Statistical Machine Learning

Lecture 6: Singular Value Decomposition (SVD)
and Principal Component Analysis (PCA)

2022-23

.

SVD

- Common data pre-processing technique to reduce high-dimensional data.
- Provides a robust justification for matrix (data arrays) approximation (reduction).
- Provides a systematic way to determine a low-dimensional approximation to a high-dimensional data in terms of dominant patterns

SVD Theorem

- *Theorem:* Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be matrix of rank $r \in \{0, 1, \dots, \min\{m, n\}\}$. There exists an $m \times m$ real orthogonal matrix $\boldsymbol{U}$ and an $n \times n$ real orthogonal matrix $\boldsymbol{V}$ such that:
$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T,$$
where
$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \sigma_r \end{pmatrix},$$
where $\boldsymbol{D} \in \mathbb{R}^{m \times n}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ and diagonal, and $\sigma_i \in \mathbb{R}$ such that $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$.

SVD Theorem

- The decomposition is also expressed using the following partition:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = (\boldsymbol{U}_1, \boldsymbol{U}_2)\begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{V}_1^T \\ \boldsymbol{V}_2^T \end{pmatrix} = \boldsymbol{U}_1 \boldsymbol{\Sigma} \boldsymbol{V}_1^T,$$

  where $\boldsymbol{U}_1$ and $\boldsymbol{V}_1$ are $m \times r$ and $n \times r$ matrices, respectively, with orthonormal columns and the $\mathbf{0}$ submatrices have compatible dimensions for the above partition to be sensible.

SVD Theorem

- *Proof:* The matrix $\boldsymbol{A}^T\boldsymbol{A}$ is an $n \times n$ symmetric matrix. Those, its eigenvalues λ are all real. Also:

$$\begin{aligned}
\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v}_i &= \lambda_i \boldsymbol{v}_i \\
\boldsymbol{v}_i^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v}_i &= \lambda_i \boldsymbol{v}_i^T\boldsymbol{v}_i \\
(\boldsymbol{A}\boldsymbol{v}_i)^T\boldsymbol{A}\boldsymbol{v}_i &= \lambda_i \left\|\boldsymbol{v}_i\right\|^2 \\
0 \leq \frac{\left\|\boldsymbol{A}\boldsymbol{v}_i\right\|^2}{\left\|\boldsymbol{v}_i\right\|^2} &= \lambda_i.
\end{aligned}$$

Let $\{\lambda_1, \ldots, \lambda_n\}$ denote the set of eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$.

SVD Theorem

- The spectral decomposition of $\boldsymbol{A}^T\boldsymbol{A}$ is then:

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T,$$

where $[\boldsymbol{\Lambda}]_{i,i} = \lambda_i$ and $[\boldsymbol{\Lambda}]_{i,j} = 0$ for all $i \neq j$. Also $\boldsymbol{V}$ is an $n \times n$ real orthonormal matrix. We can assume, without loss of generality, that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. Since the rank of the matrix $\boldsymbol{A}$ is $\rho(\boldsymbol{A}) = r$, and by using the fundamental theorem of ranks $\rho(\boldsymbol{A}) = \rho(\boldsymbol{A}) = \rho(\boldsymbol{A}\boldsymbol{A}^T) = \rho(\boldsymbol{A}^T\boldsymbol{A})$. It follows that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0.$$

SVD Theorem

- The n columns of $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots \boldsymbol{v}_n)$ are the correspondent eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$. Partition $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2)$ such that $\boldsymbol{V}_1 = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$ is $n \times r$ and $\boldsymbol{V}_1^T\boldsymbol{V}_1 = \boldsymbol{I}_{r\times r}$. Let $\boldsymbol{\Lambda}_1$ be the $r \times r$ diagonal matrix with $\lambda_1 \geq \cdots \geq \lambda_r > 0$ elements in the diagonal. The spectral decomposition produces:

$$
\begin{aligned}
\boldsymbol{A}^T\boldsymbol{A} &= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T \\
\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{V} &= \boldsymbol{V}\boldsymbol{\Lambda} \\
&= (\boldsymbol{V}_1, \boldsymbol{V}_2)\,\boldsymbol{\Lambda} \\
&= (\boldsymbol{V}_1\boldsymbol{\Lambda}_1, \boldsymbol{0})\,.
\end{aligned}
$$

SVD Theorem

- Thus

$$\begin{aligned} \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{V}_2 &= \mathbf{0} \\ (\boldsymbol{A} \boldsymbol{V}_2)^T \boldsymbol{A} \boldsymbol{V}_2 &= \mathbf{0} \\ \boldsymbol{A} \boldsymbol{V}_2 &= \mathbf{0}, \end{aligned}$$

thus each column of $\boldsymbol{V}_2$ belongs to the null space of $\boldsymbol{A}$. Let us define $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}_1^{\frac{1}{2}}$, and define the $m \times r$ matrix $\boldsymbol{U}_1$:

$$\boldsymbol{U}_1 = \boldsymbol{A} \boldsymbol{V}_1 \boldsymbol{\Sigma}^{-1}.$$

SVD Theorem

- Observe that from the spectral decomposition we also have that

$$
\begin{aligned}
\mathbf{A}^T\mathbf{A}\mathbf{V}_1 &= \mathbf{V}_1\mathbf{\Lambda}_1 = \mathbf{V}_1\mathbf{\Sigma}^2 \\
(\mathbf{A}\mathbf{V}_1)^T\mathbf{A}\mathbf{V}_1 &= \mathbf{V}_1^T\mathbf{V}_1\mathbf{\Sigma}^2 = \mathbf{\Sigma}^2 \\
\mathbf{\Sigma}^{-1}(\mathbf{A}\mathbf{V}_1)^T\mathbf{A}\mathbf{V}_1\mathbf{\Sigma}^{-1} &= \mathbf{I}_{r\times r} \\
\left(\mathbf{A}\mathbf{V}_1\mathbf{\Sigma}^{-1}\right)^T\mathbf{A}\mathbf{V}_1\mathbf{\Sigma}^{-1} &= \mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_{r\times r}.
\end{aligned}
$$

Therefore the columns of $\mathbf{U}_1$ are orthonormal.

SVD Theorem

- We also observe that:

$$\begin{aligned} \boldsymbol{U}_1 &= \boldsymbol{A}\boldsymbol{V}_1\boldsymbol{\Sigma}^{-1} \\ \boldsymbol{I}_{r\times r} = \boldsymbol{U}_1^T\boldsymbol{U}_1 &= \boldsymbol{U}_1^T\boldsymbol{A}\boldsymbol{V}_1\boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma} &= \boldsymbol{U}_1^T\boldsymbol{A}\boldsymbol{V}_1 \\ \boldsymbol{U}_1\boldsymbol{\Sigma}\boldsymbol{V}_1^T &= \boldsymbol{A}. \end{aligned}$$

Let us choose $(m-r)$ vectors $\boldsymbol{u}_k$, $k = r+1, \ldots, m$, such that these vectors are perpendicular to the columns of $\boldsymbol{U}_1$ and $\boldsymbol{u}_k^T\boldsymbol{u}_{k'} = \delta_{k,k'}$.

SVD Theorem

- So

$$\begin{aligned} \boldsymbol{U}_2 &= (\boldsymbol{u}_{r+1}, \ldots \boldsymbol{u}_m) \\ \boldsymbol{U}_2^T \boldsymbol{A} \boldsymbol{V}_1 &= \boldsymbol{U}_2^T \boldsymbol{U}_1 \boldsymbol{\Sigma} = \boldsymbol{0}. \end{aligned}$$

This implies that

$$\begin{aligned} \boldsymbol{U}^T \boldsymbol{A} \boldsymbol{V} &= \begin{pmatrix} \boldsymbol{U}_1^T \\ \boldsymbol{U}_2^T \end{pmatrix} \boldsymbol{A} (\boldsymbol{V}_1, \boldsymbol{V}_2) \\ &= \begin{pmatrix} \boldsymbol{U}_1^T \boldsymbol{A} \boldsymbol{V}_1 & \boldsymbol{U}_1^T \boldsymbol{A} \boldsymbol{V}_2 \\ \boldsymbol{U}_2^T \boldsymbol{A} \boldsymbol{V}_1 & \boldsymbol{U}_2^T \boldsymbol{A} \boldsymbol{V}_2 \end{pmatrix} \end{aligned}$$

SVD Theorem

- $$U^T A V = \begin{pmatrix} U_1^T A V_1 & 0 \\ U_2^T A V_1 & 0 \end{pmatrix} \qquad A V_2 = 0$$
$$= \begin{pmatrix} U_1^T A V_1 & 0 \\ 0 & 0 \end{pmatrix},$$

  thus

$$U^T A V = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} = D$$
$$A = U D V^T. \spadesuit$$

.

Approximation Theorem

- *Theorem*: The optimal rank-r approximation to $\boldsymbol{X}$, in a least-squares sense, is given by the rank-r SVD truncation $\tilde{\boldsymbol{X}}$:

$$\underset{\tilde{\boldsymbol{X}}:\rho(\tilde{\boldsymbol{X}})=r}{\operatorname{argmin}} \left\| \boldsymbol{X} - \tilde{\boldsymbol{X}} \right\|_F = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^T,$$

  where $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{V}}$ denote the first $r$leading columns of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, and $\tilde{\boldsymbol{\Sigma}}$ contains the leading $r \times r$ sub-block of $\boldsymbol{\Sigma}$. $\|\cdot\|_F$ is the Frobenius norm:

$$\forall \boldsymbol{M} \in \mathbb{R}^{n\times m}, \|\boldsymbol{M}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} M_{i,j}^2}.$$

Approximation Theorem

- Thus, if we consider $r' < r$

$$\tilde{\boldsymbol{X}}' = \sum_{k=1}^{r'} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^T$$

$\tilde{\boldsymbol{X}}'$ is the optimal approximation to $\boldsymbol{X}$ with r' components.

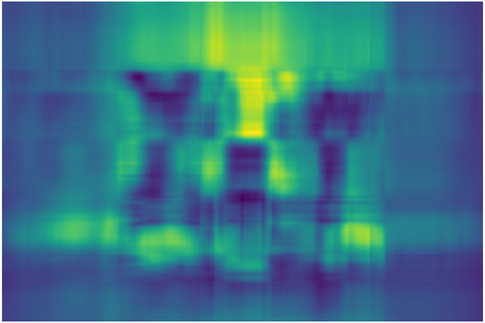
Approximation Theorem

- 

.

Approximation Theorem

- $r = 5$



Approximation Theorem

- $r = 20$

Approximation Theorem

- $r = 100$

.

Systems of Linear Equations

- The *pseudo-inverse* of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is define as the matrix $\boldsymbol{A}^{+}\mathbb{R}^{n \times m}$ such that if

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T}$$

then

$$\boldsymbol{A}^{+} \equiv \boldsymbol{V}\boldsymbol{D}^{+}\boldsymbol{U}^{T}$$

with

$$\boldsymbol{D} = \begin{cases} \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0}_{r \times (n-r)} \\ \boldsymbol{0}_{(m-r) \times r} & \boldsymbol{0}_{(m-r) \times (n-r)} \end{pmatrix} & r < \min\{m, n\} \\ \begin{pmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{0}_{(m-r) \times r} \end{pmatrix} & n = r < m \\ \begin{pmatrix} \boldsymbol{\Sigma} & , \boldsymbol{0}_{r \times (n-r)} \end{pmatrix} & m = r < n \\ \boldsymbol{\Sigma} & m = n = r \end{cases}$$

Systems of Linear Equations

- Then

$$D^+ = \begin{cases} \begin{pmatrix} \mathbf{\Sigma}^{-1} & \mathbf{0}_{r\times(m-r)} \\ \mathbf{0}_{(n-r)\times r} & \mathbf{0}_{(n-r)\times(m-r)} \end{pmatrix} & r < \min\{m,n\} \\ \begin{pmatrix} \mathbf{\Sigma}^{-1} & , \mathbf{0}_{r\times(m-r)} \end{pmatrix} & n = r < m \\ \begin{pmatrix} \mathbf{\Sigma}^{-1} \\ \mathbf{0}_{(n-r)\times r} \end{pmatrix} & m = r < n \\ \mathbf{\Sigma}^{-1} & m = n = r \end{cases} .$$

.

Systems of Linear Equations

- Consider the system $\boldsymbol{Ax} = \boldsymbol{b}$ for $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^m$ and a suitable $\boldsymbol{x} \in \mathbb{R}^n$.
- Let us explore the properties of the vector $\tilde{\boldsymbol{x}} \equiv \boldsymbol{A}^+ \boldsymbol{b}$.
- Let us consider the minimum norm defined as

$$\theta \equiv \min\{\|\boldsymbol{Ax} - \boldsymbol{b}\|_2 : \boldsymbol{x} \in \mathbb{R}^n\},$$

where

$$\|\boldsymbol{v}\|_2 = \sqrt{\boldsymbol{v}^T \boldsymbol{v}}$$

is the Euclidean norm.

Systems of Linear Equations

- The Euclidean norm is indifferent under changes of coordinates. Given a unitary matrix $\boldsymbol{P} \in \mathbb{R}^{m \times m}$ such that $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}_{m \times m}$ we have that for all $\boldsymbol{y} \in \mathbb{R}^m$ $\|\boldsymbol{y}\|_2 = \left\|\boldsymbol{P}^T\boldsymbol{y}\right\|_2$.
- Then

$$
\begin{aligned}
\tilde{\theta} &\equiv \|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{b}\|_2 \\
&= \left\|\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{b} - \boldsymbol{b}\right\|_2 \\
&= \left\|\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^+\boldsymbol{U}^T\boldsymbol{b} - \boldsymbol{b}\right\|_2 \\
&= \left\|\boldsymbol{U}\boldsymbol{D}\boldsymbol{D}^+\boldsymbol{U}^T\boldsymbol{b} - \boldsymbol{b}\right\|_2 \\
&= \left\|\boldsymbol{D}\boldsymbol{D}^+\boldsymbol{U}^T\boldsymbol{b} - \boldsymbol{U}^T\boldsymbol{b}\right\|_2 .
\end{aligned}
$$

Systems of Linear Equations

- Observe that

$$\boldsymbol{D}\boldsymbol{D}^{+} = \begin{cases} \begin{pmatrix} \boldsymbol{I}_{r\times r} & \boldsymbol{0}_{r\times(m-r)} \\ \boldsymbol{0}_{(m-r)\times r} & \boldsymbol{0}_{(m-r)\times(m-r)} \end{pmatrix} & r < \min\{m, n\} \\ \boldsymbol{I}_{r\times r} & m = r \end{cases}.$$

Systems of Linear Equations

- Therefore

$$(\boldsymbol{I}_{m\times m} - \boldsymbol{D}\boldsymbol{D}^{+})(\boldsymbol{U}^T\boldsymbol{b}) = \begin{pmatrix} \boldsymbol{0}_{r\times r} & \boldsymbol{0}_{r\times(m-r)} \\ \boldsymbol{0}_{(m-r)\times r} & \boldsymbol{I}_{(m-r)\times(m-r)} \end{pmatrix} (\boldsymbol{U}^T\boldsymbol{b})$$

$$= \begin{pmatrix} 0 \\ \vdots \\ 0 \\ [\boldsymbol{U}^T\boldsymbol{b}]_r \\ \vdots \\ [\boldsymbol{U}^T\boldsymbol{b}]_m \end{pmatrix},$$

Systems of Linear Equations

- Then

$$\tilde{\theta} = \begin{cases} \sqrt{\sum_{\ell=r+1}^{m} [\boldsymbol{U}^T \boldsymbol{b}]_\ell^2} & r < \min\{m, n\} \\ 0 & r = m \end{cases}.$$

- If $r = \rho(\boldsymbol{A}) = m$ and therefore $m \leq n$, then $\tilde{\theta} = 0$ which must be the minimum θ. Therefore $\boldsymbol{A}^+ \boldsymbol{b}$ is a solution to the linear system (with minimum norm).

Systems of Linear Equations

- If $r = \rho(\boldsymbol{A}) = m$ and therefore $m \leq n$, then $\tilde{\theta} = 0$ which must be the minimum θ. Therefore $\boldsymbol{A}^+\boldsymbol{b}$ is a solution to the linear system (with minimum norm).
- If $r < \min\{m, n\}$ we have that $\rho(\boldsymbol{A}) < \dim(\boldsymbol{b})$ and the linear system may be undetermined (therefore no solution exists). In such a case observe that for any $\boldsymbol{z} = \tilde{\boldsymbol{x}} + \boldsymbol{v}$ we have that

$$
\begin{aligned}
\theta(\boldsymbol{z}) &= \|\boldsymbol{A}(\tilde{\boldsymbol{x}} + \boldsymbol{v}) - \boldsymbol{b}\|_2 \\
&= \left\| \left(\boldsymbol{D}\boldsymbol{D}^+ - \boldsymbol{I}_{m \times m}\right) \boldsymbol{U}^T\boldsymbol{b} + \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{v} \right\|_2
\end{aligned}
$$

Systems of Linear Equations

- Observe that

$$
\begin{aligned}
\boldsymbol{U}^T \boldsymbol{A} \boldsymbol{v} &= \boldsymbol{D} \boldsymbol{V}^T \boldsymbol{v} \\
&= \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0}_{r \times (n-r)} \\ \boldsymbol{0}_{(m-r) \times r} & \boldsymbol{0}_{(m-r) \times (n-r)} \end{pmatrix} \begin{pmatrix} \{\boldsymbol{V}^T \boldsymbol{v}\}_r \\ \{\boldsymbol{V}^T \boldsymbol{v}\}_{n-r} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\Sigma} \{\boldsymbol{V}^T \boldsymbol{v}\}_r \\ \boldsymbol{0}_{m-r} \end{pmatrix}.
\end{aligned}
$$

Systems of Linear Equations

- Thus

$$\begin{aligned}\theta(z) &= \left\| \begin{pmatrix} \boldsymbol{\Sigma}\{\boldsymbol{V}^T\boldsymbol{v}\}_r \\ \boldsymbol{0}_{m-r} \end{pmatrix} - \begin{pmatrix} \boldsymbol{0}_r \\ \{\boldsymbol{U}^T\boldsymbol{b}\}_{m-r} \end{pmatrix} \right\|_2 \\ &= \sqrt{\sum_{\ell=1}^{r}(\sigma_\ell[\boldsymbol{V}^T\boldsymbol{v}]_\ell)^2 + \sum_{\ell=r+1}^{m}[\boldsymbol{U}^T\boldsymbol{b}]_\ell} \\ &\geq \tilde{\theta}\,\clubsuit.\end{aligned}$$

- Therefore the only $\boldsymbol{v}$ that satisfies the equal sign is $\boldsymbol{0}_n$.

Dominant Correlations

- There are two correlation matrices that can be computed from $\boldsymbol{X}$: $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$:

$$\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{U}\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\boldsymbol{V}^T\boldsymbol{V}\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\boldsymbol{U}^T = \boldsymbol{U}\begin{pmatrix} \boldsymbol{\Sigma}^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\boldsymbol{U}^T$$

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\boldsymbol{U}^T\boldsymbol{U}\begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T.$$

Dominant Correlations

- $\boldsymbol{X}\boldsymbol{X}^T$ is usually much larger than $\boldsymbol{X}^T\boldsymbol{X}$. Observe that we can find

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Sigma}^2$$

and from here approximate $\tilde{\boldsymbol{U}}$ (the first r columns of $\boldsymbol{U}$) as

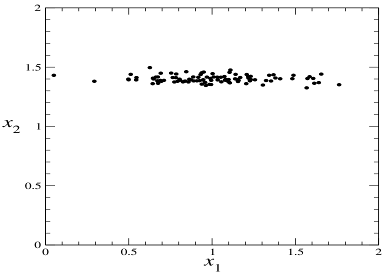
$$\tilde{\boldsymbol{U}} = \boldsymbol{X}\tilde{\boldsymbol{V}}\tilde{\boldsymbol{\Sigma}}^{-1}.$$

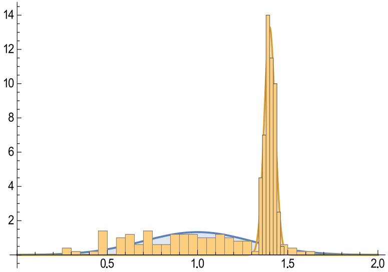
PCA

- PCA provides a hierarchical coordinate system to represent high-dimensional correlated data.
- Preprocessing of data by subtraction of the mean and reducing the variance to unity before performing SVD.
- The new coordinates are uncorrelated (orthogonal) but guard a maximum correlation with the measurements.
- Measurements are collected into a row vector. (Feature vector associated to an observable).

PCA

- Suppose we have a data set composed by 2-dimensional features such that:





- If we filter out the variations on x_2 the 'loss in information' is less than filtering out x_1.

PCA

- Given $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^{N}$, $\mathbf{x} \in \mathbb{R}^d$, then

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n,$$

  there is a matrix $\mathbf{B} \in \mathbb{R}^{d \times q}$ with $d \geq q > 0$ such that $\mathbf{z} \in \mathbb{R}^q$ defined as

$$\mathbf{z} = \mathbf{B}^T(\mathbf{x} - \boldsymbol{\mu}).$$

- Observe that $\mathbf{B}$ implements a reduction of dimensionality of the elements of the data set.

PCA

- Observe that

$$
\begin{aligned}
\mathbb{V}_{\boldsymbol{z}}[z] &= \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^T(\boldsymbol{x}-\boldsymbol{\mu})] \\
&= \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{B}^T(\boldsymbol{x}-\boldsymbol{\mu}))^2] - \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{B}^T(\boldsymbol{x}-\boldsymbol{\mu})]^2 \\
&= \mathbb{E}_{\boldsymbol{x}}[n\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\mu}] \\
&= \mathbb{E}_{\boldsymbol{x}}[x^T\boldsymbol{B}\boldsymbol{B}^T x] - \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\mu} \\
&= \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}^T x].
\end{aligned}
$$

- We want to find $\boldsymbol{B}$ that maximizes the variance of $\boldsymbol{z}$.

PCA

- We proceed sequentially. Let us define

$$V_1 = \mathbb{V}[z_1] = \frac{1}{N}\sum_{n=1}^{N} z_{1,n}^2.$$

- We need to find the direction in space $\mathbb{R}^d$, represented by the unit vector $\boldsymbol{b}_1$, such that $z_{1,n} = \boldsymbol{b}_1^T(\boldsymbol{x}_n - \boldsymbol{\mu})$ maximizes V_1. (Observe that the variances in $\boldsymbol{z}$ and in $\boldsymbol{x}$ without the shift in $\boldsymbol{\mu}$ are identical)

- Therefore

$$
\begin{aligned}
V_1 &= \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{b}_1^T\boldsymbol{x}_n\right)^2 \\
&= \boldsymbol{b}_1^T\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{b}_1 \\
&= \boldsymbol{b}_1^T\boldsymbol{S}\boldsymbol{b}_1 \\
\boldsymbol{S} &= \boldsymbol{b}_1^T\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)\boldsymbol{b}_1,
\end{aligned}
$$

where $\boldsymbol{S}$ is the data covariance matrix.

PCA

- The problem can be stated as:

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^T \boldsymbol{S} \boldsymbol{b}_1, \text{ subject to } \boldsymbol{b}_1^T \boldsymbol{b}_1 = 1.$$

- The constraint maximization problem can be defined with the aid of a Lagrange multiplier λ_1, such that:

$$\mathcal{L}_1(\boldsymbol{b}_1, \lambda_1) = \boldsymbol{b}_1^T \boldsymbol{S} \boldsymbol{b}_1 + \lambda_1 (1 - \boldsymbol{b}_1^T \boldsymbol{b}_1).$$

- The equations to be solved are:

$$\begin{aligned} \nabla_{\boldsymbol{b}_1} \mathcal{L}_1 &= 2\boldsymbol{S}\boldsymbol{b}_1 - 2\lambda_1 \boldsymbol{b}_1 = \mathbf{0} \\ \frac{\partial \mathcal{L}_1}{\partial \lambda_1} &= 1 - \boldsymbol{b}_1^T \boldsymbol{b}_1 = 0 \;\; \text{(normalization)} \end{aligned}$$

PCA

- The solutions to the problem are the eigenvectors of $\boldsymbol{S}$. We keep the one correspondent to the largest eigenvalue.
- Subsequent projections can be obtained by increasing the number of orthogonality constraints with previously found vectors, i.e.

$$\begin{aligned}
\mathcal{L}_2(\boldsymbol{b}_2, \lambda_1, \gamma_{1,2}) &= \boldsymbol{b}_2^T \boldsymbol{S} \boldsymbol{b}_2 + \lambda_2(1 - \boldsymbol{b}_2^T \boldsymbol{b}_2) + \gamma_{1,2} \boldsymbol{b}_1^T \boldsymbol{b}_2 \\
\nabla_{\boldsymbol{b}_2} \mathcal{L}_2 &= 2\boldsymbol{S}\boldsymbol{b}_2 - 2\lambda_2 \boldsymbol{b}_2 + \gamma_{1,2} \boldsymbol{b}_1 = \boldsymbol{0} \\
\frac{\partial \mathcal{L}_2}{\partial \lambda_2} &= 1 - \boldsymbol{b}_2^T \boldsymbol{b}_2 = 0 \ \ (\text{normalization}) \\
\frac{\partial \mathcal{L}_2}{\partial \gamma_{1,2}} &= \boldsymbol{b}_1^T \boldsymbol{b}_2 = 0 (\text{orthogonality}).
\end{aligned}$$

PCA

- From the first equation we obtain:

$$\begin{aligned} \boldsymbol{S}\boldsymbol{b}_2 &= \lambda_2 \boldsymbol{b}_2 - \frac{\gamma_{1,2}}{2}\boldsymbol{b}_1 \\ \boldsymbol{b}_1^T \boldsymbol{S}\boldsymbol{b}_2 &= \lambda_2 \boldsymbol{b}_1^T \boldsymbol{b}_2 - \frac{\gamma_{1,2}}{2} \\ &= (\boldsymbol{b}_2^T \boldsymbol{S}\boldsymbol{b}_1)^T = \lambda_1 \boldsymbol{b}_2^T \boldsymbol{b}_1 \end{aligned}$$

  and by the orthogonality $\boldsymbol{b}_1^T \boldsymbol{b}_2 = 0$ we have that, $\gamma_{1,2} = 0$ and

$$\boldsymbol{S}\boldsymbol{b}_2 = \lambda_2 \boldsymbol{b}_2.$$

- $\boldsymbol{b}_2$ must be the eigenvector of $\boldsymbol{S}$ corresponding to the second largest eigenvalue.

PCA

- Given the spectrum $\{\lambda_1, \ldots, \lambda_d\}$ of $\boldsymbol{S}$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$, the matrix $\boldsymbol{B}$ is then $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q)$.
- The vector z_n becomes

$$z_{n,j} = \boldsymbol{b}_j^T(\boldsymbol{x}_n - \boldsymbol{\mu}).$$

- The total fraction of the variance that is carried into the z-representation is

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{d} \lambda_j}.$$