Principal component analysis

Course of Machine Learning
Master Degree in Computer Science

University of Rome "Tor Vergata"

a.a. 2019-2020

Giorgio Gambosi

Curse of dimensionality

In general, many features: high-dimensional spaces.

- sparseness of data
- increase in the number of coefficients, for example for dimension D and order 3 of the polynomial,

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

  number of coefficients is $O(D^M)$

High dimensions lead to difficulties in machine learning algorithms (lower reliability or need of large number of coefficients) this is denoted as curse of dimensionality

Dimensionality reduction

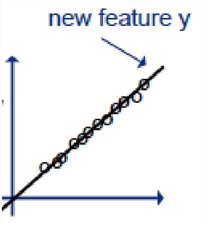
- for any given classifier, the training set size required to obtain a certain accuracy grows exponentially wrt the number of features
- it is important to bound the number of features, identifying the less discriminant ones

Dimensionality reduction

- Feature selection: identify a subset of features which are still discriminant, or, in general, still represent most dataset variance
- Feature extraction: identify a projection of the dataset onto a lower-dimensional space, in such a way to still represent most dataset variance
  - Linear projection: principal component analysis, probabilistic PCA, factor analysis
  - Non linear projection: manifold learning, autoencoders

Searching hyperplanes for the dataset

- verifying whether training set elements lie on a hyperplane (a space of lower dimensionality), apart from a limited variability (which could be seen as noise)



- principal component analysis looks for a d'-dimensional subspace ($d' < d$) such that the projection of elements onto such suspace is a “faithful” representation of the original dataset
- as “faithful” representation we mean that distances between elements and their projections are small, even minimal

PCA for $d' = 0$

- Objective: represent all d-dimensional vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ by means of a unique vector $\mathbf{x}_0$, in the most faithful way, that is so that

$$J(\mathbf{x}_0) = \sum_{i=1}^{n} ||\mathbf{x}_0 - \mathbf{x}_i||^2$$

  is minimum
- it is easy to show that

$$\mathbf{x}_0 = \mathbf{m} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

PCA for $d' = 0$

- In fact,

$$
\begin{aligned}
J(\mathbf{x}_0) &= \sum_{i=1}^{n} ||(\mathbf{x}_0 - \mathbf{m}) - (\mathbf{x}_i - \mathbf{m})||^2 \\
&= \sum_{i=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 - 2\sum_{i=1}^{n} (\mathbf{x}_0 - \mathbf{m})^T (\mathbf{x}_i - \mathbf{m}) + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 \\
&= \sum_{i=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 - 2(\mathbf{x}_0 - \mathbf{m})^T \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 \\
&= \sum_{i=1}^{n} ||\mathbf{x}_0 - \mathbf{m}||^2 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2
\end{aligned}
$$

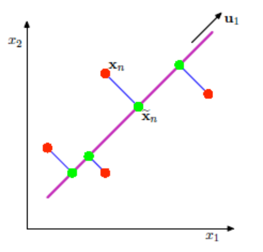
- since

$$
\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) = \sum_{i=1}^{n} \mathbf{x}_i - n \cdot \mathbf{m} = n \cdot \mathbf{m} - n \cdot \mathbf{m} = 0
$$

- the second term is independent from $\mathbf{x}_0$, while the first one is equal to zero for $\mathbf{x}_0 = \mathbf{m}$

PCA for $d' = 1$

- a single vector is too concise a representation of the dataset: anything related to data variability gets lost
- a more interesting case is the one when vectors are projected onto a line passing through $\mathbf{m}$



PCA for $d' = 1$

- let $\mathbf{u}_1$ be unit vector ($||\mathbf{u}_1|| = 1$) in the line direction: the line equation is then

$$\mathbf{x} = \alpha \mathbf{u}_1 + \mathbf{m}$$

  where α is the distance of $\mathbf{x}$ from $\mathbf{m}$ along the line
- let $\tilde{\mathbf{x}}_i = \alpha_i \mathbf{u}_1 + \mathbf{m}$ be the projection of $\mathbf{x}_i$ ($i = 1, \ldots, n$) onto the line: given $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we wish to find the set of projections minimizing the quadratic error

PCA for $d' = 1$

The quadratic error is defined as

$$
\begin{aligned}
J(\alpha_1, \ldots, \alpha_n, \mathbf{u}_1) &= \sum_{i=1}^{n} ||\tilde{\mathbf{x}}_i - \mathbf{x}_i||^2 \\
&= \sum_{i=1}^{n} ||(\mathbf{m} + \alpha_i \mathbf{u}_1) - \mathbf{x}_i||^2 \\
&= \sum_{i=1}^{n} ||\alpha_i \mathbf{u}_1 - (\mathbf{x}_i - \mathbf{m})||^2 \\
&= \sum_{i=1}^{n} +\alpha_i^2 ||\mathbf{u}_1||^2 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 - 2 \sum_{i=1}^{n} \alpha_i \mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m}) \\
&= \sum_{i=1}^{n} \alpha_i^2 + \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{m}||^2 - 2 \sum_{i=1}^{n} \alpha_i \mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m})
\end{aligned}
$$

PCA for $d' = 1$

Its derivative wrt α_k is

$$\frac{\partial}{\partial \alpha_k} J(\alpha_1, \ldots, \alpha_n, \mathbf{u}_1) = 2\alpha_k - 2\mathbf{u}_1^T(\mathbf{x}_k - \mathbf{m})$$

which is zero when $\alpha_k = \mathbf{u}_1^T(\mathbf{x}_k - \mathbf{m})$ (the orthogonal projection of $\mathbf{x}_k$ onto the line).

The second derivative turns out to be positive

$$\frac{\partial}{\partial \alpha_k^2} J(\alpha_1, \ldots, \alpha_n, \mathbf{u}_1) = 2$$

showing that what we have found is indeed a minimum.

PCA for $d' = 1$

To derive the best direction $\mathbf{u}_1$ of the line, we consider the covariance matrix of the dataset

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

By plugging the values computed for α_i into the definition of $J(\alpha_1, \ldots, \alpha_n, \mathbf{u}_1)$, we get

$$\begin{aligned}
J(\mathbf{u}_1) &= \sum_{i=1}^{n}\alpha_i^2 + \sum_{i=1}^{n}||\mathbf{x}_i - \mathbf{m}||^2 - 2\sum_{i=1}^{n}\alpha_i^2 \\
&= -\sum_{i=1}^{n}[\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})]^2 + \sum_{i=1}^{n}||\mathbf{x}_i - \mathbf{m}||^2 \\
&= -\sum_{i=1}^{n}\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T\mathbf{u}_1 + \sum_{i=1}^{n}||\mathbf{x}_i - \mathbf{m}||^2 \\
&= -n\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 + \sum_{i=1}^{n}||\mathbf{x}_i - \mathbf{m}||^2
\end{aligned}$$

PCA for $d' = 1$

- $\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})$ is the projection of $\mathbf{x}_i$ onto the line
- the product
$$\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T\mathbf{u}_1$$
is then the variance of the projection of $\mathbf{x}_i$ wrt the mean $\mathbf{m}$
- the sum
$$\sum_{i=1}^n \mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T\mathbf{u}_1 = n\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$$
is the overall variance of the projections of vectors $\mathbf{x}_i$ wrt the mean $\mathbf{m}$

PCA for $d' = 1$

Minimizing $J(\mathbf{u}_1)$ is equivalent to maximizing $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$. That is, $J(\mathbf{u}_1)$ is minimum if $\mathbf{u}_1$ is the direction which keeps the maximum amount of variance in the dataset

Hence, we wish to maximize $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$ (wrt $\mathbf{u}_1$), with the constraint $||\mathbf{u}_1|| = 1$.

By applying Lagrange multipliers this results equivalent to maximizing

$$u = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 - \lambda_1(\mathbf{u}_1^T\mathbf{u}_1 - 1)$$

This can be done by setting the first derivative wrt $\mathbf{u}_1$:

$$\frac{\partial u}{\partial \mathbf{u}_1} = 2\mathbf{S}\mathbf{u}_1 - 2\lambda_1\mathbf{u}_1$$

to 0, obtaining

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

PCA for $d' = 1$

Note that:

- u is maximized if $\mathbf{u}_1$ is an eigenvector of $\mathbf{S}$
- the overall variance of the projections is then equal to the corresponding eigenvalue

$$\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 = \mathbf{u}_1^T\lambda_1\mathbf{u}_1 = \lambda_1\mathbf{u}_1^T\mathbf{u}_1 = \lambda_1$$

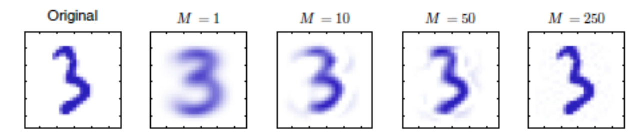
- the variance of the projections is then maximized (and the error minimized) if $\mathbf{u}_1$ is the eigenvector of $\mathbf{S}$ corresponding to the maximum eigenvalue λ_1

PCA for $d' > 1$

- The quadratic error is minimized by projecting vectors onto a hyperplane defined by the directions associated to the d' eigenvectors corresponding to the d' largest eigenvalues of $\mathbf{S}$
- If we assume data are modeled by a d-dimensional gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, PCA returns a d'-dimensional subspace corresponding to the hyperplane defined by the eigenvectors associated to the d' largest eigenvalues of $\boldsymbol{\Sigma}$
- The projections of vectors onto that hyperplane are distributed as a d'-dimensional distribution which keeps the maximum possible amount of data variability

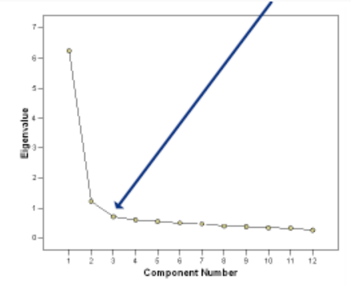
An example of PCA

- Digit recognition ($D = 28 \times 28 = 784$)



Choosing d'

Eigenvalue size distribution is usually characterized by a fast initial decrease followed by a small decrease



This makes it possible to identify the number of eigenvalues to keep, and thus the dimensionality of the projections.

Choosing d'

Eigenvalues measure the amount of distribution variance kept in the projection.

Let us consider, for each $k < d$, the value

$$r_k = \frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}$$

which provides a measure of the variance fraction associated to the k largest eigenvalues.

When $r_1 < \ldots < r_d$ are known, a certain amount p of variance can be kept by setting

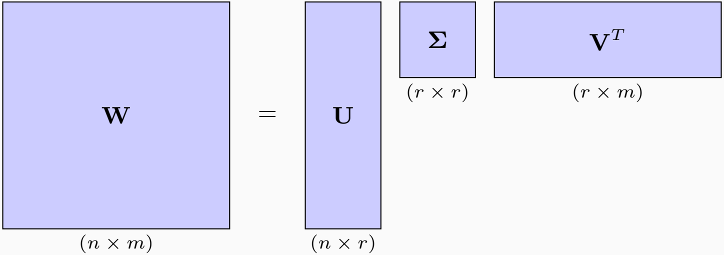
$$d' = \underset{i \in \{1,\ldots,d\}}{\text{argmin}} \; r_i > p$$

Singular Value Decomposition

Let $\mathbf{W} \in \mathbb{R}^{n \times m}$ be a matrix of rank $r \leq \min(n, m)$, and let $n > m$. Then, there exist

- $\mathbf{U} \in \mathbb{R}^{n \times r}$ orthonormal (that is, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r$)
- $\mathbf{V} \in \mathbb{R}^{m \times r}$ orthonormal (that is, $\mathbf{V}\mathbf{V}^T = \mathbf{I}_r$)
- $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ diagonal

such that $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$



SVD in greater detail

Let us consider the matrix $\mathbf{A} = \mathbf{W}^T\mathbf{W} \in \mathbb{R}^{m\times m}$. Observe that

- by definition, $\mathbf{A}$ has the same rank of $\mathbf{W}$, that is r
- $\mathbf{A}$ is symmetric: in fact, $a_{ij} = \mathbf{w}_i^T\mathbf{w}_j$ by definition, where $\mathbf{w}_k$ is the k-th column of $\mathbf{W}$; by the commutativity of vector product, $a_{ij} = \mathbf{w}_i^T\mathbf{w}_j = \mathbf{w}_j^T\mathbf{w}_i = a_{ji}$
- $\mathbf{A}$ is semidefinite positive, that is $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all non null $\mathbf{x} \in \mathbb{R}^m$: this derives from
$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T(\mathbf{W}^T\mathbf{W})\mathbf{x} = (\mathbf{W}\mathbf{x})^T(\mathbf{W}\mathbf{x}) = ||\mathbf{W}\mathbf{x}||_2 \geq 0$$

SVD in greater detail

All eigenvalues of $\mathbf{A}$ are real. In fact,

- let $\lambda \in \mathbb{C}$ be an eigenvalue of $\mathbf{A}$, and let $\mathbf{v} \in \mathbb{C}^n$ be a corresponding eigenvector: then, $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ and $\overline{\mathbf{v}}^T\mathbf{A}\mathbf{v} = \overline{\mathbf{v}}^T\lambda\mathbf{v} = \lambda\overline{\mathbf{v}}^T\mathbf{v}$
- observe that, in general, it must also be that the complex conjugates $\overline{\lambda}$ and $\overline{\mathbf{v}}$ are themselves an eigenvalue-eigenvector pair for $\mathbf{A}$: then, $\mathbf{A}\overline{\mathbf{v}} = \overline{\lambda}\overline{\mathbf{v}}$. Since $\overline{\lambda}\overline{\mathbf{v}}^T = (\overline{\lambda}\overline{\mathbf{v}})^T = (\mathbf{A}\overline{\mathbf{v}})^T = \overline{\mathbf{v}}^T\mathbf{A}^T = \overline{\mathbf{v}}^T\mathbf{A}$ by the simmetry of $\mathbf{A}$, it derives $\overline{\mathbf{v}}^T\mathbf{A}\mathbf{v} = \overline{\lambda}\overline{\mathbf{v}}^T\mathbf{v}$
- as a consequence, $\overline{\lambda}\overline{\mathbf{v}}^T\mathbf{v} = \lambda\overline{\mathbf{v}}^T\mathbf{v}$, that is $\overline{\lambda}||\mathbf{v}||^2 = \lambda||\mathbf{v}||^2$
- since $\mathbf{v} \neq \mathbf{0}$ (being an eigenvector), it must be $\overline{\lambda} = \lambda$, hence $\lambda \in \mathbb{R}$

SVD in greater detail

The eigenvectors of $\mathbf{A}$ corresponding to different eigenvalues are orthogonal

- Let $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{C}^n$ be two eigenvectors, with corresponding distinct eigenvalues λ_1, λ_2
- then, by the simmetry of $\mathbf{A}$, $\lambda_1(\mathbf{v}_1^T\mathbf{v}_2) = (\lambda_1\mathbf{v}_1)^T\mathbf{v}_2 = (\mathbf{A}\mathbf{v}_1)^T\mathbf{v}_2 = \mathbf{v}_1^T\mathbf{A}^T\mathbf{v}2 = \mathbf{v}_1^T\mathbf{A}\mathbf{v}_2 = \mathbf{v}_1^T\lambda_2\mathbf{v}_2 = \lambda_2(\mathbf{v}_1^T\mathbf{v}_2)$
- as a consequence, $(\lambda_1 - \lambda_2)\mathbf{v}_1^T\mathbf{v}_2 = 0$
- since $\lambda_1 \neq \lambda_2$, it must be $\mathbf{v}_1^T\mathbf{v}_2 = 0$, that is $\mathbf{v}_1, \mathbf{v}_2$ must be orthogonal

If an eigenvalue λ' has multiplicity $m > 1$, it is always possible to find a set of m orthonormal eigenvectors of λ'.

As a result, there exists a set of eigenvectors of $\mathbf{A}$ which provides an orthonormal base.

SVD in greater detail

All eigenvalues of a $\mathbf{A}$ are greater than zero.

- $\mathbf{A}$ is real and symmetric, then for each eigenvalue λ it must be $\lambda \in \mathbb{R}$ and there must exist an eigenvector $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$
- as a consequence, $\mathbf{v}^T(\mathbf{A}\mathbf{v}) = \lambda\mathbf{v}^T\mathbf{v}$ and

$$\lambda = \frac{\mathbf{v}^T\mathbf{A}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \frac{\mathbf{v}^T\mathbf{A}\mathbf{v}}{||\mathbf{v}||^2}$$

- $||\mathbf{v}||^2 > 0$ since $\mathbf{v}$ is an eigenvector and, since $\mathbf{A}$ is semidefinite positive, $\mathbf{v}^T\mathbf{A}\mathbf{v} \geq 0$
- as a consequence, $\lambda \geq 0$

SVD in greater detail

Overall,

- $\mathbf{A} = \mathbf{W}^T\mathbf{W}$ has r real and positive eigenvalues $\lambda_1, \ldots, \lambda_r$
- the corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are orthonormal
- $\mathbf{A}\mathbf{v}_i = (\mathbf{W}^T\mathbf{W})\mathbf{v}_i = \lambda_i\mathbf{v}_i$, $i = 1, \ldots, r$

Let us define r singular values

$$\sigma_i = \sqrt{\lambda_i} \qquad i = 1, \ldots, r$$

and let us also consider the set of vectors

$$\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{W}\mathbf{v}_i \qquad i = 1, \ldots, r$$

SVD in greater detail

- Observe that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are orthogonal, in fact:

$$\begin{aligned}\mathbf{u}_i^T\mathbf{u}_j &= \left(\frac{1}{\sigma_i}\mathbf{W}\mathbf{v}_i\right)^T\left(\frac{1}{\sigma_j}\mathbf{W}\mathbf{v}_j\right)\\ &= \frac{1}{\sigma_i\sigma_j}\mathbf{v}_i^T\mathbf{W}^T\mathbf{W}\mathbf{v}_j = \frac{1}{\sigma_i\sigma_j}\mathbf{v}_i^T(\lambda_j\mathbf{v}_j) = \frac{\sigma_j}{\sigma_i}\mathbf{v}_i^T\mathbf{v}_j\end{aligned}$$

  Hence, $\mathbf{u}_i^T\mathbf{u}_j \neq 0$ iff $\mathbf{v}_i^T\mathbf{v}_j \neq 0$, that is iff $i \neq j$.

- Moreover, $\mathbf{u}_1, \ldots, \mathbf{u}_r$ have unitary norm, in fact:

$$\begin{aligned}||\mathbf{u}_i||^2 &= \left|\left|\frac{1}{\sigma_i}\mathbf{W}\mathbf{v}_i\right|\right|^2 = \frac{1}{\lambda_i}(\mathbf{W}\mathbf{v}_i)^T(\mathbf{W}\mathbf{v}_i) = \frac{1}{\lambda_i}\mathbf{v}_i^T(\mathbf{W}^T\mathbf{W}\mathbf{v}_i)\\ &= \frac{1}{\lambda_i}\mathbf{v}_i^T(\lambda_i\mathbf{v}_i) = \frac{1}{\lambda_i}\lambda_i(\mathbf{v}_i^T\mathbf{v}_i) = 1\end{aligned}$$

SVD in greater detail

Let us also consider the following matrices

- $\mathbf{V} \in \mathbb{R}^{m\times r}$ having vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ as columns

$$\mathbf{V} = \left[\begin{array}{cccc} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_r \\ | & | & & | \end{array} \right]$$

- $\mathbf{U} \in \mathbb{R}^{n\times r}$ having vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ as columns

$$\mathbf{U} = \left[\begin{array}{cccc} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \\ | & | & & | \end{array} \right]$$

- $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ having singular values on the diagonal

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}$$

SVD in greater detail

It is easy to verify that

$$\mathbf{W}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$$

Moreover, since $\mathbf{V}$ is orthogonal, its is $\mathbf{V}^{-1} = \mathbf{V}^T$ and, as a consequence,

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

$$\mathbf{W} = \begin{bmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \\ | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} - & \mathbf{v}_1 & - \\ - & \mathbf{v}_2 & - \\ & \vdots & \\ - & \mathbf{v}_r & - \end{bmatrix}$$

PCA and SVD

- Given

$$\mathbf{X} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{bmatrix}$$

- the mean of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$\mathbf{m} = \frac{1}{n} \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{n}\mathbf{X}\mathbf{1}$$

- let $\tilde{\mathbf{X}}$ be the set of such vectors translated to have zero mean:

$$\begin{aligned} \tilde{\mathbf{X}} &= \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{bmatrix} - \begin{bmatrix} | & | & & | \\ \mathbf{m} & \mathbf{m} & \cdots & \mathbf{m} \\ | & | & & | \end{bmatrix} = \mathbf{X} - \mathbf{m}\mathbf{1}^T \\ &= \mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}\mathbf{1}^T = \mathbf{X}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) \end{aligned}$$

PCA and SVD

The correlation matrix of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is defined as:

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

where $\tilde{\mathbf{x}}_i$ is the i-th column of $\tilde{\mathbf{X}}$.
That is,

$$\mathbf{S} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$$

$\tilde{\mathbf{X}}$ has dimension $n \times d$: assuming $n > d$, we may consider its SVD

$$\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

where $\mathbf{U}\mathbf{U}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{\Sigma}$ is a diagonal matrix.

PCA and SVD

By the properties of SVD, items on the diagonal of $\mathbf{\Sigma}$ are the eigenvalues of $\mathbf{S}$ and columns of $\mathbf{V}$ are the corresponding eigenvectors.

In summary:

- To perform a PCA on $\mathbf{X}$, it is sufficient to compute the SVD of matrix

$$\tilde{\mathbf{X}} = \mathbf{X}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

- The principal components of $\mathbf{X}$ are the columns of $\mathbf{V}$, with corresponding eigenvalues given by the diagonal elements of $\mathbf{\Sigma}^2$.

Co-occurence data

Definition

- Two collections $\mathbf{V}, \mathbf{D}$ (for example, terms and documents, or customers and items)
- sequence of observations $\mathbf{W} = \{(w_1, d_1), \ldots, (w_N, d_N)\}$, with $w_i \in \mathbf{V}, d_i \in \mathbf{D}$ (for example, occurrences of terms in documents, customers accessing at item description, etc.)

Introduction to LSA

Basic assumptions

The approach of LSA (Latent Semantic Analysis) refers to three assumptions:

- sematic information can be derived from the $\mathbf{V}, \mathbf{D}$ matrix
- dimensionality reduction is a key aspect for such derivation
- "terms" and "documents" can be modeled as points (vectors) in a euclidean space

Framework

1. Dictionary $\mathbf{V}$ of $V = |\mathbf{V}|$ terms $t_1, t_2, \ldots, t_V$
2. Corpus $\mathbf{D}$ of $D = |\mathbf{D}|$ documents $d_1, d_2, \ldots, d_D$
3. Each document d_i is a sequence of N_i occurrences of terms from $\mathbf{V}$

Model

Idea

1. Each document d_i is considered as a multiset of N_i terms from $\mathbf{V}$ (hypothesis "bag of words")
2. There exists a correspondance between $\mathbf{V}$ and $\mathbf{D}$, and a vector space $\mathcal{S}$. To each term t_i a vector $\mathbf{u}_i$ is associated, hence to each document d_j it is associated a vector $\mathbf{v}_j$ in $\mathcal{S}$

Occurrence matrix

Matrix $\mathbf{W} \in \mathbb{R}^{V \times D}$: $\mathbf{W}(i,j)$ is associated to the occurrences of term t_i in document d_j. The value of $\mathbf{W}(i,j)$ depends from the measure function predefined (tf, tf-idf, entropy, etc.).

- Terms: row vectors (dimension D)
- Documents: column vectors (dimension V)

Model

Problems

1. The values V and D are very large
2. The vectors for t_i and d_j are very sparse
3. The space for terms and documents are different

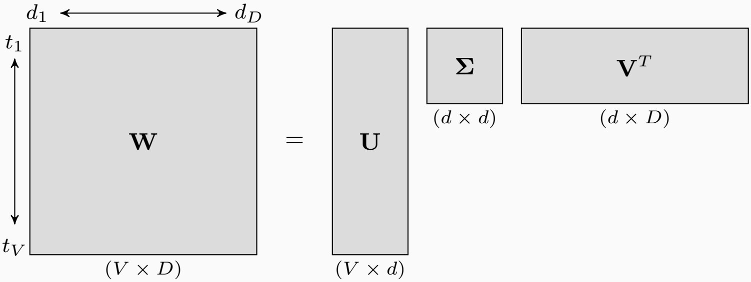
Solution

Applying singular value decomposition.
Let $\mathbf{W} \in \mathbb{R}^{n \times m}$ a matrix of rank $d \leq \min(n, m)$ and let $n > m$. Then, there exist

- $\mathbf{U} \in \mathbb{R}^{n \times d}$ orthonormal ($\mathbf{U}^T\mathbf{U} = \mathbf{I}_d$)
- $\mathbf{V} \in \mathbb{R}^{m \times d}$ orthonormal ($\mathbf{V}\mathbf{V}^T = \mathbf{I}_d$)
- $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ diagonal

such that $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$

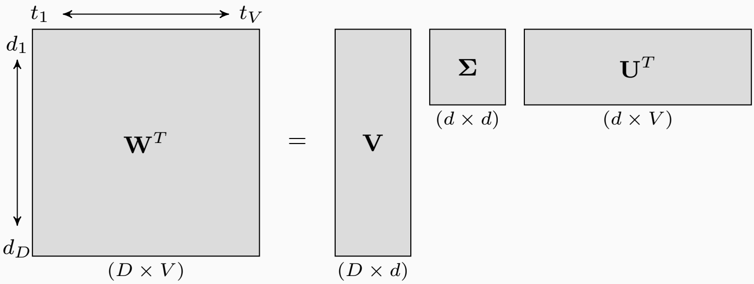
Application of SVD



Effect

The rows of $\mathbf{W}$ (terms) are projected on a d-dimensional subspace of $\mathbb{R}^D$ having the set of columns of $\mathbf{V}$ as basis: this defines for each term a new representation (row of $\mathbf{U\Sigma} \in \mathbb{R}^d$) as a vector of the coordinates with respect to this basis

Application of SVD



Effect

The rows of $\mathbf{W}^T$ (documents) are projected on a d-dimensional subspace of $\mathbb{R}^V$ having the set of columns of $\mathbf{U}$ as basis: this defines for each document a new representation (row of $\mathbf{V\Sigma} \in \mathbb{R}^d$) as a vector of the coordinates with respect to this basis

LSA

Dimensionality reduction

The dimension d of the projection space may be predefined, and less than the rank of $\mathbf{W}$. In this case,

$$\mathbf{W} \approx \overline{\mathbf{W}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Approximation

The property

$$\min_{\mathbf{A}:\text{rank}(\mathbf{A})=d} ||\mathbf{W} - \mathbf{A}||_2 = ||\mathbf{W} - \overline{\mathbf{W}}||_2$$

holds. The matrix $\overline{\mathbf{W}}$ is the matrix that best approximates $\mathbf{W}$ among all matrices of rank d according to the norm L_2 or of Frobenius

$$||\mathbf{A}||_2 = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

LSA

Effect

SVD defines a transformation from two discrete vector spaces $\mathcal{V} \in \mathbb{Z}^D$ and $\mathcal{D} \in \mathbb{Z}^V$, to one smaller continuous vector space, $\mathcal{T} \in \mathbb{R}^d$.
The dimension of $\mathcal{T}$ is less than or equal to the rank (unknown) of $\mathbf{W}$, and it is lower bounded from the amount of distortion acceptable in the projection.

Interpretation

$\hat{\mathbf{W}}$ captures the largest part of the associations between terms and documents $\mathbf{W}$, neglecting the least significative relations.

- Each term is represented as a (linear) combinations of hidden concepts, corresponding to the columns of $\mathbf{V}$: terms with projections near to each other tend to appear in the same documents (or in semantically similar documents)
- Each document is represented as a (linear) combinations of hidden topics, corresponding to the columns of $\mathbf{U}$: documents with projections near to each other tend to include the same terms (or semantically similar terms)