Mixture Models and EM:
Model-Based Clustering

CSE 446: Machine Learning
Slides by Emily Fox
University of Washington
May 22, 2019

©2017 Emily Fox

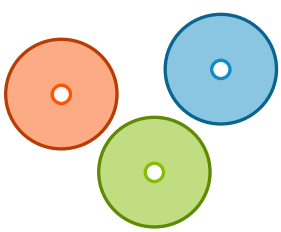
Limitations of k-means

Assigns observations to closest cluster center

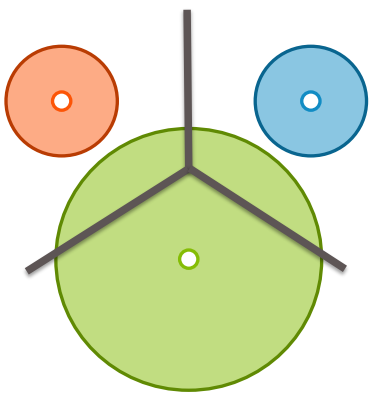
$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

Only center matters
Not cluster shapes

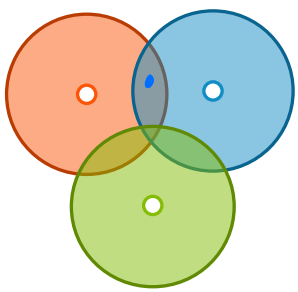
Equivalent to assuming
spherically symmetric clusters



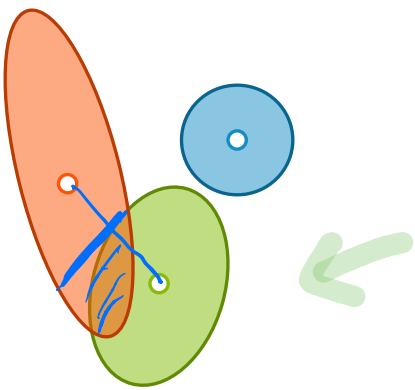
Failure modes of k-means



disparate cluster sizes



overlapping clusters

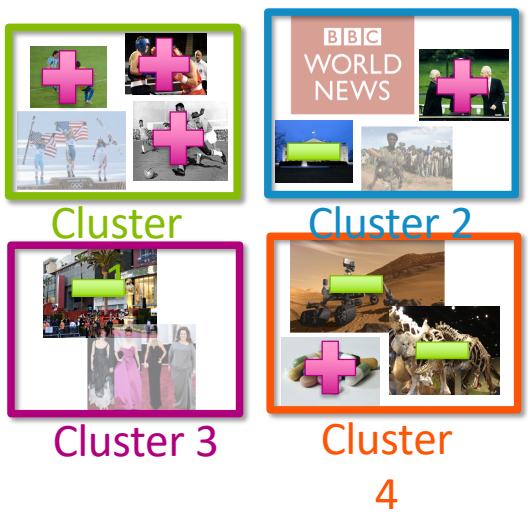


different shaped/oriented clusters

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

- Accounts for cluster shapes not just centers

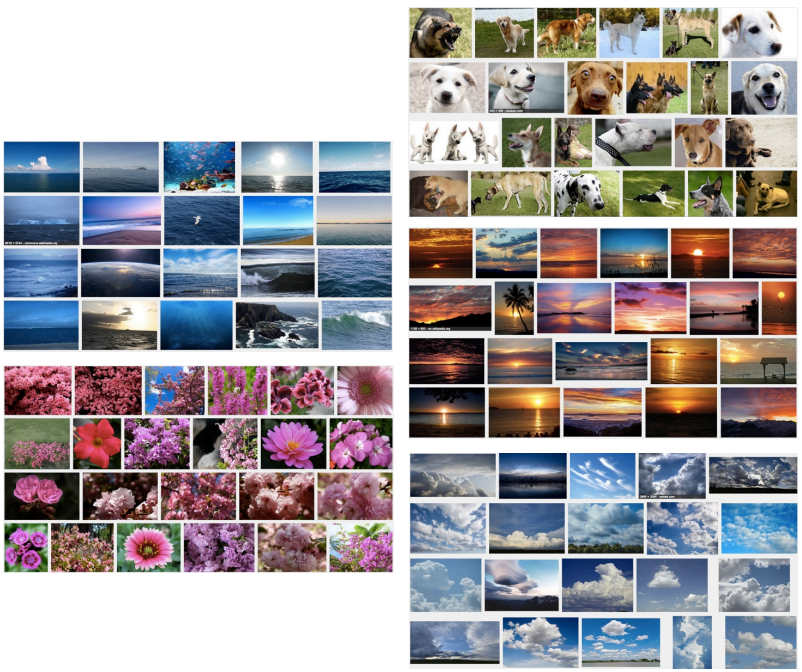


Mixture models

Motivating application: Clustering images

Discover groups of similar images

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

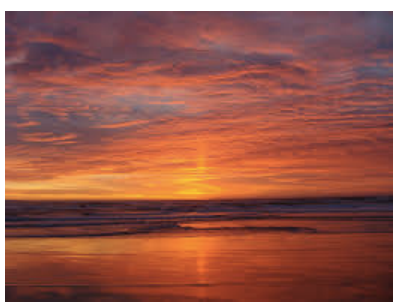


©2017 Emily Fox

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

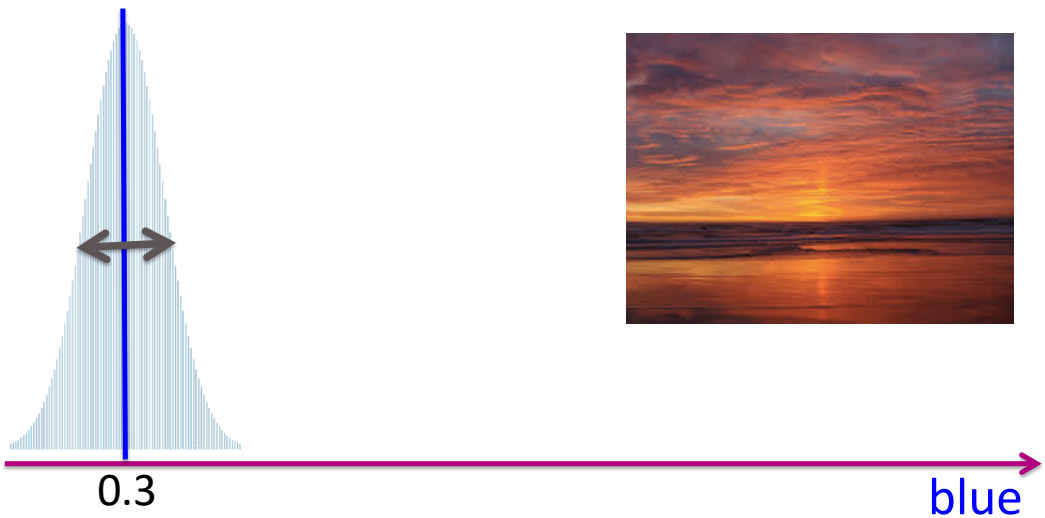
Distribution over all cloud images

Let's look at just the blue dimension

©2017 Emily Fox

Distribution over all sunset images

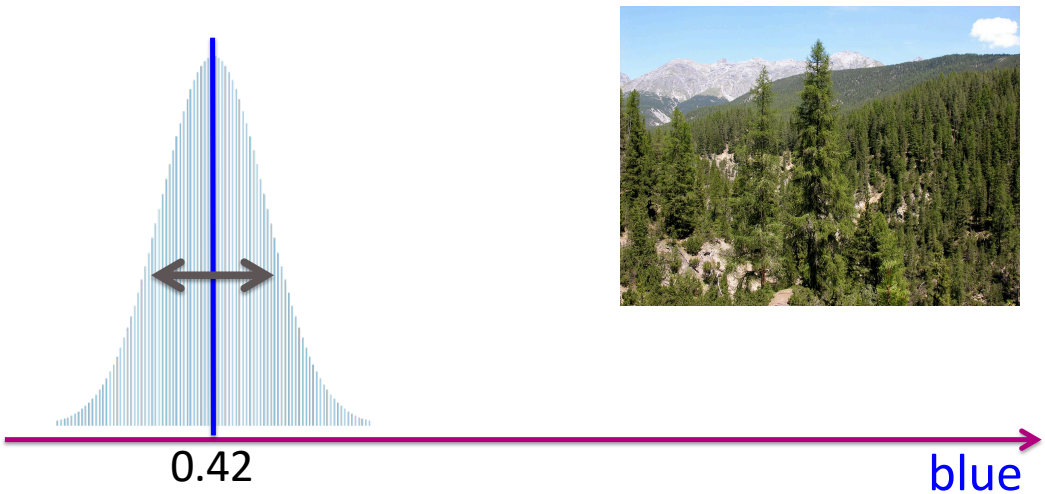
Let's look at just the blue dimension



©2017 Emily Fox

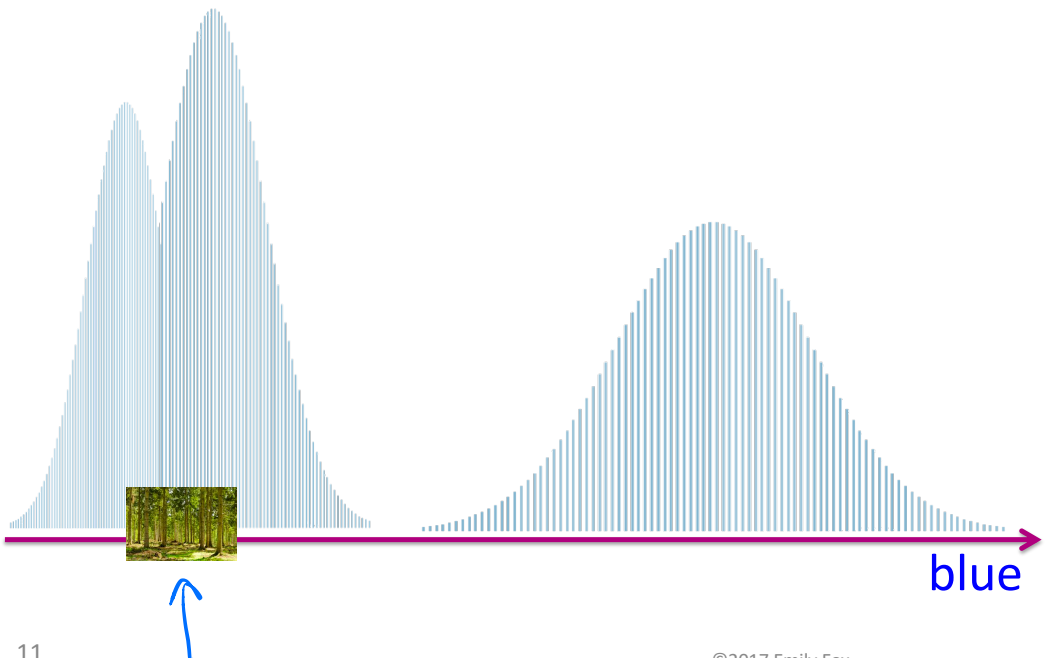
Distribution over all forest images

Let's look at just the blue dimension



©2017 Emily Fox

Distribution over all images



©2017 Emily Fox

$\vec{x}_i \in \mathbb{R}^d$

Can be distinguished along other dim

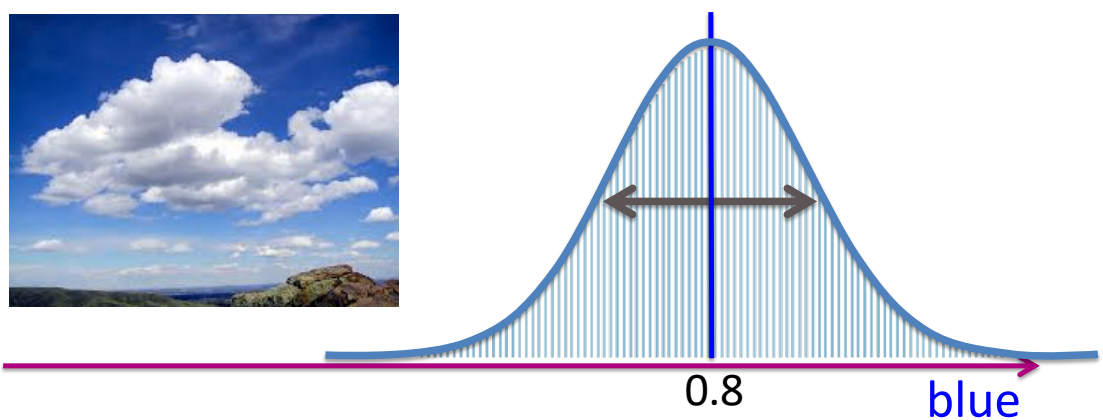
Now look at the red dimension

©2017 Emily Fox

Background: Gaussian distributions

Model for a given image type

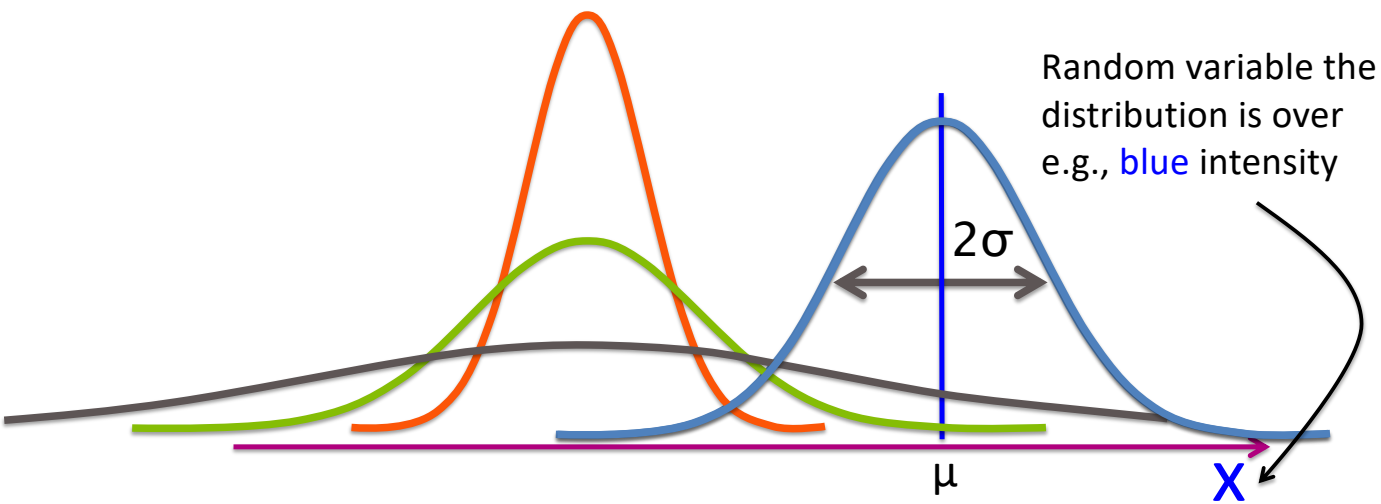
For each dim of the [R, G, B] vector, and each image type, assume a Gaussian distribution over color intensity



©2017 Emily Fox

1D Gaussians

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



©2017 Emily Fox

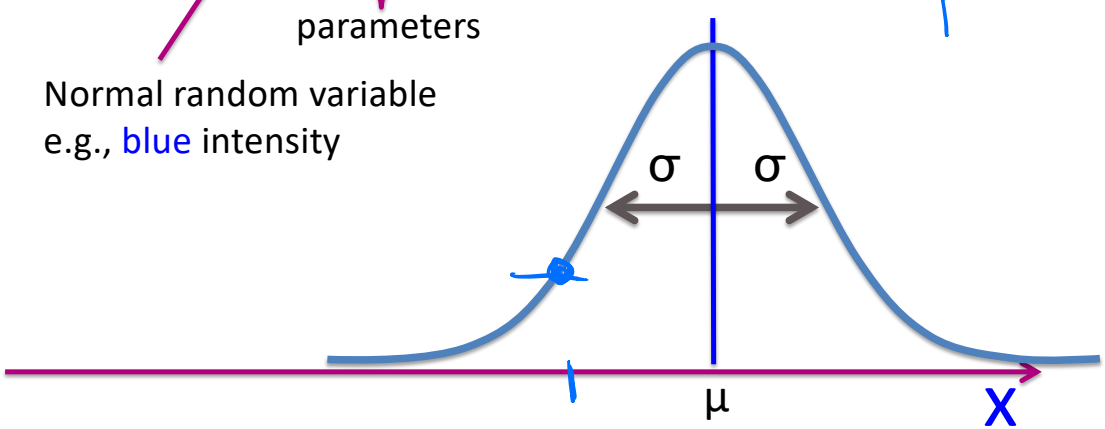
x is sample from $N(\mu, \sigma^2)$

Notation a 1D Gaussian distribution

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

parameters

Normal random variable
e.g., blue intensity



©2017 Emily Fox

$\lambda_i > 0$

Multivariate Gaussian density

$\Sigma_{ij} = \text{Cov}(x_i, x_j)$

$d \times d$

$\vec{x} = (x_1, \ldots, x_d)$ $\vec{\mu} = (\mu_1, \ldots, \mu_d)$ $\Sigma \in \mathbb{R}^{d \times d}$

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

$|A| = \det(A)$

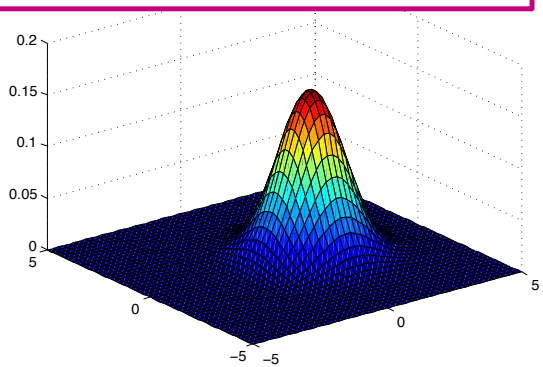
$\det(2\pi\Sigma) = (2\pi)^d \det(\Sigma)$

$\sqrt{|2\pi\Sigma|} = 2\pi^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}$

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

parameters

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



A

$x^T A x$ quadratic form

$(x-\mu)^T \Sigma^{-1} (x-\mu)$

©2017 Emily Fox

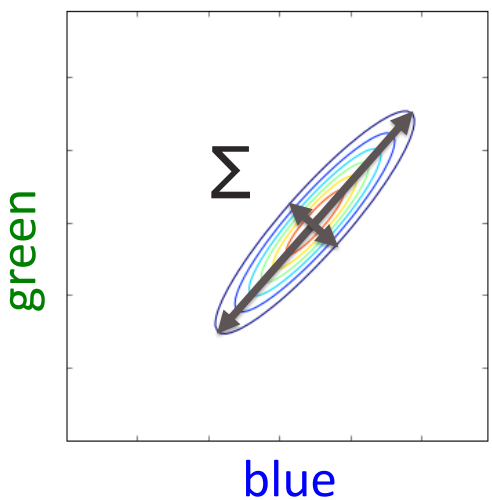
Multivariate Gaussian

Fully specified by mean μ and covariance Σ

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

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

covariance determines orientation + spread



©2017 Emily Fox

d=2

Covariance structure

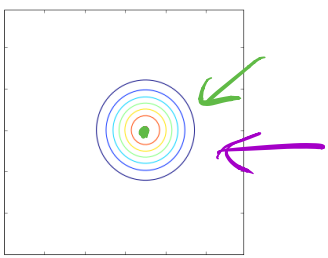
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

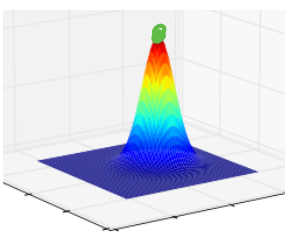
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}=0,\boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right\}$$

quadratic

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

for contour plot all points where $x^T\Sigma^{-1}x = 1$





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

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

©2017 Emily Fox

$X = (X_1 \ldots X_d)$ then X_i's are indep $N(\mu_i, \sigma_i^2)$ iff Σ diagonal.

Covariance structure

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

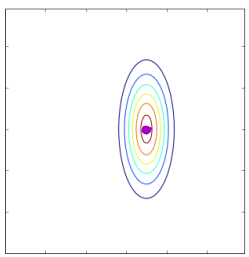
$$\Sigma = \begin{bmatrix} \sigma_B^2 & 0 \\ 0 & \sigma_G^2 \end{bmatrix}$$

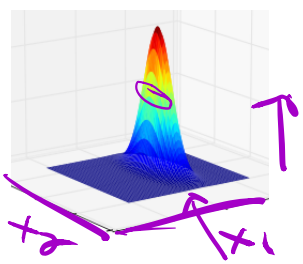
σ_B^2

σ_G^2

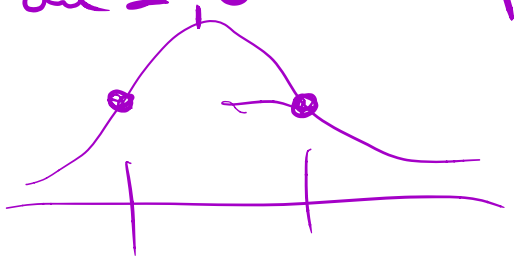
$$(x_1, x_2) \begin{pmatrix} \frac{1}{2^2} & 0 \\ 0 & \frac{1}{5^2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 1$$

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





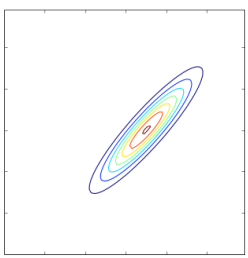
λ_1, λ_2 are $2^2, 5^2$

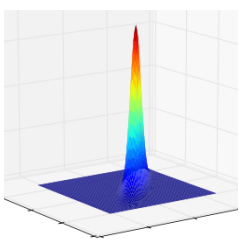


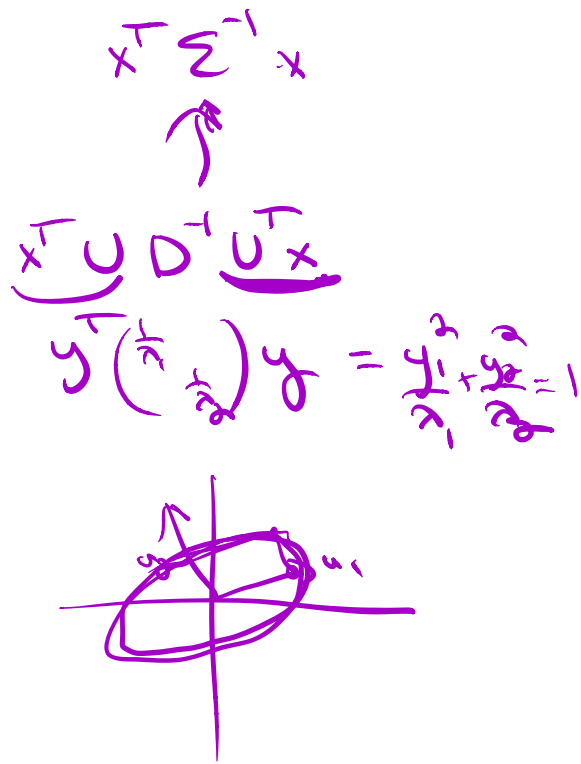
Covariance Structure

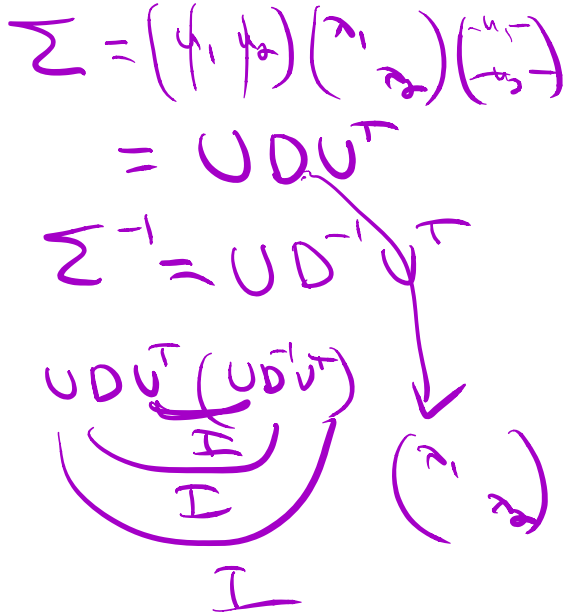
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}=0,\boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right\}$$

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









Important facts

- **Affine Property**

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$A\mathbf{X} + \mathbf{b} \sim N(A\boldsymbol{\mu} + \mathbf{b}, A\boldsymbol{\Sigma}A^T)$$

- **Constructing:** $X_1, \ldots, X_d \sim N(0,1)$ independent. Then $\mathbf{X} \sim N(0, I)$. Then

$$A\mathbf{X} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\Sigma} = AA^T$$

- **Sphering:**

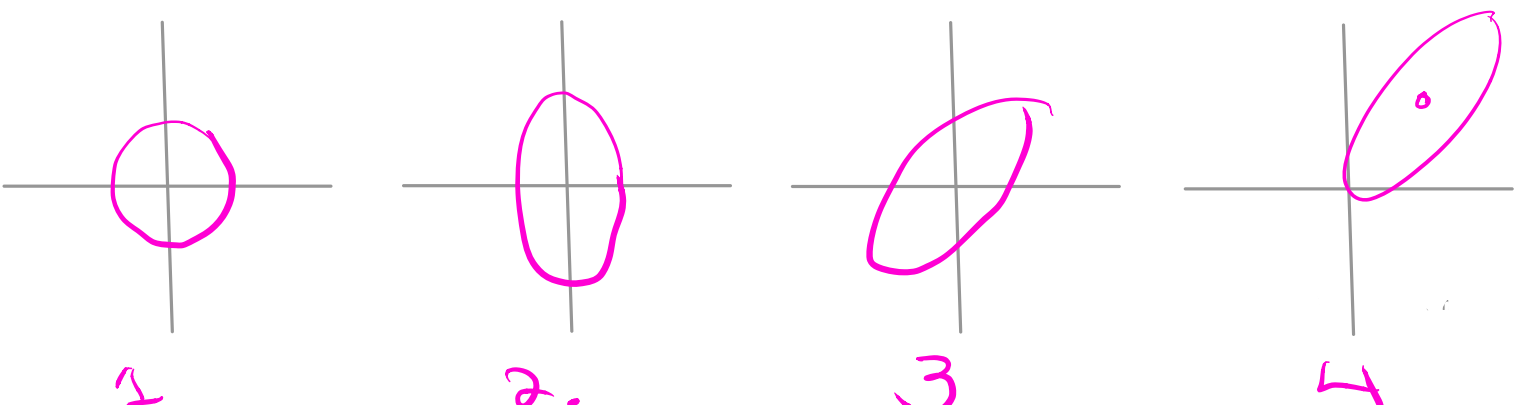
If $\boldsymbol{\Sigma}$ is psd, symmetric, then

$$\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \rightarrow \quad A^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \sim N(0, I)$$

$$\boldsymbol{\Sigma} = AA^T$$

More intuition

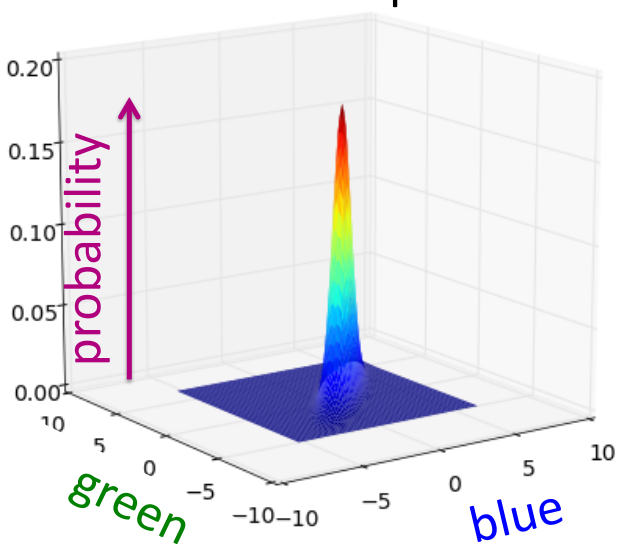
1. Start with $\mathbf{X} \sim N(0, I)$
2. (Scaling step) $D^{1/2}\mathbf{X} \sim N(0, D)$.
3. (Rotation) $UD^{1/2}\mathbf{X} \sim N(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = UDU^T$.
4. (Translation) $UD^{1/2}\mathbf{X} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$



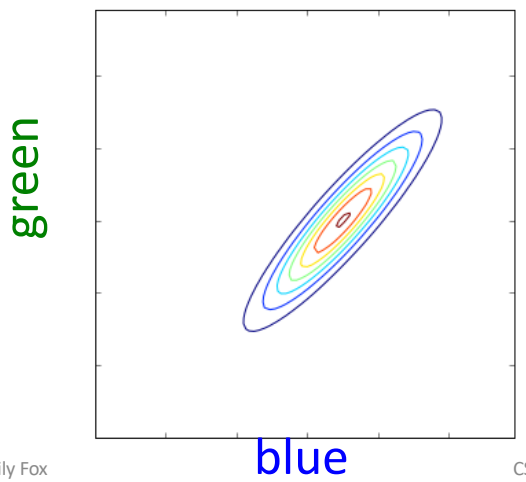
Multivariate Gaussians

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

3D mesh plot



Contour plot



©2017 Emily Fox

Summary

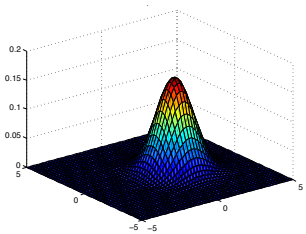
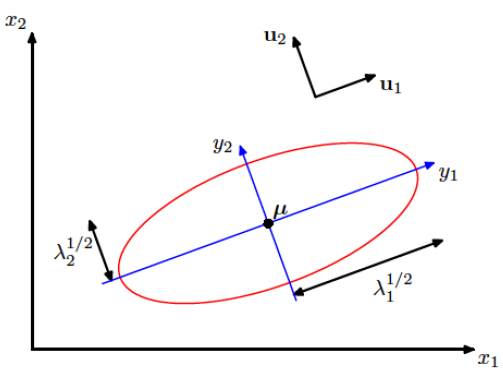
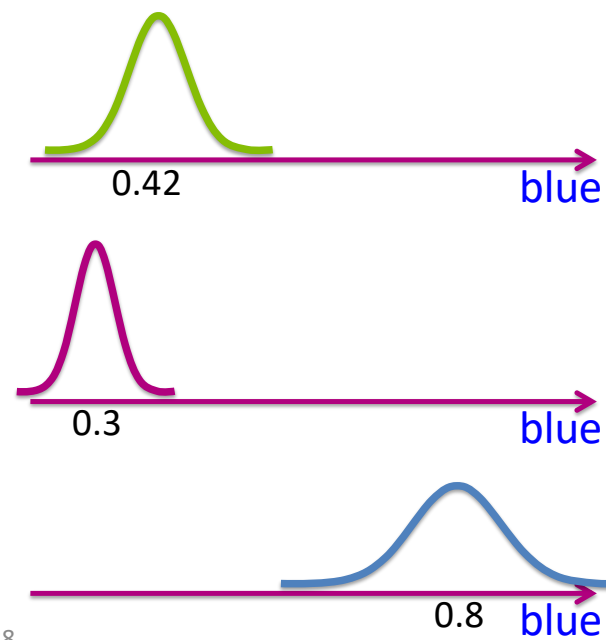


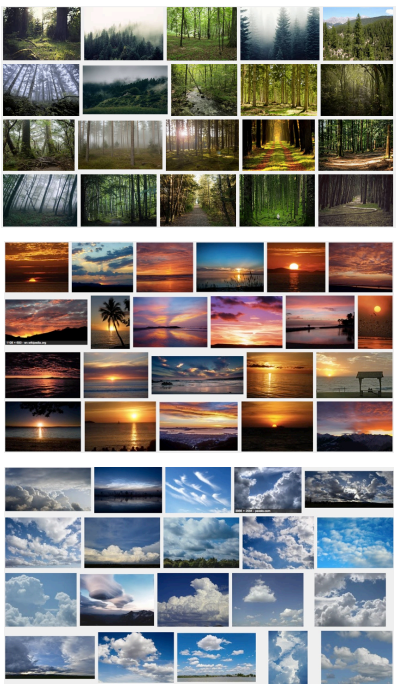
Figure 2.7 The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors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



Mixture Model (to be used for clustering)

Model as Gaussian per category/cluster





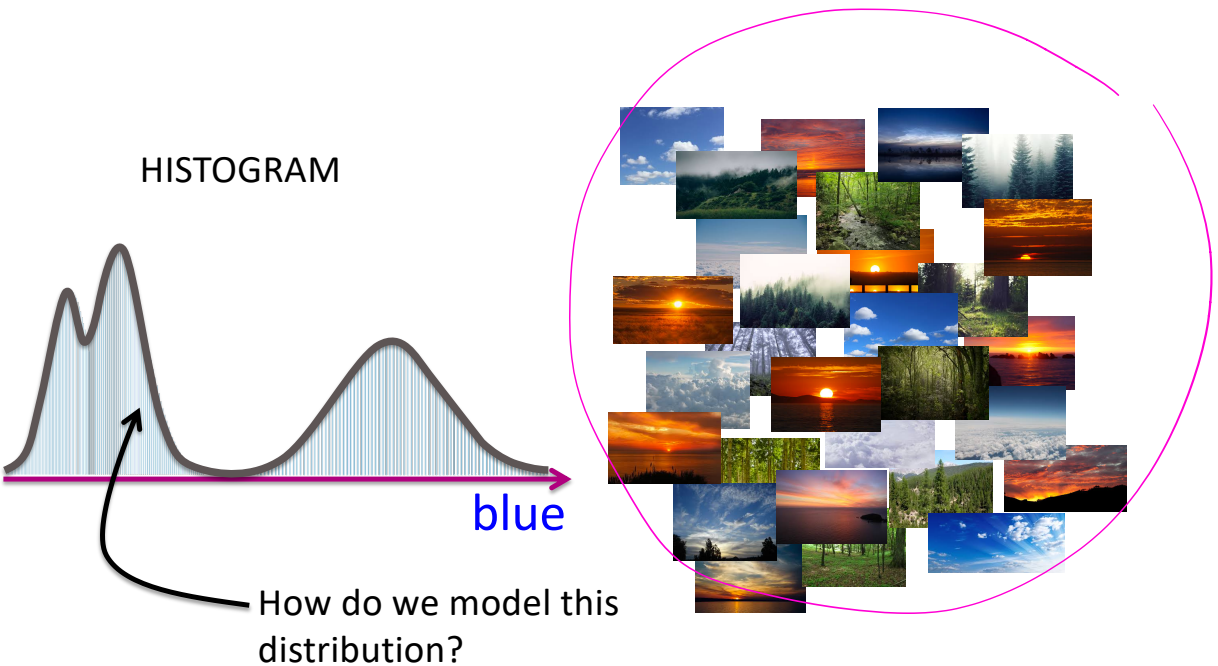
Forests

Sunsets

Clouds

©2017 Emily Fox

Jumble of unlabeled images



©2017 Emily Fox

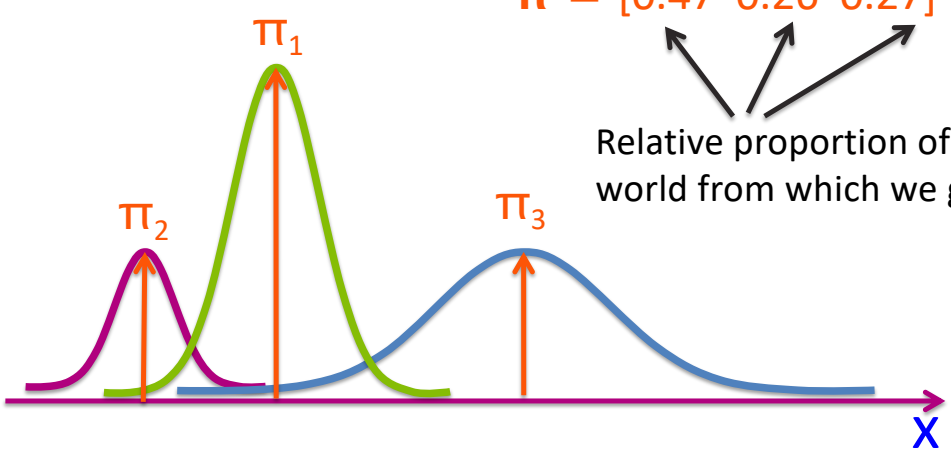
Combination of weighted Gaussians

Associate a weight π_k with each Gaussian component

π_1 π_2 π_3

$\boldsymbol{\pi} = [0.47 \;\; 0.26 \;\; 0.27]$

Relative proportion of each class in world from which we get data



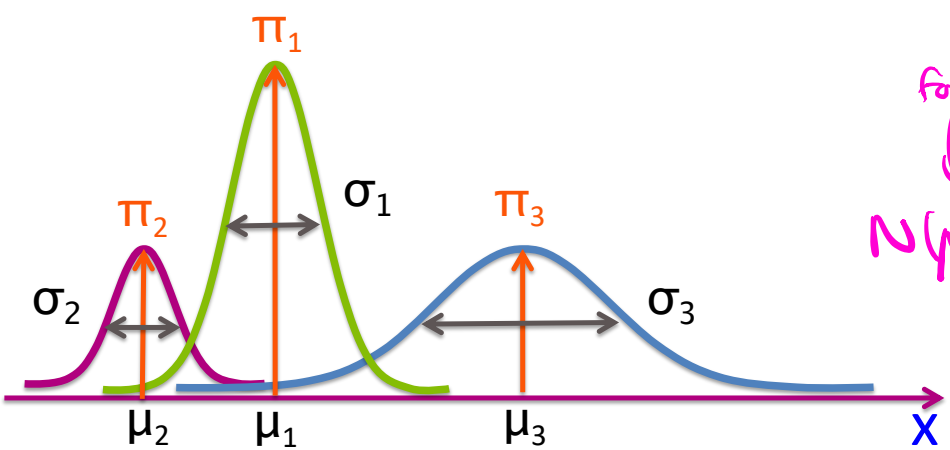
$0 \leq \pi_k \leq 1$

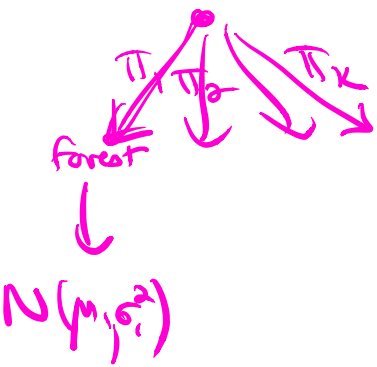
$$\sum_{k=1}^{K} \pi_k = 1$$

©2017 Emily Fox

Mixture of Gaussians (1D)

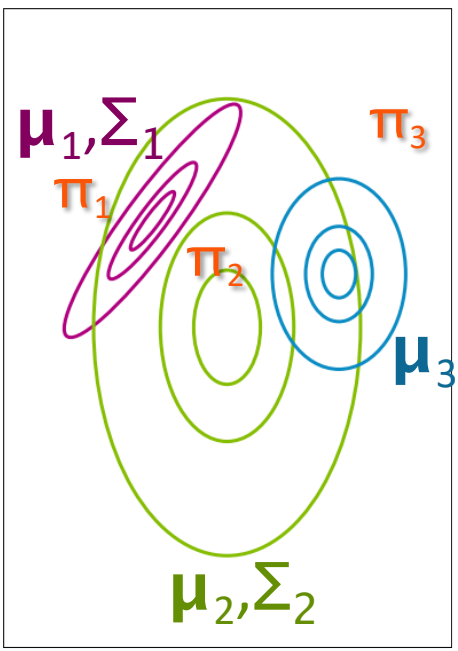
Each mixture component represents a unique cluster specified by: $\{\pi_k, \mu_k, \sigma_k^2\}$





©2017 Emily Fox

Mixture of Gaussians (general)



Each mixture component represents a unique cluster specified by:

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

©2017 Emily Fox

Mixture model

- K clusters, defined by the following parameters

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

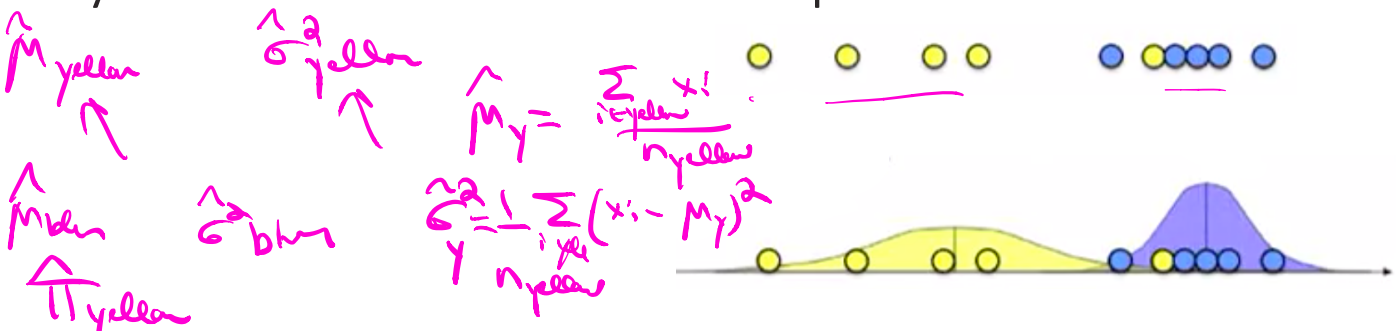
$$\sum_{j=1}^{K} \pi_j = 1.$$

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

©2017 Emily Fox

K=2 1-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

To understand better, introduce additional "latent" variables

- K clusters, defined by the following parameters

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

$z_k = \begin{cases} 1 & x \text{ comes from cluster } k \\ 0 & \text{o.w} \end{cases}$

- For each point $\mathbf{x}$, let $\mathbf{z} = (z_1, \ldots, z_K)$

indicate which cluster it was chosen from. These are called "latent variables".

$$p(z_k = 1) = \pi_k$$

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

x_i $z_i = (z_{i1} \cdots z_{iK})$

With z's in hand, we can compute many relevant quantities

given params

- Conditional distribution of **x** given **z**

$$Pr(x_i \mid z_{ik}=1) = N(x_i \mid \mu_k, \Sigma_k)$$

- Therefore:

$$Pr(x_i) = \sum_{j=1}^{K} Pr(z_j=1)\, Pr(x_i \mid z_j=1) = \sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \Sigma_j)$$

- Conditional probability

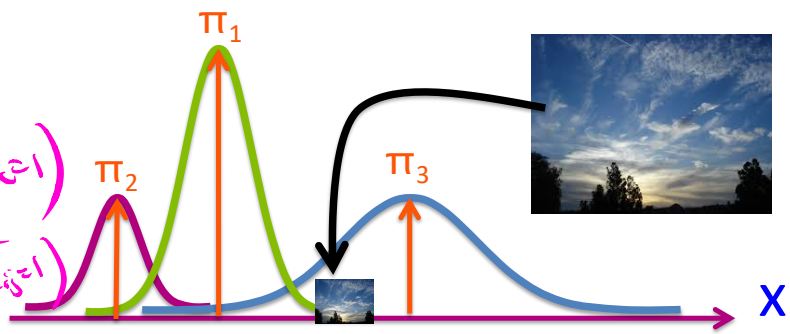
of **z** given **x**

$$Pr(z_{ik}=1 \mid x_i) = \frac{Pr(x_i \mid z_{ik}=1)\, Pr(z_{ik}=1)}{\sum_{j=1}^{K} Pr(x_i \mid z_j=1)^{Pr(z_j=1)}}$$

posterior

$$= \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum \pi_j N(x_i \mid \mu_j, \Sigma_j)}$$

π_j : prob pt comes from cluster j



©2017 Emily Fox

Mixture model cont.

- Conditional distribution of $\mathbf{x}$

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Therefore:

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

- We will also be interested in the conditional probability of **z** given **x**

$$\begin{aligned} p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \end{aligned}$$

And we can now try to calculate MLE

- Given dataset from a mixture model, find parameters

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

that maximize loglikelihood.

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

$$\prod_{i=1}^{N} p(x_i|\theta)$$

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

©2017 Emily Fox

$x_i = (x_{i1}, x_{i2}, x_{id})$

MLE

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

$\Pr(x|\theta)$

Gradient = 0 gives the following conditions:

r_{ik} = responsibility of k^{th} cluster for point i
$= \Pr(z_{ik}=1 | x_i, \theta)$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik}\, \mathbf{x}_i$$

$$N_k = \sum_{i=1}^{N} r_{ik}$$ = exp # pts in cluster k

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik}\, (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\hat{\pi}_k = \frac{N_k}{N}$$

$r_{ik} = p(z_k = 1|\mathbf{x}) = \dfrac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$

$= \dfrac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$

Does not give us a closed form ☹

$r_{ik} = \Pr(z_{ik}=1 | x_i, \theta)$
x_i comes from cluster k

Expectation Maximization Algorithm

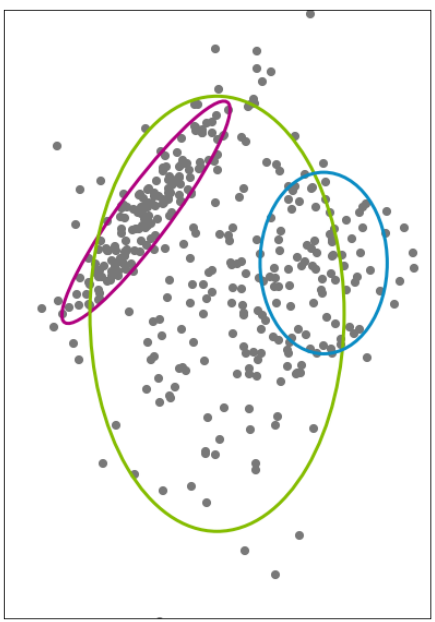
Two step approach based on following observation

- If we knew the $\mathbf{z}_i$'s, we could estimate all the parameters.
- If we knew all the parameters we could estimate the $\mathbf{z}_i$'s (or more precisely, the chance each point came from each cluster)
- EM is an iterative algorithm that alternates between these two steps.

params μ_k, Σ_k, π_k

©2017 Emily Fox

E step: estimate responsibilities

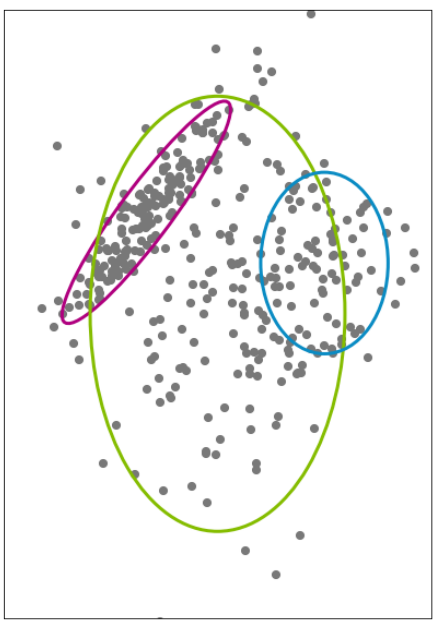


Compute $r_{ik} = \Pr(z_{ik} = 1 \mid \mathbf{x}_i, \text{params}\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\})$ (responsibilities)

©2017 Emily Fox

Responsibilities in equations

$$Pr(z_{ik}=1 \mid x_i, \theta)$$



Responsibility cluster k takes for observation i

$$r_{ik} = \frac{\pi_k \; N(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \Sigma_j)}$$

Normalized over all possible cluster assignments

©2017 Emily Fox

M step:

Given the soft assignments r_{ij}, estimate parameters

Estimating cluster parameters from soft assignments r_{ij}

©2017 Emily Fox

Estimating cluster parameters from assignments r_{ij}

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3
$x_4[1]$	$x_4[2]$	$x_4[3]$	1
$x_5[1]$	$x_5[2]$	$x_5[3]$	2
$x_6[1]$	$x_6[2]$	$x_6[3]$	2

Suppose that magically r_{ik}'s are all 0, 1, i.e. hard assignments

Estimate $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$ given data assigned to cluster k

©2017 Emily Fox

Mean/covariance MLE

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i \text{ in } k} x_i$$

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_{i \text{ in } k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

©2017 Emily Fox

Cluster proportion MLE

R	G	B	Cluster
$x_4[1]$	$x_4[2]$	$x_4[3]$	1

R	G	B	Cluster
$x_5[1]$	$x_5[2]$	$x_5[3]$	2
$x_6[1]$	$x_6[2]$	$x_6[3]$	2

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

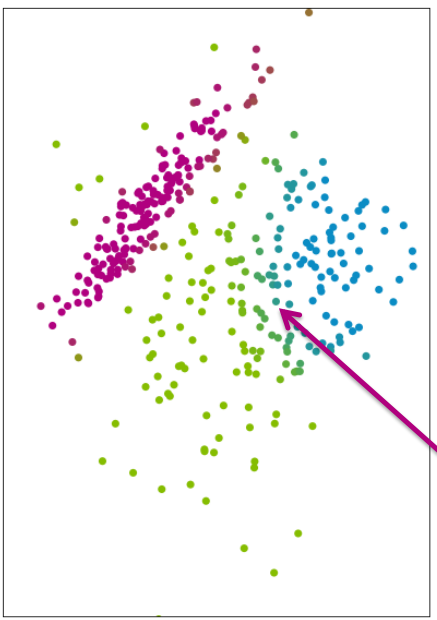
obs in cluster k

$$\hat{\pi}_k = \frac{N_k}{N}$$

total # of obs

©2017 Emily Fox

Estimating cluster parameters from soft assignments



Instead of having a full observation x_i in cluster k, just allocate a portion r_{ik}

x_i divided across all clusters, as determined by r_{ik}

©2017 Emily Fox

Maximum likelihood estimation from soft assignments

R	G	B	r_{i1}	r_{i2}	r_{i3}
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30	0.18	0.52
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.01	0.26	0.73
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.002	0.008	0.99
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.75	0.10	0.15
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.05	0.93	0.02
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.13	0.86	0.01

52% chance this obs is in cluster 3

Total weight in cluster:
(effective # of obs)

1.242	2.8	2.42

©2017 Emily Fox

Maximum likelihood estimation from soft assignments

R	G	B	Cluster 1 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30
$x_2[1]$			
$x_3[1]$			
$x_4[1]$			
$x_5[1]$			
$x_6[1]$			

R	G	B	Cluster 2 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.18
$x_2[1]$			
$x_3[1]$			
$x_4[1]$			
$x_5[1]$			
$x_6[1]$			

R	G	B	Cluster 3 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.52
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.73
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.99
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.15
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.02
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.01

©2017 Emily Fox

Cluster-specific location/shape MLE

R	G	B	Cluster 1 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.01
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.002
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.75
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.05
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.13

1.242

$$\hat{\mu}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} x_i$$

$$\hat{\Sigma}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

$$N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik}$$

Total weight in cluster k = effective # obs

Compute cluster parameter estimates with weights on each row operation

©2017 Emily Fox

MLE of cluster proportions $\hat{\pi}_k$

r_{i1}	r_{i2}	r_{i3}
0.30	0.18	0.52
0.01	0.26	0.73
0.002	0.008	0.99
0.75	0.10	0.15
0.05	0.93	0.02
0.13	0.86	0.01

$$\hat{\pi}_k = \frac{N_k^{\text{soft}}}{N}$$

$$N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik}$$

Total weight in cluster k = effective # obs

Total weight in cluster:

1.242	**2.8**	**2.42**

Total weight in dataset: 6

datapoints N

Estimate cluster proportions from relative weights

©2017 Emily Fox

M step summary

Compute cluster parameter estimates from soft assignments

$$\hat{\mu}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} x_i$$

$$\hat{\Sigma}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

$$N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik} \qquad \hat{\pi}_k = \frac{N_k^{\text{soft}}}{N}$$

©2017 Emily Fox

Expectation maximization (EM)

Expectation maximization (EM): An iterative algorithm

Motivates an iterative algorithm:

$$\theta = \{\mu_k^{(0)}, \Sigma_k^{(0)}, \pi_k^{(0)}\}$$

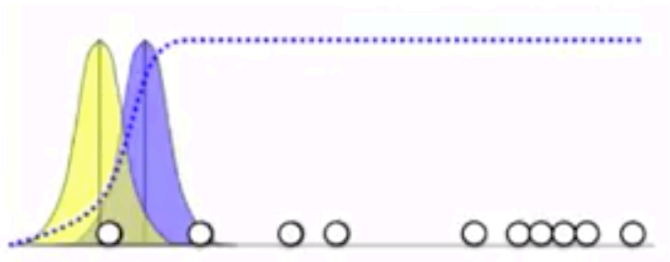
1. E-step: estimate cluster responsibilities given current parameter estimates

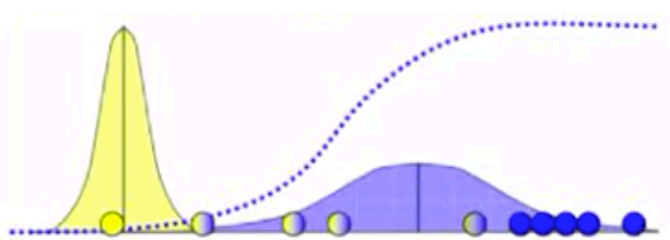
$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^{K} \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}$$

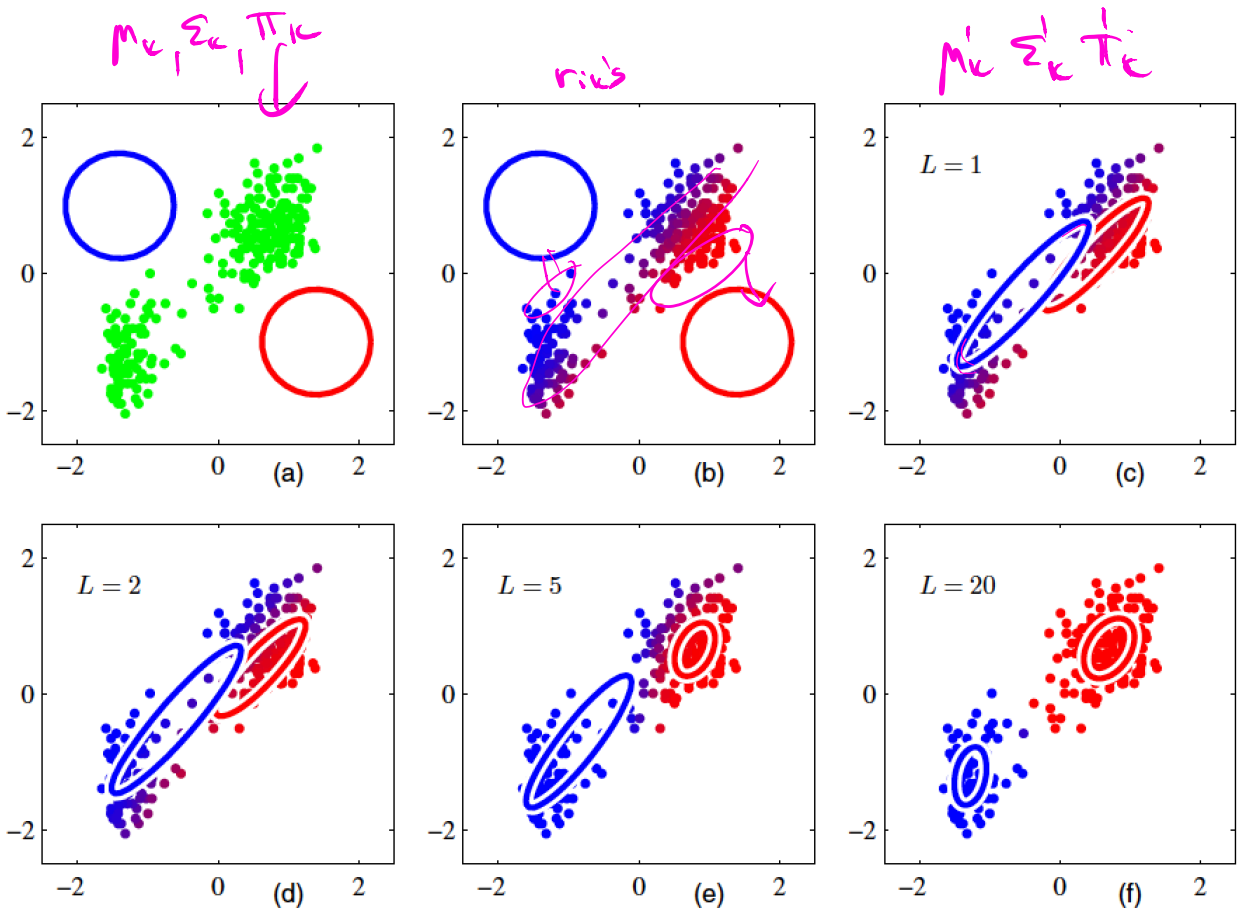
2. M-step: maximize likelihood over parameters given current responsibilities

$$\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}_k \mid \{\hat{r}_{ik}, x_i\}$$

©2017 Emily Fox

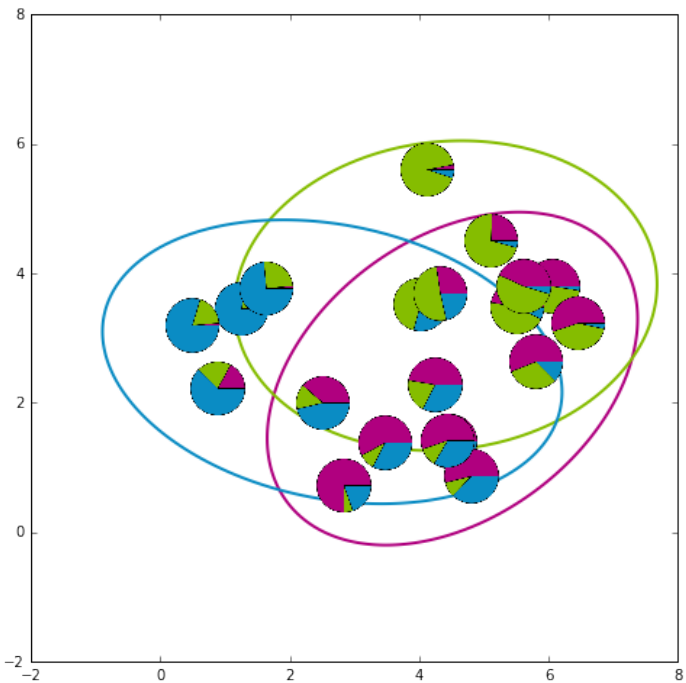






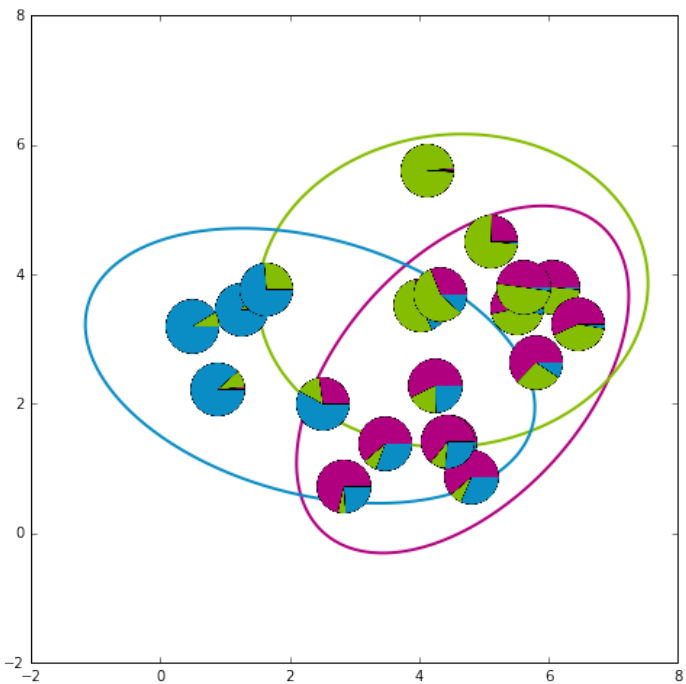
©2017 Emily Fox

EM for mixtures of Gaussians in pictures – initialization



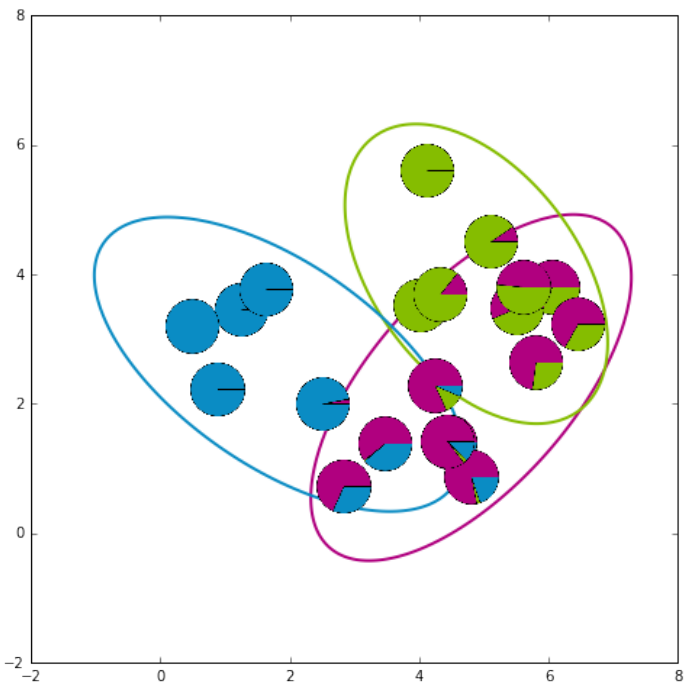
©2017 Emily Fox

EM for mixtures of Gaussians in pictures – after 1st iteration



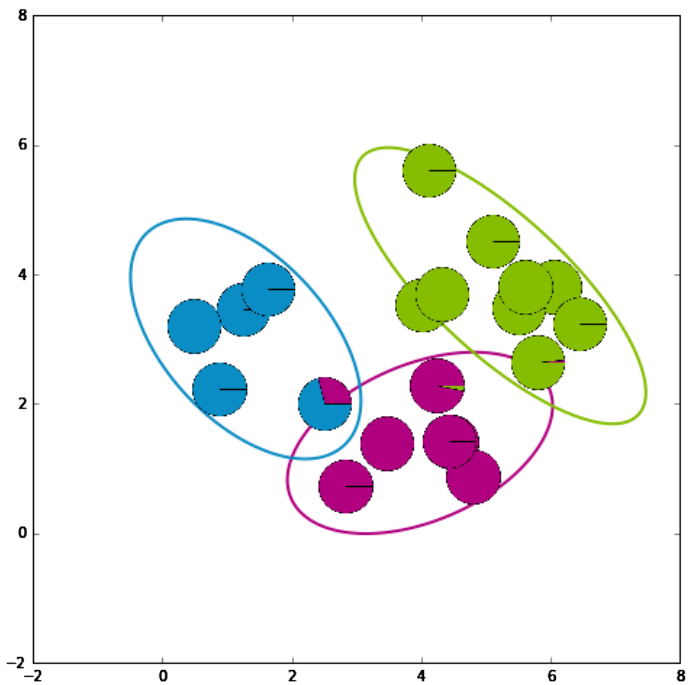
©2017 Emily Fox

EM for mixtures of Gaussians in pictures – after 2nd iteration



©2017 Emily Fox

EM for mixtures of Gaussians in pictures – converged solution



©2017 Emily Fox

The nitty gritty of EM

Convergence and initialization of EM

Convergence of EM

- EM is a coordinate-ascent algorithm
  - Can equate E-and M-steps with alternating maximizations of the objective function
- Convergence to a local maximum.

- We assess via (log) likelihood of data under current parameter and responsibility estimates

©2017 Emily Fox

Initialization

- Many ways to initialize the EM algorithm
- Important for convergence rates & quality of local maximum found
- Examples:
  - Choose K observations at random to define K "centroids". Assign other observations to nearest centroid to form initial parameter estimates.
  - Initialize from k-means solution

©2017 Emily Fox

Potential of vanilla EM to overfit

Overfitting of MLE

Maximizing likelihood can overfit to data

Imagine at K=2 example with one obs assigned to cluster 1 and others assigned to cluster 2

- What parameter values maximize likelihood?

Set center equal to point and shrink variance to 0

Likelihood goes to ∞ !

©2017 Emily Fox

Simple regularization of M-step for mixtures of Gaussians

Simple fix: Don't let variances → 0!

Add small amount to diagonal of covariance estimate

©2017 Emily Fox

Summary

What you can do now…

- Understand Gaussian mixture models (and multivariate Gaussians)
- Estimate soft assignments (responsibilities) given mixture model parameters
- Solve maximum likelihood parameter estimation using soft assignments (weighted data)
- Implement an EM algorithm for inferring soft assignments and cluster parameters
  - Determine an initialization strategy
  - Implement a variant that helps avoid overfitting issues
- Compare and contrast with k-means
  - Soft vs. hard assignments

©2017 Emily Fox