ECE 5424: Introduction to Machine Learning

Topics:

- (Finish) Expectation Maximization
- Principal Component Analysis (PCA)

Readings: Barber 15.1-15.4

Stefan Lee

Virginia Tech

Project Poster

- Poster Presentation: **Best Project Prize!**
  - Dec 6th 1:30-3:30pm
  - Goodwin Hall Atrium
  - Print poster (or bunch of slides)
    - Fedex, Library, ECE support, CS support
  - Format:
    - Portrait, 2 feet (width) x 36 inches (height)
    - See https://filebox.ece.vt.edu/~f16ece5424/project.html
- Submit poster as PDF by Dec 6th 1:30pm
  - Makes up the final portion of your project grade

Final Exam

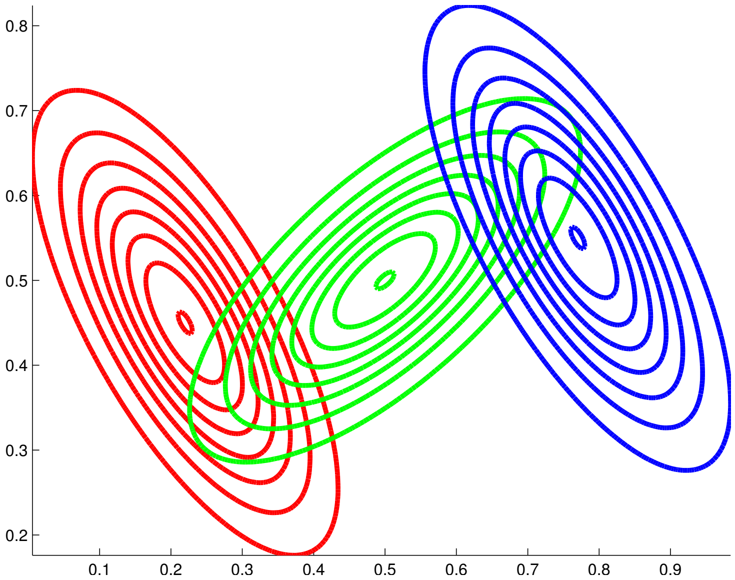
- Dec 14th in class (DURH 261); 2:05 - 4:05 pm
- Content:
  - Almost completely about material since the midterm
    - SVM, Neural Networks, Descision Trees, Ensemble Techniques, K-means, EM **(today),** Factor Analysis **(Thrusday)**
  - True/False (explain your choice like last time)
  - Multiple Choice
  - Some ‘Prove this’
  - Some ‘What would happen with algorithm A on this dataset”

Homework & Grading

- HW3 & HW4 should be graded this week
- Will release solutions this week as well

Recap of Last Time

GMM



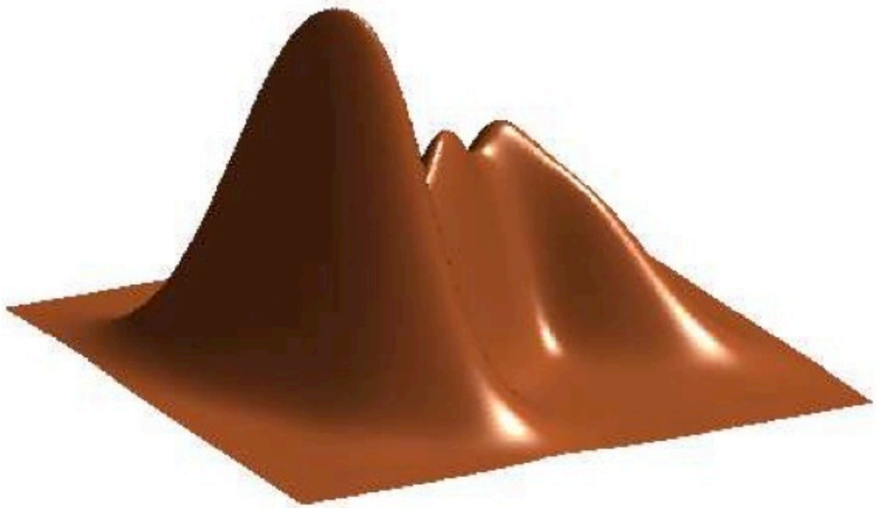


Figure Credit: Kevin Murphy

EM

- Expectation Maximization [Dempster ‘77]

- Often looks like “soft” K-means

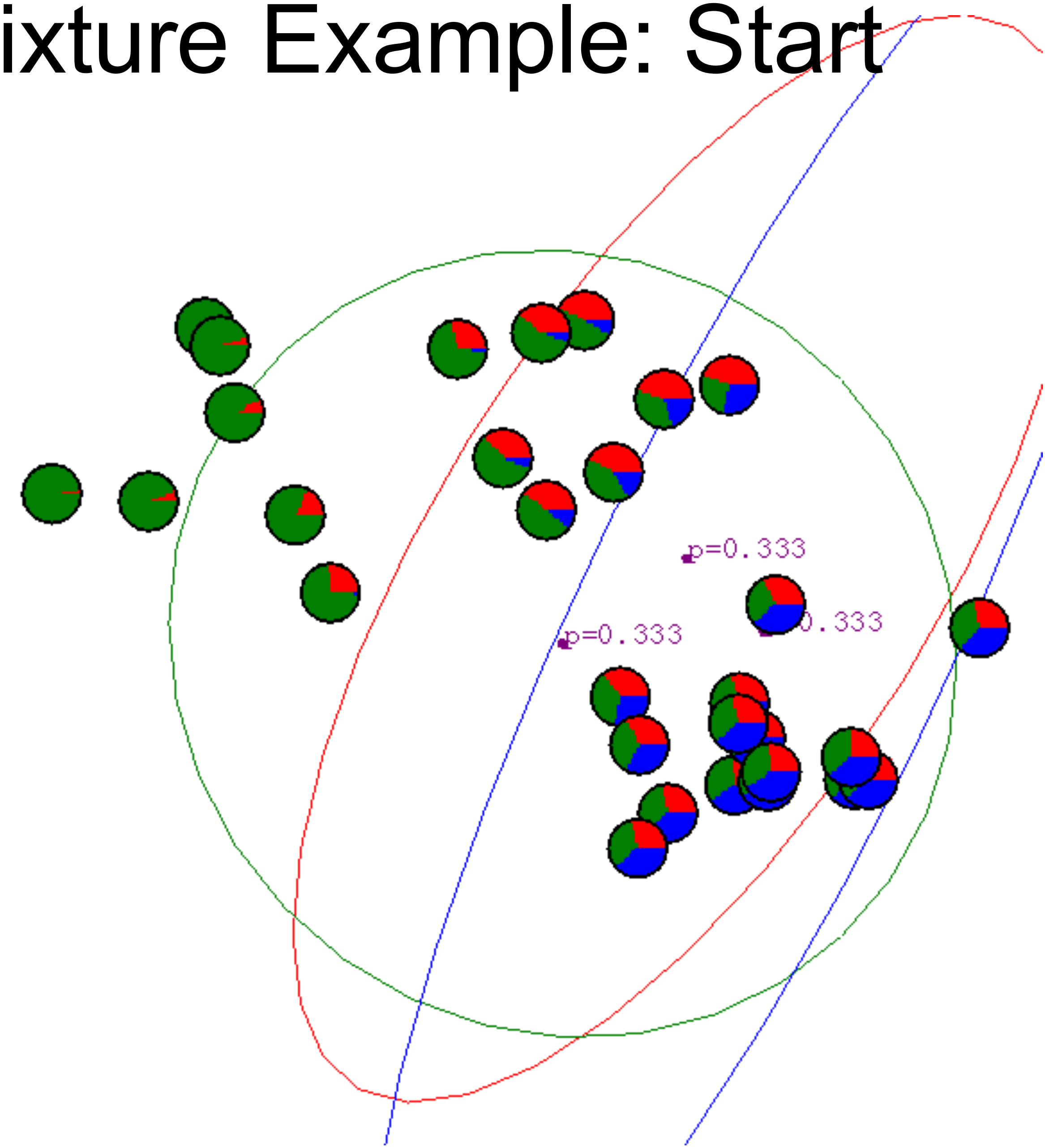
- Extremely general
- Extremely useful algorithm
  - Essentially THE goto algorithm for unsupervised learning

- Plan
  - EM for learning GMM parameters
  - EM for general unsupervised learning problems

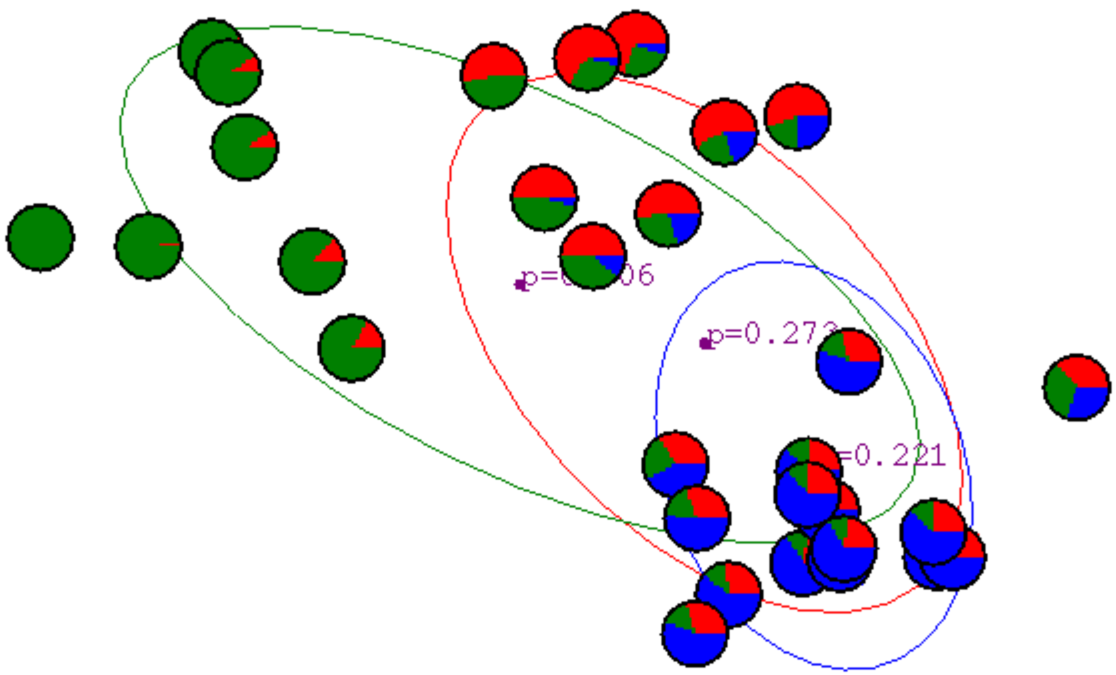
EM for Learning GMMs

- Simple Update Rules
  - E-Step: estimate $P(z_i = j \mid x_i)$
  - M-Step: maximize full likelihood weighted by posterior

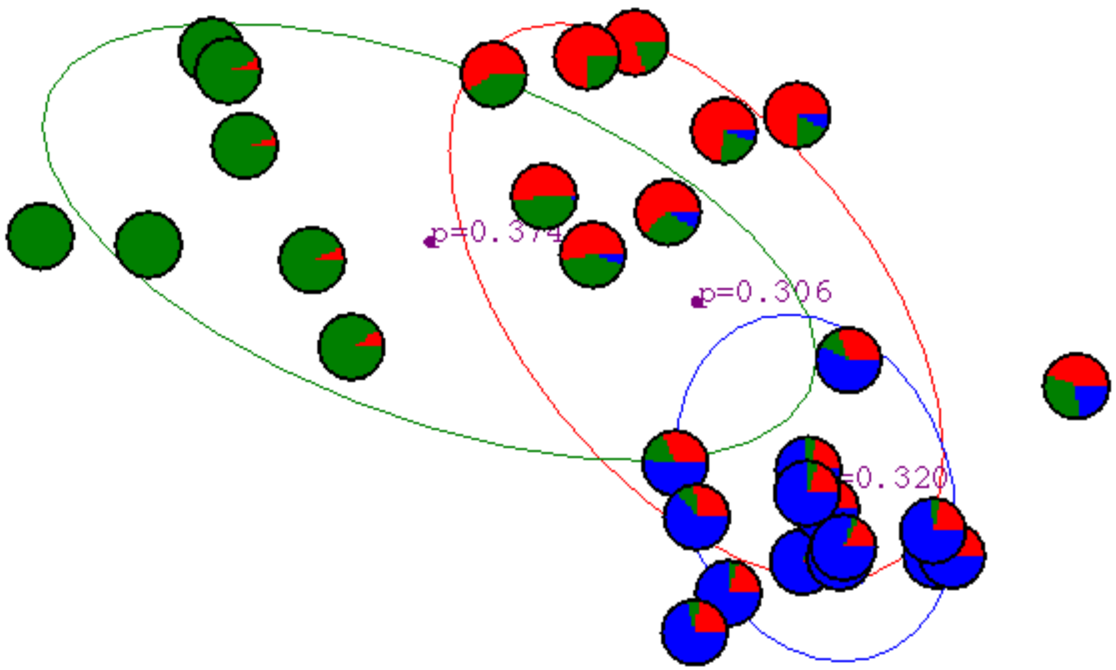
Gaussian Mixture Example: Start



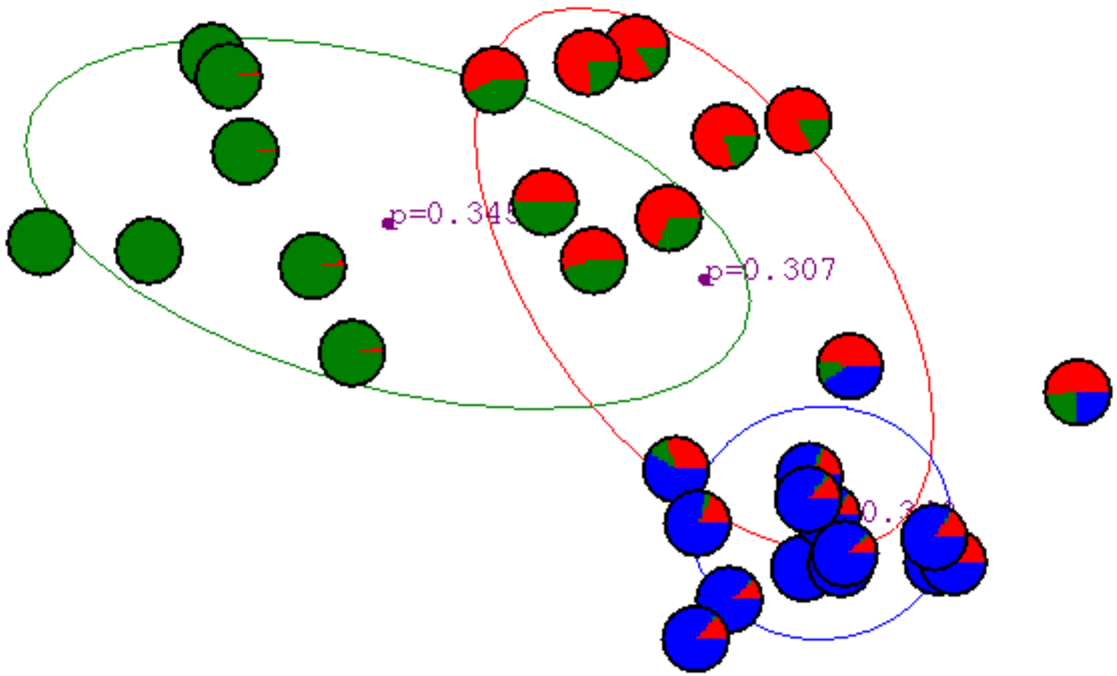
After 1st iteration



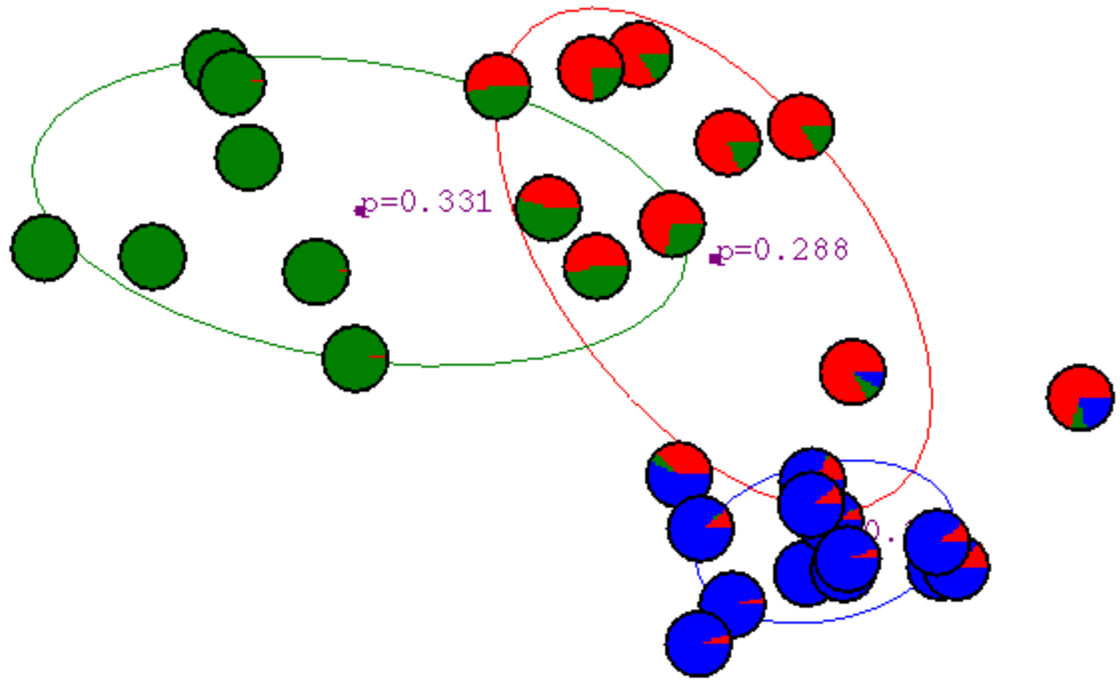
After 2nd iteration



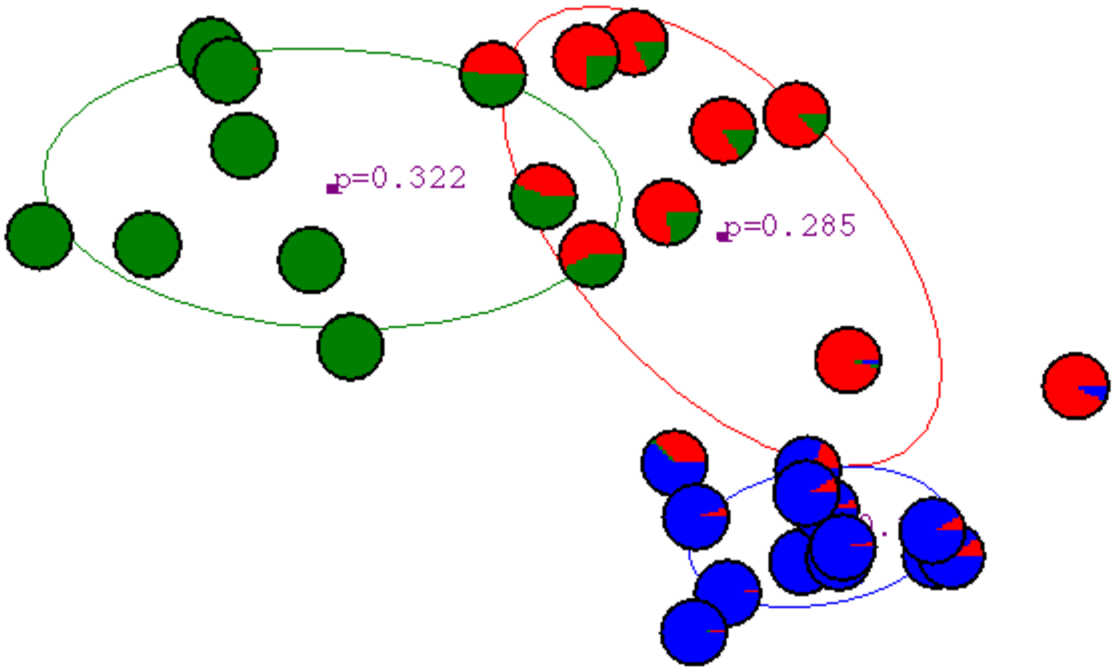
After 3rd iteration



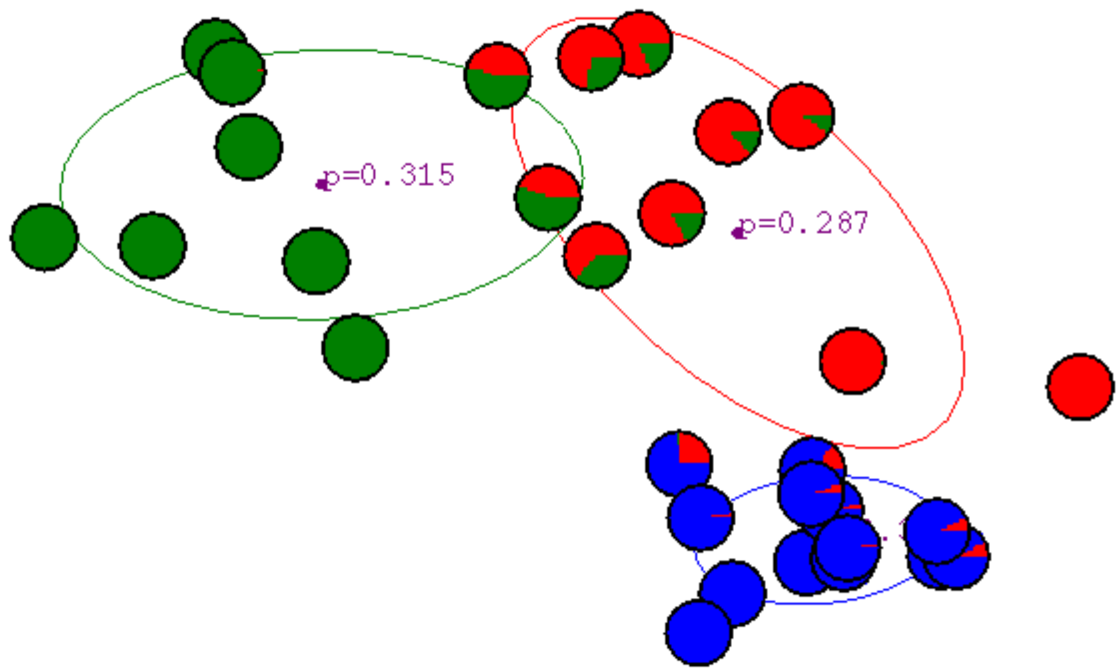
After 4th iteration



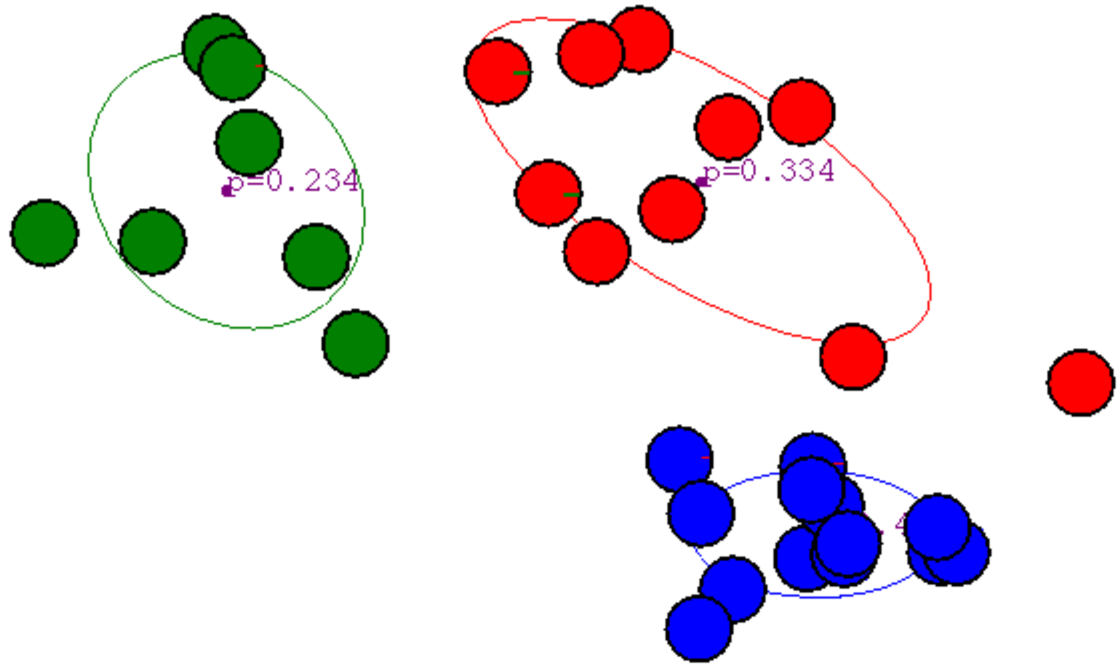
After 5th iteration



After 6th iteration



After 20th iteration



General Mixture Models

P(x | z)	P(z)	Name
Gaussian	Categorial	GMM
Multinomial	Categorical	Mixture of Multinomials
Categorical	Dirichlet	Latent Dirichlet Allocation

The general learning problem with missing data

- Marginal likelihood – **x** is observed, **z** is missing:

$$
\begin{aligned}
ll(\theta : \mathcal{D}) &= \log \prod_{i=1}^{N} P(\mathbf{x}_i \mid \theta) \\
&= \sum_{i=1}^{N} \log P(\mathbf{x}_i \mid \theta) \\
&= \sum_{i=1}^{N} \log \sum_{\mathbf{z}} P(\mathbf{x}_i, \mathbf{z} \mid \theta)
\end{aligned}
$$

Applying Jensen's inequality

$$ll(\theta : \mathcal{D}) = \sum_{i=1}^{N} \log \sum_{\mathbf{z}} Q_i(\mathbf{z}) \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i(\mathbf{z})}$$

$$ll(\theta : \mathcal{D}) \; \geq \; F(\theta, Q_i) \; = \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i(\mathbf{z})}$$

Convergence of EM

- Define potential function F(,Q):

$$ll(\theta : \mathcal{D}) \;\geq\; F(\theta, Q_i) \;=\; \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i(\mathbf{z})}$$

- EM corresponds to coordinate ascent on F
  - Thus, maximizes lower bound on marginal log likelihood

EM is coordinate ascent

$$ll(\theta : \mathcal{D}) \geq F(\theta, Q_i) = \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i(\mathbf{z})}$$

- **E-step**: Fix $\theta^{(t)}$, maximize F over Q:

$$\begin{aligned}
F(\theta, Q_i) &= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta^{(t)})}{Q_i(\mathbf{z})} \\
&= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{z} \mid \mathbf{x}_i, \theta) P(\mathbf{x}_i \mid \theta^{(t)})}{Q_i(\mathbf{z})} \\
&= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log P(\mathbf{x}_i \mid \theta^{(t)}) + \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{z} \mid \mathbf{x}_i, \theta^{(t)})}{Q_i(\mathbf{z})} \\
&= ll(\theta^{(t)} : \mathcal{D}) - \sum_{i=1}^{N} KL(Q_i(\mathbf{z}) || P(\mathbf{z} \mid \mathbf{x}_i, \theta^{(t)})
\end{aligned}$$

  - “Realigns” F with likelihood:

$$Q_i^{(t)}(\mathbf{z}) = P(\mathbf{z} \mid \mathbf{x}_i, \theta^{(t)})$$

$$F(\theta^{(t)}, Q^{(t)}) = ll(\theta^{(t)} : \mathcal{D})$$

EM is coordinate ascent

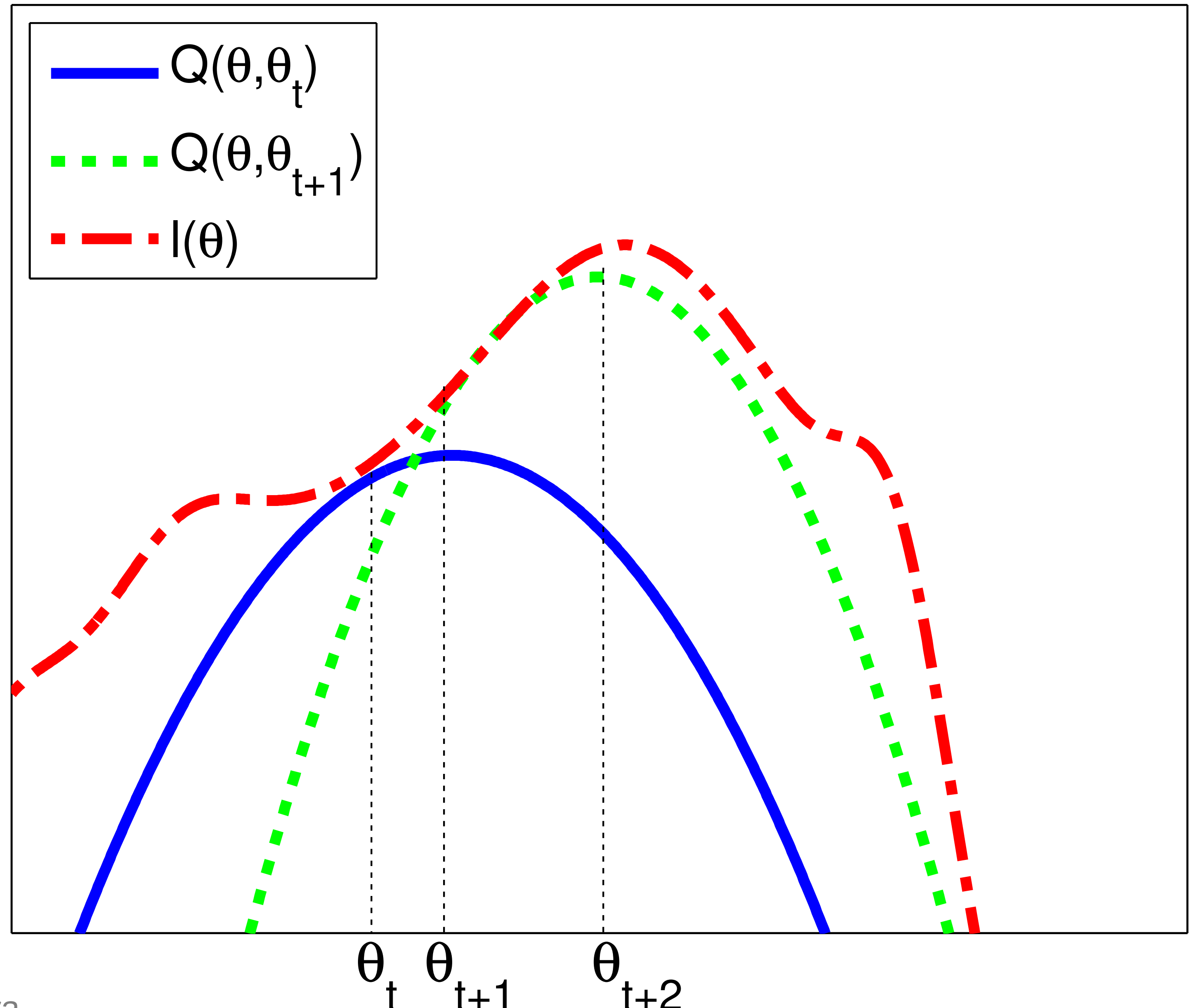
$$ll(\theta : \mathcal{D}) \;\geq\; F(\theta, Q_i) \;=\; \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i(\mathbf{z})}$$

- **M-step**: Fix $Q^{(t)}$, maximize F over

$$\begin{aligned}
F(\theta, Q_i) &= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i^{(t)}(\mathbf{z}) \log \frac{P(\mathbf{x}_i, \mathbf{z} \mid \theta)}{Q_i^{(t)}(\mathbf{z})} \\
&= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i^{(t)}(\mathbf{z}) \log P(\mathbf{x}_i, \mathbf{z} \mid \theta) - \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i^{(t)}(\mathbf{z}) \log Q_i^{(t)}(\mathbf{z}) \\
&= \sum_{i=1}^{N} \sum_{\mathbf{z}} Q_i^{(t)}(\mathbf{z}) \log P(\mathbf{x}_i, \mathbf{z} \mid \theta) + \sum_{i=1}^{N} \underbrace{H(Q_i^{(t)})}_{\text{constant}}
\end{aligned}$$

- Corresponds to weighted dataset:
  - $\langle \mathbf{x}_1, \mathbf{z}=1 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=1|\mathbf{x}_1)$
  - $\langle \mathbf{x}_1, \mathbf{z}=2 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=2|\mathbf{x}_1)$
  - $\langle \mathbf{x}_1, \mathbf{z}=3 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=3|\mathbf{x}_1)$
  - $\langle \mathbf{x}_2, \mathbf{z}=1 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=1|\mathbf{x}_2)$
  - $\langle \mathbf{x}_2, \mathbf{z}=2 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=2|\mathbf{x}_2)$
  - $\langle \mathbf{x}_2, \mathbf{z}=3 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=3|\mathbf{x}_2)$
  - …

EM Intuition



What you should know

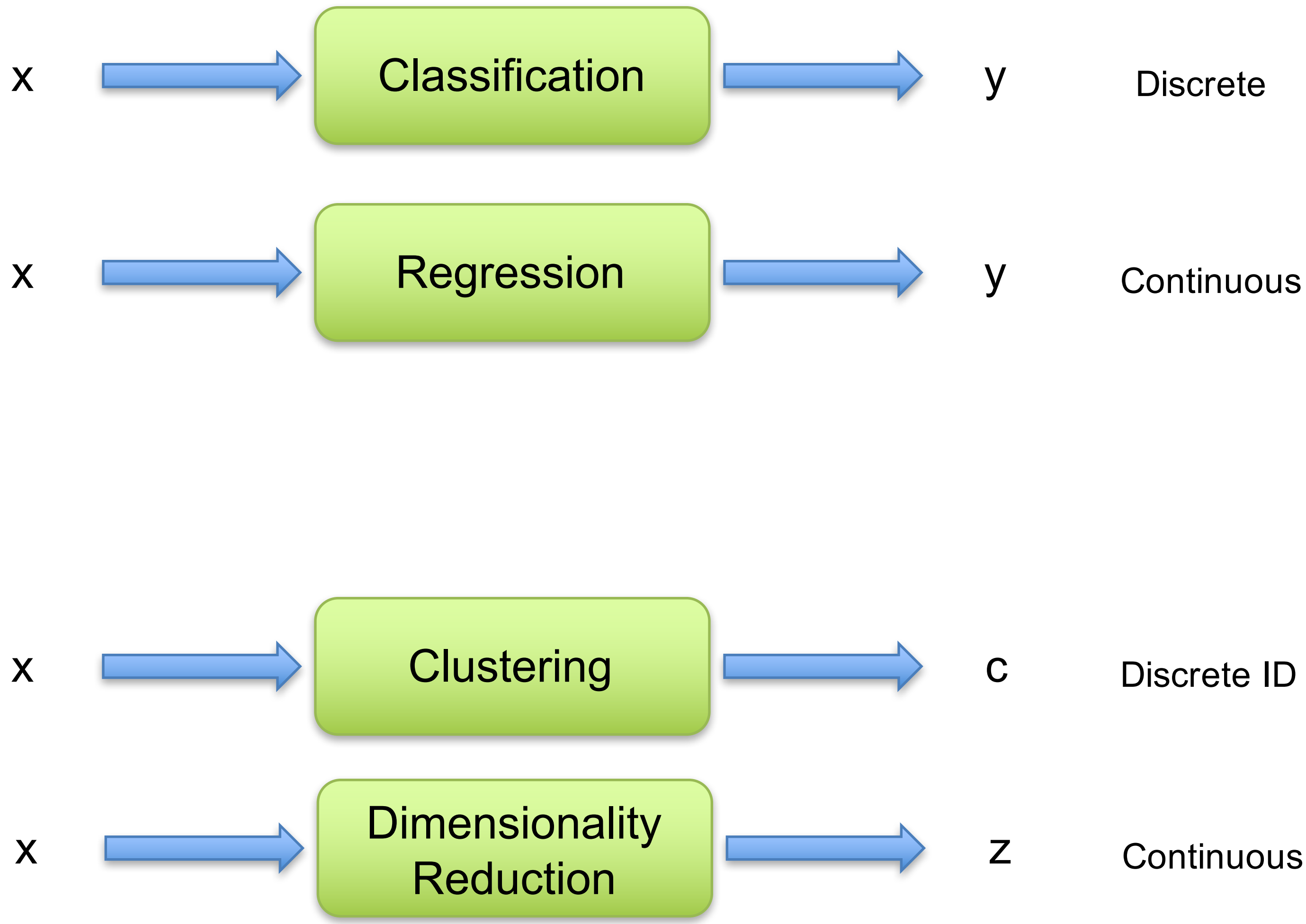
- K-means for clustering:
  - algorithm
  - converges because it's coordinate ascent

- EM for mixture of Gaussians:
  - How to "learn" maximum likelihood parameters (locally max. like.) in the case of unlabeled data

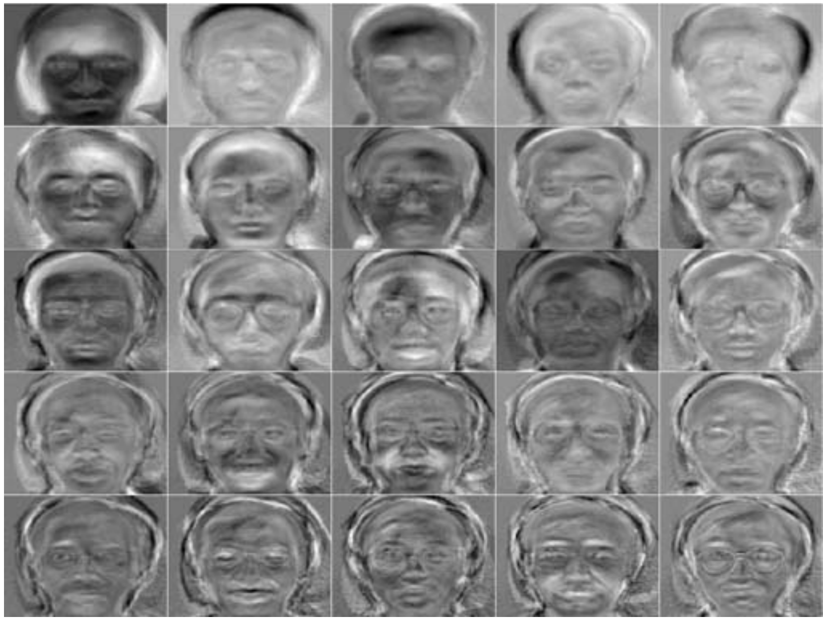
- EM is coordinate ascent
- Remember, E.M. can get stuck in local minima, and empirically it <u>DOES</u>

- General case for EM

Tasks


x
Classification
y
Discrete
x
Regression
y
Continuous
x
Clustering
c
Discrete ID
x
Dimensionality Reduction
z
Continuous

New Topic: PCA



Synonyms

- Principal Component Analysis
- Karhunen–Loève transform

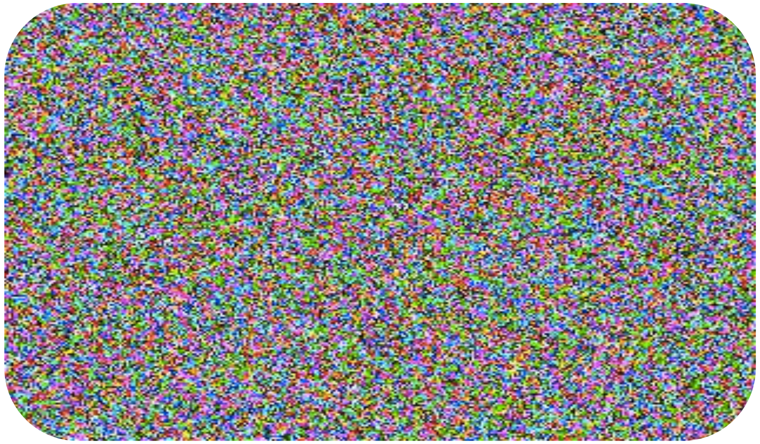
- Eigen-Faces
- Eigen-<Insert-your-problem-domain>

- PCA is a Dimensionality Reduction Algorithm

- Other Dimensionality Reduction algorithms
  - Linear Discriminant Analysis (LDA)
  - Independent Component Analysis (ICA)
  - Local Linear Embedding (LLE)
  - …

Dimensionality reduction

- Input data may have thousands or millions of dimensions!
  - e.g., images have 5M pixels



Dimensionality reduction

- Input data may have thousands or millions of dimensions!
  - e.g., images have 5M pixels

- **Dimensionality reduction**:
  represent data with fewer dimensions
  - easier learning – fewer parameters
  - visualization – hard to visualize more than 3D or 4D
  - discover “intrinsic dimensionality” of data
    - high dimensional data that is truly lower dimensional

PCA / KL-Transform

- De-correlation view
  - Make features uncorrelated
  - No projection yet

- Max-variance view:
  - Project data to lower dimensions
  - Maximize variance in lower dimensions

- Synthesis / Min-error view:
  - Project data to lower dimensions
  - Minimize reconstruction error

- All views lead to same solution

Basic PCA algorithm

- **Center data (**subtract mean)
- **Estimate covariance**
- **Find eigenvectors and values of covariance**
- **Principle components:** choose k eigenvectors with highest corresponding values

Video

- https://youtu.be/pSRA8GpWIrA?t=162

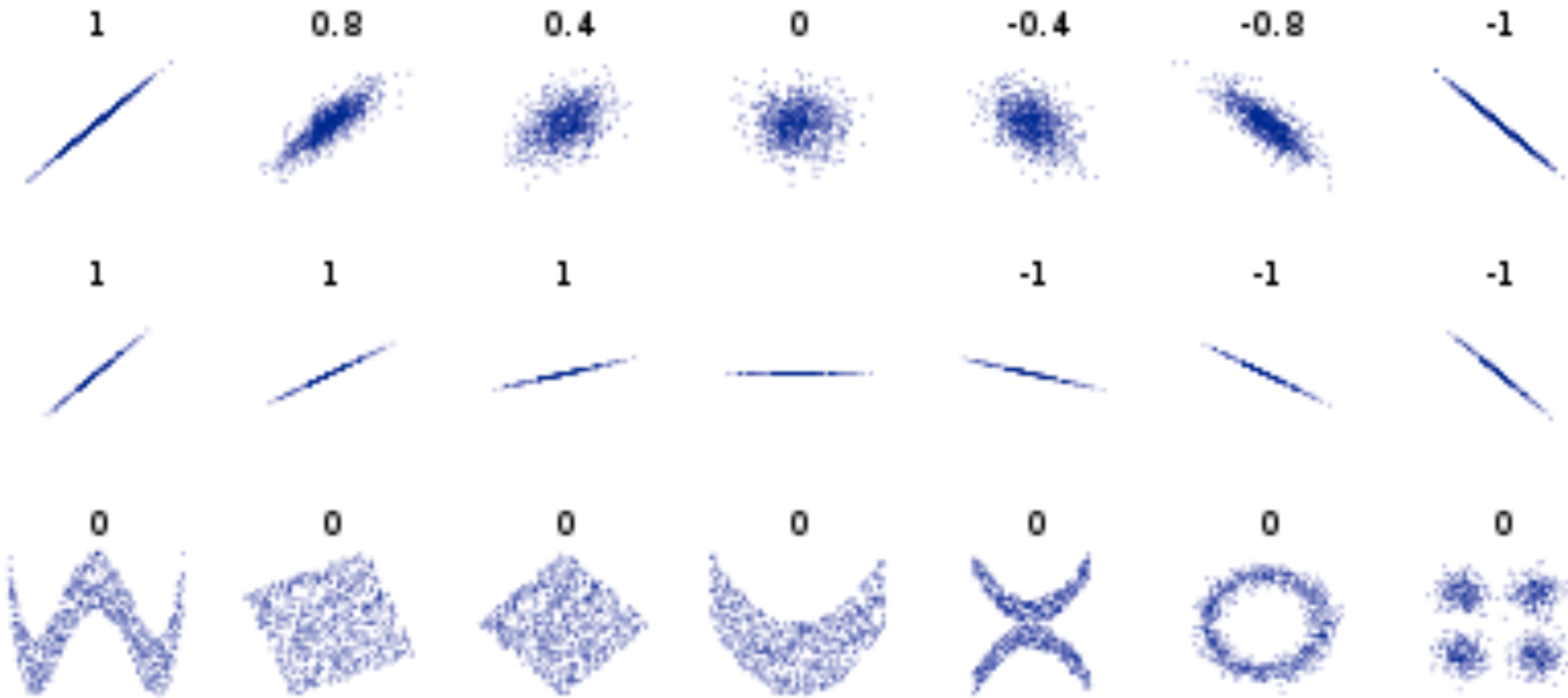
Video

- What if the dimension is high?
  - Covariance matrix is d x d
  - For high d, Eigen decomposition is very slow… $O(d^3)$

- Use Singular Value Decomposition (SVD)
  - finds k-eigenvectors
  - great implementations $O(N^2d)$

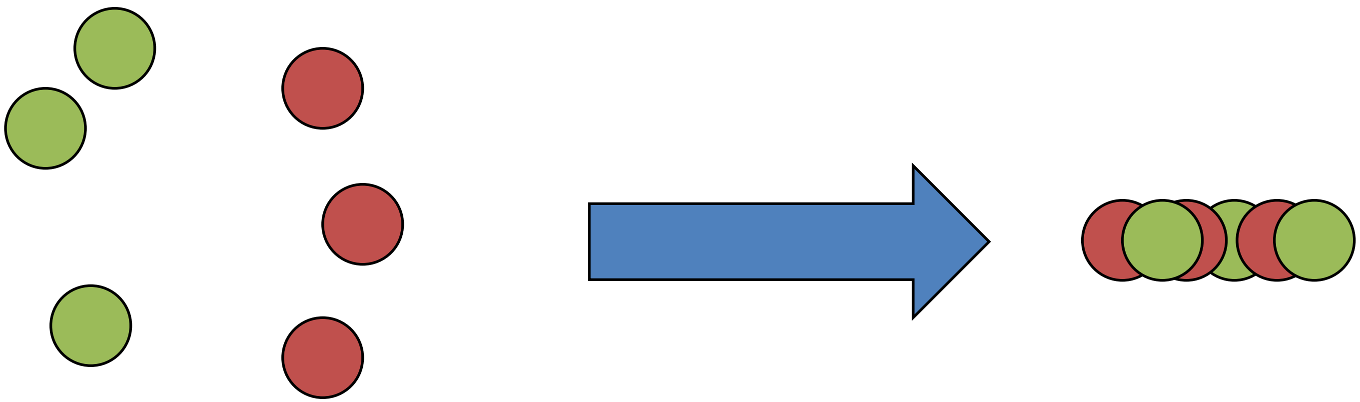
PCA Problem #1

- Only removed linear dependencies
- Correlation Cov(x,y) / (var(x), var(y))



PCA Problem #2

- Direction of maximum variance may not be use for classification



- See Linear Discriminate Analysis (if you have labels)

What you need to know

- Dimensionality Reduction
  - why and when its important
    - visualization
    - compression
    - faster learning
- Principle Component Analysis
  - KL Transform view
    - Notes have reconstruction error and max variance views too
  - Relationship to covariance matrix and eigenvectors
  - using SVD for PCA

Machine Learning Lectures are Over

- Basics of Statistical Learning
  - Loss functions, MLE, MAP, Bayesian estimation, bias-variance tradeoff, regularization, cross-validation
- Supervised Learning
  - Nearest neighbor, Naïve Bayes, Logistic Regression, Neural Networks, Support Vector Machines, Kernels, Decision Trees
  - Ensemble methods: bagging and boosting
- Unsupervised Learning
  - Clustering/Density estimation: k-means, GMMs, EM
  - Dimensionality Reduction: PCA (with SVD)

What is Left?

- Poster Session
  - Dec 6th 1:30-3:30pm in Goodwin Hall Atrium

- Final Exam
  - Dec 14th in class (DURH 261); 2:05 - 4:05 pm

- A Sincere Thank You