Finite mixture model

Examples

- Energy consumption: You have a device to measures the total energy usage. And you want to decompose this signal into a sum of components which you can then try to match to various devices in your house (e.g. computer, refrigerator, washing machine), so that you can figure out which one is wasting the most electricity.
- Biology: You have collected lots of videos of bees. You want to cluster their behaviors into meaningful groups, so that you can look at each cluster and figure out what it means.
- Text mining: You have a collection of scientific papers and you want to understand different research topics for each paper. You fit a model to a dataset of scientific papers which tries to identify different topics in the papers.

Distribution for a categorical random variable

Categorical distirbution: A generalization of Bernoulli distribution. Let $Z \sim Cat(H, p)$ represent a categorical distribution with

- $P(Z = h) = p_h$ for $h = 1, \ldots, H$ and
- $\sum_{h=1}^{H} p_h = 1$.

Example: discrete choice model

Suppose we have a set of H categories and we label these $1, \ldots, H$. Independently, consumers choose one of the H categories with the same probability. Then a reasonable model is $Z_i \overset{ind}{\sim} Cat(H, p)$.

Multinomial distribution

If we count the number of times the consumer chose each category, i.e.

$$Y_h = \Sigma_{i=1}^{n}(Z_i = h),$$

then the result is the multinomial distribution, i.e. $Y \sim Mult(n, p)$. The multinomial distribution has probability mass function

$$p(y; n, p) = \frac{n!}{y_1! \cdots y_H!} \prod_{h=1}^{H} p_h^{y_h}$$

which has

- $E[Y_i] = np_i$,
- $V[Y_i] = np_i(1 - p_i)$, and
- $Cov[Y_i, Y_j] = -np_ip_j$ for $(i \neq j)$.

A special case is $H = 2$ which is the binomial distribution.

Dirichlet distribution

The Dirichlet distribution (named after Peter Gustav Lejeune Dirichlet), i.e. $P \sim Dir(a)$, is a probability distribution for a probability vector of length H. The probability density function for the Dirichlet distribution is

$$p(P; a) = \frac{\Gamma(a_1 + \cdots + a_H)}{\Gamma(a_1) \cdots \Gamma(a_H)} \prod_{h=1}^{H} p_h^{a_h - 1}$$

where $p_h \geq 0$, $\sum_{h=1}^{H} p_h = 1$, and $a_h > 0$.

Letting $a_0 = \sum_{h=1}^{H} a_h$, then some moments are

- $E[p_h] = \frac{a_h}{a_0}$,
- $V[p_h] = \frac{a_h(a_0 - a_h)}{a_0^2(a_0+1)}$,
- $Cov(p_h, p_k) = -\frac{a_h a_k}{a_0^2(a_0+1)}$, and
- $\text{mode}(p_h) = \frac{a_h - 1}{a_0 - H}$ for $a_h > 1$.

A special case is $H = 2$ which is the beta distribution.

Conjugate prior for multinomial distribution

The Dirichlet distribution is the natural conjugate prior for the multinomial distribution. If

$$Y \sim Mult(n, \pi) \qquad \text{and} \qquad \pi \sim Dir(a)$$

then

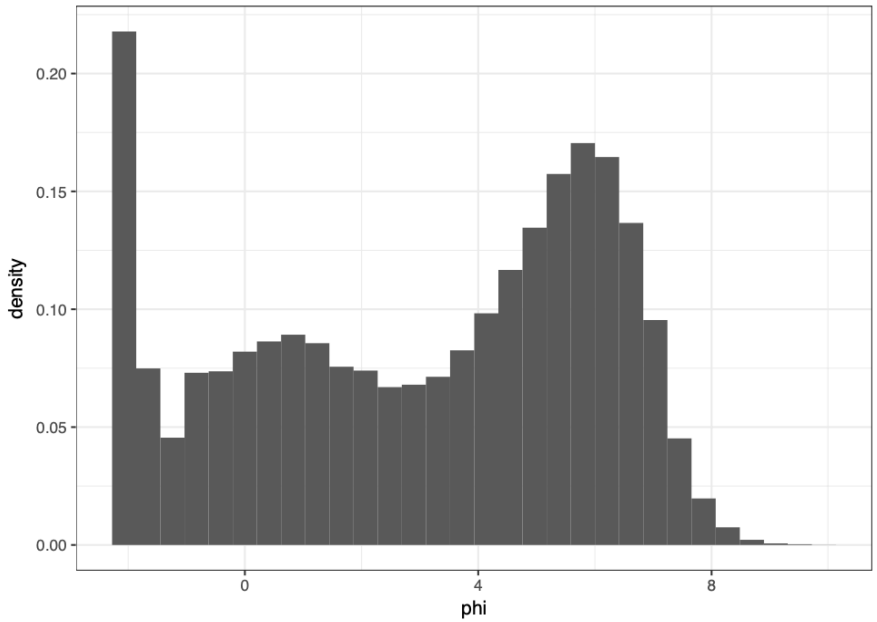
$$\pi | y \sim Dir(a + y).$$

A special case:

- $a = 1$ which is the uniform density over π,

Finite mixtures

Let's focus on modeling the univariate distribution for ϕ



A model for the marginal distribution for $Y_i = \phi_i$ is

$$Y_i \overset{ind}{\sim} \sum_{h=1}^{H} \pi_h N(\mu_h, \sigma_h^2)$$

where $\sum_{h=1}^{H} \pi_h = 1$

Finite mixtures

$$Y_i \overset{ind}{\sim} \sum_{h=1}^{H} \pi_h N(\mu_h, \sigma_h^2)$$

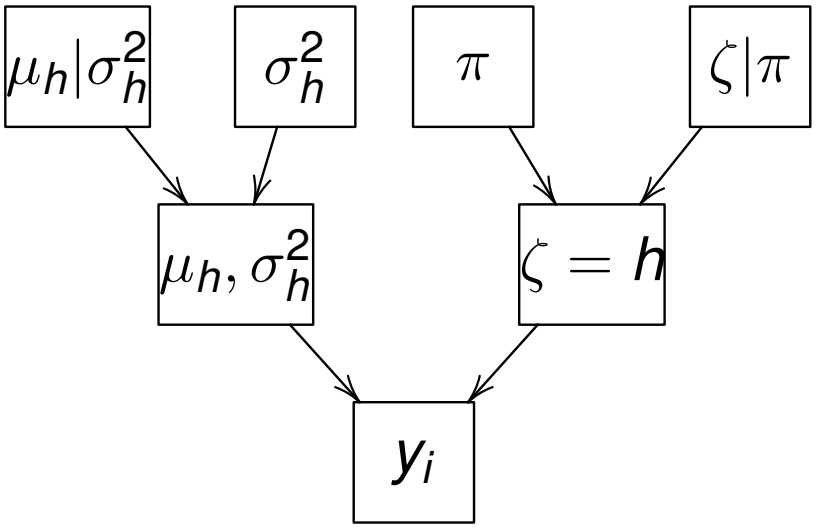
where $\sum_{h=1}^{H} \pi_h = 1$

we can introduce a latent variable $\zeta_i = h$ if observation i came from group h. Then

$$\begin{aligned} Y_i|\zeta_i = z &\overset{ind}{\sim} N(\mu_z, \sigma_z^2) \\ \zeta_i &\overset{ind}{\sim} Cat(H, \pi) \end{aligned}$$

where $\zeta \sim Cat(H, \pi)$ is a categorical random variable with $P(\zeta = h) = \pi_h$ for $h = 1, \ldots, H$ and $\pi = (\pi_1, \ldots, \pi_H)$.

Hierarchical Structure



$$p(\pi, \mu, \sigma^2, \zeta|y) \propto p(y|\zeta, \mu, \sigma^2)p(\mu|\sigma^2)p(\sigma^2)p(\pi)p(\zeta|\pi)$$

Let's assume:

$$\begin{aligned} \pi &\sim Dir(a) \\ \mu_h|\sigma_h^2 &\overset{ind}{\sim} N(m_h, v_h^2\sigma_h^2) \\ \sigma_h^2 &\overset{ind}{\sim} IG(c_h, d_h) \end{aligned}$$

Commonly, we have $m_h = m$, $v_h = v$, $c_h = c$, and $d_h = d$. (On Page 535 there are some suggested values)

Gibbs Sampling for inference

1. For $i = 1, \ldots, n$, sample ζ_i from its full conditional (as the ζ are conditionally independent across i):

$$P(\zeta_i = h | \ldots) \propto \pi_h N(y_i; \mu_h, \sigma_h^2)$$

2. Jointly sample π and μ, σ^2 as they are conditionally independent.
   1. Sample $\pi_1 \ldots \pi_H \sim Dir(a + n_1, \ldots, a + n_H)$
   2. For $h = 1, \ldots, H$, sample μ_h, σ_h^2 from their full conditional (as these are conditionally independent across h):

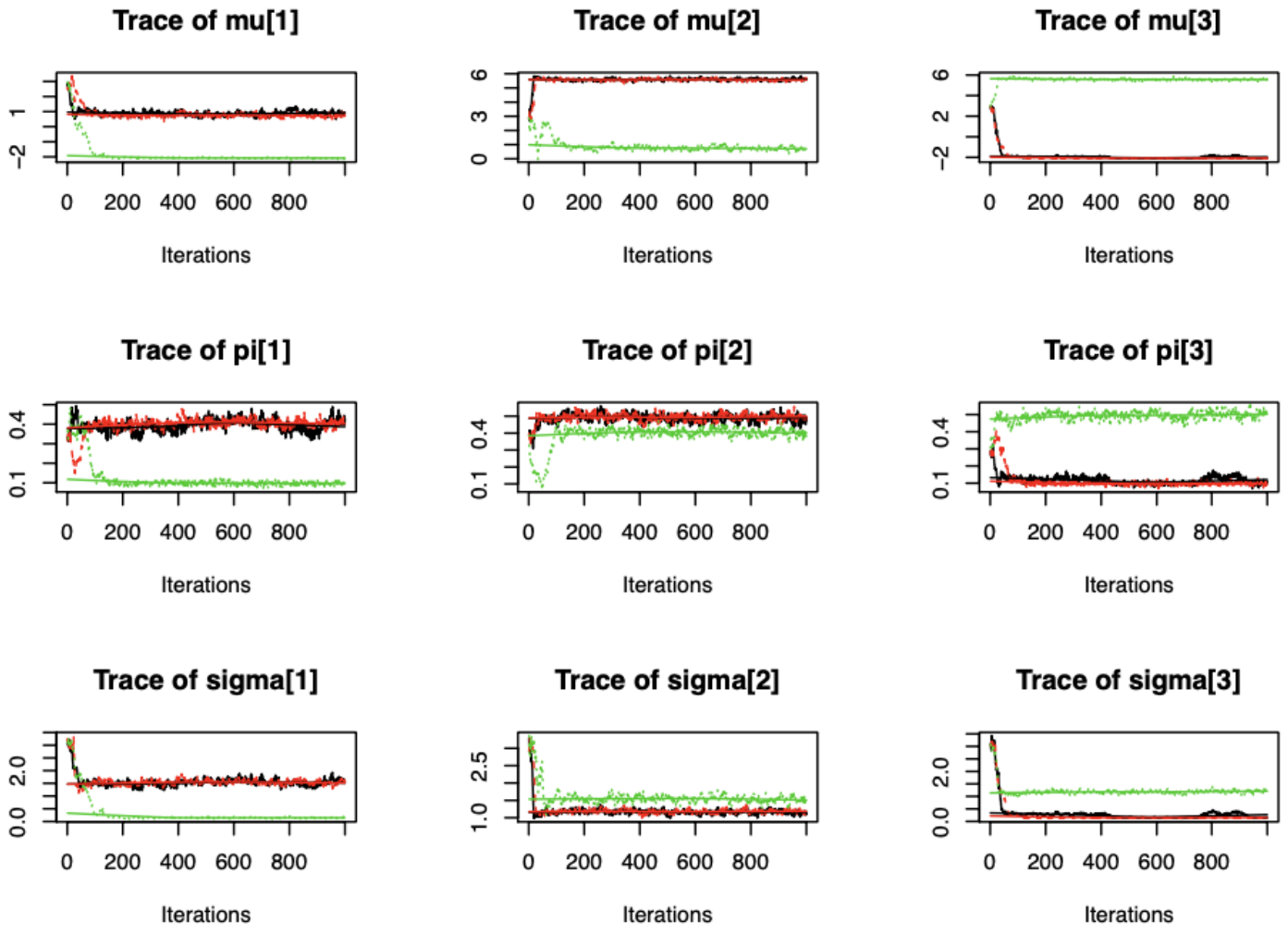
$$\sigma_h^2 \overset{ind}{\sim} IG(c_h', d_h') \quad \mu_h | \sigma_h^2 \overset{ind}{\sim} N(m_h', v_h'^2 \sigma_h^2)$$

where

$$\begin{aligned}
v_h'^2 &= (1/v_h^2 + Z_h)^{-1} \\
m_h' &= v_h'^2 (m_h/v_h^2 + Z_h \bar{y}_h) \\
c_h' &= c_d + Z_h/2 \\
d_h' &= d_h + \tfrac{1}{2}(\textstyle\sum_{i:\zeta_i=h}(y_i - \bar{y}_h)^2 + (\frac{Z_h}{1+Z_h/v_h^2})(\bar{y}_h - m_h)^2) \\
\bar{y}_h &= \tfrac{1}{Z_h} \textstyle\sum_{i:\zeta_i=h} y_i
\end{aligned}$$

Remark: Label Switching

If we let $H = 3$ And we perform Gibbs sampling for multiple draws.



Label switching

The parameters of the model are unidentified due to label-switching, i.e.

$$Y_i \overset{ind}{\sim} \sum_{h=1}^{H} \pi_h N(\mu_h, \sigma_h^2) \overset{d}{=} \sum_{h'=1}^{H} \pi_{h'} N(\mu_{h'}, \sigma_{h'}^2)$$

for some permutation h'.

One way to resolve this issue is to enforce identifiability in the prior. For example, in one-dimension, we can order the component means: $\mu_1 < \mu_2 < \cdots < \mu_H$.

Option:

$$Dir(\pi; a)(\mu_1 < \cdots < \mu_H) \prod_{h=1}^{H} N(\mu_h; m_h, v_h^2 \quad) IG(\sigma_h^2; c_h, d_h)$$

Clustering

Data (y) can then be clustered by assigning them to a group based on their posterior probabilities of group membership, i.e. for data(y) i, we assign the group according to

$$\text{argmax}_h P(\zeta_i = h|y).$$

Unfortunately clustering is extremely sensitive to the parametric model chosen, e.g. normal in this example, and the cluster could change dramatically with a different choice, e.g. t.

Choosing H

When using finite mixture models one of the key choices is to choose H, the number of clusters.

- A Bayesian approach would place a prior on H, e.g. a Poisson or truncated Poisson, and then use MCMC to estimate the posterior
- A more pragmatic approach is to start with a small H and then determine whether there is some feature of the data that is not being adequately addressed, e.g. via cross validation, model checking.
- An empirical Bayes finds an MLE (or MAP) via

$$\hat{H} = \text{argmax}_H p(y|H) = \int p(y|\pi, \mu, \sigma^2, H) p(\pi, \mu, \sigma^2|H) d\pi d\mu d\sigma^2$$

Typically this MLE (or MAP) is found via the EM algorithm.

Another notation: introduction of missing indicator variables

$$P(y_i|\theta,\lambda) = \lambda_1 f(y_i|\theta_1) + \lambda_2 f(y_i|\theta_2) + \cdots + \lambda_H f(y_i|\theta_H)$$

Now we introducing missing indicator variables $Z_i = (z_{i1}, \ldots, z_{iK})$:

$$z_{im} = \begin{cases} 1 & \textit{if } y_i \textit{ is drawn from the } m^{th} \textit{ mixture component} \\ 0 & \textit{otherwise}. \end{cases}$$

Thus we have the following two-layer model:

$$Z_i \sim mult(1, (\lambda_1, \ldots, \lambda_K))$$

Z is $n \times K$ matrix, missing data

$$y_i|Z_i \sim f(y_i|\theta_m), \quad \textit{if } z_{im} = 1 \quad Or \quad p(y_i|Z_i) = \prod_{m=1}^{K} (f(y_i|\theta_m))^{z_{im}}$$

Another notation: introduction of missing indicator variables

Therefore the complete-data likelihood is:

$$P(y, z|\theta, \lambda) = \prod_{i=1}^{n} \prod_{m=1}^{K} (\lambda_m f(y_i|\theta_m))^{z_{im}}$$

Note that the marginal distribution of y_i can be extracted from the complete-data likelihood:

$$P(y_i|\theta, \lambda) = \sum_{z_i} \prod_{m=1}^{K} (\lambda_m f(y_i|\theta_m))^{z_{im}} = \Sigma_{m=1}^{K} \lambda_m f(y_i|\theta_m)$$

MLE by the EM

Log-likelihood of complete data:

$$log(P(y, z|\theta, \lambda)) = \sum_{i=1}^{n} \sum_{m=1}^{K} z_{im}[\log \lambda_m + \log f(y_i|\theta_m)]$$

Taking expectation w.r.t $[z|y, \theta^{(t)}, \lambda^{(t)}]$:

$$E[\log P(y, z|\theta, \lambda)] = \sum_{i=1}^{n} \sum_{m=1}^{K} E(z_{im}|y, \theta^{(t)}, \lambda^{(t)})[\log \lambda_m + \log f(y_i|\theta_m)]$$

Calculate the conditional expectation:

$$E(z_{im}|y, \theta^{(t)}, \lambda^{(t)}) = P(z_{im} = 1|y, \theta^{(t)}, \lambda^{(t)})$$

$$= \frac{P(y_i|z_{im} = 1, \theta_j^{(t)})P(z_{im} = 1|\lambda^{(t)})}{\sum_{j=1}^{K} P(y_i|z_{ij} = 1, \theta_m^{(t)})P(z_{ij} = 1|\lambda^{(t)})} = \frac{\lambda_m^{(t)} f(y_i|\theta_m^{(t)})}{\sum_{j=1}^{K} \lambda_j^{(t)} f(y_i|\theta_j^{(t)})}$$

MLE by the EM

Let $w_{im}^{(t)} := \frac{\lambda_m^{(t)} f(y_i|\theta_m^{(t)})}{\sum_{j=1}^{K} \lambda_j^{(t)} f(y_i|\theta_j^{(t)})}$ weight of y_i from $f(.|\theta_m)$

- E-step: Calculate the weights $\mathbf{w}_{im}^{(t)}$ for $m = 1, \ldots, K$ and $i = 1, \ldots, n$ Then

$$Q(\theta, \lambda|\theta^{(t)}, \lambda^{(t)}) = \sum_{m=1}^{K} w_{.m}^{(t)} \log \lambda_m + \sum_{m=1}^{K} [\sum_{i=1}^{n} w_{im}^{(t)} \log f(y_i|\theta_m)]$$

- M-step: Let $\mathbf{w}_{..}^{(t)} = \sum_{m=1}^{K} w_{.m}^{(t)} = n$

$$Q_m(\theta_m|\theta^{(t)}, \lambda^{(t)}) = \sum_{i=1}^{n} w_{im}^{(t)} \log f(y_i|\theta_m)$$

Then $\lambda_m^{(t+1)} = \frac{w_{.m}^{(t)}}{\mathbf{w}_{..}^{(t)}} = \frac{w_{.m}^{(t)}}{n}$

$\theta_m^{t+1} = \text{argmax}_\theta \, Q_m(\theta_m|\theta^{(t)}, \lambda^{(t)})$

Some examples

- Mixture exponential
- Exponential family