Gaussian Mixture Model and EM (Expectation Maximization) Algorithm

Changshui Zhang
Dept. of Automation
Tsinghua University

zcs@mail.tsinghua.edu.cn

Reference

- 新教材电子版
- Jeff A. Bilmes, A Gentle Tutorial of the Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models

GMM (Gaussian Mixture Model)

- Observed data are *N* samples independently generated from the following probabilistic model

$$P(X|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(X|\theta_i)$$

where $\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$ and $\sum_{i=1}^{M} \alpha_i = 1$

$$p\left(x \middle| \omega_j, \theta_j\right) \sim N(\mu_i, \Sigma_i)$$

GMM (Gaussian Mixture Model)

- Likelihood Function and log Likelihood Function

$$p(X|\theta) = \prod_{i=1}^{N} \sum_{j=1}^{c} N\left(x_i \middle| \mu_j, \Sigma_j\right) P(\omega_j)$$

$$\ln p(X|\theta) = \sum_{i=1}^{N} \ln \sum_{j=1}^{c} N\left(x_i \middle| \mu_j, \Sigma_j\right) P(\omega_j)$$

$$P(X|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(X|\theta_i)$$

GMM (Gaussian Mixture Model)

$$\ln p(X|\theta) = \sum_{i=1}^{N} \ln \sum_{j=1}^{c} N\left(x_i \middle| \mu_j, \Sigma_j\right) P(\omega_j)$$

$$\frac{\partial \ln p(X|\theta)}{\partial \mu_k} = 0$$

$$\sum_{i=1}^{N} \frac{N(x_i|\mu_k, \Sigma_k) P(\omega_k)}{\sum_{j=1}^{c} N\left(x_i \middle| \mu_j, \Sigma_j\right) P\left(\omega_j\right)} \Sigma_k^{-1}(x_i - \mu_k) = 0$$

GMM (Gaussian Mixture Model)

$$\sum_{i=1}^{N} \frac{N(x_i|\mu_k,\Sigma_k)P(\omega_k)}{\sum_{j=1}^{c} N\left(x_i\middle|\mu_j,\Sigma_j\right)P\left(\omega_j\right)} \Sigma_k^{-1}(x_i-\mu_k)=0$$

$$P(\omega_k|x_i,\mu_k,\Sigma_k)=\frac{N(x_i|\mu_k,\Sigma_k)P(\omega_k)}{\sum_{j=1}^{c} N\left(x_i\middle|\mu_j,\Sigma_j\right)P\left(\omega_j\right)}$$

$$\mu_k=\frac{\sum_{i=1}^{N} P(\omega_k|x_i,\mu_k,\Sigma_k)x_i}{\sum_{i=1}^{N} P(\omega_k|x_i,\mu_k,\Sigma_k)}$$

GMM (Gaussian Mixture Model)

$$P(\omega_k|x_i,\mu_k,\Sigma_k) = \frac{N(x_i|\mu_k,\Sigma_k)P(\omega_k)}{\sum_{j=1}^{c} N\left(x_i\middle|\mu_j,\Sigma_j\right)P(\omega_j)}$$

$$\mu_k = \frac{\sum_{i=1}^{N} P(\omega_k|x_i,\mu_k,\Sigma_k)x_i}{\sum_{i=1}^{N} P(\omega_k|x_i,\mu_k,\Sigma_k)}$$

$$P(\omega_k) = \frac{1}{N}\sum_{i=1}^{N} P(\omega_k|x_i,\mu_k,\Sigma_k)$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} P(\omega_k|x_i,\mu_k,\Sigma_k)(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{N} P(\omega_k|x_i,\mu_k,\Sigma_k)}$$

Parameter estimation for GMM

- Initialize: $\mu_k^{new}, \Sigma_k^{new}, P^{new}(\omega_k)$
- E step: $\mu_k = \mu_k^{new}, \Sigma_k = \Sigma_k^{new}, P(\omega_k) = P^{new}(\omega_k)$

$$P(\omega_k|x_i, \mu_k, \Sigma_k) = \frac{N(x_i|\mu_k, \Sigma_k)P(\omega_k)}{\sum_{j=1}^{c} N\left(x_i|\mu_j, \Sigma_j\right)P\left(\omega_j\right)}$$

- M step: $P^{new}(\omega_k) = \frac{1}{N}\sum_{i=1}^{N} P(\omega_k|x_i, \mu_k, \Sigma_k)$

$$\mu_k^{new} = \frac{\sum_{i=1}^{N} P(\omega_k|x_i, \mu_k, \Sigma_k)x_i}{\sum_{i=1}^{N} P(\omega_k|x_i, \mu_k, \Sigma_k)}$$

$$\Sigma_k^{new} = \frac{\sum_{i=1}^{N} P(\omega_k|x_i, \mu_k, \Sigma_k)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} P(\omega_k|x_i, \mu_k, \Sigma_k)}$$

Maximum Likelihood

- Problem:

The set of parameters: Θ

A data set: $X = \{x_1, x_2, \dots, x_N\}$

A density function: $p(x|\Theta)$

Likelihood function:

$$L(\Theta|X) := p(X|\Theta) = \prod_{i=1}^{N} p(x_i|\Theta)$$

where the data is fixed.

Task: find $\Theta^* = arg\max_{\Theta} L(\Theta|X)$

Incomplete data

- Missing values

  Empty items in forms

  Forgotten data

- Hidden variables

  Can not be measured and observed.

Basic EM

- An elaborate technique of finding the maximum likelihood estimate of the parameters of a distribution from a given data set when the data is incomplete or has missing values.

- Reference: A.P.Dempster, N.M.Laird, and D.B.Rubin. Maximum-likelihood from incomplete data via the em algorithm. J.Royal Statist. Soc. Ser. B., 39, 1977

- Two main applications of the EM algorithm:

--The data indeed has missing values, due to problems with or limitations of the observation process

--The optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing (or hidden) parameters.

- We assume that data $X = \{x_1, x_2, \dots, x_N\}$ is observed and incomplete data.
- A complete data set: $Z = (X, Y)$
- A joint density function:

$$\mathrm{P}(Z|\Theta) = p(X, Y|\Theta) = p(Y|X, \Theta)p(X|\Theta)$$

- New likelihood function: complete data likelihood

$$L(\Theta|X,Y) = L(\Theta|X,Y) = p(X,Y|\Theta)$$

- It is a random variable since Θ, X

are constant and Y is a random variable.

$$L(\theta) = \log p(X|\theta) = \log\left(\sum_{Y} p(X,Y|\theta)\right)$$

Maximizing the log likelihood directly is often difficult because the log of the sum can potentially couple all of the parameters of the model.

Any distribution $q(y)$ over the hidden variables defines a lower bound on L:

$$L(\theta) \geq \sum_{y} q(y) \log p(x, y|\theta) - \sum_{y} q(y) \log q(y) = F(q, \theta)$$

$$L(\theta) = \log \sum_y p(x, y|\theta) = \log \sum_y q(y) \frac{p(x, y|\theta)}{q(y)}$$

$$\geq \sum_y q(y) \log \frac{p(x,y|\theta)}{q(y)}$$

$$= \sum_y q(y) \log p(x, y|\theta) - \sum_y q(y) \log q(y)$$

$$= F(q, \theta)$$

The inequality is known as Jensen's inequality

The EM Algorithm

The Expectation-Maximization algorithm alternates between maximizing F with respect to q and θ , respectively, holding the other fixed.

E step: $q_{[k+1]} \leftarrow \arg\max_{q} F(q, \theta_{[k]})$

M step: $\theta_{[k+1]} \leftarrow \arg\max_{\theta} F(q_{[k+1]}, \theta)$

The EM Algorithm

The maximum in the **E step** is obtained by setting: $q_{[k+1]}(y) = p(y|x, \theta_{[k]})$

$$F(q, \theta) = \sum_y q(y) \log \frac{p(x, y|\theta)}{q(y)}$$

$$= \sum_y p(y|x, \theta) \log \frac{p(y|x, \theta) p(x|\theta)}{p(y|x, \ \theta)}$$

$$= \sum_y p(y|x, \theta) \log p(x|\theta)$$

$$= \log p(x|\theta) = L(\theta) \qquad\qquad F(q, \theta) \leq L(\theta)$$

The EM Algorithm: E Step

$$q_{[k+1]}(y) = p(y|x, \theta_{[k]})$$

$$F\left(q_{[k+1]}, \theta\right) = \sum_y q_{[k+1]}(y) \log p(x, y|\theta) - \sum_y q_{[k+1]}(y) \log q_{[k+1]}(y)$$

$$Q\left(\theta_{[k]}, \theta\right) = \sum_y p(y|x, \theta_{[k]}) \log p(x, y|\theta) = E[\ln p(x, y|\theta)|x, \theta_{[k]}]$$

The EM Algorithm: M Step

- The maximum in the **M step** is obtained by maximizing the first term of $F(q, \theta)$

M step: $\theta_{[k+1]} \leftarrow \arg\max_{\theta} \sum_y p\left(y \middle| x, \theta_{[k]}\right) \log p(y, x|\theta)$

$$Q\left(\theta_{[k]}, \theta\right) = \sum_y p(y|x, \theta_{[k]}) \log p(x, y|\theta)$$
$$= E[\ln p(x, y|\theta)|x, \theta_{[k]}]$$

EM Algorithm

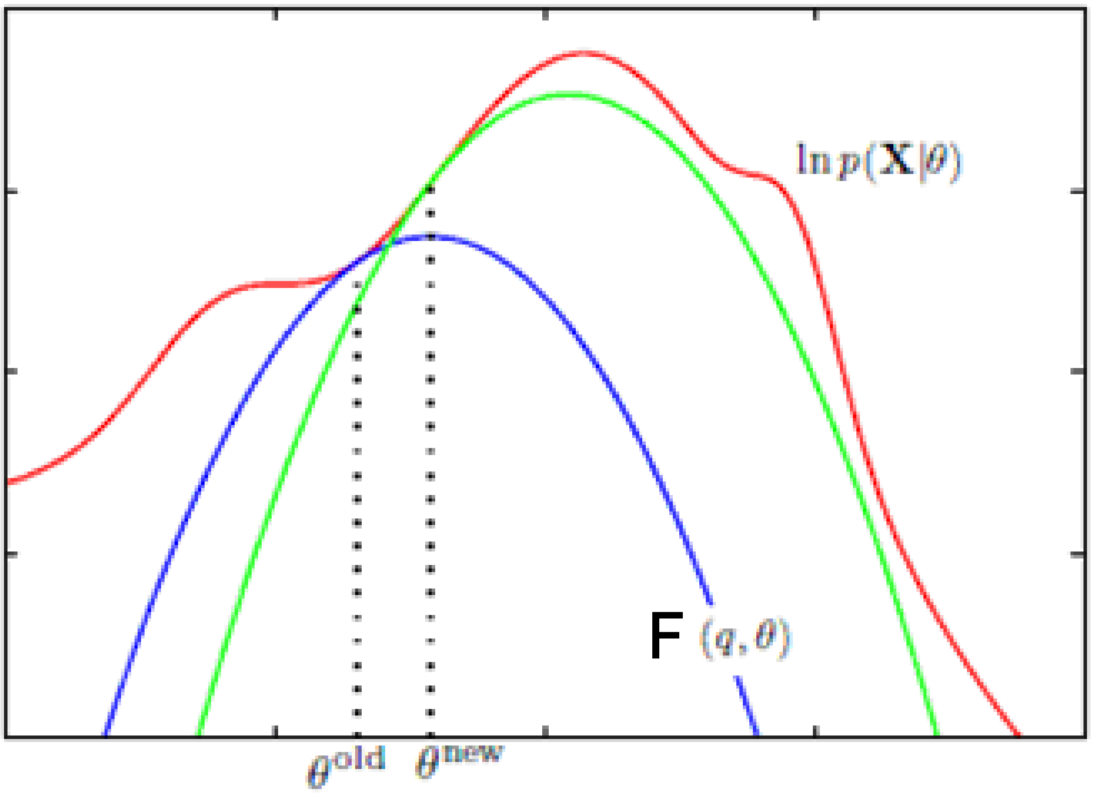
- Initialize $\theta_{[0]}$ with random/guess, set i=1
- E-step:

$$Q\left(\theta_{[k]}, \theta\right) = \sum_{y} p(y|x, \theta_{[k]}) \log p(x, y|\theta)$$
$$= E[\ln p(x, y|\theta)|x, \theta_{[k]}]$$

- M-step: $\theta_{[k+1]} \leftarrow \arg\max_{\theta} Q\left(\theta_{[k]}, \theta\right)$
- i = i+1
- repeat until convergence

EM Algorithm()

- The figure is Copied from book by Bishop



Generalized EM

- Assume $\ln p(X|\theta)$ and Q function are differentiable in θ. The EM likelihood converges to a point where $\frac{\partial}{\partial\theta}\ln p(X|\theta) = 0$

- GEM: Instead of setting $\theta_{[k+1]} = \arg\max_{\theta} Q\left(\theta_{[k]}, \theta\right)$
  Just find $\theta(n)$ such that
  $$Q\left(\theta_{[k]}, \theta_{[k+1]}\right) > Q(\theta_{[k]}, \theta)$$

- GEM also is guaranteed to converge

Finding Maximum Likelihood Mixture Densities Parameters via EM

- Observed data are *N* samples independently generated from the following probabilistic model

$$p(X|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(X|\theta_i)$$

where $\Theta = (\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M)$ and $\sum_{i=1}^{M} \alpha_i = 1$

Finding Maximum Likelihood Mixture Densities Parameters via EM

- The unknown parameters $y_i \in 1, \dots, M$ indicate that the i^{th} sample is generated by the y_i^{th} mixture component
- Though $\ln \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i=1}^{N} \ln(\sum_{j=1}^{M} \alpha_j p_j(x_i|\theta_j))$ is hard to compute, we have a much easier choice:

$$\ln(p(X,Y|\Theta)) = \sum_{i=1}^{N} \ln(P(x_i|y_i)P(y_i))$$
$$= \sum_{i=1}^{N} \ln(\alpha_{y_j} p_{y_j}(x_i|\theta_{y_j}))$$

E-step

- Guess an initial $\Theta^g = (\alpha_1^g, \ldots, \alpha_M^g, \theta_1^g, \ldots, \theta_M^g)$
- Expectation equation

$$Q(\Theta, \Theta^g) = \sum_Y \ln(p(X, Y|\Theta))p(Y|X, \Theta^g)$$
$$= \sum_y \sum_{i=1}^N \ln(\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i})) \prod_{j=1}^N p(y_j|x_j, \Theta^g)$$

$$Q\left(\theta_{[k]}, \theta\right) = \sum_y p(y|x, \theta_{[k]}) \log p(x, y|\theta)$$

E-step

$$\sum_{y}\sum_{i=1}^{N}\ln(\alpha_{y_i}p_{y_i}(x_i|\theta_{y_i}))\prod_{j=1}^{N}p(y_j|x_j,\Theta^g)$$

$$=\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{i=1}^{N}\ln(\alpha_{y_i}p_{y_i}(x_i|\theta_{y_i}))\prod_{j=1}^{N}p(y_j|x_j,\Theta^g)$$

$$=\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{i=1}^{N}\sum_{l=1}^{M}\delta_{l,y_i}\ln(\alpha_l p_l(x_i|\theta_l))\prod_{j=1}^{N}p(y_j|x_j,\Theta^g)$$

$$=\sum_{l=1}^{M}\sum_{i=1}^{N}\ln(\alpha_l p_l(x_i|\theta_l))\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\delta_{l,y_i}\prod_{j=1}^{N}p(y_j|x_j,\Theta^g)$$

E-step

$$\sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \delta_{l,y_i} \prod_{j=1}^{N} p(y_j|x_j, \Theta^g)$$

$$= \left(\sum_{y_1=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{j=1, j\neq i}^{N} p(y_j|x_j, \Theta^g) \right) p(l|x_j, \Theta^g)$$

$$= \prod_{j=1, j\neq i}^{N} \left(\sum_{y_j=1}^{M} p(y_j|x_j, \Theta^g) \right) p(l|x_j, \Theta^g)$$

$$= p(l|x_j, \Theta^g)$$

E-step

$$Q(\Theta,\Theta^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}\ln(\alpha_l p_l(x_i|\theta_l))p(l|x_j,\Theta^g)$$

$$=\sum_{l=1}^{M}\sum_{i=1}^{N}\ln(\alpha_l)p(l|x_i,\Theta^g)$$

$$+\sum_{l=1}^{M}\sum_{i=1}^{N}\ln\big(p_l(x_i|\theta_l)\big)p(l|x_i,\Theta^g)$$

- Note: The term containing θ_l and the term containing α_l are not related and can be maximized independently

M-step

- Maximize α_l: Introduce Lagrange multiplier λ

$$\frac{\partial}{\partial \alpha_l}[Q(\Theta, \Theta^g) + \lambda(\sum_l \alpha_l - 1)] = 0$$

$$\frac{\partial}{\partial \alpha_l}\left[\sum_{l=1}^{M} \sum_{i=1}^{N} \ln(\alpha_l) p(l|x_i, \Theta^g) + \lambda(\sum_l \alpha_l - 1)\right] = 0$$

$$\sum_{i=1}^{N} \frac{1}{\alpha_l} p(l|x_i, \Theta^g) + \lambda = 0$$

$$\sum_{i=1}^{N} p(l|x_i, \Theta^g) + \lambda \alpha_l = 0 \qquad l = 1, \dots, M$$

M-step

$$\sum_{i=1}^{N} p(l|x_i,\Theta^g) + \lambda\alpha_l = 0 \qquad l = 1,\dots,M$$

$$\sum_{l=1}^{M}\sum_{i=1}^{N} p(l|x_i,\Theta^g) + \sum_{l=1}^{M}\lambda\alpha_l = 0$$

$$N + \lambda = 0$$

$$\alpha_l = \frac{1}{N}\sum_{i=1}^{N} p(l|x_i,\Theta^g)$$

M-step

- Maximize θ_l : If $p(x|\theta)$ is d-dimensional Gaussian distributions with mean μ and covariance matrix Σ ,

  we can have an analytical expressions

  for μ and Σ.

M-step

$$B = \sum_{l=1}^{M} \sum_{i=1}^{N} \ln\left(p_l(x_i|\theta_l)\right) p(l|x_i, \Theta^g)$$

$$= \sum_{l=1}^{M} \sum_{i=1}^{N} \left(-\frac{\ln(|\Sigma_l|)}{2} - \frac{(x_i - \mu_l)^T \Sigma_l^{-1} (x_i - \mu_l)}{2} \right) p(l|x_i, \Theta^g)$$

$$\frac{\partial B}{\partial \mu_l} = 0, \qquad \sum_{i=1}^{N} \Sigma_l^{-1} (x_i - \mu_l) p(l|x_i, \Theta^g) = 0$$

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x, \qquad \mu_l = \frac{\sum_{i=1}^{N} x_i p(l|x_i, \Theta^g)}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)}$$

M-step

$$B = \sum_{l=1}^{M}\sum_{i=1}^{N}\left(-\frac{\ln(|\Sigma_l|)}{2} - \frac{(x_i-\mu_l)^T\Sigma_l^{-1}(x_i-\mu_l)}{2}\right)p(l|x_i,\Theta^g)$$

$$= \sum_{l=1}^{M}\left[\frac{\ln\left(|\Sigma_l^{-1}|\right)}{2}\sum_{i=1}^{N}p(l|x_i,\Theta^g) - \frac{1}{2}\sum_{i=1}^{N}p(l|x_i,\Theta^g)\, tr\left(\Sigma_l^{-1}(x_i-\mu_l)(x_i-\mu_l)^T\right)\right]$$

$$= \sum_{l=1}^{M}\left[\frac{\ln\left(|\Sigma_l^{-1}|\right)}{2}\sum_{i=1}^{N}p(l|x_i,\Theta^g) - \frac{1}{2}\sum_{i=1}^{N}p(l|x_i,\Theta^g)\, tr\left(\Sigma_l^{-1}N_{l,i}\right)\right]$$

$$N_{l,i} = (x_i-\mu_l)(x_i-\mu_l)^T$$

M-step

$$\frac{\partial B}{\partial \Sigma_l^{-1}} = 0$$

$$B = \sum_{l=1}^{M} \left[\frac{\ln(|\Sigma_l^{-1}|)}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^g) - \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^g)\, tr(\Sigma_l^{-1} N_{l,i}) \right]$$

$$\frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^g)(2\Sigma_l - diag(\Sigma_l)) - \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^g) \left(2N_{l,i} - diag\left(N_{l,i}\right)\right)$$

$$= \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^g) \left(2M_{l,i} - diag\left(M_{l,i}\right)\right)$$

$$= 2S - diag(S), \qquad M_{l,i} = \Sigma_l - N_{l,i}, \qquad S = \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^g) M_{l,i}$$

M-step

$$2S - diag(S) = 0$$

$$S = 0$$

$$S = \frac{1}{2}\sum_i^N p(l|x_i, \Theta^g)\left(\Sigma_l - N_{l,i}\right) = 0$$

$$\Sigma_l = \frac{\sum_{i=1}^N p(l|x_i, \Theta^g) N_{l,i}}{\sum_{i=1}^N p(l|x_i, \Theta^g)} = \frac{\sum_{i=1}^N p(l|x_i, \Theta^g)(x_i - \mu_l)(x_i - \mu_l)^T}{\sum_{i=1}^N p(l|x_i, \Theta^g)}$$

$$\frac{\partial \ln|A|}{\partial A} = 2A^{-1} - diag(A^{-1})$$

$$\frac{\partial tr(AB)}{\partial A} = B + B^T - diag(B)$$

M-step

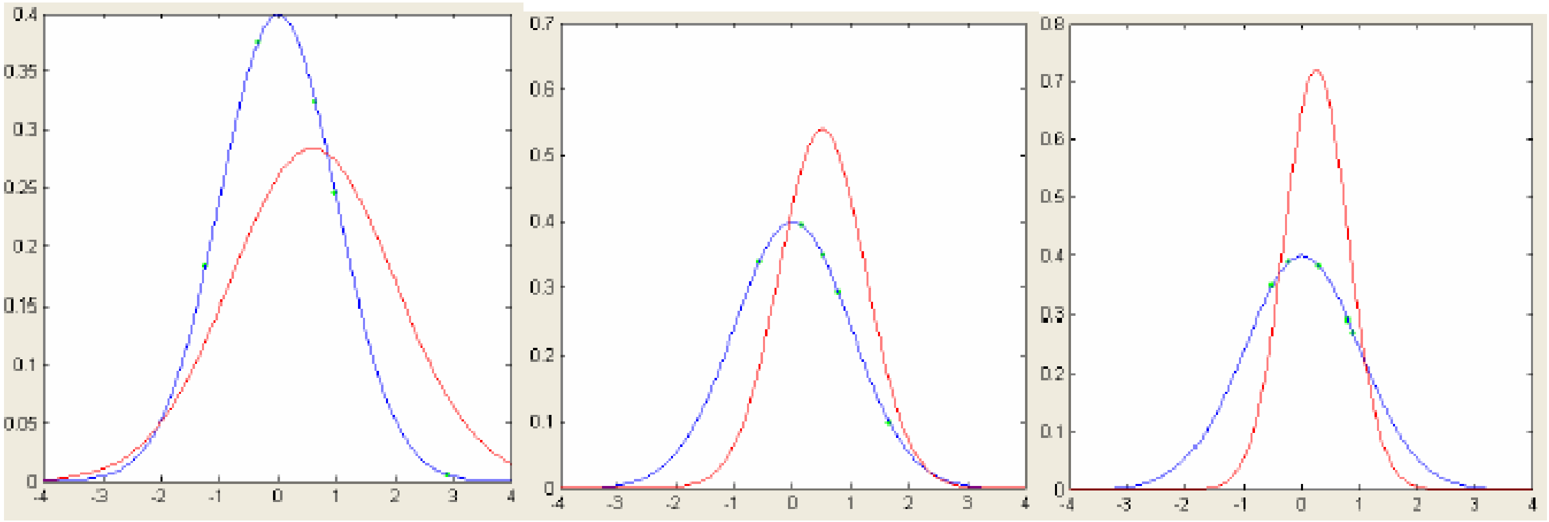
$$\alpha_l^{new} = \frac{1}{N}\sum_{i=1}^{N} p(l|x_i, \Theta^g)$$

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} x_i p(l|x_i, \Theta^g)}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)}$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} p(l|x_i, \Theta^g)(x_i - \mu_l^{new})(x_i - \mu_l^{new})^T}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)}$$

Experiments

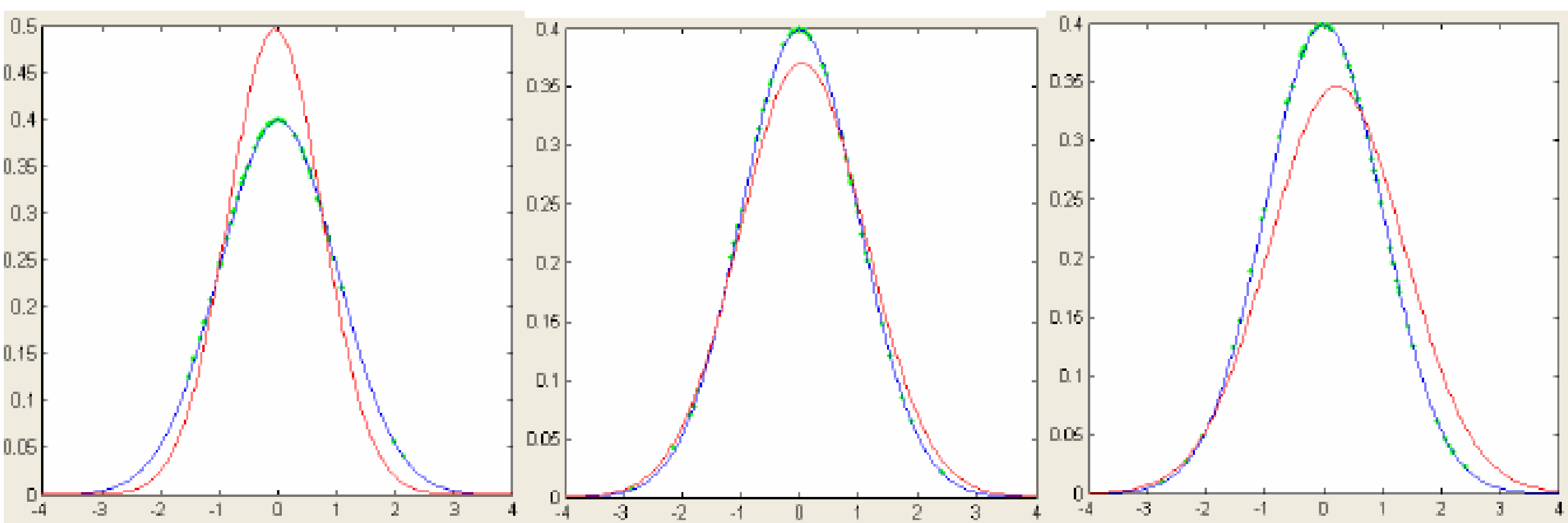
- 5 points from a Gaussian Distribution(3 times)
  - Blue: original distribution
  - Red: estimated distribution
  - Green: samples



(a) $\mu = 0.5910, \sigma^2 = 1.9690$ (b) $\mu = 0.5136, \sigma^2 = 0.5467$ (c) $\mu = 0.2471, \sigma^2 = 0.3051$

Experiments

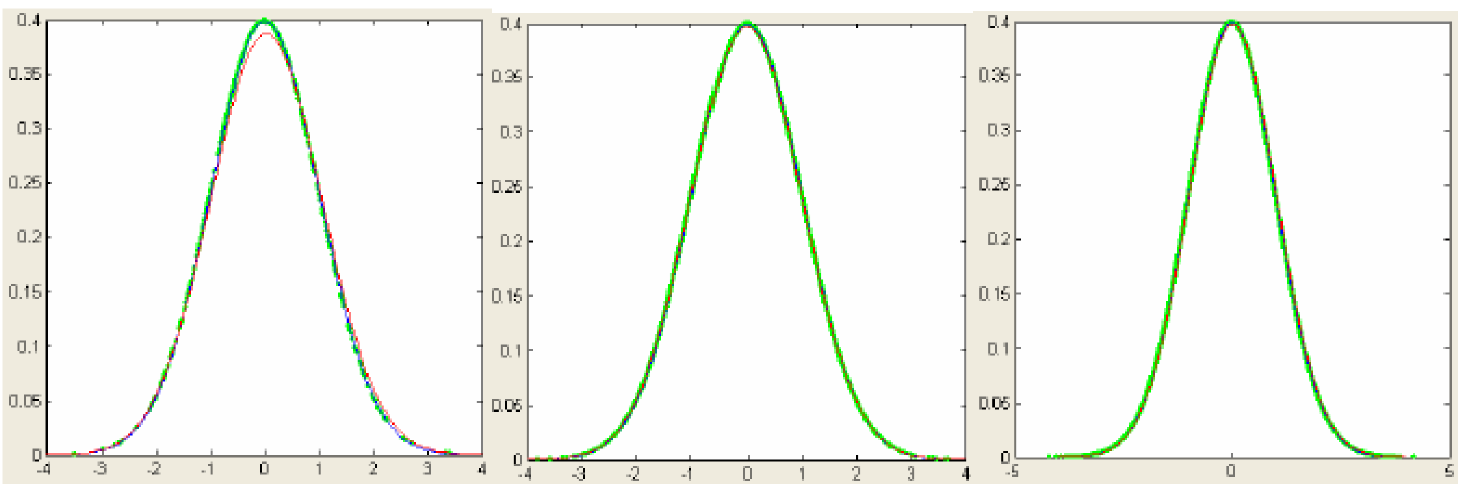
- 50 points from a Gaussian Distribution (3 times)
  - Blue: original distribution
  - Red: estimated distribution
  - Green: samples



(d) $\mu = 0.0560, \sigma^2 = 1.1030$ (e) $\mu = 0.1898, \sigma^2 = 1.1726$ (f) $\mu = 0.2174, \sigma^2 = 1.3243$

Experiments

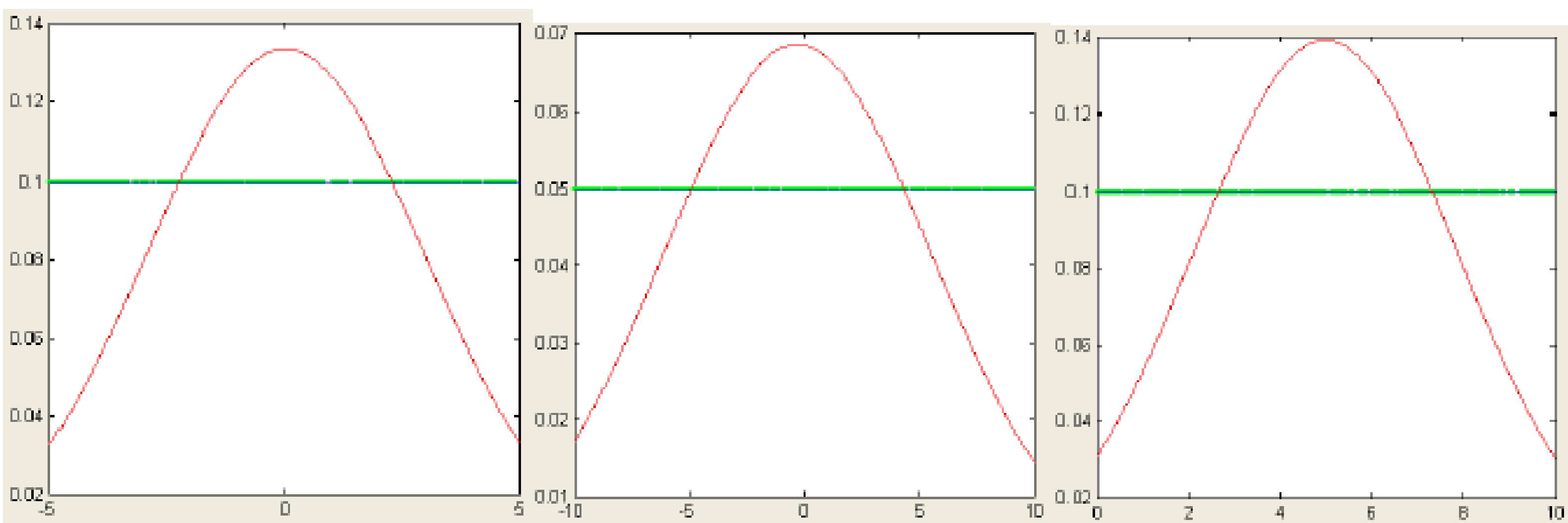
- 500,5000,50000 points from a Gaussian Distribution
  - Blue: original distribution



(g) $\mu = 0.0247, \sigma^2 = 1.0581$ (h) $\mu = -0.0095, \sigma^2 = 1.0078$ (i) $\mu = 0.0107, \sigma^2 = 1.0064$

Experiments

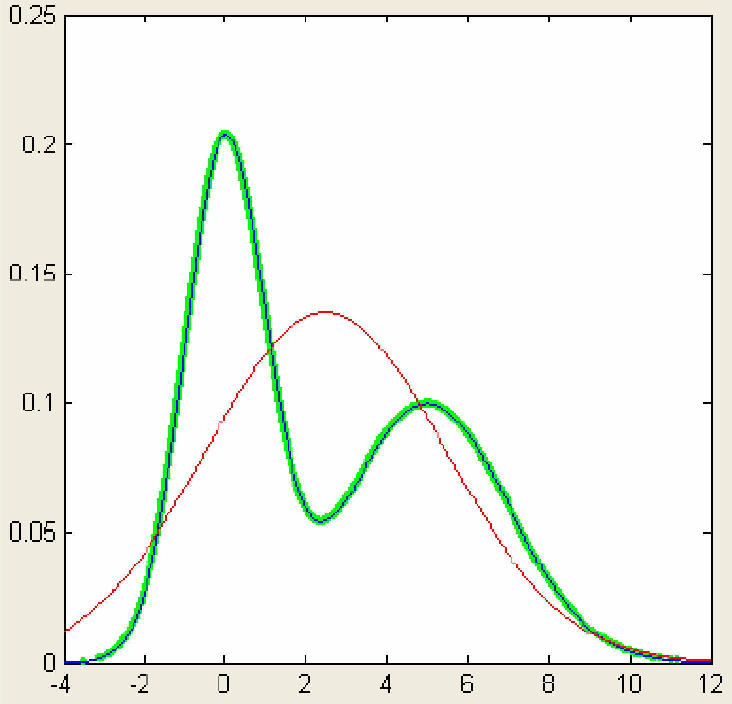
- 500 points from a uniform Distribution
  - Blue: original distribution
  - Red: estimated distribution
  - Green: samples



(a) (b) (c)

Experiments

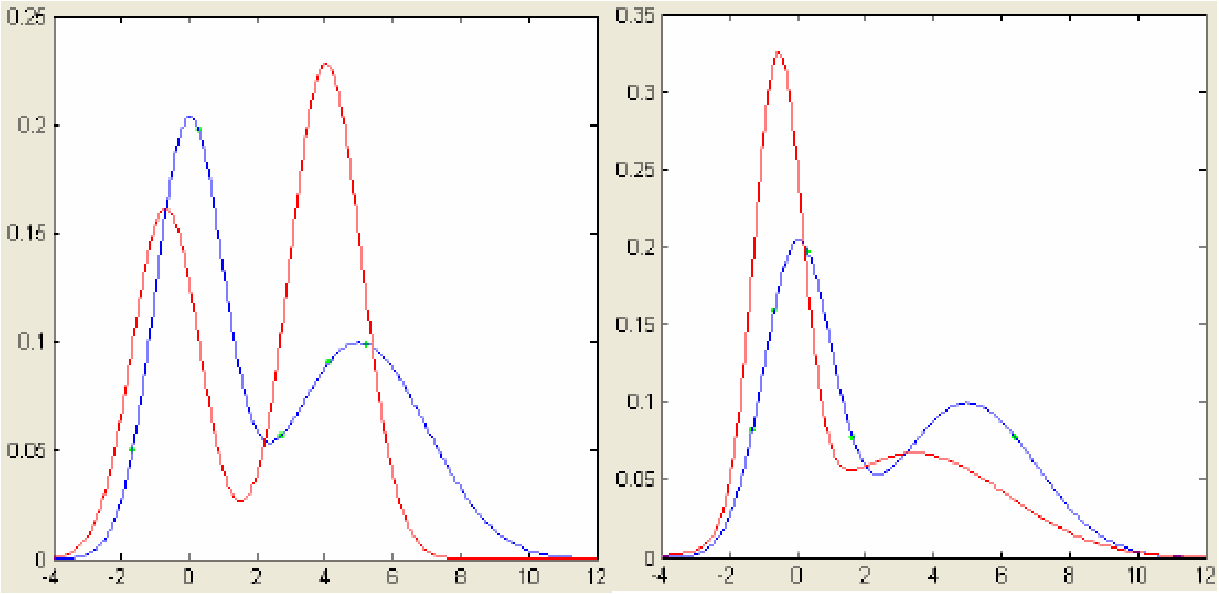
- 5000 points from a GMM
  - Blue: original distribution
  - Red: estimated distribution
  - Green: samples



(3-5) 混合高斯分布样本的正态分布最大似然估计(5000 样本)

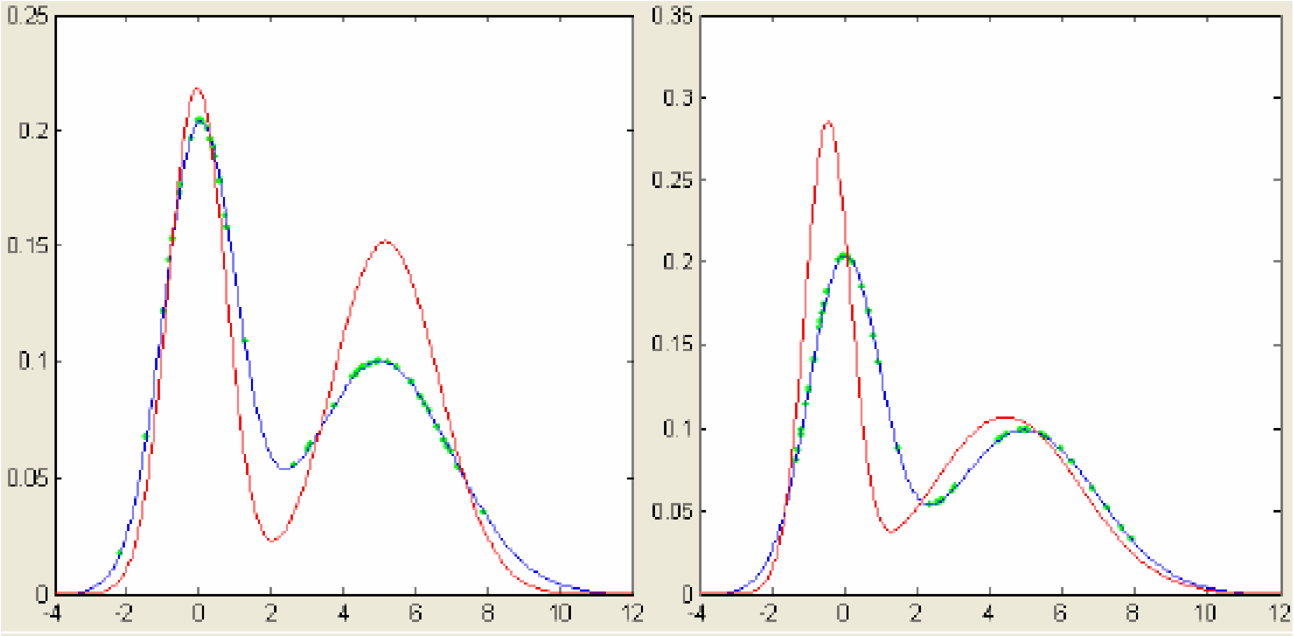
Experiments

- 5 points from a GMM (2 times)
  - Blue: original distribution
  - Red: estimated distribution
  - Green: samples



Experiments

- 50 points from a GMM (2 times)
  - Blue: original distribution
  - Red: estimated distribution
  - Green: samples



Experim

- 500 and 5000 points from a GMM (2 times)
  - Blue: original distribution
  - Red: estimated distribution
  - Green: samples

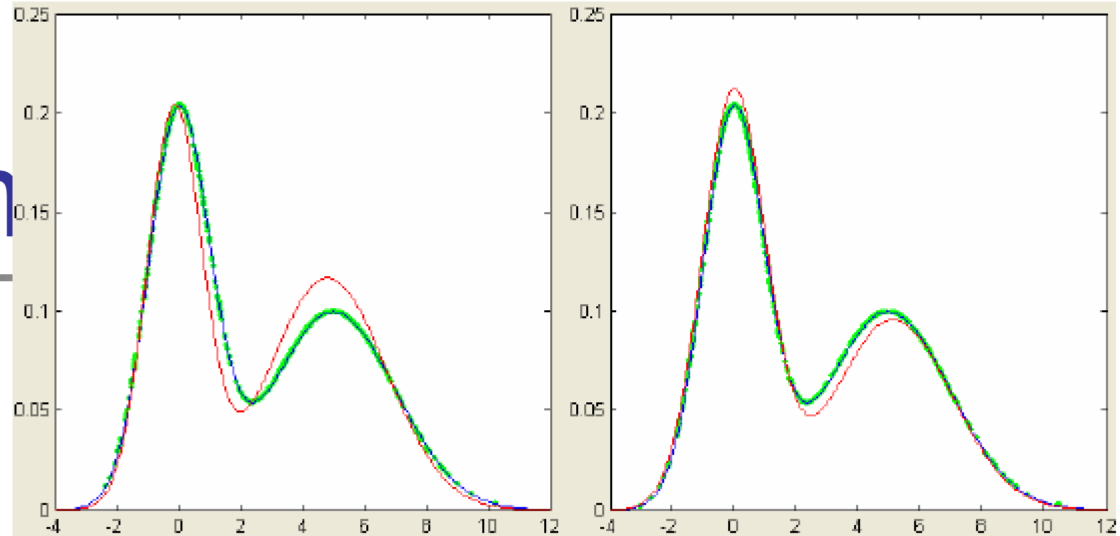


图 3-8. 一维 500 个样本的 EM 算法处理结果，蓝线为采样概率密度函数，红线为 EM 估计概率密度函数

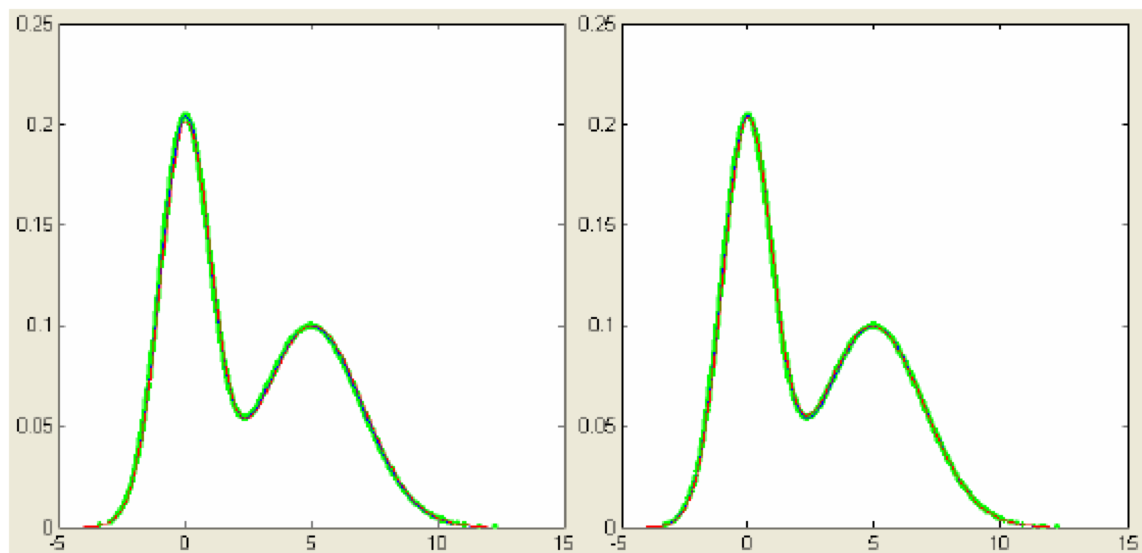
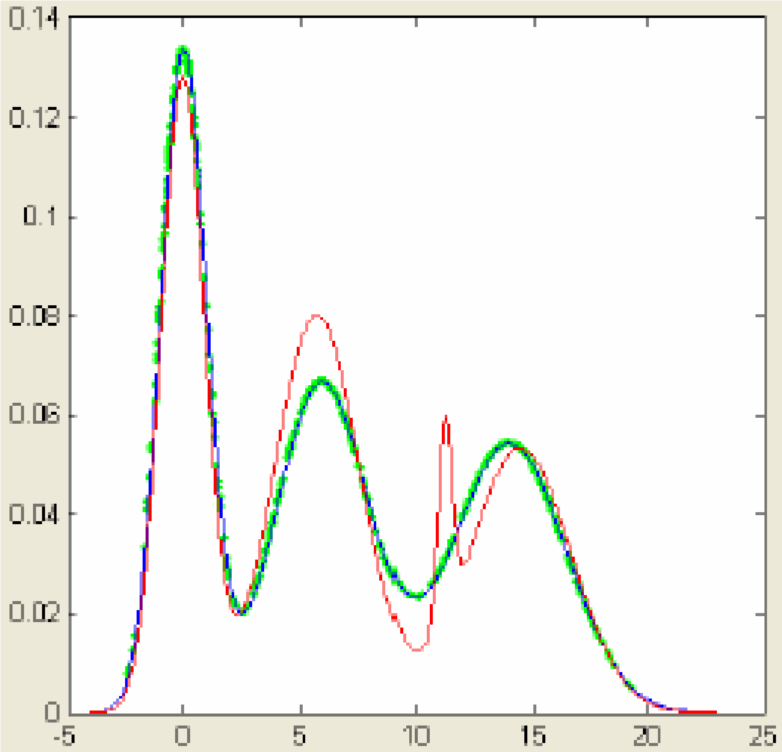
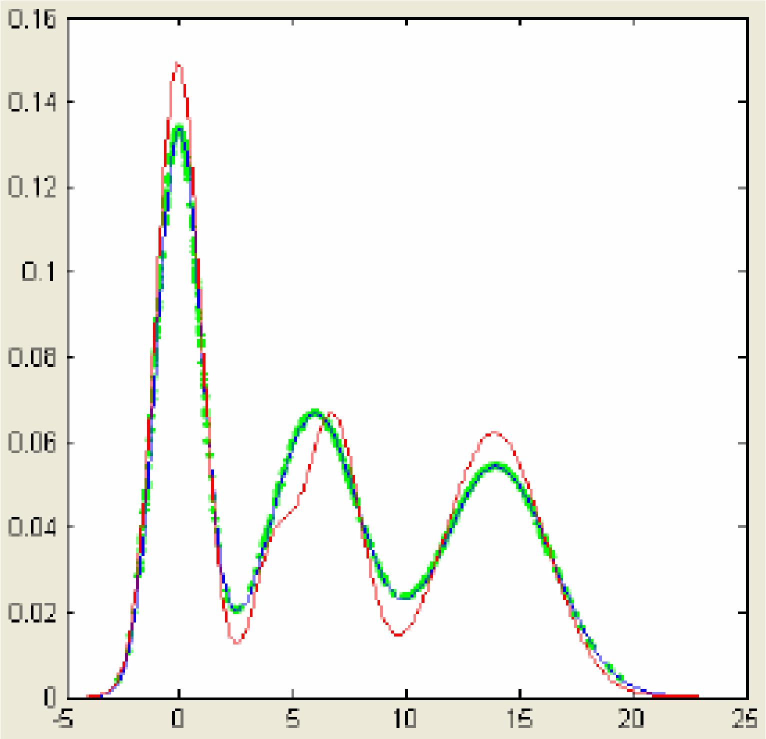


图 3-9. 一维 5000 个样本的 EM 算法处理结果，蓝线为采样概率密度函数，红线为 EM 估计概率密度函数

Experiments

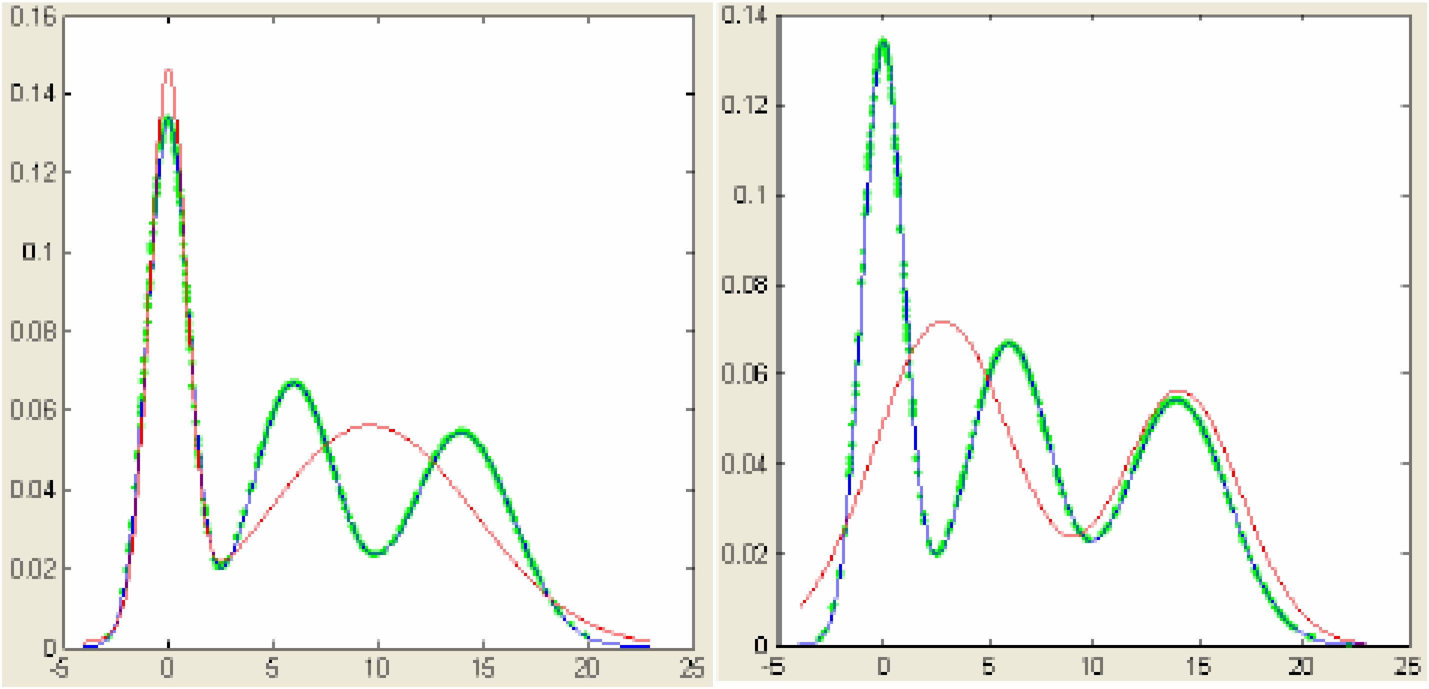
- A wrong number of components is given:





Experiments

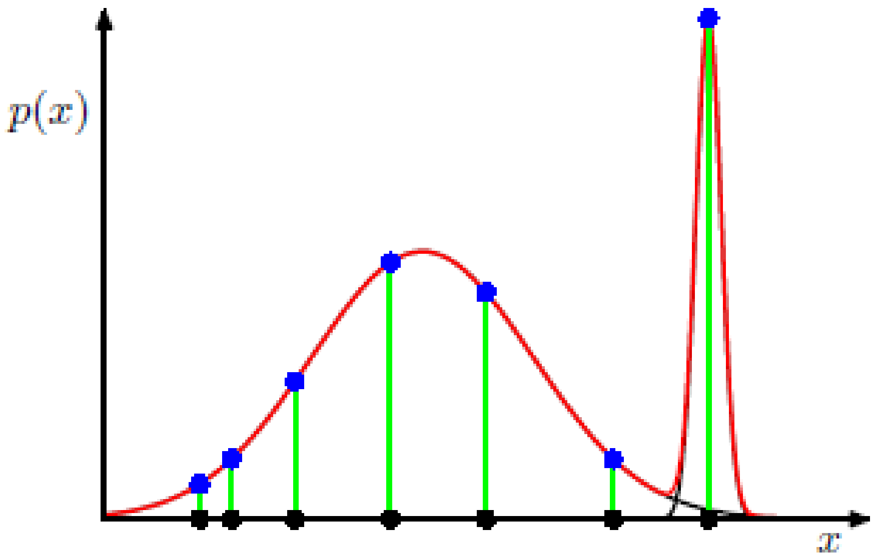
- A wrong number of components is given:



(a) (b)

Experiments

- Overfitting



- The figure is Copied from book by Duda

Research

- Parzon window on a manifold
- Stochastic EM
- Convergence of EM

- **http://www.vision.caltech.edu/welling/class/LearningSystemsB.html**

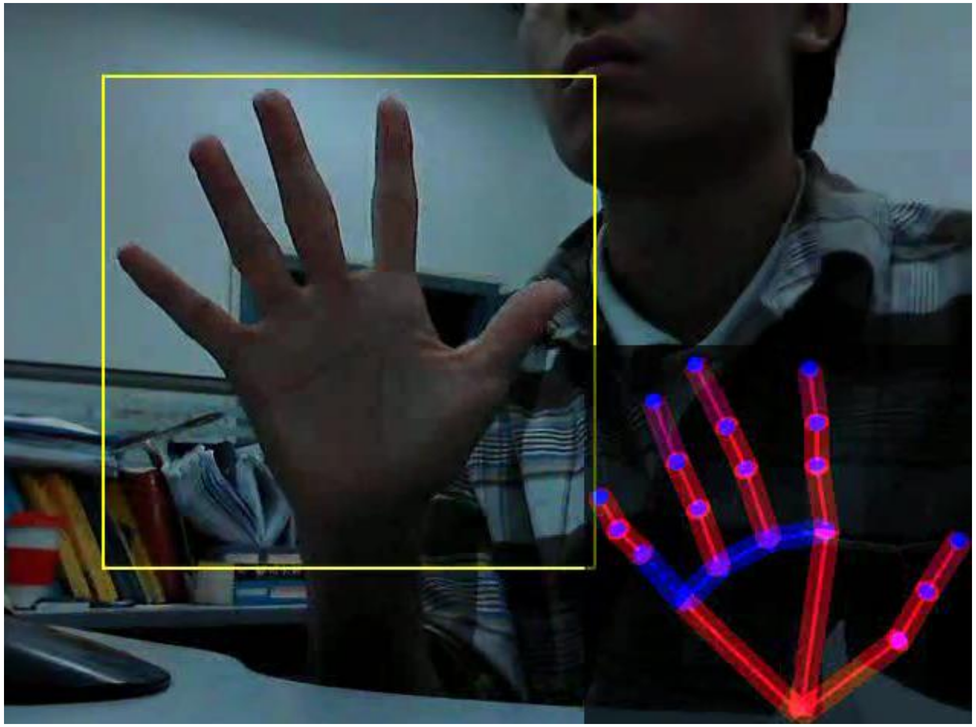
Research

- **PCA**
- **ICA**
- **Factor Analysis**
- **K-means**
- **Mixture of Gaussians**
- **Generative Topographic Mapping**

Research

- **Cluster Weighted Models**
- **Mixture of Experts**
- **Kalman Filter**
- **HMM**
- **Helmholtz Machine**
- **Boltzman Machine**
- **It will be shown that most of the learning schemes for these models can be understood as versions of the Expectation Maximization (EM) algorithm, thus providing a unified view.**

applications



Research

- How many components in GMM?
- MDL (Minimum description length) method
- Competitive EM Algorithm for Finite Mixture Models
- Merge and Splite operation and RJMCMC
- Kernel GMM

Research

EM:

- Machine Learning
- Computer Vision
- Pattern Recognition
- Bayesian Networks

Research

GMM

- background modeling in video tracking
- Speaker identification
- …

Reference:

- J. A. Bilmes et al “A Gentle Tutorial of the EM Algorithm and its application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models” 1998
- A.P.Dempster et al “Maximum-likelihood from incomplete data via the EM algorithm” 1977
- T.K.Moon ”The Expectation-Maximization Algorithm” *IEEE Trans Signal Processing* 1996

Thanks!