Segmentation/Grouping by EM

- See §16.2.1--note that we assume Gaussian for p().
- A segment is modeled as a Gaussian process that emits feature vectors (which could contain color; or color and position; or colour, texture and position).
- Segment parameters are mean and (perhaps) variance or covariance, and prior probability (was λ in the line fitting example).
- If we knew which segment each point belonged to, estimating these parameters would be easy **(this point should be familiar!)**

Segmentation/Grouping by EM

- Since we don't know which point comes from which segment, we have to use an **estimate** of the **probabilities** that a given point belongs to a given segment. We thus compute means and variances by **weighting** the standard formulas by these probabilities.

- Formally, these probabilities can be denoted

$$p(m \mid \mathbf{x}_l, \Theta^{(s)})$$

- This is the probability that $\mathbf{x}_l$ is in cluster m, given the model

- If we assume we know these estimates for the probabilities of the missing values, we can then estimate the means of the Gaussians for each segment.

Segmentation/Grouping by EM

- We estimate the mean for each segment by:

- Variances/covariances work similarly

Segmentation/Grouping by EM

- We estimate the mean for each segment by:

Iteration (step)

$$\mu_m^{(s+1)} = \frac{\sum_{l=1}^{r} \mathbf{x}_l \, p(m \mid \mathbf{x}_l, \Theta^{(s)})}{\sum_{l=1}^{r} p(m \mid \mathbf{x}_l, \Theta^{(s)})}$$

- Variances/covariances work similarly

We can sort out the chicken!

Segmentation/Grouping--E Step

- Given parameters, the probability that a given point is associated with each cluster is can be computed by:

$$p(m \mid \mathbf{x}_l, \Theta^{(s)}) = \frac{\alpha_m^{(s)} p(\mathbf{x}_l \mid \theta_m^{(s)})}{\sum_{k=1}^{M} \alpha_k^{(s)} p(\mathbf{x}_l \mid \theta_k^{(s)})}$$

- The book uses $\overline{I_{lm}}$ for $p(m \mid \mathbf{x}_l, \Theta^{(s)})$ (I suggests "indicator variable")

- (Also, my copy of the book's version of the above equation looks wrong to me-- the index l applies to points and the index for theta should refer to groups)

Segmentation/Grouping--E Step

- Given parameters, the probability that a given point is associated with each cluster is can be computed by:

$$p(m \mid \mathbf{x}_l, \Theta^{(s)}) = \frac{\alpha_m^{(s)} p(\mathbf{x}_l \mid \theta_m^{(s)})}{\sum_{k=1}^{M} \alpha_k^{(s)} p(\mathbf{x}_l \mid \theta_k^{(s)})}$$

Where does that come from?

- The book uses $\overline{I_{lm}}$ for $p(m \mid \mathbf{x}_l, \Theta^{(s)})$ (I suggests "indicator variable")

- (Also, my copy of the book's version of the above equation looks wrong to me-- the index l applies to points and the index for theta should refer to groups)

$\alpha_m^{(s)} = p(m)$ (Standard notation)

$p(m \mid \mathbf{x}_l, \Theta^{(s)}) = \dfrac{p(\mathbf{x}_l \mid m, \theta_m^{(s)}) p(m)}{p(\mathbf{x}_l \mid \theta_m^{(s)})}$ (Bayes)

$p(\mathbf{x}_l \mid \theta_m^{(s)}) = \sum_{k=1}^{M} p(\mathbf{x}_l, m, \theta_k^{(s)})$ (Marginalization)

$p(\mathbf{x}_l \mid \theta_m^{(s)}) = \sum_{k=1}^{M} p(m) p(\mathbf{x}_l \mid m, \theta_k^{(s)})$ (Definition of "|")

Therefore

$p(m \mid \mathbf{x}_l, \Theta^{(s)}) = \dfrac{\alpha_m^{(s)} p(\mathbf{x}_l \mid \theta_m^{(s)})}{\sum_{k=1}^{M} \alpha_k^{(s)} p(\mathbf{x}_l \mid \theta_k^{(s)})}$ We can do the egg!

Segmentation/Grouping by EM

- This is a lot like K-means

- Instead of binary cluster membership, each point has some probability of being in each cluster

- In addition to computing means, we generally also compute variances
  - Setting all variances equal in advance is simplest, but not so useful
  - Can assume all variances are the same ("tied")
  - Can fit different variances to each cluster (most common)
  - Can fit covariance matrices instead of variances (usually not possible if the dimension is over five or so)
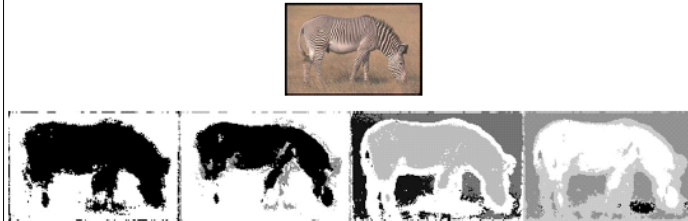
Segmentation with EM

Motion segmentation with EM (one)

- Recall the baby on the couch
- Alternative algorithms based on previous examples?

Motion segmentation with EM (one)

- Can treat background/foreground assignment as missing values!

Motion segmentation with EM (two)

- Model image sequence as consisting of regions (layers) of parametric motion
  - For example, affine motion is popular

$$\begin{pmatrix} v_x \\ v_y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$
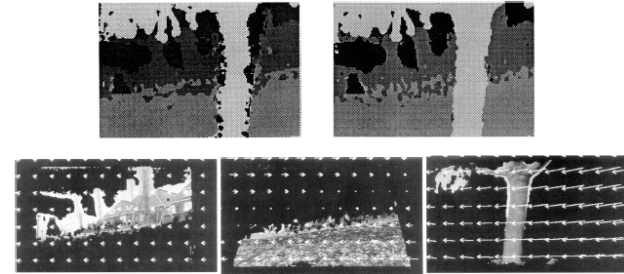
- Now we need to
  - Determine which pixels belong to which region
  - Estimate parameters
  - Yet another example of a missing value problem!

Three frames from the MPEG "flower garden" sequence

Figure from "Representing Images with layers,", by J. Wang and E.H. Adelson, IEEE
Transactions on Image Processing, 1994, c 1994, IEEE

Grey level shows region no. with highest probability

Segments and motion fields associated with them

Figure from "Representing Images with layers,", by J. Wang and E.H. Adelson, IEEE
Transactions on Image Processing, 1994, c 1994, IEEE

If we use multiple frames to estimate the appearance
of a segment, we can fill in occlusions; so we can
re-render the sequence with some segments removed.

Figure from "Representing Images with layers,", by J. Wang and E.H. Adelson, IEEE
Transactions on Image Processing, 1994, c 1994, IEEE
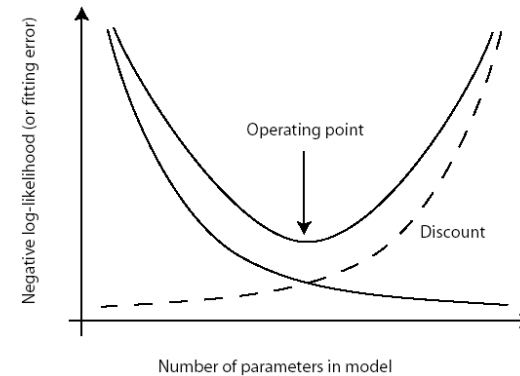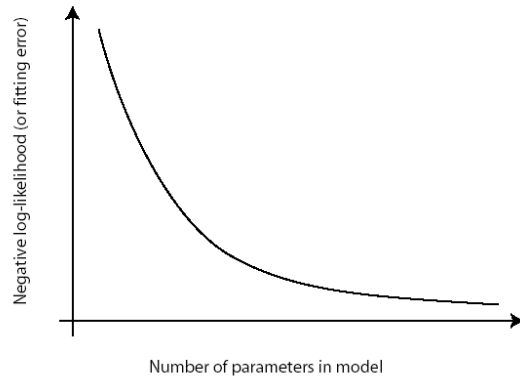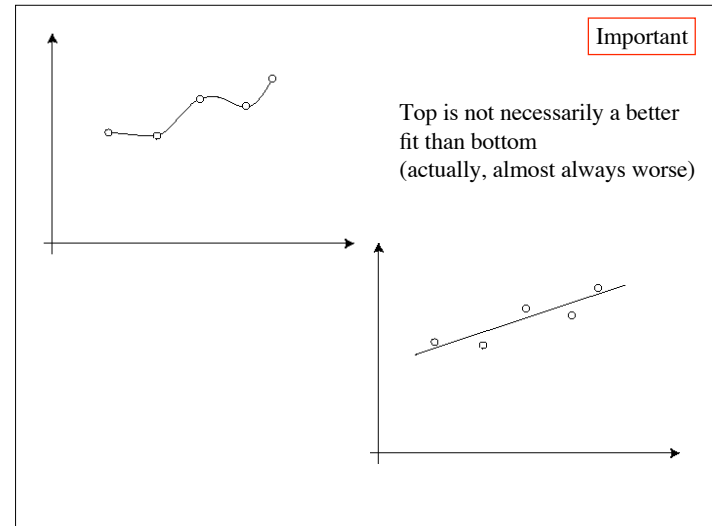
RANSAC versus EM

- Many, but not all problems that can be attacked with EM can also
  be attacked with RANSAC
  - For RANSAC, we need to be able to get a parameter estimate with a
    manageably small number of random choices.
  - RANSAC is often better

Model Selection

- In general, models with more parameters will fit a dataset better, but are poorer at prediction
- This means we can't simply look at the negative log-likelihood (or fitting error)

Top is not necessarily a better fit than bottom
(actually, almost always worse)

Number of parameters in model

Number of parameters in model

We can discount the fitting error with some term in the number of parameters in the model.

Discounts

- Let N be the number of data points, p the number of parameters

- AIC (an information criterion)
  - choose model with smallest value of

$$-2L(D;\theta^*) + 2p$$

- BIC (Bayes information criterion)
  - choose model with smallest value of

$$-2L(D;\theta^*) + p \log N$$

- Minimum description length
  - same criterion as BIC, but derived in a completely different way