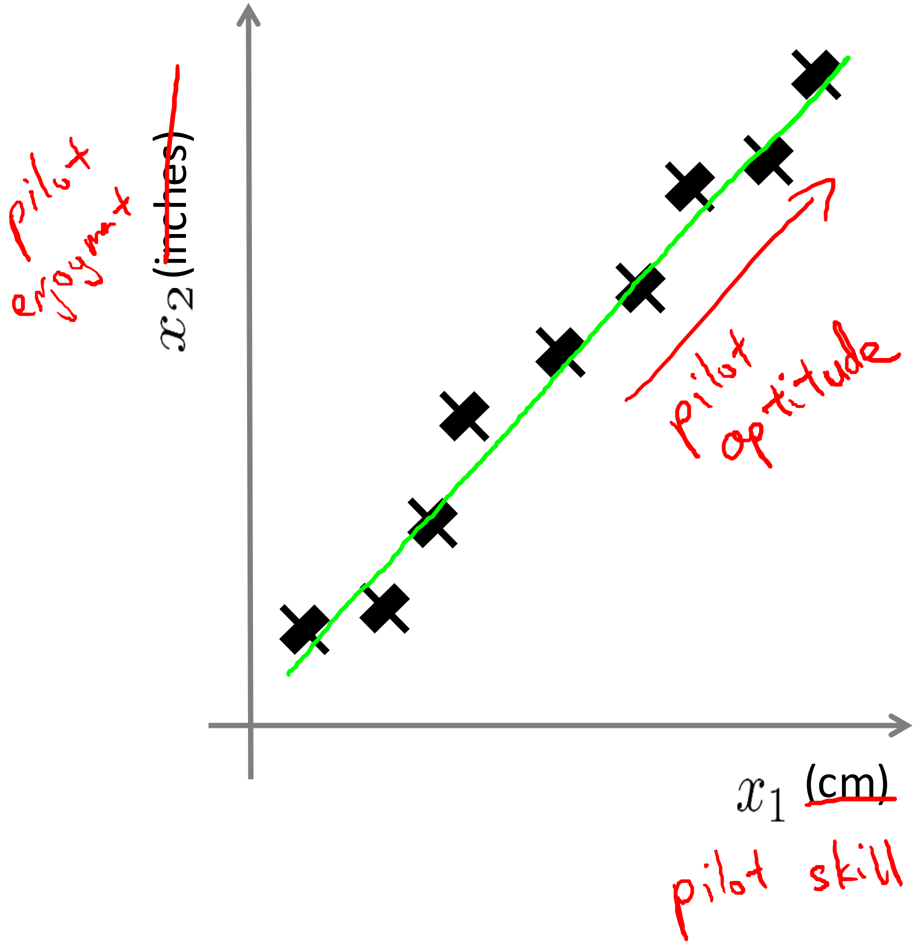
Principal Components Analysis (PCA)

PCA

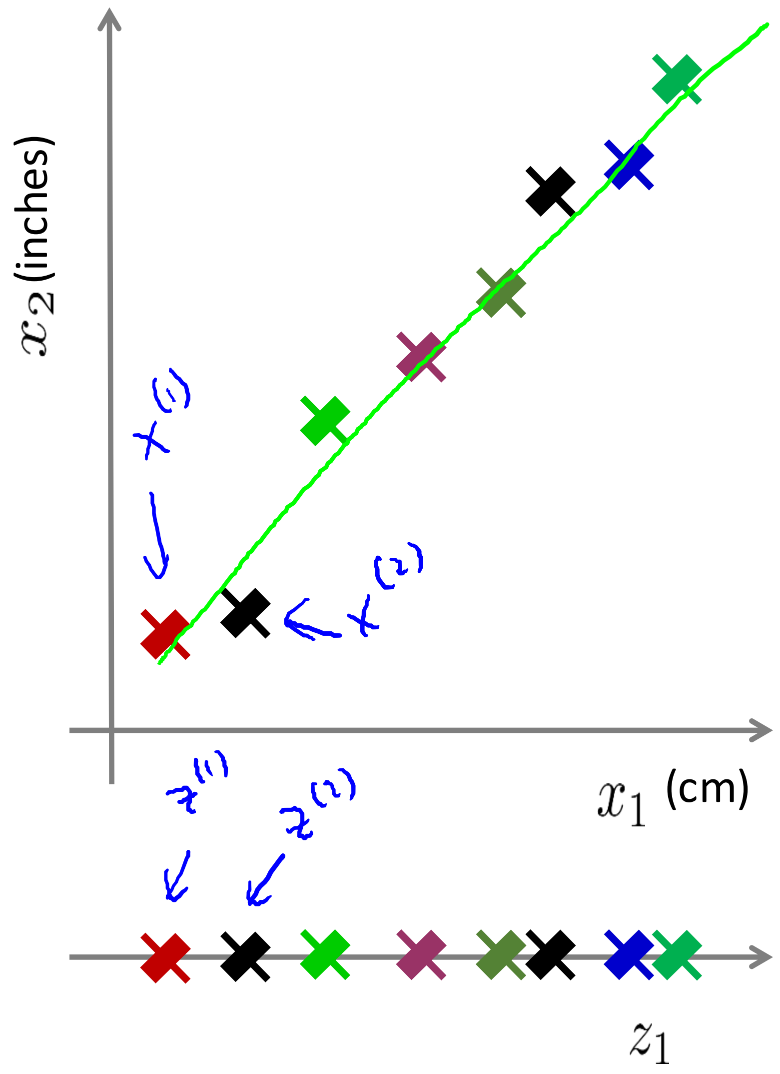
- Not all features (attributes) collected are independent.
- In fact, many attributes are highly correlated
  e.g. height and weight, total height and head height, attendance and GPA
- We can compact the data by choosing a new attribute that is a linear combination of the correlated factors

Data Compression



Reduce data from 2D to 1D

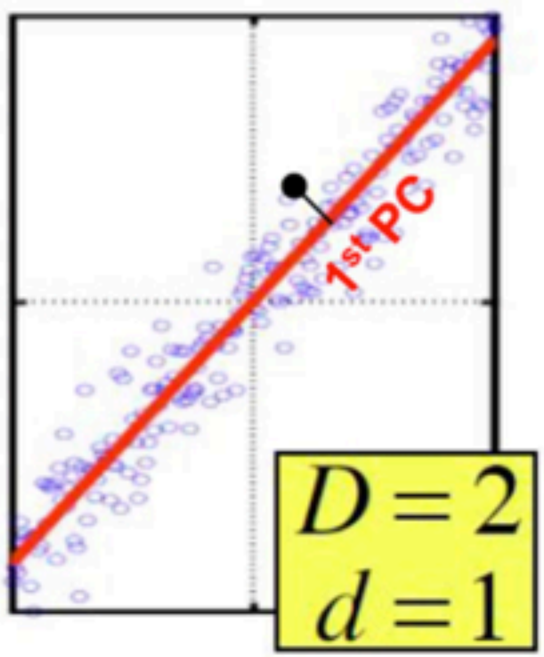
Data Compression



Reduce data from 2D to 1D

$$x^{(1)} \in \mathbb{R}^2 \rightarrow z^{(1)} \in \mathbb{R}$$
$$x^{(2)} \in \mathbb{R}^2 \rightarrow z^{(2)} \in \mathbb{R}$$
$$\vdots$$
$$x^{(m)} \in \mathbb{R}^2 \rightarrow z^{(m)} \in \mathbb{R}$$

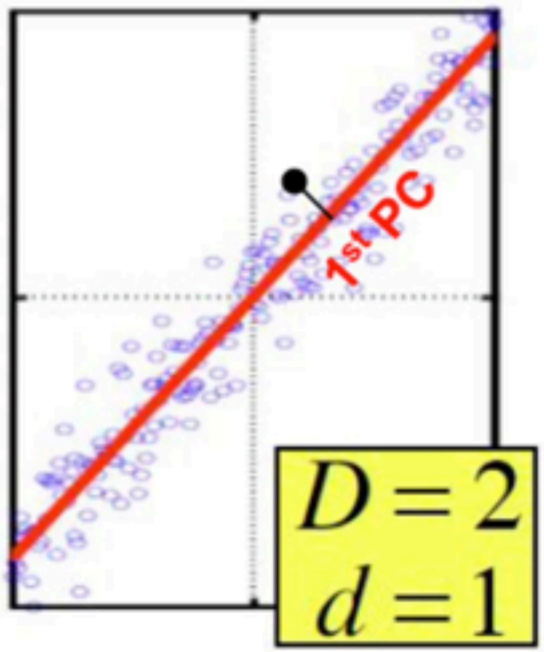
Principal components analysis



Principal Components (PC) are orthogonal directions that capture most of the variance in the data

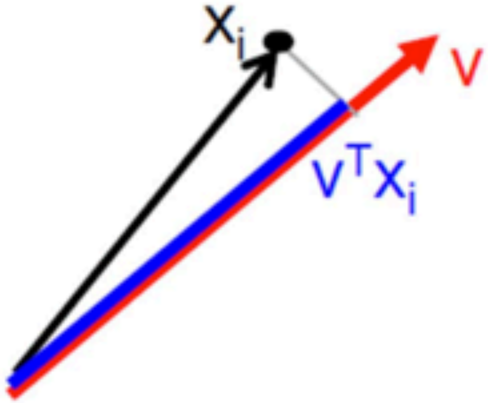
1st PC – direction of greatest variability in data

Principal components analysis



Principal Components (PC) are orthogonal directions that capture most of the variance in the data

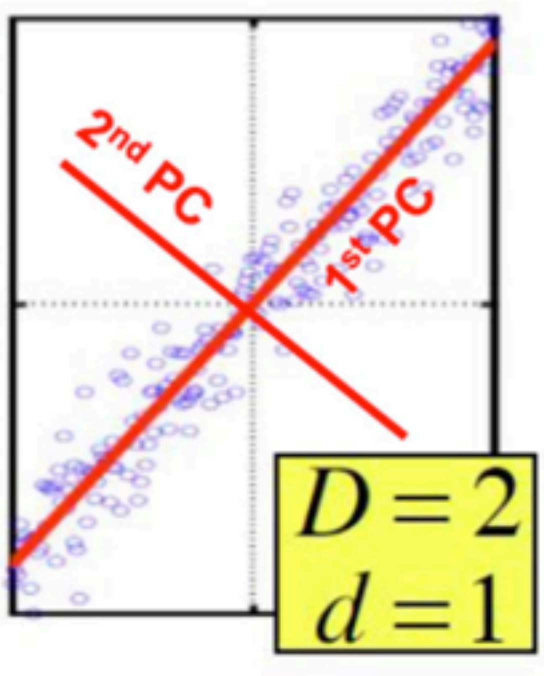
1st PC – direction of greatest variability in data



Take a data point x_i (D-dimensional vector)

Projection of x_i onto the 1st PC v is $v^T x_i$

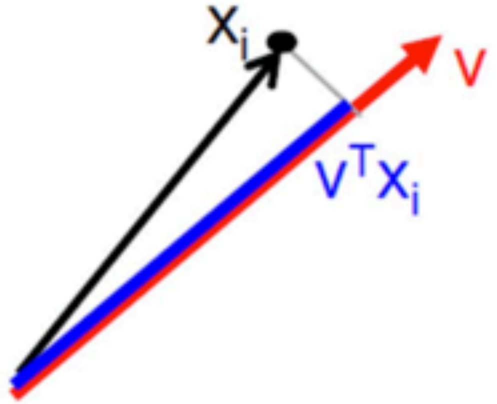
Principal components analysis



Principal Components (PC) are orthogonal directions that capture most of the variance in the data

1st PC – direction of greatest variability in data

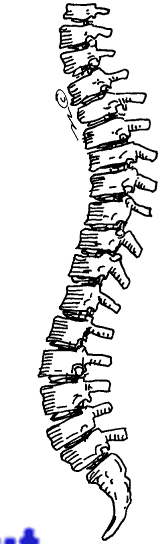
2nd PC – Next orthogonal (uncorrelated) direction of greatest variability



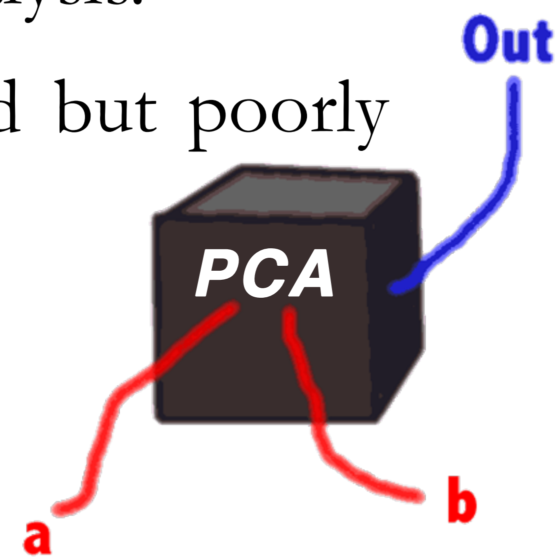
Take a data point x_i (D-dimensional vector)

Projection of x_i onto the 1st PC v is $v^T x_i$

PCA is …



- A backbone of modern data analysis.
- A black box that is widely used but poorly understood.



OK … let's dispel the magic behind this black box

PCA - Overview

- It is a mathematical tool from applied linear algebra.
- It is a simple, non-parametric method of extracting relevant information from confusing data sets.
- It provides a roadmap for how to reduce a complex data set to a lower dimension.

What do we need under our BELT?!!!

- Basics of statistical measures, e.g. variance and covariance.
- Basics of linear algebra:
  - Matrices
  - Vector space
  - Basis
  - Eigen vectors and eigen values

Variance

- A measure of the spread of the data in a data set with mean $\overline{X}$

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{\left(n - 1 \right)}$$

- Variance is claimed to be the original statistical measure of spread of data.

Covariance

- Variance – measure of the deviation from the mean for points in one dimension, e.g., heights
- Covariance – a measure of how much each of the dimensions varies from the mean with respect to each other.
- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions, e.g., number of hours studied and grade obtained.
- The covariance between one dimension and itself is the variance

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}{(n-1)}$$

$$\mathrm{cov}(X, Y) = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{(n-1)}$$

Covariance

- What is the interpretation of covariance calculations?
- Say you have a 2-dimensional data set
  - *X*: number of hours studied for a subject
  - *Y*: marks obtained in that subject
- And assume the covariance value (between X and Y) is: 104.53
- *What does this value mean*?

Covariance

- Exact value is not as important as its sign.

- A <u>positive value</u> of covariance indicates that both dimensions increase or decrease together, e.g., as the number of hours studied increases, the grades in that subject also increase.

- A <u>negative value</u> indicates while one increases the other decreases, or vice-versa, e.g., active social life vs. performance in ECE Dept.

- If <u>covariance is zero</u>: the two dimensions are independent of each other, e.g., heights of students vs. grades obtained in a subject.

Covariance

- Why bother with calculating (expensive) covariance when we could just plot the 2 values to see their relationship?

Covariance calculations are used to find relationships between dimensions in high dimensional data sets (usually greater than 3) where visualization is difficult.

Covariance Matrix

- Representing covariance among dimensions as a matrix, e.g., for 3 dimensions:

$$C = \begin{bmatrix} \operatorname{cov}(X,X) & \operatorname{cov}(X,Y) & \operatorname{cov}(X,Z) \\ \operatorname{cov}(Y,X) & \operatorname{cov}(Y,Y) & \operatorname{cov}(Y,Z) \\ \operatorname{cov}(Z,X) & \operatorname{cov}(Z,Y) & \operatorname{cov}(Z,Z) \end{bmatrix}$$

- Properties:
  - Diagonal: variances of the variables
  - $\operatorname{cov}(X,Y)=\operatorname{cov}(Y,X)$, hence matrix is symmetrical about the diagonal (upper triangular)
  - m-dimensional data will result in mxm covariance matrix

Transformation Matrices

- Consider the following:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

- The square (transformation) matrix scales (3,2)
- Now assume we take a multiple of (3,2)

$$2 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 6 \\ 4 \end{bmatrix} = \begin{bmatrix} 24 \\ 16 \end{bmatrix} = 4 \times \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

Transformation Matrices

- Scale vector (3,2) by a value 2 to get (6,4)
- Multiply by the square transformation matrix
- And we see that the result is still scaled by 4.

WHY?

A vector consists of both length and direction. Scaling a vector only changes its length and not its direction. This is an important observation in the transformation of matrices leading to formation of eigenvectors and eigenvalues.

Irrespective of how much we scale (3,2) by, the solution (under the given transformation matrix) is always a multiple of 4.

Eigenvalue Problem

- The eigenvalue problem is any problem having the following form:

$$\mathbf{A} \,.\, \mathbf{v} = \lambda \,.\, \mathbf{v}$$

A: m x m matrix

v: m x 1 non-zero vector

λ: scalar

- Any value of λ for which this equation has a solution is called the eigenvalue of A and the vector **v** which corresponds to this value is called the eigenvector of **A**.

Eigenvalue Problem

- Going back to our example:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$\mathbf{A} \;.\; \mathbf{v} \qquad = \lambda \;.\; \mathbf{v}$$

- Therefore, (3,2) is an eigenvector of the square matrix **A** and 4 is an eigenvalue of **A**
- The question is:

Given matrix A, how can we calculate the eigenvector and eigenvalues for A?

Calculating Eigenvectors & Eigenvalues

- Simple matrix algebra shows that:

$$\mathbf{A} \,.\, \mathbf{v} = \lambda \,.\, \mathbf{v}$$

$$\Leftrightarrow \qquad \mathbf{A} \,.\, \mathbf{v} - \lambda \,.\, \mathbf{I} \,.\, \mathbf{v} = \mathbf{0}$$

$$\Leftrightarrow \qquad (\mathbf{A} - \lambda \,.\, \mathbf{I}\,) .\, \mathbf{v} = \mathbf{0}$$

- Finding the roots of $|\mathbf{A} - \lambda \,.\, \mathbf{I}|$ will give the eigenvalues and for each of these eigenvalues there will be an eigenvector

Example …

Calculating Eigenvectors & Eigenvalues

- Let

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

- Then:

$$\left|A - \lambda.I\right| = \left|\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} - \lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right| = \left|\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right|$$

$$= \left|\begin{bmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{bmatrix}\right| = \left(-\lambda \times \left(-3-\lambda\right)\right) - \left(-2 \times 1\right) = \lambda^2 + 3\lambda + 2$$

- And setting the determinant to 0, we obtain 2 eigenvalues:

$\lambda_1 = -1$ and $\lambda_2 = -2$

Calculating Eigenvectors & Eigenvalues

- For λ_1 the eigenvector is:

$$(A - \lambda_1 . I).v_1 = 0$$

$$\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix} . \begin{bmatrix} v_{1:1} \\ v_{1:2} \end{bmatrix} = 0$$

$$v_{1:1} + v_{1:2} = 0 \qquad and \qquad -2v_{1:1} - 2v_{1:2} = 0$$

$$v_{1:1} = -v_{1:2}$$

- Therefore the first eigenvector is any column vector in which the two elements have equal magnitude and opposite sign.

Calculating Eigenvectors & Eigenvalues

- Therefore eigenvector v_1 is

$$v_1 = k_1 \begin{bmatrix} +1 \\ -1 \end{bmatrix}$$

where k_1 is some constant.

- Similarly we find that eigenvector v_2

$$v_2 = k_2 \begin{bmatrix} +1 \\ -2 \end{bmatrix}$$

where k_2 is some constant.

Now … Let's go back

PCA

Example of a problem

- We collected *m* parameters about 100 students:
  - Height
  - Weight
  - Hair color
  - Average grade
  - …

- We want to find the most important parameters that best describe a student.

Example of a problem

- Each student has a vector of data which describes him of length m:
  - (180,70,'purple',84,…)

- We have $n = 100$ such vectors. Let's put them in one matrix, where each column is one student vector.
- So we have a mxn matrix. This will be the input of our problem.

Which parameters can we ignore?

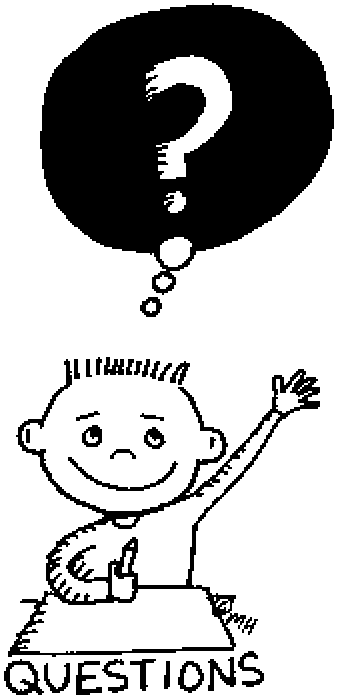
- Constant parameter (number of heads)
  - 1,1,…,1.
- Constant parameter with some noise - (thickness of hair)
  - 0.003, 0.005,0.002,….,0.0008 → low variance
- Parameter that is linearly dependent on other parameters (head size and height)
  - $Z = aX + bY$

Which parameters do we want to keep?

- Parameter that doesn't depend on others (e.g. eye color), i.e. uncorrelated ➔ low covariance.
- Parameter that changes a lot (grades)
  - The opposite of noise
  - High variance

Questions

- How we describe 'most important' features using math?
  - Variance
- How do we represent our data so that the most important features can be extracted easily?
  - Change of basis



Change of Basis !!!

- Let **X** and **Y** be $m \times n$ matrices related by a linear transformation **P**.
- **X** is the original recorded data set and **Y** is a re-representation of that data set.

$$\mathbf{PX} = \mathbf{Y}$$

- Let's define;
  - $\mathbf{p_i}$ are the rows of **P.**
  - $\mathbf{x_i}$ are the columns of **X.**
  - $\mathbf{y_i}$ are the columns of **Y.**

Change of Basis !!!

- Let **X** and **Y** be $m \times n$ matrices related by a linear transformation **P**.
- **X** is the original recorded data set and **Y** is a re-representation of that data set.

$$\mathbf{PX} = \mathbf{Y}$$

- <u>What does this mean?</u>
  - **P** is a matrix that transforms **X** into **Y**.
  - Geometrically, **P** is a rotation and a stretch (scaling) which again transforms **X** into **Y**.
  - The rows of **P**, $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m\}$ are a set of new basis vectors for expressing the columns of **X.**

Change of Basis !!!

- Lets write out the explicit dot products of **PX.**

$$\mathbf{PX} = \begin{bmatrix} \mathbf{p_1} \\ \vdots \\ \mathbf{p_m} \end{bmatrix} \begin{bmatrix} \mathbf{x_1} & \cdots & \mathbf{x_n} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{p_1 \cdot x_1} & \cdots & \mathbf{p_1 \cdot x_n} \\ \vdots & \ddots & \vdots \\ \mathbf{p_m \cdot x_1} & \cdots & \mathbf{p_m \cdot x_n} \end{bmatrix}$$

- We can note the form of each column of **Y**.

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{p_1 \cdot x_i} \\ \vdots \\ \mathbf{p_m \cdot x_i} \end{bmatrix}$$

- We can see that each coefficient of $\mathbf{y_i}$ is a dot-product of $\mathbf{x_i}$ with the corresponding row in **P**.

In other words, the j^{th} coefficient of y_i is a projection onto the j^{th} row of P.

Therefore, the *rows* of P are a new set of basis vectors for representing the *columns* of X.

Questions Remaining !!!

- Assuming linearity, the problem now is to find the appropriate *change of basis*.
- The *row* vectors $\{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_m}\}$ in this transformation will become the *principal components* of **X**.
- Now,
  - What is the best way to *re-express* **X**?
  - What is the good choice of basis **P**?
- Moreover,
  - *what features we would like* ***Y*** *to exhibit?*

Covariance Matrix

- Assuming zero mean data (subtract the mean), consider the indexed vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ which are the *rows* of an *mxn* matrix $\mathbf{X}$.

- Each row corresponds to all measurements of a particular measurement type or attribute($\mathbf{x}_i$).

- Each column of $\mathbf{X}$ corresponds to a set of measurements from particular instance or example.

- We now arrive at a definition for the covariance matrix $\mathbf{S}_\mathbf{X}$.

$$\mathbf{S}_\mathbf{X} \equiv \frac{1}{n-1}\mathbf{X}\mathbf{X}^T \quad \text{where} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_\mathbf{m} \end{bmatrix}$$

Covariance Matrix

$$\mathbf{S_X} \equiv \frac{1}{n-1}\mathbf{X}\mathbf{X}^T \quad \text{where} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_m} \end{bmatrix}$$

- The ij^{th} element of the variance is the dot product between the vector of the i^{th} measurement type with the vector of the j^{th} measurement type.
  - $\mathbf{S_X}$ is a square symmetric $m \times m$ matrix.
  - The diagonal terms of $\mathbf{S_X}$ are the variance of particular measurement types.
  - The off-diagonal terms of $\mathbf{S_X}$ are the covariance between measurement types.

Diagonalize the Covariance Matrix

Our goals are to find the covariance matrix that:

1. Minimizes redundancy, measured by covariance. (off-diagonal), i.e. we would like each variable to co-vary as little as possible with other variables.

2. Maximizes the signal, measured by variance. (the diagonal)

Since covariance is non-negative, the optimized covariance matrix will be a diagonal matrix.

Diagonalize the Covariance Matrix

PCA Assumptions

- PCA assumes that all basis vectors $\{\mathbf{p_1}, \ldots, \mathbf{p_m}\}$ are orthonormal (i.e. $\mathbf{p_i} \cdot \mathbf{p_j} = \delta_{ij}$).
- Hence, in the language of linear algebra, PCA assumes $\mathbf{P}$ is an orthonormal matrix.
- Secondly, PCA assumes the directions with the largest variances are the most "important" or in other words, most principal.
- *Why are these assumptions easiest?*

Solving PCA: Eigen Vectors of Covariance Matrix

- We will derive our first algebraic solution to PCA using linear algebra. This solution is based on an important property of *eigenvector decomposition*.

- Once again, the data set is $\mathbf{X}$, an $m \times n$ matrix, where m is the number of measurement types and n is the number of data trials.

- The goal is summarized as follows:

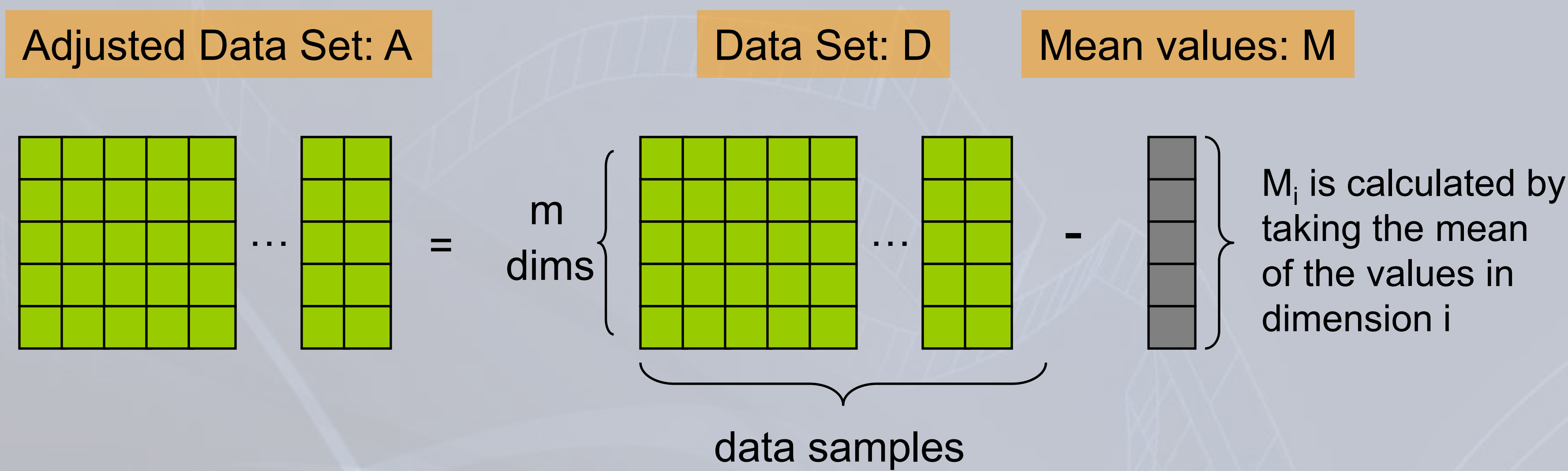
  – Find some orthonormal matrix $\mathbf{P}$ where $\mathbf{Y} = \mathbf{PX}$ such that is $\mathbf{S_Y} \equiv \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T$ diagonalized. The rows of $\mathbf{P}$ are the *principal components* of $\mathbf{X}$.

PCA Process – STEP 1

- Subtract the mean from each of the dimensions
- This produces a data set whose mean is zero.
- Subtracting the mean makes variance and covariance calculation easier by simplifying their equations.
- The variance and co-variance values are not affected by the mean value.
- Suppose we have two measurement types $\mathbf{X}_1$ and $\mathbf{X}_2$, hence $m = 2$, and ten samples each, hence $n = 10$.

Principal Component Analysis (PCA)

Steps in PCA: #1 Calculate Adjusted Data Set



PCA Process – STEP 1

X_1	X_2				X'_1	X'_2
2.5	2.4				0.69	0.49
0.5	0.7				-1.31	-1.21
2.2	2.9				0.39	0.99
1.9	2.2				0.09	0.29
3.1	3.0	$\Rightarrow$	$\overline{X_1} = 1.81$, $\overline{X_2} = 1.91$	$\Rightarrow$	1.29	1.09
2.3	2.7				0.49	0.79
2.0	1.6				0.19	-0.31
1.0	1.1				-0.81	-0.81
1.5	1.6				-0.31	-0.31
1.2	0.9				-0.71	-1.01

http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf

PCA Process – STEP 2

- Calculate the covariance matrix

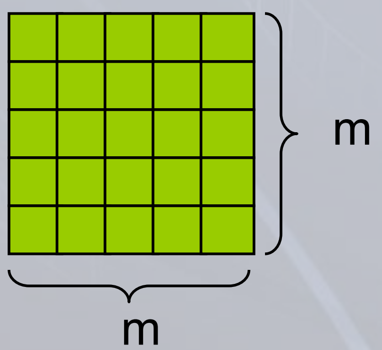
$$\text{cov} = \begin{bmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{bmatrix}$$

- Since the non-diagonal elements in this covariance matrix are positive, we should expect that both the $\mathbf{X}_1$ and $\mathbf{X}_2$ variables increase together.
- Since it is symmetric, we expect the eigenvectors to be orthogonal.

Principal Component Analysis (PCA)

Steps in PCA: #2 Calculate Co-variance matrix, C, from Adjusted Data Set, A

Co-variance Matrix: C



$C_{ij} = \text{cov}(i,j)$

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Note: Since the means of the dimensions in the adjusted data set, A, are 0, the covariance matrix can simply be written as:

$$C = (A\,A^T) / (n-1)$$

PCA Process – STEP 3

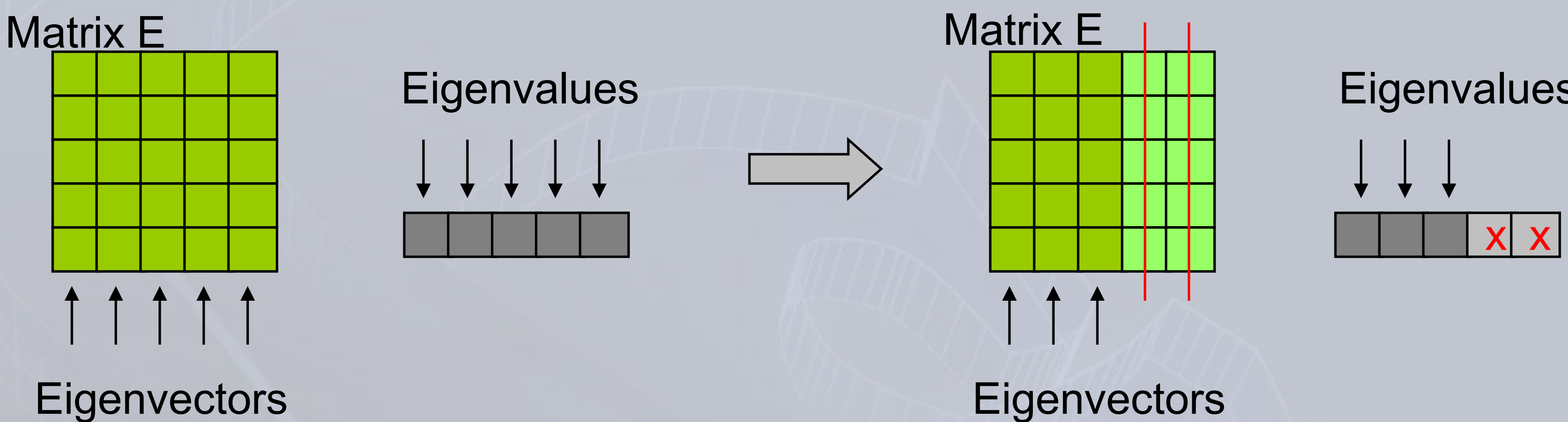
- Calculate the eigenvectors and eigenvalues of the covariance matrix

$$eigenvalues = \begin{bmatrix} 0.490833989 \\ 1.28402771 \end{bmatrix}$$

$$eigenvectors = \begin{bmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{bmatrix}$$

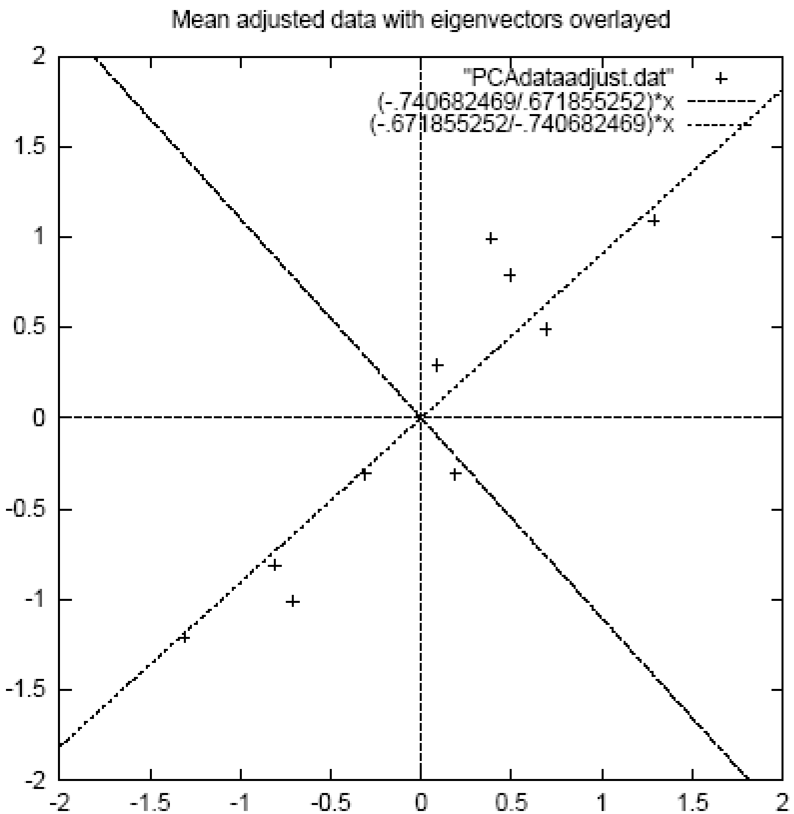
Principal Component Analysis (PCA)

Steps in PCA: #3 Calculate eigenvectors and eigenvalues of C



If some eigenvalues are 0 or very small, we can essentially discard those eigenvalues and the corresponding eigenvectors, hence reducing the dimensionality of the new basis.

PCA Process – STEP 3



A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

Eigenvectors are plotted as diagonal dotted lines on the plot. (note: they are perpendicular to each other).

One of the eigenvectors goes through the middle of the points, like drawing a line of best fit.

The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

PCA Process – STEP 4

- Reduce dimensionality and form *feature vector*

The eigenvector with the *highest* eigenvalue is the *principal component* of the data set.

In our example, the eigenvector with the largest eigenvalue is the one that points down the middle of the data.

Once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives the components in order of significance.

PCA Process – STEP 4

Now, if you'd like, you can decide to *ignore* the components of lesser significance.

You do lose some information, but if the eigenvalues are small, you don't lose much

- *m* dimensions in your data
- calculate *m* eigenvectors and eigenvalues
- choose only the first *r* eigenvectors
- final data set has only *r* dimensions.

PCA Process – STEP 4

- When the λ_i's are sorted in descending order, the proportion of variance explained by the *r* principal components is:

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{m} \lambda_i} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_r}{\lambda_1 + \lambda_2 + \ldots + \lambda_p + \ldots + \lambda_m}$$

- If the dimensions are highly correlated, there will be a small number of eigenvectors with large eigenvalues and *r* will be much smaller than *m*.
- If the dimensions are not correlated, *r* will be as large as *m* and PCA does not help.

PCA Process – STEP 4

- Feature Vector

$$\text{FeatureVector} = (\lambda_1\ \lambda_2\ \lambda_3\ \ldots\ \lambda_r)$$

(take the eigenvectors to keep from the ordered list of eigenvectors, and form a matrix with these eigenvectors in the columns)

We can either form a feature vector with both of the eigenvectors:

$$\begin{bmatrix} -0.677873399 & -0.735178656 \\ -0.735178656 & 0.677873399 \end{bmatrix}$$

or, we can choose to leave out the smaller, less significant component and only have a single column:

$$\begin{bmatrix} -0.677873399 \\ -0.735178656 \end{bmatrix}$$

PCA Process – STEP 5

- Derive the new data

FinalData = RowFeatureVector x RowZeroMeanData

RowFeatureVector is the matrix with the eigenvectors in the columns *transposed* so that the eigenvectors are now in the rows, with the most significant eigenvector at the top.

RowZeroMeanData is the mean-adjusted data *transposed*, i.e., the data items are in each column, with each row holding a separate dimension.

PCA Process – STEP 5

- FinalData is the final data set, with data items in columns, and dimensions along rows.

- What does this give us?

  The original data *solely in terms of the vectors we chose*.

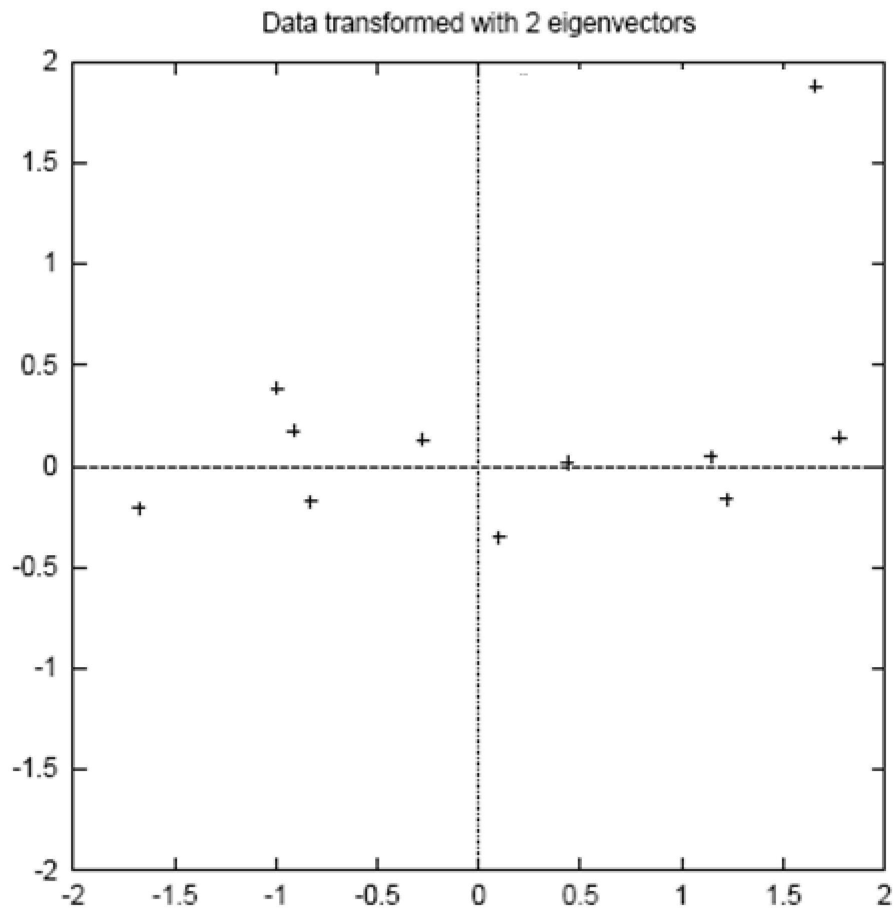
- We have changed our data from being in terms of the axes $\mathbf{X_1}$ and $\mathbf{X_2}$, to now be in terms of our 2 eigenvectors.

PCA Process – STEP 5

FinalData (transpose: dimensions along columns)

$newX_1$	$newX_2$
-0.827870186	-0.175115307
1.77758033	0.142857227
-0.992197494	0.384374989
-0.274210416	0.130417207
-1.67580142	-0.209498461
-0.912949103	0.175282444
0.0991094375	-0.349824698
1.14457216	0.0464172582
0.438046137	0.0177646297
1.22382956	-0.162675287

PCA Process – STEP 5



The table of data by applying the PCA analysis using both eigenvectors, and a plot of the new data points.

Singular Value Decomposition

Singular Value Decomposition

$$A_{mn} = U_{mm} S_{mn} V_{nn}^T$$

where $U^T U = I, V^T V = I$; the columns of U are orthonormal eigenvectors of AA^T, the columns of V are orthonormal eigenvectors of $A^T A$, and S is a diagonal matrix containing the square roots of eigenvalues from U or V in descending order.

Singular Value Decomposition (SVD)

SVD of a Matrix

$M = U \; S \; V^T$

U and V are orthogonal matrices, and S is a diagonal matrix consisting of singular values.

Applications of PCA

- Exploratory data analysis
- Data preprocessing, dimensionality reduction
- Data compression, data reconstruction
  - (lossy) data compression technique
  - The table describing the data with first k-principal components is smaller than original data table

References

- J. SHLENS, "TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS," 2005 http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf
- Introduction to PCA, http://dml.cs.byu.edu/~cgc/docs/dm/Slides/