UNIVERSIDADE FEDERAL DO CEARÁ

DEPARTAMENTO DE COMPUTAÇÃO

LogIA

Aprendizagem de Máquina Probabilística

César Lincoln Cavalcante Mattos

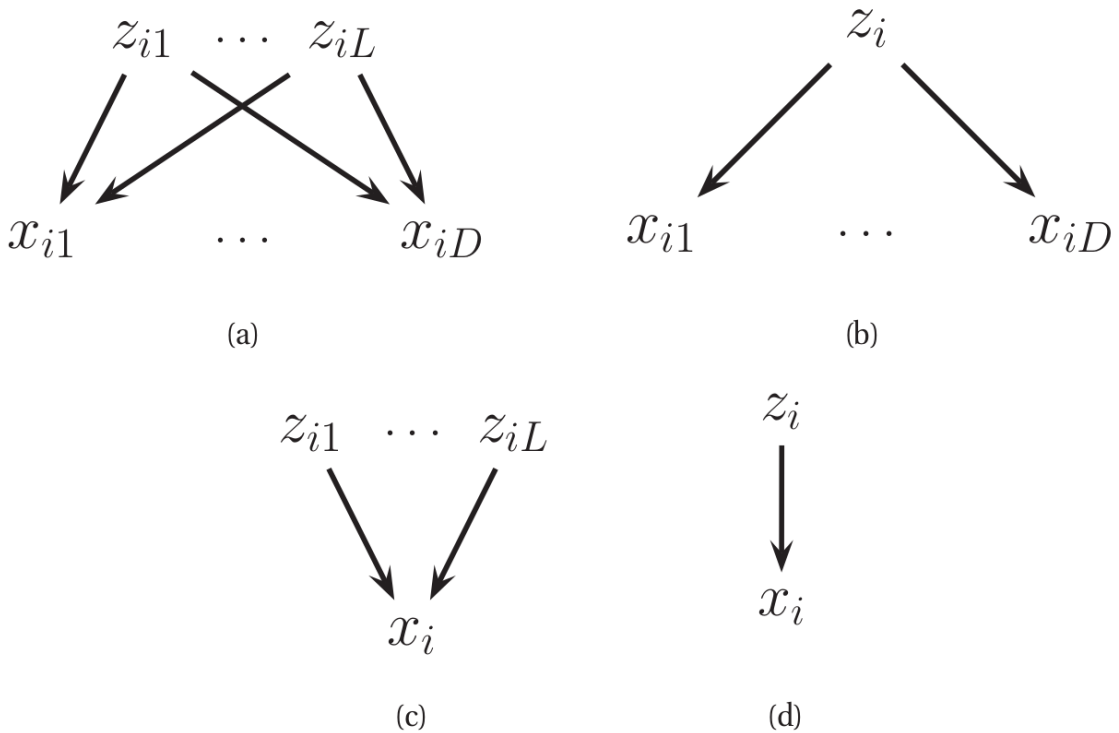
2025

Agenda

1. Modelos de misturas
2. Algoritmo Expectation-Maximization
   - Algoritmo EM para Mistura de Gaussianas (GMMs)
   - Algoritmo EM para GMMs com estimação MAP
   - Seleção de modelos para GMM
   - Algoritmo EM para Mistura de Especialistas
   - Algoritmo EM como um limiar inferior
   - Variantes do algoritmo EM
   - Algoritmo EM para dados faltantes
3. Tópicos adicionais
4. Referências

Modelos de variáveis latentes

- Uma maneira de modelar variáveis correlacionadas é através de **modelos de variáveis latentes** (*latent variable models*, LVM).
- Considera que as observações foram geradas por "causas" ocultas comuns.
- Pode ser visto como uma maneira de obter um "gargalo" (*bottleneck*) que representa os dados de maneira comprimida.



Modelos de misturas

- Considere um modelo generativo com variáveis latentes discretas:

$$z \sim p(z), \text{ em que } z \in \{1, \ldots, K\},$$
$$\boldsymbol{x} \sim p(\boldsymbol{x}|z).$$

Modelos de misturas

- Considere um modelo generativo com variáveis latentes discretas:

$$z \sim p(z), \text{ em que } z \in \{1, \ldots, K\},$$
$$\boldsymbol{x} \sim p(\boldsymbol{x}|z).$$

- Escolhemos uma priori $p(z) = \mathrm{Cat}(\boldsymbol{\pi})$, ou seja, $p(z = k) = \pi_k$.
- A verossimilhança $p(\boldsymbol{x}|z = k)$ relaciona a variável latente z e a observação $\boldsymbol{x}$.

Modelos de misturas

- Considere um modelo generativo com variáveis latentes discretas:

$$z \sim p(z), \text{ em que } z \in \{1, \ldots, K\},$$
$$\boldsymbol{x} \sim p(\boldsymbol{x}|z).$$

- Escolhemos uma priori $p(z) = \mathrm{Cat}(\boldsymbol{\pi})$, ou seja, $p(z = k) = \pi_k$.
- A verossimilhança $p(\boldsymbol{x}|z = k)$ relaciona a variável latente z e a observação $\boldsymbol{x}$.
- Um modelo de mistura pode ser obtido ao marginalizar z:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\boldsymbol{x}|z = k)p(z = k)$$
$$= \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|z = k),$$

em que $0 \leq \pi_k \leq 1$, $\sum_{k=1}^{K} \pi_k = 1$ e $\boldsymbol{\theta}$ são parâmetros do modelo.

Modelos de misturas para agrupamento

- Ao usar um modelo de misturas nas observações $\boldsymbol{x}_i|_{i=1}^N$ disponíveis, podemos encontrar qual componente (ou grupo/*cluster*) a gerou calculando a posteriori $p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta})$.

Modelos de misturas para agrupamento

- Ao usar um modelo de misturas nas observações $\boldsymbol{x}_i|_{i=1}^N$ disponíveis, podemos encontrar qual componente (ou grupo/*cluster*) a gerou calculando a posteriori $p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta})$.
- No chamado **soft clustering**, computamos a **responsabilidade** r_{ik} do grupo k via regra de Bayes:

$$r_{ik} \triangleq p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{p(z_i = k|\boldsymbol{\theta})p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\theta})}{\sum_{k'=1}^{K} p(z_i = k'|\boldsymbol{\theta})p(\boldsymbol{x}_i|z_i = k', \boldsymbol{\theta})}.$$

- Note que a diferença em relação a um classificador generativo é que z_i não é observado.

Modelos de misturas para agrupamento

- Ao usar um modelo de misturas nas observações $\boldsymbol{x}_i|_{i=1}^N$ disponíveis, podemos encontrar qual componente (ou grupo/*cluster*) a gerou calculando a posteriori $p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta})$.
- No chamado **soft clustering**, computamos a **responsabilidade** r_{ik} do grupo k via regra de Bayes:

$$r_{ik} \triangleq p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{p(z_i = k|\boldsymbol{\theta})p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\theta})}{\sum_{k'=1}^{K} p(z_i = k'|\boldsymbol{\theta})p(\boldsymbol{x}_i|z_i = k', \boldsymbol{\theta})}.$$

- Note que a diferença em relação a um classificador generativo é que z_i não é observado.
- No caso do chamado **hard clustering**, temos:

$$\hat{z}_i = \arg\max_k r_{ik} = \arg\max_k [\log p(z_i = k|\boldsymbol{\theta}) + \log p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\theta})].$$

Mistura de especialistas

- Modelos de misturas também podem ser usados no contexto de modelos discriminantes para regressão ou classificação.
- Uma **mistura de especialistas** (*mixture of experts*, MoE) é obtida a partir de submodelos (especialistas) em certas regiões do espaço de entrada.

Mistura de especialistas

- Modelos de misturas também podem ser usados no contexto de modelos discriminantes para regressão ou classificação.
- Uma **mistura de especialistas** (*mixture of experts*, MoE) é obtida a partir de submodelos (especialistas) em certas regiões do espaço de entrada.
- Para K modelos de regressão linear, temos:

$$
\begin{aligned}
p(y_i|\boldsymbol{x}_i, z_i = k, \boldsymbol{\theta}) &= \mathcal{N}(y_i|\boldsymbol{w}_k^\top \boldsymbol{x}_i, \sigma_k^2), \quad 1 \leq k \leq K, \\
p(z_i|\boldsymbol{x}_i, \boldsymbol{\theta}) &= \mathrm{Cat}(z_i, \mathcal{S}(\boldsymbol{V}^\top \boldsymbol{x}_i)), \\
p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) &= \sum_{k=1}^{K} p(y_i|\boldsymbol{x}_i, z_i = k, \boldsymbol{\theta}) p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta}),
\end{aligned}
$$

em que $p(z_i|\boldsymbol{x}_i, \boldsymbol{\theta})$ age como uma *gating function* e $\mathcal{S}(\cdot)$ é uma função softmax parametrizada por $\boldsymbol{V}$.

Mistura de Gaussianas

- O modelo de mistura de Gaussianas (*Gaussian mixture model*, GMM) é um dos mais usados para representar dados contínuos:

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(z_i = k)\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Mistura de Gaussianas

- O modelo de mistura de Gaussianas (*Gaussian mixture model*, GMM) é um dos mais usados para representar dados contínuos:

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(z_i = k)\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

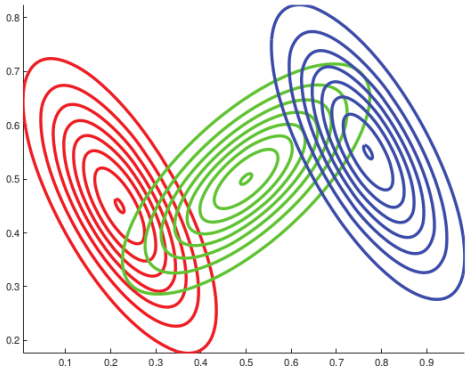
- Alternativamente, como z_i possui uma distribuição categórica, podemos representá-la por um vetor $\boldsymbol{z}_i \in \{0, 1\}^K$ em que $\sum_k z_{ik} = 1$:

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \underbrace{p(z_{ik} = 1)}_{\pi_k} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

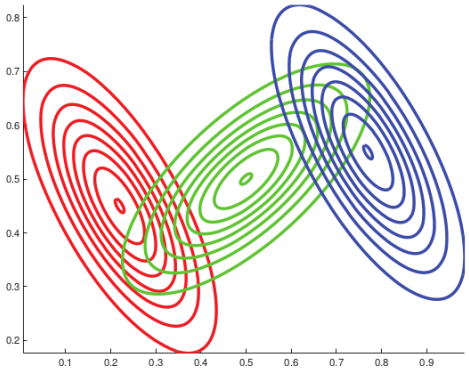
Mistura de Gaussianas

- Um GMM com componentes D-dimensionais suficientes pode aproximar qualquer distribuição contínua em $\mathbb{R}^D$.

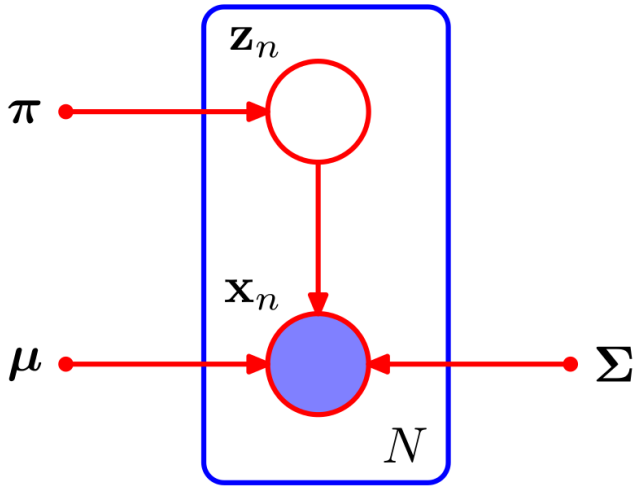


Mistura de Gaussianas

- Um GMM com componentes D-dimensionais suficientes pode aproximar qualquer distribuição contínua em $\mathbb{R}^D$.



- As variáveis do modelo e suas relações podem ser representadas por um **modelo gráfico probabilístico**:



Algoritmo Expectation-Maximization

- ML ou MAP não é aplicável diretamente ao GMM (ou qualquer outro LVM), pois as variáveis latentes não são observadas.

Algoritmo Expectation-Maximization

- ML ou MAP não é aplicável diretamente ao GMM (ou qualquer outro LVM), pois as variáveis latentes não são observadas.
- Em geral, a verossimilhança dos dados observados é dada por:

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{z_i} p(\boldsymbol{x}_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}),$$

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{z_i} p(\boldsymbol{x}_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}),$$

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{z_i} p(\boldsymbol{x}_i, z_i|\boldsymbol{\theta}) \right],$$

em que todos os parâmetros foram reunidos em $\boldsymbol{\theta}$.

Algoritmo Expectation-Maximization

- ML ou MAP não é aplicável diretamente ao GMM (ou qualquer outro LVM), pois as variáveis latentes não são observadas.
- Em geral, a verossimilhança dos dados observados é dada por:

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{z_i} p(\boldsymbol{x}_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}),$$

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{z_i} p(\boldsymbol{x}_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}),$$

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{z_i} p(\boldsymbol{x}_i, z_i|\boldsymbol{\theta}) \right],$$

em que todos os parâmetros foram reunidos em $\boldsymbol{\theta}$.

- **Problema**: as variáveis latentes não são observadas e $\mathcal{L}(\boldsymbol{\theta})$ não pode ser calculada.

Algoritmo Expectation-Maximization

- **Ideia**: Considerar que os valores z_i são conhecidos e definir a verossimilhança dos dados completos (conjunta):

$$\mathcal{L}_c(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{N} \log p(\boldsymbol{x}_i, z_i|\boldsymbol{\theta}).$$

Algoritmo Expectation-Maximization

- **Ideia**: Considerar que os valores z_i são conhecidos e definir a verossimilhança dos dados completos (conjunta):

$$\mathcal{L}_c(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{N} \log p(\boldsymbol{x}_i, z_i|\boldsymbol{\theta}).$$

- **Ideia**: Computar $\mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})]$ em relação à sua posteriori:

$$\begin{aligned}\mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})] &= \sum_{k=1}^{K} p(z_i = k|\boldsymbol{X}, \boldsymbol{\theta})\mathcal{L}_c(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} p(z_i = k|\boldsymbol{X}, \boldsymbol{\theta}) \log p(\boldsymbol{x}_i, z_i|\boldsymbol{\theta}).\end{aligned}$$

Algoritmo Expectation-Maximization

- **Ideia**: Considerar que os valores z_i são conhecidos e definir a verossimilhança dos dados completos (conjunta):

$$\mathcal{L}_c(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{N} \log p(\boldsymbol{x}_i, z_i|\boldsymbol{\theta}).$$

- **Ideia**: Computar $\mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})]$ em relação à sua posteriori:

$$\begin{aligned}\mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})] &= \sum_{k=1}^{K} p(z_i = k|\boldsymbol{X}, \boldsymbol{\theta})\mathcal{L}_c(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} p(z_i = k|\boldsymbol{X}, \boldsymbol{\theta}) \log p(\boldsymbol{x}_i, z_i|\boldsymbol{\theta}).\end{aligned}$$

- **Ideia**: Como buscamos uma solução de ML, otimizamos $\mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})]$ em função de $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_{ML} = \arg\max \mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})].$$

Algoritmo Expectation-Maximization

- **Problema**: Temos uma relação cruzada de dependência entre $\mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})]$ e $\boldsymbol{\theta}$.

Algoritmo Expectation-Maximization

- **Problema**: Temos uma relação cruzada de dependência entre $\mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})]$ e $\boldsymbol{\theta}$.
- **Ideia**: Seguimos uma estratégia iterativa, resultando no algoritmo EM (**Expectation-Maximization**):

Algoritmo Expectation-Maximization

- **Problema**: Temos uma relação cruzada de dependência entre $\mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})]$ e $\boldsymbol{\theta}$.
- **Ideia**: Seguimos uma estratégia iterativa, resultando no algoritmo EM (**Expectation-Maximization**):
  - **Passo E**: O valor esperado na iteração atual t é dado por:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})|\boldsymbol{X}, \boldsymbol{\theta}^{(t-1)}],$$

    em que explicitou-se a dependência dos dados de treino $\boldsymbol{X}$ e dos parâmetros da iteração anterior $\boldsymbol{\theta}^{(t-1)}$.

Algoritmo Expectation-Maximization

- **Problema**: Temos uma relação cruzada de dependência entre $\mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})]$ e $\boldsymbol{\theta}$.
- **Ideia**: Seguimos uma estratégia iterativa, resultando no algoritmo EM (**Expectation-Maximization**):
  - **Passo E**: O valor esperado na iteração atual t é dado por:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})|\boldsymbol{X}, \boldsymbol{\theta}^{(t-1)}],$$

    em que explicitou-se a dependência dos dados de treino $\boldsymbol{X}$ e dos parâmetros da iteração anterior $\boldsymbol{\theta}^{(t-1)}$.
  - **Passo M**: O valor atualizado dos parâmetros é obtido via otimização da função auxiliar $Q(\cdot)$:

$$\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}).$$

Agenda

Algoritmo EM para GMMs

- No caso de um GMM com K componentes, temos:

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(\boldsymbol{x}_i, z_i|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{N} \log p(\boldsymbol{x}_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{N} \log \prod_{k=1}^{K} [\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{\mathbb{I}(z_i=k)} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}(z_i = k) [\log \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)],
\end{aligned}$$

  em que o expoente $\mathbb{I}(z_i = k)$ foi incluído por considerarmos que $\boldsymbol{x}_i$ foi observado da k-ésima componente da mistura.
- Note que $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}|_{k=1}^{K}$.

Algoritmo EM para GMMs

- A função auxiliar é obtida tomando a esperança com relação a z_i:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) &= \mathbb{E}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k)[\log \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}[\mathbb{I}(z_i = k)][\log \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \underbrace{p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})}_{r_{ik}}[\log \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik} \log \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),
\end{aligned}
$$

em que r_{ik} é o **coeficiente de responsabilidade** da componente k pela observação i.

Algoritmo EM para GMMs

- Portanto, o **passo E** consiste em computar os coeficientes de responsabilidade r_{ik}:

$$r_{ik} \triangleq p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}) = \frac{p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\theta}^{(t-1)})p(z_i = k)}{\sum_{k'=1}^{K} p(\boldsymbol{x}_i|z_i = k', \boldsymbol{\theta}^{(t-1)})p(z_i = k')}$$
$$= \frac{\pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t-1)} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{k'}^{(t-1)}, \boldsymbol{\Sigma}_{k'}^{(t-1)})}.$$

Algoritmo EM para GMMs

- No **passo M**, otimizamos $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$ com relação a $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}|_{k=1}^{K}$.
- Para π_k, como $\sum_k \pi_k = 1$, derivamos o Lagrangiano abaixo:

$$\frac{\partial}{\partial \pi_k}\left[Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) + \lambda\left(1 - \sum_{k=1}^{K}\pi_k\right)\right] = 0$$

$$\frac{\partial}{\partial \pi_k}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}\log\pi_k + \lambda\left(1 - \sum_{k=1}^{K}\pi_k\right)\right] = 0$$

$$\sum_{i=1}^{N}\left[\frac{1}{\pi_k}r_{ik} - \lambda\right] = 0$$

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

- Como $\sum_k \pi_k = 1$, temos $\lambda = \frac{\sum_k \sum_i r_{ik}}{N} = 1$ e $\pi_k = \frac{\sum_{i=1}^{N} r_{ik}}{N}$.

Algoritmo EM para GMMs

- Precisamos ainda otimizar $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$ em relação a $\boldsymbol{\mu}_k$ e $\boldsymbol{\Sigma}_k$:

$$Q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)}) = \sum_{i=1}^{N} r_{ik} \log \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\propto \sum_{i=1}^{N} r_{ik} \left[-\frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right].$$

- A última expressão corresponde à log-verossimilhança de uma Gaussiana em que cada termo é ponderado por r_{ik}, logo:

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} r_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} r_{ik}},$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} r_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top}{\sum_{i=1}^{N} r_{ik}} = \frac{\sum_{i=1}^{N} r_{ik} \boldsymbol{x}_i \boldsymbol{x}_i^\top}{\sum_{i=1}^{N} r_{ik}} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top.$$

Algoritmo EM para GMMs

Resumo do algoritmo

1. Escolha $K, \pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}, \forall k$.
2. Faça $t = 1$ e repita até convergir:
   - Passo E:

$$r_{ik} \triangleq p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}) = \frac{\pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t-1)} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{k'}^{(t-1)}, \boldsymbol{\Sigma}_{k'}^{(t-1)})}.$$

   - Passo M:

$$\pi_k^{(t)} = \frac{\sum_{i=1}^{N} r_{ik}}{N},$$

$$\boldsymbol{\mu}_k^{(t)} = \frac{\sum_{i=1}^{N} r_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} r_{ik}}, \quad \boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_{i=1}^{N} r_{ik} \boldsymbol{x}_i \boldsymbol{x}_i^\top}{\sum_{i=1}^{N} r_{ik}} - \boldsymbol{\mu}_k^{(t)} (\boldsymbol{\mu}_k^{(t)})^\top.$$

Algoritmo K-médias

- Considera-se que $\pi_k = \frac{1}{K}, \boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}, \forall k$, são fixos.
- Considera-se ainda que $r_{ik} \triangleq p(z_i = k | \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}) \approx \mathbb{I}(k = \hat{z}_i)$, em que $\hat{z}_i = \arg\max_k p(z_i = k | \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$

Resumo do algoritmo

1. Escolha $K, \boldsymbol{\mu}_k^{(0)}, \forall k$.
2. Faça $t = 1$ e repita até convergir:
   - Atribuição dos padrões:

$$\hat{z}_i = \arg\max_k p(z_i = k | \boldsymbol{x}_i, \boldsymbol{\mu}_k^{(t-1)}) = \arg\min_k \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t-1)} \right\|^2 .$$

   - Atualização dos centróides:

$$\boldsymbol{\mu}_k^{(t)} = \frac{\sum_i \mathbb{I}(\hat{z}_i = k) \boldsymbol{x}_i}{\sum_i \mathbb{I}(\hat{z}_i = k)} = \frac{1}{N_k} \sum_{i:\hat{z}_i = k} \boldsymbol{x}_i .$$

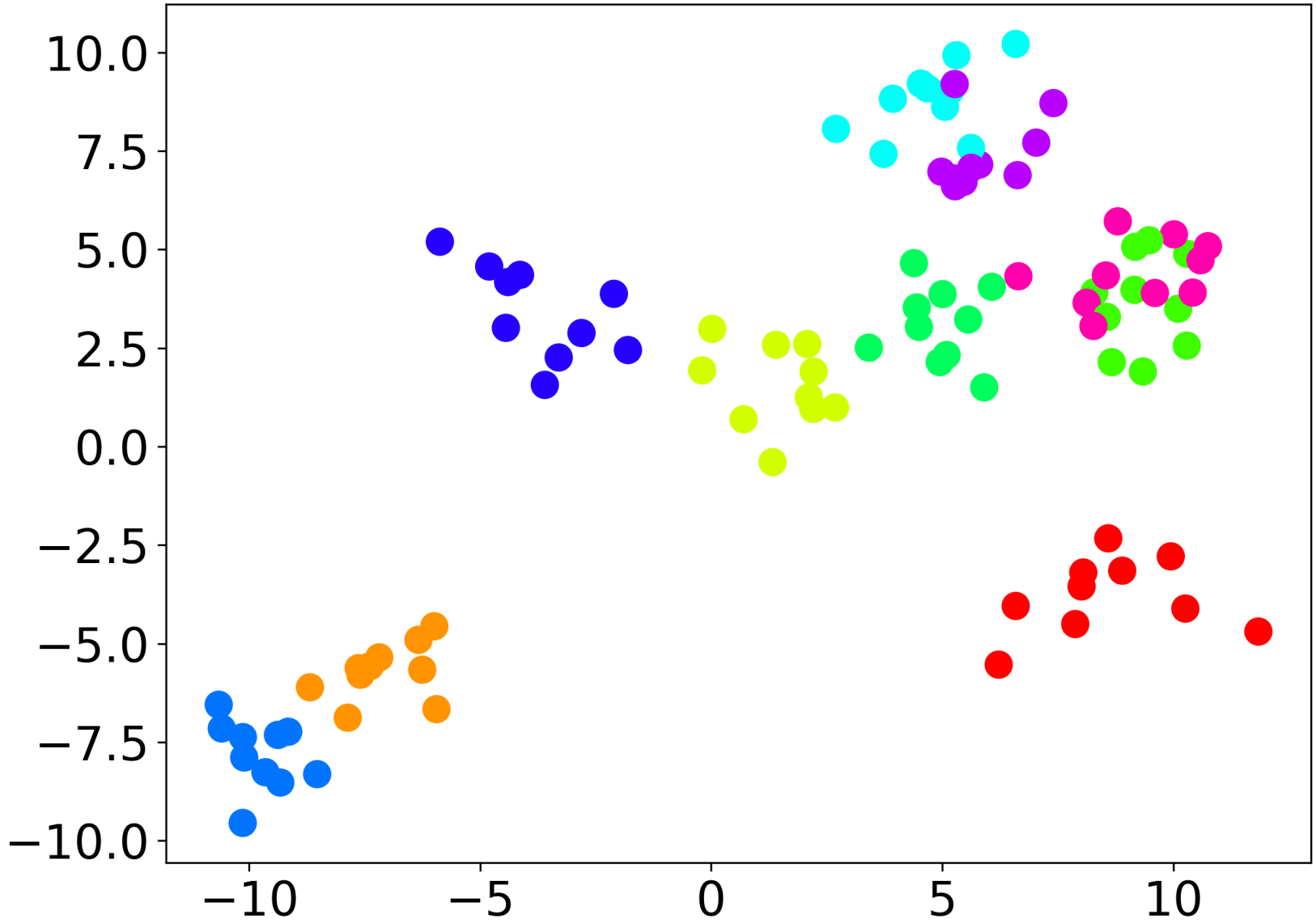
Algoritmo K-médias

- Podemos executar o algoritmo K-médias com diferentes inicializações $\boldsymbol{\mu}_k^{(0)}$ para os centróides e escolher aquela com menor **erro de reconstrução**:

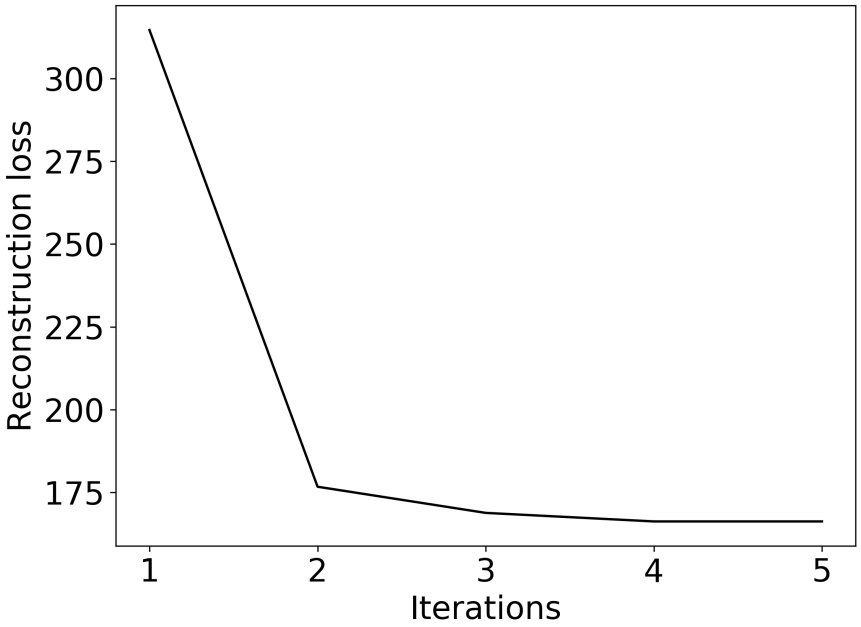
$$E_{\text{rec.}} = \sum_{k=1}^{K} \sum_{i:\hat{z}_i=k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2$$

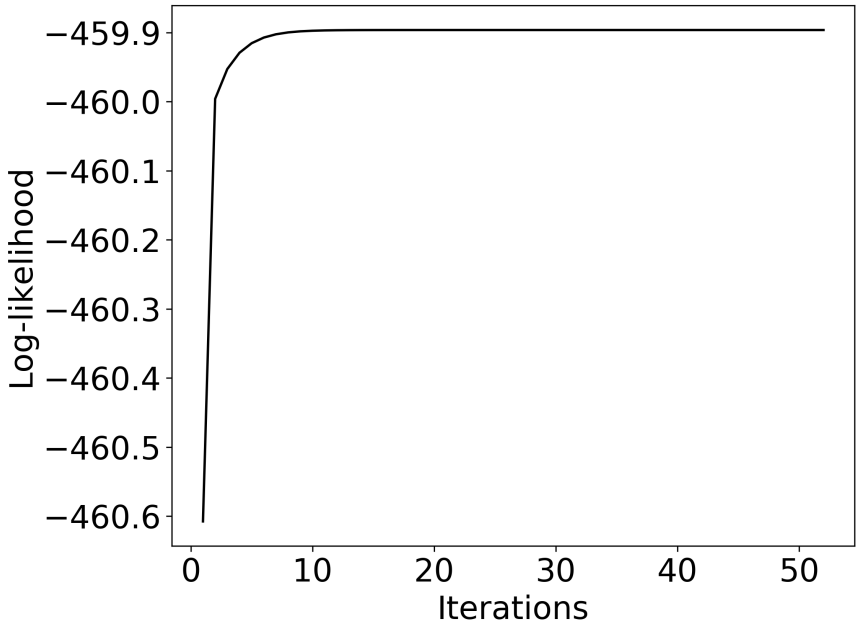
- O K-médias++ é uma heurística eficiente para a inicialização:
  1. Escolha o primeiro centróide aleatoriamente entre os dados.
  2. Repita $K-1$ vezes: escolha o próxima centróide entre os dados com probabilidade proporcional à distância ao seu centróide mais próximo.
- Podemos usar o algoritmo K-médias para inicializar as componentes do GMM:
  - $\rightarrow$ Os vetores de média são iniciados com os centróides.
  - $\rightarrow$ As matrizes de covariância são iniciadas com as matrizes de covariância estimadas dos grupos encontrados.

K-médias e GMM

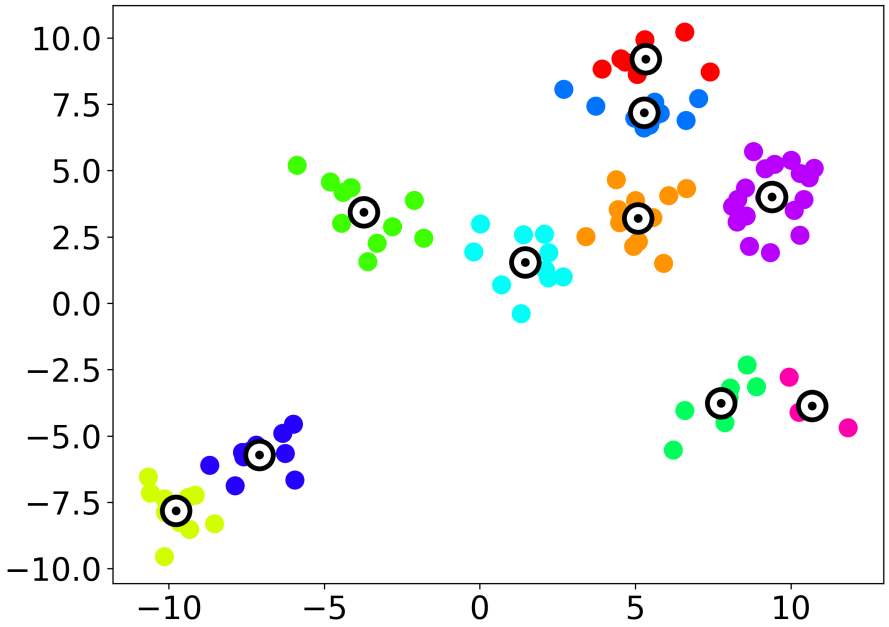


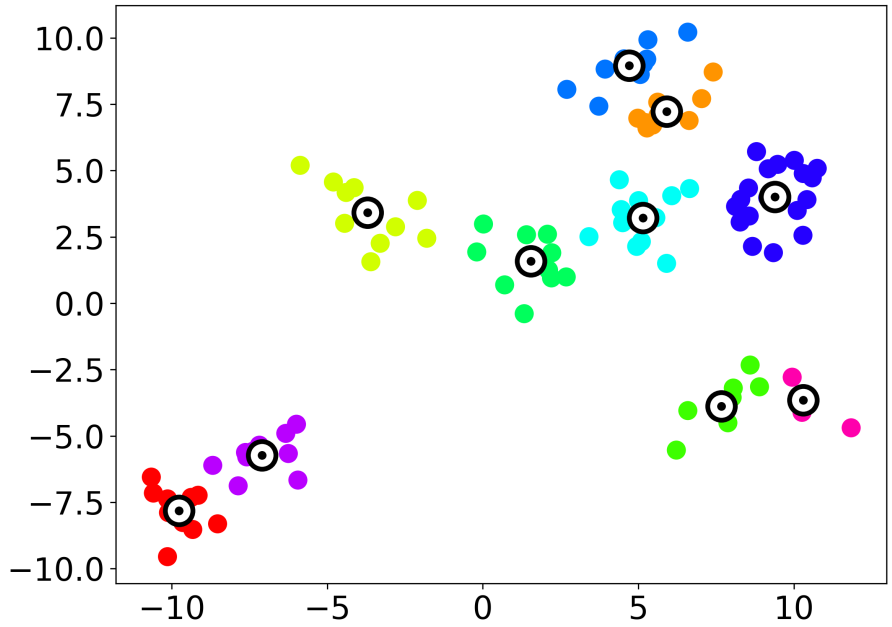
K-médias (esquerda) e GMM (direita)



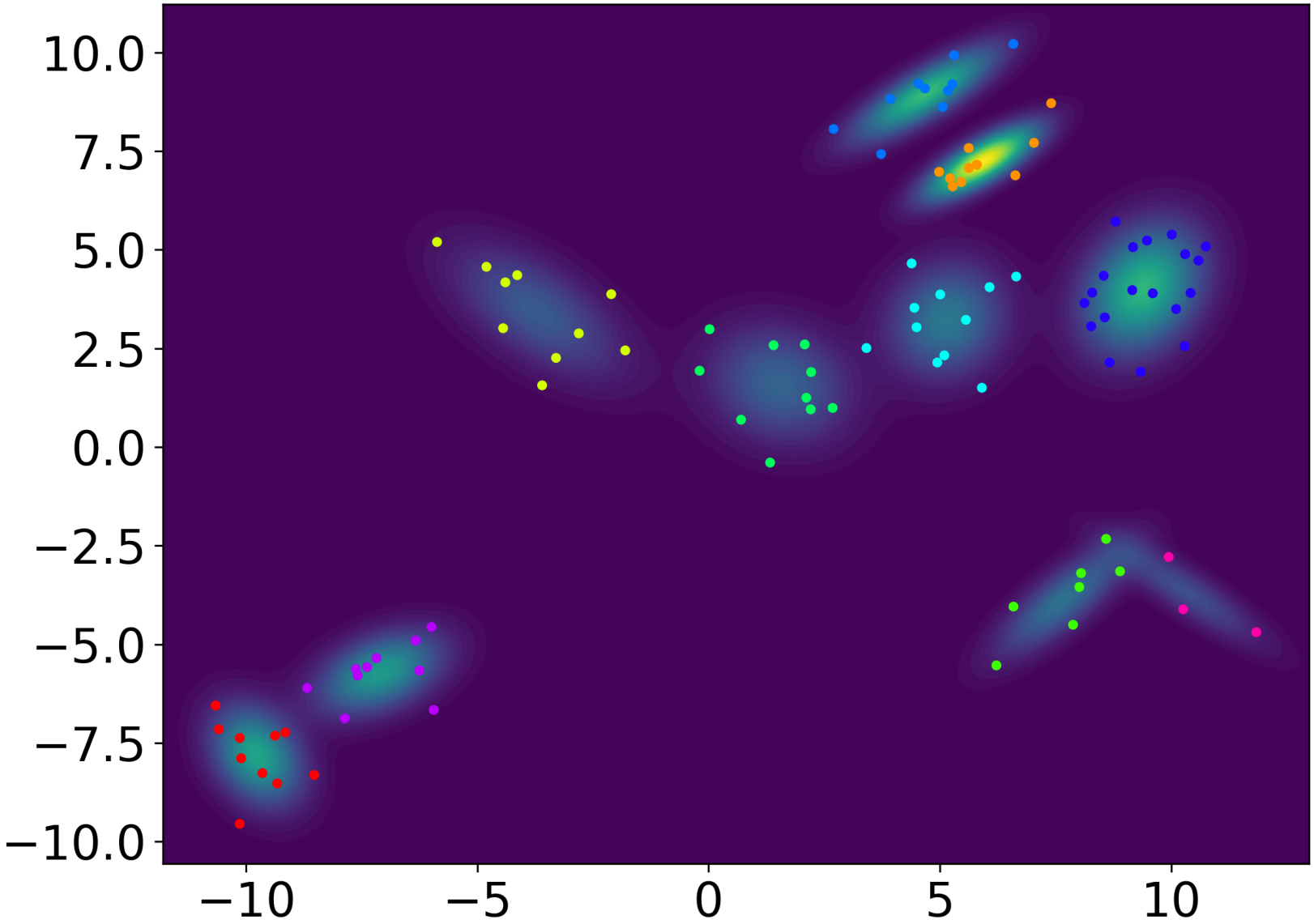


K-médias (esquerda) e GMM (direita)

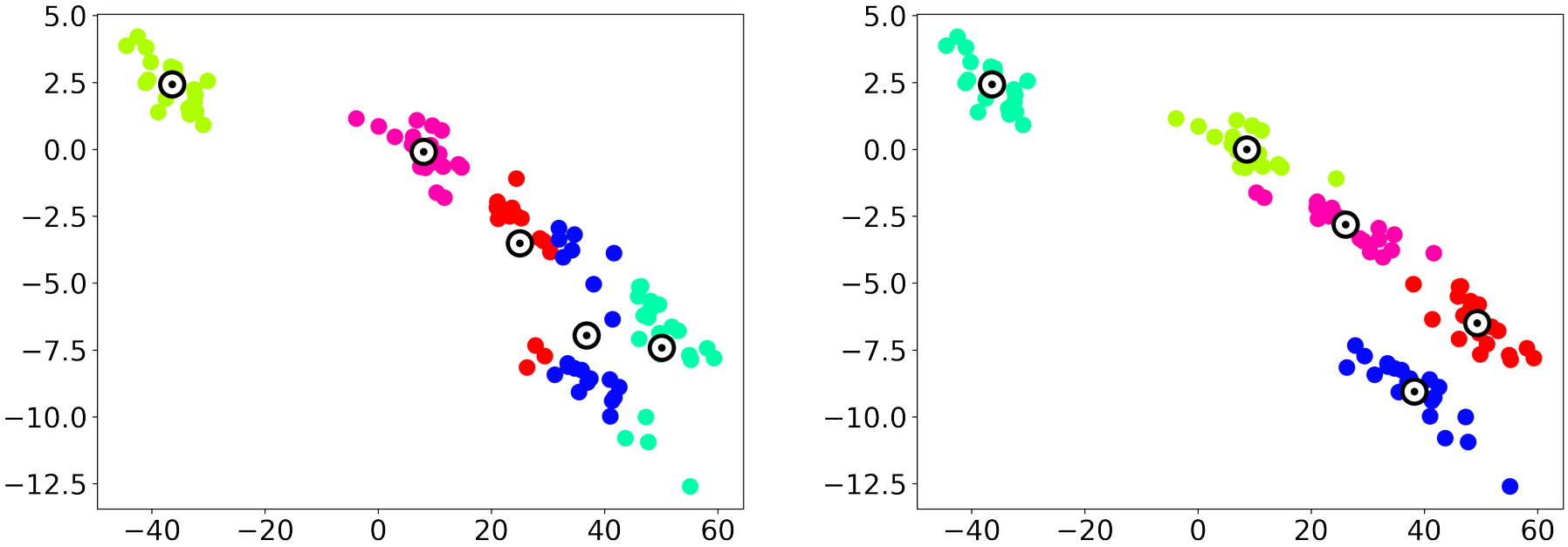




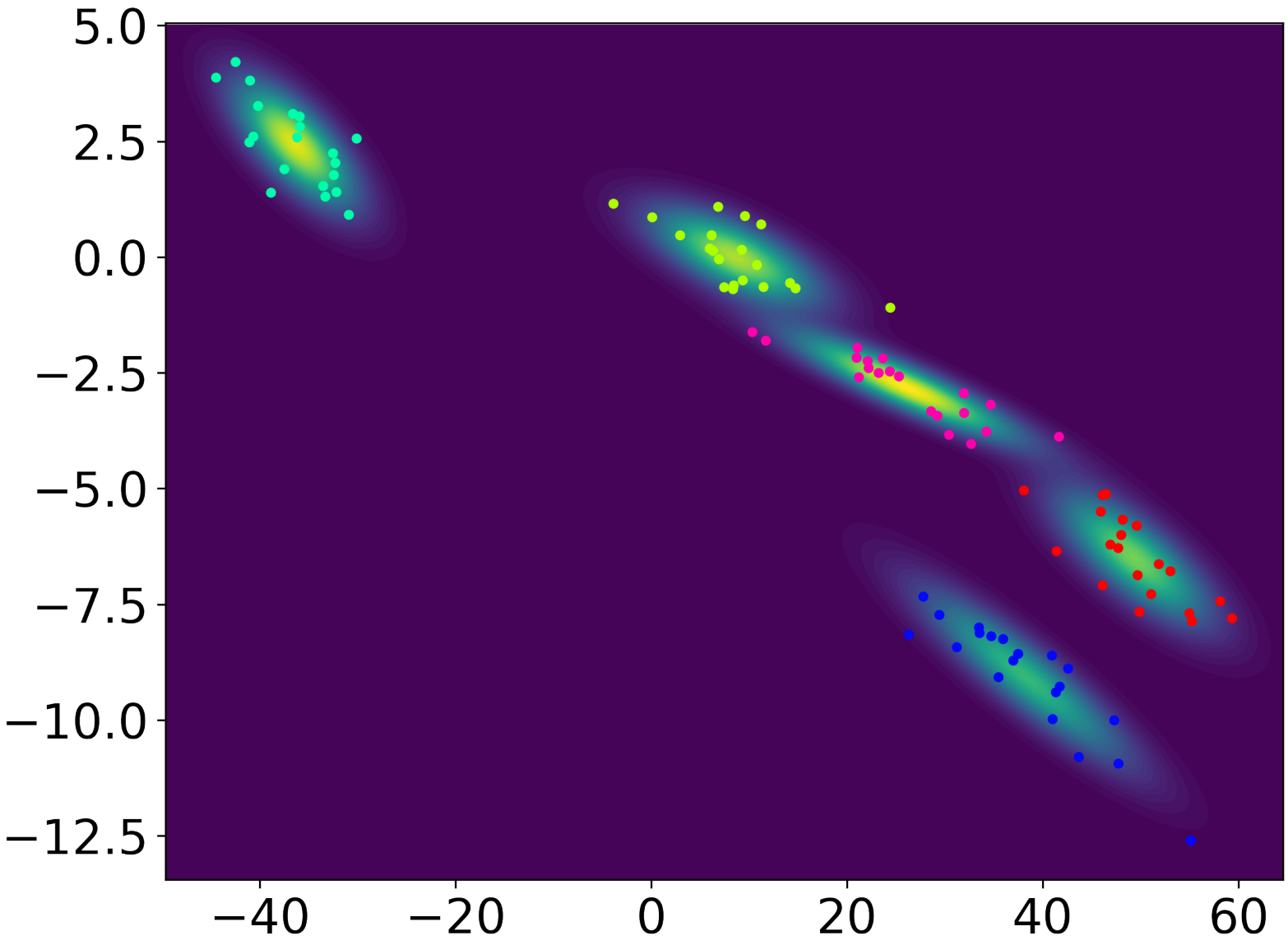
GMM



K-médias (esquerda) e GMM (direita)



GMM



Agenda

Algoritmo EM para GMMs com solução MAP

- Podemos obter uma solução MAP com o algoritmo EM considerando prioris para os parâmetros do modelo.
- A função auxiliar seria dada por:

$$Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik} \log \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$
$$\underbrace{+ \log p(\boldsymbol{\pi}) + \sum_{k=1}^{K} \log p(\boldsymbol{\theta}_k)}_{\text{termos das prioris}} .$$

- Menor chance de ocorrer singularidades na otimização, em que uma Gaussiana é colocada exatamente sobre uma observação.
- Para as probabilidades $\boldsymbol{\pi}$ das componentes da mistura, a distribuição de Dirichlet é conjugada:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}).$$

Algoritmo EM para GMMs com solução MAP

- Para as médias $\boldsymbol{\mu}_k$ e as covariâncias $\boldsymbol{\Sigma}_k$, escolhemos a priori conjugada em que média e covariância são dependentes:

$$\begin{aligned} p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k)p(\boldsymbol{\Sigma}_k) \\ &= \mathrm{NIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{m}_0, \kappa_0, \nu_0, \boldsymbol{S}_0) \\ &\triangleq \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma}_k)\mathrm{IW}(\boldsymbol{\Sigma}_k|\boldsymbol{S}_0, \nu_0), \end{aligned}$$

  onde definimos a distribuição Normal-inverse-Wishart (NIW).

Algoritmo EM para GMMs com solução MAP

- Para as médias $\boldsymbol{\mu}_k$ e as covariâncias $\boldsymbol{\Sigma}_k$, escolhemos a priori conjugada em que média e covariância são dependentes:

$$\begin{aligned} p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k)p(\boldsymbol{\Sigma}_k) \\ &= \mathrm{NIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{m}_0, \kappa_0, \nu_0, \boldsymbol{S}_0) \\ &\triangleq \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma}_k)\mathrm{IW}(\boldsymbol{\Sigma}_k|\boldsymbol{S}_0, \nu_0), \end{aligned}$$

  onde definimos a distribuição Normal-inverse-Wishart (NIW).
- Escolhas usuais para os hiperparâmetros a partir dos dados são $\boldsymbol{S}_0 = \frac{1}{K^{2/D}}\boldsymbol{S}_{\bar{x}}, \nu_0 = D + 2, \boldsymbol{\mu}_0 = \bar{\boldsymbol{x}}, \kappa_0 = 0.01$ (ou $\kappa_0 \to 0$), em que $\bar{\boldsymbol{x}}$ e $\boldsymbol{S}_{\bar{x}}$ são a média e a covariância amostral.
- Somente o passo M do algoritmo EM será modificado!

Algoritmo EM para GMMs com solução MAP

Resumo do algoritmo

1. Escolha $K, \alpha_k, \boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{S}_0, \pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}, \forall k$.
2. Faça $t = 1$ e repita até convergir:
   - Passo E:

$$r_{ik} \triangleq p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}) = \frac{\pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t-1)} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{k'}^{(t-1)}, \boldsymbol{\Sigma}_{k'}^{(t-1)})}.$$

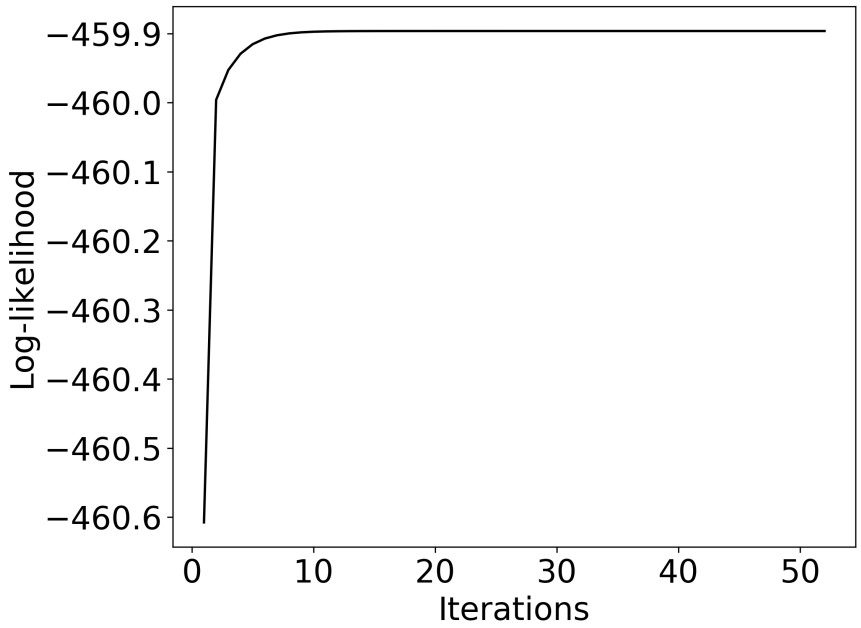
   - Passo M:

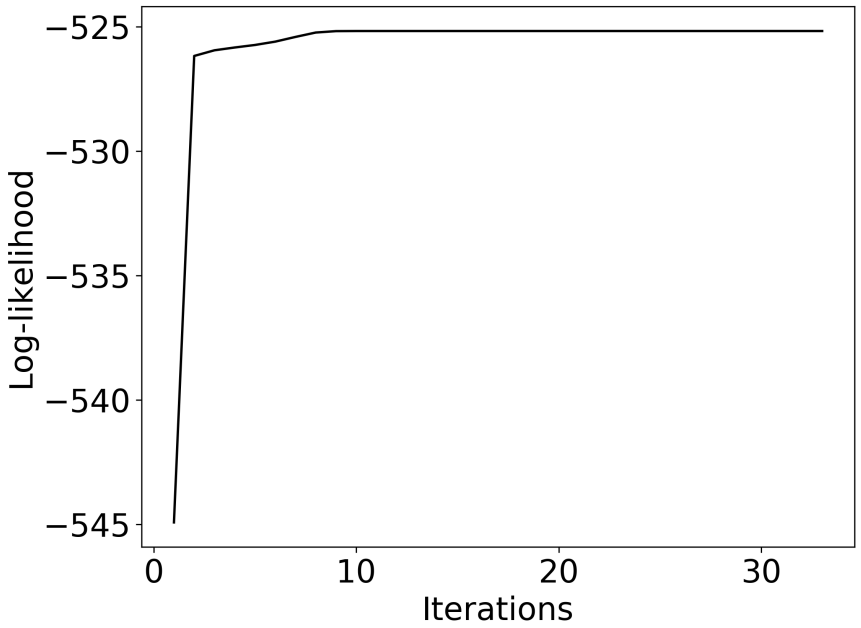
$$\pi_k^{(t)} = \frac{\alpha_k - 1 + \sum_i r_{ik}}{N - K + \sum_k \alpha_k},$$

$$\bar{\boldsymbol{x}}_k \triangleq \frac{\sum_i r_{ik} \boldsymbol{x}_i}{\sum_i r_{ik}}, \quad \boldsymbol{\mu}_k^{(t)} = \frac{\kappa_0 \boldsymbol{m}_0 + \bar{\boldsymbol{x}}_k \sum_i r_{ik}}{\kappa_0 + \sum_i r_{ik}},$$

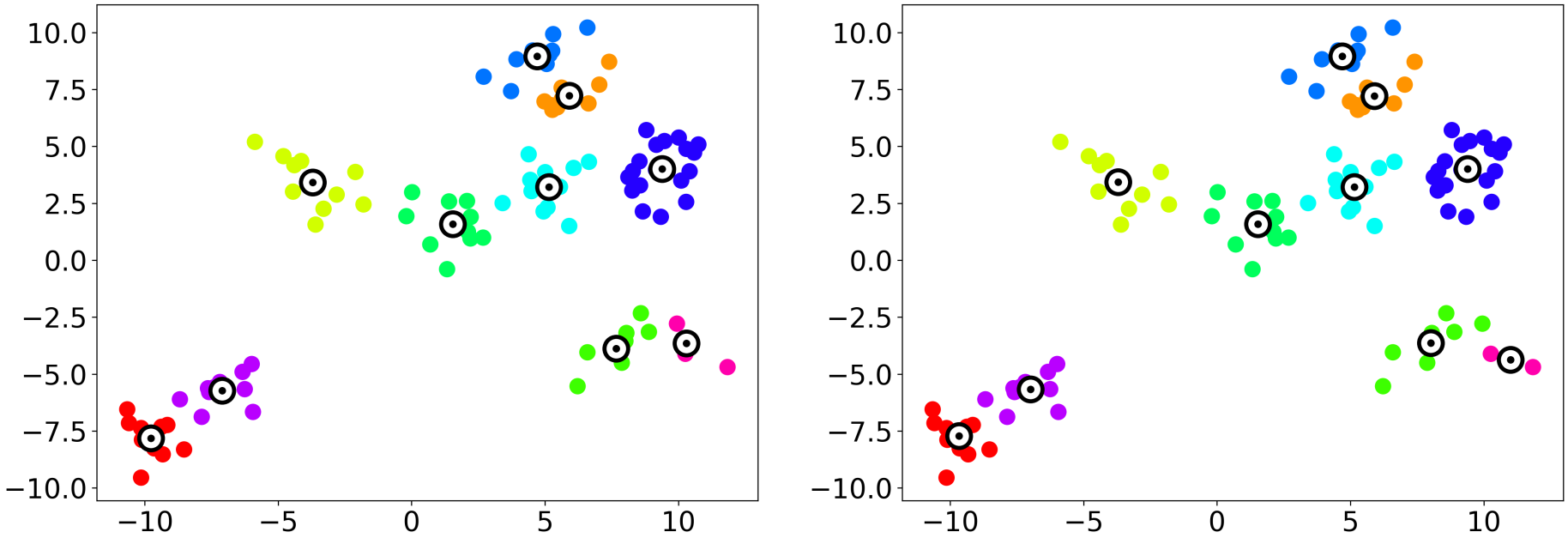
$$\boldsymbol{\Sigma}_k^{(t)} = \frac{\boldsymbol{S}_0 + \sum_i r_{ik} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^\top + \frac{\kappa_0 \sum_i r_{ik}}{\kappa_0 + \sum_i r_{ik}} (\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^\top}{\nu_0 + D + 2 + \sum_i r_{ik}}.$$

GMM-ML (esquerda) e GMM-MAP (direita)

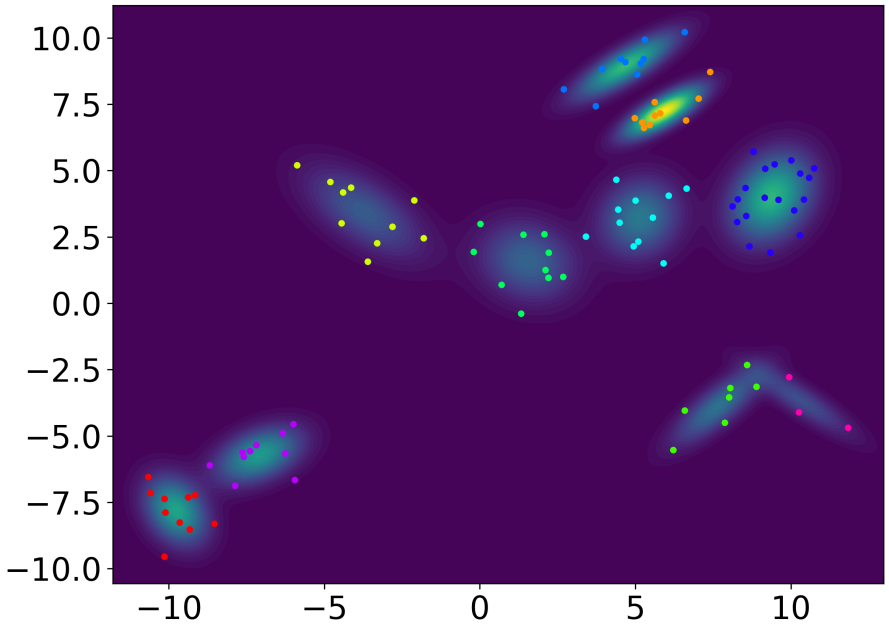


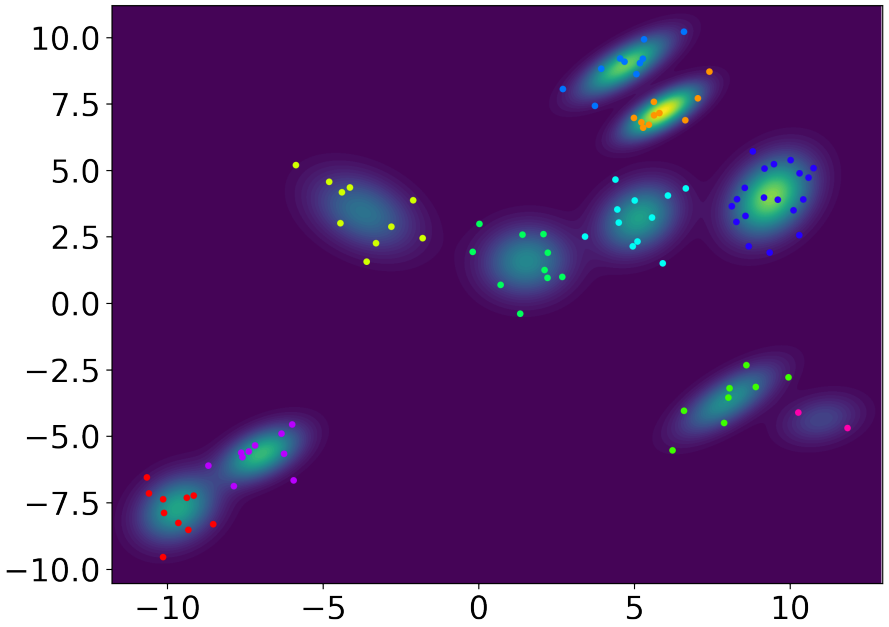


GMM-ML (esquerda) e GMM-MAP (direita)

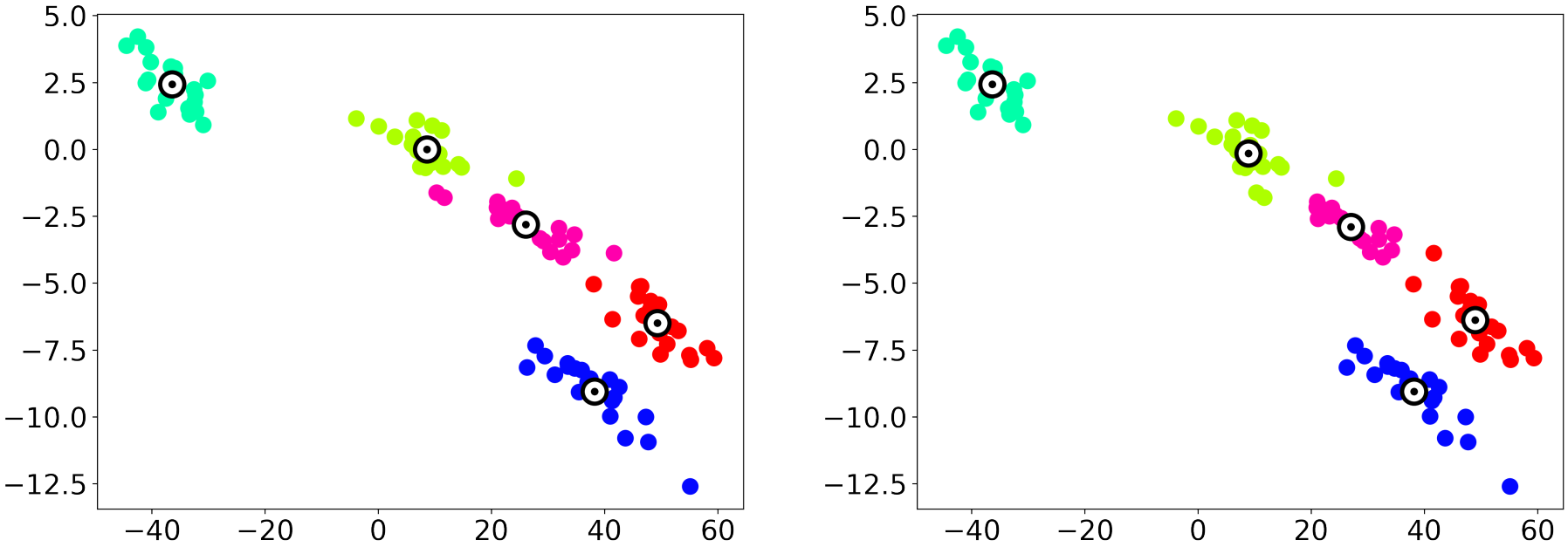


GMM-ML (esquerda) e GMM-MAP (direita)

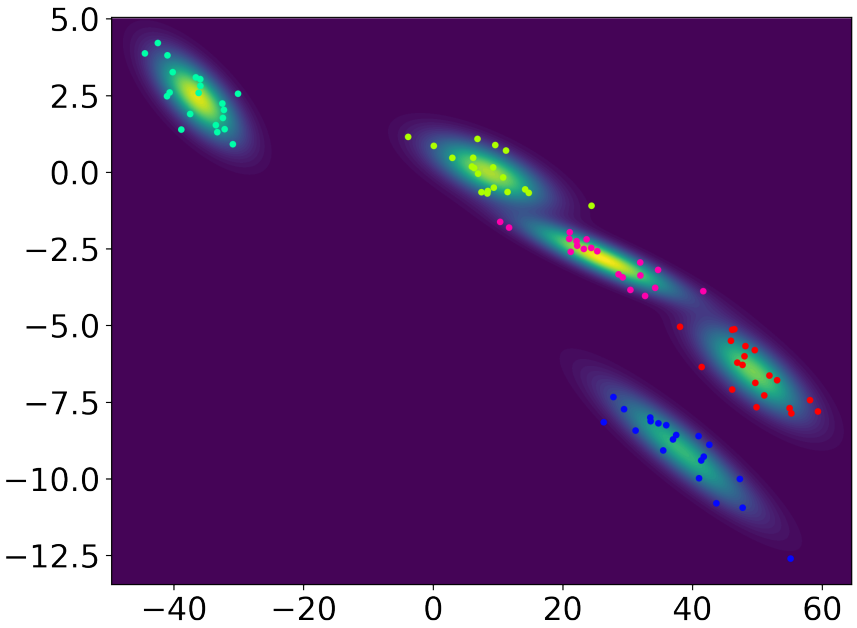


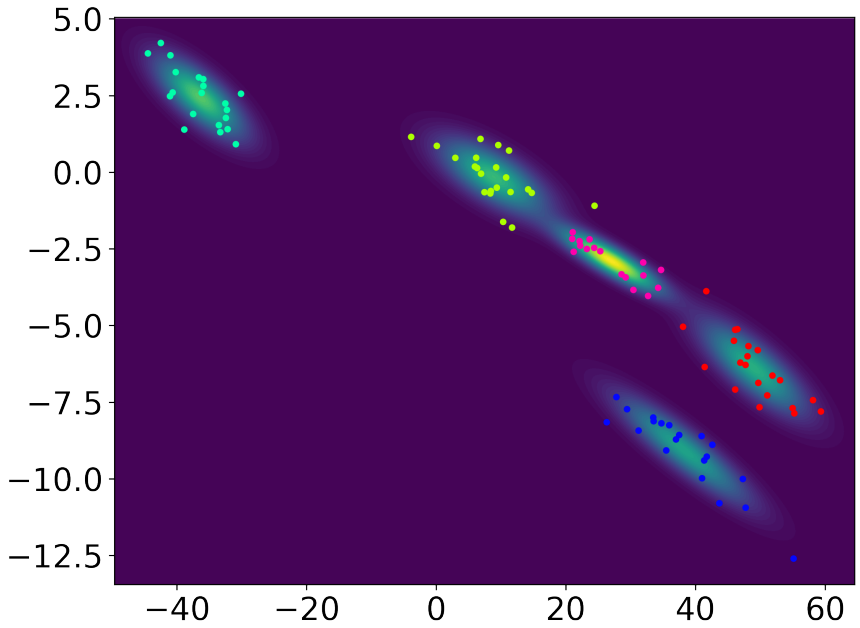


GMM-ML (esquerda) e GMM-MAP (direita)



GMM-ML (esquerda) e GMM-MAP (direita)





Agenda

Seleção de modelos para GMM

- Além dos parâmetros do GMM ou outro modelo de mistura, precisamos escolher o número K de componentes.
- Duas abordagens práticas:
  - $\rightarrow$ Calcular o **BIC (Bayesian Information Criterion)** ou o **AIC (Akaike Information Criterion)** para diferentes valores de K:

$$\mathrm{BIC}(\mathcal{D}|K) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{1}{2} M \log N,$$
$$\mathrm{AIC}(\mathcal{D}|K) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - M,$$

em que M é o número de parâmetros e N é o total de observações.

Seleção de modelos para GMM

- Além dos parâmetros do GMM ou outro modelo de mistura, precisamos escolher o número K de componentes.
- Duas abordagens práticas:
  - $\rightarrow$ Calcular o **BIC (Bayesian Information Criterion)** ou o **AIC (Akaike Information Criterion)** para diferentes valores de K:

$$\mathrm{BIC}(\mathcal{D}|K) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{1}{2} M \log N,$$
$$\mathrm{AIC}(\mathcal{D}|K) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - M,$$

em que M é o número de parâmetros e N é o total de observações.

- Em um GMM, temos
$$M = \underbrace{K-1}_{\pi_k|_{k=1}^K} + \underbrace{K \times D}_{\boldsymbol{\mu}_k|_{k=1}^K} + \underbrace{K \times D \times (D+1)/2}_{\boldsymbol{\Sigma}_k|_{k=1}^K}.$$

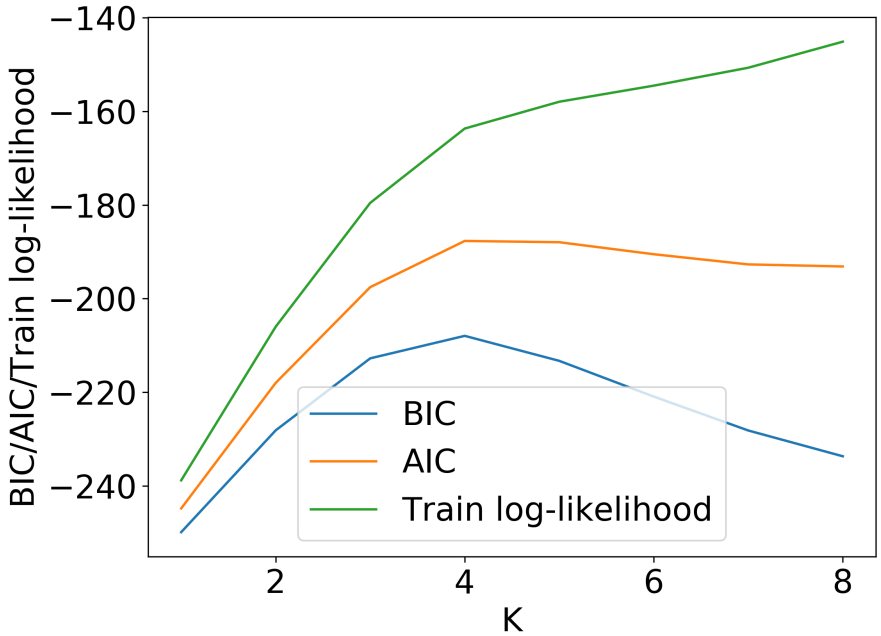
Seleção de modelos para GMM

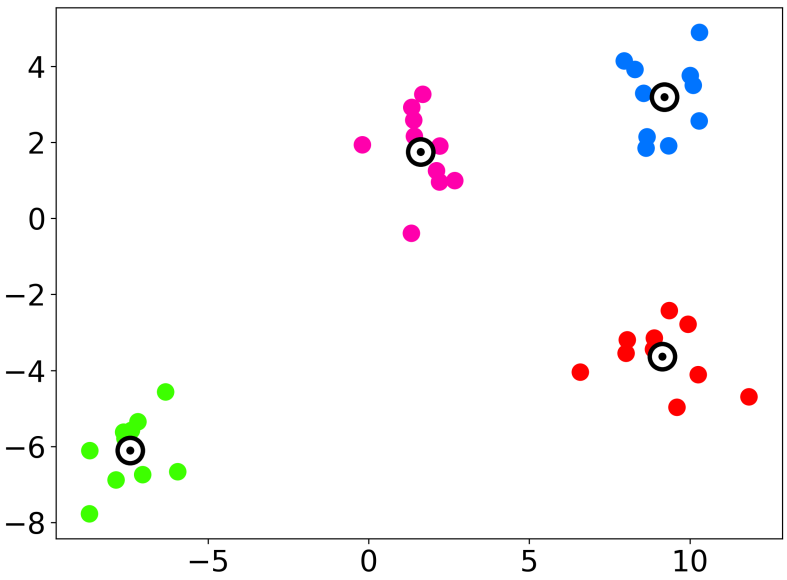
- Além dos parâmetros do GMM ou outro modelo de mistura, precisamos escolher o número K de componentes.
- Duas abordagens práticas:
  - → Calcular o **BIC (Bayesian Information Criterion)** ou o **AIC (Akaike Information Criterion)** para diferentes valores de K:

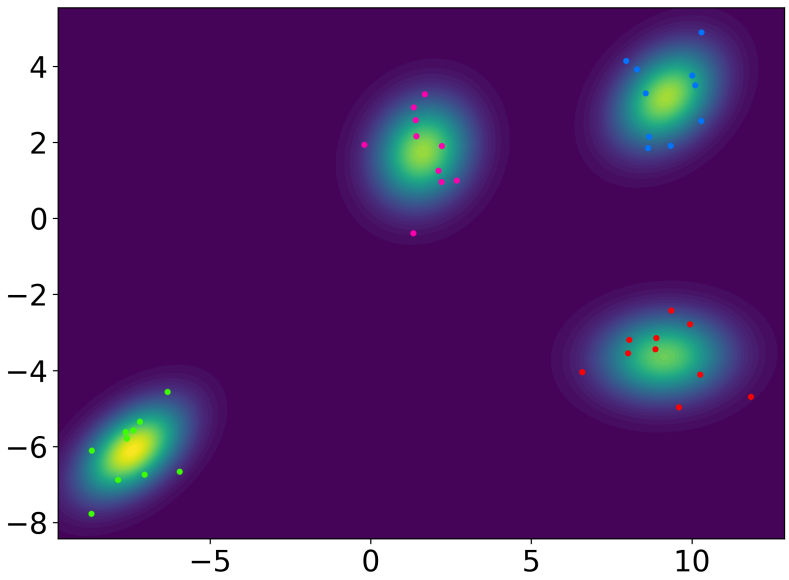
$$\mathrm{BIC}(\mathcal{D}|K) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{1}{2} M \log N,$$
$$\mathrm{AIC}(\mathcal{D}|K) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - M,$$

    em que M é o número de parâmetros e N é o total de observações.
    - Em um GMM, temos
    $$M = \underbrace{K-1}_{\pi_k|_{k=1}^K} + \underbrace{K \times D}_{\boldsymbol{\mu}_k|_{k=1}^K} + \underbrace{K \times D \times (D+1)/2}_{\boldsymbol{\Sigma}_k|_{k=1}^K}.$$
  - → Calcular a log-verossimilhança $\log p(\mathcal{D}|\hat{\boldsymbol{\theta}})$ em **dados não usados no treinamento** para diferentes valores de K.

Seleção de modelos via AIC/BIC







Agenda

Algoritmo EM para Mistura de Especialistas

- No caso de uma mistura de especialistas representados por modelos lineares, temos a seguinte log-verossimilhança para os dados completos:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \log[\pi_{ik} \mathcal{N}(y_i | \boldsymbol{w}_k^\top \boldsymbol{x}_i, \sigma_k^2)],$$

$$\pi_{ik} \triangleq \mathcal{S}(\boldsymbol{V}^\top \boldsymbol{x}_i)_k,$$

$$r_{ik} \propto \pi_{ik}^{(t-1)} \mathcal{N}(y_i | \boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t-1)}, (\sigma_k^{(t-1)})^2).$$

- O passo E (cálculo dos r_{ik}) continua o mesmo, apenas trocando π_k por π_{ik}.

Algoritmo EM para Mistura de Especialistas

- O passo M com relação aos parâmetros $\boldsymbol{w}_k$ e σ_k^2 é dado por:

$$
\begin{aligned}
Q(\boldsymbol{\theta}_k, \boldsymbol{\theta}^{(t-1)}) &= \sum_{i=1}^{N} r_{ik} \log \mathcal{N}(y_i | \boldsymbol{w}_k^\top \boldsymbol{x}_i, \sigma_k^2) \\
&= \sum_{i=1}^{N} r_{ik} \left\{ -\frac{1}{\sigma_k^2} (y_i - \boldsymbol{w}_k^\top \boldsymbol{x}_i)^2 \right\}
\end{aligned}
$$

Algoritmo EM para Mistura de Especialistas

- O passo M com relação aos parâmetros $\boldsymbol{w}_k$ e σ_k^2 é dado por:

$$Q(\boldsymbol{\theta}_k, \boldsymbol{\theta}^{(t-1)}) = \sum_{i=1}^{N} r_{ik} \log \mathcal{N}(y_i | \boldsymbol{w}_k^\top \boldsymbol{x}_i, \sigma_k^2)$$

$$= \sum_{i=1}^{N} r_{ik} \left\{ -\frac{1}{\sigma_k^2} (y_i - \boldsymbol{w}_k^\top \boldsymbol{x}_i)^2 \right\}$$

- A expressão acima corresponde a um problema de mínimos quadrados ponderado, como no IRLS, tendo solução na forma:

$$\boldsymbol{w}_k = (\boldsymbol{X}^\top \boldsymbol{R}_k \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{R}_k \boldsymbol{y},$$
$$\boldsymbol{R}_k = \mathrm{diag}(r_{1k}, r_{2k}, \ldots, r_{Nk}).$$

- Para a variância, temos:

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} r_{ik} (y_i - \boldsymbol{w}_k^\top \boldsymbol{x}_i)^2}{\sum_{i=1}^{N} r_{ik}}$$

Algoritmo EM para Mistura de Especialistas

- No caso do parâmetro $\boldsymbol{V}$ da função softmax, temos a seguinte componente:

$$Q(\boldsymbol{V}, \boldsymbol{\theta}^{(t-1)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \log \pi_{ik}.$$

- Como $\pi_{ik} = \mathcal{S}(\boldsymbol{V}^\top \boldsymbol{x}_i)_k$, a expressão acima corresponde à função custo de um modelo de regressão logística multiclasse.
- A única diferença de uma regressão softmax é o uso de "soft-labels" r_{ik} em vez de saídas 1-hot-encoding.

Algoritmo EM para Mistura de Especialistas

Resumo do algoritmo

1. Escolha K, $\boldsymbol{V}^{(0)}$, $\boldsymbol{w}_k^{(0)}$, $(\sigma_k^2)^{(0)}$, $\forall k$.
2. Faça $t = 1$ e repita até convergir:
   - Passo E:

$$\pi_{ik}^{(t-1)} = \mathcal{S}((\boldsymbol{V}^{(t-1)})^\top \boldsymbol{x}_i)_k,$$

$$r_{ik} = \frac{\pi_{ik}^{(t-1)} \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{k'=1}^{K} \pi_{ik'}^{(t-1)} \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_{k'}^{(t-1)}, \boldsymbol{\Sigma}_{k'}^{(t-1)})}.$$

   - Passo M:

$$\boldsymbol{w}_k^{(t)} = (\boldsymbol{X}^\top \boldsymbol{R}_k \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{R}_k \boldsymbol{y}, \quad \boldsymbol{R}_k = \text{diag}(r_{1k}, \ldots, r_{Nk}),$$

$$(\sigma_k^2)^{(t)} = \frac{\sum_i r_{ik} (y_i - (\boldsymbol{w}_k^{(t)})^\top \boldsymbol{x}_i)^2}{\sum_i r_{ik}}, \quad \boldsymbol{V}^{(t)} \leftarrow \text{“soft” softmax.}$$

Agenda

Algoritmo EM como um limiar inferior

- Reescrevemos a log-verossimilhança dos dados observados de um LVM com variáveis latentes $\boldsymbol{z}_i$ discretas (como uma mistura):

$$\mathcal{L}(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{N} \log p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{\boldsymbol{z}_i} p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) \right].$$

- Considere uma distribuição arbitrária $q(\boldsymbol{z}_i)$:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{q(\boldsymbol{z}_i)} \right].$$

Algoritmo EM como um limiar inferior

- Reescrevemos a log-verossimilhança dos dados observados de um LVM com variáveis latentes $\boldsymbol{z}_i$ discretas (como uma mistura):

$$\mathcal{L}(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{N} \log p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{\boldsymbol{z}_i} p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) \right].$$

- Considere uma distribuição arbitrária $q(\boldsymbol{z}_i)$:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{q(\boldsymbol{z}_i)} \right].$$

- **Problema**: A expressão acima é difícil de trabalhar por causa do log fora do somatório.

Algoritmo EM como um limiar inferior

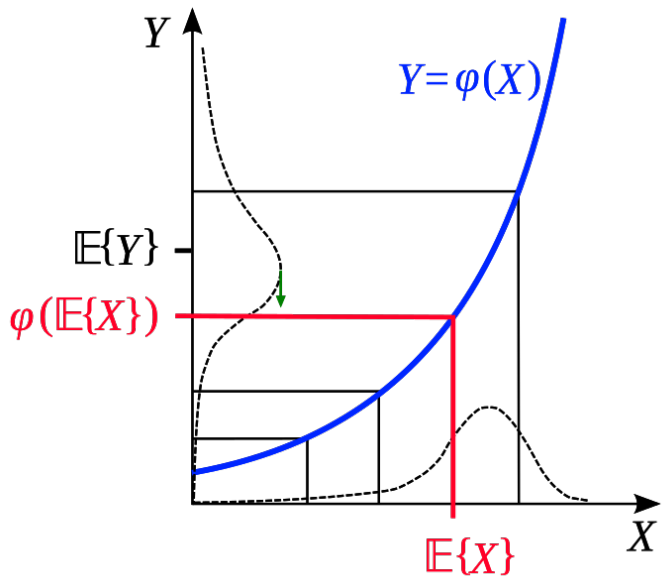
- **Ideia**: Usamos a **desigualdade de Jensen**, definida por:

$$\phi\left(\sum_{i=1}^{N} \lambda_i \boldsymbol{x}_i\right) \leq \sum_{i=1}^{N} \lambda_i \phi(\boldsymbol{x}_i),$$

  para uma função $\phi(\cdot)$ convexa qualquer.

- No contexto de probabilidade, fazemos $\lambda_i = p(\boldsymbol{x}_i)$ e $\sum_i \lambda_i = 1$:

$$\phi(\mathbb{E}\{x\}) \leq \mathbb{E}\{\phi(x)\}).$$



Algoritmo EM como um limiar inferior

- A função $\log(\cdot)$ é côncava, então usamos o inverso da desigualdade de Jensen:

$$\log(\mathbb{E}\{x\}) \geq \mathbb{E}\{\log(x)\}).$$

- Substituindo na definição de $\mathcal{L}(\boldsymbol{\theta})$:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log \left[\sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{q(\boldsymbol{z}_i)} \right] \\
&\geq \sum_{i} \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \log \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{q(\boldsymbol{z}_i)} \\
&\geq \sum_{i} \mathbb{E}_{q(\boldsymbol{z}_i)}[\log p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})] + \mathbb{H}[q(\boldsymbol{z}_i)] \triangleq \mathcal{L}(\boldsymbol{\theta}, q),
\end{aligned}$$

em que $\mathbb{H}[q(\boldsymbol{z}_i)] = -\mathbb{E}_{q(\boldsymbol{z}_i)}[\log q(\boldsymbol{z}_i)]$ é um termo de entropia.

Algoritmo EM como um limiar inferior

- A escolha da distribuição q (agora um "parâmetro") deve maximizar o limiar inferior (lower bound) obtido:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, q) &= \sum_i \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \log \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{q(\boldsymbol{z}_i)} \\
&= \sum_i \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \log \frac{p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) p(\boldsymbol{x}_i|\boldsymbol{\theta})}{q(\boldsymbol{z}_i)} \\
&= \sum_i \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \log \frac{p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta})}{q(\boldsymbol{z}_i)} + \sum_i \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \log p(\boldsymbol{x}_i|\boldsymbol{\theta}) \\
&= -\sum_i \mathrm{KL}(q(\boldsymbol{z}_i) \| p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta})) + \underbrace{\sum_i \log p(\boldsymbol{x}_i|\boldsymbol{\theta})}_{\mathcal{L}(\boldsymbol{\theta})},
\end{aligned}$$

em que usamos a definição da divergência de Kullback-Leibler:

$$\mathrm{KL}(q(\boldsymbol{z}) \| p(\boldsymbol{z})) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}.$$

Algoritmo EM como um limiar inferior

- A divergência de KL é sempre não-negativa e somente igual a zero para distribuições idênticas.
- Teremos $\mathcal{L}(\boldsymbol{\theta}, q) = \mathcal{L}(\boldsymbol{\theta})$ para $q(\boldsymbol{z}_i) = p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$.
- Como os parâmetros $\boldsymbol{\theta}$ são desconhecidos, usamos sua estimativa $\boldsymbol{\theta}^{(t-1)}$ até a iteração anterior.
- Assim, fazemos $q^{(t)}(\boldsymbol{z}_i) = p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$, que é o resultado do **passo E** do algoritmo EM:

$$\mathcal{L}(\boldsymbol{\theta}, q^{(t)}) = \sum_i \mathbb{E}_{q^{(t)}(\boldsymbol{z}_i)}[\log p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})] + \mathbb{H}[q^{(t)}(\boldsymbol{z}_i)].$$

- O primeiro termo de $\mathcal{L}(\boldsymbol{\theta}, q^{(t)})$ é a log-verossimilhança esperada dos dados completos.
- O segundo termo independe de $\boldsymbol{\theta}$, logo o **passo M** torna-se:

$$\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} \sum_i \mathbb{E}_{q^{(t)}(\boldsymbol{z}_i)}[\log p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})].$$

Algoritmo EM como um limiar inferior

- Como usamos $q^{(t)}(\boldsymbol{z}_i) = p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$, o termo KL some:

$$\mathcal{L}(\boldsymbol{\theta}^{(t-1)}, q^{(t)}) = -\sum_i \underbrace{\mathrm{KL}(q^{(t)}(\boldsymbol{z}_i)||p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}))}_{0} + \mathcal{L}(\boldsymbol{\theta}^{(t-1)}),$$

$$\mathcal{L}(\boldsymbol{\theta}^{(t-1)}, q^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t-1)}).$$

- A cada passo E do algoritmo EM o limiar inferior "toca" $\mathcal{L}(\boldsymbol{\theta}^{(t-1)})$, então maximizá-lo equivale a maximizar $\mathcal{L}(\boldsymbol{\theta}^{(t-1)})$.
- Cada atualização do algoritmo EM monotonicamente incrementa a log-verossimilhança dos dados observados:

$$\mathcal{L}(\boldsymbol{\theta}^{(t)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(t)}, q^{(t)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(t-1)}, q^{(t)}) = \mathcal{L}(\boldsymbol{\theta}^{(t-1)}).$$

Agenda

Variantes do algoritmo EM

- Os passos E e M do algoritmo EM são dados por:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \mathbb{E}[\mathcal{L}_c(\boldsymbol{\theta})|\mathcal{D}, \boldsymbol{\theta}^{(t-1)}] = \sum_i \mathbb{E}_{q^{(t)}(\boldsymbol{z}_i)}[\log p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})],$$

$$\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} \left[Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) + \mathbb{H}[q^{(t)}(\boldsymbol{z}_i)] \right],$$

  em que o termo $\mathbb{H}[q^{(t)}(\boldsymbol{z}_i)]$ pode ser ignorado no algoritmo original por não depender de $\boldsymbol{\theta}$.
- Note que para obter o melhor limiar inferior, tivemos que fazer $q^{(t)}(\boldsymbol{z}_i) = p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$.
- No entanto, a distribuição $p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$ pode não ser tratável/analítica e requerer inferência aproximada no passo E.
- Além disso, o passo M pode não admitir solução analítica.

Variantes do algoritmo EM

- **EM variacional**: Escolhemos uma distribuição $q^{(t)}(\boldsymbol{z}_i)$ parametrizada e a otimizamos em conjunto com os outros parâmetros do modelo.

Variantes do algoritmo EM

- **EM variacional**: Escolhemos uma distribuição $q^{(t)}(\boldsymbol{z}_i)$ parametrizada e a otimizamos em conjunto com os outros parâmetros do modelo.
- **Monte Carlo EM (MCEM)**: Geramos amostras via aproximação de Monte Carlo (como Markov Chain Monte Carlo, MCMC) para computar a esperança em relação a $p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$.

Variantes do algoritmo EM

- **EM variacional**: Escolhemos uma distribuição $q^{(t)}(\boldsymbol{z}_i)$ parametrizada e a otimizamos em conjunto com os outros parâmetros do modelo.
- **Monte Carlo EM (MCEM)**: Geramos amostras via aproximação de Monte Carlo (como Markov Chain Monte Carlo, MCMC) para computar a esperança em relação a $p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$.
- **EM estocástico**: Semelhante ao MCEM, mas admitindo amostragens estocásticas de $p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$.

Variantes do algoritmo EM

- **EM variacional**: Escolhemos uma distribuição $q^{(t)}(\boldsymbol{z}_i)$ parametrizada e a otimizamos em conjunto com os outros parâmetros do modelo.
- **Monte Carlo EM (MCEM)**: Geramos amostras via aproximação de Monte Carlo (como Markov Chain Monte Carlo, MCMC) para computar a esperança em relação a $p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$.
- **EM estocástico**: Semelhante ao MCEM, mas admitindo amostragens estocásticas de $p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)})$.
- **EM generalizado**: Quando o passo M não é analítico, podemos fazer atualizações parciais nos parâmetros, por exemplo via gradiente ascendente.

Agenda

Algoritmo EM para dados faltantes

- Considere a situação em que os dados disponíveis $\boldsymbol{X}$ possuem uma parte visível, $\boldsymbol{X}_v = \{x_{id} : O_{id} = 1\})$, e uma parte faltante, $\boldsymbol{X}_h = \{x_{id} : O_{id} = 0\}$, em que $O_{id} = 1$ indica que a dimensão d do padrão i é observada.
- Considerando a hipótese "missing at random" (o dado faltante independe do seu valor, mas depende dos observados), temos:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(\boldsymbol{x}_{iv}, |\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \left[\sum_{\boldsymbol{x}_{ih}} p(\boldsymbol{x}_{iv}, \boldsymbol{x}_{ih} | \boldsymbol{\theta}) \right].\end{aligned}$$

- Como $\mathcal{L}(\boldsymbol{\theta})$ é difícil de otimizar diretamente, podemos usar o algoritmo EM.

Algoritmo EM para dados faltantes

- Considere a tarefa de estimar os parâmetros de uma Gaussiana multivariada com dados faltantes.
- A log-verossimilhança esperada dos dados completos será:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) &= \mathbb{E}\left[\sum_{i=1}^{N} \log \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right] \\
&\propto -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_i \mathbb{E}[(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})] \\
&\propto -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\sum_i \mathbb{E}[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top]) \\
&\propto -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}), \\
\boldsymbol{A} &\triangleq \sum_i \left(\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top] - 2\boldsymbol{\mu}_i\mathbb{E}[\boldsymbol{x}_i]^\top + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top\right).
\end{aligned}
$$

- Para concluir o passo E, precisamos computar $\mathbb{E}[\boldsymbol{x}_i]$ e $\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top]$.

Algoritmo EM para dados faltantes

- Pela propriedade de condicionamento da Gaussiana, temos:

$$\begin{aligned}\boldsymbol{x}_{ih}|\boldsymbol{x}_{iv},\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{m}_i,\boldsymbol{S}_i),\\ \boldsymbol{m}_i &= \boldsymbol{\mu}_h + \boldsymbol{\Sigma}_{hv}\boldsymbol{\Sigma}_v^{-1}(\boldsymbol{x}_{iv}-\boldsymbol{\mu}_v),\\ \boldsymbol{S}_i &= \boldsymbol{\Sigma}_h - \boldsymbol{\Sigma}_{hv}\boldsymbol{\Sigma}_v^{-1}\boldsymbol{\Sigma}_{vh},\end{aligned}$$

em que $\boldsymbol{\mu}^{(t-1)} = [\boldsymbol{\mu}_v, \boldsymbol{\mu}_h]^\top$ e $\boldsymbol{\Sigma}^{(t-1)} = \begin{bmatrix} \boldsymbol{\Sigma}_v & \boldsymbol{\Sigma}_{vh} \\ \boldsymbol{\Sigma}_{hv} & \boldsymbol{\Sigma}_h \end{bmatrix}$.

- Assim:

$$\mathbb{E}[\boldsymbol{x}_i] = [\mathbb{E}[\boldsymbol{x}_{ih}]; \boldsymbol{x}_{iv}] = [\boldsymbol{m}_i; \boldsymbol{x}_{iv}],$$

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top] &= \mathbb{E}\left[\begin{bmatrix}\boldsymbol{x}_{ih}\\ \boldsymbol{x}_{iv}\end{bmatrix}[\boldsymbol{x}_{ih}^\top, \boldsymbol{x}_{iv}^\top]\right]\\ &= \begin{bmatrix}\mathbb{E}[\boldsymbol{x}_{ih}\boldsymbol{x}_{ih}^\top] & \mathbb{E}[\boldsymbol{x}_{ih}]\boldsymbol{x}_{iv}^\top\\ \boldsymbol{x}_{iv}\mathbb{E}[\boldsymbol{x}_{ih}]^\top & \boldsymbol{x}_{iv}\boldsymbol{x}_{iv}^\top\end{bmatrix} = \begin{bmatrix}\boldsymbol{S}_i + \boldsymbol{m}_i\boldsymbol{m}_i^\top & \boldsymbol{m}_i\boldsymbol{x}_{iv}^\top\\ \boldsymbol{x}_{iv}\boldsymbol{m}_i^\top & \boldsymbol{x}_{iv}\boldsymbol{x}_{iv}^\top\end{bmatrix},\end{aligned}$$

em que usamos $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \mathrm{cov}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\top$.

Algoritmo EM para dados faltantes

- No passo M precisamos resolver a equação $\nabla Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \mathbf{0}$.
- Seguimos as regras de atualização usuais para a Gaussiana:

$$\boldsymbol{\mu}^{(t)} = \frac{1}{N} \sum_{i}^{N} \mathbb{E}[\boldsymbol{x}_i],$$

$$\boldsymbol{\Sigma}^{(t)} = \left(\frac{1}{N-1} \sum_{i}^{N} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top] \right) - \boldsymbol{\mu}^{(t)} (\boldsymbol{\mu}^{(t)})^\top.$$

- Note que poderíamos incluir prioris para os parâmetros e obter uma solução MAP.

Tópicos adicionais

- Mistura de Bernoullis.
- Mistura de distribuições t de Student.
- Algoritmo EM incremental.
- Mistura hierárquica de especialistas.

Referências bibliográficas

- **Cap. 11** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Cap. 9** - BISHOP, Christopher M. **Pattern recognition and machine learning**, 2006.