

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

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

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

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

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

- $\forall p, q$ distributions, $D(p \parallel q) \geq 0$
- $D(p \parallel q) = 0$ 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

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



- $H_{\text{diff}}(p) \not\geq 0$
 - $H_{\text{diff}}(p)$ 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-words 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

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

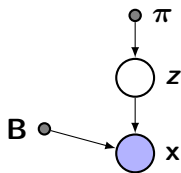
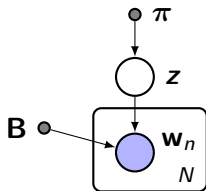
Multinomial mixture model (Unigram mixture)

- K topics
- \mathbf{z} component indicator vector
- $\mathbf{z} = (z_1, \dots, z_K)^\top \in \{0, 1\}^K$
- $\mathbf{z} \sim \mathcal{M}(1, (\pi_1, \dots, \pi_K))$
- $p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$
- $\mathbf{w}_n | \{z_k = 1\} \sim \mathcal{M}(1, (b_{1k}, \dots, b_{dk}))$
- $p(w_{nj} = 1 | z_k = 1) = b_{jk}$

- $$p(\mathbf{w}, \mathbf{z}) = \left[\prod_{n=1}^N \prod_{j=1}^d \prod_{k=1}^K b_{jk}^{w_{nj} z_k} \right] \cdot \prod_{k=1}^K \pi_k^{z_k}$$

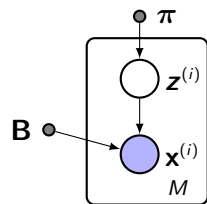
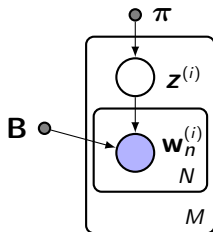
- $$p(\mathbf{x}, \mathbf{z}) \propto \left[\prod_{j=1}^d \prod_{k=1}^K b_{jk}^{x_j z_k} \right] \cdot \prod_{k=1}^K \pi_k^{z_k}$$

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



The same model written jointly for all documents

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



Relationship between $p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})$ and $p(\mathbf{w}^{(i)}, \mathbf{z}^{(i)})$

On the previous slides, we only wrote $p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})$ up to a constant factor. Indeed, given that $\mathbf{x}^{(i)}$ is the bag-of-words representation of document i and that it does not take into account the order in which the words have appeared in the document we have

$$p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \binom{N}{x_1^{(i)}, \dots, x_d^{(i)}} \prod_k \left[\pi_k^{z_k^{(i)}} \prod_j b_{jk}^{x_j^{(i)} z_k^{(i)}} \right]$$

$$p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \binom{N}{x_1^{(i)}, \dots, x_d^{(i)}} p(\mathbf{w}^{(i)}, \mathbf{z}^{(i)})$$

So that, with $c^{(i)} = \log \binom{N}{x_1^{(i)}, \dots, x_d^{(i)}}$ with

$$\log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \log p(\mathbf{w}^{(i)}, \mathbf{z}^{(i)}) + c^{(i)}$$

In the following slides we will write $\tilde{c} = \sum_{i=1}^M c^{(i)}$.

Applying maximum likelihood to the multinomial mixture

Let $\mathcal{Z} = \{z \in \{0, 1\}^K \mid \sum_{k=1}^K z_k = 1\}$, and $C = \binom{N}{x_1, \dots, x_d}$

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

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:

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

Applying maximum likelihood to the multinomial mixture

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

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

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

→ Seems a chicken and egg problem...

- In addition, we want to solve

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

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

Principle of the Expectation-Maximization Algorithm

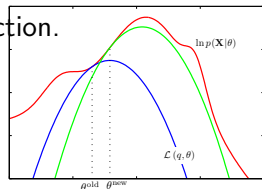
$$\begin{aligned}\log p(\mathbf{x}; \boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{q(\mathbf{z})} \\ &\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},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

- 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

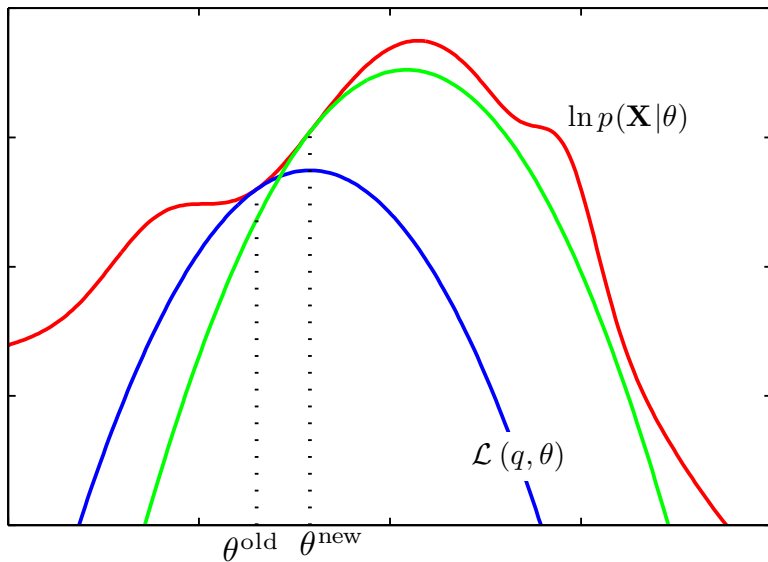
$$\mathcal{L}(q, \boldsymbol{\theta}) = \log p(\mathbf{x}; \boldsymbol{\theta}) - KL(q || p(\cdot | \mathbf{x}; \boldsymbol{\theta}))$$

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

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



A graphical idea of the EM algorithm



Expectation Maximization algorithm

Initialize $\theta = \theta_0$

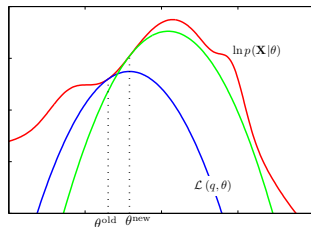
WHILE (Not converged)

Expectation step

① $q(z) = p(z \mid \mathbf{x}; \theta^{(t-1)})$

②

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



Maximization step

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

$$\theta^{\text{old}} = \theta^{(t-1)}$$

$$\theta^{\text{new}} = \theta^{(t)}$$

ENDWHILE

Expected complete log-likelihood

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

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

Expectation step for the Multinomial mixture

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

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

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

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

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

Maximization step for the Multinomial mixture

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

This yields the updates:

$$b_{jk}^{(t)} = \frac{\sum_i x_j^{(i)} q_{ik}^{(t)}}{N \sum_i q_{ik}^{(t)}}$$

and

$$\pi_k^{(t)} = \frac{\sum_i q_{ik}^{(t)}}{M}$$

Final EM algorithm for the Multinomial mixture model

Initialize $\theta = \theta_0$

WHILE (Not converged)

Expectation step

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

Maximization step

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

ENDWHILE

The EM algorithm for the Gaussian mixture model

Gaussian mixture model

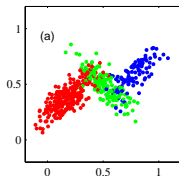
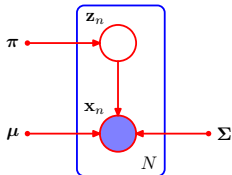
- K components
- z component indicator
- $\mathbf{z} = (z_1, \dots, z_K)^\top \in \{0, 1\}^K$
- $\mathbf{z} \sim \mathcal{M}(1, (\pi_1, \dots, \pi_K))$

- $$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

- $$p(\mathbf{x}|\mathbf{z}; (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k) = \sum_{k=1}^K z_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- $$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

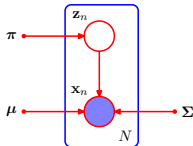
- Estimation:
$$\underset{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}{\operatorname{argmax}} \log \left[\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$$



EM Algorithm for the Gaussian mixture model

Soit $\theta^t = (\pi^t, (\mu_k^t, \Sigma_k^t)_k)$.

$$\prod_{i=1}^n p(\mathbf{z}^i, \mathbf{x}^i; \theta) = \prod_{i=1}^n \prod_{k=1}^K \pi_k^{z_k^i} \left(\mathcal{N}(\mathbf{x}^i; \mu_k, \Sigma_k) \right)^{z_k^i}$$



E step:

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

$$q_k^i = P(z_k^i = 1 | \mathbf{x}^i; \theta^t) = \frac{p(\mathbf{x}^i | z_k^i = 1; \theta^t) P(z_k^i = 1; \theta^t)}{p(\mathbf{x}^i; \theta^t)} = \frac{\pi_k^t \mathcal{N}(\mathbf{x}^i; \mu_k^t, \Sigma_k^t)}{\sum_{\ell} \pi_{\ell}^t \mathcal{N}(\mathbf{x}^i; \mu_{\ell}^t, \Sigma_{\ell}^t)}$$

$$\begin{aligned} \mathbb{E}_q[\log p(\mathbf{z}, \mathbf{x} | \theta)] &= \mathbb{E}_q \left[\sum_{i,k} z_k^i (\log \pi_k + \log \mathcal{N}(\mathbf{x}^i; \mu_k, \Sigma_k)) \right] \\ &= \sum_{i,k} q_k^i \log \pi_k - \frac{1}{2} q_k^i (\mathbf{x}_i - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) - \frac{1}{2} q_k^i \log((2\pi)^d |\Sigma_k|) \end{aligned}$$

EM Algorithm for the Gaussian mixture model II

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

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

M step:

$$\max_{\pi, (\mu_k, \Sigma_k)_k} Q((\pi, (\mu_k, \Sigma_k)_k), \theta^t) \quad \text{s.t.} \quad \sum_k \pi_k = 1$$

After calculations:

$$n_k^{t+1} = \sum_i q_k^i$$

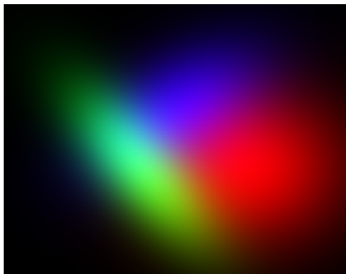
$$\pi_k^{t+1} = \frac{n_k^{t+1}}{n}$$

$$\mu_k^{t+1} = \frac{1}{n_k^{t+1}} \sum_i q_k^i x_i$$

$$\Sigma_k^{t+1} = \frac{1}{n_k^{t+1}} \sum_i q_k^i (x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^\top$$

EM Algorithm for the Gaussian mixture model III

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



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

