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

$$\pi(y_i|\lambda, \theta) = \lambda_1 \pi_1(y_i|\theta_1) + \lambda_2 \pi_2(y_i|\theta_2) + \cdots + \lambda_k \pi_k(y_i|\theta_k).$$

That is, $\pi_j(y_i|\theta_j)$ is a density for a jth "component", $j = 1, \dots, k$, and $\pi(y_i|\lambda, \theta)$ is called a k component **mixture density** with **mixture weights** $\lambda_1, \dots, \lambda_k$ (as these weights specify a probability distribution).

Often in textbooks one is just given this mixture density (together with some unobserved auxillary variables as defined on the next slide) and one uses the so-called EM-algorithm when finding what one <u>hopes</u> is the maximum likelihood estimate of (λ, θ).

Here we use instead a "fully Bayesian approach".

A hierarchical model

Conditional 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 with

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

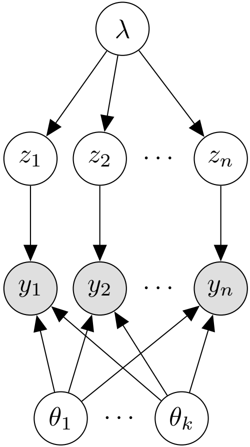
Then conditional on both (θ, λ) and

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

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 "**missing data**").

We

- call $Z_1, \ldots, Z_n$ **auxiliary/dummy variables**,
- 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

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

Prior

We (typically) assume a priori that

- θ and λ are independent;
- $\theta_1, \ldots, \theta_k$ are independent;
- $\theta_j \sim \pi_j$ (a prior density depending on the problem at hand), $j = 1, \ldots, k$;
- e.g. $\lambda = (\lambda_1, \ldots, \lambda_k)$ could be uniformly distributed on the $(k-1)$-dimensional simplex

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

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

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

  (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

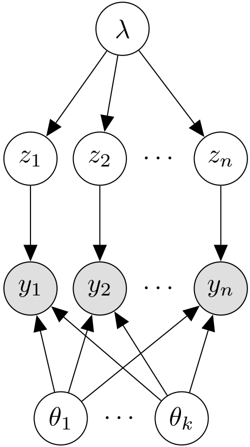
$$\pi(\lambda_1, \ldots, \lambda_{k-1} | \alpha) \propto \prod_{j=1}^{k} \lambda_j^{\alpha_j - 1}$$

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 Δ_{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



- $\lambda = (\lambda_1, \ldots, \lambda_k) \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_k)$, $\lambda_j \geq 0$, $j = 1, \ldots, k$, and $\sum_{j=1}^{k} \lambda_j = 1$.
- Given λ:

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

- Given λ, θ, z: y_i has density $\pi_{z_i}(y_i|\theta_{z_i})$, $i = 1, \ldots, n$.
- $\theta_j \sim \pi_j$, $j = 1, \ldots, k$.

Posterior

As y is the data, the unknown are the missing data z and the parameters λ and θ – we include all of them into the posterior!

The posterior density is

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

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

Full conditional for each z_i

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

$$\pi(z_i | y, \lambda, \theta, z_{-i}) \propto \pi_{z_i}(y_i | \theta_{z_i}) \lambda_{z_i}.$$

Thus

$$\pi(z_i | y, \lambda, \theta, z_{-i}) = \frac{\pi_{z_i}(y_i | \theta_{z_i}) \lambda_{z_i}}{\sum_{j=1}^{k} \pi_j(y_i | \theta_j) \lambda_j}, \quad z_i \in \{1, \dots, k\},$$

which is a simple distribution to sample from.

Full conditional for each θ_j

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

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

This is equivalent to the posterior density for the case of independent observations from $\pi_j(\cdot|\theta_j)$ (i.e. when restricted to observations for the jth component).

For example, if the mixture component density $\pi_j(y_j|\theta_j)$ is normal and we choose a prior density $\pi_j(\theta_j)$ as in earlier lectures, we know how to sample from this full conditional.

Full conditional for λ

The (joint) full conditional distribution of λ is

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

where $n_j(z)$ is the number of dummy variables 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 $Y_1, \dots, Y_n$ with unknown mixture weights $\lambda = (\lambda_1, \dots, \lambda_k)$ and unknown parameters $\theta = (\theta_1, \dots, \theta_k)$ by considering auxillary variables $Z_1, \dots, Z_k$ which are included into the posterior together with (θ, λ).

For the posterior simulations we may use a Metropolis within Gibbs sampler, where we alternate between updating from the full conditionals of

$$
\begin{aligned}
&z_i|\cdots, \quad i = 1, \dots, n, && \text{(this step is very easy – use a Gibbs type update);}\\
&\theta_j|\cdots, \quad j = 1, \dots, k, && \text{(Gibbs or random walk Metroplis or... type update);}\\
&\lambda|\cdots && \text{(this step is very easy – use a Gibbs type update).}
\end{aligned}
$$

This is now followed by an exercise...