CS434a/541a: Pattern Recognition
Prof. Olga Veksler

Lecture 7

Curse of Dimensionality,
Dimensionality Reduction with PCA

Dimensionality on the Course Road Map

a lot is known

1. Bayesian Decision theory (rare case)
   - Know probability distribution of the categories
   - Do not even need training data
   - Can design optimal classifier
2. ML and Bayesian parameter estimation
   - Need to estimate Parameters of probability dist.
   - Need training data
3. Non-Parametric Methods
   - No probability distribution, labeled data
4. Linear discriminant functions and Neural Nets
   - The shape of discriminant functions is known
   - Need to estimate parameters of discriminant functions
5. Unsupervised Learning and Clustering
   - No probability distribution and unlabeled data

little is known

affects all these methods

Today

- Problems of high dimensional data, "the curse of dimensionality"
  - running time
  - overfitting
  - number of samples required
- Dimensionality Reduction Methods
  - Principle Component Analysis (today)
  - Fisher Linear Discriminant (next time)

Curse of Dimensionality: Complexity

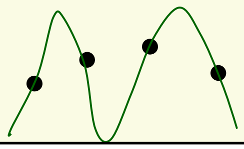
- Complexity (running time) increases with dimension ***d***
- A lot of methods have at least O(***nd***^{2}) complexity, where ***n*** is the number of samples
  - For example if we need to estimate covariance matrix
- So as ***d*** becomes large, O(***nd***^{2}) complexity may be too costly

Curse of Dimensionality: Overfitting

- If d is large, n, the number of samples, may be too small for accurate parameter estimation
- For example, covariance matrix has d^2 parameters:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_d^2 \end{bmatrix}$$

- For accurate estimation, n should be much bigger than d^2
- Otherwise model is too complicated for the data, ***overfitting***:

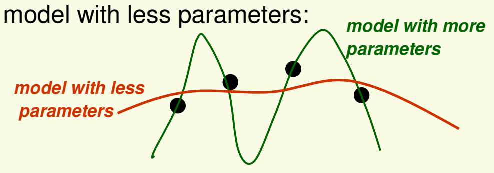


Curse of Dimensionality: Overfitting

- Paradox: If $n < d^2$ we are better off assuming that features are uncorrelated, even if we know this assumption is wrong
- In this case, the covariance matrix has only d parameters:

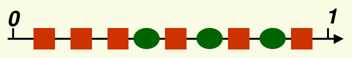
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_d^2 \end{bmatrix}$$

- We are likely to avoid overfitting because we fit a model with less parameters:



Curse of Dimensionality: Number of Samples

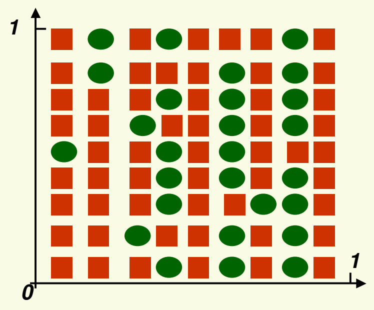
- Suppose we want to use the nearest neighbor approach with $k = 1$ (***1NN***)
- Suppose we start with only one feature



- This feature is not discriminative, i.e. it does not separate the classes well
- We decide to use 2 features. For the 1NN method to work well, need a lot of samples, i.e. samples have to be dense
- To maintain the same density as in 1D (9 samples per unit length), how many samples do we need?

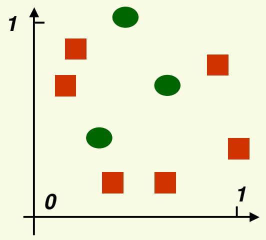
Curse of Dimensionality: Number of Samples

- We need 9^2 samples to maintain the same density as in ***1D***



Curse of Dimensionality: Number of Samples

- Of course, when we go from 1 feature to 2, no one gives us more samples, we still have 9



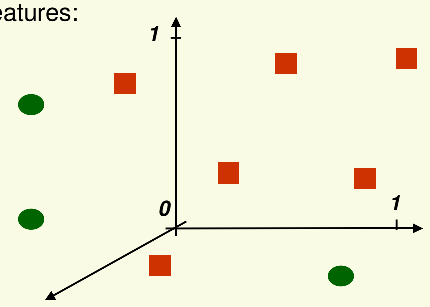
- This is way too sparse for ***1NN*** to work well

Curse of Dimensionality: Number of Samples

- In general, if ***n*** samples is dense enough in ***1D***
- Then in ***d*** dimensions we need n^d samples!
- And n^d grows really really fast as a function of ***d***
- Common pitfall:
  - If we can't solve a problem with a few features, adding more features seems like a good idea
  - However the number of samples usually stays the same
  - The method with more features is likely to perform worse instead of expected better

Curse of Dimensionality: Number of Samples

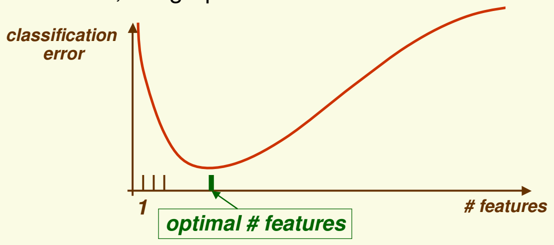
- Things go from bad to worse if we decide to use 3 features:



- If ***9*** was dense enough in 1D, in 3D we need $9^3 = 729$ samples!

Curse of Dimensionality: Number of Samples

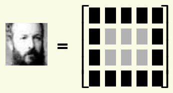
- For a fixed number of samples, as we add features, the graph of classification error:



- Thus for each fixed sample size ***n***, there is the optimal number of features to use

The Curse of Dimensionality

- We should try to avoid creating lot of features
- Often no choice, problem starts with many features
- Example: Face Detection
  - One sample point is ***k*** by ***m*** array of pixels



  - Feature extraction is not trivial, usually every pixel is taken as a feature
  - Typical dimension is 20 by 20 = 400
  - Suppose ***10*** samples are dense enough for 1 dimension. Need only $\boldsymbol{10^{400}}$ samples

Dimensionality Reduction

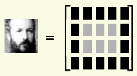
- High dimensionality is challenging and redundant
- It is natural to try to reduce dimensionality
- Reduce dimensionality by feature combination: combine old features ***x*** to create new features ***y***

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \to f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}\right) = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = y \quad \text{with } k < d$$

- For example, $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \to \begin{bmatrix} x_1 + x_2 \\ x_3 + x_4 \end{bmatrix} = y$
- Ideally, the new vector ***y*** should retain from ***x*** all information important for classification

The Curse of Dimensionality

- Face Detection, dimension of one sample point is ***km***



- The fact that we set up the problem with ***km*** dimensions (features) does not mean it is really a ***km***-dimensional problem
- Space of all ***k*** by ***m*** images has ***km*** dimensions
- Space of all ***k*** by ***m*** faces must be much smaller, since faces form a tiny fraction of all possible images
- Most likely we are not setting the problem up with the right features
- If we used better features, we are likely need much less than ***km***-dimensions

Dimensionality Reduction

- The best ***f***(***x***) is most likely a non-linear function
- Linear functions are easier to find though
- For now, assume that ***f***(***x***) is a linear mapping
- Thus it can be represented by a matrix ***W***:

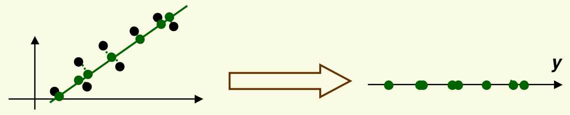
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \Rightarrow W \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} w_{11} & \cdots & w_{1d} \\ \vdots & & \vdots \\ w_{k1} & \cdots & w_{kd} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} \quad \text{with } k < d$$

Feature Combination

- We will look at 2 methods for feature combination
  - Principle Component Analysis (PCA)
  - Fischer Linear Discriminant (next lecture)

PCA

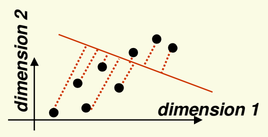
- After the data is projected on the best line, need to transform the coordinate system to get 1D representation for vector ***y***



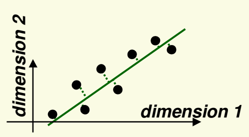
- Note that new data ***y*** has the same variance as old data ***x*** in the direction of the green line
- PCA preserves largest variances in the data. We will prove this statement, for now it is just an intuition of what PCA will do

Principle Component Analysis (PCA)

- Main idea: seek most accurate data representation in a lower dimensional space
- Example in 2-D
  - Project data to 1-D subspace (a line) which minimize the projection error



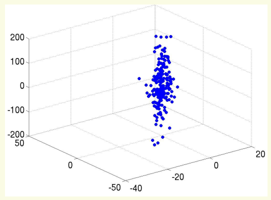
large projection errors, bad line to project to



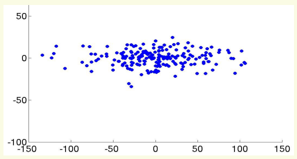
small projection errors, good line to project to

- Notice that the the good line to use for projection lies in the direction of largest variance

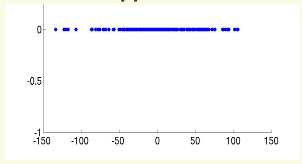
PCA: Approximation of Elliptical Cloud in 3D



best 2D approximation

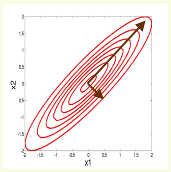


best 1D approximation



PCA

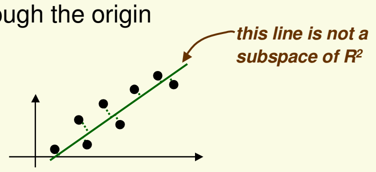
- What is the direction of largest variance in data?
- Recall that if ***x*** has multivariate distribution N(μ,Σ), direction of largest variance is given by eigenvector corresponding to the largest eigenvalue of Σ



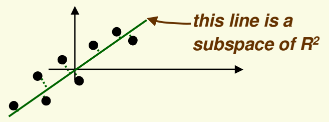
- This is a hint that we should be looking at the covariance matrix of the data (note that PCA can be applied to distributions other than Gaussian)

PCA: Linear Algebra for Derivation

- Recall that subspace ***W*** contains the zero vector, i.e. it goes through the origin



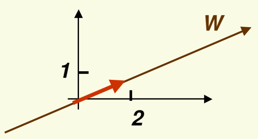
- For derivation, it will be convenient to project to subspace ***W***: thus we need to shift everything



PCA: Linear Algebra for Derivation

- Let ***V*** be a ***d*** dimensional linear space, and ***W*** be a ***k*** dimensional linear subspace of ***V***
- We can always find a set of ***d*** dimensional vectors {$e_1, e_2, \ldots, e_k$} which forms an orthonormal basis for ***W***
  - $<e_i,e_j> = 0$ if ***i*** is not equal to ***j*** and $<e_i,e_i> = 1$
- Thus any vector in ***W*** can be written as

$$\alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_k e_k = \sum_{i=1}^{k} \alpha_i e_i \quad \textit{for scalars } \alpha_1, \ldots, \alpha_k$$



Let ***V*** = ***R***^{2} and ***W*** be the line x-2y=0. Then the orthonormal basis for W is

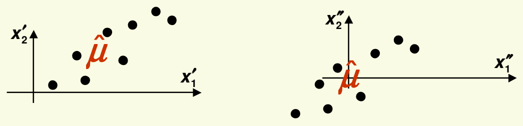
$$\left\{\begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}\right\}$$

PCA Derivation: Shift by the Mean Vector

- Before PCA, subtract sample mean from the data

$$x - \frac{1}{n}\sum_{i=1}^{n} x_i = x - \hat{\mu}$$

- The new data has zero mean: E(***X***-E(***X***)) = E(***X***)-E(***X***) = 0
- All we did is change the coordinate system



- Another way to look at it:
  - first step of getting ***y*** is to subtract the mean of ***x***

$$x \rightarrow y = f(x) = g(x - \hat{\mu})$$

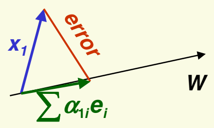
PCA: Derivation

- We want to find the most accurate representation of data $D=\{x_1,x_2,\dots,x_n\}$ in some subspace W which has dimension $k < d$
- Let $\{e_1,e_2,\dots,e_k\}$ be the orthonormal basis for W. Any vector in W can be written as $\sum_{i=1}^{k}\alpha_i e_i$
- Thus x_1 will be represented by some vector in W

$$\sum_{i=1}^{k}\alpha_{1i} e_i$$

- Error this representation:

$$error = \left\| x_1 - \sum_{i=1}^{k}\alpha_{1i} e_i \right\|^2$$



PCA: Derivation

- To minimize J, need to take partial derivatives and also enforce constraint that $\{e_1,e_2,\dots,e_k\}$ are orthogonal

$$J(e_1,\dots,e_k,\alpha_{11},\dots\alpha_{nk}) = \sum_{j=1}^{n}\left\| x_j - \sum_{i=1}^{k}\alpha_{ji} e_i \right\|^2$$

- Let us simplify J first

$$J(e_1,\dots,e_k,\alpha_{11},\dots\alpha_{nk}) = \sum_{j=1}^{n}\left\| x_j \right\|^2 - 2\sum_{j=1}^{n} x_j^t\left(\sum_{i=1}^{k}\alpha_{ji} e_i\right) + \sum_{j=1}^{n}\sum_{i=1}^{k}\alpha_{ji}^2$$

$$= \sum_{j=1}^{n}\left\| x_j \right\|^2 - 2\sum_{j=1}^{n}\sum_{i=1}^{k}\alpha_{ji} x_j^t e_i + \sum_{j=1}^{n}\sum_{i=1}^{k}\alpha_{ji}^2$$

PCA: Derivation

- To find the total error, we need to sum over all x_j's
- Any x_j can be written as $\sum_{i=1}^{k}\alpha_{ji} e_i$
- Thus the total error for representation of all data D is:

$$\underbrace{J(e_1,\dots,e_k,\alpha_{11},\dots\alpha_{nk})}_{\text{unknowns}} = \overset{\text{sum over all data points}}{\sum_{j=1}^{n}} \underbrace{\boxed{\left\| x_j - \sum_{i=1}^{k}\alpha_{ji} e_i \right\|^2}}_{\text{error at one point}}$$

PCA: Derivation

$$J(e_1,\dots,e_k,\alpha_{11},\dots\alpha_{nk}) = \sum_{j=1}^{n}\left\| x_j \right\|^2 - 2\sum_{j=1}^{n}\sum_{i=1}^{k}\alpha_{ji} x_j^t e_i + \sum_{j=1}^{n}\sum_{i=1}^{k}\alpha_{ji}^2$$

- First take partial derivatives with respect to α_{ml}

$$\frac{\partial}{\partial\alpha_{ml}} J(e_1,\dots,e_k,\alpha_{11},\dots\alpha_{nk}) = -2x_m^t e_l + 2\alpha_{ml}$$

- Thus the optimal value for α_{ml} is

$$-2x_m^t e_l + 2\alpha_{ml} = 0 \;\Rightarrow\; \alpha_{ml} = x_m^t e_l$$

PCA: Derivation

$$J(e_1,\dots,e_k,\alpha_{11},\dots\alpha_{nk}) = \sum_{j=1}^{n}\|x_j\|^2 - 2\sum_{j=1}^{n}\sum_{i=1}^{k}\alpha_{ji}x_j^te_i + \sum_{j=1}^{n}\sum_{i=1}^{k}\alpha_{ji}^2$$

- Plug the optimal value for $\alpha_{ml} = x^t_m e_l$ back into J

$$J(e_1,\dots,e_k) = \sum_{j=1}^{n}\|x_j\|^2 - 2\sum_{j=1}^{n}\sum_{i=1}^{k}(x_j^te_i)x_j^te_i + \sum_{j=1}^{n}\sum_{i=1}^{k}(x_j^te_i)^2$$

- Can simplify J

$$J(e_1,\dots,e_k) = \sum_{j=1}^{n}\|x_j\|^2 - \sum_{j=1}^{n}\sum_{i=1}^{k}(x_j^te_i)^2$$

PCA: Derivation

$$J(e_1,\dots,e_k) = \underbrace{\sum_{j=1}^{n}\|x_j\|^2}_{\text{constant}} - \sum_{i=1}^{k}e_i^tSe_i$$

- Minimizing J is equivalent to maximizing $\sum_{i=1}^{k}e_i^tSe_i$
- We should also enforce constraints $e_i^te_i = 1$ for all i
- Use the method of Lagrange multipliers, incorporate the constraints with undetermined $\lambda_1,\dots,\lambda_k$
- Need to maximize new function u

$$u(e_1,\dots,e_k) = \sum_{i=1}^{k}e_i^tSe_i - \sum_{j=1}^{k}\lambda_j(e_j^te_j - 1)$$

PCA: Derivation

$$J(e_1,\dots,e_k) = \sum_{j=1}^{n}\|x_j\|^2 - \sum_{j=1}^{n}\sum_{i=1}^{k}(x_j^te_i)^2$$

- Rewrite J using $(a^tb)^2 = (a^tb)(a^tb) = (b^ta)(a^tb) = b^t(aa^t)b$

$$J(e_1,\dots,e_k) = \sum_{j=1}^{n}\|x_j\|^2 - \sum_{i=1}^{k}e_i^t\left(\sum_{j=1}^{n}(x_jx_j^t)\right)e_i$$
$$= \sum_{j=1}^{n}\|x_j\|^2 - \sum_{i=1}^{k}e_i^tSe_i$$

- Where $S = \sum_{j=1}^{n}x_jx_j^t$
- S is called the scatter matrix, it is just n-1 times the sample covariance matrix we have seen before

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j-\hat{\mu})(x_j-\hat{\mu})^t$$

PCA: Derivation

- If x is a vector and $f(x) = f(x_1,\dots,x_d)$ is a function, to simplify notation, define

$$\frac{d}{dx}f(x) = \begin{bmatrix}\frac{\partial f}{\partial x_1}\\ \vdots \\ \frac{\partial f}{\partial x_d}\end{bmatrix}$$

- It can be shown that $\frac{d}{dx}(x^tx) = 2x$
- If A is a symmetric matrix, it can be shown that

$$\frac{d}{dx}(x^tAx) = 2Ax$$

PCA: Derivation

$$u(e_1,\dots,e_k)=\sum_{i=1}^{k}e_i^tSe_i-\sum_{j=1}^{k}\lambda_j\left(e_j^te_j-1\right)$$

- Compute the partial derivatives with respect to e_m

$$\frac{\partial}{\partial e_m}u(e_1,\dots,e_k)=2Se_m-2\lambda_m e_m=0$$

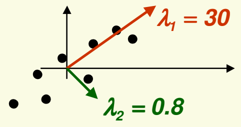
Note: *e_m is a vector, what we are really doing here is taking partial derivatives with respect to each element of e_m and then arranging them up in a linear equation*

- Thus λ_m and e_m are eigenvalues and eigenvectors of scatter matrix S

$$Se_m=\lambda_m e_m$$

PCA

- The larger the eigenvalue of S, the larger is the variance in the direction of corresponding eigenvector



- This result is exactly what we expected: project x into subspace of dimension k which has the largest variance
- This is very intuitive: restrict attention to directions where the scatter is the greatest

PCA: Derivation

$$J(e_1,\dots,e_k)=\sum_{j=1}^{n}\left\|x_j\right\|^2-\sum_{i=1}^{k}e_i^tSe_i$$

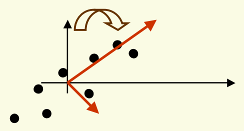
- Let's plug e_m back into J and use $Se_m=\lambda_m e_m$

$$J(e_1,\dots,e_k)=\sum_{j=1}^{n}\left\|x_j\right\|^2-\sum_{i=1}^{k}\lambda_i\left\|e_i\right\|^2=\underbrace{\sum_{j=1}^{n}\left\|x_j\right\|^2}_{constant}-\sum_{i=1}^{k}\lambda_i$$

- Thus to minimize J take for the basis of W the k eigenvectors of S corresponding to the k largest eigenvalues

PCA

- Thus PCA can be thought of as finding new orthogonal basis by rotating the old axis until the directions of maximum variance are found



PCA as Data Approximation

- Let $\{e_1, e_2, \ldots, e_d\}$ be all d eigenvectors of the scatter matrix S, sorted in order of decreasing corresponding eigenvalue
- Without any approximation, for any sample x_i:

$$x_i = \sum_{j=1}^{d} \alpha_j e_j = \underbrace{\alpha_1 e_1 + \ldots + \alpha_k e_k}_{\text{approximation of } x_i} + \overbrace{\alpha_{k+1} e_{k+1} \ldots + \alpha_d e_d}^{\text{error of approximation}}$$

- coefficients $\alpha_m = x^t_i e_m$ are called *principle components*
  - The larger k, the better is the approximation
  - Components are arranged in order of importance, more important components come first
- Thus PCA takes the first k most important components of x_i as an approximation to x_i

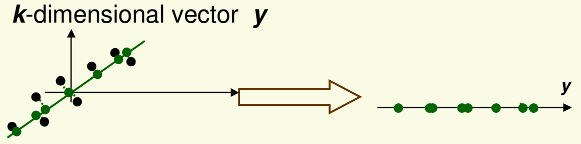
Recipe for Dimension Reduction with PCA

Data $D=\{x_1, x_2, \ldots, x_n\}$. Each x_i is a d-dimensional vector. Wish to use PCA to reduce dimension to k

1. Find the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Subtract sample mean from the data $z_i = x_i - \hat{\mu}$
3. Compute the scatter matrix $S = \sum_{i=1}^{n} z_i z_i^t$
4. Compute eigenvectors $e_1, e_2, \ldots, e_k$ corresponding to the k largest eigenvalues of S
5. Let $e_1, e_2, \ldots, e_k$ be the columns of matrix $E = [e_1 \cdots e_k]$
6. The desired y which is the closest approximation to x is $y = E^t z$

PCA: Last Step

- Now we know how to project the data
- Last step is to change the coordinates to get final k-dimensional vector y

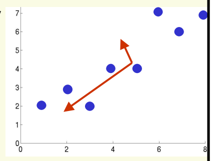


- Let matrix $E = [e_1 \cdots e_k]$
- Then the coordinate transformation is $y = E^t x$
- Under E^t, the eigenvectors become the standard basis: $E^t e_i = \begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_k \end{bmatrix} e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$

PCA Example Using Matlab

- Let $D = \{(1,2),(2,3),(3,2),(4,4),(5,4),(6,7),(7,6),(9,7)\}$
- Convenient to arrange data in array

$$X = \begin{bmatrix} 1 & 2 \\ \vdots & \vdots \\ 9 & 7 \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_8 \end{bmatrix}$$



- Mean $\mu = mean(X) = [4.6 \;\; 4.4]$
- Subtract mean from data to get new data array Z

$$Z = X - \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} = X - repmat(\mu, 8, 1) = \begin{bmatrix} -3.6 & -4.4 \\ \vdots & \vdots \\ 4.4 & 2.6 \end{bmatrix}$$

- Compute the scatter matrix S

$$S = 7 * cov(Z) = [-3.6 \;\; -4.4]\begin{bmatrix} -3.6 \\ -4.4 \end{bmatrix} + \ldots + [4.4 \;\; 2.6]\begin{bmatrix} 4.4 \\ 2.6 \end{bmatrix} = \begin{bmatrix} 57 & 40 \\ 40 & 34 \end{bmatrix}$$

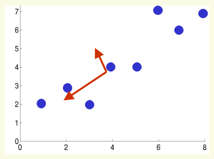
*matlab uses unbiased estimate for covariance, so S=(n-1)*cov(Z)*

PCA Example Using Matlab

- Use [V,D] =eig(***S***) to get eigenvalues and eigenvectors of ***S***

$$\lambda_1 = 87 \text{ and } e_1 = \begin{bmatrix} -0.8 \\ -0.6 \end{bmatrix}$$

$$\lambda_2 = 3.8 \text{ and } e_2 = \begin{bmatrix} 0.6 \\ -0.8 \end{bmatrix}$$

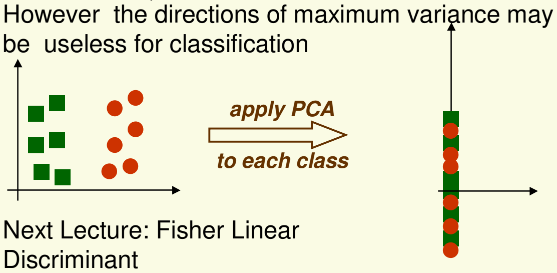


- Projection to 1D space in the direction of e_1

$$Y = e_1^t Z^t = \left(\begin{bmatrix} -0.8 & -0.6 \end{bmatrix} \begin{bmatrix} -3.6 & \cdots & 4.4 \\ -4.4 & \cdots & 2.6 \end{bmatrix} \right) = \begin{bmatrix} 4.3 & \cdots & -5.1 \end{bmatrix}$$
$$= \begin{bmatrix} y_1 & \cdots & y_8 \end{bmatrix}$$

Drawbacks of PCA

- PCA was designed for accurate *data representation*, not for data classification
  - Preserves as much variance in data as possible
  - If directions of maximum variance is important for classification, will work
- However the directions of maximum variance may be useless for classification



- Next Lecture: Fisher Linear Discriminant
  - preserve direction useful for discrimination