Unigram mixtures and the EM algorithm

Guillaume Obozinski

Swiss Data Science Center

African Masters of Machine Intelligence, 2018-2019, AIMS, Kigali

Outline

- Brief review of entropy and Kullback-Leibler
- Bag of word model
- Mixture of unigrams
- Abstract EM scheme
- Application to the mixture of unigrams

Review: Entropy

Let X a r.v. with values in the finite set $\mathcal{X}$ and $p(x) = P(X = x)$.

Quantity of information of the observation x

$$I(x) := \log \frac{1}{p(x)}$$

Definition of entropy

$$H(X) := E\left[I(X)\right] = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

Remarks:

- Convention: $0 \log 0 = 0$
- H defined either with natural log or the log in base 2 (i.e. $\log_2$).
- $\log_2$ is better for coding interpretations
- In this course we will use the natural logarithm.

Review: Kullback-Leibler divergence

Definition

Let p and q be two finite distributions on $\mathcal{X}$ finite. The Kullback-Leibler divergence is defined by

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad = E_{X \sim p}\left[\log \frac{p(X)}{q(X)}\right]$$

⚠ The KL divergence is *not* a distance: it is not symmetric.

- $\forall p, q$ distributions, $\quad D(p \parallel q) \geq 0$
- $D(p \parallel q) = 0$ if and only if $p = q$
- If $\exists x \in \mathcal{X}$ with $q(x) = 0$ and $p(x) \neq 0$ then $D(p \parallel q) = +\infty$.

Review: Differential entropy and KL

Let X be a r.v. with distribution P and density p w.r.t. a measure μ.

Differential entropy:

$$H_{\text{diff}}(p) = -\int_{\mathcal{X}} p(x) \log(p(x)) d\mu(x)$$

Differential Kullback Leibler Divergence

$$\begin{aligned} D_{\text{diff}}(p \parallel q) &= \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x) \\ &= E_{X \sim p}\left[\log \frac{p(X)}{q(X)}\right] \end{aligned}$$

⚠

- $H_{\text{diff}}(p) \ngeq 0$
- $H_{\text{diff}}(p)$ depends on the reference measure μ.

$\Rightarrow$ $H_{\text{diff}}(p)$ does not capture intrinsic properties of P.

- However, $D_{\text{diff}}(p \parallel q)$ does not depend on μ.

The bag-of-word model, a vector-space representation of documents

Given

- a vocabulary of size d,

Represent a document consisting of N words

$$(w_1, \dots, w_N)$$

as x the vector of counts, or the vector of frequencies of the number of appearances of each of the words (possibly corrected with *tf-idf*):

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{N}_+^d, \quad \text{or } [0,1]_+^d, \quad \text{or } \mathbb{R}^d.$$

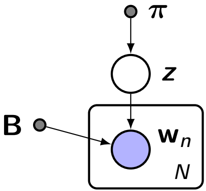
Document collection

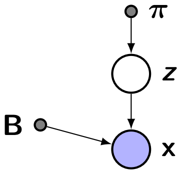
$$X = \begin{bmatrix} | & & | \\ x^{(1)} & \dots & x^{(M)} \\ | & & | \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & & x_1^{(M)} \\ \vdots & \ddots & \vdots \\ x_d^{(1)} & & x_d^{(M)} \end{bmatrix} \in \mathbb{R}^{d \times M}$$

Multinomial mixture model (Unigram mixture)

- K topics
- $\boldsymbol{z}$ component indicator vector
- $\boldsymbol{z} = (z_1, \ldots, z_K)^\top \in \{0,1\}^K$
- $\boldsymbol{z} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
- $\mathbf{w}_n \mid \{z_k = 1\} \sim \mathcal{M}(1, (b_{1k}, \ldots, b_{dk}))$
- $p(w_{nj} = 1 \mid z_k = 1) = b_{jk}$
- $$p(\mathbf{w}, \boldsymbol{z}) = \Big[\prod_{n=1}^{N} \prod_{j=1}^{d} \prod_{k=1}^{K} b_{jk}^{w_{nj} z_k}\Big] \cdot \prod_{k=1}^{K} \pi_k^{z_k}$$
- $$p(\mathbf{x}, \boldsymbol{z}) \propto \Big[\prod_{j=1}^{d} \prod_{k=1}^{K} b_{jk}^{x_j z_k}\Big] \cdot \prod_{k=1}^{K} \pi_k^{z_k}$$

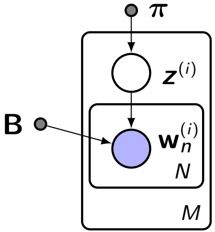
with $x_j = \sum_{n=1}^{N} w_{nj}$.

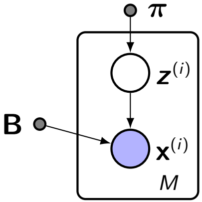




The same model written jointly for all documents

- $\boldsymbol{z}^{(i)}$ component indicator vector
- $\boldsymbol{z}^{(i)} = (z_1^{(i)}, \ldots, z_K^{(i)})^\top \in \{0,1\}^K$
- $\boldsymbol{z}^{(i)} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\boldsymbol{z}^{(i)}) = \prod_{k=1}^{K} \pi_k^{z_k^{(i)}}$
- $\mathbf{w}_n^{(i)} \mid \{z_k^{(i)}=1\} \sim \mathcal{M}(1, (b_{1k}, \ldots, b_{dk}))$
- $p(w_{nj}^{(i)} = 1 \mid z_k^{(i)} = 1) = b_{jk}$
- $\prod_{i=1}^{M} p(\mathbf{w}^{(i)}, \boldsymbol{z}^{(i)}) = \prod_{i,k} \left[\pi_k^{z_k^{(i)}} \prod_{n,j} b_{jk}^{w_{nj}^{(i)} z_k^{(i)}} \right]$
- $\prod_{i=1}^{M} p(\mathbf{x}^{(i)}, \boldsymbol{z}^{(i)}) = \prod_{i,k} \left[\pi_k^{z_k^{(i)}} \prod_{j} b_{jk}^{x_j^{(i)} z_k^{(i)}} \right]$





Applying maximum likelihood to the multinomial mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\boldsymbol{z}\in\mathcal{Z}} p(\mathbf{x},\boldsymbol{z}) = \sum_{\boldsymbol{z}\in\mathcal{Z}} \prod_{k=1}^K \left[\prod_{j=1}^d b_{jk}^{x_j z_k}\right] \pi_k^{z_k} = \sum_{k=1}^K \left[\prod_{j=1}^d b_{jk}^{x_j}\right] \pi_k$$

Issue

- The marginal log-likelihood $\ell(\mathbf{B},\boldsymbol{\pi}) = \sum_i \log(p(\mathbf{x}^{(i)}))$ is now complicated
- No hope to find a simple solution to the maximum likelihood problem
- By contrast the complete log-likelihood has a rather simple form:

$$\tilde{\ell}(\mathbf{B},\boldsymbol{\pi}) = \sum_{i=1}^M \log p(\mathbf{x}^{(i)},\boldsymbol{z}^{(i)}) = \sum_{i,j,k} x_j^{(i)}\, z_k^{(i)} \log(b_{jk}) + \sum_{i,k} z_k^{(i)} \log(\pi_k)$$

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\mathbf{B}, \boldsymbol{\pi}) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \boldsymbol{z}^{(i)}) = \sum_{i,j,k} x_j^{(i)}\, z_k^{(i)} \log(b_{jk}) + \sum_{i,k} z_k^{(i)} \log(\pi_k)$$

- If we knew $\boldsymbol{z}^{(i)}$ we could maximize $\tilde{\ell}(\mathbf{B}, \boldsymbol{\pi})$.
- If we knew $\mathbf{B}$ and $\boldsymbol{\pi}$, we could find the best $\boldsymbol{z}^{(i)}$ since we could compute the true a posteriori on $\boldsymbol{z}^{(i)}$ given $\mathbf{x}^{(i)}$:

$$p(z_k = 1 \mid \mathbf{x}; \mathbf{B}, \boldsymbol{\pi}) = \frac{\pi_k \prod_{j=1}^{d} b_{jk}^{x_j}}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j=1}^{d} b_{jk'}^{x_j}}$$

$\rightarrow$ Seems a chicken and egg problem...

- In addition, we want to solve

$$\max_{\mathbf{B}, \boldsymbol{\pi}} \sum_i \log \left(\sum_{\boldsymbol{z}^{(i)}} p(\mathbf{x}^{(i)}, \boldsymbol{z}^{(i)}) \right) \quad \text{and not} \quad \max_{\substack{\mathbf{B}, \boldsymbol{\pi}, \\ \boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(M)}}} \sum_i \log p(\mathbf{x}^{(i)}, \boldsymbol{z}^{(i)})$$

- Can we still use the intuitions above to construct an algorithm maximizing the marginal likelihood?

Principle of the Expectation-Maximization Algorithm

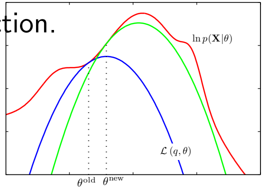
$$\log p(\mathbf{x};\boldsymbol{\theta}) = \log \sum_{\boldsymbol{z}} p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta}) = \log \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z})}$$
$$\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z})}$$
$$= \mathbb{E}_q[\log p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta})] + H(q) =: \mathcal{L}(q,\boldsymbol{\theta})$$

- This shows that $\mathcal{L}(q,\boldsymbol{\theta}) \leq \log p(\mathbf{x};\boldsymbol{\theta})$
- Moreover: $\boldsymbol{\theta} \mapsto \mathcal{L}(q,\boldsymbol{\theta})$ is often a **concave** function.
- Finally it is possible to show that

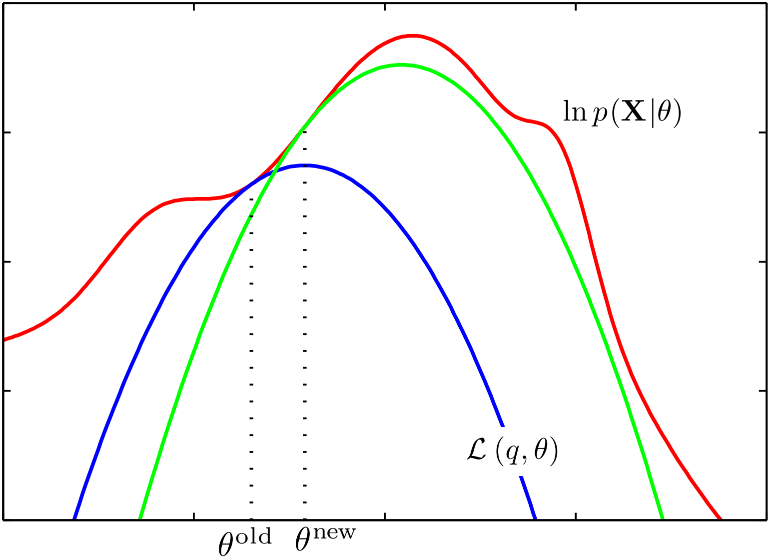
$$\mathcal{L}(q,\boldsymbol{\theta}) = \log p(\mathbf{x};\boldsymbol{\theta}) - KL(q||p(\cdot|\mathbf{x};\boldsymbol{\theta}))$$

So that if we set $q(\boldsymbol{z}) = p(\boldsymbol{z} \mid \mathbf{x};\boldsymbol{\theta}^{(t)})$ then

$$L(q,\boldsymbol{\theta}^{(t)}) = p(\mathbf{x};\theta^{(t)}).$$



A graphical idea of the EM algorithm



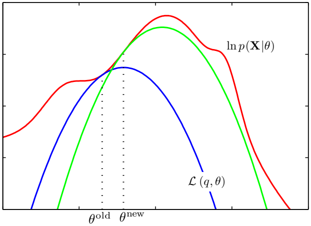
Expectation Maximization algorithm

Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

WHILE (Not converged)

Expectation step

1. $q(\boldsymbol{z}) = p(\boldsymbol{z} \mid \mathbf{x}; \boldsymbol{\theta}^{(t-1)})$
2. $$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_q\big[\log p(\mathbf{x}, \boldsymbol{z}; \boldsymbol{\theta}^{(t-1)})\big] + H(q)$$



Maximization step

1. $\boldsymbol{\theta}^{(t)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_q\big[\log p(\mathbf{x}, \boldsymbol{z}; \boldsymbol{\theta}^{(t-1)})\big]$

$$\begin{aligned} \boldsymbol{\theta}^{\text{old}} &= \boldsymbol{\theta}^{(t-1)} \\ \boldsymbol{\theta}^{\text{new}} &= \boldsymbol{\theta}^{(t)} \end{aligned}$$

ENDWHILE

Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$, we have

$$
\begin{aligned}
\mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\mathbf{B}, \boldsymbol{\pi})\right] &= \mathbb{E}_{q^{(t)}}\left[\log p(\boldsymbol{X}, \boldsymbol{Z}; \mathbf{B}, \boldsymbol{\pi})\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \boldsymbol{z}^{(i)}; \mathbf{B}, \boldsymbol{\pi})\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i,j,k} x_j^{(i)}\, z_k^{(i)} \log(b_{jk}) + \sum_{i,k} z_k^{(i)} \log(\pi_k)\right] \\
&= \sum_{i,j,k} x_j^{(i)}\, \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log(b_{jk}) + \sum_{i,k} \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log(\pi_k) \\
&= \sum_{i,j,k} x_j^{(i)}\, q_{ik}^{(t)} \log(b_{jk}) + \sum_{i,k} q_{ik}^{(t)} \log(\pi_k)
\end{aligned}
$$

Expectation step for the Multinomial mixture

We computed previously $q_i^{(t)}(\boldsymbol{z}^{(i)})$, which is a multinomial distribution defined by

$$q_i^{(t)}(\boldsymbol{z}^{(i)}) = p(\boldsymbol{z}^{(i)} | \mathbf{x}^{(i)}; \mathbf{B}^{(t-1)}, \boldsymbol{\pi}^{(t-1)})$$

Abusing notation we will denote $(q_{i1}^{(t)}, \ldots, q_{iK}^{(t)})$ the corresponding vector of probabilities defined by

$$q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$$

$$q_{ik}^{(t)} = p(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{B}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}) = \frac{\pi_k^{(t-1)} \prod_{j=1}^d \left[b_{jk}^{(t-1)}\right]^{x_j^{(i)}}}{\sum_{k'=1}^K \pi_{k'}^{(t-1)} \prod_{j=1}^d \left[b_{jk'}^{(t-1)}\right]^{x_j^{(i)}}}$$

Maximization step for the Multinomial mixture

$$(\mathbf{B}^t, \boldsymbol{\pi}^t) = \underset{\mathbf{B},\boldsymbol{\pi}}{\operatorname{argmax}}\, \mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\mathbf{B}, \boldsymbol{\pi})\right]$$

This yields the updates:

$$\boxed{b_{jk}^{(t)} = \frac{\sum_i x_j^{(i)}\, q_{ik}^{(t)}}{N \sum_i q_{ik}^{(t)}}} \quad \text{and} \quad \boxed{\pi_k^{(t)} = \frac{\sum_i q_{ik}^{(t)}}{M}}$$

Final EM algorithm for the Multinomial mixture model

Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

WHILE (Not converged)

Expectation step

$$q_{ik}^{(t)} \leftarrow \frac{\pi_k^{(t-1)} \prod_{j=1}^{d} \left[b_{jk}^{(t-1)}\right]^{x_j^{(i)}}}{\sum_{k'=1}^{K} \pi_{k'}^{(t-1)} \prod_{j=1}^{d} \left[b_{jk'}^{(t-1)}\right]^{x_j^{(i)}}}$$

Maximization step

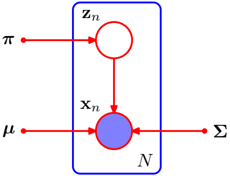
$$b_{jk}^{(t)} \leftarrow \frac{\sum_i x_j^{(i)} \, q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}} \qquad \text{and} \qquad \pi_k^{(t)} \leftarrow \frac{\sum_i q_{ik}^{(t)}}{M}$$

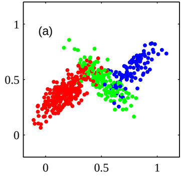
ENDWHILE

The EM algorithm for the Gaussian mixture model

Gaussian mixture model

- K components
- $\boldsymbol{z}$ component indicator
- $\boldsymbol{z} = (z_1, \ldots, z_K)^\top \in \{0,1\}^K$
- $\boldsymbol{z} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
- $p(\mathbf{x}|\boldsymbol{z}; (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k) = \sum_{k=1}^{K} z_k \, \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Estimation: $\underset{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}{\operatorname{argmax}} \log \left[\sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$

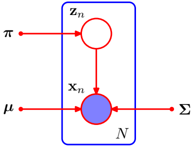




EM Algorithm for the Gaussian mixture model

Soit $\boldsymbol{\theta}^t = (\pi^t, (\boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}_k^t)_k)$.

$$\prod_{i=1}^{n} p(\boldsymbol{z}^i, \mathbf{x}^i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_k^i} \Big(\mathcal{N}(x^i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\Big)^{z_k^i}$$



E step:

$p(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^n | \mathbf{x}^1, \ldots, \mathbf{x}^n; \boldsymbol{\theta}^t) = \prod_{i=1}^n p(\boldsymbol{z}^i | \mathbf{x}^i; \boldsymbol{\theta}^t)$

$$q_k^i = P(z_k^i = 1 | x^i; \boldsymbol{\theta}^t) = \frac{p(x^i | z_k^i = 1; \boldsymbol{\theta}^t)\, P(z_k^i = 1; \boldsymbol{\theta}^t)}{p(x^i; \boldsymbol{\theta}^t)} = \frac{\pi_k^t \,\mathcal{N}(x^i; \boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}_k^t)}{\sum_\ell \pi_\ell^t \,\mathcal{N}(x^i; \boldsymbol{\mu}_\ell^t, \boldsymbol{\Sigma}_\ell^t)}$$

$$\mathbb{E}_q[\log p(\boldsymbol{z}, \mathbf{x} | \boldsymbol{\theta})] = \mathbb{E}_q\Big[\sum_{i,k} z_k^i \left(\log \pi_k + \log \mathcal{N}(x^i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)\Big]$$

$$= \sum_{i,k} q_k^i \log \pi_k - \frac{1}{2} q_k^i (x_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (x_i - \boldsymbol{\mu}_k) - \frac{1}{2} q_k^i \log((2\pi)^d |\boldsymbol{\Sigma}_k|)$$

EM Algorithm for the Gaussian mixture model II

Let $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) := \sum_{i=1}^{n} \sum_{z^{(i)} \in \mathcal{Z}} \log p(x^{(i)}, z^{(i)}; \boldsymbol{\theta})\, p(z^{(i)} \mid x^{(i)}; \boldsymbol{\theta}^{(t)})$, then

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \sum_{i,k} q_k^i \log \pi_k - \frac{1}{2} q_k^i (x_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (x_i - \boldsymbol{\mu}_k) - \frac{1}{2} q_k^i \log((2\pi)^d |\boldsymbol{\Sigma}_k|)$$

M step:

$$\max_{\pi, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k} Q\Big(\big(\pi, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k\big), \boldsymbol{\theta}^t\Big) \qquad \text{s.t.} \qquad \sum_k \pi_k = 1$$

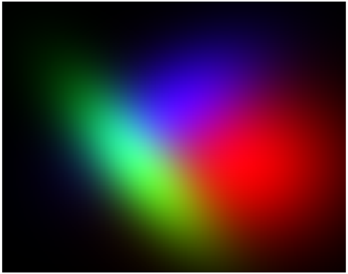
After calculations:

$$\boxed{n_k^{t+1} = \sum_i q_k^i} \qquad \boxed{\pi_k^{t+1} = \frac{n_k^{t+1}}{n}} \qquad \boxed{\boldsymbol{\mu}_k^{t+1} = \frac{1}{n_k^{t+1}} \sum_i q_k^i x_i}$$

$$\boxed{\boldsymbol{\Sigma}_k^{t+1} = \frac{1}{n_k^{t+1}} \sum_i q_k^i (x_i - \boldsymbol{\mu}_k^{t+1})(x_i - \boldsymbol{\mu}_k^{t+1})^\top}$$

EM Algorithm for the Gaussian mixture model III

$p(\mathbf{x}|z)$



$p(z|\mathbf{x})$