

EM-algorithm for mixture of generalized Gaussian distribution (GGD)

1 Definitions and problem statement

Density of GGD:

$$GND(x|\mu, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x-\mu|/\alpha)^\beta}.$$

Mixture of GGDs:

$$p(x|\theta) = \sum_{k=1}^K \pi_k p_k(x|\theta_k)$$

Maximum a posteriori estimation (MAP):

$$p(\theta|D) \propto \prod_{i=1}^N p(x_i|\theta) \cdot \prod_{k=1}^K p(\theta_k|\omega) \xrightarrow{\theta} \max,$$

with independent components:

$$p_k(x) = \prod_{j=1}^d GND(x^j|\mu_k^j, \alpha_k^j, \beta),$$

$$\log p_k(x_i) = d(\log \beta - \log \Gamma(1/\beta)) - \sum_{j=1}^d \log(2\alpha_k^j) - \sum_{j=1}^d \left(\frac{|x_i^j - \mu_k^j|}{\alpha_k^j} \right)^\beta.$$

Prior distribution on rectangle side lengths:

$$p(\alpha_k|a, b) = \prod_{j=1}^d p(\alpha_k^j|a, b),$$

$$p(\alpha|a, b) \propto \alpha^{-1-a} e^{-b/\alpha^\beta}.$$

2 EM-algorithm

2.1 Definitions

- $X = (x_i^j)$ – matrix $N \times d$;
- $p = (p_k(x_i)) = (p_1, \dots, p_K)$ – matrix $N \times K$;

- $\pi = (\pi_k) = (\pi_1, \dots, \pi_K)^T$ – matrix $K \times 1$;
- $\mu = (\mu_k^j)$ – matrix $K \times d$;
- $\alpha = (\alpha_k^j)$ – matrix $K \times d$;
- $r = (r_{ik})$ – matrix $N \times K$.

2.2 E-step

$$r_{ik} = \frac{\pi_k^t p_k(x_i | \theta_k^t)}{\sum_{k'=1}^K \pi_{k'}^t p_{k'}(x_i | \theta_{k'}^t)};$$

In matrix language:

1. $\tilde{r} = p \cdot \text{diag}(\pi) = (p_1 \pi_1, \dots, p_K \pi_K) = (\tilde{r}_1, \dots, \tilde{r}_N)^T$ – matrix $N \times K$;
2. $u = \log \tilde{r} = (\log p_1 + \log \pi_1, \dots, \log p_K + \log \pi_K) = (u_1, \dots, u_N)^T$.
3. $r = (\tilde{r}_1 / \sum \tilde{r}_1, \dots, \tilde{r}_N / \sum \tilde{r}_N)^T$;
4. $\log r = (u_1 - \log(\sum e^{u_1}), \dots, u_N - \log(\sum e^{u_N}))$.

2.3 M-step

Stop condition:

$$\frac{|p(\theta^{t+1}|D) - p(\theta^t|D)|}{p(\theta^t|D)} < \varepsilon$$

$$\pi_k = \frac{\sum_{i=1}^N r_{ik}}{N}.$$

$$\mu_k^j = \arg \min m_k^j(t_k^j) = \sum_{i=1}^N r_{ik} |x_i^j - t_k^j|^\beta.$$

Легко видеть, что для минимизации $l(\mu, \alpha)$ достаточно сначала минимизировать её по μ , а затем по α . Минимизация по μ эквивалентна минимизации $m(\mu)$, которая является выпуклой функцией, принимающей на $\pm\infty$ значение $+\infty$, а значит, достигающей единственной точки минимума. Более того, очевидно, что данная точка минимума лежит на отрезке $[\min x_i, \max x_i]$. Значит, функционал $m(\mu)$ может быть эффективно минимизирован, например, методом золотого сечения. Обозначим минимальное значение $m(\mu)$ как m_* , после чего остается минимизировать функцию

$$l(\alpha) = (N + a + 1) \log \alpha + \frac{m_* + b}{\alpha^\beta},$$

что может быть легко проделано аналитически и приводит к следующему результату:

$$\alpha_*^\beta = \frac{\beta(m_* + b)}{N + a + 1} = \left(1 - \frac{a + 1}{N + a + 1}\right) \frac{\beta m_*}{N} + \frac{a + 1}{N + a + 1} \cdot \frac{\beta b}{a + 1}.$$

Перепараметризуя гиперпараметры априорного распределения с помощью формул

$$\alpha_0^\beta = \frac{\beta b}{a + 1}, \lambda = \frac{a + 1}{N + a + 1},$$

и используя обозначение

$$\alpha_{MLE} = \left(\frac{\beta m_*}{N} \right)^{1/\beta},$$

получаем окончательное выражение:

$$\alpha_* = \left((1 - \lambda) \alpha_{MLE}^\beta + \lambda \alpha_0^\beta \right)^{1/\beta}.$$