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 ☺

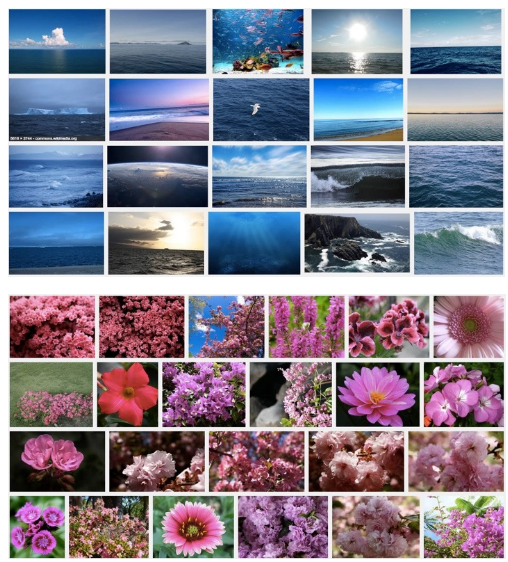
Goal for today

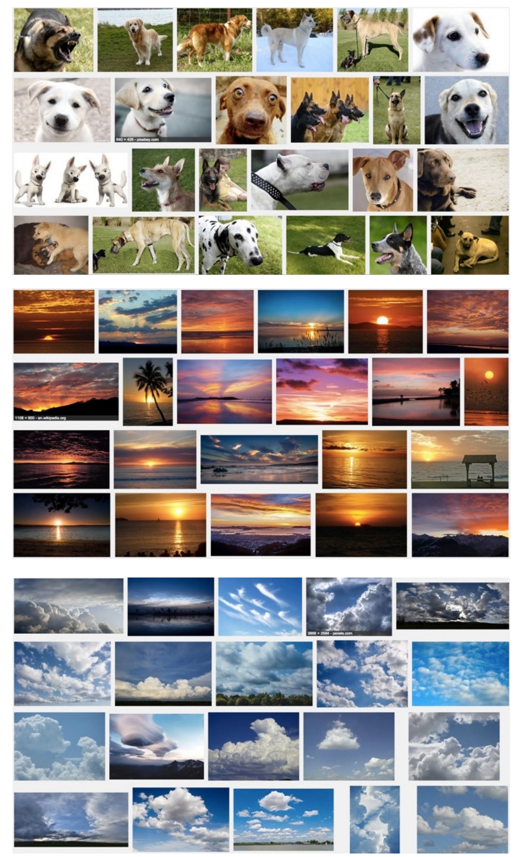
- Introduce you to a fundamental machine learning problem: clustering.
- Give you a very gentle introduction to multivariate Gaussian distributions.

Motivating application: Clustering images

Discover groups of similar images

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





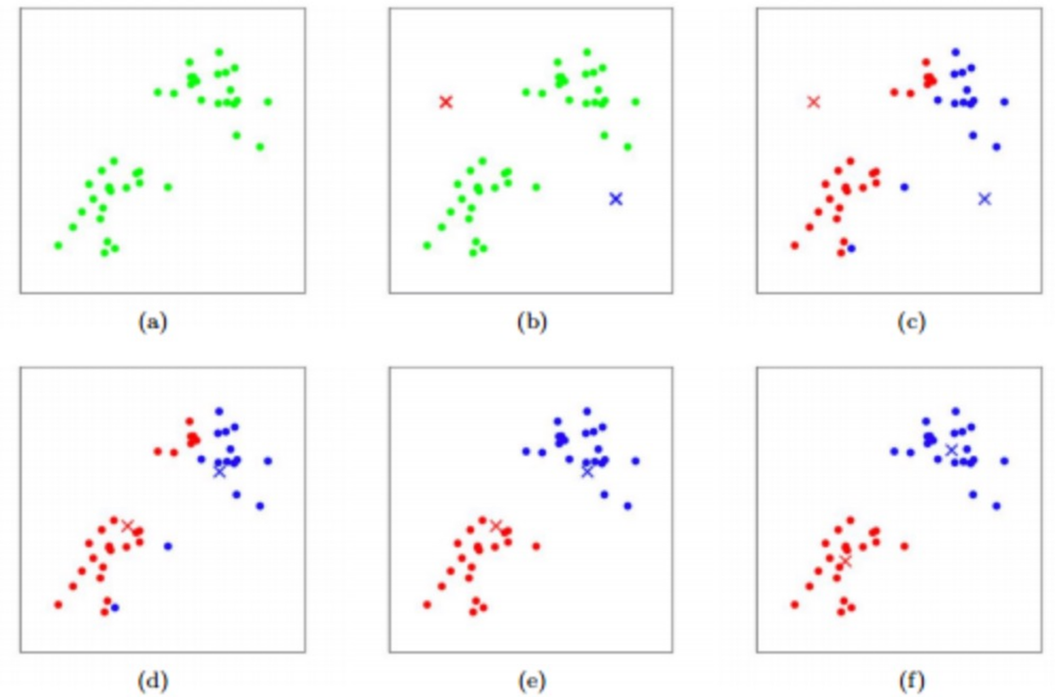
Another example: clustering documents

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

So how are these data represented?

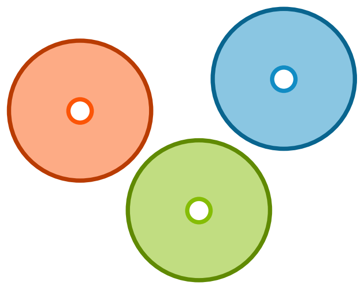
- As points in d-dimensional space (where d is typically large).

How to approach clustering? One way: k-means



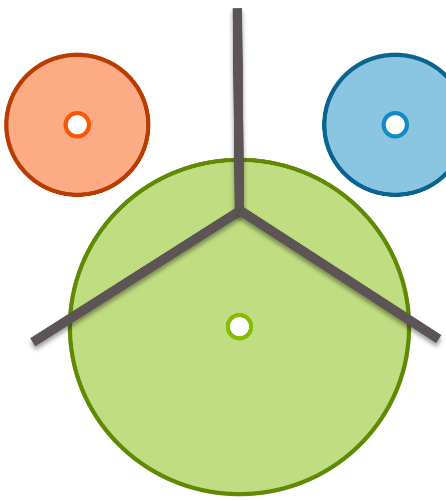
Only center matters
Not cluster shapes

Equivalent to assuming spherically symmetric clusters

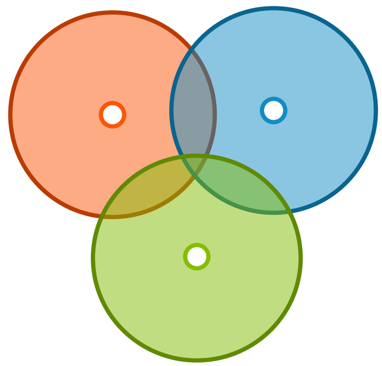


5 K-means picture From Stanford CS 221

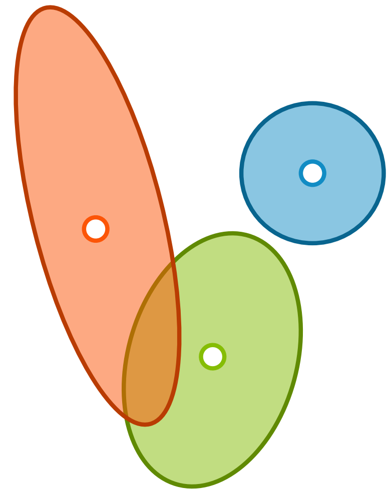
Failure modes of k-means clustering



disparate cluster sizes



overlapping clusters

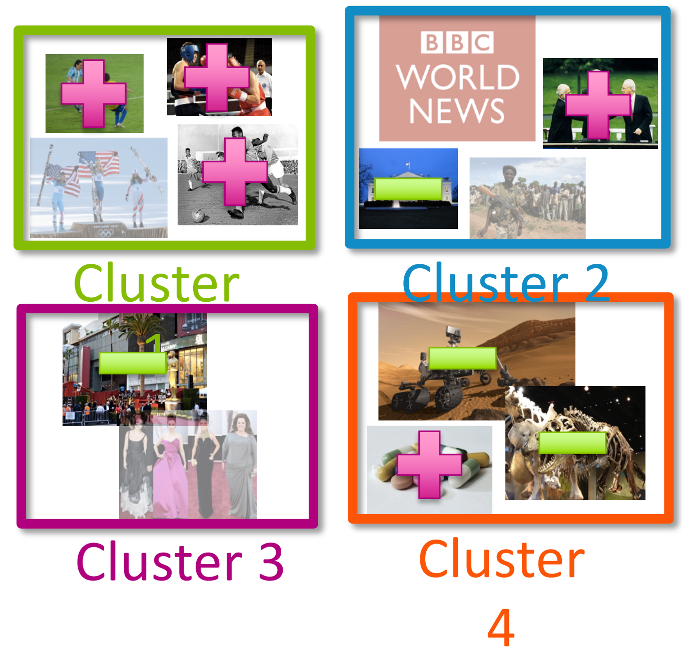


different
shaped/oriented
clusters

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

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.

Background: Multivariate Gaussian distributions

Overly simple image representation

Consider average red, green, blue pixel intensities

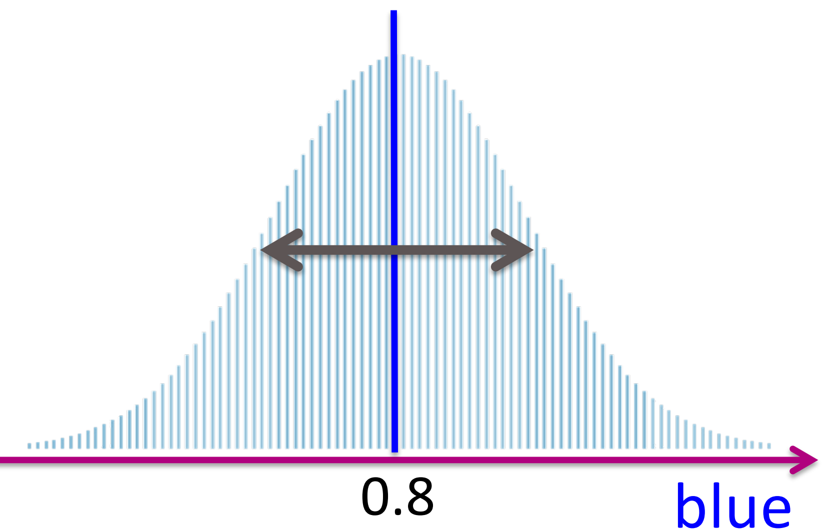
[R = 0.05, G = 0.7, B = 0.9]

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

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

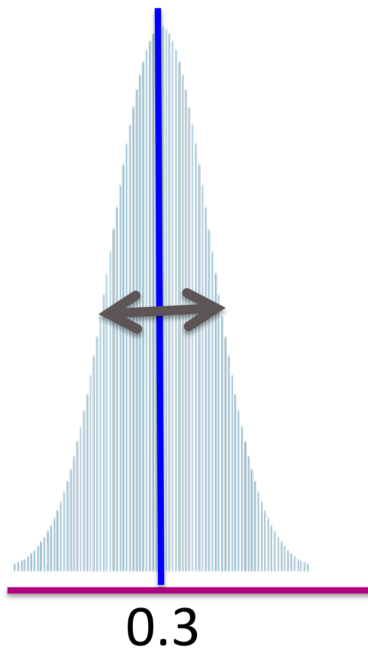
Distribution over all cloud images

Let's look at just the blue dimension



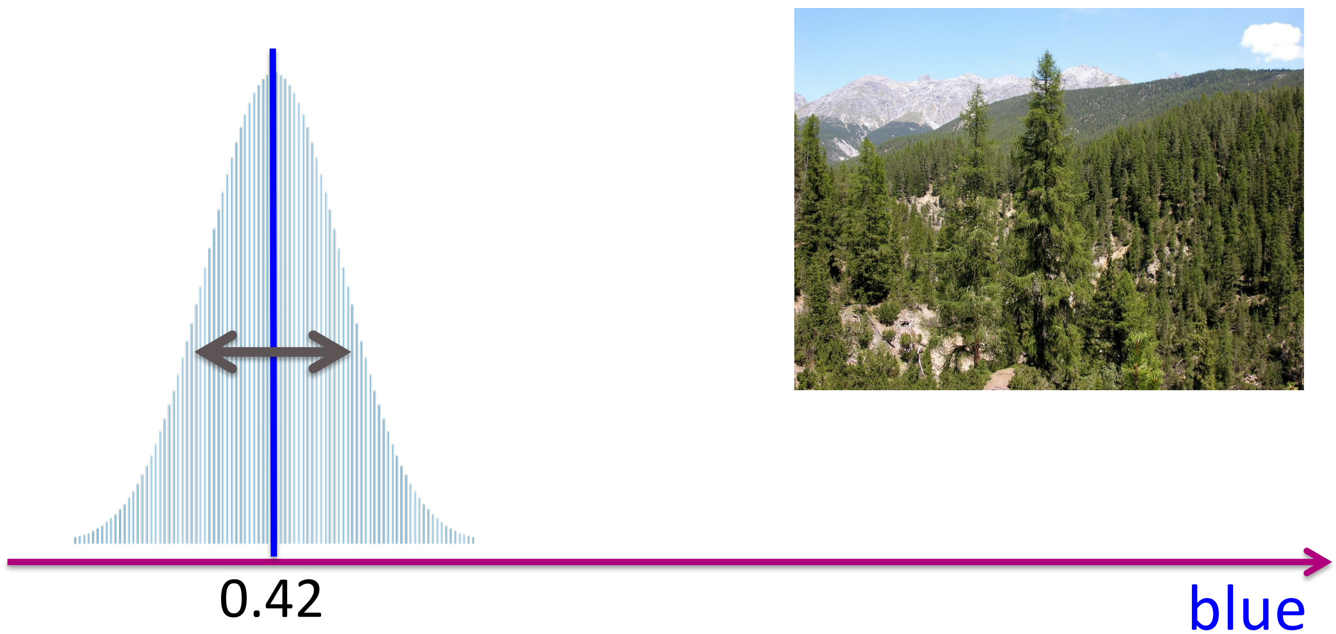
Distribution over all sunset images

Let's look at just the blue dimension

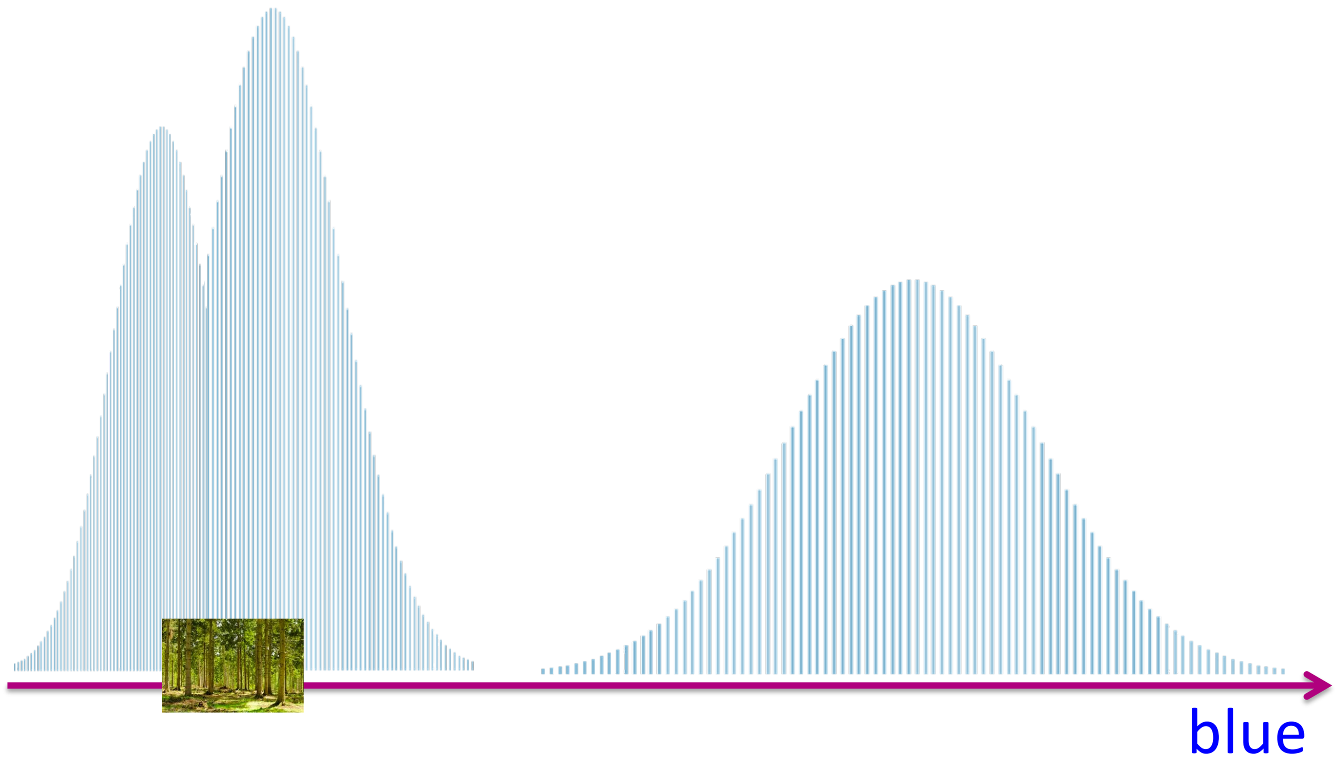


Distribution over all forest images

Let's look at just the blue dimension



Distribution over all images

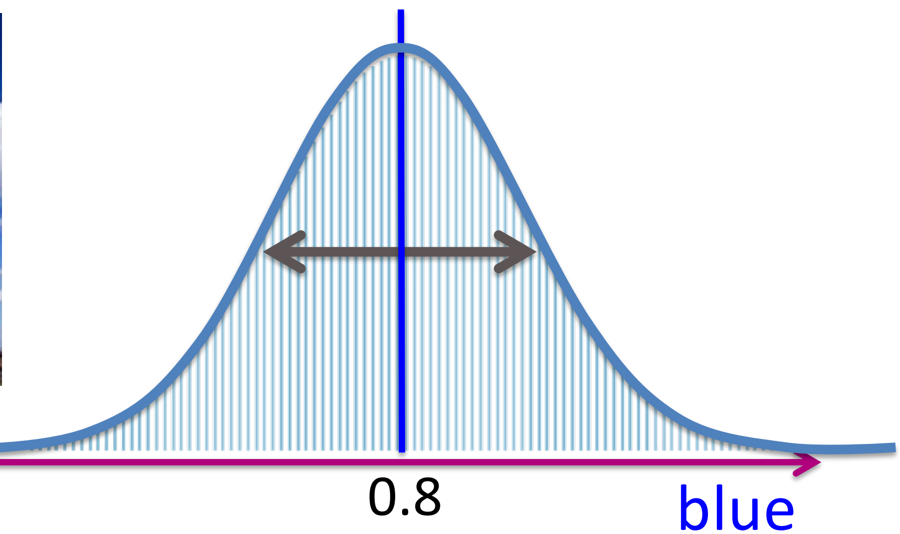


Can be distinguished along other dim

Now look at the red dimension

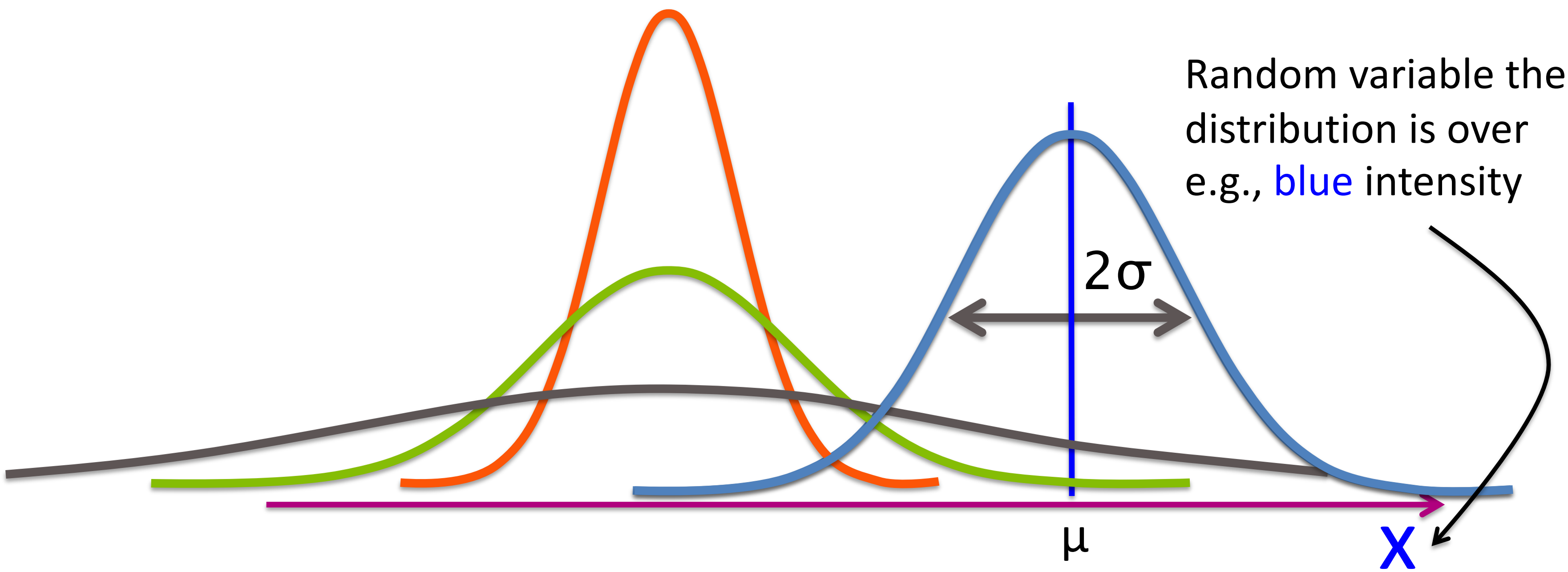
Model for a given image type

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



1D Gaussians

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

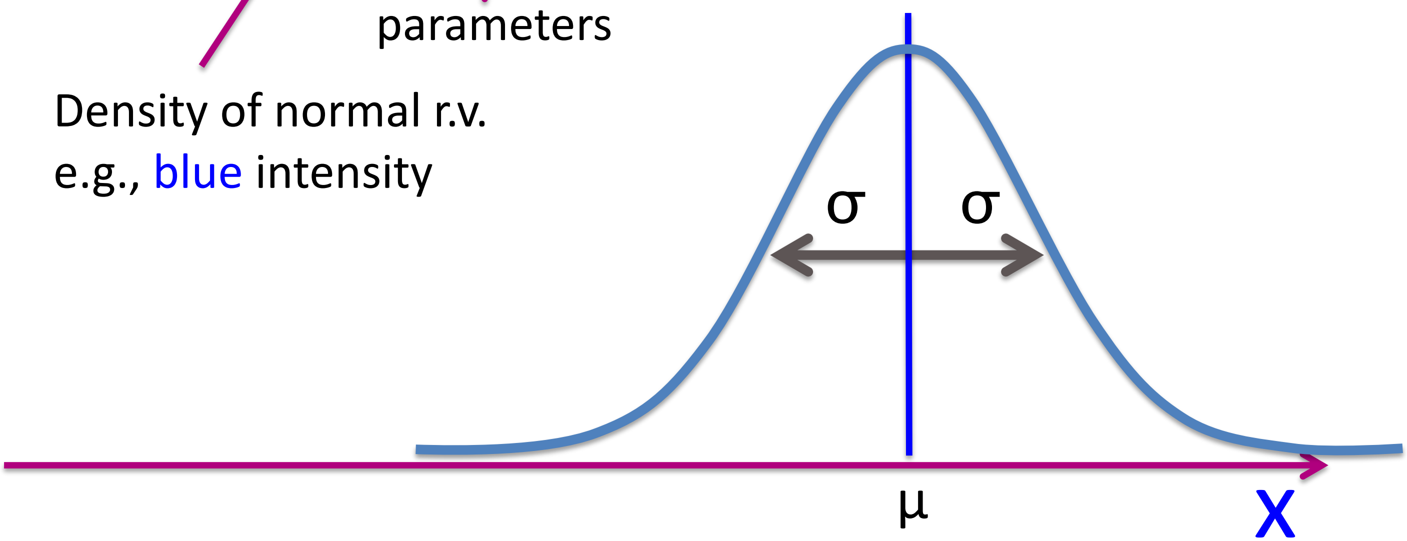


Density of 1D Gaussian distribution

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

parameters

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



Covariance Matrix

Problem 1 on your current homework

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

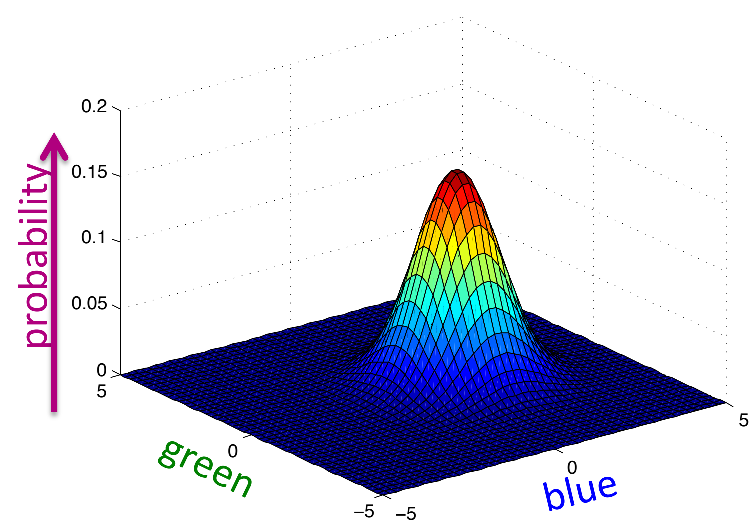
Given set of random variables $X_1, X_2, \dots, X_n$

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



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

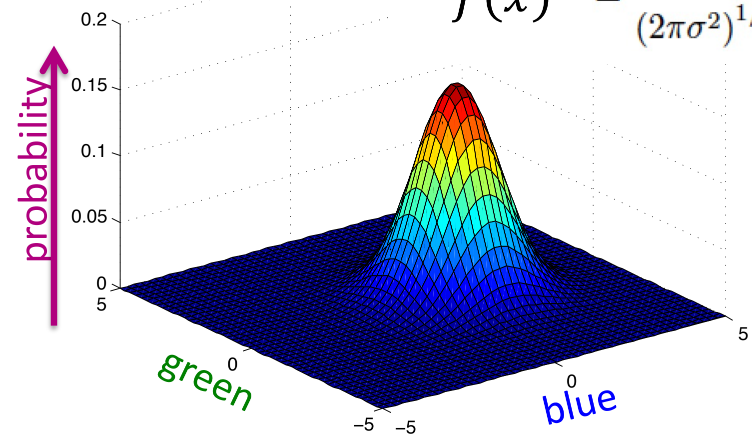
$$\Sigma = \begin{pmatrix} \sigma_{blue}^2 & \sigma_{blue,green} \\ \sigma_{green,blue} & \sigma_{green}^2 \end{pmatrix}$$

covariance determines orientation + spread

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

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

1-dimensional

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

Multi-dimensional

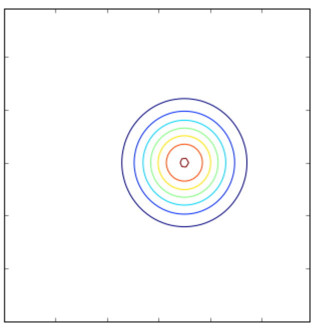
$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

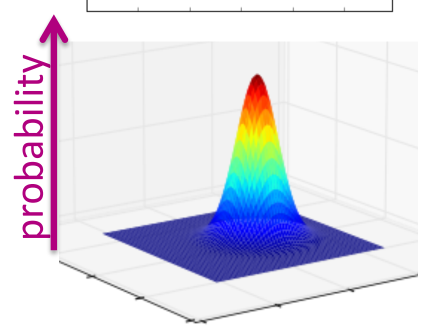
Covariance structure

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

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

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

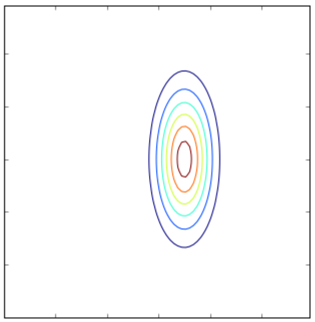


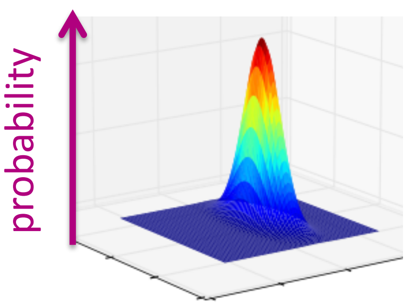


Covariance structure

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

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

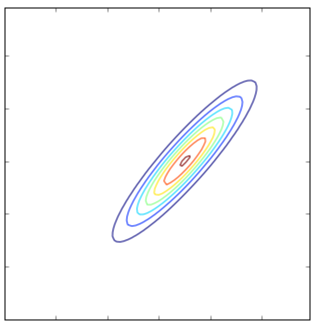


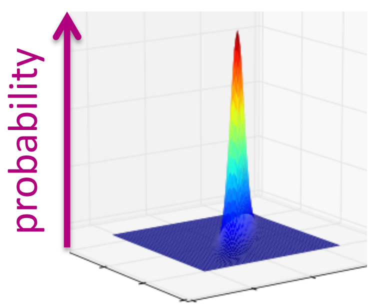


Covariance Structure

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

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





Advanced…

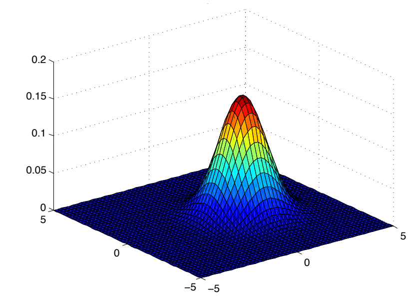
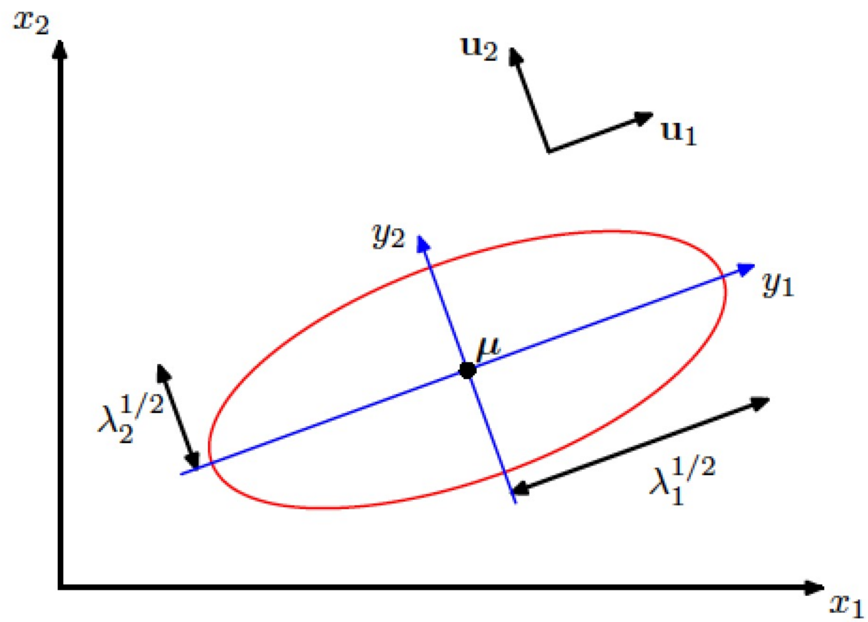


Figure 2.7 The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors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



Multivariate Gaussian density

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

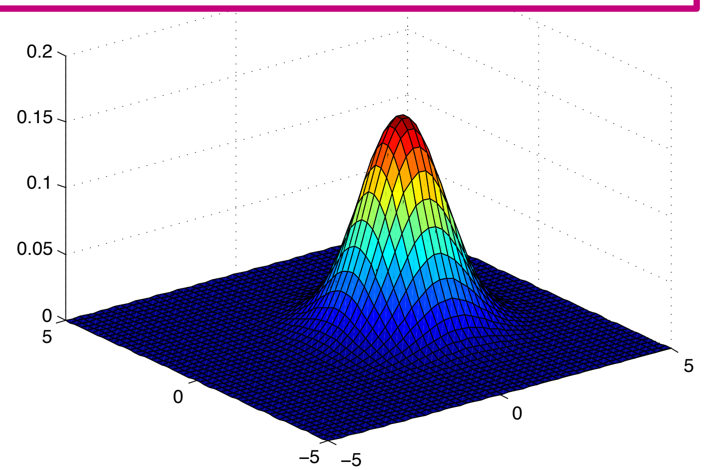
$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

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

parameters

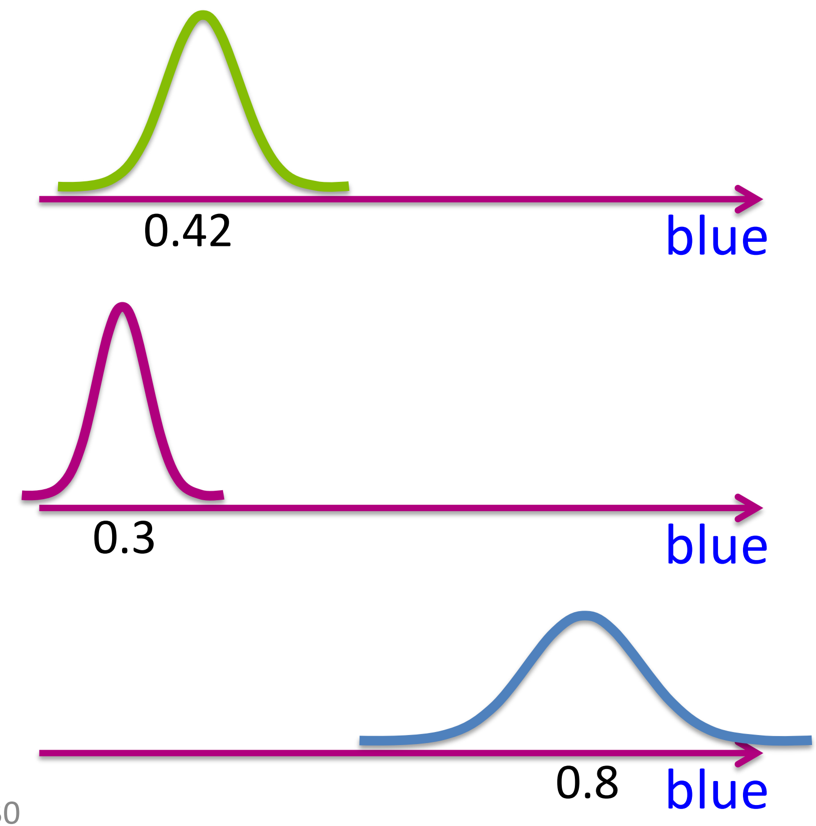
Random vector

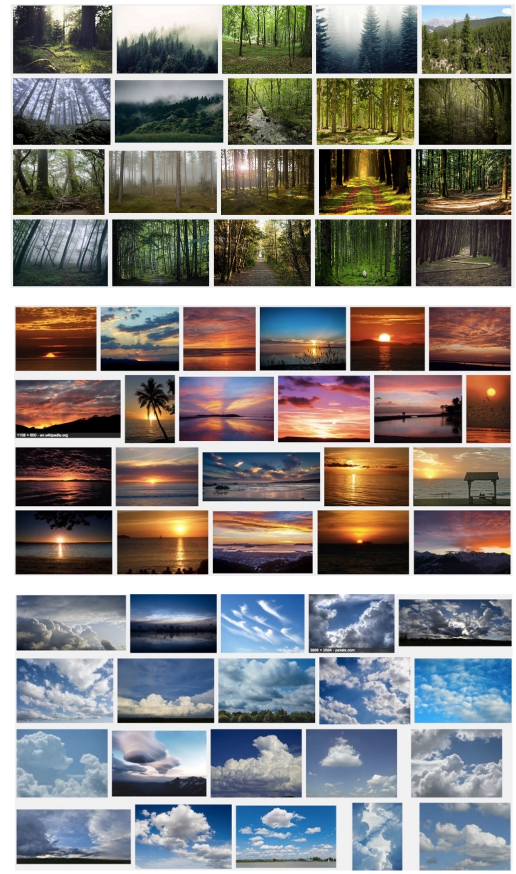
e.g., [R, G, B] intensities



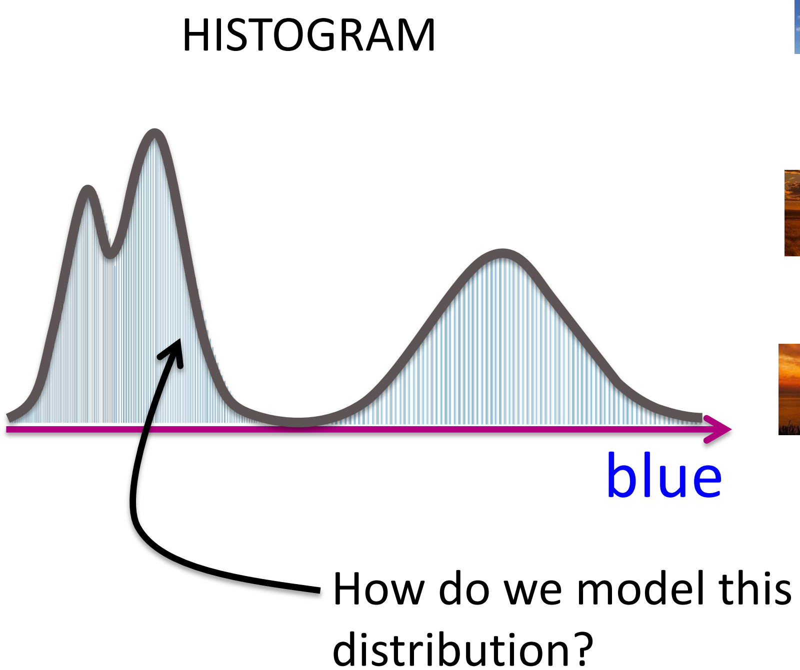
Clustering using mixture model

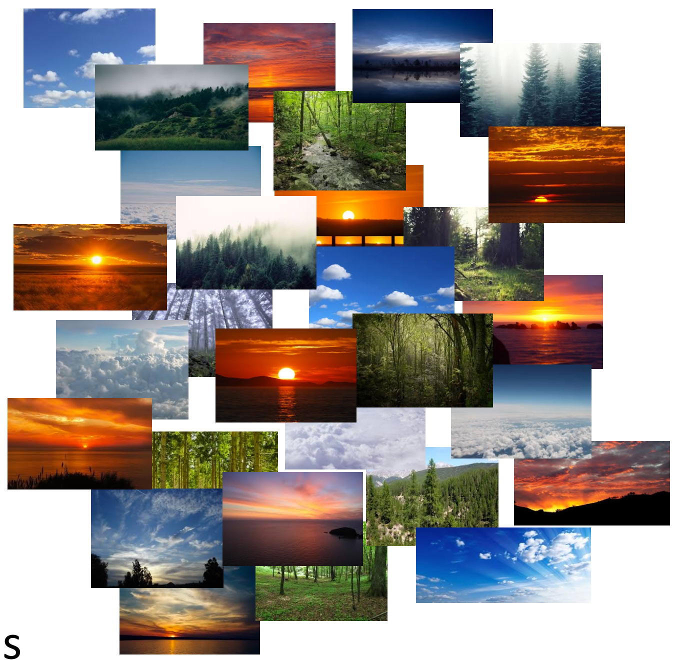
Model as Gaussian per category/cluster





Jumble of unlabeled images



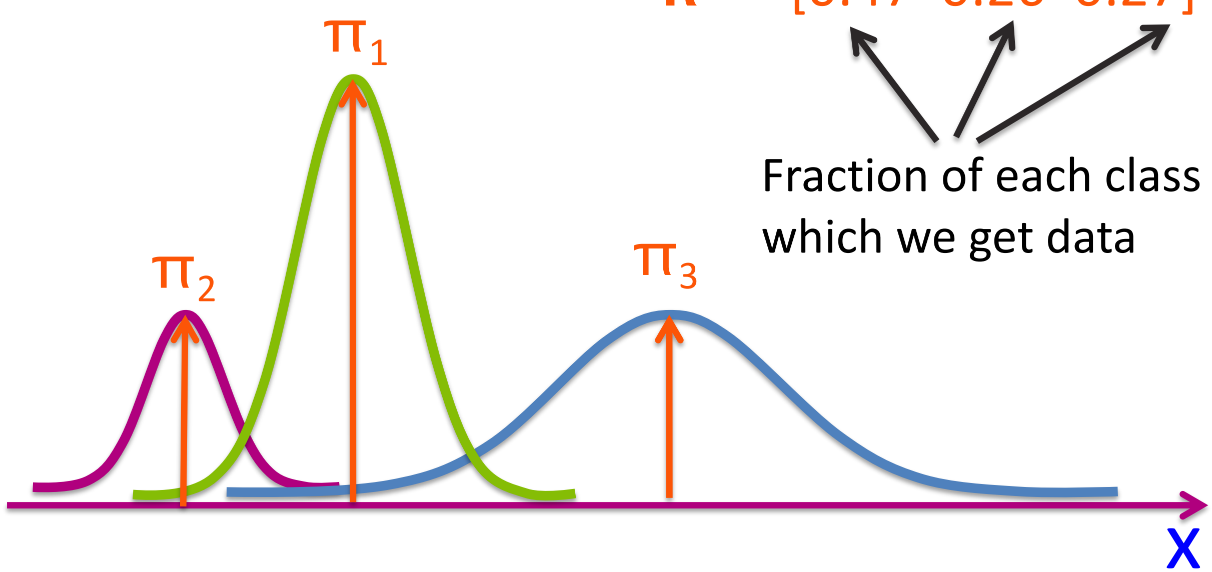


Combination of weighted Gaussians

Associate a weight π_k with each Gaussian component

$$\pi = \begin{matrix} \pi_1 & \pi_2 & \pi_3 \\ [0.47 & 0.26 & 0.27] \end{matrix}$$

Fraction of each class in world from which we get data

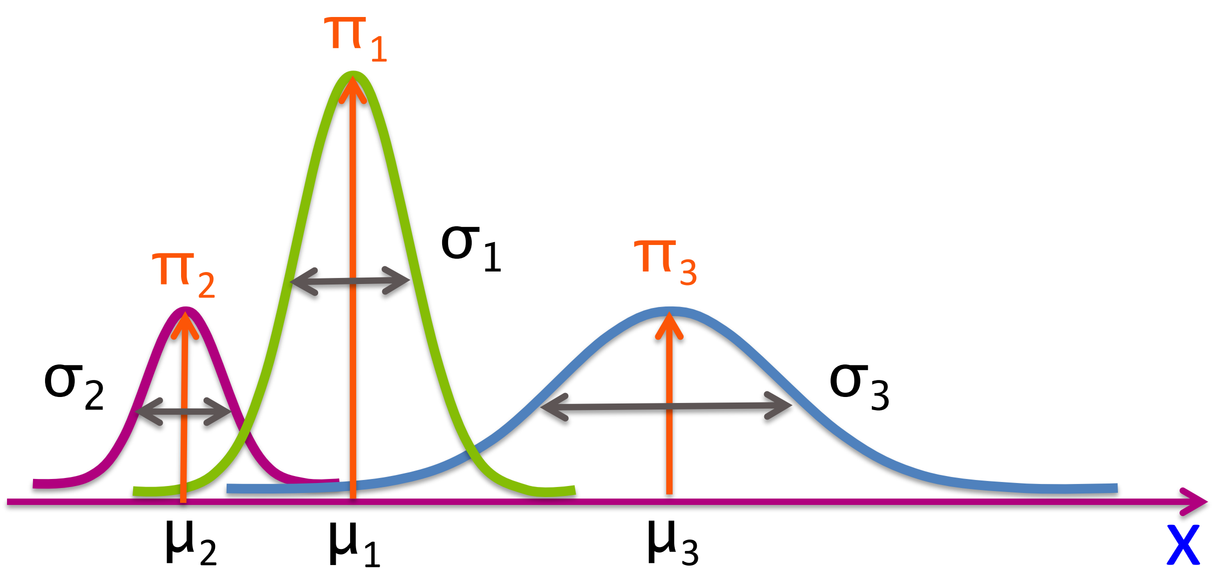


$$0 \leq \pi_k \leq 1$$

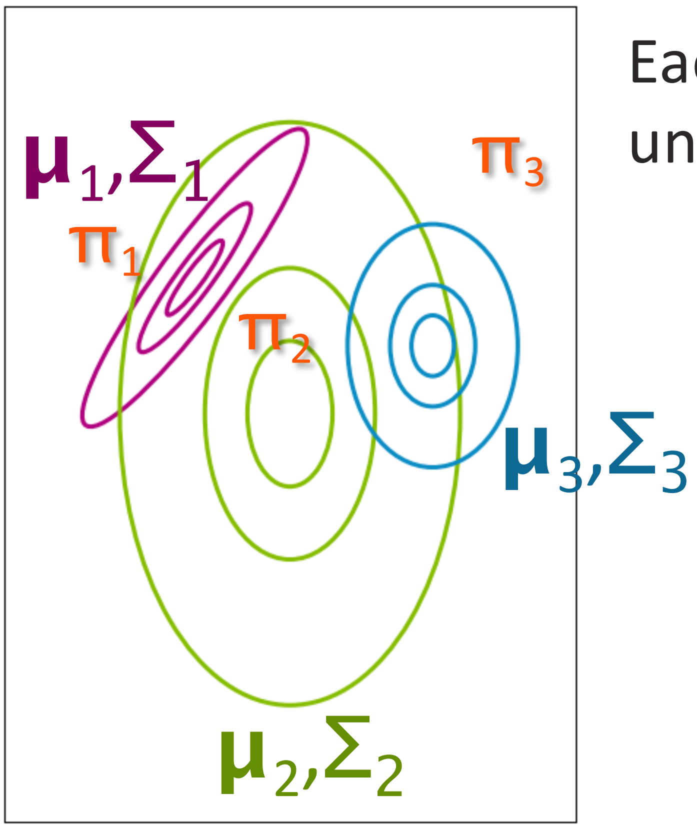
$$\sum_{k=1}^{K} \pi_k = 1$$

Mixture of Gaussians (1D)

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



Mixture of Gaussians (general)



Each mixture component represents a unique cluster specified by:

$$\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$$

Mixture model

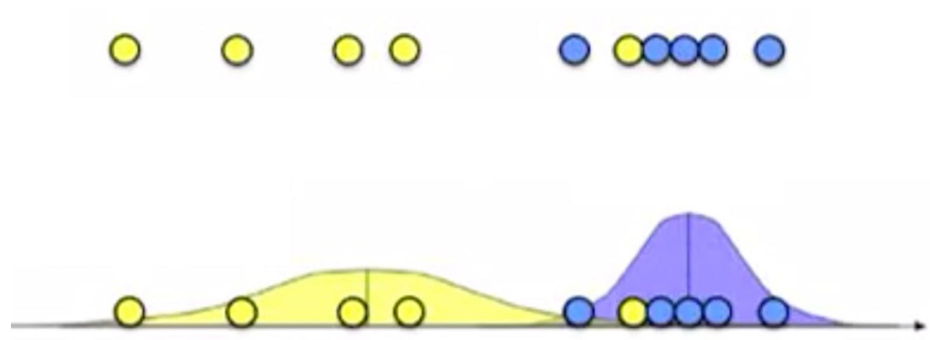
- K clusters, defined by the following **unknown** parameters

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

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

Two1-D Gaussians, with unknown mean and variance

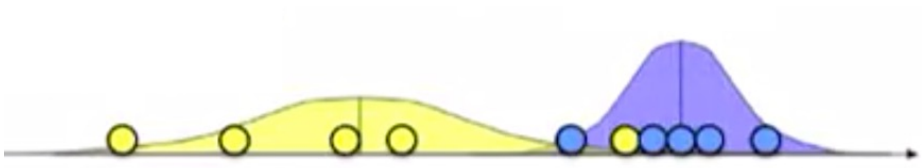
- Easy if know the source of each data point.



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

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

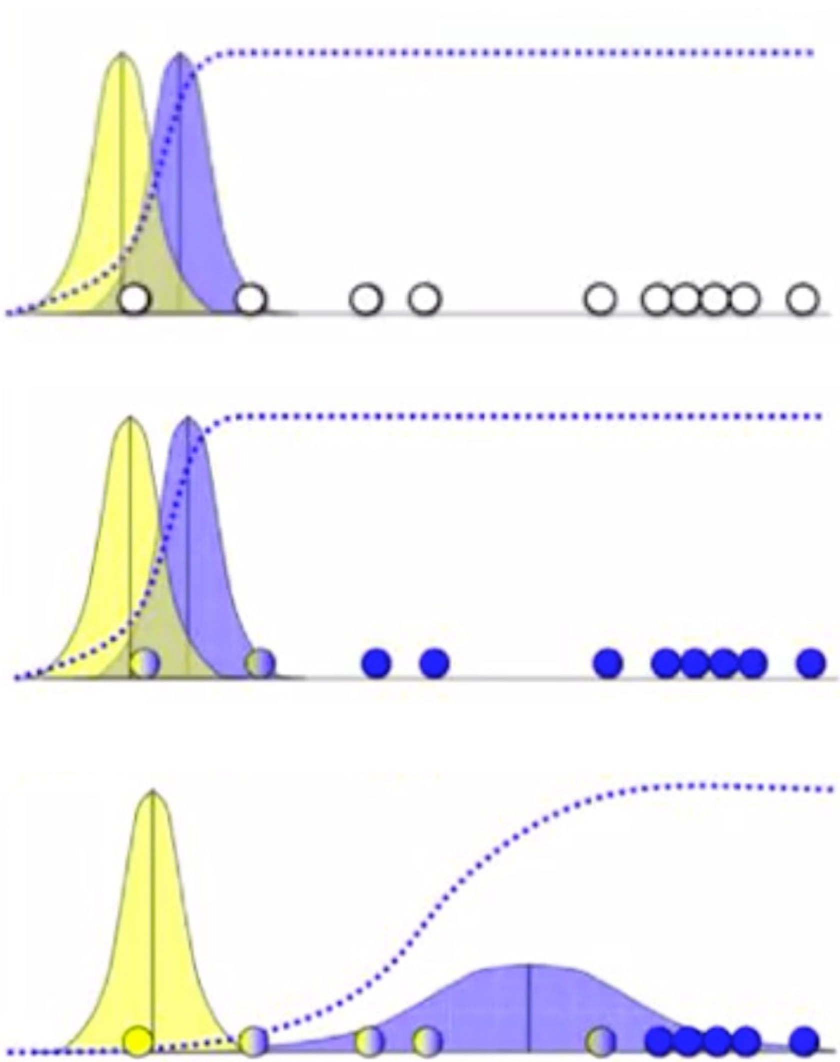
$$\sum_{j=1}^{k} \pi_j = 1.$$

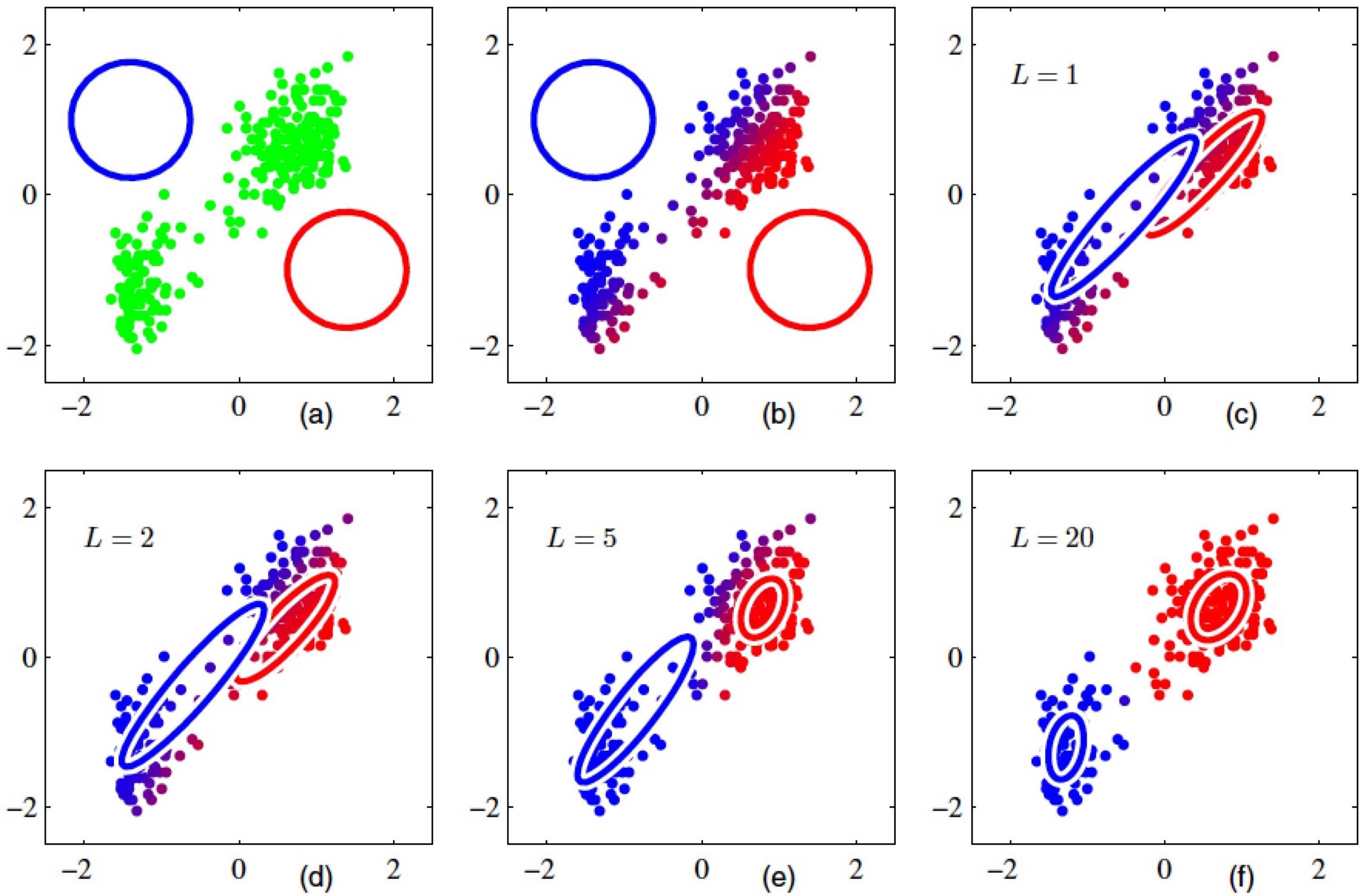
- Problem: Assume that the data comes from such a distribution, and recover the parameters of the distribution.
- Determine, for each point, the likelihood of it belonging to cluster j, for each j.
- **PROBLEM: no closed form solution**

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.





(a)
(b)
(c)
(d)
(e)
(f)
$L = 1$
$L = 2$
$L = 5$
$L = 20$