EXPECTATION MAXIMIZATION

J. Elder

CSE 6390/PSYC 6225 Computational Modeling of Visual Perception

Credits

- Some of these slides were sourced and/or modified from:
  - Christopher Bishop, Microsoft UK
  - Simon Prince, University College London

Mixtures of Gaussians

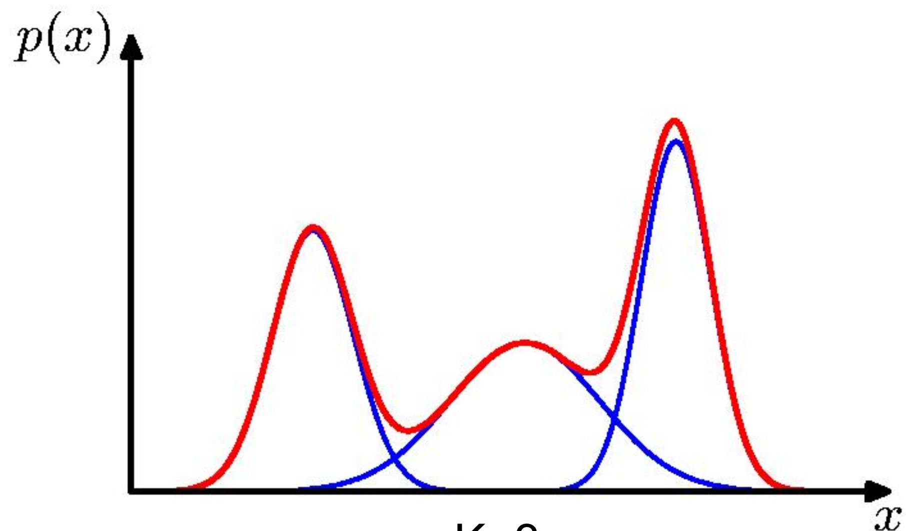
- Combine simple models into a complex model:

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

Component

Mixing coefficient

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



K=3

Mixtures of Gaussians

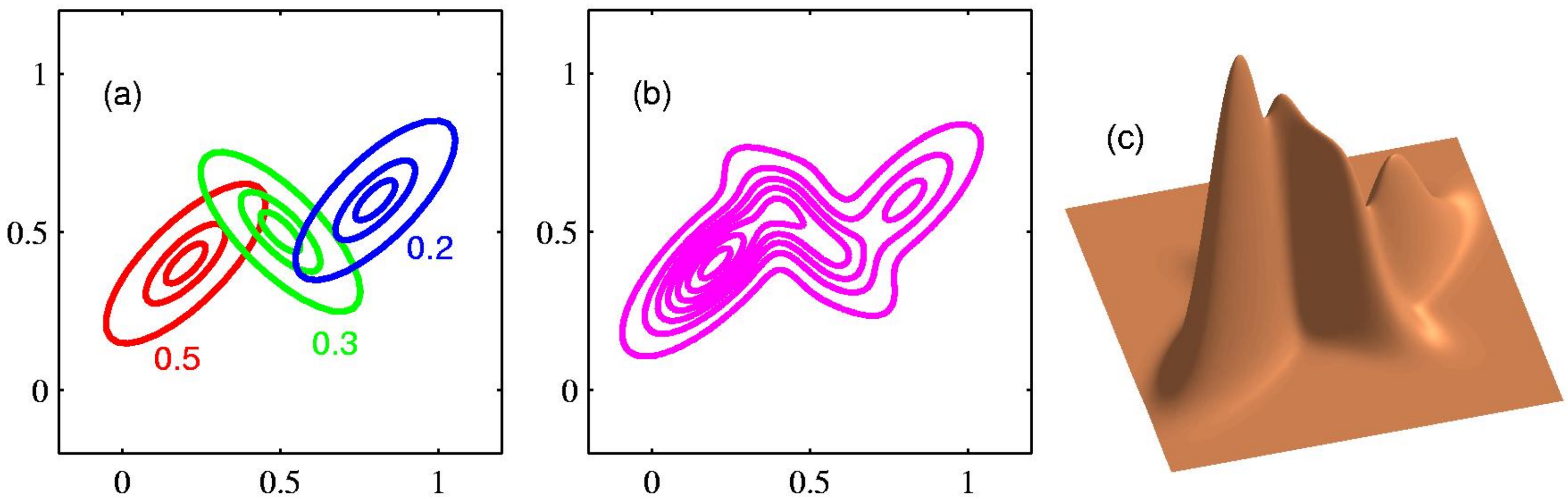
- Determining parameters μ, σ and π using maximum log likelihood

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

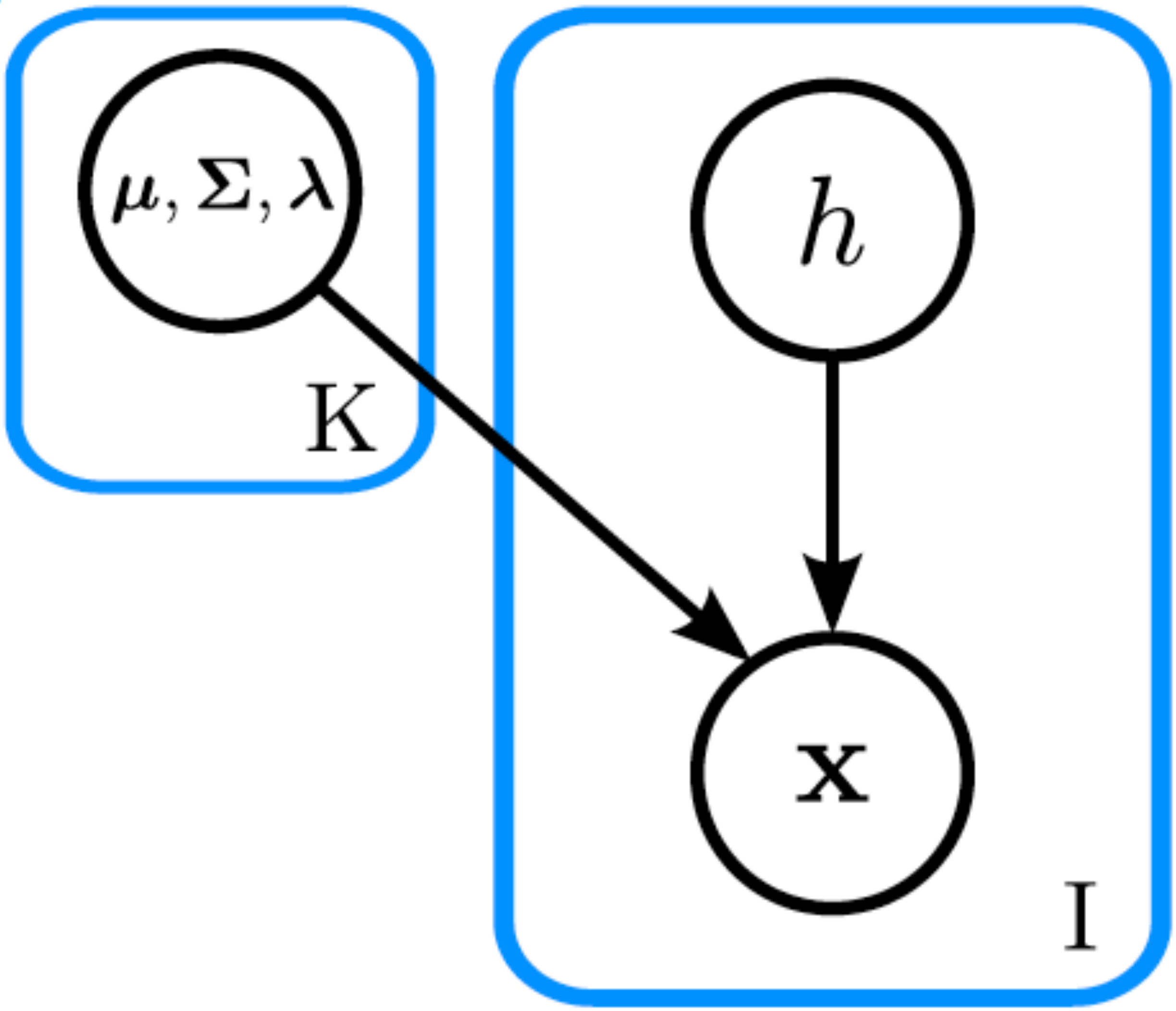
Log of a sum; no closed form maximum.

- Solution: use standard, iterative, numeric optimization methods or the *expectation maximization* algorithm (Chapter 9).

Mixtures of Gaussians



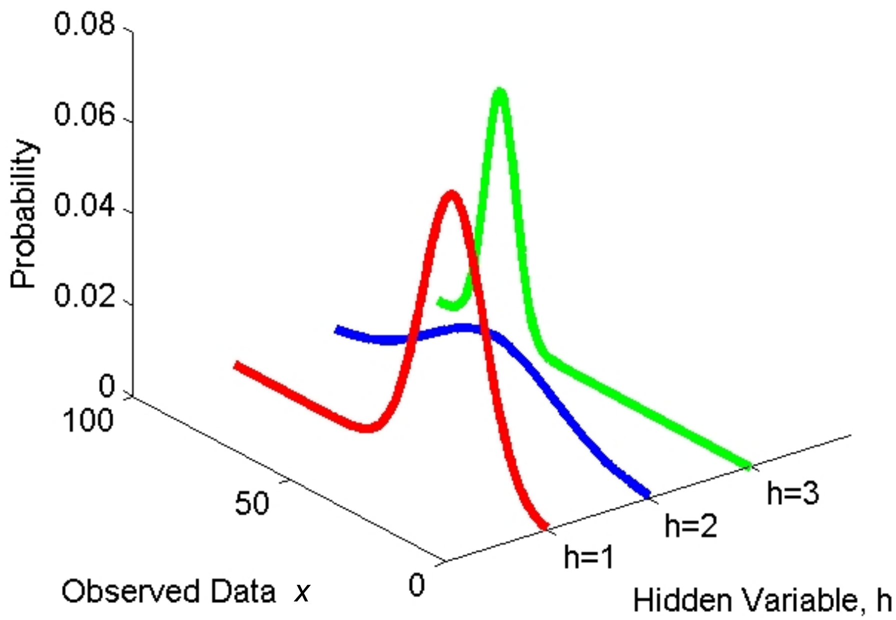
Graphical Model for Gaussian Mixture

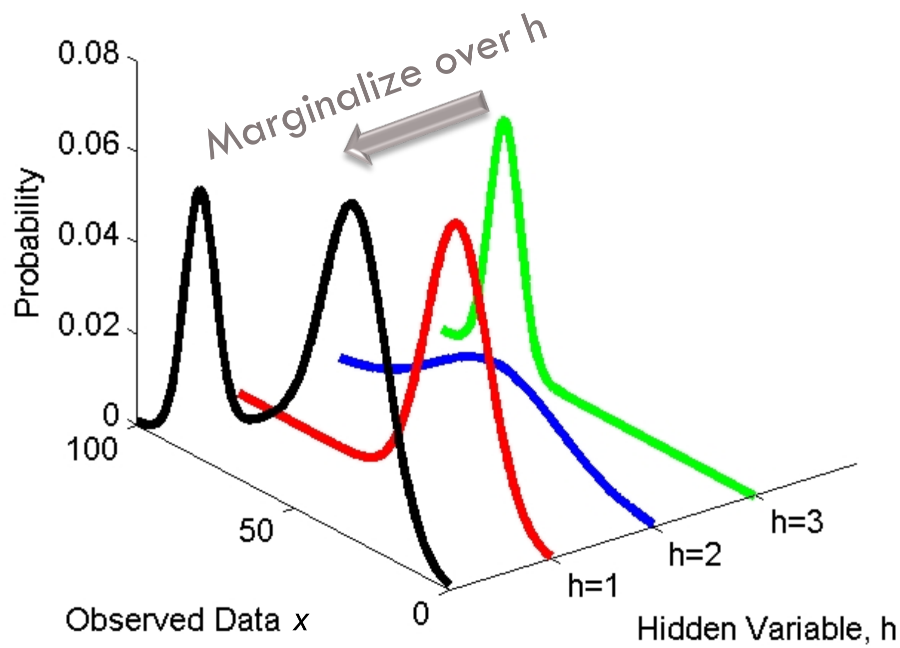


Hidden Variable Interpretation

$$Pr(x|w_{1\ldots K}, \mu_{1\ldots K}\sigma^2_{1\ldots K}) = \sum_{k=1}^{K} w_k \mathcal{G}_x\left[\mu_k, \sigma^2_k\right]$$

$$= \sum_{k=1}^{K} Pr(h=k)Pr(x|h=k)$$

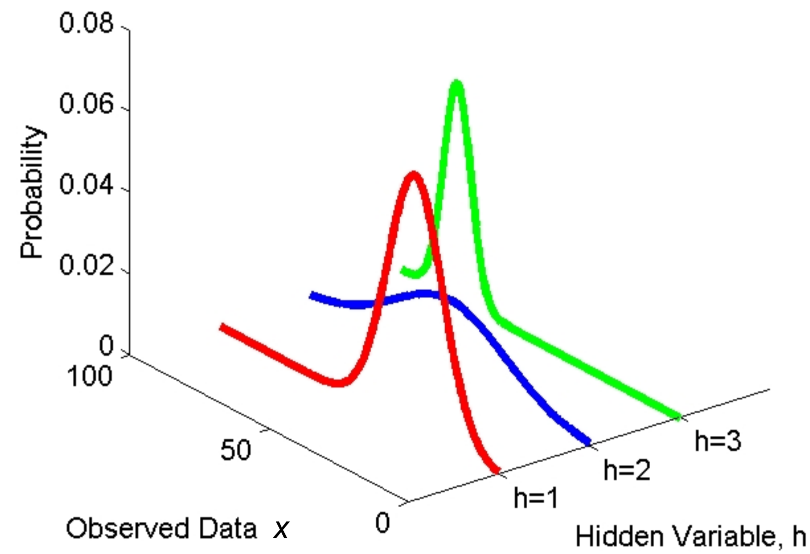




Hidden Variable Interpretation

ASSUMPTIONS

- for each training datum x_i there is a hidden variable h_i.
- h_i represents which Gaussian x_i came from
- hence h_i takes discrete values



OUR GOAL:

To estimate the parameters θ:

The means μ_k

The covariances Σ_k

The weights (mixing coefficients) w_k

for all K components of the model.

THING TO NOTICE:

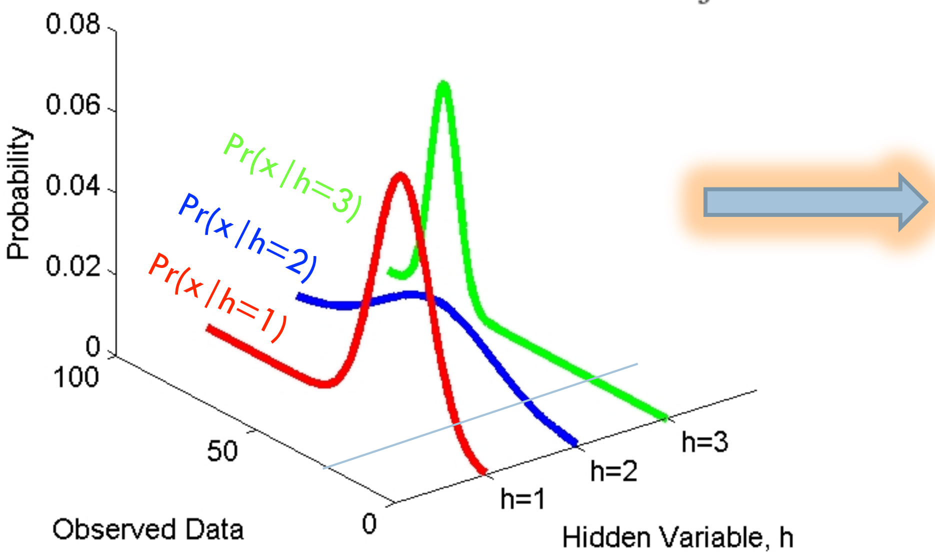
If we knew the hidden variables h_i for the training data it would very easy to estimate parameters θ – just estimate individual Gaussians separately.

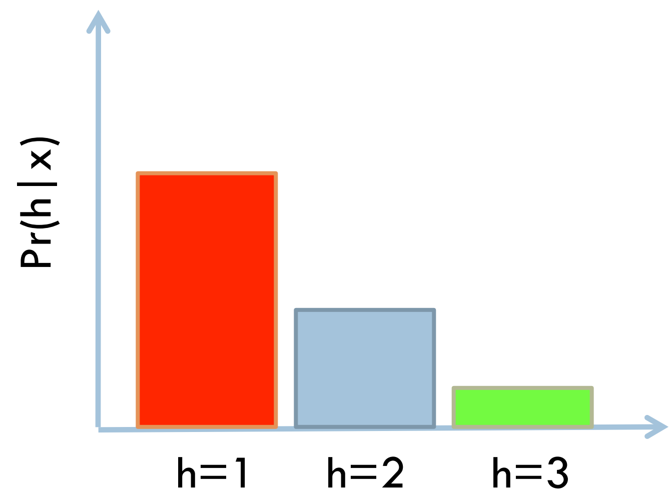
Hidden Variable Interpretation

THING TO NOTICE #2:

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

$$Pr(h_i = k|x_i, \theta) = \frac{Pr(x_i|h_i = k, \theta)Pr(h_i = k)}{\sum_{j=1}^{K} Pr(x_i|h_i = j, \theta)Pr(h_i = j)}$$





Expectation Maximization

Chicken and egg problem:

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

Solution: Expectation Maximization (EM) algorithm

(Dempster, Laird and Rubin 1977)

Alternate between:

1. Expectation Step (E-Step)
   - For fixed θ find posterior distribution over $h_{1...N}$
2. Maximization Step (M-Step)
   - Given these distributions, maximize lower bound on likelihood w.r.t. θ

Expectation Maximization

$$Pr(\mathbf{x}|\boldsymbol{\theta}) \qquad = \int Pr(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta}) d\mathbf{h} = \int Pr(\mathbf{x}|\mathbf{h}, \boldsymbol{\theta}) Pr(\mathbf{h}) d\mathbf{h}$$

We introduce a probability distribution q(**h**) over the hidden variables **h**.

EM works by defining a lower bound $B\left[q(\mathbf{h}),\theta\right]$ on the log likelihood $\log P(\mathbf{x} \mid \theta)$.

and then iteratively increasing this lower bound by alternately updating

$q(\mathbf{h})$ (E-step)

and

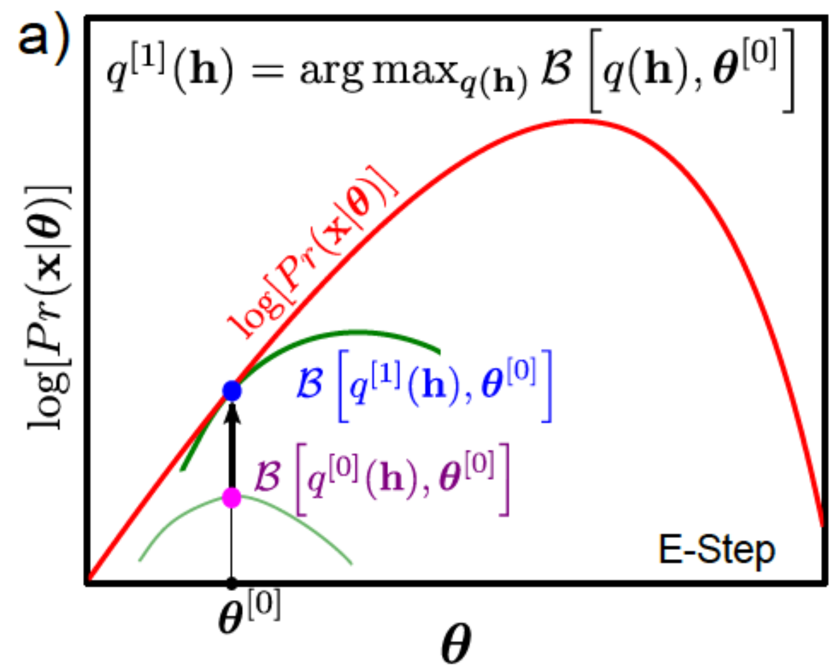
θ (M-step)

E-Step

Fix θ.

Find $q(\mathbf{h})$ that maximizes lower bound:

$$q_i^{[t]}[\mathbf{h}] = \underset{q_i[\mathbf{h}]}{\arg\max}\, \mathcal{B}[q_i[\mathbf{h}], \theta^{[t-1]}].$$

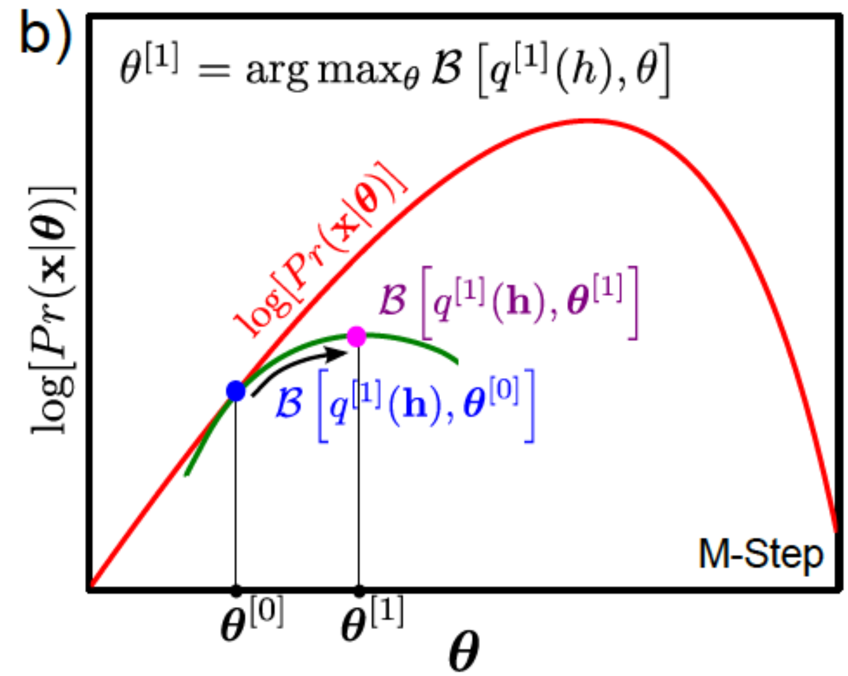


M-Step

Fix $q(\mathbf{h})$.

Find θ that maximizes lower bound:

$$\boldsymbol{\theta}^{[t]} = \underset{\boldsymbol{\theta}}{\arg\max}\, \mathcal{B}[q^{[t]}(h), \boldsymbol{\theta}].$$



Lower Bound on Likelihood

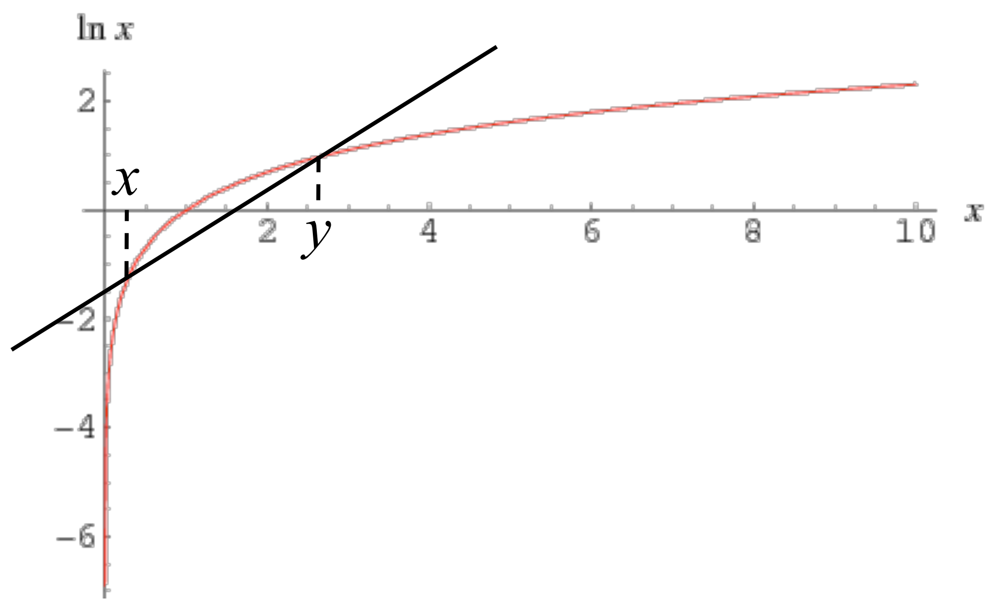
$$\mathcal{B}[q_i(\mathbf{h}_i), \boldsymbol{\theta}] = \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i | \boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_{1 \ldots I}$$

Note that log is a concave function, i.e.,

$$\log(tx + (1-t)y) \geq tf(x) + (1-t)f(y) \quad \forall x, y > 0,\ 0 \leq t \leq 1$$

or equivalently

$$\frac{\partial^2}{\partial x^2} \log x \leq 0$$



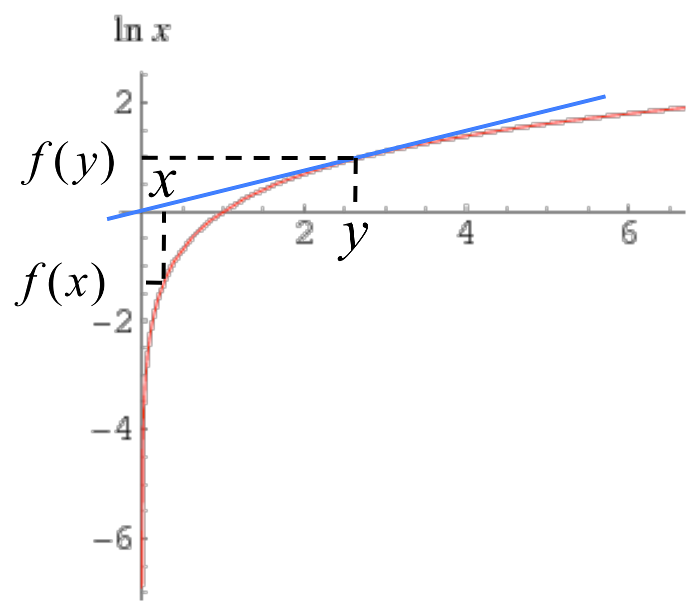
Jensen's Inequality

If $f(x)$ is a concave function, then $E[f(x)] \leq f\left(E[x]\right)$

Proof: Note that $f(y) - f(x) \geq (y - x) f'(y) \quad \forall y > 0$

Choose $y = E[x]$. Then

$f\left(E[x]\right) - f(x) \geq (E[x] - x) f\left(E[x]\right)$



Now, taking expectations of both sides,

$f\left(E[x]\right) - E\left[f(x)\right] \geq 0$

$\rightarrow E\left[f(x)\right] \leq f\left(E[x]\right)$

Lower Bound on Likelihood

Thus

$$\mathcal{B}[q_i(\mathbf{h}_i), \boldsymbol{\theta}] = \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log\left[\frac{Pr(\mathbf{x}, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right] d\mathbf{h}_{1\ldots I}$$

$$\leq \sum_{i=1}^{I} \log\left[\int q_i(\mathbf{h}_i) \frac{Pr(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)} d\mathbf{h}\right]$$

$$= \sum_{i=1}^{I} \log\left[\int Pr(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta}) d\mathbf{h}\right].$$

$$= \log P(\mathbf{x} \mid \theta)$$

E-Step

- How do we maximize the bound in the E-step?

$$\mathcal{B}[q_i(\mathbf{h}_i), \boldsymbol{\theta}] = \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i | \boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_{1\ldots I}$$

$$= \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{h}_i | \mathbf{x}_i, \boldsymbol{\theta}) Pr(\mathbf{x}_i, \boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_{1\ldots I}$$

$$= \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[Pr(\mathbf{x}_i, \boldsymbol{\theta}) \right] d\mathbf{h}_{1\ldots I} - \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{q_i(\mathbf{h}_i)}{Pr(\mathbf{h}_i | \mathbf{x}_i, \boldsymbol{\theta})} \right] d\mathbf{h}_{1\ldots I}$$

$$= \underbrace{\sum_{i=1}^{I} \log \left[Pr(\mathbf{x}_i, \boldsymbol{\theta}) \right]}_{\text{Log-Likelihood}} - \underbrace{\sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{q_i(\mathbf{h}_i)}{Pr(\mathbf{h}_i | \mathbf{x}_i, \boldsymbol{\theta})} \right] d\mathbf{h}_{1\ldots I}}_{\text{Kullback-Leibler Divergence}}$$

Kullback-Leibler Divergence

$$D_{KL}(P \| Q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

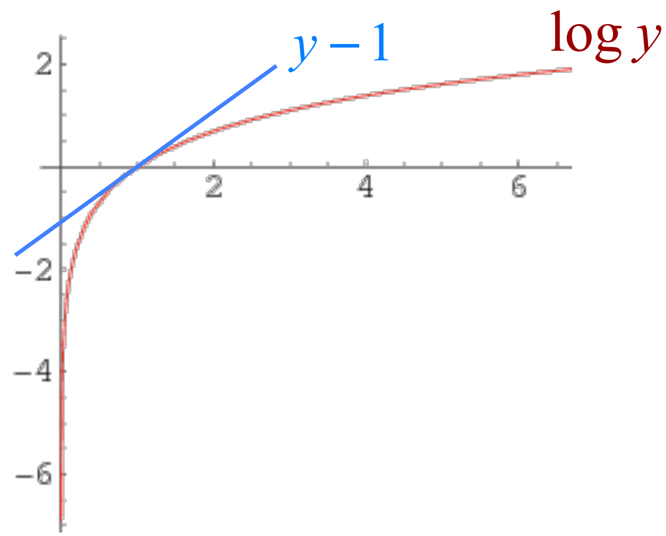
- An information-theoretic measure of the deviation between two distributions
- The KL divergence is strictly non-negative.
  - Proof:

$$D_{KL}(P \| Q) = \int p(x) \log \frac{p(x)}{q(x)} dx = -\int p(x) \log \frac{q(x)}{p(x)} dx$$

Note that $\log y \leq y - 1 \quad \forall y > 0$

Thus $\int p(x) \log \frac{q(x)}{p(x)} dx \leq \int p(x) \left(\frac{q(x)}{p(x)} - 1 \right) dx = \int \big(q(x) - p(x) \big) dx = 0$

Thus $D_{KL}(P \| Q) = \int p(x) \log \frac{p(x)}{q(x)} dx \geq 0.$



E-Step

- Thus the bound is maximized when the KL-D is 0.

i.e., when $\underbrace{q(\mathbf{h}) = P(\mathbf{h} \mid \mathbf{x}, \theta)}$

"Responsibility": h_i is responsible for explaining x_i.

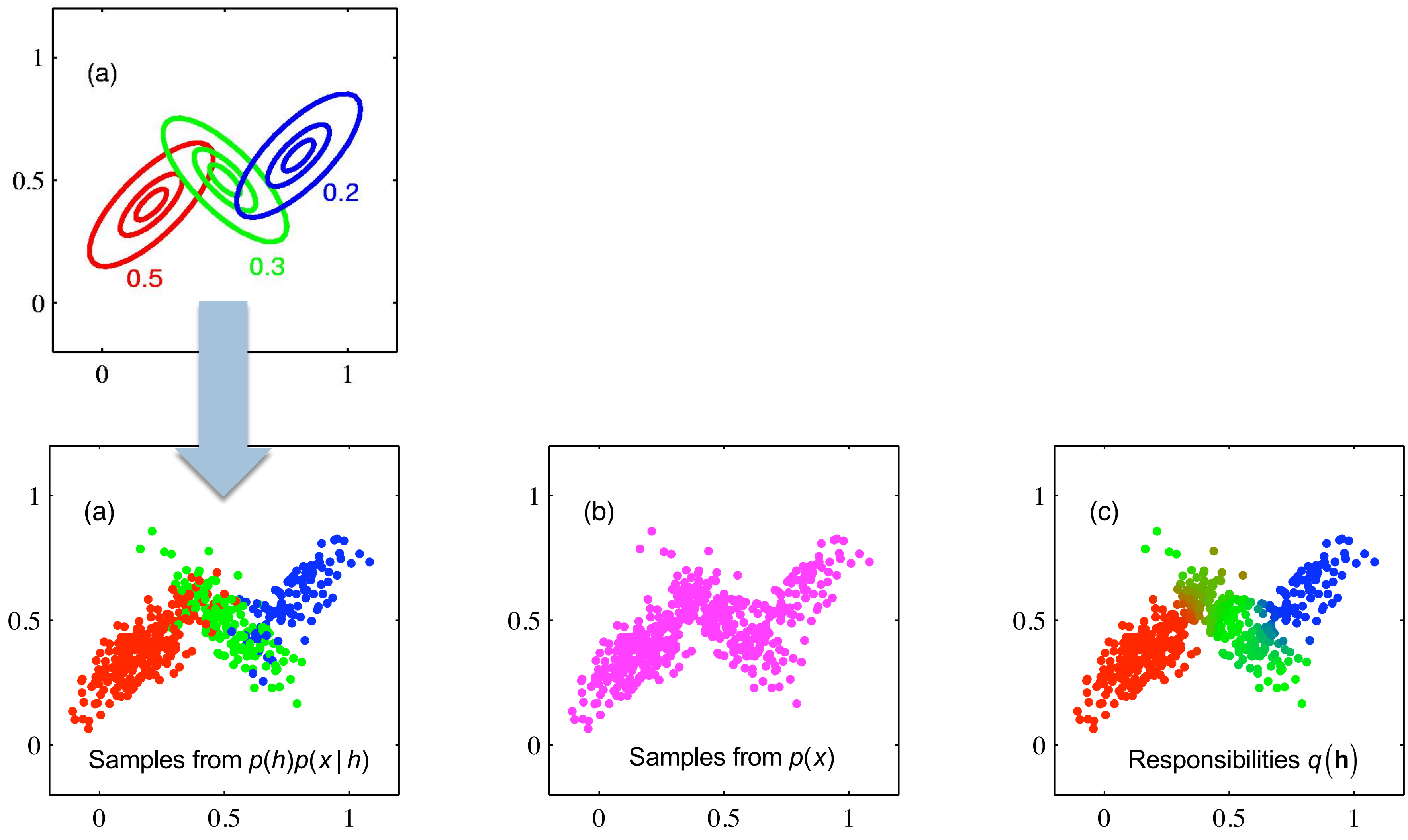
$$\mathcal{B}[q_i(\mathbf{h}_i), \boldsymbol{\theta}] = \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i | \boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_{1\ldots I}$$

$$= \underbrace{\sum_{i=1}^{I} \log \left[Pr(\mathbf{x}_i, \boldsymbol{\theta}) \right]}_{\text{Log-Likelihood}} - \underbrace{\sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{q_i(\mathbf{h}_i)}{Pr(\mathbf{h}_i | \mathbf{x}_i, \boldsymbol{\theta})} \right] d\mathbf{h}_{1\ldots I}}_{\text{Kullback-Leibler Divergence}}$$

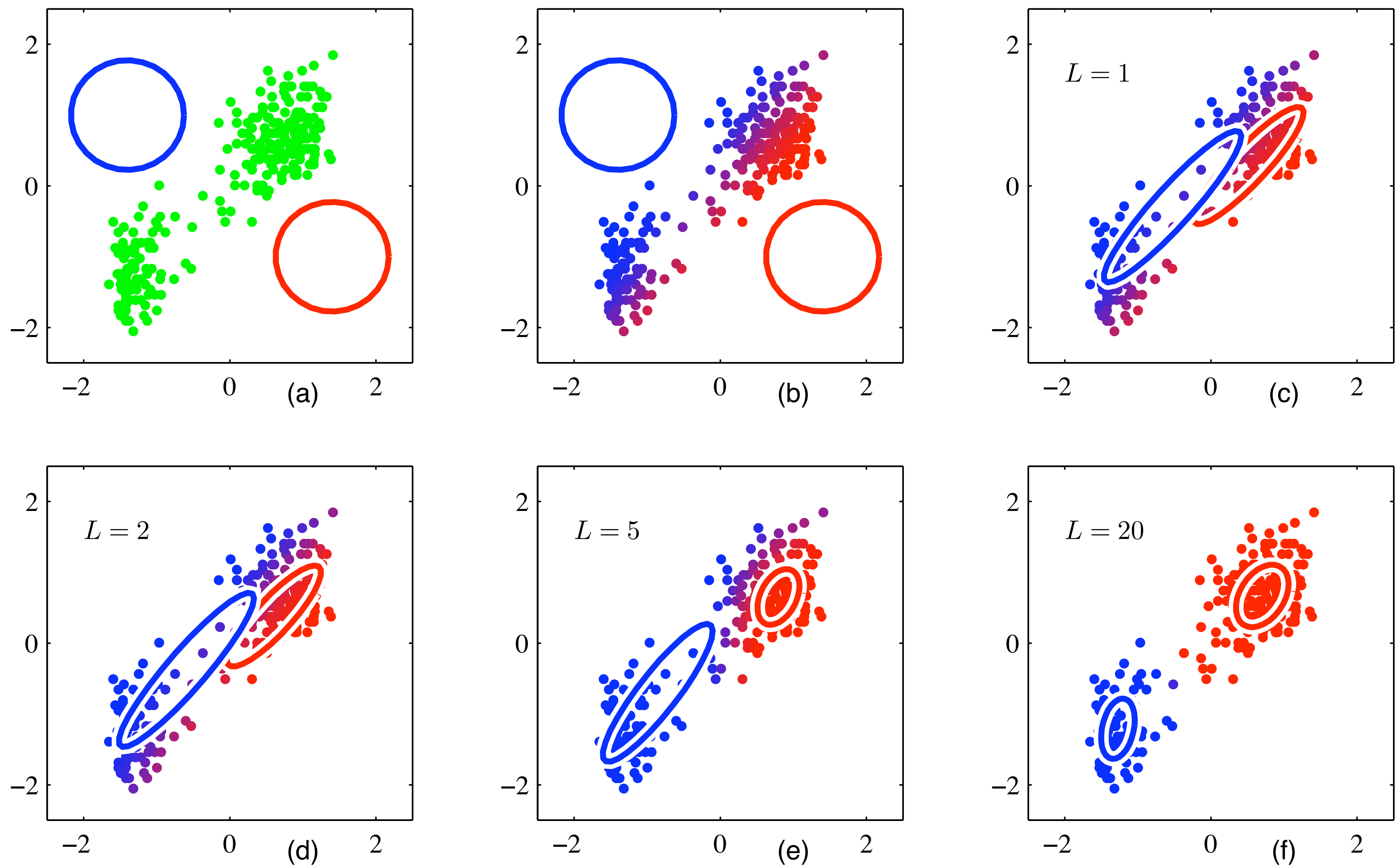
M-Step

$$
\begin{aligned}
\boldsymbol{\theta}^{[t]} &= \underset{\boldsymbol{\theta}}{\arg\max}\, \mathcal{B}[q_i^{[t]}(\mathbf{h}_i), \boldsymbol{\theta}] \\
&= \underset{\boldsymbol{\theta}}{\arg\max} \sum_{i=1}^{I} \int q_i^{[t]}(\mathbf{h}_i) \log\left[\frac{Pr(\mathbf{x}, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right] d\mathbf{h}_{1\ldots I} \\
&= \underset{\boldsymbol{\theta}}{\arg\max} \sum_{i=1}^{I} \int q_i^{[t]}(\mathbf{h}_i) \log\left[Pr(\mathbf{x}, \mathbf{h}_i|\boldsymbol{\theta})\right] - q_i(\mathbf{h}_i) \log\left[q_i(\mathbf{h}_i)\right] d\mathbf{h}_{1\ldots I} \\
&= \underset{\boldsymbol{\theta}}{\arg\max} \sum_{i=1}^{I} \int q_i^{[t]}(\mathbf{h}_i) \log\left[Pr(\mathbf{x}, \mathbf{h}_i|\boldsymbol{\theta})\right] d\mathbf{h}_{1\ldots I}.
\end{aligned}
$$

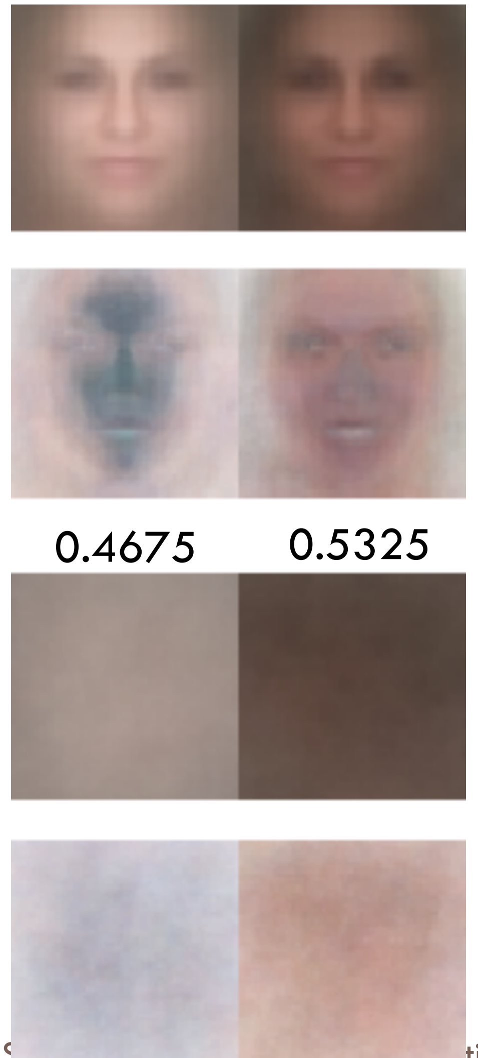
Gaussian Mixtures



Old Faithful Example



Face Detection Example: 2 Components

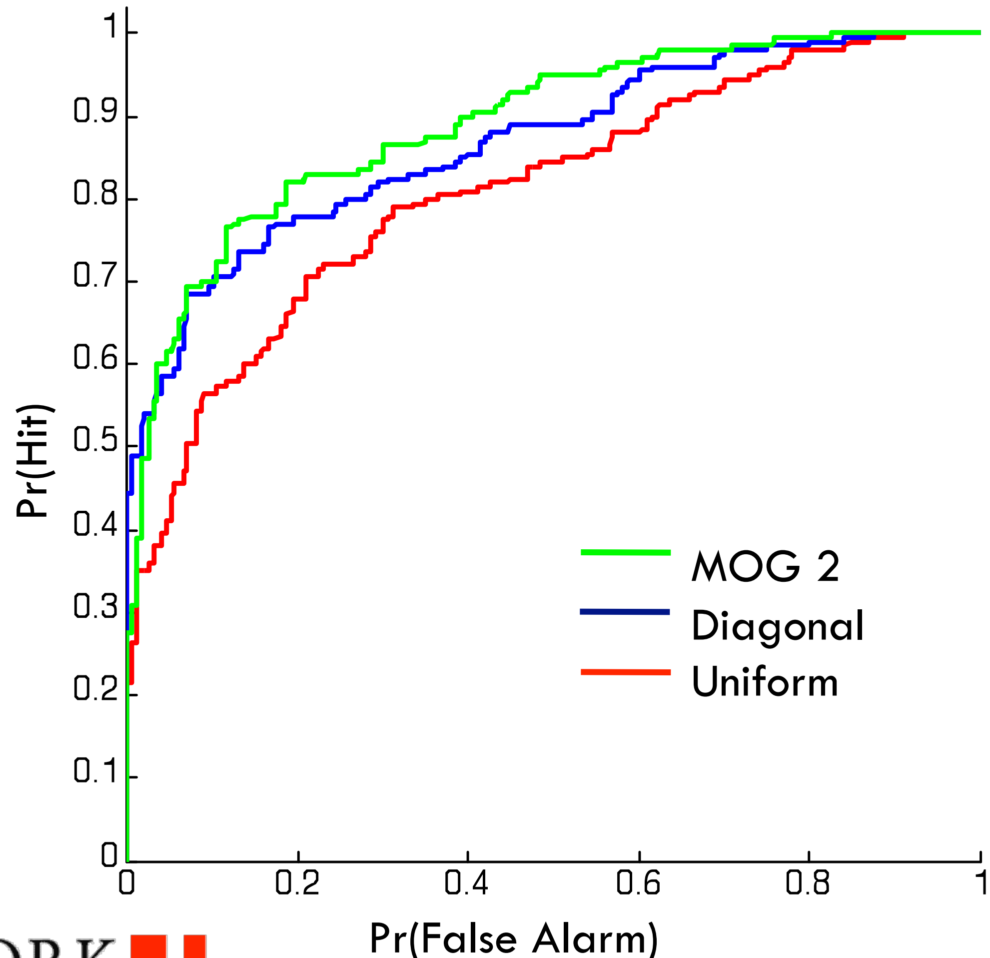


The face model and non-face model have divided the data into two clusters. In each case, these clusters have roughly equal weights.

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.

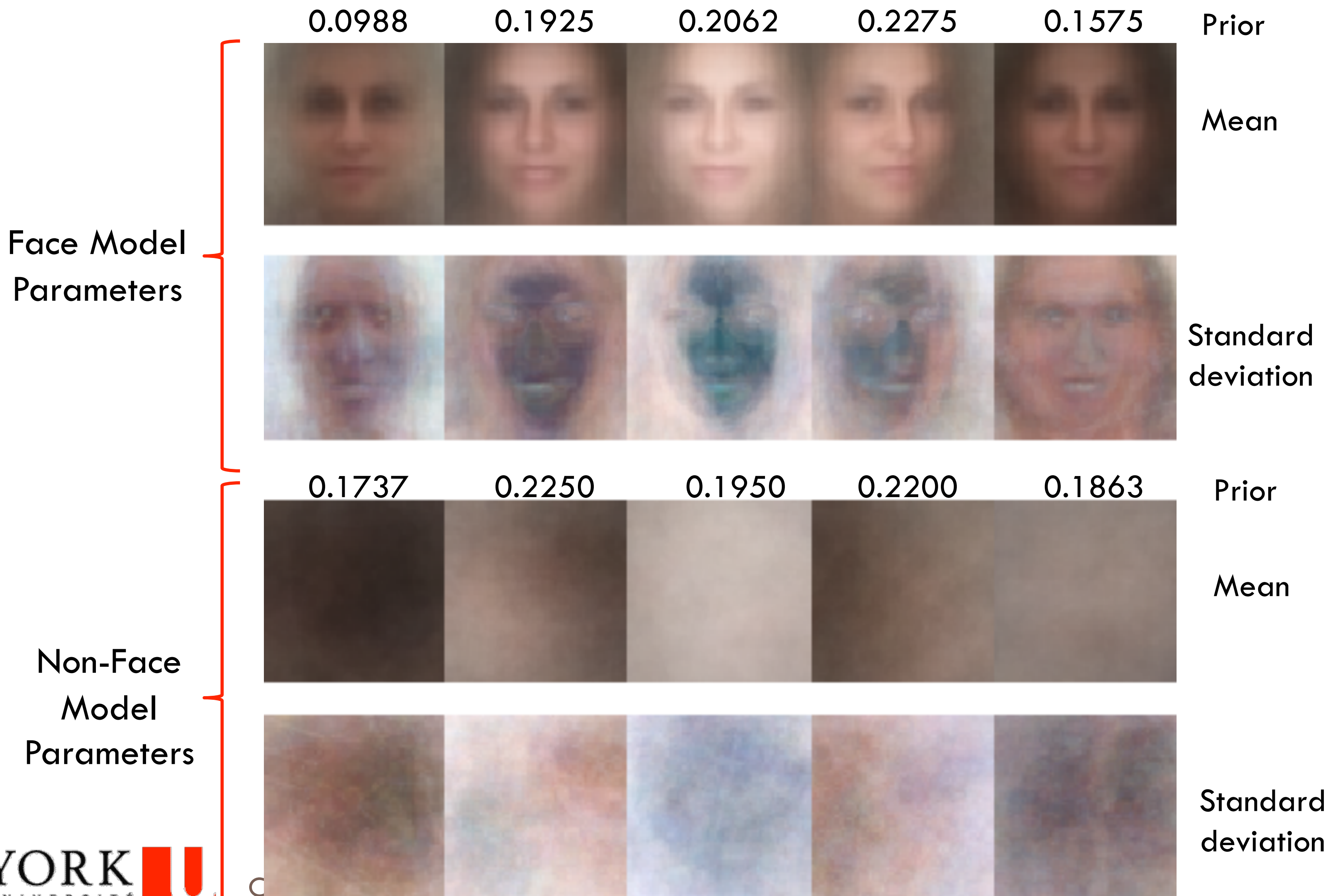
Results for MOG 2 Model



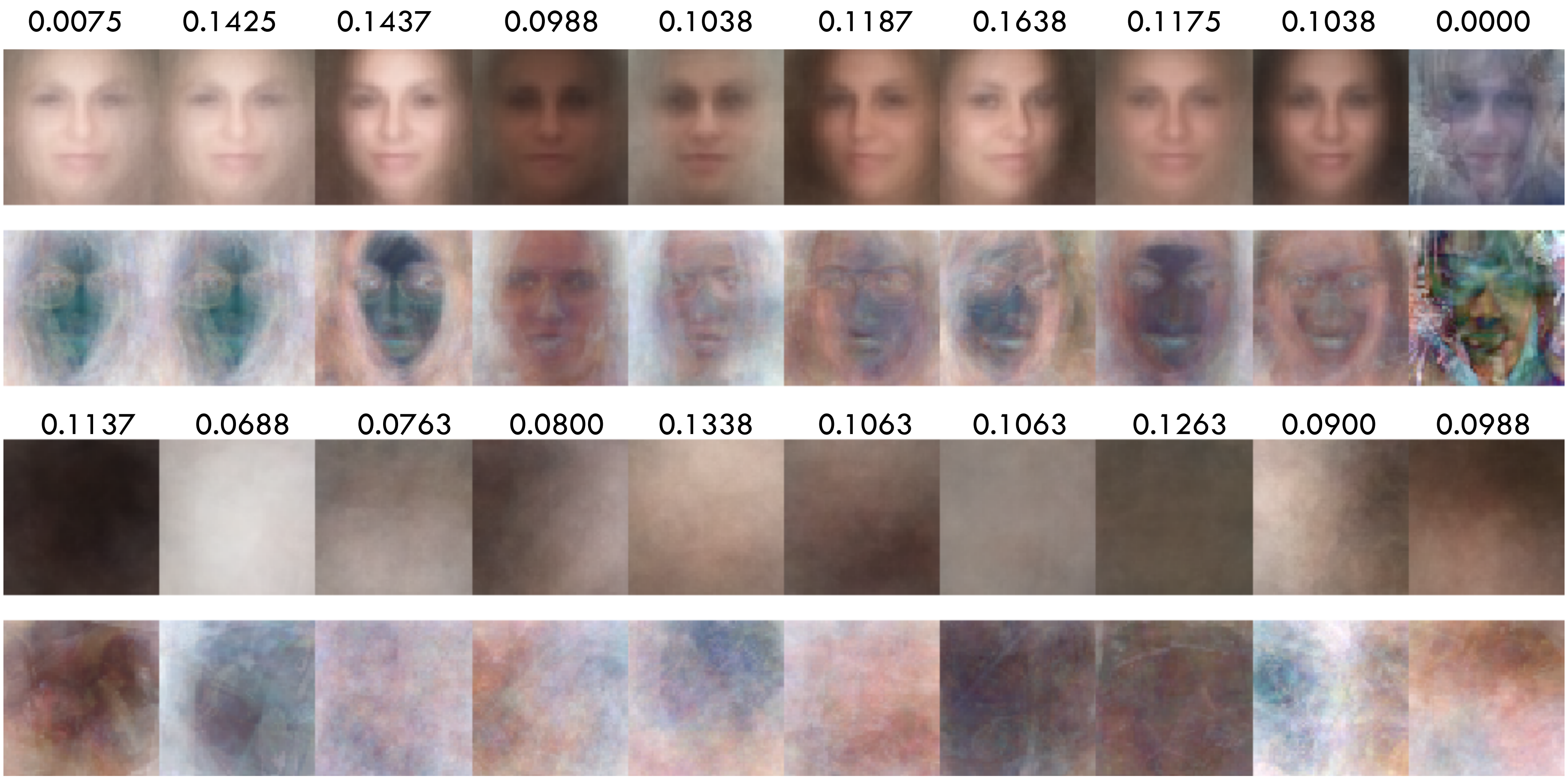
Performance improves relative to a single Gaussian model, although it is not dramatic.

We have a better description of the data likelihood.

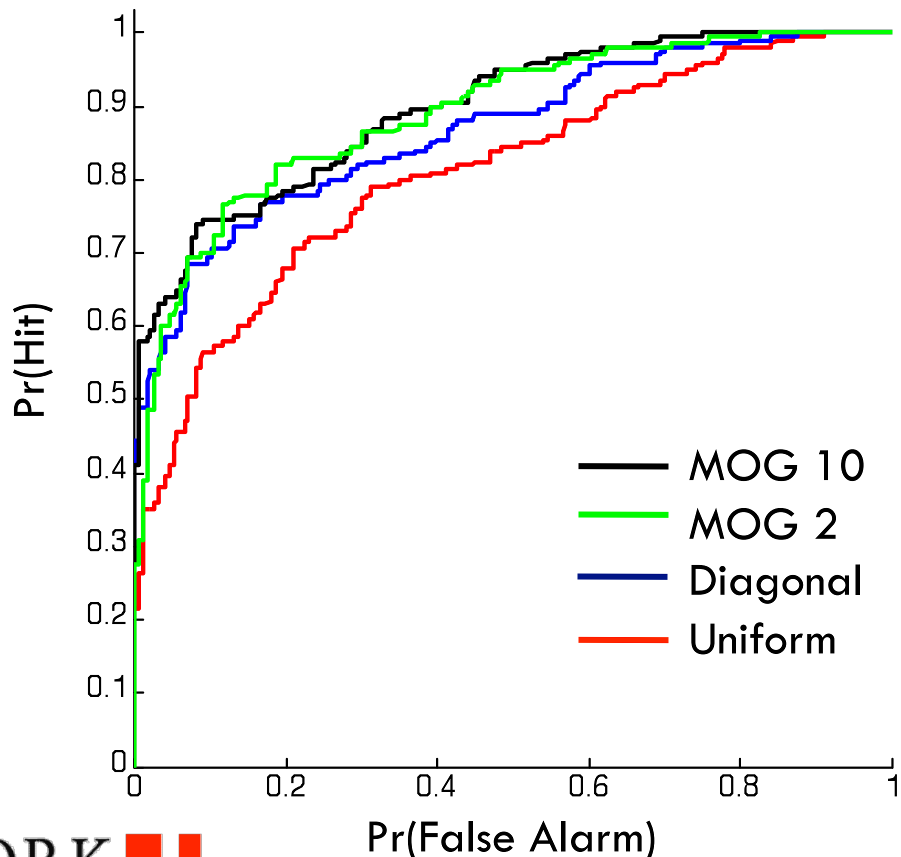
MOG 5 Components



MOG 10 Components



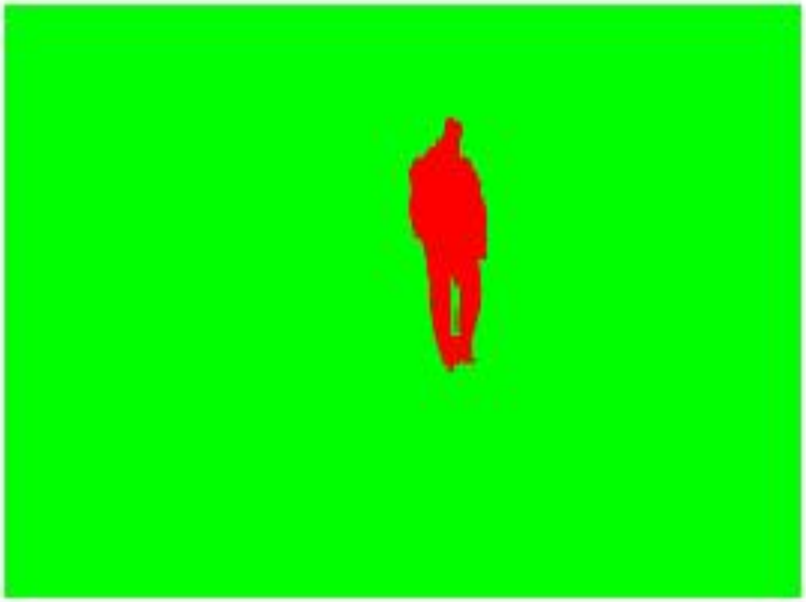
Results for Mog 10 Model



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

Background Subtraction

Test Image



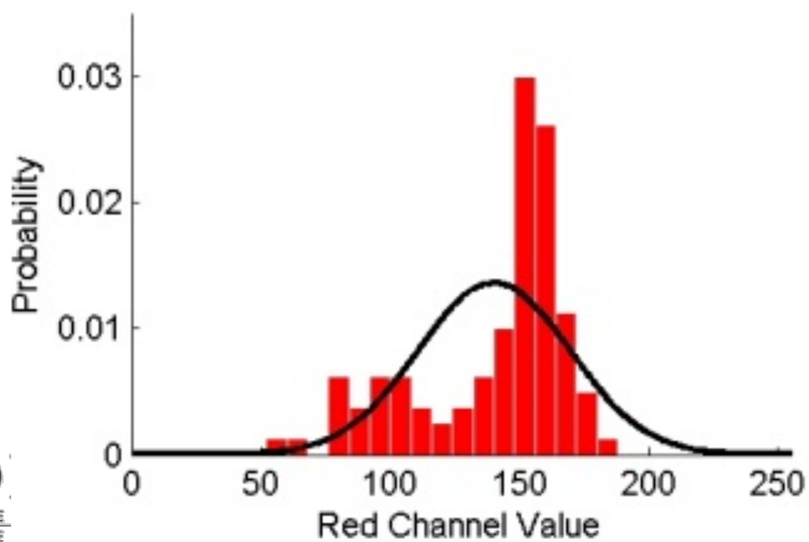
Desired Segmentation

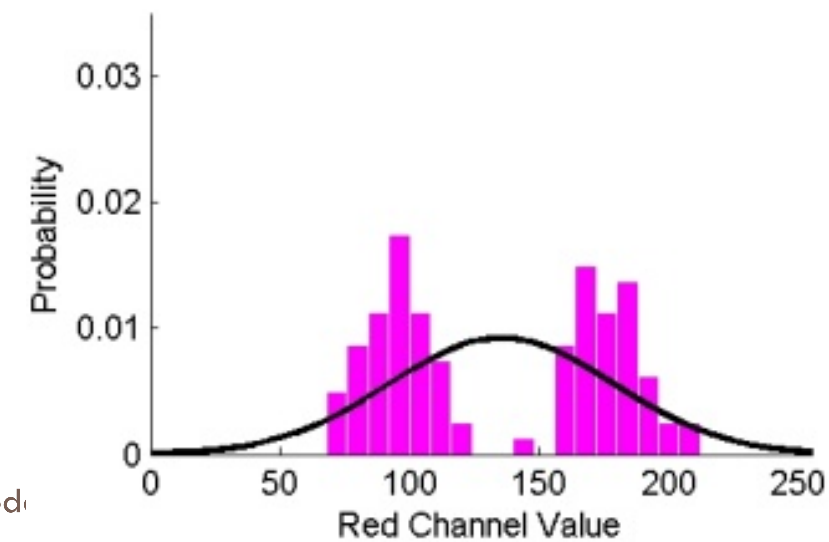
GOAL : (i) Learn background model (ii) use this to segment regions where the background has been occluded

What if the scene isn't static?

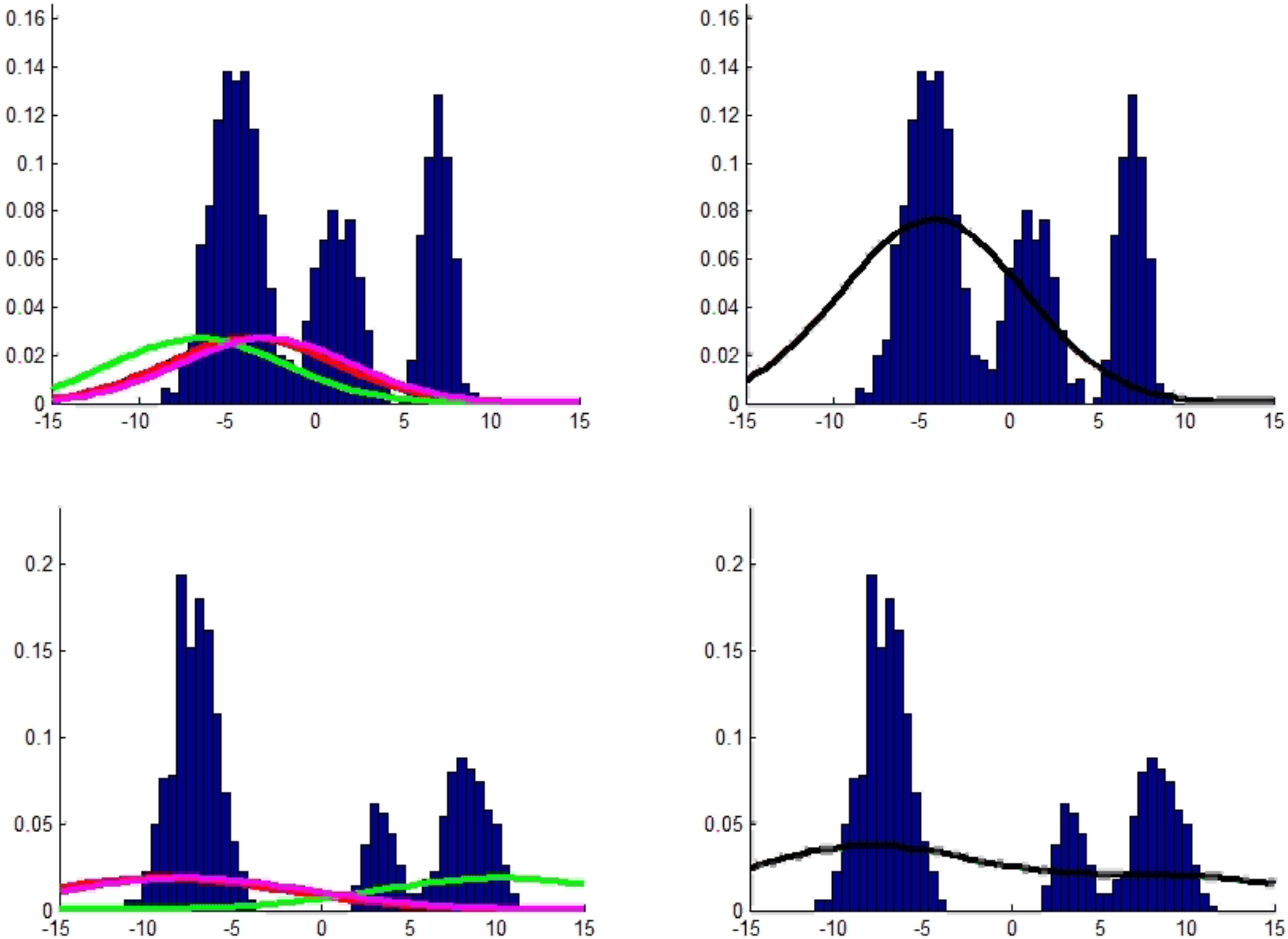
Gaussian is no longer a good fit to the data.

Not obvious exactly what probability model would fit better.





One Dimensional Example



Final Fitted Models