Parametric Models Part II: Expectation-Maximization and Mixture Density Estimation

Selim Aksoy
Department of Computer Engineering
Bilkent University
saksoy@cs.bilkent.edu.tr

Missing Features

- Suppose that we have a Bayesian classifier that uses the feature vector $\mathbf{x}$ but a subset $\mathbf{x}_{\boldsymbol{g}}$ of $\mathbf{x}$ are observed and the values for the remaining features $\mathbf{x}_{\boldsymbol{b}}$ are missing.
- How can we make a decision?
  - Throw away the observations with missing values.
  - Or, substitute $\mathbf{x}_{\boldsymbol{b}}$ by their average $\bar{\mathbf{x}}_{\boldsymbol{b}}$ in the training data, and use $\mathbf{x} = (\mathbf{x}_{\boldsymbol{g}}, \bar{\mathbf{x}}_{\boldsymbol{b}})$.
  - Or, marginalize the posterior over the missing features, and use the resulting posterior

$$P(w_i|\mathbf{x}_{\boldsymbol{g}}) = \frac{\int P(w_i|\mathbf{x}_{\boldsymbol{g}}, \mathbf{x}_{\boldsymbol{b}})\, p(\mathbf{x}_{\boldsymbol{g}}, \mathbf{x}_{\boldsymbol{b}})\, d\mathbf{x}_{\boldsymbol{b}}}{\int p(\mathbf{x}_{\boldsymbol{g}}, \mathbf{x}_{\boldsymbol{b}})\, d\mathbf{x}_{\boldsymbol{b}}}.$$

Expectation-Maximization

- We can also extend maximum likelihood techniques to allow learning of parameters when some training patterns have missing features.
- The *Expectation-Maximization (EM)* algorithm is a general iterative method of finding the maximum likelihood estimates of the parameters of a distribution from training data.

Expectation-Maximization

- There are two main applications of the EM algorithm:
  - Learning when the data is incomplete or has missing values.
  - Optimizing a likelihood function that is analytically intractable but can be simplified by assuming the existence of and values for additional but missing (or hidden) parameters.
- The second problem is more common in pattern recognition applications.

Expectation-Maximization

- Assume that the observed data $\mathcal{X}$ is generated by some distribution.
- Assume that a complete dataset $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ exists as a combination of the observed but incomplete data $\mathcal{X}$ and the missing data $\mathcal{Y}$.
- The observations in $\mathcal{Z}$ are assumed to be i.i.d. from the joint density

$$p(\mathbf{z}|\mathbf{\Theta}) = p(\mathbf{x}, \mathbf{y}|\mathbf{\Theta}) = p(\mathbf{y}|\mathbf{x}, \mathbf{\Theta})p(\mathbf{x}|\mathbf{\Theta}).$$

Expectation-Maximization

- We can define a new likelihood function

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

called the complete-data likelihood where $L(\mathbf{\Theta}|\mathcal{X})$ is referred to as the incomplete-data likelihood.

- The EM algorithm:
  - First, finds the expected value of the complete-data log-likelihood using the current parameter estimates (expectation step).
  - Then, maximizes this expectation (maximization step).

Expectation-Maximization

- Define

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

  as the expected value of the complete-data log-likelihood w.r.t. the unknown data $\mathcal{Y}$ given the observed data $\mathcal{X}$ and the current parameter estimates $\mathbf{\Theta}^{(i-1)}$.

- The expected value can be computed as

$$E\left[\log p(\mathcal{X}, \mathcal{Y}|\mathbf{\Theta})|\mathcal{X}, \mathbf{\Theta}^{(i-1)}\right] = \int \log p(\mathcal{X}, \mathbf{y}|\mathbf{\Theta})\, p(\mathbf{y}|\mathcal{X}, \mathbf{\Theta}^{(i-1)})\, d\mathbf{y}.$$

- This is called the *E-step*.

Expectation-Maximization

- Then, the expectation can be maximized by finding optimum values for the new parameters $\boldsymbol{\Theta}$ as
$$\boldsymbol{\Theta}^{(i)} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i-1)}).$$
- This is called the *M-step*.
- These two steps are repeated iteratively where each iteration is guaranteed to increase the log-likelihood.
- The EM algorithm is also guaranteed to converge to a local maximum of the likelihood function.

Generalized Expectation-Maximization

- Instead of maximizing $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i-1)})$, the *Generalized Expectation-Maximization* algorithm finds some set of parameters $\boldsymbol{\Theta}^{(i)}$ that satisfy

$$Q(\boldsymbol{\Theta}^{(i)}, \boldsymbol{\Theta}^{(i-1)}) > Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i-1)})$$

  at each iteration.

- Convergence will not be as rapid as the EM algorithm but it allows greater flexibility to choose computationally simpler steps.

Mixture Densities

- A mixture model is a linear combination of m densities

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{j=1}^{m} \alpha_j p_j(\mathbf{x}|\boldsymbol{\theta_j})$$

  where $\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_m, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_m})$ such that $\alpha_j \geq 0$ and $\sum_{j=1}^{m} \alpha_j = 1$.

- $\alpha_1, \ldots, \alpha_m$ are called the mixing parameters.
- $p_j(\mathbf{x}|\boldsymbol{\theta_j})$, $j = 1, \ldots, m$ are called the component densities.

Mixture Densities

- Suppose that $\mathcal{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ is a set of observations i.i.d. with distribution $p(\mathbf{x}|\boldsymbol{\Theta})$.
- The log-likelihood function of $\boldsymbol{\Theta}$ becomes

$$\log L(\boldsymbol{\Theta}|\mathcal{X}) = \log \prod_{i=1}^{n} p(\mathbf{x_i}|\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} \alpha_j p_j(\mathbf{x_i}|\boldsymbol{\theta_j}) \right).$$

- We cannot obtain an analytical solution for $\boldsymbol{\Theta}$ by simply setting the derivatives of $\log L(\boldsymbol{\Theta}|\mathcal{X})$ to zero because of the logarithm of the sum.

Mixture Density Estimation via EM

- Consider $\mathcal{X}$ as incomplete and define hidden variables $\mathcal{Y} = \{y_i\}_{i=1}^n$ where y_i corresponds to which mixture component generated the data vector $\mathbf{x_{\boldsymbol{i}}}$.
- In other words, $y_i = j$ if the i'th data vector was generated by the j'th mixture component.
- Then, the log-likelihood becomes

$$\begin{aligned} \log L(\boldsymbol{\Theta}|\mathcal{X},\mathcal{Y}) &= \log p(\mathcal{X},\mathcal{Y}|\boldsymbol{\Theta}) \\ &= \sum_{i=1}^{n} \log(p(\mathbf{x_{\boldsymbol{i}}}|y_i,\boldsymbol{\theta_i})p(y_i|\boldsymbol{\theta_i})) \\ &= \sum_{i=1}^{n} \log(\alpha_{y_i} p_{y_i}(\mathbf{x_{\boldsymbol{i}}}|\boldsymbol{\theta_{y_i}})). \end{aligned}$$

Mixture Density Estimation via EM

- Assume we have the initial parameter estimates $\boldsymbol{\Theta}^{(g)} = (\alpha_1^{(g)}, \ldots, \alpha_m^{(g)}, \boldsymbol{\theta}_{\boldsymbol{1}}^{(g)}, \ldots, \boldsymbol{\theta}_{\boldsymbol{m}}^{(g)})$.
- Compute

$$p(y_i|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)}) = \frac{\alpha_{y_i}^{(g)} p_{y_i}(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{y_i}}^{(g)})}{p(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\Theta}^{(g)})} = \frac{\alpha_{y_i}^{(g)} p_{y_i}(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{y_i}}^{(g)})}{\sum_{j=1}^{m} \alpha_j^{(g)} p_j(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{j}}^{(g)})}$$

and

$$p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\Theta}^{(g)}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)}).$$

Mixture Density Estimation via EM

- Then, $Q(\mathbf{\Theta}, \mathbf{\Theta}^{(g)})$ takes the form

$$
\begin{aligned}
Q(\mathbf{\Theta}, \mathbf{\Theta}^{(g)}) &= \sum_{\mathbf{y}} \log p(\mathcal{X}, \mathbf{y} | \mathbf{\Theta}) p(\mathbf{y} | \mathcal{X}, \mathbf{\Theta}^{(g)}) \\
&= \sum_{j=1}^{m} \sum_{i=1}^{n} \log(\alpha_j p_j(\mathbf{x_i} | \boldsymbol{\theta_j})) p(j | \mathbf{x_i}, \mathbf{\Theta}^{(g)}) \\
&= \sum_{j=1}^{m} \sum_{i=1}^{n} \log(\alpha_j) p(j | \mathbf{x_i}, \mathbf{\Theta}^{(g)}) \\
&\qquad + \sum_{j=1}^{m} \sum_{i=1}^{n} \log(p_j(\mathbf{x_i} | \boldsymbol{\theta_j})) p(j | \mathbf{x_i}, \mathbf{\Theta}^{(g)}).
\end{aligned}
$$

Mixture Density Estimation via EM

- We can maximize the two sets of summations for α_j and $\boldsymbol{\theta}_j$ independently because they are not related.
- The estimate for α_j can be computed as

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)})$$

where

$$p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)}) = \frac{\alpha_j^{(g)} p_j(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{j}}^{(g)})}{\sum_{t=1}^{m} \alpha_t^{(g)} p_t(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{t}}^{(g)})}.$$

Mixture of Gaussians

- We can obtain analytical expressions for $\boldsymbol{\theta_j}$ for the special case of a Gaussian mixture where $\boldsymbol{\theta_j} = (\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$ and

$$\begin{aligned} p_j(\mathbf{x}|\boldsymbol{\theta_j}) &= p_j(\mathbf{x}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j}) \\ &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma_j}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_j})^T \boldsymbol{\Sigma_j}^{-1}(\mathbf{x}-\boldsymbol{\mu_j})\right]. \end{aligned}$$

- Equating the partial derivative of $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(g)})$ with respect to $\boldsymbol{\mu_j}$ to zero gives

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{j}} = \frac{\sum_{i=1}^n p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)})\mathbf{x_i}}{\sum_{i=1}^n p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)})}.$$

Mixture of Gaussians

- We consider five models for the covariance matrix $\boldsymbol{\Sigma_j}$:
  - $\boldsymbol{\Sigma_j} = \sigma^2 \mathbf{I}$

$$\hat{\sigma}^2 = \frac{1}{nd} \sum_{j=1}^{m} \sum_{i=1}^{n} p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)}) \|\mathbf{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}}\|^2$$

  - $\boldsymbol{\Sigma_j} = \sigma_j^2 \mathbf{I}$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)}) \|\mathbf{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}}\|^2}{d \sum_{i=1}^{n} p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)})}$$

Mixture of Gaussians

- Covariance models continued:
  - $\mathbf{\Sigma_j} = \text{diag}(\{\sigma_{jk}^2\}_{k=1}^d)$

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^n p(j|\mathbf{x_i}, \mathbf{\Theta}^{(g)})(\mathbf{x_{ik}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}_k})^2}{\sum_{i=1}^n p(j|\mathbf{x_i}, \mathbf{\Theta}^{(g)})}$$

  - $\mathbf{\Sigma_j} = \mathbf{\Sigma}$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^n p(j|\mathbf{x_i}, \mathbf{\Theta}^{(g)})(\mathbf{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})(\mathbf{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})^T$$

  - $\mathbf{\Sigma_j} = \text{arbitrary}$

$$\hat{\mathbf{\Sigma}}_{\boldsymbol{j}} = \frac{\sum_{i=1}^n p(j|\mathbf{x_i}, \mathbf{\Theta}^{(g)})(\mathbf{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})(\mathbf{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})^T}{\sum_{i=1}^n p(j|\mathbf{x_i}, \mathbf{\Theta}^{(g)})}$$

Mixture of Gaussians

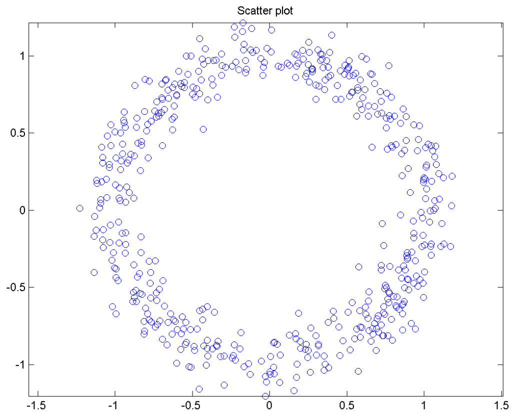
- Summary:
  - Estimates for α_j, $\boldsymbol{\mu_j}$ and $\boldsymbol{\Sigma_j}$ perform both expectation and maximization steps simultaneously.
  - EM iterations proceed by using the current estimates as the initial estimates for the next iteration.
  - The priors are computed from the proportion of examples belonging to each mixture component.
  - The means are the component centroids.
  - The covariance matrices are calculated as the sample covariance of the points associated with each component.

Mixture of Gaussians

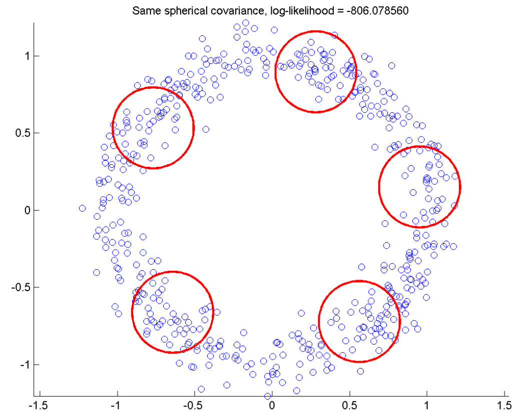
- Questions:
  - How can we find the number of components in the mixture?
  - How can we find the initial estimates for $\mathbf{\Theta}$?
  - How do we know when to stop the iterations?
    - Stop if the change in log-likelihood between two iterations is less than a threshold.
    - Or, use a threshold for the number of iterations.

Examples

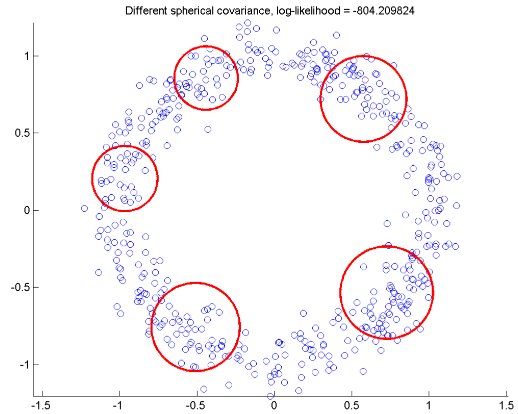
- Mixture of Gaussians examples
- 1-D Bayesian classification examples
- 2-D Bayesian classification examples



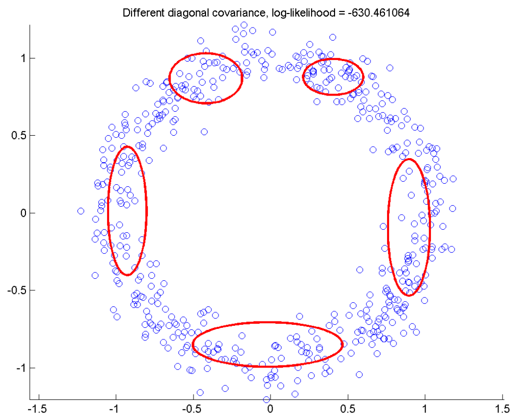
(a) Scatter plot.



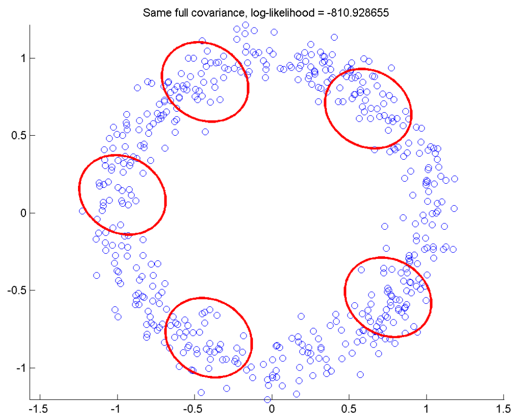
(b) Same spherical covariance, log-likelihood = -806.08.



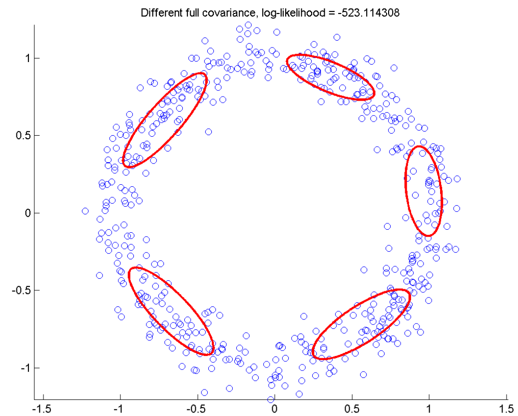
(c) Different spherical covariance, log-likelihood = -804.21.



(d) Different diagonal covariance, log-likelihood = -630.46.

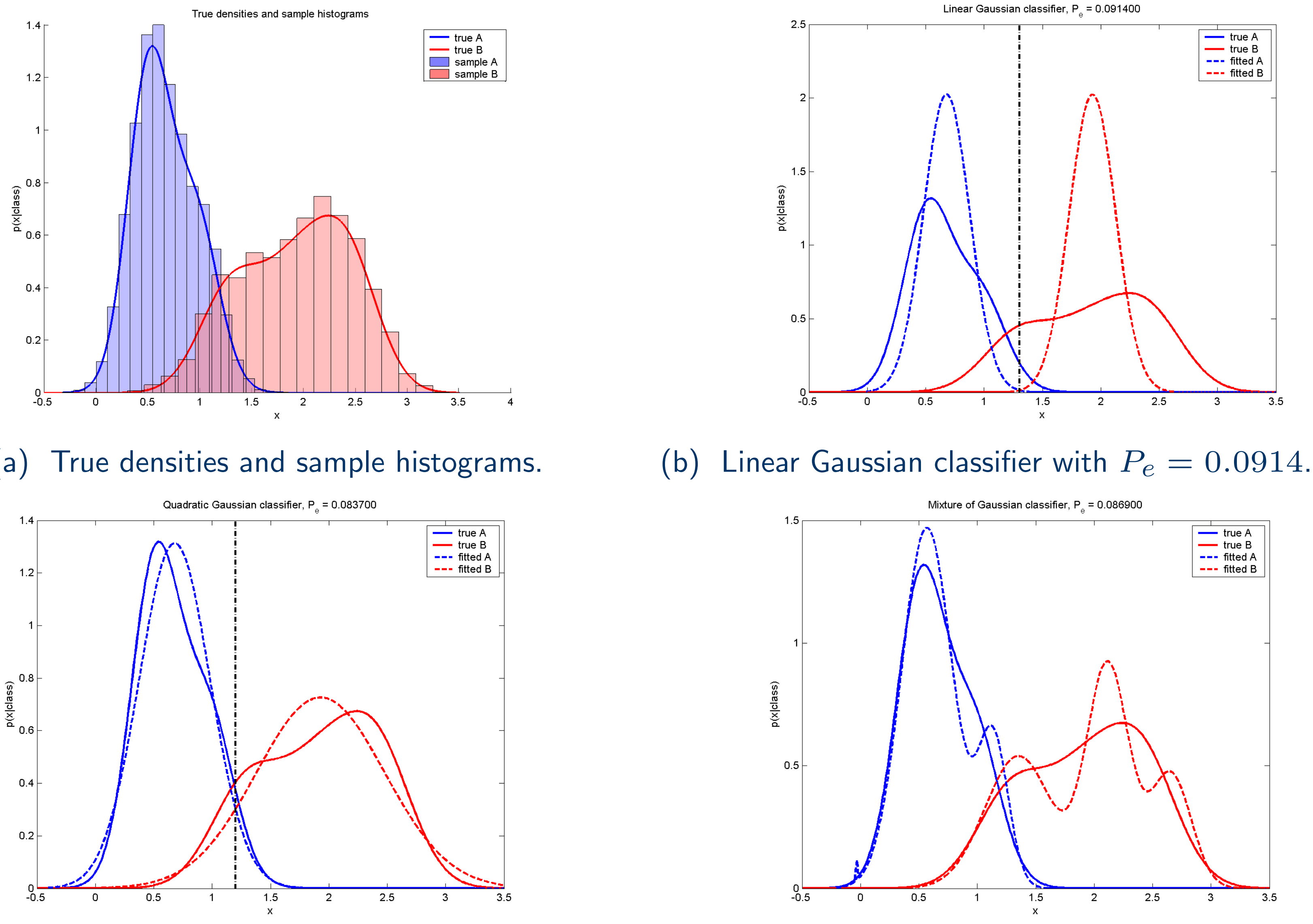


(e) Same arbitrary covariance, log-likelihood = -810.93.



(f) Different arbitrary covariance, log-likelihood = -523.11.

Figure 1: Fitting mixtures of 5 Gaussians to data from a circular distribution.



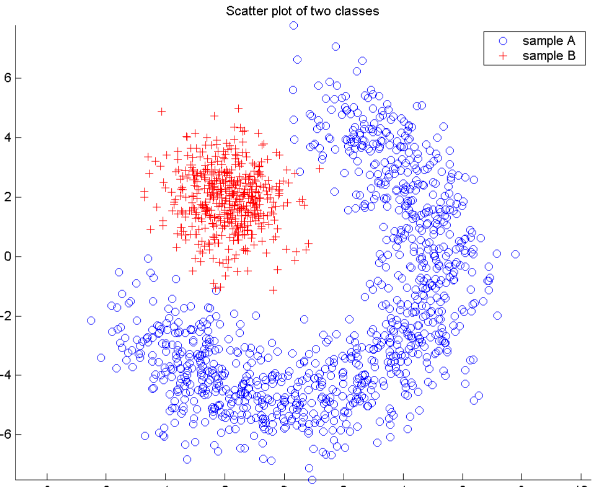
(a) True densities and sample histograms.

(b) Linear Gaussian classifier with $P_e = 0.0914$.

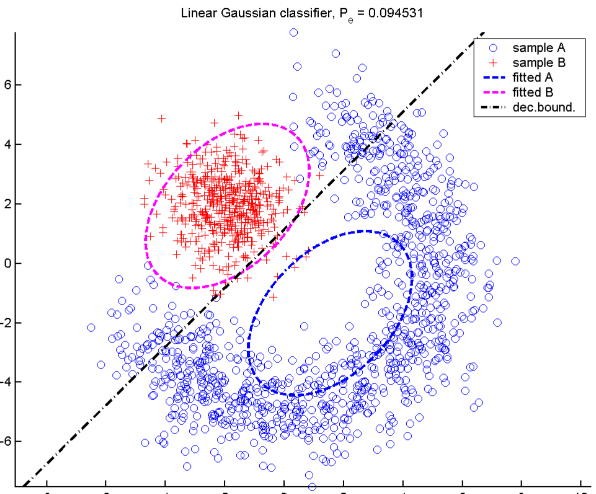
(c) Quadratic Gaussian classifier with $P_e = 0.0837$.

(d) Mixture of Gaussian classifier with $P_e = 0.0869$.

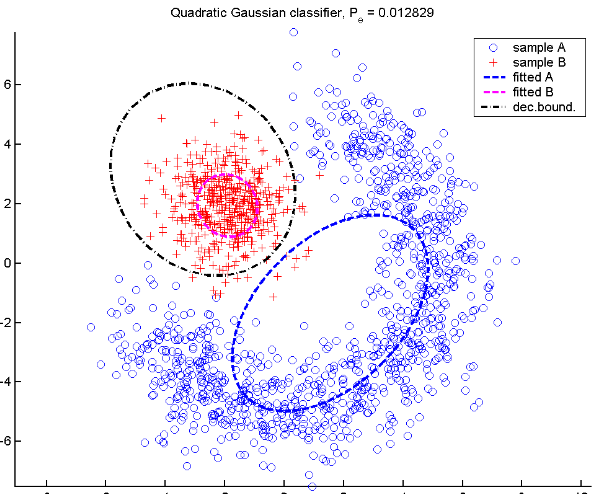
Figure 2: 1-D Bayesian classification examples where the data for each class come from a mixture of three Gaussians. Bayes error is $P_e = 0.0828$.



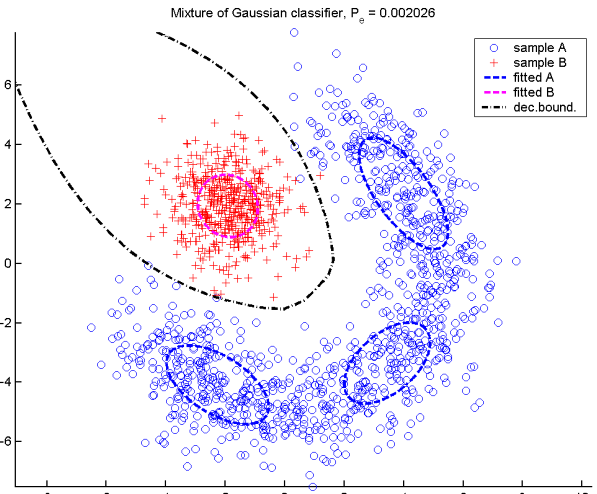
(a) Scatter plot.



(b) Linear Gaussian classifier with $P_e = 0.094531$.

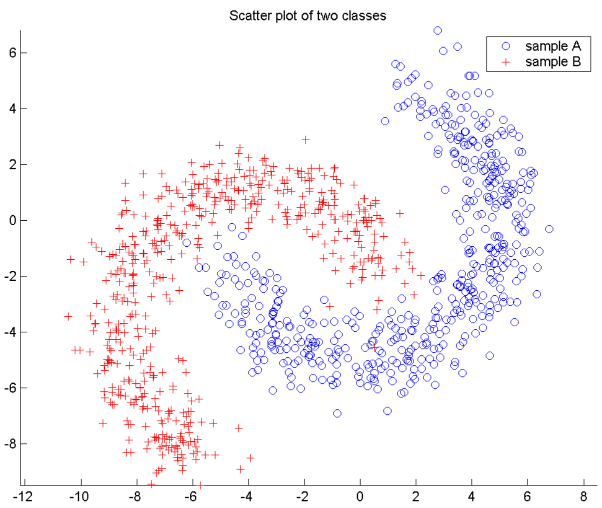


(c) Quadratic Gaussian classifier with $P_e = 0.012829$.

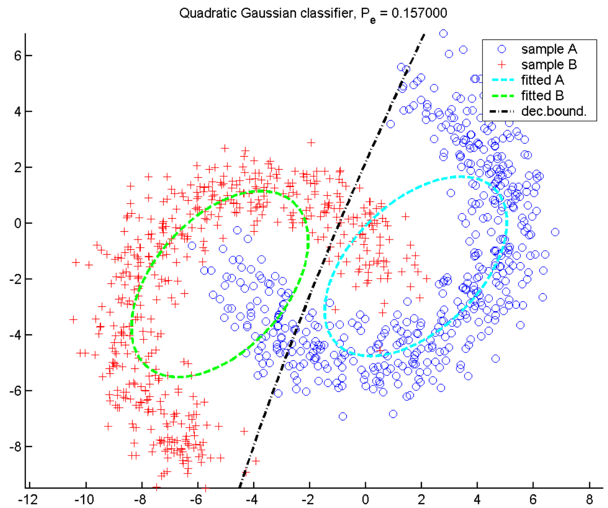


(d) Mixture of Gaussian classifier with $P_e = 0.002026$.

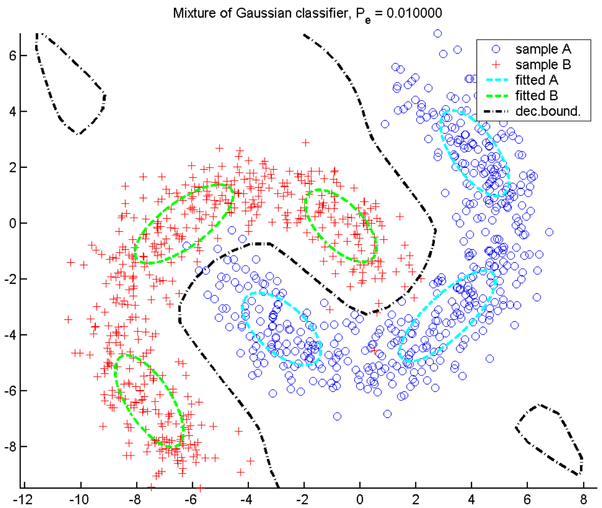
Figure 3: 2-D Bayesian classification examples where the data for the classes come from a banana shaped distribution and a bivariate Gaussian.



(a) Scatter plot.



(b) Quadratic Gaussian classifier with $P_e = 0.1570$.



(c) Quadratic Gaussian classifier with $P_e = 0.0100$.

Figure 4: 2-D Bayesian classification examples where the data for each class come from a banana shaped distribution.