CSE 312

Foundations of Computing II

Lecture 27: Multivariate Gaussians, clustering and EM

Anna R. Karlin

Slide Credit: Emily Fox

Incorporating some of my own from CSE 446 ☺

Quiz 3: Monday Dec 6

Final: Monday Dec 13

Details on both at top of webpage under announcements

Goal for today

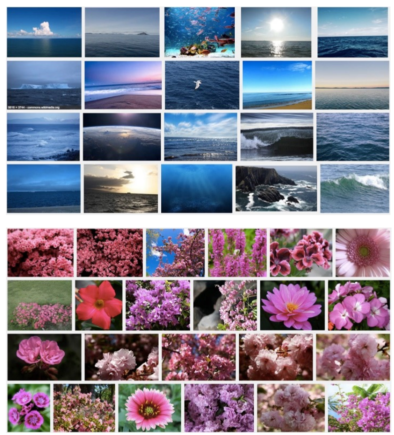
- Introduce you to a fundamental machine learning problem: clustering.
- Give you a very gentle introduction to multivariate Gaussian distributions.

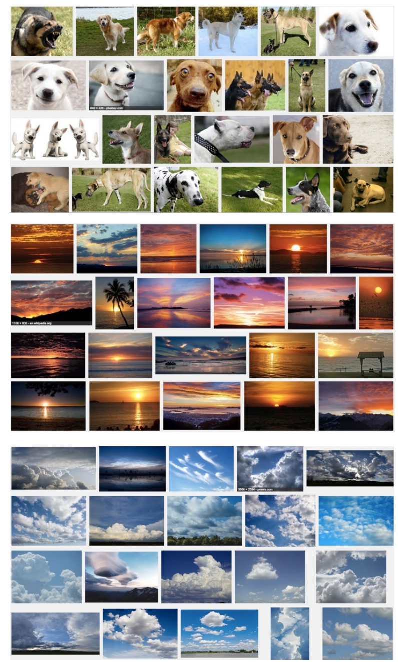
©2017 Emily Fox

Motivating application: Clustering images

Discover groups of similar images

- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- …





©2017 Emily Fox

Another example: clustering documents

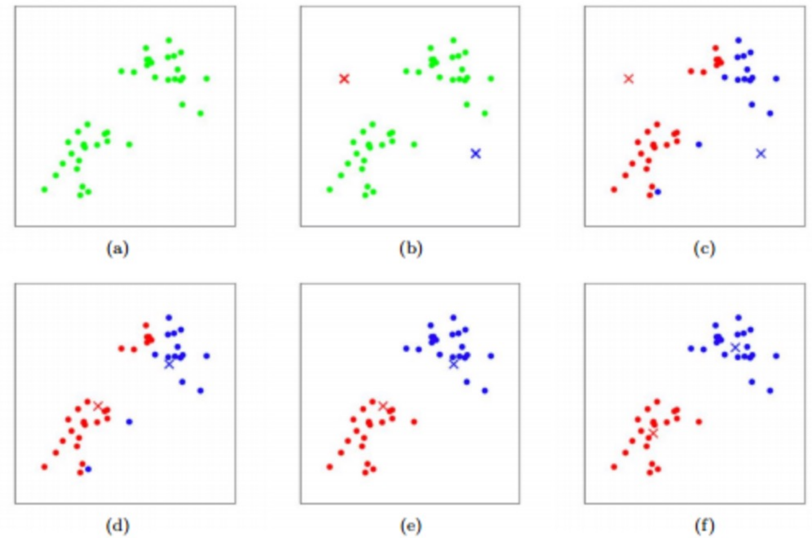
E.g. into international news, sports, culture, etc.

So how are these data represented?

- As points in d-dimensional space (where d is typically large).

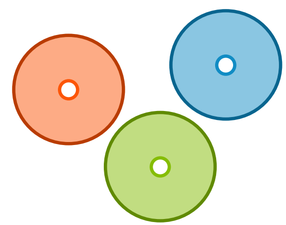
©2017 Emily Fox

How to approach clustering? One way: k-means



Only center matters
Not cluster shapes

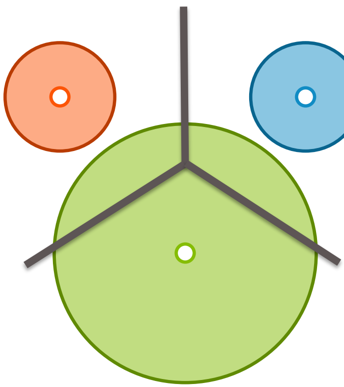
Equivalent to assuming spherically symmetric clusters



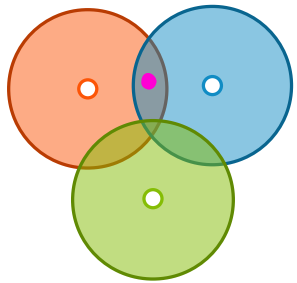
 K-means picture From Stanford CS 221

©2017 Emily Fox

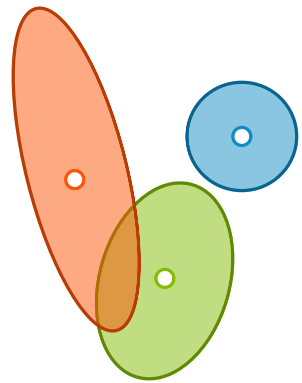
Failure modes of k-means clustering



disparate cluster sizes



overlapping clusters

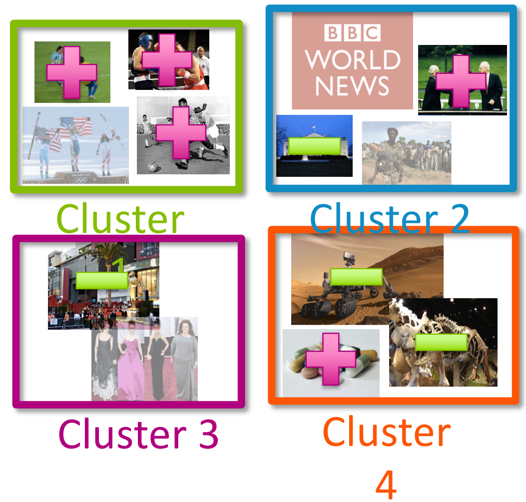


different shaped/oriented clusters

©2017 Emily Fox

Motivates probabilistic model: Mixture model

- Take uncertainty in assignment into account
  e.g., when clustering documents, might want to say 54% chance document is world news, 45% science, 1% sports, and 0% entertainment



- Allow for cluster shapes not just centers

- Enables learning different weightings of dimensions
  - e.g., how much to weight each word in the vocabulary when computing cluster assignment

©2017 Emily Fox

Mixture model

- k clusters, defined by probability distribution over **Gaussian** random variables.

- π_i, μ_i, σ_i^2 for each cluster. $$\sum_{j=1}^{k} \pi_j = 1.$$

- Problem: Assume that the data comes from such a distribution, and recover the parameters of the distribution.
- Determine, for each point, the likelihood of it belonging to cluster j, for each j.

©2017 Emily Fox

Background: Multivariate Gaussian distributions

Overly simple image representation

Consider average red, green, blue pixel intensities

[R = 0.05, G = 0.7, B = 0.9]

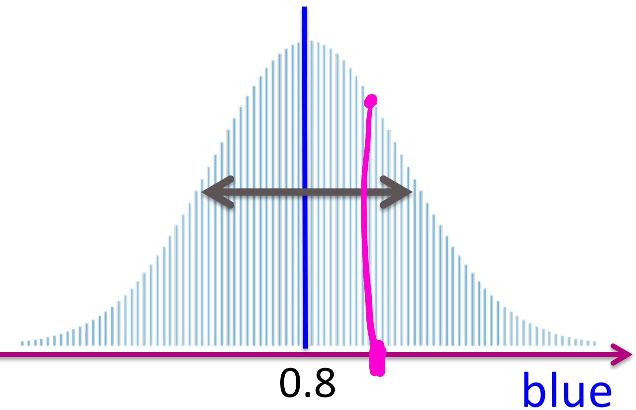
[R = 0.85, G = 0.05, B = 0.35]

[R = 0.02, G = 0.95, B = 0.4]

©2017 Emily Fox

Distribution over all cloud images

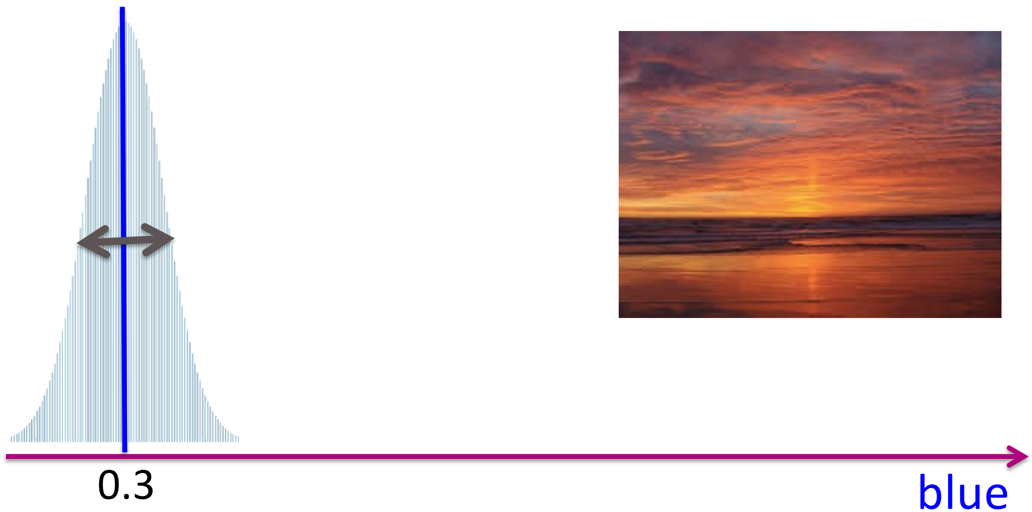
Let's look at just the blue dimension



©2017 Emily Fox

Distribution over all sunset images

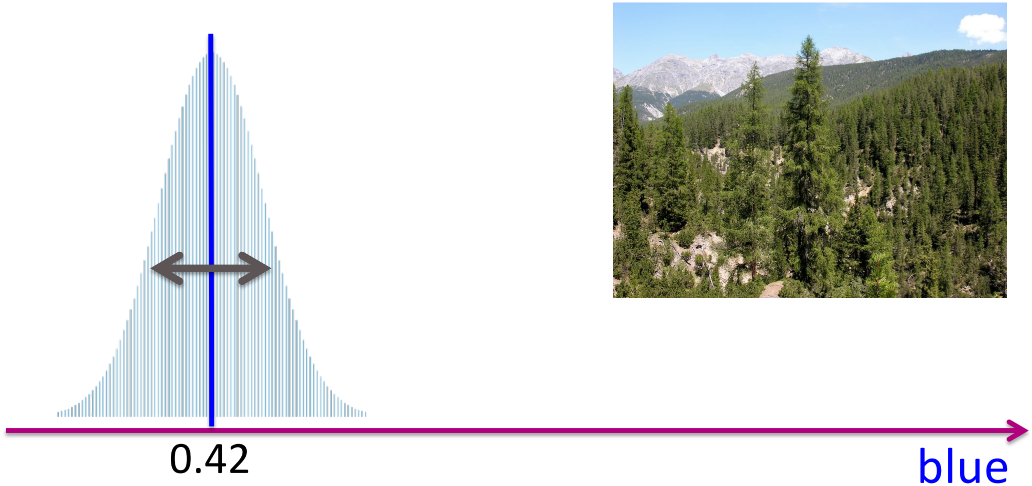
Let's look at just the blue dimension



©2017 Emily Fox

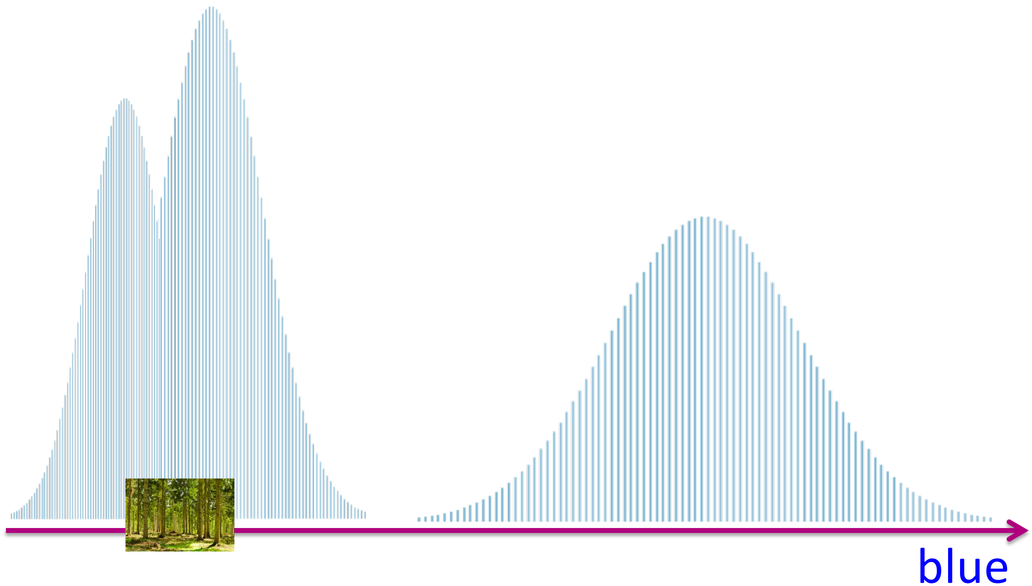
Distribution over all forest images

Let's look at just the blue dimension



©2017 Emily Fox

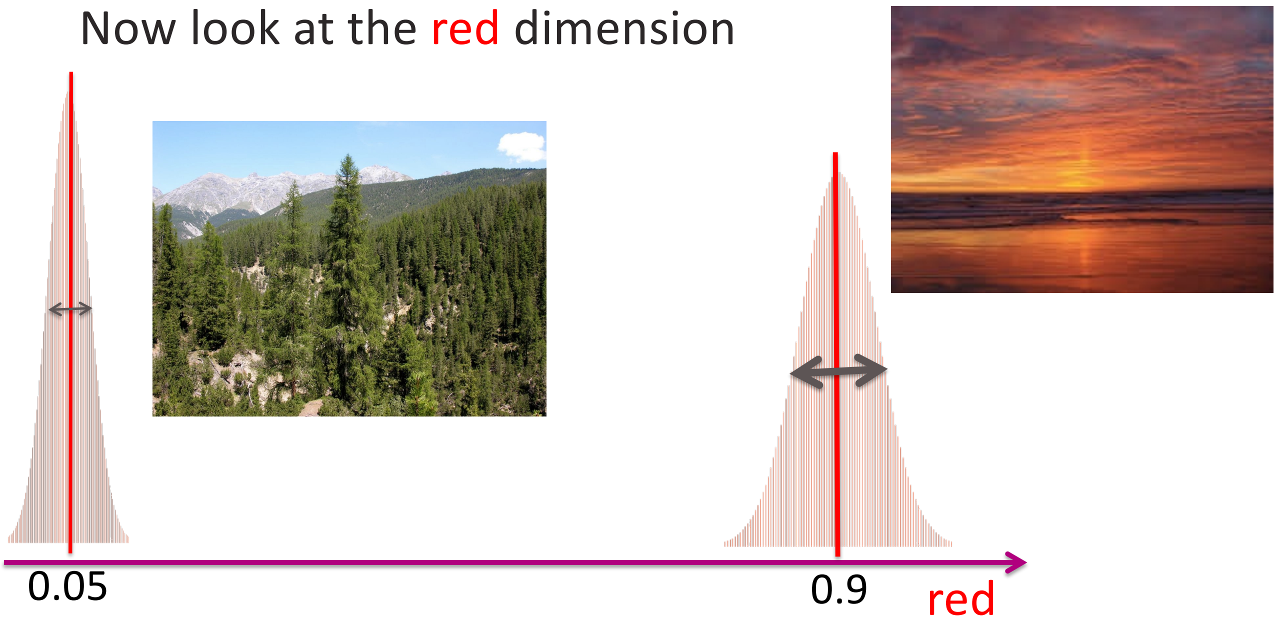
Distribution over all images



©2017 Emily Fox

Can be distinguished along other dim

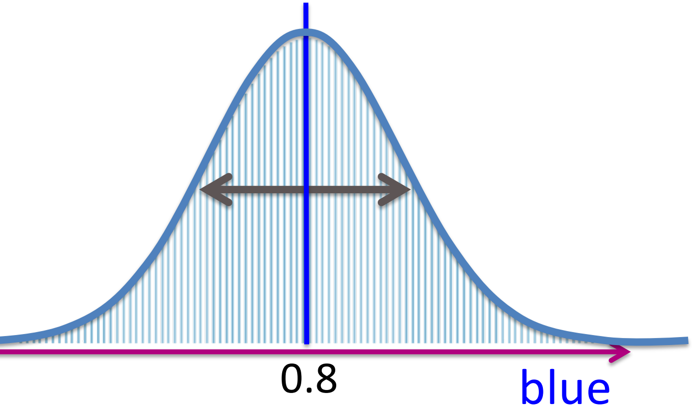
Now look at the red dimension



©2017 Emily Fox

Model for a given image type

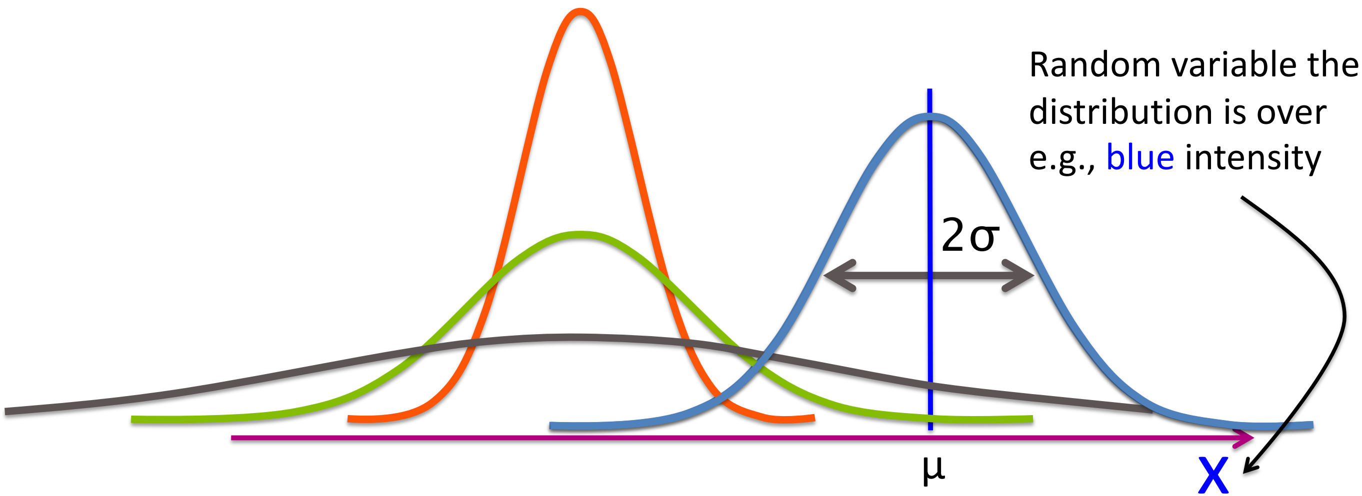
For each dim of the [R, G, B] vector, and each image type, assume a Gaussian distribution over color intensity



©2017 Emily Fox

1D Gaussians

Fully specified by mean μ and variance σ^2 (or st. dev. σ)



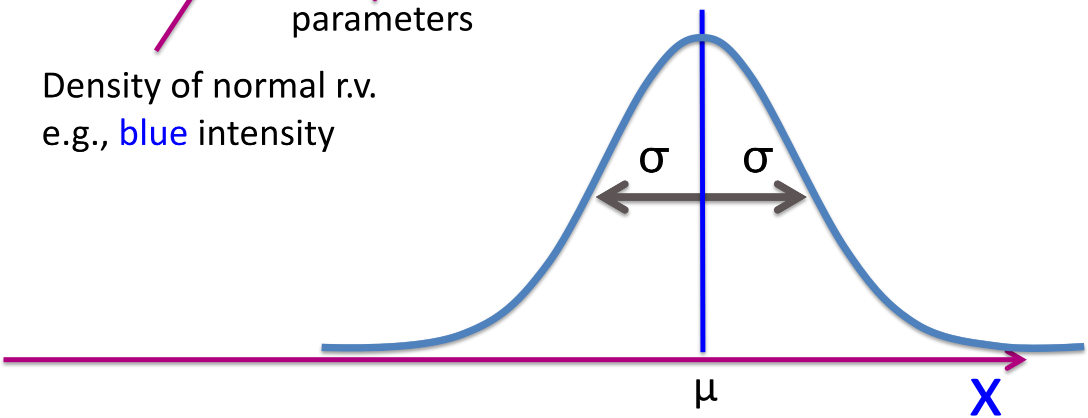
©2017 Emily Fox

Density of 1D Gaussian distribution

$$f(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

parameters

Density of normal r.v.
e.g., blue intensity



©2017 Emily Fox

Covariance Matrix

$$\sum_{(x,y)} (x - E(X))(y - E(Y)) \Pr(X=x, Y=y)$$

Problem 1 on your current homework

$$\mathbf{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])] = \mathbb{E}[\boldsymbol{XY}] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Y}]$$

$= \mathrm{Cov}(Y, X)$

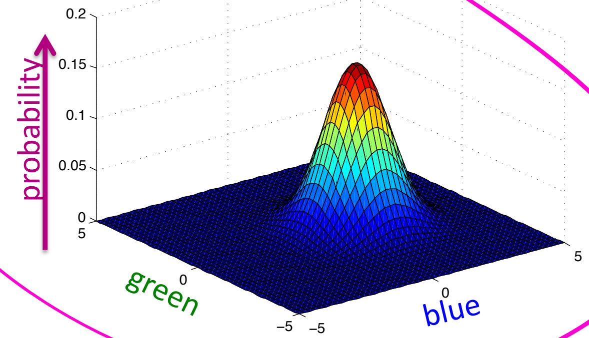
Given set of random variables $X_1, X_2, \dots, X_n$

The **"covariance matrix"**

$$\Sigma = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \mathrm{Cov}(X_i, X_j) & \vdots \\ \mathrm{Cov}(X_n, X_1) & \cdots & \mathrm{Cov}(X_n, X_n) \end{bmatrix}$$

Bivariate Gaussian (2 dimensions)

Fully specified by means μ and covariance matrix Σ



$$\mu = [\mu_{blue}, \mu_{green}]$$

$$\Sigma = \begin{pmatrix} \sigma_{blue}^2 & \sigma_{blue,green} \\ \sigma_{green,blue} & \sigma_{green}^2 \end{pmatrix}$$

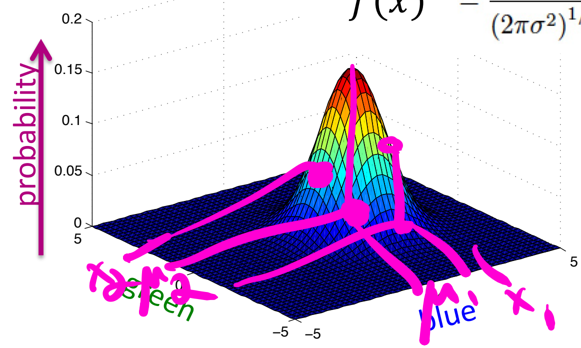
covariance determines orientation + spread

©2017 Emily Fox

Multivariate Gaussian (example – bivariate)

Fully specified by means μ and covariance matrix Σ

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mu = [\mu_{blue}, \mu_{green}]$$

$$\Sigma = \begin{pmatrix} \sigma_{blue}^2 & \sigma_{blue,green} \\ \sigma_{green,blue} & \sigma_{green}^2 \end{pmatrix}$$

covariance determines orientation + spread

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

©2017 Emily Fox

Multi-dimensional

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

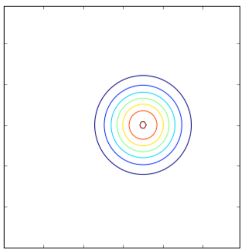
1-dimensional

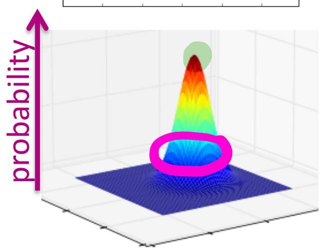
$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

©2017 Emily Fox

Covariance structure

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$





$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$$f(\mathbf{x}|\boldsymbol{\mu}=0, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right\}$$

(x_1, x_2), (x_1', x_2')

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}$$

$$(x_1 \quad x_2) \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{x_1^2}{\sigma^2} + \frac{x_2^2}{\sigma^2} = 1$$

$$x_1^2 + x_2^2 = \sigma^2$$

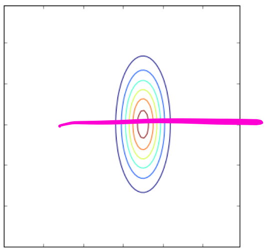
©2017 Emily Fox

Covariance structure

$$f\left(\mathbf{x} \mid \boldsymbol{\mu}=0, \boldsymbol{\Sigma}\right)=\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right\}$$

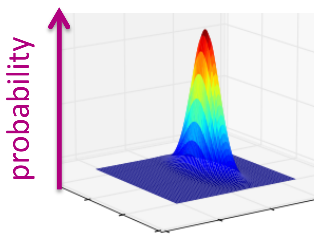
$$\Sigma = \begin{pmatrix} \sigma_B^2 & 0 \\ 0 & \sigma_G^2 \end{pmatrix}$$

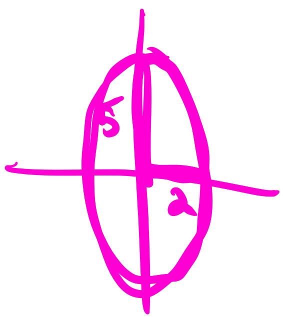
$$\begin{pmatrix} 2^2 & 0 \\ 0 & 5^2 \end{pmatrix} \qquad \Sigma^{-1} = \begin{pmatrix} \frac{1}{2^2} & 0 \\ 0 & \frac{1}{5^2} \end{pmatrix}$$



$$(x_1, x_2)\begin{pmatrix} \frac{1}{2^2} & \\ & \frac{1}{5^2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\frac{x_1^2}{2^2} + \frac{x_2^2}{5^2} = 1$$

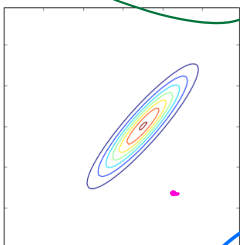


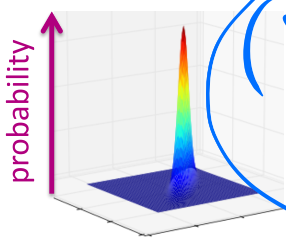


Covariance Structure

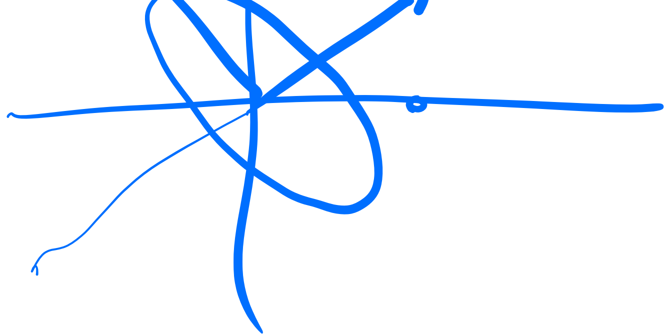
$$f\left(\mathbf{x}|\boldsymbol{\mu}=0,\boldsymbol{\Sigma}\right)=\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right\}$$

$$\Sigma = \begin{bmatrix} \sigma_B^{\ 2} & \sigma_{B,G} \\ \sigma_{G,B} & \sigma_G^{\ 2} \end{bmatrix}$$





©2017 Emily Fox



©2017 Emily Fox

Advanced...

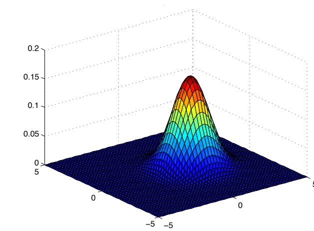
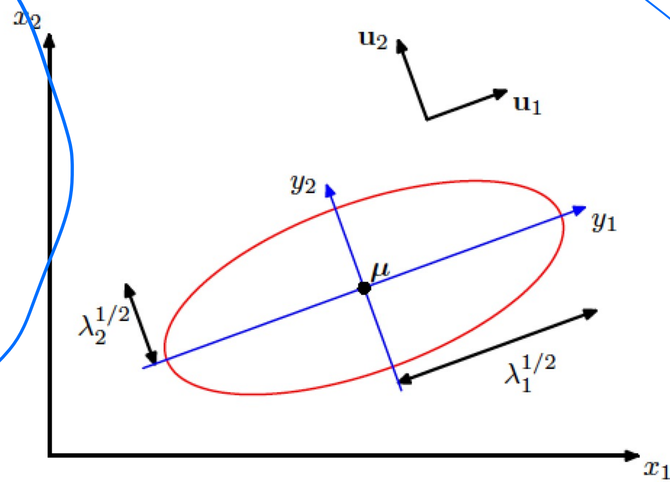


Figure 2.7 The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i.



Bishop

Multivariate Gaussian density

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

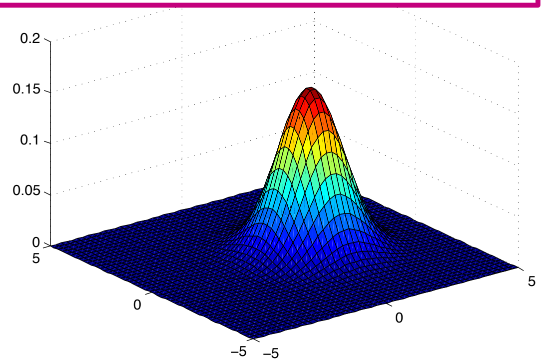
$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$

parameters

Random vector

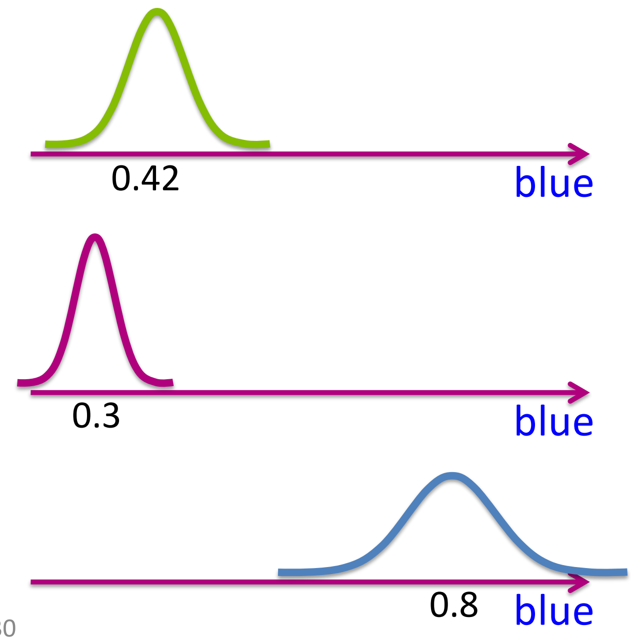
e.g., [R, G, B] intensities

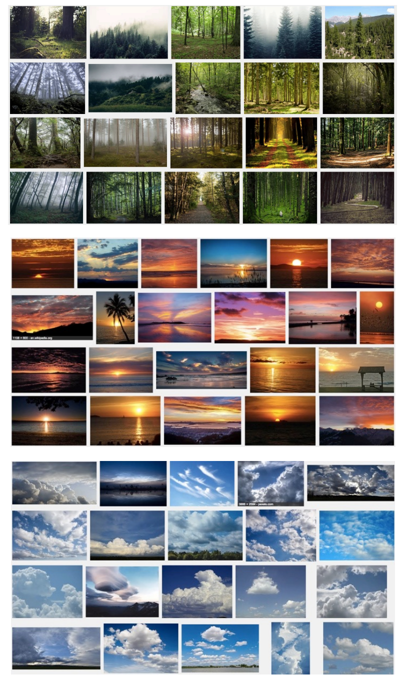


©2017 Emily Fox

Clustering using mixture model

Model as Gaussian per category/cluster





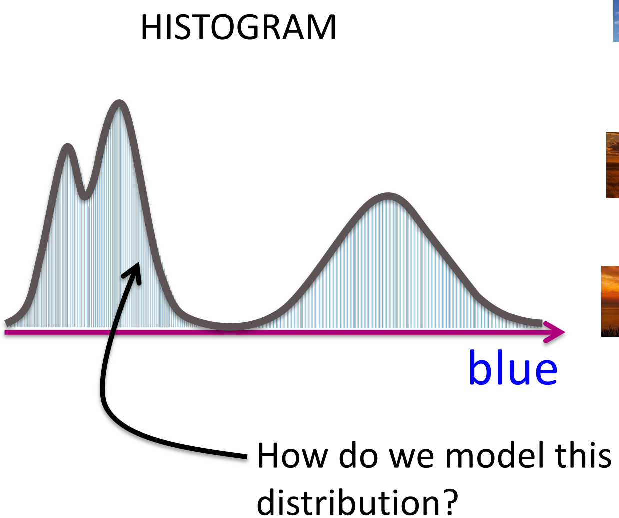
Forests

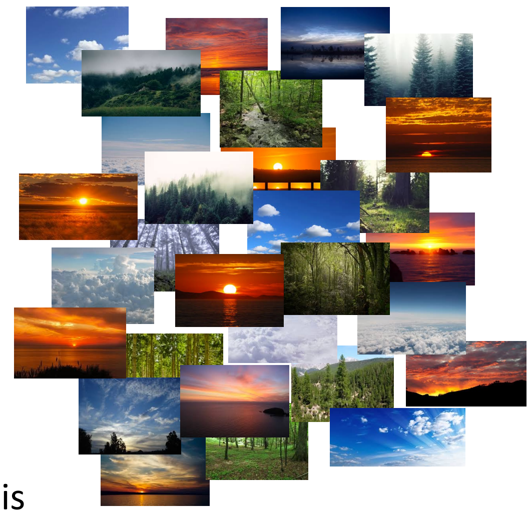
Sunsets

Clouds

©2017 Emily Fox

Jumble of unlabeled images





©2017 Emily Fox

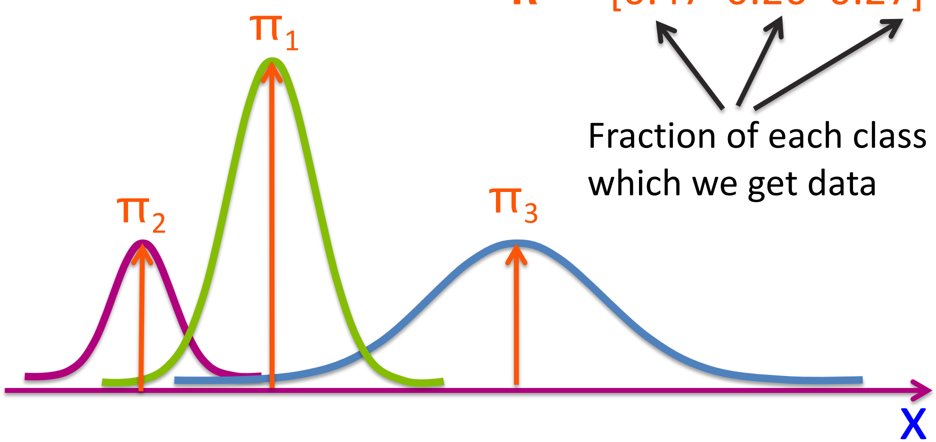
Combination of weighted Gaussians

Associate a weight π_k with each Gaussian component

$\pi_1 \quad \pi_2 \quad \pi_3$

$\pi = [0.47 \;\; 0.26 \;\; 0.27]$

Fraction of each class in world from which we get data



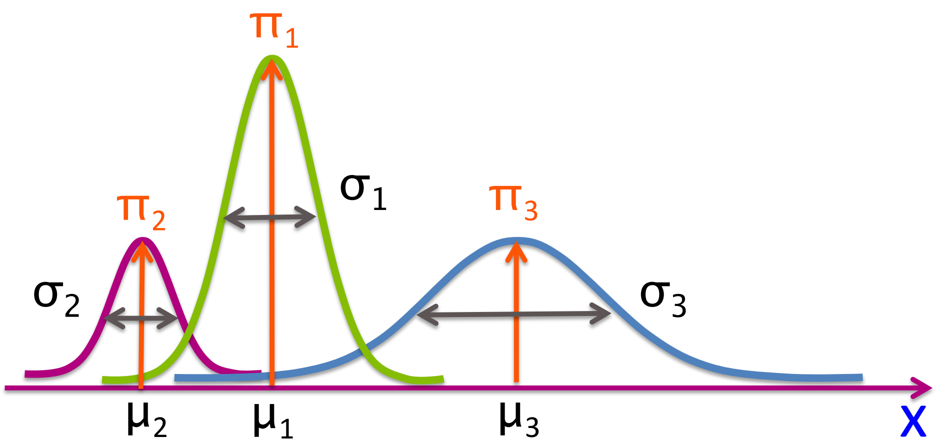
$0 \leq \pi_k \leq 1$

$\sum_{k=1}^{K} \pi_k = 1$

©2017 Emily Fox

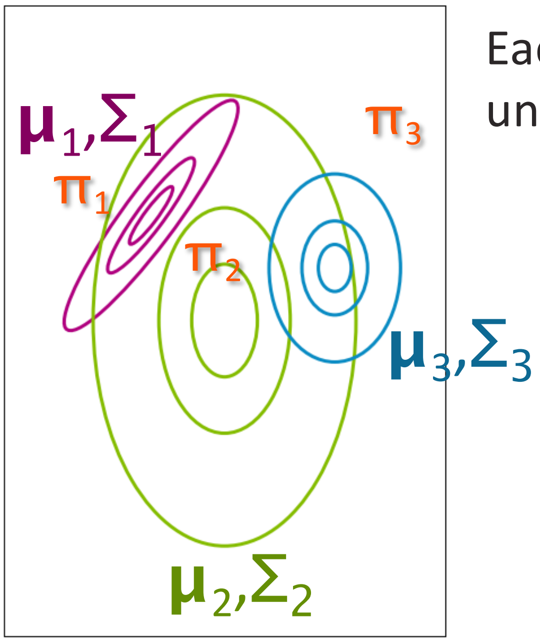
Mixture of Gaussians (1D)

Each mixture component represents a unique cluster specified by: $\{\pi_k, \mu_k, \sigma_k^2\}$



©2017 Emily Fox

Mixture of Gaussians (general)



Each mixture component represents a unique cluster specified by:

$$\{\pi_k, \mu_k, \Sigma_k\}$$

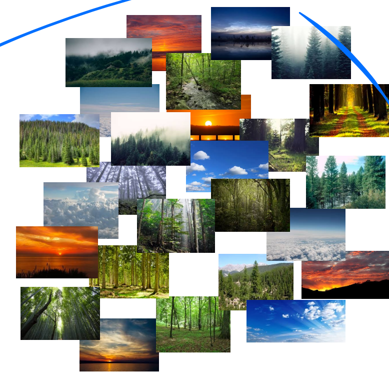
©2017 Emily Fox

Mixture model

- K clusters, defined by the following **unknown** parameters

$$\Theta = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^{k}$$

$$\sum_{j=1}^{k} \pi_j = 1.$$

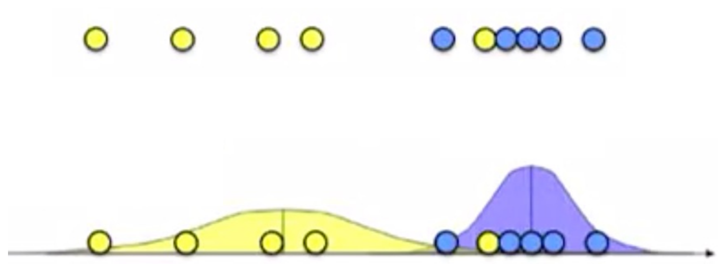


- Problem: Assume that the data comes from such a distribution, and recover the parameters of the distribution (e.g. MLE)
- Determine, for each point, the likelihood of it belonging to cluster j, for each j.

©2017 Emily Fox

Two1-D Gaussians, with unknown mean and variance

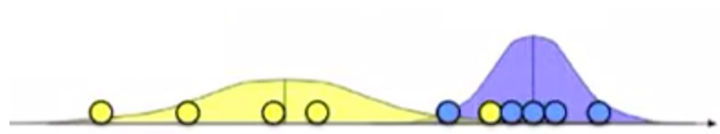
- Easy if know the source of each data point.



Images dues to Victor Lavrenko

Two1-D Gaussians, with unknown mean and variance

- Easy if know the source of each data point.



- What if we don’t know the source?

Images dues to Victor Lavrenko

Mixture model

- K clusters, defined by the following parameters

$$\Theta = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^{k} \qquad \sum_{j=1}^{k} \pi_j = 1.$$

- Problem: Assume that the data comes from such a distribution, and recover the parameters of the distribution.
- Determine, for each point, the likelihood of it belonging to cluster j, for each j.
- **PROBLEM: no closed form solution**

©2017 Emily Fox

Expectation Maximization Algorithm

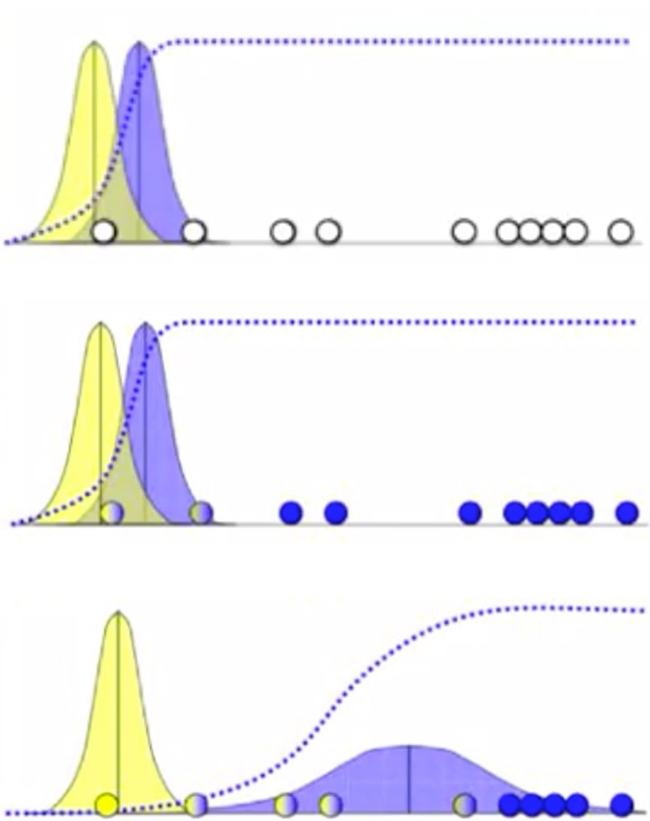
Two step approach based on following observation

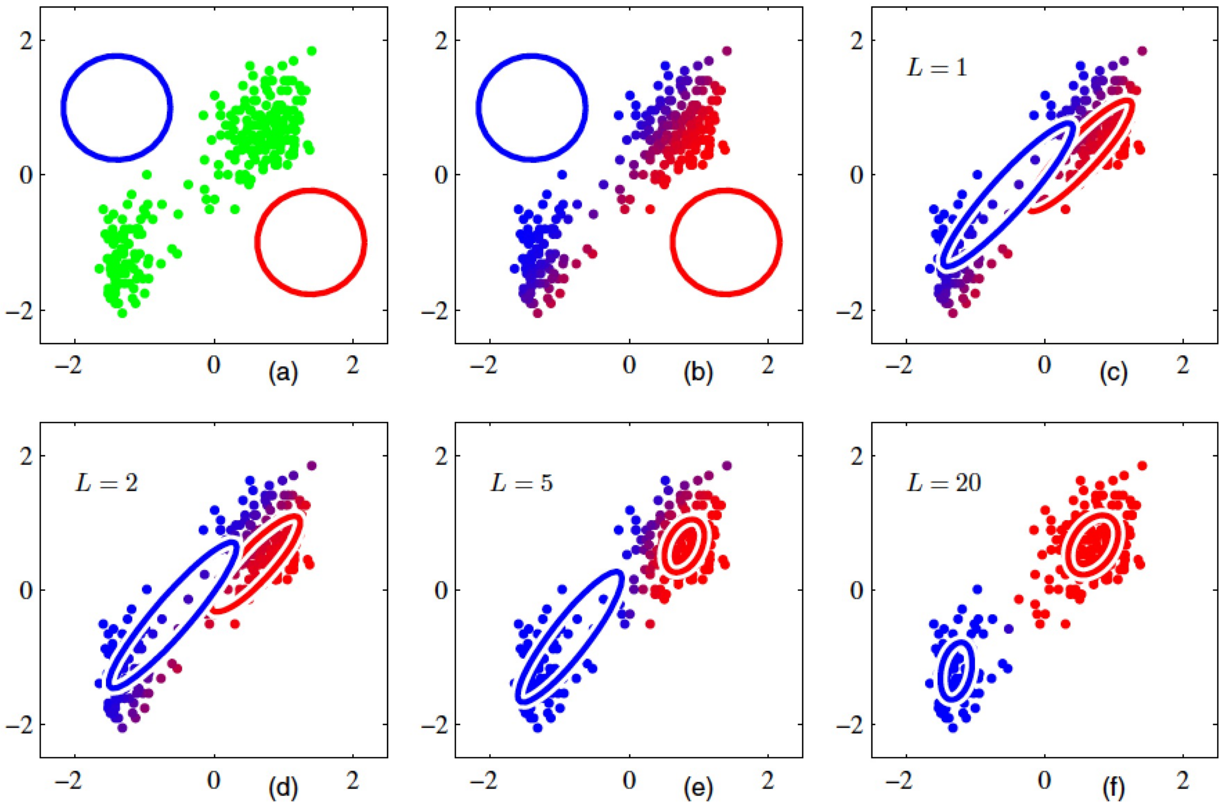
- If we knew which cluster each sample was from, we could estimate all the parameters.

- If we knew all the parameters we could estimate the chance each point came from each cluster.

- EM is an iterative algorithm that alternates between these two steps.

©2017 Emily Fox





©2017 Emily Fox