



UNIVERSIDADE  
FEDERAL DO CEARÁ



# Aprendizagem de Máquina Probabilística

César Lincoln Cavalcante Mattos

2024

# Agenda

- 1 Inferência variacional
- 2 Aproximação de mean field
- 3 Inferência variacional para Gaussiana univariada
- 4 Inferência variacional para regressão linear
- 5 Inferência variacional para mistura de Gaussianas
- 6 Tópicos adicionais
- 7 Referências

trabalho

# Inferência aproximada

- Definimos uma **priori**  $p(\mathbf{z})$  para a variável latente.
- Definimos uma **verossimilhança**  $p(\mathbf{x}|\mathbf{z})$ .
- Expressões de interesse (considerando  $\mathbf{z}$  contínuo):

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}, \quad \text{(verossimilhança marginal)}$$

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}, \quad \text{(posteriori)}.$$

# Inferência aproximada

- Definimos uma **priori**  $p(\mathbf{z})$  para a variável latente.
- Definimos uma **verossimilhança**  $p(\mathbf{x}|\mathbf{z})$ .
- Expressões de interesse (considerando  $\mathbf{z}$  contínuo):

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}, \quad \text{(verossimilhança marginal)}$$

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}, \quad \text{(posteriori)}.$$

- A análise é semelhante quando temos parâmetros  $\mathbf{w} \sim p(\mathbf{w})$  em vez (ou além) de  $\mathbf{z}$ .

# Inferência aproximada

- Definimos uma **priori**  $p(\mathbf{z})$  para a variável latente.
- Definimos uma **verossimilhança**  $p(\mathbf{x}|\mathbf{z})$ .
- Expressões de interesse (considerando  $\mathbf{z}$  contínuo):

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}, \quad \text{(verossimilhança marginal)}$$

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}, \quad \text{(posteriori)}.$$

- A análise é semelhante quando temos parâmetros  $\mathbf{w} \sim p(\mathbf{w})$  em vez (ou além) de  $\mathbf{z}$ .
- Soluções analíticas só são possíveis para distribuições da família exponencial (Gaussiana, Gamma, Bernoulli, Categórica...).

# Inferência aproximada

- Definimos uma **priori**  $p(z)$  para a variável latente.
- Definimos uma **verossimilhança**  $p(x|z)$ .
- Expressões de interesse (considerando  $z$  contínuo):

$$p(x) = \int p(x|z)p(z)dz, \quad (\text{verossimilhança marginal})$$

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}, \quad (\text{posteriori}).$$

- A análise é semelhante quando temos parâmetros  $w \sim p(w)$  em vez (ou além) de  $z$ .
- Soluções analíticas só são possíveis para distribuições da família exponencial (Gaussiana, Gamma, Bernoulli, Categórica...).
- Nos demais casos devemos realizar inferência aproximada:
  - **Aproximação de Laplace** (vimos para regressão logística).
  - **Inferência variacional** (veremos agora).
  - **Markov Chain Monte Carlo** (não veremos neste curso).



# Inferência variacional

- Seja um modelo probabilístico em que  $\mathbf{X}$  denota  $N$  observações e  $\mathbf{Z}$  denota  $N$  variáveis latentes.
- **Problema:** Queremos aproximar a posteriori  $p(\mathbf{Z}|\mathbf{X})$  por uma distribuição arbitrária  $q(\mathbf{Z})$ .


# Inferência variacional

- Seja um modelo probabilístico em que  $\mathbf{X}$  denota  $N$  observações e  $\mathbf{Z}$  denota  $N$  variáveis latentes.
- **Problema:** Queremos aproximar a posteriori  $p(\mathbf{Z}|\mathbf{X})$  por uma distribuição arbitrária  $q(\mathbf{Z})$ .
- **Ideia:** Podemos quantificar a qualidade da aproximação pela divergência de Kullback-Leibler:

$$\begin{aligned}\text{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X})) &= \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} d\mathbf{Z} \\ &= \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{X}) d\mathbf{Z} \\ &= \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{Z}|\mathbf{X})] \\ &= \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})} \left[ \log \frac{p(\mathbf{X}, \mathbf{Z})}{p(\mathbf{X})} \right] \\ &= \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z})] + \log p(\mathbf{X}).\end{aligned}$$

**Entropia** **Igual no EM**

Boys





## Inferência variacional

- A divergência KL é sempre não-negativa, sendo igual a zero somente quando as distribuições são idênticas. Assim:

$$\text{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X})) \geq 0$$

$$\mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z})] + \log p(\mathbf{X}) \geq 0$$

$$\underbrace{\log p(\mathbf{X})}_{\text{evidência}} \geq \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z})] - \underbrace{\mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})]}_{\text{entropia}} = \underbrace{\mathcal{L}(q(\mathbf{Z}))}_{\text{ELBO}},$$

em que  $\mathcal{L}(q(\mathbf{Z}))$  é o *evidence lower bound* (ELBO).

- Note que mesmo não-analítica, a evidência  $\log p(\mathbf{X})$  do modelo pode ser maximizada ao maximizarmos o ELBO.

# Inferência variacional

- A divergência KL é sempre não-negativa, sendo igual a zero somente quando as distribuições são idênticas. Assim:

$$\text{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X})) \geq 0$$

$$\mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z})] + \log p(\mathbf{X}) \geq 0$$

$$\underbrace{\log p(\mathbf{X})}_{\text{evidência}} \geq \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z})] - \underbrace{\mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})]}_{\text{entropia}} = \underbrace{\mathcal{L}(q(\mathbf{Z}))}_{\text{ELBO}},$$

**Conjunta Esperada**

em que  $\mathcal{L}(q(\mathbf{Z}))$  é o *evidence lower bound* (ELBO).

- Note que mesmo não-analítica, a evidência  $\log p(\mathbf{X})$  do modelo pode ser maximizada ao maximizarmos o ELBO.
- Alternativamente, o ELBO pode ser escrito como:

$$\begin{aligned}\mathcal{L}(q(\mathbf{Z})) &= \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] \\ &= \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}|\mathbf{Z})] + \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] \\ &= \underbrace{\mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}|\mathbf{Z})]}_{\text{ajuste às observações}} - \underbrace{\text{KL}(q(\mathbf{Z})\|p(\mathbf{Z}))}_{\text{regularização pela priori}}.\end{aligned}$$

# Inferência variacional

- O ELBO depende da distribuição  $q(\mathbf{Z})$ , chamada de **distribuição variacional**, que deve ter uma forma que facilite os cálculos.
- Os parâmetros que definem  $q(\mathbf{Z})$  são **parâmetros variacionais**, pois não fazem parte do modelo, mas do algoritmo de inferência.
- Note que a abordagem variacional converte o problema de inferência não-analítica em um **problema de otimização**.

# Inferência variacional

- O ELBO depende da distribuição  $q(\mathbf{Z})$ , chamada de **distribuição variacional**, que deve ter uma forma que facilite os cálculos.
- Os parâmetros que definem  $q(\mathbf{Z})$  são **parâmetros variacionais**, pois não fazem parte do modelo, mas do algoritmo de inferência.
- Note que a abordagem variacional converte o problema de inferência não-analítica em um **problema de otimização**.
- Na Física, o ELBO possui uma relação com a energia em um sistema:

$$\begin{aligned}\mathcal{L}(q(\mathbf{Z})) &= \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})], \\ \text{VFE} &= -\mathcal{L}(q(\mathbf{Z})) \\ \text{VFE} &= \underbrace{\mathbb{E}_{q(\mathbf{Z})}[-\log p(\mathbf{X}, \mathbf{Z})]}_{\text{expected energy}} - \underbrace{\mathcal{H}_{q(\mathbf{Z})}}_{\text{entropia}},\end{aligned}$$

em que VFE é a chamada *variational free energy*.

# Inferência variacional

- Escolhemos minimizar o **KL reverso**:

$$\text{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X})) = \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} d\mathbf{Z}.$$

- Mas poderíamos ter escolhido minimizar o **KL direto**:

$$\text{KL}(p(\mathbf{Z}|\mathbf{X})\|q(\mathbf{Z})) = \int p(\mathbf{Z}|\mathbf{X}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} d\mathbf{Z}.$$

- Qual a diferença?

# Inferência variacional

- Escolhemos minimizar o **KL reverso**:

$$\text{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X})) = \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} d\mathbf{Z}.$$

- Mas poderíamos ter escolhido minimizar o **KL direto**:

$$\text{KL}(p(\mathbf{Z}|\mathbf{X})\|q(\mathbf{Z})) = \int p(\mathbf{Z}|\mathbf{X}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} d\mathbf{Z}.$$

- Qual a diferença?
- Onde  $p(\mathbf{Z}|\mathbf{X}) \rightarrow 0$ , o KL reverso só é definido para  $q(\mathbf{Z}) \rightarrow 0$ .
  - Queremos que  $q(\mathbf{Z})$  seja próximo de zero nessas regiões;
  - Logo, o **KL reverso força zeros** (*zero forcing*).

# Inferência variacional

- Escolhemos minimizar o **KL reverso**:

$$\text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z} | \mathbf{X})) = \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z} | \mathbf{X})} d\mathbf{Z}.$$

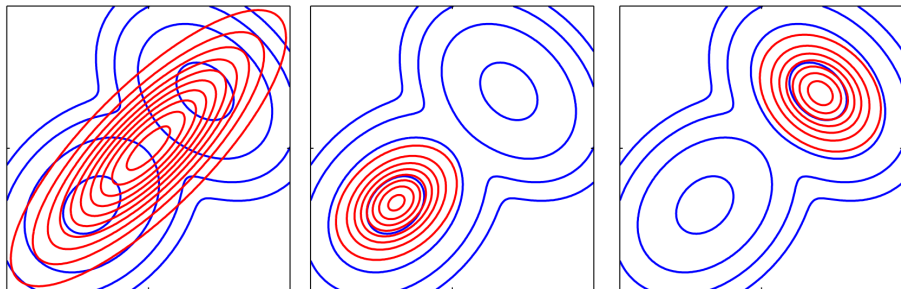
- Mas poderíamos ter escolhido minimizar o **KL direto**:

$$\text{KL}(p(\mathbf{Z} | \mathbf{X}) \| q(\mathbf{Z})) = \int p(\mathbf{Z} | \mathbf{X}) \log \frac{p(\mathbf{Z} | \mathbf{X})}{q(\mathbf{Z})} d\mathbf{Z}.$$

- Qual a diferença?
- Onde  $p(\mathbf{Z} | \mathbf{X}) \rightarrow 0$ , o KL reverso só é definido para  $q(\mathbf{Z}) \rightarrow 0$ .
  - Queremos que  $q(\mathbf{Z})$  seja próximo de zero nessas regiões;
  - Logo, o **KL reverso força zeros** (*zero forcing*).
- Onde  $p(\mathbf{Z} | \mathbf{X}) > 0$ , o KL direto só é definido para  $q(\mathbf{Z}) > 0$ .
  - Queremos que  $q(\mathbf{Z})$  seja maior que zero nessas regiões;
  - Logo, o **KL direto evita zeros** (*zero avoiding*).

# Inferência variacional

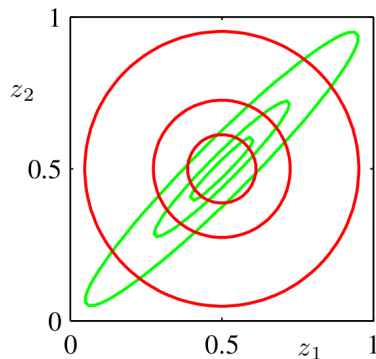
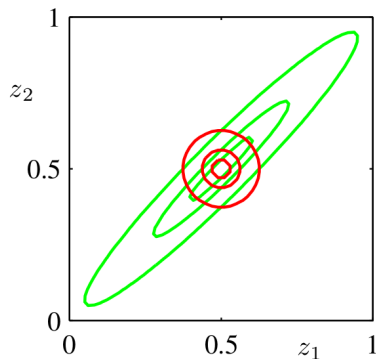
- No **KL reverso** a aproximação  $q$  tende a **subestimar** o suporte da distribuição original, se ajustando a uma de suas modas.
- No **KL direto**,  $q$  **superestima** o suporte original, buscando cobrir todas as modas.
- Nos exemplos abaixo,  $p$  (em azul) é uma distribuição bimodal e  $q$  (em vermelho) é uma Gaussiana. O primeiro cenário usou o KL direto, nos outros foi usado o KL reverso.





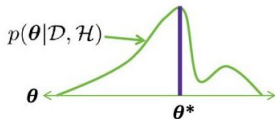
# Inferência variacional

- Considere agora uma distribuição  $p$  (em verde) Gaussiana alongada aproximada por uma distribuição  $q$  (em vermelho) dada pelo produto de duas Gaussianas univariadas.
- No cenário da esquerda, usou-se o KL reverso, no da direita, usou-se o KL direto.



# Diferentes métodos de inferência aproximada

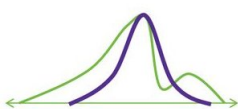
Tenta não colocar  
massa onde não tem



MAP (maximum *a posteriori*)

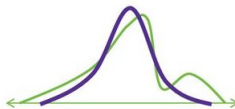


Mean field / Variational Bayes (VB)



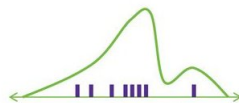
Laplace

Gaussiana  
centrada na moda



Expectation Propagation  
(EP)

Tenta colocar massa  
na maior região possível



Markov chain  
Monte Carlo  
(MCMC)

# Agenda

- 1 Inferência variacional
- 2 Aproximação de mean field
- 3 Inferência variacional para Gaussiana univariada
- 4 Inferência variacional para regressão linear
- 5 Inferência variacional para mistura de Gaussianas
- 6 Tópicos adicionais
- 7 Referências

# Aproximação de mean field

- O procedimento de inferência variacional em si pode ser exato.
- Quando escolhemos uma forma específica, mas restrita, para a distribuição variacional  $q(\mathbf{Z})$  o método torna-se aproximado.

# Aproximação de mean field

- O procedimento de inferência variacional em si pode ser exato.
- Quando escolhemos uma forma específica, mas restrita, para a distribuição variacional  $q(\mathbf{Z})$  o método torna-se aproximado.
- Uma das aproximações mais comuns é a chamada *mean field*, em que a distribuição variacional é completamente fatorada:

$$q(\mathbf{Z}) = \prod_{i=1}^N q_i(\mathbf{z}_i).$$

# Aproximação de mean field

- O procedimento de inferência variacional em si pode ser exato.
- Quando escolhemos uma forma específica, mas restrita, para a distribuição variacional  $q(\mathbf{Z})$  o método torna-se aproximado.
- Uma das aproximações mais comuns é a chamada *mean field*, em que a distribuição variacional é **completamente fatorada**:

$$q(\mathbf{Z}) = \prod_{i=1}^N q_i(\mathbf{z}_i).$$

- Alternativamente, seguiríamos uma aproximação de *mean field* estruturada, em que algumas correlações são mantidas, por grupos:

$$q(\mathbf{Z}) = \prod_{g=1}^G q_g(\mathbf{Z}_g),$$

em que  $\mathbf{Z}_g$  denota as variáveis latentes do subgrupo  $g$ .

# Aproximação de mean field

- Durante a inferência variacional desejamos minimizar  $\text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z} | \mathbf{X}))$ .
- Como vimos, isso envolve maximizar o ELBO. Considerando, uma aproximação de mean field, temos (denotando  $q_i \triangleq q_i(\mathbf{z}_i)$ ):

elbo

$$\begin{aligned}\mathcal{L}(q(\mathbf{Z})) &= \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] \\ &= \int \prod_i q_i \left[ \log p(\mathbf{X}, \mathbf{Z}) - \sum_k \log q_k \right] d\mathbf{Z} \\ &= \int \prod_i q_i \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int \prod_i q_i \sum_k \log q_k d\mathbf{Z}.\end{aligned}$$

Real

# Aproximação de mean field

- Escolhemos uma componente  $q_j$  e a isolamos das demais, ignorando os termos que não a contêm.
- Denotando  $\mathcal{L} \triangleq \mathcal{L}(q(\mathbf{Z}))$ , temos:

$$\begin{aligned}\mathcal{L} &= \int \prod_i q_i \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int \prod_i q_i \sum_k \log q_k d\mathbf{Z} \\&= \int q_j \int \prod_{i \neq j} q_i \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q_j \int \prod_{i \neq j} q_i \left[ \log q_j + \sum_{k \neq j} \log q_k \right] d\mathbf{Z} \\&= \int q_j \mathbb{E}_{i \neq j} [\log p(\mathbf{X}, \mathbf{Z})] dz_j - \int q_j \log q_j dz_j + \text{const.},\end{aligned}$$

em que  $\mathbb{E}_{i \neq j}$  denota esperança com relação a todos os termos  $q(z_i)$ ,  $i \neq j$ , e const. reúne os termos independentes de  $z_j$ .



# Aproximação de mean field

- Reorganizamos os termos para obter uma divergência de KL:

$$\begin{aligned}\mathcal{L} &= \int q_j \underbrace{\mathbb{E}_{i \neq j} [\log p(\mathbf{X}, \mathbf{Z})]}_{\tilde{u}} d\mathbf{z}_j - \int q_j \log q_j d\mathbf{z}_j + \text{const.} \\ &= \mathbb{E}_{q_j} [\tilde{u}] - \mathbb{E}_{q_j} [\log q_j] + \text{const.} \\ &= \mathbb{E}_{q_j} [\log \exp(\tilde{u} + \text{const.})] - \mathbb{E}_{q_j} [\log q_j] + \text{const.} \\ &= -\text{KL}(q_j \parallel \exp(\mathbb{E}_{i \neq j} [\log p(\mathbf{X}, \mathbf{Z})] + \text{const.})) + \text{const.},\end{aligned}$$

em que  $\tilde{u}$  indica uma distribuição não normalizada.

# Aproximação de mean field

- Reorganizamos os termos para obter uma divergência de KL:

$$\begin{aligned}\mathcal{L} &= \int q_j \underbrace{\mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z})]}_{\tilde{u}} d\mathbf{z}_j - \int q_j \log q_j d\mathbf{z}_j + \text{const.} \\ &= \mathbb{E}_{q_j}[\tilde{u}] - \mathbb{E}_{q_j}[\log q_j] + \text{const.} \\ &= \mathbb{E}_{q_j}[\log \exp(\tilde{u} + \text{const.})] - \mathbb{E}_{q_j}[\log q_j] + \text{const.} \\ &= -\text{KL}(q_j \parallel \exp(\mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z})] + \text{const.})) + \text{const.},\end{aligned}$$

em que  $\tilde{u}$  indica uma distribuição não normalizada.

- Maximizamos  $\mathcal{L}$  ao minimizarmos o termo KL, ou seja:

$$\log q_j = \mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

# Aproximação de mean field

## Inferência variacional via aproximação de mean field

- ① **Inicialize** os parâmetros variacionais de  $q_i^{(0)}, \forall i$ ;
- ② Repita até convergir:
  - **Mantenha constante** as componentes  $q_{i \neq j}$  e atualize  $q_j$ :

$$\log q_j^{(t+1)} = \mathbb{E}_{i \neq j} [\log p(\mathbf{X}, \mathbf{Z})] + \text{const.}, \quad \forall j.$$

# Aproximação de mean field

## Inferência variacional via aproximação de mean field

- 1 Inicialize os parâmetros variacionais de  $q_i^{(0)}, \forall i$ ;
- 2 Repita até convergir:
  - Mantenha constante as componentes  $q_{i \neq j}$  e atualize  $q_j$ :

$$\log q_j^{(t+1)} = \mathbb{E}_{i \neq j} [\log p(\mathbf{X}, \mathbf{Z})] + \text{const.}, \quad \forall j.$$

- Note que as formas das distribuições  $q_i$  não foram fixadas, sendo determinadas a partir do procedimento de inferência. Essa é a chamada abordagem variacional de “**forma livre**” (*free form*).

# Aproximação de mean field

## Inferência variacional via aproximação de mean field

- 1 Inicialize os parâmetros variacionais de  $q_i^{(0)}, \forall i$ ;
- 2 Repita até convergir:
  - Mantenha constante as componentes  $q_{i \neq j}$  e atualize  $q_j$ :

$$\log q_j^{(t+1)} = \mathbb{E}_{i \neq j} [\log p(\mathbf{X}, \mathbf{Z})] + \text{const.}, \quad \forall j.$$

- Note que as formas das distribuições  $q_i$  não foram fixadas, sendo determinadas a partir do procedimento de inferência. Essa é a chamada abordagem variacional de “**forma livre**” (*free form*).
- O valor do ELBO não precisa ser computado diretamente, mas avaliá-lo pode ser útil para:
  - Verificar a convergência da otimização;
  - Verificar se a implementação está correta, pois o ELBO deve ser monotonicamente crescente ao longo das iterações;
  - Realizar seleção de modelos, usando o ELBO como uma aproximação da evidência.

# Variational Bayes

- O procedimento visto para variáveis latentes é chamado de *mean field variational EM*:

$$p(\mathbf{Z}|\mathcal{D}) \approx q(\mathbf{Z}) = \prod_{i=1}^N q_i(\mathbf{z}_i).$$

# Variational Bayes

- O procedimento visto para variáveis latentes é chamado de *mean field variational EM*:

$$p(\mathbf{Z}|\mathcal{D}) \approx q(\mathbf{Z}) = \prod_{i=1}^N q_i(\mathbf{z}_i).$$

- Para modelos sem variáveis latentes  $\mathbf{Z}$  mas com parâmetros  $\boldsymbol{\theta}$ , temos o chamado *mean field variational Bayes*:

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx q(\boldsymbol{\theta}) = \prod_{k=1}^K q_k(\boldsymbol{\theta}_k).$$

# Variational Bayes

- O procedimento visto para variáveis latentes é chamado de *mean field variational EM*:

$$p(\mathbf{Z}|\mathcal{D}) \approx q(\mathbf{Z}) = \prod_{i=1}^N q_i(\mathbf{z}_i).$$

- Para modelos sem variáveis latentes  $\mathbf{Z}$  mas com parâmetros  $\boldsymbol{\theta}$ , temos o chamado *mean field variational Bayes*:

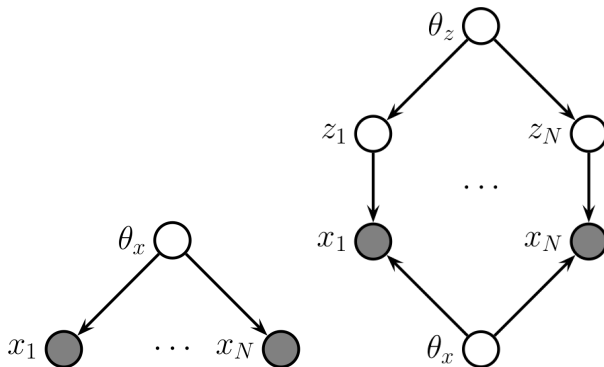
$$p(\boldsymbol{\theta}|\mathcal{D}) \approx q(\boldsymbol{\theta}) = \prod_{k=1}^K q_k(\boldsymbol{\theta}_k).$$

- No caso de variáveis latentes e parâmetros, temos o *mean field variational Bayes EM*:

$$p(\boldsymbol{\theta}, \mathbf{Z}|\mathcal{D}) \approx q(\boldsymbol{\theta}, \mathbf{Z}) = \prod_{k=1}^K q_k(\boldsymbol{\theta}_k) \prod_{i=1}^N q_i(\mathbf{z}_i).$$



# Variational Bayes



- À esquerda, temos variáveis latentes **globais**.
- À direita, temos variáveis latentes **globais** e **locais**.

# Agenda

- 1 Inferência variacional
- 2 Aproximação de mean field
- 3 Inferência variacional para Gaussiana univariada**
- 4 Inferência variacional para regressão linear
- 5 Inferência variacional para mistura de Gaussianas
- 6 Tópicos adicionais
- 7 Referências

# Inferência variacional para Gaussiana univariada

- A **inferência dos parâmetros** de uma distribuição Gaussiana a partir dos dados é analítica para prioris conjugadas.
- No entanto, detalharemos o procedimento variacional por questões didáticas e por permitir prioris não conjugadas.
- Considerando uma Gaussiana univariada  $\mathcal{N}(x|\mu, \tau^{-1})$ , buscamos a **posteriori**  $p(\mu, \tau^{-1}|\mathcal{D})$  dadas as observações  $\mathcal{D}$ .

# Inferência variacional para Gaussiana univariada

- A **inferência dos parâmetros** de uma distribuição Gaussiana a partir dos dados é analítica para prioris conjugadas.
- No entanto, detalharemos o procedimento variacional por questões didáticas e por permitir prioris não conjugadas.
- Considerando uma Gaussiana univariada  $\mathcal{N}(x|\mu, \tau^{-1})$ , buscamos a **posteriori**  $p(\mu, \tau^{-1}|\mathcal{D})$  dadas as observações  $\mathcal{D}$ .
- A **verossimilhança** será dada por:

$$p(\mathcal{D}|\mu, \tau) = \prod_{i=1}^N \mathcal{N}(x_i|\mu, \tau^{-1}).$$

# Inferência variacional para Gaussiana univariada

- A **inferência dos parâmetros** de uma distribuição Gaussiana a partir dos dados é analítica para prioris conjugadas.
- No entanto, detalharemos o procedimento variacional por questões didáticas e por permitir prioris não conjugadas.
- Considerando uma Gaussiana univariada  $\mathcal{N}(x|\mu, \tau^{-1})$ , buscamos a **posteriori**  $p(\mu, \tau^{-1}|\mathcal{D})$  dadas as observações  $\mathcal{D}$ .
- A **verossimilhança** será dada por:

$$p(\mathcal{D}|\mu, \tau) = \prod_{i=1}^N \mathcal{N}(x_i|\mu, \tau^{-1}).$$

- As **priori** dos parâmetros serão as conjugadas usuais:

$$p(\mu, \tau) = \mathcal{N}(\mu|\mu_0, (\kappa_0\tau)^{-1})\text{Ga}(\tau|a_0, b_0).$$

# Inferência variacional para Gaussiana univariada

- A **inferência dos parâmetros** de uma distribuição Gaussiana a partir dos dados é analítica para prioris conjugadas.
- No entanto, detalharemos o procedimento variacional por questões didáticas e por permitir prioris não conjugadas.
- Considerando uma Gaussiana univariada  $\mathcal{N}(x|\mu, \tau^{-1})$ , buscamos a **posteriori**  $p(\mu, \tau^{-1}|\mathcal{D})$  dadas as observações  $\mathcal{D}$ .
- A **verossimilhança** será dada por:

$$p(\mathcal{D}|\mu, \tau) = \prod_{i=1}^N \mathcal{N}(x_i|\mu, \tau^{-1}).$$

- As **priori** dos parâmetros serão as conjugadas usuais:

$$p(\mu, \tau) = \mathcal{N}(\mu|\mu_0, (\kappa_0\tau)^{-1})\text{Ga}(\tau|a_0, b_0).$$

- A **posteriori** será aproximada por uma forma fatorada:

$$q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau).$$

# Inferência variacional para Gaussiana univariada

- A distribuição  $q_\mu(\mu)$  ótima é obtida marginalizando  $\tau$  da log-conjunta (ignorando os termos que não contêm  $\mu$ ):

$$\begin{aligned}\log q_\mu(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathcal{D}|\mu, \tau)p(\mu|\tau)p(\tau)] + \text{cte.} \\ &= \mathbb{E}_{q_\tau}[\log p(\mathcal{D}|\mu, \tau) + \log p(\mu|\tau)] + \text{cte.} \\ &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2} \left\{ \kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^N (x_i - \mu)^2 \right\} + \text{cte.} \\ &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2} \left\{ \mu^2(\kappa_0 + N) - 2\mu \left( \kappa_0\mu_0 + \sum_{i=1}^N x_i \right) \right\} + \text{cte.}\end{aligned}$$

# Inferência variacional para Gaussiana univariada

- A distribuição  $q_\mu(\mu)$  ótima é obtida marginalizando  $\tau$  da log-conjunta (ignorando os termos que não contêm  $\mu$ ):

$$\begin{aligned}\log q_\mu(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathcal{D}|\mu, \tau)p(\mu|\tau)p(\tau)] + \text{cte.} \\ &= \mathbb{E}_{q_\tau}[\log p(\mathcal{D}|\mu, \tau) + \log p(\mu|\tau)] + \text{cte.} \\ &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2} \left\{ \kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^N (x_i - \mu)^2 \right\} + \text{cte.} \\ &= -\frac{\mathbb{E}_{q_\tau}[\tau]}{2} \left\{ \mu^2(\kappa_0 + N) - 2\mu \left( \kappa_0\mu_0 + \sum_{i=1}^N x_i \right) \right\} + \text{cte.}\end{aligned}$$

- Como temos uma forma quadrática para  $\mu$ , a forma ótima para  $q_\mu(\mu)$  é uma Gaussiana:

$$\log q_\mu(\mu) = \log \mathcal{N}(\mu|\mu_N, \kappa_N^{-1}) \propto -\frac{\kappa_N}{2}(\mu^2 - 2\mu\mu_N + \mu_N^2).$$



# Inferência variacional para Gaussiana univariada

- Assim, a forma ótima de  $q_\mu(\mu)$  será dada por:

$$\begin{aligned}q_\mu(\mu) &= \mathcal{N}(\mu | \mu_N, \kappa_N^{-1}), \\ \mu_N &= \frac{\kappa_0 \mu_0 + \sum_{i=1}^N x_i}{\kappa_0 + N}, \\ \kappa_N &= (\kappa_0 + N) \mathbb{E}_{q_\tau}[\tau].\end{aligned}$$

- O valor de  $\mathbb{E}_{q_\tau}[\tau]$  poderá ser calculado quando obtivermos a forma de  $q_\tau(\tau)$ .

# Inferência variacional para Gaussiana univariada

- A distribuição  $q_\tau(\tau)$  ótima é obtida marginalizando  $\mu$  na log-conjunta (ignorando os termos que não contêm  $\tau$ ):

$$\begin{aligned}\log q_\tau(\tau) &= \mathbb{E}_{q_\mu}[\log p(\mathcal{D}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)] + \text{cte.} \\ &= \frac{N}{2} \log \tau + \frac{1}{2} \log \tau + (a_0 - 1) \log \tau - b_0 \tau \\ &\quad - \frac{\tau}{2} \mathbb{E}_{q_\mu} \left[ \sum_{i=1}^N (x_i - \mu)^2 + \kappa_0 (\mu - \mu_0)^2 \right] + \text{cte.} \\ &= \left( \frac{N+1}{2} + a_0 - 1 \right) \log \tau \\ &\quad - \tau \left\{ b_0 + \frac{1}{2} \mathbb{E}_{q_\mu} \left[ \sum_{i=1}^N (x_i - \mu)^2 + \kappa_0 (\mu - \mu_0)^2 \right] \right\} + \text{cte.}\end{aligned}$$

## Inferência variacional para Gaussiana univariada

- A distribuição  $q_\tau(\tau)$  ótima é obtida marginalizando  $\mu$  na log-conjunta (ignorando os termos que não contêm  $\tau$ ):

$$\begin{aligned}\log q_\tau(\tau) &= \mathbb{E}_{q_\mu}[\log p(\mathcal{D}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)] + \text{cte.} \\ &= \frac{N}{2} \log \tau + \frac{1}{2} \log \tau + (a_0 - 1) \log \tau - b_0 \tau \\ &\quad - \frac{\tau}{2} \mathbb{E}_{q_\mu} \left[ \sum_{i=1}^N (x_i - \mu)^2 + \kappa_0 (\mu - \mu_0)^2 \right] + \text{cte.} \\ &= \left( \frac{N+1}{2} + a_0 - 1 \right) \log \tau \\ &\quad - \tau \left\{ b_0 + \frac{1}{2} \mathbb{E}_{q_\mu} \left[ \sum_{i=1}^N (x_i - \mu)^2 + \kappa_0 (\mu - \mu_0)^2 \right] \right\} + \text{cte.}\end{aligned}$$

- Reconhecemos o logaritmo de uma distribuição gamma:

$$\log q_\tau(\tau) = \log \text{Ga}(\tau|a_N, b_N) \propto (a_N - 1) \log \tau - b_N \tau.$$

# Inferência variacional para Gaussiana univariada

- Assim, a forma ótima de  $q_\tau(\tau)$  será dada por:

$$\begin{aligned}q_\tau(\tau) &= \text{Ga}(\tau | a_N, b_N), \\a_N &= a_0 + \frac{N+1}{2}, \\b_N &= b_0 + \frac{1}{2} \mathbb{E}_{q_\mu} \left[ \kappa_0 (\mu - \mu_0)^2 + \sum_{i=1}^N (x_i - \mu)^2 \right] \\&= b_0 + \frac{\kappa_0}{2} (\mathbb{E}_{q_\mu} [\mu^2] - 2\mathbb{E}_{q_\mu} [\mu] \mu_0 + \mu_0^2) \\&\quad + \frac{1}{2} \sum_{i=1}^N (x_i^2 - 2x_i \mathbb{E}_{q_\mu} [\mu] + \mathbb{E}_{q_\mu} [\mu^2]).\end{aligned}$$

# Inferência variacional para Gaussiana univariada

- Agora podemos computar as esperanças necessárias:

$$\mathbb{E}_{q_\tau}[\tau] = \frac{a_N}{b_N},$$

$$\mathbb{E}_{q_\mu}[\mu] = \mu_N,$$

$$\mathbb{E}_{q_\mu}[\mu^2] = \frac{1}{\kappa_N} + \mu_N^2$$

# Inferência variacional para Gaussiana univariada

- As atualizações das distribuições passam a ser:

$$q_{\mu}(\mu) = \mathcal{N}(\mu | \mu_N, \kappa_N^{-1}),$$

$$\mu_N = \frac{\kappa_0 \mu_0 + \sum_{i=1}^N x_i}{\kappa_0 + N}, \quad \kappa_N = (\kappa_0 + N) \frac{a_N}{b_N},$$

$$q_{\tau}(\tau) = \text{Ga}(\tau | a_N, b_N),$$

$$a_N = a_0 + \frac{N + 1}{2},$$

$$b_N = b_0 + \frac{\kappa_0}{2} \left( \frac{1}{\kappa_N} + \mu_N^2 - 2\mu_N \mu_0 + \mu_0^2 \right) \\ + \frac{1}{2} \sum_{i=1}^N \left( x_i^2 - 2x_i \mu_N + \frac{1}{\kappa_N} + \mu_N^2 \right).$$

- Note que  $\mu_N$  e  $a_N$  são fixos, mas  $\kappa_N$  e  $b_N$  precisam ser atualizados iterativamente.

# Inferência variacional para Gaussiana univariada

- A aproximação variacional da posteriori será dada por:

$$\begin{aligned}q(\mu, \tau) &= q_\mu(\mu) q_\tau(\tau) \\&= \mathcal{N}(\mu | \mu_N, \kappa_N^{-1}) \text{Ga}(\tau | a_N, b_N) \\&= \frac{\sqrt{\kappa_N}}{\sqrt{2\pi}} e^{-\frac{\tau(x-\mu_N)^2}{2}} \frac{b_N^{a_N}}{\Gamma(a_N)} \tau^{a_N-1} e^{-b_N \tau} \\&= \frac{b_N^{a_N} \sqrt{\kappa_N}}{\Gamma(a_N) \sqrt{2\pi}} \tau^{a_N-1} e^{-b_N \tau} e^{-\frac{\tau(x-\mu_N)^2}{2}}.\end{aligned}$$

- A posteriori analítica seria uma distribuição Gaussiana-gamma (ou normal-gamma):

$$\begin{aligned}\text{NormalGamma}(\mu, \tau | m, \kappa, a, b) \\&= \mathcal{N}(\mu | m, (\kappa \tau)^{-1}) \text{Ga}(\tau | a, b) \\&= \frac{b^a \sqrt{\kappa}}{\Gamma(a) \sqrt{2\pi}} \tau^{a-\frac{1}{2}} e^{-b\tau} e^{-\frac{\kappa \tau (x-m)^2}{2}}.\end{aligned}$$

# Inferência variacional para Gaussiana univariada

## Algoritmo variational para uma Gaussiana

- 1 Escolha os hiperparâmetros das priors:  $\mu_0, \kappa_0, a_0, b_0$ .
- 2 Calcule os valores ótimos para  $\mu_N$  e  $a_N$ :

$$\mu_N = \frac{\kappa_0 \mu_0 + \sum_{i=1}^N x_i}{\kappa_0 + N}, \quad a_N = a_0 + \frac{N + 1}{2}.$$

- 3 Inicialize os parâmetros variacionais:  $\kappa_N^{(0)}, b_N^{(0)}$ .
- 4 Repita até convergir (índices das iterações omitidos):

$$\kappa_N = (\kappa_0 + N) \frac{a_N}{b_N},$$

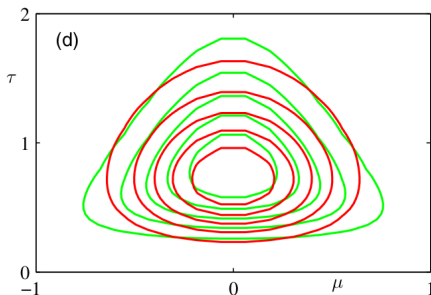
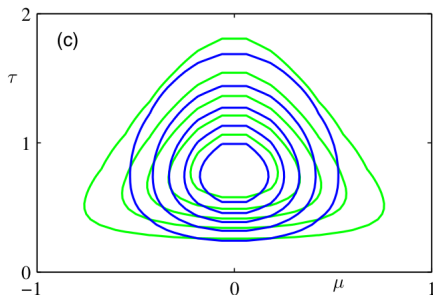
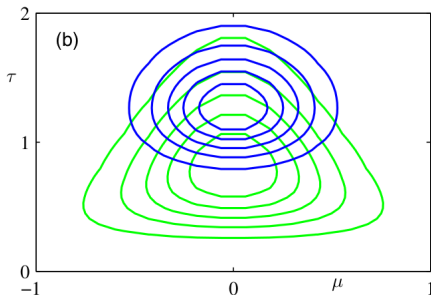
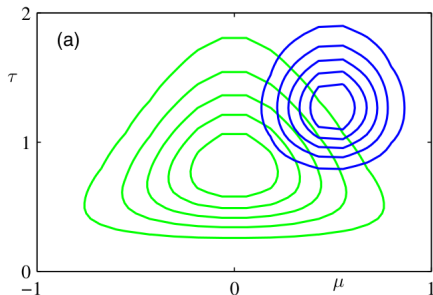
$$b_N = b_0 + \frac{\kappa_0}{2} \left[ \frac{1}{\kappa_N} + (\mu_N - \mu_0)^2 \right] + \frac{1}{2} \sum_{i=1}^N \left[ \frac{1}{\kappa_N} + (x_i - \mu_N)^2 \right]$$

- 5 Defina  $p(\mu, \tau | \mathcal{D}) \approx q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$  em que:

$$q_\mu(\mu) = \mathcal{N}(\mu | \mu_N, \kappa_N^{-1}), \quad q_\tau(\tau) = \text{Ga}(\tau | a_N, b_N).$$



# Inferência para os parâmetros de $\mathcal{N}(x|\mu, \tau^{-1})$



# Agenda

- 1 Inferência variacional
- 2 Aproximação de mean field
- 3 Inferência variacional para Gaussiana univariada
- 4 Inferência variacional para regressão linear
- 5 Inferência variacional para mistura de Gaussianas
- 6 Tópicos adicionais
- 7 Referências

# Inferência variacional para regressão linear

- Consideramos anteriormente o **modelo linear** abaixo:

$$\underbrace{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})}_{\text{verossimilhança}} = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \tau^{-1}\mathbf{I}),$$
$$\underbrace{p(\mathbf{w})}_{\text{priori}} = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}),$$

em que  $\boldsymbol{\theta}$  reúne todos os parâmetros do modelo.

- Já encontramos antes a solução analítica para  $p(\mathbf{w}|\mathcal{D})$ , mas fixamos  $\tau, \alpha$  ou os otimizamos via ML-II.

# Inferência variacional para regressão linear

- Consideramos anteriormente o **modelo linear** abaixo:

$$\underbrace{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})}_{\text{verossimilhança}} = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \tau^{-1}\mathbf{I}),$$
$$\underbrace{p(\mathbf{w})}_{\text{priori}} = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}),$$

em que  $\boldsymbol{\theta}$  reúne todos os parâmetros do modelo.

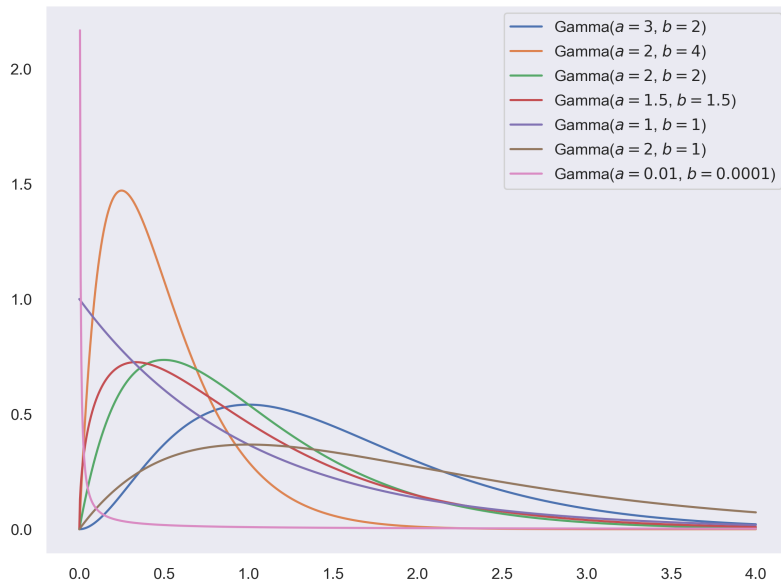
- Já encontramos antes a solução analítica para  $p(\mathbf{w}|\mathcal{D})$ , mas fixamos  $\tau, \alpha$  ou os otimizamos via ML-II.
- Definimos a **priori** abaixo para todos os parâmetros:

$$p(\mathbf{w}, \tau, \alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, (\tau\alpha)^{-1})\text{Ga}(\tau|a_0^\tau, b_0^\tau)\text{Ga}(\alpha|a_0^\alpha, b_0^\alpha).$$

- Buscamos a **aproximação variacional** abaixo para a **posteriori**:

$$p(\mathbf{w}, \tau, \alpha|\mathcal{D}) \approx q(\mathbf{w}, \tau, \alpha).$$

# Inferência variacional para regressão linear



# Inferência variacional para regressão linear

- Encontramos as regras de atualização de cada componente marginalizando as demais na log-conjunta:

$$q(\mathbf{w}, \tau, \alpha) = \mathcal{N}(\mathbf{w} | \mathbf{w}_N, \tau^{-1} \mathbf{V}_N) \text{Ga}(\tau | a_N^\tau, b_N^\tau) \text{Ga}(\alpha | a_N^\alpha, b_N^\alpha),$$

$$\mathbf{V}_N = (\mathbf{A} + \mathbf{X}^\top \mathbf{X})^{-1},$$

$$\mathbf{A} = \frac{a_N^\alpha}{b_N^\alpha} \mathbf{I},$$

$$\mathbf{w}_N = \mathbf{V}_N \mathbf{X}^\top \mathbf{y},$$

$$a_N^\tau = a_0^\tau + \frac{N}{2},$$

$$b_N^\tau = b_0^\tau + \frac{1}{2} (\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \mathbf{w}_N^\top \mathbf{A} \mathbf{w}_N),$$

$$a_N^\alpha = a_0^\alpha + \frac{D}{2},$$

$$b_N^\alpha = b_0^\alpha + \frac{1}{2} \left( \frac{a_N^\tau}{b_N^\tau} \mathbf{w}_N^\top \mathbf{w}_N + \text{Tr}(\mathbf{V}_N) \right).$$

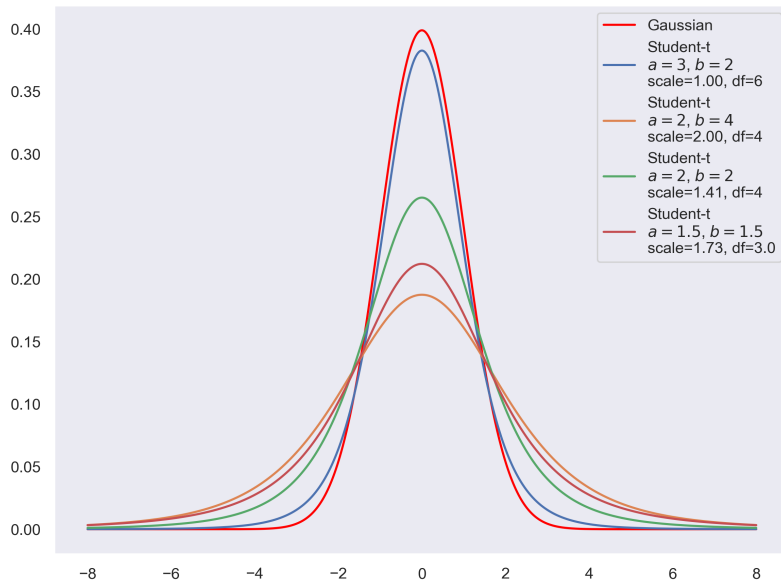
# Inferência variacional para regressão linear

- Aproximamos a **distribuição preditiva** substituindo a posteriori pela distribuição variacional (note que não precisamos marginalizar novamente  $\alpha$ ):

$$\begin{aligned} p(y_*|\mathbf{x}_*) &= \int p(y_*|\mathbf{x}_*, \mathbf{w}, \tau) p(\mathbf{w}, \tau | \mathcal{D}) d\mathbf{w} d\tau \\ &\approx \int p(y_*|\mathbf{x}_*, \mathbf{w}, \tau) q(\mathbf{w}, \tau) d\mathbf{w} d\tau \\ &\approx \int \mathcal{N}(y_* | \mathbf{w}^\top \mathbf{x}_*, \tau^{-1}) \mathcal{N}(\mathbf{w} | \mathbf{w}_N, \tau^{-1} \mathbf{V}_N) \text{Ga}(\tau | a_N^\tau, b_N^\tau) d\mathbf{w} d\tau \\ &\approx \int \mathcal{N}(y_* | \boldsymbol{\mu}_N^\top \mathbf{x}_*, \tau^{-1} (1 + \mathbf{x}_*^\top \mathbf{V}_N \mathbf{x}_*)) \text{Ga}(\tau | a_N^\tau, b_N^\tau) d\mathbf{w} d\tau \\ &\approx \mathcal{T}\left(y_* \middle| \boldsymbol{\mu}_N^\top \mathbf{x}_*, \frac{b_N^\tau}{a_N^\tau} (1 + \mathbf{x}_*^\top \mathbf{V}_N \mathbf{x}_*), 2a_N^\tau\right), \end{aligned}$$

em que  $\mathcal{T}(y_* | \mu, \sigma^2, \nu)$  denota uma distribuição **t de Student** de média  $\mu$ , escala  $\sigma$  e grau de liberdade  $\nu$ . Para  $\nu > 2$ , a variância é dada por  $\sigma^2 \frac{\nu}{\nu-1} = \frac{b_N^\tau}{a_N^\tau - 1} (1 + \mathbf{x}_*^\top \mathbf{V}_N \mathbf{x}_*)$ .

# Inferência variacional para regressão linear





# Inferência variacional para regressão linear

- Podemos calcular o ELBO para monitorar o algoritmo:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{q(\boldsymbol{\theta})}[\log p(\mathbf{y}, \boldsymbol{\theta} | \mathbf{X})] - \mathbb{E}_{q(\boldsymbol{\theta})}[\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q(\mathbf{w}, \tau)}[\log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \tau)] + \mathbb{E}_{q(\mathbf{w}, \tau, \alpha)}[\log p(\mathbf{w}, \tau | \alpha)] \\ &\quad + \mathbb{E}_{q(\alpha)}[\log p(\alpha)] - \mathbb{E}_{q(\mathbf{w}, \tau)}[\log q(\mathbf{w}, \tau)] - \mathbb{E}_{q(\alpha)}[\log q(\alpha)] \\ &= -\frac{N}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^N \left( \frac{a_N^\tau}{b_N^\tau} (y_i - \mathbf{w}_N^\top \mathbf{x}_i)^2 + \mathbf{x}_i^\top \mathbf{V}_N \mathbf{x}_i \right) \\ &\quad + \frac{1}{2} \log |\mathbf{V}_N| + \frac{D}{2} \\ &\quad - \log \Gamma(a_0^\tau) + a_0^\tau \log b_0^\tau - b_0^\tau \frac{a_N^\tau}{b_N^\tau} + \log \Gamma(a_N^\tau) - a_N^\tau \log b_N^\tau + a_N^\tau \\ &\quad - \log \Gamma(a_0^\alpha) + a_0^\alpha \log b_0^\alpha + \log \Gamma(a_N^\alpha) - a_N^\alpha \log b_N^\alpha,\end{aligned}$$

em que  $\Gamma(\cdot)$  denota a função gamma.

# Inferência variacional para regressão linear e ARD

- Alternativamente, poderíamos considerar o modelo linear abaixo:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \tau^{-1}\mathbf{I}),$$

$$p(\mathbf{w}, \tau, \boldsymbol{\alpha}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \text{diag}(\boldsymbol{\alpha})^{-1})\text{Ga}(\tau|a_0^\tau, b_0^\tau) \prod_{d=1}^D \text{Ga}(\alpha_d|a_0^\alpha, b_0^\alpha).$$

# Inferência variacional para regressão linear e ARD

- Alternativamente, poderíamos considerar o modelo linear abaixo:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \tau^{-1}\mathbf{I}),$$

$$p(\mathbf{w}, \tau, \boldsymbol{\alpha}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \text{diag}(\boldsymbol{\alpha})^{-1})\text{Ga}(\tau|a_0^\tau, b_0^\tau) \prod_{d=1}^D \text{Ga}(\alpha_d|a_0^\alpha, b_0^\alpha).$$

- Temos um hiperparâmetro  $\alpha_d$  para cada dimensão de  $\mathbf{w}$ .
  - **Framework ARD** (*automatic relevance determination*): capaz de identificar as variáveis **mais relevantes** ao problema.

# Inferência variacional para regressão linear e ARD

- A nova forma ótima para a **distribuição variacional** será:

$$q(\mathbf{w}, \tau, \boldsymbol{\alpha}) = \mathcal{N}(\mathbf{w} | \mathbf{w}_N, \tau^{-1} \mathbf{V}_N) \text{Ga}(\tau | a_N^\tau, b_N^\tau) \prod_{d=1}^D \text{Ga}(\alpha_d | a_N^\alpha, b_{N_d}^\alpha),$$

$$\mathbf{V}_N = (\mathbf{A} + \mathbf{X}^\top \mathbf{X})^{-1},$$

$$\mathbf{A} = \text{diag}(a_N^\alpha / b_N^\alpha),$$

$$\mathbf{w}_N = \mathbf{V}_N \mathbf{X}^\top \mathbf{y},$$

$$a_N^\tau = a_0^\tau + \frac{N}{2},$$

$$b_N^\tau = b_0^\tau + \frac{1}{2} (\|\mathbf{y} - \mathbf{X} \mathbf{w}_N\|^2 + \mathbf{w}_N^\top \mathbf{A} \mathbf{w}_N),$$

$$a_N^\alpha = a_0^\alpha + \frac{1}{2},$$

$$b_{N_d}^\alpha = b_0^\alpha + \frac{1}{2} \left( \frac{a_N^\tau}{b_N^\tau} w_{Nd}^2 + (\mathbf{V}_N)_{dd} \right).$$

# Inferência variacional para regressão linear e ARD

- A **distribuição preditiva** é a mesma do caso sem ARD, pois a posteriori de  $\alpha$  não aparece no cálculo:

$$p(y_*|\mathbf{x}_*) \approx \mathcal{T}\left(y_*|\boldsymbol{\mu}_N^\top \mathbf{x}_*, \frac{b_N^\tau}{a_N^\tau}(1 + \mathbf{x}_*^\top \mathbf{V}_N \mathbf{x}_*), 2a_N^\tau\right),$$

em que  $\mathcal{T}(\mu, \sigma^2, \nu)$  denota uma distribuição **t de Student** de média  $\mu$ , escala  $\sigma$  e grau de liberdade  $\nu$ . Para  $\nu > 2$ , a variância é dada por  $\frac{b_N^\tau}{a_N^\tau - 1}(1 + \mathbf{x}_*^\top \mathbf{V}_N \mathbf{x}_*)$ .

# Inferência variacional para regressão linear e ARD

- O cálculo do ELBO é parecido com o caso sem ARD, com uma alteração na última linha da expressão:

$$\begin{aligned}\mathcal{L} = & -\frac{N}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^N \left( \frac{a_N^\tau}{b_N^\tau} (y_i - \mathbf{w}_N^\top \mathbf{x}_i)^2 + \mathbf{x}_i^\top \mathbf{V}_N \mathbf{x}_i \right) \\ & + \frac{1}{2} \log |\mathbf{V}_N| + \frac{D}{2} \\ & - \log \Gamma(a_0^\tau) + a_0^\tau \log b_0^\tau - b_0^\tau \frac{a_N^\tau}{b_N^\tau} + \log \Gamma(a_N^\tau) - a_N^\tau \log b_N^\tau + a_N^\tau \\ & + \sum_{d=1}^D \left[ -\log \Gamma(a_0^\alpha) + a_0^\alpha \log b_0^\alpha + \log \Gamma(a_N^\alpha) - a_N^\alpha \log b_{N_d}^\alpha \right],\end{aligned}$$

em que  $\Gamma(\cdot)$  denota a função gamma.

# Inferência variacional para regressão linear e ARD

- Compare a predição obtida via **inferência variacional**:

$$p(y_*|\mathbf{x}_*) \approx \mathcal{T} \left( y_* | \boldsymbol{\mu}_N^\top \mathbf{x}_*, \frac{b_N^\tau}{a_N^\tau} (1 + \mathbf{x}_*^\top \mathbf{V}_N \mathbf{x}_*), 2a_N^\tau \right).$$

- Com a predição da **regressão linear Bayesiana** obtida antes:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \mathbf{S}_0),$$

$$p(\mathbf{y} | \mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{y} | \mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}),$$

$$p(\mathbf{w} | \mathcal{D}) = \mathcal{N}(\mathbf{w} | \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\mu} = (\mathbf{S}_0 \mathbf{X}^\top \mathbf{X} + \sigma^2 \mathbf{I})^{-1} \mathbf{S}_0 \mathbf{X}^\top \mathbf{y},$$

$$\boldsymbol{\Sigma} = \mathbf{S}_0 - (\mathbf{S}_0 \mathbf{X}^\top \mathbf{X} + \sigma^2 \mathbf{I})^{-1} \mathbf{S}_0 \mathbf{X}^\top \mathbf{X} \mathbf{S}_0,$$

$$p(y_* | \mathbf{x}_*) = \mathcal{N}(y_* | \boldsymbol{\mu}^\top \mathbf{x}_*, \mathbf{x}_*^\top \boldsymbol{\Sigma} \mathbf{x}_* + \sigma^2).$$

# Inferência variacional para regressão linear e ARD

- Compare a predição obtida via **inferência variacional**:

$$p(y_*|\mathbf{x}_*) \approx \mathcal{T} \left( y_* | \boldsymbol{\mu}_N^\top \mathbf{x}_*, \frac{b_N^\tau}{a_N^\tau} (1 + \mathbf{x}_*^\top \mathbf{V}_N \mathbf{x}_*), 2a_N^\tau \right).$$

- Com a predição da **regressão linear Bayesiana** obtida antes:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \mathbf{S}_0),$$

$$p(\mathbf{y} | \mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{y} | \mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}),$$

$$p(\mathbf{w} | \mathcal{D}) = \mathcal{N}(\mathbf{w} | \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\mu} = (\mathbf{S}_0 \mathbf{X}^\top \mathbf{X} + \sigma^2 \mathbf{I})^{-1} \mathbf{S}_0 \mathbf{X}^\top \mathbf{y},$$

$$\boldsymbol{\Sigma} = \mathbf{S}_0 - (\mathbf{S}_0 \mathbf{X}^\top \mathbf{X} + \sigma^2 \mathbf{I})^{-1} \mathbf{S}_0 \mathbf{X}^\top \mathbf{X} \mathbf{S}_0,$$

$$p(y_* | \mathbf{x}_*) = \mathcal{N}(y_* | \boldsymbol{\mu}^\top \mathbf{x}_*, \mathbf{x}_*^\top \boldsymbol{\Sigma} \mathbf{x}_* + \sigma^2 \mathbf{I}).$$

- Não dependemos mais de  $\mathbf{S}_0$  e  $\sigma^2$ , mas dos parâmetros de suas **distribuições variacionais**!



# Inferência variacional para regressão linear e ARD

## Resumo do algoritmo

- ① Escolha os hiperparâmetros das priors ( $D$  é a dimensão de  $\mathbf{x}_i$ ):

$$p(\mathbf{w}, \tau, \alpha) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \text{diag}(\alpha)^{-1}) \text{Ga}(\tau | a_0^\tau, b_0^\tau) \prod_{d=1}^D \text{Ga}(\alpha_d | a_0^\alpha, b_0^\alpha).$$

- ② Inicialize os parâmetros variacionais da aproximação:

$$q(\mathbf{w}, \tau, \alpha) = q(\mathbf{w} | \tau) q(\tau) q(\alpha) = \mathcal{N}(\mathbf{w} | \mathbf{w}_N, \tau^{-1} \mathbf{V}_N) \text{Ga}(\tau | a_N^\tau, b_N^\tau) \prod_{d=1}^D \text{Ga}(\alpha_d | a_N^\alpha, b_{Nd}^\alpha).$$

- ③ Considerando  $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ , repita até convergir:

→ Atualize os parâmetros de  $q(\mathbf{w}, \tau)$ :

$$\mathbf{A} = \text{diag}(a_N^\alpha / b_N^\alpha), \quad \mathbf{V}_N = (\mathbf{A} + \mathbf{X}^\top \mathbf{X})^{-1}, \quad \mathbf{w}_N = \mathbf{V}_N \mathbf{X}^\top \mathbf{y},$$
$$a_N^\tau = a_0^\tau + \frac{N}{2}, \quad b_N^\tau = b_0^\tau + \frac{1}{2} (\|\mathbf{y} - \mathbf{X} \mathbf{w}_N\|^2 + \mathbf{w}_N^\top \mathbf{A} \mathbf{w}_N).$$

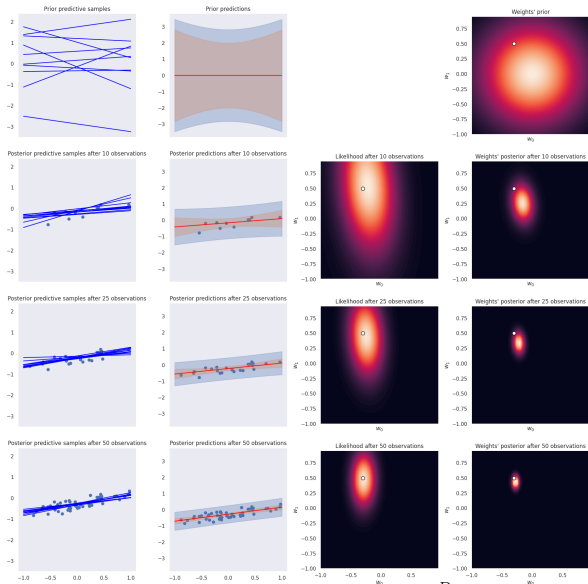
→ Atualize os parâmetros de  $q(\alpha)$ :

$$a_N^\alpha = a_0^\alpha + \frac{1}{2}, \quad b_{Nd}^\alpha = b_0^\alpha + \frac{1}{2} \left( \frac{a_N^\tau}{b_N^\tau} w_{Nd}^2 + (\mathbf{V}_N)_{dd} \right).$$

- ④ Predições para novos padrões  $\mathbf{x}_*$  são dadas por:

$$p(y_* | \mathbf{x}_*) \approx \mathcal{T} \left( y_* | \boldsymbol{\mu}_N^\top \mathbf{x}_*, \frac{b_N^\tau}{a_N^\tau} (1 + \mathbf{x}_*^\top \mathbf{V}_N \mathbf{x}_*), 2a_N^\tau \right).$$

# Inferência variacional para regressão linear



$$p(\mathbf{w}, \tau, \alpha) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \text{diag}(\alpha)^{-1}) \text{Ga}(\tau | 2, 1) \prod_{d=1}^D \text{Ga}(\alpha_d | 2, 1).$$

# Agenda

- 1 Inferência variacional
- 2 Aproximação de mean field
- 3 Inferência variacional para Gaussiana univariada
- 4 Inferência variacional para regressão linear
- 5 Inferência variacional para mistura de Gaussianas
- 6 Tópicos adicionais
- 7 Referências

# Inferência variacional para mistura de Gaussianas

- No GMM, a **verossimilhança** dos dados observados é dada por

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

- Usamos a **verossimilhança dos dados completos** (verossimilhança conjunta):

$$p(\mathbf{X}, \mathbf{z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{i=1}^N \prod_{k=1}^K \pi_k^{\mathbb{I}(z_i=k)} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\mathbb{I}(z_i=k)}.$$

- Modificaremos a notação, denotando  $z_{ik} = 1$ , caso  $\mathbf{x}_i$ , pertença à componente  $k$  e  $z_{ik} = 0$ , caso contrário.
- Além disso, faremos  $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k^{-1}$ .
- Assim:

$$p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{i=1}^N \prod_{k=1}^K \pi_k^{z_{ik}} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{ik}}.$$

# Inferência variacional para mistura de Gaussianas

- Escolhemos uma **priori conjugada** para os parâmetros:

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \text{Dir}(\boldsymbol{\pi} | \alpha_0 \mathbf{1}) \prod_k \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) \text{Wi}(\boldsymbol{\Lambda}_k | \boldsymbol{L}_0, \nu_0),$$

em que usamos priori iguais para todas as componentes e os subscritos 0 indicam os seus hiperparâmetros.

# Inferência variacional para mistura de Gaussianas

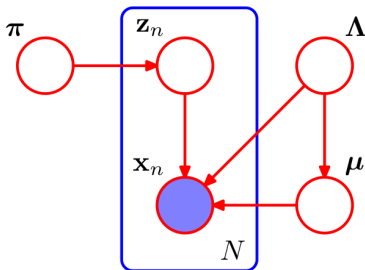
- Escolhemos uma **priori conjugada** para os parâmetros:

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \text{Dir}(\boldsymbol{\pi} | \alpha_0 \mathbf{1}) \prod_k \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) \text{Wi}(\boldsymbol{\Lambda}_k | \mathbf{L}_0, \nu_0),$$

em que usamos priori iguais para todas as componentes e os subscritos 0 indicam os seus hiperparâmetros.

- A **distribuição conjunta** envolvendo todas as variáveis será:

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &= p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\mathbf{Z} | \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu} | \boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}). \end{aligned}$$



# Inferência variacional para mistura de Gaussianas

- Escolhemos uma **distribuição variacional** fatorizada entre as variáveis latentes e os parâmetros (essa é a única aproximação do procedimento de inferência!):

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}).$$

# Inferência variacional para mistura de Gaussianas

- Escolhemos uma **distribuição variacional** fatorizada entre as variáveis latentes e os parâmetros (essa é a única aproximação do procedimento de inferência!):

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}).$$

- A componente  $q(\mathbf{Z})$  é atualizada da maneira usual, marginalizando as demais variáveis a partir da log-verossimilhança dos dados completos (a log-conjunta):

$$q(\mathbf{Z}) = \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}}[\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const.}$$

- Lembre que todos os termos que não envolvem  $\mathbf{Z}$  serão “absorvidos” pelo termo constante.



# Inferência variacional para mistura de Gaussianas

- Prosseguimos a partir da **distribuição conjunta** anterior:

$$\begin{aligned}\log q(\mathbf{Z}) &\propto \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}} [\log p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\mathbf{Z} | \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu} | \boldsymbol{\Lambda}) p(\boldsymbol{\Lambda})] \\ &\propto \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\ &\propto \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}} \left[ \log \prod_{i=1}^N \prod_{k=1}^K \pi_k^{z_{ik}} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{ik}} \right] \\ &\propto \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}} \left[ \sum_{i=1}^N \sum_{k=1}^K z_{ik} \log \pi_k + z_{ik} \log \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) \right] \\ &\propto \sum_{i=1}^N \sum_{k=1}^K \left\{ z_{ik} \mathbb{E}_{\boldsymbol{\pi}} [\log \pi_k] + z_{ik} \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} [\log \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})] \right\} .\end{aligned}$$

- Como esperado, precisamos computar esperanças com relação aos parâmetros.

# Inferência variacional para mistura de Gaussianas

- Prosseguimos encontrando a distribuição variacional  $q(\boldsymbol{\pi})$ , integrando os outros valores na log-conjunta:

$$\begin{aligned}\log q(\boldsymbol{\pi}) &\propto \mathbb{E}_{\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}}[\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu} | \boldsymbol{\Lambda}) p(\boldsymbol{\Lambda})] \\ &\propto \mathbb{E}_{\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}}[\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\pi})] \\ &\propto \mathbb{E}_{\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}} \left[ \log \prod_{i=1}^N \prod_{k=1}^K \pi_k^{z_{ik}} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{ik}} \text{Dir}(\boldsymbol{\pi} | \alpha_0 \mathbf{1}) \right] \\ &\propto \sum_{i=1}^N \sum_{k=1}^K \mathbb{E}_{\mathbf{Z}}[z_{ik}] \log \pi_k + (\alpha_0 - 1) \sum_{k=1}^K \log \pi_k.\end{aligned}$$

- A última expressão equivale a uma nova distribuição de Dirichlet:

$$\log q(\boldsymbol{\pi}) \propto \sum_k (\alpha_k - 1) \log \pi_k,$$

$$q(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha}), \quad \alpha_k = \alpha_0 + \underbrace{\sum_i \mathbb{E}_{\mathbf{Z}}[z_{ik}]}_{N_k} = \alpha_0 + N_k.$$

# Inferência variacional para mistura de Gaussianas

- Para os parâmetros  $\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k$ , temos:

$$\log q(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k) q(\boldsymbol{\Lambda}_k) \propto \mathbb{E}_{\mathbf{Z}, \boldsymbol{\pi}} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu} | \boldsymbol{\Lambda}) p(\boldsymbol{\Lambda})].$$

- Isolando os termos contendo  $\boldsymbol{\mu}_k$  e  $\boldsymbol{\Lambda}_k$ , a expressão final possui a mesma forma da priori, uma distribuição Gaussiana-Wishart:

$$q(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k) q(\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \text{Wi}(\boldsymbol{\Lambda}_k | \mathbf{L}_k, \nu_k),$$

$$\beta_k = \beta_0 + N_k, \quad N_k = \sum_i r_{ik}, \quad r_{ik} = \mathbb{E}_{\mathbf{Z}}[z_{ik}],$$

$$\mathbf{m}_k = \frac{1}{\beta_k} (\beta_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}_k),$$

$$\mathbf{L}_k^{-1} = \mathbf{L}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^\top,$$

$$\nu_k = \nu_0 + N_k + 1,$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_i r_{ik} \mathbf{x}_i, \quad \mathbf{S}_k = \frac{1}{N_k} \sum_i r_{ik} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top.$$

# Inferência variacional para mistura de Gaussianas

- Voltamos à distribuição variacional  $q(\mathbf{Z})$ :

$$\log q(\mathbf{Z}) \propto \sum_{i=1}^N \sum_{k=1}^K \left\{ z_{ik} \mathbb{E}_{\boldsymbol{\pi}}[\log \pi_k] + z_{ik} \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}[\log \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})] \right\}.$$

- Como vimos,  $q(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha})$ . Assim, o primeiro termo será:

$$\mathbb{E}_{\pi_k}[\log \pi_k] = \psi(\alpha_k) - \psi\left(\sum_{k'} \alpha_{k'}\right),$$

em que  $\psi(\cdot)$  é a função digamma:  $\psi(a) = \frac{d}{da} \log \Gamma(a)$ .

# Inferência variacional para mistura de Gaussianas

- A esperança do segundo termo pode ser detalhada:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}[\log \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})] &= -\frac{D}{2} \log(2\pi) + \frac{1}{2} \mathbb{E}_{\boldsymbol{\Lambda}}[\log |\boldsymbol{\Lambda}_k|] \\ &\quad - \frac{1}{2} \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}[(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k (\mathbf{x}_i - \boldsymbol{\mu}_k)],\end{aligned}$$

em que  $D$  é a dimensão de  $\mathbf{x}_i$ .

- Vimos que

$$q(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k) q(\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \text{Wi}(\boldsymbol{\Lambda}_k | \mathbf{L}_k, \nu_k).$$

- Assim, as esperanças podem ser calculadas:

$$\mathbb{E}_{\boldsymbol{\Lambda}}[\log |\boldsymbol{\Lambda}_k|] = D \log 2 + \log |\mathbf{L}_k| + \sum_{d=1}^D \psi\left(\frac{\nu_k + 1 - d}{2}\right),$$

$$\mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}[(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k (\mathbf{x}_i - \boldsymbol{\mu}_k)] = \frac{D}{\beta_k} + \nu_k (\mathbf{x}_i - \mathbf{m}_k)^\top \mathbf{L}_k (\mathbf{x}_i - \mathbf{m}_k).$$

# Inferência variacional para mistura de Gaussianas

- Substituindo os termos calculados, encontramos a expressão de atualização para  $q(\mathbf{Z})$ :

$$q(\mathbf{Z}) = \prod_{i=1}^N \prod_{k=1}^K q(z_{ik}),$$

$$\begin{aligned}\log q(z_{ik}) &\propto z_{ik} (\mathbb{E}_{\boldsymbol{\pi}}[\log \pi_k] + \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}[\log \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})]) \\ &\propto z_{ik} \log \rho_{ik},\end{aligned}$$

$$q(\mathbf{Z}) \propto \prod_{i=1}^N \prod_{k=1}^K \rho_{ik}^{z_{ik}},$$

em que  $\log \rho_{ik} \triangleq \mathbb{E}_{\boldsymbol{\pi}}[\log \pi_k] + \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}[\log \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})]$ .

- Para cada valor  $i$  temos  $z_{ik} = 1$  para somente um valor de  $k$ . Além disso,  $\sum_k z_{ik} = 1$ . Assim, a normalização será:

$$q(\mathbf{Z}) = \prod_{i=1}^N \prod_{k=1}^K r_{ik}^{z_{ik}}, \quad r_{ik} = \frac{\rho_{ik}}{\sum_j \rho_{ij}} = \mathbb{E}[z_{ik}].$$

# Inferência variacional para mistura de Gaussianas

- A **distribuição preditiva** é aproximada substituindo a posteriori pela **distribuição variacional**:

$$\begin{aligned} p(\mathbf{x}_*|\mathbf{X}) &= \sum_{\mathbf{z}_*} \int p(\mathbf{x}_*|\mathbf{z}_*, \boldsymbol{\theta}) p(\mathbf{z}_*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} \\ &\approx \sum_k \int \pi_k \mathcal{N}(\mathbf{x}_*|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) q(\boldsymbol{\pi}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) d\boldsymbol{\pi} d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k \\ &\approx \sum_k \frac{\alpha_k}{\sum_{k'} \alpha_{k'}} \mathcal{T}\left(\mathbf{x}_* \middle| \mathbf{m}_k, \frac{1 + \beta_k}{(\nu_k + 1 - D)\beta_k} \mathbf{L}_k^{-1}, \nu_k + 1 - D\right), \end{aligned}$$

em que  $\mathcal{T}(\cdot)$  denota uma distribuição **t de Student multivariada**. A matriz de covariância é dada por

$$\frac{1 + \beta_k}{(\nu_k + 1 - D)\beta_k} \mathbf{L}_k^{-1}.$$

- Note que a preditiva é uma mistura de distribuições t de Student, mas para  $N$  grande converge para uma mistura de Gaussianas.

# Inferência variacional para mistura de Gaussianas

- O ELBO pode ser calculado para monitorar a convergência do algoritmo (apesar dos cálculos serem longos...):

$$\mathcal{L} = \sum_{\mathbf{Z}} \int q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \log \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)}{q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda}.$$

- Como usamos uma priori conjugada, já sabíamos do início o formato das distribuições variacionais: Categórica para  $\mathbf{Z}$ , Dirichlet para  $\boldsymbol{\pi}$  e Gaussiana-Wishart para  $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ .



# Inferência variacional para mistura de Gaussianas

- O ELBO pode ser calculado para monitorar a convergência do algoritmo (apesar dos cálculos serem longos...):

$$\mathcal{L} = \sum_{\mathbf{Z}} \int q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \log \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)}{q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda}.$$

- Como usamos uma priori conjugada, já sabíamos do início o formato das distribuições variacionais: Categórica para  $\mathbf{Z}$ , Dirichlet para  $\boldsymbol{\pi}$  e Gaussiana-Wishart para  $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ .
- Note que poderíamos escrever o ELBO diretamente em função dos parâmetros variacionais e otimizá-lo, encontrando expressões de atualização idênticas às que derivamos anteriormente (usaremos essa estratégia nas próximas aulas!).

# Inferência variacional para mistura de Gaussianas

- O cálculo do ELBO pode ainda ser usado para aproximar a distribuição do **número de componentes**  $K$ .
- Após treinar vários modelos com diversos valores para  $K$ , temos:

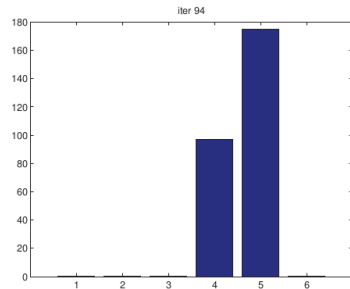
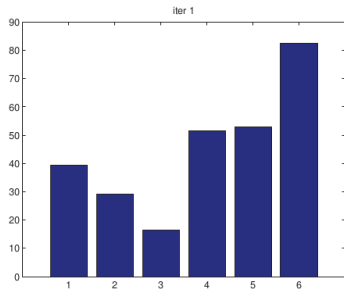
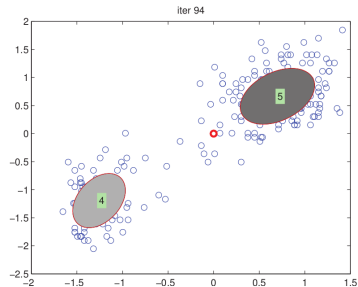
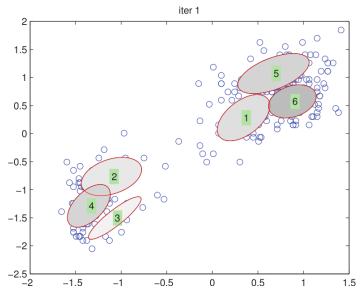
$$p(K|\mathcal{D}) \approx \frac{\exp(\mathcal{L}(K))}{\sum_{K'} \exp(\mathcal{L}(K'))}.$$

- No entanto, como existem  $K!$  permutações possíveis para  $K$  componentes que apresentam o mesmo valor do ELBO, devemos usar o seguinte **limiar modificado**:

$$\mathcal{L}'(K) = \mathcal{L}(K) + \log(K!).$$

- Alternativamente, poderíamos escolher um valor de  $K$  relativamente grande e usar  $\alpha_0 \ll 1$ . Com isso, o procedimento de otimização **“poda”** componentes associados a poucos padrões.

# Inferência variacional para mistura de Gaussianas



Exemplo de GMM variacional para  $\alpha_0 = 0.001$ . Abaixo, a evolução dos parâmetros  $\alpha_k$ .

# Inferência variacional para GMMs

## Resumo do algoritmo

- 1 Escolha  $K, \alpha_0, \mathbf{m}_0, \beta_0, \nu_0, \mathbf{L}_0, \alpha_k^{(0)}, \mathbf{m}_k^{(0)}, \beta_k^{(0)}, \nu_k^{(0)}, \mathbf{L}_k^{(0)}, \forall k$ .
- 2 Faça  $t = 1$  e repita até convergir:
  - Passo E:

$$\begin{aligned} \log \rho_{ik} &= \psi(\alpha_k) - \psi\left(\sum_{k'} \alpha_{k'}\right) - \frac{D}{2} \log(2\pi) \\ &\quad + \frac{1}{2} \left[ D \log 2 + \log |\mathbf{L}_k| + \sum_{d=1}^D \psi\left(\frac{\nu_k + 1 - d}{2}\right) \right] \\ &\quad - \frac{1}{2} \left[ \frac{D}{\beta_k} + \nu_k (\mathbf{x}_i - \mathbf{m}_k)^\top \mathbf{L}_k (\mathbf{x}_i - \mathbf{m}_k) \right], \\ r_{ik} &= \frac{\rho_{ik}}{\sum_j \rho_{ij}}. \end{aligned}$$

# Inferência variacional para GMMs

## Resumo do algoritmo

- 1 Escolha  $K, \alpha_0, \mathbf{m}_0, \beta_0, \nu_0, \mathbf{L}_0, \alpha_k^{(0)}, \mathbf{m}_k^{(0)}, \beta_k^{(0)}, \nu_k^{(0)}, \mathbf{L}_k^{(0)}, \forall k$ .
- 2 Faça  $t = 1$  e repita até convergir:

- Passo M:

$$N_k = \sum_i r_{ik}, \quad \alpha_k = \alpha_0 + N_k, \quad \beta_k = \beta_0 + N_k, \quad \nu_k = \nu_0 + N_k + 1,$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_i r_{ik} \mathbf{x}_i, \quad \mathbf{m}_k = \frac{1}{\beta_k} (\beta_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}_k),$$

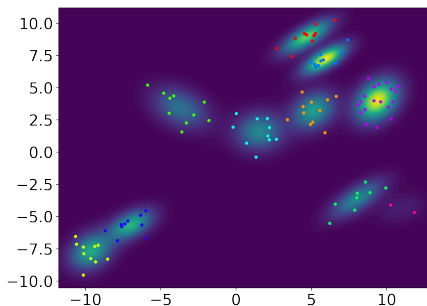
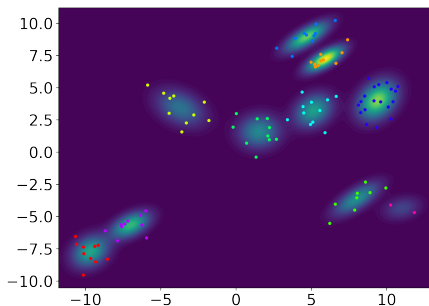
$$\mathbf{S}_k = \frac{1}{N_k} \sum_i r_{ik} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top,$$

$$\mathbf{L}_k^{-1} = \mathbf{L}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^\top.$$

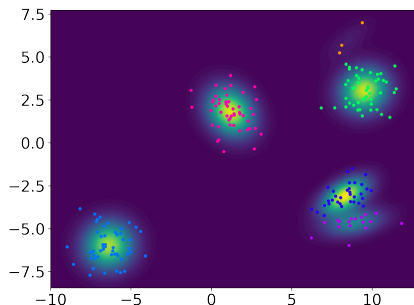
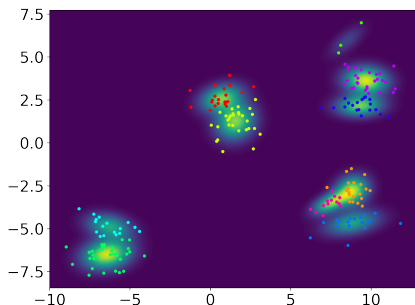
- 3 Dado um padrão  $\mathbf{x}_*$ , a verossimilhança marginal é computada por:

$$p(\mathbf{x}_* | \mathbf{X}) \approx \sum_k \frac{\alpha_k}{\sum_{k'} \alpha_{k'}} \mathcal{T} \left( \mathbf{x}_* \middle| \mathbf{m}_k, \frac{1 + \beta_k}{(\nu_k + 1 - D) \beta_k} \mathbf{L}_k^{-1}, \nu_k + 1 - D \right).$$

# GMM-MAP (esquerda) e Var-GMM (direita), $K = 10$



# GMM-MAP (esquerda) e Var-GMM (direita), $K = 10$



# Agenda

- 1 Inferência variacional
- 2 Aproximação de mean field
- 3 Inferência variacional para Gaussiana univariada
- 4 Inferência variacional para regressão linear
- 5 Inferência variacional para mistura de Gaussianas
- 6 Tópicos adicionais
- 7 Referências



# Tópicos adicionais

- Inferência variacional para regressão logística;
- Inferência variacional para regressão softmax;
- SVI - *stochastic variational inference* (ainda veremos!);
- Algoritmo Expectation-Propagation (EP).

# Tópicos adicionais

- Inferência variacional para regressão logística;
- Inferência variacional para regressão softmax;
- SVI - *stochastic variational inference* (ainda veremos!);
- Algoritmo Expectation-Propagation (EP).
- Reparameterization trick (ainda veremos!)

$$\mathcal{L} = \mathbb{E}_{p(\epsilon)} [\log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})], \quad \mathbf{z} = g(\epsilon, \phi, \mathbf{x}).$$

- Black box variational inference

$$\nabla_{\phi} \mathcal{L} = \mathbb{E}_{q_{\phi}} \left[ \underbrace{\nabla_{\phi} \log q_{\phi}(\mathbf{z}|\mathbf{x})}_{\text{score function}} (\log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})) \right].$$

# Agenda

- 1 Inferência variacional
- 2 Aproximação de mean field
- 3 Inferência variacional para Gaussiana univariada
- 4 Inferência variacional para regressão linear
- 5 Inferência variacional para mistura de Gaussianas
- 6 Tópicos adicionais
- 7 Referências

# Referências bibliográficas

- **Cap. 21** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Cap. 10** - BISHOP, Christopher M. **Pattern recognition and machine learning**, 2006.