Probabilistic Fitting

- Generative probabilistic model
  - Tells a story about how stochastic data comes to be
  - Darts fall around the center of the board, but where exactly?
  - Consider a model with parameters, θ
  - Consider an observation, x_i
  - We denote the probability of seeing x_i under the model by:

$$p(x_i \mid \Theta)$$

Read "given" or "conditioned on"

Restricts to the case of θ

Defined by $P(A \mid B) = \frac{P(A,B)}{P(B)}$

Probabilistic Fitting

- Multiple observations
  - Suppose we have multiple observations, in a vector **x**
  - What is the probability of **x**?

- If observations are independent then probability is the product of the individual observations
  - Essentially a definition, but it is consistent with intuition
  - The observations are conditionally independent **given** the model

- So, the probability of **x** is then:

$$p(\mathbf{x} \mid \Theta) = \prod p(x_i \mid \Theta)$$

Probabilistic Fitting

- So, given the model, we have the probability of observing the data

$$p(\mathbf{x} \mid \Theta) = \prod p(x_i \mid \Theta)$$

- But what we really want is the probability of the model (parameters) given the data!

- Bayes rule comes to the rescue!

Bayes Rule

- Bayes rule: $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$
- Proof $$P(A,B) = P(B \mid A)P(A) = P(A \mid B)P(B)$$
- With our notation: $$P(\Theta \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \Theta)P(\Theta)}{P(\mathbf{x})}$$

likelihood function for the parameters

prior probability (often taken to be uniform)

$$P(\Theta \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \Theta) P(\Theta)}{P(\mathbf{x})}$$

posterior probability

normalizer, often is not of interest

Common special case

$$P(\Theta \mid \mathbf{x}) \propto P(\mathbf{x} \mid \Theta)$$

Know the words in **red**

Probabilistic Fitting

- If we assume **uniform** prior, then we can find the posterior density for the parameters by:

$$p(\Theta \mid \mathbf{x}) \propto p(\mathbf{x} \mid \Theta)$$

- Now the objective is to find the parameters Θ such that this *likelihood* is maximum

- Note--this is the same as finding the parameters which minimize the **negative log likelihood**

Probabilistic fitting with independence and uniform prior

Finding the "best" model under simple circumstances

$$\underset{\Theta}{\text{maximize}}\ p(\Theta \mid \mathbf{x}) \qquad \text{(one definition of best } \Theta\text{)}$$

$$\underset{\Theta}{\text{maximize}}\ p(\mathbf{x} \mid \Theta) \qquad \text{(by Bayes rule, uniform prior)}$$

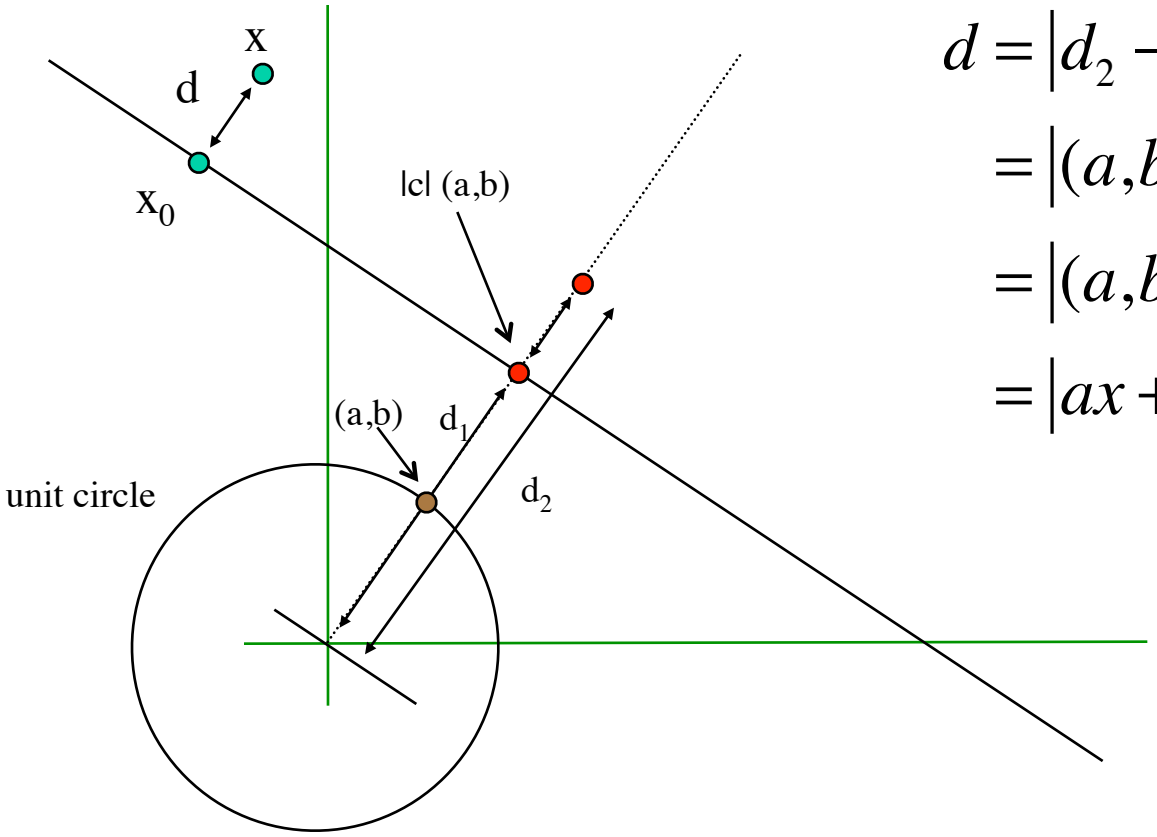
$$\underset{\Theta}{\text{minimize}}\ -\log(p(\mathbf{x} \mid \Theta)) \qquad \text{(log is monotonic increasing)}$$

$$\underset{\Theta}{\text{minimize}}\ -\log\left(\prod p(x_i \mid \Theta)\right) \qquad \text{(by independence)}$$

$$\underset{\Theta}{\text{minimize}}\ -\sum \log(p(x_i \mid \Theta)) \qquad \text{(high school math)}$$

Lines (again!)

- ax+by+c=0 where a^2+b^2=1
- Distance squared from (x,y) to this line is (ax+by+c)^2



distance of line
to origin is |c|

$$d = |d_2 - d_1|$$
$$= |(a,b)\bullet\mathbf{x} - (a,b)\bullet\mathbf{x_0}|$$
$$= |(a,b)\bullet\mathbf{x} + c|$$
$$= |ax + by + c|$$

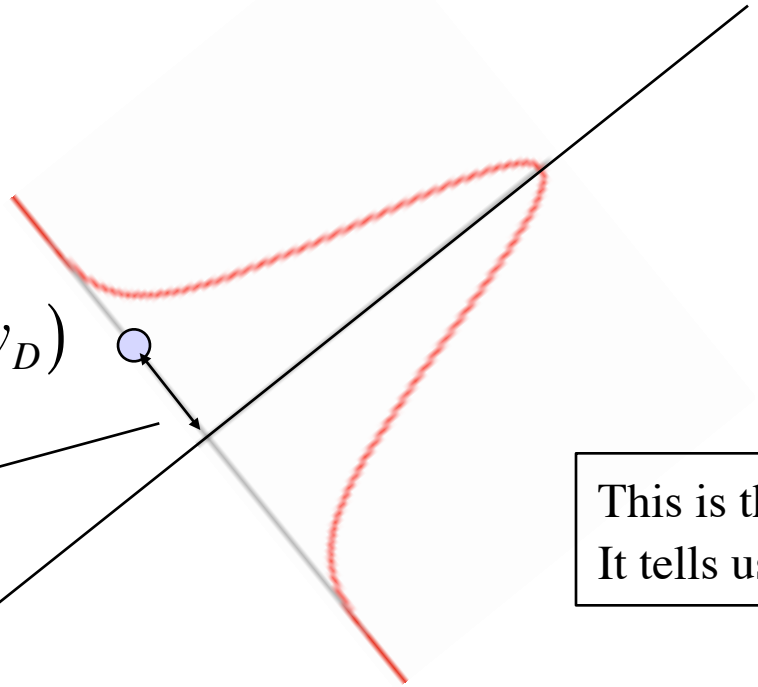
- **Generative model** for lines: Choose point on line, and then, with probability proportional to p(d), **normally distributed** (Gaussian), go a distance d from the line.

- Now the probability of an observed (x,y) is given by

$$p((x,y)\,|\,\Theta) \propto \exp(-\frac{(ax+by+c)^2}{2\sigma^2})$$

Lines

Convenient formula for line
$ax+by+c=0$
where $a^2+b^2=1$



$$d^2 = (ax_D + by_D + c)^2$$

This is the generative model
It tells us P(data | model)

$$p((x_D, y_D) \mid \Theta) \propto \exp\left(-\frac{(ax_D + by_D + c)^2}{2\sigma^2}\right)$$

We have the probability density of the observed (x,y) given by

$$p((x,y)\,|\,\Theta) \propto \exp(-\frac{(ax+by+c)^2}{2\sigma^2})$$

The negative log is

$$\frac{(ax+by+c)^2}{2\sigma^2}$$

And the negative log likelihood of multiple observations is

$$\frac{1}{2\sigma^2}\sum_i (ax_i+by_i+c)^2$$

From the previous slide, we had that the negative log likelihood of multiple observations is given by

$$\frac{1}{2\sigma^2}\sum_i (ax_i + by_i + c)^2 \qquad \text{(where } a^2 + b^2 = 1\text{)}$$

This should be recognizable as homogeneous least squares

Thus we have shown that least squares is maximum likelihood estimation under normality (Gaussian) error statistics!

Back to  and  problems

Segmentation/Grouping by EM

- We assume that the observed data is from multiple hidden processes (e.g., clusters)

- A generative statistical model for the data
  - Choose a “cause”, e.g., a cluster, according to $p(c)$.
  - Given the cluster, sample its probability model $p(X|c)$.

- For concreteness, assume Gaussians
  - This is a Gaussian Mixture Model (GMM)

Details optional

Gaussian Mixture Model (GMM)

- Generative process
  - Chose a mixture component (cluster), m, with probability $p(m)$
  - For the component m, consult the particular Gaussian distribution
  - Generate a sample from that distribution

- This models the distribution

$$p(\mathbf{x}) = \sum_m p(m)p(\mathbf{x} \mid m) \qquad \text{where} \qquad p(\mathbf{x} \mid m) = \mathbb{N}(\boldsymbol{\mu}_m, \Sigma)$$

Example here is a Gaussian

Details optional

Gaussian Mixture Model (GMM)

- Generative process
  - Chose a mixture component (cluster), m, with probability $p(m)$
  - For the component m, consult the particular Gaussian distribution
  - Generate a sample from that distribution

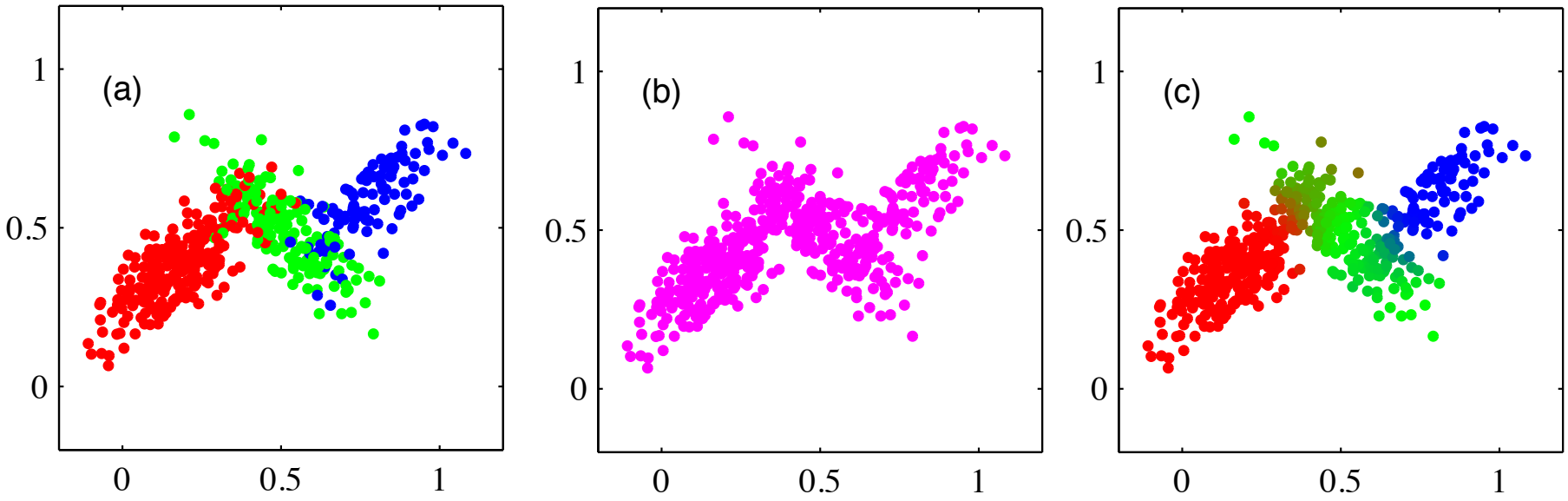
- This models the distribution

$$p(\mathbf{x}) = \sum_m p(m)p(\mathbf{x} \mid m) \qquad \text{where} \qquad p(\mathbf{x} \mid m) = \mathbb{N}\left(\boldsymbol{\mu}_m, \Sigma\right)$$

- And for multiple points

$$p(\{\mathbf{x}_i\}) = \prod_i \left(\sum_m p(m)p(\mathbf{x} \mid m) \right)$$

GMM illustrated



Truth

Data

Clustering according to the model

Segmentation/Grouping by EM

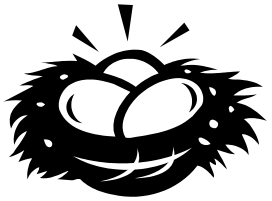
- Statistical clustering estimates the parameters of the assumed underlying model
  - In the GMM this is means and variances of the Gaussians
- The generative process tells us about the data, given the model
  - Clustering reverses this
  - Given the data (evidence) what is the model
- This is an example of Bayesian inference
  - Unlike the line case, there is no closed form solution for the minimum
  - One way to approach this problem is EM (Expectation Maximization)
    - EM is the classic “chicken” and “egg” algorithm

Segmentation/Grouping by EM

- Using EM for statistical clustering breaks the problem into to two parts
  - Who goes with which model (correspondence)
  - What are the model parameters (e.g., means and variances)

- If we knew which segment each point belonged to, estimating these parameters would be easy **(this point should be familiar!)**
  - EM relaxes this to be a probability estimate that each point is with each cluster **(soft assignments).**

- If we know the parameters, we can compute the probabilities that a point belongs to each cluster (**soft clustering**).



EM flow chart

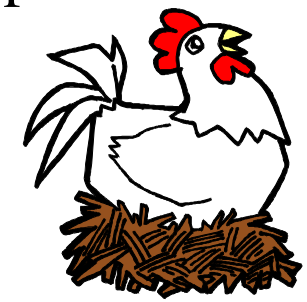
Design probability model

Guess soft correspondence

OR

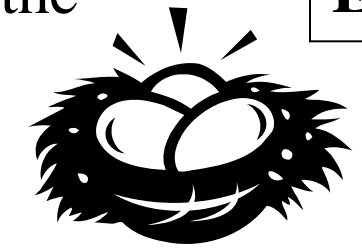
Guess model parameters

Assume (soft) correspondence are **fixed**. Update model parameters using max likelihood



M

Assume model is **fixed**. Find correspondence probabilities (the extent each point is in each cluster).



E

Segmentation/Grouping by EM

- Since we don't know which point comes from which segment, we use an **estimate** of the **probabilities** that a given point belongs to a given segment.
- Formally, these probabilities can be denoted $p(m \mid \mathbf{x}_i, \Theta^{(m)})$
- This is the probability of being in cluster m, given the data, $\mathbf{x}_i$, and the model for m.
- 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.
- Specifically, we compute means and variances by **weighting** the standard formulas by these probabilities.

Details optional

Segmentation/Grouping by EM

- We estimate the mean for each segment by:

Iteration (step)

$$\mu_m^{(s+1)} = \frac{\sum_{i=1}^{r} \mathbf{x}_i \, p(m \mid \mathbf{x}_i, \Theta_m^{(s)})}{\sum_{i=1}^{r} p(m \mid \mathbf{x}_i, \Theta_m^{(s)})}$$

This sum is also the mixture component weight, *p(m)*.

- Variances/covariances work similarly

We can sort out the chicken!

Details optional

Segmentation/Grouping--E Step

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

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

- The book uses $\overline{I_{lm}}$ for $p(m \mid \mathbf{x}_i, \Theta^{(s)})$ (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)

Details optional

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

Details optional

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

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

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

We can do the egg!

Need to know

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 ("tied parameters") is simplest, but often having different variances is important.
  - 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

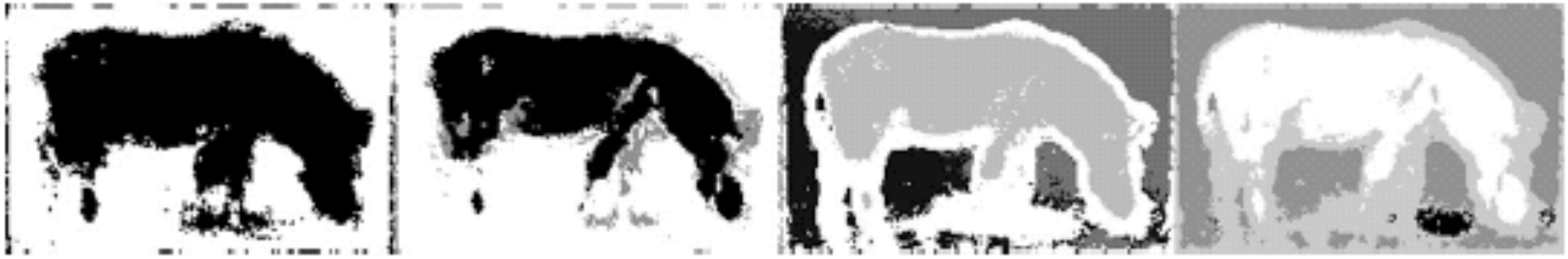


Figure from “Color and Texture Based Image Segmentation Using EM and Its Application to Content Based Image Retrieval”, S.J. Belongie et al., Proc. Int. Conf. Computer Vision, 1998, c1998, IEEE

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.
  - If applicable, RANSAC is often better because
    - Getting the right error model is hard
    - Fitting using EM is very prone to local minimum
      - Alternative is sampling methods
      - Very effective but technical and expensive

Need to know

Statistical clustering main points

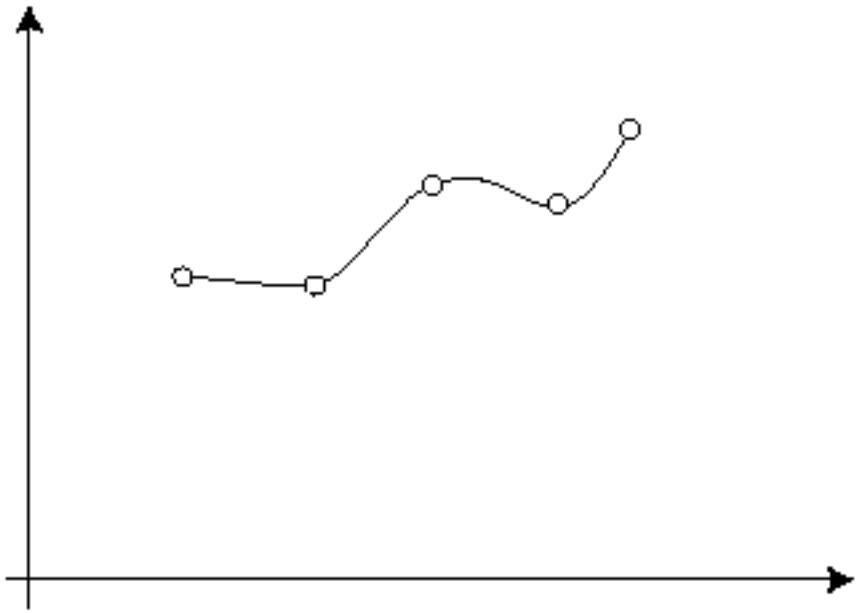
- Data points come from a specific statistical generative process
- Clustering models assume that there a multiple components
  - The components need not be of the same form
    - Important advantage of probabilistic methods is that disparate models are properly connected via probability theory
- We cluster by *fitting* the model to training data
- This has no close form solution. We use fancy methods to estimate the fit.
- The most common principled method is Expectation Maximization (EM)
- EM breaks our clustering problem into two parts that alternate
  - Estimating the *expected* value of missing values (correspondence)
  - Estimating *maximum* likelihood of the parameters given the missing value assignment probabilities

Important

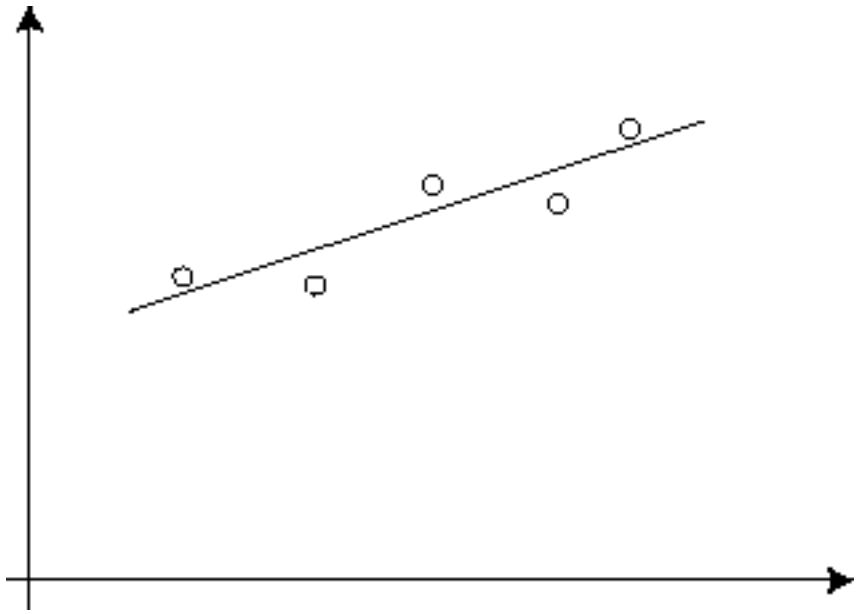
Model Selection

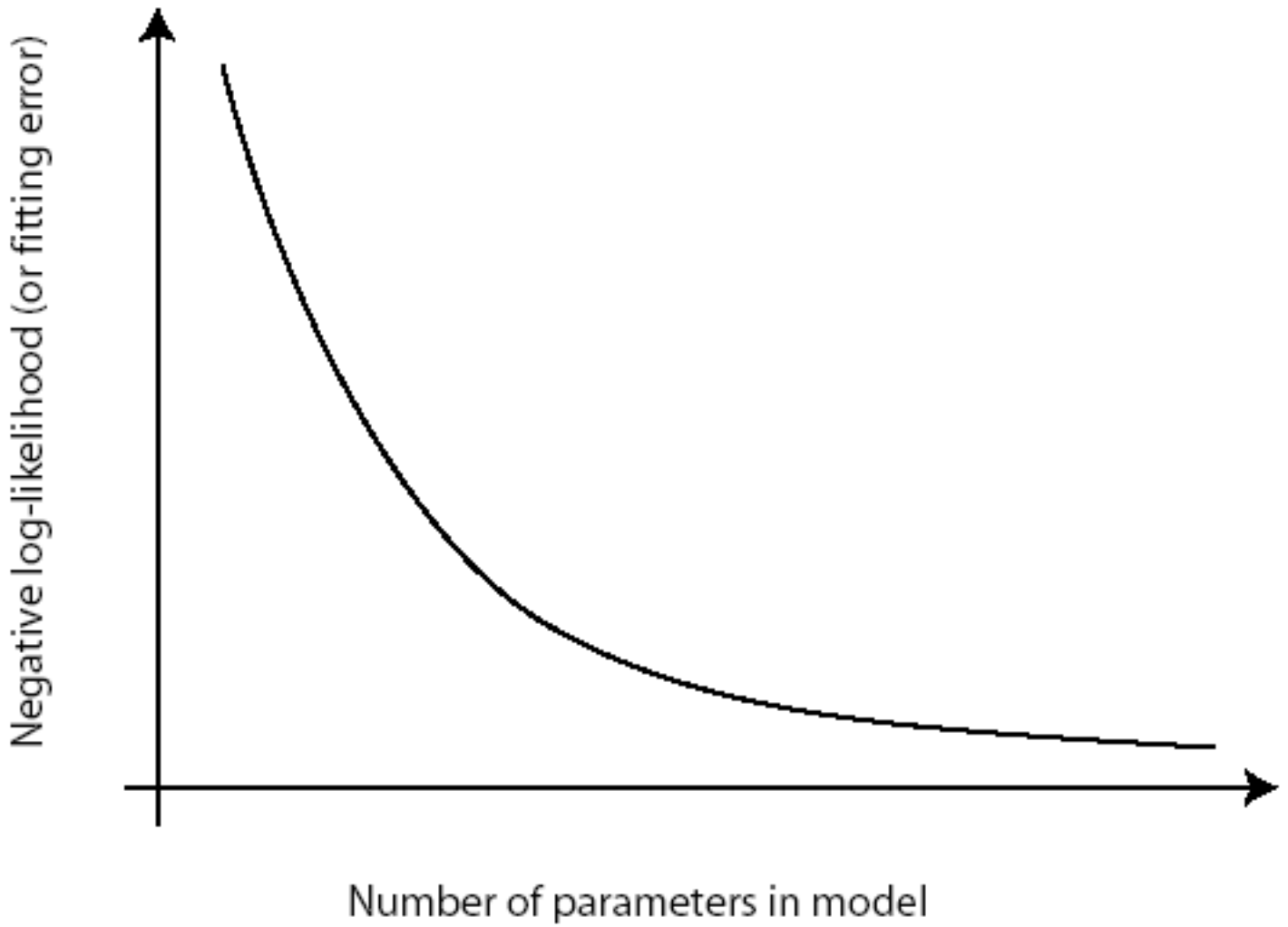
- Examples of model selection
  - Choose the number of clusters
  - Choose Gaussian verses Poison
  - Choose a quadratic curve instead of a line
- In general, models with more parameters will fit a dataset better, but are poorer at prediction
- This means we can't simply look at the negative log-likelihood (or fitting error)

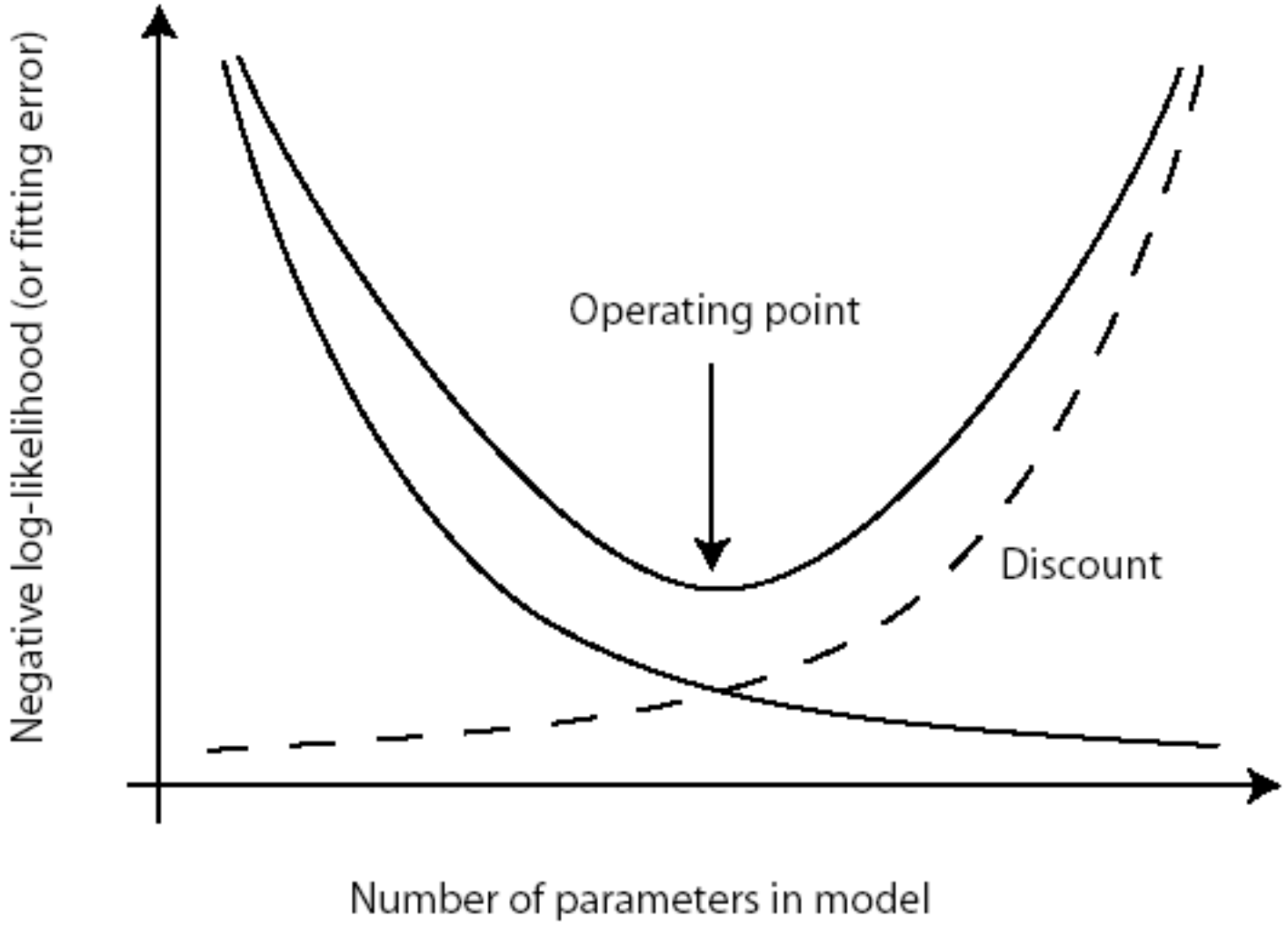
Important



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




Negative log-likelihood (or fitting error)
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 $-2L\left(D;\theta^{*}\right)+2p$
- BIC (Bayes information criterion)
  - choose model with smallest value of $-2L\left(D;\theta^{*}\right)+p\log N$
- Minimum description length
  - same criterion as BIC, but derived in a completely different way
- Now that you know about "discounts", be aware that they usually don't work very well!
  - Assumptions used to justify them are too restrictive for most problems.

Important

Cross-validation

- Split data set into two pieces, fit to one, and compute negative log-likelihood on the **other**
- One set is “training data”, the other is “testing data” or “held out data”
- Average over different splits
- This estimates the quality of your model
  - Often (rightfully so) used to compare algorithms
- If you are doing model selection, then you choose the model with the smallest value of this average
  - This works because adding parameters causes over fitting of the training data which gives worse performance on test data
  - However, it ignores priors over models