Valeria Edefonti¹ / Giovanni Parmigiani²

Combinatorial Mixtures of Multiparameter Distributions: An Application to Bivariate Data

¹ Department of Clinical Sciences and Community Health, University of Milan, via A. Vanzetti, 5, Milan, MI 20133, Italy, E-mail: valeria.edefonti@unimi.it

² Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA; Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Abstract:

We introduce *combinatorial mixtures* – a flexible class of models for inference on mixture distributions whose components 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 computational 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 be characterized by differences in location, scale, correlations or any of the combinations. We illustrate our approach using publicly available data on molecular subtypes of lung and prostate cancers.

Keywords: Bayesian inference, clustering, Markov chain Monte Carlo, mixture models with unknown number of components

DOI: 10.1515/ijb-2015-0064

1 Introduction

Since the beginning of the last century [1, 2], finite mixture distributions have received attention as tools for modelling population heterogeneity or for building flexible finite-parameter distributions. Monographs on finite mixtures include the classical Titterton, Smith, and Makov, McLachlan and Basford, and McLachlan and Peel [3–5]. Böhning and Seidel [6] is a review with emphasis on nonparametric maximum likelihood, whereas Marin et al. [7] is an introduction from a Bayesian perspective. For a complete treatment of several aspects of finite mixture models from both perspectives, see Frühwirth-Schnatter [8].

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, proposing a parametric solution to this class of problems.

Our approach is to allow each element of the component-specific parameter vector 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 or protein levels, which 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

Valeria Edefonti is the corresponding author.

© 2017 Walter de Gruyter GmbH, Berlin/Boston.

This content is free.

characterized by variability in otherwise tightly controlled processes, or by the lack of otherwise strong correlations [9, 10]. Also, several mechanisms can coexist. In addition, in recent years, it has been recognized that bimodally distributed genes can identify promising candidates for clinically relevant prognostic or predictive signatures [11]. In breast cancer, ER (estrogen receptor), PR (progesterone receptor), and HER2 (human epidermal growth factor receptor 2) genes can be considered molecular switches, with distinct on/off expression states defining clinically unique subtypes with different prognoses and different response to different therapeutic strategies [12]. Another relevant example of application of mixture models is in the modelling of the shape of the hazard function via two- or three-component mixtures in tumor dormancy phenomena [13]. In these contexts, 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 mechanisms are likely to characterize the subgroups. Our paper adds a new tool to Bayesian mixture models that contributes to answering these questions.

Combinatorial mixtures relates to product partition [14–17] and Dirichlet process mixture (DPM) [18, 19] models, which allow for parameters to be clustered in subsets. Variants of the standard formulations of these models may be considered that are similar in spirit to *combinatorial mixtures*. However, our approach is more oriented to provide experts with direct tools for elicitation of *a priori* information and visualization of *a posteriori* evidence from the fitted mixture models. This is a key feature as mixture models have been increasingly used in recent years to model problems with a direct application in clinical practice.

The remainder of the paper is organized as follows. In Section 2 we introduce the notation and missing data formulation for Bayesian mixtures. In Section 3 we propose a general formulation for *combinatorial mixtures*. In Section 4 *combinatorial mixtures* are introduced in the context of normal mixture models. The performance of the methodology is then illustrated in an application to gene expression data from lung and prostate cancer patients in Section 5.

2 Bayesian methods for mixtures: notation and missing data formulation

In a finite mixture model, observations $x^n = (x_1, \dots, x_n)$ are assumed to be conditionally independent as follows

$$x_i|K, \theta, \omega \sim p(x_i|K, \theta, \omega) = \sum_{k=1}^K \omega_k p(x_i|\theta_k), \quad i = 1, \dots, n, \quad (1)$$

where K is the number of components, $\omega = (\omega_1, \dots, \omega_K)$ are the mixture weights – constrained to be non-negative and to sum to unity – and $\theta = (\theta_1, \dots, \theta_K)$ is the component-specific parameter vector. A *fully* Bayesian analysis [48] requires specification of a prior distribution on the parameters (K, θ, ω) . The number of components and the mixture components parameters are modeled jointly and inference about these quantities is based on their posterior distributions.

In the missing data formulation of mixture models, each observation x_i , $i = 1, \dots, n$, comes from a specific, but unknown, component, z_i , of the mixture. Model (1) can be written in terms of the missing data, with $z_1, \dots, z_n$ assumed to be realizations of conditionally independent and identically distributed discrete random variables, $z_1, \dots, z_n$, with probability mass function:

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

Conditional on the z s, $x_1, \dots, x_n$ are independent observations from densities:

$$p(x_i|z_i = k, \theta, \omega) = p(x_i|\theta_k), \quad i = 1, \dots, n, \quad k = 1, \dots, K.$$

Integrating out $z_1, \dots, z_n$ yields model (1).

Group belonging indicators may be known, with the so called *supervised* model representing K distinct subpopulations with specified distributions, as opposed to the general *unsupervised* case with unknown class indicators. Either settings are amenable to the modelling strategy that we propose. In the *supervised* case, the main goal of the analysis is simply that of generating hypotheses about which mechanisms are likely to characterize different and known subpopulations in the data.

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, assume the following mixture model for the $(J \times 1)$ column vector, $x_i, J \geq 1$:

$$x_i | \boldsymbol{\theta}, \boldsymbol{\omega} \sim p(x_i | \boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{k=1}^{K^*} \omega_k p(x_i | \boldsymbol{\theta}_k), \quad i = 1, \dots, n, \quad (2)$$

where the k -th component-specific parameter $\boldsymbol{\theta}_k = (\theta_k^1, \dots, \theta_k^d, \dots, \theta_k^D)^T$ is a column vector listing all the D parameters characterizing the k -th 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.

After assuming prior independence between parameters $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$, and the standard Dirichlet distribution as the prior for the weights: $\boldsymbol{\omega} \sim \text{Dir}(a_0)$, $a_0 = (a_{1,0}, \dots, a_{K^*,0})$, one may specify the following prior distribution on any $\boldsymbol{\theta}^d = (\theta_1^d, \dots, \theta_k^d, \dots, \theta_{K^*}^d)$, $d = 1, \dots, D$, that allows for degeneracy along equality constraints:

$$\boldsymbol{\theta}^d = (\theta_1^d, \dots, \theta_k^d, \dots, \theta_{K^*}^d) | \gamma^d, H \text{ i.i.d.} \sim \prod_{k: \theta_k^d \in \tilde{\boldsymbol{\theta}}^d(\gamma^d)} p(\theta_k^d | H), \quad d = 1, \dots, D, \quad (3)$$

and

$$\gamma^d | \boldsymbol{\pi}^d \text{ i.i.d.} \sim \text{Multi}(1, \boldsymbol{\pi}^d), \quad \boldsymbol{\pi}^d : \pi_l^d \geq 0, \quad \sum_{l=0}^{B_{K^*}-1} \pi_l^d = 1, \quad d = 1, \dots, D, \quad (4)$$

where $\tilde{\boldsymbol{\theta}}^d \in \text{Unique}(\boldsymbol{\theta}^d)$ – meaning that $\tilde{\boldsymbol{\theta}}^d$ is (one of) the U_d -dimensional row vectors similar to $\boldsymbol{\theta}^d$ but with the duplicate elements suppressed – $\tilde{\boldsymbol{\theta}}^d(\gamma^d)$ indicates that the product is carried out over the different elements of $\boldsymbol{\theta}^d$, as induced by the $\boldsymbol{\theta}^d$ -sharing pattern variable γ^d , H is a set of suitable hyperparameters, $\text{Multi}(\cdot | 1, \boldsymbol{\pi}^d)$ indicates a multinomial distribution with parameters equal to 1 and $\boldsymbol{\pi}^d$, and B_{K^*} is the Bell exponential number representing the number of possible degeneracy patterns. Note that:

- The product is carried out over U_d probability distributions;
- Each value of γ^d is associated with more than one U_d -dimensional $\text{Unique}(\boldsymbol{\theta}^d)$ vector; we indicate with $\tilde{\boldsymbol{\theta}}^d$ one of these vectors (see Section 4.1 for details on our specific choice).

With an alternative notation, if we indicate with γ_l^d any value of γ^d , eqs (3) and (4) may be jointly written as:

$$\boldsymbol{\theta}^d = (\theta_1^d, \dots, \theta_k^d, \dots, \theta_{K^*}^d) | H \text{ i.i.d.} \sim \sum_{l=0}^{B_{K^*}-1} \pi_l^d \prod_{k: \theta_k^d \in \tilde{\boldsymbol{\theta}}^d(\gamma_l^d)} p(\theta_k^d | H), \quad d = 1, \dots, D, \quad (5)$$

where the mixture form of the prior distribution is in evidence.

Well-known mixture modeling scenarios included in our formulation are the following ones:

- No sharing: for every d , $U_d = K^*$;
- Complete sharing: for every d , $U_d = 1$, and therefore $K = 1$;
- Partial sharing (mixed version of the previous cases): $U_d = K^*$ for some d s and $U_d = 1$ for some others.

However, extra scenarios are provided by our approach in the “Partial sharing” set-up in the following directions:

- For every d , $1 < U_d < K^*$;
- For different d s, subgroups of shared parameters may be different.

Consider the following example. 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 mixtures* class may put *a priori* some mass in the following way: say parameter vectors θ^d and $\theta^{d'}$ assume the following values $\theta^d = (\lambda_1, \lambda_2, \lambda_1, \lambda_3, \lambda_1)$ and $\theta^{d'} = (\lambda'_1, \lambda'_2, \lambda'_3, \lambda'_2, \lambda'_4)$. In this case:

$$\tilde{\theta}^d = (\lambda_1, \lambda_2, \lambda_3) \quad \text{and} \quad \tilde{\theta}^{d'} = (\lambda'_1, \lambda'_2, \lambda'_3, \lambda'_4),$$

meaning that some elements are shared for each dimension. The collection of sets with shared components indices, $E_h^d = \{k \in \{1, \dots, K^*\} : \theta_k^d = \tilde{\theta}_h^d\}$, 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.

The Bell number, B_n , is defined as the number of partitions of a set A_0 of size n and, therefore, it counts all the possible comparisons involving the elements of θ^d across the K^* groups. For any number of groups, the Bell number is finite, though it could be very large.

Prior independence between parameters θ_k^d , for fixed d , is assumed, but it is not a key aspect of *combinatorial mixtures*. Details on variance-covariance matrix decompositions and on the prior distributions of the remaining parameters are given in the context of normal mixtures in the next section.

4 Combinatorial mixtures of normal distributions

4.1 Model specification

Data $x^n = (x_1, \dots, x_n)$ are assumed to be independent observations from a mixture density with K^* bivariate normal components:

$$x_i | \mu, \Sigma, \omega \sim \sum_{k=1}^{K^*} \omega_k N_2(x_i | \mu_k, \Sigma_k), \quad x_i = \begin{bmatrix} x_i^1 \\ x_i^2 \end{bmatrix}, \quad i = 1, \dots, n,$$

where:

$$\mu = (\mu_1, \dots, \mu_{K^*}), \quad \Sigma = (\Sigma_1, \dots, \Sigma_{K^*}), \quad \mu_k = \begin{bmatrix} \mu_{k1} \\ \mu_{k2} \end{bmatrix}, \quad \Sigma_k = \begin{bmatrix} \sigma_{11k}^2 & \sigma_{12k} \\ \sigma_{21k} & \sigma_{22k}^2 \end{bmatrix}, \\ k = 1, \dots, K^*,$$

and $N_2(\cdot | \mu_k, \Sigma_k)$ indicates a bivariate normal distribution with expectation vector μ_k and variance-covariance matrix Σ_k , Σ_k positive definite.

In this case, $J = 2$, $D = 5$, and $d \in \{m_1, m_2, v_1, v_2, c\}$, meaning that we potentially allow all γ^d -induced patterns of sharing of means (in the following indicated with m_1 and m_2 for $j = 1$ and $j = 2$, respectively), variances (similarly, indicated with v_1 and v_2) and covariances (indicated with c), respectively, among the K^* components.

For the modeling of the variance-covariance structure, we adopt the following direct decomposition proposed by Barnard et al. [20]:

$$\Sigma_k = \text{diag}(S_k) R_k \text{diag}(S_k), \quad S_k = \begin{bmatrix} S_{k1} \\ S_{k2} \end{bmatrix}, \quad R_k = \begin{bmatrix} 1 & r_k \\ r_k & 1 \end{bmatrix}, \quad k = 1, \dots, K^*,$$

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

$$S = (S_1, \dots, S_{K^*}), \quad R = (R_1, \dots, R_{K^*}).$$

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.

After the following prior independence assumptions on the new parameter set (ω, μ, S, R) :

$$\omega \perp (\mu, S, R), \quad \mu \perp (S, R), \quad S \perp R,$$

combinatorial mixtures for normal mixture models may be specified through the following priors on the component-specific parameters:

$$\begin{aligned} \mu^j &= (\mu_{1j}, \dots, \mu_{K^*j}) | \gamma_j^m, \eta^2 \text{ i.i.d. } \sim \prod_{k: \mu_{kj} \in \tilde{\mu}^j(\gamma_j^m)} N(\mu_{kj} | \eta^2), \quad j = 1, 2, \\ \log S^j &= (\log S_{1j}, \dots, \log S_{K^*j}) | \gamma_j^v, e^2 \text{ i.i.d. } \sim \prod_{k: S_{kj} \in \tilde{S}^j(\gamma_j^v)} N(\log S_{kj} | e^2), \\ & \quad j = 1, 2, \end{aligned}$$

$$r = (r_1, \dots, r_{K^*}) | \gamma^c \text{ i.i.d. } \sim \prod_{k: r_k \in \tilde{r}(\gamma^c)} U_{(-1,1)}(r_k),$$

and on the γ^d indicators of sharing of means, standard deviations, and/or correlations among the mixture components:

$$\gamma_j^m | \pi^m \text{ i.i.d. } \sim \text{Multi}(1, \pi^m), \quad \gamma_j^v | \pi^v \text{ i.i.d. } \sim \text{Multi}(1, \pi^v), \quad \gamma^c | \pi^c \sim \text{Multi}(1, \pi^c), \quad j = 1, 2,$$

or, alternatively, in the univariate case, we opt for a slightly different specification:

$$\sigma^2 | \gamma^v, c, d \sim \prod_{k: \sigma_k^2 \in \tilde{\sigma}^2(\gamma^v)} IG(\sigma_k^2 | c, d) \quad \text{and} \quad \gamma^v | \pi^v \sim \text{Multi}(1, \pi^v).$$

We assume prior independence between each component-specific parameter, μ , S and r , and the γ s corresponding to the other ones. In addition, we have that $\log(S_{kj} | 0, e^2) \sim N(S_{kj} | 0, e^2)$ indicates that S_{kj} is distributed according to a lognormal distribution with parameters equal to 0 and e^2 , $IG(\cdot | c, d)$ indicates an inverse-gamma distribution with parameters c and d , $c, d > 0$ (using the parameterization in which the mean is $d/(c-1)$, $c > 1$), $U_{(-1,1)}(\cdot)$ indicates a uniform distribution on $(-1, 1)$, and η^2 , c , d and e^2 are known, π^m , π^v , π^c are known, non-negative and they sum to unity. Note that for R_k be positive definite, $r_k \neq \pm 1$. We specify our priors on the standard deviations and correlations following practical suggestions in Barnard et al. [20], who suggested the independent log-normal priors to be more flexible than the scaled inverted chi-squared distributions in one of their applications, too. In addition, prior means of 0 are assumed for the normal and lognormal distributions of the component-specific means and standard deviations, whereas the corresponding variances are chosen to express a vague prior information. However, our approach applies equally to different forms of the component-specific distribution and different choices of the hyperparameters.

This model specification derives from that of Section 3 after posing, for $j = 1, 2$, and $k = 1, \dots, K^*$:

$$\theta^d = \theta^{m_j} = \mu^j, \quad \theta_k^d = \theta_k^{m_j} = \mu_{kj}, \quad \theta^d = \theta^{v_j} = S^j, \quad \theta_k^d = \theta_k^{v_j} = S_{kj},$$

$$\theta^d = \theta^c = r, \quad \theta_k^d = \theta_k^c = r_k,$$

$$\gamma^d = \gamma^{m_j} = \gamma_j^m, \quad \gamma^d = \gamma^{v_j} = \gamma_j^v, \quad \pi^d = \pi^{m_j} = \pi_j^m = \pi^m, \quad \pi^d = \pi^{v_j} = \pi_j^v = \pi^v.$$

Now, assume $K^* = 3$. For any d , we can re-express the corresponding γ^d values looking at the (3×3) matrix, P^d , showing all the possible pairwise comparisons between elements in θ^d across the three components, where $p_{k,k'} = 0$ iff $\theta_k^d = \theta_{k'}^d$, $p_{k,k'} = 1$ iff $\theta_k^d \neq \theta_{k'}^d$ for $k, k' = 1, 2, 3$, d fixed. From the corresponding upper triangular matrix one can obtain a (1×3) vector after deleting the diagonal elements (which are all equal to 0) and juxtaposing the remaining rows of the matrix. For instance, we represent the matrix

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad \text{with the vector} \quad (1, 1, 1),$$

and, consistently, we pose (see Section 3):

$$\gamma_l^d = 4 \text{ iff } \gamma_l^d = \{p_{1,2}, p_{1,3}, p_{2,3}\} = \{1, 1, 1\} = 111,$$

and, similarly, for the remaining γ_l^d s:

$$0 = 000, \quad 1 = 110, \quad 2 = 101, \quad 3 = 011.$$

For example, in the univariate set-up, we have:

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

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

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

where the following shorthand notation:

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

stands for:

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

with y_1, y_2, y_3 univariate random variables, y_3 assuming values in $\{0, 1\}$.

Note that we implicitly opt for the following specification of $\tilde{\theta}^d$ (see Section 3 for details):

$$\begin{aligned} \text{if } \gamma^d = 000 & \quad \text{then } \tilde{\theta}^d = (\theta_1^d) \\ \text{if } \gamma^d = 110 & \quad \text{then } \tilde{\theta}^d = (\theta_1^d, \theta_2^d) \\ \text{if } \gamma^d = 101 & \quad \text{then } \tilde{\theta}^d = (\theta_1^d, \theta_2^d) \\ \text{if } \gamma^d = 011 & \quad \text{then } \tilde{\theta}^d = (\theta_1^d, \theta_3^d) \\ \text{if } \gamma^d = 111 & \quad \text{then } \tilde{\theta}^d = (\theta_1^d, \theta_2^d, \theta_3^d), \end{aligned}$$

although other specifications of *a priori* collapsed components, like:

$$\begin{aligned}
&\text{if } \gamma^d = 000 \quad \text{then } \tilde{\theta}^d = (\theta_1^d) \\
&\text{if } \gamma^d = 110 \quad \text{then } \tilde{\theta}^d = (\theta_1^d, \theta_3^d) \\
&\text{if } \gamma^d = 101 \quad \text{then } \tilde{\theta}^d = (\theta_1^d, \theta_2^d) \\
&\text{if } \gamma^d = 011 \quad \text{then } \tilde{\theta}^d = (\theta_1^d, \theta_3^d) \\
&\text{if } \gamma^d = 111 \quad \text{then } \tilde{\theta}^d = (\theta_1^d, \theta_2^d, \theta_3^d),
\end{aligned}$$

are equivalent with respect to our formulation of *combinatorial mixtures*. This renaming of the γ^d s values equally applies to the bivariate case. Although his usefulness is limited to $K^* = 3$ or $K^* = 4$, it has a relevant practical motivation, as practitioners may directly identify the number and indexes of collapsed components for any variable.

4.2 Computation

Bayesian inference for mixture models is carried out via Markov Chain Monte Carlo (MCMC) methods (see Gottardo and Raftery [21] for a general framework where to use MCMC algorithms when the target distribution is a mixture of mutually singular distributions (i.e. of several distributions of different dimensions), like in our case). A sampled realization of the Markov chain $\{\Phi^{(t)}\}$, with posterior distribution $p(\mathbf{x}^n)$ as its stationary distribution, is produced generating iteratively the parameters, $^{(t)}$, and the missing data, $\mathbf{z}^{n(t)} = (z_1^{(t)}, \dots, z_n^{(t)})$, according to $p(\mathbf{x}^n, \mathbf{z}^{n(t)})$ and $p(\mathbf{z}^n | \mathbf{x}^n, ^{(t+1)})$ respectively.

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

1. Updating the vectors $(\boldsymbol{\mu}^j, \gamma_j^m)$, $j = 1, 2$;
2. Updating the vectors $(\mathbf{S}^j, \gamma_j^v)$, $j = 1, 2$;
3. Updating the vector $(\mathbf{r}, \gamma^c)$;
4. Updating the weights $\boldsymbol{\omega}$;
5. Updating the allocation variables vector $\mathbf{z}^n$.

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 $\mathbf{z}^n$ (see Appendix Section A.1 and Section A.2).

Move types 1., 2., and 3. follow from *combinatorial mixtures* assumptions and involve a change in dimension. We applied the Gibbs sampler for drawing $(\boldsymbol{\mu}^j, \gamma_j^m)$, $j = 1, 2$. At each iteration, we updated γ_j^m integrating out $\boldsymbol{\mu}^j$ from the posterior distribution of $(\boldsymbol{\mu}^j, \gamma_j^m)$; then we updated $\boldsymbol{\mu}^j$ conditioning on γ_j^m using conjugacy [22]. We applied three Metropolis-Hastings steps for drawing $(\mathbf{S}^j, \gamma_j^v)$, $j = 1, 2$, and $(\mathbf{r}, \gamma^c)$ (see Appendix Section A.1 and Section A.2 for details). Move types 4. and 5. follow [22].

In the univariate normal mixture model, vector $(\boldsymbol{\sigma}^2, \gamma^v)$ is updated via Gibbs sampler, as described for $(\boldsymbol{\mu}^j, \gamma_j^m)$ in the bivariate set-up.

4.3 Label switching

Meaningful parametric inference in mixture models requires to tackle the label switching problem [8, 23–25], which is the invariance of representation eq. (1) under permutations of the class labels. In a Bayesian context, this issue becomes more acute, because class labels may permute during the simulation run. If symmetric priors (where *a priori* all components have the same distribution) are placed upon the parameters of a mixture model, then the resulting posterior distribution is invariant to permutations in the labelling of the parameters. As a result, the usual practices of summarizing joint posterior distributions by marginal distributions and estimating quantities of interest by their posterior means are inappropriate.

An elementary solution to this issue [26] is to summarize inference 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. As a by-product, we implement a

graphical representation of the corresponding matrix of relative frequencies (named the “O’Hagan” matrix in the following), to visualize clusters in the data.

However, when the goals of the analysis include both parameter estimation and clustering and the prior structure is symmetric (as in our case), an effective posterior inference requires to use some of the relabelling methods that have been proposed to solve the label switching issue in Bayesian inference (see Jasra et al. [27], Zhu and Fan [28]). In the current paper, we implement four available relabelling methods, the “Data-based” [29], the “ECR iterative” version 1 and 2 [30], and the “Stephens” Kullback-Leibler based relabelling algorithm [24]. We also propose a fifth relabelling algorithm, the “Class president”, which adapts an idea of Chung et al. [31]. *A priori*, we identify $K^* - 1$ “class president” observations. When all K classes have at least one observation, each president is assigned to a different class with probability one. When postprocessing, we relabel observations consistent with this constraint: at each iteration, if a president’s allocation variable has changed, we replace the allocation variables of its sample, as well as those of all samples who are in the same group at that iteration, with the original label of the president. Chung et al. [31] considers a fixed number of components. We also need to consider the cases in which the number of components becomes smaller than K^* and the cases in which presidents may be assigned to the same class. To choose presidents, if $K^* = 3$, we use two samples at the extremes of the empirical distribution of the data. This labels the two most extreme groups while the third, if present, is labeled by exclusion.

In our application, we separately apply each of the relabelling algorithm to each subset of the parameter space that is jointly sampled in the MCMC procedure (i.e. (μ^j, γ_j^m) , (S^j, γ_j^v) , with $j = 1, 2, (r, \gamma^c)$, and ω). In the case of a component-specific parameter subset, we apply the relabelling algorithm to the component-specific parameter set (say, μ^1) and we consistently re-define the corresponding γ^d indicator (say, γ_1^m) in accordance with the permuted component-specific parameters. We finally derive the posterior estimates of the parameters from the MCMC chains after each relabelling algorithm is applied.

Combinatorial mixtures apply, as a special case, to the supervised context, with (obviously) no label switching issues.

4.4 Implications of combinatorial mixtures

In the following, we describe two aspects that follow from our definition of *combinatorial mixtures*. We refer to the univariate case with $K^* = 3$ for simplicity.

Parameters γ^m and γ^v are connected with the unknown number of mixture components K . We, indeed, derive the corresponding prior distribution on K from the joint prior distribution of (γ^m, γ^v) 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:

$$\begin{aligned}
 P\{K = 1\} &= \pi_{000}^m \pi_{000}^v, \\
 P\{K = 2\} &= \pi_{000}^m \pi_{110}^v + \pi_{000}^m \pi_{101}^v + \pi_{000}^m \pi_{011}^v + \pi_{110}^m \pi_{000}^v + \pi_{110}^m \pi_{110}^v + \\
 &\quad + \pi_{101}^m \pi_{000}^v + \pi_{101}^m \pi_{101}^v + \pi_{011}^m \pi_{000}^v + \pi_{011}^m \pi_{011}^v, \\
 P\{K = 3\} &= \pi_{000}^m \pi_{111}^v + \pi_{110}^m \pi_{101}^v + \pi_{110}^m \pi_{011}^v + \pi_{110}^m \pi_{111}^v + \pi_{101}^m \pi_{110}^v + \\
 &\quad + \pi_{101}^m \pi_{011}^v + \pi_{101}^m \pi_{111}^v + \pi_{011}^m \pi_{110}^v + \pi_{011}^m \pi_{101}^v + \pi_{011}^m \pi_{111}^v + \\
 &\quad + \pi_{111}^m \pi_{000}^v + \pi_{111}^m \pi_{110}^v + \pi_{111}^m \pi_{101}^v + \pi_{111}^m \pi_{011}^v + \pi_{111}^m \pi_{111}^v.
 \end{aligned}$$

In addition, in the presence of a symmetric prior structure and with no constraints posed by relabelling algorithms on single parameters or combinations of them, label switching applies equally to inference on γ^m and γ^v : combinations of (γ^m, γ^v) with different names actually correspond to the same mixture model for the data. Figure 1 depicts in a stylized fashion all the available combinations of (γ^m, γ^v) for $K^* = 3$ and marks the equivalent ones with the same color. The original twenty-five possible (γ^m, γ^v) combinations are mapped into ten effective ones, identified using the same color. The corresponding density estimation plots and number of components are superimposed. Corresponding prior (posterior) probabilities are summed when reporting results.

Figure 1 Mixture model templates corresponding to each (γ^m, γ^v) pair for *combinatorial mixtures* priors in a normal mixture model with K^* equal to 3. The number of components and a typical shape are superimposed. (a) Color-coded cells identify the ten mixture models with the same shape; (b) The grey cells identify the twelve extra models that *combinatorial mixtures* allow to fit, in comparison with traditional approaches to mixture models.

4.5 Comparison with available mixture models

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. In the univariate set-up, Figure 1 shows the thirteen white cells available to traditional approaches, and highlights in grey the twelve extra models that *combinatorial mixtures* allow. The thirteen 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 twelve grey cells correspond to the following scenarios:

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

The same argument applies to single variables in the bivariate set-up.

Traditional mixture models with an unknown number of components generally specify a prior directly on the unknown number of mixture components K . We previously showed that parameters γ^m and γ^v are connected with K (see Section 4.4 for an example in the univariate set-up), so there is no difference in specifying a prior on K or specifying a prior on the γ indicators. In addition, traditional bivariate mixture models do not generally assume any decomposition of the variance-covariance matrix. The available options are sharing/no sharing of the variance-covariance matrix across components. The former option implies sharing of both variances and covariances: $\gamma_1^v = 000, \gamma_2^v = 000$, and $\gamma^c = 000$ in our context. The latter option includes all the (very different) remaining cases. The adoption of a decomposition allows to share the variances, without necessarily sharing the covariances (or viceversa), and represents the easiest way to a free modeling of the variance-covariance structure in traditional mixture models too. However, even when a decomposition is assumed for the variance-covariance matrix, one still has the following possibilities: complete sharing/no sharing of the variances or complete sharing/no sharing of the covariances across components. In any of these cases, all the variances (or covariances) should be either equal or different across components. *Combinatorial mixtures* go further in this direction of modeling each variance or covariance in a freer 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 [9].

Combinatorial mixtures relates to 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 is 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 [14, 15] of Crowley [16], where we have two or three sets of partitions, one for the means, one for the variances/standard deviations, and, eventually, one for the correlations. Our priors were specified directly on equality events, though this will induce priors on partition sets E_{μ}^d . A similar comment applies to nonparametric product partition models as proposed by Dahl [17], which require 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.

Mixtures of Dirichlet Process priors [32], as typically implemented for normal models (see Escobar and West [18] and West and Turner [19]), allow for parameters to be clustered in subsets [33]. In the original form, 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. However, one can specify (say, in the univariate case) two separate Dirichlet process priors, G and H , with $\mu_i | G \sim G$ and $\sigma_i^2 | H \sim H$, $i = 1, \dots, n$, with a formulation that results in two separate partitions of the n observations, one based on equality of means and the other based on equality of variances.

5 Gene expression in lung and prostate cancers

Figure 2 Histograms of the expression levels for the genes TRIM29, MSN, and ITGB5 in the lung cancer dataset.

Next, we fit the normal mixture model of Section 4.1 to data on the molecular classification of lung and prostate cancers [34–37]. The former dataset comes from the web-based information supporting the published manuscript Garber et al. [38]. The study is performed by scientists at the Dana Farber Cancer Center 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. The latter dataset comes from the effort of “The Cancer Genome Atlas Research Network” who has comprehensively characterized 333 primary prostate carcinomas using seven genomic platforms, to gain further insight into the molecular-genetic heterogeneity of primary prostate cancer and to establish a molecular taxonomy of the disease [39]. Here, we consider the mRNA expression data available for download at <https://tcga-data.nci.nih.gov> (last access: October, 30th, 2016). For these data, gene level RPKM values from RNA-seq were log2 transformed and filtered to remove low-variability genes (bottom 25% removed, based on interdecile range). For both datasets, our objectives are: (a) estimating the number of subgroups in the sample; (b) making inference about the assignment of samples to these subgroups; and (c) generating hypotheses about

which mechanisms are likely to characterize the subgroups. As we were seeking a novel molecular classification of tumor subtypes, we removed the 17 normal samples from the lung cancer dataset.

We searched for genes whose expression data provide interesting scenarios to fit *combinatorial mixtures*. After visual inspection of the heatmap, we carried out some preliminary analyses according to the Fraley and Raftery approach [40] and the R package mclust [41]. In addition, we plotted the scatter graphs of a few promising pairs of genes to improve the presentation of the bivariate application of the mixture model. In the lung cancer dataset, we selected genes with HUGO names: TRIM29, MSN, ITGB5, and CSTB. In the prostate cancer dataset, we selected the ERG and STOM genes. In the univariate analyses, the selected genes showed evidence of one, two, or three groups in the data, with possibly unequal variances in the groups. In the bivariate analyses, both pairs of genes showed evidence of two groups.

In the following, we present separate univariate models for the TRIM29, MSN, and ITGB5 genes from the lung cancer dataset. Results for CSTB and STOM expression data are equivalent to those of the ITGB5 gene and of limited interest for our analysis; results for the ERG expression data are equivalent to those of the MSN gene (data not shown). They are not described in the current application. We then present two applications of the bivariate model to the MSN and CSTB genes from the lung cancer dataset and to the ERG and STOM genes from the prostate cancer dataset. In the former case, we expect the component-specific correlation coefficients between genes to be close to 0, whereas, in the latter case, we expect them to be markedly different in their sign.

Figure 3 Two-dimensional kernel density estimation plot for the bivariate distribution of the MSN and CSTB genes in the lung cancer dataset and of the ERG and STOM genes in the prostate cancer dataset.

In the univariate analysis, Figure 2 shows the histograms of the expression levels for the three genes selected from the lung cancer dataset. TRIM29 and MSN genes are promising for classification of lung cancer patients, showing long tails on the right and on the left side of the distribution, respectively. Accordingly, results from the Fraley and Raftery approach identify as the best BIC (Bayesian Information Criterion) model for TRIM29 a three-component normal mixture with unequal variances, and for MSN a two-component normal mixture with unequal variances. On the other hand, the ITGB5 distribution seems more similar to a single normal distribution, as suggested by the BIC values in the Fraley and Raftery approach too. In the bivariate application, Figure 3 shows the two-dimensional kernel density estimation plots representing the joint distribution of the MSN and CSTB genes in the lung cancer dataset (left) and that of the ERG and STOM genes in the prostate cancer dataset (right). For both pairs of genes, the plot suggests the existence of two groups. Results from the Fraley and Raftery approach point to a two-component “diagonal, varying volume and shape” mixture model for the former pair and to a two-component “ellipsoidal, varying volume, shape, and orientation” for the latter one, according to the best BIC values [41]. In addition, the assumption of a maximum number of components equal to 3 is not a stringent limitation in the examined set-ups.

For each application, we run one chain at a time, doing approximately 50 simulations that appear to converge to the same distribution. We report results from a representative chain among them, corresponding to a total of 2,000,000 iterations with a burn-in of 500,000 iterations. Standard convergence diagnostics, including the use of the autocorrelation function, and those by Geweke, Heidelberger and Welch, and Raftery and Lewis, pointed to our final choice for the total number of iterations and number of burn-in iterations [42].

In both applications, all our runs start with the γ s being equal to 000 and the starting values for the means, standard deviations, and correlations being equal to the corresponding sample values, in the hypothesis of one group in the data.

We consider Bayesian estimation in the case where we do not have strong prior information on the parameters.

In the univariate application, the following choice of hyperparameters results in weakly informative priors on the means μ_1, μ_2, μ_3 , and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$: $\eta^2 = 5000, c = 0.75, d = 0.15$. Similarly, in the bivariate application, we choose the following hyperparameters for the prior distributions of the means and standard deviations: 1. $\eta^2 = 5000, e^2 = 4$, for the lung cancer dataset; 2. $\eta^2 = 3600, e^2 = 2.25$, for the prostate cancer dataset. The hyperparameters of the priors on the standard deviations (variances) are chosen looking at the range of the standard

deviations (variances) of typical microarrays (Affymetrix Hu95A) and RNA-Seq (Illumina HiSeq) gene expression data and adding some extra variability. In both the applications, the Dirichlet prior on the weights, ω , is symmetric with $a_0 = (1, 1, 1)$ and the multinomial priors on the γ s are vague in the sense of giving the same *a priori* probability to all the possible values of π^m , π^v , and π^c : $\pi^m = \pi^v = \pi^c = (0.2, 0.2, 0.2, 0.2, 0.2)$. The effect of different *a priori* scenarios for ω and the γ s is investigated in the sensitivity analysis presented in Section 5.3 and Appendix Section A.3 for the univariate model.

For the implementation of the “Class president” relabelling algorithm, we choose the following class presidents among the available samples:

- Univariate application:
 - TRIM29: AD119=5.18; AD295=8.00;
 - MSN: COID3580=8.46; SQ8T1=8.98;
 - ITGB5: AD183=7.42; AD10=7.69.
- Bivariate application:
 - MSN and CSTB: COID3580=(8.46, 8.43); SQ13T1=(10.10, 10.05);
 - ERG and STOM: ID183=(9.99, 27.36); ID159=(86.35, 58.17),

where the samples are indicated as in the original dataset, the corresponding gene expression levels are provided for each gene under consideration, and, in the bivariate application, the former expression level is for gene MSN (or ERG) and the latter for gene CSTB (or STOM).

We present the estimated joint distributions (%) of (γ^m, γ^v) for each gene, together with the marginal distribution of γ^c in the bivariate case. We marked in boldface the frequencies of those combinations that are only available with *combinatorial mixtures*, in comparison with traditional approaches. We plot the marginal posterior distributions of the component-specific parameters, together with those of the corresponding γ indicators. We summarize posterior inference on the relabelled chains through the posterior medians of the parameters calculated on the entire chain and conditional on the ten effective cells showing the same color. Moreover, we report the “O’Hagan” matrices to present clusters of subjects 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 samples.

Calculations are performed using the open-source statistical computing environment R [43, 44], packages MCMCpack [45], label.switching [46], and coda [47] and a specialized code reflecting the procedure described in Section 4.2 and the results provided in Appendix Section A.1 and Section A.2.

5.1 Application of the univariate model

Figure 4 Joint posterior probabilities (%) of (γ^m, γ^v) for each gene in the univariate application to the lung cancer dataset, as obtained by one run of the simulation. For a correct interpretation, one has to sum those frequencies indicated in the cells with the same color. The frequencies of those extra combinations fitted using *combinatorial mixtures*, in comparison with traditional approaches, are emphasized using boldface.

Figure 4 shows the estimated joint posterior probabilities of (γ^m, γ^v) for the selected genes from the lung cancer dataset. The TRIM29 chain spends $\sim 51\%$ of the sweeps on $(111, 111)$ (grey cell), which represents a three-component mixture with different means and different variances. The second most probable combination ($\sim 39\%$ of the sweeps) is given by $(111, 110)$ or $(111, 101)$ or $(111, 001)$ (purple cells). The corresponding mixture model has three components, with three different means and two variances. The third most probable combination (5% of the sweeps), $(110, 111)$ or $(101, 111)$ or $(011, 111)$ (red cells), implies a mixture model with three components, with two different means and three different variances. *A posteriori*, a mixture model with three components with at least two different means or variances is suitable for TRIM29. The chain spends $\sim 47\%$ of the sweeps on the twelve combinations of (γ^m, γ^v) that would not be possible to select using traditional approaches to mixture models. The majority of the 47% is given by two (purple and red) of the three top combinations. As traditional approaches would place a higher portion of the overall mass on the $(111, 111)$ combination, as compared to *combinatorial mixtures*, our approach allows for a gain in parsimony in this case. This gain is represented by the purple and red cells, where either a mean or a variance is collapsed to another one.

The MSN chain spends $\sim 33\%$ of the sweeps on a mixture model with two components, with one shared mean and two different variances (orange cells). The second most probable combination ($\sim 19\%$) is given by $(111, 110)$ or $(111, 101)$ or $(111, 001)$ (purple cells), which represent a three-component mixture model with three different means and two different variances. The third most probable combination ($\sim 19\%$) 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, with two of them sharing the means and the other two sharing the variances. *A posteriori*, a two- or three-component mixture model is suitable for MSN. The chain spends $\sim 50\%$ of the iterations on the twelve 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 (purple and light-blue cells). Compared to traditional three-component mixtures, *combinatorial mixtures* still allow for a more parsimonious solution.

The ITGB5 chain spends $\sim 55\%$ of the sweeps 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 combination (19%) is given by the light-blue cells $((\gamma^m, \gamma^v) = (110, 101)$, 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 $(\gamma^m, \gamma^v) = (011, 101)$), corresponding to a three-component mixture model, with two components sharing the means and the other two sharing the variances. The third most probable combination ($\sim 11\%$) is $(000, 111)$ (brown cell), corresponding to a three-component mixture model with one shared mean and three different variances. The chain spends $\sim 66\%$ of the sweeps on $\gamma^m = 000$ and groups are eventually created by differences in variances. Combinations where $\gamma^v = 000$ are not supported by our model. The chain spends $\sim 31\%$ of the sweeps on the twelve combinations of (γ^m, γ^v) that would not be possible to select using traditional approaches, with most of these sweeps represented by the light-blue combinations. Traditional approaches are not restrictive in this case.

Figure 5 Trace plots of the parameters for the TRIM29 gene in the lung cancer dataset, as estimated by one run of the simulation. Component means and corresponding standard deviations for each iteration are color-coded in the same way across plots.

TRIM29		Median value of the parameter									
Cell color	Relabelling method	mu1	mu2	mu3	sigma1	sigma2	sigma3	w1	w2	w3	% time spent
Grey	Stephens	5.32	6.25	8.11	0.08	0.17	0.14	0.68	0.18	0.14	50.63
	ECR (iterative vs. 1 and 2)	5.32	6.25	8.11	0.08	0.17	0.14	0.68	0.18	0.14	50.63
	Data based	5.32	6.25	8.11	0.08	0.18	0.14	0.69	0.18	0.14	50.63
	Class president	5.32	6.25	8.11	0.08	0.17	0.14	0.68	0.18	0.14	50.63
Purple	Stephens	5.31	6.20	8.13	0.09	0.06	0.07	0.66	0.21	0.13	38.82
	ECR (iterative vs. 1 and 2)	5.31	6.20	8.13	0.09	0.06	0.07	0.66	0.21	0.13	38.82
	Data based	5.31	6.20	8.13	0.08	0.06	0.07	0.68	0.19	0.13	38.82
	Class president	5.31	6.20	8.12	0.09	0.06	0.07	0.66	0.21	0.13	38.82
Red	Stephens	5.51	5.51	7.99	0.83	0.03	0.14	0.09	0.78	0.12	5.00
	ECR (iterative vs. 1 and 2)	5.51	5.51	7.99	0.80	0.03	0.14	0.10	0.78	0.12	5.00
	Data based	5.51	5.51	7.99	0.83	0.03	0.14	0.09	0.78	0.12	5.00
	Class president	5.51	5.51	7.98	0.76	0.03	0.16	0.11	0.78	0.12	5.00

Relabelling method		Median value of the parameter								
		mu1	mu2	mu3	sigma1	sigma2	sigma3	w1	w2	w3
Stephens		5.32	6.21	8.11	0.09	0.07	0.12	0.67	0.20	0.13
ECR (iterative vs. 1-2)		5.32	6.21	8.11	0.09	0.07	0.12	0.67	0.20	0.13
Data based		5.32	6.21	8.11	0.08	0.15	0.12	0.69	0.18	0.13
Class president		5.32	6.22	8.11	0.08	0.07	0.12	0.67	0.20	0.13

Figure 6 Estimated posterior medians of the parameters for each of the five relabelling algorithms implemented in the univariate application to the lung cancer dataset for the TRIM29 gene. These estimates were calculated on the overall chain and conditional to those cases where the chain spends more simulation time (cumulative % of simulation time spent $\geq 90\%$).

Figure 5 and Figure 6 provide additional results for the TRIM29 gene. Figure 5 shows the trace plots of the parameters μ_1, μ_2, μ_3 and $\sigma_1^2, \sigma_2^2, \sigma_3^2$, respectively, plotted along with those of the corresponding γ s. An iteration every one-hundred is reported and points smaller than the 5th percentile and bigger than the 95th percentile are suppressed to allow efficient display of the results. Component means and corresponding standard deviations are color-coded across plots. Figure 6 shows the estimated posterior medians of the parameters for each of the five relabelling algorithms implemented. These estimates were calculated on the overall chain and conditional to those cases where the chain spends more simulation time (cumulative % of simulation time spent $\geq 90\%$). The estimated posterior distributions are highly compatible with the explorative plots presented in this Section. Suppose (γ^m, γ^v) at iteration t implies a three-component mixture, with three different means and standard deviations $((\gamma^m, \gamma^v) = (111, 111)$ – grey cell in Figure 4). According to Figure 5 and Figure 6, the three means have estimated posterior medians at 5.32, 6.25 and 8.11, while the corresponding estimated medians for the standard deviations are 0.08, 0.17, and 0.14, respectively. If the combination (111, 110) or (111, 101) or (111, 011) (purple cells in Figure 4) is given at the next iteration, the posterior medians of the means are similar to those of the previous case, but one standard deviation is shared between two components (posterior median ~ 0.06), and the corresponding mixture model has three components with two different standard deviations. Similarly, if the combination (110, 111) or (101, 111) or (011, 111) (red cells in Figure 4) is given at the next iteration, two means collapsed, with a shared posterior median at ~ 5.50 . The corresponding standard deviations are fairly different, with *a posteriori* medians of 0.03 and ~ 0.80 . The remaining mean has a median slightly smaller than the 8.11 in the grey case, but still close to 8.00, and the median of the corresponding standard deviation is still ~ 0.14 . The component (5.32, 0.08) has a weight close to 0.70. The remaining ones have weights ranging from 0.10 to 0.20. The alternative relabelling algorithms provide similar posterior medians of the parameters. Similar considerations hold for the estimated posterior distributions of the parameters in the case of MSN and ITGB5 genes (data not shown).

Figure 7 “O’Hagan” matrices from the univariate analysis of the lung cancer dataset. Relative frequencies of occurrence of two samples 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 7 depicts the “O’Hagan” matrix for each of the three genes. For TRIM29 expression data, we identify three groups of patients. The first on the left is the one with the smallest mean and standard deviation and the highest weight (posterior medians of the mean, standard deviation, and weight 5.32, 0.08, and ~ 0.70 in Figure 5 and Figure 6). The intermediate block represents the component with a mean of approximately 6 (posterior medians of the mean, standard deviation, and weight ~ 6.20 , 0.12, and ~ 0.20 in Figure 5 and Figure 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 a mean around 8 (posterior medians of the mean, standard deviation, and weight 8.11, ~ 0.12 , and 0.13 in Figure 5 and Figure 6). The transition from the second to the third component is less smooth than the previous one.

For MSN expression data, we identify three alternative scenarios. In the first one (dark and light cross in the center of the plot) there is one central group of patients characterized by a posterior median of the shared mean of 9.71, posterior medians of the standard deviations of ~ 0.01 and ~ 1 and posterior medians of the weights of ~ 0.90 and ~ 0.10 , respectively (orange and brown cells in Figure 4). In a few iterations, a third group is fitted with the same shared mean, a posterior median of the standard deviation of ~ 0.10 (brown cell) or more (orange cells); its weight ranges between ~ 0.05 and ~ 0.25 , and the weight of the (9.71, ~ 0.01) component is correspondingly lower than in the previous case. In the second scenario (three separate dark and light blocks), there are three groups of patients characterized by posterior medians of the three means given by ~ 8.25 , 9.65, and 10.37, posterior medians of the standard deviations of ~ 0.10 , 0.05, and 0.05, and posterior medians of the weights of ~ 0.20 , ~ 0.40 , and ~ 0.40 , respectively (purple cells in Figure 4). In the third scenario (dark and light cross at the right of the previous cross), there is still a central group of patients characterized by the highest posterior median of the shared mean of ~ 9.88 (sharing between two components), posterior medians of the standard deviations of ~ 0.01 and ~ 0.70 (light-blue cells) or ~ 1 (red cells), and posterior medians of the weights of ~ 0.80 (or ~ 0.90) and ~ 0.10 , respectively (light-blue and red cells) (see Figure 4 for cell colors). Another group is rarely fitted with a smaller posterior median of the mean, which may be either between ~ 8.25 and ~ 9.00 or ~ 9.70 (data not shown).

For ITGB5 expression data, we identify one central group of patients characterized by a posterior median of the mean of ~ 7.50 , a posterior median of the standard deviation ranging between 0.63 and 0.73 and the smallest weight of ~ 0.10 . Partially overlapping with this one, there is a second central group of patients characterized by a similar posterior median of the mean, a definitely smaller posterior median of the standard deviation of ~ 0.02 and the highest posterior median of the weight. The posterior median of this weight ranges from ~ 0.50 to ~ 0.90 depending on the possible presence/absence of a third group characterized by a different mean (posterior median of the mean equal to ~ 7.60 or ~ 7.70 , with corresponding standard deviation being

equal to ~ 0.02) or standard deviation (posterior median of the standard deviation equal to ~ 0.10) (data not shown).

In conclusion, for each univariate set-up, *combinatorial mixtures* allow to: 1. identify a variable number of distinct groups of samples with sensible posterior medians of the component-specific parameters and corresponding weights, and 2. list a small set of visited competing scenarios of groups and corresponding parameters which account, if any (see MSN gene), for uncertainty in classification. The main conclusions are in accordance with those from the Fraley and Raftery approach. However, with *combinatorial mixtures*, the chains are allowed to spend some simulation time on more parsimonious solutions. For instance, the TRIM29 gene spend 40% of the simulation time on the red cells, where one of the three variances is shared, as well as 50% of the simulation time on the expected grey cells that represent the more general model with three different means and variances. We also gain some extra flexibility in the modelling of the component-specific variances when both the approaches agree that the mean is shared across one or two components, as with the ITGB5 gene.

5.2 Application of the bivariate model

Figure 8 Joint posterior probabilities (%) of (γ_j^m, γ_j^v) , $j = 1, 2$, and posterior probability (%) of γ^c for the MSN and CSTB genes in the bivariate application to the lung cancer dataset, as obtained by one run of the simulation. For a correct interpretation, one has to sum those frequencies indicated in the cells with the same color. The frequencies of those extra combinations fitted using *combinatorial mixtures*, in comparison with traditional approaches, are emphasized using boldface.

In this subsection, we report results derived from the application of the normal mixture model of Section 4.1 to data on the molecular classification of lung and prostate cancers. We start reporting detailed results for the joint modelling of the MSN and CSTB genes in the lung cancer dataset.

Figure 8 shows the estimated joint posterior probabilities (%) of (γ_j^m, γ_j^v) , $j = 1, 2$, and the estimated posterior probability (%) of γ^c . The MSN marginal chain spends $\sim 80\%$ of the iterations on either (110, 111) or (101, 111) or (001, 111) (red cells). The second most probable combination ($\sim 9\%$) is (111, 111) (grey cell). *A posteriori*, a mixture model with three components – with two or three different means and three different standard deviations – is suitable for MSN. The CSTB marginal chain spends $\sim 65\%$ of the iterations on either (110, 111) or (101, 111) or (001, 111) (red cells). The second most probable combination ($\sim 20\%$) is given by the six light-blue cases: $(\gamma_2^m, \gamma_2^v) = (110, 101)$, or $(\gamma_2^m, \gamma_2^v) = (110, 011)$, or $(\gamma_2^m, \gamma_2^v) = (101, 110)$, or $(\gamma_2^m, \gamma_2^v) = (101, 011)$, or $(\gamma_2^m, \gamma_2^v) = (011, 110)$, or $(\gamma_2^m, \gamma_2^v) = (011, 101)$. The third most probable combination ($\sim 7\%$) is (111, 111) (grey cell). *A posteriori*, a mixture model with three components with two/three different means and two/three different standard deviations is 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 85% of the iterations on nine out of the twelve 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 85% 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 case of one shared correlation across components ($\sim 70\%$) (dark-brown cells) or on the intermediate cases ($\sim 25\%$) of two different correlations (other cells). Traditional mixture models where a decomposition of the variance-covariance matrix is assumed allow to fit all the five cases listed in Figure 8(c), depending on how many different components are assumed; so, no cases are emphasized using boldface.

Figure 9 Trace plots of the parameters from the bivariate mixture model applied to the MSN and CSTB genes from the lung cancer dataset, as estimated by one run of the simulation. Component means, corresponding standard deviations, and correlations for each iteration are color-coded in the same way across plots.

MSN GENE		Median value of the parameter									
Cell color	Relabelling method	mu1_1	mu2_1	mu3_1	sigma1_1	sigma2_1	sigma3_1	w1	w2	w3	% time spent
Red	Stephens	8.49	9.96	10.03	0.89	0.49	0.47	0.20	0.04	0.76	81.76
	ECR (iterative vs. 1-2)	8.49	9.96	10.03	0.89	0.49	0.47	0.20	0.04	0.76	81.76
	Data based	8.49	9.96	10.03	0.89	0.50	0.47	0.20	0.04	0.76	81.76
	Class president	8.50	9.96	10.03	0.88	0.51	0.46	0.20	0.04	0.76	81.76
Grey	Stephens	8.46	9.08	10.02	0.87	0.49	0.47	0.21	0.01	0.79	8.97
	ECR (iterative vs. 1-2)	8.46	9.08	10.02	0.87	0.49	0.47	0.21	0.01	0.79	8.97
	Data based	8.46	9.09	10.03	0.87	0.50	0.47	0.21	0.01	0.79	8.97
	Class president	8.47	9.06	10.02	0.87	0.50	0.47	0.21	0.01	0.79	8.97

CSTB GENE		Median value of the parameter									
Cell color	Relabelling method	mu1_2	mu2_2	mu3_2	sigma1_2	sigma2_2	sigma3_2	w1	w2	w3	% time spent
Red	Stephens	8.72	9.86	9.89	0.53	0.59	0.57	0.20	0.04	0.76	65.03
	ECR (iterative vs. 1-2)	8.72	9.86	9.89	0.53	0.59	0.57	0.20	0.04	0.76	65.03
	Data based	8.72	9.86	9.89	0.53	0.60	0.56	0.20	0.04	0.76	65.03
	Class president	8.73	9.86	9.89	0.53	0.60	0.56	0.20	0.04	0.76	65.03
Light-Blue	Stephens	8.74	9.87	9.90	0.55	0.56	0.56	0.21	0.04	0.76	20.66
	ECR (iterative vs. 1-2)	8.74	9.87	9.90	0.55	0.56	0.56	0.21	0.04	0.76	20.66
	Data based	8.74	9.87	9.90	0.55	0.57	0.56	0.20	0.04	0.76	20.66
	Class president	8.74	9.87	9.90	0.55	0.57	0.55	0.20	0.04	0.75	20.66
Grey	Stephens	8.73	9.22	9.90	0.53	0.49	0.56	0.20	0.00	0.79	6.55
	ECR (iterative vs. 1-2)	8.73	9.21	9.90	0.53	0.49	0.56	0.20	0.00	0.79	6.55
	Data based	8.73	9.22	9.90	0.53	0.49	0.56	0.20	0.00	0.79	6.55
	Class president	8.73	9.21	9.90	0.53	0.51	0.56	0.20	0.00	0.79	6.55

		Median value of the parameter						
Cell color	Relabelling method	r1	r2	r3	w1	w2	w3	% time spent
Red-purple	Stephens	0.01	0.01	0.01	0.2	0.03	0.76	72.64
	ECR (iterative vs. 1-2)	0.01	0.01	0.01	0.2	0.03	0.76	72.64
	Data based	0.01	0.01	0.01	0.2	0.03	0.76	72.64
	Class president	0.01	0.01	0.01	0.2	0.04	0.76	72.64
Ocher	Stephens	-0.04	-0.05	-0.04	0.2	0.04	0.76	24.38
	ECR (iterative vs. 1-2)	-0.04	-0.05	-0.04	0.2	0.04	0.76	24.38
	Data based	-0.04	-0.05	-0.04	0.2	0.04	0.76	24.38
	Class president	-0.04	-0.05	-0.04	0.2	0.04	0.76	24.38

		Median value of the parameter						
Relabelling method		mu1_1	mu2_1	mu3_1	mu1_2	mu2_2	mu3_2	
Stephens		8.49	9.95	10.03	8.73	9.86	9.89	
ECR (iterative vs. 1-2)		8.49	9.95	10.03	8.73	9.86	9.89	
Data based		8.49	9.95	10.03	8.73	9.86	9.89	
Class president		8.49	9.95	10.03	8.73	9.86	9.89	
		sigma1_1	sigma2_1	sigma3_1	sigma1_2	sigma2_2	sigma3_2	
Stephens		0.89	0.49	0.47	0.54	0.57	0.56	
ECR (iterative vs. 1-2)		0.89	0.49	0.47	0.54	0.57	0.56	
Data based		0.89	0.50	0.47	0.54	0.58	0.56	
Class president		0.88	0.51	0.47	0.54	0.58	0.56	
		r1	r2	r3	w1	w2	w3	
Stephens		0	0	0	0.2	0.04	0.76	
ECR (iterative vs. 1-2)		0	0	0	0.2	0.04	0.76	
Data based		0	0	0	0.2	0.04	0.76	
Class president		0	0	0	0.2	0.04	0.76	

Figure 10 Estimated posterior medians of the parameters for each of the five relabelling algorithms implemented in the bivariate application to the lung cancer dataset. These estimates were calculated on the overall chain and conditional to those cases where the chain spends more simulation time (cumulative % of simulation time spent $\geq 90\%$).

Figure 9 shows the trace plots of the parameters μ^1 , μ^2 , S^1 , S^2 , and r , along with those of the corresponding γ s, for the MSN and CSTB genes in the lung cancer dataset. An iteration every one-hundred is reported, and

points smaller than the 5th percentile and bigger than the 95th percentile are removed from the plot to allow efficient display of the results. Corresponding estimates for each iteration are represented using the same color across plots. Figure 10 shows the estimated posterior medians of the parameters for each of the five relabelling algorithms implemented in the bivariate application to the lung cancer dataset. These estimates were calculated on the overall chain and conditional to those cases where the chain spends more simulation time (cumulative % of simulation time spent $\geq 90\%$). According to Figure 9 and Figure 10, the MSN gene has a marginal posterior distribution characterized by two means, with posterior medians around 8.50 and 10.00. The posterior medians of the corresponding standard deviations are ~ 0.90 and ~ 0.50 , respectively. For most of the simulation time ($\sim 80\%$ – red cells in Figure 8), the mean centered in 10 is shared between two components and is associated with two similar standard deviations. The corresponding weight has a posterior median of 0.80, which comes from the sum of the two corresponding medians of the posterior weights. The posterior median of the weight of the other component is 0.20. For a few iterations ($\sim 10\%$ – grey cells in Figure 8), a third different mean is fitted, with a posterior median of ~ 9.00 , but the corresponding weight is close to 0.

Similarly, the CSTB gene has a marginal posterior distribution characterized by two means, with posterior medians around 8.70 and 9.90, and corresponding posterior medians of the standard deviations of ~ 0.55 and ~ 0.60 , respectively. For most of the simulation time ($\sim 85\%$ – red and light-blue cells in Figure 8), the mean centered in 9.90 is shared between two components and is associated with slightly higher (but similar) medians of the standard deviation, as compared to the component with posterior median of the mean at 8.70. The corresponding weight has a posterior median of 0.80. The posterior median of the weight of the other component is 0.20. For a few iterations ($\sim 7\%$ – grey cells in Figure 8), a third different mean is still fitted, with a posterior median of ~ 9.20 , but the corresponding weight is equal to 0. As to the corresponding correlation coefficients, if $\gamma^c = 000$, the component correlations are collapsed at a posterior median of 0.01; if $\gamma^c = 110$ or $\gamma^c = 101$ or $\gamma^c = 011$, the component correlations have posterior medians of -0.04 and -0.05 , with -0.04 coming more frequently from sharing of components. In either case, the correlation coefficients between the two genes are approximately 0 in any component of the mixture model. *A posteriori*, the component with posterior medians of the mean vector around (8.50, 8.75), posterior medians of the standard deviations of (0.90, 0.55) and posterior medians of the correlation equal to 0 has a posterior median of the weight of 0.20. The component with posterior medians of the mean vector around (10.00, 9.90), posterior medians of the standard deviations around (0.50, 0.60) and posterior medians of the correlation equal to 0 has a posterior median of the weight of ~ 0.80 . There is no difference in the posterior medians estimated by the different relabelling methods.

Figure 11 "O'Hagan" matrix for the bivariate normal mixture model fitted to the lung cancer dataset. Relative frequencies of occurrence of two samples being in the same group are plotted in a black-to-white color scale and sorted according to a non-decreasing ordering of the MSN gene. Dark and light blocks, with proportions similar to the estimated weights of the mixture, identify different groups of observations.

The "O'Hagan" matrix is reported in Figure 11. The plot identifies two distinct groups with different mean vectors. The group on the left has the smallest mean vector (posterior medians of the mean vector around (8.50, 8.75) in Figure 9 and Figure 10) and weight (posterior medians of the weight equal to 0.20 in Figure 9 and Figure 10). The transition from the first to the second group is characterized by some uncertainty in the classification of intermediate points. The group on the right of the picture is the one with posterior medians of the mean vector around (10.00, 9.90) and the biggest weight of 0.80 (see Figure 9 and Figure 10). In conclusion, *combinatorial mixtures* allow to identify two groups of samples with very different weights – 0.20 and 0.80, respectively – located at around (8.50, 8.75) and (10.00, 9.90), with corresponding medians of the standard deviations equal

to (0.90, 0.55) and (0.50, 0.60). The posterior correlation coefficients between the two genes are close to 0 in both groups for most of the simulation time.

ERG GENE		Median value of the parameter									
Cell color	Relabelling method	mu1_1	mu2_1	mu3_1	sigma1_1	sigma2_1	sigma3_1	w1	w2	w3	% time spent
Purple	Stephens	14.55	42.71	82.46	3.56	16.86	16.86	0.53	0.08	0.39	56.28
	ECR (iterative vs. 1-2)	14.55	42.71	82.46	3.56	16.86	16.86	0.53	0.08	0.39	56.28
	Data based	14.55	42.71	82.46	3.56	16.86	16.86	0.53	0.08	0.39	56.28
	Class president	14.55	42.71	82.46	3.56	16.86	16.86	0.53	0.08	0.39	56.28
Grey	Stephens	14.54	48.07	82.98	3.58	20.08	16.10	0.54	0.10	0.37	42.65
	ECR (iterative vs. 1-2)	14.54	48.07	82.96	3.58	19.96	16.12	0.54	0.10	0.37	42.65
	Data based	14.54	48.07	82.97	3.58	20.01	16.11	0.54	0.10	0.37	42.65
	Class president	14.54	48.07	82.88	3.58	19.51	16.21	0.54	0.10	0.37	42.65

STOM GENE		Median value of the parameter									
Cell color	Relabelling method	mu1_2	mu2_2	mu3_2	sigma1_2	sigma2_2	sigma3_2	w1	w2	w3	% time spent
Orange	Stephens	45.36	45.36	45.36	12.73	7.85	12.73	0.54	0.08	0.38	64.72
	ECR (iterative vs. 1-2)	45.36	45.36	45.36	12.73	7.85	12.73	0.54	0.08	0.38	64.72
	Data based	45.36	45.36	45.36	12.73	7.85	12.73	0.54	0.08	0.38	64.72
	Class president	45.36	45.36	45.36	12.73	7.85	12.72	0.54	0.08	0.38	64.72
Brown	Stephens	45.36	45.36	45.36	12.80	7.60	12.62	0.54	0.08	0.38	12.30
	ECR (iterative vs. 1-2)	45.36	45.36	45.36	12.80	7.60	12.61	0.54	0.08	0.38	12.30
	Data based	45.36	45.36	45.36	12.80	7.60	12.62	0.54	0.08	0.38	12.30
	Class president	45.36	45.36	45.36	12.80	7.60	12.61	0.54	0.08	0.38	12.30
White	Stephens	45.22	45.22	45.22	12.38	12.38	12.38	0.54	0.09	0.38	11.28
	ECR (iterative vs. 1-2)	45.22	45.22	45.22	12.38	12.38	12.38	0.54	0.09	0.38	11.28
	Data based	45.22	45.22	45.22	12.38	12.38	12.38	0.54	0.09	0.38	11.28
	Class president	45.22	45.22	45.22	12.38	12.38	12.38	0.54	0.09	0.38	11.28
Green	Stephens	45.02	47.64	45.02	12.66	7.90	12.66	0.53	0.08	0.39	4.21
	ECR (iterative vs. 1-2)	45.02	47.64	45.02	12.66	7.90	12.66	0.53	0.08	0.39	4.21
	Data based	45.02	47.64	45.01	12.66	7.90	12.66	0.53	0.08	0.39	4.21
	Class president	45.02	47.62	45.02	12.66	7.90	12.66	0.53	0.08	0.39	4.21

Cell color		Median value of the parameter						
Relabelling method		r1	r2	r3	w1	w2	w3	% time spent
Ocher	Stephens	0.74	-0.56	-0.56	0.54	0.08	0.38	99.98
	ECR (iterative vs. 1-2)	0.74	-0.56	-0.56	0.54	0.08	0.38	99.98
	Data based	0.74	-0.56	-0.56	0.54	0.08	0.38	99.98
	Class president	0.74	-0.56	-0.56	0.54	0.08	0.38	99.98

Relabelling method		Median value of the parameter					
		mu1_1	mu2_1	mu3_1	mu1_2	mu2_2	mu3_2
Stephens		14.55	43.98	82.63	45.32	45.4	45.31
ECR (iterative vs. 1-2)		14.55	43.98	82.62	45.32	45.4	45.31
Data based		14.55	43.98	82.63	45.32	45.4	45.31
Class president		14.55	43.98	82.59	45.32	45.4	45.31
		sigma1_1	sigma2_1	sigma3_1	sigma1_2	sigma2_2	sigma3_2
Stephens		3.57	17.15	16.60	12.68	8.34	12.66
ECR (iterative vs. 1-2)		3.57	17.14	16.61	12.68	8.34	12.66
Data based		3.57	17.14	16.61	12.68	8.35	12.66
Class president		3.57	17.09	16.64	12.68	8.35	12.65
		r1	r2	r3	w1	w2	w3
Stephens		0.74	-0.56	-0.56	0.54	0.08	0.38
ECR (iterative vs. 1-2)		0.74	-0.56	-0.56	0.54	0.08	0.38
Data based		0.74	-0.56	-0.56	0.54	0.08	0.38
Class president		0.74	-0.56	-0.56	0.54	0.08	0.38

Figure 12 Estimated posterior medians of the parameters for each of the five relabelling algorithms implemented in the bivariate application to the prostate cancer dataset. These estimates were calculated on the overall chain and conditional to those cases where the chain spends more simulation time (cumulative % of simulation time spent $\geq 90\%$).

Now, we briefly present the results of the application of *combinatorial mixtures* to the ERG and STOM bivariate distribution.

Figure 12 shows the estimated posterior medians of the parameters for each of the five relabelling algorithms implemented in the bivariate application to the prostate cancer dataset. The ERG gene has a marginal posterior distribution characterized by three means. No matter of the visited color, two of them have posterior medians being equal to 14.55 and ~ 83 . The third mean has a posterior estimate of ~ 43 or ~ 48 , depending if the purple or the grey cell is visited. For $\sim 56\%$ (purple cells) of the simulation time, the highest standard deviation (posterior median of the standard deviation being equal to 16.90) is shared between the components with the two highest posterior medians of the mean, whereas the lowest one is smaller (posterior median of the standard deviation being equal to 3.56). For $\sim 43\%$ (grey cells) of the simulation time, a third standard deviation is fitted (posterior median being equal to ~ 20) and it corresponds to the posterior estimate of the mean of 48; the shared posterior median of the standard deviation around 16.90 is slightly smaller ($\sim 16.10\%$) in this case. The component with posterior medians of the mean and standard deviation being equal to 14.55 and 3.60 have a posterior median of the weight close to 0.55; the component with posterior medians of the mean and standard deviation being around 83 and 16 (or 17) have a posterior median of the weight close to 0.40. The component with the intermediate posterior estimate of the mean has the remaining weight of ~ 0.10 .

The STOM gene has a marginal posterior distribution characterized by one mean, with posterior median around 45.30. The corresponding posterior medians of the standard deviations are either one, two, or three ($\sim 11\%$, $\sim 65\%$, and 12.30% of the simulation time, respectively), depending on the visited cell combination. When there is one standard deviation, its posterior estimate is at 12.38; where there are two of them, the posterior estimates are at 7.85 and 12.73; when there are three of them, their posterior estimates are still similar to the previous ones, being equal to 7.60, 12.60, and 12.80. The component with a posterior median of the standard deviation around 7.85 has a posterior median of the weight slightly smaller than 10%, whereas the weights of the remaining components (with posterior medians of the standard deviations around 12.73) sum up and contribute to the highest weight. For a few iterations ($\sim 4\%$ – green cells), two different means are fitted, with

posterior medians of 45.02 and 47.60, but the corresponding estimates of the standard deviations and weights are still similar to the previous cases. Two components share the same correlation coefficient between the two genes, with a posterior median of -0.60 ; the remaining component has a strong positive correlation, with a posterior median of 0.75 . *A posteriori*, the component with posterior medians of the mean vector around $(14.55, 45.32)$, posterior medians of the standard deviations of $(3.57, 12.68)$ and posterior median of the correlation being equal to 0.74 has a posterior median of the weight of 0.54 . The component with posterior medians of the mean vector around $(82.63, 45.31)$, posterior medians of the standard deviations around $(16.60, 12.65)$ and posterior median of the correlation being equal to -0.56 has a posterior median of the weight of 0.38 . The component with posterior medians of the mean vector around $(43.98, 45.40)$, posterior medians of the standard deviations around $(17.15, 8.35)$ and posterior median of the correlation being equal to -0.56 has a posterior median of the weight of 0.08 . There are limited differences in the posterior medians estimated by the different relabelling methods.

Figure 13 "O'Hagan" matrix for the bivariate normal mixture model fitted to the prostate cancer dataset. Relative frequencies of occurrence of two samples being in the same group are plotted in a black-to-white color scale and sorted according to a non-decreasing ordering of the ERG gene. Dark and light blocks, with proportions similar to the estimated weights of the mixture, identify different groups of observations.

In addition, we present the "O'Hagan" matrix in Figure 13. The plot suggests the existence of two extreme groups of prostate cancer samples. The group on the left has the smallest mean for the ERG gene (posterior medians of the mean vector around $(14.55, 45.32)$ in Figure 12), a positive correlation of 0.74 and the highest weight (posterior medians of the weight equal to 0.54 in Figure 12). The group on the right of the picture has the highest mean for the ERG gene (posterior medians of the mean vector around $(82.63, 45.32)$, the negative correlation of -0.56 and the posterior median of the weight of 0.38 (see Figure 12). The transition between the two groups is managed by fitting the third group with posterior median of the weight of 0.08 . In conclusion, *combinatorial mixtures* allow to identify two groups of samples with similar weights – 0.54 and 0.38 , respectively – located at around $(14.55, 45.32)$ and $(82.63, 45.32)$, with corresponding medians of the standard deviations equal to $(3.57, 12.68)$ and $(16.60, 12.65)$. These groups are characterized by very different posterior correlation coefficients between the two genes. A third group with a negligible weight allows to model the expression data of the samples in between the previous groups.

5.3 Sensitivity to model specification and other issues

Reporting substantive results requires exploring sensitivity to model specification, especially regarding the prior assumptions, and verifying the proper convergence of the MCMC.

Sensitivity of the posterior distribution of the 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 [48] and Stephens [49].

Here, we are concerned on sensitivity of the number of components to different values of the π s. We present the results for the univariate application. To improve prior elicitation and interpretation of the sensitivity analysis, we provide this in terms of the equivalent prior on the number of components, K , introduced in Section 4.4. The default *vague* set-up proposed in the analysis, $\pi^m = (0.2, 0.2, 0.2, 0.2, 0.2)$ and $\pi^v = (0.2, 0.2, 0.2, 0.2, 0.2)$,

is equivalent to assume: $P\{1 \text{ subgroup}\}=0.04$, $P\{2 \text{ subgroups}\}=0.36$ and $P\{3 \text{ subgroups}\}=0.60$ (say, case 3 hereafter). Three groups are *a priori* suggested by our default set-up. For this reason, in the sensitivity analysis we assess the effects on the posterior estimates of other two scenarios, which *a priori* favor either the case of one group or all three possibilities in almost the same way:

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

Appendix Figure 14 summarizes results of the sensitivity analysis to different values of π^m and π^v in the univariate application for the available genes. Each table presents the 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 general conclusions of our analysis with regard to the number of components are that it appears to be robust to the prior specification. For each of the three genes, *a posteriori* the number of groups with the highest probability is the same, no matter of the prior distribution chosen. Deviations of the probabilities from each other are within a sensible range and are compatible with the prior structure imposed.

In addition, we report results from the sensitivity analysis assessing the effect of different choices of a_0 , the hyperparameter of the Dirichlet prior on the weights, in the univariate case. Alternative scenarios are: $a_0 = 0.1$ and $a_0 = 4$ [50, 51], while the baseline is $a_0 = 1$ (see, for example, Diebolt and Robert [22]). Appendix Figure 15 shows the percentage of simulation time spent by the different chains on the most-visited colors (cumulative % of simulation time spent $\geq 90\%$) for each gene and a_0 value. Our analysis still appears to be robust to the prior specification for the weights. Even in the case of overfitting mixtures, like with MSN and ITGB5 genes, the color ranking is essentially unchanged across the different scenarios. The percentage of time spent on each color and the posterior estimates of the parameters reflect the following results [50, 51]:

- $a_{k,0} = 0.1$ favors for overfitting mixtures empty groups,
- $a_{k,0} = \frac{D}{2} + 3 = 4$ favors similar non-empty components.

In detail, as $a_{k,0}$ increases, the MSN and ITGB5 chains spend an increasingly higher percentage of time on the most-visited orange cell (a two-component mixture model with one mean and two different variances). When the other colors are visited and $a_0 = 0.1$, the third component with an extra mean or standard deviation is generally empty; when $a_0 = 4$, there generally exist similar components which share part of the overall weight.

An essential element of the performance of the MCMC is its ability to move between different values of each γ , that is to mix over the number of components. Plots of the changes in the γ s against the number of sweeps were provided in our application. They show that the MCMC mixes well over the γ s. In addition, the percentage of simulation time spent by the chains is similar for those cells indicated with the same color and this reassures against potential convergence issues. We detected no influence of different sets of starting values in either univariate or bivariate analysis.

6 Discussion

We introduce a class of parametric 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 of the vector, it is thus possible that some components share the same value of the parameter, while others do not. Moreover, for different dimensions, subgroups of shared parameters may be different, either because of a different cardinality or of 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.

Bayesian inference and computation approaches were illustrated in a setting based on the normal model and applied to data on molecular subtypes of cancer. Because cancer subtypes may be characterized by differences in location, scale, correlations or any of the combinations, effective procedures for carrying out simultaneously variable selection and clustering (see, for instance, Kim et al. [52]) may benefit from a flexible set of different

criteria to address these issues. *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 example, in nutritional epidemiology, the identification of clusters of subjects at increased cancer risk might benefit from methods where 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 higher correlations between nutrients [53] may be suboptimal [54] and cancer risk/protection associated with the identified clusters may be underestimated due to this issue.

Although it would be possible to achieve a *combinatorial mixture* similar to ours by using DPMs based on separate Dirichlet process priors for the means and variances parameters, our approach is a more direct and, we hope, intuitive extension of popular mixture approaches. In addition, the γ indicators (as presented in Figure 1) can be a tool to guide experts in their elicitation of prior information. In detail, specifying a prior on the different cell colors, instead on the number of components, may help to focus the available prior knowledge on single gamma combinations. Similarly, the visualization of the corresponding fitted mixture models through the same figure, together with the posterior estimates, may facilitate the communication of results [13].

A challenge to the implementation of *combinatorial mixtures* is that the number of K^* -way comparisons between elements of the component-specific parameter vectors increases rapidly with the maximum number of groups. For $K^* = 3$ the Bell number is $B(K^*) = 5$, for $K^* = 4$, it is $B(K^*) = 15$, and, for $K^* = 7$, it is $B(K^*) = 877$. It follows that *combinatorial mixtures* are an effective tool for clustering in potentially complicated scenarios, but when the number of potential groups in the data is relatively small. In addition, the paper currently deals with the bivariate case ($D=2$) only. This is a limitation, as a generalization of *combinatorial mixtures* to higher dimensions would enhance applicability [55].

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. In addition, we suggest to consider the simplest independence prior structure for the means, variances/standard deviations and correlations. It seems to us that for most purposes of the model there is a case for keeping to this simplest independence prior structure and defining weakly informative priors. However, modelling dependence might be relevant in biological applications. 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 previously explored dependence among component-specific parameters, conditional to $(\gamma^m, \gamma^v) \neq (000, 000)$, in a univariate supervised set-up with two known groups. Our prior set-up presented a re-parametrisation of the dependent parameters (say, means or variances) in terms of two independent parameters, a general parameter and a difference (or a ratio for the variances) between the original dependent parameters. We then assumed a mixture of singular distributions for the prior on the independent parameters and derived the corresponding posterior distribution via MCMC. For further details on how to deal with dependence among component-specific parameters and how to apply *combinatorial mixtures* in a supervised context, we refer to Edefonti's Ph.D. Thesis [56], which is available upon request.

Acknowledgement

The results show here are in whole or part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>, last access: October, 30th, 2016.

Funding

Università degli Studi di Milano (Grant / Award Number: 'Young Investigator Grant Program 2014', 'Young Investigator Grant Program 2015'), National Cancer Institute (Grant / Award Number: 'P30CA006516').

A Appendix

A.1 Markov Chain Monte Carlo algorithm: details for the univariate model

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

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

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

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

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

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

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

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

are used hereafter.

From the factorization of the joint distribution of all the variables:

$$\begin{aligned} p(\boldsymbol{\omega}, \boldsymbol{\theta}, z^n, x^n) &= p(\boldsymbol{\omega})p(\boldsymbol{\theta})p(z^n|\boldsymbol{\omega})p(x^n|\boldsymbol{\theta}, z^n) \\ &= p(\boldsymbol{\omega})p(\boldsymbol{\mu}|\gamma^m)p(\gamma^m)p(\boldsymbol{\sigma}^2|\gamma^v)p(\gamma^v)p(z^n|\boldsymbol{\omega})p(x^n|\boldsymbol{\theta}, z^n), \end{aligned}$$

we get:

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

with:

$$p(\boldsymbol{\mu} | -(\boldsymbol{\mu}), x^n, z^n) = \begin{cases} N(\mu_1 | \mu_{123}^*, \eta_{123}^{2*}) & \text{if } \gamma^m = 0 \\ N(\mu_1 | \mu_1^*, \eta_1^{2*})N(\mu_2 | \mu_{23}^*, \eta_{23}^{2*}) & \text{if } \gamma^m = 1 \\ N(\mu_1 | \mu_{13}^*, \eta_{13}^{2*})N(\mu_2 | \mu_2^*, \eta_2^{2*}) & \text{if } \gamma^m = 2 \\ N(\mu_1 | \mu_{12}^*, \eta_{12}^{2*})N(\mu_3 | \mu_3^*, \eta_3^{2*}) & \text{if } \gamma^m = 3 \\ N(\mu_1 | \mu_1^*, \eta_1^{2*})N(\mu_2 | \mu_2^*, \eta_2^{2*})N(\mu_3 | \mu_3^*, \eta_3^{2*}) & \text{if } \gamma^m = 4 \end{cases},$$

where:

$$\mu_{123}^* = \eta_{123}^{2*} \left(\sum_{k=1}^3 \frac{n_k}{\sigma_k^2} \bar{x}_k \right), \quad \eta_{123}^{2*} = \left(\frac{1}{\eta^2} + \sum_{k=1}^3 \frac{n_k}{\sigma_k^2} \right)^{-1},$$

$$\mu_k^* = \eta_k^{2*} \frac{n_k}{\sigma_k^2} \bar{x}_k, \quad \eta_k^{2*} = \left(\frac{1}{\eta^2} + \frac{n_k}{\sigma_k^2} \right)^{-1}, \quad k = 1, 2, 3,$$

$$\begin{aligned} \mu_{k'k''}^* &= \eta_{k'k''}^{2*} \left(\frac{n_{k'}}{\sigma_{k'}^2} \bar{x}_{k'} + \frac{n_{k''}}{\sigma_{k''}^2} \bar{x}_{k''} \right), \quad \eta_{k'k''}^{2*} = \left(\frac{1}{\eta^2} + \frac{n_{k'}}{\sigma_{k'}^2} + \frac{n_{k''}}{\sigma_{k''}^2} \right)^{-1}, \\ k', k'' &= 1, 2, 3, \quad k' \neq k'', \quad k', k'' \neq k, \end{aligned}$$

and:

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

where π^{m*} is such that:

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

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

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

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

where:

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

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

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

with:

$$p(\sigma^2 | - (\sigma^2), x^n, z^n) = \begin{cases} IG(\sigma_1^2 | c_{123}^*, d_{123}^*), & \text{if } \gamma^v = 0 \\ IG(\sigma_1^2 | c_1^*, d_1^*) IG(\sigma_{23}^2 | c_{23}^*, d_{23}^*) & \text{if } \gamma^v = 1 \\ IG(\sigma_1^2 | c_{13}^*, d_{13}^*) IG(\sigma_2^2 | c_2^*, d_2^*) & \text{if } \gamma^v = 2 \\ IG(\sigma_1^2 | c_{12}^*, d_{12}^*) IG(\sigma_3^2 | c_3^*, d_3^*) & \text{if } \gamma^v = 3 \\ IG(\sigma_1^2 | c_1^*, d_1^*) IG(\sigma_2^2 | c_2^*, d_2^*) IG(\sigma_3^2 | c_3^*, d_3^*) & \text{if } \gamma^v = 4 \end{cases},$$

where:

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

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

$$c_{k'k''}^* = c + \frac{n_{k'} + n_{k''}}{2}, \quad d_{k'k''}^* = d + \frac{1}{2} \left(\sum_{i:z_i=k'} (x_i - \mu_{k'})^2 + \sum_{i:z_i=k''} (x_i - \mu_{k''})^2 \right), \quad k', k'' = 1, 2, 3, \quad k' \neq k'', \quad k', k'' \neq k,$$

and:

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

where π^{v*} is such that:

$$\begin{aligned}\pi_0^{v*} &= \frac{\pi_0^v}{F} \frac{\Gamma(c_{123}^*)}{(d_{123}^*)^{c_{123}^*}}, \\ \pi_k^{v*} &= \frac{\pi_k^v}{F} \frac{d^c}{\Gamma(c)} \frac{\Gamma(c_k^*)}{(d_k^*)^{c_k^*}} \frac{\Gamma(c_{k'k''}^*)}{(d_{k'k''}^*)^{c_{k'k''}^*}}, \quad k, k', k'' = 1, 2, 3, \quad k' \neq k'', \quad k', k'' \neq k, \\ \pi_4^{v*} &= \frac{\pi_4^v}{F} \left(\frac{d^c}{\Gamma(c)} \right)^2 \prod_{k=1}^3 \left(\frac{\Gamma(c_k^*)}{(d_k^*)^{c_k^*}} \right), \quad k = 1, 2, 3,\end{aligned}$$

and:

$$\begin{aligned}F &= \pi_0^v \frac{\Gamma(c_{123}^*)}{(d_{123}^*)^{c_{123}^*}} + \pi_1^v \frac{d^c}{\Gamma(c)} \frac{\Gamma(c_1^*)}{(d_1^*)^{c_1^*}} \frac{\Gamma(c_{23}^*)}{(d_{23}^*)^{c_{23}^*}} + \pi_2^v \frac{d^c}{\Gamma(c)} \frac{\Gamma(c_2^*)}{(d_2^*)^{c_2^*}} \frac{\Gamma(c_{13}^*)}{(d_{13}^*)^{c_{13}^*}} + \\ &+ \pi_3^v \frac{d^c}{\Gamma(c)} \frac{\Gamma(c_3^*)}{(d_3^*)^{c_3^*}} \frac{\Gamma(c_{12}^*)}{(d_{12}^*)^{c_{12}^*}} + \pi_4^v \left(\frac{d^c}{\Gamma(c)} \right)^2 \prod_{k=1}^3 \left(\frac{\Gamma(c_k^*)}{(d_k^*)^{c_k^*}} \right).\end{aligned}$$

Finally, for the weights one has:

$$p(\omega | - (\omega), x^n, z^n) = \text{Dir}(a_{1,0} + n_1, a_{2,0} + n_2, a_{3,0} + n_3),$$

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

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

A.2 Markov Chain Monte Carlo algorithm: details for the bivariate model

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

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

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

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

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

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

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

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

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

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

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

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

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

are used hereafter.

From the factorization of the joint distribution of all the variables:

$$\begin{aligned} p(\omega, \theta, z^n, x^n) &= p(\omega)p(\theta)p(z^n|\omega)p(x^n|\theta, z^n) \\ &= p(\omega)p(\mu^1|\gamma_1^m)p(\gamma_1^m)p(\mu^2|\gamma_2^m)p(\gamma_2^m)p(S^1|\gamma_1^v)p(\gamma_1^v)p(S^2|\gamma_2^v)p(\gamma_2^v) \\ &\quad p(r|\gamma^c)p(\gamma^c)p(z^n|\omega)p(x^n|\theta, z^n), \end{aligned}$$

we get:

$$p(\mu^j, \gamma_j^m | - (\mu^j, \gamma_j^m), x^n, z^n) = p(\mu^j | - (\mu^j), x^n, z^n)p(\gamma_j^m | - (\mu^j, \gamma_j^m), x^n, z^n),$$

$$j = 1, 2, \\ \text{with:}$$

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

where:

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

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

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

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

$$\mu_{k'k''j}^* = \eta_{k'k''j}^{2*} (p_{k'j} q_{k''} + p_{k''j} q_{k'})^{-1}, \quad \eta_{k'k''j}^{2*} = \left(\frac{1}{\eta^2} + p_{k'j} + p_{k''j} \right)^{-1},$$

$k', k'' = 1, 2, 3, k' \neq k'', k', k'' \neq k, j = 1, 2,$
and:

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

where π_j^{m*} is such that:

$$\pi_{0j}^{m*} = \frac{L_j \pi_0^m}{M} \eta_{123j}^* \exp \left\{ \frac{1}{2} \left(\frac{\mu_{123j}^*}{\eta_{123j}^*} \right)^2 \right\},$$

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

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

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

where:

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

and:

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

$j = 1, 2.$

For the weights one has:

$$p(\omega | -(\omega), x^n, z^n) = \text{Dir}(a_{1,0} + n_1, a_{2,0} + n_2, a_{3,0} + n_3),$$

and for the generic group label, $z_i, i = 1, \dots, n: p(z_i = k | \omega_k, \{\theta\}, x_i) =$

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

$k = 1, 2, 3$.

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

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

$j = 1, 2$,

which assume five different expressions for each j depending on the value of γ_j^v . The proposal distribution is built accordingly. If (y^j, y_4^j) indicates the proposed candidate for the $(t+1)$ -th iteration, with $y^j = (y_{1j}, y_{2j}, y_{3j})$, one may sample y_4^j from a multinomial distribution with parameters $N = 1, p = (0.2, 0.2, 0.2, 0.2, 0.2)$. Conditioning on y_4^j , one may sample the corresponding vector 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, $S^{j(t)}$, and having a standard deviation $\epsilon = 0.05$ for the lung cancer dataset or $\epsilon = 0.10$ for the prostate cancer dataset. A constraint is added to guarantee that the proposed standard deviations, y^j , are non-negative. This gives the following expression:

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

with:

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

$j = 1, 2$.

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

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

which assumes five different expressions depending on the value of γ^c . The proposal distribution is built as before. If (y, y_4) indicates the proposed candidate for the $(t+1)$ -th iteration, with $y = (y_1, y_2, y_3)$, y_4 is sampled from a multinomial distribution with parameters $N = 1$ and $p = (0.2, 0.2, 0.2, 0.2, 0.2)$. Conditioning on y_4 , y is sampled as $p(y^j | y_4^j)$ before. However, we now assume the following mixture distribution as our building block distribution for y_i :

$$p(y_i | \alpha_0, \beta_0) = \beta_0 * \frac{1}{\alpha_0 - \frac{1}{\alpha_0}} + (1 - \beta_0) * \left(-\frac{1}{\alpha_0 - \frac{1}{\alpha_0}} \right) * r_i^{(t)}. \quad (6)$$

This comes from the choice of updating each correlation as: $y_i = \alpha * \beta * r_i^{(t)}$, where $\alpha \sim U_{(1/\alpha_0, \alpha_0)}$ and

$$\beta \sim \begin{cases} -1 & 1 - \beta_0 \\ 1 & \beta_0 \end{cases},$$

with: $i = 1, 2, 3, \alpha_0 = 1.10$ for the application on the lung cancer dataset or $\alpha_0 = 1.01$ in the prostate cancer dataset, and $\beta_0 = 0.95$, for both sets of data. A constraint is also added to guarantee that the proposed correlations, y , lie in $(-1, 1)$. This gives the following expression:

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

with:

$$p(y|y_4) \propto \begin{cases} p(y_1 | \alpha_0, \beta_0), & \text{if } y_4 = 0 \\ p(y_1 | \alpha_0, \beta_0)p(y_2 | \alpha_0, \beta_0) & \text{if } y_4 = 1 \\ p(y_1 | \alpha_0, \beta_0)p(y_2 | \alpha_0, \beta_0) & \text{if } y_4 = 2 \\ p(y_1 | \alpha_0, \beta_0)p(y_3 | \alpha_0, \beta_0) & \text{if } y_4 = 3 \\ p(y_1 | \alpha_0, \beta_0)p(y_2 | \alpha_0, \beta_0)p(y_3 | \alpha_0, \beta_0) & \text{if } y_4 = 4 \end{cases},$$

where $p(y_i | \alpha_0, \beta_0)$, $i = 1, 2, 3$, was defined as in eq. (6).

A.3 Results of the sensitivity analysis

TRIM29						
	case 1		case 2		case 3	
	prior	posterior	prior	posterior	prior	posterior
1 subgroup	64	0.03	36	0.04	4	0.02
2 subgroups	25	0.29	39	0.57	36	1.16
3 subgroups	11	99.67	25	99.40	60	98.82

MSN						
	case 1		case 2		case 3	
	prior	posterior	prior	posterior	prior	posterior
1 subgroup	64	0.01	36	0.01	4	0.00
2 subgroups	25	46.19	39	43.73	36	33.43
3 subgroups	11	53.80	25	56.27	60	66.57

ITGB5						
	case 1		case 2		case 3	
	prior	posterior	prior	posterior	prior	posterior
1 subgroup	64	0.01	36	0.00	4	0.00
2 subgroups	25	71.07	39	67.49	36	57.19
3 subgroups	11	28.92	25	32.57	60	42.81

Figure 14 Results of the sensitivity analysis to different values of the hyperparameter of the priors on the sharing pattern indicators γ^m and γ^v , π^m and π^v : prior and posterior probabilities (%) of observing one, two or three groups for TRIM29, MSN, and ITGB5 genes in the lung cancer dataset under cases 1, 2, and 3. Case 1 is such that: $P\{1 \text{ subgroup}\}=64\%$, $P\{2 \text{ subgroups}\} \approx 25\%$ and $P\{3 \text{ subgroups}\} \approx 11\%$; case 2 is such that: $P\{1 \text{ subgroup}\}=36\%$, $P\{2 \text{ subgroups}\}=39\%$ and $P\{3 \text{ subgroups}\}=25\%$; case 3 is such that: $P\{1 \text{ subgroup}\}=4\%$, $P\{2 \text{ subgroups}\}=36\%$ and $P\{3 \text{ subgroups}\}=60\%$.

TRIM29						
	a0=0.1		a0=1		a0=4	
	color	% time spent	color	% time spent	color	% time spent
1	grey	49.58	grey	50.63	grey	50.58
2	purple	38.29	purple	38.82	purple	38.75
3	red	5.12	red	5.00	red	4.92

MSN						
	a0=0.1		a0=1		a0=4	
	color	% time spent	color	% time spent	color	% time spent
1	orange	28.90	orange	32.93	orange	49.27
2	light-blue	22.77	purple	18.68	purple	14.43
3	purple	16.40	light-blue	18.59	light-blue	11.96
4	red	13.61	red	12.24	brown	10.01
5	brown	11.38	brown	9.17	red	8.16

ITGB5						
	a0=0.1		a0=1		a0=4	
	color	% time spent	color	% time spent	color	% time spent
1	orange	45.02	orange	55.16	orange	62.79
2	light-blue	24.18	light-blue	18.88	light-blue	16.54
3	brown	14.16	brown	11.08	brown	8.65
4	red	8.51	purple	8.22	purple	6.77

Figure 15 Results of the sensitivity analysis to different values of the hyperparameter of the Dirichlet prior on the weights, a_0 : percentages of simulation time spent on the most-visited colors (cumulative % of simulation time spent $\geq 90\%$) for TRIM29, MSN, and ITGB5 genes in the lung cancer dataset under the following scenarios: 1. $a_0 = 0.1$, which favours for overfitting mixtures empty groups; 2. $a_0 = 1$, which is the standard uniform prior; 3. $a_0 = \frac{D}{2} + 3 = 4$, which favours similar non-empty components.

References

- [1] Newcomb S. A generalized theory of the combination of observations so as to obtain the best result. *Am J Math.* 1886;8:343–366.
- [2] Pearson K. Contribution to the mathematical theory of evolution. *Phil Trans R Soc Lond A* (1887–1895). 1894;185:71–110.
- [3] Titterton DM, Smith AF, Makov UE. Statistical analysis of finite mixture distributions. New York: Wiley, 1985.
- [4] McLachlan GJ, Basford KE. Mixture models: inference and applications to clustering. New York: Marcel Dekker, 1988.
- [5] McLachlan GJ, Peel D. Finite mixture models. New York: John Wiley & Sons, 2000.
- [6] Böhning D, Seidel W. Editorial: recent developments in mixture models. *Comput Stat Data Anal.* 2003;41:349–357.
- [7] Marin J, Mengersen K, Robert CP. Bayesian modelling and inference on mixtures of distributions. In: Dey D, Rao CR, editors. *Handbook of statistics Vol. 25*. Amsterdam: Elsevier-Sciences, 2005:459–507.
- [8] Frühwirth-Schnatter S. Finite mixture and Markov switching models. New York: Springer Science & Business Media, 2006.
- [9] Dettling M, Gabrielson E, Parmigiani G. Searching for differentially expressed gene combinations. *Genome Biol.* 2005;6:R88.
- [10] Shedden K, Taylor J. Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas. In: Shoemaker JS, Lin SM, editors. *Methods of microarray data analysis IV*. New York: Springer, 2004:121–132.
- [11] Hellwig B, Hengstler JG, Schmidt M, Gehrmann MC, Schormann W, Rahnenführer J. Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics.* 2010;11:276.
- [12] Kernagis DN, Hall AH, Datto MB. Genes with bimodal expression are robust diagnostic targets that define distinct subtypes of epithelial ovarian cancer with different overall survival. *J Mol Diagn.* 2012;14:214–222.
- [13] Demicheli R, Retsky MW, Hrushesky WJ, Baum M. Tumor dormancy and surgery-driven interruption of dormancy in breast cancer: learning from failures. *Nat Rev Clin Oncol.* 2007;4:699–710.
- [14] Hartigan JA. Partition models. *Commun Stat Theory Methods.* 1990;19:2745–2756.
- [15] Barry D, Hartigan JA. Product partition models for change point problems. *Ann Stat.* 1992;20:260–279.
- [16] Crowley EM. Product partition models for normal means. *J Am Stat Assoc.* 1997;92:192–198.
- [17] Dahl DB. Modal clustering in a class of product partition models. *Bayesian Anal.* 2009;4:243–264.
- [18] Escobar MD, West M. Bayesian density estimation and inference using mixtures. *J Am Stat Assoc.* 1995;90:577–588.
- [19] West M, Turner DA. Deconvolution of mixtures in analysis of neural synaptic transmission. *Statistician.* 1994;43:31–43.
- [20] Barnard J, McCulloch RE, Meng XL. Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Stat Sin.* 2000;10:1281–1311.

- [21] Gottardo R, Raftery AE. Markov Chain Monte Carlo with mixtures of mutually singular distributions. *J Comput Graph Stat.* 2008;17:949–975.
- [22] Diebolt J, Robert CP. Estimation of finite mixture distributions through Bayesian sampling. *J R Stat Soc Ser B.* 1994;56:363–375.
- [23] Kadane JB. The role of identification in Bayesian theory. In: Fienberg SE, Zellner A, editors. *Studies in Bayesian econometrics and statistics.* North-Holland Publishing: Elsevier, 1975:175–191.
- [24] Stephens M. Dealing with label-switching in mixture models. *J R Stat Soc Ser B.* 2000;62:795–809.
- [25] Frühwirth-Schnatter S. Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J Am Stat Assoc.* 2001;96:194–209.
- [26] O'Hagan A. Contribution to the discussion of Richardson and Green (1997), on Bayesian analysis of mixtures with an unknown number of components. *J R Stat Soc Ser B.* 1997;59:772.
- [27] Jasra A, Holmes CC, Stephens DA. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Stat Sci.* 2005;20:50–67.
- [28] Zhu W, Fan Y. Relabelling algorithms for mixture models with applications for large data sets. *J Stat Comput Simul.* 2016;86:394–413.
- [29] Rodriguez CE, Walker S. Label switching in Bayesian mixture models: deterministic relabeling strategies. *J Comput Graph Stat.* 2014;23:25–45.
- [30] Papastamoulis P, Iliopoulos G. An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *J Comput Graph Stat.* 2010;19:313–331.
- [31] Chung H, Loken E, Schafer J L. Difficulties in drawing inferences with finite-mixture models. *Am Stat.* 2004;58:152–158.
- [32] Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann Stat.* 1974;2:1152–1174.
- [33] Blackwell D, MacQueen JB. Ferguson distributions via Polya urn schemes. *Ann Stat.* 1973;1:353–355.
- [34] Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med.* 2002;8:816–824.
- [35] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Nat Acad Sci USA.* 2001;98:13790–5.
- [36] Miura K, Bowman ED, Simon R, Peng AC, Robles AI, Jones RT, et al. Laser capture microdissection and microarray expression analysis of lung adenocarcinoma reveals tobacco smoking- and prognosis-related molecular profiles. *Cancer Res.* 2002;62:3244–3250.
- [37] Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, et al. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res.* 2002;62:3005–3008.
- [38] Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z. Diversity of gene expression in adenocarcinoma of the lung. *Proc Nat Acad Sci USA.* 2001;98:13784–9 Pacyna- Gengelbach M.
- [39] The Cancer Genome Atlas Research Network et al. The molecular taxonomy of primary prostate cancer. *Cell.* 2015;163:1011–1025.
- [40] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc.* 2002;97:611–631.
- [41] Fraley C, Raftery AE, Murphy TB, Scrucca L. mclust Version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation 2012.
- [42] Robert C, Casella G. *Introducing Monte Carlo Methods with R.* New York: Springer Science & Business Media, 2009.
- [43] R Core Team. *R: a language and environment for statistical computing.* 2015. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- [44] Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat.* 1996;5:299–314.
- [45] Martin AD, Quinn KM, Park JH. MCMCpack: Markov Chain Monte Carlo in R. *J Stat Soft.* 2011;42:22 <http://www.jstatsoft.org/v42/i09/>.
- [46] Papastamoulis P. label.switching: relabelling MCMC outputs of mixture models. R package version 1.3. 2014 <http://CRAN.R-project.org/package=label.switching>.
- [47] Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R News.* 2006;6:7–11.
- [48] Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components, with discussion. *J R Stat Soc Ser B.* 1997;59:731–792.
- [49] Stephens M. Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods. *Ann Stat.* 2000;28:40–74.
- [50] Rousseau J, Mengersen K. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J R Stat Soc Ser B Stat Methodol.* 2011;73:689–710.
- [51] Frühwirth-Schnatter S. Dealing with label switching under model uncertainty. In: Mengersen K, Robert CP, Titterton DM, editors. *Mixtures estimation and application..* New York: John Wiley & Sons, Ltd, 2011:213–239.
- [52] Kim S, Tadesse MC, Vannucci M. Variable selection in clustering via Dirichlet process mixture models. *Biometrika.* 2006;93:877–893.
- [53] Edefonti V, Randi G, Decarli A, La Vecchia C, Bosetti C, Franceschi S, et al. Clustering dietary habits and the risk of breast and ovarian cancers. *Ann Oncol.* 2009;20:581–590.
- [54] Chang WC. On using principal components before separating a mixture of two multivariate normal distributions. *Appl Stat.* 1983;32:267–275.
- [55] McLachlan G, Flack L, Ng S, Wang K. Clustering of gene expression data via normal mixture models. In: Yakovlev YA, Klebanov L, Gaile D, editors. *Statistical Methods for Microarray Data Analysis: Methods and Protocols* 2013. New York: Springer 103–119.
- [56] Edefonti V. Integrating supervised and unsupervised learning in genomics applications. Ph.D. thesis, Bocconi University, Milan, Italy 2006.