Mixture Model Challenge

Václav Šmídl

April 8, 2022

Recapitulation

- What is latent variable of mixture model?

Recapitulation

- What is latent variable of mixture model?
- What is the conditional likelihood of observation of a mixture model?

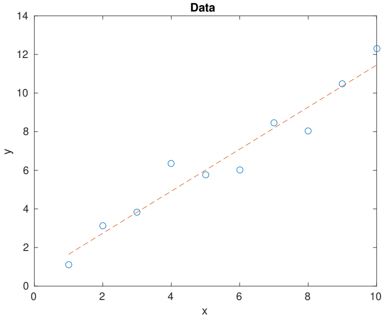
Recapitulation

- What is latent variable of mixture model?
- What is the conditional likelihood of observation of a mixture model?
- What is the advantage of one-hot representation?

Review of building blocks

Linear regression:

$$
\begin{aligned}
y_i &= \boldsymbol{x}_i\theta + e_i, \\
e_i &\sim \mathcal{N}(0, \sigma_e), \\
p(y_i|x_i, \theta) &= \mathcal{N}(\boldsymbol{x}_i\theta, \sigma_e),
\end{aligned}
$$



Review of building blocks

Linear regression likelihood:

$p(\theta_i)$

$$\begin{aligned} y_i &= \boldsymbol{x}_i\theta + e_i, \\ e_i &\sim \mathcal{N}(0, \sigma_e), \\ p(y_i|x_i, \theta) &= \mathcal{N}(\boldsymbol{x}_i\theta, \sigma_e), \end{aligned}$$

Prior (ridge):

$$p(\theta_i) = \mathcal{N}(0, \alpha)$$

Review of building blocks

Linear regression likelihood:

$$\begin{aligned} y_i &= \boldsymbol{x}_i\theta + e_i, \\ e_i &\sim \mathcal{N}(0, \sigma_e), \\ p(y_i|x_i, \theta) &= \mathcal{N}(\boldsymbol{x}_i\theta, \sigma_e), \end{aligned}$$

Prior (ridge):

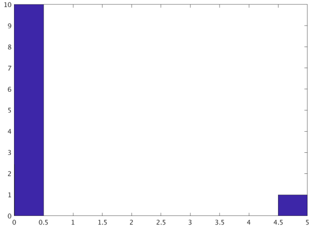
$$p(\theta_i) = \mathcal{N}(0, \alpha)$$

Prior Sparse:

$$\theta = [0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0]$$

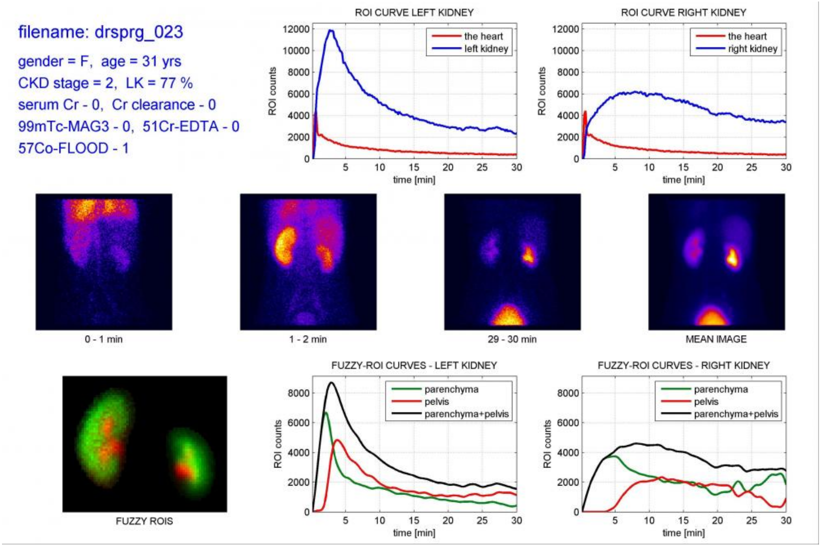
$$p(\theta_i) = \begin{cases} St(\nu, 0, \sigma_0) \\ 0.9\mathcal{N}(0, \sigma_0) + 0.1\mathcal{N}(0, \sigma_1) \\ 0.9\mathcal{N}(0, \sigma_0) + 0.1\mathcal{N}(0, 100\sigma_0) \end{cases}$$

$p(\theta_i)$



Patlak Rutland plot

Sequence of scintigraphic images of kidneys.

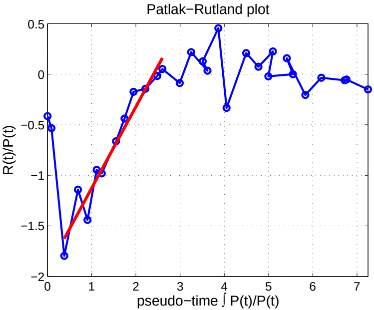


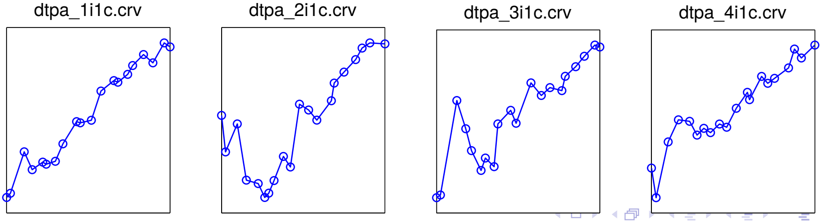
Challenge: Patlak Ruland plot

Patlak Rutland plot is a ratio of the parenchyma curve over the cumulative sum of the heart curve.

Diagnostically important is the **slope** of the linear part, which:

- typically starts around 1min
- typically ends around 3min
- with outliers
- manual fit is the norm





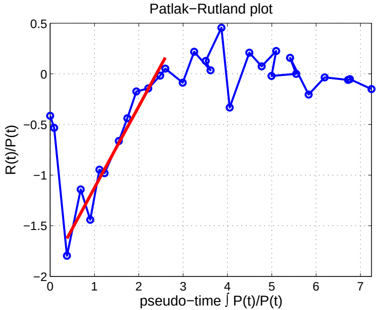
Models?

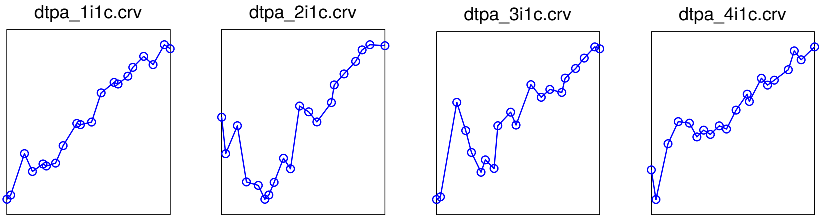
- no heuristics
- probabilistic model providing either:
  - estimate of $\hat{a}$
  - posterior distribution $p(a)$

  where a is the slope of:

$$y_i = ax_i + b + e_i$$

for points belonging to it.



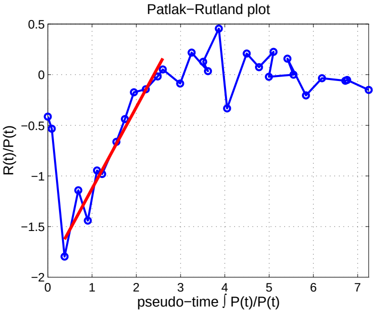


Model 1: Robust regression

Model

$$y_i = ax_i + b + e_i$$

is valid for all $x_i \in [1, 3]$ with heavy-tailed noise $p(e_i)$.



Model 1: Robust regression

Model

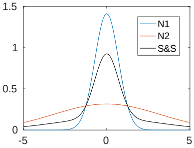
$$y_i = ax_i + b + e_i$$

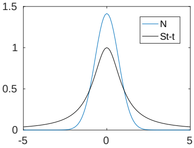
is valid for all $x_i \in [1, 3]$ with heavy-tailed noise $p(e_i)$.
Possibilities:

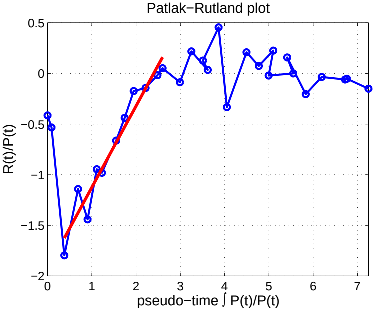
- Spike and slab,

$$p(e_i) = \alpha\mathcal{N}(0, v_1) + (1-\alpha)\mathcal{N}(0, v_2),$$

- Student-t $p(e_i) = \mathcal{S}t(\nu, 0, \sigma)$,







Latent representations

Spike and slab

$$p(e_i|v_1, v_2) = \alpha\mathcal{N}(0, v_1) + (1-\alpha)\mathcal{N}(0, v_2),$$

is the marginal of

$$\begin{aligned} p(e_i, l_i|v_1, v_2) &= p(e_i|l_i, v_1, v_2)p(l_i) \\ p(e_i|l_i = \epsilon_1, v_1) &= \mathcal{N}(0, v_1), \\ p(e_i|l_i = \epsilon_2, v_2) &= \mathcal{N}(0, v_2), \end{aligned}$$

Student-t

$p(e_i) = \mathcal{S}t(\nu, 0, \sigma)$, is the marginal of

$$\begin{aligned} p(e_i) &= \int p(e_i|\beta_i)p(\beta_i) \\ p(e_i|\beta_i) &= \mathcal{N}(0, \beta_i), \\ p(\beta_i) &= i\mathcal{G}(\gamma_0, \delta_0), \end{aligned}$$

Difference between S&S and Student

- minimum and maximum variance?
- the number of “clusters”?

Spike&Slab

Spike and slab prior of the noise implies:

$$p(e_i, l_i) = \mathcal{N}(0, v_1)^{l_{1,i}} \mathcal{N}(0, v_2)^{l_{2,i}},$$
$$p(y_i, l_i | x_i) = \mathcal{N}(ax_i + b, v_1)^{l_{1,i}} \mathcal{N}(ax_i + b, v_2)^{l_{2,i}},$$

Do the point outside of the line follow the $ax_i + b$ curve?

Spike&Slab

Spike and slab prior of the noise implies:

$$p(e_i, l_i) = \mathcal{N}(0, v_1)^{l_{1,i}} \mathcal{N}(0, v_2)^{l_{2,i}},$$
$$p(y_i, l_i | x_i) = \mathcal{N}(ax_i + b, v_1)^{l_{1,i}} \mathcal{N}(ax_i + b, v_2)^{l_{2,i}},$$

Do the point outside of the line follow the $ax_i + b$ curve?

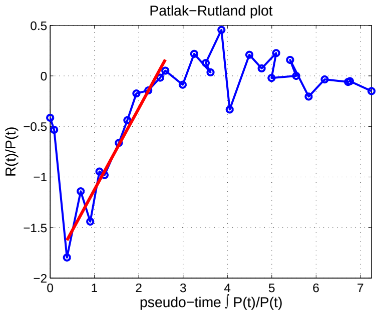
What model do they follow?

$$p(y_i, l_i | x_i) = \mathcal{N}(ax_i + b, v_1)^{l_{1,i}} \, p(x_i | l_i = \epsilon_2)^{l_{2,i}},$$

Possibilities:

$$p(x_i | l_i = \epsilon_2) = \mathcal{U}(y_{min}, y_{max}),$$
$$p(x_i | l_i = \epsilon_2) = \mathcal{N}(ax, v(x_i, \psi)),$$
$$\vdots$$

Be creative.



Tutorial model

Recall matrix notation from linear regression:

$$y_i = \boldsymbol{x}_i^\top \theta, \qquad y_i = \begin{bmatrix} x_i & 1 \end{bmatrix} \begin{bmatrix} a & b \end{bmatrix}^\top,$$

With one-hot label $l_i = [1, 0]$ if in regression $l_i = [0, 1]$ if outside

$$p(y_i|x_i, l_i) = \mathcal{N}(\boldsymbol{x}_i^\top \theta, \omega^{-1})^{l_{1,i}} U(-2, 1)^{l_{2,i}},$$

$$p(l_i|\alpha) = \alpha_1^{l_{1,i}} \alpha_2^{l_{2,i}}, \qquad p(\boldsymbol{\alpha}) = Di(z_0)$$

$$p(\theta) = \mathcal{N}(0, \tau I_2), \quad p(\omega) = \Gamma(\gamma_0, \delta_0)$$

Joint distribution:

$$\mathcal{L} = \sum_{i=1}^{n} \log p(y_i, l_i|x_i, \theta, \omega, \alpha) p(\theta, \omega, \alpha) = \frac{1}{2} \sum_i l_{1,i} \left(\log(\omega/2\pi) - \omega(y_i - \boldsymbol{x}_i^\top \theta)^2 \right)$$

$$+ \sum_i l_{2,i} \log \frac{1}{3} + \sum_i \sum_k l_{k,i} \log \alpha_k + (z_0 - 1) \sum_k \log \alpha_k +$$

$$+ \frac{2}{2} \log \tau - \frac{1}{2} \tau \theta^\top \theta + (\gamma_0 - 1) \log \omega - \delta_0 \omega,$$

$$\text{s.t} \sum_k l_{i,k} = 1, \ \sum_k \alpha_k = 1,$$

Categorical distribution

Categorical distribution $i \in \{c_1, c_2 \ldots c_K\}$

$$p(i|\boldsymbol{\alpha}) \quad = \quad \prod_{k=1}^{K} \alpha_k^{[i=c_k]}$$

Categorical distribution

Categorical distribution $i \in \{c_1, c_2 \ldots c_K\}$

$$p(i|\boldsymbol{\alpha}) = \prod_{k=1}^{K} \alpha_k^{[i=c_k]}$$

$$\sum \alpha_k = 1$$

where $[i = c_k]$ is the Iverson bracket.
With one-hot $I \in \{[1, 0 \ldots 0], [0, 1, \ldots 0] \ldots [0, 0 \ldots, 1]\}$

$$p(I|\alpha) = \prod_k \alpha_k^{I_k}$$

$$\log p(I|\alpha) = \sum_k I_k \log \alpha_k \qquad \mathrm{E}(I) = \alpha,\ \hat{I} = \alpha$$

In mixtures,

$$\log p(I|\ldots) \propto \sum_k I_k \mathcal{E}_k(\cdots),$$

Categorical distribution

Categorical distribution $i \in \{c_1, c_2 \ldots c_K\}$

$$p(i|\boldsymbol{\alpha}) = \prod_{k=1}^{K} \alpha_k^{[i=c_k]}$$

$$\sum \alpha_k = 1$$

where $[i = c_k]$ is the Iverson bracket.
With one-hot $I \in \{[1, 0 \ldots 0], [0, 1, \ldots 0] \ldots [0, 0 \ldots, 1]\}$

$$p(I|\alpha) = \prod_k \alpha_k^{I_k}$$

$$\log p(I|\alpha) = \sum_k I_k \log \alpha_k \qquad \mathrm{E}(I) = \alpha,\ \hat{I} = \alpha$$

In mixtures,

$$\log p(I|\ldots) \propto \sum_k I_k \mathcal{E}_k(\cdots),$$

$$\hat{I}_k = \frac{\exp \mathcal{E}_k(\cdots)}{\sum \exp \mathcal{E}_k(\cdots)}$$

Expectation maximization (EM) algorithm: M-step

Expectation over l: $l_{1,i} \to \hat{l}_{1,i}$, $l_{2,i} \to \hat{l}_{2,i}$

$$\begin{aligned}
\mathrm{E}_l(\mathcal{L}) = &\frac{1}{2}\sum_i \hat{l}_{1,i}\log\omega - \frac{1}{2}\sum_i \hat{l}_{1,i}\omega(y_i - \boldsymbol{x}_i^\top\theta)^2 \\
&+ \sum_i \hat{l}_{2,i}\log\frac{1}{3} + \sum_i\sum_k \hat{l}_{k,i}\log\alpha_k + (z_0 - 1)\sum_k \log\alpha_k + \\
&+ \frac{2}{2}\log\tau - \frac{1}{2}\tau\theta^\top\theta + (\gamma_0 - 1)\log\omega - \delta_0\omega + \text{n.c.}, \qquad \text{s.t. : } \sum_k \alpha_k = 1,
\end{aligned}$$

In θ (via moment matching of $q(\theta) = \mathcal{N}(\hat{\theta}, \Sigma_\theta)$):

$$\begin{aligned}
\mathcal{L} &= -\frac{1}{2}\textstyle\sum_i \hat{l}_{1,i}\omega(y_i - \boldsymbol{x}_i^\top\theta)^2 - \frac{1}{2}\tau\theta^\top\theta \\
&= -\frac{1}{2}\textstyle\sum_i \hat{l}_{1,i}\omega(y_i^2 - 2y_i\boldsymbol{x}_i^\top\theta + \theta^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\theta) - \frac{1}{2}\tau\theta^\top\theta \\
&= -\frac{1}{2}\left(\omega\textstyle\sum_i \hat{l}_{1,i}y_i^2 - 2\omega\sum_i \hat{l}_{1,i}y_i\boldsymbol{x}_i^\top\theta + \theta^\top\left(\omega\sum_i \hat{l}_{1,i}\boldsymbol{x}_i\boldsymbol{x}_i^\top + \tau I_2\right)\theta\right) \\
\hat{\theta} &= \left(\omega\textstyle\sum_i \hat{l}_{1,i}\boldsymbol{x}_i\boldsymbol{x}_i^\top + \tau I_2\right)^{-1}\left(\omega\sum_i \hat{l}_{1,i}y_i\boldsymbol{x}_i^\top\right)
\end{aligned}$$

Expectation maximization (EM) algorithm: M-step

Expectation over I: $I_{1,i} \to \hat{I}_{1,i}$, $I_{2,i} \to \hat{I}_{2,i}$

$$\mathcal{L} = \frac{1}{2}\sum_i \hat{I}_{1,i}\log\omega - \frac{1}{2}\sum_i \hat{I}_{1,i}\omega(y_i - \boldsymbol{x}_i^\top\theta)^2$$
$$+ \sum_i \hat{I}_{2,i}\log\frac{1}{3} + \sum_i\sum_k \hat{I}_{k,i}\log\alpha_k + (z_0 - 1)\sum_k \log\alpha_k +$$
$$+ \frac{2}{2}\log\tau - \frac{1}{2}\tau\theta^\top\theta + (\gamma_0 - 1)\log\omega - \delta_0\omega + \text{n.c.}, \qquad \text{s.t.}: \sum_k \alpha_k = 1,$$

In ω, (via moment matching of $q(\omega) = \mathcal{G}(\gamma, \delta)$):

$$\mathcal{L} = (\gamma - 1)\log\omega - \delta\omega$$
$$\gamma = \frac{1}{2}\sum_i \hat{I}_{1,i} + \gamma_0$$
$$\delta = \frac{1}{2}\textstyle\sum_i \hat{I}_{1,i}(y_i - \boldsymbol{x}_i^\top\hat{\theta})^2 + \delta_0$$
$$\hat{\omega} = \gamma/\delta$$

Expectation maximization (EM) algorithm: M-step

Expectation over l: $l_{1,i} \to \hat{l}_{1,i}$, $l_{2,i} \to \hat{l}_{2,i}$

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{2}\sum_i \hat{l}_{1,i} \log \omega - \frac{1}{2}\sum_i \hat{l}_{1,i}\omega(y_i - \boldsymbol{x}_i^\top \theta)^2 \\
&+ \sum_i \hat{l}_{2,i} \log \frac{1}{3} + \sum_i \sum_k \hat{l}_{k,i} \log \alpha_k + (z_0 - 1)\sum_k \log \alpha_k + \\
&+ \frac{2}{2}\log \tau - \frac{1}{2}\tau\theta^\top\theta + (\gamma_0 - 1)\log \omega - \delta_0 \omega + \text{n.c.}, \qquad \text{s.t.} : \sum_k \alpha_k = 1,
\end{aligned}
$$

In α (via moment matching of $q(\alpha) = \mathcal{D}i(z)$):

$$
\begin{aligned}
\mathcal{L} &= \sum_k \sum_i \hat{l}_{k,i} \log \alpha_k + (z_0 - 1)\sum_k \log \alpha_k, \qquad \text{s.t.} : \sum_k \alpha_k = 1,, \\
&= \sum_k z_k \log \alpha_k + \lambda(\sum_k \alpha_k - 1) \\
z_k &= \sum_i \hat{l}_{k,i} + z_0 \\
\hat{\alpha}_k &= z_k / \sum_k z_k
\end{aligned}
$$

Expectation maximization (EM) algorithm: E-step

Probability of l_i given estimates

$$\mathcal{L}(l_i) = \frac{1}{2} l_{1,i} \log \omega - \frac{1}{2} l_{1,i} \omega (y_i - \boldsymbol{x}_i^\top \theta)^2 - l_{1,i} \frac{1}{2} \log(2\pi)$$
$$+ l_{2,i} \log \frac{1}{3} + \sum_k l_{k,i} \log \alpha_k + (z_0 - 1) \sum_k \log \alpha_k + \text{n.c.}, \qquad \text{s.t. : } \sum_k l_k = 1$$

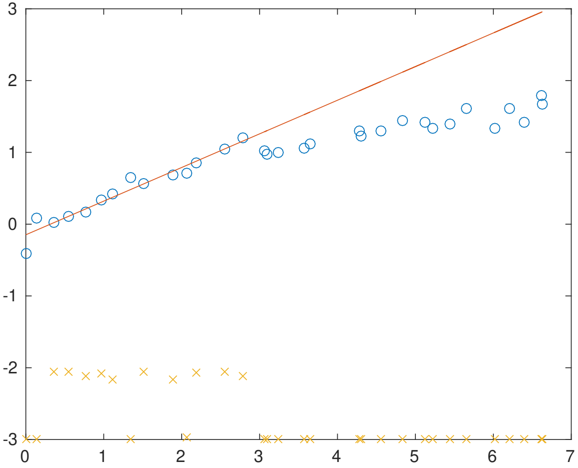
Template:

$$p(l_i | w) \propto \prod_k w_k^{l_{k,i}}, \qquad \sum_k w_k = 1, \qquad \sum_k l_{k,i} = 1 (\text{in our case})$$

Matching (check with the Bayes rule):

$$\log w_1 = \frac{1}{2} \log \omega - \frac{1}{2} \omega (y_i - \boldsymbol{x}_i^\top \theta)^2 - \frac{1}{2} \log(2\pi) + \log \alpha_1$$
$$\log w_2 = \log \frac{1}{3} + \log \alpha_2$$
$$\hat{l}_1 = w_1 / (w_1 + w_2)$$

EM algorithm results



- sensitive to initizalization

Modeling choices

How to encode our knowledge that the linear part is somewhere between 1 and 3min? hard constraints should be avoided.

Modeling choices

How to encode our knowledge that the linear part is somewhere between 1 and 3min? hard constraints should be avoided.

1. Prior on the latent l:
$$p(l|x) = \alpha_0(x),$$
2. Other options? Variance of the noise?

Assignment

Load data Patlak.mat
35 studies with:

- xpr x axis
- ypr y axis
- name name of the study
- int_start ignore
- int_end ignore

Assignment	points
find slope of linear part for all 35 studies	
own code (WLS + uniform)	5
own code (WLS + novel proposition)	10