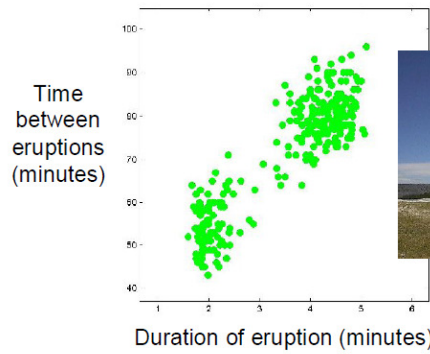
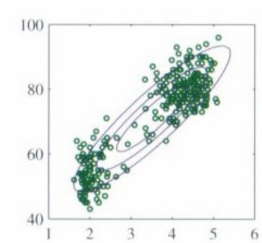
More on the Gaussian distribution

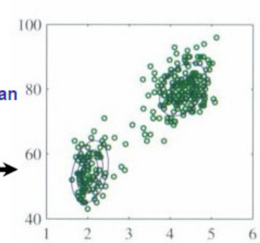
Old Faithful Data Set





Bad description of bimodal data using the inherently unimodal Gaussian

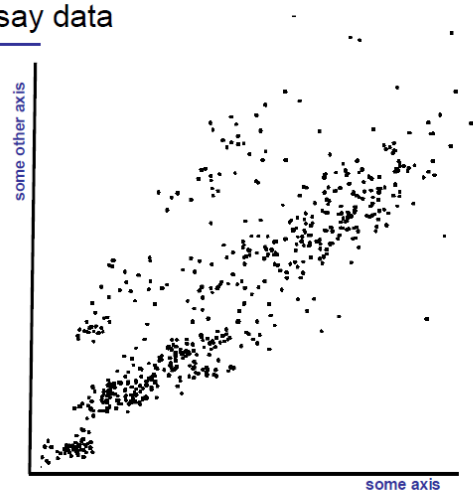
However, if we are willing to use more than one Gaussian, we can fit one to each mode or cluster of the data.



Biological assessment data

Some Bio Assay data

Sometimes it may not even be easy to tell how many "bumps" there are…



Idea: Use a Mixture of Gaussians

- Linear super-position of Gaussians

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

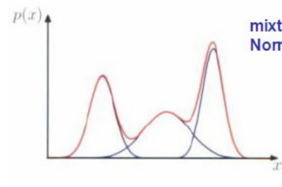
- Normalization and positivity require

$$\sum_{k=1}^{K} \pi_k = 1 \qquad 0 \leqslant \pi_k \leqslant 1$$

- Can interpret the mixing coefficients as prior probabilities

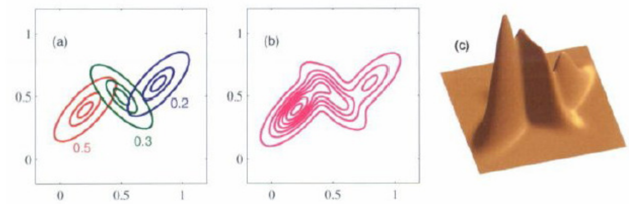
$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k)$$

Mixture of Gaussians

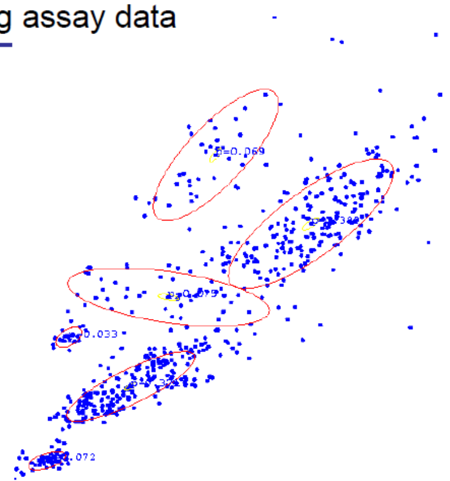


mixture of 3 one-dimensional Normal distributions

mixture of 3 two-dimensional Gaussians



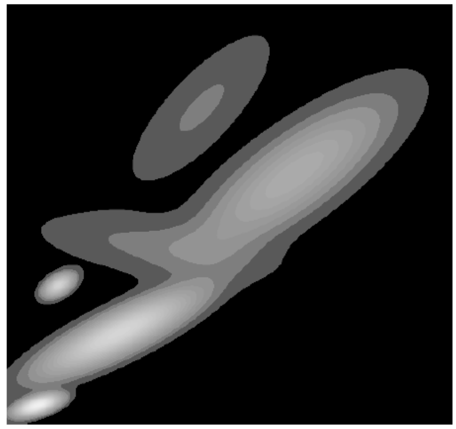
GMM describing assay data



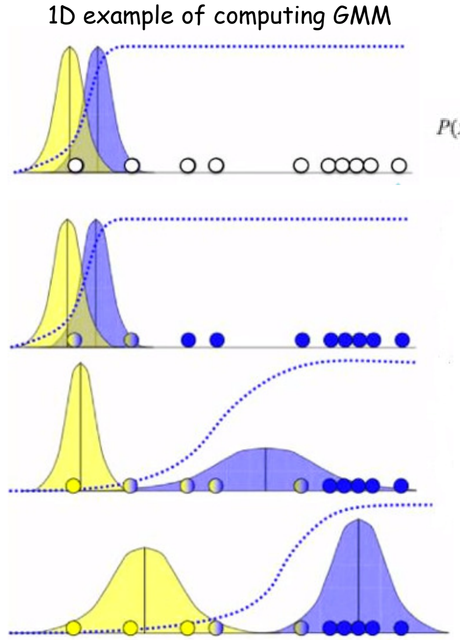
GMM density function

Note: now we have a continuous estimate of the density, so can estimate a value at any point.

Also, could draw constant-probability contours if we wanted to.



1D example of computing GMM



$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$

Matching Histograms/Distributions

- Histogram matching has many applications
  - sensed signal segmentation
  - sensed signal tracking
  - content-based retrieval – in a database of sensed signal, finding out those signals that have histograms similar to the histogram of query signal.
- Before matching two histograms, they are converted into normalized histograms
- Several metrics exist for matching normalized histograms:
  - Bhattacharyya coefficient
  - Kullback-Liebler divergence
  - Diffusion distance

Bhattacharyya Distance

- Considered two normalized histograms P and Q.

Let b_i be the i^{th} bin of P (respectively Q), $i = 1, \cdots, B$.

The Bhattacharyya coefficient d^{bt} between P and Q is given by:

$$d^{bt} = \sqrt{1 - \sum_{i=1}^{B} \sqrt{P(b_i) Q(b_i)}}$$

This can also be written as:

$$d = \left(d^{bt}\right)^2 - 1 = -\sum_{i=1}^{B} \sqrt{P(b_i) Q(b_i)}$$

Bhattacharyya distance is then defined as:

$D = \ln d$

Bhattacharyya Distance – Cont.

A computationally simple form of the above results:

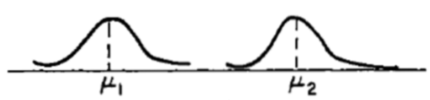
$$d = \frac{1}{4}\ln\left\{\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2}+\frac{\sigma_q^2}{\sigma_p^2}+2\right)\right\}+\frac{1}{4}\left\{\frac{(\mu_p-\mu_q)^2}{\sigma_p^2+\sigma_q^2}\right\}$$

where:

σ_p^2 is the variance of the p-th distribution

μ_p is the mean of the p-th distribution

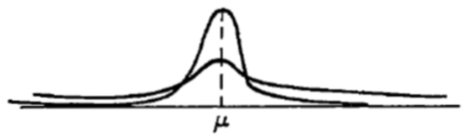
d is the Bhattacharyya distance between p and q distributions



(a)

When the variances are equal but means are not, the first term of the Bhattacharyya measure will be zero but the second term will be nonzero.

The second term will be large if the variance is small under this condition, implying that a large difference in means accompanied by small variances, is a desirable quality in a feature for distinguishing between two clusters.

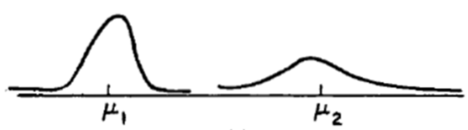


(b)

The above situation is the reverse, that is, the means are equal but the variance is not.

If the variances are significantly different, the feature is still considered of potential usefulness in separating the clusters.

Thus in this situation, the second term of the Bhattacharyyya distance will be zero but the first term will be nonzeron.



(c)

In this case, both the mean and variance are unequal and both terms of the measure will be nonzero.

For example, the feature rejection criterion would be to only retain those features with large Bhattacharyya value.

Regression

•Linear (straight-line) relationships between two quantitative variables are easy to understand and quit common.

•Correlation measures the direction and strength of these relationships.

•When scatter plot shows a linear relationship, we would like to summarize the overall pattern by drawing a line on the scatter plot.

•A regression line is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes.

•We often use a regression line to predict the value of *y* for a given value of *x*.

•We will use the line to predict *y* from *x*, so the prediction errors we make are errors in *y*, the vertical direction in the scatterplot.

•We want *vertical* distances of the points from the line to be as small as possible.

•There are many ways to make the collection of vertical distances "*as small as possible*".

•The most common is the *least-squares method*.

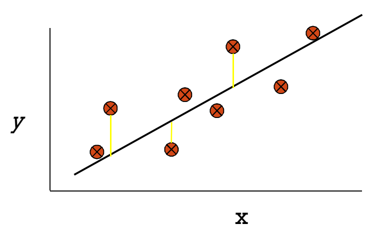
Least-Squares Regression Line

- The least-squares regression line of *y* on *x* is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
- We have data on an explanatory variable *x* and a response variable *y* for *n measures* (individuals).
- From the data, calculate the means μ_x and μ_y and the standard deviations σ_x and σ_y of the two variables, and their correlation *r*.
- The least-squares regression line is :

$$\hat{y} = a + bx$$

$$\text{with slop: } b = r\frac{\sigma_x}{\sigma_y}$$

$$\text{and intercept: } a = \mu_y - b\mu_x$$



Which line should we use?

Choose an objective function

For simple linear regression we choose sum squared error (SSE)

Sum (*actual*$_i$ -*predicted*$_i$)^2 = Sum (*residue*$_i$)^2

Thus, find the line which minimizes the sum of the squared residues (e.g. least squares)

How do we "learn" parameters

- For the 2D problem (line) we have defined coefficients for the bias and the independent variable (i.e. *y*-intercept and slope)

$$\hat{y} = a + bx$$

- Least Squares problem is defined as:

$$\varepsilon = \sum_i (y_i - \hat{y}_i)^2 = \sum_i [y_i - (a + bx_i)]^2$$

- The estimation of the parameters is obtained using basic results from calculus and, specifically, uses the property that a quadratic expression reaches its minimum value when its derivatives vanish.

- Taking the derivative of ε with respect to a and b and setting them to zero, we have:

$$\frac{\partial \varepsilon}{\partial a} = \frac{\partial \sum (y_i - a - bx_i)^2}{\partial a} = 0$$

$$= -2(\sum y_i - a - b\sum x_i) = -2(n\mu_y - na - nb\mu_x)$$

$$\Rightarrow a = \mu_y - b\mu_x$$

$$\frac{\partial \varepsilon}{\partial b} = \frac{\partial \sum (y_i - a - bx_i)^2}{\partial b} = 0$$

$$= 2b\sum x_i^2 + 2a\sum x_i - 2\sum y_i x_i = 0.$$

$$= -2\sum x_i (y_i - a - bx_i) = -2\sum x_i (y_i - \mu_y + b\mu_x - bx_i) = 0$$

$$b\sum x_i (x_i - \mu_x) = \sum x_i (y_i - \mu_y)$$

$$\Rightarrow b = \frac{\sum (y_i - \mu_y)(x_i - \mu_x)}{\sum (x_i - \mu_x)^2}$$

Facts about least-squares regression

•The distinction between explanatory and response variables is essential in regression. Least-squares regression looks at the distances of the data points from the line only in the y-direction. If we reverse the roles of the two variables, we get a different least-squares regression line.

•There is a close connection between correlation and the slope of the least-square line. For example, A change in one standard deviation in *x* corresponds to a change of *r* standard deviation in *y*.

- The least-squares regression line always passes through the point μ_x, μ_y.

Outliers and influential observation in regression

•An outliers is an observation that lies outside the overall pattern of the other observations.

•Points that are outliers in the y direction of a scatterplot have large regression residuals.

•An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation.

•Points in the x-direction of a scatterplot are often **influential** for the least-squares regression line.

Multivariate Linear Regression

- We have seen how to solve a simple linear regression model.
- A dependent variable guided by a single independent variable.
- In general, one dependent variable may be influenced by many independent variables.
- Multivariate Regression is a method of modeling multiple responses, or dependent variables, with a single set of predictor variables.

Let y be a function defined as a linear combination of two independent variables x_1 and x_2 : $\hat{y} \approx b_1 x_1 + b_2 x_2$

Let $Q(b_1, b_2) = \varepsilon.\varepsilon = \varepsilon^2 =$
$(y - b_1 x_1 - b_2 x_2)(y - b_1 x_1 - b_2 x_2)$

$$\frac{\partial Q}{\partial b_1} = -x_1.(y - b_1 x_1 - b_2 x_2) - (y - b_1 x_1 - b_2 x_2).x_1$$
$$\frac{\partial Q}{\partial b_2} = -x_2.(y - b_1 x_1 - b_2 x_2) - (y - b_1 x_1 - b_2 x_2).x_2$$

Setteng the derivatives equal to zero, we have:

$$(x_1.y) = (x_1.x_1)b_1 + (x_1.x_2)b_2$$
$$(x_2.y) = (x_2.x_1)b_1 + (x_2.x_2)b_2$$

Or,

$$\begin{bmatrix}(x_1.y)\\(x_2.y)\end{bmatrix}=\begin{bmatrix}(x_1.x_1) & (x_1.x_2)\\(x_2.x_1) & (x_2.x_2)\end{bmatrix}\begin{bmatrix}b_1\\b_2\end{bmatrix}$$

This matrix representation can be generalized, for example for the case of dimensions of vector *y* and each *x* be of dimension three.

$$X=[\overline{x}_1 \quad \overline{x}_2]=\begin{bmatrix}x_{11} & x_{12}\\x_{21} & x_{22}\\x_{31} & x_{32}\end{bmatrix}$$

$$X^T=\begin{bmatrix}x_{11} & x_{21} & x_{31}\\x_{12} & x_{22} & x_{32}\end{bmatrix}$$

For example,

$$\begin{bmatrix}(x_1.x_1) & (x_1.x_2)\\(x_2.x_1) & (x_2.x_2)\end{bmatrix}=\begin{bmatrix}x_{11} & x_{21} & x_{31}\\x_{12} & x_{22} & x_{32}\end{bmatrix}\begin{bmatrix}x_{11} & x_{12}\\x_{12} & x_{22}\\x_{13} & x_{32}\end{bmatrix}$$ (an example of covariance matrix)

For example, the following equation:

$$\begin{bmatrix}(x_1.y)\\(x_2.y)\end{bmatrix}=\begin{bmatrix}(x_1.x_1) & (x_1.x_2)\\(x_2.x_1) & (x_2.x_2)\end{bmatrix}\begin{bmatrix}b_1\\b_2\end{bmatrix}$$

can be written as:

$$\begin{bmatrix}x_{11} & x_{21} & x_{31}\\x_{12} & x_{22} & x_{32}\end{bmatrix}\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix}=\begin{bmatrix}x_{11} & x_{21} & x_{31}\\x_{12} & x_{22} & x_{32}\end{bmatrix}\begin{bmatrix}x_{11} & x_{12}\\x_{12} & x_{22}\\x_{13} & x_{32}\end{bmatrix}\begin{bmatrix}b_1\\b_2\end{bmatrix}$$

Or,

$$X^TY=\left(X^TX\right)B$$

The above equation leads to an analytic solution for B using an inverse matrix.

$$B=\left(X^TX\right)^{-1}X^TY$$

The above equation is the central result of least-squares analysis.