A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data [☆]

Arnošt Komárek*

Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

Abstract

An R package mixAK is introduced which implements routines for a semiparametric density estimation through normal mixtures using the Markov chain Monte Carlo (MCMC) methodology. Besides producing the MCMC output, the package computes posterior summary statistics for important characteristics of the fitted distribution or computes and visualizes the posterior predictive density. For the estimated models, penalized expected deviance (PED) and deviance information criterion (DIC) is directly computed which allows for a selection of mixture components. Additionally, multivariate right-, left- and interval-censored observations are allowed. For univariate problems, the reversible jump MCMC algorithm has been implemented and can be used for a joint estimation of the mixture parameters and the number of mixture components. The core MCMC routines have been implemented in C++ and linked to R to ensure a reasonable computational speed. We briefly review implemented algorithms and illustrate the use of the package on three real examples of different complexity.

Key words: Density estimation, Deviance information criterion, Markov chain Monte Carlo, Penalized expected deviance, Reversible jump; Software.

1. Introduction

It has been proven on several places that normal mixtures are a suitable semiparametric structure to model unknown distributions and at the same time are a natural tool for clustering or modelling heterogeneity, see Titterington et al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000) or

[☆] This paper is supplemented by R package mixAK available in *CRAN* at `http://www.R-project.org`. Full code for examples from the paper is available as package vignettes. The code from appendix is electronically available at journal's webpage.

*Corresponding address: Katedra pravděpodobnosti a matematické statistiky, Matematicko-fyzikální fakulta Univerzity Karlovy v Praze, Sokolovská 83, CZ–186 75 Praha 8, Czech Republic. Tel.: (+420)221 913 282, fax: (+420)222 323 316.

Email address: arnost.komarek@mff.cuni.cz (Arnošt Komárek)

Böhning et al. (2007) for a comprehensive discussion of the topic. Starting with a paper by Diebolt and Robert (1994), in which the Gibbs algorithm has been introduced for normal mixtures with pre-specified number of mixture components, their use for Bayesian density estimation became relatively routine. Subsequently, Richardson and Green (1997) showed how the reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithm of Green (1995) can be used for a joint estimation of mixture parameters and a number of mixture components in a univariate case. Since then, several attempts have been made to develop a RJ-MCMC algorithm for multivariate normal mixtures, see, e.g., Dellaportas and Papageorgiou (2006). Alternatively, selection of the number of mixture components can be based on comparison of models fitted with different numbers of components by the means of some joint measure of model complexity and fit. Popular such measures, especially in Bayesian modelling, are the deviance information criterion (DIC, Spiegelhalter et al., 2002) or penalized expected deviance (PED) suggested recently by Plummer (2008).

At this point, mixtures with a pre-specified number of components can be easily handled by WinBUGS (Lunn et al., 2000) however DIC or PED can be only manually and with some programming effort computed for mixtures in the current version. Additionally, reversible jump, birth death or other trans-dimensional MCMC algorithms for mixture problems do not seem to be fully implemented in any of nowadays standard publicly available packages like R (R Development Core Team, 2009) or WinBUGS. See also conclusion section of this paper, where we discuss some WinBUGS suggestions due to Lunn et al. (2005). The purpose of this paper is to present a new R package called `mixAK` which allows for the initial analysis of mixtures either with pre-specified number of components where selection of the number of components is based on DIC or PED, or with the number of components estimated jointly with the remaining model parameters using the RJ-MCMC algorithm. To make the applicability of the package even broader, the package allows to estimate the density if the (part of) data are right-, left-, or generally interval-censored. To ensure a reasonable computational speed, the core parts of the MCMC routines have been implemented in C++ and linked to R.

The rest of the paper is organized as follows. In Section 2, the normal mixture as a model for the unknown density is described together with Bayesian specification of the problem. Section 3 provides an overview of the MCMC methods used in the package `mixAK` to sample from the corresponding posterior distribution. Section 4 describes possibility how to initialize two chains to start MCMC. The posterior inference is discussed in Section 5. A practical analysis using the package is illustrated on three real data examples in Section 6. The paper is finalized by concluding remarks in Section 7 where we also discuss limitations of the package.

2. Model

2.1. Basic model for a density

Let $\boldsymbol{Y}_1,\,\ldots,\,\boldsymbol{Y}_n$ be p-dimensional i.i.d. random vectors with a density $g_y(\boldsymbol{y})$. To incorporate right-, left- and interval-censoring, we will assume that we observe $\lfloor \boldsymbol{l}_1,\,\boldsymbol{u}_1\rfloor,\,\ldots,\,\lfloor \boldsymbol{l}_n,\,\boldsymbol{u}_n\rfloor$, where

$$\lfloor \boldsymbol{l}_i,\,\boldsymbol{u}_i\rfloor = \begin{pmatrix}\lfloor l_{i,1},\,u_{i,1}\rfloor\\ \vdots \\ \lfloor l_{i,p},\,u_{i,p}\rfloor\end{pmatrix} \qquad (i=1,\ldots,n). \tag{1}$$

In the expression (1), $-\infty \leq l_{i,j} < \infty$ are lower limits of observed intervals, $-\infty < u_{i,j} \leq \infty$ are upper limits of observed intervals ($l_{i,j} \leq u_{i,j}$, $i=1,\ldots,n$, $j=1,\ldots,p$) and $\lfloor\ \rfloor$ is open-, closed-, or half-open interval according to the context. Note that $y_{i,j} = l_{i,j} = u_{i,j}$ if the observation is not censored, $-\infty = l_{i,j} < u_{i,j} < \infty$ indicates left-censored observation, $-\infty < l_{i,j} < u_{i,j} = \infty$ indicates right-censored observation and $-\infty < l_{i,j} < u_{i,j} < \infty$ indicates interval-censored observation. Censoring leading to observed intervals is assumed to be non-informative. Density $g_y(\boldsymbol{y})$ is modelled as the following shifted and scaled normal mixture:

$$\begin{aligned} g_y(\boldsymbol{y}) &= |\boldsymbol{S}|^{-1}\sum_{k=1}^{K} w_k\varphi\bigl(\boldsymbol{S}^{-1}(\boldsymbol{y}-\boldsymbol{m})\,\big|\,\boldsymbol{\mu}_k,\,\boldsymbol{\Sigma}_k\bigr)\\ &= \sum_{k=1}^{K} w_k\varphi(\boldsymbol{y}\,|\,\boldsymbol{m}+\boldsymbol{S}\boldsymbol{\mu}_k,\,\boldsymbol{S}\boldsymbol{\Sigma}_k\boldsymbol{S}'), \end{aligned} \tag{2}$$

where $\varphi(\cdot\,|\,\boldsymbol{\mu}_k,\,\boldsymbol{\Sigma}_k)$ is a density of the normal distribution $\mathcal{N}_p(\boldsymbol{\mu}_k,\,\boldsymbol{\Sigma}_k)$, $\boldsymbol{w} = (w_1,\ldots,w_K)'$, $0 \leq w_k \leq 1$, $\sum_{k=1}^{K} w_k = 1$, is a vector of unknown mixture weights, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k : k=1,\ldots,K\}$ are unknown mixture means, and $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_k : k=1,\ldots,K\}$ are unknown mixture variance-covariance matrices. Further, $\boldsymbol{m} = (m_1,\ldots,m_p)'$ is a fixed shift vector and $\boldsymbol{S} = \mathrm{diag}(s_1,\ldots,s_p)$ is a fixed scale matrix. Inclusion of the shift vector $\boldsymbol{m}$ and the scale matrix $\boldsymbol{S}$ in the model is here mainly due to a possibility to improve the mixing and numerical stability of the MCMC algorithm described in section 3, especially needed when different margins are measured in considerably different scales. Finally, let $\boldsymbol{Q}$=$\{\boldsymbol{Q}_1$,$\ldots$,$\boldsymbol{Q}_K\}$ be mixture precision matrices, i.e., $\boldsymbol{Q}_k = \boldsymbol{\Sigma}_k^{-1}$ ($k=1,\ldots,K$). In the remainder of the paper, let $\boldsymbol{y}^* = \boldsymbol{S}^{-1}(\boldsymbol{y}-\boldsymbol{m})$ be shifted and scaled values of $\boldsymbol{y}$ and let

$$g_{y^*}(\boldsymbol{y}^*) = \sum_{k=1}^{K} w_k\varphi(\boldsymbol{y}^*\,|\,\boldsymbol{\mu}_k,\,\boldsymbol{\Sigma}_k) \tag{3}$$

be a density of $\boldsymbol{Y}^* = \boldsymbol{S}^{-1}(\boldsymbol{Y}-\boldsymbol{m})$. Finally, let $\boldsymbol{l}_i^* = \boldsymbol{S}^{-1}(\boldsymbol{l}_i-\boldsymbol{m})$ and $\boldsymbol{u}_i^* = \boldsymbol{S}^{-1}(\boldsymbol{u}_i-\boldsymbol{m})$ ($i=1,\ldots,n$) be shifted and scaled lower and upper limits of observed intervals.

2.2. Bayesian specification

Bayesian technique and MCMC are used to estimate unknown parameters and make a corresponding inference in the package `mixAK`. Hence a prior distribution has to be specified for model parameters. Let $f(\cdot)$ and $f(\cdot\,|\,\cdot)$, respectively, be a generic symbol for a (conditional) probability density and let

$$\boldsymbol{\theta} = (w_1,\,\ldots,\,w_K,\,\boldsymbol{\mu}_1',\ldots,\boldsymbol{\mu}_K',\,\mathrm{vec}(\boldsymbol{Q}_1)',\ldots,\mathrm{vec}(\boldsymbol{Q}_K)',\,\boldsymbol{\gamma}')' \tag{4}$$

be a vector of unknown parameters provided the number of mixture components is known or fixed in advance. In expression (4), $\boldsymbol{\gamma} = (\gamma_1,\,\ldots,\,\gamma_p)'$ is the variance hyperparameter and its meaning will be explained below. Further, let [data] denote a set $\{\boldsymbol{l}_i,\,\boldsymbol{u}_i:\ i = 1,\ldots,n\}$ of observed intervals. Following expression (2), the *observed data* likelihood of the model equals

$$\begin{aligned} L_\theta(\boldsymbol{\theta},\,K) &= f\bigl([\text{data}]\,\big|\,\boldsymbol{\theta},\,K\bigr)\\ &= |\boldsymbol{S}|^{-n}\prod_{i=1}^{n}\Bigl\{\oint_{l_{i,1}}^{u_{i,1}}\cdots\oint_{l_{i,p}}^{u_{i,p}}\sum_{k=1}^{K}w_k\varphi\bigl(\boldsymbol{S}^{-1}(\boldsymbol{y}_i-\boldsymbol{m})\,\big|\,\boldsymbol{\mu_k},\,\boldsymbol{\Sigma}_k\bigr)dy_{i,p}\cdots dy_{i,1}\Bigr\}, \end{aligned} \tag{5}$$

with the convention that $\oint_l^u f(y)dy = \int_l^u f(y)dy$, whenever $l < u$ and $\oint_l^u f(y)dy = f(l) = f(u)$ whenever $l = u$ (uncensored observation).

The following prior specifications are implemented in the package `mixAK` and specified default values are used if not determined by the user. Note that default values attempt to use weakly informative prior distribution.

Mixture weights $\boldsymbol{w}$: a Dirichlet distribution $\mathrm{D}(\delta,\ldots,\delta)$, i.e.,

$$f(\boldsymbol{w}\,|\,K) = \bigl\{\Gamma(\delta)\bigr\}^{-K}\Gamma(K\delta)\prod_{k=1}^{K}w_k^{\delta-1}, \tag{6}$$

where δ is a fixed hyperparameter. A default value is $\delta = 1$.

Mixture means $\boldsymbol{\mu}$ and precision matrices $\boldsymbol{Q}$: it is possible to choose

1. *Semiconjugate independent* Normal and Wishart prior independently for the K components, i.e.,

$$\begin{aligned} f(\boldsymbol{\mu},\,\boldsymbol{Q}\,|\,\gamma,\,K) &= \prod_{k=1}^{K}\bigl\{f(\boldsymbol{\mu}_k)\,f(\boldsymbol{Q}_k\,|\,\gamma)\bigr\}\\ &\propto \prod_{k=1}^{K}\Bigl[|\boldsymbol{D}_k|^{-\frac{1}{2}}\,\exp\bigl\{-\frac{1}{2}(\boldsymbol{\mu}_k-\boldsymbol{\xi}_k)'\boldsymbol{D}_k^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\xi}_k)\bigr\}\\ &\qquad\times\ |\boldsymbol{\Xi}|^{-\frac{\zeta}{2}}|\boldsymbol{Q}_k|^{\frac{\zeta-p-1}{2}}\,\exp\bigl\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Xi}^{-1}\boldsymbol{Q}_k)\bigr\}\Bigr]. \end{aligned} \tag{7}$$

That is, $f(\boldsymbol{\mu}_k)$ is a density of the normal $\mathcal{N}_p(\boldsymbol{\xi}_k,\,\boldsymbol{D}_k)$ and $f(\boldsymbol{Q}_k\,|\,\boldsymbol{\gamma})$ is a density of the Wishart $\mathrm{W}_p(\zeta,\,\boldsymbol{\Xi})$, where $\boldsymbol{\Xi} = \mathrm{diag}(\gamma_1,\ldots,\gamma_p)$. The semiconjugate prior is also the default choice in the package `mixAK`.

2. *Natural-conjugate* Normal-Wishart prior independently for the K components, i.e.,

$$\begin{aligned} f(\boldsymbol{\mu},\, \boldsymbol{Q} \,|\, \gamma,\, K) &= \prod_{k=1}^{K} \bigl\{ f(\boldsymbol{\mu}_k \,|\, \boldsymbol{Q}_k)\; f(\boldsymbol{Q}_k \,|\, \boldsymbol{\gamma}) \bigr\} \\ &\propto \prod_{k=1}^{K} \Bigl[c_k^{\frac{p}{2}} |\boldsymbol{Q}_k|^{\frac{1}{2}}\; \exp\bigl\{ -\frac{c_k}{2} (\boldsymbol{\mu}_k - \boldsymbol{\xi}_k)' \boldsymbol{Q}_k (\boldsymbol{\mu}_k - \boldsymbol{\xi}_k) \bigr\} \\ &\qquad \times\; |\boldsymbol{\Xi}|^{-\frac{\zeta}{2}} |\boldsymbol{Q}_k|^{\frac{\zeta-p-1}{2}}\; \exp\bigl\{ -\frac{1}{2} \mathrm{tr}(\boldsymbol{\Xi}^{-1} \boldsymbol{Q}_k) \bigr\} \Bigr]. \end{aligned} \tag{8}$$

That is, $f(\boldsymbol{\mu}_k \,|\, \boldsymbol{Q}_k)$ is a density of the normal $\mathcal{N}_p(\boldsymbol{\xi}_k,\, c_k^{-1}\, \boldsymbol{Q}_k^{-1})$ and $f(\boldsymbol{Q}_k \,|\, \boldsymbol{\gamma})$ is a density of the Wishart $\mathrm{W}_p(\zeta,\, \boldsymbol{\Xi})$ with $\boldsymbol{\Xi} = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$.

In prior distributions (7) and (8), $\boldsymbol{\xi}_1,\, \ldots,\, \boldsymbol{\xi}_K$ are prior means, $\zeta > p - 1$ are Wishart prior degrees of freedom and $\boldsymbol{\Xi}$ is Wishart prior scale matrix. Further, $\boldsymbol{D}_1,\, \ldots,\, \boldsymbol{D}_K$ in (7) are prior variance-covariance matrices and $c_1 > 0,\, \ldots,\, c_K > 0$ in (8) are prior precision parameters. Let $y^*_{\mathrm{min},j} = \min\{y^*_{i,j} : i = 1, \ldots, n\}$, $y^*_{\mathrm{max},j} = \max\{y^*_{i,j} : i = 1, \ldots, n\}$, $R_j = y^*_{\mathrm{max},j} - y^*_{\mathrm{min},j}$ $(j = 1, \ldots, p)$ be minimum, maximum and range of the shifted- and scaled-data in margin j where in the case of censoring unobserved values of y^* are replaced by their reasonable initial values (e.g., midpoints of observed intervals). Possible choices for prior hyperparameters are then $\xi_{k,j} = 0.5(y^*_{\mathrm{min},j} + y^*_{\mathrm{max},j})$, $\boldsymbol{D}_k = \mathrm{diag}(d_{k,1}, \ldots, d_{k,p})$, $d_{k,j} = R_j^2$ $(k = 1, \ldots, K,\; j = 1, \ldots, p)$ which follows suggestions of Richardson and Green (1997). Weakly informative Wishart prior is obtained, e.g., with $\zeta = p+1$. Default values for precisions $c_1,\, \ldots,\, c_K$ with a natural-conjugate prior are $c_k = 1$ $(k = 1, \ldots, K)$.

Variance hyperparameter $\boldsymbol{\gamma}$ As noted by Richardson and Green (1997) in the univariate case $(p = 1)$, it seems restrictive to suppose that knowledge of the range of the data (R_j) implies much about the size of $\boldsymbol{Q}_k$ and suggest to add an additional hierarchical level and use a gamma hyperprior for $\boldsymbol{\Xi}$. Multivariately, we will assume that $\boldsymbol{\Xi} = \mathrm{diag}(\gamma_1,\, \ldots,\, \gamma_p)$ and a priori $\gamma_j^{-1} \sim \mathrm{Gamma}(g_j,\, h_j)$ independently for $j = 1, \ldots, p$ which is the prior construction used also by Dellaportas and Papageorgiou (2006). That is,

$$f(\gamma_1^{-1}, \ldots, \gamma_p^{-1}) = \prod_{j=1}^{p} \Bigl\{ \frac{h_j^{g_j}}{\Gamma(g_j)}\, (\gamma_j^{-1})^{g_j - 1} \exp(-h_j \gamma_j^{-1}) \Bigr\}. \tag{9}$$

Following Richardson and Green (1997), a weak prior is obtained with g_j being a small positive value and h_j being a small multiple of $1/R_j^2$. Our default values are $g_j = 0.2$, $h_j = 10/R_j^2$ $(j = 1, \ldots, K)$.

Suppose, we want to estimate the number of mixture components K jointly with the remaining mixture parameters. A possible approach is to follow Richardson and Green (1997) where a prior distribution is assumed for K and the inference is based on a sample from the joint posterior distribution of $\boldsymbol{\theta}$ and K

obtained using the RJ-MCMC of Green (1995). In the package `mixAK`, two prior distributions have been implemented for K.

1. *Uniform* prior on $\{1, \ldots, K_{max}\}$, i.e.,
$$f(K) \equiv \mathrm{P}(K = K^*) = \frac{1}{K_{max}} \qquad (K^* = 1, \ldots, K_{max}), \tag{10}$$
where K_{max} is a chosen maximal number of mixture components.
2. *Truncated Poisson* prior, i.e.,
$$f(K) \equiv \mathrm{P}(K = K^*) \propto \frac{\lambda^{K^*} \exp(-\lambda)}{K^*!} \qquad (K = 1, \ldots, K_{max}), \tag{11}$$
where K_{max} is a chosen maximal number of mixture components and λ chosen untruncated Poisson mean.

Default approach in package `mixAK` is with K fixed in advance (i.e., $\mathrm{P}(K = K_{max}) = 1$) in which case the number of mixture components must be chosen according to the value of PED or DIC. Note that at this moment, joint estimation of K and $\boldsymbol{\theta}$ using RJ-MCMC is implemented for univariate data ($p = 1$) only.

The prior distribution of our model is then hierarchically specified as
$$f(\boldsymbol{\theta},\, K) = f(\boldsymbol{\theta}\,|\,K) \,\times\, f(K), \tag{12}$$
where $f(\boldsymbol{\theta}\,|\,K) = f(\boldsymbol{w}\,|\,K) \times f(\boldsymbol{\mu},\, \boldsymbol{Q}\,|\,\boldsymbol{\gamma},\, K) \times f(\boldsymbol{\gamma})$ follows from (6) − (9) and $f(K)$ follows from (10) or (11). Using the Bayes' formula, the posterior distribution combines (5) and (12) into
$$f\bigl(\boldsymbol{\theta},\, K \,\big|\, [\text{data}]\bigr) \propto L_\theta(\boldsymbol{\theta},\, K) \,\times\, f(\boldsymbol{\theta},\, K). \tag{13}$$

2.3. Latent parameters

Computation of the posterior distribution is simplified by introduction of latent (additional) parameters as generally explained by Tanner and Wong (1987). Let $\boldsymbol{\psi}$ be the vector of latent parameters. A joint prior is specified for $(\boldsymbol{\psi}',\, \boldsymbol{\theta}',\, K)'$ hierarchically as
$$f(\boldsymbol{\psi},\, \boldsymbol{\theta},\, K) = f(\boldsymbol{\psi}\,|\,\boldsymbol{\theta},\, K) \times f(\boldsymbol{\theta}\,|\,K) \times f(K), \tag{14}$$
the sample from the posterior distribution $f\bigl(\boldsymbol{\psi},\, \boldsymbol{\theta},\, K \,\big|\, [\text{data}]\bigr)$ is obtained using the (RJ-)MCMC methodology and the inference is based on a sample from the (marginal) posterior distribution $f\bigl(\boldsymbol{\theta},\, K \,\big|\, [\text{data}]\bigr)$ which is directly available as a subset of the complete sample. In this context, the first term in the decomposition (14), $f(\boldsymbol{\psi}\,|\,\boldsymbol{\theta},\, K)$, is sometimes called as the *complete data* likelihood.

In mixture context with censored data, one often considers two sorts of latent parameters: (i) component allocations denoted by $\boldsymbol{r} = (r_1, \ldots, r_n)'$, $r_i \in \{1, \ldots, K\}$, (ii) unobserved values of (shifted and scaled) censored data, i.e., values of $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_n^*$ for those observations which are censored. For convenience in notation we will provide explanation for the situation when all observations are censored and hence $\boldsymbol{\psi} = (\boldsymbol{y}_1^{*\prime}, \ldots, \boldsymbol{y}_n^{*\prime}, r_1, \ldots, r_n)'$. It is easily seen that the complete data likelihood is decomposed as

$$f(\boldsymbol{\psi} \,|\, \boldsymbol{\theta},\, K) = \prod_{i=1}^{n} f(\boldsymbol{y}_i^*,\, r_i \,|\, \boldsymbol{\theta},\, K) = \prod_{i=1}^{n} \bigl\{ f(\boldsymbol{y}_i^* \,|\, r_i,\, \boldsymbol{\theta},\, K) \,\times\, f(r_i \,|\, \boldsymbol{\theta},\, K) \bigr\}, \quad (15)$$

where

$$\begin{aligned} f(\boldsymbol{y}_i^* \,|\, r_i,\, \boldsymbol{\theta},\, K) &= f(\boldsymbol{y}_i^* \,|\, r_i,\, \boldsymbol{\theta}) = \varphi(\boldsymbol{y}_i^* \,|\, \boldsymbol{\mu}_{\boldsymbol{r_i}},\, \boldsymbol{\Sigma}_{r_i}) \\ f(r_i \,|\, \boldsymbol{\theta},\, K) &\equiv \mathrm{P}(r_i = k \,|\, \boldsymbol{\theta},\, K) = w_k \quad (k = 1, \ldots, K). \end{aligned}$$

Together with the likelihood

$$\begin{aligned} L_{\psi,\theta}(\boldsymbol{\psi},\, \boldsymbol{\theta},\, K) &= f\bigl([\text{data}] \,|\, \boldsymbol{\psi},\, \boldsymbol{\theta},\, K\bigr) \\ &= \prod_{i=1}^{n} f(\boldsymbol{l}_i^*,\, \boldsymbol{u}_i^* \,|\, \boldsymbol{y}_i^*) \propto \prod_{i=1}^{n} I\bigl(y_{i,1}^* \in \lfloor l_{i,1}^*,\, u_{i,1}^* \rfloor, \ldots, y_{i,p}^* \in \lfloor l_{i,p}^*,\, u_{i,p}^* \rfloor\bigr), \end{aligned}$$

where I denotes an indicator function, this leads to the joint posterior distribution

$$f\bigl(\boldsymbol{\psi},\, \boldsymbol{\theta},\, K \,\big|\, [\text{data}]\bigr) \propto L_{\psi,\theta}(\boldsymbol{\psi},\, \boldsymbol{\theta},\, K) \,\times\, f(\boldsymbol{\psi},\, \boldsymbol{\theta},\, K) \quad (16)$$

for which the marginalization over $\boldsymbol{\psi}$ leads to the desirable (marginal) posterior (13).

3. Markov chain Monte Carlo

The sample $\{\boldsymbol{\psi}^{(t)},\, \boldsymbol{\theta}^{(t)},\, K^{(t)} \,:\, t = 1, \ldots, T\}$ from the posterior distribution is obtained using a (reversible jump) Markov chain Monte Carlo simulation in which one iteration consists of the following move types depending on whether the number of mixture components K is prespecified or not. With K prespecified, the algorithm iterates between

1. updating the latent (censored) observations, see subsection 3.1;
2. updating the mixture related parameters, see subsection 3.2.

With K random (allowed currently only when $p = 1$), we iterate between

1. updating the latent (censored) observations, see subsection 3.1;
2. update of mixture related parameters as with fixed K as described in subsection 3.2;
3. split-combine move, see subsection 3.3;
4. birth-death move, see subsection 3.3.

Improvement in the mixing of the chains can be achieved if within one MCMC sweep, only one of the steps (2)–(4) is performed, each with a given probability π_a^{mix}, π_b^{mix} or π_c^{mix}, respectively ($\pi_a^{mix} + \pi_b^{mix} + \pi_c^{mix} = 1$). This is also the user's option in the package `mixAK`.

3.1. Update of latent (censored) observations

For each i, when $\boldsymbol{y}_i^*$ contains some censored components, it is updated by sampling from the full conditional distribution which is a (multivariate) normal distribution $\mathcal{N}_p(\boldsymbol{\mu}_{r_i},\,\boldsymbol{\Sigma}_{r_i})$ constrained by the limits of the observed intervals $\boldsymbol{l}_i^*$ and $\boldsymbol{u}_i^*$. In a univariate case, this is done by the inverse cdf sampling, in a multivariate case a method described by Geweke (1991) is used.

3.2. Update of mixture related parameters without a change of K

The moves where the mixture parameters are updated, however when the number of mixture components K remains unaltered, follow largely the proposal of Diebolt and Robert (1994). In fact, all parameters are updated in blocks using a Gibbs kernel by sampling from the full conditional distribution. For details, see the supplement of the paper available as the vignette of the package `mixAK`.

3.3. Moves allowing a change of the number of mixture components

For univariate data ($p = 1$), moves allowing a change of the number of mixture components follow the RJ-MCMC approach taken in Richardson and Green (1997) and the reader is referred therein for details.

4. Initial values to start MCMC

To start the MCMC simulation, initial values are required for the model parameters. In the package `mixAK`, all initial values can either be supplied by the user or generated automatically by the program. Up to two chains can automatically be initialized using two different strategies. Initial values for latent (censored) observations from chain 1 are also used to determine data driven priors described in Sec. 2.2. In a sequel, let $\hat{s}_j^*$ be sample standard deviations of uncensored shifted and scaled observations, lower bounds of shifted and scaled right-censored observations, midpoints of shifted and scaled interval-censored observations, and upper bounds of shifted and scaled left-censored observations in the j-th margin ($j = 1,\dots,p$).

Latent (censored) observations For uncensored observation, both chains start indeed from $y_{i,j}^* = l_{i,j}^* = u_{i,j}^*$. For right-censored observation, chain 1 starts from $y_{i,j}^* = l_{i,j}^* + \hat{s}_j^*$ and chain 2 from $y_{i,j}^* = l_{i,j}^* + |z_{i,j}^*|$, where $z_{i,j}^*$ is sampled from $\mathcal{N}\bigl(0,\,(\hat{s}_j^*)^2\bigr)$. Similarly, for left-censored observation, chain 1 starts from $y_{i,j}^* = u_{i,j}^* - \hat{s}_j^*$ and chain 2 from $y_{i,j}^* = u_{i,j}^* - |z_{i,j}^*|$, where $z_{i,j}^*$ is sampled from $\mathcal{N}\bigl(0,\,(\hat{s}_j^*)^2\bigr)$. For interval-censored observation, chain 1 starts from the midpoint $y_{i,j}^* = 0.5(l_{i,j}^* + u_{i,j}^*)$ and chain 2 from a value chosen uniformly at random from interval $(l_{i,j}^*,\,u_{i,j}^*)$ ($i = 1,\dots,n,\ j = 1,\dots,p$).

Number of mixture components When the number of mixture components is estimated using the RJ-MCMC, the first chain is initialized with $K = 1$ and the second chain with $K = \min(2,\,K_{max})$.

Mixture weights The first chain starts with $w_1 = \cdots = w_K = K^{-1}$, where K is the initial value for the number of mixture components. The initial weights for the second chain are sampled from the prior Dirichlet distribution $\mathrm{D}(\delta, \ldots, \delta)$.

Mixture means For chain 1, initial mixture means in the j-th margin are chosen equidistantly in the interval $(y^*_{\min,j},\, y^*_{\max,j})$ $(j = 1, \ldots, p)$. For chain 2, initial means in the j-th margin are independently sampled from the normal distributions with means equidistantly splitted in the interval $(y^*_{\min,j},\, y^*_{\max,j})$ and standard deviations equal to $\hat{s}^*_j/K$, where K is the initial number of mixture components for the second chain.

Mixture precision matrices The initial mixture precision matrices for the first chain are all the same and are equal to inverted sample variance-covariance matrix computed from initial values of (latent) observations. For chain 2, the initial mixture precision matrices are all diagonal. Diagonal of the initial value of matrix $\boldsymbol{Q}_k$ is equal to inverted sample variances of margins of (latent) observations multiplied by z_k^{-1}, where z_k is sampled from the uniform distribution $\mathcal{U}(0.1,\, 1.1)$.

Component allocations For both chains, observation is allocated to component showing the highest posterior probability given the initial mixture and initial value of latent (censored) observation.

Variance hyperparameter The initial value of $\boldsymbol{\gamma}$ for the first chain is equal to the diagonal of the sample variance-covariance matrix computed from initial values of (latent) observations multiplied by the prior hyperparameter ζ. In the second chain, the same approach is exploited together with multiplication of each initial value of γ_j $(j = 1, \ldots, p)$ by a random variate sampled independently from the uniform distribution $\mathcal{U}(0,\, p)$.

5. Posterior inference

5.1. Label switching problem

It is well known that in mixture problems the posterior distributions (13) or (16) are invariant against the switching of the labelling of mixture components. For example, in Diebolt and Robert (1994), Richardson and Green (1997) artificial identifiability constraints of the type $\mu_1 < \cdots < \mu_K$ are used in a univariate setting to ensure identifiability of the posterior distribution. The problem becomes rather complicated in higher dimensions since the number of identifiability constraints on the parameter space is very large. For a general discussion of this problem, see Jasra et al. (2005). In the MCMC implemented in the package `mixAK`, identification of the posterior distribution can be achieved, e.g., by using relabelling techniques retrospectively, by post-processing the MCMC output (see Stephens, 2000b). However, several important quantities related to the posterior inference, e.g. those of the posterior predictive inference (section

5.2) are invariant to label switching and hence relabelling is redundant in the situations when solely the predictive inference is of interest.

5.2. Predictive density

Very often and especially in situations when the normal mixture is used as a convenient semiparametric structure to model the unknown distribution, the estimate of the density (2) or (3) is of primary interest. A suitable estimator is given by the posterior predictive densities $\mathrm{E}\bigl[g_y(\boldsymbol{y})\bigm|[\text{data}]\bigr]$ or $\mathrm{E}\bigl[g_{y^*}(\boldsymbol{y}^*)\bigm|[\text{data}]\bigr]$ which can easily be approximated from the MCMC output as

$$\hat{g}_y(\boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T} g_y\bigl(\boldsymbol{y}\bigm|\boldsymbol{\theta}^{(t)},\,K^{(t)}\bigr) \approx \mathrm{E}\bigl[g_y(\boldsymbol{y})\bigm|[\text{data}]\bigr], \tag{17}$$

where

$$g_y\bigl(\boldsymbol{y}\bigm|\boldsymbol{\theta}^{(t)},\,K^{(t)}\bigr) = |\boldsymbol{S}|^{-1}\sum_{k=1}^{K^{(t)}} w_k^{(t)}\varphi\bigl(\boldsymbol{S}^{-1}(\boldsymbol{y}-\boldsymbol{m})\bigm|\boldsymbol{\mu}_k^{(t)},\,\boldsymbol{\Sigma}_k^{(t)}\bigr).$$

The expression for $\hat{g}_{y^*}(\boldsymbol{y}^*) \approx \mathrm{E}\bigl[g_{y^*}(\boldsymbol{y}^*)\bigm|[\text{data}]\bigr]$ is analogous. In the package `mixAK`, functions have been implemented to compute the values of all uni- and bivariate marginal densities derived from $\hat{g}_y$ or $\hat{g}_{y^*}$ in a prespecified grid of $\boldsymbol{y}$ or $\boldsymbol{y}^*$ values and visualize them on plots.

5.3. Convergence of the chains

Due to the label switching, we know in advance that the posterior distribution of the K-component model has $K!$ symmetric modes and converging MCMC should visit all of them. This knowledge may be exploited when checking convergence by exploring the chains for mixture weights, means and variances before any possible relabelling, see, e.g. Jasra et al. (2005). Additionally, it is possible to compare posterior distributions of component weights, means or variances which should be identical among components (see subsection 6.2 for illustration).

Another strategy is to base the convergence diagnostic on quantities which are invariant against label switching. For example, moments or quantiles of the mixtures (2) and (3) satisfy this condition. In the package `mixAK`, first two moments, i.e.,

$$\mathrm{E}(\boldsymbol{Y}^*) = \sum_{k=1}^{K} w_k\boldsymbol{\mu}_k, \qquad \mathrm{E}(\boldsymbol{Y}) = \boldsymbol{m} + \boldsymbol{S}\mathrm{E}(\boldsymbol{Y}^*),$$

$$\mathrm{var}(\boldsymbol{Y}^*) = \sum_{k=1}^{K} w_k\Bigl[\boldsymbol{\Sigma}_k + \bigl\{\boldsymbol{\mu}_k - \mathrm{E}(\boldsymbol{Y}^*)\bigr\}\bigl\{\boldsymbol{\mu}_k - \mathrm{E}(\boldsymbol{Y}^*)\bigr\}'\Bigr], \quad \mathrm{var}(\boldsymbol{Y}) = \boldsymbol{S}\mathrm{var}(\boldsymbol{Y}^*)\boldsymbol{S}'.$$

are computed and stored at each iteration of the MCMC and their traceplots and other tools can be used to check the convergence.

Additionally, deviance based quantities are stored at each iteration of the MCMC and can also be used to evaluate the convergence. Namely, the following quantities are computed and stored for $t = 1, \dots, T$:

$$D_{obs}^{(t)} = -2 \sum_{i=1}^{n} \log\Bigl\{ g_y\bigl(\boldsymbol{y}_i^{(t)} \,\big|\, \boldsymbol{\theta}^{(t)},\, K^{(t)}\bigr)\Bigr\}, \tag{18}$$

$$\ell_{compl,0}^{(t)} = \sum_{i=1}^{n} \log\Bigl\{ |\boldsymbol{S}|^{-1} \varphi\bigl(\boldsymbol{S}^{-1}(\boldsymbol{y}_i^{(t)} - \boldsymbol{m}) \,\big|\, \boldsymbol{\mu}_{r_i^{(t)}}^{(t)},\, \boldsymbol{\Sigma}_{r_i^{(t)}}^{(t)}\bigr)\Bigr\}, \tag{19}$$

$$\ell_{compl,1}^{(t)} = \sum_{i=1}^{n} \log(w_{r_i^{(t)}}), \tag{20}$$

$$D_{compl}^{(t)} = -2(\ell_{compl,0}^{(t)} + \ell_{compl,1}^{(t)}). \tag{21}$$

Note that if there are no censored observations, $D_{obs}^{(t)}$ in (18) is twice the observed data logarithmic likelihood (5). In the following, $D_{obs}^{(t)}$ will be called *observed deviance* even in the presence of censoring. Further, $\ell_{compl,0}^{(t)}$ in (19) and $\ell_{compl,1}^{(t)}$ in (20) form together the complete data logarithmic likelihood (15) and hence $D_{compl}^{(t)}$ in (21) is twice the complete data logarithmic likelihood. In the following $D_{compl}^{(t)}$ will be called *complete deviance.*

5.4. Deviance information criterion and penalized expected deviance

A general approach to comparison of complex models based on the samples from the posterior distribution has been suggested by Spiegelhalter et al. (2002) who introduced the deviance information criterion (DIC). In the discussion section of their paper, Richardson showed how the DIC based on the predictive density could be used to discriminate between mixture models with different numbers of components. Other versions of DIC for mixture and in general missing data models have been discussed by Celeux et al. (2006). In package `mixAK`, the Richardson's version of DIC, denoted as DIC_3 in Celeux et al. (2006) has been implemented and can be used to compare the mixture models with different numbers of components, especially in multivariate situations when the reversible jump MCMC allowing for a joint estimation of mixture parameters and the number of mixture components has not been implemented. That is, our DIC is computed as

$$\text{DIC} = \overline{D} + p_D, \qquad p_D = \overline{D} - \tilde{D},$$

where $\overline{D}$ is the approximation to the posterior mean of the deviance, where the posterior expectation is taken with respect to $\boldsymbol{\theta}$, K (if random) and $\boldsymbol{y}$ if there is censoring present. Further, $\tilde{D}$ is the deviance evaluated in the "estimate" to the model parameters and p_D is the effective dimension. The deviance D is based on the normal mixture (2) and a predictive density is taken as the "estimate" to the model parameters. Hence using the MCMC sample

$$\overline{D} = \frac{1}{T}\sum_{t=1}^{T} D_{obs}^{(t)}, \qquad \tilde{D} = -2\,\log\Biggl[\frac{1}{T}\sum_{t=1}^{T}\prod_{i=1}^{n}\Bigl\{ g_y\bigl(\boldsymbol{y}_i^{(t)} \,\big|\, \boldsymbol{\theta}^{(t)},\, K^{(t)}\bigr)\Bigr\}\Biggr].$$

In rejoinder to discussion, Celeux et al. (2006) however conclude that it remains controversial to apply DIC beyond the exponential family case due to lack in theoretical foundation. Possible solution to this problem is offered by Plummer (2008) who suggests to use penalized loss function for Bayesian model comparison and shows that DIC is an approximation to a penalized loss function based on the deviance, with a penalty derived from a cross-validation argument. Particularly in mixture context, Plummer (2008) recommends to use penalized expected deviance (PED) which can be computed using the package `mixAK` as well. In principle, $2n$ separate MCMC runs, with a single observation deleted in each of two runs are required to compute the PED. This computationally demanding task can be avoided by the use of importance sampling where only two parallel chains, $\{\boldsymbol{\theta}^{(1,t)},\, K^{(1,t)} : t = 1,\dots,T\}$, $\{\boldsymbol{\theta}^{(2,t)},\, K^{(2,t)} : t = 1,\dots,T\}$ of genuine parameters of interest are needed. Consequently, PED is computed as

$$\mathrm{PED} = \hat{D}_e + \hat{p}_{opt},$$

where

$$\begin{aligned}\hat{D}_e &= \frac{1}{2T}\sum_{t=1}^{T}\bigl(D_{obs}^{(1,t)} + D_{obs}^{(2,t)}\bigr),\\ D_{obs}^{(c,t)} &= -2\sum_{i=1}^{n}\log\Bigl\{g_y\bigl(\boldsymbol{y}_i^{(c,t)}\,\big|\,\boldsymbol{\theta}^{(c,t)},\, K^{(c,t)}\bigr)\Bigr\} \quad (c = 1,2).\end{aligned}$$

In the case of censoring, $\boldsymbol{y}_i^{(c,t)}$ is sampled from the normal mixture given by $\boldsymbol{\theta}^{(c,t)}$, $K^{(c,t)}$ truncated on the observed intervals $\lfloor \boldsymbol{l}_i,\, \boldsymbol{u}_i \rfloor$. Further, in both censored and uncensored cases, replicated observations $\boldsymbol{y}_i^{(rep1,t)}$ and $\boldsymbol{y}_i^{(rep2,t)}$ are sampled from the (untruncated) normal mixture given by $\boldsymbol{\theta}^{(1,t)}$, $K^{(1,t)}$ and $\boldsymbol{\theta}^{(2,t)}$, $K^{(2,t)}$, respectively and used to calculate the optimism $\hat{p}_{opt}$ of $\hat{D}_e$ as

$$\begin{aligned}\hat{p}_{opt} &= \sum_{i=1}^{n}\hat{p}_{opt_i},\\ \hat{p}_{opt_i} &= \Bigl(\sum_{t=1}^{T} w_i^{(t)}\Bigr)^{-1}\sum_{t=1}^{T} w_i^{(t)}\Biggl[\log\biggl\{\frac{g_y\bigl(\boldsymbol{y}_i^{(rep1,t)}\,\big|\,\boldsymbol{\theta}^{(1,t)},\, K^{(1,t)}\bigr)}{g_y\bigl(\boldsymbol{y}_i^{(rep1,t)}\,\big|\,\boldsymbol{\theta}^{(2,t)},\, K^{(2,t)}\bigr)}\biggr\}\\ &\qquad + \log\biggl\{\frac{g_y\bigl(\boldsymbol{y}_i^{(rep2,t)}\,\big|\,\boldsymbol{\theta}^{(2,t)},\, K^{(2,t)}\bigr)}{g_y\bigl(\boldsymbol{y}_i^{(rep2,t)}\,\big|\,\boldsymbol{\theta}^{(1,t)},\, K^{(1,t)}\bigr)}\biggr\}\Biggr],\end{aligned}$$

where

$$w_i^{(t)} = \Bigl\{g_y\bigl(\boldsymbol{y}_i^{(1,t)}\,\big|\,\boldsymbol{\theta}^{(1,t)},\, K^{(1,t)}\bigr)\cdot g_y\bigl(\boldsymbol{y}_i^{(2,t)}\,\big|\,\boldsymbol{\theta}^{(2,t)},\, K^{(2,t)}\bigr)\Bigr\}^{-1} \qquad (i = 1,\dots,n,\ t = 1,\dots,T)$$

are importance sampling weights.

Another common possibility for Bayesian model comparison is provided by the Bayes factor (Kass and Raftery, 1995) which however has some practical limitations (see, e.g., Plummer, 2008) and it is currently not implemented in the package `mixAK`.

6. Examples

All examples in this section have been run on Intel Core 2 Duo 3 GHz CPU with 3.25 GB RAM. Convergence of the MCMC has been evaluated using the R package `coda` (Plummer et al., 2007). Selected part of the R code for the examples is shown in the appendix. Additional output and more detailed explanation on how to use the package are offered in package vignettes.

6.1. 1 dimension: Galaxy data

The galaxy data which give velocities (in km/sec) of 82 distant galaxies, diverging from our own galaxy were in the context of mixture modelling introduced by Roeder (1990). Richardson and Green (1997) estimated the velocity density using the RJ-MCMC and we repeat their analysis using the package `mixAK`. Hence the following prior distributions, their parameters and parameters of the proposal densities have been used: uniform prior on K with $K_{max} = 30$, $\delta = 1$, semiconjugate prior on $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ with $\boldsymbol{\xi}_k = 21.73$, $\boldsymbol{D}_k = 630.5121$ for all k, $\zeta = 4$, $g_1 = 0.2$, $h_1 = 0.008$, $a_1 = b_1 = 2$, $a_2 = b_2 = 2$, $a_3 = b_3 = 1$. The data were neither shifted nor scaled before running the RJ-MCMC, i.e., $\boldsymbol{m} = 0$, $\boldsymbol{S} = 1$. We report results based on 500 000 iterations of 1:10 thinned RJ-MCMC obtained after a burn-in period of 100 000 iterations which took about 11 min.

During the course of the RJ-MCMC the chain visited models with the number of mixture components K ranging from 1 to 19 with the highest posterior probabilities of 0.11, 0.21, 0.25, 0.19, 0.11, and 0.05 for $K = 4$, 5, 6, 7, 8, and 9, respectively. For the remaining values of K, the posterior probability was lower than 0.04, see Figure 1. The split-combine move has been accepted in 15% of cases whereas the birth-death move in 17% of cases. Consequently, a good mixing of the chain with respect to the number of mixture components has been obtained as is illustrated on the traceplot of K in Figure 1. From the posterior

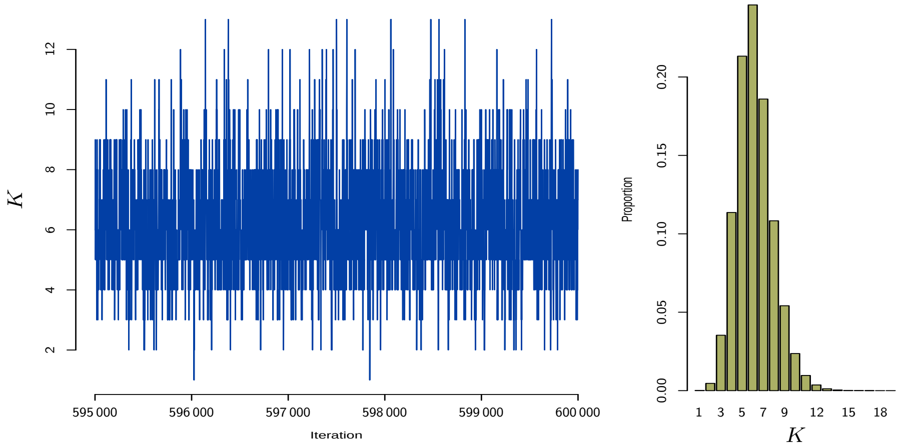

Figure 1: Galaxy data: traceplot (last 5 000 iterations) and histogram for the number of mixture components K.

output, we further show for the purpose of comparison the posterior predictive density (17) and additionally, conditional (given K) posterior predictive densities of the velocity for $K = 4, \ldots, 9$ in Figure 2 which corresponds to Fig. 2 (c) in Richardson and Green (1997).

6.2. 2 dimensions: Old Faithful data

For our second example, we consider the Old Faithful data (version from Härdle, 1991) analyzed using mixtures, e.g., by Stephens (2000b) or Dellaportas and Papageorgiou (2006). In sequence, we fitted one to ten component bivariate mixture to the data using the package `mixAK` with the following prior distributions: $\delta = 1$, semiconjugate prior on $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ with $\boldsymbol{\xi}_k = (-0.1207,\,-0.1028)'$, $\boldsymbol{D}_k = \mathrm{diag}(9.4033,\,15.1983)$, $\zeta = 3$, $g_1 = g_2 = 0.2$, $\boldsymbol{h} = (1.0635,\,0.6580)'$. Due to the fact that the two margins are measured in two quite different scales, we have shifted and scaled observations in both margins by corresponding sample means and standard deviations, i.e., $\boldsymbol{m} = (3.488,\,70.897)'$, $\boldsymbol{S} = \mathrm{diag}(1.141,\,13.595)'$. Reported results are based on 500 000 iterations of 1:10 thinned MCMC obtained after a burn-in period of 100 000 iterations. Sampling of one chain took between 6 min for a model with $K = 1$ up to 54 min for a model with $K = 10$.

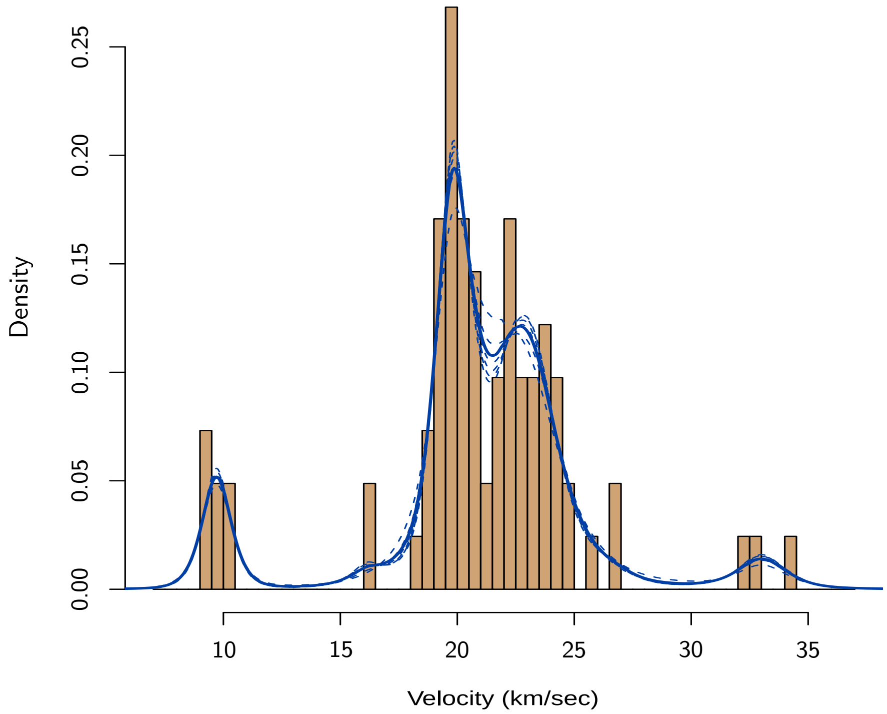

Figure 2: Galaxy data: histogram and predictive densities. Solid line: overall (unconditional) predictive density, dashed lines: conditional predictive densities for $K = 4, \ldots, 9$.

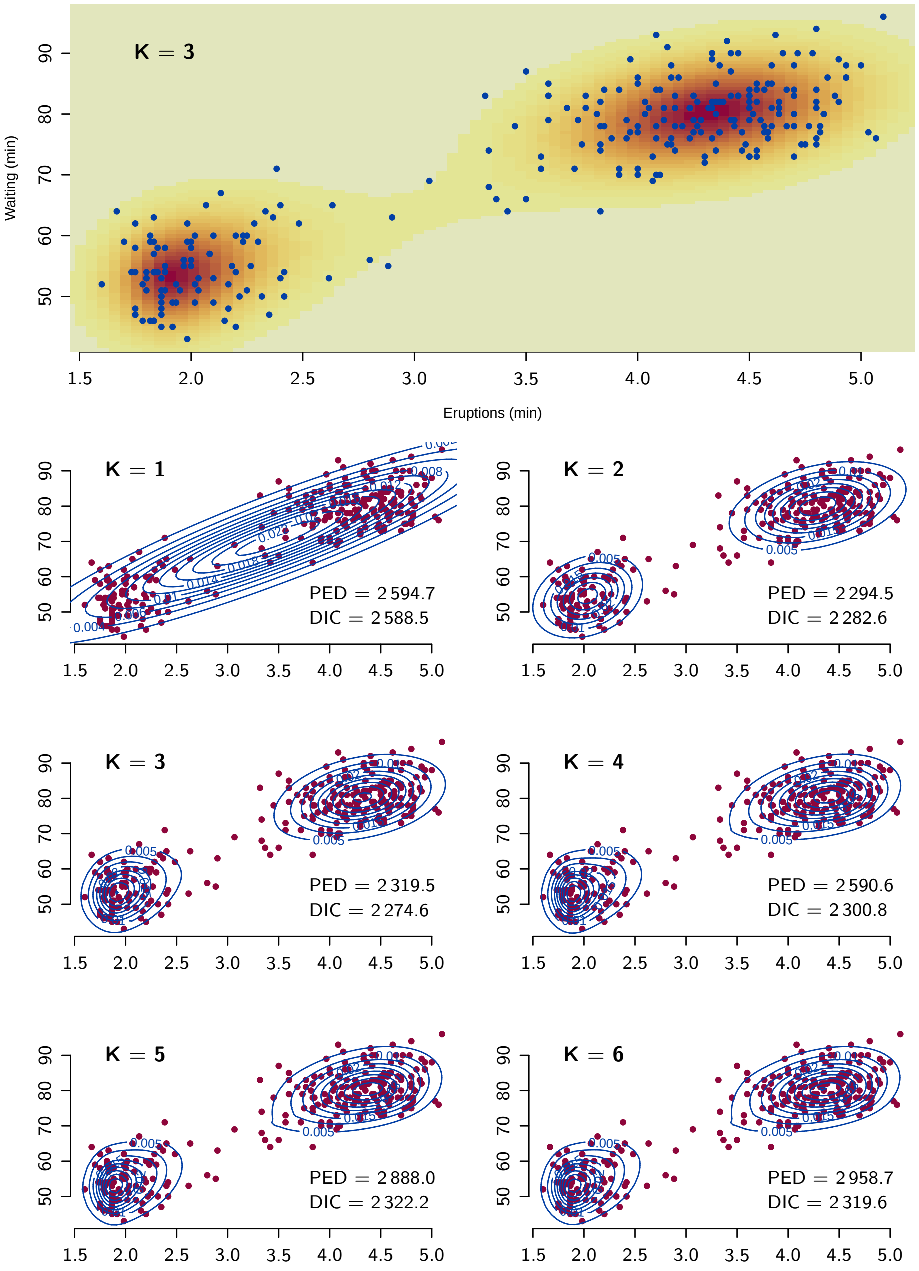

Figure 3: Old Faithful data: scatterplot and predictive densities for different values of K and obtained values of PED and DIC.

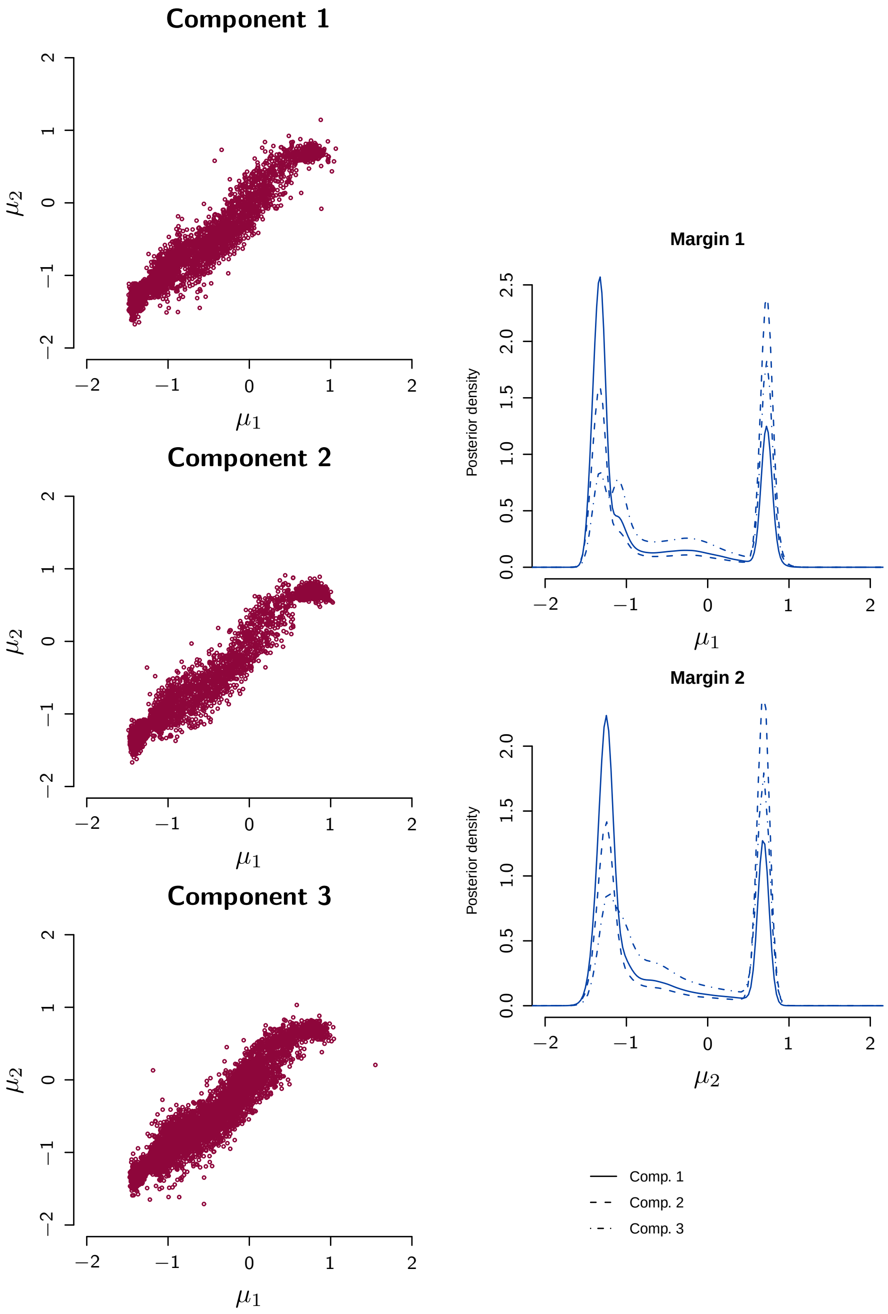

Figure 4: Old Faithful data: scatterplots of sampled mixture means (randomly selected 10 000 iterations) – left panel and estimated posterior densities of marginal mixture means of three components – right panel.

Table 1: Signal Tandmobiel® data. Posterior medians and 95% credible intervals for the mean and standard deviation of the emergence times.

	Proportion censoring			Mean of emergence		Std. dev. of emergence	
				Poster.	95% cred.	Poster.	95% cred.
Tooth	Left	Interval	Right	median	interval	median	interval
11	55.6%	39.4%	5.1%	7.00	(6.97, 7.03)	0.73	(0.71, 0.76)
12	14.7%	72.8%	12.6%	8.11	(8.08, 8.14)	0.93	(0.90, 0.96)
13	0.3%	47.3%	52.4%	11.30	(11.25, 11.36)	1.34	(1.27, 1.42)
14	0.7%	61.7%	37.6%	10.55	(10.50, 10.59)	1.30	(1.26, 1.35)
15	0.4%	42.0%	57.6%	11.52	(11.45, 11.59)	1.53	(1.45, 1.63)
16	86.0%	12.5%	1.5%	6.38	(6.35, 6.40)	0.58	(0.56, 0.60)

Predictive densities based on the models with $K = 1, \ldots, 6$ are shown in Figure 3 together with the values of PED and DIC. It is clear that at least two components are needed. Values of PED are quite similar for $K = 2$ and $K = 3$. The same is true for values of DIC. This coincides with previous results of others, see Fig. 8 (c) in Stephens (2000b) where according to the chosen prior model with $K = 2$ or 3 or 4 reached the highest posterior probability. In the analysis of Dellaportas and Papageorgiou (2006), highest posterior probabilities of 0.3035 and 0.5854 have been obtained for $K = 2$ and $K = 3$, respectively.

Further, Figure 4 shows scatterplots of sampled mixture means in a three-component model and estimated posterior densities of mixture means without imposing any identifiability constraints or relabelling techniques. All three scatterplots are quite similar as well as estimated marginal densities and hence there is no serious indication that the chain would not visit all the modes of the posterior distribution.

6.3. 6 dimensions with interval censoring: Signal Tandmobiel® data

Our third example considers the data from the Signal Tandmobiel® study (Vanobbergen et al., 2000) which was a dental longitudinal study conducted in Flanders in 1996–2001 involving 4 430 children born in 1989. We will analyze the emergence times of first six permanent teeth from the maxillary right quadrant of the mouth (teeth 11, 12, 13, 14, 15, 16 in the European dental notation). In the course of the study children underwent annual dental examinations when emergence (among other things) was recorded. Hence, the emergence times are interval-censored with observed intervals of length of approximately 1 year or left-censored if the tooth was already present at the first examination or right-censored if the tooth has not emerged by the end of the study, see Table 1 for the amount of different types of censoring in the data. However, due to the fact that clinically, the permanent teeth hardly emerge before the age of 5 years (Ekstrand et al., 2003) we have changed all left-censored observations into interval-censored ones with the lower limit of the observed intervals equal to 5 years for the purpose of computation.

Six-dimensional mixture estimated using the package `mixAK` is used as a tool for semiparametric density estimation and consequent inference on some characteristics of the joint emergence distribution of several teeth. In a sequel, models with fixed number of components $K = 1, \ldots, 10$ have been fitted, resulting PED, DIC and other quantities compared. The following prior distributions have been used: $\delta = 1$, semiconjugate prior on $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ with $\boldsymbol{\xi}_k = (8.43, 9.51, 9.77, 9.75, 9.79, 7.68)'$, $\boldsymbol{D}_k = \mathrm{diag}(31.9, 59.5, 67.9, 67.3, 68.6, 18.1)$ for all k, $\zeta = 7$, $g_1 = \cdots = g_6 = 0.2$, $\boldsymbol{h} = (0.31, 0.17, 0.15, 0.15, 0.15, 0.55)'$. The data were neither shifted nor scaled before running the MCMC, i.e., $\boldsymbol{m} = (0, \ldots, 0)'$, $\boldsymbol{S} = \mathrm{diag}(1, \ldots, 1)$. Results based on 20 000 iterations of the 1:10 thinned MCMC obtained after a burn-in period of 10 000 iterations are reported. Sampling of one chain took between 95 min for a model with $K = 1$ up to 135 min for a model with $K = 10$.

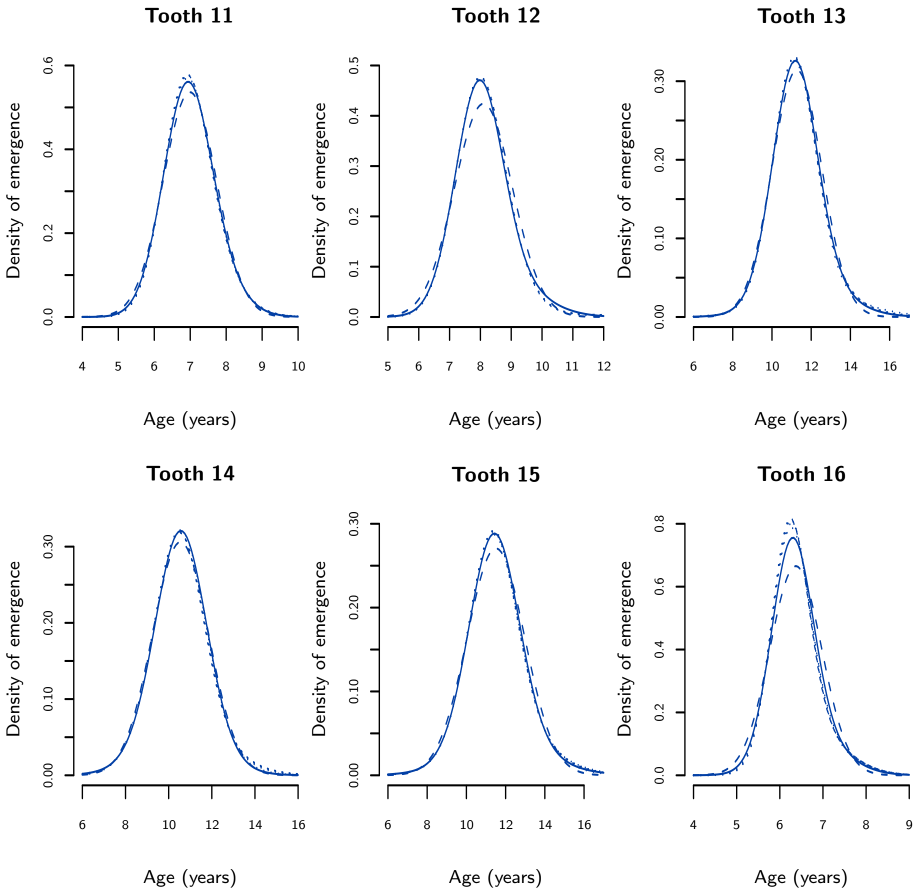

Figure 5: Signal Tandmobiel® data. Marginal predictive densities based on the models with different values of K, solid line for $K = 2$, dashed line for $K = 1$, dotted lines for $K = 3, \ldots, 10$.

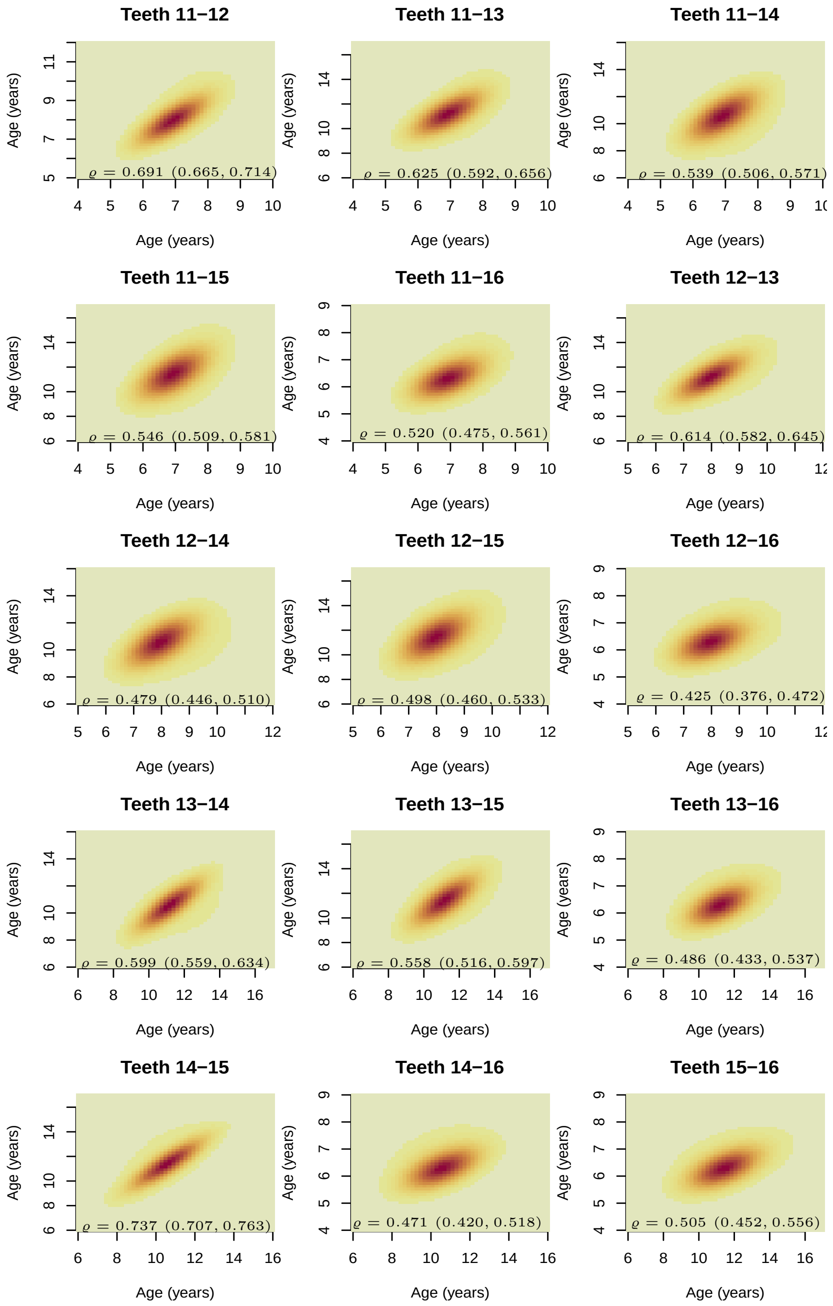

Figure 6: Signal Tandmobiel® data. Bivariate joint predictive densities based on the model with $K = 2$ and posterior median with 95% credible interval for the correlation between the two emergence times.

PED for a model with $K = 1$ reached the value of 78 853, dropped to 75 237 for $K = 2$ and started to increase through 75 858 ($K = 3$), 76 673 ($K = 4$), 77 355 ($K = 5$), 79 529 ($K = 6$), 79 987 ($K = 7$), 81 536 ($K = 8$), 82 283 ($K = 9$), 83 304 ($K = 10$). Graphical examination of the marginal predictive densities, see Figure 5, revealed that marginally the remarkable change (corresponding also to a considerable decrease of PED) in the shape of the estimated densities happens only when we switch from a model with $K = 1$ to a model with two or more mixture components. In the light of above findings, we will use the model with two mixture components as a suitable semiparametric structure to fit the density of the six emergence times.

For clinicians, it is useful to have an information concerning the timing and distribution of the emergence times, as well as the idea on how the emergence times of different teeth relate to each other. Estimated marginal densities of the emergence for different teeth have already been shown in Figure 5. From the MCMC output, it is quite easy to calculate posterior summary statistics for, e.g., the mean and the standard deviation of the emergence time of each tooth, see Table 1 where we report posterior medians and 95% credible intervals. Further, the idea on how the emergence times of different teeth relate to each other can be obtained from the posterior predictive densities of each pair of teeth or from the posterior summary statistics for the Pearson correlation coefficient, all shown in Figure 6. All mentioned quantities can be obtained in a straightforward manner using the functions of the package mixAK.

7. Concluding remarks

In this paper, we have introduced an R package which can be used in a straightforward manner for a density estimation using normal mixtures where at the same time, multivariate (interval-)censored observations can be considered. The package provides not only the core part of the estimation, i.e. MCMC, but also some routines for a consequent processing of the chains. That is, posterior summaries for several mixture related parameters are computed and posterior predictive densities can be easily computed and visualized. For mixture models of an arbitrary dimension, selection of the number of mixture components can be based on the penalized expected deviance or deviance information criterion, directly produced by the package routines.

For univariate problems, the package implements also the reversible jump MCMC algorithm allowing for a joint selection of the number of mixture components. According to our best knowledge, this has not been implemented in any of standardly used software packages. Even though quite recently, Lunn et al. (2005) presented a generic methodology for RJ-MCMC and implemented it in a WinBUGS framework. However, in the mixture context, they consider only birth-death moves. Due to the fact that their methodology aims to be generic for transdimensional models, split-combine moves, specific in a mixture setting and rather crucial for the success of the RJ-MCMC in this context are not considered.

It should be highlighted that it is not the ambition of the package to cover all nowadays available approaches to the (Bayesian) analysis of mixtures. Rather, it is the main intention of the package to allow inexperienced users easily get started with Bayesian mixture analyzes or allow for initial analyzes before switching to often time consuming coding of more advanced methods more suitable for a particular problem. The most important limitations of the package include the following. Firstly, in the selection of prior distributions, the user is limited to these described in subsection 2.2 and hence except the choice between the independent and natural conjugate priors for mixture means and variances, any sensitivity analysis is limited to changes of prior hyperparameters. Possible extension of the list of prior distributions and implementation of slightly different models is only possible by extension of the C++ code. Nevertheless, the list of offered prior distributions cover these mostly appearing in the literature on Bayesian analyzes of mixtures and should suffice for most initial analyzes. Further, it has been shown that usage of latent component allocations and corresponding Gibbs algorithm decelerate convergence of the algorithm by the drastic increase in the dimensionality of the sampling space (see, e.g., Celeux et al., 2000; Cappé et al., 2003; Jasra et al., 2005). Alternatives not requiring the use of latent allocations include, e.g., tempering MCMC (Neal, 1996) used by Celeux et al. (2000) and Jasra et al. (2005) or population and evolutionary MCMC (EMC, Liang and Wong, 2001) used by Jasra et al. (2007). In contrast to the Gibbs algorithm, both tempering MCMC and EMC require selection of several tuning parameters and hence their use might be too complicated to get started. Finally, there are other algorithms available for sampling from distributions of varying dimension than RJ-MCMC implemented in the package mixAK. For example, birth-and-death MCMC (BD-MCMC) introduced by Stephens (2000a) is a valuable competitor to RJ-MCMC. See also Cappé et al. (2003) for the link between RJ-MCMC and BD-MCMC and Sisson (2005) for an overview of available algorithms for transdimensional sampling.

Acknowledgments

I want to thank associate editor and two anonymous referees whose comments led to a considerable improvement of the manuscript. The work on this paper has been supported by grants GAČR 201/09/P077 (*Czech Science Foundation*) and MSM 0021620839 (*Ministry of Education, Youth and Sports of the Czech Republic*).

Data collection for the Signal Tandmobiel® project introduced in subsection 6.3 was supported by Unilever, Belgium. The Signal Tandmobiel® project comprises the following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven), K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

A. R package

This appendix shows briefly how to use the package `mixAK` to obtain the results presented in Section 6. Detailed explanation can be found in package vignettes.

A.1. Galaxy data

- Load the data:

```
> data("Galaxy", package = "mixAK")
```

- Specify prior and parameters for densities of $\boldsymbol{u}$:

```
> GalaxyPrior <- list(priorK = "uniform", Kmax = 30, delta = 1,
+      priormuQ = "independentC", xi = 21.73, D = 630.5121,
+      zeta = 2 * 2, g = 0.2, h = 0.016 / 2)
> parRJMCMC <- list(par.u1 = c(2, 2), par.u2 = c(2, 2), par.u3 = c(1, 1))
```

- Run MCMC:

```
> GalaxyModel <- NMixMCMC(y0 = Galaxy, prior = GalaxyPrior,
+      RJMCMC = parRJMCMC,
+      nMCMC = c(burn = 100000, keep = 500000, thin = 10, info = 10000),
+      scale = list(shift = 0, scale = 1), PED=TRUE)
```

- Basic posterior summary and predictive density (computed from chain 1):

```
> print(GalaxyModel)
> GalaxyPDens <- NMixPredDensMarg(GalaxyModel[[1]])
> plot(GalaxyPDens)
```

A.2. Old Faithful data

- Load the data:

```
> data("Faithful", package = "mixAK")
```

- Specify prior for a model with $K = 3$:

```
> FaithfulPrior <- list(priorK = "fixed", Kmax = 3, delta = 1,
+      priormuQ = "independentC", xi = c(-0.1207, -0.1028),
+      D = diag(c(9.4033, 15.1983)),
+      zeta = 3, g = 0.2, h = c(1.0635, 0.6580))
```

- Run MCMC with shifted and scaled data:

```
> FaithfulModel <- NMixMCMC(y0 = Faithful, prior = FaithfulPrior,
+      nMCMC = c(burn = 100000, keep = 500000, thin = 10, info = 10000),
+      PED = TRUE)
```

- Basic posterior summary (including PED, DIC) and predictive density (computed from chain 1):

```
> print(FaithfulModel)
> FaithfulPDens <- NMixPredDensJoint2(FaithfulModel[[1]])
> plot(FaithfulPDens)
```

- Scatterplot of sampled mixture means in component 2 (chain 1):

```
> j <- 2
> plot(FaithfulModel[[1]]$mu[, (j - 1) * 2 + 1],
+      FaithfulModel[[1]]$mu[, j * 2])
```

A.3. Signal Tandmobiel® data

- Load the data, select only needed columns:

```
> data("TandmobEmer", package = "mixAK")
> y0 <- TandmobEmer[, paste("EBEG.", 10 + 1:6, sep="")]
> y1 <- TandmobEmer[, paste("EEND.", 10 + 1:6, sep="")]
> censor <- TandmobEmer[, paste("CENSOR.", 10 + 1:6, sep="")]
```

- Specify prior for a model with $K = 2$:

```
> TandmobPrior <- list(priorK = "fixed", Kmax = 2, delta = 1,
+      priormuQ = "independentC",
+      xi = c(8.43, 9.60, 9.77, 9.76, 9.80, 7.98),
+      D = diag(c(31.9, 62.5, 67.9, 67.4, 68.9, 18.1)),
+      zeta = 7, g = 0.2, h = c(0.31, 0.16, 0.15, 0.15, 0.15, 0.55))
```

- Run MCMC:

```
> TandmobModel <- NMixMCMC(y0 = y0, y1 = y1, censor = censor,
+      prior = TandmobPrior,
+      nMCMC = c(burn = 10000, keep = 20000, thin = 10, info = 1000),
+      scale = list(shift = 0, scale = 1), PED = TRUE)
```

- Basic posterior summary (including PED, DIC), marginal univariate and pairwise bivariate predictive densities (computed from chain 1):

```
> print(TandmobModel)
> TandmobPDensUni <- NMixPredDensMarg(TandmobModel[[1]])
> plot(TandmobPDensUni)
> TandmobPDensBi <- NMixPredDensJoint2(TandmobModel[[1]])
> plot(TandmobPDensBi)
```

References

Böhning, D., Seidel, W., Alfó, M., Garel, B., Patilea, V., Walther, G., 2007. Editorial: Advances in mixture models. Computational Statistics and Data Analysis 51, 5205–5210.

Cappé, O., Robert, C. P., Rydén, T., 2003. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. Journal of the Royal Statistical Society, Series B 65, 679–700.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., 2006. Deviance information criteria for missing data models (with discussion). Bayesian Analysis 1, 651–706.

Celeux, G., Hurn, M., Robert, C. P., 2000. Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association 95, 957–970.

Dellaportas, P., Papageorgiou, I., 2006. Multivariate mixtures of normals with unknown number of components. Statistics and Computing 16, 57–68.

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

Ekstrand, K. R., Christiansen, J., Christiansen, M. E., 2003. Time and duration of eruption of first and second permanent molars: a longitudinal investigation. Community Dentistry and Oral Epidemiology 31, 344–350.

Geweke, J., 1991. Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities. Computer Sciences and Statistics 23, 571–578.

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

Härdle, W., 1991. Smoothing Techniques with Implementation in S. Springer Verlag, New York.

Jasra, A., Holmes, C. C., Stephens, D. A., 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science 20, 50–67.

Jasra, A., Stephens, D. A., Holmes, C. C., 2007. Population-based reversible jump Markov chain Monte Carlo. Biometrika 94, 787–807.

Kass, R. E., Raftery, A. E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Liang, F., Wong, W. H., 2001. Real parameter evolutionary Monte Carlo with applications to Bayesian mixture models. Journal of the American Statistical Asociation 96, 653–666.

Lunn, D. J., Best, N., Whittaker, J., 2005. Generic reversible jump MCMC using graphical models. Tech. Rep. EPH-2005-01, Department of Epidemiology and Public Health, Imperial College, London.

Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing 10, 325–337.

McLachlan, G. J., Basford, K. E., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, Inc., New York.

McLachlan, G. J., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, New York.

Neal, R., 1996. Sampling from multimodal distributions using tempered transitions. Statistics and Computing 6, 353–366.

Plummer, M., 2008. Penalized loss functions for Bayesian model comparison. Biostatistics 9, 523–539.

Plummer, M., Best, N., Cowles, K., Vines, K., 2007. coda: Output analysis and diagnostics for MCMC. R package version 0.13-1.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org`

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

Roeder, K., 1990. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. Journal of the American Statistical Association 85, 617–624.

Sisson, S., 2005. Transdimensional Markov chains: A decade of progress and future perspectives. Journal of the American Statistical Association 100, 1077–1089.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A., 2002. Bayesian measures of model complexity and fit (with Discussion). Journal of the Royal Statistical Society, Series B 64, 583–639.

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

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

Tanner, M. A., Wong, W. H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–550.

Titterington, D. M., Smith, A. F. M., Makov, U. E., 1985. Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons, Chichester.

Vanobbergen, J., Martens, L., Lesaffre, E., Declerck, D., 2000. The Signal-Tandmobiel® project – a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. European Journal of Paediatric Dentistry 2, 87–96.