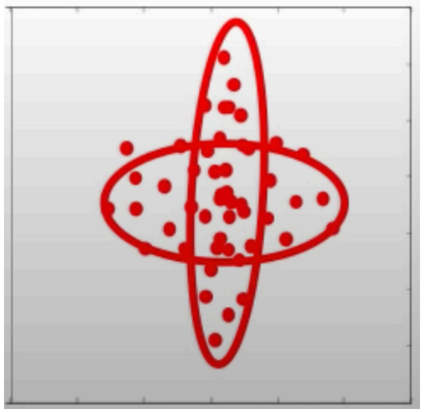
Segmentation and Clustering (Part 2)

Today's Lecture

- **Segmentation and grouping**
  - Gestalt principles
  - Image segmentation
- **Segmentation as clustering**
  - k-Means
  - Feature spaces
- **Probabilistic clustering**
  - Gaussian and Multivariate Gaussian distributions
  - Mixture of Gaussians
  - Expectation-Maximization (EM)
- **Model-free clustering**
  - Mean-Shift clustering
- **Graph theoretic segmentation**
  - Normalised cuts

Probabilistic Clustering

- **K-means algorithm**
  - **Assigned each point x to *exactly* one cluster**
    - **'Hard assignment'**
  - **What if clusters are overlapping?**
    - **Hard to tell which cluster is right**
    - **Maybe we should try to remain uncertain**
  - **What if cluster has a non-circular shape?**



- **Gaussian Mixture models**
  - **Clusters modeled as Gaussians**
    - **Not just by their mean**
  - **EM algorithm: assign data to cluster with some *probability***
  - **Gives probability model of x ("generative")**

Univariate Normal (Gaussian) Distribution

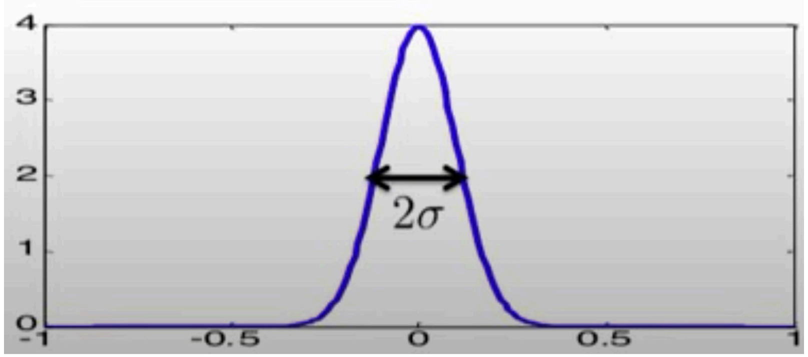
$$Pr(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-0.5(x-\mu)^2/\sigma^2\right]$$

Or, usual notation:

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

Univariate (1D) Gaussian or normal distribution describes a single continuous variable.

Takes 2 parameters: μ and σ^2>0



Parameters: mean μ and variance σ^2>0 (standard deviation σ)

Maximum Likelihood estimates

$$\hat{\mu} = \frac{1}{N}\sum_i x^{(i)}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x^{(i)} - \hat{\mu})^2$$

Multivariate Normal Distribution

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

Multivariate normal distribution describes multiple continuous variables. Takes two parameters

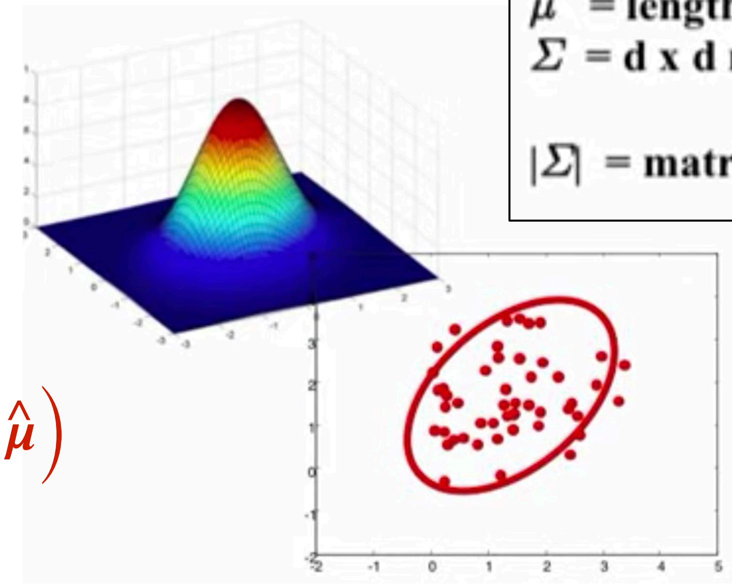
- a **vector** containing mean position, μ
- a symmetric “positive definite” **covariance matrix** Σ

μ = length-d row vector
Σ = d x d matrix

$|\Sigma|$ = matrix determinant

$$\hat{\boldsymbol{\mu}} = \frac{1}{m}\sum_{j}\boldsymbol{x}^{(j)}$$

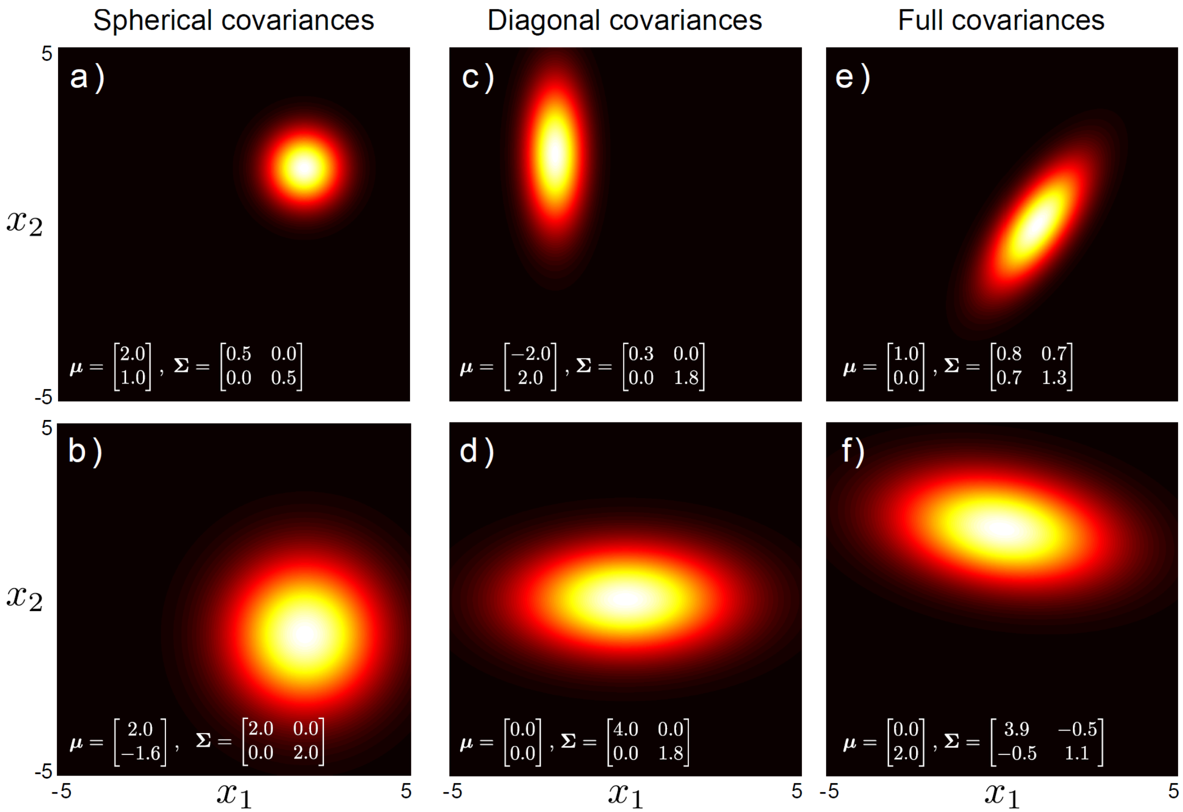
$$\hat{\Sigma} = \frac{1}{m}\sum_{j}\left(\boldsymbol{x}^{j}-\hat{\boldsymbol{\mu}}\right)^{T}\left(\boldsymbol{x}^{j}-\hat{\boldsymbol{\mu}}\right)$$



Types of covariance

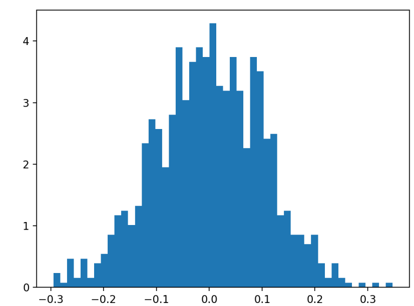
Covariance matrix has three forms, termed spherical, diagonal and full

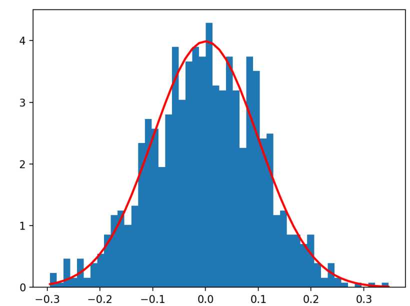
$$\boldsymbol{\Sigma}_{spher} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \quad \boldsymbol{\Sigma}_{diag} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \boldsymbol{\Sigma}_{full} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$$



slide credit: Simon J.D. Prince

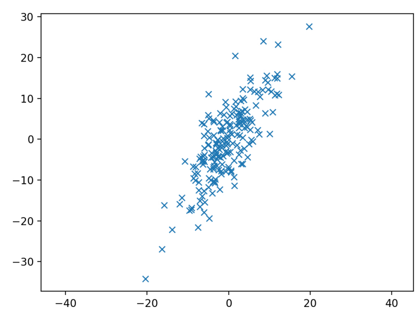
Probabilistic Clustering

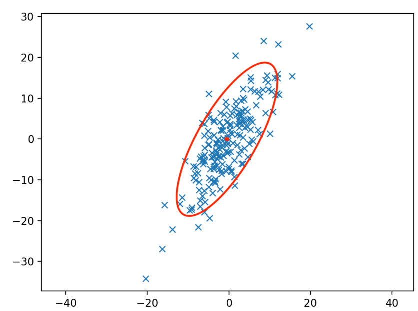




$\mathcal{N}\left(x;\, \mu,\ \sigma\right)$

Gaussian (normal) distribution gives probability model of x ("**generative**")

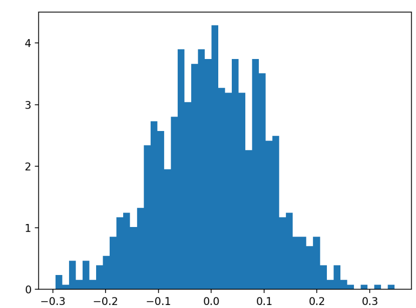


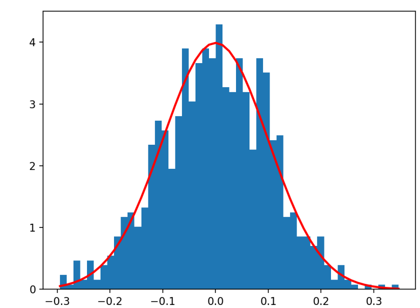


$\mathcal{N}\left(\boldsymbol{x};\, \boldsymbol{\mu},\ \Sigma\right)$

Probabilistic Clustering

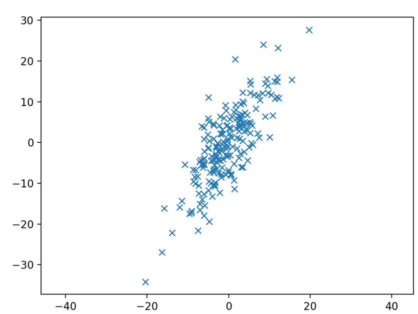
- **Basic idea**
  - Instead of treating the data as a bunch of points, assume that they are all generated by sampling a continuous function.
  - This function is called a generative model
  - Defined by a vector of parameters $\boldsymbol{\theta}$

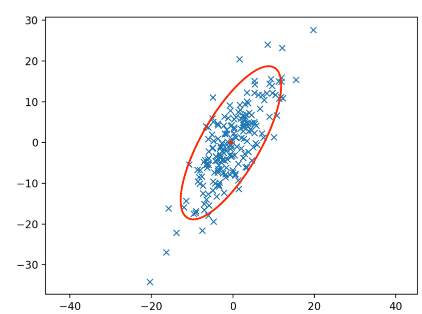




$$\mathcal{N}\left(x;\, \mu,\ \sigma\right)$$

Gaussian (normal) distribution gives probability model of x (“**generative**”)

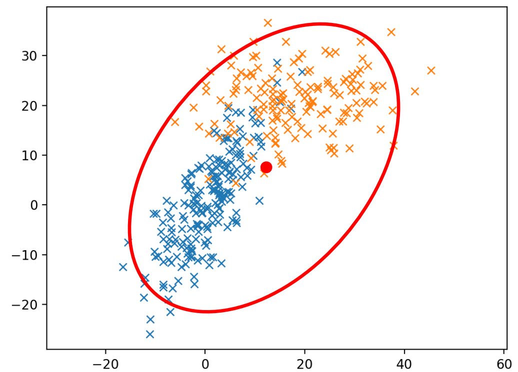


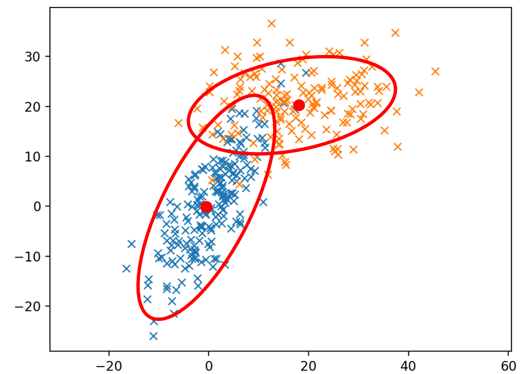


$$\mathcal{N}\left(\boldsymbol{x};\, \boldsymbol{\mu},\ \Sigma\right)$$

Single Gaussian vs Mixture of Gaussians Model

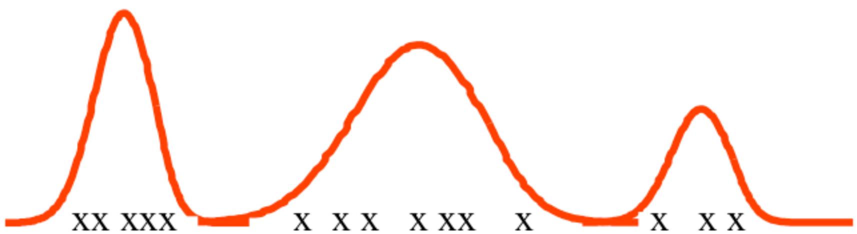
A normal distribution is not good enough! Need a way to make more complex distributions





We model each cluster using one of these Gaussians “bells”

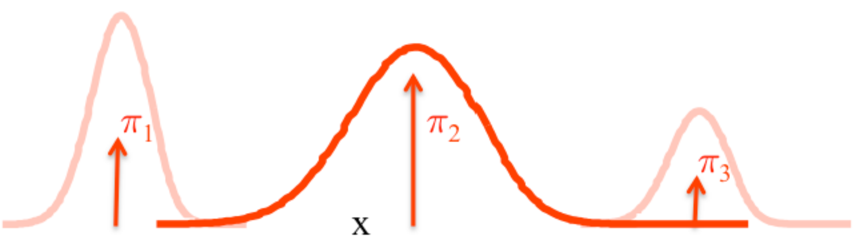
- Start with parameters describing each cluster
- Mean μ_c , variance σ_c , "size" π_c
- Probability distribution: $p(x) = \sum_c \pi_c \, \mathcal{N}(x \; ; \; \mu_c, \sigma_c)$



- Start with parameters describing each cluster
- Mean μ_c , variance σ_c , "size" π_c
- Probability distribution: $p(x) = \sum_c \pi_c \, \mathcal{N}(x \; ; \; \mu_c, \sigma_c)$

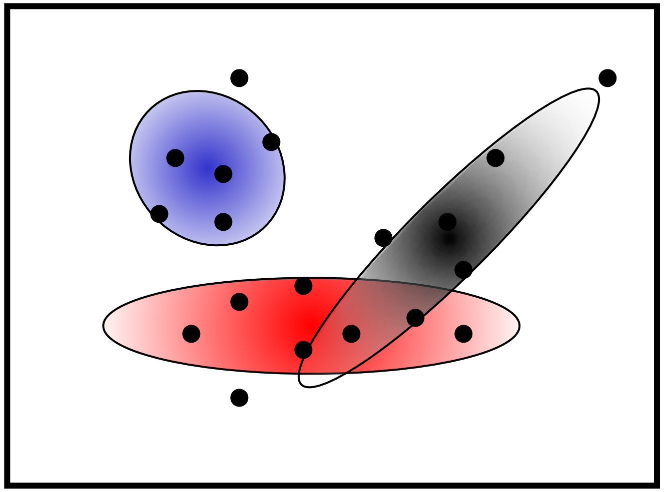
$p(z = c) = \pi_c$ Select a mixture component with probability π

$p(x|z = c) = \mathcal{N}(x \; ; \; \mu_c, \sigma_c)$ Sample from that component's Gaussian



Slide credit: A. Ihler

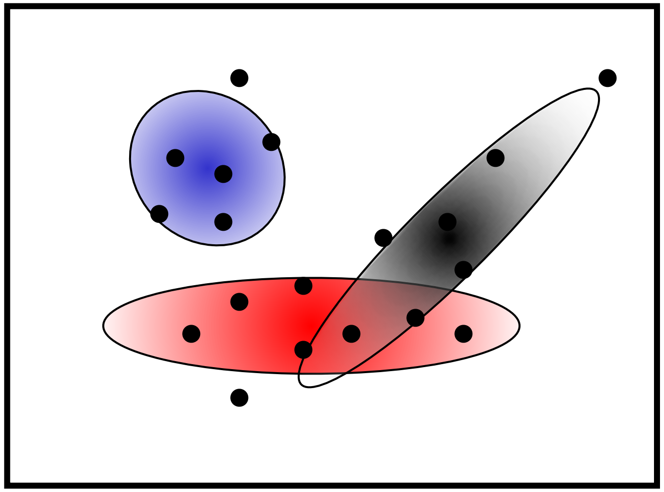
Mixture of Gaussians



- **One generative model is a mixture of Gaussians (MoG)**
  - K Gaussian blobs with means μ_b covariance matrices V_b, dimension d
    - Blob *b* defined by: $$P(x|\mu_b, V_b) = \frac{1}{\sqrt{(2\pi)^d |V_b|}} e^{-\frac{1}{2}(x-\mu_b)^T V_b^{-1}(x-\mu_b)}$$
  - Blob *b* is selected with probability α_b
  - The likelihood of observing x is a weighted mixture of Gaussians

Slide credit: Steve Seitz

Mixture of Gaussians

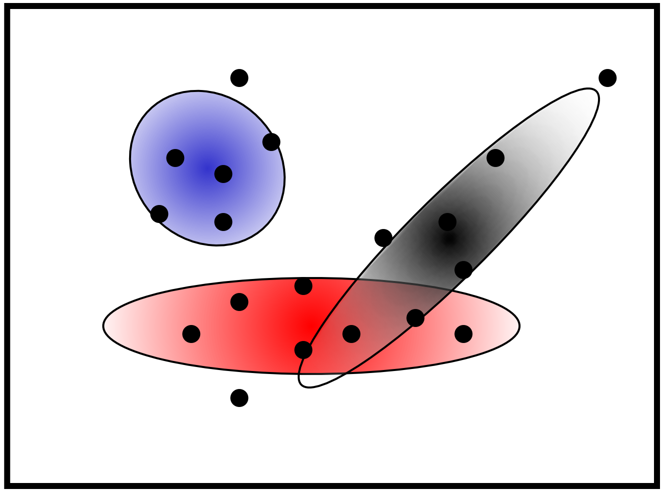


- **One generative model is a mixture of Gaussians (MoG)**
  - K Gaussian blobs with means μ_b covariance matrices V_b, dimension d
    - Blob *b* defined by: $$P(x|\mu_b, V_b) = \frac{1}{\sqrt{(2\pi)^d |V_b|}} e^{-\frac{1}{2}(x-\mu_b)^T V_b^{-1}(x-\mu_b)}$$
  - Blob *b* is selected with probability α_b
  - The likelihood of observing x is a weighted mixture of Gaussians

$$P(x|\theta) = \sum_{b=1}^{K} \alpha_b P(x|\theta_b) , \quad \theta = [\mu_1, \ldots, \mu_n, V_1, \ldots, V_n]$$

Slide credit: Steve Seitz

Expectation Maximization (EM)



- **Goal**
  - Find blob parameters $\boldsymbol{\theta}$ that maximize the likelihood function:

$$P(data|\theta) = \prod_x P(x|\theta)$$

- **Approach:**
  1. E-step: given current guess of blobs, compute ownership of each point
  2. M-step: given ownership probabilities, update blobs to maximise likelihood function
  3. Repeat until convergence

Slide credit: Steve Seitz

EM Details

- **E-step**
  - Compute probability that point $\boldsymbol{x}$ is in blob *b*, given current guess of $\boldsymbol{\theta}$

$$P(b|x, \mu_b, V_b) = \frac{\alpha_b P(x|\mu_b, V_b)}{\sum_{i=1}^{K} \alpha_i P(x|\mu_i, V_i)}$$

EM Details

- **E-step**
  - Compute probability that point $\boldsymbol{x}$ is in blob *b*, given current guess of $\boldsymbol{\theta}$

$$P(b|x, \mu_b, V_b) = \frac{\alpha_b P(x|\mu_b, V_b)}{\sum_{i=1}^{K} \alpha_i P(x|\mu_i, V_i)}$$

- **M-step**

EM Details

- **E-step**
  - Compute probability that point $\boldsymbol{x}$ is in blob *b*, given current guess of $\boldsymbol{\theta}$

$$P(b|x, \mu_b, V_b) = \frac{\alpha_b P(x|\mu_b, V_b)}{\sum_{i=1}^{K} \alpha_i P(x|\mu_i, V_i)}$$

- **M-step**
  - Compute probability that blob *b* is selected

EM Details

- **E-step**
  - Compute probability that point $\boldsymbol{x}$ is in blob *b*, given current guess of $\boldsymbol{\theta}$

$$P(b|x, \mu_b, V_b) = \frac{\alpha_b P(x|\mu_b, V_b)}{\sum_{i=1}^{K} \alpha_i P(x|\mu_i, V_i)}$$

- **M-step**
  - Compute probability that blob *b* is selected

$$\alpha_b^{new} = \frac{1}{N} \sum_{i=1}^{N} P(b|x_i, \mu_b, V_b)$$

(N data points)

EM Details

- **E-step**
  - Compute probability that point $\boldsymbol{x}$ is in blob *b*, given current guess of $\boldsymbol{\theta}$

$$P(b|x, \mu_b, V_b) = \frac{\alpha_b P(x|\mu_b, V_b)}{\sum_{i=1}^{K} \alpha_i P(x|\mu_i, V_i)}$$

- **M-step**
  - Compute probability that blob *b* is selected

$$\alpha_b^{new} = \frac{1}{N} \sum_{i=1}^{N} P(b|x_i, \mu_b, V_b)$$

(N data points)

MANCHESTER 1824

The University of Manchester

EM Details

- **E-step**
  - Compute probability that point $\boldsymbol{x}$ is in blob *b*, given current guess of $\boldsymbol{\theta}$

$$P(b|x,\mu_b,V_b) = \frac{\alpha_b P(x|\mu_b,V_b)}{\sum_{i=1}^{K} \alpha_i P(x|\mu_i,V_i)}$$

- **M-step**
  - Compute probability that blob *b* is selected

$$\alpha_b^{new} = \frac{1}{N}\sum_{i=1}^{N} P(b|x_i,\mu_b,V_b)$$

  - Mean of blob *b* $\quad \mu_b^{new} = \frac{\sum_{i=1}^{N} x_i P(b|x_i,\mu_b,V_b)}{\sum_{i=1}^{N} P(b|x_i,\mu_b,V_b)}$ **(N data points)**

MANCHESTER 1824

The University of Manchester

EM Details

- **E-step**
  - Compute probability that point $\boldsymbol{x}$ is in blob *b*, given current guess of $\boldsymbol{\theta}$

$$P(b|x,\mu_b,V_b) = \frac{\alpha_b P(x|\mu_b,V_b)}{\sum_{i=1}^{K} \alpha_i P(x|\mu_i,V_i)}$$

- **M-step**
  - Compute probability that blob *b* is selected

$$\alpha_b^{new} = \frac{1}{N}\sum_{i=1}^{N} P(b|x_i,\mu_b,V_b)$$

  - Mean of blob *b* $\mu_b^{new} = \frac{\sum_{i=1}^{N} x_i P(b|x_i,\mu_b,V_b)}{\sum_{i=1}^{N} P(b|x_i,\mu_b,V_b)}$ **(N data points)**

EM Details

- **E-step**
  - Compute probability that point $\boldsymbol{x}$ is in blob *b*, given current guess of $\boldsymbol{\theta}$

$$P(b|x, \mu_b, V_b) = \frac{\alpha_b P(x|\mu_b, V_b)}{\sum_{i=1}^{K} \alpha_i P(x|\mu_i, V_i)}$$

- **M-step**
  - Compute probability that blob *b* is selected

$$\alpha_b^{new} = \frac{1}{N} \sum_{i=1}^{N} P(b|x_i, \mu_b, V_b)$$

  - Mean of blob *b* $\quad \mu_b^{new} = \frac{\sum_{i=1}^{N} x_i P(b|x_i, \mu_b, V_b)}{\sum_{i=1}^{N} P(b|x_i, \mu_b, V_b)}$ **(N data points)**
  - Covariance of blob *b*

EM Details

- **E-step**
  - Compute probability that point $\boldsymbol{x}$ is in blob *b*, given current guess of $\boldsymbol{\theta}$

$$P(b|x, \mu_b, V_b) = \frac{\alpha_b P(x|\mu_b, V_b)}{\sum_{i=1}^{K} \alpha_i P(x|\mu_i, V_i)}$$

- **M-step**
  - Compute probability that blob *b* is selected

$$\alpha_b^{new} = \frac{1}{N} \sum_{i=1}^{N} P(b|x_i, \mu_b, V_b)$$

  - Mean of blob *b* $\mu_b^{new} = \frac{\sum_{i=1}^{N} x_i P(b|x_i, \mu_b, V_b)}{\sum_{i=1}^{N} P(b|x_i, \mu_b, V_b)}$ **(N data points)**

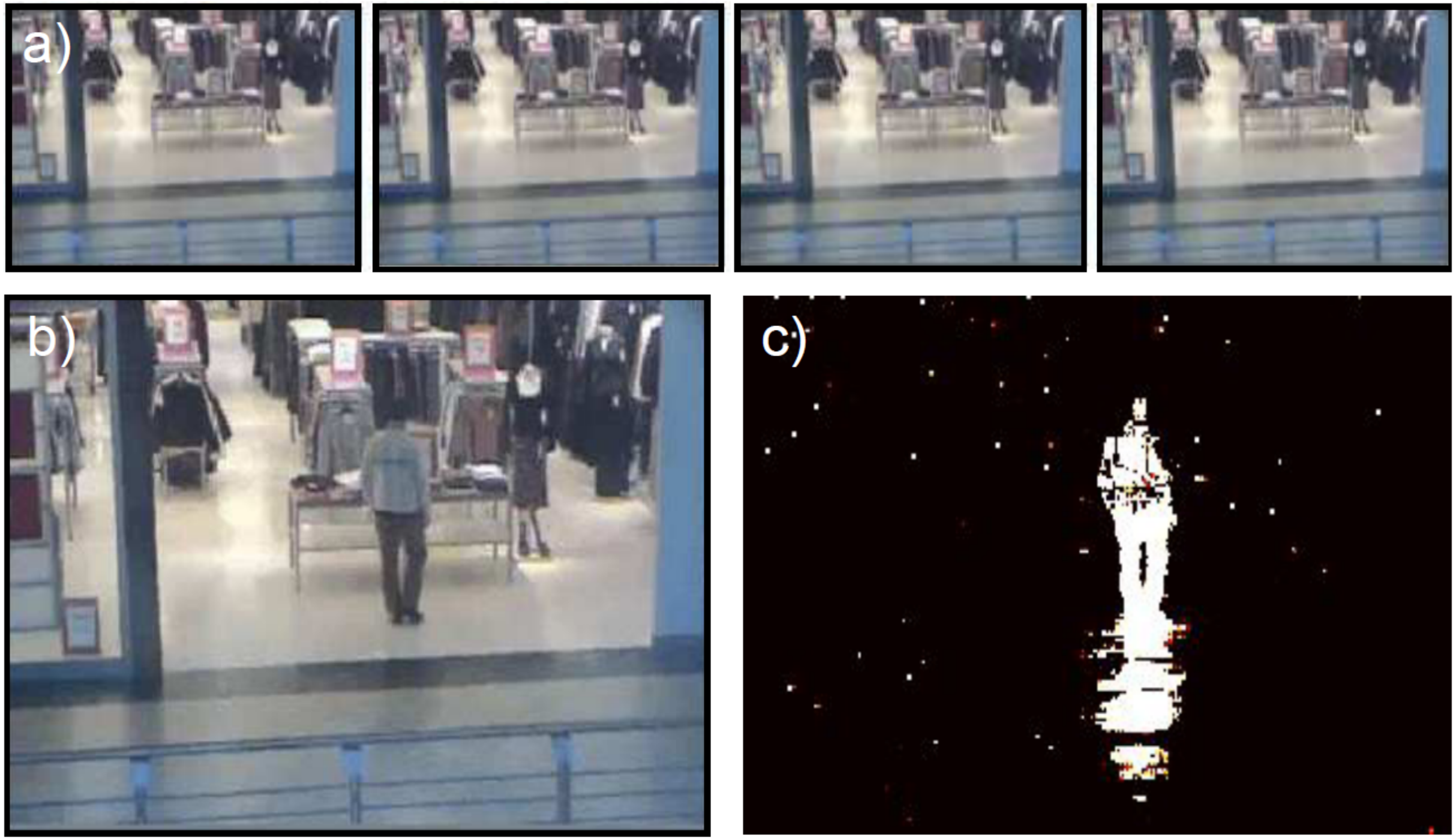
  - Covariance of blob *b*

$$V_b^{new} = \frac{\sum_{i=1}^{N} (x_i - \mu_b^{new})(x_i - \mu_b^{new})^T P(b|x_i, \mu_b, V_b)}{\sum_{i=1}^{N} P(b|x_i, \mu_b, V_b)}$$

Applications of EM

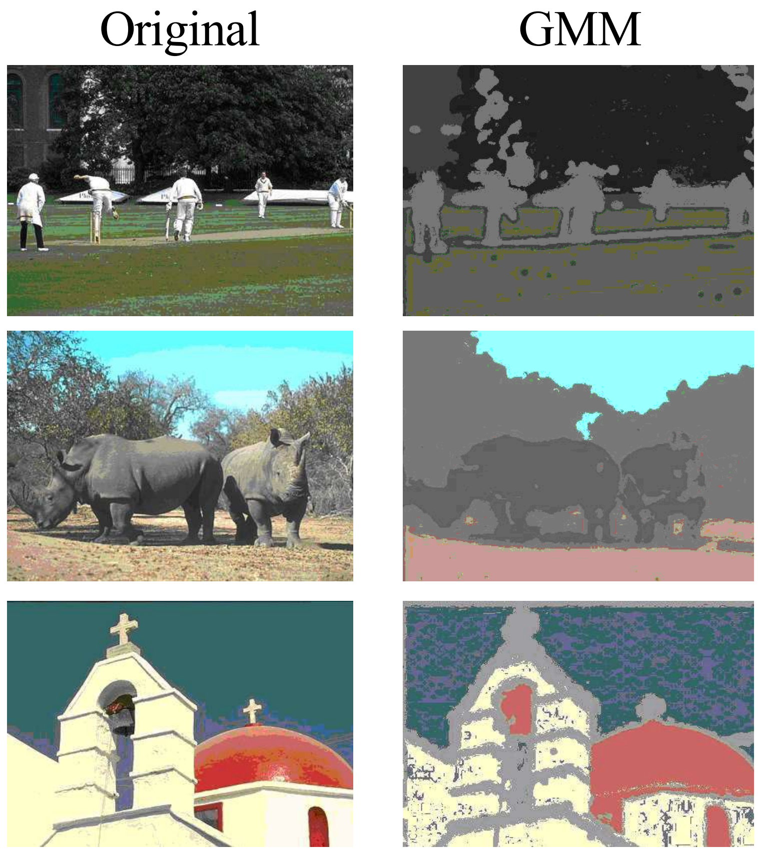
- **Turns out this is useful for all sorts of problems**
  - Any clustering problem
  - Any model estimation problem
  - Missing data problems
  - Finding outliers
  - Segmentation problems
    - Segmentation based on color
    - Segmentation based on motion
    - Foreground/background separation
  - ...

Application: Background subtraction



slide credit: Simon J.D. Prince

Segmentation with EM



Segmentation examples using the GMM for K = 5 components.

Sfikas et al, IEEE ICIP 2007

Summary: Mixtures of Gaussians, EM

- **Pros**
  - Probabilistic interpretation
  - Soft assignments between data points and clusters
  - Generative model, can predict novel data points
  - Relatively compact storage

Summary: Mixtures of Gaussians, EM

- **Pros**
  - Probabilistic interpretation
  - Soft assignments between data points and clusters
  - Generative model, can predict novel data points
  - Relatively compact storage
- **Cons**
  - Local minima

Summary: Mixtures of Gaussians, EM

- **Pros**
  - Probabilistic interpretation
  - Soft assignments between data points and clusters
  - Generative model, can predict novel data points
  - Relatively compact storage
- **Cons**
  - Local minima
  - Initialization
    - Often a good idea to start with some k-means iterations.

Summary: Mixtures of Gaussians, EM

- **Pros**
  - Probabilistic interpretation
  - Soft assignments between data points and clusters
  - Generative model, can predict novel data points
  - Relatively compact storage
- **Cons**
  - Local minima
  - Initialization
    - Often a good idea to start with some k-means iterations.
  - Need to know number of components

Summary: Mixtures of Gaussians, EM

- **Pros**
  - Probabilistic interpretation
  - Soft assignments between data points and clusters
  - Generative model, can predict novel data points
  - Relatively compact storage
- **Cons**
  - Local minima
  - Initialization
    - Often a good idea to start with some k-means iterations.
  - Need to know number of components
  - Need to choose generative model

Summary: Mixtures of Gaussians, EM

- **Pros**
  - Probabilistic interpretation
  - Soft assignments between data points and clusters
  - Generative model, can predict novel data points
  - Relatively compact storage
- **Cons**
  - Local minima
  - Initialization
    - Often a good idea to start with some k-means iterations.
  - Need to know number of components
  - Need to choose generative model
  - Numerical problems are often a nuisance

Next Lecture

- **Segmentation and grouping**
  - Gestalt principles
  - Image segmentation
- **Segmentation as clustering**
  - k-Means
  - Feature spaces
- **Probabilistic clustering**
  - Mixture of Gaussians, EM
- **Model-free clustering**
  - Mean-Shift clustering
- **Graph theoretic segmentation**
  - Normalised cuts

Reading

Reading

- Forsyth, Ponce: Chapter 14