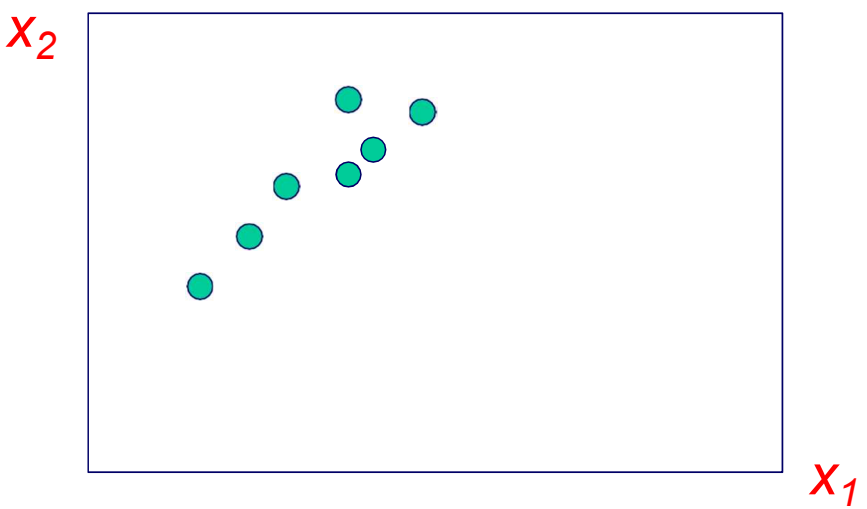
Component Analysis and Discriminants

Reducing dimensionality when:

1. Classes are disregarded
   Principal Component Analysis (PCA)
2. Classes are considered
   Discriminant Analysis



Component Analysis vs Discriminants

Two classical approaches for finding effective linear transformations

- PCA (Principal Component Analysis)
  "Projection that best represents data in least-square sense"

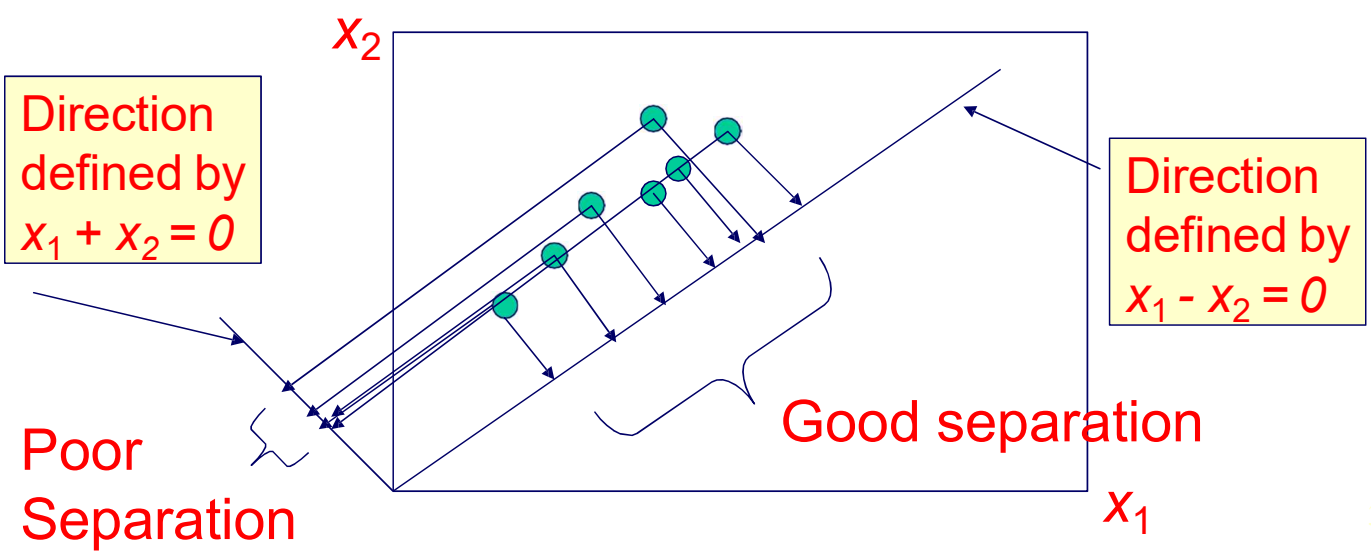
- MDA (Multiple Discriminant Analysis)
  "Projection that best separates the data in a least-squares sense"

PCA: Linear Projections of Data

- Excessive dimensionality $\mathbf{x} = [x_1, x_2, \ldots x_d]$ causes
  - Computational difficulty and overfitting
  - Visualization issues
- Solution:
  - Combine features to reduce dimensionality
  - Linear combinations, *e.g.*, $2x_1+3x_2+x_3$
    - are simple to compute and tractable
    - Project high dimensional data onto a lower dimensional space

Projection to a lower dimensional space

- Allow computer to search for interesting directions



Linear Projection

- Equation of a plane that passes through origin: $x_1 + 2x_2 + 4x_3 = 0$

$$[1\ 2\ 4]\ \mathrm{x}\ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1 + 2x_2 + 4x_3 = 0$$

Can be written as $\mathbf{w}^T\mathbf{x} = 0$ where $\mathbf{w} = [1\ 2\ 4]^T$

- Projection of a point x along a line with projection weights **a** is given by:

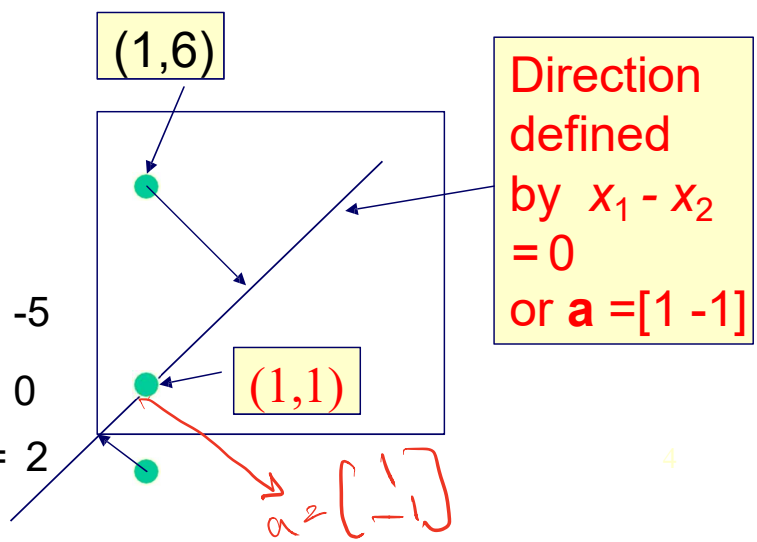
$$\mathrm{a}^T\mathrm{x} = \sum_{j=1}^{d} \mathrm{a}_j x_j$$

- Example:

Projection of (1,6) along [1 -1]: 1 - 6 = -5

Projection of (1,1) along [1 -1]: 1 - 1 = 0

Projection of (1,-1) along [1 -1]: 1 + 1 = 2

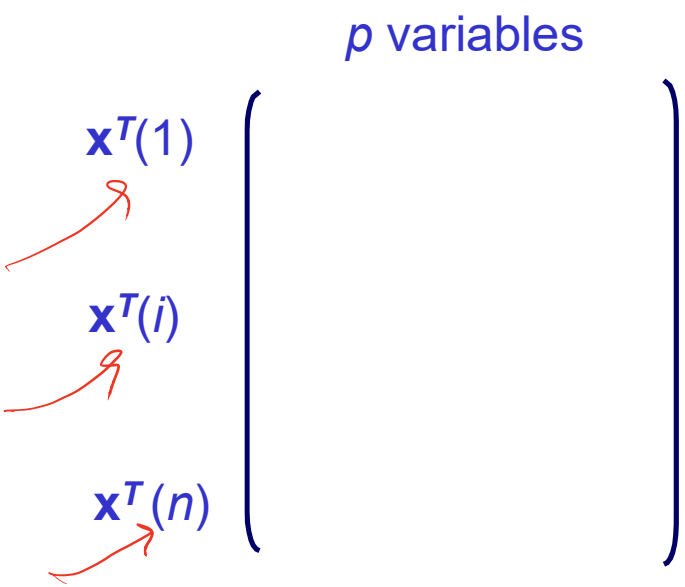


Principal Component Analysis

- We will look at Data Matrix, Scatter Matrix and Covariance Matrix to arrive at best projection of data

Data Matrix

Let **X** be a *n x p data matrix* of *n* samples



x(*i*) is a *p* x 1 column vector

Each row of matrix is of the form $\mathbf{x}^T(i)$

Assume **X** is centralized, so that the mean m_i of each variable is subtracted for that variable $\mathbf{m}=[m_1,m_2,\ldots,m_p]^T$

Scatter Matrix

$$\mathbf{S}_{p\times p} = \sum_{i=1}^{n} \Big(\mathbf{x}(i) - \mathbf{m}\Big)\Big(\mathbf{x}(i) - \mathbf{m}\Big)^{T}$$

- The Scatter Matrix **S** is (*n*-1) times the sample Covariance Matrix
- Relationship Between Data Matrix and Scatter Matrix:
  - **S** = **X** ^{T} **X** is the *p* x *p* scatter matrix of the data since **X** has zero mean

Projection

Let **a** be a $p \times 1$ column vector of projection weights that result in the largest variance when the data matrix **X** are projected along **a.**

The projection of any data vector **x** is the linear combination:

$$\mathbf{a}^T\mathbf{x} = \sum_{j=1}^{p} a_j x_j$$

Projected values of all data vectors in data matrix **X** onto **a** can be expressed as **Xa** which is an $n \times 1$ column vector.

Variance along Projection

We define the Sample Variance along **a** as the variance of the sample containing $\mathbf{a}^T\mathbf{x}(i)$'s

$$\mathrm{Var}_{\mathbf{a}}\left(\mathbf{X}\right) \propto \left(\mathbf{Xa}\right)^T\left(\mathbf{Xa}\right) = a^T X^T X a$$

$$= \mathbf{a}^T\mathbf{S}\mathbf{a}$$

where **S** is is the scatter matrix of **S**.

$$Xa = \begin{bmatrix} x(1)^T a \\ x(2)^T a \\ \vdots \\ x(n)^T a \end{bmatrix}$$

$$\text{Var(sample)} \propto \sum_{i=1}^{n} \left(x(i)^T a\right)^2$$

$$= \frac{1}{n-1}(Xa)^T(Xa)$$

$$= \frac{1}{n-1}\|Xa\|^2$$

Variance along Projection

Therefore, the variance is a function of both **X** and **a**.

Maximizing variance along **a** is not well-defined since we can increase it without limit by increasing the size of the components of **a**.

Find a subject to

$\|a\|^2 = 1$

Optimization Problem

Impose a normalization constraint on **a** i.e,

$\mathbf{a}^T\mathbf{a} = 1$

Optimization problem is to maximize

$$u(\mathbf{a}) = \mathbf{a}^T\mathbf{S}\mathbf{a}$$

Variance Criterion

subject to

$$\mathbf{a}^T\mathbf{a} = 1$$

Normalization Criterion

Optimization Problem

Impose a normalization constraint on **a** i.e,
$\mathbf{a}^T\mathbf{a} = 1$
Optimization problem is to maximize

$$u(\mathbf{a}) = \mathbf{a}^T\mathbf{S}\mathbf{a} - \lambda\left(\mathbf{a}^T\mathbf{a} - 1\right)$$

Variance Criterion

Normalization Criterion

Where λ is a Lagrange multiplier.

Solution: Differentiating wrt **a** yields

$$\nabla_{\mathbf{a}} u(\mathbf{a}) = 0 \qquad 2\mathbf{S}\mathbf{a} - 2\lambda\mathbf{a} = 0$$

which reduces to

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = 0$$

Characteristic Equation of S!

Candidate a's : eigenvalues of S

Optimization Problem

Impose a normalization constraint on **a** i.e,
$\mathbf{a}^T\mathbf{a} = 1$
Optimization problem is to maximize

$$u(\mathbf{a}) = \mathbf{a}^T\mathbf{S}\mathbf{a} - \lambda\left(\mathbf{a}^T\mathbf{a} - 1\right)$$

Variance Criterion

Normalization Criterion

Where λ is a Lagrange multiplier.

Solution: Differentiating wrt λ yields

$a^T a - 1 = 0 \Rightarrow a^T a = 1$

$\Rightarrow$ Orthonormal eigenvalues of S are solution candidates

Characteristic Equation

Given a p x p matrix **M**, a very important class of linear equations is of the form

$$\mathbf{M}_{pxp}\,\mathbf{x}_{px1} = \lambda\mathbf{x}_{px1}$$

which can be rewritten as $(\mathbf{M}-\lambda\mathbf{I})\mathbf{x} = \mathbf{0}$.

If **M** is real and symmetric there are p possible linearly independent and orthonormal solution vectors (vectors **x** that satisfy the characteristic equation) called eigenvectors, $\mathbf{e}_1, .., \mathbf{e}_p$ and associated eigenvalues $\lambda_1, ..., \lambda_p$, for which $\det(\mathbf{M}-\lambda\mathbf{I})=0$.

Principal Components

If the matrix **M** is the Scatter matrix **S**, the Characteristic Equation is:

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = 0$$
$$\det(\mathbf{S} - \lambda\mathbf{I}) = 0$$

The roots are eigenvalues of **S.**
Therefore, the candidate solutions to the optimization problem are eigenvectors of **S**, and their corresponding values.

Principal Components

If the matrix **M** is the Scatter matrix **S**, the Characteristic Equation is:

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = 0$$
$$\det(\mathbf{S} - \lambda\mathbf{I}) = 0$$

The roots are eigenvalues of **S.**
The corresponding eigenvectors are called the principal components.
First principal component is the eigenvector associated with the largest eigenvalue of **S**.

Variance along each eigenvector

The variance along each eigenvector $\mathbf{e}_j$ is calculated as:

$$\mathrm{Var}_{\mathbf{e}_j}(\mathbf{X}) \propto (\mathbf{X}\mathbf{e}_j)^T (\mathbf{X}\mathbf{e}_j)$$

$$= e_j^T X^T X e_j = e_j^T S e_j = e_j^T \lambda_j e_j = \lambda_j e_j^T e_j$$

$$= \lambda_j$$

Variance along each eigenvector

- The variance along each eigenvector $\mathbf{e}_j$ is the eigenvalue associated with that eigenvector. Therefore, the solution to the previous optimization problem is $\mathbf{e}_{max}$, the eigenvector associated with the largest eigenvalue, λ_{max}.
- The maximum variance is λ_{max}.

Variance along each eigenvector

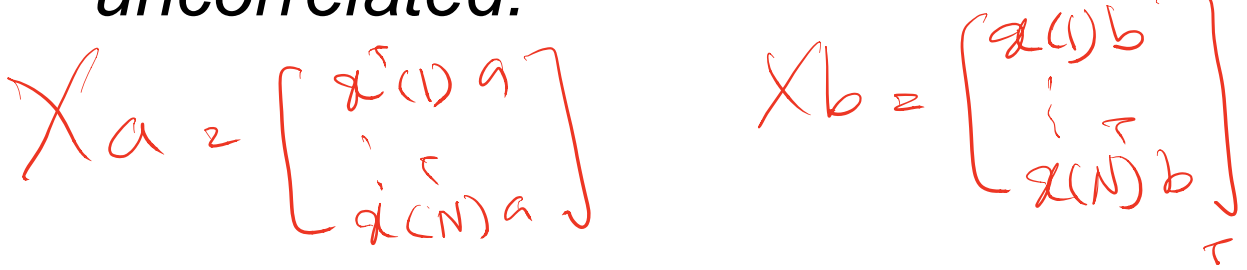
- If we want to find the direction that is *constrained to be orthogonal* to $\mathbf{e}_{max}$ which has the highest variance, we have *p*-1 candidates: the other eigenvalues of **S**.
- Obviously, it would be the direction of the eigenvector that has the second largest eigenvalue.

Variance along each eigenvector

- If we want to find the direction that is *constrained to be orthogonal* to both previous directions, which has the highest variance, we have *p*-2 candidates: the other eigenvalues of **S**.
- Obviously, it would be the direction of the eigenvector that has the third largest eigenvalue.

Variance along each eigenvector

- What does the orthogonality constraint do?
- Assume that **a** and **b** are two vectors that are orthogonal, i.e. $\mathbf{a}^T\mathbf{b} = 0$. We show that the samples of data points projected along **a** and **b** are *uncorrelated.*



$$\text{Trace}\left((Xa)^T(Xb)\right) = \text{Tr}(a^T X^T X b)$$

not needed

$$= \sum_i (x^{(i)T} a)(x^{(i)T} b)$$

$$= \text{Tr}(a^T S b) = \text{Tr}(S b a^T)$$

$$= \text{Tr}(S)\,\text{Tr}(b a^T) = \text{Tr}(S)\,\text{Tr}(a^T b)$$

$a^T b$ is a scalar, so $\text{Tr}(a^T b) = a^T b$; this is not true!

We must show that

$$= \sum_i [x^{(i)T} a][a^{(i)T} b]$$

$$(Xa)^T(Xb) = 0$$

which is

equivalent to $\sum_i (x^{(i)T} a)(x^{(i)T} b)$

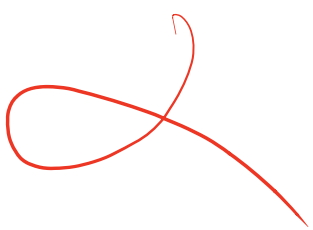
$= 0$

but $(Xa)^T(Xb) = a^T X^T X b$

$= a^T S b = a^T \lambda_b b = \lambda_b a^T b = 0$

Variance along each eigenvector

- This means **Xa** (vector of the samples $\mathbf{a}^T\mathbf{x}(i)$) and **Xb** (vector of the samples $\mathbf{b}^T\mathbf{x}(i)$) are orthogonal:



a, b orthogonal eigenvectors of S

Variance along each eigenvector

- We have *p* linearly independent and orthogonal eigenvectors, so we can continue doing this and find *p* directions that have the highest variance subject to being orthogonal to all other directions.

Variance along each eigenvector

- This means the samples obtained from projection of the data points along each direction are uncorrelated from one another!

Principal Directions

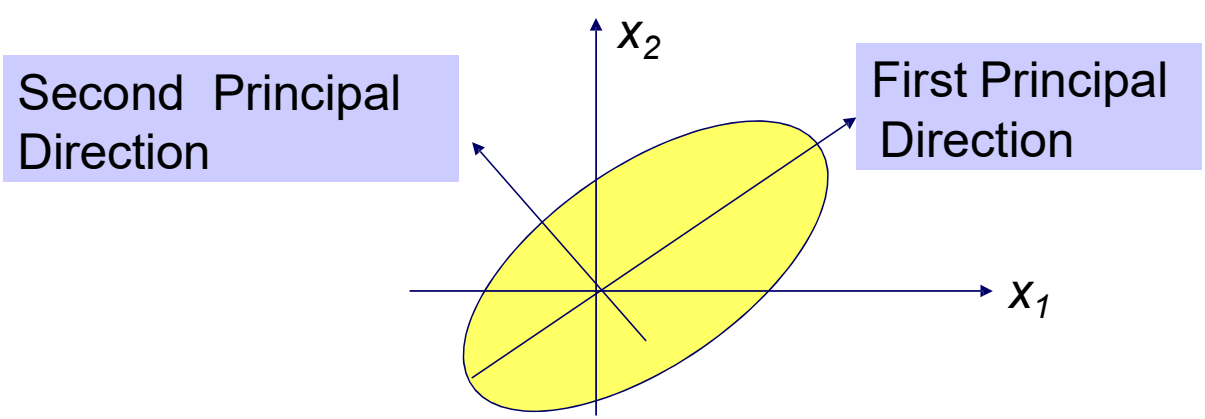
- Therefore, we are looking for directions that have the highest possible variance, which give us projected **x**(*i*)'s that are uncorrelated from one another.
- Those directions are called principal directions or principal component loading vectors.

First Principal Direction

- First principal component is the eigenvector associated with the largest eigenvalue of **S**.
- It is the direction along which **X** has the largest variation.
- Second Principal direction is:
  - In a direction orthogonal to first. Therefore, it is uncorrelated with the first principal direction.
  - Has second largest eigenvalue.

Other Principal Directions

- All principal directionss are in direction orthogonal to one another, so they are uncorrelated.
- The variation along each principal directions depends on how large its eigenvalue is.



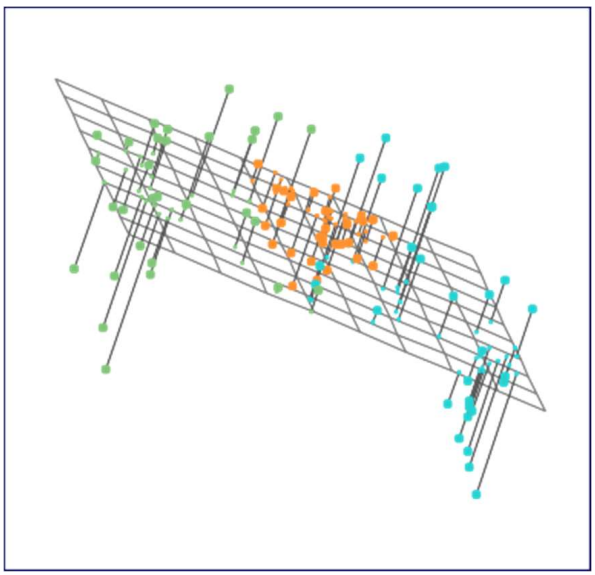
Principal Components

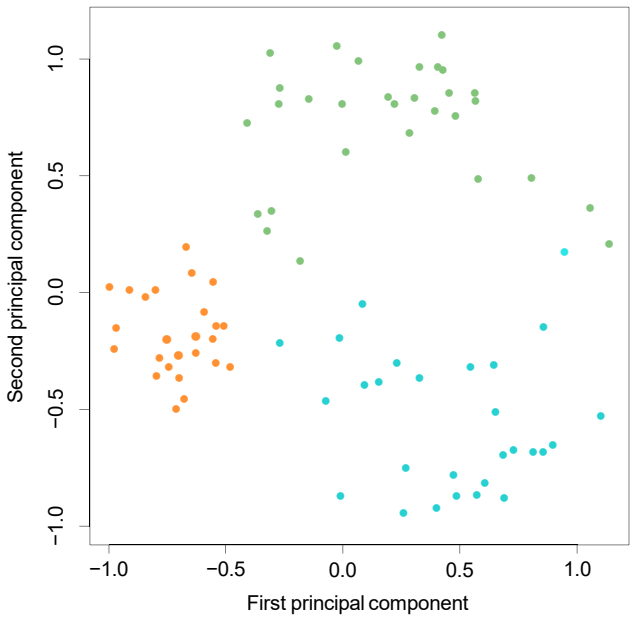
- The component of each data point along each principal direction is called a principal components or principal component score of that data point.
- We call the component of each data point along the i^{th} principal direction, its ith principal component.
- Principal components are found by projecting data points onto principal directions.

Alternative Derivation of PCA

- DHS Text gives another way to derive the fact that eigenvectors of scatter matrix are principal components
- By setting up the squared error criterion function of the data and the vector **e** that minimizes it satisfies the characteristic equation
- Can be generalized to from one-dimensional projection to a dimension *p'* which are the eigenvectors of **S** forming the principal components

Another Interpretation of Principal Components





The first three principal components of a data set span the three-dimensional hyperplane that is closest to the *n* observations,

PCA find the hyperplane closest to the observations

- The first principal component loading vector (direction) has a very special property: it defines the line in *p*-dimensional space that is *closest* to the *n* observations (using average squared Euclidean distance as a measure of closeness)

PCA find the hyperplane closest to the observations

- The notion of principal directions as the dimensions that are closest to the *n* observations extends beyond just the first principal direction.
- For instance, the first two principal directions of a data set span the plane that is closest to the *n* observations, in terms of average squared Euclidean distance.

Projection into *k* Eigenvectors

Variance of data projected into first *k* eigenvectors is proportional to

$$\sum_{j=1}^{k} \lambda_j$$

That is because:

$$\mathrm{Var}_{\mathbf{e}_j}\left(\mathbf{X}\right) \propto \left(\mathbf{X}\mathbf{e}_j\right)^T \left(\mathbf{X}\mathbf{e}_j\right)$$

$$= \lambda_j$$

Projection into *k* Eigenvectors

and:

$$\mathrm{Var}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_k}(\mathbf{X}) \propto$$

Projections along eigenvectors:

$$T = [X\mathbf{e}_1 \quad X\mathbf{e}_2 \quad \cdots \quad X\mathbf{e}_k]$$

$$= X V_k$$

$$V_k [\mathbf{e}_1 \cdots \mathbf{e}_k]$$

Var of Samples $\propto \sum_i \sum_j T_{ij}^2$ ✓

$$= \|T\|_F^2 = \left(\sqrt{\mathrm{Tr}(T^T T)}\right)^2 \quad \mathrm{Tr}(T^T T)$$

$$= \mathrm{Tr}\left[(XV_k)^T (XV_k)\right]$$

$$= \mathrm{Tr}\left(V_k^T X^T X V_k\right) = \mathrm{Tr}\left(V_k^T S V_k\right)$$

$$\mathrm{Tr}\left([v_1 \cdots v_k]^T [Sv_1 \cdots Sv_k]\right)$$

$$= \mathrm{Tr}\left([v_1 \cdots v_k]^T [\lambda_1 v_1 \cdots \lambda_k v_k]\right)$$

$$= \mathrm{Tr}\left(\begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix} [\lambda_1 v_1 \cdots \lambda_k v_k]\right)$$

$$= Tr\left(\begin{bmatrix} \lambda_1 v_1^T v_1 & \lambda_2 v_1^T v_2 & 0 & \cdots & 0 \\ \underbrace{\quad}_{1} & \underbrace{\quad}_{0} & & & \\ & \lambda_2 & & & \\ & & \ddots & & 0 \\ & & & & \lambda_k \end{bmatrix} \right)$$

$$= \sum_{i=1}^{k} \lambda_k$$

If we use all eigenvalues

$$\text{Var}_{e_1 \cdots e_p} \propto \sum_{k=1}^{p} \lambda_p$$

On the other hand

$$\text{Var}_{e_1 \cdots e_p} \propto Tr\left(V^T S V\right)$$

$$= Tr\left(\underbrace{V V^T}_{I_{p \times p}} S\right) = Tr(S)$$

$= Tr(X^T X) = \|X\|_F^2 = \sum_{ij} X_{ij}^2$

$\propto Var(\text{all } q(i)s) = \sum_{j=1}^{p} Var(X_j)$

remember that variance has a normalization factor $\frac{1}{n-1}$

Variance of the whole data

= Variance of the principal components $\propto \sum_{i=1}^{p} \lambda_i$

Projection into *k* Eigenvectors

- Variance of data projected into first *k* eigenvectors is proportional to

$$\sum_{j=1}^{k} \lambda_j$$

- The proportion of variance lost in approximating true data matrix **X** using only first *k* eigenvectors is

$$\frac{\sum_{j=k+1}^{p} \lambda_j}{\sum_{l=1}^{p} \lambda_l}$$

Usually 5 to 10 principal components capture 90% of variance in the data

How Many Principal Components Should One Use?

When Using PCA for classification, one can consider the number of components used as a hyper parameter and use cross-validation to select the number of components.

How Many Principal Components Should One Use?

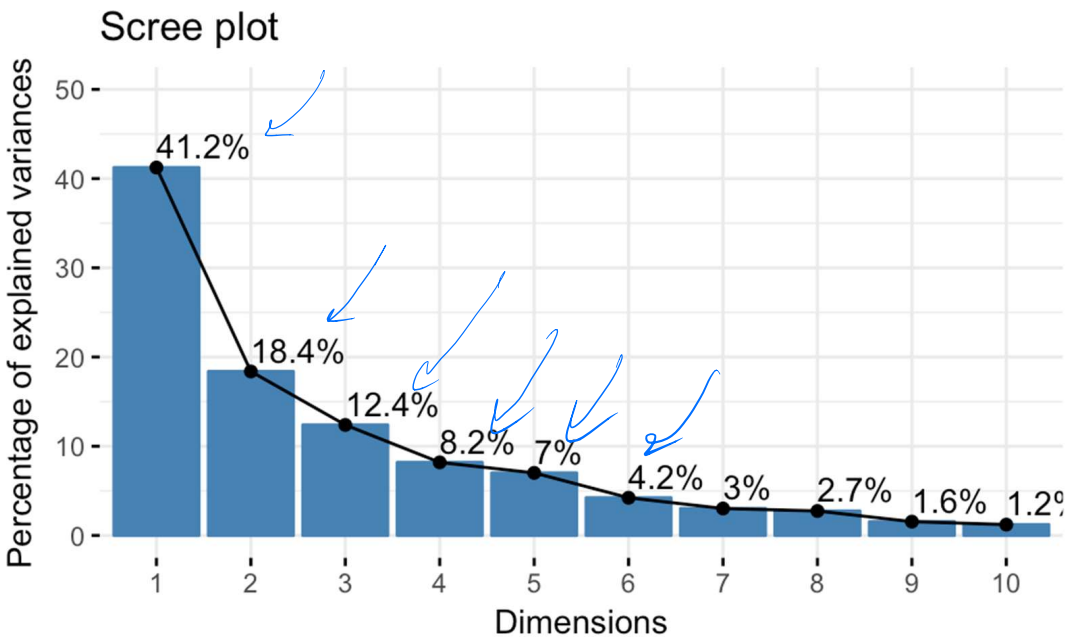
An alternative method to determine the number of principal components merely from unlabeled **X** is to look at a Scree Plot, which is the plot of eigenvalues ordered from largest to the smallest.

How Many Principal Components Should One Use?

- The number of component is determined at the point, beyond which the remaining eigenvalues are all relatively small and of comparable size (Jollife 2002, Peres-Neto, Jackson, and Somers (2005)).

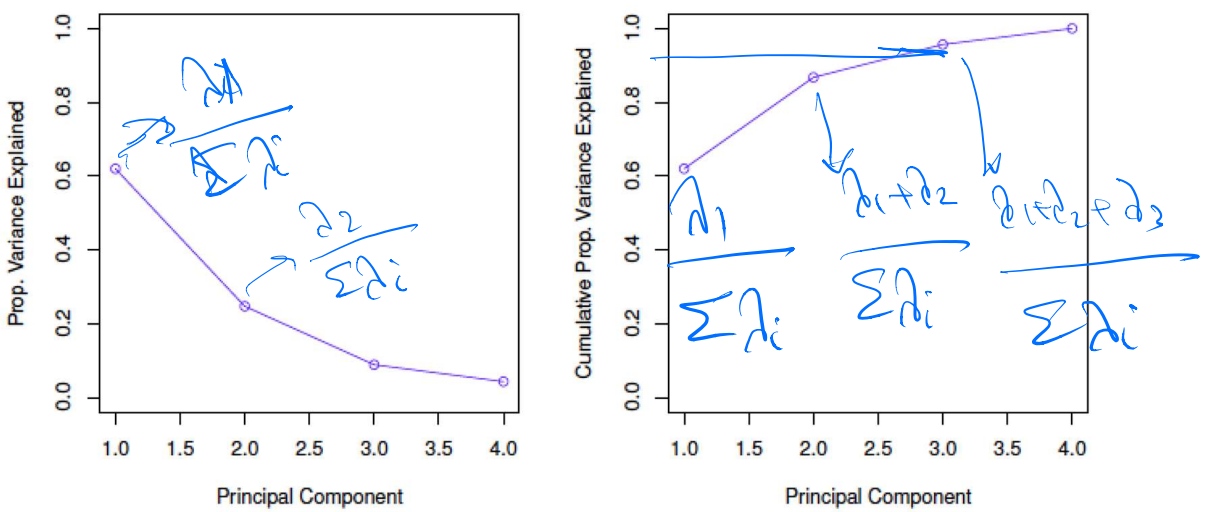
How Many Principal Components Should One Use?

Scree plot



Proportion Variance Explained: continued

- The PVEs sum to one. We sometimes display the cumulative PVEs.



PCA and SVD

- Recall that *any real matrix* $\mathbf{X}_{n\times p}$ can be decomposed into the following form:

$$\mathbf{U}\Sigma\mathbf{V}^T$$

which is called the Singular Value Decomposition (SVD) of **X**.

$\mathbf{U}_{n\times n}$ and $\mathbf{V}_{p\times p}$ are orthogonal matrices.

PCA and SVD

- The matrix $\mathbf{\Sigma}$ is of the form

where $\sigma_1, \ldots, \sigma_r$ are called the singular values of $\mathbf{X}$ and $r = \text{rank}(X)$

Obviously, $r \leq \min(n, p)$

PCA and SVD

- Defining the matrix **Σ'** as:

$$\Sigma' = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$$

We can rewrite the SVD in its *compact form*

$$V = [V_1 \; V_2]$$

$$[U_1 \; U_2] \begin{bmatrix} \Sigma' & 0 \\ 0 & 0 \end{bmatrix} V^T$$

$$[U_1 \Sigma' \; 0] \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 \Sigma' V_1^T$$

PCA and SVD

- On the other hand, we know the principal components are the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Plugging in $\mathbf{X} = \mathbf{U\Sigma V}^T$ and $\mathbf{X} = \mathbf{U}_1\mathbf{\Sigma'V}_1^T$:

$$X^TX = (U\Sigma V^T)^T(U\Sigma V^T)$$

$$= V\Sigma^T U^T U \Sigma V^T = V\Sigma^T\Sigma V^T$$

$$U^TU = I$$

$$= V \begin{bmatrix} \Sigma'^2 & 0 \\ 0 & 0 \end{bmatrix} V^T$$

$$= V_1 \Sigma'^2 V_1^T$$

λ_i's are eigenvalues of $X^T X$

but we showed that they are

also σ_2^2.

PCA and SVD

- Meaning that right singular vectors **V** are principal directions and that squared singular values are proportional to the variance of data along each principal component:

Columns of V are the eigenvectors of $X^T X$, which means they are principal directions.

Projections of $x^{(i)}$'s on the principal directions are principal components

$$XV = U\Sigma V^T V = U\Sigma$$

($n \times n$, $n \times p$, $p \times r$, $n \times r$)

$$\Rightarrow XV_1 = U_1 \Sigma' V_1^T V_1 = U_1 \Sigma'$$

($n \times p$, $p \times r$, $n \times r$, $r \times r$, $r \times p$, $p \times r$, $r \times r$, $n \times r$)

$V_{r \times p}$

Principal components

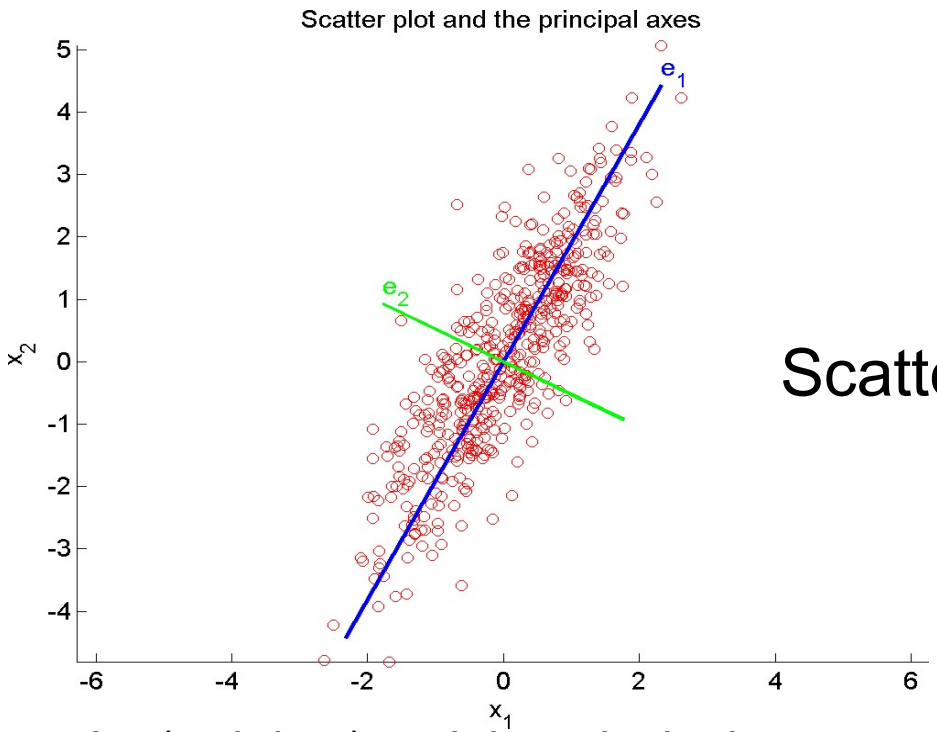
PCA and SVD

- The projections of data points onto principal directions, i.e. principal components/scores are obtained from $\mathbf{XV}_1$ (or $\mathbf{XV}$):

PCA and SVD

- Note that the number of meaningful principal components (with nonzero λ_i) depends on the rank of the matrix **X**.
- One can use the full SVD and find *p* directions, but if the rank of the data matrix is less than *p* [r ≤ min(*p*,*n*)] then *n-r* principal directions will have a zero eigenvalues. How do you interpret that?

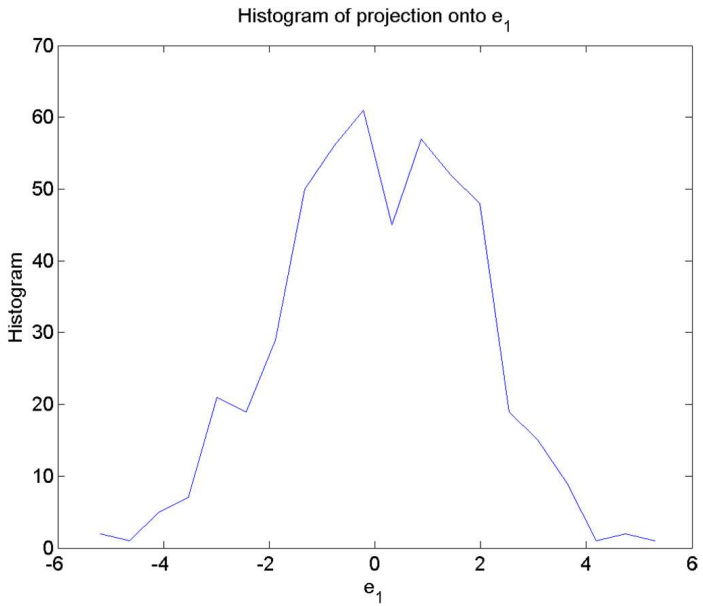
Examples



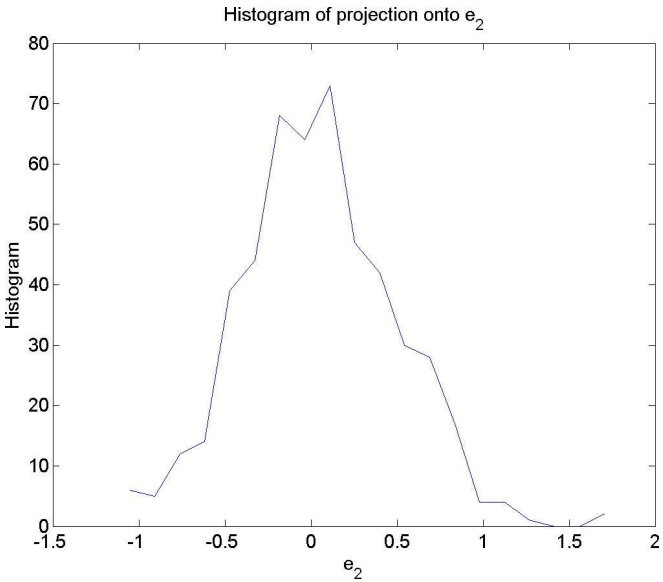
Scatter plot.

Scatter plot (red dots) and the principal axes for a bivariate sample. The blue line shows the axis e_1 with the greatest variance and the green line shows the axis e_2 with the smallest variance. Features are now uncorrelated.

Examples



Projection onto e_1.

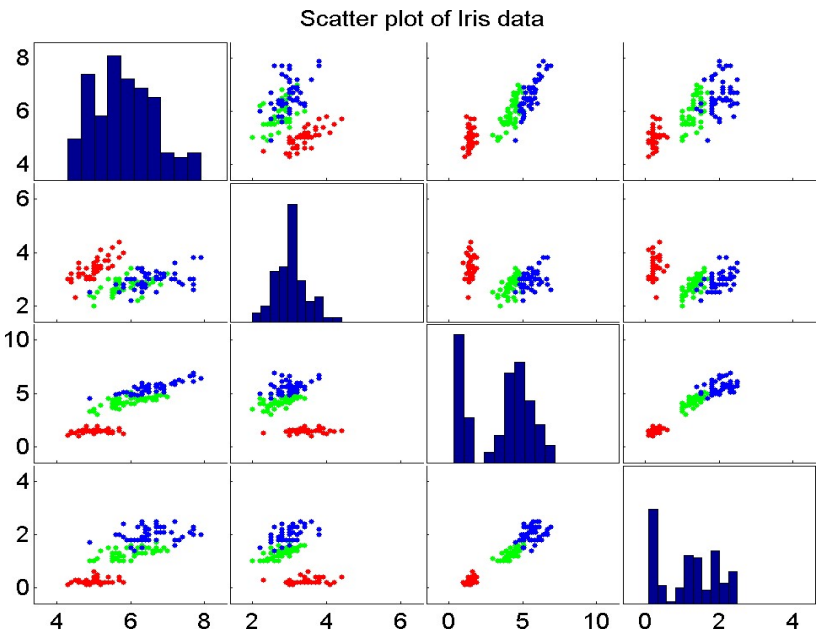


(c) Projection onto e_2.

Examples

4 variables
3 species
50 samples/class

- Setosa
- Versicolor
- Virginica



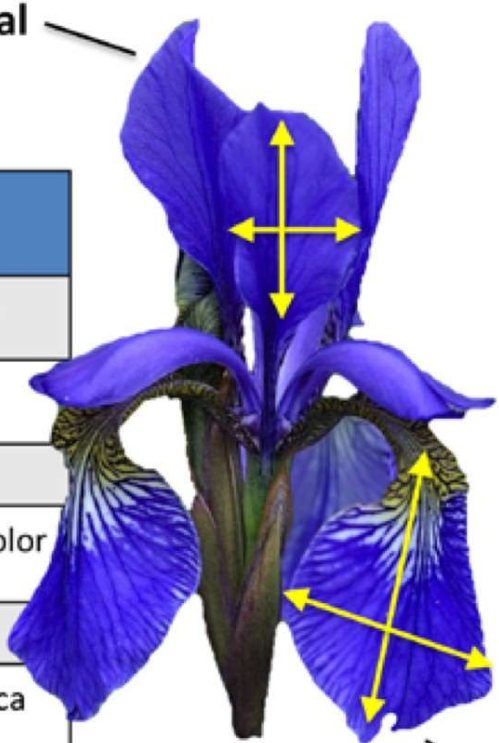
Scatter plot of the iris data. Diagonal cells show the histogram for each feature. Other cells show scatters of pairs of features x_1, x_2, x_3, x_4 in top-down and left-right order. Red, green and blue points represent samples for the setosa, versicolor and virginica classes, respectively.

Samples
(instances, observations)

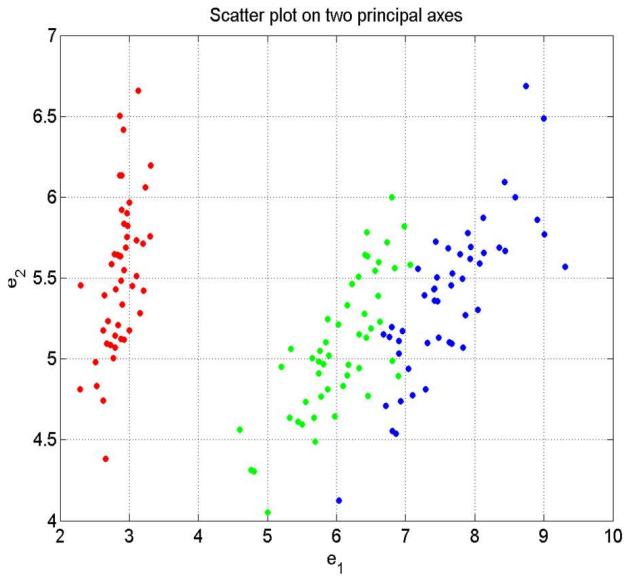
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

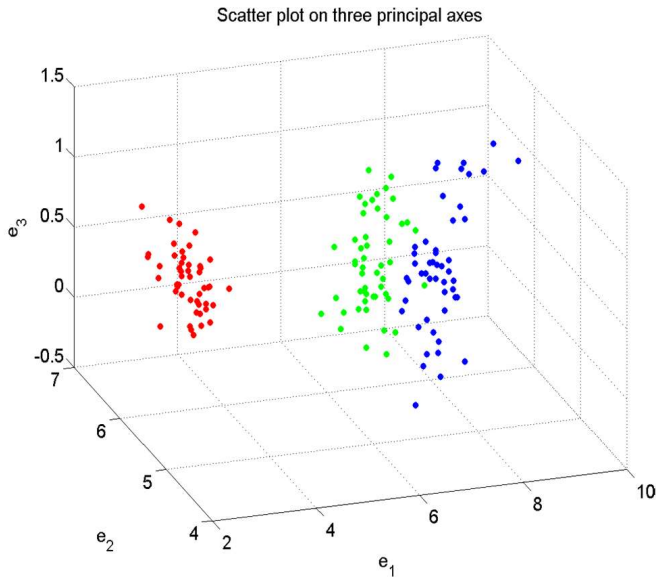
Features
(attributes, measurements, dimensions)

Class labels
(targets)



Examples





Scatter plot of the projection of the iris data onto the first two and the first three principal axes. Red, green and blue points represent samples for the setosa, versicolor and virginica classes, respectively.