains seem promising, (Jasra et al., 2006). An alternative search algorithm, the *shotgun stochastic search* method, which is close in spirit to MCMC but aims to search rapidly for the most probable models has been recently proposed by Hans et al. (2007). Note that it is a discussion point whether to report models with high posterior probabilities and their associated variables, or to extract marginal information about the selected variables by looking at their marginal posterior probabilities of inclusion, Sha et al. (2004).

We end this section by referring to the recent work of Tadesse et al. (2005) and Kim et al. (2006) where variable selection and clustering of the samples are performed simultaneously. This joint modeling is motivated by the remark that using a high dimensional vector of gene expression to uncover clusters among the samples might not be effective and on the contrary can tend to mask existing structure, while a more parsimonious model that selects a only a small subset of covariates to inform the clustering is more easily interpretable. Such joint modeling was discussed in a Bayesian context in two related papers, which differ in their model for the clustering structure. Tadesse et al. (2005) formulate their clustering structure using a finite mixture of multivariate normals with a variable number of components and use reversible jump techniques to explore different structures, whilst Kim et al. (2006) exploit the computational benefits of DP mixtures.

### 5.2 *Bayesian shrinkage with sparsity priors*

An alternative approach to BVS for selecting a small number of regressors is to use a hierarchical formulation of the regression problem with a prior on the regression coefficients that favors sparsity. Effectively, a large number of regression coefficients are essentially set to zero by having very small posterior values. Note that different choices of prior and hierarchical structures can be interpreted as different choice of penalties if one adopts the point of view of penalized likelihood, framework into which ridge regression and other shrinkage methods can be cast.

A general formulation that encompasses many of the models which have been proposed is that of scale mixture of normals. In this formulation, the regression coefficients $\beta_g$ are given independent normal priors: $\beta_g \sim N(0, \tau_g)$, and the variances $\tau_g$ are themselves given a hyperprior distribution: $\tau_g \sim p(\tau_g)$. Choice of this prior distribution leads to different kind of sparsity for the $\beta$s, but all priors used have in common the desirable feature that the integrated prior for $\beta$ is a heavy tail distribution with a peak around zero, thus favoring only a small number to be substantially different from zero. Note that a Laplace prior for $\beta_g$ which corresponds to a Lasso type penalization can also be written as a scale mixture of normals with $p(\tau_g)$ being a one parameter exponential distribution, (Griffin & Brown, 2005).

Bae & Mallick (2004) implement three different choices of prior for $\tau_g$ in the context of gene expression studies: an inverse gamma with 2 hyper parameters that are chosen in order to favor large variances, a Laplace prior with one parameter and a Jeffreys improper prior (which implies an improper prior of the form $\frac{1}{\beta_g}$ for $\beta_g$). They found that Jeffreys prior induces more sparseness than the Laplace prior and yields good performance. There is currently a lot of interest in using general families of scale mixtures and in calibrating them for efficient inference in high dimensional set-ups (Griffin & Brown, 2005).

Sparsity priors can also be used in the context of latent factor models (West, 2003; Lucas et al., 2006). Modeling high dimensional data via latent factor models is a powerful dimension reduction technique that allows the identification of patterns of covariation among genes. In their application of factor models to the analysis of gene expression data, West (2003) and Lucas et al. (2006) further structure the factor loading matrix to encourage sparsity via a mixture prior with point mass at zero. A biological interpretation of the factors as potentially representing biological pathways is then derived by examining the list of genes most weighted on each factor.

**Acknowledgments** The authors would like to thank their colleagues Marta Blangiardo, Natalia Bochkina, Anne-Mette Hein and Peter Green for many insightful discussions on the Bayesian modelling of microarray data. This chapter was completed while SR was associated with the progam "Stochastic Computation in the Biological Sciences" at the Isaac Newton Institute for Mathematical Sciences, the support of which is gratefully acknowledged. The support of BBSRC "Exploiting Genomics" grant 28EGM16093 is gratefully acknowledged.

## References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A.

(1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS* **96**, 6745–6750.

Bae, K. and Mallick, B. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* **20**, 3423–3430.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

Bhattacharjee, M., Pritchard, C. C., Nelson, P. S., and Arjas, E. (2004). Bayesian integrated functional analysis of microarray data. *Bioinformatics* **20**, 2943–2953.

Bochkina, N. and Richardson, S. (2006). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics, tentatively accepted*.

Broët, P., Lewin, A., Richardson, S., Dalmasso, C., and Magdelenat, H. (2004). A mixture model based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics* **20(16)**, 2562–2571.

Broët, P. and Richardson, S. (2006). Bayesian hierarchical model for identifying change in gene expression from microarray experiments. *Bioinformatics* **9**, 671–683.

Broët, P., Richardson, S., and Radvanyi, F. (2002). Bayesian hierarchical model for identifying change in gene expression from microarray experiments. *Journal of Computational Biology* **9**, 671–683.

Brown, P., Vannucci, M., and Fearn, T. (1998). Multivariate Bayes variable selection and prediction. *Journal of the Royal Statistical Society, Series B* **60(3)**, 627–641.

Brown, P., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B* **64(3)**, 519–536.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. 1995, Measurement Error in Nonlinear Models (Chapman and Hall/CRC).

Ceccarelli, M. and Antoniol, G. (2006). A Deformable Grid-Matching Approach for Microarray Images. *IEEE Transactions on Image Processing* **15**, 3178–3188.

Chipman, H., George, E., and McCulloch, R. (2001). The Practical Implementation of Bayesian Model Selection (incl. discussion). *IMS Lecture Notes - Monograph Series* **38**, 67–134.

Clyde, M. (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics 6*, ed. J. Bernardo, J. Berger, A. Dawid, & A. Smith, Proceedings of the Sixth Valencia International Meeting (Oxford University Press), 157–185.

Dahl, D. (2006). Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. In *Bayesian Inference for Gene Expression and Proteomics*, ed. K.-A. Do, P. Müller, & M. Vannucci (Cambridge University Press), 201–218.

Do, K., Müller, P., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Series C* **54**, 627–644.

Do, K.-A., Müller, P., and Vannucci, M. 2006, Bayesian Inference for Gene Expression and Proteomics (Cambridge University Press).

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Frigessi, A., van de Wiel, M., Holden, M., Svendsrud, D., Glad, I., and Lyng, H. (2005). Genome-wide estimation of transcript concentrations from spotted cDNA microarray data. *Nucleic Acids Res.* **33(17)**, e143.

Garrett-Mayer, E. and Scharpf, R. (2006). Models for probability of under- and overexpression: the POE scale. In *Bayesian Inference for Gene Expression and Proteomics*, ed. K.-A. Do, P. Müller, & M. Vannucci (Cambridge University Press), 137–154.

George, E. and McCulloch, R. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association* **88**, 881–889.

George, E. and McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339-373.

Gottardo, R., Besag, J., Stephens, M., and Murua, A. (2006a). Probabilistic segmentation and intensity estimation for microarray images. *Biostatistics* **7(1)**, 85–99.

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. E. (2006b). Quality Control and Robust Estimation for cDNA Microarrays With Replicates. *Journal of the American Statistical Association* **101**, DOI 10.1198.

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. E. (2006c). Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples. *Biometrics* **62**, 10–18.

Griffin, J. and Brown, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. *Technical report* http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic /griffin/personal/vspaper.pdf.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search in regression with many predictors. *Journal of the American Statistical Society, to appear* .

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006a). A quantative study of gene regulation involved in the immune response of anopheles mosquitoes: an application of Bayesian hierarchical clustering. *J. Amer. Stat. Soc.* **101**, 18–29.

Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J., and Dimopoulos, G. (2006b). Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. *PNAS* **102**, 16939–16944.

Hein, A.-M. K., Richardson, S., Causton, H. C., Ambler, G. K., and Green, P. J. (2005). BGX: a fully Bayesian gene expression index for Affymetrix GeneChip data. *Biostatistics* **6(3)**, 349–373.

Hein, A.-M. K. and Richardson, S. (2006). A powerful method for detecting differentially expressed genes from GeneChip arrays with no replicates. *BMC Bioinformatics***7**, 353.

Holmes, C. and Held, L. (2006). Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis* **1**, 145–168.

House, L., Clyde, M., and Huang, Y.-C., T. (2006). Bayesian Identification of Differential Gene Expression Induced by Metals in Human Bronchial Epithelial Cells. *Bayesian Analysis* **1**, 105–120.

Ibrahim, J.G., Chen, M.-H. C., and Gray, R.J. (2002). Bayesian Models for Gene Expression With DNA Microarray Data. *Journal of the American Statistical Association* **97**, 88–99.

Ishwaran, H. and Rao, J. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* **98**, 438–455.

Ishwaran, H. and Rao, J. (2005a). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics* **33**, 730–773.

Ishwaran, H. and Rao, J. (2005b). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association* **100**, 438–455.

Jasra, A., Stephens, D., and Holmes, C. (2006). Population based reversible jump Markov chain Monte CarloTechnical Report. Imperial College, UK.

Kendziorski, C., Newton, M., Lan, H., and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899–3914.

Kim, S., Tadesse, M., and Vanucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika, in press* .

Lau, J. and Green, P. (2006a). Bayesian model based clustering procedures. *Journal of Computational and Graphical Statistics, tentatively accepted* .

Lau, J. and Green, P. (2006b). Bayesian clustering using an asymmetric Dirichlet process, with application to gene expression profile classification. *Technical report, University of Bristol, UK* .

Lee, K., Sha, N., Dougherty, E., Vannucci, M., and Mallick, B. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics* **19(1)**, 90–97.

Lewin, A., Richardson, S., Marshall, C., Glazier, A., and Aitman, T. (2006). Bayesian Modelling of Differential Gene Expression. *Biometrics* **62**, 1–9.

Lin, Y., Reynolds, P., and Feingold, E. (2003). An Empirical Bayesian Method for Differential Expression Studies using One-channel Microarray Data. *Statistical Applications in Genetics and Molecular Biology* **2**, Article 8.

Lönnstedt, I. and Speed, T. (2003). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics*, ed. K.-A. Do, P. Müller, & M. Vannucci (Cambridge University Press), 155–176.

Medvedovic, M., Yeung, K., and Bumgarner, R. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* **20**, 1222–1232.

Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023–1032.

Müller, P., Parmigiani, G., and Rice, K. (2007). FDR and Bayesian multiple comparison rules. In *Bayesian Statistics 8*, ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Oxford University Press), to appear.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001). On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology* **8**, 37–52.

Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* **5**, 155–176.

Parmigiani, G., Garrett, E., Anbazhagan, R., and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society, Series B* **64**, 717–736.

Ramoni, M. F., Sebastiani, P., and Kohane, I. S. (2002). Cluster analysis of gene expression dynamics. *PNAS* **99**, 9121–9126.

Reilly, C., Wang, C. and Rutherford, M. (2003). A method for normalizing microarrays using genes that are not differentially expressed. *Journal of the American Statistical Association* **98**, 868–878.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

Schadt, E., Li, C., Su, C., and Wong, W. (2000). Analyzing High-Density Oligonucleotide Gene Expression Array Data. *Journal of Cellular Biochemistry* **80**, 192–202.

Sethurman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**, 639–650.

Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, N., Buckley, C., and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular singatures of disease stage. *Biometrics* **60**, 812–819.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, 3.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.

Tadesse, M., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **100**, 602–617.

Vogl, C., Sanchez-Cabo, F., Stocker, G., Hubbard, S., and Wolkenhauer, O. (2005). A fully Bayesian model to cluster gene-expression profiles. *Bioinformatics* **21**, ii 130 – ii 136.

Wakefield, J. C., Zhou, C., and Self, S. G. (2003). Modelling gene expression data over time: Curve clustering with informative prior distributions. In *Bayesian Statistics 7*, ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Oxford University Press), 721–732.

Walker, S., Damine, P., Laud, P., and Smith, A. (1999). Bayesian nonparametric inference for distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B* **61**, 485–527.

West, M. (2003). Bayesian Factor Regression Models in the "Large p, Small n" Paradigm. In *Bayesian Statistics 7*, ed. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, & M. West, Proceedings of the Seventh Valencia International Meeting (Oxford University Press), 733–742.

Zhou, C. and Wakefield, J. C. (2006). A Bayesian mixture model for partitioning gene expression data. *Biometrics* **62**, 515–525.

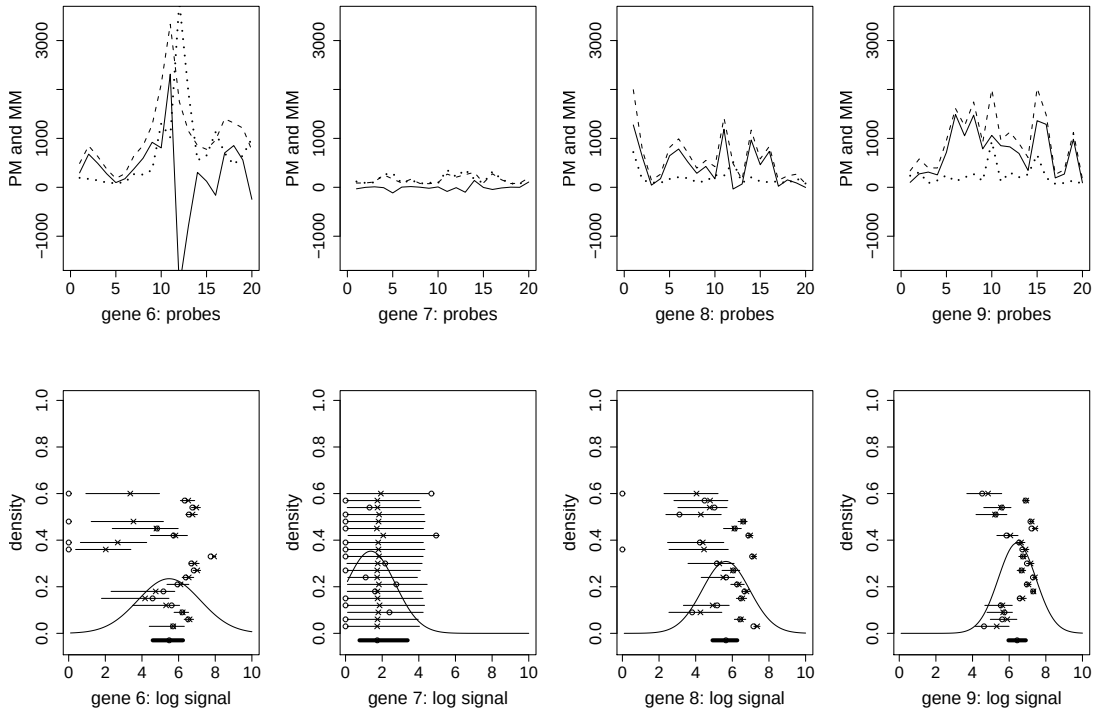


**Figure 1.** Upper panel: probe set response for four genes from an oligonucleotide microarray. Each probe set consists of 20 probe pairs. Solid lines show PM-MM, dashed lines show PM, dotted lines show MM. Lower panel: summaries of posterior distributions related to expression of the four genes, from the model in Hein et al. (2005). The 95% equal-tailed credibility intervals of the $S_{gjcr}$ are shown as horizontal lines (shifted vertically) and should be read off the $x$-axis. The bold line shows the 95% equal-tailed credibility interval for $\mu_{gc}$. Circles show the observed log(PM-MM) values (plotted at zero when MM > PM). Curves show $TN(\hat{\mu}_{gc}, \hat{\sigma}^2_{gc})$, with $\hat{\mu}_{gc}$ and $\hat{\sigma}^2_{gc}$ equal to the posterior means. Figure reproduced with permission from Hein et al. (2005).

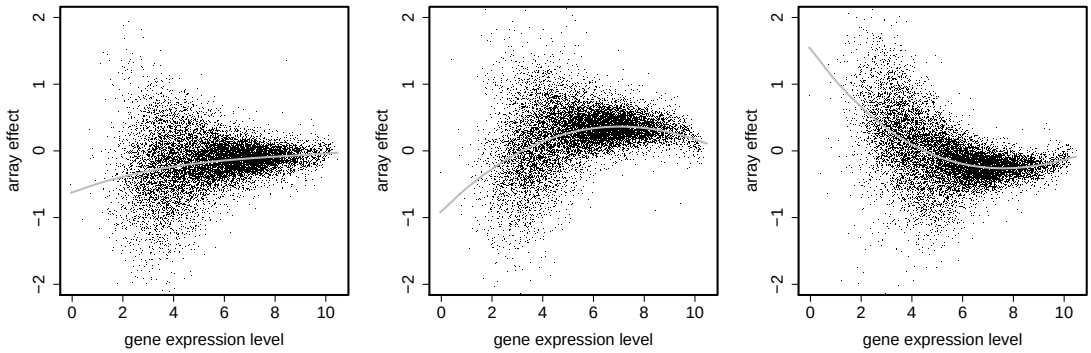


**Figure 2.** Array effects as a function of expression level, for a wildtype mouse expression data set of 3 arrays, as presented in Lewin et al. (2006). Each plot shows the array effect from 1 array (curve) with the data residuals from the mean across arrays (points).

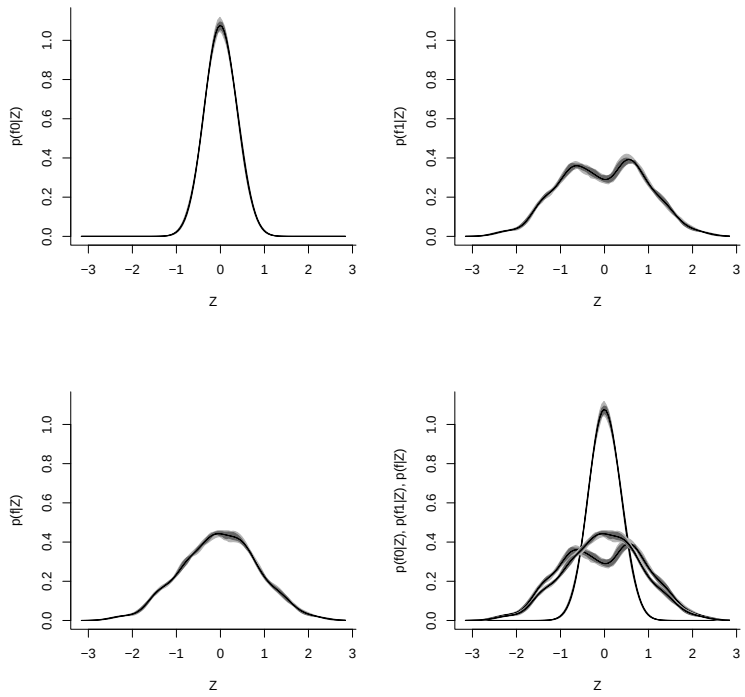


**Figure 3.** Illustration of the posterior distributions for the mixture densities in the Do et al. (2005) model, found for the Alon et al. (1999) cancer data set. The first three panels show $f_0$, $f_1$ and $f$; the fourth panel shows all three. Each curve is a draw from the posterior distribution of the relevant mixture component. Figure courtesy of Peter Müller.

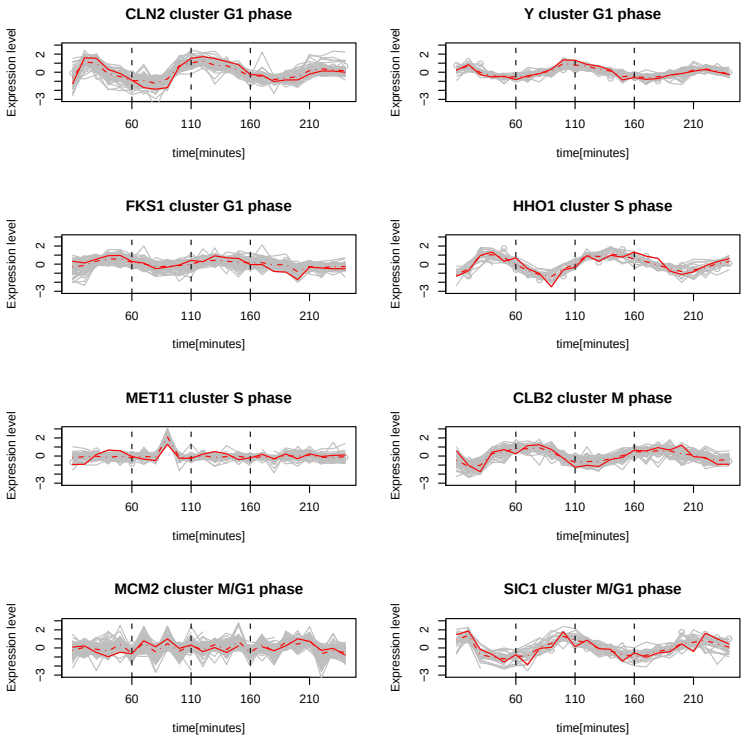


**Figure 4.** Clusters of cell-cycle regulated genes found in the Spellman et al. (1998) data using the model of Vogl et al. (2005). Figure reproduced with permission from Vogl et al. (2005).

| | | |
|---|---|---|
| Extracting Signal from Observed Intensities | | |
| Frigessi et al. (2005) | TransCount | http://alba.uio.no/base/local/demotranscount/index.html |
| Hein et al. (2005) | BGX | http://www.bgx.org.uk/software.html |
| Differential Expression | | |
| Baldi & Long (2001) | Cyber-T | http://visitor.ics.uci.edu/genex/cybert/ |
| Broët et al. (2002) | nmix | http://www.bgx.org.uk/software.html |
| Broët et al. (2004) | gmix | http://www.bgx.org.uk/software.html |
| Do et al. (2005) | BayesMIX | http://odin.mdacc.tmc.edu/∼kim/bayesmix/ |
| Efron et al. (2001) | EBAM | http://bioconductor.fhcrc.org/packages/2.0/bioc/html/siggenes.html |
| Gottardo et al. (2006b) | rama | http://www.stat.ubc.ca/∼raph/Software/BiocRPackages/BiocRPackages.html |
| Gottardo et al. (2006c) | bridge | http://www.stat.ubc.ca/∼raph/Software/BiocRPackages/BiocRPackages.html |
| Ibrahim et al. (2002) | | code available from author: mhchen@stat.uconn.edu |
| Ishwaran & Rao (2003, 2005a,b) | BAM | http://www.bamarray.com/ |
| Lewin et al. (2006) | BayesDE | http://www.bgx.org.uk/software.html |
| Lönnstedt & Speed (2003) | SMA | http://www.stat.berkeley.edu/∼terry/zarray/Software/smacode.html |
| Newton et al. (2001, 2004) | EBarrays | http://www.stat.wisc.edu/%7Enewton/research/arrays.html |
| Parmigiani et al. (2002) | POE | http://astor.som.jhmi.edu/poe/ |
| Reilly et al. (2003) | | http://www.biostat.umn.edu/ cavanr/geneNormRepHier.txt |
| Smyth (2004) | limma | http://bioinf.wehi.edu.au/limma/ |
| Clustering Profiles | | |
| Gottardo et al. (2006c) | bridge | http://www.stat.ubc.ca/∼raph/Software/BiocRPackages/BiocRPackages.html |
| Heard et al. (2006a,b) | | http://stats.ma.ic.ac.uk/naheard/public_html/ |
| Kendziorski et al. (2003) | EBarrays | http://www.stat.wisc.edu/%7Enewton/research/arrays.html |
| Vogl et al. (2005) | | http://genome.tugraz.at/BayesianClustering/ |
| Zhou & Wakefield (2006) | | http://faculty.washington.edu/jonno/cv.html |

**Table 1**

*Web pages containing software or code for the models described in this chapter*