Last updated: November 6, 2012

MIXTURE MODELS AND EM

J. Elder

CSE 4404/5327 Introduction to Machine Learning and Pattern Recognition

Credits

- Some of these slides were sourced and/or modified from:
  - Christopher Bishop, Microsoft UK
  - Simon Prince, University College London
  - Sergios Theodoridis, University of Athens & Konstantinos Koutroumbas, National Observatory of Athens

Mixture Models and EM: Topics

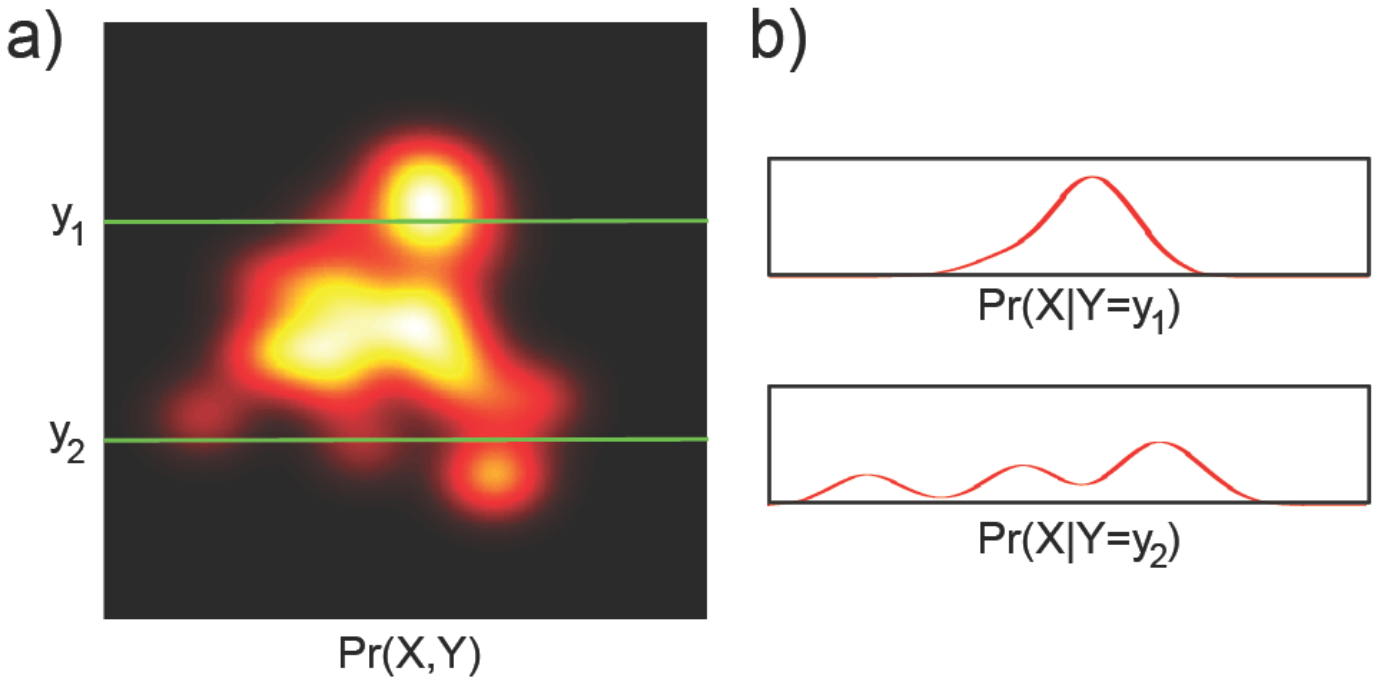
1. Intuition
2. Equations
3. Examples
4. Applications

Mixture Models and EM: Topics

1. **Intuition**
2. Equations
3. Examples
4. Applications

Motivation

- What do we do if a distribution is not well-approximated by a standard parametric model?



Mixtures of Gaussians

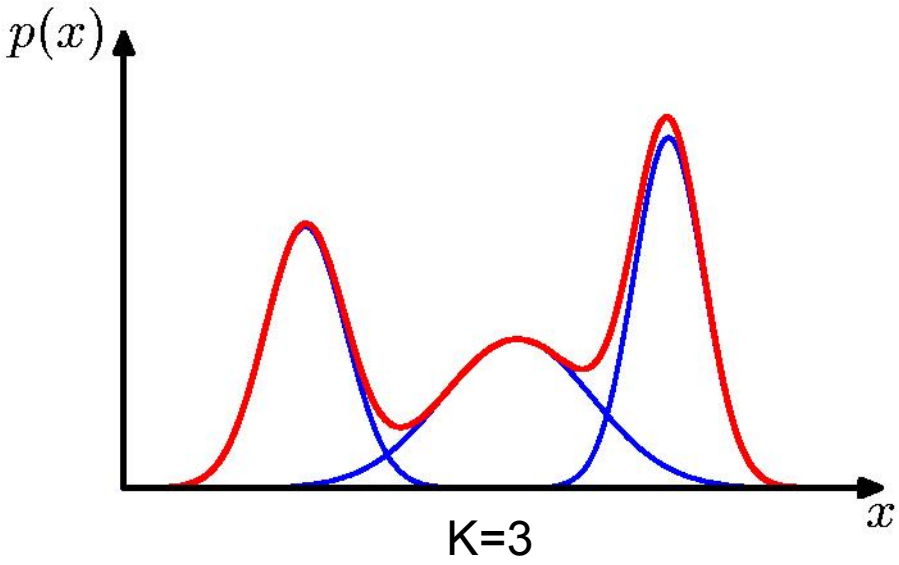
- Combine simple models into a complex model:

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

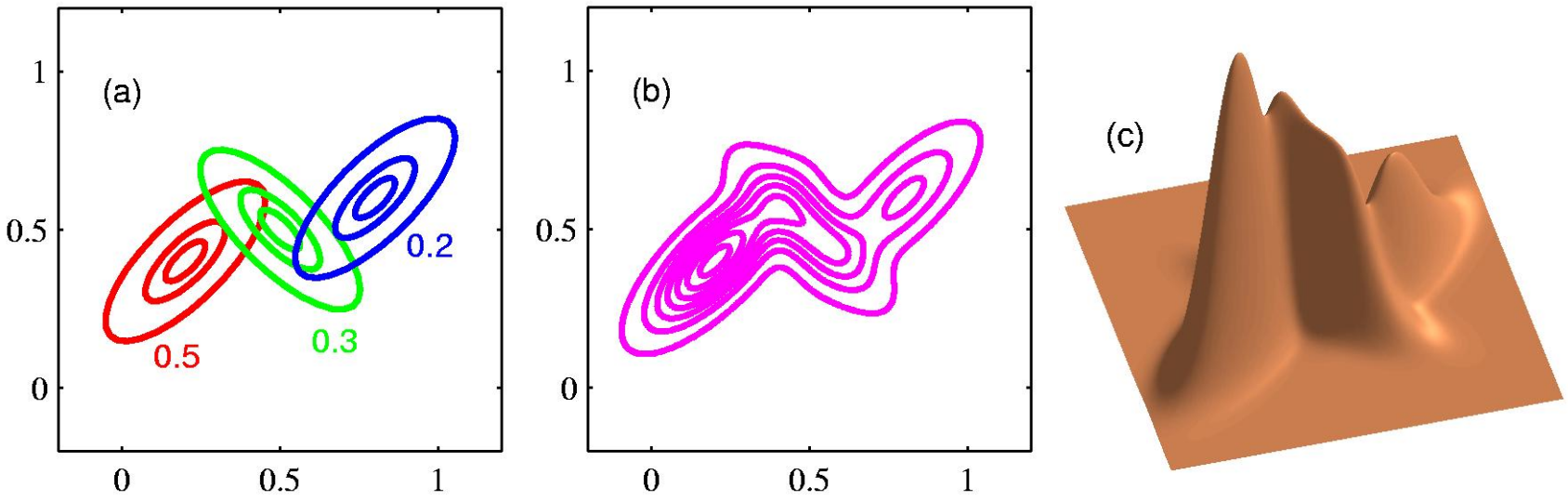
Component

Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



Mixtures of Gaussians



Mixtures of Gaussians

- Determining parameters μ, σ and π using maximum log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Log of a sum; no closed form maximum.

- Solution: use standard, iterative, numeric optimization methods or the ***expectation maximization*** algorithm.

Hidden Variable Interpretation

$$p(\mathbf{x_n}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x_n}; \mu_k, \Sigma_k)$$

- Assumptions
  - for each training observation $\mathbf{x_n}$ there is a hidden variable z_n.
  - $z_n = 1,\ldots,K$ represents which Gaussian $\mathbf{x_n}$ came from
- With this interpretation, we have:

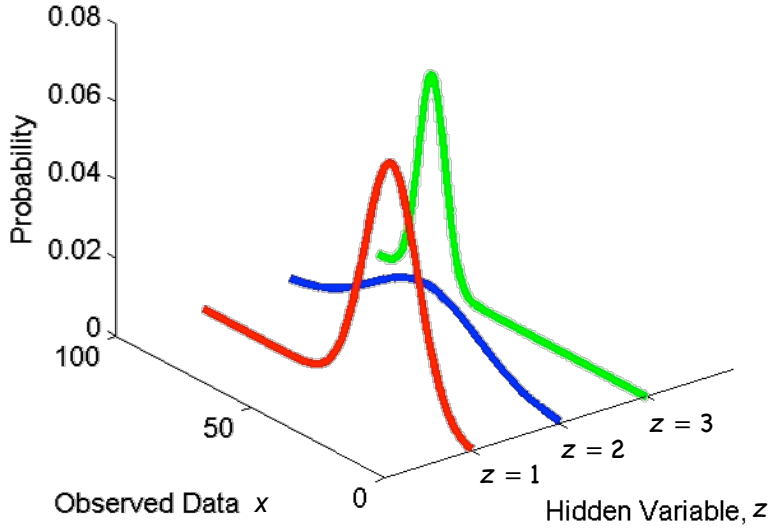
$$p(\mathbf{x_n}) = \sum_{k=1}^{K} p(\mathbf{x_n} \mid z_n = k)\, p(z_n = k)$$

where

$$p(z_n = k) = \pi_k$$

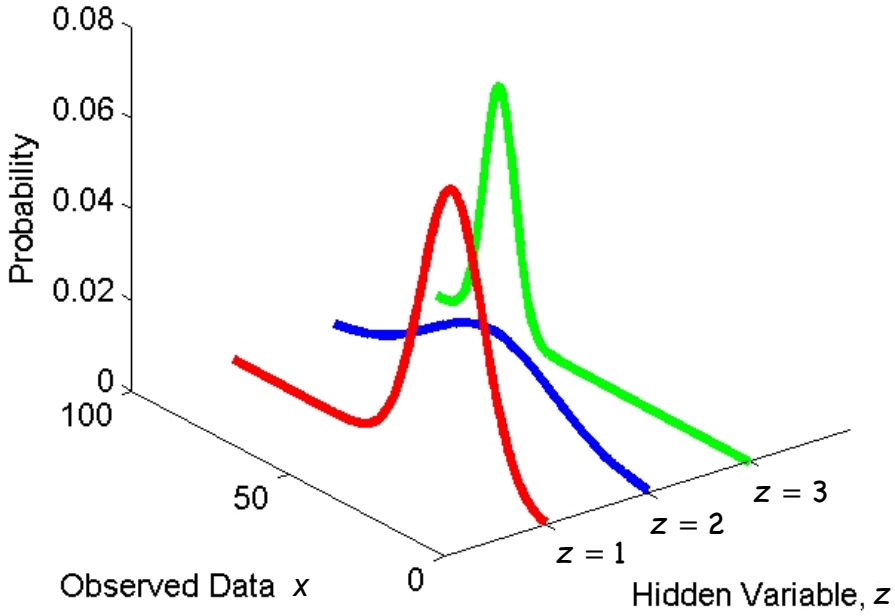
and

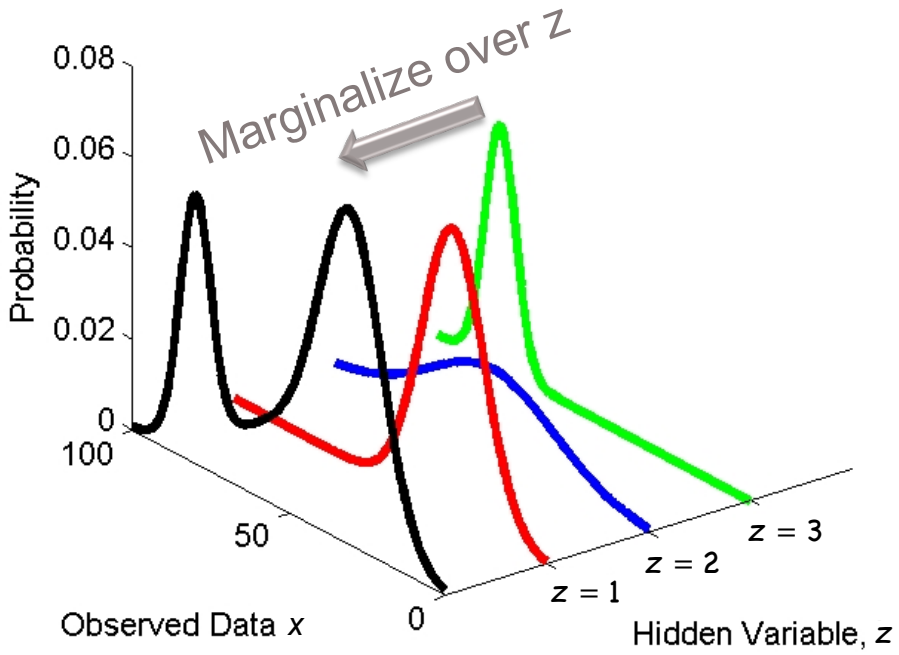
$$p(\mathbf{x_n} \mid z_n = k) \sim \mathcal{N}(\mathbf{x_n}; \mu_k, \Sigma_k)$$



Hidden Variable Interpretation

$$p\left(\mathbf{x}_n \mid \pi_1,\ldots\pi_K,\mu_1,\ldots\mu_K,\Sigma_1,\ldots\Sigma_K\right)=\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\mathbf{x}_n;\mu_k,\Sigma_k\right)=\sum_{k=1}^{K}p\left(z_n=k\right)p\left(\mathbf{x}_n \mid z_n=k\right)$$





Hidden Variable Interpretation

$$p\left(\mathbf{x}_n \mid \pi_1,\ldots\pi_K,\mu_1,\ldots\mu_K,\Sigma_1,\ldots\Sigma_K\right)=\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\mathbf{x}_n;\mu_k,\Sigma_k\right)=\sum_{k=1}^{K}p\left(z_n=k\right)p\left(\mathbf{x}_n \mid z_n=k\right)$$

OUR GOAL

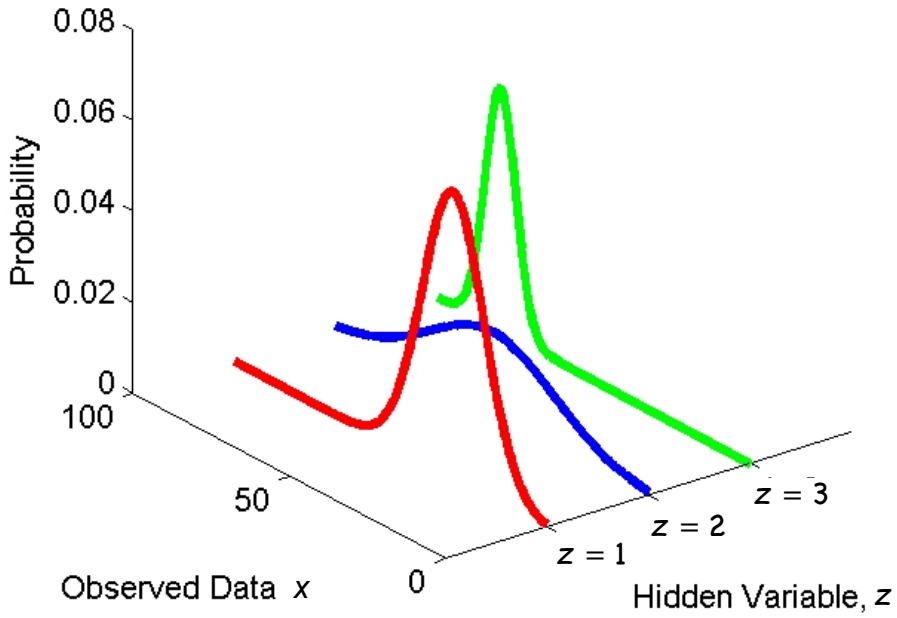
To estimate the parameters θ:

The means μ_k

The covariances Σ_k

The weights (mixing coefficients) π_k

for all K components of the model.



THING TO NOTICE

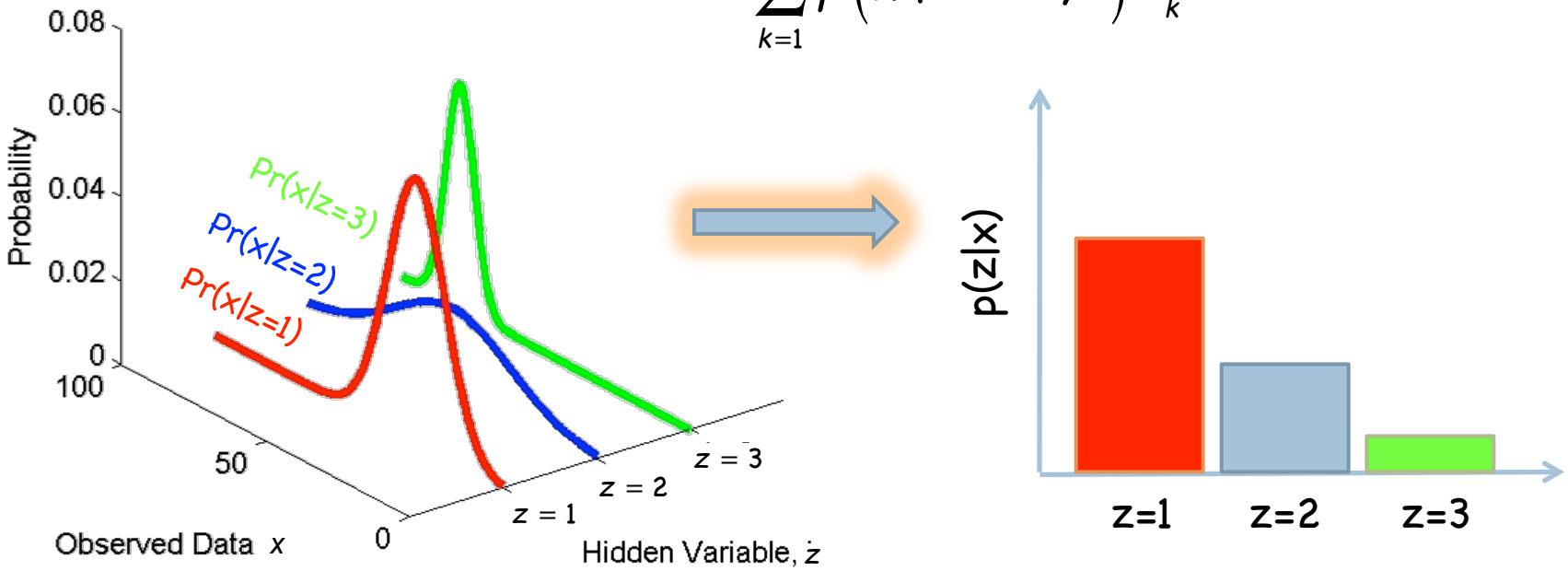
If we knew the hidden variables z_n for the training data it would be easy to estimate parameters θ – just estimate individual Gaussians separately.

Hidden Variable Interpretation

THING TO NOTICE #2:

If we knew the parameters θ it would very easy to estimate the posterior distribution over each hidden variable z_n using Bayes' rule:

$$p(z = k \mid \mathbf{x}, \theta) = \frac{p(\mathbf{x} \mid z = k, \theta)\pi_k}{\sum_{k=1}^{K} p(\mathbf{x} \mid z = k, \theta)\pi_k}$$



Expectation Maximization

Chicken and egg problem:

- could find $z_{1...N}$ if we knew θ
- could find θ if we knew $z_{1...N}$

Solution: Expectation Maximization (EM) algorithm

(Dempster, Laird and Rubin 1977)

EM for Gaussian mixtures can be viewed as alternation between 2 steps:

1. Expectation Step (E-Step)
   - For fixed θ find posterior distribution over responsibilities $z_{1...N}$
2. Maximization Step (M-Step)
   - Now use these posterior probabilities to re-estimate θ

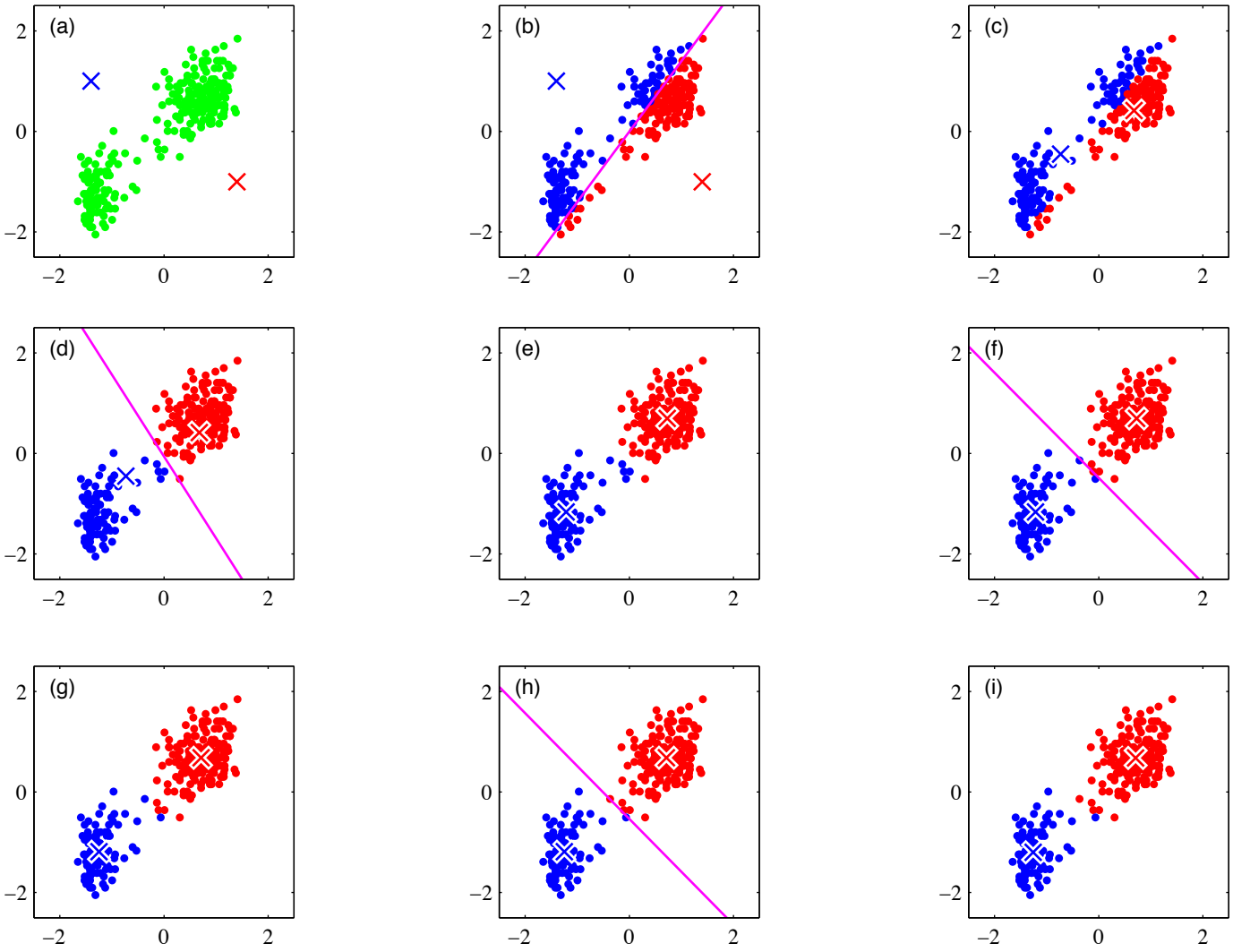
K-Means: A Poor-Man's Approximation

- The K-means algorithm is very similar to EM, except that
  - We assume the mixing coefficients are the same for each component.
  - We assume each component is isotropic with common variance.
  - We do not admit any uncertainty about the responsibilities at each iteration.

K-Means: A Poor-Man's Approximation

- K-Means consists of the following steps:
  1. Randomly select the mean for each component.
  2. Associate each input with the nearest component mean.
  3. Recompute the means based upon the new association.
- Steps 2 and 3 are alternated until convergence.

Example



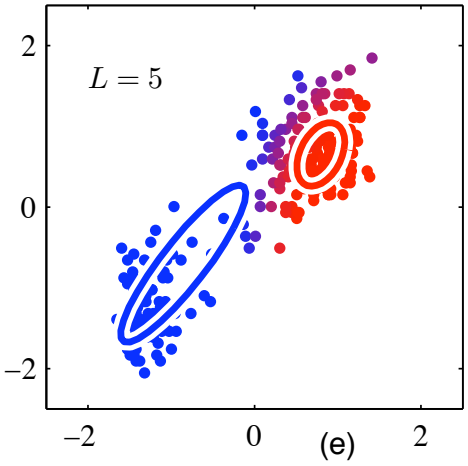
Mixture Models and EM: Topics

1. Intuition
2. **Equations**
3. Examples
4. Applications

Responsibilities

- The ***responsibility*** r_{nk} of component k for observation $\mathbf{x}_n$ is the posterior probability that component k generated $\mathbf{x}_n$.

$$\text{Responsibility } r_{nk} \triangleq P\left(z_n = k \mid \mathbf{x}_n; \theta\right)$$



- Responsibilities Update Equation:

$$\text{Responsibility } r_{nk} = \frac{p\left(\mathbf{x}_n \mid z_n = k; \theta\right)\pi_k(t)}{\sum_{k=1}^{K} p\left(\mathbf{x}_n \mid z_n = k; \theta\right)\pi_k(t)}$$

- Let N_k = effective number of observations explained by component k:

$$N_k \triangleq \sum_{n=1}^{N} r_{nk}$$

Example: Mixture of Gaussians

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- Differentiating does not allow us to isolate the parameters analytically.
- However, it does generate equations that can be used for iterative estimation of these parameters by EM.

Example: Mixture of Gaussians

- Parameter Update Equations:

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\left(\mathbf{x}_n - \mu_k\right)\left(\mathbf{x}_n - \mu_k\right)^t$$

$$\pi_k = \frac{1}{N}\sum_{k=1}^{N} r_{nk}$$

Mixture of (Multivariate) Gaussians

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- Here we will derive the formula for the mixing coefficents π_k.
- **Textbook Problem 11.2: Derive the EM formulae for the mean μ_k and covariance Σ_k.**
  - **Please do this at home!**
  - **When solving for Σ_k, recall that:**

$$\frac{d}{dA}|A| = |A|A^{-t} \quad \textbf{and} \quad \frac{d}{dA}x^tAx = xx^t$$

Initialization

- EM typically takes longer to converge than K-Means, and the computations are more elaborate.
- It is thus common to use the solution to K-means to initialize the parameters for EM.

Convergence

- EM is guaranteed to monotonically increase the likelihood.
- However, since in general the likelihood is non-convex, we are not guaranteed to find the globally optimal parameters.

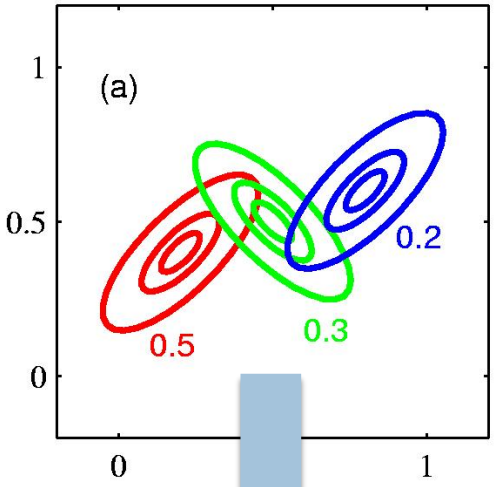
End of Lecture

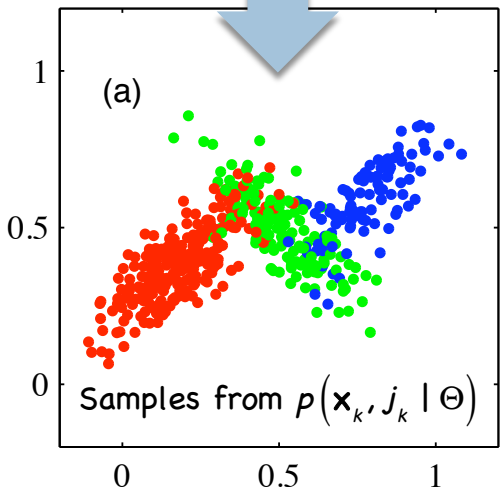
October 24, 2012

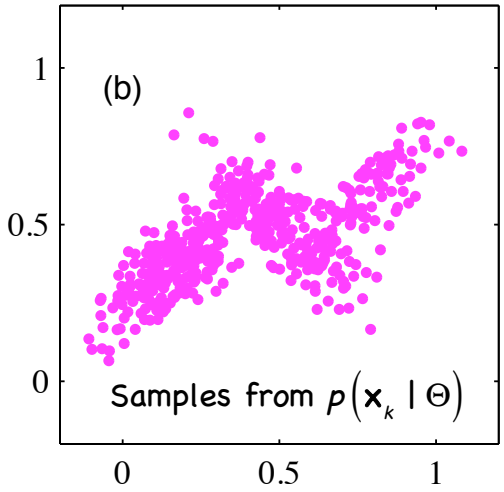
Mixture Models and EM: Topics

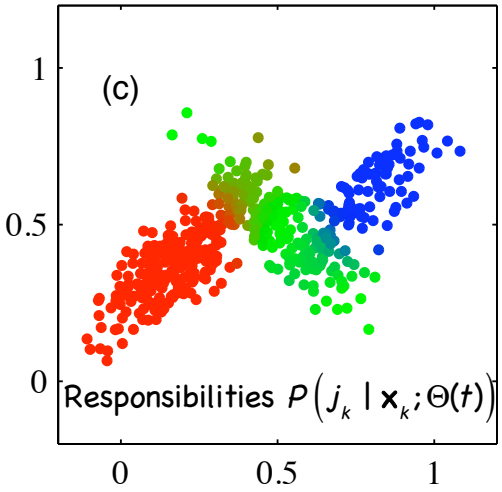
1. Intuition
2. Equations
3. **Examples**
4. Applications

Bivariate Gaussian Mixture Example

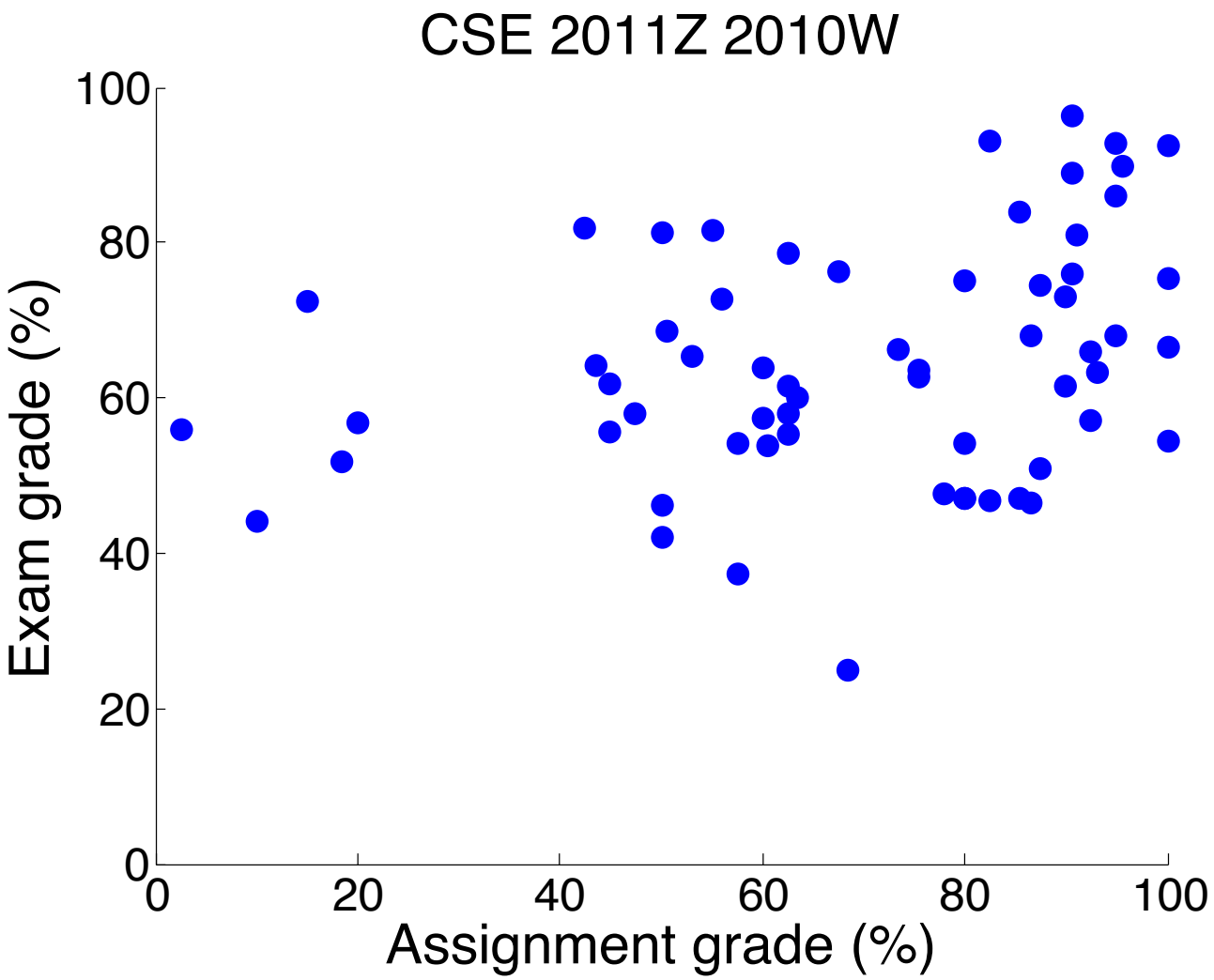






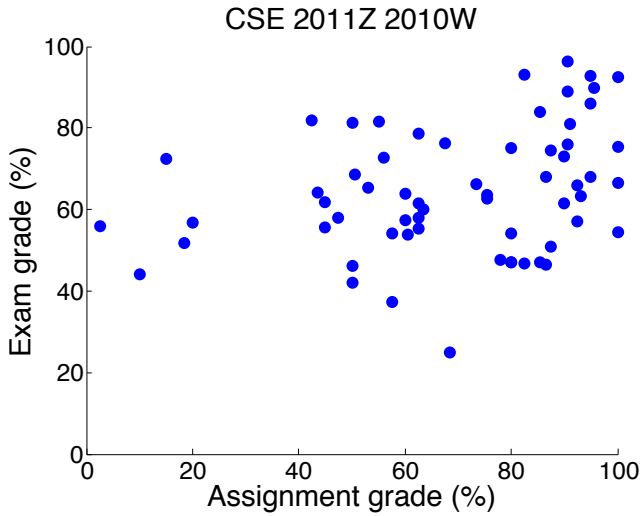


2-Component Bivariate MATLAB Example

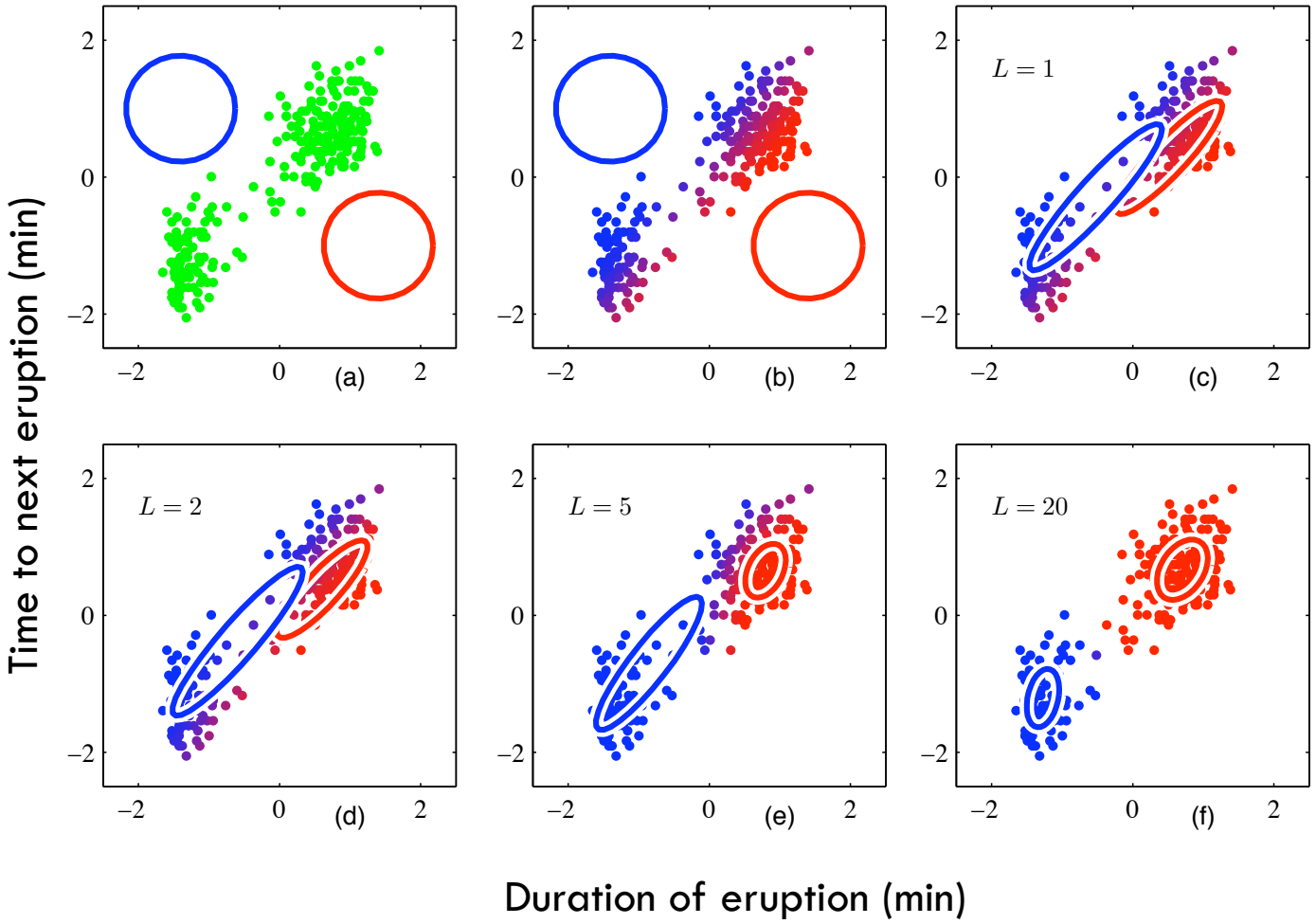


2-Component Bivariate MATLAB Example

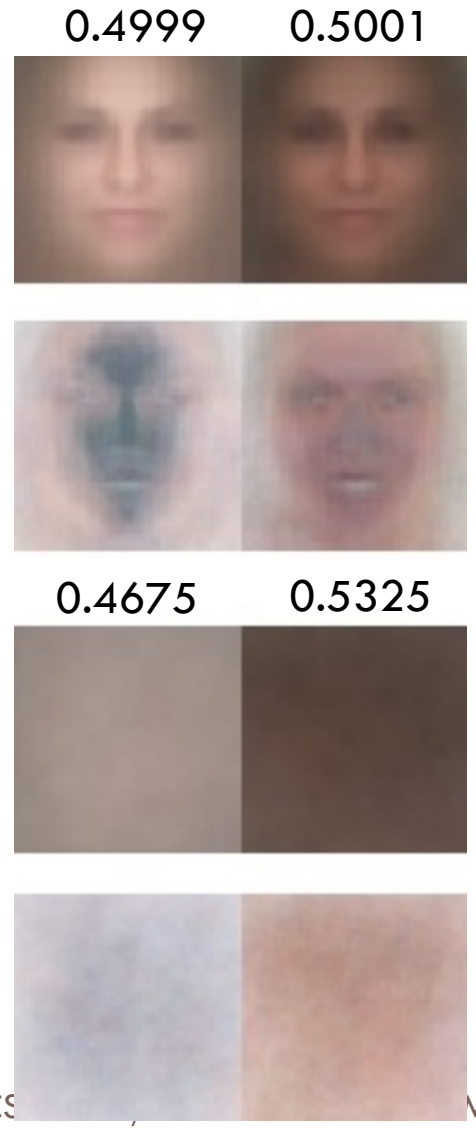

```
%update responsibilities
    for i=1:k
        p(:,i)=alphas(i).*mvnpdf(x,mu(i,:),squeeze(S(i,:,:)));
    end
    p=p./repmat((sum(p,2)),1,k);


%update parameters
    for i=1:k
        Nk=sum(p(:,i));
        mu(i,:)=p(:,i)'*x/Nk;
        dev=x-repmat(mu(i,:),N,1);
        S(i,:,:)=(repmat(p(:,i),1,D).*dev)'*dev/Nk;
        alphas(i)=Nk/N;
    end
```



Old Faithful Example



Face Detection Example: 2 Components



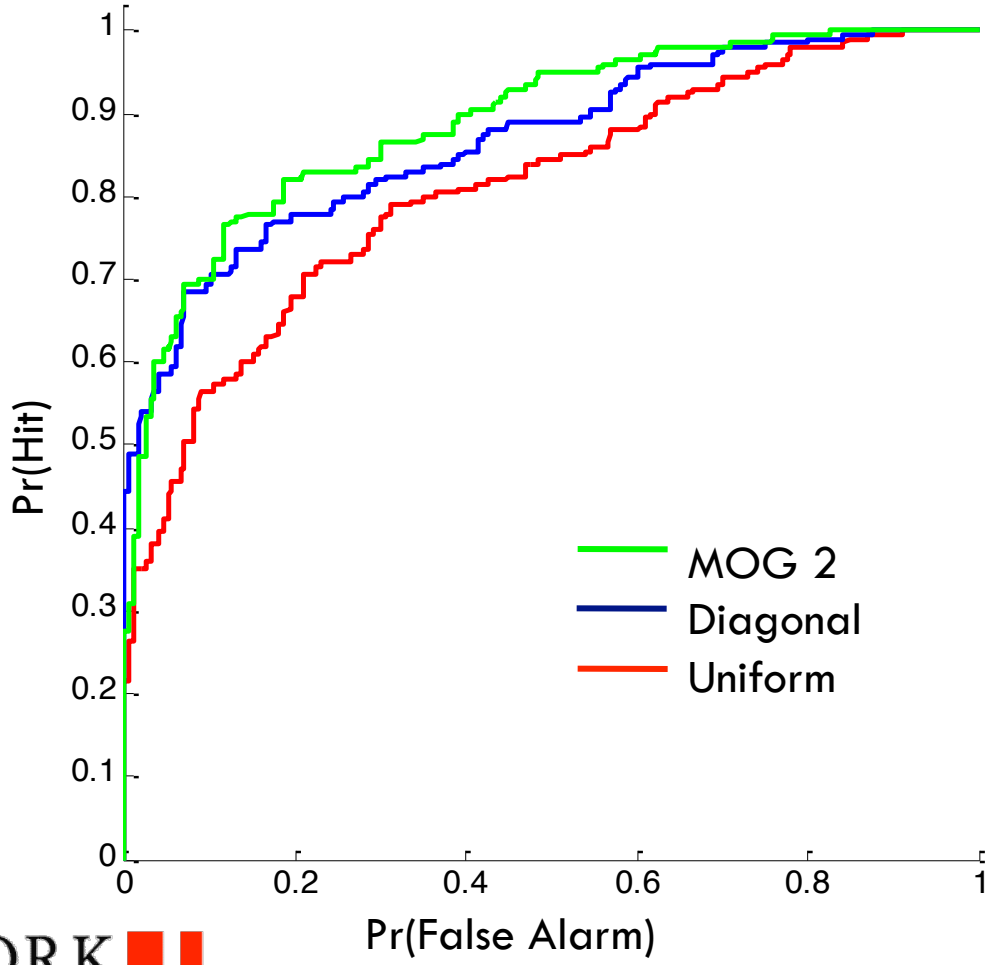
Each component is still assumed to have diagonal covariance.

The face model and non-face model have divided the data into two clusters. In each case, these clusters have roughly equal weights.

The primary thing that these seem to have captured is the photometric (luminance) variation.

Note that the standard deviations have become smaller than for the single Gaussian model as any given data point is likely to be close to one mean or the other.

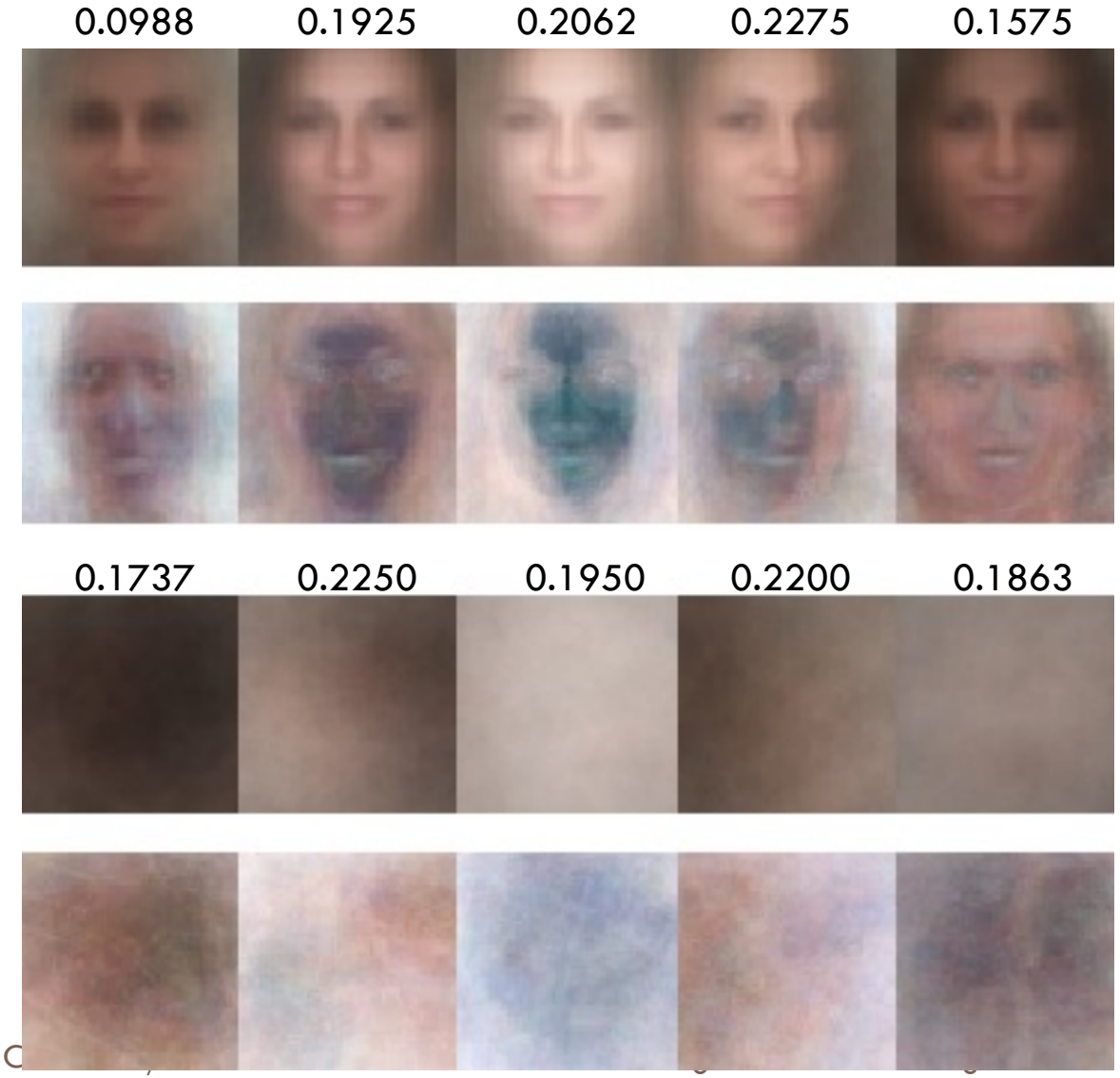
Results for MOG 2 Model



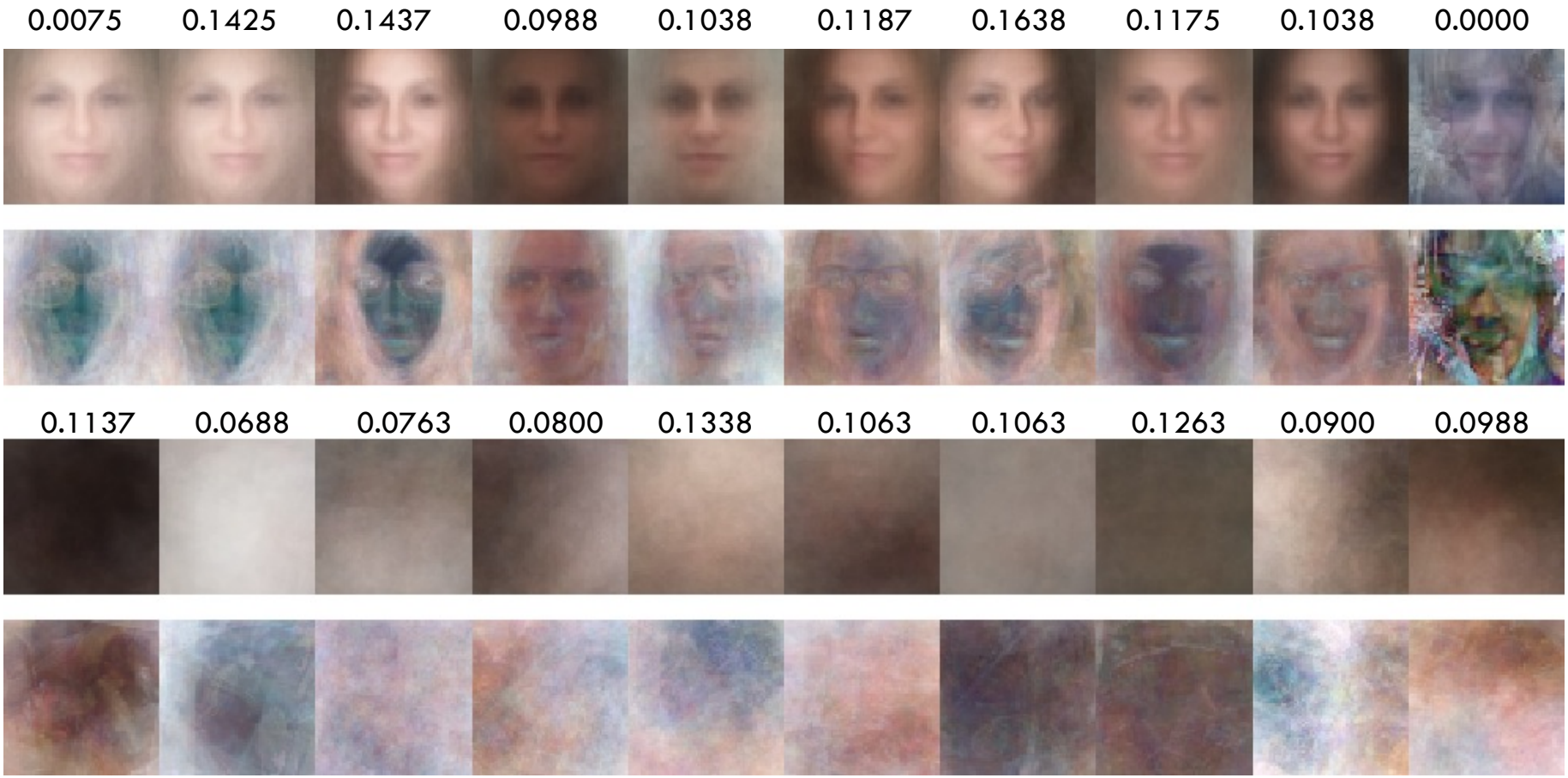
Performance improves relative to a single Gaussian model, although it is not a dramatic improvement.

We have a better description of the data likelihood.

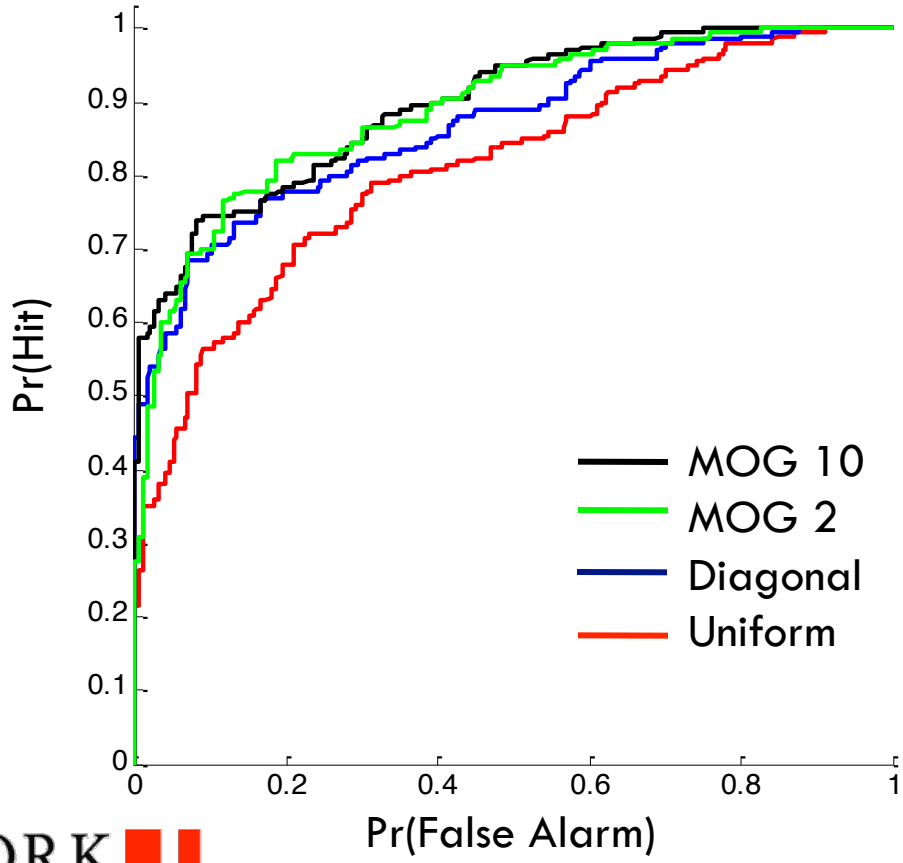
MOG 5 Components



MOG 10 Components



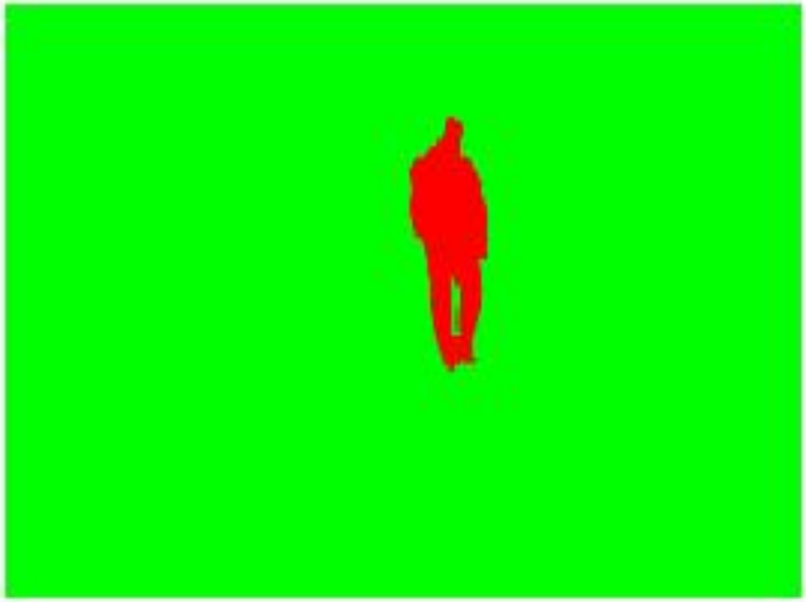
Results for Mog 10 Model



Performance improves slightly more, particularly at low false alarm rates.

Background Subtraction

Test Image



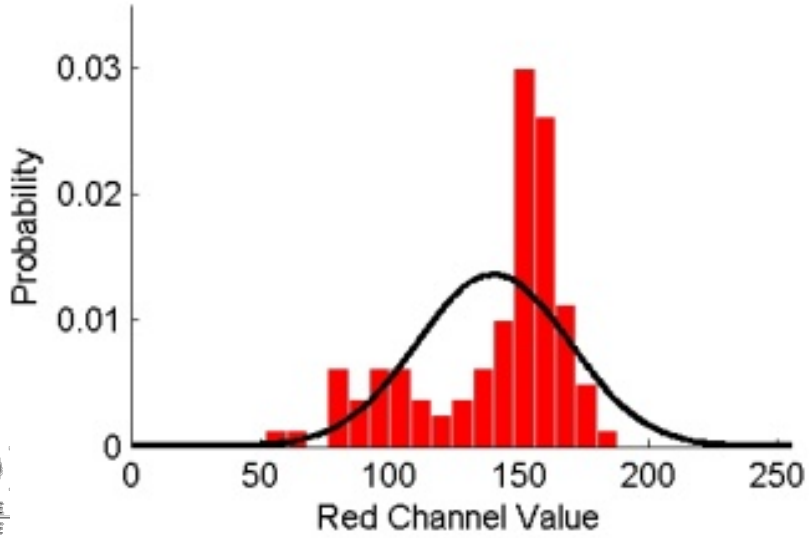
Desired Segmentation

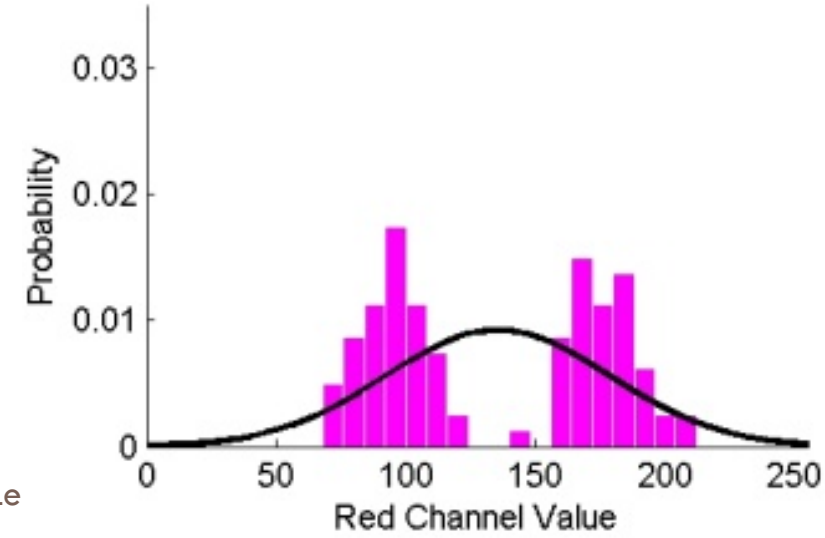
GOAL : (i) Learn background model (ii) use this to segment regions where the background has been occluded

What if the scene isn't static?

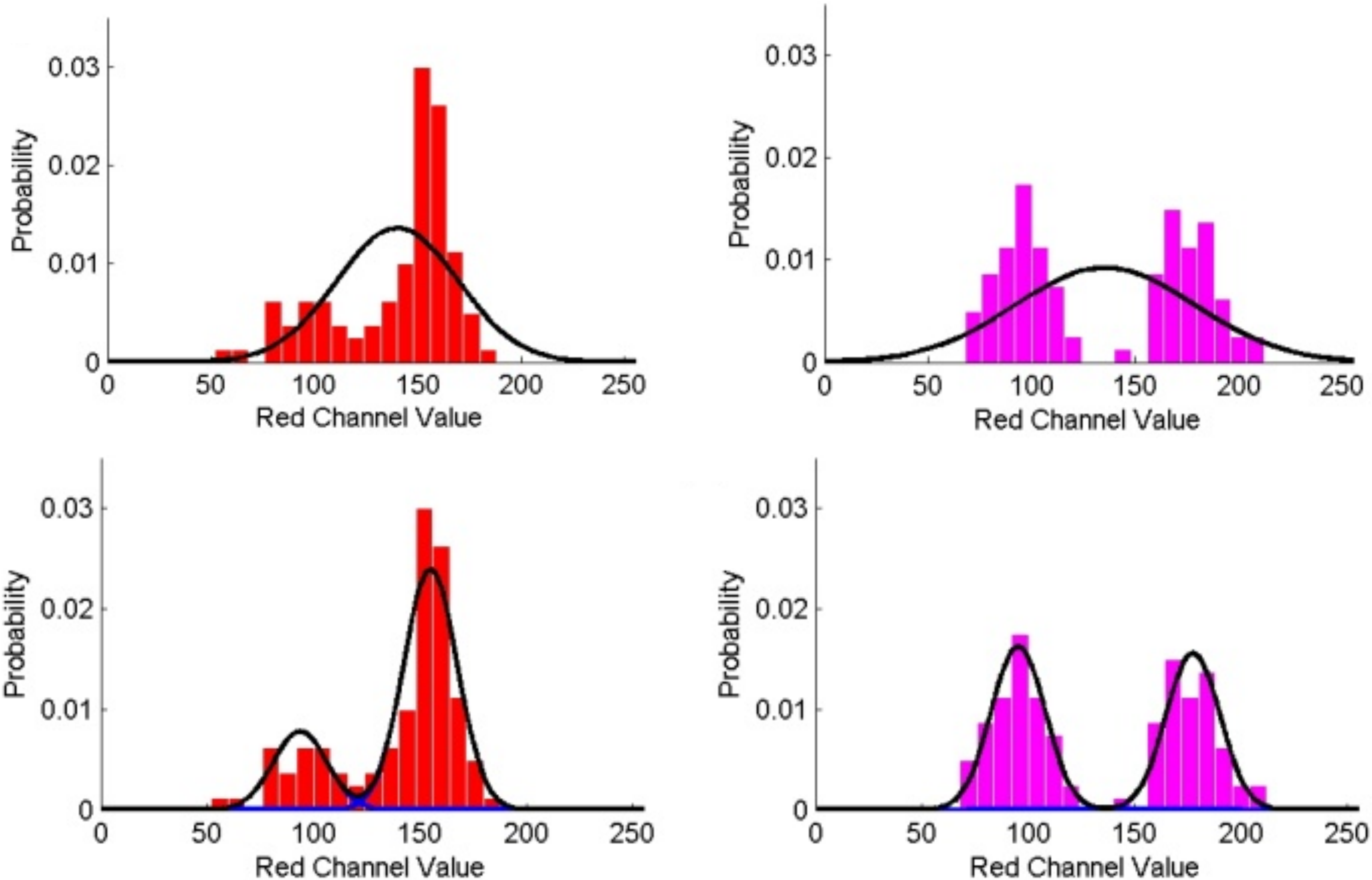
Gaussian is no longer a good fit to the data.

Not obvious exactly what probability model would fit better.





Background Mixture Model



Background Subtraction Example