San José State University

Math 250: Mathematical Data Visualization

Principal Component Analysis (PCA)

— A First Dimensionality Reduction Approach

Dr. Guangliang Chen

Outline of the presentation:

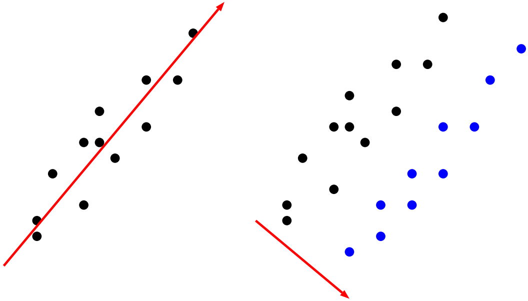
- The 1D maximum-variance projection problem
- The general maximum-variance projection problem
- PCA: procedure, interpretation, numerical issues, and applications
- Dimension reduction by orthogonal best-fit linear subspaces (PCA without centering -> useful on frequency data)

Introduction

- Many data sets have very high dimensions nowadays, which causes a significant challenge in data storage and modeling.
- We need a way to reduce the dimension of the data in order to reduce memory requirement while increasing speed.
- If we discard some dimensions, will that degrade the performance?
- The answer can be no, as long as we do it carefully by **preserving the information that is needed by the task**. In fact, it may even lead to better accuracy in many cases.

An important fact

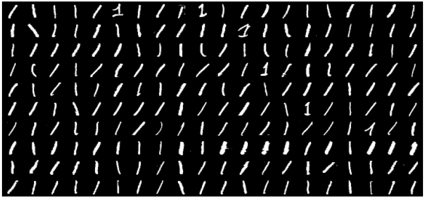
"Useful" information of a high dimensional data set is often contained in only a small number of dimensions.



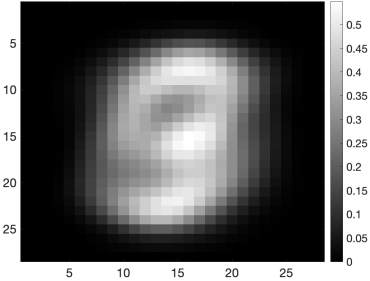
An example: the MNIST handwritten digits

Though in $\mathbb{R}^{784}$, the number of degrees of freedom (i.e., parameters) each digit has is much less than 784.

For example, the images of digit 1 mainly differ in slope and thickness.



The data set also has useless dimensions such as the boundary pixels.



Principal Component Analysis (PCA)

Different dimentionality reduction algorithms preserve different kinds of information.

This course covers the following methods:

- **Principal Component Analysis (PCA)**: variance
- **Multidimensional Scaling (MDS)**: distance
- **Laplacian Eigenmaps**: local affinity
- **Linear Discriminant Analysis (LDA)**: separation among classes

Principal Component Analysis (PCA)

The following dimensionality reduction methods are not covered in this course, but are often covered in a multivariate statistical analysis course:

- PCA and LDA (from a statistical point of view)
- Factor analysis
- Canonical correlation analysis

For example,

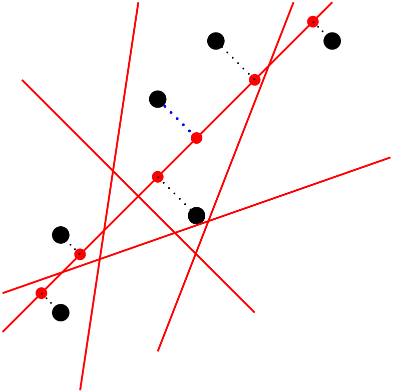
- SJSU – *Math 257: Multivariate Data Analysis*
- PSU – *STAT 505: Applied Multivariate Statistical Analysis*[1]

[1] `https://online.stat.psu.edu/stat505/`

The 1D maximum-variance projection problem

Assume a set of n data points in d dimensions, i.e., $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, where d is large.

We would like to find a line onto which the orthogonal projections of the data would have the largest possible amount of variance.



Principal Component Analysis (PCA)

Mathematical derivation

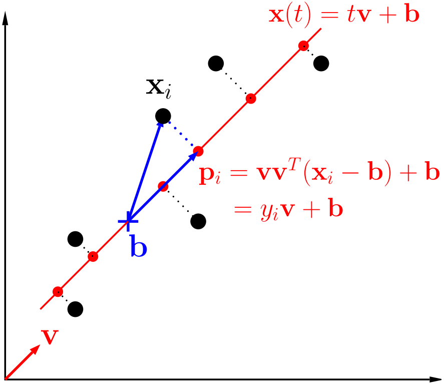
Given data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, find a line $\mathcal{S}$ parametrized by

$$\mathbf{x}(t) = t \cdot \mathbf{v} + \mathbf{b},$$

where $\mathbf{v}, \mathbf{b} \in \mathbb{R}^d$ and $\|\mathbf{v}\| = 1$, such that the 1D orthogonal projections of the data points onto the line

$$y_i = \mathbf{v}^T(\mathbf{x}_i - \mathbf{b}), \qquad 1 \le i \le n$$

have the largest possible variance.


$\mathbf{x}(t) = t\mathbf{v} + \mathbf{b}$
$\mathbf{x}_i$
$\mathbf{p}_i = \mathbf{v}\mathbf{v}^T(\mathbf{x}_i - \mathbf{b}) + \mathbf{b}$
$= y_i\mathbf{v} + \mathbf{b}$
$\mathbf{b}$
$\mathbf{v}$

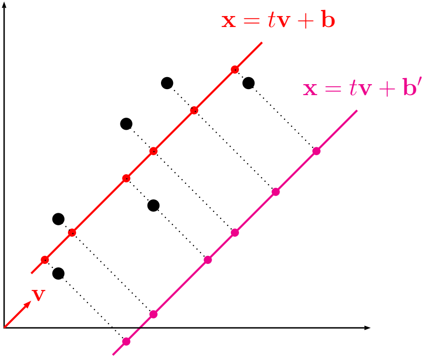
Principal Component Analysis (PCA)

Observe that for parallel lines, i.e.,

$$\mathbf{x}(t) = t\mathbf{v} + \mathbf{b}$$
$$\mathbf{x}(t) = t\mathbf{v} + \mathbf{b}'$$

where $\mathbf{b} \neq \mathbf{b}' \in \mathbb{R}^d$, the orthogonal projections of the data onto them are different, but the amount of variance is the same!

This implies that for the purpose of preserving variance, the choice of $\mathbf{b}$ can be arbitrary (only the vector $\mathbf{v} \in \mathbb{R}^d$ matters).



$\mathbf{x} = t\mathbf{v} + \mathbf{b}$
$\mathbf{x} = t\mathbf{v} + \mathbf{b}'$
$\mathbf{v}$

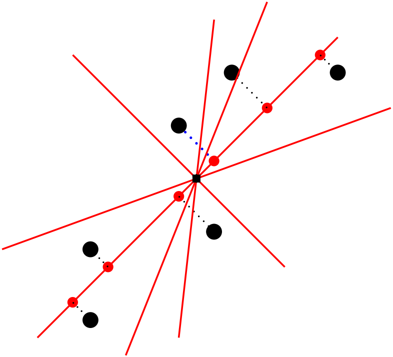
Principal Component Analysis (PCA)

To make the problem tractable, we fix (for all $\mathbf{v}$)

$$\mathbf{b} = \bar{\mathbf{x}} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

so that we only need to compare candidate lines that pass through $\bar{\mathbf{x}}$, the centroid of the data set.

We have thus eliminated the variable $\mathbf{b}$ and only need to focus on the unit-vector variable $\mathbf{v}$ (representing the direction of the line) when trying to maximize the projection variance.



Principal Component Analysis (PCA)

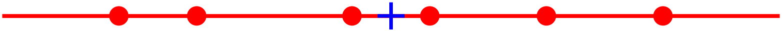
It turns out that with such a fixed choice of $\mathbf{b}$, the projection coefficients of the data onto any direction $\mathbf{v}$

$$y_i = \mathbf{v}^T(\mathbf{x}_i - \bar{\mathbf{x}}), \quad 1 \le i \le n$$

are always automatically centered:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \mathbf{v}^T \cdot \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}}) = \mathbf{v}^T \cdot \mathbf{0} = 0.$$

This will make things like variance very simple to compute.



Since we now have $\bar{y} = 0$, the variance of the projections of the data is simply

$$\frac{1}{n-1}\sum_{i=1}^{n} y_i^2$$

We call $\sum_{i=1}^{n} y_i^2$ the **scatter** of the projections.

We can correspondingly formulate the 1D maximum-variance projection problem as follows:

$$\boxed{\max_{\mathbf{v}:\, \|\mathbf{v}\|=1} \underbrace{\sum_{i=1}^{n} y_i^2}_{\text{scatter}}, \qquad \text{where} \quad y_i = \mathbf{v}^T(\mathbf{x}_i - \bar{\mathbf{x}}).}$$

Principal Component Analysis (PCA)

To solve the problem, we rewrite the objective function as follows:

$$\begin{aligned}\sum_{i=1}^{n} y_i^2 &= \sum \underbrace{\mathbf{v}^T(\mathbf{x}_i-\bar{\mathbf{x}})}_{y_i}\underbrace{(\mathbf{x}_i-\bar{\mathbf{x}})^T\mathbf{v}}_{y_i}\\ &= \sum \mathbf{v}^T\left[(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T\right]\mathbf{v}\\ &= \mathbf{v}^T\underbrace{\left[\sum(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T\right]}_{:=\mathbf{S}\,(d\times d\ \text{matrix})}\mathbf{v}\\ &= \mathbf{v}^T\mathbf{S}\mathbf{v}.\end{aligned}$$

Remark. The matrix $\mathbf{S}$ is called the sample scatter matrix of the data. It is square, symmetric, and positive semidefinite, because it is a sum of such matrices!

Principal Component Analysis (PCA)

Accordingly, we have obtained the following (Rayleigh quotient) problem

$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{S} \mathbf{v}$$

which can be easily solved.

Theorem 0.1. Given a set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^d$ with centroid $\bar{\mathbf{x}} = \frac{1}{n} \sum \mathbf{x}_i$, the optimal direction for projecting the data (in order to have maximum variance) is the largest eigenvector of the sample covariance matrix $\mathbf{S} = \sum (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$:

$$\max_{\mathbf{v}:\,\|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{S} \mathbf{v} = \underbrace{\lambda_1}_{\text{max scatter}}, \qquad \text{achieved when } \mathbf{v} = \pm \mathbf{v}_1.$$

Computing

The theorem requires constructing a $d \times d$ matrix from the given data

$$\mathbf{S} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

which can be a significant challenge for large data sets in high dimensions.

- It takes $\mathcal{O}(d^2)$ memory to store $\mathbf{S}$;
- The time complexity of obtaining $\mathbf{S}$ is $\mathcal{O}(nd^2)$.

We show that the eigenvectors of $\mathbf{S}$ can be efficiently computed from the Singular Value Decomposition (SVD) of the centered data matrix.

Principal Component Analysis (PCA)

Denote the original and centered data matrices (rows are data points) by

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}^T \in \mathbb{R}^{n\times d}$$

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \widetilde{\mathbf{x}}_1 & \cdots & \widetilde{\mathbf{x}}_n \end{bmatrix}^T \in \mathbb{R}^{n\times d}, \qquad \text{where} \quad \widetilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}, \ \forall i$$

Then

$$\mathbf{S} = \sum_{i=1}^n \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^T = [\widetilde{\mathbf{x}}_1 \ldots \widetilde{\mathbf{x}}_n] \cdot \begin{bmatrix} \widetilde{\mathbf{x}}_1^T \\ \vdots \\ \widetilde{\mathbf{x}}_n^T \end{bmatrix} = \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}} \in S_{0+}^d(\mathbb{R}).$$

Principal Component Analysis (PCA)

Thus, the maximum-variance direction $\mathbf{v}_1$ can be computed as the top right singular vector of the centered data matrix $\widetilde{\mathbf{X}}$:

$$\widetilde{\mathbf{X}} \approx \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T \quad \longleftarrow \ \text{rank-1 SVD}$$

The amount of scatter captured by the 1-dimensional projection line,

$$\mathbf{x}(t) = t\mathbf{v}_1 + \bar{\mathbf{x}}, \quad t \in \mathbb{R}$$

is the following (note that it is highest possible):

$$\lambda_1 = \sigma_1^2.$$

Principal Component Analysis (PCA)

We derive a few more useful formulas (in matrix form): Let

$$\mathbf{y} = (y_i)_{1 \le i \le n} \in \mathbb{R}^n, \qquad \text{where} \quad y_i = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{v}_1$$

the projection coefficients of the centered data, and

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_n \end{bmatrix}^T \in \mathbb{R}^{n \times d}, \qquad \text{where} \quad \mathbf{p}_i = y_i \mathbf{v}_1 + \bar{\mathbf{x}}$$

the full coordinates of the projections in the original space ($\mathbb{R}^d$).

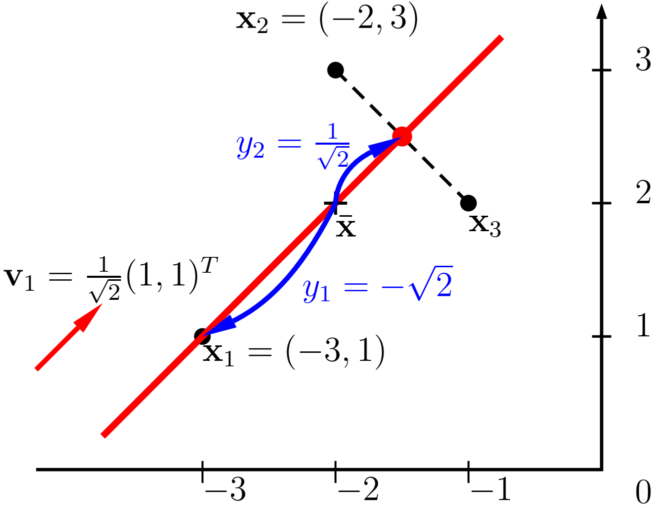
Then in matrix form, we can write

$$\begin{aligned} \mathbf{y} &= \widetilde{\mathbf{X}} \mathbf{v}_1 = \sigma_1 \mathbf{u}_1 \\ \mathbf{P} &= \widetilde{\mathbf{X}} \mathbf{v}_1 \mathbf{v}_1^T + \mathbf{1} \bar{\mathbf{x}}^T = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \mathbf{1} \bar{\mathbf{x}}^T \end{aligned}$$

Example 0.1. Find the maximum-variance direction for the following data set of 3 points in $\mathbb{R}^2$:

$$\mathbf{x}_1 = \begin{pmatrix} -3 \\ 1 \end{pmatrix}, \ \mathbf{x}_2 = \begin{pmatrix} -2 \\ 3 \end{pmatrix}, \ \mathbf{x}_3 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

Principal Component Analysis (PCA)

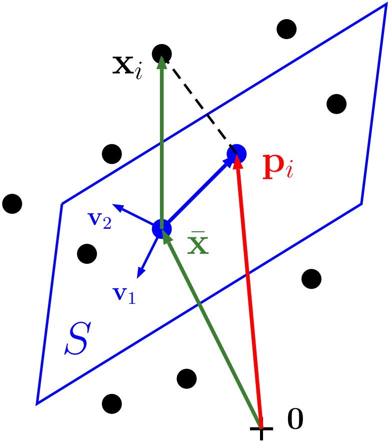


The general maximum-variance projection problem

Given a data set, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, and a positive integer k, we would like to find a k-D plane for orthogonally projecting the data which can preserve the most variance:

$$\mathbf{x}(\boldsymbol{\alpha}) = \mathbf{V}_k \cdot \boldsymbol{\alpha} + \mathbf{b}, \quad \boldsymbol{\alpha} \in \mathbb{R}^k.$$

Here, $\mathbf{V}_k \in \mathbb{R}^{d\times k}$ and $\mathbf{V}_k^T \mathbf{V}_k = \mathbf{I}_k$. For the same reason we fix $\mathbf{b} = \bar{\mathbf{x}}$, and focus on finding the optimal $\mathbf{V}_k$.



Principal Component Analysis (PCA)

For any fixed choice of $\mathbf{V}_k$, the orthogonal projections of the data points onto the plane S are given by

$$\mathbf{p}_i = \mathbf{V}_k \underbrace{\mathbf{V}_k^T(\mathbf{x}_i - \bar{\mathbf{x}})}_{\mathbf{y}_i \in \mathbb{R}^k} + \bar{\mathbf{x}} = \mathbf{V}_k \mathbf{y}_i + \bar{\mathbf{x}}, \qquad 1 \le i \le n$$

Note that the projection coefficients $\{\mathbf{y}_i\}_{i=1}^n$ are also centered:

$$\sum_{i=1}^n \mathbf{y}_i = \mathbf{V}_k^T \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) = \mathbf{V}_k^T \mathbf{0} = \mathbf{0}.$$

As a result, the total scatter of the projected points is

$$\sum_{i=1}^n \|\mathbf{y}_i - \mathbf{0}\|^2 = \sum_{i=1}^n \|\mathbf{y}_i\|^2 = \sum_{i=1}^n \mathbf{y}_i^T \mathbf{y}_i.$$

Principal Component Analysis (PCA)

Our goal is thus to solve the following problem

$$\max_{\mathbf{V}_k^T\mathbf{V}_k=\mathbf{I}} \sum_{i=1}^n \mathbf{y}_i^T\mathbf{y}_i, \qquad \text{where} \quad \mathbf{y}_i = \mathbf{V}_k^T(\mathbf{x}_i - \bar{\mathbf{x}}).$$

We need to rewrite the sum into an explicit expression in $\mathbf{V}_k$:

$$\begin{aligned}\sum_{i=1}^n \mathbf{y}_i^T\mathbf{y}_i &= \sum_{i=1}^n \underbrace{(\mathbf{x}_i-\bar{\mathbf{x}})^T\mathbf{V}_k}_{\mathbf{y}_i^T}\underbrace{\mathbf{V}_k^T(\mathbf{x}_i-\bar{\mathbf{x}})}_{\mathbf{y}_i} \qquad (\mathbf{y}_i^T\mathbf{y}_i = \mathrm{trace}(\mathbf{y}_i\mathbf{y}_i^T))\\ &= \sum_{i=1}^n \mathrm{trace}\left[\mathbf{V}_k^T(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T\mathbf{V}_k\right]\\ &= \mathrm{trace}\left(\mathbf{V}_k^T\left[\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T\right]\mathbf{V}_k\right) = \mathrm{trace}\left(\mathbf{V}_k^T\mathbf{S}\mathbf{V}_k\right).\end{aligned}$$

Accordingly, we have obtained the following trace maximization problem

$$\max_{\mathbf{V}_k^T\mathbf{V}_k=\mathbf{I}} \operatorname{trace}\left(\mathbf{V}_k^T\mathbf{S}\mathbf{V}_k\right)$$

To better understand the problem, write $\mathbf{V}_k = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_k\end{bmatrix}$. Then

- The objective function can be written as

$$\mathbf{v}_1^T\mathbf{S}\mathbf{v}_1 + \cdots + \mathbf{v}_k^T\mathbf{S}\mathbf{v}_k.$$

  It is a sum of the scatter of the projection onto each direction $\mathbf{v}_i$.

- The constraint, $\mathbf{V}_k^T\mathbf{V}_k = \mathbf{I}$, requires unit-norm and orthogonality:

$$\mathbf{v}_i^T\mathbf{v}_j = \begin{cases} 1, & i = j \\ 0 & i \neq j \end{cases}$$

Principal Component Analysis (PCA)

The following result states that the k-dimensional maximum-variance projection plane can be directly found from the spectral decomposition of the sample scatter matrix.

Theorem. Given a set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and a positive integer k, let $\lambda_1 \geq \cdots \geq \lambda_k$ be the largest k eigenvalues of the sample scatter matrix $\mathbf{S}$ with corresponding (unit-norm) eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^d$. Then the maximum-variance projection plane of dimension k is the plane through the centroid $\bar{\mathbf{x}}$ and with orthonormal basis $\mathbf{V}_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$, and the total amount of scatter the projections have is $\lambda_1 + \cdots + \lambda_k$, i.e.,

$$\max_{\mathbf{V}_k^T \mathbf{V}_k = \mathbf{I}} \operatorname{trace}\left(\mathbf{V}_k^T \mathbf{S} \mathbf{V}_k\right) = \lambda_1 + \cdots + \lambda_k, \qquad \text{when } \mathbf{V}_k = [\mathbf{v}_1 \ldots \mathbf{v}_k].$$

Principal Component Analysis (PCA)

Question: How much scatter does the data have in total?

Computing

Similarly, the maximum-variance directions $\mathbf{V}_k = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_k \end{bmatrix}$ can be computed as the top k right singular vectors of $\widetilde{\mathbf{X}}$:

$$\widetilde{\mathbf{X}} \approx \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T \quad \longleftarrow \text{ rank-}k \text{ SVD}$$

The amount of scatter captured by the k-dimensional projection plane,

$$\mathbf{x}(\boldsymbol{\alpha}) = \mathbf{V}_k \cdot \boldsymbol{\alpha} + \bar{\mathbf{x}}, \quad \boldsymbol{\alpha} \in \mathbb{R}^k$$

is the following (note that it is highest possible):

$$\lambda_1 + \cdots + \lambda_k = \sigma_1^2 + \cdots + \sigma_k^2.$$

Principal Component Analysis (PCA)

Let

$$\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \cdots & \mathbf{y}_n\end{bmatrix}^T \in \mathbb{R}^{n\times k}, \qquad \text{where} \quad \mathbf{y}_i = \mathbf{V}_k^T(\mathbf{x}_i - \bar{\mathbf{x}})$$

$$\mathbf{P} = \begin{bmatrix}\mathbf{p}_1 & \cdots & \mathbf{p}_n\end{bmatrix}^T \in \mathbb{R}^{n\times d}, \qquad \text{where} \quad \mathbf{p}_i = \mathbf{V}_k\mathbf{y}_i + \bar{\mathbf{x}}$$

be the matrices of projection coefficients and projection points, respectively.

Then

$$\begin{aligned}\mathbf{Y} &= \widetilde{\mathbf{X}}\mathbf{V}_k = \mathbf{U}_k\mathbf{\Sigma}_k,\\ \mathbf{P} &= \widetilde{\mathbf{X}}\mathbf{V}_k\mathbf{V}_k^T + \mathbf{1}\bar{\mathbf{x}}^T = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T + \mathbf{1}\bar{\mathbf{x}}^T\end{aligned}$$

Principal component analysis (PCA)

Let $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}^T \in \mathbb{R}^{n\times d}$ be a given high dimensional data set.

The process of identifying the maximum-variance directions,

$$\mathbf{V}_k = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_k \end{bmatrix} \in \mathbb{R}^{d\times k},$$

as well as the corresponding projection coefficients,

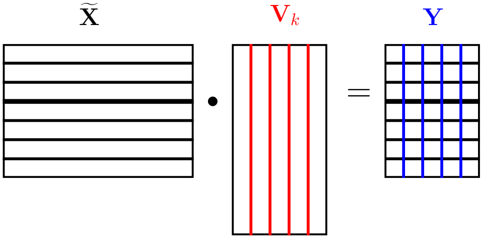
$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_n \end{bmatrix}^T \in \mathbb{R}^{n\times k},$$

is called *principal component analysis (PCA)*.

Principal Component Analysis (PCA)

We say that

- The unit vector $\mathbf{v}_j$, for each $1 \leq j \leq k$, is the j**th principal direction** of the data;
- The projection coefficients $\mathbf{Y} \in \mathbb{R}^{n\times k}$, are **the first** k **principal components** of the data:



Principal Component Analysis (PCA)

- The ith row of $\mathbf{Y}$, i.e.,

  $$\mathbf{y}_i^T = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{V}_k \in \mathbb{R}^k,$$

  is the vector of **the first k principal components of $\mathbf{x}_i$**;

- The jth column of $\mathbf{Y}$, i.e.,

  $$\mathbf{Y}(:, j) = \widetilde{\mathbf{X}} \mathbf{v}_j = \sigma_j \mathbf{u}_j,$$

  is **the jth principal component of the data $\mathbf{X}$**.

Remark. Different principal components of the data must be uncorrelated.

To see this, first note (again) that each principal component of the data (columns of $\mathbf{Y}$) has been centered:

$$\mathbf{1}^T\mathbf{Y} = \mathbf{1}^T\widetilde{\mathbf{X}}\mathbf{V}_k = \mathbf{0}^T\mathbf{V}_k = \mathbf{0}^T$$

Now, the pairwise dot products between the principal components are

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{\Sigma}_k^T\mathbf{U}_k^T\mathbf{U}_k\mathbf{\Sigma}_k = \mathbf{\Sigma}_k^2$$

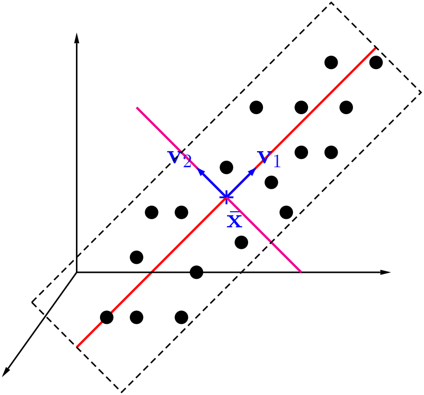
This shows that

- each principal component has scatter σ_j^2, and
- the different components are uncorrelated.

Principal Component Analysis (PCA)

Remark. PCA is a change of coordinate system by using the maximum-variance directions of the data!

- The new origin is set at the centroid of the data set, $\bar{\mathbf{x}}$;
- The new coordinate axes are set along the principal directions of the data, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$;
- The new coefficients are the principal components.



An SVD-based algorithm for PCA

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and integer k (with $0 < k < d$)

Output: Top k principal directions $\mathbf{V}_k = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_k\end{bmatrix}$ and corresponding principal components $\mathbf{Y} \in \mathbb{R}^{n\times k}$.

Steps:

1. Center data: $\widetilde{\mathbf{X}} = [\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}]^T$ where $\bar{\mathbf{x}} = \frac{1}{n}\sum \mathbf{x}_i$
2. Perform rank-k SVD: $\widetilde{\mathbf{X}} \approx \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$
3. Return: $\mathbf{Y} = \widetilde{\mathbf{X}}\mathbf{V}_k = \mathbf{U}_k\mathbf{\Sigma}_k$ (the latter is more efficient to compute because of the diagonal matrix $\mathbf{\Sigma}_k$)

MATLAB implementation of PCA

```
Xtilde = X - mean(X,1);
[U,S,V] = svds(Xtilde, k); % k is the reduced dimension
Y = U .* diag(S)’;
```

Example 0.2. Perform PCA (with $k = 2$), by hand and also in Matlab, on the following data set:

$$\mathbf{x}_1 = \begin{pmatrix} -3 \\ 1 \end{pmatrix}, \ \mathbf{x}_2 = \begin{pmatrix} -2 \\ 3 \end{pmatrix}, \ \mathbf{x}_3 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

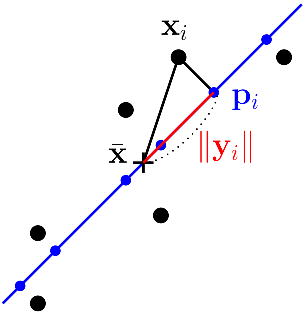
Connection to orthogonal least-squares fitting

It turns out that the following two planes coincide:

(1) **PCA plane**: which maximizes the projection variance,

(2) **Orthogonal best-fit plane**: which minimizes the orthogonal least-squares fitting error.

Mathematical justification:



$$\underbrace{\sum \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2}_{\text{total scatter}} = \underbrace{\sum \|\mathbf{y}_i\|^2}_{\text{proj. var.}} + \underbrace{\sum \|\mathbf{x}_i - \mathbf{p}_i\|^2}_{\text{ortho. fitting error}}$$

Other interpretations of PCA

The PCA plane also tries to preserve, as much as possible, the Euclidean distances between the given data points:

$$\|\mathbf{y}_i - \mathbf{y}_j\|_2 \approx \|\mathbf{x}_i - \mathbf{x}_j\|_2 \quad \text{for all pairs } i \neq j$$

More on this when we get to the MDS part.

PCA can also be regarded as a feature extraction method:

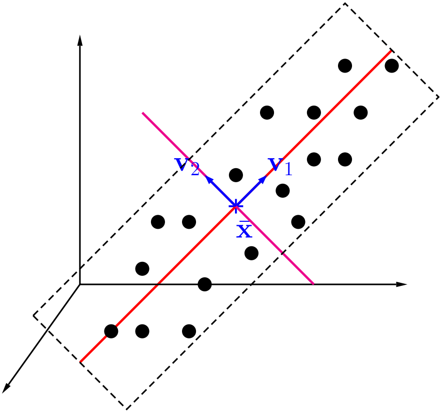
$$\mathbf{v}_j = \frac{1}{\lambda_j}\mathbf{S}\mathbf{v}_j = \frac{1}{\lambda_j}\widetilde{\mathbf{X}}^T(\widetilde{\mathbf{X}}\mathbf{v}_j) \in \mathrm{Col}(\widetilde{\mathbf{X}}^T), \quad \text{for all } j < \mathrm{rank}(\widetilde{\mathbf{X}})$$

This shows that each $\mathbf{v}_j$ is a linear combination of the centered data points (and also a linear combination of the original data points).

Application to data visualization

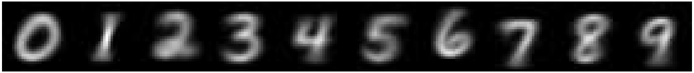
Given a high dimensional data set $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, one can visualize the data by

- projecting the data onto a 2 or 3 dimensional PCA plane,
- and plotting the principal components as new coordinates



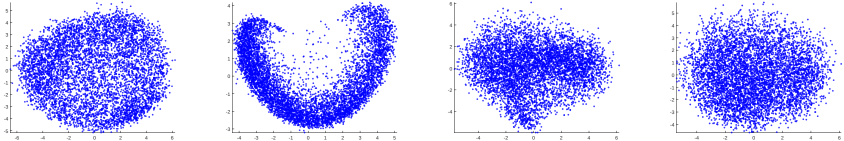
2D visualization of MNIST handwritten digits

1. The "average" writer



2. The full appearance of each digit class

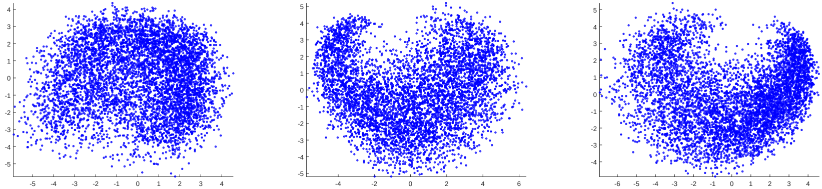
0 - 3



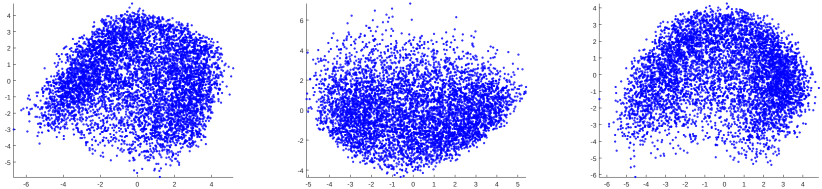
(cont'd on next page)

Principal Component Analysis (PCA)

4-6

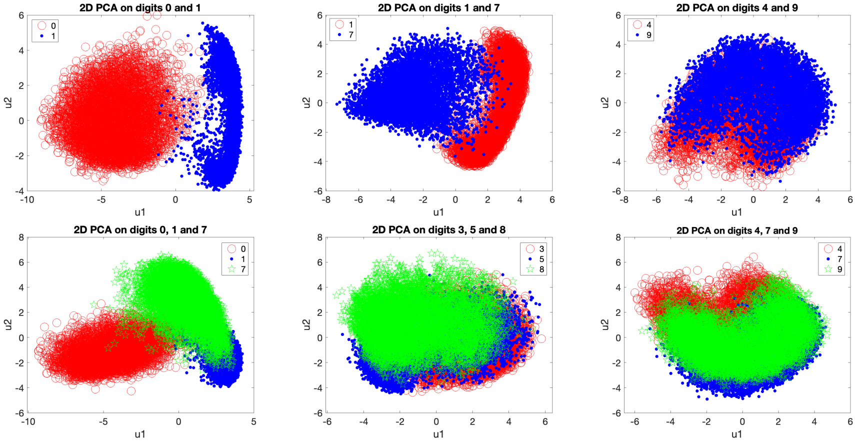


7-9



Principal Component Analysis (PCA)

3. Groups of digits



2D PCA on digits 0 and 1

2D PCA on digits 1 and 7

2D PCA on digits 4 and 9

2D PCA on digits 0, 1 and 7

2D PCA on digits 3, 5 and 8

2D PCA on digits 4, 7 and 9

How to set the parameter k in other settings?

In general, we select the dimension k (as small as possible) such that the top k principal components explain a certain fraction of the total scatter of the data:

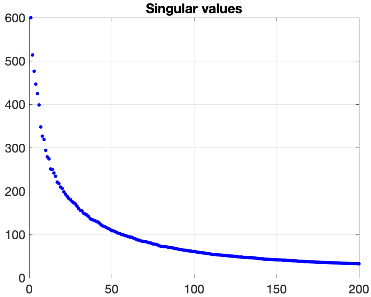
$$\underbrace{\sum_{i=1}^{k} \sigma_i^2}_{\text{explained scatter}} \quad / \quad \underbrace{\sum_{i=1}^{r} \sigma_i^2}_{\text{total scatter}} \quad > \quad p.$$

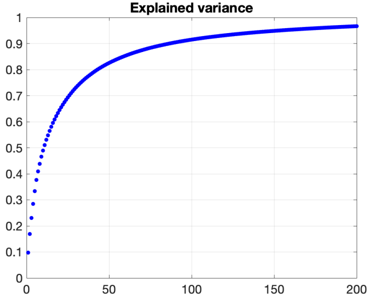
Common values of p are $.95$ (the most commonly used), or $.99$ (more conservative, less reduction), or $.90, .80$ (more aggressive).

However, in practical contexts, it is possible to get much lower than this threshold while maintaining or even improving the accuracy.

Principal Component Analysis (PCA)

Example: MNIST handwritten digits





Principal Component Analysis (PCA)

Note that

$$\sum_{i=1}^{r} \sigma_i^2 = \|\widetilde{\mathbf{X}}\|_F^2,$$

so there is no need to compute all singular values of $\widetilde{\mathbf{X}}$.

Matlab implementation:

$Xtilde = X - mean(X);$
$s = svds(Xtilde, 200);$
$fracs = cumsum(s.\hat{}2)/norm(Xtilde,' fro')\hat{}2;$
$k = find(fracs > 0.95, 1,' first');$
$[U, S] = svds(Xtilde, k)$

Feature scaling

PCA is a variance-preserving projection method. Since variance is determined by distance, PCA is sensitive to the units used by the different features, which can cause them to have arbitrary magnitudes.

A common scaling method is to **standardize each dimension (feature) to have mean zero and standard deviation 1**, so that they are on comparable scales.

In MATLAB, it is implemented in the function 'normalize':

Xnorm = *normalize(X);* % X is an $n \times d$ data matrix.

Out-of-sample extension for PCA

Suppose we have carried out PCA on a given data set (e.g., training data):

- $\widetilde{\mathbf{X}} = [\mathbf{x}_1 \ldots \mathbf{x}_n]^T - [\bar{\mathbf{x}} \ldots \bar{\mathbf{x}}]^T$ where $\bar{\mathbf{x}} = \frac{1}{n}\sum \mathbf{x}_i$
- $\widetilde{\mathbf{X}} \approx \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T$

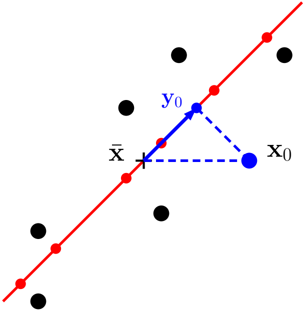
Now there is a new point $\mathbf{x}_0$ (e.g., a test point). How can we extend PCA to $\mathbf{x}_0$?

Two options:

- (naive) Add the new point to the data set and re-run PCA (maybe more accurate, but time-consuming)

- (better) Just use the PCA plane that has already been obtained to project the new point directly:

$$\mathbf{y}_0 = \mathbf{V}_k^T \cdot (\mathbf{x}_0 - \bar{\mathbf{x}})$$



Concluding remarks on PCA

PCA projects the (centered) data onto a k-dim plane that

- maximize the amount of variance in the projection domain,
- minimizes the orthogonal least-squares fitting error

As a dimension reduction and feature extraction method, it is

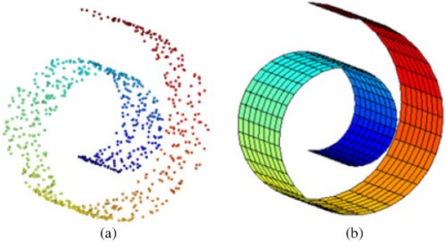
- unsupervised (blind to labels),
- nonparameteric (model-free), and
- very popular!

Principal Component Analysis (PCA)

Lastly, PCA is a linear projection method:

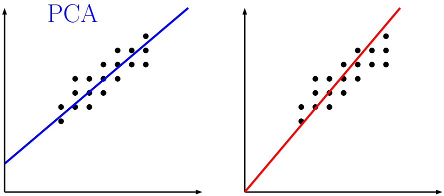
$$\mathbf{y}_0 = \mathbf{V}_k^T(\mathbf{x}_0 - \bar{\mathbf{x}})$$

For nonlinear (manifold) data, PCA will need to use a dimension higher than the manifold dimension (in order to preserve most of the variance).



Dimension reduction via orthogonal best-fit linear subspaces

PCA fits an orthogonal least squares plane to the data through its centroid and reduces the data to a set of principal components. It maximizes the amount of projection variance among all planes of the same dimension.



In contrast, the orthogonal best-fit linear subspace minimizes the orthogonal (squared) fitting error among all linear subspaces of the same dimension, and preserves the most amount of scatter of the data *relative to the origin*.

Principal Component Analysis (PCA)

To see this, first note that the total scatter of the given data *relative to the origin* is

$$\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 = \|\mathbf{X}\|_F^2.$$

For a candidate k-dimensional linear subspace with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d\times k}$, the orthogonal projections of the data have scatter (relative to the origin)

$$\begin{aligned}\sum_{i=1}^{n} \|\mathbf{V}\mathbf{V}^T\mathbf{x}_i\|^2 &= \|\mathbf{V}\mathbf{V}^T\mathbf{X}^T\|_F^2 \\ &= \operatorname{trace}(\mathbf{V}\mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V}\mathbf{V}^T) = \operatorname{trace}(\mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V}),\end{aligned}$$

where we used the property $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

Principal Component Analysis (PCA)

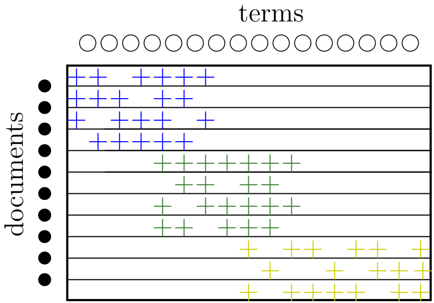
In order to maximize the amount of preserved scatter, we must set $\mathbf{V}$ to the matrix consisting of the top k eigenvectors of $\mathbf{X}^T\mathbf{X}$, which are the top right singular vectors of $\mathbf{X}$.

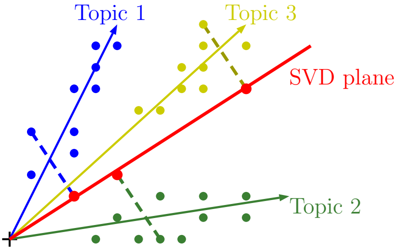
This shows that the k-dimensional orthogonal best-fit linear subspaces preserves the most amount of scatter of the data around the origin and the maximum amount is

$$\sum_{i=1}^{k} \sigma_i^2(\mathbf{X})$$

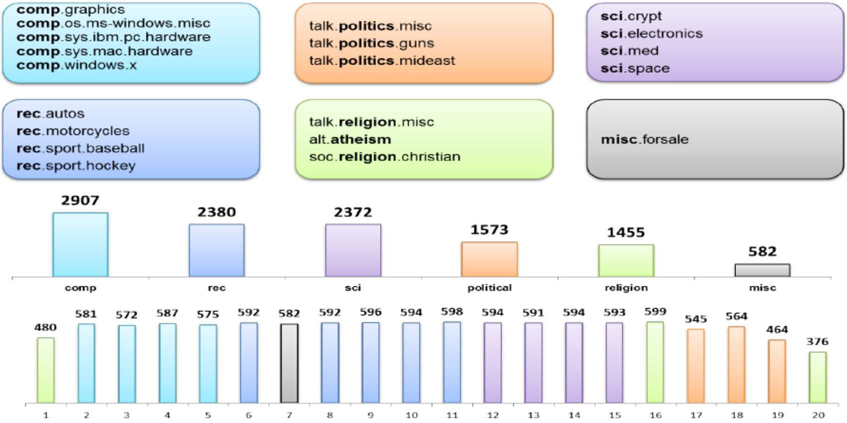
The orthogonal best-fit linear subspace also provides a way to reduce the dimensionality of the data by $\mathbf{Y} = \mathbf{X}\mathbf{V}$. We refer to this method as **uncentered PCA**.

Orthogonal best-fit linear subspaces are useful for modeling frequency data (such as document collection) and reducing their high dimensionality by SVD directly (no centering):





Application: Visualization of 20 newsgroups data



Principal Component Analysis (PCA)

Summary information:

- 18,774 documents partitioned nearly evenly across 20 different newsgroups.
- A total of 61,118 unique words (including stopwords) present in the corpus.

A significant challenge:

- The stopwords dominate in most documents in terms of frequency and make the newsgroups very hard to be .

Principal Component Analysis (PCA)

A fake document-term matrix:

	the	an	zzzz	math	design	car	cars
doc 1	8	12	1	4	2		
doc 2	7	10		3	4		
doc 3	9	15		5	2		
doc 4	5	9			2	2	2
doc 5	9	7			3	3	1
doc 6	1	1				2	

We will not use any text processing software to perform stopword removal (or other kinds of language processing such as stemming), but rather rely on the following statistical operations (in the shown order) on the document-term frequency matrix $\mathbf{X}$ to deal with stopwords:

1. Convert all the frequency counts into binary (0/1) form

	the	an	zzzz	matrix	design	car	cars
doc 1	1	1	1	1	1		
doc 2	1	1		1	1		
doc 3	1	1		1	1		
doc 4	1	1			1	1	1
doc 5	1	1			1	1	1
doc 6	1	1				1	

Principal Component Analysis (PCA)

2. Remove words that occur either in exactly one document (rare words or typos) or in “too many“ documents (stopwords or common words)

	math	design	car	cars
doc 1	1	1		
doc 2	1	1		
doc 3	1	1		
doc 4		1	1	1
doc 5		1	1	1
doc 6			1	
6	3	5	3	1

3. Apply the inverse document frequency (IDF) weighting to the remaining columns of $\mathbf{X}$:

$$\mathbf{X}(:,j) \leftarrow w_j \cdot \mathbf{X}(:,j), \qquad w_j = \log(n/n_j),$$

where n_j is the number of documents that contain the j-th word

	math	design	car	cars
doc 1	0.6931	0.1823		
doc 2	0.6931	0.1823		
doc 3	0.6931	0.1823		
doc 4		0.1823	0.6931	1.0986
doc 5		0.1823	0.6931	1.0986
doc 6			0.6931	

4. Rescale the rows of $\mathbf{X}$ to have unit norm in order to remove the documents' length information

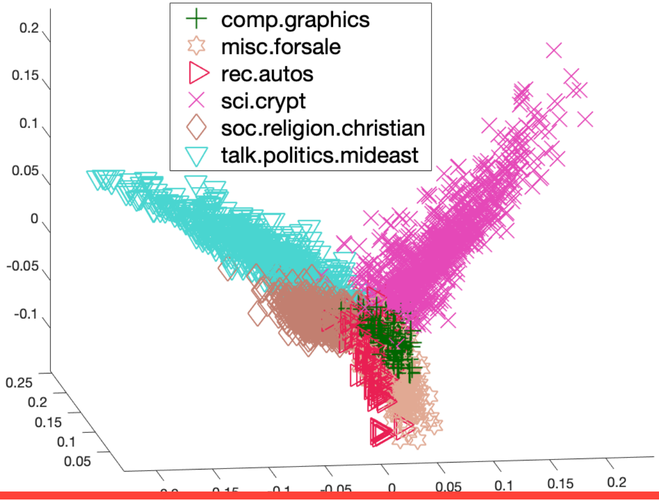
	math	design	car	cars
doc 1	0.9671	0.2544		
doc 2	0.9671	0.2544		
doc 3	0.9671	0.2544		
doc 4		0.1390	0.5284	0.8375
doc 5		0.1390	0.5284	0.8375
doc 6			1	

By applying the above procedure (a particular TF-IDF weighting scheme[2]) to the 20newsgroups data and keeping only the words with frequencies between 2 and 939 (average cluster size), we obtain a matrix of 18,768 nonempty documents and 55,571 unique words, with average row sparsity 73.4.

For ease of demonstration, we focus on six newsgroups in the processed data set (one from each category) and project them by SVD into a 3-dimensional plane through the origin for visualization.

[2] Full name: term frequency inverse document frequency. See `https://en.wikipedia.org/wiki/Tf-idf`

Principal Component Analysis (PCA)



Principal Component Analysis (PCA)

We also display the top 20 words that are the most "relevant" to the underlying topic of each class.

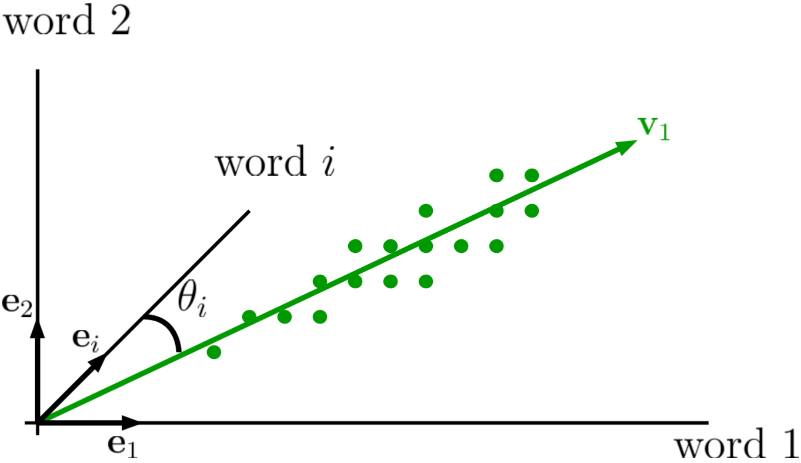
To rank the words based on relevance to each newsgroup, we first compute the top right singular vector $\mathbf{v}_1$ of a fixed newsgroup (without centering), which represents the dominant direction of the cluster.

Each keyword i corresponds to a distinct dimension of the data and is represented by $\mathbf{e}_i$.

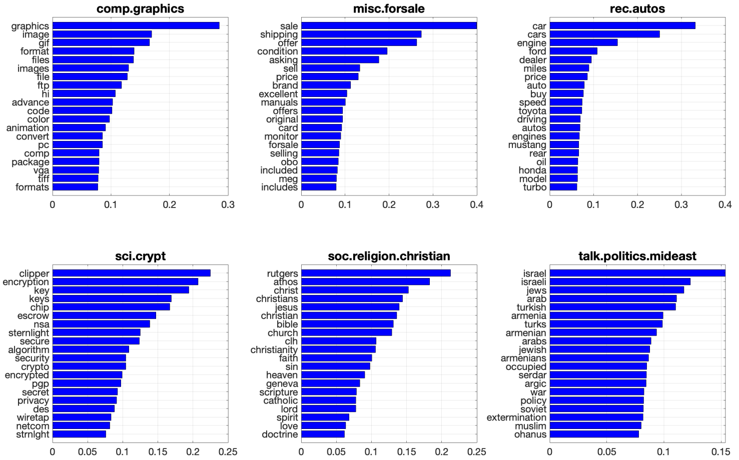
The following score can then be used to measure and compare the relevance of each keyword:

$$\mathrm{score}(i) = \cos\theta_i = \langle \mathbf{v}_1, \mathbf{e}_i \rangle = v_1(i), \quad i = 1, \ldots, 55570$$

Principal Component Analysis (PCA)



Principal Component Analysis (PCA)



Latent Semantic Analysis (LSA)

The preceding process used in reducing the dimension of documents data

- TF-IDF processing
- Dimension reduction by SVD (no centering)
- Cosine similarity

is called latent semantic analysis (see e.g., the Wikipedia page on it).

It is a technique in natural language processing of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.