# Bayesian statistics, simulation and software

## Module 11: A mixture model

Jesper Møller and Ege Rubak

Department of Mathematical Sciences
Aalborg University

# Mixture model

Conditional on parameters

$$\theta = (\theta_1, \dots, \theta_k), \qquad \lambda = (\lambda_1, \dots, \lambda_k) \quad \text{with } \lambda_1, \dots, \lambda_k \geq 0, \ \sum_j^k \lambda_j = 1,$$

suppose that $Y_1, \dots, Y_n$ are IID random variables with density

$$\pi(y_i|\lambda, \theta) = \lambda_1 \pi_1(y_i|\theta_1) + \lambda_2 \pi_2(y_i|\theta_2) + \dots + \lambda_k \pi_k(y_i|\theta_k), \quad i = 1, \dots, n,$$

where $\pi_j(y_i|\theta_j)$ is a density for a $j$th "component" which is selected with probability $\lambda_j$, $j = 1, \dots, k$. E.g. $\pi_j(y_i|\theta_j) \sim \mathcal{N}(\mu_j, \tau_j)$ and $\theta_j = (\mu_j, \tau_j)$.

We call $\pi(y_i|\lambda, \theta)$ a $k$ component **mixture density** with **mixture weights** $\lambda_1, \dots, \lambda_k$ (they specify a probability distribution).

# Two applications/purposes

$k$ component mixture density with mixture weights $\lambda_1, \ldots, \lambda_k$:

$$\pi(y_i|\lambda, \theta) = \lambda_1 \pi_1(y_i|\theta_1) + \lambda_2 \pi_2(y_i|\theta_2) + \cdots + \lambda_k \pi_k(y_i|\theta_k), \quad i = 1, \ldots, n, IID.$$

1 Cluster analysis: Want to group the $n$ observations into (at most $k$) clusters corresponding to the unknown selection of components.
2 Density estimation: View it as a flexible model for modelling densities (if $k = \infty$ it is often called nonparametric density estimation when considering the posterior distribution of $\pi(\cdot|\lambda, \theta)$ – we let $k < \infty$).

# Mixture model

Often in textbooks one is just given a mixture density (together with some unobserved axillary variables as defined on the next slide) and one uses the 'EM-algorithm for missing data' when finding what one <u>hopes</u> is the MLE of $(\lambda, \theta)$.

Instead we will use a "fully Bayesian approach": Its posterior distribution provides information not only about $(\theta, \lambda)$ (i.e., density estimation) but also about the selection of components (cluster analysis).

# A hierarchical model introducing auxiliary variables

Conditioned on parameters

$$\theta = (\theta_1, \ldots, \theta_k), \qquad \lambda = (\lambda_1, \ldots, \lambda_k) \quad \text{with } \lambda_1, \ldots, \lambda_k \geq 0, \ \sum_j^k \lambda_j = 1,$$

suppose that $Z_1, \ldots, Z_n$ are IID random variables ('auxiliary variables') with

$$P(Z_i = j | \lambda, \theta) = \lambda_j, \qquad j = 1, \ldots, k, \ i = 1, \ldots, n.$$

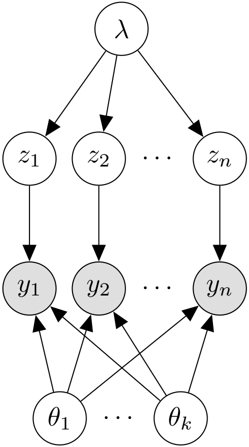
Then conditioned on both $(\theta, \lambda)$ and

$$Z = (Z_1, \ldots, Z_n) = z = (z_1, \ldots, z_n),$$

we can assume that $Y_1, \ldots, Y_n$ are independent and each $Y_i$ has (conditional) density

$$\pi(y_i | \lambda, \theta, z) = \pi_{z_i}(y_i | \theta_{z_i}).$$

# Missing data problem




Notice we have only observed $Y_1 = y_1, \ldots, Y_n = y_n$ (the **data**), i.e., the corresponding realization $Z_1 = z_1, \ldots, Z_n = z_n$ is not observed (the auxiliary variables are '**missing data**').

We refer to $y = (y_1, \ldots, y_n)$ and $z = (z_1, \ldots, z_n)$ as the **full data**.

# "Full likelihood = likelihood for data and missing data"

We have

$$\pi(y_i|\lambda,\theta,z) = \pi_{z_i}(y_i|\theta_{z_i}) = \prod_{j=1}^{k} \pi_j(y_i|\theta_j)^{1[z_i=j]}$$

and

$$P(Z_i = z_i|\lambda,\theta) = P(Z_i = z_i|\lambda) = \lambda_{z_i} = \prod_{j=1}^{k} \lambda_j^{1[z_i=j]},$$

so the joint density for the observations and missing variables (the so-called "**full likelihood**") is

$$\pi(y,z|\lambda,\theta) = \prod_{i=1}^{n} \pi_{z_i}(y_i|\theta_{z_i}) P(Z_i = z_i|\lambda) = \prod_{i=1}^{n}\prod_{j=1}^{k} \left(\pi_j(y_i|\theta_j)\lambda_j\right)^{1[z_i=j]}.$$

# Prior

We (typically) assume a priori that

- $\theta$ and $\lambda$ are independent;
- $\theta_1, \dots, \theta_k$ are independent;
- $\theta_j \sim \pi_j$, $j = 1, \dots, k$ (densities depending on the problem at hand, e.g. $\theta_j = (\mu_j, \tau_j) \sim \mathcal{N}(\mu_{j0}, \tau_{j0}) \times Gamma(\alpha_j, \beta_j)$ or see exercise);
- e.g. $\lambda = (\lambda_1, \dots, \lambda_k)$ could be uniformly distributed on the $(k-1)$-dimensional simplex

$$\Delta_{k-1} = \{(p_1, \dots, p_k) \in [0,1]^k : \sum_{j=1}^{k} p_j = 1\}$$

  (the set of probability distributions on $\{1, 2, \dots, k\}$); this is an example of a so-called Dirichlet$(1, \dots, 1)$-distribution;
- let us assume

$$\lambda \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$$

  (see next slide).

# Dirichlet distribution

**Definition:** Let $k \geq 2$ be an integer. A $k$-dimensional random vector $\lambda = (\lambda_1, \ldots, \lambda_k)$ follows a *Dirichlet distribution* with parameters $\alpha = (\alpha_1, \ldots, \alpha_k) \in (0, \infty)^k$ if $(\lambda_1, \ldots, \lambda_{k-1})$ has density

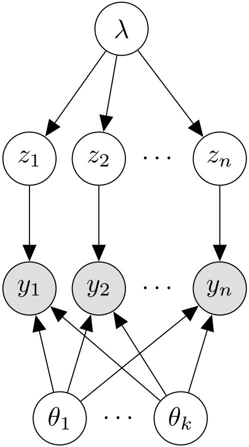
$$\pi(\lambda_1, \ldots, \lambda_{k-1} | \alpha) \propto \prod_{j=1}^{k} \lambda_j^{\alpha_j - 1}$$

where $\lambda_j \in [0, 1]$ for $j = 1, \ldots, k-1$ so that $\lambda_k := 1 - \sum_{j=1}^{k-1} \lambda_j \in [0, 1]$.

- Uniform on $\Delta_{k-1}$ if $\alpha_1 = \ldots = \alpha_k = 1$.
- Dirichlet$(\alpha_1, \alpha_2) = Be(\alpha_1, \alpha_2)$ (the case $k = 2$).
- Simulation is easy: If $X_1 \sim \Gamma(\alpha_1, 1), \ldots, X_k \sim \Gamma(\alpha_k, 1)$ are independent and $S = X_1 + \ldots + X_k$, then

$$\left( \frac{X_1}{S}, \ldots, \frac{X_k}{S} \right) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k).$$

# Graphical representation




- $\lambda = (\lambda_1, \ldots, \lambda_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$.
- Given $\lambda$:

$$P(z_i = j|\lambda) = \lambda_j, \quad j = 1, \ldots, k, \quad i = 1, \ldots, n$$

- Given $\lambda, \theta, z$:

$$y_i \text{ has density } \pi_{z_i}(y_i|\theta_{z_i}),\ i = 1, \ldots, n.$$

- $\theta_j \sim \pi_j$ for $j = 1, \ldots, k$ are independent (a priori assumption).

# Posterior

As $y$ is the data, the unknown variables are the missing data $z$ and the parameter vectors $\lambda$ and $\theta$ – we include all of them into the posterior!

The posterior density is

$$\begin{aligned}\pi(z,\lambda,\theta|y) &\propto \pi(y,z|\lambda,\theta)\pi(\lambda,\theta) = \pi(y,z|\lambda,\theta)\pi(\lambda)\pi(\theta)\\ &\propto \left\{\prod_{i=1}^{n}\prod_{j=1}^{k}\left(\pi_j(y_i|\theta_j)\lambda_j\right)^{1[z_i=j]}\right\}\left\{\prod_{j=1}^{k}\lambda_j^{\alpha_j-1}\right\}\left\{\prod_{j=1}^{k}\pi_j(\theta_j)\right\}.\end{aligned}$$

Looks complicated but we can easily handle all the full conditions – see next slides.

# Full conditional for each $z_i$

For each $i = 1, \ldots, n$, setting $z_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$ we have

$$\pi(z_i | y, \lambda, \theta, z_{-i}) \propto \pi_{z_i}(y_i | \theta_{z_i}) \lambda_{z_i}, \quad z_i \in \{1, \ldots, k\},$$

which is a simple distribution to sample from (just use the R-command 'sample').

# Full conditional for each $\theta_j$

For each $j = 1, \ldots, k$, setting $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$ we have

$$\pi(\theta_j | \theta_{-j}, \lambda, y, z) \propto \pi_j(\theta_j) \prod_{i: z_i = j} \pi_j(y_i | \theta_j).$$

This is equivalent to the posterior density for the case of independent observations from $\pi_j(\cdot|\theta_j)$ (i.e., when considering only the observations selected from the $j$th component).

For example, if the mixture component density $\pi_j(y_j|\theta_j)$ is normal, with $\theta_j$ being the mean and/or the precision parameter(s), and we choose a prior density $\pi_j(\theta_j)$ as in earlier lectures, we know how to sample from this full conditional: it is

- a normal distribution if $\theta_j$ is the mean parameter $\sim \mathcal{N}$-distribution,
- a gamma distribution if $\theta_j$ is the precision parameter $\sim Gamma$-distribution,
- a normal $\times$ gamma distribution if $\theta_j$ is the mean and precision parameters $\sim \mathcal{N} \times Gamma$-distribution.

# Full conditional for $\lambda$

The (joint) full conditional distribution of $\lambda$ is

$$\pi(\lambda|\theta, y, z) \propto \prod_{j=1}^{k} \lambda_j^{n_j(z)+\alpha_j-1} \quad \sim \mathsf{Dirichlet}(n_1(z)+\alpha_1, \dots, n_k(z)+\alpha_k),$$

where $n_j(z)$ is the number of auxiliary variables $z_i$ which are equal to $j$.
So it is easy to simulate from this full conditional.

# Conclusion

It is possible to make a fully Bayesian analysis of a mixture model for IID data $Y_1, \ldots, Y_n$ with unknown mixture weights $\lambda = (\lambda_1, \ldots, \lambda_k)$ and unknown parameters $\theta = (\theta_1, \ldots, \theta_k)$ by considering auxillary variables $Z_1, \ldots, Z_k$ which are included into the posterior together with $(\theta, \lambda)$.

For the posterior simulations we may use a Metropolis within Gibbs sampler, where a sweep consists of updating

$z_i | \cdots, \quad i = 1, \ldots, n,$ (easy: use a Gibbs type update);

$\theta_j | \cdots, \quad j = 1, \ldots, k,$ (easy: Gibbs type update if standard priors are used – else make e.g. a random walk Metroplis type update);

$\lambda | \cdots$ (easy: use a Gibbs type update).

Note that we fixed $k$ – in more advanced work $k$ is also treated as an unknown parameter...

This is now followed by an exercise, considering different cases with known values of $k$.