Finite and Sparse Finite Mixtures of Multivariate Gaussians

The aim in this practical is to fit Bayesian finite mixture models to two different data sets. Two different scenarios are considered:

- The number of data clusters is known.
- The number of data clusters is unknown and has to be estimated by the procedure.

Depending on these two scenarios the prior on the mixture weights and the specification of K has to be chosen appropriately.

The finite mixture models considered assume that in each component the data follows a multivariate normal distribution. A r-dimensional observation $\boldsymbol{y}_i$ is assumed to follow the mixture distribution given by:

$$p(\boldsymbol{y}_i|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k f_N(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

The following priors are assumed:

$$\begin{aligned}
\boldsymbol{\eta} &\sim \mathcal{D}(e_0, \ldots, e_0), \\
\boldsymbol{\mu}_k &\sim \mathcal{N}_r(\boldsymbol{b}_0, \boldsymbol{B}_0), \quad \forall k = 1, \ldots, K, \\
\boldsymbol{\Sigma}_k^{-1} &\sim \mathcal{W}_r(c_0, \boldsymbol{C}_0), \quad \forall k = 1, \ldots, K, \\
\boldsymbol{C}_0 &\sim \mathcal{W}_r(g_0, \boldsymbol{g}_0).
\end{aligned}$$

The parameters of the priors for the component-specific parameters are specified in the following way. For the prior on the mean values we use

$$\begin{aligned}
\boldsymbol{b}_0 &= (\text{median}(\boldsymbol{y}_{\cdot j}))_{j=1,\ldots,r}, \\
\boldsymbol{B}_0 &= \text{diag}(R_1^2, \ldots, R_r^2),
\end{aligned}$$

where median$(\cdot)$ determines the median of a vector, diag$(\cdot)$ creates a diagonal matrix, and R_j is the range of the observations in the jth variable.

For the prior on the variance-covariance matrices we use

$$\begin{aligned}
c_0 &= 2 + \frac{r-1}{2}, \\
g_0 &= 0.2 + \frac{r-1}{2}, \\
\boldsymbol{G}_0 &= 100 \cdot \frac{g_0}{c_0} \text{diag}(1/R_1^2, \ldots, 1/R_r^2).
\end{aligned}$$

In the exercises different settings for e_0 and K will be investigated.

Different R scripts are provided which implement functions for estimating a finite mixture of multivariate Gaussian distributions. These scripts consist of:

- `estimation-mixture.R`: provides function `sampling()` implementing the Gibbs sampler;
- `identification-mixture.R`: provides function `identifying()` for re-solving the label switching issue by clustering the means in the point process representation;
- `plot-functions.R`: provides functions `Traceplot_draws`, `Traceplot_Nk` and `pointProcessRepresentation` to obtain diagnostic plots for the posterior draws.

Exercise 1:

In this exercise we assume that the number of data clusters is known. We fit a finite mixture with 3 multivariate Gaussian components to the Iris data set by Anderson (1935) and used by Fisher (1936).

The data set can be loaded in R using:

```
> data("iris", package = "datasets")
```

We extract the numeric variables to be used as observations $\boldsymbol{y}_i$ and use the information in variable `Species` as a known classification.

```
> y <- as.matrix(iris[, 1:4])
> z <- as.integer(iris$Species)
```

The script `analysis-iris.R` contains the complete R code to fit a finite mixture with 3 multivariate Gaussian components to `y` using Bayesian estimation together with the necessary post-processing to obtain an identified model and an assess the obtained clustering by comparing it to the known classification.

- Run the code in the script `analysis-iris.R`. At the beginning of the file the additional files are sourced which define the necessary functions and also required add-on packages are loaded. Ensure that you have all necessary package installed. The required packages are all available from CRAN and can be installed using `install.packages()`.
- Go through the code step-by-step and identify in which part (1) the data is loaded, (2) the prior parameters and MCMC settings are specified, (3) Gibbs sampling is performed, (4) the post-processing is done and (5) the obtained clustering is assessed.

  Note that setting `par(ask = TRUE)` ensures that in an interactive session the user is asked before a new plot is created.

Exercise 2:

In this exercise we assume that the number of data clusters is unknown. We fit a sparse finite mixture with multivariate Gaussian components to the Iris data set and estimate the number of clusters based on the posterior of the number of filled components.

- Modify the analysis script to specify $K = 10$ and $e_0 = 0.01$ and run the analysis.
- Investigate the diagnostic plots and assess the posterior of the number of filled components. Compare the diagnostic plots of the mean values in the point process representation before empty components are removed, after a suitable number of filled components is selected and only draws with a suitable number of filled components are retained and finally after label-switching was resolved for these draws.

Exercise 3:

In this exercise we want to fit a finite mixture with 2 multivariate Gaussian components to the Old Faithful data set by Azzalini and Bowman (1990) and Härdle (1991).

The data set can be loaded in R using:

```
> data("faithful", package = "datasets")
```

The data set only contains numeric variables which we want to use as observations $\boldsymbol{y}_i$. No known classification is available for this data set.

```
> y <- as.matrix(faithful)
> z <- NULL
```

- Modify the code in the script `analysis-iris.R` to use the Faithful data set and specify $K = 2$.
- Go through the code step-by-step and assess how applying the method to this data set performs.

Exercise 4:

In this exercise we want to fit a sparse finite mixture with multivariate Gaussian components to the Old Faithful data set and estimate the number of clusters based on the posterior of the number of filled components.

- Modify the analysis script to use the Old Faithful data set and specify $K = 10$ and $e_0 = 0.01$ and run the analysis.
- Investigate the diagnostic plots and assess the posterior of the number of filled components. Compare the diagnostic plots of the mean values in the point process representation before empty components are removed, after a suitable number of filled components is selected and only draws with a suitable number of filled components are retained and finally after label-switching was resolved for these draws.
- Select different values for K and e_0 as well as the burn-in and the recorded draws and investigate how the results change.