Mixture Models and Expectation-Maximization

Yaqiang Yao

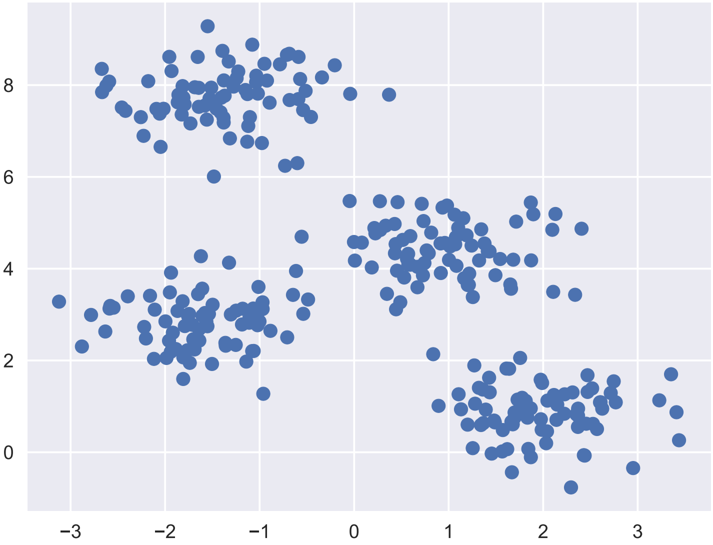
School of Computer Science and Technology
University of Science and Technology of China
Hefei China

April 9, 2018

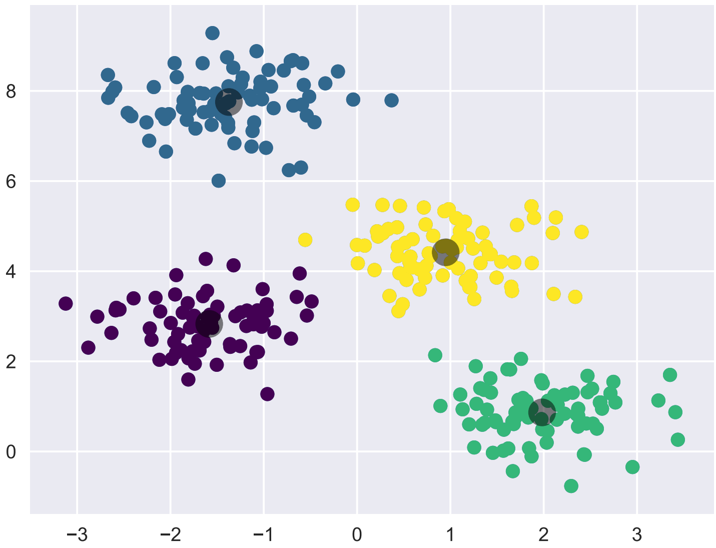
Outline

Outline

Motivation



Motivation



K-Means Clustering: Distortion Measure

- Dataset $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$
- Partition in K clusters
- Introduce a set of vectors (Cluster prototype): $\boldsymbol{\mu_k}$
- Introduce a corresponding set of binary indicator variable $r_{nk} \in \{0, 1\}$ (1-of-K coding scheme), such that

$$r_{nk} = \begin{cases} 1 & \text{if } \boldsymbol{x}_n \text{ is assigned to cluster } k \\ 0 & \text{otherwise} \end{cases}$$

- Distortion measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2$$

K-Means Clustering: Expectation Maximization

- Find values for $\{r_{nk}\}$ and $\{\boldsymbol{\mu}_k\}$ to minimize

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2$$

- Interactive procedure
  1. Minimize J w.r.t r_{nk}, keep $\boldsymbol{\mu}_k$ fixed (Expectation)

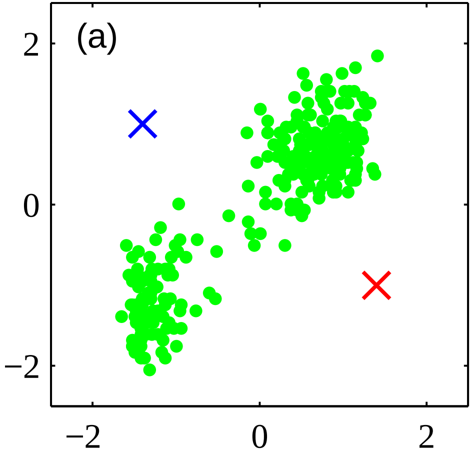
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2 \\ 0 & \text{otherwise} \end{cases}$$

  2. Minimize J w.r.t $\boldsymbol{\mu}_k$, keep r_{nk} fixed (Maximization)

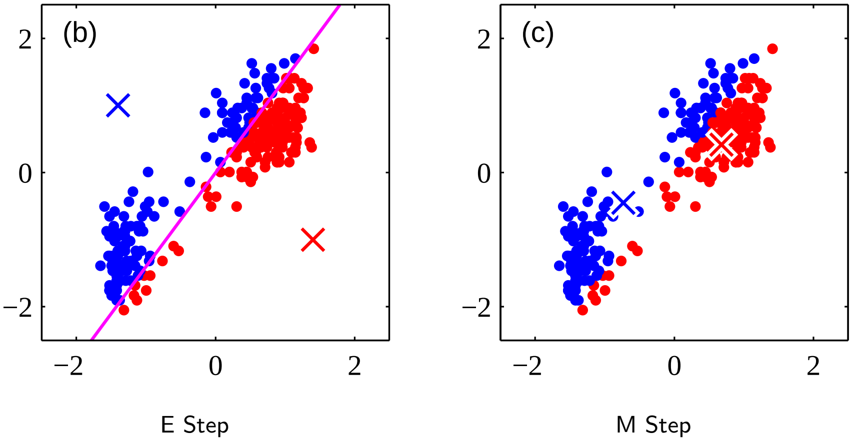
$$2 \sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0$$

$$\Longrightarrow \boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \boldsymbol{x}_n}{\sum_n r_{nk}}$$

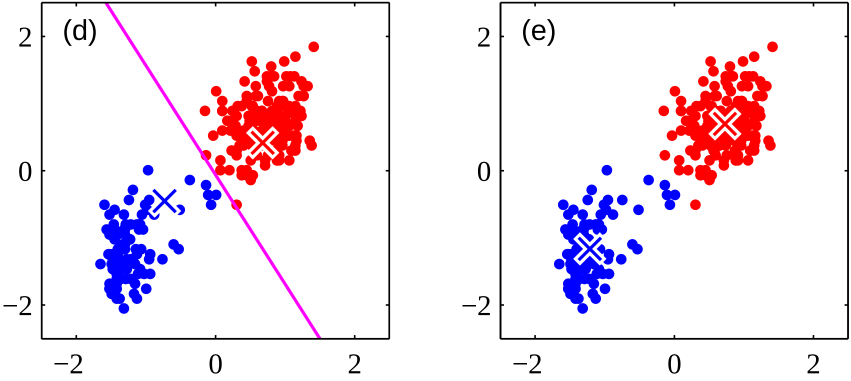
K-Means Clustering: Example



K-Means Clustering: Example cont.



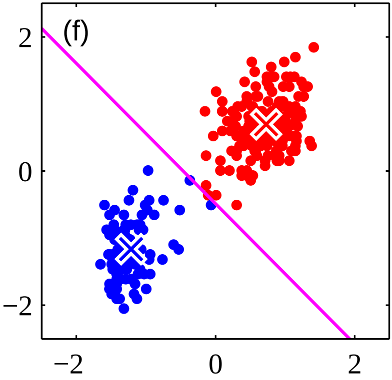
K-Means Clustering: Example cont.



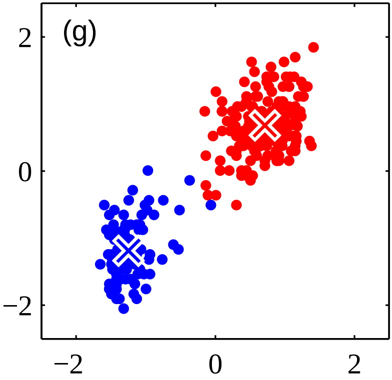
E Step

M Step

K-Means Clustering: Example cont.

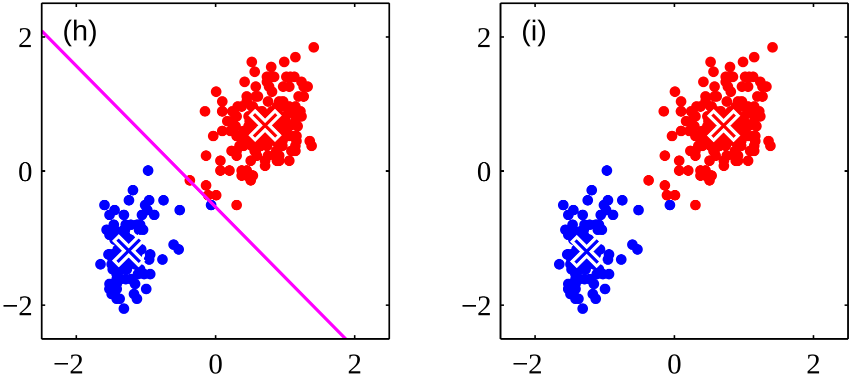


E Step



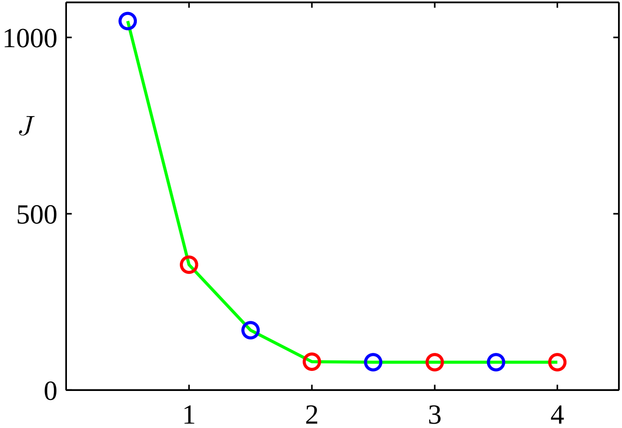
M Step

K-Means Clustering: Example cont.



E Step   M Step

K-Means Clustering: Example cont.



- Each E or M step reduces the value of the objective function J
- Convergence to a global or local maximum

K-Means Clustering: Concluding Remarks

1. Direct implementation of K-Means can be slow
2. Online version:

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \eta_n(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{\text{old}})$$

3. K-mediods, general distortion measure

$$\tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\boldsymbol{x}_n, \boldsymbol{\mu}_k)$$

   where $\mathcal{V}(\cdot)$ is any kind of dissimilarity measure.

4. Hard assignment: $r_{nk} = 1$ and $r_{nj} = 0$ for $j \neq k$.

Image Segmentation and Compression Example

Original image (96,615 colors)

100%

Image Segmentation and Compression Example

Quantized image (10colors, K-Means)

23.2%

Image Segmentation and Compression Example

Quantized image (3colors, K-Means)

8.7%

Image Segmentation and Compression Example

Quantized image (2colors, K-Means)

4.1%

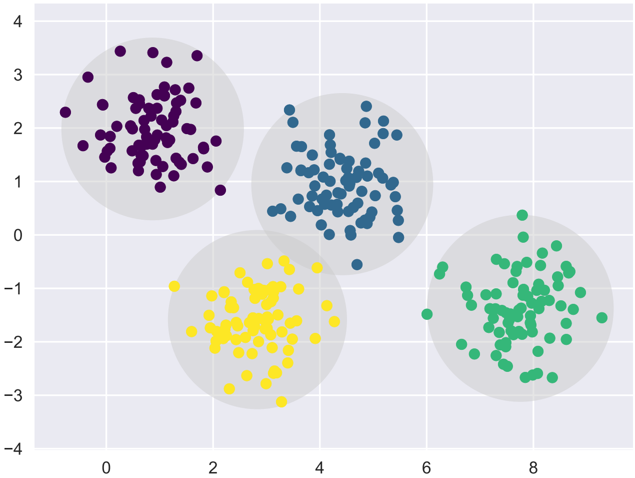
Outline

Weakness of K-Means

- Non-probabilistic nature
- Use simple distance-from-cluster-center to assign cluster membership

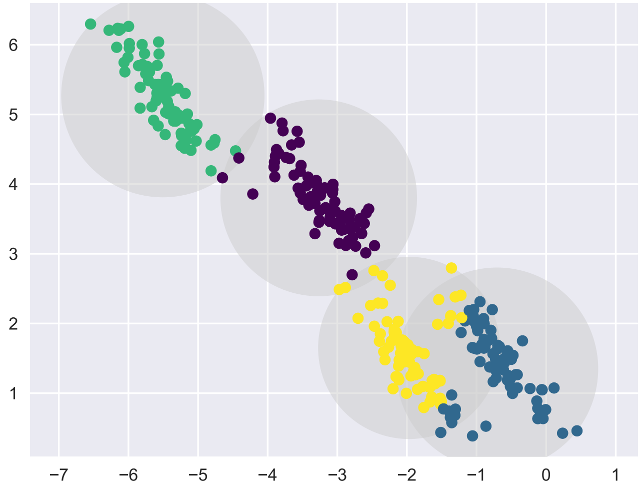
Weakness of K-Means

- Non-probabilistic nature
- Use simple distance-from-cluster-center to assign cluster membership



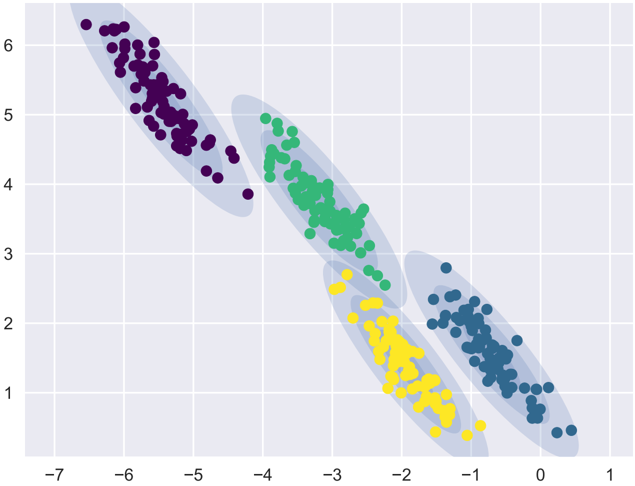
Weakness of K-Means

- Non-probabilistic nature
- Use simple distance-from-cluster-center to assign cluster membership



Weakness of K-Means

- Non-probabilistic nature
- Use simple distance-from-cluster-center to assign cluster membership



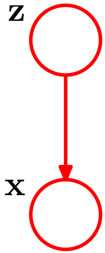
Mixture of Gaussians: Latent Variables

- Gaussian Mixture Distribution:

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

  where
  $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$

- Introduce latent variable $\boldsymbol{z}$



1. $\boldsymbol{z}$ is binary 1-of-K coding variable
2. $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$

Mixture of Gaussians: Conditional Distribution

- The marginal distribution of $\boldsymbol{z}$ (prior distribution)

$$p(z_k = 1) = \pi_k$$

where $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$

- Since $\boldsymbol{z}$ uses a 1-of-K representation

$$p(\boldsymbol{z}) = \prod_k \pi_k^{z_k}$$

Mixture of Gaussians: Conditional Distribution

- The marginal distribution of $\boldsymbol{z}$ (prior distribution)

$$p(z_k = 1) = \pi_k$$

where $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$

- Since $\boldsymbol{z}$ uses a 1-of-K representation

$$p(\boldsymbol{z}) = \prod_k \pi_k^{z_k}$$

- Given a particular value of $\boldsymbol{z}$, the conditional distribution of $\boldsymbol{x}$ is a Gaussian

$$p(\boldsymbol{x}|z_k = 1) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Another form of the above conditional distribution

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

Mixture of Gaussians: Marginal Distribution

- Joint distribution

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$$

- Marginal distribution

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

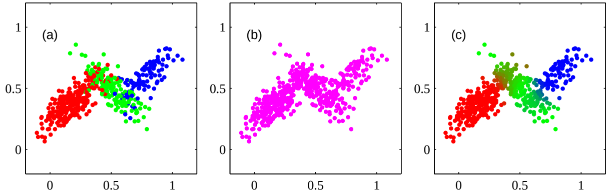
- The use of the joint probability $p(\boldsymbol{x}, \boldsymbol{z})$, leads to significant simplifications.

Mixture of Gaussians: Posterior Distribution

- The responsibility of component k to generate observation $\boldsymbol{x}$

$$\gamma(z_k) \equiv p(z_k = 1|\boldsymbol{x}) = \frac{p(z_k = 1)p(\boldsymbol{x}|z_k = 1)}{\sum_k p(z_k = 1)p(\boldsymbol{x}|z_k = 1)}$$
$$= \frac{\pi_k p(\boldsymbol{x}|z_k = 1)}{\sum_k \pi_k p(\boldsymbol{x}|z_k = 1)}$$

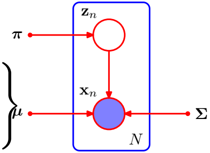
- Generate random samples with ancestral sampling
    1. $\hat{\boldsymbol{z}} \sim p(\boldsymbol{z})$
    2. $\boldsymbol{x} \sim p(\boldsymbol{x}|\hat{\boldsymbol{z}})$



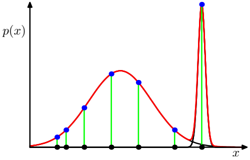
Mixture of Gaussians: Posterior Distribution

- Log Likelihood

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$



- Singularity: when a mixture component collapses on a datapoint
- Identifiability: for a ML solution in a K-component mixture there are $K!$ equivalent solutions.



EM for Gaussian Mixtures

- Maximum of log likelihood: derivatives of $\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ w.r.t parameters to 0.
- For the $\boldsymbol{\mu}_k$:

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

$$\Longrightarrow \boldsymbol{\mu}_k = \frac{1}{\sum_n \gamma(z_{nk})} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n$$

- For the $\boldsymbol{\Sigma}_k$:

$$\boldsymbol{\Sigma}_k = \frac{1}{\sum_n \gamma(z_{nk})} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x} - \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)^T$$

EM for Gaussian Mixtures cont.

- For the π_k:
  1. Take account of the constraint $\sum_k \pi_k = 1$
  2. Lagrange multiplier

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

$$\implies 0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$$

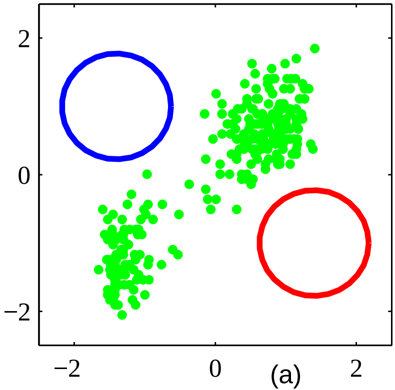
$$\implies \lambda \sum_{k=1}^{K} \pi_k = -\sum_{n=1}^{N} \frac{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = -N$$

$$\implies \pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} \equiv N_k$$

- N_k can be interpreted as the effective number of points assigned to cluster k.

EM for Gaussian Mixtures: Example

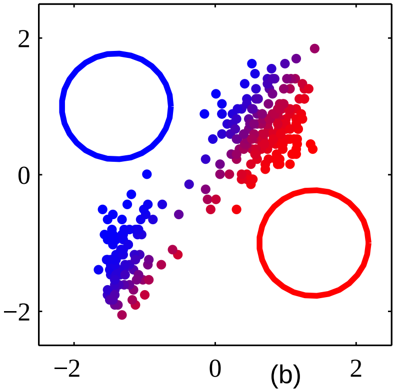
- No closed form solutions: $\gamma(z_{nk})$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood
- Alternate between estimating the current $\gamma(z_{nk})$ and updating the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$



(a)

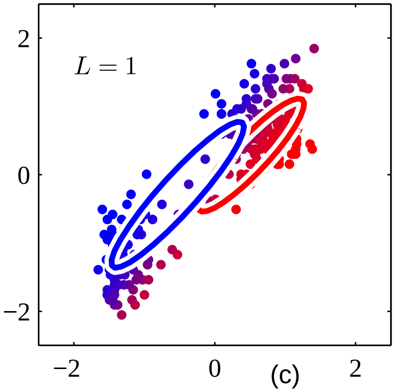
EM for Gaussian Mixtures: Example

- No closed form solutions: $\gamma(z_{nk})$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood
- Alternate between estimating the current $\gamma(z_{nk})$ and updating the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$



EM for Gaussian Mixtures: Example

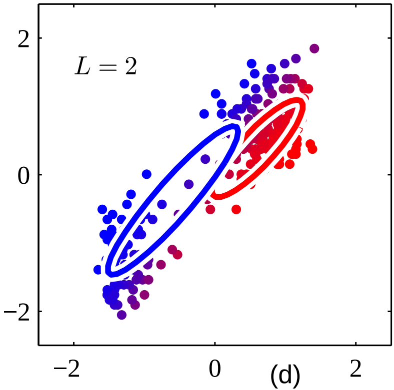
- No closed form solutions: $\gamma(z_{nk})$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood
- Alternate between estimating the current $\gamma(z_{nk})$ and updating the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$



(c)

EM for Gaussian Mixtures: Example

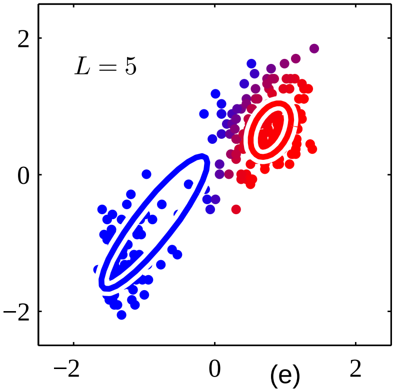
- No closed form solutions: $\gamma(z_{nk})$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood
- Alternate between estimating the current $\gamma(z_{nk})$ and updating the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$



(d)

EM for Gaussian Mixtures: Example

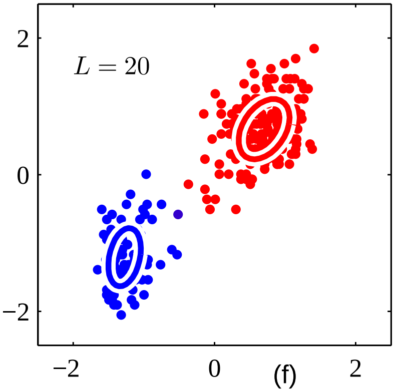
- No closed form solutions: $\gamma(z_{nk})$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood
- Alternate between estimating the current $\gamma(z_{nk})$ and updating the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$



(e)

EM for Gaussian Mixtures: Example

- No closed form solutions: $\gamma(z_{nk})$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood
- Alternate between estimating the current $\gamma(z_{nk})$ and updating the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$



(f)

EM for Gaussian Mixtures: Summary

1. Initialize $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$ and evaluate log-likelihood
2. E-Step Evaluate responsibilities

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

3. M-Step Re-estimate parameters, using current responsibilities:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\boldsymbol{x} - \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

4. Evaluate log-likelihood $lnp(\boldsymbol{X} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\})$ and check for convergence (go to step 2).

Outline

An Alternative View of EM: Latent Variables

- Let $\boldsymbol{X}$ observed data, $\boldsymbol{Z}$ latent variables, $\boldsymbol{\theta}$ parameters
- Goal: maximize marginal log-likelihood of observed data

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right\}$$

- Optimization problematic due to log-sum

An Alternative View of EM: Latent Variables

- Let $\boldsymbol{X}$ observed data, $\boldsymbol{Z}$ latent variables, $\boldsymbol{\theta}$ parameters
- Goal: maximize marginal log-likelihood of observed data

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right\}$$

- Optimization problematic due to log-sum
- Assume straightforward maximization for complete data

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

- Latent $\boldsymbol{Z}$ is known only through $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$
- Consider expectation of complete data log-likelihood

An Alternative View of EM: Algorithm

1. **Initialization:** Choose initial set of parameters $\boldsymbol{\theta}^{old}$
2. **E-step:** use current parameters $\boldsymbol{\theta}^{old}$ to compute $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$ to find expected complete-data log-likelihood for general $\boldsymbol{\theta}$

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

3. **M-step:** determine $\boldsymbol{\theta}^{new}$ by maximizing above formula

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

4. **Check convergence:** stop, or $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$ and go to **E-step**

An Alternative View of EM: Gaussian Mixtures Revisited

- Complete-data (log-) likelihood

$$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

- Expectation

$$\mathbb{E}[z_{nk}] = \gamma(z_{nk})$$

$$\mathbb{E}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

Outline

The EM algorithm in General

- Let $\boldsymbol{X}$ observed data, $\boldsymbol{Z}$ latent variables, $\boldsymbol{\theta}$ parameters
- Goal: maximize marginal log-likelihood of observed data

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right\}$$

- Maximization of $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ is simple, but difficult for $p(\boldsymbol{X}|\boldsymbol{\theta})$
- Given any $q(\boldsymbol{Z})$, we decompose the data log-likelihood

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q(\boldsymbol{Z}) \| p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}))$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}$$

$$KL(q(\boldsymbol{Z}) \| p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \geq 0^{1}$$

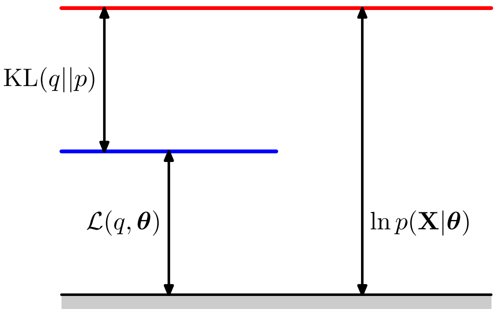
[1] Need Verification

The EM algorithm in General: The EM Bound

- $\mathcal{L}(q, \boldsymbol{\theta})$ **is a lower bound on the data log-likelihood**

$$\mathcal{L}(q, \boldsymbol{\theta}) = \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - KL(q(\boldsymbol{Z}) \| p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})) \leq \ln p(\boldsymbol{X}|\boldsymbol{\theta})$$

- **The The EM algorithm performs coordinate ascent on** $\mathcal{L}$
  1. E-step: maximizes $\mathcal{L}$ w.r.t. q for fixed $\boldsymbol{\theta}$
  2. M-step: maximizes $\mathcal{L}$ w.r.t. $\boldsymbol{\theta}$ for fixed q

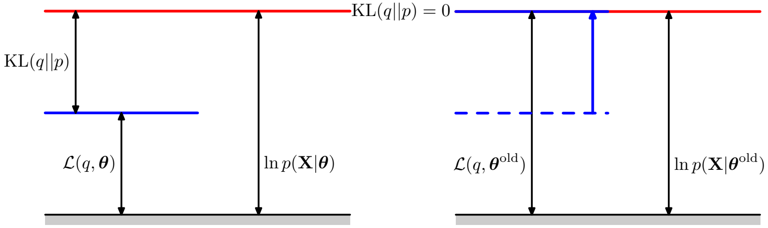


The EM algorithm in General: The E-step

- E-step: maximizes $\mathcal{L}$ w.r.t. q for fixed $\boldsymbol{\theta}$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - KL(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}))$$

- $\mathcal{L}$ maximized for $q(\boldsymbol{Z}) \leftarrow p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$

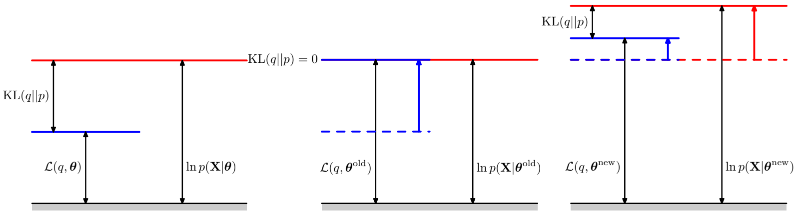


The EM algorithm in General: The M-step

- M-step: maximizes $\mathcal{L}(q, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ for fixed q

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{Z} q(\boldsymbol{Z}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) - \sum_{Z} q(\boldsymbol{Z}) \ln q(\boldsymbol{Z})$$

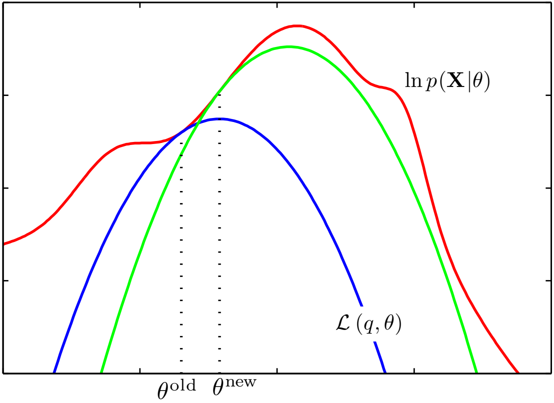
- $\mathcal{L}$ maximized for $\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$



The EM algorithm in General: Picture in Parameter Space

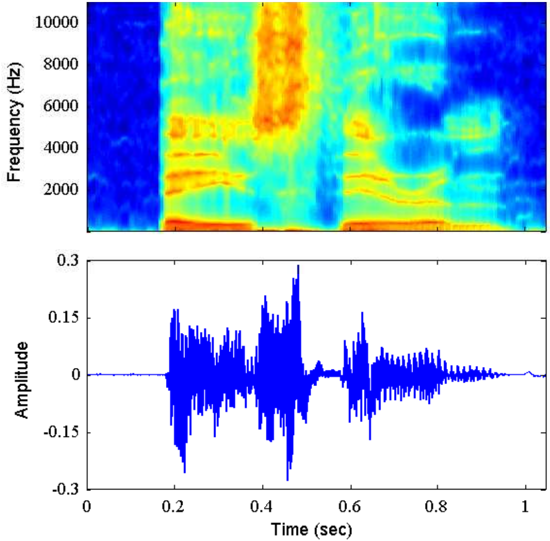
E-step resets bound $\mathcal{L}(q, \boldsymbol{\theta})$ on $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$, it is

- tight as $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$
- tangential at $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$
- convex in $\boldsymbol{\theta}$ for exponential family mixture components



Outline

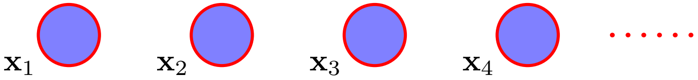
Motivation



Frequency (Hz)
10000
8000
6000
4000
2000
Amplitude
0.3
0.15
0
-0.15
-0.3
0
0.2
0.4
0.6
0.8
1
Time (sec)

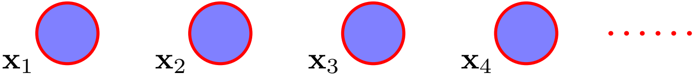
Markov Model

- Simplest way: model as independent:

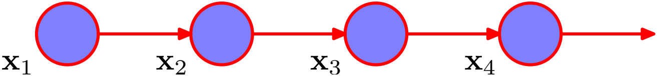


Markov Model

- Simplest way: model as independent:

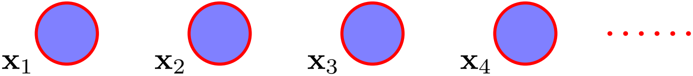


- Better to link observations, e.g. first-order Markov model, condition on previous observation

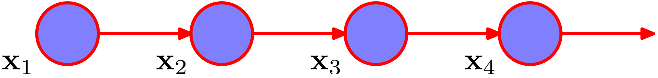


Markov Model

- Simplest way: model as independent:



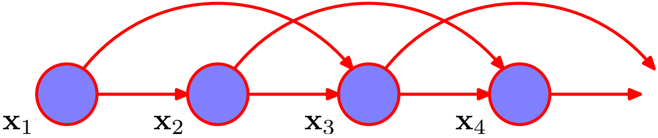
- Better to link observations, e.g. first-order Markov model, condition on previous observation



- Joint distribution

$$
\begin{aligned}
p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) &= \prod_{n=1}^{N} p(\boldsymbol{x}_n | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}) \\
&= p(\boldsymbol{x}_1) \prod_{n=2}^{N} p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1})
\end{aligned}
$$

Exercise 1



Q: Show that second-order Markov chain described by the joint distribution

$$p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) = p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1) \prod_{n=3}^{N} p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2})$$

satisfies the conditional independence property

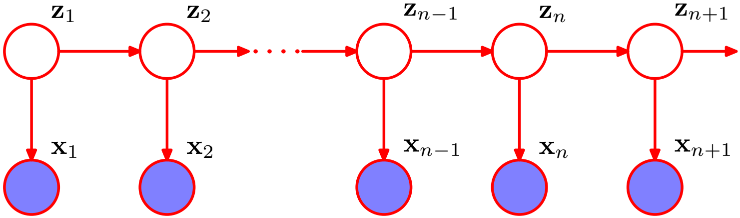
$$p(\boldsymbol{x}_n|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2})$$

Hidden Markov Model

- Introduce additional latent variables to permit a rich class of models to be constructed out of simple components

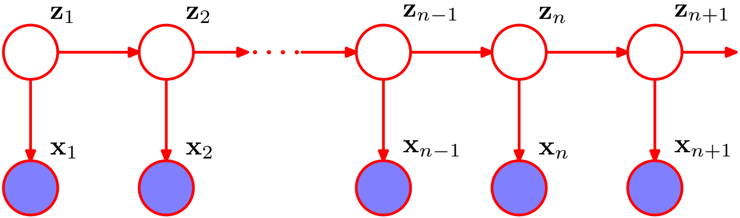
Hidden Markov Model

- Introduce additional latent variables to permit a rich class of models to be constructed out of simple components
- For each observation x_n, introduce a corresponding latent variable z_n



Hidden Markov Model

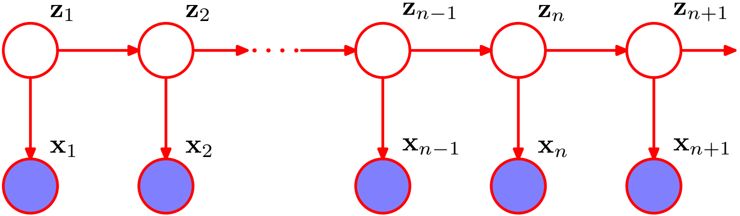
- Introduce additional latent variables to permit a rich class of models to be constructed out of simple components
- For each observation $\boldsymbol{x}_n$, introduce a corresponding latent variable $\boldsymbol{z}_n$



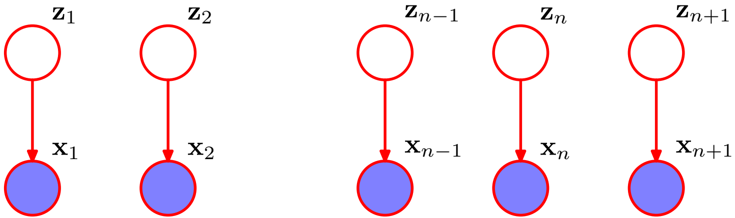
- Joint distribution

$$p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_N) = p(\boldsymbol{z}_1) \left[\prod_{n=2}^{N} p(\boldsymbol{z}_n | \boldsymbol{z}_{n-1}) \right] \prod_{n=1}^{N} p(\boldsymbol{x}_n | \boldsymbol{z}_n)$$

Relationship with Mixture Model

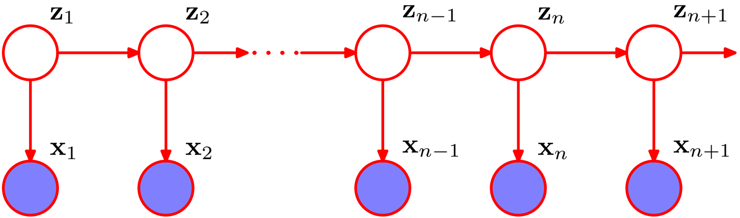


Relationship with Mixture Model



- Examining a single time slice of the model, it corresponds to a mixture distribution with component densities given by $p(\boldsymbol{x}|\boldsymbol{z})$

Relationship with Mixture Model



- Examining a single time slice of the model, it corresponds to a mixture distribution with component densities given by $p(\boldsymbol{x}|\boldsymbol{z})$
- HMM can be interpreted as an extension of a mixture model
- The choice of mixture component for each observation is not selected independently but depends on the choice of component for the previous observation

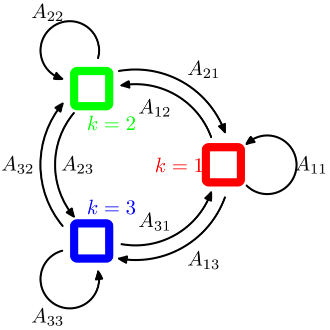
Transition Probabilities

- Elements of transition matrix

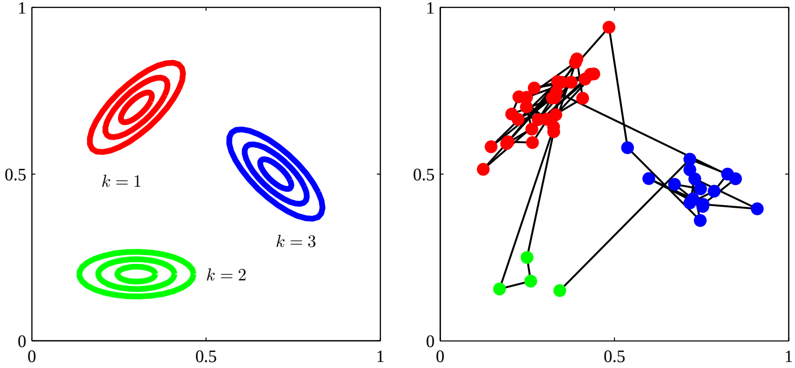
$$A_{jk} \equiv p(z_{nk} = 1 | z_{n-1,j} = 1)$$

- Constraints: $0 \leq A_{jk} \leq 1$ with $\sum_k A_{jk} = 1$
- Conditional distribution

$$p(\boldsymbol{z}_n | \boldsymbol{z}_{n-1}, \boldsymbol{A}) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j}, z_{nk}}$$



Sampling from HMMs



Emission Distribution

Successive Observations

Maximumm Likelihood for HMM

- **E-Step:** forward-backward algorithm

$$\gamma(z_{tk}) = \mathbb{E}[z_{tk}] = \sum_{\boldsymbol{z}} \gamma(\boldsymbol{z}) z_{tk}$$

$$\xi(z_{t-1,i}, z_{tj}) = \mathbb{E}[z_{t-1,i}, z_{tj}] = \sum_{\boldsymbol{z}} \gamma(\boldsymbol{z}) z_{t-1,i} z_{tj}$$

- **M-Step:** expected complete-data log likelihood function

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{t=2}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} \xi(z_{t-1,i}, z_{t,j}) \ln A_{ij} \\ &\quad + \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma(z_{tk}) \ln p(\boldsymbol{x}_t|\boldsymbol{\phi}_k) \end{aligned}$$

Exercise 2

Q: Verify the M-step equations

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{i=1}^{K} \gamma(z_{1i})}$$

$$A_{ij} = \frac{\sum_{t=2}^{T} \xi(z_{t-1,i}, z_{tj})}{\sum_{k=1}^{K} \sum_{t=2}^{T} \xi(z_{t-1,i}, z_{tk})}$$

by maximization of the expected complete-data log likelihood function, using appropriate Lagrange multipliers to enforce the summation constraints on the components of $\boldsymbol{\pi}$ and $\boldsymbol{A}$.