EECS 545: Machine Learning

Lecture 16. Unsupervised Learning: EM & PCA

Honglak Lee and Michał Dereziński

3/09/2022

Outline

Unsupervised Learning for Clustering:

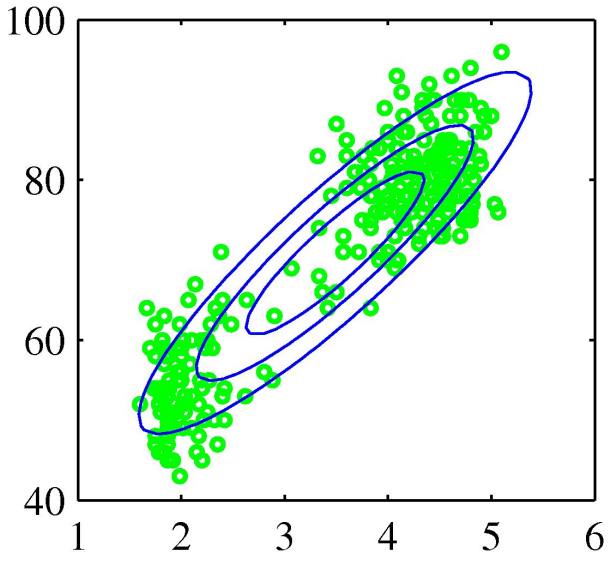
- Recap: K-Means
- Recap: Expectation Maximization
- Gaussian Mixtures with EM

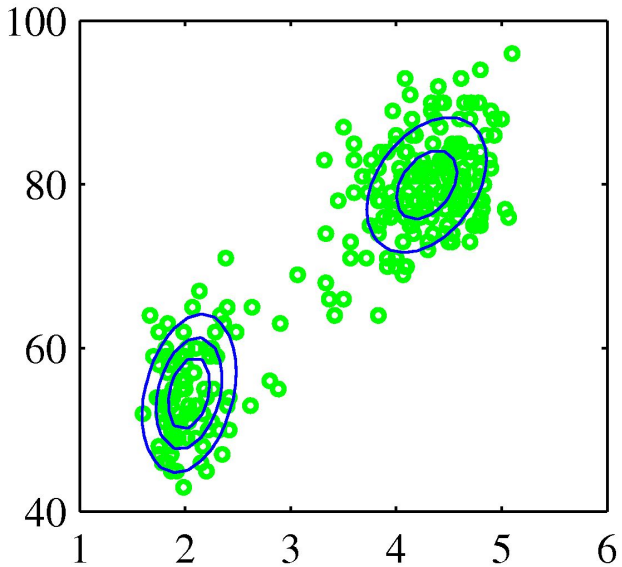
Unsupervised Learning for Finding Subspace:

- Principal Component Analysis

Clustering Pixels

- How do we find clusters of pixels?





The K-Means Algorithm

- Recap: Minimize *J*; sum of squared distance of points from the center of its own cluster.

$$J = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \|\mathbf{x}^{(n)} - \mu_k\|^2$$

- $r_{nk} \in \{0,1\}$ is a indicator variable.
  - $\boldsymbol{r_{nk}} = 1$ if $\mathbf{x}^{(n)}$ is in cluster k and $\boldsymbol{r_{nj}} = 0$ for all $j \neq k$.
- EM: repeat the following updates until convergence

1. $r := \arg\min_r J(r, \mu)$
2. $\mu := \arg\min_\mu J(r, \mu)$

The K-Means Algorithm

- Set the cluster centers arbitrarily.
- Repeat until convergence:
  - **Cluster assignment ("E-Step"): assign each point to closest center.**

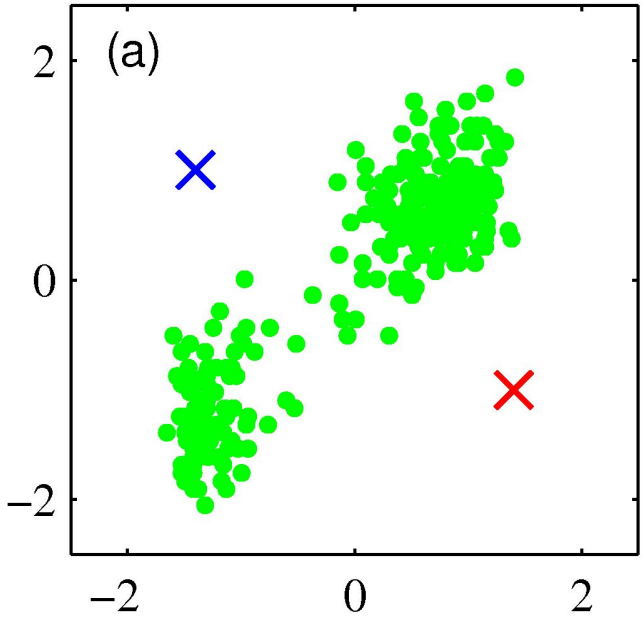
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}^{(n)} - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

  - **Parameter update ("M-Step"): update the centers**

$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}^{(n)}}{\sum_n r_{nk}}$$

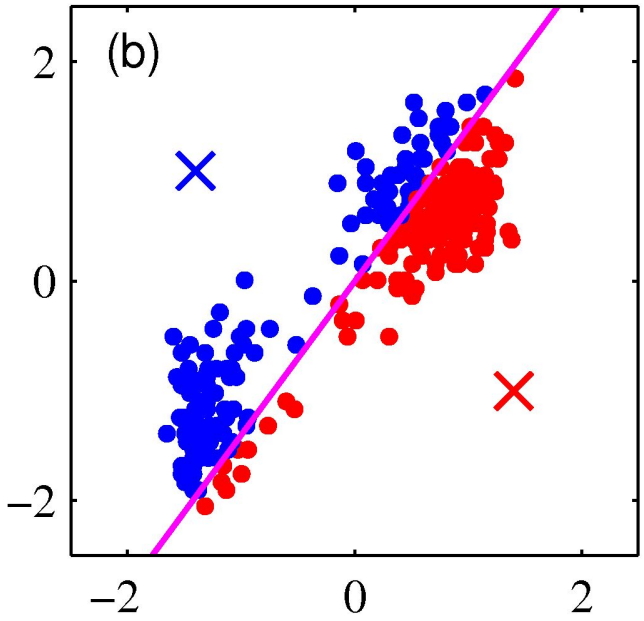
K-Means Clustering

- Select K. Pick random means.
  - Here K=2.



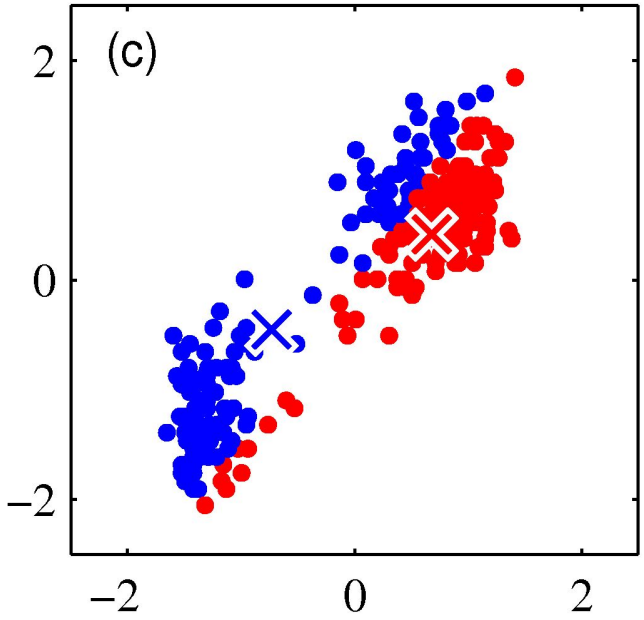
The E Step

- Assign each point to the nearest center.



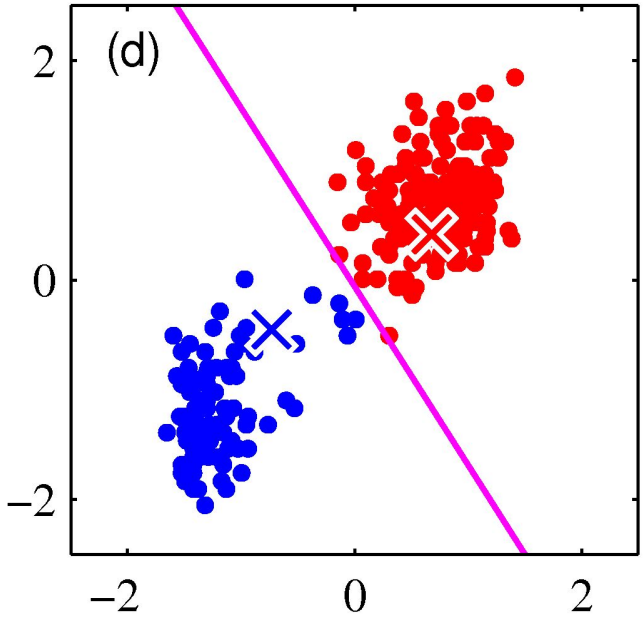
The M-Step

- Compute new centers for each cluster.



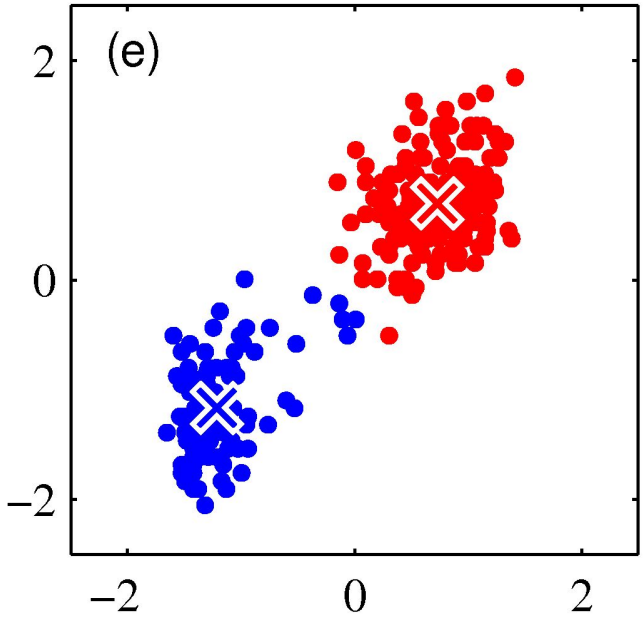
The E-Step Again

- Re-assign points to the now-nearest center.



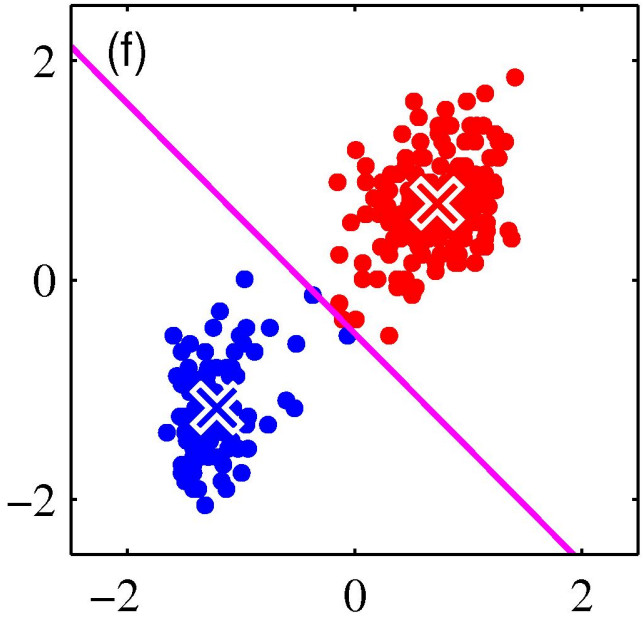
The M-Step Again

- Compute centers for the new clusters.



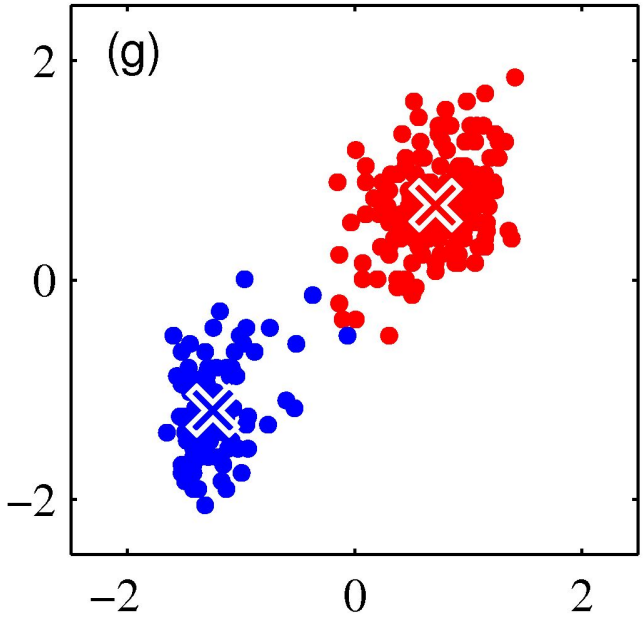
Another E-Step

- Reassign the pixels to centers.



Another M-Step

- New centers.



Expectation Maximization

Expectation Maximization

- Parameter learning when the data is not fully observed.
  - Suppose that we have observed variables X, and hidden variables Z
- Main idea:

  (E-step) Run inference about Z given X: Q=P(Z|X)

  (M-step) Update parameters by treating Q as observation!
- Example:
  - K-Means (a special case of Gaussian mixtures)
  - Gaussian mixtures

Recap: The EM Algorithm in General

- We have shown that:

$$\begin{aligned}\log p(\mathbf{X}|\theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})} + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\theta)} \\ &= \mathcal{L}(q,\theta) + KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X})) \\ &\geq \mathcal{L}(q,\theta)\end{aligned}$$

Evidence Lower bound (ELBO) or variational lower bound

with equality holding if and only if

$$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X})$$

- For a fixed θ, what is the *q* that maximizes *L(q, θ)*?
- $p(\mathbf{Z}|\mathbf{X})$ because all other *q* result in strictly less than log $p(\mathbf{X}|\theta)$.

Recap: The EM Algorithm in General

- We also note that for a fixed q, $L(q, \theta)$ can be decomposed into two terms:
  - A weighted sum of log $p(\mathbf{X}, \mathbf{Z}|\theta)$. This is tractable and can be optimized wrt θ
  - Entropy of $q(\mathbf{Z})$ which is independent of θ since q is fixed.

$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{\log p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})} \\ &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})\end{aligned}$$

Thus, we can find θ that maximize $L(q, \theta)$ when q is fixed.

The EM Algorithm

- Initialize random parameters θ
- Repeat until convergence:
  - "E - step": Set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$ ← Do this for each training sample
  - "M - step": Update θ via the following maximization

$$\text{argmax}_\theta \; \mathcal{L}(q, \theta) = \text{argmax}_\theta \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\theta)$$

Note: This is for a single example for conceptual simplicity of presentation.
We need to aggregate the lower bound for all the training samples. See next slide.

- Note we have assumed that $p(\mathbf{Z}|\mathbf{X}, \theta)$ is tractable (i.e., find exact posterior $p(\mathbf{Z}|\mathbf{X}, \theta)$) .

Q. What if its not?

The EM Algorithm (Multiple data-points)

- Variational lower bound for a single example **x**

$$\log p(\mathbf{x}|\theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}|\theta)}{q(\mathbf{z})} + KL(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}, \theta))$$

$$\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}|\theta)}{q(\mathbf{z})}$$

- Lower bound of the log-likelihood of the entire training data $\mathcal{D} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$

$$\log p(\mathcal{D}|\theta) = \sum_n \log p(\mathbf{x}^{(n)}|\theta) = \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}^{(n)}|\theta)}{q^{(n)}(\mathbf{z})} + \sum_n KL(q^{(n)}(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}^{(n)}, \theta))$$

$$\geq \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}^{(n)}|\theta)}{q^{(n)}(\mathbf{z})}$$

The EM Algorithm (Multiple data-points)

$$\log p(\mathcal{D}|\theta) = \sum_n \log p(\mathbf{x}^{(n)}|\theta) = \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}^{(n)}|\theta)}{q^{(n)}(\mathbf{z})} + \sum_n KL(q^{(n)}(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}^{(n)}, \theta))$$

$$\geq \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}^{(n)}|\theta)}{q^{(n)}(\mathbf{z})}$$

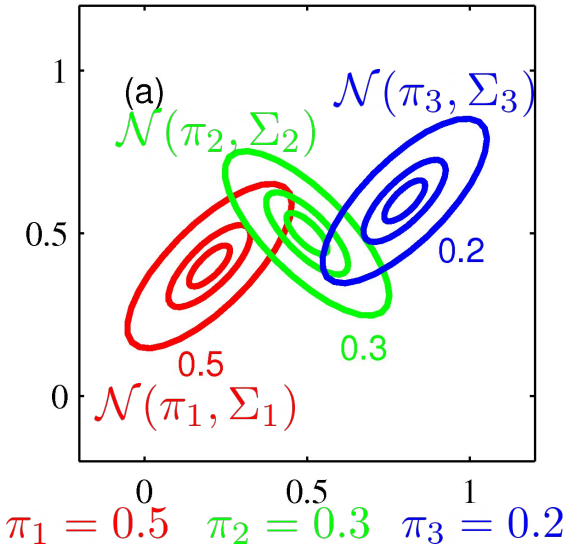
- Initialize random parameters θ
- Repeat until convergence:
  - "E - step": Set $q^{(n)}(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}^{(n)}, \theta)$, for each training sample n
  - "M - step": Update θ via the following maximization

$$\text{argmax}_\theta \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}^{(n)}|\theta)$$

Mixtures of Gaussians (recap)

- Let z in $\{0,1\}^K$ be a 1-of-K random variable;

$$p(z_k = 1) = \pi_k \qquad \sum_{k=1}^{K} \pi_k = 1$$



- Generate x from Gaussian given the selected cluster assignment z

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

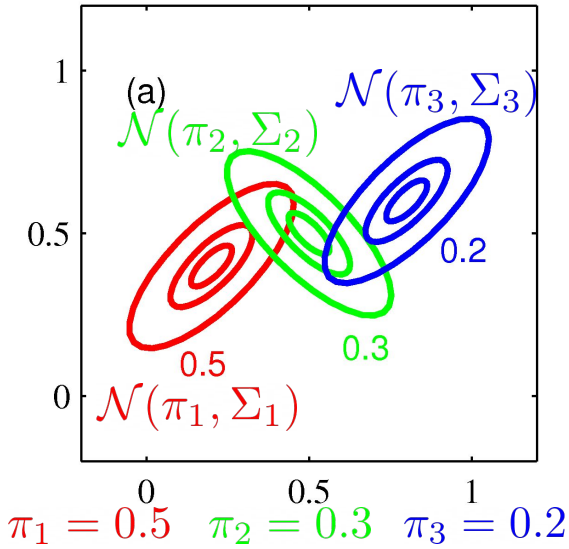
$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

Mixtures of Gaussians (recap)

- In other words, generate (sample) z then x:

$$p(z_k = 1) = \pi_k \qquad \sum_{k=1}^{K} \pi_k = 1$$

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



- Joint and marginal distributions:

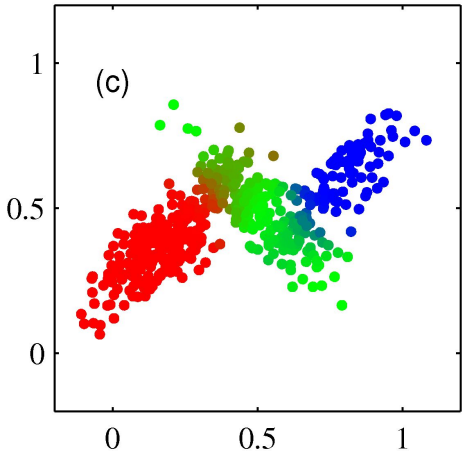
$$p(\mathbf{x}, \mathbf{z}) = \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

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

Mixtures of Gaussians: E-step

- Need to calculate $p(\mathbf{Z}|\mathbf{X})$ i.e., *soft assignments*
- Responsibility is the degree (posterior prob.) to which each Gaussian explains an observation x.

$$q^{(n)}(\mathbf{z}_k) = p(\mathbf{z}_k|\mathbf{x}^{(n)}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} =: \gamma(\mathbf{z}_{nk})$$



Q. Verify this! (Hint: Use Bayes Rule)

Mixtures of Gaussians: M-step

General formula for M-step:

$$\text{argmax}_\theta \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}^{(n)} | \theta)$$

Plug in for GMM: $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid k \in \{1 \dots K\}\}$

$$\text{argmax}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{s.t.} \sum_{k=1}^{K} \pi_k = 1$$

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

Mixtures of Gaussians: M-step

Let's first simplify the expression for J

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

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \left(\log \pi_k + \log \frac{1}{(2\pi)^{m/2}(\det \boldsymbol{\Sigma}_k)^{1/2}} - \frac{1}{2}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) \right)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \left((2\pi)^{m/2}(\det \boldsymbol{\Sigma}_k)^{1/2} \right)$$

$$- \frac{1}{2} \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2} \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \det \boldsymbol{\Sigma}_k$$

$$- \frac{1}{2} \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + const$$

Mixtures of Gaussians: M-step

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \det \boldsymbol{\Sigma}_k$$
$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + const$$

- Maximize J wrt $\boldsymbol{\mu}_k$ by differentiating wrt $\boldsymbol{\mu}_k$ and setting gradient to 0:

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) = 0$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \mathbf{x}^{(n)}}{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}$$

Mixtures of Gaussians: M-step

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \det \boldsymbol{\Sigma}_k$$
$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + const$$

- To find $\boldsymbol{\Sigma}_k$, we use change of variables: $\mathbf{M}_k = \boldsymbol{\Sigma}_k^{-1}$

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{M}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \det \mathbf{M}_k$$
$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T \mathbf{M}_k(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + const$$

Mixtures of Gaussians: M-step

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{M}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log\det \mathbf{M}_k$$
$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T \mathbf{M}_k (\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + const$$

- Maximize J wrt $\mathbf{M}_k$ by differentiating wrt $\mathbf{M}_k$ and setting gradient to 0:

$$\frac{\partial J}{\partial \mathbf{M}_k} = \frac{1}{2}\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})\mathbf{M}_k^{-1} - \frac{1}{2}\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T = 0$$

$$\boldsymbol{\Sigma}_k = \mathbf{M}_k^{-1} = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}$$

Mixtures of Gaussians: M-step

$$J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \det \boldsymbol{\Sigma}_k$$
$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k) + const$$

- Finally we need to $\max_{\boldsymbol{\pi}} \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k \quad \text{s.t.} \sum_{k=1}^{K} \pi_k = 1$

- Use Lagrange multipliers

$$L(\boldsymbol{\pi}, \alpha) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \alpha \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

Mixtures of Gaussians: M-step

- Finally we need to $\max_{\boldsymbol{\pi}} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k \quad \text{s.t.} \sum_{k=1}^{K} \pi_k = 1$
- Use Lagrange multipliers

$$L(\boldsymbol{\pi}, \alpha) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{nk}) \log \pi_k - \alpha \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

- Setting $\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \frac{1}{\pi_k} - \alpha = 0$ gives $\pi_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}{\alpha}$
- Using the constraint we get

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}{\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})} = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})}{N}$$

Mixtures of Gaussians: M-step

- The mean of a cluster is the weighted mean, weighted by the responsibilities.

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \mathbf{x}^{(n)}}{N_k}$$

  - N_k is the effective number of points in cluster k

$$N_k = \sum_{n=1}^{N} \gamma(\mathbf{z}_{nk}) \qquad \pi_k = \frac{N_k}{N}$$

- Likewise for covariance:

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(\mathbf{z}_{nk})(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k)^T}{N_k}$$

EM for Gaussian Mixtures

- Initialize means, covariances, and mixing coefficients for the K Gaussians.
- E Step: Given the coefficients, evaluate the responsibilities (or posterior of z_{nk}'s given $x^{(n)}$).

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(n)}|\mu_j, \Sigma_j)} = P(z_k = 1|\mathbf{x}^{(n)})$$

EM for Gaussian Mixtures

- M Step: Given the responsibilities, re-evaluate the coefficients.

$$\pi_k^{\text{new}} = \frac{N_k}{N} = \frac{\sum_n \gamma(z_{nk})}{N}$$

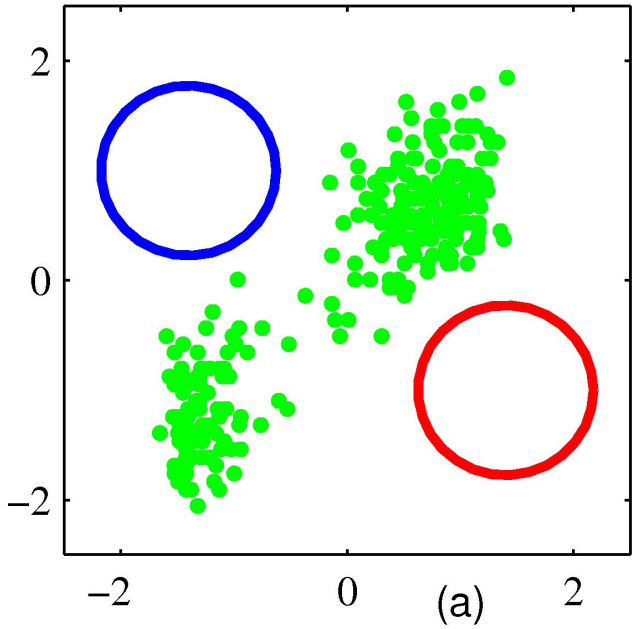
$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}^{(n)}$$

$$\sum_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}^{(n)} - \mu_k^{\text{new}})(\mathbf{x}^{(n)} - \mu_k^{\text{new}})^T$$

- Stop when either coefficients or log likelihood converges.

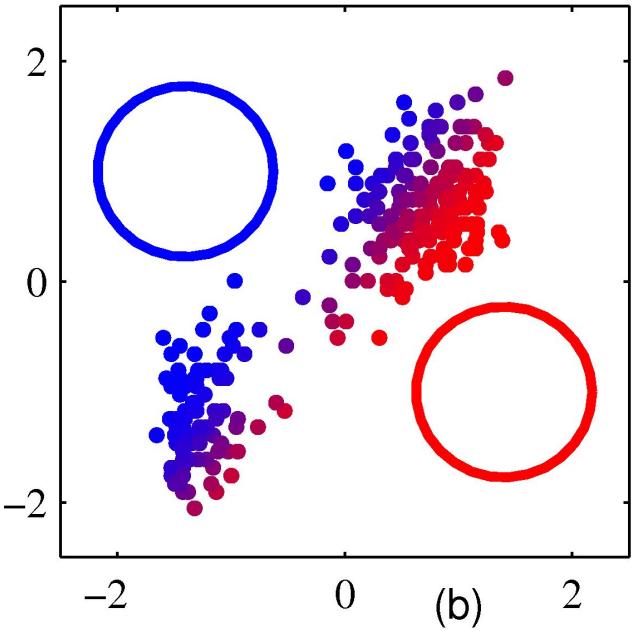
EM Example

- Initialize parameters: means, covariances, and mixing coefficients.



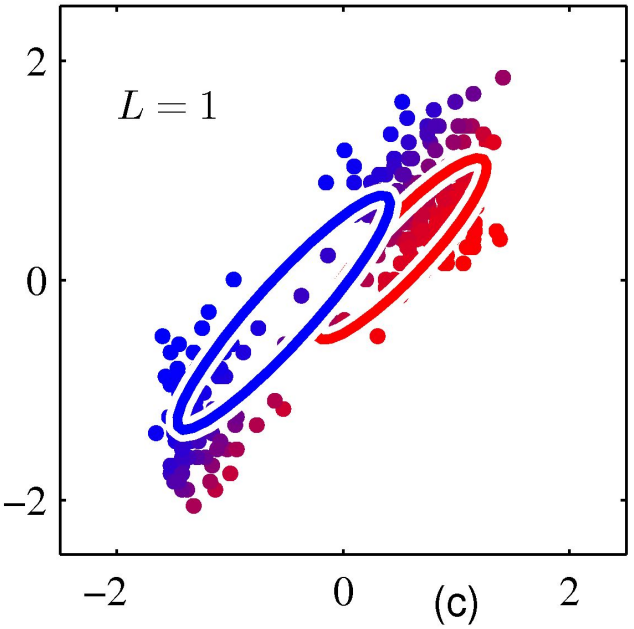
EM Example

- First E Step



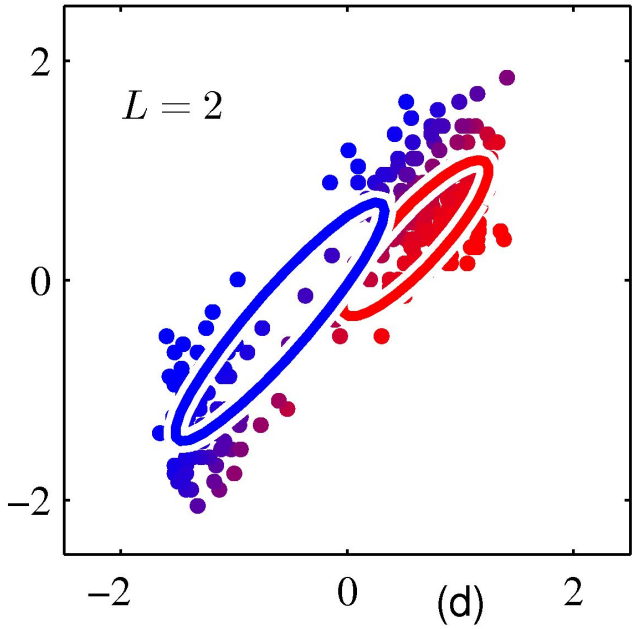
EM Example

- First M Step



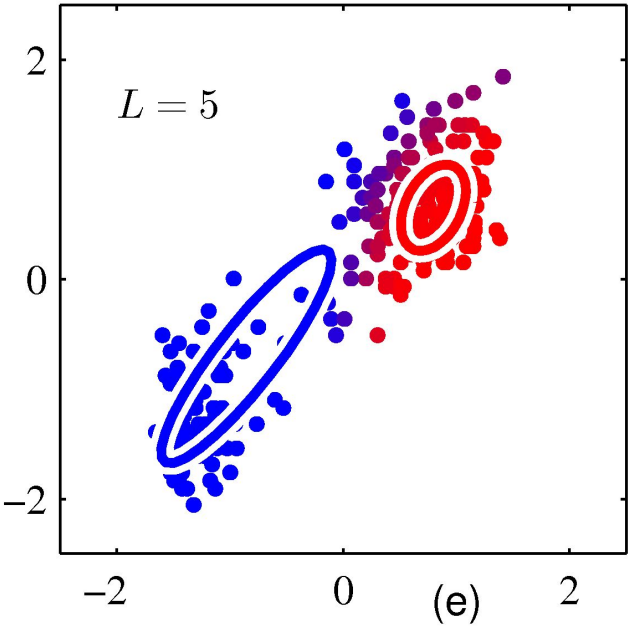
EM Example

- Second E and M Steps



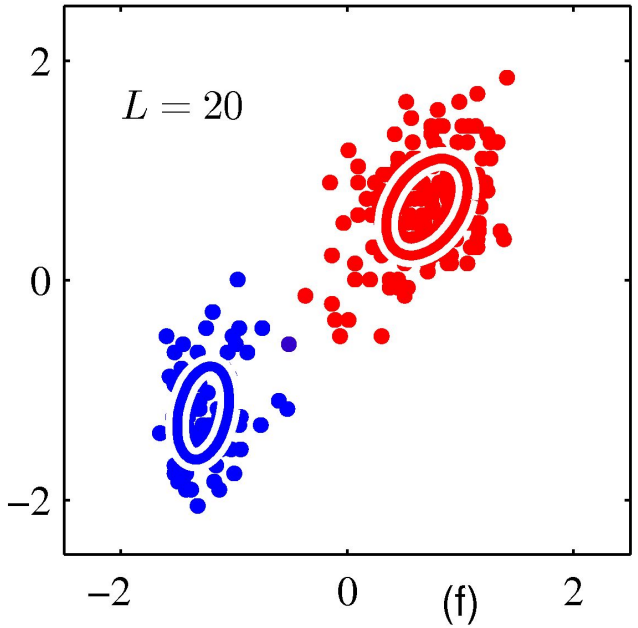
EM Example

- Three more E-M cycles



EM Example

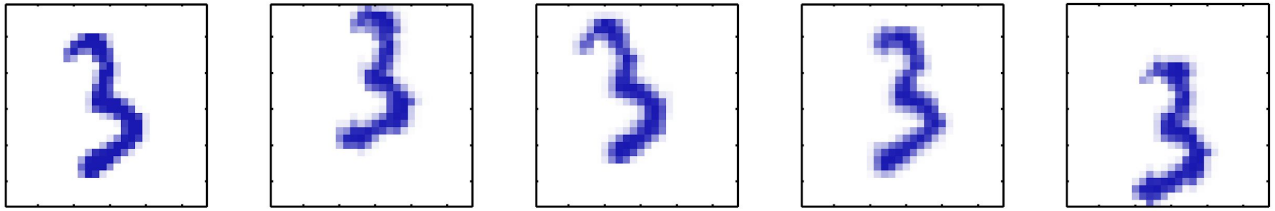
- Fifteen E-M cycles later



Principal Component Analysis

High-Dimensional Data

- . . . may have low-dimensional structure.



- The data is 100x100-dimensional.
- But there are only three degrees of freedom, so it lies on a 3-dimensional subspace.
  - (on a non-linear manifold, in this case)

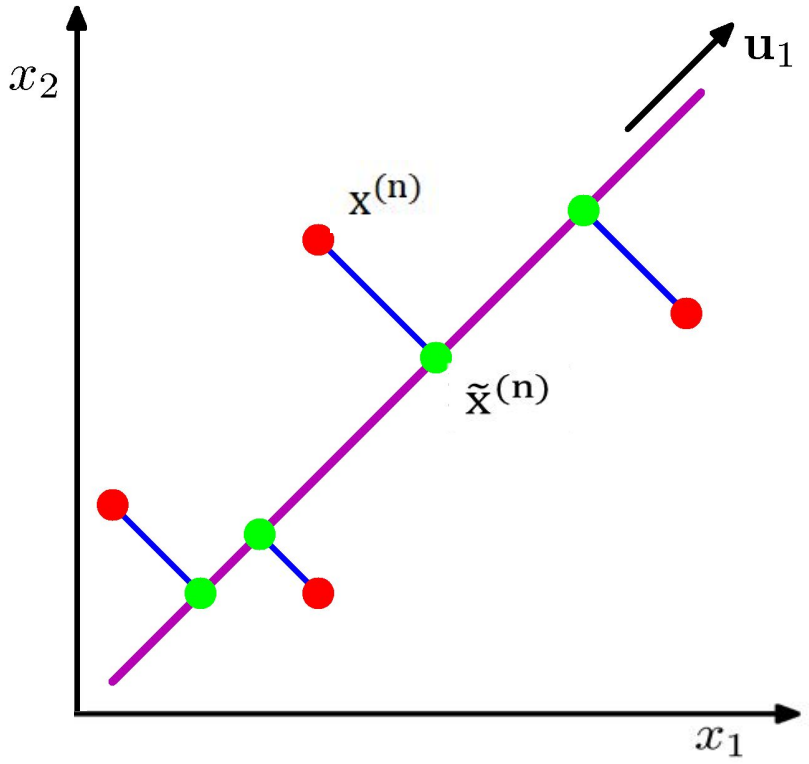
Principal Component Analysis

- Given a set **x** = $\{x^{(n)}\}$ of observations
  - in a space of dimension D,
  - find a subspace of dimension $M < D$
  - that captures most of its variability.
- PCA can be described as either:
  - maximizing the variance of the projection, or
  - minimizing the squared approximation error.

Two Descriptions of PCA

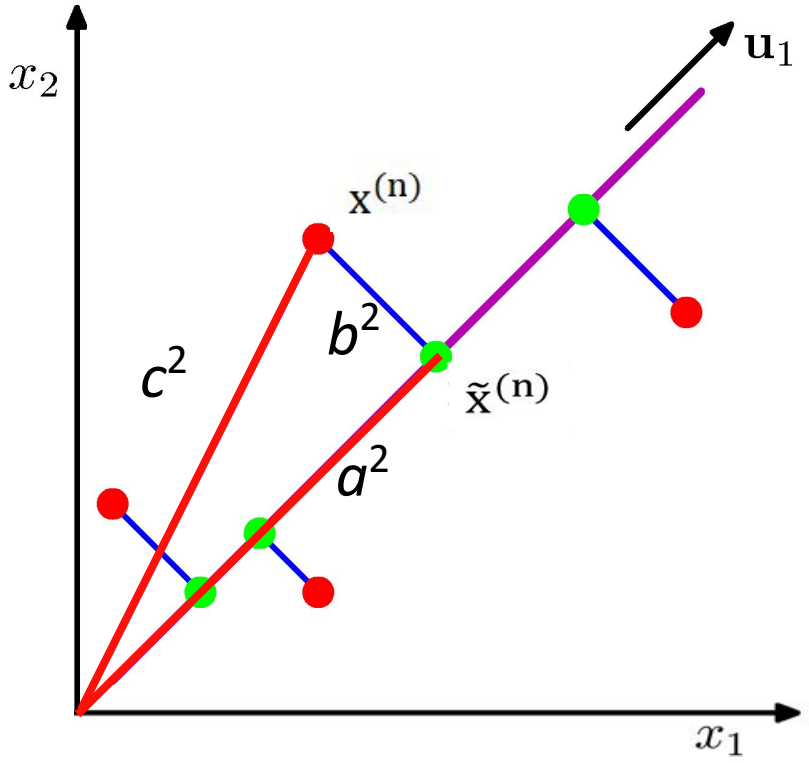
Approximate with the projection:

- Maximize variance, or
- Minimize squared error



Equivalent Descriptions

- With mean at the origin $c_i^2 = a_i^2 + b_i^2$
- With constant $\sum_i c_i^2$
  - Minimizing $\sum_i b_i^2$
  - Maximizes $\sum_i a_i^2$
  - And vice versa



First Principal Component

- Given data points $\{\mathbf{x}^{(n)}\}$ in D-dim space,
  - Mean $\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^{(n)}$
  - Data covariance $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}^{(n)} - \overline{\mathbf{x}})(\mathbf{x}^{(n)} - \overline{\mathbf{x}})^T$
    - D x D matrix
- Let u_1 be the principal component we want.
  - Length 1: $\mathbf{u}_1^T\mathbf{u}_1 = 1$
  - Projection of $\mathbf{x}^{(n)}$: $\mathbf{u}_1^T\mathbf{x}^{(n)}$

First Principal Component

- Maximize the projection variance:

$$\frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_1^T\mathbf{x}^{(n)} - \mathbf{u}_1^T\overline{\mathbf{x}})^2 = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$$

- Use a Lagrange multiplier to enforce $\mathbf{u}_1^T\mathbf{u}_1 = 1$
- Maximize: $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T\mathbf{u}_1)$

- Derivative is zero when $\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$
  - That is $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 = \lambda_1$
- So u_1 is eigenvector with largest eigenvalue.

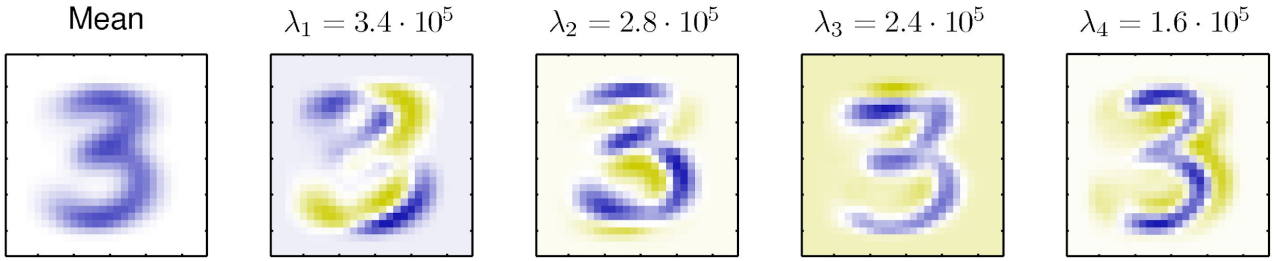
PCA by Maximizing Variance

- Repeat to find the *M* eigenvectors of the data covariance matrix S corresponding to the *M* largest eigenvalues.

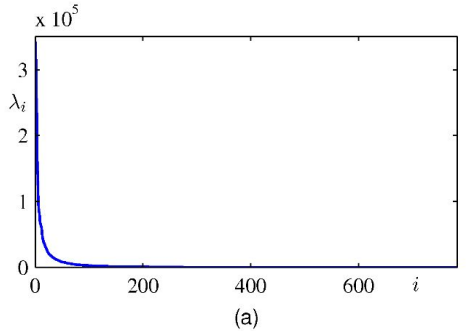
- We can do the same thing by minimizing the squared error of the projection.

Digit Image Example

- The mean and first four PCA eigenvectors.

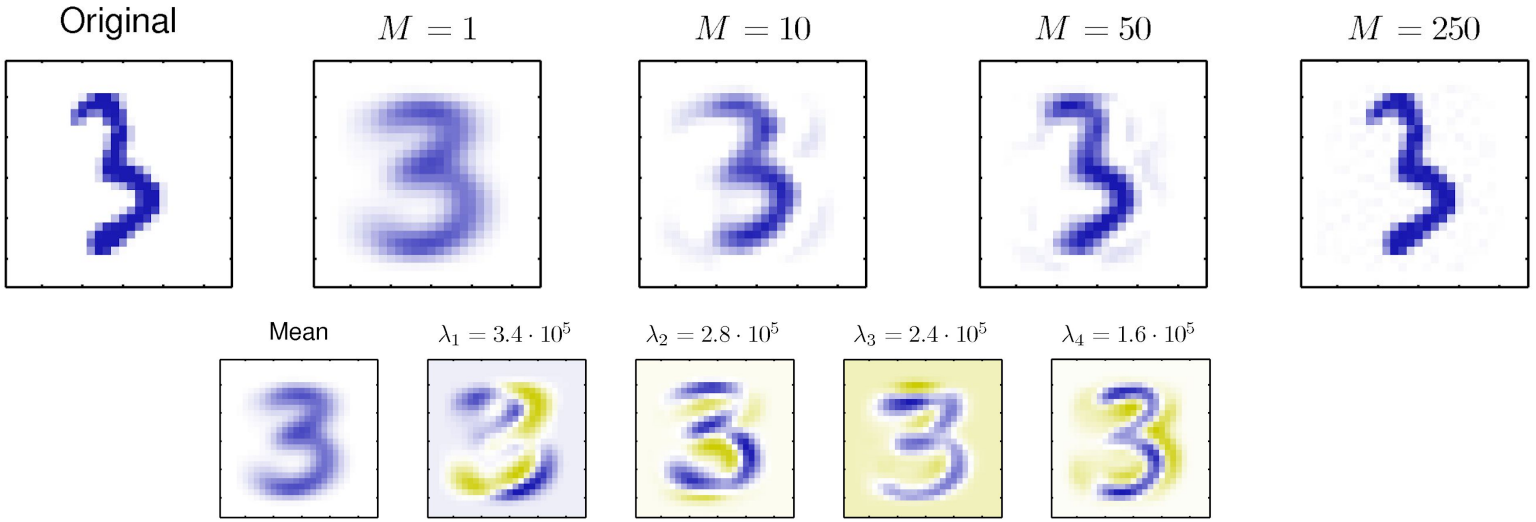


- The eigenvalue spectrum:



Reconstructing the Image

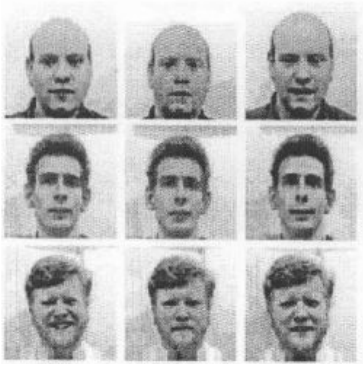
- Compress the image representation by using only first M eigenvectors, and discarding the less important information.



Learning features via PCA

- Example: Eigenfaces

Training face images



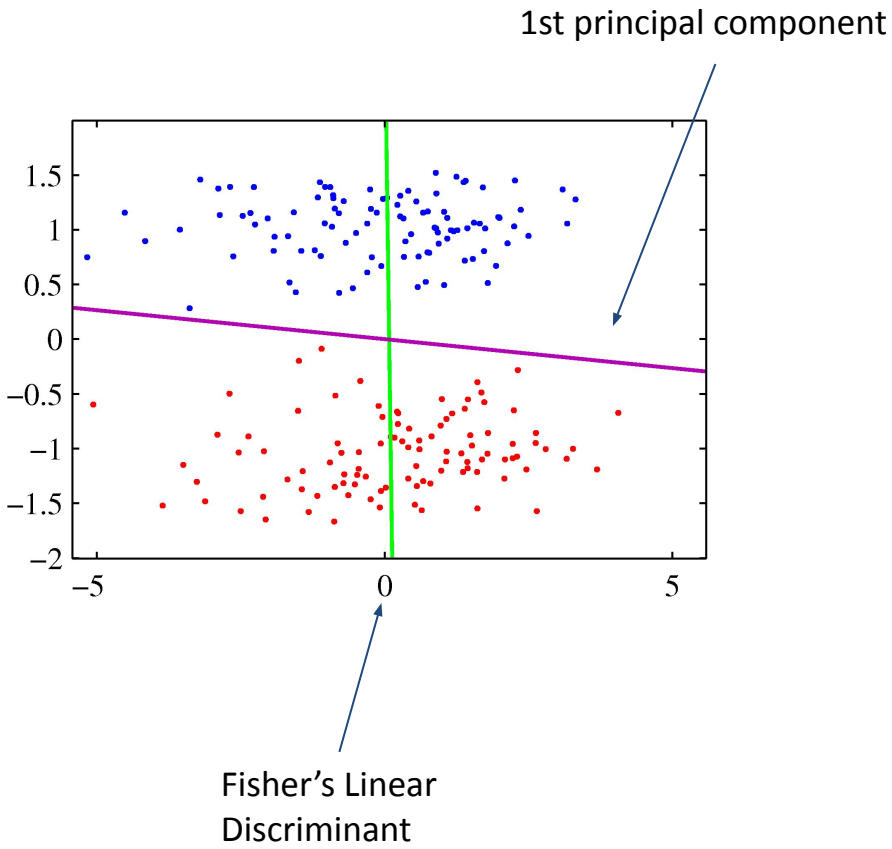
Learned PCA bases

Test example

Images from www.cse.unr.edu/~bebis/CS485/Lectures/Eigenfaces.ppt

Limits to PCA

- Maximizing variance is not always the best way to make the structure visible.
- PCA vs Fisher's linear discriminant



Probabilistic PCA

- We can view PCA as solving a probabilistic latent variable problem.
- Describe a distribution *p*(x) in *D*-dimensional space, in terms of a latent variable z in *M*-dimensional space.

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mu + \epsilon \qquad p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \qquad \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

- W is a *D* by *M* linear transformation from z to x

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I})$$

Probabilistic PCA

- Given the generative model

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mu + \epsilon$$

- we can infer

$$E[\mathbf{x}] = E[(\mathbf{W}\mathbf{z} + \mu + \epsilon)] = \mu$$

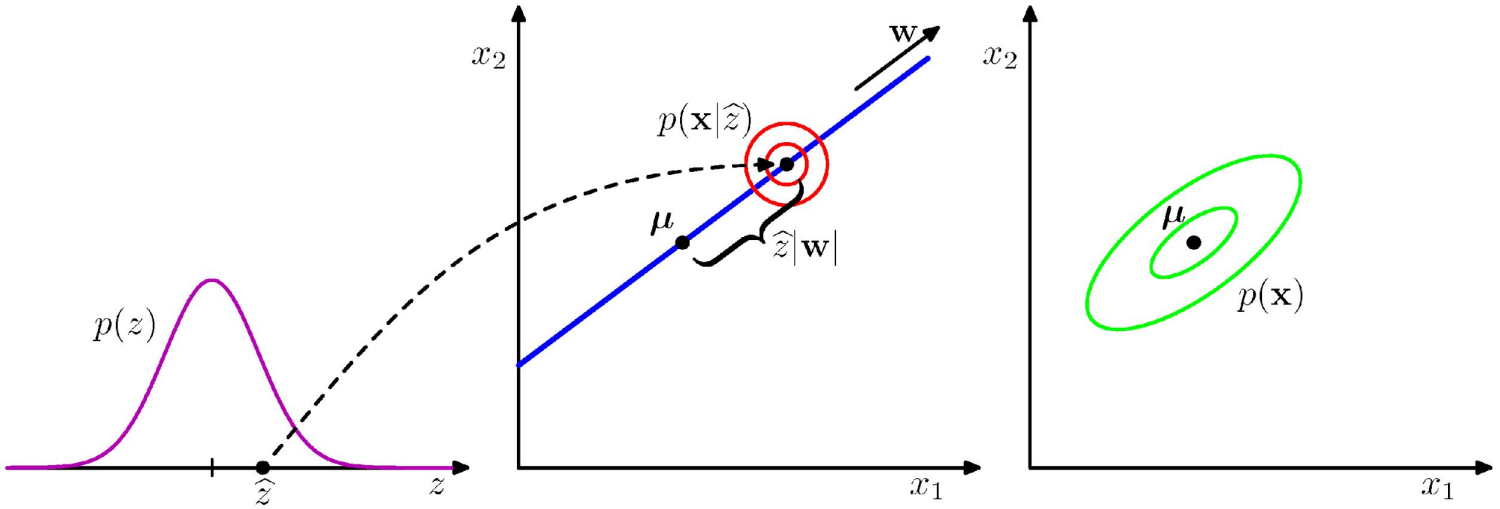
$$\begin{aligned} cov[\mathbf{x}] &= E[(\mathbf{W}\mathbf{z} + \epsilon)(\mathbf{W}\mathbf{z} + \epsilon)^T] \\ &= E[\mathbf{W}\mathbf{z}\mathbf{z}^T\mathbf{W}^T] + E[\epsilon\epsilon^T] = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I} \end{aligned}$$

Probabilistic PCA

- The generative model

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mu + \epsilon$$

can be illustrated



Likelihood of Probabilistic PCA

- (Marginal) likelihood

$$\ln p(\mathbf{x}|\mathbf{W},\mu,\sigma^2) = \sum_i p\left(x^{(i)}\middle|\mathbf{W},\mu,\sigma^2\right)$$

$$= -\frac{ND}{2}\ln 2\pi - \frac{N}{2}\ln|C| - \frac{1}{2}\sum_i \left(x^{(i)}-\mu\right)^T C^{-1}\left(x^{(i)}-\mu\right)$$

$$where\ C = \mathbf{W}\mathbf{W}^T + \sigma^2 I$$

- We can simply maximize this likelihood function with respect to $\mathbf{W}, \mu, \sigma$.

Maximum Likelihood Parameters

- Mean: $\mu = \bar{\mathbf{x}}$
- Noise: $\sigma^2_{ML} = \frac{1}{D-M} \sum_{i=M+1}^{D} \lambda_i$
- W: $\mathbf{W}_{ML} = \mathbf{U}_M(\mathbf{L}_M - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$

where

- $\mathbf{L}_M$ is diag with the *M* largest eigenvalues
- $\mathbf{U}_M$ is the M corresponding eigenvectors
- **R** is an arbitrary *M* by *M* orthogonal matrix (rotation matrix) (i.e., **z** can be defined by rotating "back")

Maximum likelihood by EM

- Latent variable model

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$

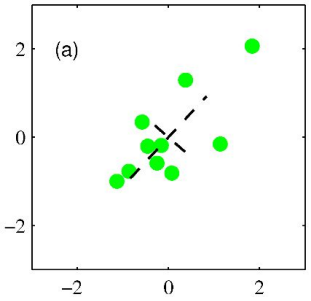
$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I})$$

- E-step: Estimate the posterior Q(z)=P(z|x)
  - Use linear Gaussian
- M-step: Maximize the data-completion likelihood given Q(z):

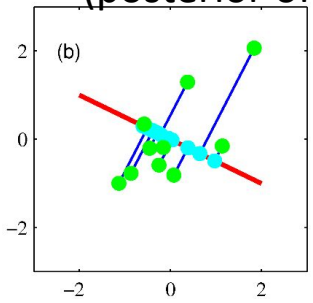
$$\text{maximize}_{\theta=\{\mathbf{W},\mu,\sigma\}} \sum_i \sum_{z^{(i)}} Q(z^{(i)}) \log P_\theta(x^{(i)}, z^{(i)})$$

Finding PCA params by EM

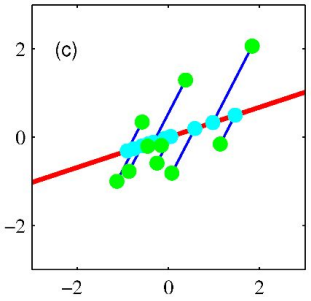
Data points



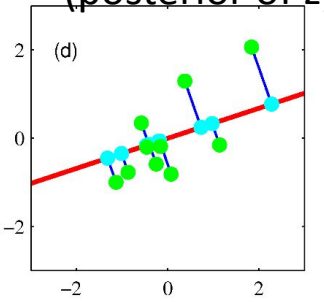
E-step: projection (posterior of z)



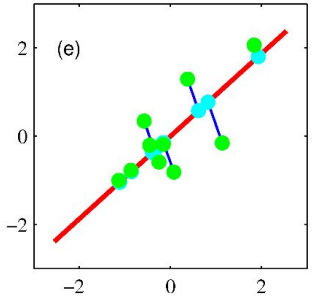
M-step: update W



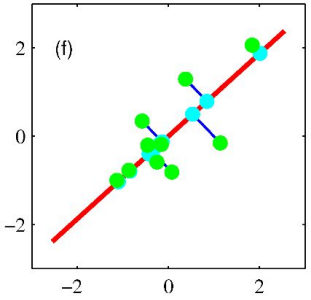
E-step: projection (posterior of z)



M-step: update W



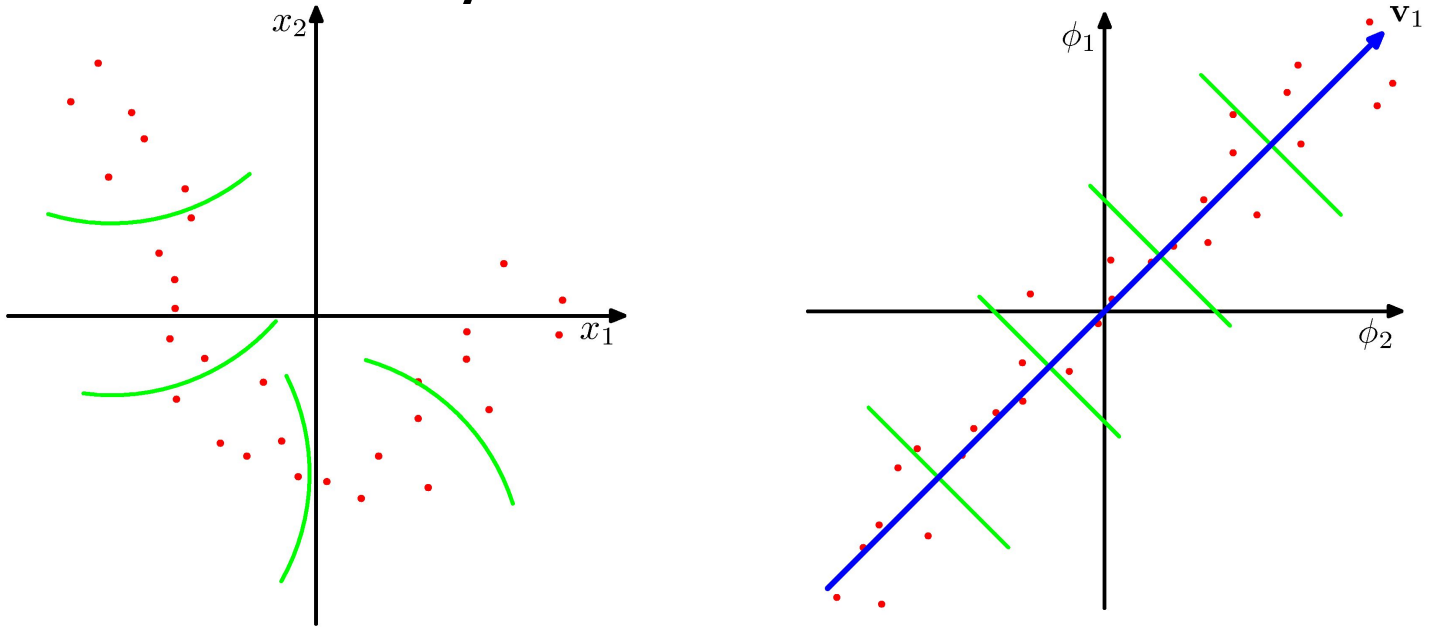
After convergence



- Illustrating EM on simulated data

Kernel PCA

- Suppose the regularity that allows dimensionality reduction is non-linear.



Kernel PCA

- As with regression and classification, we can transform the raw input data $\{\mathbf{x}^{(n)}\}$ to a set of feature values

$$\{\, \mathbf{x}^{(n)} \,\} \rightarrow \{\; \phi(\mathbf{x}^{(n)}) \,\}$$

- Linear PCA (on the nonlinear feature space) gives us a linear subspace in the feature value space, corresponding to nonlinear structure in the data space.

Kernel PCA

- Define a kernel, to avoid having to evaluate the feature vectors explicitly.

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$$

- Express PCA in terms of the kernel,
  - Some care is required to centralize the data.

$$K_{nm} = \phi(\mathbf{x}^{(n)})^T \phi(\mathbf{x}^{(m)}) = k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)})$$

Kernel PCA

- Assume that $\{\phi(\mathbf{x}^{(n)})\}$ have zero mean.
- Sample covariance matrix: $S = \frac{1}{N}\sum_{n=1}^{N} \phi(\mathbf{x}^{(n)})\phi(\mathbf{x}^{(n)})^T = \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}$
- Let **v** be an eigenvector for S

$$S\mathbf{v} = \lambda\mathbf{v} \implies \lambda\mathbf{v} = \mathbf{\Phi}^T\left(\frac{1}{N}\mathbf{\Phi}\mathbf{v}\right)$$

$$\therefore \mathbf{v} = \mathbf{\Phi}^T\boldsymbol{\alpha} \quad \text{for some } \boldsymbol{\alpha} \in \mathbb{R}^N$$

- Thus, $\lambda\mathbf{v} = S\mathbf{v} \implies \lambda\mathbf{\Phi}^T\boldsymbol{\alpha} = \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Phi}^T\boldsymbol{\alpha} = \frac{1}{N}\mathbf{\Phi}^T K\boldsymbol{\alpha}$
- Multiply Φ on both sides and cancel out $K=\Phi\Phi^T$

$$\lambda N\boldsymbol{\alpha} = K\boldsymbol{\alpha} \quad \rightarrow \alpha \text{ is a eigenvector of K}$$

Kernel PCA

- We thus have $v = \Phi^T \alpha$, where α is eigenvector of kernel matrix K.
- Now, $\|\mathbf{v}\| = 1 \implies \boldsymbol{\alpha}^T K \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \lambda_K \boldsymbol{\alpha} = 1 \implies \|\boldsymbol{\alpha}\| = \lambda_K^{-1/2}$
- It is often infeasible to obtain v (depends on dim of Φ), but we can compute projections:

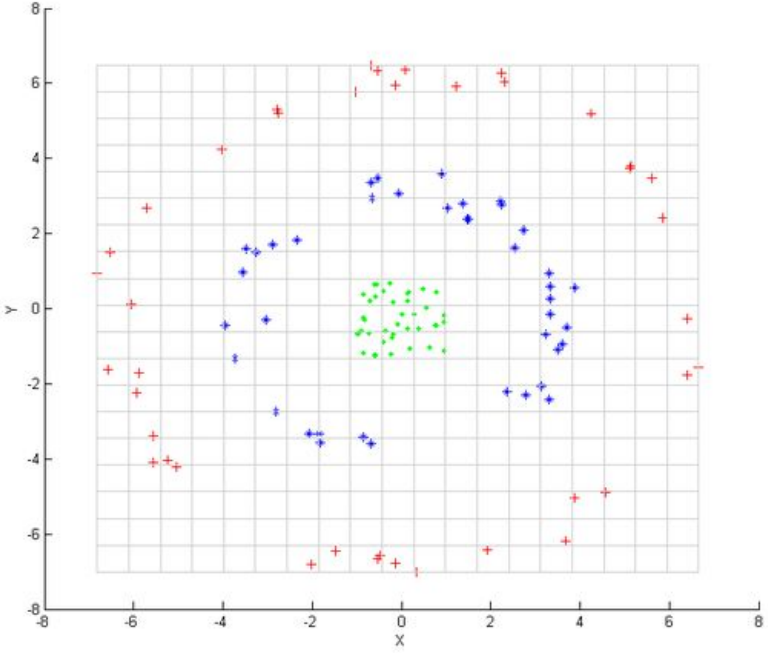
$$\mathbf{v}^T \phi(\mathbf{x}) = \boldsymbol{\alpha}^T \boldsymbol{\Phi} \phi(\mathbf{x}) = \boldsymbol{\alpha}^T k(\mathbf{x}) \text{ where } k(\mathbf{x}) = \begin{bmatrix} k(\mathbf{x}^{(1)}, \mathbf{x}) & \dots k(\mathbf{x}^{(N)}, \mathbf{x}) \end{bmatrix}$$

- Finally, some care is required to centralize data (to ensure that features have zero mean):

$$K' = K - \mathbf{1}_N K - K \mathbf{1}_N + \mathbf{1}_N K \mathbf{1}_N$$
$$\text{where} \quad \mathbf{1}_N \in \mathbb{R}^{N \times N} \quad \text{is a matrix of ones.}$$

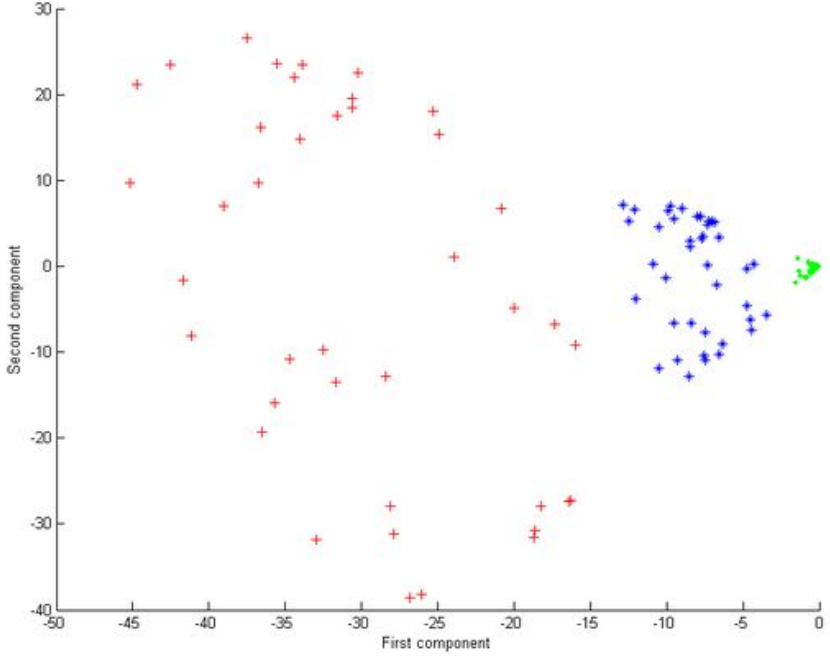
Kernel PCA

Linear PCA operates only in the given (in this case two-dimensional) space, in which these concentric point clouds are not linearly separable.



Data

The first principal component is enough to distinguish the three different groups

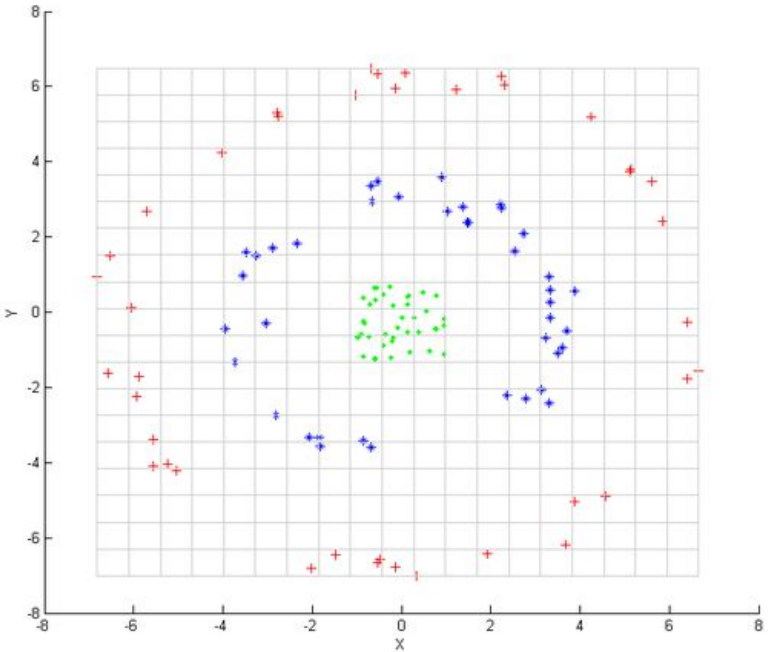


Kernel PCA with

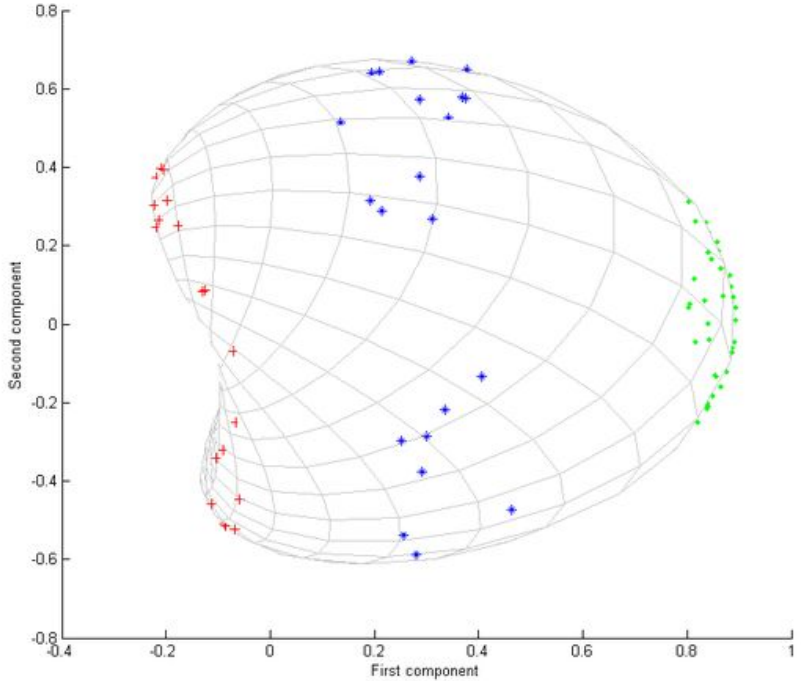
$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^2$$

Source: https://en.wikipedia.org/wiki/Kernel_principal_component_analysis

Kernel PCA



Data



Kernel PCA with Gaussian kernel

Source: https://en.wikipedia.org/wiki/Kernel_principal_component_analysis

Thank you!

Quiz: Click here or scan QR code

Next class: Generative models