Multidimensional Gaussian distribution and classification with Gaussians

Guido Sanguinetti

Informatics 2B— Learning and Data Lecture 9
6 March 2012

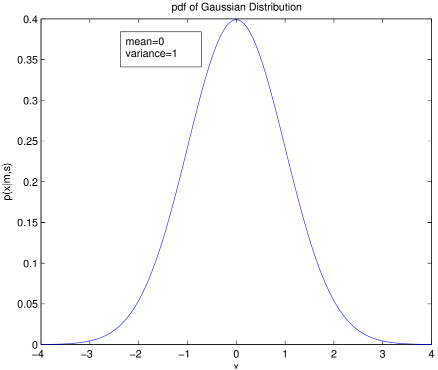
Overview

Today's lecture

Gaussians

- The multidimensional Gaussian distribution
- Bayes theorem and probability density functions
- The Gaussian classifier

One-dimensional Gaussian with zero mean and unit variance ($\mu = 0$, $\sigma^2 = 1$):



$$p(x|\mu, \sigma^2) = N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

The multidimensional Gaussian distribution

- The d-dimensional vector $\mathbf{x}$ is multivariate Gaussian if it has a probability density function of the following form:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

  The pdf is parameterized by the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.

The multidimensional Gaussian distribution

- The d-dimensional vector $\mathbf{x}$ is multivariate Gaussian if it has a probability density function of the following form:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

  The pdf is parameterized by the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.

- The 1-dimensional Gaussian is a special case of this pdf

The multidimensional Gaussian distribution

- The d-dimensional vector $\mathbf{x}$ is multivariate Gaussian if it has a probability density function of the following form:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

  The pdf is parameterized by the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.
- The 1-dimensional Gaussian is a special case of this pdf
- The argument to the exponential $0.5(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is referred to as a *quadratic form*.

Covariance matrix

- The mean vector $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$:

$$\boldsymbol{\mu} = E[\mathbf{x}]$$

Covariance matrix

- The mean vector $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$:

$$\boldsymbol{\mu} = E[\mathbf{x}]$$

- The covariance matrix $\boldsymbol{\Sigma}$ is the expectation of the deviation of $\mathbf{x}$ from the mean:

$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

Covariance matrix

- The mean vector $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$:

$$\boldsymbol{\mu} = E[\mathbf{x}]$$

- The covariance matrix $\boldsymbol{\Sigma}$ is the expectation of the deviation of $\mathbf{x}$ from the mean:

$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

- $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric matrix:

$$\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] = E[(x_j - \mu_j)(x_i - \mu_i)] = \Sigma_{ji}$$

Covariance matrix

- The mean vector $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$:

$$\boldsymbol{\mu} = E[\mathbf{x}]$$

- The covariance matrix $\boldsymbol{\Sigma}$ is the expectation of the deviation of $\mathbf{x}$ from the mean:

$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

- $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric matrix:

$$\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] = E[(x_j - \mu_j)(x_i - \mu_i)] = \Sigma_{ji}$$

- The sign of the covariance helps to determine the relationship between two components:

Covariance matrix

- The mean vector $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$:

$$\boldsymbol{\mu} = E[\mathbf{x}]$$

- The covariance matrix $\boldsymbol{\Sigma}$ is the expectation of the deviation of $\mathbf{x}$ from the mean:

$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

- $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric matrix:

$$\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] = E[(x_j - \mu_j)(x_i - \mu_i)] = \Sigma_{ji}$$

- The sign of the covariance helps to determine the relationship between two components:
  - If x_j is large when x_i is large, then $(x_j - \mu_j)(x_i - \mu_i)$ will tend to be positive;

Covariance matrix

- The mean vector $\boldsymbol{\mu}$ is the expectation of $\mathbf{x}$:

$$\boldsymbol{\mu} = E[\mathbf{x}]$$

- The covariance matrix $\boldsymbol{\Sigma}$ is the expectation of the deviation of $\mathbf{x}$ from the mean:

$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

- $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric matrix:

$$\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] = E[(x_j - \mu_j)(x_i - \mu_i)] = \Sigma_{ji}$$

- The sign of the covariance helps to determine the relationship between two components:
  - If x_j is large when x_i is large, then $(x_j - \mu_j)(x_i - \mu_i)$ will tend to be positive;
  - If x_j is small when x_i is large, then $(x_j - \mu_j)(x_i - \mu_i)$ will tend to be negative.

Correlation matrix

The covariance matrix is not scale-independent: Define the correlation coefficient:

$$\rho(x_j, x_k) = \rho_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}}$$

Correlation matrix

The covariance matrix is not scale-independent: Define the correlation coefficient:

$$\rho(x_j, x_k) = \rho_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}}$$

- Scale-independent (ie independent of the measurement units) and location-independent, ie:

$$\rho(x_j, x_k) = \rho(ax_j + b, sx_k + t)$$

Correlation matrix

The covariance matrix is not scale-independent: Define the correlation coefficient:

$$\rho(x_j, x_k) = \rho_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}}$$

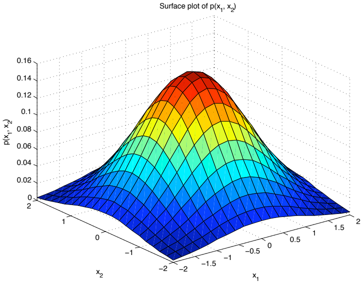
- Scale-independent (ie independent of the measurement units) and location-independent, ie:

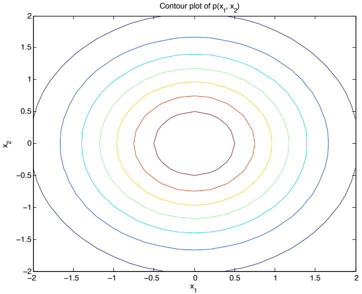
$$\rho(x_j, x_k) = \rho(ax_j + b, sx_k + t)$$

- The correlation coefficient satisfies $-1 \leq \rho \leq 1$, and

$$\begin{aligned} \rho(x, y) &= +1 \qquad \text{if } y = ax + b \quad a > 0 \\ \rho(x, y) &= -1 \qquad \text{if } y = ax + b \quad a < 0 \end{aligned}$$

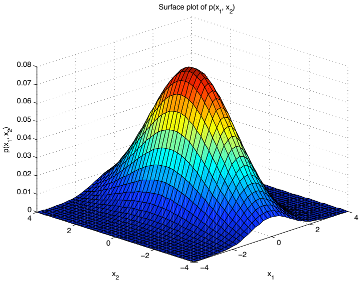
Spherical Gaussian

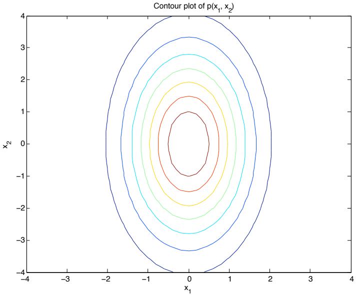




$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \rho_{12} = 0$$

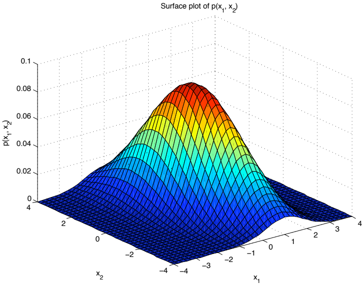
Diagonal Covariance Gaussian

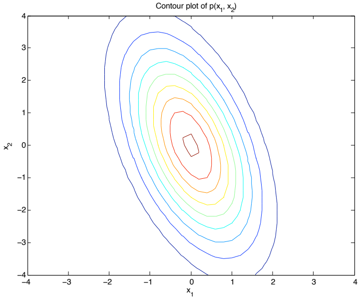




$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \qquad \rho_{12} = 0$$

Full covariance Gaussian





$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} \qquad \rho_{12} = -0.5$$

Parameter estimation

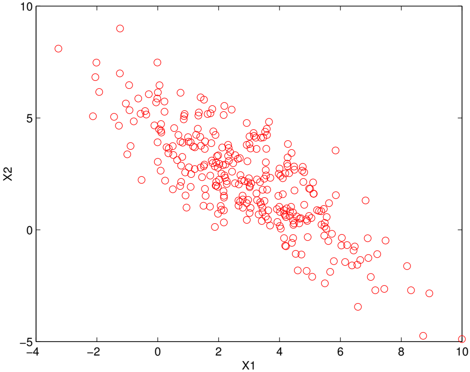
- It is possible to show that the mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ that maximize the likelihood of the training data are given by:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^n$$

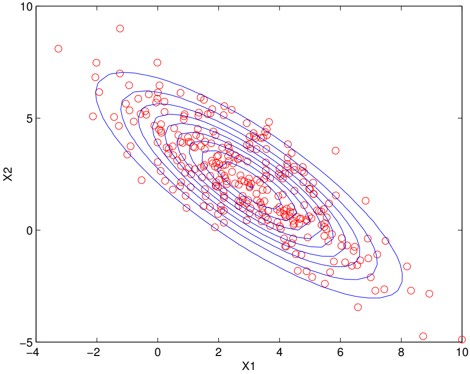
$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}^n - \hat{\boldsymbol{\mu}})(\mathbf{x}^n - \hat{\boldsymbol{\mu}})^T$$

- The mean of the distribution is estimated by the sample mean and the covariance by the sample covariance

Example data



Maximum likelihood fit to a Gaussian



Bayes theorem and probability densities

- Rules for probability densities are similar to those for probabilities:

$$p(x, y) = p(x|y)p(y)$$
$$p(x) = \int p(x, y)dy$$

Bayes theorem and probability densities

- Rules for probability densities are similar to those for probabilities:

$$p(x, y) = p(x|y)p(y)$$
$$p(x) = \int p(x, y)dy$$

- We may mix probabilities of discrete variables and probability densities of continuous variables:

$$p(x, Z) = p(x|Z)P(Z)$$

Bayes theorem and probability densities

- Rules for probability densities are similar to those for probabilities:

$$p(x, y) = p(x|y)p(y)$$
$$p(x) = \int p(x, y)dy$$

- We may mix probabilities of discrete variables and probability densities of continuous variables:

$$p(x, Z) = p(x|Z)P(Z)$$

- Bayes' theorem for continuous data x and class C:

$$P(C|x) = \frac{p(x|C)P(C)}{p(x)}$$
$$P(C|x) \propto p(x|C)P(C)$$

Bayes theorem and univariate Gaussians

- If $p(x \mid C)$ is Gaussian with mean μ_c and variance σ_c^2:

$$\begin{aligned} P(C \mid x) &\propto p(x \mid C)P(C) \\ &\propto N(x; \mu_c, \sigma_c^2)P(C) \\ &\propto \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(\frac{-(x-\mu_c)^2}{2\sigma_c^2}\right) P(C) \end{aligned}$$

Bayes theorem and univariate Gaussians

- If $p(x \mid C)$ is Gaussian with mean μ_c and variance σ_c^2:

$$\begin{aligned} P(C \mid x) &\propto p(x \mid C) P(C) \\ &\propto N(x; \mu_c, \sigma_c^2) P(C) \\ &\propto \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(\frac{-(x-\mu_c)^2}{2\sigma_c^2} \right) P(C) \end{aligned}$$

- Taking logs, we have the log likelihood $LL(x \mid C)$:

$$\begin{aligned} LL(x \mid C) &= \ln p(x \mid \mu_c, \sigma_c^2) \\ &= \frac{1}{2} \left(-\ln(2\pi) - \ln \sigma_c^2 - \frac{(x-\mu_c)^2}{\sigma_c^2} \right) \end{aligned}$$

Bayes theorem and univariate Gaussians

- If $p(x \mid C)$ is Gaussian with mean μ_c and variance σ_c^2:

$$\begin{aligned} P(C \mid x) &\propto p(x \mid C)P(C) \\ &\propto N(x; \mu_c, \sigma_c^2)P(C) \\ &\propto \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(\frac{-(x-\mu_c)^2}{2\sigma_c^2}\right) P(C) \end{aligned}$$

- Taking logs, we have the log likelihood $LL(x \mid C)$:

$$\begin{aligned} LL(x \mid C) &= \ln p(x \mid \mu_c, \sigma_c^2) \\ &= \frac{1}{2}\left(-\ln(2\pi) - \ln \sigma_c^2 - \frac{(x-\mu_c)^2}{\sigma_c^2}\right) \end{aligned}$$

- The log posterior probability $LP(C \mid x)$ is:

$$\begin{aligned} LP(C \mid x) &\propto LL(x \mid C) + LP(C) \\ &\propto \frac{1}{2}\left(-\ln(2\pi) - \ln \sigma_c^2 - \frac{(x-\mu_c)^2}{\sigma_c^2}\right) + \ln P(C) \end{aligned}$$

Example: 1-dimensional Gaussian classifier

- Two classes, S and T, with some observations:

Class S	10	8	10	10	11	11
Class T	12	9	15	10	13	13

Example: 1-dimensional Gaussian classifier

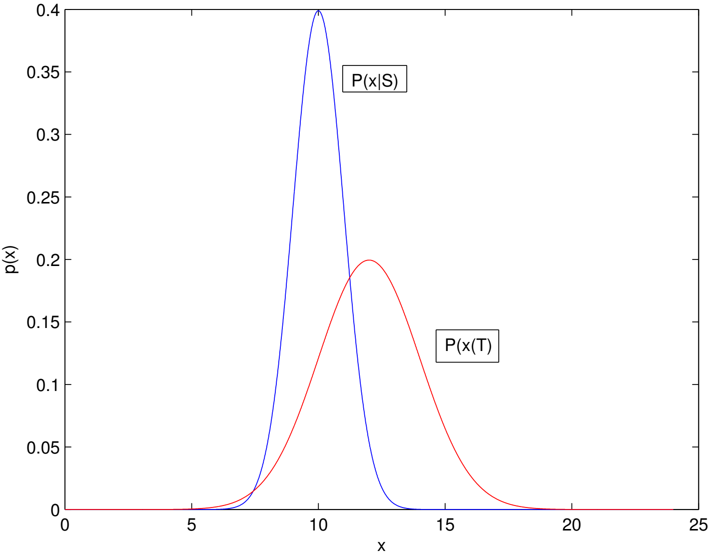
- Two classes, S and T, with some observations:

Class S	10	8	10	10	11	11
Class T	12	9	15	10	13	13

- Assume that each class may be modelled by a Gaussian. The mean and variance of each pdf are estimated by the sample mean and sample variance:

$$\mu(S) = 10 \qquad \sigma^2(S) = 1$$
$$\mu(T) = 12 \qquad \sigma^2(T) = 4$$

Gaussian pdfs for S and T



Example: 1-dimensional Gaussian classifier

- Two classes, S and T, with some observations:

Class S	10	8	10	10	11	11
Class T	12	9	15	10	13	13

- Assume that each class may be modelled by a Gaussian. The mean and variance of each pdf are estimated by the sample mean and sample variance:

$$\mu(S) = 10 \qquad \sigma^2(S) = 1$$
$$\mu(T) = 12 \qquad \sigma^2(T) = 4$$

- The following unlabelled data points are available:

$$x^1 = 10 \qquad x^2 = 11 \qquad x^3 = 6$$

To which class should each of the data points be assigned? Assume the two classes have equal prior probabilities.

Log odds

- Take the log odds (posterior probability ratios):

$$\ln \frac{P(S|X=x)}{P(T|X=x)} = -\frac{1}{2}\left(\frac{(x-\mu_S)^2}{\sigma_S^2} - \frac{(x-\mu_T)^2}{\sigma_T^2} + \ln \sigma_S^2 - \ln \sigma_T^2\right) + \ln P(S) - \ln P(T)$$

Log odds

- Take the log odds (posterior probability ratios):

$$\ln \frac{P(S|X=x)}{P(T|X=x)} = -\frac{1}{2}\left(\frac{(x-\mu_S)^2}{\sigma_S^2} - \frac{(x-\mu_T)^2}{\sigma_T^2} + \ln \sigma_S^2 - \ln \sigma_T^2\right) + \ln P(S) - \ln P(T)$$

- In the example the priors are equal, so:

$$\begin{aligned}\ln \frac{P(S|X=x)}{P(T|X=x)} &= -\frac{1}{2}\left(\frac{(x-\mu_S)^2}{\sigma_S^2} - \frac{(x-\mu_T)^2}{\sigma_T^2} + \ln \sigma_S^2 - \ln \sigma_T^2\right) \\ &= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right)\end{aligned}$$

Log odds

- Take the log odds (posterior probability ratios):

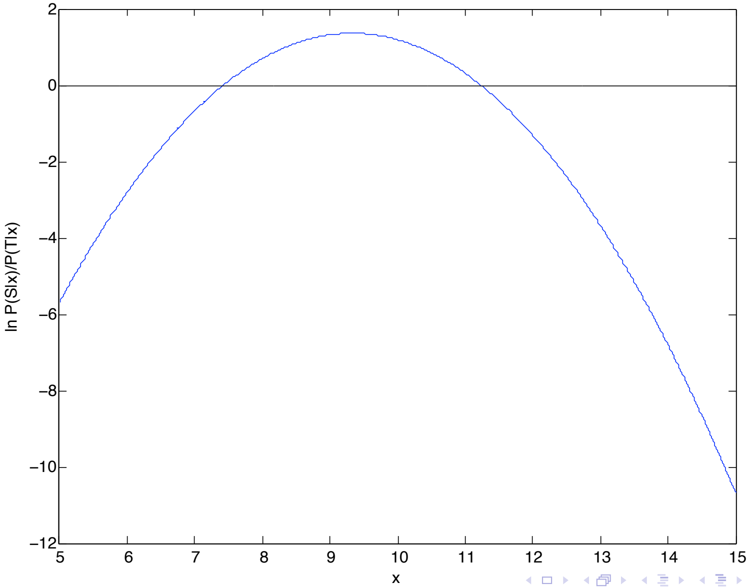
$$\ln \frac{P(S|X=x)}{P(T|X=x)} = -\frac{1}{2}\left(\frac{(x-\mu_S)^2}{\sigma_S^2} - \frac{(x-\mu_T)^2}{\sigma_T^2} + \ln \sigma_S^2 - \ln \sigma_T^2\right) + \ln P(S) - \ln P(T)$$

- In the example the priors are equal, so:

$$\ln \frac{P(S|X=x)}{P(T|X=x)} = -\frac{1}{2}\left(\frac{(x-\mu_S)^2}{\sigma_S^2} - \frac{(x-\mu_T)^2}{\sigma_T^2} + \ln \sigma_S^2 - \ln \sigma_T^2\right)$$
$$= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right)$$

- If log odds are less than 0 assign to T, otherwise assign to S.

Log odds



Example: unequal priors

- Now, assume $P(S) = 0.3$, $P(T) = 0.7$. Including this prior information, to which class should each of the above test data points (x^1, x^2, x^3) be assigned?

Example: unequal priors

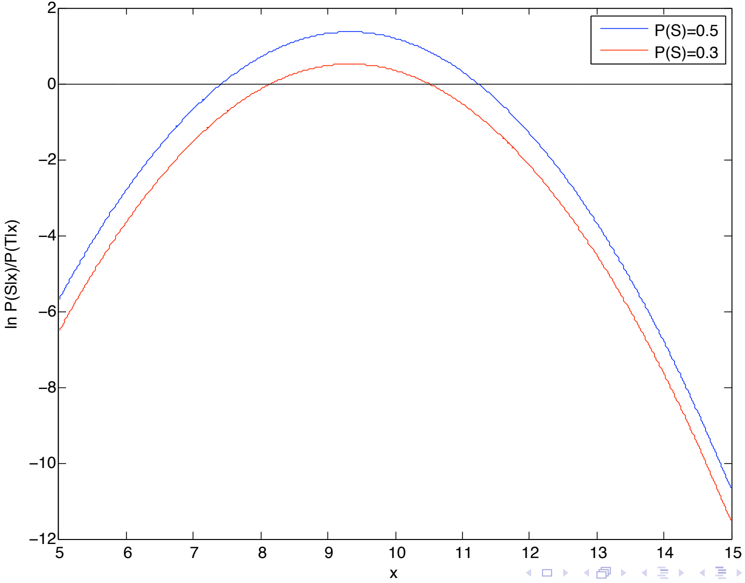
- Now, assume $P(S) = 0.3$, $P(T) = 0.7$. Including this prior information, to which class should each of the above test data points (x^1, x^2, x^3) be assigned?
- Again compute the log odds:

$$\ln \frac{P(S|X=x)}{P(T|X=x)} = -\frac{1}{2}\left(\frac{(x-\mu_S)^2}{\sigma_S^2} - \frac{(x-\mu_T)^2}{\sigma_T^2} + \ln \sigma_S^2 - \ln \sigma_T^2\right) + \ln P(S) - \ln P(T)$$

$$= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right) + \ln P(S) - \ln P(T)$$

$$= -\frac{1}{2}\left((x-10)^2 - \frac{(x-12)^2}{4} - \ln 4\right) + \ln(3/7)$$

Log odds



Multivariate Gaussian classifier

- Multivariate Gaussian (in d dimensions):

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Multivariate Gaussian classifier

- Multivariate Gaussian (in d dimensions):

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- Log likelihood:

$$LL(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Multivariate Gaussian classifier

- Multivariate Gaussian (in d dimensions):

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- Log likelihood:

$$LL(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

- If $p(\mathbf{x} \mid C) \sim p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the log posterior probability is:

$$\ln P(C|\mathbf{x}) \propto -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \ln P(C)$$

Example

- 2-dimensional data from three classes (A, B, C).

Example

- 2-dimensional data from three classes (A, B, C).
- The classes have equal prior probabilities.

Example

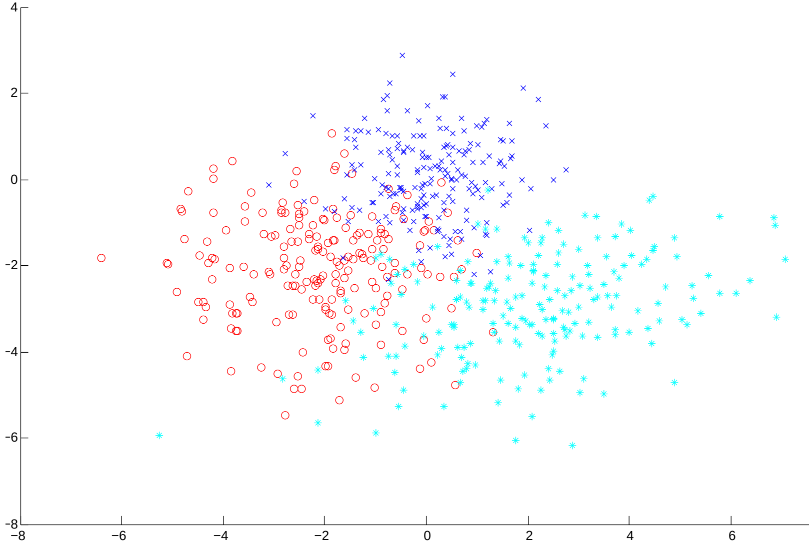
- 2-dimensional data from three classes (A, B, C).
- The classes have equal prior probabilities.
- 200 points in each class

Example

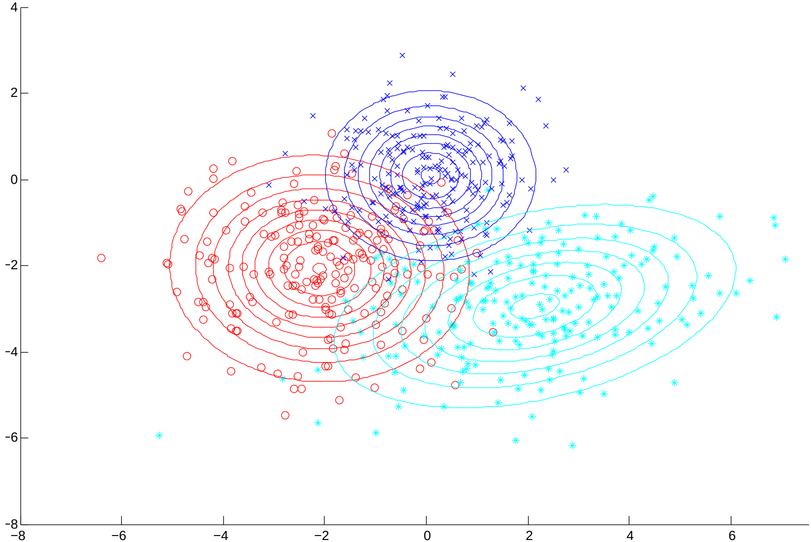
- 2-dimensional data from three classes (A, B, C).
- The classes have equal prior probabilities.
- 200 points in each class
- Load into Matlab ($n \times 2$ matrices, each row is a data point) and display using a scatter plot:

```
xa = load('trainA.dat');
xb = load('trainB.dat');
xc = load('trainC.dat');
hold on;
scatter(xa(:, 1), xa(:,2), 'r', 'o');
scatter(xb(:, 1), xb(:,2), 'b', 'x');
scatter(xc(:, 1), xc(:,2), 'c', '*');
```

Training data

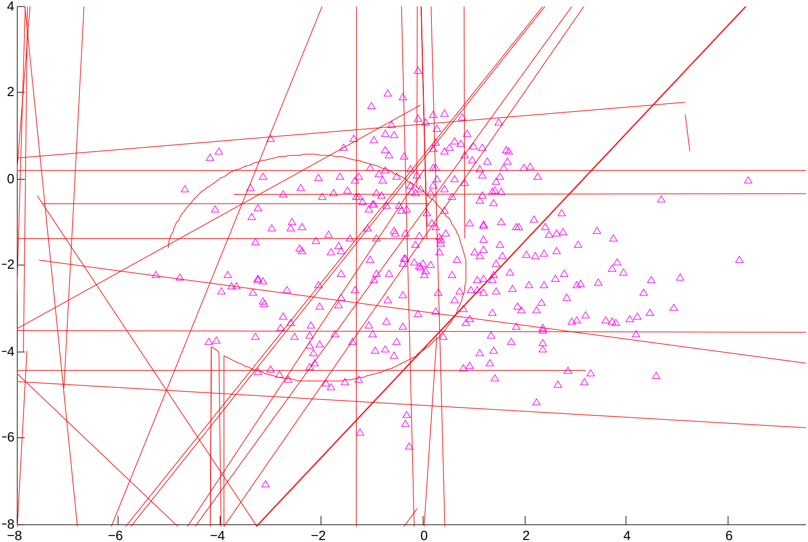


Gaussians estimated from training data

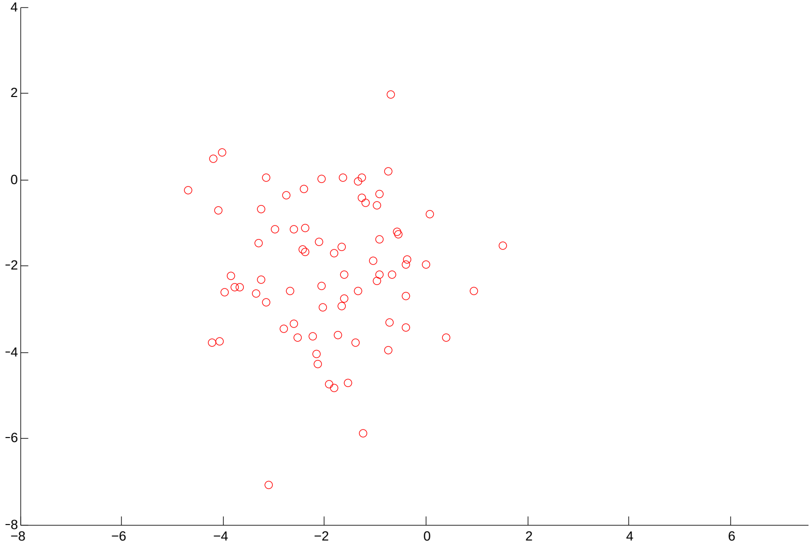


Testing data

Testing data — with estimated class distributions

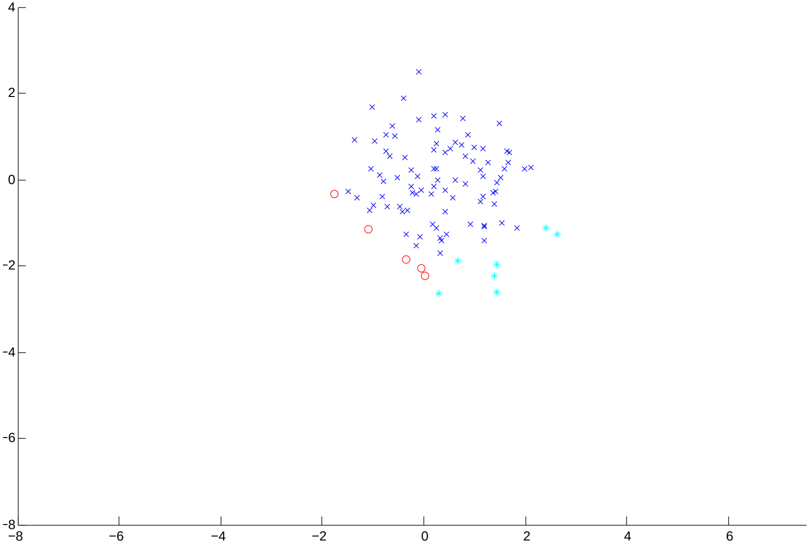


Testing data — with true classes indicated



Classifying test data from class A

Classifying test data from class B



Classifying test data from class C

Results

- Analyze results by percent correct, and in more detail with a confusion matrix

Results

- Analyze results by percent correct, and in more detail with a confusion matrix
  - Rows of a confusion matrix correspond to the predicted classes (classifier outputs)

Results

- Analyze results by percent correct, and in more detail with a confusion matrix
  - Rows of a confusion matrix correspond to the predicted classes (classifier outputs)
  - Columns correspond to the true class labels

Results

- Analyze results by percent correct, and in more detail with a confusion matrix
  - Rows of a confusion matrix correspond to the predicted classes (classifier outputs)
  - Columns correspond to the true class labels
  - Element (r, c) is the number of patterns from true class c that were classified as class r

Results

- Analyze results by percent correct, and in more detail with a confusion matrix
  - Rows of a confusion matrix correspond to the predicted classes (classifier outputs)
  - Columns correspond to the true class labels
  - Element (r, c) is the number of patterns from true class c that were classified as class r
  - Total number of correctly classified patterns is obtained by summing the numbers on the leading diagonal

Results

- Analyze results by percent correct, and in more detail with a confusion matrix
  - Rows of a confusion matrix correspond to the predicted classes (classifier outputs)
  - Columns correspond to the true class labels
  - Element (r, c) is the number of patterns from true class c that were classified as class r
  - Total number of correctly classified patterns is obtained by summing the numbers on the leading diagonal
- Confusion matrix in this case:

Test Data		True class A	True class B	True class C
Predicted class	A	77	5	9
	B	15	88	2
	C	8	7	89

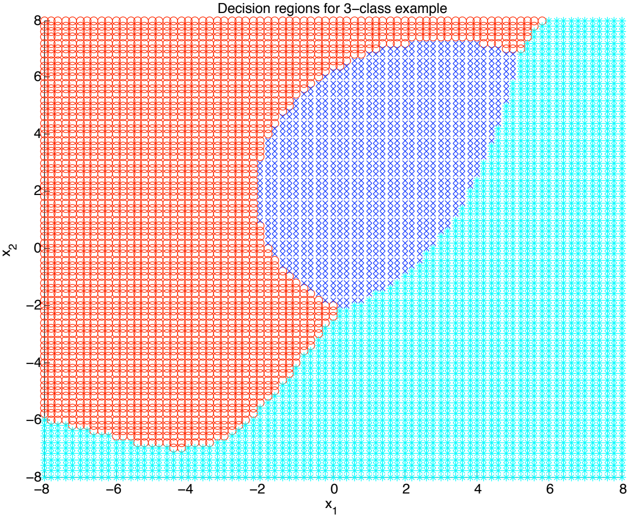
Results

- Analyze results by percent correct, and in more detail with a confusion matrix
  - Rows of a confusion matrix correspond to the predicted classes (classifier outputs)
  - Columns correspond to the true class labels
  - Element (r, c) is the number of patterns from true class c that were classified as class r
  - Total number of correctly classified patterns is obtained by summing the numbers on the leading diagonal
- Confusion matrix in this case:

		True class		
Test Data		A	B	C
Predicted	A	77	5	9
class	B	15	88	2
	C	8	7	89

- Overall proportion of test patterns correctly classified is $(77 + 88 + 89)/300 = 254/300 = 0.85$.

Decision Regions



Example: Classifying spoken vowels

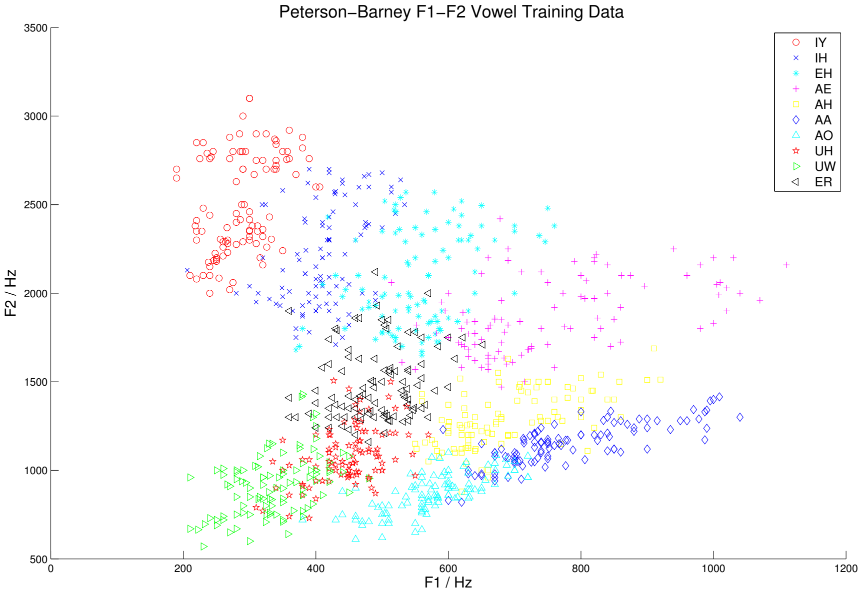
- 10 Spoken vowels in American English
- Vowels can be characterised by *formant frequencies* — resonances of vocal tract
  - there are usually three or four identifiable formants
  - first two formants written as F1 and F2
- Peterson-Barney data — recordings of spoken vowels by American men, women, and children
  - two examples of each vowel per person
  - for this example, data split into training and test sets
  - children's data not used in this example
  - different speakers in training and test sets
- (see `http://en.wikipedia.org/wiki/Vowel` for more)
- Classify the data using a Gaussian classifier
- Assume equal priors

The data

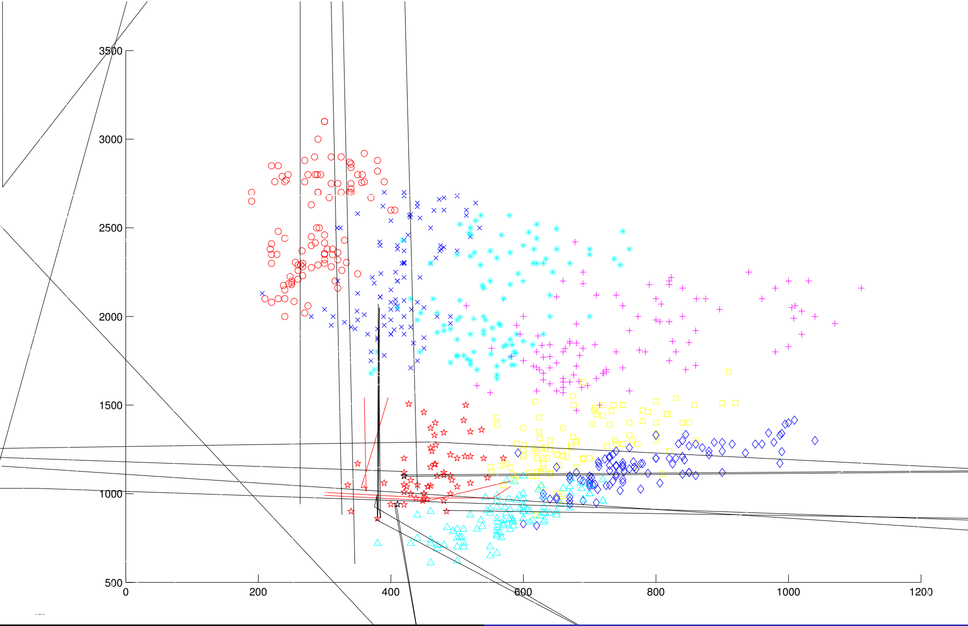
Ten steady-state vowels, frequencies of F1 and F2 at their centre:

- **IY** — "bee"
- **IH** — "big"
- **EH** — "red"
- **AE** — "at"
- **AH** — "honey"
- **AA** — "heart"
- **AO** — "frost"
- **UH** — "could"
- **UW** — "you"
- **ER** — "bird"

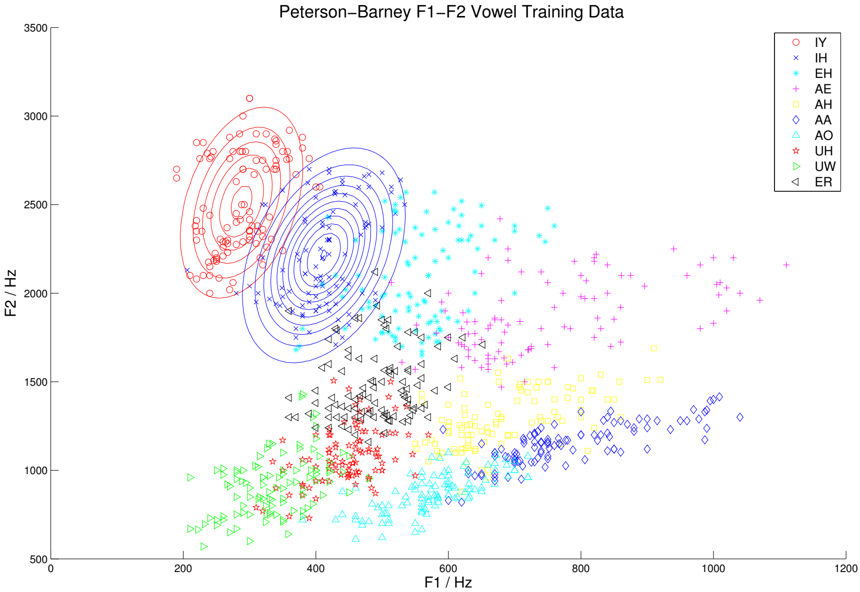
Vowel data — 10 classes



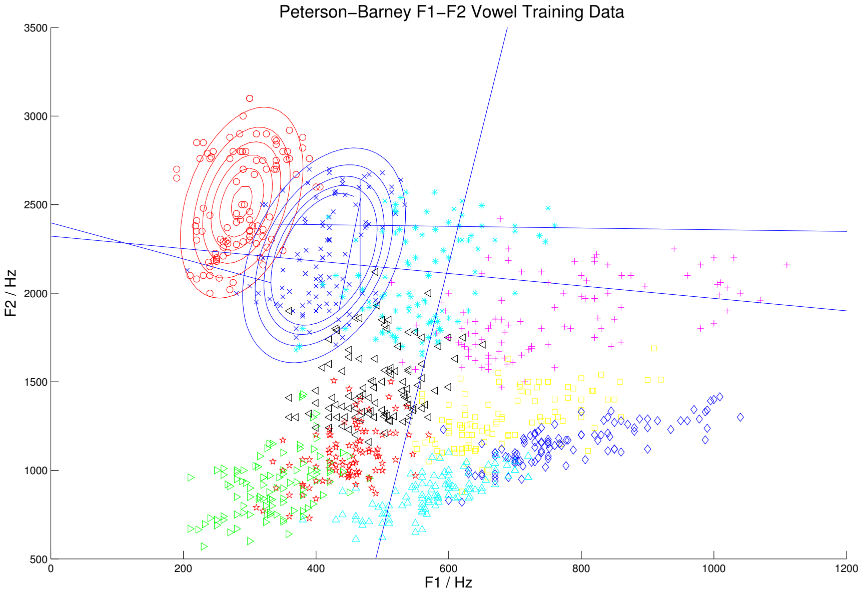
Gaussian for class 1 (IY)



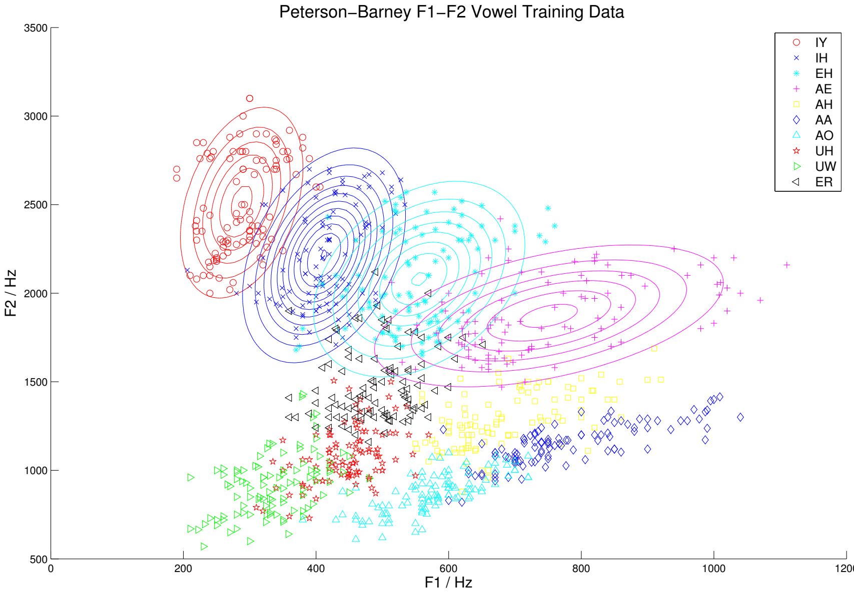
Gaussian for class 2 (IH)



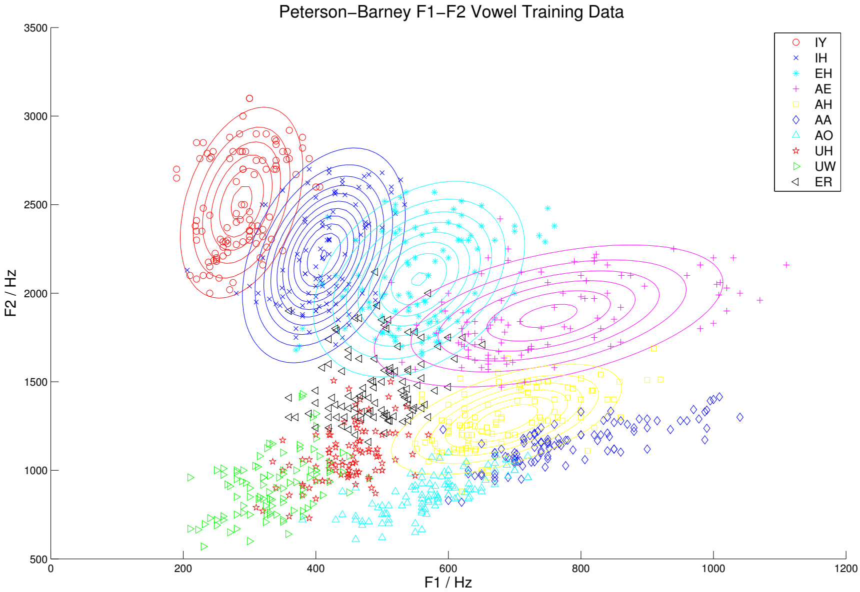
Gaussian for class 3 (EH)



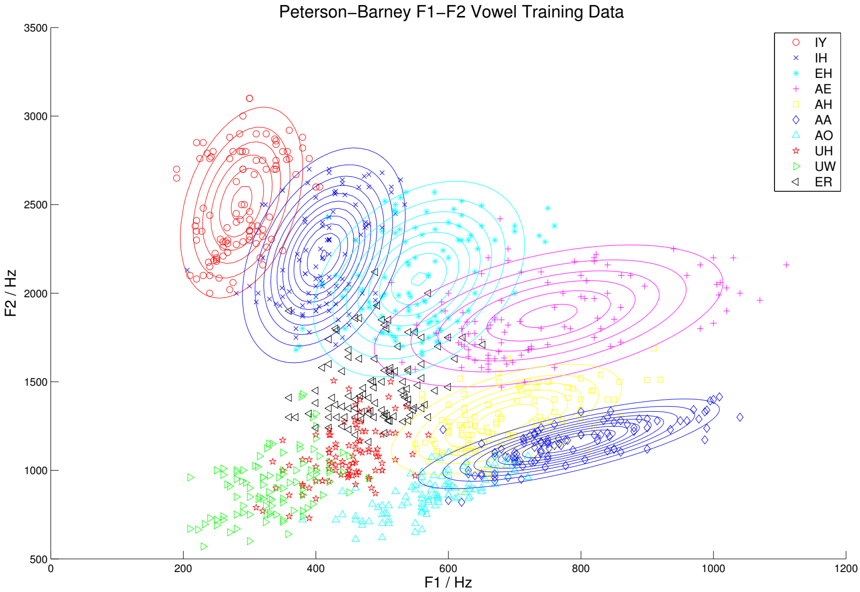
Gaussian for class 4 (AE)



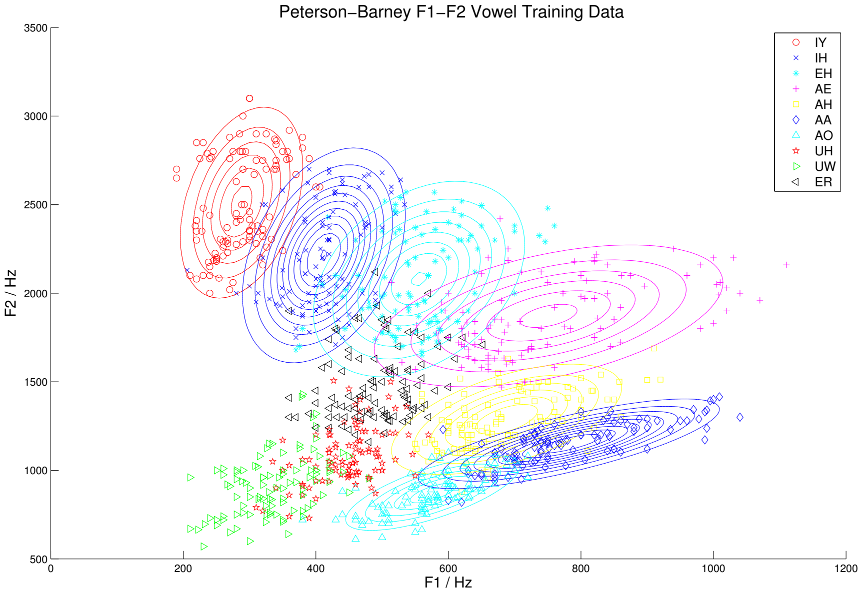
Gaussian for class 5 (AH)



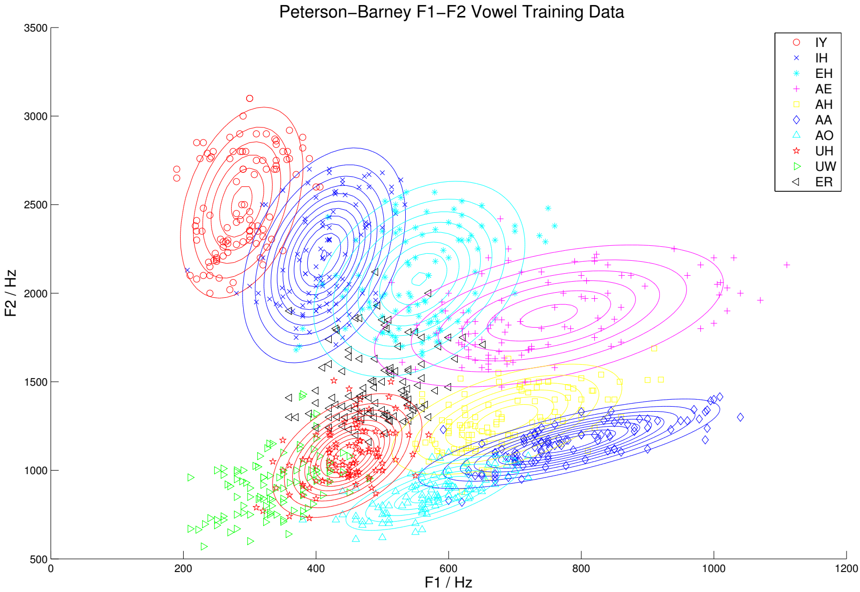
Gaussian for class 6 (AA)



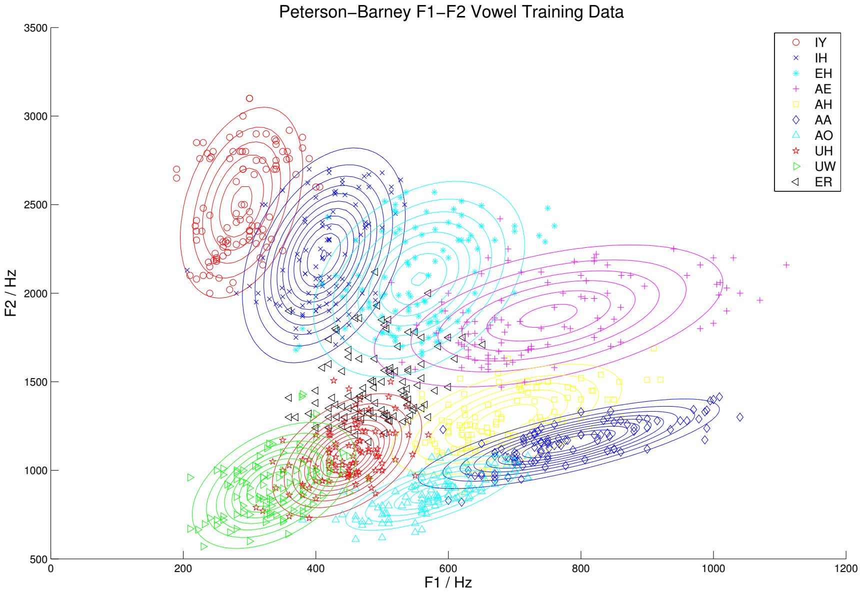
Gaussian for class 7 (AO)



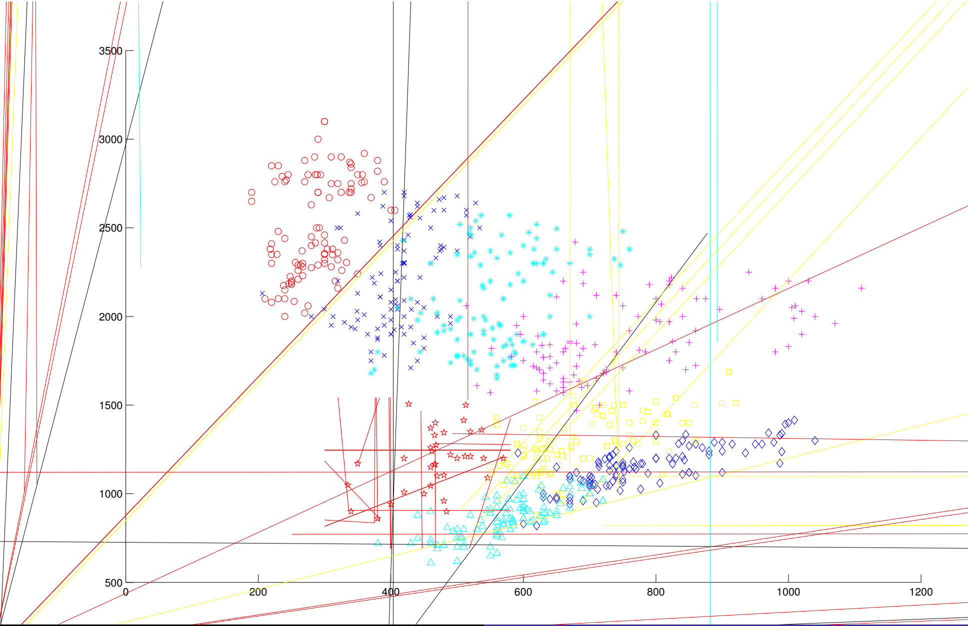
Gaussian for class 8 (UH)



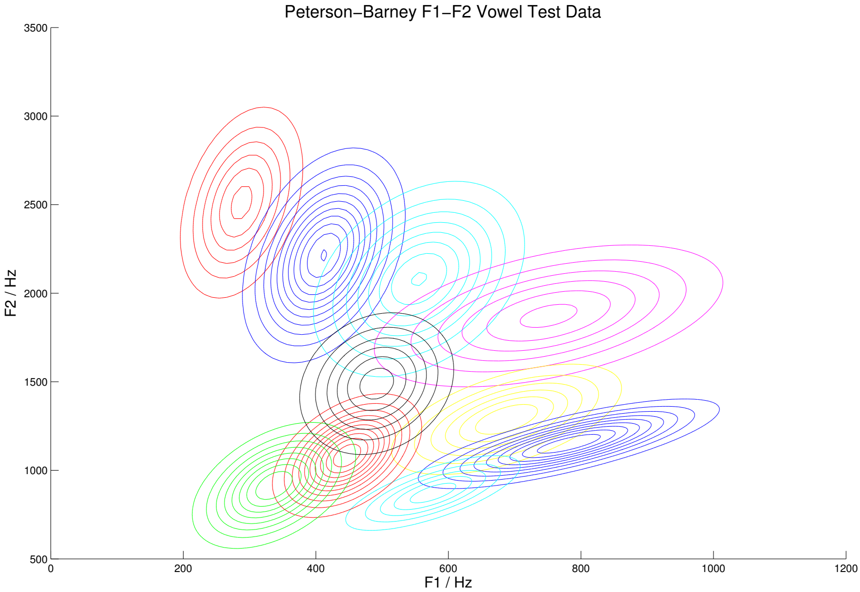
Gaussian for class 9 (UW)



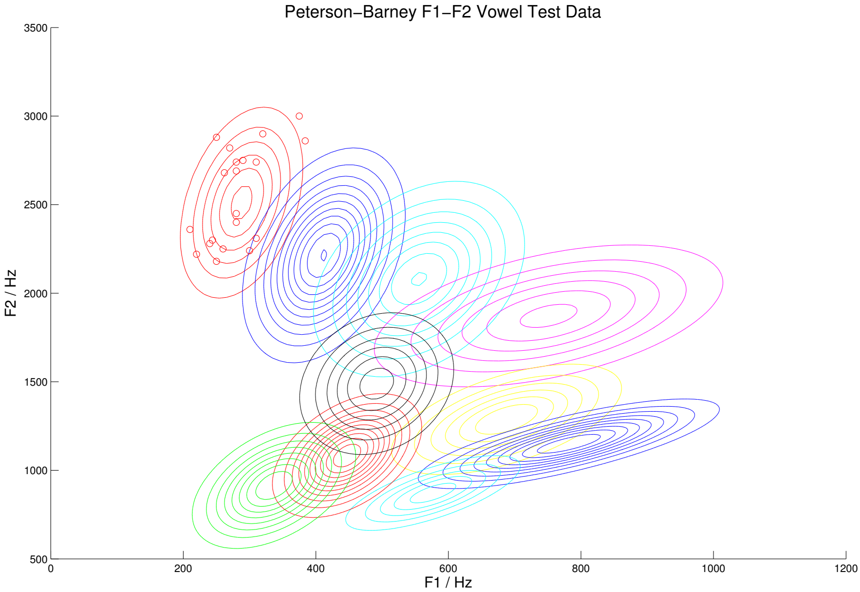
Gaussian for class 10 (ER)



Gaussians for each class



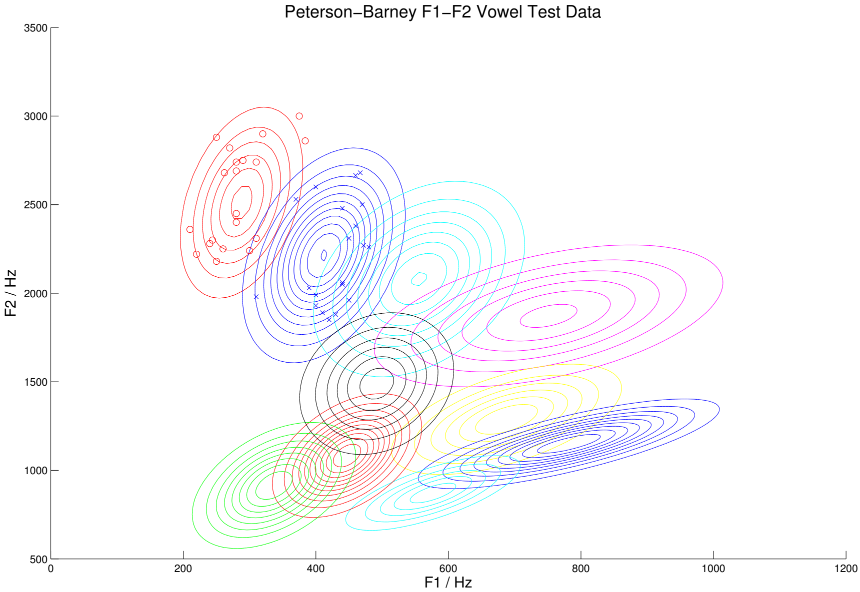
Test data for class 1 (IY)



Confusion matrix

	True class
	IY
IY	20
IH	0
EH	0
AE	0
AH	0
AA	0
AO	0
UH	0
UW	0
ER	0
% corr.	100

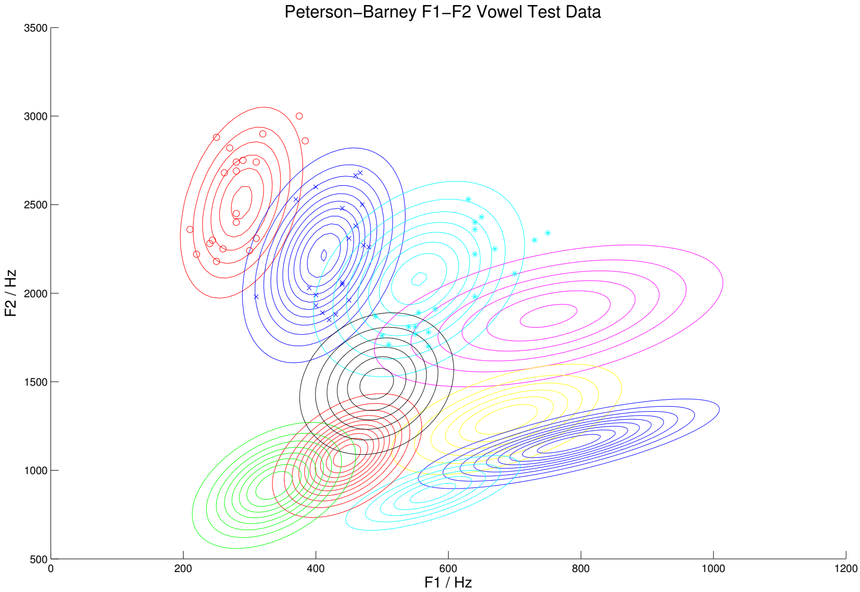
Test data for class 2 (IH)



Confusion matrix

	True class	
	IY	IH
IY	20	0
IH	0	20
EH	0	0
AE	0	0
AH	0	0
AA	0	0
AO	0	0
UH	0	0
UW	0	0
ER	0	0
% corr.	100	100

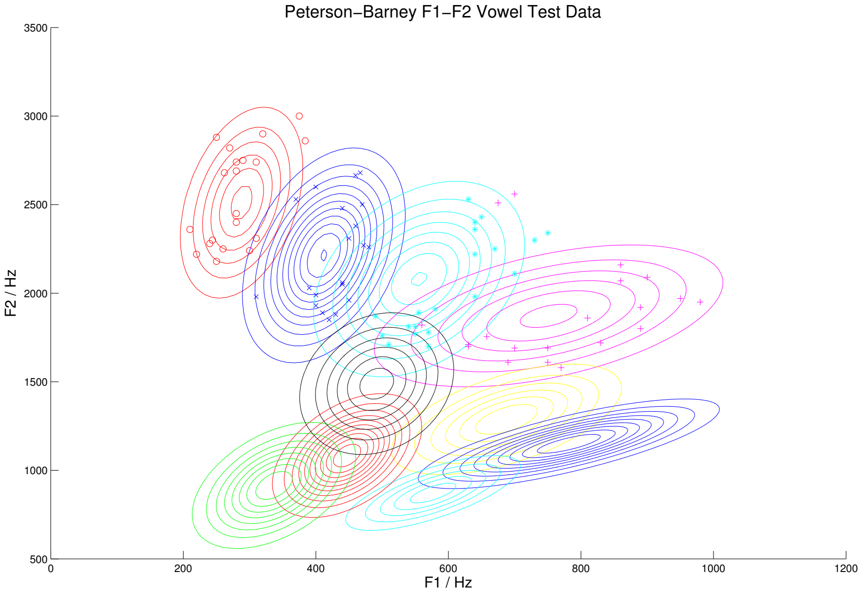
Test data for class 3 (EH)



Confusion matrix

	True class		
	IY	IH	EH
IY	20	0	0
IH	0	20	0
EH	0	0	15
AE	0	0	1
AH	0	0	0
AA	0	0	0
AO	0	0	0
UH	0	0	0
UW	0	0	0
ER	0	0	4
% corr.	100	100	75

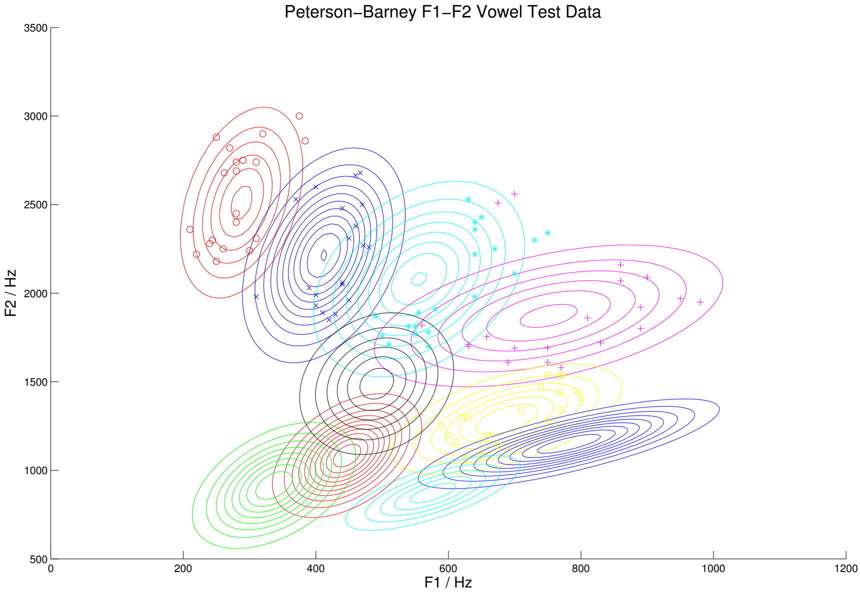
Test data for class 4 (AE)



Confusion matrix

	True class			
	IY	IH	EH	AE
IY	20	0	0	0
IH	0	20	0	0
EH	0	0	15	3
AE	0	0	1	16
AH	0	0	0	1
AA	0	0	0	0
AO	0	0	0	0
UH	0	0	0	0
UW	0	0	0	0
ER	0	0	4	0
% corr.	100	100	75	80

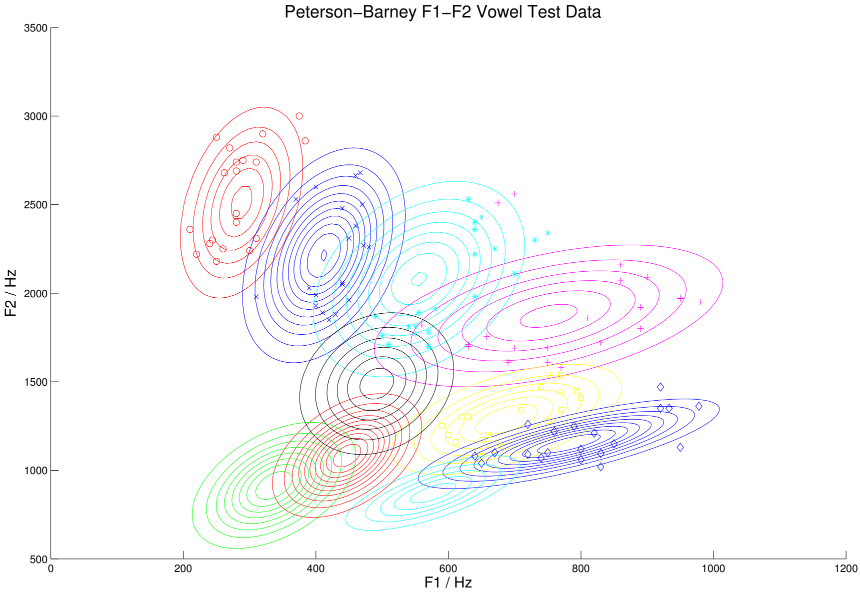
Test data for class 5 (AH)



Confusion matrix

	True class				
	IY	IH	EH	AE	AH
IY	20	0	0	0	0
IH	0	20	0	0	0
EH	0	0	15	3	0
AE	0	0	1	16	0
AH	0	0	0	1	18
AA	0	0	0	0	2
AO	0	0	0	0	0
UH	0	0	0	0	0
UW	0	0	0	0	0
ER	0	0	4	0	0
% corr.	100	100	75	80	90

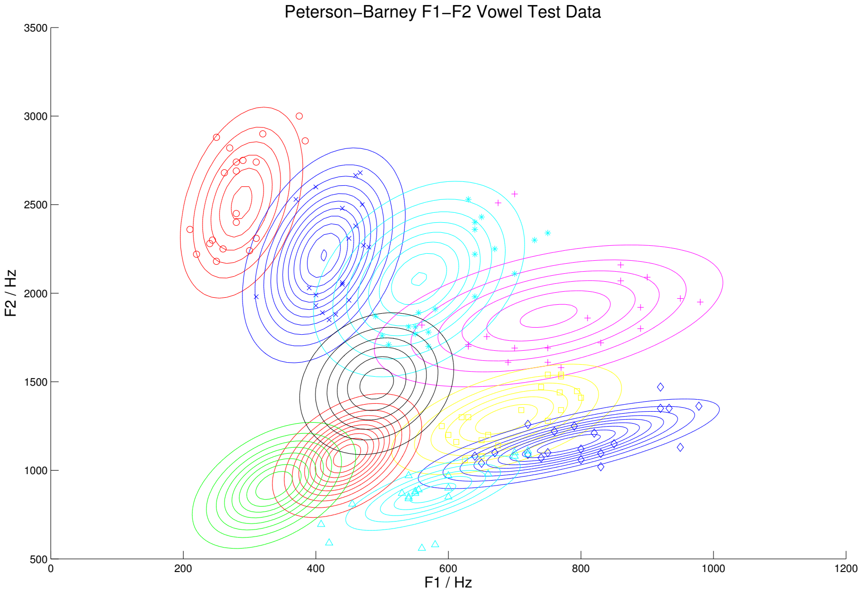
Test data for class 6 (AA)



Confusion matrix

	True class					
	IY	IH	EH	AE	AH	AA
IY	20	0	0	0	0	0
IH	0	20	0	0	0	0
EH	0	0	15	3	0	0
AE	0	0	1	16	0	0
AH	0	0	0	1	18	2
AA	0	0	0	0	2	17
AO	0	0	0	0	0	1
UH	0	0	0	0	0	0
UW	0	0	0	0	0	0
ER	0	0	4	0	0	0
% corr.	100	100	75	80	90	85

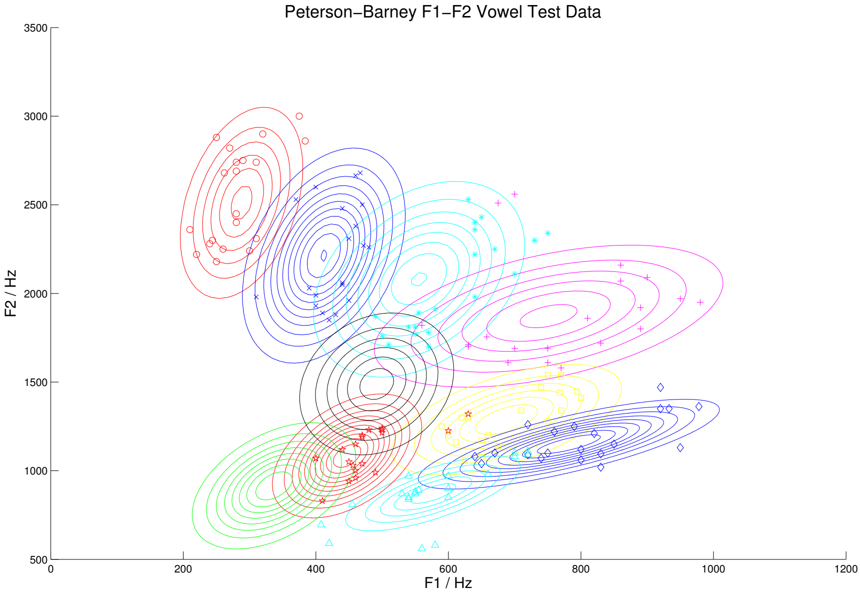
Test data for class 7 (AO)



Confusion matrix

	True class						
	IY	IH	EH	AE	AH	AA	AO
IY	20	0	0	0	0	0	0
IH	0	20	0	0	0	0	0
EH	0	0	15	3	0	0	0
AE	0	0	1	16	0	0	0
AH	0	0	0	1	18	2	0
AA	0	0	0	0	2	17	4
AO	0	0	0	0	0	1	16
UH	0	0	0	0	0	0	0
UW	0	0	0	0	0	0	0
ER	0	0	4	0	0	0	0
% corr.	100	100	75	80	90	85	80

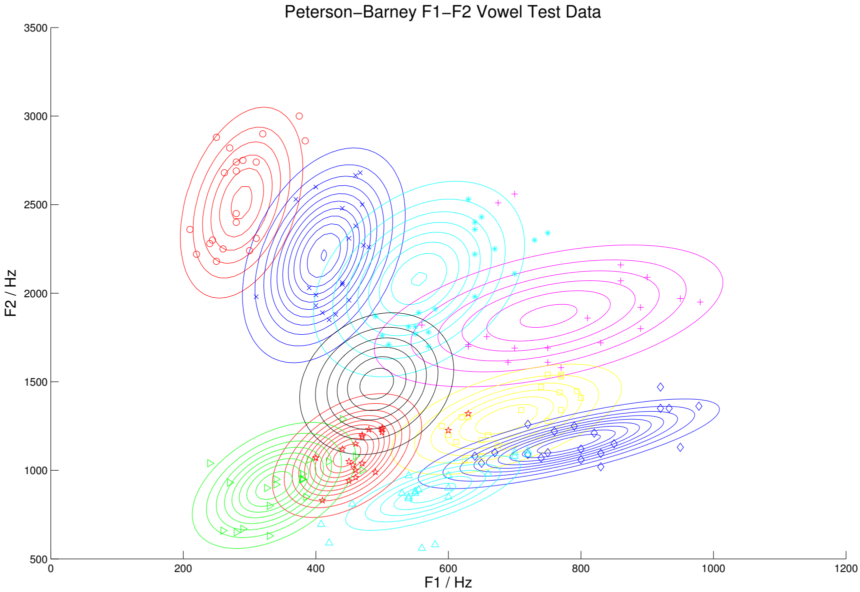
Test data for class 8 (UH)



Confusion matrix

	True class							
	IY	IH	EH	AE	AH	AA	AO	UH
IY	20	0	0	0	0	0	0	0
IH	0	20	0	0	0	0	0	0
EH	0	0	15	3	0	0	0	0
AE	0	0	1	16	0	0	0	0
AH	0	0	0	1	18	2	0	2
AA	0	0	0	0	2	17	4	0
AO	0	0	0	0	0	1	16	0
UH	0	0	0	0	0	0	0	18
UW	0	0	0	0	0	0	0	0
ER	0	0	4	0	0	0	0	0
% corr.	100	100	75	80	90	85	80	90

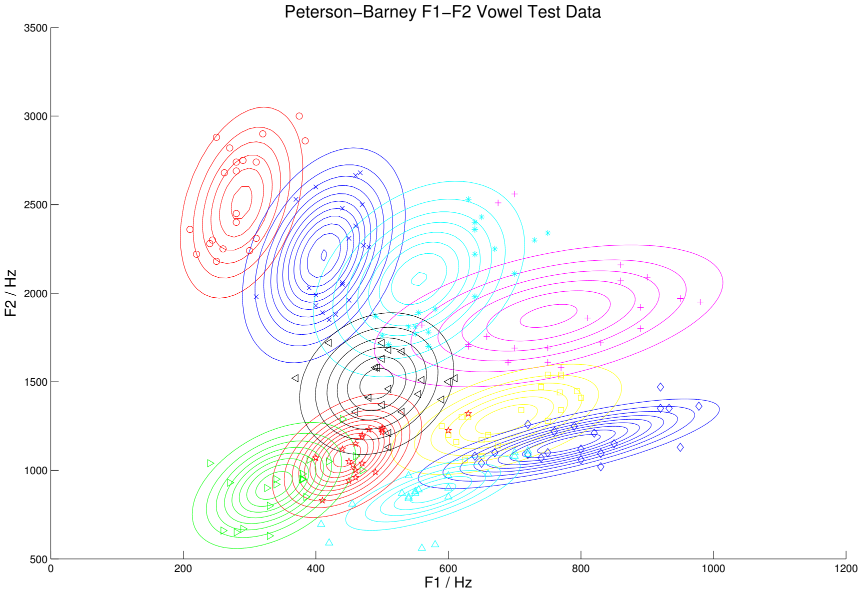
Test data for class 9 (UW)



Confusion matrix

	True class								
	IY	IH	EH	AE	AH	AA	AO	UH	UW
IY	20	0	0	0	0	0	0	0	0
IH	0	20	0	0	0	0	0	0	0
EH	0	0	15	3	0	0	0	0	0
AE	0	0	1	16	0	0	0	0	0
AH	0	0	0	1	18	2	0	2	0
AA	0	0	0	0	2	17	4	0	0
AO	0	0	0	0	0	1	16	0	0
UH	0	0	0	0	0	0	0	18	5
UW	0	0	0	0	0	0	0	0	15
ER	0	0	4	0	0	0	0	0	0
% corr.	100	100	75	80	90	85	80	90	75

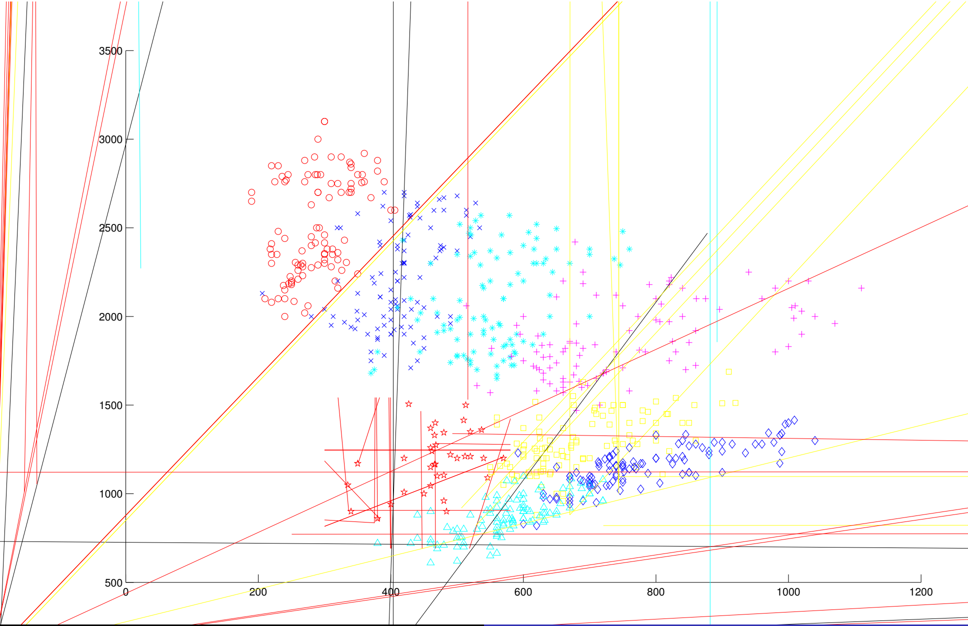
Test data for class 10 (ER)



Final confusion matrix

	True class									
	IY	IH	EH	AE	AH	AA	AO	UH	UW	ER
IY	20	0	0	0	0	0	0	0	0	0
IH	0	20	0	0	0	0	0	0	0	0
EH	0	0	15	3	0	0	0	0	0	0
AE	0	0	1	16	0	0	0	0	0	0
AH	0	0	0	1	18	2	0	2	0	0
AA	0	0	0	0	2	17	4	0	0	0
AO	0	0	0	0	0	1	16	0	0	0
UH	0	0	0	0	0	0	0	18	5	2
UW	0	0	0	0	0	0	0	0	15	0
ER	0	0	4	0	0	0	0	0	0	18
% corr.	100	100	75	80	90	85	80	90	75	90
	Total: 86.5% correct									

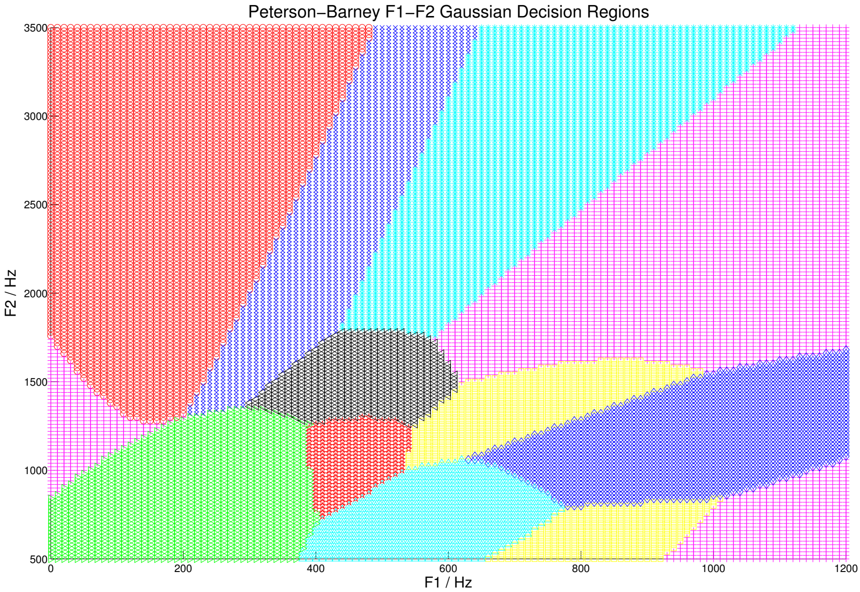
Classifying the training set



Training set confusion matrix

	True class									
	IY	IH	EH	AE	AH	AA	AO	UH	UW	ER
IY	99	8	0	0	0	0	0	0	0	0
IH	3	85	15	0	0	0	0	0	0	3
EH	0	7	69	11	0	0	0	0	0	11
AE	0	0	5	86	4	0	0	0	0	4
AH	0	0	0	3	87	8	3	2	0	1
AA	0	0	0	0	4	82	10	0	0	0
AO	0	0	0	0	5	12	86	2	0	0
UH	0	0	0	0	0	0	2	73	19	10
UW	0	0	0	0	0	0	1	15	79	1
ER	0	2	13	2	2	0	0	10	4	72
%	97.1	83.3	67.6	84.3	85.3	80.4	84.3	71.6	77.5	70.6
	Total: 80.2% correct									

Decision Regions



Summary

- Using Bayes' theorem with pdfs
- The Gaussian classifier: 1-dimensional and multi-dimensional
- Vowel classification example