Summer School in Statistics for
Astronomers and Physicists III
June 4-9, 2007

Mixture Models and The EM Algorithm

Tom Hettmansperger
Department of Statistics
Penn State University

Mixtures of Normal Distributions

- Consider the problem of analyzing data that comes from a mixture of two or more normal populations.
- We **do not** have labels that indicate which population a particular data point comes from.
- If there are two populations then we have 5 parameters: a mixing proportion, 2 means, and 2 variances.
- We wish to estimate these 5 parameters and provide standard errors to assess their precision.

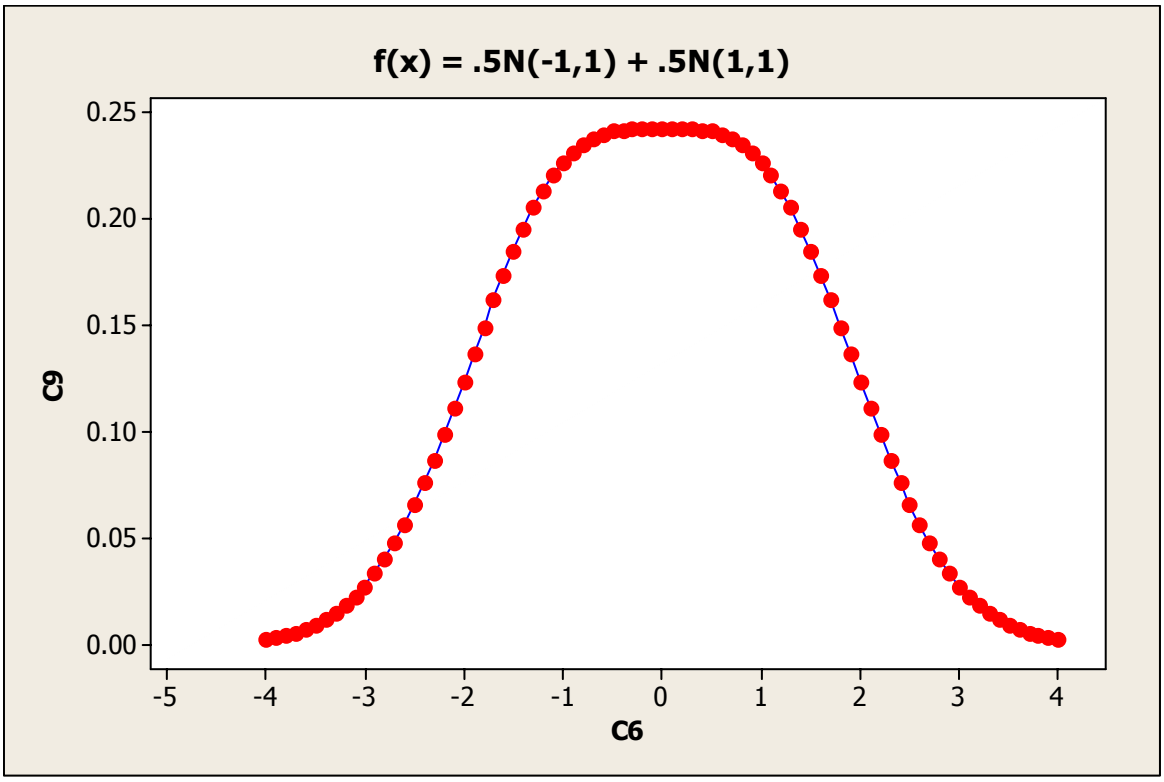
We write the model as:

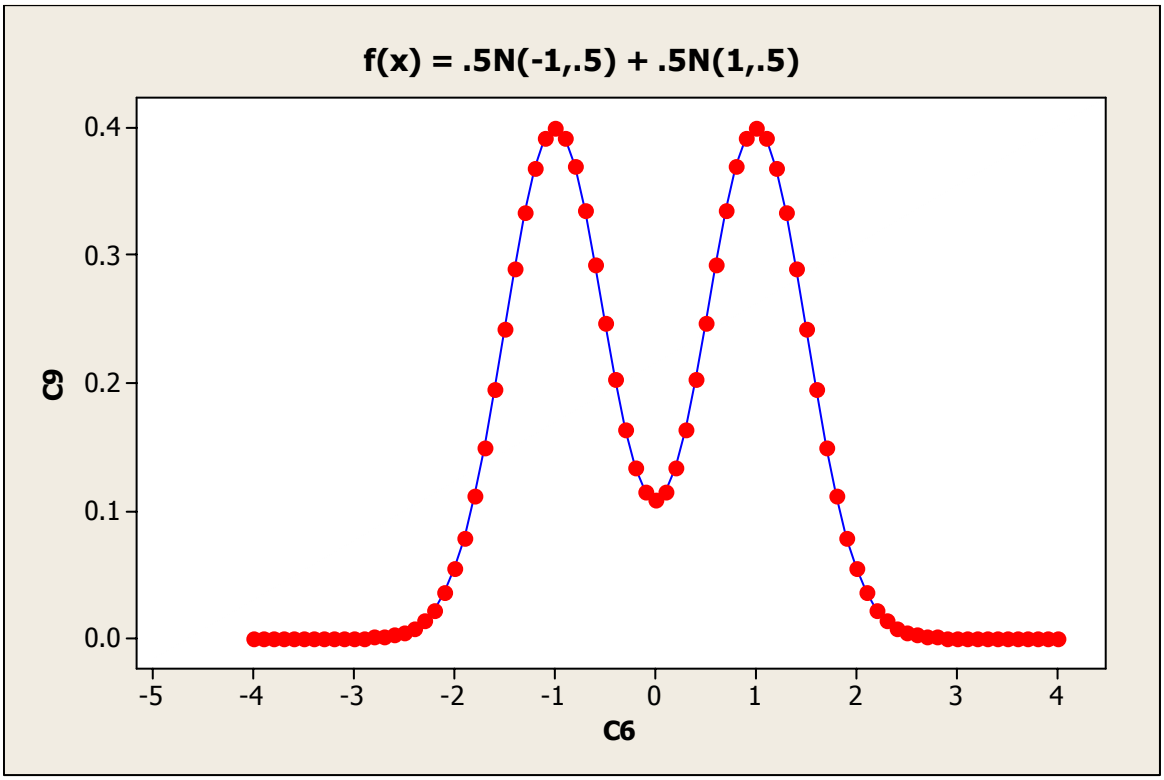
$$f(x) = \lambda f_1(x) + (1 - \lambda) f_2(x)$$

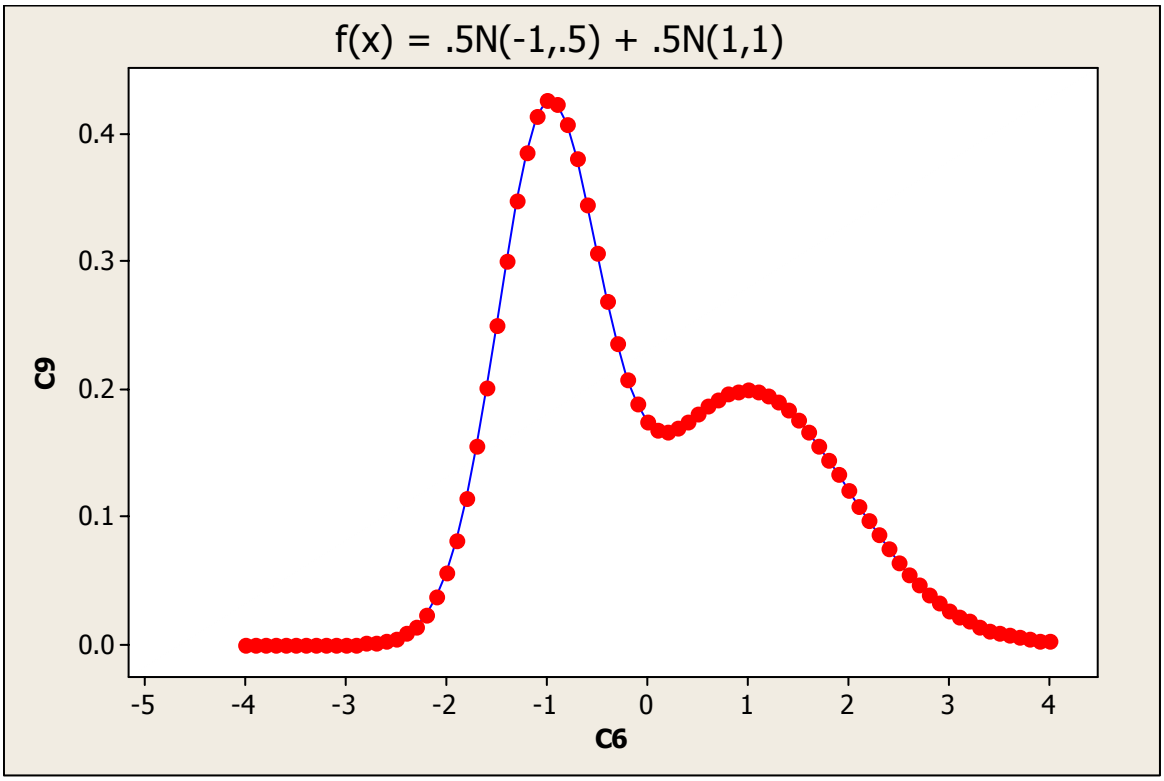
Where $f_1(x)$ and $f_2(x)$ are the two normal pdfs.

Some examples of possible shapes

determined by mixing normal distributions.




f(x) = .5N(-1,.5) + .5N(1,.5)
C9
0.0
0.1
0.2
0.3
0.4
-5
-4
-3
-2
-1
0
1
2
3
4
C6


f(x) = .5N(-1,.5) + .5N(1,1)
C9
0.0
0.1
0.2
0.3
0.4
-5
-4
-3
-2
-1
0
1
2
3
4
C6

We will use **maximum likelihood** to estimate the parameters. Given a sample $x_1, \ldots, x_n$ compute:

$$L(\Psi) = \prod_{i=1}^{n} \{\lambda \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp(-\frac{1}{2\sigma_1^2}(x_i - \mu_1)^2 +$$

$$(1-\lambda)\frac{1}{\sqrt{2\pi}\,\sigma_2} \exp(-\frac{1}{2\sigma_2^2}(x_i - \mu_2)^2\}$$

where $\Psi^T = (\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ is the parameter vector.

This likelihood function is quite difficult to maximize directly. Hence, we will introduce a new calculation device called the **EM algorithm**.

Suppose, for the moment, that we know which subpopulation the various data points come from.

Let $z_i = 1$ if x_i is from the first population and 0 otherwise. Then $P(z_i = 1) = \lambda$.

The **complete data** is given by $(x_1, z_1), \ldots, (x_n, z_n)$.

The joint distribution of (X_i, Z_i) is given by

$$g(x,z) = f_1^{\,z}(x) f_2^{\,1-z}(x) \lambda^z (1-\lambda)^{1-z}$$

where $f_1 and\, f_2$ are the two normal pdfs.

The **complete data likelihood** is given by:

$$L_c(\Psi) = \prod_{i=1}^{n}\left[\lambda \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp(-\frac{1}{2\sigma_1^2}(x_i-\mu_1)^2\right]^{z_i} \times$$

$$\left[(1-\lambda) \frac{1}{\sqrt{2\pi}\,\sigma_2} \exp(-\frac{1}{2\sigma_2^2}(x_i-\mu_2)^2\right]^{1-z_i}$$

$$\log L_c(\Psi) =$$

$$-\sum_{i=1}^{n} z_i \frac{1}{2\sigma_1^2}(x_i-\mu_1)^2 - \sum_{i=1}^{n}(1-z_i)\frac{1}{2\sigma_2^2}(x_i-\mu_2)^2$$

$$+\log\lambda \sum_{i=1}^{n} z_i + \log(1-\lambda)\sum_{i=1}^{n}(1-z_i)$$

$$-\log\sigma_1 \sum_{i=1}^{n} z_i - \log\sigma_2 \sum_{i=1}^{n}(1-z_i) + K$$

$$\frac{\partial \log L_c}{\partial \mu_1} \Rightarrow \sum_{i=1}^{n} z_i (x_i - \mu_1) = 0 \text{ and } \hat{\mu}_1 = \frac{\Sigma z_i x_i}{\Sigma z_i}$$

$$\frac{\partial \log L_c}{\partial \sigma_1^2} \Rightarrow \hat{\sigma}_1^2 = \frac{\Sigma z_i (x_i - \hat{\mu}_1)^2}{\Sigma z_i}$$

$$\frac{\partial \log L_c}{\partial \lambda} \Rightarrow \hat{\lambda} = \frac{\Sigma z_i}{n}$$

Likewise for $\hat{\mu}_2$ and $\hat{\sigma}_2^2$.

Call this set of 5 estimates $Eqns\ (1)$.

Easy to compute.

The problem: we don't know z_i!

The solution:

replace $\log L_c(\Psi)$ by $\log E(L_c(\Psi) \mid data)$

This only requires $E(Z_i \mid x_i) = P(Z_i = 1 \mid x_i)$

Recall Bayes formula:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

$$E(Z_i \mid x_i) = P(Z_i = 1 \mid x_i) = \frac{\lambda f_1(x_i)}{\lambda f_1(x_i) + (1-\lambda) f_2(x_i)}$$

where $f_1(x_i) = f(x_i; \mu_1, \sigma_1^2)$ and
$f_2(x_i) = f(x_i; \mu_2, \sigma_2^2)$ are the 2 normal pdfs.

Call this $Eqn\ (2)$

How it works:

Begin with a set of initial values:

$$\lambda^0,\ \mu_1^0, \mu_2^0, \sigma_1^0, \text{ and } \sigma_2^0$$

E-step: use $Eqn\ (2)$ to compute: z_i^0

M-step: use z_i^0 and $Eqns\ (1)$ to compute

$$\lambda^1,\ \mu_1^1, \mu_2^1, \sigma_1^1, \text{ and } \sigma_2^1$$

Iterate between the E and M steps until convergence.

$$\widehat{\Psi} = (\hat{\lambda},\ \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$$

Did we **solve the wrong problem**?

We have maximized $L_c(\Psi)$ not the regular likelihood $L(\Psi)$.

Dempster, Laird, and Rubin (J. Royal Statist Soc B, 1977, p1-38) show that the solution to the complete data problem never decreases the regular likelihood.

Hence, we can use the EM algorithm to find the regular maximum likelihood estimates.

Dave Hunter will discuss the pitfalls.

Standard errors of the maximum likelihood estimates are estimated by estimating the information matrix.

The square roots of the diagonal elements of the inverse of the estimate of the **information matrix** divided by square root of n are the estimates of the standard errors.

The **bootstrap** can also be used to approximate the standard errors. This requires iterations (for the EM algorithm) inside the bootstrap iterations and can be computationally expensive.

Example: "We give here two small portions of the spectrum of a bright quasar described in the following study:

HIGH-RESOLUTION STIS/HUBBLE SPACE TELESCOPE AND HIRES/KECK SPECTRA OF THREE WEAK Mg ii ABSORBERS TOWARD PG 1634+7061

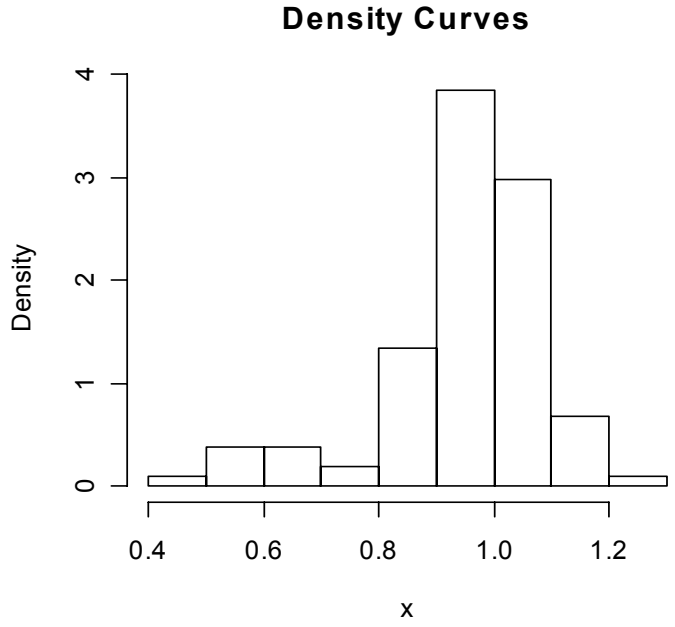
Jane C. Charlton, Jie Ding, Stephanie G. Zonak, Christopher W. Churchill, Nicholas A. Bond, and Jane R. Rigby

We give regions around the 3-times-ionized silicon line Si IV 1394 and the 3-times-ionized carbon line C IV 1551 for the z=0.653411 absorption system..."

http://astrostatistics.psu.edu/datasets/QSO_absorb.html

Data are the normalized intensities of the quasar light.

n = 104



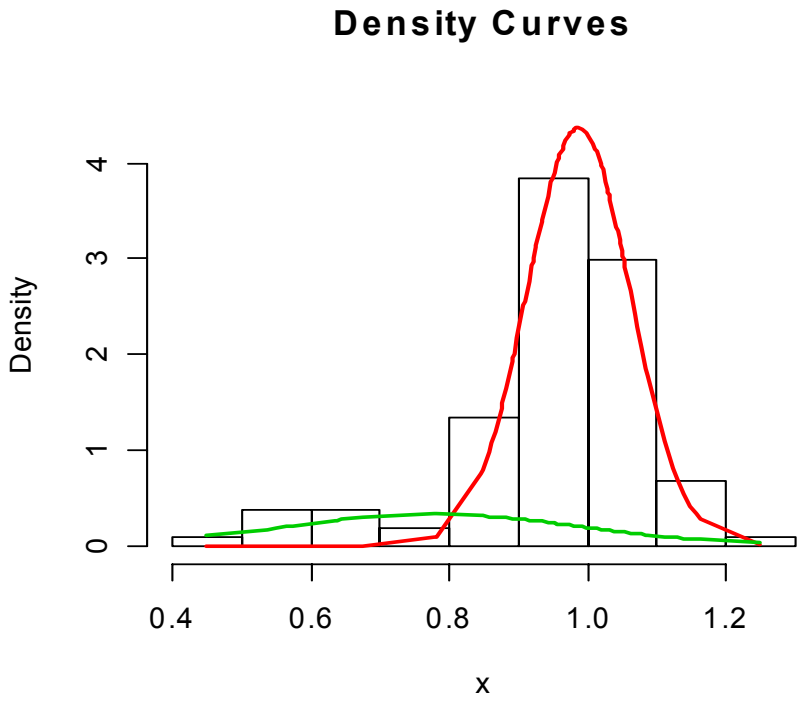
loglik = 57.4898

BIC = 2loglik - k×ln(n) = 105.69

mean = .95, Stdev = .14

We will consider fitting normal mixture models with 2, 3, and 4 components.

First consider a **2 component mixture**:



Mean: 0.99 0.78

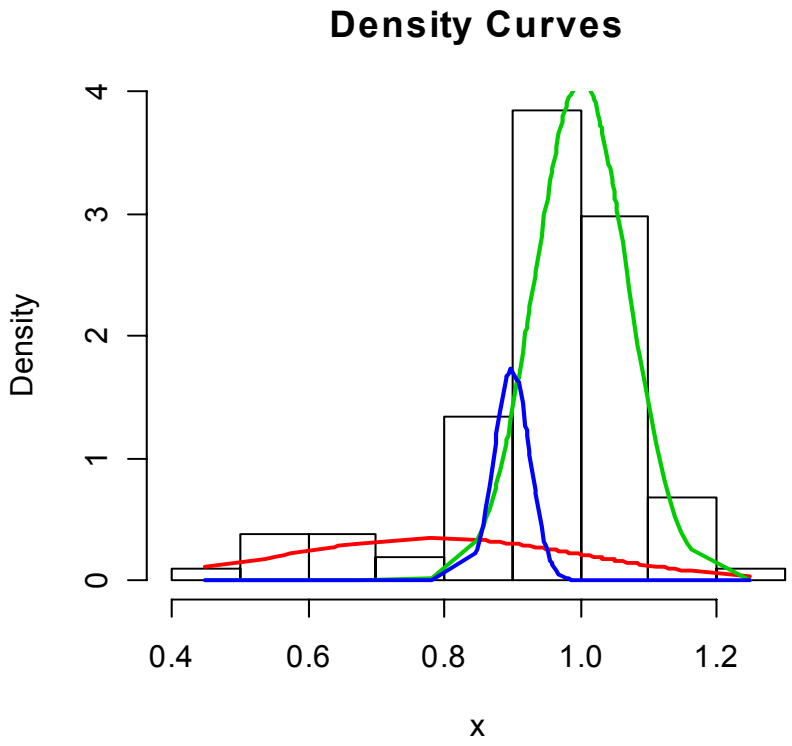
StDev: 0.075 0.22

lambda: 0.82 0.18

loglik = 78.91

BIC = 2loglik - k×ln(n) = 134.6

Consider next a **3 component mixture**:



Mean: 0.786 1.000 0.899

StDev: 0.218 0.069 0.026

lambda: 0.19 0.70 0.11

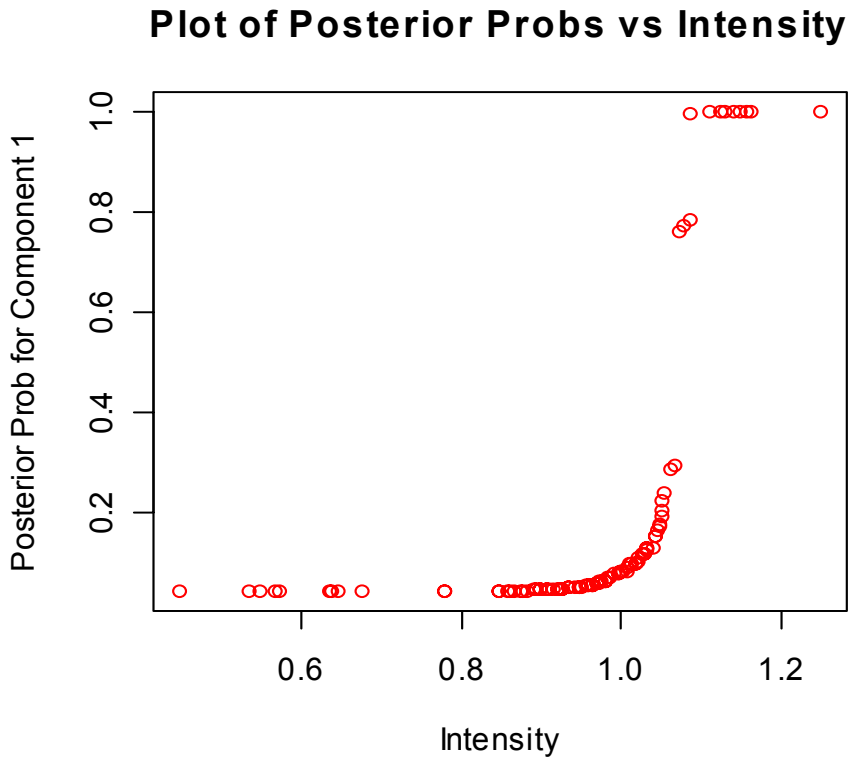
loglik = 80

BIC = 122.84

Recall for the 2 component mixture:

- Mean: 0.99 0.78
- StDev: 0.075 0.22
- lambda: 0.82 0.18

The EM algorithm also provides posterior probabilities that each data point comes from the first component.



Some references:

Finite Mixture Models by McLachlan and Peel

(Excellent book published in 2000 and covers a wide variety of topics in parametric mixture models with a lot of examples.)

Mixtools, a library of R functions developed by Derek Young.

(Excellent set of functions that covers a wide variety of applications including univariate parametric models, repeated measures, and regression. Model fitting and graphics.)