Combinatorial Mixtures of Multiparameter Distributions

BY VALERIA EDEFONTI
Istituto di Statistica Medica e Biometria "Giulio A. Maccacaro",
Università degli Studi di Milano, 20133 Milan, Italy
valeria.edefonti@unimi.it
AND GIOVANNI PARMIGIANI
The Sidney Kimmel Comprehensive Cancer Center,
Johns Hopkins University, Baltimore, Maryland, 21205, U.S.A.
gp@jhu.edu

August 21, 2009

Abstract

We introduce *combinatorial mixtures* —a flexible class of models for inference on mixture distributions whose component have multidimensional parameters. The key idea is to allow each element of the component-specific parameter vectors to be shared by a subset of other components. This approach allows for mixtures that range from very flexible to very parsimonious, and unifies inference on component-specific parameters with inference on the number of components. We develop Bayesian inference and computation approaches for this class of distributions, and illustrate them in an application. This work was originally motivated by the analysis of cancer subtypes: in terms of biological measures of interest, subtypes may characterized by differences in location, scale, correlations or any of the combinations. We illustrate our approach using data on molecular subtypes of lung cancer.

Some key words: Bayesian inference, Markov chain Monte Carlo, Clustering.

1 Introduction

Since the beginning of the last century, (Newcomb (1886) and Pearson (1894)) finite mixture distributions have received attention both as tools for modeling population heterogeneity, and as a practical way of building flexible finite-parameter distributions. Monographs on finite mixtures include the classical Titterington et al. (1985), McLachlan and Basford (1988) and McLachlan and Peel (2000). Böhning and Seidel (2003) is a recent review with emphasis on nonparametric maximum likelihood, while Marin et al. (2005) is an introduction from a Bayesian perspective.

One of the important remaining challenges of mixture modeling is to develop approaches that achieve a practical compromise between flexibility and parsimony, especially for mixtures whose component distributions are themselves characterized by multiple parameters. In this setting, one has the option of allowing each component to have its own parameter vector, or to share a subset of the vector elements across components. For example, in the context of normal mixtures, it is common to assume either component-specific locations and variances, or a common variance, or a common mean. These three choices are extremes of a richer and useful set of patterns in which one shares some of the parameters in some of the components. Here we develop this idea formally.

Our approach is to allow each element of component-specific parameter vectors to be either different or equal to that of other components. A positive probability is put on every possible combination of equalities, whence the name *combinatorial mixtures*. This partial sharing allows for greater generality and flexibility in comparison with traditional approaches to mixture modeling, while still allowing to assign mass to models that are more parsimonious than the general mixture case, in which no sharing takes place. One of the implications of our setting is that, once a maximum number of components is specified, inference on the parameters and the number of components is subsumed by the inference on combinatorial patterns. If there is complete sharing among two components, then the effective number of components is reduced by one. Therefore assigning a prior on sharing patterns implies assigning a prior on the effective number of mixture components.

This development was originally motivated by applications in molecular biology, where one deals with continuous measures, such as RNA levels, or protein levels, that vary across unknown biological subtypes. In some cases, subtypes are characterized by an increase in the level of the marker measured, while in others they are characterized by variability in otherwise tightly controlled processes, or by the lack of otherwise strong correlations (Dettling et al. (2005); Shedden and Taylor (2004)). Also, several mechanisms can coexist. In this context, the main goals of a mixture model analysis are to a) estimate the number of subgroups in a sample; b) make inferences about the assignment of samples to these subgroups; and c) generate hypotheses about which of the mechanisms above is likely to characterize the subgroups. Our paper adds a new tool to Bayesian mixture models that allows to answer all three of these questions.

The paper is structured as follows. In Section 2 we review relevant Bayesian methods. In Section 3 we propose a general formulation for combinatorial mixtures. In Section 4 combinatorial mixtures are worked out in detail for univariate and bivariate normal mixtures. The performance of the methodology is then illustrated in an application to gene expression in lung cancer.

2 Bayesian Methods for Mixtures

2.1 Notation and Missing Data Formulation

In a finite mixture model, observations $\boldsymbol{x}^n = (x_1, \ldots, x_n)$ are assumed to be conditionally independent from density

$$x_i|K, \boldsymbol{\theta}, \boldsymbol{\omega} \sim p(x_i|K, \boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{k=1}^{K} \omega_k p(x_i|\theta_k), \qquad i = 1, \ldots, n, \tag{1}$$

where K is the number of components, $\boldsymbol{\omega} = (w_1, \ldots, w_K)$ are the mixture weights —constrained to be non-negative and to sum to unity— and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ is the component specific parameter vector. A *fully* Bayesian analysis (Richardson and Green (1997)) is completed by a prior distribution on the parameters $(K, \boldsymbol{\theta}, \boldsymbol{\omega})$. The number of components and the mixture components parameters are modeled jointly and inference about these quantities is based on their posterior distributions.

The missing data formulation of mixture models has played an essential role in both Bayesian and non-Bayesian approaches. In this formulation, each observation $x_i,\ i = 1, \ldots, n$, is assumed to arise from a specific but unknown component, z_i, of the mixture. Model (1) can be written in terms of the missing data, with $z_1, \ldots, z_n$ assumed to be realizations of conditionally independent and identically distributed discrete random variables, $z_1, \ldots, z_n$, with probability mass function:

$$p(z_i = k|\boldsymbol{\theta}, \boldsymbol{\omega}) = \omega_k, \qquad \text{for } i = 1, \ldots, n,\ k = 1, \ldots, K.$$

Conditional on the zs, $x_1, \ldots, x_n$ are independent observations from densities:

$$p(x_i|z_i = k, \boldsymbol{\theta}, \boldsymbol{\omega}) \ = \ p(x_i|\theta_k), \qquad i = 1, \ldots, n,\ k = 1, \ldots, K.$$

Integrating out $z_1, \ldots, z_n$ yields model (1). The identity of the group from which each observation arises may also be known. The model will then be composed by K distinct subpopulations having their own distributions. We will refer to that special case as *supervised*, as opposed to the *unsupervised* case in which the class indicators are unknown. Both supervised and unsupervised settings are compatible with the analysis that we propose.

2.2 Mixtures of Dirichlet Processes

Escobar and West (1995) and West and Turner (1994) used Dirichlet process mixture (DPM) models (Antoniak (1974)) to build flexible mixtures of normal distributions. The basic normal mixture model assumes that data $\boldsymbol{x}^n = (x_1, \ldots, x_n)$ are conditionally independent and normally

distributed, $x_i|\theta_i \sim N(\mu_i, \sigma_i^2)$ with means μ_i and variances σ_i^2 forming the parameters $\theta_i = (\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$. It supposes further that the variability among the θ_is is represented by the mixing distribution $G(\cdot)$ on $\Re \times \Re^+$. If $G(\cdot)$ is uncertain and modeled as a Dirichlet process (DP), then the data come from a Dirichlet mixture of normals. In particular, these papers assume $G \sim DP(\alpha G_0)$, where α is a positive scalar and $G_0(\cdot)$ a specified bivariate distribution function over $\Re \times \Re^+$. Relevant to the development in this paper is the discreteness of $G(\cdot)$ under the Dirichlet process assumption. In any sample $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ of size n from $G(\cdot)$ there is positive probability of coincident values (Blackwell and MacQueen (1973)). However, a significant constraint in this setting is that the event that two components have the same mean but different variances, or vice-versa, is assigned probability zero.

2.3 Unknown number of components

Recent Bayesian work has further investigated inference on the number of components K. Carlin and Chib (1995) and Raftery (1996) estimated the marginal likelihoods of K components and suggested using Bayes factors to compare K versus $K + 1$ components. Mengersen and Robert (1996) also took a testing perspective, but used the Kullback–Leibler divergence as a measure of distance between mixtures with K and $K + 1$ components. Nobile (1994), Nobile (2005), Phillips and Smith (1996), Richardson and Green (1997), Roeder and Wasserman (1997) and Stephens (2000) all worked with a prior distribution on the number of components and obtained Markov Chain Monte Carlo (MCMC) estimates of the posterior. Nobile (1994), Nobile (2005) and Roeder and Wasserman (1997) estimated the marginal likelihoods of each model separately and then formed an estimate of the posterior of K using Bayes' theorem. Roeder and Wasserman (1997) proposed to approximate the marginal likelihoods using the Schwarz criterion.

Although their methods differ considerably, Phillips and Smith (1996), Richardson and Green (1997) and Stephens (2000) shared a common approach consisting of running an MCMC sampler on a composite model, with jumps between submodels that allow the sampler to change the number of components in the mixture. Then the posterior of K can be estimated by the relative amount of simulation time spent by the sampler in each submodel. The approach proposed in the current paper shares this idea as well.

Phillips and Smith (1996) considered the birth and death of mixture components using an iterative jump-diffusion sampling algorithm. A Markov process in continuous time is generated from a jump component which makes discrete transitions between models at random times and a diffusion component which samples values for the model-specific parameters between the jumps.

Richardson and Green (1997) applied the reversible jump MCMC method of Green (1995). Moves between models are achieved by periodically proposing combine/split moves that rely on moment matching, and rejecting those with the appropriate probability to ensure that the chain possesses the desired stationary distribution. A proposed model is generated by randomly choosing a combine or a split move. The combine move selects two components randomly and proposes merging them into one, whereas the split move suggests splitting a randomly chosen component into two new ones.

Stephens (2000) presented an alternative method of constructing an ergodic Markov chain with appropriate stationary distribution, based on a continuous time Markov birth-death process. The relationship between relative rates of births and deaths and stationarity is used to construct an easily simulated process in which births occur at a constant rate from the prior and deaths occur

at a rate that is very low for components which are critical in explaining the data and very high for components which do not help explain the data.

Finally, Nobile (1994), Nobile (2005), Casella et al. (2004), and Steele et al. (2003) have chosen to work in terms of the allocation variables only, after integrating out the component-specific parameters analytically.

2.4 Product partition models

Our proposal is also related to product partition models, introduced by Hartigan (1990) and Barry and Hartigan (1992). These consider a random partition, $\rho = \{A_1, \ldots, A_K\}$, of a set $A_0 = \{1, \ldots, n\}$ of objects, where a partition of A_0 is defined as a set of nonempty, pairwise disjoint subsets of A_0 whose union is A_0. They assume that observations in different components of a partition of the data are independent. The likelihood of $\boldsymbol{x}^n = (x_1, \ldots, x_n)$ for a partition ρ is the product over the components:

$$p(\boldsymbol{x}^n|\rho) \propto \prod_{k=1}^{K} p(\boldsymbol{x}_{A_k}),$$

where $\boldsymbol{x}_{A_k}$ is the vector of observations corresponding to the elements of the component A_k. The component likelihood $p(\boldsymbol{x}_{A_k})$ - that is the likelihood contribution from a component A_k- is defined for any non-empty component $A_k \subset A_0$ and can take any number of forms. The prior distribution for a partition ρ is also taken to be a product over the partition components:

$$p(\rho) = \prod_{k=1}^{K} c(A_k),$$

where $c(A_k) \geqslant 0$ is termed cohesion and is defined for each non-empty $A_k \subset A_0$. The resulting posterior distribution is also a product partition model with posterior cohesions $p(\boldsymbol{x}_{A_k})c(A_k)$. In the context of cluster analysis, a set partition defines a clustering for the observed data. The partition components define groups of data referred to as clusters.

A component density for observations in each component can be introduced in the context of the so called *parametric product partition models*. Crowley (1997) applied product partition models to the normal means problem: $x_i|\mu_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$, though the case where variances are unknown is not addressed. Given $\mu_1, \ldots, \mu_n$, the x_is are independent. However, a prior distribution is chosen for the μs that allows some set of them to be equal, the set of equal μ values defining the partition ρ. Let μ^{A_k} denote the common value of μ_i for $i \in A_k$, when all μ values in A_k are equal. The component density $p_{A_k}(\mu^{A_k})$ is the conditional density of μ^{A_k} given that A_k is a component. The prior distribution is defined in three steps:

- a prior distribution with cohesions $c(A_k)$ is defined on the set of all the partitions ρ by assuming: $c(A_k) = (n_{A_k} - 1)!/m^{n_{A_k}-1}$, where n_{A_k} is the number of objects in set A_k and m is a parameter that must be estimated;

- given $\rho = \{A_1, \ldots, A_K\}$, the parameter values $\mu^{A_1}, \ldots, \mu^{A_K}$ are independent with density $p_{A_k}(\mu^{A_k})$, specifically $\mu^{A_k} \sim N(\mu_0, n\sigma_0^2/n_{A_k})$, where μ_0 and σ_0^2 are parameters that have to be estimated;
- given ρ, $\mu_i = \mu^{A_k}$, whenever $i \in A_k$.

Dahl (2009) showed that univariate conjugate DPM models for a known parametric family of distributions, that is $x_i \mid \theta_i \sim p(\cdot|\theta_i)$, $\theta_i \mid G \sim G$, $G \sim DP(\alpha G_0)$, may be expressed as *nonparametric product partition models*, by using an alternative parameterization of $\boldsymbol{\theta}$ and by integrating out analytically the component model parameters. This is also true for the univariate normal-normal DPM model, where $p(\cdot|\theta)$ is the (univariate) normal distribution with mean θ and known variance σ^2 and G_0 is the normal distribution with known mean m and known variance τ^2. In this case, the link with product partition models (and an additional technical condition) allow to apply a deterministic algorithm to find the global maximum a posteriori clustering of the posterior clustering distribution and the maximum likelihood clustering. However, in this setting the variances are required to be constant and known for each component. By contrast, unknown variances are an integral part of our model.

3 Combinatorial Mixtures

With the term *combinatorial mixtures* we refer to a general class of mixture models in which elements of the parameter vector can be shared across any subset of the components, and positive mass is put on every possible combination of sharing patterns. To define this class, consider the mixture model:

$$\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{\omega} \sim p(\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{k=1}^{K^*} \omega_k p(\boldsymbol{x}_i|\boldsymbol{\theta}_k), \qquad i = 1, \ldots, n, \tag{2}$$

where each observation is potentially multivariate, that is $\boldsymbol{x}_i$ is a $(J \times 1)$ column vector, $J \geq 1$. The component-specific parameter $\boldsymbol{\theta}_k = (\theta_k^1, \ldots, \theta_k^D)^T$ is a column vector listing all the D parameters characterizing the component distribution. Here K^* represents the maximum number of components, and it is fixed, though as we will see, the actual number of components K is still an unknown parameter.

For a given d, define $\boldsymbol{\theta}^d = (\theta_1^d, \ldots, \theta_{K^*}^d)$. The combinatorial class includes, for each d, the possibility that any subset of the components shares the same value of the d-th component. It is useful to define $\tilde{\boldsymbol{\theta}}^d = \text{Unique}(\boldsymbol{\theta}^d)$, a U_d-dimensional row vector similar to $\boldsymbol{\theta}^d$ but with the duplicate elements suppressed. Also, for any d and $h = 1, \ldots, U_d$, let $\tilde{\theta}_h^d$ indicate the h-th element of $\tilde{\theta}^d$, and let E_h^d be the set including component indices with the same value as the h-th element. In symbols $E_h^d = \{k \in \{1, \ldots, K^*\} : \theta_k^d = \tilde{\theta}_h^d\}$.

Our formulation encompasses three important special cases commonly encountered in mixture modeling. Each can be identified by considering the values of U_d. The fully component-specific

parameters case occurs when each component has its own parameter vector, that is, no sharing takes place, and for every d, $U_d = K^*$. The common parameter, or completely degenerate, case occurs when all components have the same parameters, that is for every d, n, $U_d = 1$, and effectively $K = 1$. Finally, cases in which each component has both component-specific and common parameters occur when $U_d = K^*$ for some ds and $U_d = 1$ for others.

However, more general scenarios are compatible with our formalization. First, for any $\boldsymbol{\theta}^d$, it is possible to have proper subgroups of shared components, in which case $1 < U_d < K^*$. For the d-th element of the parameter vector, some components share the same value, while others do not. Second, for different $\boldsymbol{\theta}^d$s, subgroups of shared elements can be different.

A simple example can illustrate the generality of the approach. Let the maximum number of groups, K^*, be equal to 5 and the dimensionality of the parameter space, D, be equal to 2. An element of the combinatorial mixture class could be as follows: say parameter vectors $\boldsymbol{\theta}^d$ and $\boldsymbol{\theta}^{d'}$, across the K^* groups, be constrained to be, respectively: $\boldsymbol{\theta}^d = (\alpha_1, \alpha_2, \alpha_1, \alpha_3, \alpha_1)$ and $\boldsymbol{\theta}^{d'} = (\alpha'_1, \alpha'_2, \alpha'_3, \alpha'_2, \alpha'_4)$. In this case:

$$\tilde{\boldsymbol{\theta}}^{\boldsymbol{d}} = \text{Unique}(\boldsymbol{\theta}^{\boldsymbol{d}}) = (\alpha_1, \alpha_2, \alpha_3) \qquad \text{and} \qquad \tilde{\boldsymbol{\theta}}^{\boldsymbol{d'}} = \text{Unique}(\boldsymbol{\theta}^{\boldsymbol{d'}}) = (\alpha'_1, \alpha'_2, \alpha'_3, \alpha'_4),$$

meaning that some elements are shared in each of the dimensions. Specifically, the collection of sets with shared components indices are given by:

$$E_1^d = \{1, 3, 5\}, \quad E_2^d = \{2\}, \quad E_3^d = \{4\},$$

and:

$$E_1^{d'} = \{1\}, \quad E_2^{d'} = \{2, 4\}, \quad E_3^{d'} = \{3\}, \quad E_4^{d'} = \{5\}$$

respectively. Dimensions d and d' have a different number of shared elements, and the sharing occurs in different mixture components.

Prior specifications can proceed by assigning a prior directly to the space of parameters, by allowing for degeneracy along equality constraints. For any d, we can identify all the possible degeneracy patterns in the following way. Consider the $(K^* \times K^*)$ matrix, C^d, showing all the possible pairwise comparisons between elements in $\boldsymbol{\theta}^d$ across the K^* components. Each cell equals either 0 or 1, where: $c_{k,k'} = 0$ iff $\theta_k^d = \theta_{k'}^d$, $c_{k,k'} = 1$ iff $\theta_k^d \neq \theta_{k'}^d$ for $k, k' = 1, \ldots, K^*$, and d fixed. The matrix is symmetric and has 1s on the diagonal. From the corresponding upper triangular matrix one can extract a $(1 \times \frac{K^*(K^*-1)}{2})$ vector obtained juxtaposing the rows of the upper diagnonal matrix, after removal of the diagonal elements. For instance, for $K^* = 3$, one can represent the matrix

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \qquad \text{by the vector} \qquad (1, 0, 1).$$

Each degeneracy pattern for a given d is uniquely identified by a different value of the vector.
The Bell number, $B(K^*)$, represents the number of possible degeneracy patterns. $B(n)$ is defined as the number of partitions of a set A_0 of size n and, therefore, it accounts for all the possible comparisons involving the elements of $\boldsymbol{\theta}^d$ across the K^* groups. For example, for fixed d and $K^* = 3$ we have $B(K^*) = 5$, with possible partitions given by: $\{\{1, 2, 3\}\}$, $\{\{1\}, \{2, 3\}\}$, $\{\{1, 3\}, \{2\}\}$, $\{\{1, 2\}, \{3\}\}$, $\{\{1\}, \{2\}, \{3\}\}$, and corresponding values of the vector given by:

$$(0, 0, 0) \text{ or } (1, 1, 0) \text{ or } (1, 0, 1) \text{ or } (0, 1, 1) \text{ or } (1, 1, 1).$$

For simplicity, we can introduce a single categorical random variable γ^d that captures the values of the vector in the following way:

$$\gamma^d = 000 \text{ or } \gamma^d = 110 \text{ or } \gamma^d = 101 \text{ or } \gamma^d = 011 \text{ or } \gamma^d = 111.$$

For any number of groups, the Bell number is finite, though it could be very large. In the easiest independence set-up, a discrete prior distribution can be assigned to each γ^d, for a given d, for the a priori representation of sharing patterns for the corresponding $\boldsymbol{\theta}^d$ element of the parameter vector. A more detailed illustration of the priors on the remaining parameters is given in the context of normal mixtures in the next section.

4 Combinatorial Mixtures of Normal Distributions

4.1 Univariate Model Specification

The idea of combinatorial mixtures can be detailed considering the case of univariate normal mixtures. Data $\boldsymbol{x}^n = (x_1, \ldots, x_n)$ are assumed to be independent observations from a mixture density with K^* normal components:

$$p(x_i|\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\omega}) = \sum_{k=1}^{K^*} \omega_k N(x_i|\mu_k, \sigma_k^2), \tag{3}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{K^*})$, $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_{K^*}^2)$ are vectors containing component specific parameters, $\boldsymbol{\omega} = (w_1, \ldots, w_{K^*})$ are constrained to be non-negative and to sum to unity and $N(\cdot|\mu, \sigma^2)$ indicates a normal distribution with expectation μ and variance σ^2, $\sigma^2 > 0$. In this case $D = 2$ and $d \in \{m, v\}$, meaning that we allow all patterns of sharing of means and variances, respectively, among the K^* components. If two components share both mean and variance they collapse into one, so the effective number of components K is a function of the pattern.

The class of combinatorial normal mixture models above can be seen as a multi-partition generalization of the product partition model of Crowley (1997), where we have two sets of partitions, one for the means and one for the variances. Our priors will be specified directly on equality events, though this will induce priors on partition sets E_h^d.

Mixtures of Dirichlet Process priors, as typically implemented for normal models, allow for parameters to be clustered in subsets. However, the number of possible patterns is smaller than that of combinatorial mixtures, as single bidimensional parameters, $\theta_i = (\mu_i, \sigma_i^2)$ are shared.

Bayesian inference requires a joint prior distribution on the unknown parameters $(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$. The following shorthand notation is used. Assume y_1, y_2, y_3 are generic univariate random variables, with y_3 assuming values in $\{0, 1\}$. Then:

$$y_1, y_2|y_3 \sim \begin{cases} p(y_1) & \text{if } y_3 = 0 \\ p(y_1)p(y_2) & \text{if } y_3 = 1 \end{cases},$$

stands for:

$$\begin{array}{ll} \text{if} \quad y_3 = 0 & y_1 = y_2 \,|y_3 \;\sim\; p(y_1); \\ \text{if} \quad y_3 = 1 & y_1, y_2|y_3 \;\sim\; p(y_1)p(y_2). \end{array}$$

It is straightforward to extend this to the case where one has three generic random variables, y_1, y_2, y_3, and a categorical one, y_4. Each of the five cases corresponds to one of the possible comparisons between y_1, y_2 and y_3.

Assume $K^* = 3$. A factorization of the model comes from the following assumptions on $(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$:

$$\boldsymbol{w} \;\perp\; (\boldsymbol{\mu}, \boldsymbol{\sigma^2}), \qquad \boldsymbol{w} = (w_1, w_2, w_3), \qquad w_k \geq 0, \; \sum_{k=1}^{3} w_k = 1,$$

$$\boldsymbol{w} \;\sim\; Dir(\boldsymbol{a_0}), \qquad \boldsymbol{a_0} = (a_{1,0}, \ldots, a_{3,0}),$$

$$\boldsymbol{\mu} \;\perp\; \boldsymbol{\sigma^2}, \qquad \boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3), \qquad \boldsymbol{\sigma^2} = (\sigma_1^2, \sigma_2^2, \sigma_3^2),$$

$$\mu_1, \mu_2, \mu_3|\gamma^m, \eta^2 \;\sim\; \begin{cases} N(\mu_1|\, 0, \eta^2) & \text{if } \gamma^m = 000 \\ N(\mu_1|\, 0, \eta^2)N(\mu_2|\, 0, \eta^2) & \text{if } \gamma^m = 110 \\ N(\mu_1|\, 0, \eta^2)N(\mu_2|\, 0, \eta^2) & \text{if } \gamma^m = 101 \\ N(\mu_1|\, 0, \eta^2)N(\mu_3|\, 0, \eta^2) & \text{if } \gamma^m = 011 \\ N(\mu_1|\, 0, \eta^2)N(\mu_2|\, 0, \eta^2)N(\mu_3|\, 0, \eta^2) & \text{if } \gamma^m = 111 \end{cases},$$

$$\gamma^m|\boldsymbol{\pi}^m \;\sim\; \text{Multi}(1, \boldsymbol{\pi}^m),$$

where γ^m is the random variable that represents the sharing patterns for the mean vector (see Section 3), $Dir(\boldsymbol{\alpha})$ indicates a Dirichlet distribution with parameter vector $\boldsymbol{\alpha}$, and Multi$(\cdot|N, \boldsymbol{p})$ indicates a Multinomial distribution with parameters N and $\boldsymbol{p}$. Similarly, for the variances:

$$\sigma_1^2, \sigma_2^2, \sigma_3^2|\gamma^v, c, d \;\sim\; \begin{cases} IG(\sigma_1^2|\, c, d), & \text{if } \gamma^v = 000 \\ IG(\sigma_1^2|\, c, d)IG(\sigma_2^2|\, c, d) & \text{if } \gamma^v = 110 \\ IG(\sigma_1^2|\, c, d)IG(\sigma_2^2|\, c, d) & \text{if } \gamma^v = 101 \\ IG(\sigma_1^2|\, c, d)IG(\sigma_3^2|\, c, d) & \text{if } \gamma^v = 011 \\ IG(\sigma_1^2|\, c, d)IG(\sigma_2^2|\, c, d)IG(\sigma_3^2|\, c, d) & \text{if } \gamma^v = 111 \end{cases},$$

$$\gamma^v|\boldsymbol{\pi^v} \;\sim\; \text{Multi}(1, \boldsymbol{\pi}^v),$$

and:

$$\boldsymbol{\mu} \perp \gamma^v, \qquad \boldsymbol{\sigma^2} \perp \gamma^m,$$

where $\boldsymbol{a_0}$, η^2, $\boldsymbol{\pi}^m$, c, d and $\boldsymbol{\pi}^v$ are known, and $IG(\cdot|\alpha,\beta)$ indicates an inverse-gamma distribution with parameters α and β, $\alpha,\beta > 0$ (using the parameterization in which the mean is $\beta/(\alpha-1)$, $\alpha > 1$). For the distribution of (μ_1,μ_2,μ_3), we assume prior means of 0 and vague variances. However, the approach applies equally to nonzero means.

The following factorization of the joint distribution of all the variables summarizes conditional independence assumptions:

$$p(\boldsymbol{\omega},\boldsymbol{z}^n,\boldsymbol{\theta},\boldsymbol{x}^n) \;=\; p(\boldsymbol{\omega})p(\boldsymbol{\theta})p(\boldsymbol{z}^n|\boldsymbol{\omega})p(\boldsymbol{x}^n|\boldsymbol{\theta},\boldsymbol{z}^n),$$

where: $\boldsymbol{\theta} = (\boldsymbol{\mu},\gamma^m,\boldsymbol{\sigma^2},\gamma^v)$, and:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\mu}|\gamma^m)p(\gamma^m)p(\boldsymbol{\sigma^2}|\gamma^v)p(\gamma^v).$$

These will be used in deriving the full conditional distributions, discussed in the Appendix.

Parameters γ^m and γ^v are connected with the unknown number of mixture components K. From the joint prior distribution of (γ^m,γ^v) it is possible to derive the corresponding prior distribution on K by listing the (γ^m,γ^v) combinations associated with any number of components and summing their probabilities. In the case of this section, this leads to:

$$Pr\{K=1\} = \pi^m_{000}\pi^v_{000},$$

$$Pr\{K=2\} = \pi^m_{000}\pi^v_{110}+\pi^m_{000}\pi^v_{101}+\pi^m_{000}\pi^v_{011}+\pi^m_{110}\pi^v_{000}+\pi^m_{110}\pi^v_{110}+\pi^m_{101}\pi^v_{000}+\pi^m_{101}\pi^v_{101}+\pi^m_{011}\pi^v_{000}+\pi^m_{011}\pi^v_{011},$$

$$\begin{aligned} Pr\{K=3\} \;=\; & \pi^m_{000}\pi^v_{111} + \pi^m_{110}\pi^v_{101} + \pi^m_{110}\pi^v_{011} + \pi^m_{110}\pi^v_{111} + \pi^m_{101}\pi^v_{110} + \pi^m_{101}\pi^v_{011} + \pi^m_{101}\pi^v_{111} + \pi^m_{011}\pi^v_{110} + \\ & + \pi^m_{011}\pi^v_{101} + \pi^m_{011}\pi^v_{111} + \pi^m_{111}\pi^v_{000} + \pi^m_{111}\pi^v_{110} + \pi^m_{111}\pi^v_{101} + \pi^m_{111}\pi^v_{011} + \pi^m_{111}\pi^v_{111}. \end{aligned}$$

Meaningful parametric inference in mixture models requires to tackle the label switching problem (Kadane (1975), Stephens (2000b), Frühwirth-Schnatter (2001)), that is the invariance of the likelihood under relabeling of the mixture components. In a Bayesian context this invariance can lead to symmetric and highly multimodal posterior distributions. The usual practices of summarizing joint posterior distributions by marginal distributions, and estimating quantities of interest by their posterior means, are often inappropriate. However, our main goals here are closer to clustering than to pure parameter estimation. We are mainly interested in a) estimating the number of subgroups in a sample; b) making inferences about the assignment of samples to these subgroups; and c) generating hypotheses about which mechanisms are likely to characterize the subgroups. For this reason, we follow O'Hagan (1997) which proposed to summarize inferences by looking at pairs of observations and counting how often they are assigned to the same mixture component. In this way, one obtains a nearness measure that can be used descriptively or provide the basis for clustering.

We implement a graphical representation of the corresponding relative frequencies matrix, named the "O'Hagan" matrix from now on, to visualize clusters in the data.

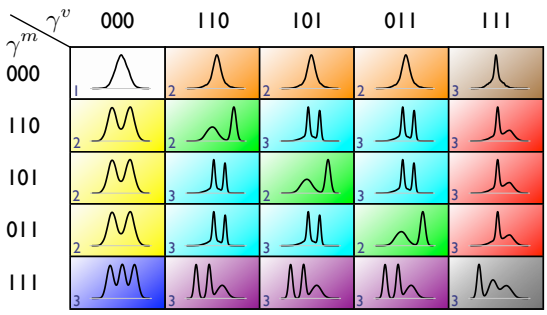
Our solution does not impose constraints on either single parameters or combinations of them. If no constraints are put on the parameters, and combinatorial mixtures are assumed, interesting challenges arise when making inferences on γ^m and γ^v. Apparently different combinations of (γ^m, γ^v) may actually correspond to the same mixture model for the data. Figure 1(a) depicts in a stylized fashion all the possible (γ^m, γ^v) combinations, highlighting equivalent ones with the same color. The original twenty-five possible cases for (γ^m, γ^v) are mapped into ten different ones sharing the same color. Corresponding density estimation plots are superimposed to depict a generic template for how the fitted model should look like, and the correponding number of components is added to each combination. Corresponding prior (posterior) probabilities are summed when reporting results.

Combinatorial mixtures allow for greater generality and flexibility in comparison with traditional approaches to mixture modeling, while still allowing to assign mass to models that are more parsimonious than the general mixture case in which no sharing takes place. Figure 1(b) shows in white the 13 cells available to traditional approaches, and highlights in grey the 12 extra models that combinatorial mixtures allow. The 13 white cells correspond to the following scenarios:

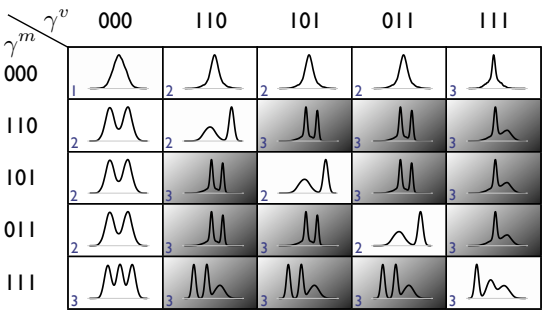
- complete sharing: $(\gamma^m, \gamma^v) = (000, 000)$;
- partial sharing: means:
  - $K = 2$: $(\gamma^m, \gamma^v) = (000, 110)$, $(\gamma^m, \gamma^v) = (000, 101)$, $(\gamma^m, \gamma^v) = (000, 011)$;
  - $K = 3$: $(\gamma^m, \gamma^v) = (000, 111)$;
- partial sharing: variances:
  - $K = 2$: $(\gamma^m, \gamma^v) = (110, 000)$, $(\gamma^m, \gamma^v) = (101, 000)$, $(\gamma^m, \gamma^v) = (011, 000)$;
  - $K = 3$: $(\gamma^m, \gamma^v) = (111, 000)$;
- no sharing:
  - $K = 2$: $(\gamma^m, \gamma^v) = (110, 110)$, $(\gamma^m, \gamma^v) = (101, 101)$, $(\gamma^m, \gamma^v) = (011, 011)$;
  - $K = 3$: $(\gamma^m, \gamma^v) = (111, 111)$.

The 12 grey cells correspond to the following scenarios:

- three-component mixture, with two components sharing the means and two others sharing the variances (light-blue cells in Figure 1(a));
- three-component mixture, with two different means and three different variances (red cells in Figure 1(a));
- three-component mixture, with three different means and two different variances (violet cells in Figure 1(a)).



(a) Same colored cells identify the 10 models with the same shape.



(b) The grey cells identify the 12 extra models that combinatorial mixtures allow to fit, in comparison with traditional approaches to mixture models.

Figure 1: Mixture model templates corresponding to each (γ^m, γ^v) pair for combinatorial mixtures priors in a normal mixture model. The number of components and a typical shape are also noted.

In genomics applications, for example, the additional options can be critical in interpreting cancer clusters, as those may arise from changes in location, scale or correlations, or any of the combinations. In nutritional epidemiology applications, the identification of clusters of subjects having an increased risk of cancer might be more effective if differences in the correlation structure of the nutrients are accounted for directly in the clustering procedure. The usual practice of performing a principal component analysis before a cluster analysis to account for known high correlations between nutrients may be suboptimal (Chang (1983)) and cancer risk/protection associated with the identified clusters may be low due to this problem.

Combinatorial mixtures apply, as a special case, in the supervised context as well. However, the label switching problem does not apply in this case.

Prior independence between parameters is not essential to combinatorial mixtures.

4.2 Bivariate Model Specification

In the following, the idea of combinatorial mixtures is detailed considering the case of bivariate normal mixtures. However, the methodology is generic and applies much more widely. Data $\boldsymbol{x}^n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are assumed to be independent observations from a mixture density with K^* bivariate normal components:

$$\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\omega} \sim \sum_{k=1}^{K^*} \omega_k N_2(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \Sigma_k), \qquad \boldsymbol{x}_i = \begin{bmatrix} x_i^1 \\ x_i^2 \end{bmatrix}, \; i = 1, \ldots, n,$$

where:

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K^*}),\ \boldsymbol{\Sigma} = (\Sigma_1, \ldots, \Sigma_{K^*}), \qquad \boldsymbol{\mu}_k = \begin{bmatrix} \mu_{k1} \\ \mu_{k2} \end{bmatrix},\ \Sigma_k = \begin{bmatrix} \sigma^2_{11_k} & \sigma_{12_k} \\ \sigma_{21_k} & \sigma^2_{22_k} \end{bmatrix},\ k = 1, \ldots, K^*,$$

$$\boldsymbol{w} = (w_1, \ldots, w_{K^*}), \quad w_k \geq 0,\ \sum_{k=1}^{K^*} w_k = 1,$$

and $N_2(\cdot|\boldsymbol{\mu}, \Sigma)$ indicates a bivariate normal distribution with expectation vector $\boldsymbol{\mu}$ and variance-covariance matrix Σ, Σ positive definite matrix. In this case $J = 2$, $D = 5$ and $d \in \{m_1, m_2, v_1, v_2, c\}$, meaning that we allow all patterns of sharing of means, variances and covariances, respectively, among the K^* components.

For the modeling of the variance-covariance structure an attractive approach is proposed by Barnard et al. (2000). They introduced a direct decomposition which separates the standard deviations and correlations and proposed choices for the priors in terms of standard deviations and correlations. This separation has a relevant practical motivation as most practitioners are trained to think in terms of standard deviations and correlations; the standard deviations are on the original scale, and the correlations are scale free. We adopt this approach in the following:

$$\Sigma_k = \text{diag}(\boldsymbol{S}_k) R_k \text{diag}(\boldsymbol{S}_k), \qquad \boldsymbol{S}_k = \begin{bmatrix} S_{k1} \\ S_{k2} \end{bmatrix}, \quad R_k = \begin{bmatrix} 1 & r_k \\ r_k & 1 \end{bmatrix}, \qquad k = 1, \ldots, K^*,$$

where each $\boldsymbol{S}_k$ is a vector of standard deviations, $\text{diag}(\boldsymbol{S}_k)$ is a diagonal matrix with diagonal elements $\boldsymbol{S}_k$, each R_k is a correlation matrix, and accordingly:

$$\boldsymbol{S} = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{K^*}), \qquad \boldsymbol{R} = (R_1, \ldots, \boldsymbol{R}_{K^*}).$$

Assume $K^* = 3$. A factorization of the model comes from the following assumptions on $(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{S}, \boldsymbol{R})$:

$$\boldsymbol{w} \perp (\boldsymbol{\mu}, \boldsymbol{S}, \boldsymbol{R}), \qquad \boldsymbol{\mu} \perp (\boldsymbol{S}, \boldsymbol{R}), \qquad \boldsymbol{S} \perp \boldsymbol{R},$$

$$\boldsymbol{w} \ \sim \ Dir(\boldsymbol{a}_0), \qquad \boldsymbol{a}_0 = (a_{1,0}, a_{2,0}, a_{3,0}),$$

$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3 :\ \boldsymbol{\mu}^j = (\mu_{1j}, \mu_{2j}, \mu_{3j})|\gamma^m_j, \eta^2$ i.i.d. $\sim$

$$\begin{cases} N(\mu_{1_j}|\ 0, \eta^2) & \text{if } \gamma^m_j = 000 \\ N(\mu_{1_j}|\ 0, \eta^2) N(\mu_{2_j}|\ 0, \eta^2) & \text{if } \gamma^m_j = 110 \\ N(\mu_{1_j}|\ 0, \eta^2) N(\mu_{2_j}|\ 0, \eta^2) & \text{if } \gamma^m_j = 101 \\ N(\mu_{1_j}|\ 0, \eta^2) N(\mu_{3_j}|\ 0, \eta^2) & \text{if } \gamma^m_j = 011 \\ N(\mu_{1_j}|\ 0, \eta^2) N(\mu_{2_j}|\ 0, \eta^2) N(\mu_{3_j}; 0, \eta^2) & \text{if } \gamma^m_j = 111 \end{cases}, \qquad j = 1, 2,$$

$$\boldsymbol{\gamma}^m : \ \gamma_j^m | \boldsymbol{\pi}^m \text{ i.i.d. } \sim \text{ Multi}(1, \boldsymbol{\pi}^m), \qquad \boldsymbol{\pi}^m = \begin{bmatrix} P(\mu_{1j} = \mu_{2j} = \mu_{3j}) \\ P(\mu_{1j} \neq \mu_{2j} = \mu_{3j}) \\ P(\mu_{1j} = \mu_{3j} \neq \mu_{2j}) \\ P(\mu_{1j} = \mu_{2j} \neq \mu_{3j}) \\ P(\mu_{1j} \neq \mu_{2j} \neq \mu_{3j}) \end{bmatrix}, \ j = 1, 2,$$

where $\boldsymbol{\gamma}^m$ is the random vector that contains the sharing patterns for $\boldsymbol{\mu}^1$ and $\boldsymbol{\mu}^2$ (see Section 3) and we have posed: $\gamma^{m_j} = \gamma_j^m$, $j = 1, 2$. Similarly, for the standard deviations we choose independent log normal priors having the following structure:

$$\log \boldsymbol{S}_1, \log \boldsymbol{S}_2, \log \boldsymbol{S}_3 : \ \log \boldsymbol{S}^j = (\log S_{1j}, \log S_{2j}, \log S_{3j}) | \gamma_j^v, e^2 \text{ i.i.d. } \sim$$

$$\begin{cases} N(\log S_{1j} | \ 0, e^2), & \text{if } \gamma_j^v = 000 \\ N(\log S_{1j} | \ 0, e^2) N(\log S_{2j} | \ 0, e^2) & \text{if } \gamma_j^v = 110 \\ N(\log S_{1j} | \ 0, e^2) N(\log S_{2j} | \ 0, e^2) & \text{if } \gamma_j^v = 101 \\ N(\log S_{1j} | \ 0, e^2) N(\log S_{3j} | \ 0, e^2) & \text{if } \gamma_j^v = 011 \\ N(\log S_{1j} | \ 0, e^2) N(\log S_{2j} | \ 0, e^2) N(\log S_{3j} | \ 0, e^2) & \text{if } \gamma_j^v = 111 \end{cases}, \ j = 1, 2,$$

$$\boldsymbol{\gamma}^v : \ \gamma_j^v | \boldsymbol{\pi}^v \text{ i.i.d. } \sim \text{ Multi}(1, \boldsymbol{\pi}^v), \qquad \boldsymbol{\pi}^v = \begin{bmatrix} P(S_{1j} = S_{2j} = S_{3j}) \\ P(S_{1j} \neq S_{2j} = S_{3j}) \\ P(S_{1j} = S_{3j} \neq S_{2j}) \\ P(S_{1j} = S_{2j} \neq S_{3j}) \\ P(S_{1j} \neq S_{2j} \neq S_{3j}) \end{bmatrix}, \qquad j = 1, 2,$$

whereas, for each correlation coefficient r_k in R_k, we have:

$$R_1, R_2, R_3 : \ \boldsymbol{r} = (r_1, r_2, r_3) | \gamma^c \ \sim \ \begin{cases} UN_{(-1,1)}(r_1), & \text{if } \gamma^c = 000 \\ UN_{(-1,1)}(r_1) \, UN_{(-1,1)}(r_2), & \text{if } \gamma^c = 110 \\ UN_{(-1,1)}(r_1) \, UN_{(-1,1)}(r_2), & \text{if } \gamma^c = 101 \\ UN_{(-1,1)}(r_1) \, UN_{(-1,1)}(r_3), & \text{if } \gamma^c = 011 \\ UN_{(-1,1)}(r_1) \, UN_{(-1,1)}(r_2) \, UN_{(-1,1)}(r_3), & \text{if } \gamma^c = 111 \end{cases},$$

$$\gamma^c | \boldsymbol{\pi}^c \ \sim \ \text{Multi}(1, \boldsymbol{\pi}^c), \qquad \boldsymbol{\pi}^c = \begin{bmatrix} P(r_1 = r_2 = r_3) \\ P(r_1 \neq r_2 = r_3) \\ P(r_1 = r_3 \neq r_2) \\ P(r_1 = r_2 \neq r_3) \\ P(r_1 \neq r_2 \neq r_3) \end{bmatrix},$$

and we assume independence between each component-specific parameter, $\boldsymbol{\mu}$, $\boldsymbol{S}$ and $\boldsymbol{r}$, and the $\boldsymbol{\gamma}$s corresponding to the other ones.

Finally, $\boldsymbol{a_0}$, η^2, $\boldsymbol{\pi}^m$, e^2, $\boldsymbol{\pi}^v$, $\boldsymbol{\pi}^c$ are known, $\log(Y | \mu, \sigma^2) \ \sim \ N(y | \mu, \sigma^2)$ indicates that Y is distributed according to a lognormal distribution with parameters μ and σ^2, and $UN_{(a,b)}(\cdot)$ indicates a uniform

distribution on (a,b). Note that for R_k be positive definite, $r_k \neq \pm 1$. Prior means of 0 are assumed for the normal and lognormal distributions of $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3)$ and $(\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3)$, respectively, as the corresponding variances are chosen to express a vague prior information. However, the approach applies equally to nonzero means.

The following factorization of the joint distribution of all the variables summarizes conditional independence assumptions:

$$p(\boldsymbol{\omega}, \boldsymbol{z}^n, \boldsymbol{\theta}, \boldsymbol{x}^n) = p(\boldsymbol{\omega})p(\boldsymbol{\theta})p(\boldsymbol{z}^n|\boldsymbol{\omega})p(\boldsymbol{x}^n|\boldsymbol{\theta}, \boldsymbol{z}^n),$$

where: $\boldsymbol{\theta} = (\boldsymbol{\mu}^1, \gamma_1^m, \boldsymbol{\mu}^2, \gamma_2^m, \boldsymbol{S}^1, \gamma_1^v, \boldsymbol{S}^2, \gamma_2^v, \boldsymbol{r}, \gamma^c)$, and:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\mu}^1|\gamma_1^m)p(\gamma_1^m)p(\boldsymbol{\mu}^2|\gamma_2^m)p(\gamma_2^m)p(\boldsymbol{S}^1|\gamma_1^v)p(\gamma_1^v)p(\boldsymbol{S}^2|\gamma_2^v)p(\gamma_2^v)p(\boldsymbol{r}|\gamma^c)p(\gamma^c),$$

and will be used in deriving the full conditional distributions, discussed in the Appendix.

Traditional bivariate mixture models do not generally assume any decomposition of the variance-covariance matrix. The available traditional options are either sharing the variance-covariance matrix across components or no sharing of this matrix across components. The first option implies sharing of both variances and covariances, and then $\gamma^c = 000$ is always automatically selected when both $\gamma_1^v = 000$ and $\gamma_2^v = 000$ occur. The second option includes all the possible remaining cases, and is very general. The adoption of a decomposition allows to share one or both the standard deviations, without necessarily sharing the correlations (or viceversa), and represents an easy way to guarantee a freerer modeling of the variance-covariance structure in traditional mixture models too. When a decomposition is assumed for the variance-covariance matrix, available traditional options are still the following ones: complete sharing of the variances/no sharing of the variances and complete sharing of the covariances/no sharing of the covariances across components. In any of these cases, all the variances (covariances) should be either equal or different across components. Combinatorial mixtures allow to go further in the direction of modeling each variance or correlation in a freerer way, at the expense of dealing with extra complexity in model specification. In detail, the bivariate normal extension might allow to model an interesting phenomenon observed in microarray analysis when two variables have the same mean and variance but opposite correlations in diseased and normal samples (Dettling et al. (2005)).

4.3 Computation

Bayesian inference for mixture models may be performed using MCMC methods, generating iteratively the parameters and the missing data. Diebolt and Robert (1994) described it for the case where the number of mixtures components is known. We gave an overview of methods for our case, where the number of components is unknown, in Subsection 2.3.
As we assume priors that are mixtures of mutually singular distributions, we also refer to Gottardo and Raftery (2008), which deals specifically with MCMC methods for cases where the target distribution is a mixture of mutually singular distributions. The authors introduced a framework

for such measures to which the general theory of MCMC applies, as it does to general dominating measures. The idea is to find a common dominating measure that allows the use of traditional Metropolis-Hastings algorithms, and then Gibbs sampler, if the full conditionals are available. We will apply this to mixture models where priors are mixtures of singular distributions.

The Markov chain $\left\{\Phi^{(t)}\right\}$ with posterior distribution $p(\boldsymbol{\phi}|\boldsymbol{x}^n)$ as its stationary distribution is constructed in the following way. A sampled realization of the Markov chain is produced generating iteratively the parameters, $\boldsymbol{\phi}^{(t)}$, and the missing data, $\boldsymbol{z}^{n(t)} = (z_1^{(t)}, \ldots, z_n^{(t)})$ according to $p(\boldsymbol{\phi}|\boldsymbol{x}^n, \boldsymbol{z}^{n(t)})$ and $p(\boldsymbol{z}^n|\boldsymbol{x}^n, \boldsymbol{\phi}^{(t+1)})$ respectively.

For the univariate normal mixture model, we have: $p(\boldsymbol{\phi}|\boldsymbol{x}^n) = p(\boldsymbol{\mu}, \gamma^m, \boldsymbol{\sigma^2}, \gamma^v, \boldsymbol{\omega}|\boldsymbol{x}^n)$ and $\boldsymbol{\phi}^{(t)} = (\boldsymbol{\mu}^{(t)}, \gamma^{m(t)}, \boldsymbol{\sigma^2}^{(t)}, \gamma^{v(t)}, \boldsymbol{\omega}^{(t)})$, and we make use of four move types:

1. updating the vector $(\boldsymbol{\mu}, \gamma^m)$;
2. updating the vector $(\boldsymbol{\sigma^2}, \gamma^v)$;
3. updating the weights $\boldsymbol{\omega}$;
4. updating the allocation variables vector $\boldsymbol{z}^n$.

Full conditional distributions of the variables given all the others exist in a closed form and are introduced in the Appendix. Move types 1. and 2. follow from combinatorial mixtures assumptions. They involve a change in dimension. In detail, $(\boldsymbol{\mu}, \gamma^m)$ $(\boldsymbol{\sigma^2}, \gamma^v)$ are drawn according to the described strategy. At the t-th iteration, $t \neq 1$:

1. sample the parameter regulating the dimension of the parameter space, $\gamma^{m(t)}$, $\gamma^{v(t)}$;
2. conditioning on $\gamma^{m(t)}$, $\gamma^{v(t)}$ (and on all the remaining parameters), sample the corresponding location parameter, $\boldsymbol{\mu}^{(t)}$, $\boldsymbol{\sigma^2}^{(t)}$.

Updating $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ for fixed values of γ^m, γ^v, and the remaining parameters, makes use of conjugacy and follows Diebolt and Robert (1994). Updating γ^m, γ^v requires integrating out the corresponding location parameters in the vector, to get the posterior probability associated with each possible value of the γs. Move types 3. and 4. are now standard, largely following Diebolt and Robert (1994) as well.

For the bivariate normal mixture model, we have:
$p(\boldsymbol{\phi}|\boldsymbol{x}^n) = p(\boldsymbol{\mu}^1, \gamma_1^m, \boldsymbol{\mu}^2, \gamma_2^m, \boldsymbol{S}^1, \gamma_1^v, \boldsymbol{S}^2, \gamma_2^v, \boldsymbol{r}, \gamma^c, \boldsymbol{\omega}|\boldsymbol{x}^n)$ and:
$\boldsymbol{\phi}^{(t)} = (\boldsymbol{\mu}^{1(t)}, \gamma_1^{m(t)}, \boldsymbol{\mu}^{2(t)}, \gamma_2^{m(t)}, \boldsymbol{S}^{1(t)}, \gamma_1^{v(t)}, \boldsymbol{S}^{2(t)}, \gamma_2^{v(t)}, \boldsymbol{r}^{(t)}, \gamma^{c(t)}, \boldsymbol{\omega}^{(t)})$, and we make use of seven move types:

1. updating the vectors $(\boldsymbol{\mu}^j, \gamma_j^m)$, $\quad j = 1, 2$;
2. updating the vectors $(\boldsymbol{S}^j, \gamma_j^v)$, $\quad j = 1, 2$;

3. updating the vector $(\boldsymbol{r}, \gamma^c)$;

4. updating the weights $\boldsymbol{\omega}$;

5. updating the allocation variables vector $\boldsymbol{z}^n$.

Full conditional distributions of the variables given all the others exist in a closed form for $(\boldsymbol{\mu}^j, \gamma_j^m)$, $j = 1, 2$, $\boldsymbol{\omega}$ and $\boldsymbol{z}^n$. They are introduced in the Appendix.

Move types 1. 2. and 3. follow from combinatorial mixtures assumptions and involve a change in dimension. We applied the Gibbs sampler for drawing $(\boldsymbol{\mu}^j, \gamma_j^m)$, $j = 1, 2$, according to the same strategy described in the univariate case. We applied three Metropolis-Hastings steps for drawing $(\boldsymbol{S}^j, \gamma_j^m)$, $j = 1, 2$, and $(\boldsymbol{r}, \gamma^c)$. We referr to the Appendix for details. Move types 4. and 5. still follow Diebolt and Robert (1994).

4.4 Gene Expression in Lung Cancer

Next, the normal mixture models of Subsections 4.1 and 4.2 are applied to data on the molecular classification of lung cancer. DNA microarrays are part of a class of biotechnologies that allows the monitoring of thousands of genes simultaneously under different biological or experimental conditions. They may be used to characterize the molecular variation among tumors. This may lead to a more reliable classification of lung cancers and to the identification of potentially promising genes for that classification. Using gene array measurements of expression profiles, several groups have reported findings suggesting that distinctive molecular profiles could lead to refinement of classification and prognostication of lung cancer (Garber et al. (2001), Beer et al. (2002), Bhattacharjee et al. (2001), Miura et al. (2002) and Wigle et al. (2002)).
Here we consider a dataset from the web-based information supporting the published manuscript Garber et al. (2001). The study is performed by scientists at the Dana-Farber Cancer Institute and the Massachusetts Institute of Technology and used Affymetrix oligonucleotide arrays Hu95A representing 12,600 transcripts to profile 203 samples, including 186 lung tumor samples of various histologic patterns and 17 normal samples. For our purposes, the normal samples were removed from the study. A selection is made on the total number of genes and **four** biologically promising candidates are identified for further analyses. They were selected as genes involved in the distinction between BRCA-1 and sporadic (with no detected mutation in BRCA-1 gene) breast cancers. Their HUGO names are: "ITGB5", "MSN", "TRIM29" and "CSTB".

TRIM29 (tripartite motif-containing 29) may act as a transcriptional regulatory factor involved in carcinogenesis and/or differentiation. Its basic function is mediating estrogen action in various target organs. It is reported to be involved in prostate carcinoma, where is low expressed. It may also function in the suppression of radiosensitivity. MSN or Moesin (for membrane-organizing extension spike protein) is important in cell-cell recognition and signaling and for cell movement. It is found to be involved in a case of ALK+ anaplasic large cell lymphoma. It is reported to have a role in discriminating between patients susceptible to locoregional lymph node metastasis and other patients, in oral squamous cell carcinoma. ITGB5 (integrin beta 5) codes for a protein involved in cell-extracellular adhesion and cell-cell-adhesion. It is implicated in cellular processes like cell communication and motility. Less is known about involvement of ITGB5 in cancer, except

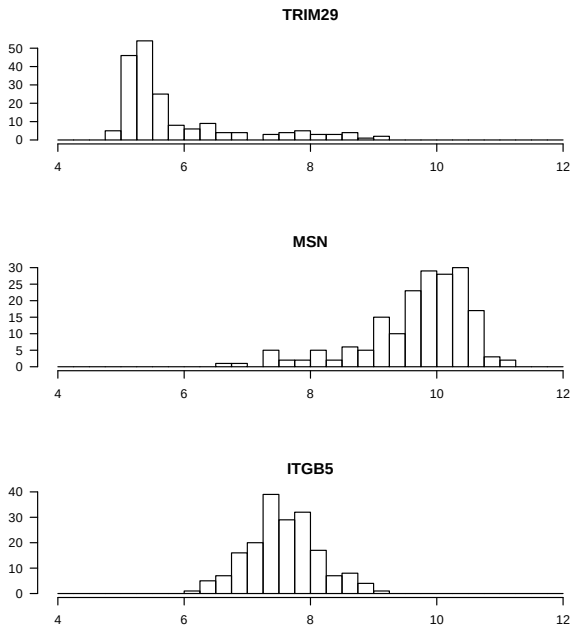


Figure 2: Histograms of the expression levels for the genes under study.

for squamous cell carcinoma of the toungue, though it is involved in vascular development, a process critical for cancer growth. CSTB (cystatin B) and other cystatins regulate tumour-associated cysteine proteases are reported to be upregulated in non-small cell lung tumours, and thus able to counteract harmful tumour-associated proteolytic activity. CSTB is, specifically, overexpressed in most hepatocellular carcinomas and is also elevated in the serum of a large proportion of hepatocellular carcinoma patients.

We next fit two combinatorial mixture models using these four variables. We present separate univariate models for TRIM29, MSN and ITGB5. Results for CSTB are similar to the ones for ITGB5. We then present an application of the bivariate model to the MSN and CSTB genes. Calculations are performed using the open-source statistical computing environment R (R Development Core Team (2006), R Development Core Team (2008), Ihaka and Gentleman (1996)), its library MCMCpack (Martin and Quinn (2005)) and a specialized code reflecting the procedure described in Subsection 4.3 and the results summarized in the Appendix.
We consider Bayesian estimation in the case where we do not have strong prior information on the parameters. There are cases where subjective priors are preferable, and our prior setting could be modified accordingly. However, it seems that for most purposes of the model there is a case for keeping to the simplest independence prior structure for the means, variances/standard deviations and correlations, and defining weakly informative priors.

Because of the label switching problem, we do not report posterior estimates of individual parameters. We present the estimated joint distributions (%) of (γ^m, γ^v) for each gene, and the marginal distribution of γ^c in the bivariate case, as those parameters are not involved in the label switching problem. We marked in boldface the frequencies of those combinations that are only

available with combinatorial mixtures, in comparison with traditional approaches. Moreover, we report the "O'Hagan" matrices to check directly on clusters in the data. Relative frequencies of the co-occurrence of two samples in the same group are plotted in a black-to-white color scale and sorted according to a non-decreasing ordering of the raw data (non-decreasing ordering of the first variable, in the case of bivariate data). Dark and light blocks, with proportions similar to the estimated weights of the mixture, identify different groups of observations.

4.4.1 Application of the Univariate Model

Figure 2 shows the univariate distribution of the expression levels for the three genes considered in this application. TRIM29 and MSN seem to be promising for lung cancer classification as well. TRIM29 has a long right tail suggesting the presence of more than one group of patients. MSN has an interesting left tail too. On the other hand, ITGB5 distribution seems to be similar to a single univariate Normal.

The following choice of hyperparameters results in weakly informative priors on the means μ_1, μ_2, μ_3 and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$: $\eta^2 = 5000$, $c = 0.75$, $d = 0.15$. The Dirichlet prior on the weights, $\boldsymbol{\omega}$, is symmetric with $\boldsymbol{a_0} = (1, 1, 1)$. Finally, there are different ways of being vague in assigning prior distributions on the γs. The Multinomial priors are vague in such a way that they give the same a priori probability to all the possible values of $\boldsymbol{\pi}^m$ and $\boldsymbol{\pi}^v$ respectively: $\boldsymbol{\pi}^m = \boldsymbol{\pi}^v = (0.2, 0.2, 0.2, 0.2, 0.2)$. Different hyperparameters scenarios will be considered in the sensitivity analysis presented in Subsection 4.4.3.

For each gene, we run one chain at a time, doing several simulations. We report results from one of these chains, corresponding to a total of 20,000 iterations with a burn-in of 2,000 iterations. From visual inspection of the chains, we conclude that these numbers are adequate for reliable results. All our runs start with $\gamma^m = 000$ and $\gamma^v = 000$.

The estimated joint posterior probabilities of (γ^m, γ^v) for each gene are shown in Figure 3. The TRIM29 chain spends 40.12% of the iterations on either $(111, 110)$ or $(111, 101)$ or $(111, 001)$ (violet cells). The corresponding mixture model has three components, with three different means and two variances. The second most probable pair is $(111, 111)$ (grey cell), which represents a three-component mixture with different means and different variances. The third most probable combination, $(110, 110)$ or $(101, 101)$ or $(011, 011)$ (green cells), implies a mixture model with two components, one of which comes from sharing both means and variances between two components. A posteriori a mixture model with either two or three components with different means and variances seems suitable for TRIM29. The chain spends about 48% of the iterations on the 12 combinations of (γ^m, γ^v) that would not be possible to select using traditional approaches to mixture models. The majority of the 48% is due to the most probable combination (violet cells). As traditional approaches may place a larger portion of mass on the $(111, 111)$ combination, combinatorial mixtures allow for a gain in parsimony in this case.

The MSN chain spends 44.25% of the iterations on a mixture model with two effective components, one of which comes from complete sharing between two components (green cells). The second most probable combination is given by either $(110, 111)$ or $(101, 111)$ or $(001, 111)$ (red cells), which represent a three-component mixture model with two different means and three different variances.

γ^m \ γ^v	000	110	101	011	111
000	0.00	0.00	0.00	0.00	0.00
110	0.00	5.93	**0.18**	**0.00**	**1.73**
101	0.00	**0.21**	7.29	**0.21**	**2.40**
011	0.00	**0.27**	**0.28**	5.53	**2.10**
111	0.00	**20.37**	**15.67**	**4.07**	33.62

(a) TRIM29

γ^m \ γ^v	000	110	101	011	111
000	0.00	0.00	0.00	0.00	0.00
110	1.90	12.35	**2.98**	**3.01**	**7.95**
101	1.15	**2.42**	13.95	**2.80**	**8.43**
011	1.69	**4.31**	**3.85**	17.95	**11.57**
111	0.20	**0.83**	**0.75**	**0.89**	0.92

(b) MSN

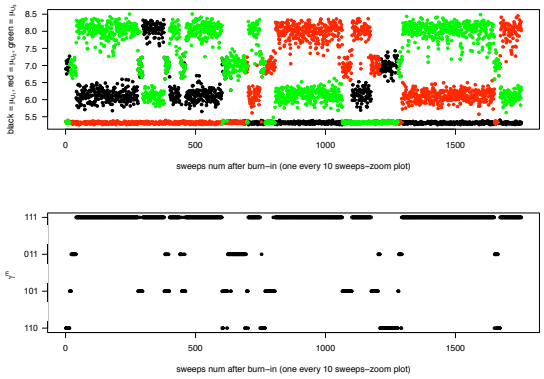
γ^m \ γ^v	000	110	101	011	111
000	26.05	17.55	18.70	18.00	13.12
110	0.63	0.35	**0.39**	**0.37**	**0.32**
101	0.59	**0.40**	0.44	**0.37**	**0.23**
011	0.63	**0.41**	**0.50**	0.44	**0.39**
111	0.03	**0.00**	**0.02**	**0.02**	0.00

(c) ITGB5

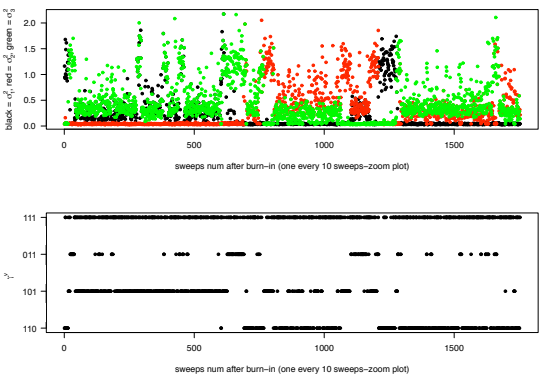
Figure 3: Joint posterior probabilities (%) of (γ^m, γ^v) for the three genes as obtained by one run of the simulation. For correct interpretation, one needs to sum frequencies in cells with the same color. The frequencies of those cases that are gained using combinatorial mixtures, in comparison with traditional approaches, are emphasized using boldface.

The third most probable combination happens on the six light-blue cases: $(\gamma^m, \gamma^v) = (110, 101)$, or $(\gamma^m, \gamma^v) = (110, 011)$, or $(\gamma^m, \gamma^v) = (101, 110)$, or $(\gamma^m, \gamma^v) = (101, 011)$, or $(\gamma^m, \gamma^v) = (011, 110)$, or $(\gamma^m, \gamma^v) = (011, 101)$. The corresponding mixture has three components, two of them sharing means while two others share the variances. A two- or three-component mixture model seems suitable for MSN. The chain spends about 50% of the iterations on the 12 combinations of (γ^m, γ^v) that would not be possible to select using traditional approaches. The majority of this 50% is due to the second and the third most probable combinations (red and light-blue cells). Compared to traditional two-component mixtures, combinatorial mixtures allow to fit the extra case of three different variances for two components with different means, thus achieving extra flexibility in modeling the variances. Compared to traditional three-component mixtures, combinatorial mixtures allow for a more parsimonious solution.

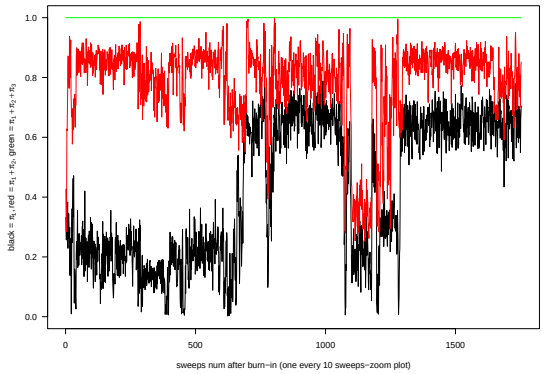
The ITGB5 chain spends 54.25% of the iterations on $(000, 110)$ or $(000, 101)$ or $(000, 011)$, which imply a mixture model with two components having the same mean but two different variances (orange cells). The second most probable case is $(000, 000)$ (white cell), corresponding to a single univariate normal distribution, while the third one is $(000, 111)$ (brown cell), corresponding to a three-component mixture model with one shared mean and three different variances. The chain spends almost all the iterations on $\gamma^m = 000$ and groups are eventually created only by differences in variances. The chain spends about 4% of the iterations on the 12 combinations of (γ^m, γ^v) that would not be possible to select using traditional approaches. Each cell contributes almost equally



(a) μ_1, μ_2, μ_3 plotted along with γ^m



(b) $\sigma_1^2, \sigma_2^2, \sigma_3^2$ plotted along with γ^v



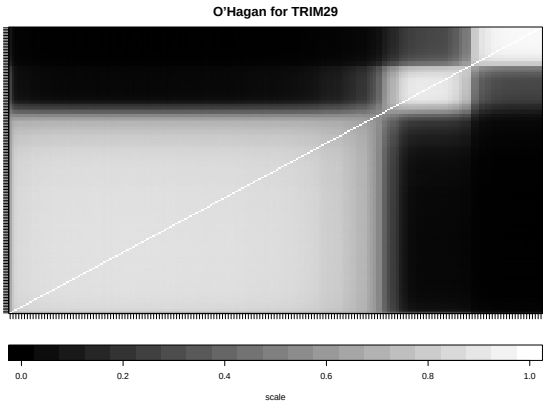
(c) $\omega_1, \omega_2, \omega_3$: adaptation of the stacked area plot

Figure 4: Marginal chains of the parameters for TRIM29 as obtained by the simulation. Component means and corresponding variances and weights for each iteration are color-coded in the same way across plots.

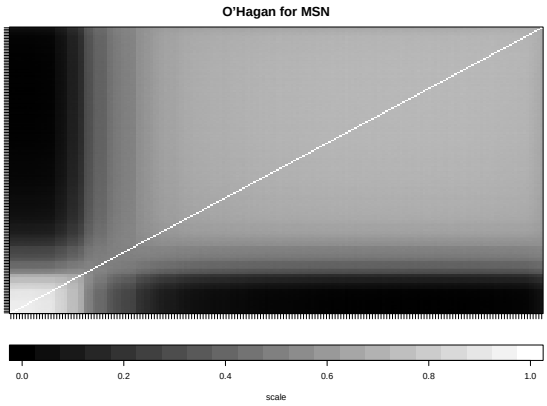
to this 4%. Traditional approaches are not restrictive in this case.

Additional results for TRIM29 are in Figure 4. The estimated marginal posterior distributions of the parameters μ_1, μ_2, μ_3 and $\sigma_1^2, \sigma_2^2, \sigma_3^2$, respectively, are plotted along with the corresponding γs. One iteration every ten is reported and points smaller than the 0.2th percentile and bigger than the 99.8th percentile are not shown. Component means and corresponding variances and weights for each iteration are represented using the same color across plots. The weights are shown adapting the usual stacked area plot, where each line represents the cumulative sum of the corresponding weights. The estimated posterior distributions are highly compatible with the explorative plots in Subsection 4.4.1. Suppose (γ^m, γ^v) at iteration t implies a three-component mixture, with three different means and two different variances ((111, 110) or (111, 101) or (111, 001)). According to the plots, the three means are estimated at 5.4, 6 and 8, while the corresponding variances are 0.09 for the first component and 0.4 for the remaining ones. If $(\gamma^m, \gamma^v) = (111, 111)$ at the next iteration, a third higher variance is estimated, as the corresponding mixture model has three components with

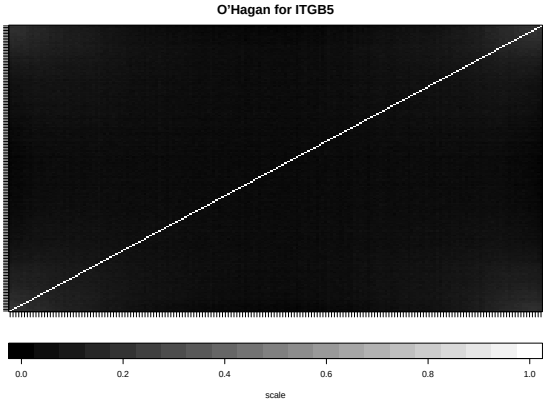
different variances. The component with mean 5.4 and variance approximately 0.09 has a weight close to 0.7. The remaining ones have weights equal to 0.15 respectively. If γ^m and γ^v assume the same intermediate value ((γ^m, γ^v) equal to $(110, 110)$, or $(101, 101)$, or $(011, 011)$), estimated means are approximately at 5.4 and 7 and corresponding variances between 1 and 1.5. Similar considerations hold for the marginal chains of the parameters in the case of MSN and ITGB5 genes.



(a) TRIM29



(b) MSN



(c) ITGB5

Figure 5: "O'Hagan" matrices for the three genes. Relative frequencies of occurrence of two being in the same group are plotted in a black-to-white color scale and sorted according to a non-decreasing ordering of the raw data. Dark and light blocks, with proportions similar to the estimated weights of the mixture, identify different groups of observations.

Figure 5 depicts the "O'Hagan" matrices for each of the three genes. In the case of TRIM29, we identify three groups of patients. The first on the left is the one with the smallest mean and variance and the highest weight. The intermediate block represents the component with a mean of approximately 6. The transition from the first to the second component is very smooth indicating uncertainty in the classification of intermediate points. The component on the right of the picture is the one with mean around 8. The transition from the second to the third component is less smooth than the previous one. Both the second and the third group have a weight of 0.15. In the

case of MSN, we identify two groups of patients. From the left, the first is well defined and has mean approximately at 8.5, variance equal to 0.8 and the smallest weight (data not shown). The transition from the first to the second component is very smooth. The component on the right of the picture has a mean at about 10 and weight at about 0.85 (data not shown). The group is not so well defined. No groups are evident for ITGB5.

4.4.2 Application of the Bivariate Model

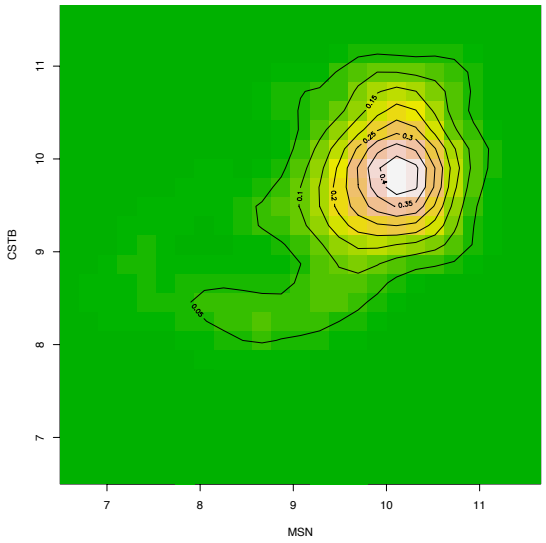


Figure 6: Image of a two-dimensional kernel density estimation plot for the bivariate distribution of the MSN and CSTB genes. Contour lines are added to the existing image.

In the following, we propose the results of the application of the bivariate combinatorial mixture model to the MSN and CSTB genes. Figure 6 shows the image of a two-dimensional kernel density estimation plot representing the bivariate distribution of these genes. The kernel density estimation is obtained with an axis-aligned bivariate normal kernel and evaluated on a square grid. Contour lines are added to the existing image. The plot suggests the existence of two or more groups.

The following choice of hyperparameters results in weakly informative priors on the means, standard deviations and correlations, respectively: $\eta^2 = 5000$, $e^2 = 4$, $a = -1, b = 1$. The Dirichlet prior on the weights, $\boldsymbol{\omega}$, is still symmetric with $\boldsymbol{a_0} = (1, 1, 1)$. The Multinomial priors are such that: $\boldsymbol{\pi}^m = \boldsymbol{\pi}^v = \boldsymbol{\pi}^c = (0.2, 0.2, 0.2, 0.2, 0.2)$.

We run one chain for our data, doing several simulations. We report results from one of the these chains, corresponding to a total of 40,000 iterations with a burn-in of 2,000 iterations. We choose the starting values for each run according to a preliminary k-means clustering algorithm. The number of groups for the k-means is specified looking at some exploratory plots. If there is no evidence of more than one group in the data according to all the possible criteria (differences in means, standard deviations, correlations), the starting points are chosen in the following way:

$\gamma_j^m = \gamma_j^v = \gamma^c = 000$, $j = 1, 2$, and means, standard deviations and correlations, respectively, are assumed to be equal to the corresponding sample values.

γ_1^m \ γ_1^v	000	110	101	011	111
000	0.00	0.00	0.00	0.00	0.00
110	0.01	3.77	**1.82**	**0.61**	**37.57**
101	0.02	**1.02**	5.87	**0.20**	**38.69**
011	0.00	**0.11**	**0.04**	1.00	**5.10**
111	0.00	**0.18**	**0.16**	**0.01**	3.82

(a) MSN ($j = 1$)

γ_2^m \ γ_2^v	000	110	101	011	111
000	0.00	0.00	0.13	0.09	3.97
110	0.02	0.10	**0.26**	**0.91**	**4.79**
101	1.03	**5.32**	6.77	**7.02**	**51.90**
011	0.13	**0.81**	**1.15**	1.02	**10.84**
111	0.02	**0.11**	**0.16**	**0.58**	2.85

(b) CSTB ($j = 2$)

γ^c	000	110	101	011	111
	6.64	15.96	6.10	31.09	40.20

(c) correlation between MSN and CSTB

Figure 7: Joint posterior probabilities (%) of (γ_j^m, γ_j^v), $j = 1, 2$, and posterior probability (%) of γ^c, as obtained by one run of the simulation. For correct interpretation, one needs to sum frequencies in cells with the same color. The frequencies of those cases that are gained using combinatorial mixtures, in comparison with traditional approaches, are emphasized using boldface.

The estimated joint posterior probabilities (%) of (γ_j^m, γ_j^v), $j = 1, 2$, and the estimated posterior probability (%) of γ^c are shown in Figure 7. The MSN marginal chain spends more than 80% of the iterations on either $(110, 111)$ or $(101, 111)$ or $(001, 111)$ (red cells). The second most probable combination is: $(110, 110)$ or $(101, 101)$ or $(011, 011)$ (green cells). A posteriori a mixture model with either two or three components with different means and standard deviations seems suitable for MSN. The CSTB marginal chain spends more than 65% of the iterations on either $(110, 111)$ or $(101, 111)$ or $(001, 111)$ (red cells) too. The second most probable combination happens on the six light-blue cases: $(\gamma^m, \gamma^v) = (110, 101)$, or $(\gamma^m, \gamma^v) = (110, 011)$, or $(\gamma^m, \gamma^v) = (101, 110)$, or $(\gamma^m, \gamma^v) = (101, 011)$ or $(\gamma^m, \gamma^v) = (011, 110)$, or $(\gamma^m, \gamma^v) = (011, 101)$. A posteriori a mixture model with three components with different means and standard deviations seems suitable for CSTB.

In both cases, there is not so much evidence in favor of those cases allowed by traditional mixture models. The chain spends more than 80% of the iterations on the 12 combinations of (γ_1^m, γ_1^v) and (γ_2^m, γ_2^v) that would not be possible to select using traditional approaches to mixture models. The majority of this 80% is due to the most probable combinations (red and light-blue cells). Combinatorial mixtures might end up either in extra flexibility in modeling standard deviations or in a more parsimonious solution, depending if one assumes the traditional two-component or the three-component mixture for comparison.

Finally, the chain spends almost all of the iterations either on the intermediate cases of two different correlations (pink cells) or on the case of three different correlations across components (dark-green cells). Traditional mixture models where a decomposition of the variance-covariance matrix is assumed allow to fit all the five cases listed in Table 7(c), depending on how many different components are assumed; so, no cases are emphasized using boldface.

Figure 8 shows the estimated marginal posterior distributions of the parameters $\boldsymbol{\mu}^1$, $\boldsymbol{\mu}^2$, $\boldsymbol{S}^1$, $\boldsymbol{S}^2$ and $\boldsymbol{r}$, along with the corresponding γs, and the weights $\boldsymbol{\omega}$. One iteration every ten is reported, and points smaller than the 5th percentile and bigger than the 95th percentile are not shown. Corresponding estimates for each iteration are represented using the same color across plots. The MSN gene shows two means that are around 8.3 and 10. Either sharing of the 8.3 mean ($\gamma_1^m = 110$) or sharing of the 10 mean ($\gamma_1^m = 101$) are allowed. If $\gamma_1^m = 110$, the 8.3 means are associated with the two bigger standard deviations (0.6 and 0.8) and the 10 mean with the smaller one of 0.4. If $\gamma_1^m = 101$, the corresponding standard deviations are around 0.8 for the 8.3 mean and 0.4, 0.6 for the 10 means. In any case, the collapsed means admit two standard deviations, one of which is able to cover the arm that connects the two components with different means (see Figure 6). The CSTB gene shows two means that are around 8.6 and 9.9. The 8.6 mean is generally not shared across components. The corresponding standard deviations are around 0.3-0.5, and 0.6 and 0.8, respectively. The component with mean around 10 accounts for the connecting cluster in Figure 6. If $\gamma^c = 111$, the component correlations are 0.1, -0.4 and 0.8. If $\gamma^c = 011$, the component correlations are 0.2 and 0.8, with 0.2 coming from sharing of correlations between the red and the black components. The weights are shown in terms of cumulative sums. The component with mean vector around (10,10) and the smallest standard deviations has a correlation of 0.1 and a weight close to 0.65. The component with mean vector around (8.3,8.6), the smallest standard deviations and correlation equal to -0.4 has a weight close to 0.20. The remaning weight of around 0.15 belongs to the the connecting cluster. This cluster is fitted using the biggest standard deviation corresponding to the shared means for MSN, the component with mean equal to 10 and standard deviation equal to 0.8 for CSTB, and the remaining correlation of 0.8.

The "O'Hagan" matrix is reported in Figure 9. The plot identifies two distinct groups with different mean vectors. The first group on the left has the smallest mean vector and its weight is equal to 0.20. The transition from the first to the second group reveals a great uncertainty in the classification of intermediate points. These points are likely to belong to the connecting cluster in Figure 6, which has a weight of 0.15. The group on the right of the picture is the one with mean vector around (10,10) and the biggest weight.

4.4.3 Sensitivity to Model Specification and Other Issues

Reporting substantive results requires exploring sensitivity to model specification, especially regarding the prior, and verifying the proper convergence of the MCMC. Sensitivity of the posterior distribution of both the γs and the remaining parameters needs to be investigated. For an extensive treatment on sensitivity analysis in the context of normal mixture models with an unknown number of components, see Richardson and Green (1997). For further development of sensitivity analysis issues for those models, one can refer to Stephens (2000). Our main concern here is to show how the inference on the number of components is affected by different values for $\boldsymbol{\pi}^m$ and $\boldsymbol{\pi}^v$. We do so in terms of the equivalent prior on the number of components, K, introduced in Subsection 4.1. Elicitation of potentially interesting priors is easier, as the induced prior on K allows to define a different way of being vague. Interpretation of the sensitivity analyses may be easier as well. The default *vague* set-up proposed in the analysis, $\boldsymbol{\pi}^m = (0.2, 0.2, 0.2, 0.2, 0.2)$ and $\boldsymbol{\pi}^v = (0.2, 0.2, 0.2, 0.2, 0.2)$, can be interpreted in terms of its implications on the number of chosen subgroups, and it corresponds to: P{1 subgroup}=0.04, P{2 subgroups}=0.36 and P{3 subgroups}=0.60 (say, case 3 hereafter). Three groups are a priori favored by our default set-up.

For this reason, we consider for the sensitivity analysis two other scenarios, which a priori favor either the case of one group or all three possibilities in almost the same way. They are respectively such that:

- $\boldsymbol{\pi}^m = \boldsymbol{\pi}^v = (0.8, 0.05, 0.05, 0.05, 0.05) \leftrightarrow$ P{1 subgroup}=0.64, P{2 subgroups} ≈ 0.25 and P{3 subgroups} ≈ 0.11 (say, case 1 hereafter);
- $\boldsymbol{\pi}^m = \boldsymbol{\pi}^v = (0.6, 0.1, 0.1, 0.1, 0.1) \leftrightarrow$ P{1 subgroup}=0.36, P{2 subgroups}=0.39 and P{3 subgroups}=0.25 (say, case 2 hereafter).

Results of sensitivity analysis to different values of $\boldsymbol{\pi}^m$ and $\boldsymbol{\pi}^v$ are summarized in Figure 10 for each gene. Each table presents posterior probabilities of observing one, two or three groups, respectively, given the three different a priori set-ups. The prior probabilities corresponding to each scenario are added in the left column for comparison. The sensitivity of the inference on the number of components to the prior specification of $\boldsymbol{\pi}^m$ and $\boldsymbol{\pi}^v$ is low.

An essential element of the performance of the MCMC is its ability to move between different values of γ^m and γ^v, that is to mix over the number of components. A plot of the changes in γ^m and γ^v against the number of sweeps for TRIM29 is presented in Figure 4(c) and Figure 4(b) respectively. It shows that the MCMC mixes well over the γs. Similar plots were obtained for MSN and ITGB5. We detected no influence of starting values. Within the range of weak priors that we have been using, we have observed good mixing patterns in all our runs.

5 Discussion

We introduce a class of mixture models that we call combinatorial mixtures for mixture distributions whose components have multidimensional parameters. The approach allows each element of the component-specific parameter vector to be shared by any subset of other components. For any dimension, it is thus possible that some components share the same value of the element, while others do not. Moreover, for different dimensions, subgroups of shared elements can be different, either because of a different cardinality or different shared components. From a Bayesian perspective, prior specification has a key role in building combinatorial mixtures: inference can proceed by assigning a prior directly to the space of parameters, allowing for degeneracy along equality constraints for each element. Although our focus is on Bayesian analysis, combinatorial mixtures can also be useful for non-Bayesian modeling.

Free and partial sharing of components allows for greater generality and flexibility in comparison with traditional approaches to mixture modeling, while still allowing to eventually prefer models that are more parsimonious than the general *no sharing* case. This allows for potentially relevant efficiency gains in the overall estimate procedure. One of the implications of our setting is that, once a maximum number of components is specified, inference on the parameters and the number of components is subsumed by the inference on combinatorial patterns.

Bayesian inference and computation approaches were illustrated in a setting based on the normal

model, and applied to data on molecular subtypes of lung cancer. Because cancer subtypes may be characterized by differences in location, scale, correlations or any of the combinations, effective procedures for addressing simultaneously variable selection and clustering (see, for instance, Kim et al. (2006)) may rely on a flexible set of different criteria. Combinatorial mixtures allow to cover all the possible comparisons between relevant parameters across groups, and are potentially useful for applications in a range of application areas. For more extensive exhaustive application of combinatorial mixtures to both supervised and unsupervised contexts one can refer to Edefonti's Ph.D. Thesis (Edefonti (2006)), which is available upon request.

Combinatorial mixtures generalize both product partition and DPM models. Both parametric and nonparametric product partition models are probability models for the estimation of a set of parameters, a subset of which are allowed to be equal. However, they are usually proposed and implemented focusing on one dimension only. Our class of combinatorial normal mixture models can be seen as a multi-partition generalization of the product partition model of Crowley (1997), where we have two sets of partitions, one for the means and one for the variances. Our priors were specified directly on equality events, though this will induce priors on partition sets E_h^d. A similar comment applies to nonparametric product partition models as proposed by Dahl (2009), which requires for the univariate normal-normal DPM model that the variance is constant and known for each component. However, the requirement is imposed to satisfy a technical condition (Condition 1) concerning the partition likelihood and related to the mode-finding algorithm applicable to those models. The paper does not investigate the general case where variances are not constant and unknown.
Dirichlet Process priors, as typically implemented for normal models, allow for parameters to be clustered in subsets. However, the number of possible patterns is smaller than that of combinatorial mixtures, as only bidimensional parameters, $\theta_i = (\mu_i, \sigma_i^2)$ are potentially shared.

A challenge to the implementation of combinatorial mixtures is that the number of K^*-way comparisons between the elements of component-specific parameter vectors increases rapidly with the maximum number of groups. For $K^* = 3$ the Bell number is $B(K^*) = 5$, for $K^* = 7$, it is $B(K^*) = 877$.

With regard to the priors, we suggested independence across parameters because there is a case for it for most purposes. However, dependence across priors might be desiderable in some cases. For example, mean, variance and correlation vectors may be dependent. Conditional to the corresponding γs, means, variances and correlations may be assumed as dependent across components. The first scenario might be relevant in biological applications too. For example, if the mean in one group is bigger than that of another, one might expect their variances to be different as well. Accounting for this kind of dependence poses challenges that are not fully explored in the paper. We explored the scenario of dependence of component parameters conditional to $(\gamma^m, \gamma^v) \neq (000, 000)$ in a univariate supervised set-up with two known groups. For both component means and variances, we proposed a reparametrization of the dependent parameters in terms of two independent parameters, a general parameter and a difference/ratio between the original dependent parameters. We assumed a mixture of singular distributions for the prior on the independent parameters and derived the corresponding posterior distribution via MCMC. For further details we refer to Edefonti's Ph.D. Thesis (Edefonti (2006)). Finally, in choosing a prior for the hyperparameters of the component-specific parameter prior distributions, one may consider the dependence of the component-specific parameters on the hyperparameters.

6 Appendix

6.1 Univariate Model

Full conditional distributions for the univariate normal mixture model are proposed in the following. The shorthand notations:

$$-(\boldsymbol{\mu},\gamma^m)=\{\boldsymbol{\sigma^2},\gamma^v,\boldsymbol{\omega}\},\qquad -(\boldsymbol{\sigma^2},\gamma^v)=\{\boldsymbol{\mu},\gamma^m,\boldsymbol{\omega}\},$$

$$-(\boldsymbol{\mu})=\{\gamma^m,\boldsymbol{\sigma^2},\gamma^v,\boldsymbol{\omega}\},\qquad -(\boldsymbol{\sigma^2})=\{\gamma^v,\boldsymbol{\mu},\gamma^m,\boldsymbol{\omega}\},$$

$$-(\boldsymbol{\omega})=\{\boldsymbol{\theta}\}=\{\boldsymbol{\mu},\gamma^m,\boldsymbol{\sigma^2},\gamma^v\},$$

$$n_k=\#\{i:z_i=k\},\quad \bar{x}_k=\frac{1}{n_k}\sum_{i:z_i=k}x_i,\qquad k=1,2,3,$$

$$0=000,\ 1=110,\ 2=101,\ 3=011,\ 4=111,$$

$$\pi_0^d=\pi_{000}^d,\ \pi_1^d=\pi_{110}^d,\ \pi_2^d=\pi_{101}^d,\ \pi_3^d=\pi_{011}^d,\ \pi_4^d=\pi_{111}^d,\qquad d=m,v,$$

$$\pi_0^{d^*}=\pi_{000}^{d^*},\ \pi_1^{d^*}=\pi_{110}^{d^*},\ \pi_2^{d^*}=\pi_{101}^{d^*},\ \pi_3^{d^*}=\pi_{011}^{d^*},\ \pi_4^{d^*}=\pi_{111}^{d^*},\qquad d=m,v,$$

are used hereafter. From the joint distribution derived in Subsection 4.1, we get:

$$p(\boldsymbol{\mu},\gamma^m|-(\boldsymbol{\mu},\gamma^m),\boldsymbol{x}^n,\boldsymbol{z}^n)=p(\boldsymbol{\mu}|-(\boldsymbol{\mu}),\boldsymbol{x}^n,\boldsymbol{z}^n)p(\gamma^m|-(\boldsymbol{\mu},\gamma^m),\boldsymbol{x}^n,\boldsymbol{z}^n),$$

with:

$$p(\boldsymbol{\mu}|-(\boldsymbol{\mu}),\boldsymbol{x}^n,\boldsymbol{z}^n)=\begin{cases}N(\mu_1|\,\mu^*_{123},\eta^{2^*}_{123}) & \text{if } \gamma^m=0\\ N(\mu_1|\,\mu^*_1,\eta^{2^*}_1)N(\mu_2|\,\mu^*_{23},\eta^{2^*}_{23}) & \text{if } \gamma^m=1\\ N(\mu_1|\,\mu^*_{13},\eta^{2^*}_{13})N(\mu_2|\,\mu^*_2,\eta^{2^*}_2) & \text{if } \gamma^m=2\\ N(\mu_1|\,\mu^*_{12},\eta^{2^*}_{12})N(\mu_3|\,\mu^*_3,\eta^{2^*}_3) & \text{if } \gamma^m=3\\ N(\mu_1|\,\mu^*_1,\eta^{2^*}_1)N(\mu_2|\,\mu^*_2,\eta^{2^*}_2)N(\mu_3|\,\mu^*_3,\eta^{2^*}_3) & \text{if } \gamma^m=4\end{cases},$$

where:

$$\mu^*_{123}=\eta^{2^*}_{123}\left(\sum_{k=1}^3\frac{n_k}{\sigma^2_k}\bar{x}_k\right),\qquad \eta^{2^*}_{123}=\left(\frac{1}{\eta^2}+\sum_{k=1}^3\frac{n_k}{\sigma^2_k}\right)^{-1},$$

$$\mu^*_k=\eta^{2^*}_k\frac{n_k}{\sigma^2_k}\bar{x}_k,\qquad \eta^{2^*}_k=\left(\frac{1}{\eta^2}+\frac{n_k}{\sigma^2_k}\right)^{-1},\qquad k=1,2,3,$$

$$\mu^*_{k'k''}=\eta^{2^*}_{k'k''}\left(\frac{n_{k'}}{\sigma^2_{k'}}\bar{x}_{k'}+\frac{n_{k''}}{\sigma^2_{k''}}\bar{x}_{k''}\right),\qquad \eta^{2^*}_{k'k''}=\left(\frac{1}{\eta^2}+\frac{n_{k'}}{\sigma^2_{k'}}+\frac{n_{k''}}{\sigma^2_{k''}}\right)^{-1},$$

$$k^{'}, k^{''} = 1, 2, 3,\ k^{'} \neq k^{''},\ k^{'}, k^{''} \neq k,$$

and:

$$p(\gamma^m| -(\boldsymbol{\mu}, \gamma^m), \boldsymbol{x}^n, \boldsymbol{z}^n) = \text{Multi}(1, \boldsymbol{\pi}^{m^*}),$$

where $\boldsymbol{\pi}^{m^*}$ is such that:

$$\pi_0^{m*} = \frac{\pi_0^m}{D}\eta_{123}^* \exp\left\{\frac{1}{2}\left(\frac{\mu_{123}^*}{\eta_{123}^*}\right)^2\right\},$$

$$\pi_k^{m*} = \frac{\pi_k^m}{D}\frac{1}{\eta}\eta_k^* \exp\left\{\frac{1}{2}\left(\frac{\mu_k^*}{\eta_k^*}\right)^2\right\}\eta_{k^{'}k^{''}}^* \exp\left\{\frac{1}{2}\left(\frac{\mu_{k^{'}k^{''}}^*}{\eta_{k^{'}k^{''}}^*}\right)^2\right\}, \quad k, k^{'}, k^{''} = 1, 2, 3,\ k^{'} \neq k^{''},\ k^{'}, k^{''} \neq k,$$

$$\pi_4^{m*} = \frac{\pi_4^m}{D}\frac{1}{\eta^2}\prod_{k=1}^{3}\left(\eta_k^* \exp\left\{\frac{1}{2}\left(\frac{\mu_k^*}{\eta_k^*}\right)^2\right\}\right), \quad k = 1, 2, 3,$$

where:

$$\begin{aligned} D = {} & \pi_0^m \eta_{123}^* \exp\left\{\frac{1}{2}\left(\frac{\mu_{123}^*}{\eta_{123}^*}\right)^2\right\} + \pi_1^m \frac{1}{\eta}\eta_1^* \exp\left\{\frac{1}{2}\left(\frac{\mu_1^*}{\eta_1^*}\right)^2\right\}\eta_{23}^* \exp\left\{\frac{1}{2}\left(\frac{\mu_{23}^*}{\eta_{23}^*}\right)^2\right\} + \\ & + \pi_2^m \frac{1}{\eta}\eta_2^* \exp\left\{\frac{1}{2}\left(\frac{\mu_2^*}{\eta_2^*}\right)^2\right\}\eta_{13}^* \exp\left\{\frac{1}{2}\left(\frac{\mu_{13}^*}{\eta_{13}^*}\right)^2\right\} + \\ & + \pi_3^m \frac{1}{\eta}\eta_3^* \exp\left\{\frac{1}{2}\left(\frac{\mu_3^*}{\eta_3^*}\right)^2\right\}\eta_{12}^* \exp\left\{\frac{1}{2}\left(\frac{\mu_{12}^*}{\eta_{12}^*}\right)^2\right\} + \pi_4^m \frac{1}{\eta^2}\prod_{k=1}^{3}\left(\eta_k^* \exp\left\{\frac{1}{2}\left(\frac{\mu_k^*}{\eta_k^*}\right)^2\right\}\right). \end{aligned}$$

Moreover, from the joint distribution it follows for the variances:

$$p(\boldsymbol{\sigma^2}, \gamma^v| -(\boldsymbol{\sigma^2}, \gamma^v), \boldsymbol{x}^n, \boldsymbol{z}^n) = p(\boldsymbol{\sigma^2}| -(\boldsymbol{\sigma^2}), \boldsymbol{x}^n, \boldsymbol{z}^n) p(\gamma^v| -(\boldsymbol{\sigma^2}, \gamma^v), \boldsymbol{x}^n, \boldsymbol{z}^n),$$

with:

$$p(\boldsymbol{\sigma^2}|-(\boldsymbol{\sigma^2}),\boldsymbol{x}^n,\boldsymbol{z}^n)=\begin{cases} IG(\sigma_1^2|\ c^*_{123},d^*_{123}), & \text{if } \gamma^v=0\\ IG(\sigma_1^2|\ c^*_1,d^*_1)IG(\sigma_2^2|\ c^*_{23},d^*_{23}) & \text{if } \gamma^v=1\\ IG(\sigma_1^2|\ c^*_{13},d^*_{13})IG(\sigma_2^2|\ c^*_2,d^*_2) & \text{if } \gamma^v=2\\ IG(\sigma_1^2|\ c^*_{12},d^*_{12})IG(\sigma_3^2|\ c^*_3,d^*_3) & \text{if } \gamma^v=3\\ IG(\sigma_1^2|\ c^*_1,d^*_1)IG(\sigma_2^2|\ c^*_2,d^*_2)IG(\sigma_3^2|\ c^*_3,d^*_3) & \text{if } \gamma^v=4 \end{cases},$$

where:

$$c^*_{123}=c+\frac{n}{2},\qquad d^*_{123}=d+\frac{1}{2}\sum_{k=1}^{3}\sum_{i:z_i=k}(x_i-\mu_k)^2,$$

$$c^*_k=c+\frac{n_k}{2},\qquad d^*_k=d+\frac{1}{2}\sum_{i:z_i=k}(x_i-\mu_k)^2,\qquad k=1,2,3,$$

$$c^*_{k'k''}=c+\frac{n_{k'}+n_{k''}}{2},\ \ d^*_{k'k''}=d+\frac{1}{2}\left(\sum_{i:z_i=k'}(x_i-\mu_{k'})^2+\sum_{i:z_i=k''}(x_i-\mu_{k''})^2\right),$$

$$k',k''=1,2,3,\ k'\neq k'',\ k',k''\neq k,$$

and:

$$p(\gamma^v|\ -(\boldsymbol{\sigma^2},\gamma^v),\boldsymbol{x}^n,\boldsymbol{z}^n)=\text{Multi}(1,\boldsymbol{\pi}^{v^*}),$$

where $\boldsymbol{\pi}^{v^*}$ is such that:

$$\pi_0^{v*}=\frac{\pi_0^v}{F}\frac{\Gamma(c^*_{123})}{(d^*_{123})^{c^*_{123}}},$$

$$\pi_k^{v*}\ =\ \frac{\pi_k^v}{F}\frac{d^c}{\Gamma(c)}\frac{\Gamma(c^*_k)}{\left(d^*_k\right)^{c^*_k}}\frac{\Gamma(c^*_{k'k''})}{\left(d^*_{k'k''}\right)^{c^*_{k'k''}}},\quad k,k',k''=1,2,3,\ k'\neq k'',\ k',k''\neq k,$$

$$\pi_4^{v*}=\frac{\pi_4^v}{F}\left(\frac{d^c}{\Gamma(c)}\right)^2\prod_{k=1}^{3}\left(\frac{\Gamma(c^*_k)}{\left(d^*_k\right)^{c^*_k}}\right),\quad k=1,2,3,$$

and:

$$
\begin{aligned}
F &= \pi_0^v \frac{\Gamma(c^*_{123})}{(d^*_{123})^{c^*_{123}}} + \pi_1^v \frac{d^c}{\Gamma(c)} \frac{\Gamma(c^*_1)}{(d^*_1)^{c^*_1}} \frac{\Gamma(c^*_{23})}{(d^*_{23})^{c^*_{23}}} + \pi_2^v \frac{d^c}{\Gamma(c)} \frac{\Gamma(c^*_2)}{(d^*_2)^{c^*_2}} \frac{\Gamma(c^*_{13})}{(d^*_{13})^{c^*_{13}}} + \pi_3^v \frac{d^c}{\Gamma(c)} \frac{\Gamma(c^*_3)}{(d^*_3)^{c^*_3}} \frac{\Gamma(c^*_{12})}{(d^*_{12})^{c^*_{12}}} + \\
&+ \pi_4^v \left(\frac{d^c}{\Gamma(c)} \right)^2 \prod_{k=1}^{3} \left(\frac{\Gamma(c^*_k)}{\left(d^*_k\right)^{c^*_k}} \right).
\end{aligned}
$$

Finally, for the weights one has:

$$
p(\boldsymbol{\omega}| -(\boldsymbol{\omega}), \boldsymbol{x}^n, \boldsymbol{z}^n) = Dir(a_{0,1} + n_1, a_{0,2} + n_2, a_{0,2} + n_3),
$$

and for the generic group label, z_i, $i = 1, \ldots, n$:

$$
p(z_i = k|\ \omega_k, -(\boldsymbol{\omega}), x_i) = \frac{\omega_k \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma_k^2}(x_i - \mu_k)^2 \right\}}{\sum_{k=1}^{3} \omega_k \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma_k^2}(x_i - \mu_k)^2 \right\}}, \qquad k = 1, 2, 3.
$$

6.2 Bivariate Model

Available full conditional distributions for the bivariate normal mixture model are proposed in the following. The shorthand notations:

$$
-(\boldsymbol{\mu}^j, \gamma_j^m) = \{\boldsymbol{\mu}^{j'}, \gamma_{j'}^m, \boldsymbol{S}^j, \gamma_j^v, \boldsymbol{S}^{j'}, \gamma_{j'}^v, \boldsymbol{r}, \gamma^c, \boldsymbol{\omega}\}, \quad -(\boldsymbol{S}^j, \gamma_j^v) = \{\boldsymbol{\mu}^j, \gamma_j^m, \boldsymbol{\mu}^{j'}, \gamma_{j'}^m, \boldsymbol{S}^{j'}, \gamma_{j'}^v, \boldsymbol{r}, \gamma^c, \boldsymbol{\omega}\},
$$

$$
-(\boldsymbol{\mu}^j) = \{\gamma_j^m, \boldsymbol{\mu}^{j'}, \gamma_{j'}^m, \boldsymbol{S}^j, \gamma_j^v, \boldsymbol{S}^{j'}, \gamma_{j'}^v, \boldsymbol{r}, \gamma^c, \boldsymbol{\omega}\}, \qquad -(\boldsymbol{S}^j) = \{\boldsymbol{\mu}^j, \gamma_j^m, \boldsymbol{\mu}^{j'}, \gamma_{j'}^m, \gamma_j^v, \boldsymbol{S}^{j'}, \gamma_{j'}^v, \boldsymbol{r}, \gamma^c, \boldsymbol{\omega}\},
$$

$$
j, j' = 1, 2,\ j \neq j',
$$

$$
-(\boldsymbol{r}, \gamma^c) = \{\boldsymbol{\mu}^1, \gamma_1^m, \boldsymbol{\mu}^2, \gamma_2^m, \boldsymbol{S}^1, \gamma_1^v, \boldsymbol{S}^2, \gamma_2^v, \boldsymbol{\omega}\}, \qquad -(\boldsymbol{r}) = \{\boldsymbol{\mu}^1, \gamma_1^m, \boldsymbol{\mu}^2, \gamma_2^m, \boldsymbol{S}^1, \gamma_1^v, \boldsymbol{S}^2, \gamma_2^v, \gamma^c, \boldsymbol{\omega}\},
$$

$$
-(\boldsymbol{\omega}) = \{\boldsymbol{\theta}\} = \{\boldsymbol{\mu}^1, \gamma_1^m, \boldsymbol{\mu}^2, \gamma_2^m, \boldsymbol{S}^1, \gamma_1^v, \boldsymbol{S}^2, \gamma_2^v, \boldsymbol{r}, \gamma^c\},
$$

$$
n_k = \#\{i : z_i = k\}, \quad \bar{x}_{kj} = \frac{1}{n_k} \sum_{i:z_i=k} x_i^j, \qquad k = 1, 2, 3,\ \ j = 1, 2,
$$

$$
0 = 000,\ 1 = 110,\ 2 = 101,\ 3 = 011,\ 4 = 111,
$$

$$
\pi_0^d = \pi_{000}^d,\ \pi_1^d = \pi_{110}^d,\ \pi_2^d = \pi_{101}^d,\ \pi_3^d = \pi_{011}^d,\ \pi_4^d = \pi_{111}^d, \qquad d = m, v, c,
$$

$$
\pi_0^{d^*} = \pi_{000}^{d^*},\ \pi_1^{d^*} = \pi_{110}^{d^*},\ \pi_2^{d^*} = \pi_{101}^{d^*},\ \pi_3^{d^*} = \pi_{011}^{d^*},\ \pi_4^{d^*} = \pi_{111}^{d^*}, \qquad d = m, v, c,
$$

are used hereafter. From the joint distribution derived at the end of Subsection 4.2, we get:

$$p(\boldsymbol{\mu}^j, \gamma_j^m | -(\boldsymbol{\mu}^j, \gamma_j^m), \boldsymbol{x}^n, \boldsymbol{z}^n) = p(\boldsymbol{\mu}^j | -(\boldsymbol{\mu}^j), \boldsymbol{x}^n, \boldsymbol{z}^n) p(\gamma_j^m | -(\boldsymbol{\mu}^j, \gamma_j^m), \boldsymbol{x}^n, \boldsymbol{z}^n), \qquad j = 1, 2,$$

with:

$$p(\boldsymbol{\mu}^j | -(\boldsymbol{\mu}^j), \boldsymbol{x}^n, \boldsymbol{z}^n) = \begin{cases} N(\mu_{1j} |\, \mu^*_{123j}, {\eta^2}^*_{123j}) & \text{if } \gamma_j^m = 0 \\ N(\mu_{1j} |\, \mu^*_{1j}, {\eta^2}^*_{1j}) N(\mu_{2j} |\, \mu^*_{23j}, {\eta^2}^*_{23j}) & \text{if } \gamma_j^m = 1 \\ N(\mu_{1j} |\, \mu^*_{13j}, {\eta^2}^*_{13j}) N(\mu_{2j} |\, \mu^*_{2j}, {\eta^2}^*_{2j}) & \text{if } \gamma_j^m = 2 \\ N(\mu_{1j} |\, \mu^*_{12j}, {\eta^2}^*_{12j}) N(\mu_{3j} |\, \mu^*_{3j}, {\eta^2}^*_{3j}) & \text{if } \gamma_j^m = 3 \\ N(\mu_{1j} |\, \mu^*_{1j}, {\eta^2}^*_{1j}) N(\mu_{2j} |\, \mu^*_{2j}, {\eta^2}^*_{2j}) N(\mu_{3j} |\, \mu^*_{3j}, {\eta^2}^*_{3j}) & \text{if } \gamma_j^m = 4 \end{cases},$$

where:

$$q_k = \frac{S_{kj'} \bar{x}_{kj} + S_{kj} r_k (\mu_{kj'} - \bar{x}_{kj'})}{S_{kj'}}, \qquad p_{kj} = \frac{n_k}{(1 - r_k^2) {S_{kj}}^2}, \qquad k = 1, 2, 3, \quad j, j^{'} = 1, 2,\ j \neq j^{'},$$

$$\mu^*_{123j} = {\eta^2}^*_{123j} \left(\sum_{k=1}^{3} p_{kj} q_k \right), \qquad {\eta^2}^*_{123j} = \left(\frac{1}{\eta^2} + \sum_{k=1}^{3} p_{kj} \right)^{-1}, \qquad j = 1, 2,$$

$$\mu^*_{kj} = {\eta^2}^*_{kj} p_{kj} q_k, \qquad {\eta^2}^*_{kj} = \left(\frac{1}{\eta^2} + p_{kj} \right)^{-1}, \qquad k = 1, 2, 3, \qquad j = 1, 2,$$

$$\mu^*_{k'k''j} = {\eta^2}^*_{k'k''j} \left(p_{k'j} q_{k'} + p_{k''j} q_{k''} \right), \qquad {\eta^2}^*_{k'k''j} = \left(\frac{1}{\eta^2} + p_{k'j} + p_{k''j} \right)^{-1},$$

$$k^{'}, k^{''} = 1, 2, 3,\ k^{'} \neq k^{''},\ k^{'}, k^{''} \neq k,\ j = 1, 2,$$

and:

$$p(\gamma_j^m | -(\boldsymbol{\mu}^j, \gamma_j^m), \boldsymbol{x}^n, \boldsymbol{z}^n) = \text{Multi}(1, \boldsymbol{\pi}^{m^*}), \qquad j = 1, 2,$$

where $\boldsymbol{\pi}_j^{m^*}$ is such that:

$$\pi_{0j}^{m*} = \frac{L_j \pi_0^m}{H} \eta^*_{123j} \exp\left\{ \frac{1}{2} \left(\frac{\mu^*_{123j}}{\eta^*_{123j}} \right)^2 \right\},$$

$$\pi_{kj}^{m*} = \frac{L_j \pi_k^m}{H} \frac{1}{\eta} \eta^*_{kj} \exp\left\{ \frac{1}{2} \left(\frac{\mu^*_{kj}}{\eta^*_{kj}} \right)^2 \right\} \eta^*_{k'k''j} \exp\left\{ \frac{1}{2} \left(\frac{\mu^*_{k'k''j}}{\eta^*_{k'k''j}} \right)^2 \right\},$$

$$k, k^{'}, k^{''} = 1, 2, 3,\ k^{'} \neq k^{''},\ k^{'}, k^{''} \neq k,$$

$$\pi_{4j}^{m*} = \frac{L_j \pi_4^m}{H} \frac{1}{\eta^2} \prod_{k=1}^{3} \left(\eta_{kj}^* \exp\left\{ \frac{1}{2} \left(\frac{\mu_{kj}^*}{\eta_{kj}^*} \right)^2 \right\} \right), \quad k = 1, 2, 3,$$

where:

$$\begin{aligned} H = L_j \Bigg[& \pi_0^m \eta_{123j}^* \exp\left\{ \frac{1}{2} \left(\frac{\mu_{123j}^*}{\eta_{123j}^*} \right)^2 \right\} + \pi_1^m \frac{1}{\eta} \eta_{1j}^* \exp\left\{ \frac{1}{2} \left(\frac{\mu_{1j}^*}{\eta_{1j}^*} \right)^2 \right\} \eta_{23j}^* \exp\left\{ \frac{1}{2} \left(\frac{\mu_{23j}^*}{\eta_{23j}^*} \right)^2 \right\} + \\ & + \pi_2^m \frac{1}{\eta} \eta_{2j}^* \exp\left\{ \frac{1}{2} \left(\frac{\mu_{2j}^*}{\eta_{2j}^*} \right)^2 \right\} \eta_{13j}^* \exp\left\{ \frac{1}{2} \left(\frac{\mu_{13j}^*}{\eta_{13j}^*} \right)^2 \right\} + \\ & + \pi_3^m \frac{1}{\eta} \eta_{3j}^* \exp\left\{ \frac{1}{2} \left(\frac{\mu_{3j}^*}{\eta_{3j}^*} \right)^2 \right\} \eta_{12j}^* \exp\left\{ \frac{1}{2} \left(\frac{\mu_{12j}^*}{\eta_{12j}^*} \right)^2 \right\} + \pi_4^m \frac{1}{\eta^2} \prod_{k=1}^{3} \left(\eta_{kj}^* \exp\left\{ \frac{1}{2} \left(\frac{\mu_{kj}^*}{\eta_{kj}^*} \right)^2 \right\} \right) \Bigg], \end{aligned}$$

and:

$$\begin{aligned} L_j = & \left(\frac{1}{\eta} \right) \left(\frac{1}{2\pi} \right)^n \prod_{k=1}^{3} \left(\frac{1}{S_{k1} S_{k2} \sqrt{1 - r_k^2}} \right)^{n_k} * \\ * & \exp\left\{ -\frac{1}{2(1 - r_k^2)} \left[\sum_{k=1}^{3} \sum_{i:z_i=k} \left(\frac{(x_i^{j^{'}} - \mu_{kj^{'}})^2}{{S_{kj^{'}}}^2} + \frac{S_{kj^{'}} (x_i^j)^2 - 2 S_{kj} r_k (x_i^j x_i^{j^{'}} - x_i^j \mu_{kj^{'}})}{{S_{kj}}^2 S_{kj^{'}}} \right) \right] \right\}, \end{aligned}$$

j=1,2.

For the weights one has:

$$p(\boldsymbol{\omega} | -(\boldsymbol{\omega}), \boldsymbol{x}^n, \boldsymbol{z}^n) = Dir(a_{0,1} + n_1, a_{0,2} + n_2, a_{0,2} + n_3),$$

and for the generic group label, z_i, $i = 1, \ldots, n$:

$p(z_i = k | \, \omega_k, \{\boldsymbol{\theta}\}, \boldsymbol{x}_i) =$

$$= \frac{\omega_k \frac{1}{2\pi S_{k1} S_{k2}\sqrt{1-r_k^2}} \exp\left\{-\frac{1}{2(1-r_k^2)}\left(\frac{(x_i^1-\mu_{k1})^2}{{S_{k1}}^2} + \frac{(x_i^2-\mu_{k2})^2}{{S_{k2}}^2} - 2r_k\frac{(x_i^1-\mu_{k1})(x_i^2-\mu_{k2})}{S_{k1}S_{k2}}\right)\right\}}{\sum_{k=1}^{3}\omega_k \frac{1}{2\pi S_{k1} S_{k2}\sqrt{1-r_k^2}} \exp\left\{-\frac{1}{2(1-r_k^2)}\left(\frac{(x_i^1-\mu_{k1})^2}{{S_{k1}}^2} + \frac{(x_i^2-\mu_{k2})^2}{{S_{k2}}^2} - 2r_k\frac{(x_i^1-\mu_{k1})(x_i^2-\mu_{k2})}{S_{k1}S_{k2}}\right)\right\}},$$

$$k = 1, 2, 3.$$

We refer to the Metropolis-Hastings algorithm for the update of $(\boldsymbol{S}^j, \gamma_j^v)$, $j = 1, 2$, and $(\boldsymbol{r}, \gamma^c)$. For $(\boldsymbol{S}^j, \gamma_j^v)$, $j = 1, 2$, the target distributions $\pi(\boldsymbol{S}^{j(t)}, \gamma_j^{v(t)})$, where $(\boldsymbol{S}^{j(t)}, \gamma_j^{v(t)})$ indicates the current position of the chain at the t-th iteration, $t > 1$, are represented by the marginal posterior distributions of $(\boldsymbol{S}^j, \gamma_j^v)$:

$$p(\boldsymbol{S}^j, \gamma_j^v | -(\boldsymbol{S}^j, \gamma_j^v), \boldsymbol{x}^n, \boldsymbol{z}^n) = p(\boldsymbol{S}^j|\gamma_j^v)p(\gamma_j^v)p(\boldsymbol{x}^n|\{\boldsymbol{\theta}\}, \omega, \boldsymbol{z}^n) \propto p(\boldsymbol{S}^j|\gamma_j^v)p(\boldsymbol{x}^n|\{\boldsymbol{\theta}\}, \omega, \boldsymbol{z}^n),$$

$$j = 1, 2,$$

which assume five different expressions for each j depending on the value of γ_j^v. The proposal distribution is built accordingly. If $(\boldsymbol{y}^j, y_4^j)$ indicates the proposed candidate for the (t+1)-th iteration, with $\boldsymbol{y}^j = (y_{1j}, y_{2j}, y_{3j})$, one may sample y_4^j from a Multinomial distribution with parameters $N = 1$, $\boldsymbol{p} = (0.2, 0.2, 0.2, 0.2, 0.2)$. Conditioning on y_4^j, one may sample the corresponding vector $\boldsymbol{y}^j$ using the usual structure of degeneracy along equality constraints proposed with the prior distributions. The building block distribution within this structure is assumed to be a Normal distribution centered at the current position of the chain, $\boldsymbol{S}^{j(t)}$, and having a standard deviation of $\epsilon = 0.05$. A constraint is added to guarantee that the proposed standard deviations, $\boldsymbol{y}^j$, are non-negative. This gives the following expression:

$$q((\boldsymbol{S}^{j(t)}, \gamma_j^{v(t)}), (\boldsymbol{y}^j, y_4^j)) = p(y_4^j)p(\boldsymbol{y}^j|y_4^j) \propto p(\boldsymbol{y}^j|y_4^j),$$

with:

$$p(\boldsymbol{y}^j|y_4^j) \propto \begin{cases} N(y_{1j}|\ {S_{1j}}^{(t)}, \epsilon^2), & \text{if } y_4^j = 0 \\ N(y_{1j}|\ {S_{1j}}^{(t)}, \epsilon^2)N(y_{2j}|\ {S_{2j}}^{(t)}, \epsilon^2) & \text{if } y_4^j = 1 \\ N(y_{1j}|\ {S_{1j}}^{(t)}, \epsilon^2)N(y_{2j}|\ {S_{2j}}^{(t)}, \epsilon^2) & \text{if } y_4^j = 2 \\ N(y_{1j}|\ {S_{1j}}^{(t)}, \epsilon^2)N(y_{3j}|\ {S_{3j}}^{(t)}, \epsilon^2) & \text{if } y_4^j = 3 \\ N(y_{1j}|\ {S_{1j}}^{(t)}, \epsilon^2)N(y_{2j}|\ {S_{2j}}^{(t)}, \epsilon^2)N(y_{3j}|\ {S_{3j}}^{(t)}, \epsilon^2) & \text{if } y_4^j = 4 \end{cases}, \ j = 1, 2.$$

For the updating of $(\boldsymbol{r}, \gamma^c)$, the target distribution $\pi(\boldsymbol{r}^{(t)}, \gamma^{c(t)})$, where $(\boldsymbol{r}^{(t)}, \gamma^{c(t)})$ indicates the current position of the chain at the t-th iteration, $t > 1$, is represented by the marginal posterior distribution of $(\boldsymbol{r}, \gamma^c)$:

$$p(\boldsymbol{r}, \gamma^c | -(\boldsymbol{r}, \gamma^c), \boldsymbol{x}^n, \boldsymbol{z}^n) = p(\boldsymbol{r}|\gamma^c)p(\gamma^c)p(\boldsymbol{x}^n|\{\boldsymbol{\theta}\}, \omega, \boldsymbol{z}^n) \propto p(\boldsymbol{r}|\gamma^c)p(\boldsymbol{x}^n|\{\boldsymbol{\theta}\}, \omega, \boldsymbol{z}^n),$$

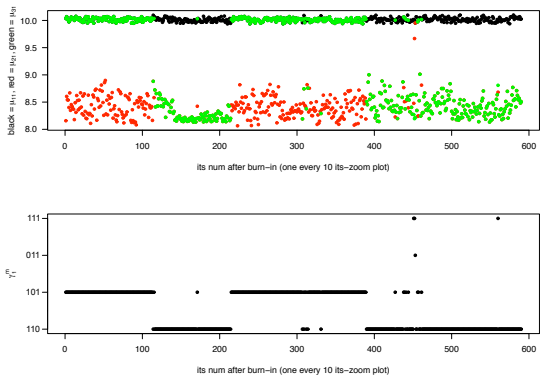
which assumes five different expressions depending on the value of γ^c. The proposal distribution is built as before. If $(\boldsymbol{y}, y_4)$ indicates the proposed candidate for the (t+1)-th iteration, with $\boldsymbol{y} = (y_1, y_2, y_3)$, y_4 is sampled from a Multinomial distribution with parameters $N = 1$ and $\boldsymbol{p} = (0.2, 0.2, 0.2, 0.2, 0.2)$. Conditioning on y_4, $\boldsymbol{y}$ is sampled as $p(\boldsymbol{y}^j|y_4^j)$ before. We assume a translated Beta distribution as our building block distribution. A constraint is added to guarantee that the proposed correlations, $\boldsymbol{y}$, lie in $(-1, 1)$. This gives the following expression:

$$q((\boldsymbol{r}^{(t)}, \gamma^{c(t)}), (\boldsymbol{y}, y_4)) = p(y_4)p(\boldsymbol{y}|y_4) \propto p(\boldsymbol{y}|y_4),$$

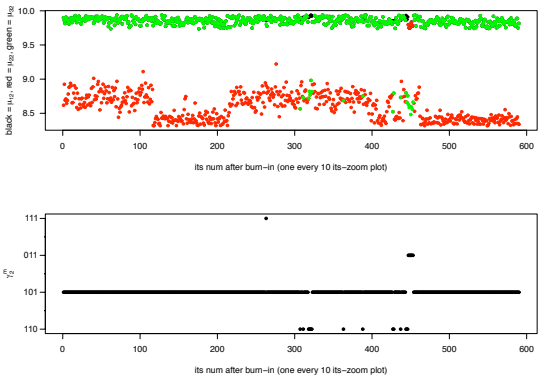
with:

$$p(\boldsymbol{y}|y_4) \propto \begin{cases} Be_{(-1,1)}(y_1|\,\alpha, \beta), & \text{if } y_4 = 0 \\ Be_{(-1,1)}(y_1|\,\alpha, \beta)Be_{(-1,1)}(y_2|\,\alpha, \beta) & \text{if } y_4 = 1 \\ Be_{(-1,1)}(y_1|\,\alpha, \beta)Be_{(-1,1)}(y_2|\,\alpha, \beta) & \text{if } y_4 = 2 \\ Be_{(-1,1)}(y_1|\,\alpha, \beta)Be_{(-1,1)}(y_3|\,\alpha, \beta) & \text{if } y_4 = 3 \\ Be_{(-1,1)}(y_1|\,\alpha, \beta)Be_{(-1,1)}(y_2|\,\alpha, \beta)Be_{(-1,1)}(y_3|\,\alpha, \beta) & \text{if } y_4 = 4 \end{cases},$$

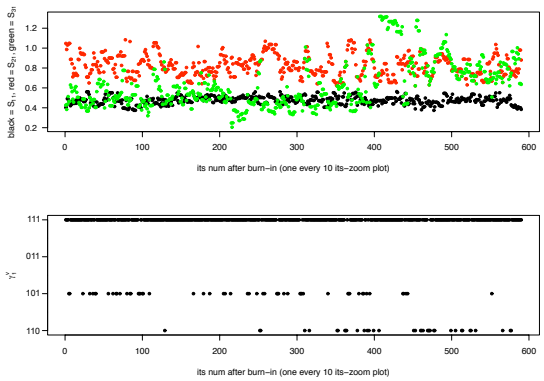
where $Be_{(-1,1)}(\cdot|\,\alpha, \beta)$ indicates a Beta distribution on the support $(-1, 1)$ with parameters α, β, $\alpha, \beta > 0$. We choose α and β in such a way that the proposal distribution is centered at the current position of the chain, $\boldsymbol{r}^{(t)}$, and $\alpha + \beta = 5$.



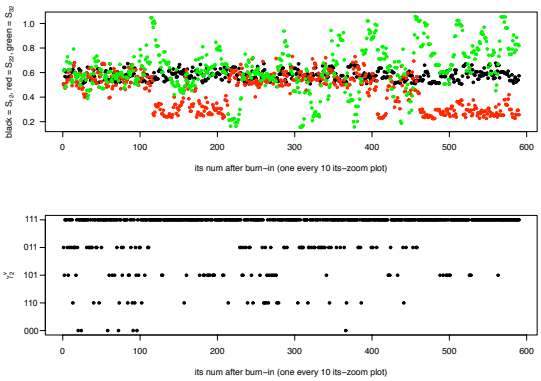
(a) $\boldsymbol{\mu}^1 = (\mu_{11}, \mu_{21}, \mu_{31})$ plotted along with γ_1^m



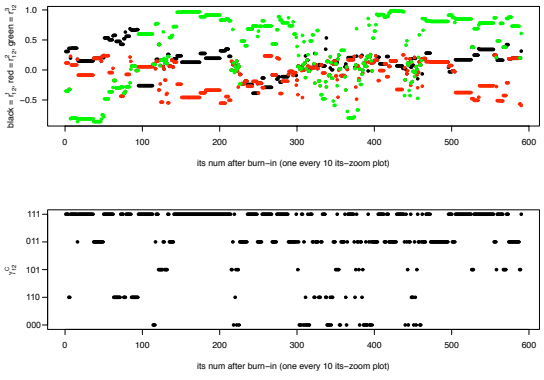
(b) $\boldsymbol{\mu}^2 = (\mu_{12}, \mu_{22}, \mu_{32})$ plotted along with γ_2^m



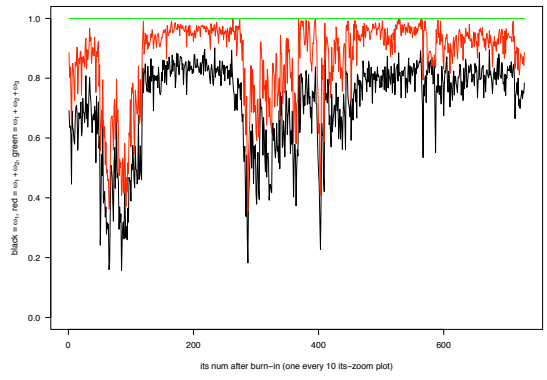
(c) $\boldsymbol{S}^1 = (S_{11}, S_{21}, S_{31})$ plotted along with γ_1^v



(d) $\boldsymbol{S}^2 = (S_{12}, S_{22}, S_{32})$ plotted along with γ_2^v



(e) $\boldsymbol{r} = (r_1, r_2, r_3)$ plotted along with γ^c



(f) $\omega_1, \omega_2, \omega_2$: adaptation of the stacked area plot

Figure 8: Marginal chains of the parameters for the bivariate mixture model as obtained by the simulation. Component means and corresponding standard deviations, correlations and weights for each iteration are color-coded in the same way across plots.

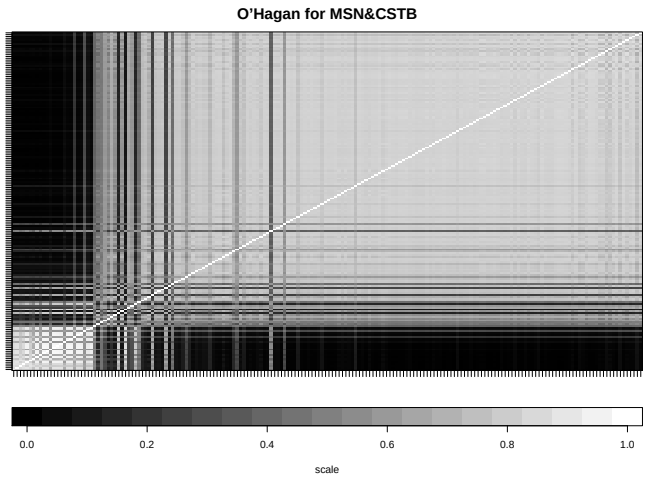


Figure 9: "O'Hagan" matrix for the bivariate normal mixture model. Relative frequencies of occurrence of two being in the same group are plotted in a black-to-white color scale and sorted according to a non-decreasing ordering of the first variable, the MSN gene. Dark and light blocks, with proportions similar to the estimated weights of the mixture, identify different groups of observations.

TRIM29

	case 1		case 2		case 3	
	prior	posterior	prior	posterior	prior	posterior
1 subgroup	0.64	0.001449928	0.36	0.002349883	0.04	0.001049948
2 subgroups	0.25	0.3113844	0.39	0.4865257	0.36	0.2018399
3 subgroups	0.11	0.6871656	0.25	0.5111244	0.6	0.7971101

MSN

	case 1		case 2		case 3	
	prior	posterior	prior	posterior	prior	posterior
1 subgroup	0.64	0.1093945	0.36	0.1131943	0.04	0.07564622
2 subgroups	0.25	0.6259187	0.39	0.5084246	0.36	0.4711264
3 subgroups	0.11	0.2646868	0.25	0.3783811	0.6	0.4532273

ITGB5

	case 1		case 2		case 3	
	prior	posterior	prior	posterior	prior	posterior
1 subgroup	0.64	0.8486576	0.36	0.6842658	0.04	0.2633368
2 subgroups	0.25	0.1234438	0.39	0.2532373	0.36	0.5750712
3 subgroups	0.11	0.02789861	0.25	0.06249688	0.6	0.1615919

Figure 10: Results of sensitivity analysis to different values of $\boldsymbol{\pi}^m$ and $\boldsymbol{\pi}^v$: prior and posterior probabilities of observing one, two or three groups for each gene under cases 1,2,3. Case 1 is such that: P{1 subgroup}=0.64, P{2 subgroups} $\approx$ 0.25 and P{3 subgroups} $\approx$ 0.11; case 2 is such that: P{1 subgroup}=0.36, P{2 subgroups}=0.39 and P{3 subgroups}=0.25; case 3 is such that: P{1 subgroup}=0.04, P{2 subgroups}=0.36 and P{3 subgroups}=0.60.

References

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2(6)**, 1152–1174.

Barnard, J., McCulloch, R. E. and Meng, X. L. (2000) Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica*, **10**, 1281–1311.

Barry, D. and Hartigan, J. A. (1992) Product partition models for change point problems. *The Annals of Statistics*, **20**, 260–279.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B. and Hanash, S. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, **8**, 816–824.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J. and Meyerson, M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. National Academy of Science USA*, **98**, 13790–13795.

Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, **1**, 353–355.

Böhning, D. and Seidel, W. (2003) Editorial: recent developments in mixture models. *Computational Statistics and Data Analysis*, **41**, 349–357.

Carlin, B. P. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, **57**, 473–484.

Casella, G., Robert, C. P. and Wells, M. T. (2004) Mixture models, latent variables and partitioned importance sampling. *Statistical Methodology*, **57,1-2**, 1–18.

Chang, W. C. (1983) On using principal components before separating a mixture of two multivariate Normal distributions. *Applied Statistics*, **32**, 267-275.

Crowley, E. M. (1997) Product partition models for normal means. *Journal of the American Statistical Association*, **92**, 192–198.

Dahl, D. B. (2009) Modal clustering in a class of product partition models. *Bayesian Analysis*, **4(2)**, 243–264.

Dettling, M., Gabrielson, E. and Parmigiani, G. (2005) Searching for differentially expressed gene combinations. *Genome Biology*, **6(10)**, R88.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 363–375.

Edefonti, V. (2006) Integrating supervised and unsupervised learning in genomics applications. *PhD Thesis*. Bocconi University, Milan, Italy.

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

Frühwirth-Schnatter, S. (2001) Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**, 194–209.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D. and Petersen, I. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. National Academy of Science USA*, **98(24)**, 13784–13789.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gottardo, R. and Raftery A. E. (2008) Markov chain Monte Carlo with mixtures of mutually singular distributions. *Journal of Computational and Graphical Statistics*, **17(4)**, 917–943.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hartigan, J. A. (1990) Partition models. *Communications in Statistics, Part A Theory and Methods*, **19**, 2745–2756.

Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5(3)**, 299–314.

Kadane, J. B. (1975) The role of identification in Bayesian theory. In *Studies in Bayesian econometrics and statistics* (eds. S. E. Fienberg and A. Zellner), pp. 175–191. Amsterdam: North-Holland Publishing Co..

Kim, S., Tadesse, M. G. and Vannucci, M. (2006) Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **93(4)**, 877–893.

Marin, J. M., Mengersen, K. and Robert, C. P. (2005) Bayesian modelling and inference on mixtures of distributions. In *Handbook of Statistics*, (eds. D. Dey and C. R. Rao), vol. 25(16), pp. 459–507. Amsterdam: Elsevier-Sciences.

Martin, A. D. and Quinn, K. M.(2005) MCMCpack: Markov chain Monte Carlo (MCMC) Package *Manuscript*, URL http://mcmcpack.wustl.edu

McLachlan G. J. and Basford K. E. (1988) *Mixture models: inference and applications to clustering*. New York: Marcel Dekker.

McLachlan G. J. and Peel D. (2000) *Finite mixture models*. New York: John Wiley & Sons.

Mengersen, K. L. and Robert, C. P. (1996) Testing for mixtures: a Bayesian entropic approach. In *Bayesian Statistics 5*, (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 255–276. Oxford: Oxford University Press.

Miura, K., Bowman, E. D., Simon, R., Peng, A. C., Robles, A. I., Jones, R. T., Katagiri, T., He, P., Mizukami, H., Charboneau, L., Kikuchi, T., Liotta, L. A., Nakamura, Y. and Harris, C. C. (2002) Laser capture microdissection and microarray expression analysis of lung adenocarcinoma reveals tobacco smoking- and prognosis-related molecular profiles. *Cancer Research*, **62**, 3244–50.

Newcomb, S. (1886) A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, **8**, 343–366.

Nobile, A. (1994) Bayesian analysis of finite mixture distributions. *PhD Thesis.* Carnegie Mellon University, Pittsburgh, USA.

Nobile, A. (2005) Bayesian finite mixtures: a note on prior specification and posterior computation. *Tech. Rep. 05-3, Department of Statistics, University of Glasgow.*

O'Hagan, A. (1997) Contribution to the discussion of Richardson and Green (1997), On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59**, 772.

Pearson, K. (1894) Contribution to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. A*, **185**, **4**, 71–110.

Phillips, D. B. and Smith, A. F. M. (1996) Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice*, (eds. W. R. Gilks, D. J. Spiegelhalter and S. Richardson), pp. 215–239. London: Chapman and Hall.

Raftery, A. E. (1996) Hypothesis testing and model selection. In *Markov Chain Monte Carlo in practice*, (eds. W. R. Gilks, D. J. Spiegelhalter and S. Richardson), pp. 163–187. London: Chapman and Hall.

R Development Core Team (2006) R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0*, URL http://www.R-project.org.

R Development Core Team (2008) R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0*, URL http://www.R-project.org.

Richardson, S. and Green, P. J. (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.

Roeder, K. and Wasserman, L. (1997) Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

Shedden, K. and Taylor, J. (2004) Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas. In *Methods of Microarray Data Analysis IV*, (ed. J. S. Shoemaker and S. M. Lin), pp. 121–132. Springer.

Steele, R. J., Raftery, A. E. and Emond, M. J. (2003) Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). *Journal of Computational and Graphical Statistics*, **15**, 712–734.

Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *The Annals of Statistics*, **28**, 40–74.

Stephens, M. (2000) Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B*, **62**, 795–809.

Titterington D. M., Smith A. F. M. and Makov U. E. (1985) *Statistical analysis of finite mixture distributions.* New York: Wiley.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

West, M., and Turner, D. A. (1994) Deconvolution of mixtures in analysis of neural synaptic transmission. *The Statistician*, **43**, 31–43.

Wigle, D. A., Jurisica, I., Radulovich, N., Pintilie, M., Rossant, J., Liu, N., Lu, C., Woodgett, J., Seiden, I., Johnston, M., Keshavjee, S., Darling, G., Winton, T., Breitkreutz, B. J., Jorgenson, P., Tyers, M., Shepherd, F. A. and Tsao, M. S. (2002) Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research*, **62**, 3005–3008.