Parametric Density Estimation using Gaussian Mixture Models

An Application of the EM Algorithm

Based on tutorials by Jeff A. Bilmes and by Ludwig Schwardt

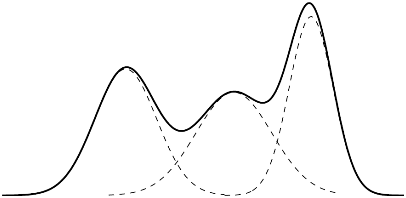
Pardis Noorzad – Amirkabir University of Technology – Bahman 12, 1389

Outline

What does EM do?

- **iterative** maximization of the likelihood function
  - when data has missing values or
  - when maximization of the likelihood is difficult
  - **data augmentation**
- $\mathbf{X}$ is incomplete data
- $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ is the complete data set

EM and MLE



- **maximum likelihood estimation**
  - estimates density parameters
  - $\Longrightarrow$ EM app: **parametric density estimation**
  - when density is a **mixture of Gaussians**

The density estimation problem

Definition

Our **parametric density estimation** problem:

$\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$

$f(\mathbf{x}_i|\mathbf{\Theta})$, i.i.d. assumption

$f(\mathbf{x}_i|\mathbf{\Theta}) = \sum_{k=1}^K \alpha_k p(\mathbf{x}_i|\theta_k)$

K components (given)

$\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K)$

Constraint on mixing probabilities α_i

Sum to unity

$$\int_{\mathbb{R}^D} p(\mathbf{x}_i|\theta_k) d\mathbf{x}_i = 1$$

$$\begin{aligned} 1 &= \int_{\mathbb{R}^D} f(\mathbf{x}_i|\boldsymbol{\Theta}) d\mathbf{x}_i \\ &= \int_{\mathbb{R}^D} \sum_{k=1}^{K} \alpha_k p(\mathbf{x}_i|\theta_k) d\mathbf{x}_i \\ &= \sum_{k=1}^{K} \alpha_k. \end{aligned}$$

The maximum likelihood problem

Definition

Our **MLE** problem:

$p(\mathbf{X}|\mathbf{\Theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{\Theta}) = \mathcal{L}(\mathbf{\Theta}|\mathbf{X})$

$\mathbf{\Theta}^* = \arg\max_{\mathbf{\Theta}} \mathcal{L}(\mathbf{\Theta}|\mathbf{X})$

the log-likelihood: $\log(\mathcal{L}(\mathbf{\Theta}|\mathbf{X}))$

$\log(\mathcal{L}(\mathbf{\Theta}|\mathbf{X})) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \alpha_k p(\mathbf{x}_i, \theta_k)\right)$

difficult to optimize b/c it contains log of sum

The EM formulation

Definition

Our **EM** problem:

$$\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$$

$$\mathcal{L}(\mathbf{\Theta}|\mathbf{Z}) = \mathcal{L}(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y}) = p(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta})$$

$$y_i \in 1 \ldots K$$

$y_i = k$ if the ith sample was generated by the kth component

$\mathbf{Y}$ is a random vector

$$\log(\mathcal{L}(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y})) = \sum_{i=1}^{N} \log(\alpha_{y_i} p(\mathbf{x}_i|\theta_{y_i}))$$

The EM formulation
Continued

Definition

The E- and M-steps with the auxiliary function Q:

E-step: $Q(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)}) = E[\log(p(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}))|\mathbf{X}, \mathbf{\Theta}^{(i-1)}]$

M-step: $\mathbf{\Theta}^{(i)} = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)})$

The EM formulation

The Q function

$$
\begin{aligned}
Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) &= E[\log(p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Theta}))|\mathbf{X}, \boldsymbol{\Theta}^{(t)}] \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N} \log(\alpha_k p(\mathbf{x}_i|\theta_k))p(k|\mathbf{x}_i, \boldsymbol{\Theta}^{(t)}) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N} \log(\alpha_k)p(k|\mathbf{x}_i, \boldsymbol{\Theta}^{(t)}) \\
&+ \sum_{k=1}^{K}\sum_{i=1}^{N} \log(p(\mathbf{x}_i|\theta_k))p(k|\mathbf{x}_i, \boldsymbol{\Theta}^{(t)})
\end{aligned}
$$

The EM formulation

Solving for α_k

We use the Lagrange multiplier to enforce $\sum_k \alpha_k = 1$.

$$\frac{\partial}{\partial \alpha_k}[\sum_k \sum_i \log(\alpha_k) p(k|\mathbf{x}_i, \boldsymbol{\Theta}^{(t)}) + \lambda(\sum_k \alpha_k - 1)] = 0$$

$$\sum_i \frac{1}{\alpha_k} p(k|\mathbf{x}_i, \boldsymbol{\Theta}^{(t)}) + \lambda = 0$$

Summing both sides over k, we get $\lambda = -N$.

$$\alpha_k = \frac{1}{N} \sum_i p(k|\mathbf{x}_i, \boldsymbol{\Theta}^{(t)}).$$

GMM

Solving for $\mathbf{m}_k$ and $\mathbf{\Sigma}_k$

Gaussian components

$$p(\mathbf{x}_i|\mathbf{x}_k, \mathbf{\Sigma}_k) = (2\pi)^{-\frac{D}{2}}|\mathbf{\Sigma}_k|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{x}_i - \mathbf{m}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}_i - \mathbf{m}_k))$$

$$\mathbf{m}_k = \frac{\sum_i \mathbf{x}_i p(k|\mathbf{x}_i, \mathbf{\Theta}^{(t)})}{\sum_i p(k|\mathbf{x}_i, \mathbf{\Theta}^{(t)})}$$

$$\mathbf{\Sigma}_k = \frac{p(k|\mathbf{x}_i, \mathbf{\Theta}^{(t)})(\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T}{p(k|\mathbf{x}_i, \mathbf{\Theta}^{(t)})}$$

Choice of Σ_k
Full covariance

Fits data best, but costly in high-dimensional space.

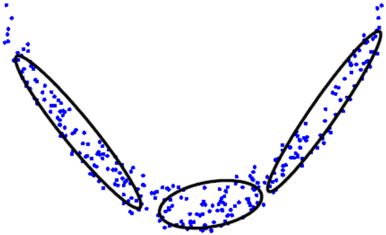


Figure: full covariance

Choice of Σ_k
Diagonal covariance

A tradeoff between cost and quality.

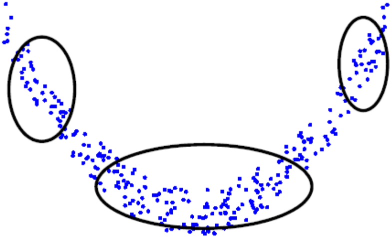


Figure: diagonal covariance

Choice of Σ_k

Spherical covariance, $\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}$

Needs many components to cover data.

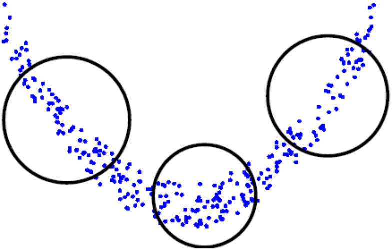


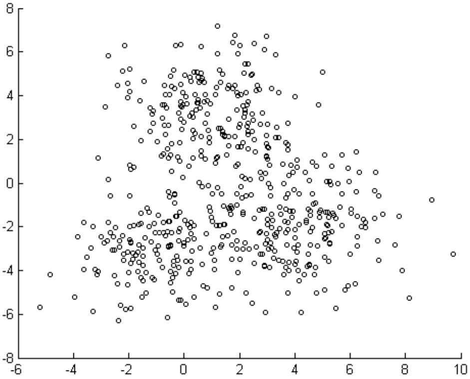
Figure: spherical covariance

Decision should be based on the size of training data set

- so that all parameters may be tuned

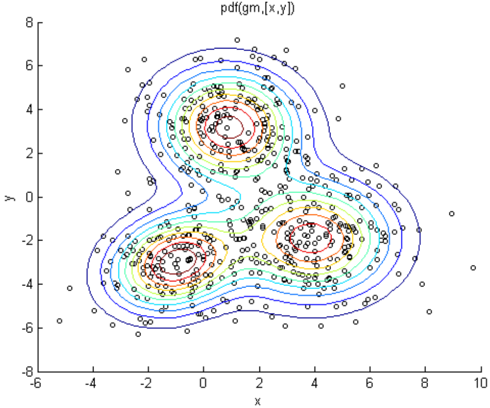
Data

data generated from a mixture of three bivariate Gaussians



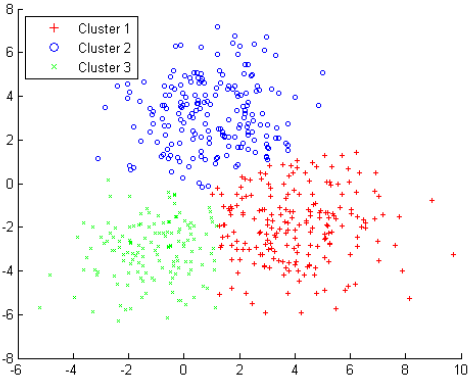
PDFs

estimated pdf contours (after 43 iterations)



Clusters

$\mathbf{x}_i$ assigned to the cluster k corresponding to the highest $p(k|\mathbf{x}_i, \mathbf{\Theta})$



A problem with missing data
Life-testing experiment

- lifetime of light bulbs follows an exponential distribution with mean θ
- two separate experiments
- N bulbs tested until failed
  - failure times recorded as $x_1, \ldots, x_N$
- M bulbs tested
  - failure times not recorded
  - only number of bulbs r, that failed at time t
  - missing data are failure times $u_1, \ldots, u_M$

EM formulation

Life-testing experiment

- $\log(\mathcal{L}(\theta|x,u)) = -N(\log\theta + \bar{x}/\theta) - \sum_i^M (\log\theta + u_i/\theta)$

- expected value for bulb still burning: $t + \theta$

- one that burned out: $\theta - \frac{te^{-t}/\theta^{(k)}}{1-e^{-t}/\theta^{(k)}} = \theta - th^{(k)}$

- E-step:

$$\begin{aligned} Q(\theta,\theta^{(k)}) &= -(N+M)\log\theta \\ &\quad - \frac{1}{\theta}(N\bar{x} + (M-r)(t+\theta^{(k)}) + r(\theta^{(k)} - th^{(k)})) \end{aligned}$$

- M-step: $\frac{1}{N+M}(N\bar{x} + (M-r)(t+\theta^{(k)}) + r(\theta^{(k)} - th^{(k)}))$

Thank you for your attention. Any questions?