Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística - ME524

# Algoritmo EM para mistruras

Grupo
Victor Dalla 206493, Mariana Ferreira 183670
Prof. Dra. Mariana Motta

Campinas
2019

**Resumo**

Texto resumo.

# Introdução

- comentar misturas e abordagem bayesiana
- ferramentas computacionais

Os estatísticos que trabalham com análise e modelagem de dados atualmente estão em uma posição luxuosa de conseguir estimar, prever e inferir sobre sistemas complexos de interesse, graças a métodos computacionais cada vez mais poderosos e robustos. Modelos robustos, como os modelos de mistura, constituem uma fascinante ilustração desses aspectos: enquanto dentro de uma família paramétrica, eles oferecem aproximações maleáveis em ambientes não paramétricos e, embora baseados em distribuições padrões, eles representam desafios computacionais altamente complexos.
As distribuições de misturas compreendem um número finito ou infinito de componentes, possivelmente de diferentes tipos de distribuições, que descrevem as características dos dados. Facilitam, assim, uma descrição muito mais cuidadosa dos sistemas complexos. Por exemplo, na genética, a localização de características quantitativas em um cromossomo e a interpretação de microarranjos relacionam-se a misturas.
Abordagens Bayesianas à modelagem de misturas têm atraído grande interesse entre pesquisadores e praticantes. O paradigma Bayesiano permite que declarações de probabilidade sejam feitas diretamente sobre os parâmetros desconhecidos e opiniões prévias a serem incluídas na análise e modelagem do modelo. Essa estrutura também permite que a dificuldade de um modelo de mistura seja decomposta em um conjunto de estruturas mais simples, através do uso de variáveis latentes.

## Mistura finita

A descrição de uma mistura de distribuições é simples: qualquer combinação convexa de outras distribuições $f_i$ é uma mistura, como mostra a combinação abaixo:

$$\sum_{i=1}^{k} p_i f_i(x), \quad \sum_{i=1}^{k} p_i = 1, \quad k > 1$$

Na maioria dos casos, as distribuições $f_i$ são de uma família paramétrica, com parâmetro desconhecido $\theta_i$, levando ao modelo de mistura paramétrica:

$$\sum_{i=1}^{k} p_i f\left(x|\theta_i\right)$$

Além disso, o comportamento da cauda de uma mistura é sempre descrito por um ou dois de seus componentes e que, portanto, deve refletir a escolha da família paramétrica $f\left(.|\theta_i\right)$. Note também que a representação de misturas como combinações convexas de distribuições implica na propriedade de cálculo dos momentos:

$$\mathbb{E}\left[X^m\right] = \sum_{i=1}^{k} p_i \mathbb{E}^{f_i}\left[X^m\right]$$

A verossimilhança $\mathbb{L}(\underline{\theta}, \underline{p}|\underline{x}) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f\left(x_i|\theta_j\right)$ de uma mistura de k distribuições tem $k^n$ termos, o que impossibilita alguma solução analítica.

## Dados faltantes

```
# \[ X_{i}\left|Z_{i}=z \sim f\left(x |
# \theta_{z}\right), \qquad Z_{i} \sim
# \mathscr{M}_{k}\left(1 ; p_{1}, \ldots,
# p_{k}\right)\right. \]
```

## Algoritmo

## Simulação

- mistura de normais com apenas as médias desconhecidas
- apresentar os problemas com abordagem EM

Mistura: $pN\left(\mu_1, 1\right) + (1-p)N\left(\mu_2, 1\right)$, $(\mu_1, \mu_2, p) = (0, 2.5, 0.7)$

```
library(dplyr)
```

```
rmixnorm <- function(n, mean = list(0), sd = list(1), p = list(0.5)) {
    norm <- c(do.call(rmultinom, list(1, length(p), p)))
    rnorm(n, mean[[norm]], sd[[norm]])
}
```

## Resultado