Dimensionality reduction

Victor Kitov

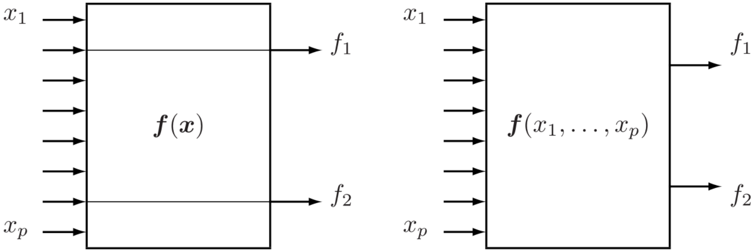
Skoltech
Skolkovo Institute of Science and Technology

November-December 2015.

Table of Contents

Definition

Feature selection / Feature extraction



(a) feature selector (b) feature extractor

Feature extraction: find transformation of original data which extracts most relevant information for machine learning task.

We will consider unsupervised dimensionality reduction methods, which try to preserve geometrical properties of the data.

Applications of dimensionality reduction

Applications:

- visualization in 2D or 3D
- reduce operational costs (less memory, disc, CPU usage on data transfer)
- remove multi-collinearity to improve performance of machine-learning models

Categorization

Supervision in dimensionality reduction:

- supervised (such as Fisher's direction)
- unsupervied

Mapping to reduced space:

- linear
- non-linear

Fisher's direction

- A direction of best separation between the classes in the following sence:
$$\frac{\left\| w^T(m_1 - m_2)\right\|^2}{w^T S_W w} \to \min_w$$
- Notation:
  - m_i is the mean of samples, belonging to class i
  - S_w is within class covariance, defined by
$$S_W = \frac{N_1}{N}\Sigma_1 + \frac{N_2}{N}\Sigma_2$$
Σ_i is covariance of observations in class i, N_i - number of observations in class i and N is the total number of observations.
- Optimal solution:
$$w \propto S_W^{-1}(m_1 - m_2)$$

Table of Contents

2. Principal component analysis
   - Definition
   - Derivation
   - Application details

Definition

Linear transformation of data, using orthogonal matrix $A = [a_1; a_2; ...a_D] \in \mathbb{R}^{DxD}$, $a_i \in \mathbb{R}^D$:

$$\xi = A^T x$$

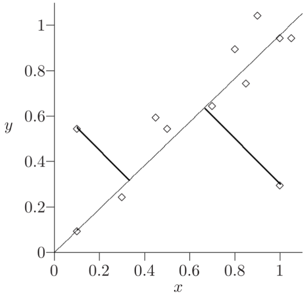
Equivalent ways to derive PCA:

1. Find orthogonal transform A yielding new variables ξ_i having stationary values for their variance and uncorrelated ξ_j
2. Find line of best fit, plane of best fit, etc.
    - fit is the sum of squares of perpendicular distances.
3. Find line, plane, etc. preserving most of the variability of the data.
    - variability is a sum of squared projections

We will use definition 1 for derivation and then prove that it is equivalent to definitions 2 and 3.

Example: line of best fit

- In PCA sum of squared of perpendicular distances to line is minimized.



- Not invariant to scale - features should be standardized.
- Method works for $\mathbb{E}x = 0$. So sample observations should be shifted to guarantee: $\sum_{n=1}^{N} x_n^D = 0,\ d = 1, 2, ..D.$

2 Principal component analysis

- Definition
- Derivation
- Application details

Covariance matrix properties

$\Sigma = cov[x] \in \mathbb{R}^{DxD}$ is symmetric positive semidefinite matrix

- has $\lambda_1, \lambda_2, ...\lambda_D$ eigenvalues, satisfying: $\lambda_i \in \mathbb{R}$, $\lambda_i \geq 0$.
- if eigenvalues are unique, corresponding eigenvectors are also unique
- always exists a set of orthogonal eigenvectors $z_1, z_2, ...z_D$: $\Sigma z_i = \lambda_i z_i$.

later we will assume that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D \geq 0$.

Derivation: 1st component

Consider first component:

$$\xi_1 = \sum_{j=1}^{D} a_{1j} x_j$$

Optimization problem:

$$\begin{cases} \mathrm{Var}\xi_1 \to \max_a \\ |a_1|^2 = a_1^T a_1 = 1 \end{cases}$$

Variance is equal:

$$\begin{aligned} Var[\xi_1] &= \mathsf{E}[\xi_1^2] - (\mathsf{E}\xi_1)^2 = \mathsf{E}[a_1^T x x^T a_1] - \mathsf{E}[a_1^T x]\mathsf{E}[x^T a_1] \\ &= a_1^T \left(\mathsf{E}[x x^T] - \mathsf{E}[x]\mathsf{E}[x^T] \right) a_1 = a_1^T \Sigma a_1 \end{aligned}$$

Derivation: 1st component

Optimization problem is equivalent to finding unconditional stationary value of

$$L(a_1,\nu) = a_1^T\Sigma a_1 - \nu(a_1^T a_1 - 1) \to extr_{a_1,\nu}$$

$$\frac{1}{2}\frac{\partial L}{\partial a_1} = 0 : \ \Sigma a_1 - \nu a_1 = 0$$

a_1 is selected from a set of eigenvectors of A. Since

$$Var[\xi_1] = a_1^T\Sigma a_1 = \lambda_i a_1^T a_1$$

a_1 is the eigenvector, corresponding to largest eigenvalue λ_i. Eigenvector is not unique if λ_{max} is a repeated root of characteristic equation: $|\Sigma - \nu I|$=0.

Derivation: 2nd component

$$\xi_2 = a_2^T x$$

$$\begin{cases} \mathrm{Var}[\xi_2] = a_2^T \Sigma a_2 \to \max_{a_2} \\ a_2^T a_2 = |a_2|^2 = 1 \\ cov[\xi_1, \xi_2] = a_2^T \Sigma a_1 = \lambda_1 a_2^T a_1 = 0 \end{cases}$$

Lagrangian (assuming $\lambda_1 > 0$)

$$L(a_2, \nu, \eta) = a_2^T \Sigma a_2 - \nu(a_2^T a_2 - 1) - \eta a_2^T a_1 \to extr_{a_2, \nu, \eta}$$

$$\frac{\partial L}{\partial a_2} = 0 : \; 2\Sigma a_2 - 2\nu a_2 - \eta a_1 = 0$$

$$a_1^T \frac{\partial L}{\partial a_2} = 2 a_1^T \Sigma a_2 - \eta = 0$$

Derivation: 2nd component

Since $a_1^T \Sigma a_2 = a_2^T \Sigma a_1 = 0$, we obtain $\eta = 0$. Then we have that:

$$\Sigma a_2 = \nu a_2$$

so a_2 is eigenvector of Σ, and since we maximize

$$\mathrm{Var}[\xi_2] = a_2^T \Sigma a_2 = \lambda_i a_2^T a_2$$

this should be eigenvector, corresponding to second largest eigenvalue λ_2.

Derivation: k-th component

$$\xi_k = a_k^T x$$

$$\begin{cases} Var[\xi_k] = a_k^T \Sigma a_k \to \max_{a_k} \\ a_k^T a_k = |a_k|^2 = 1 \\ cov[\xi_k, \xi_j] = a_k^T \Sigma a_j = \lambda_j a_k^T a_j = 0, \quad j = 1, 2, ...k-1. \end{cases}$$

Lagrangian (assuming $\lambda_j > 0$, $j = 1, 2, ...k-1$)

$$L(a_k, \nu, \eta) = a_k^T \Sigma a_k - \nu(a_k^T a_k - 1) - \sum_{i=1}^{k-1} \eta_i a_k^T a_i \to extr_{a_2, \nu, \eta}$$

$$\frac{\partial L}{\partial a_k} = 0 : \; 2\Sigma a_k - 2\nu a_k - \sum_{i=1}^{k-1} \eta_i a_i = 0$$

$$\forall j = 1, 2, ...k-1 : \; a_j^T \frac{\partial L}{\partial a_2} = 2 a_j^T \Sigma a_k - \eta_j = 0$$

Derivation: k-th component

Since $a_j^T \Sigma a_k = a_k^T \Sigma a_j = 0$, we obtain $\eta_j = 0$ for all $j = 1, 2, ...k-1$, so

$$\Sigma a_k = \nu a_k$$

a_k is then the eigenvector.
Variance of ξ_i is

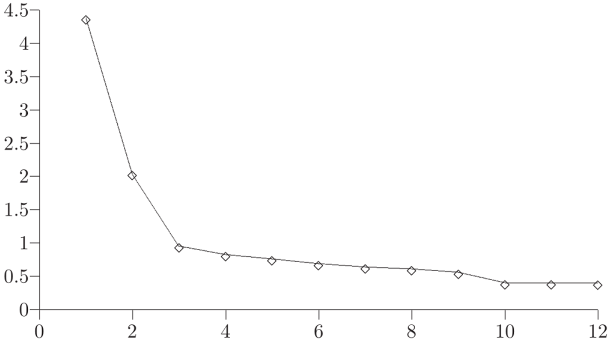
$$Var[\xi_k] = a_k^T \Sigma a_k = \lambda_i a_k^T a_k = \lambda_i$$

so a_k should be the eigenvector corresponding to the k-th largest eigenvalue λ_k.

2 Principal component analysis
- Definition
- Derivation
- Application details

Number of components

- Data visualization: 2 or 3 components.
- Take most significant components until their variance falls sharply down:



Number of components

Remind that $A=[a_1|a_2|...|a_D]$, $A^TA=I$, $\xi=A^Tx$.
Denote $S_k=[\xi_1,\xi_2,...\xi_k,0,0,...,0]\in\mathbb{R}^D$

$$\mathbb{E}[\|S_k\|^2]=\mathbb{E}[\xi_1^2+\xi_2^2+...+\xi_k^2]=\sum_{i=1}^{k}\text{var}\,\xi_i=\sum_{i=1}^{k}\lambda_i$$

$$\begin{aligned}\mathbb{E}[\|S_D\|^2] &= \mathbb{E}[\xi^T\xi]= \\ &= \mathbb{E}x^TAA^Tx=\mathbb{E}\left[x^Tx\right]=\mathbb{E}[\|x\|^2]\end{aligned}$$

Select such k^* that

$$\frac{\mathbb{E}[\|S_k\|^2]}{\mathbb{E}[\|x\|^2]}=\frac{\mathbb{E}[\|S_k\|^2]}{\mathbb{E}[\|S_D\|^2]}=\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{D}\lambda_i}>threshold$$

We may select k^* to account for 90%, 95% or 99% of total variance.

Transformation $\xi \rightleftarrows x$

Dependence between original and transformed features:

$$\xi = A^T(x-\mu),\; x = A\xi + \mu,$$

where μ is the mean of the original non-shifted data.
Taking first r components - $A_r = [a_1|a_2|...|a_r]$, we get the image of the reduced transformation:

$$\xi_r = A_r^T(x-\mu)$$

ξ_r will correspond to

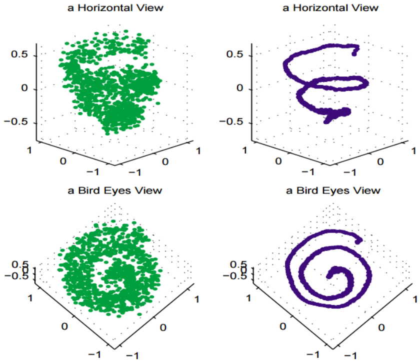
$$x_r = A\begin{pmatrix}\xi_r\\ 0\end{pmatrix} + \mu = A_r\xi_r + \mu$$

$$x_r = A_rA_r^T(x-\mu) + \mu$$

$A_rA_r^T$ is projection matrix with rank r.

Application - data filtering

Local linear projection method:



X. Huo and Jihong Chen (2002). Local linear projection (LLP). First IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh, NC, October. http://www.gensips.gatech.edu/proceedings/.

Properties of PCA

- Depends on scaling of individual features.
- Assumes that each feature has zero mean.

- Covariance matrix replaced with sample-covariance.
- Does not require distribution assumptions about x.

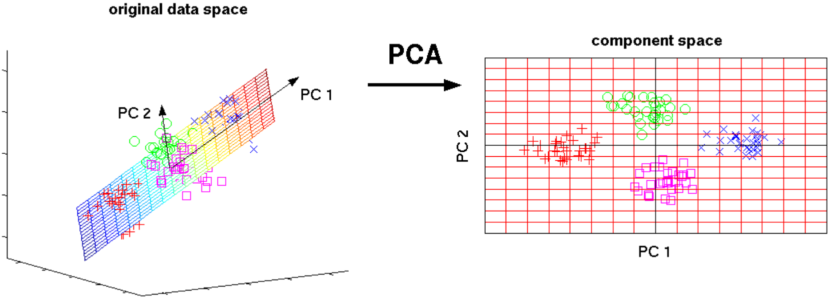
Example

Faces database:

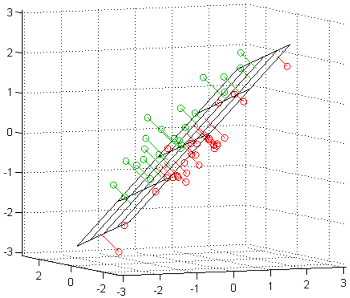
Eigenfaces

PCA for visualization

Uncorrelatedness does not imply independence.



Best hyperplane fit



Subspace L_k or rank k best fits points $x_1, x_2, ...x_D$ if sum of squared distances of these points to this plane is maximized over all planes of rank k.

Best hyperplane fit

For point x_i denote p_i the projection on plane L_k and h_i - orthogonal component. Then $\|x_i\|^2 = \|p_i\|^2 + \|h_i\|^2$.
For set of points:

$$\sum_i \|x_i\|^2 = \sum_i \|p_i\|^2 + \sum_i \|h_i\|^2$$

Since sum of squares is constant, minimization of $\sum_i \|h_i\|^2$ is equivalent to maximization of $\sum_i \|p_i\|^2$.

Another view on PCA directions

k-th step optimization problem for $\xi_k = a_k^T x$:

$$\begin{cases} Var[\xi_k] = a_k^T \Sigma a_k \to \max_{a_k} \\ a_k^T a_k = |a_k|^2 = 1 \\ cov[\xi_k, \xi_j] = a_k^T \Sigma a_j = \lambda_j a_k^T a_j = 0, \quad j = 1, 2, ...k-1. \end{cases}$$

can be equivalently represented as:

$$\begin{cases} \|Xa_k\|^2 \to \max_{a_k} \\ \|a_k\| = 1 \\ a_k \perp a_1, a_k \perp a_2, ...a_k \perp a_{k-1} \text{ if } k \geq 2 \end{cases} \tag{1}$$

since maximization of $\|Xa_k\|^2$ is equivalent to maximization of $\frac{1}{N}\|Xa_k\|^2 = \frac{1}{N}(Xa_k)^T (Xa_k) = \frac{1}{N} a_k^T X^T X a_k = a_k^T \Sigma a_k$.

Property of PCA

Theorem 1

For $1 \le k \le r$ let L_r be the subspace spanned by $a_1, a_2, ...a_r$. Then for each k L_k is the best-fit k-dimensional subspace for X.

Proof: use induction. For $r = 1$ the statement is true by definition since projection maximization is equivalent to distance minimization.

Suppose theorem holds for $r - 1$. Let L_r be the plane of best-fit of dimension with $\dim L = r$. We can always choose a orthonormal basis of L_r $b_1, b_2, ...b_r$ so that

$$\begin{cases} \|b_r\| = 1 \\ b_r \perp a_1, b_r \perp a_2, ...b_r \perp a_{r-1} \end{cases}$$

by setting b_r perpendicular to projections of $a_1, a_2, ...a_{r-1}$ on L_r.

Property of PCA

Consider the sum of squared projections:

$$\|Xb_1\|^2 + \|Xb_2\|^2 + ... + \|Xb_{r-1}\|^2 + \|Xb_r\|^2$$

By induction proposition:

$$\|Xb_1\|^2+\|Xb_2\|^2+...+\|Xb_{r-1}\|^2 \le \|Xa_1\|^2+\|Xa_2\|^2+...+\|Xa_{r-1}\|^2$$

and

$$\|Xb_r\|^2 \le \|Xa_r\|^2$$

since b_r by (31) satisfies constraints of optimization problem (30) and a_r is its optimal solution.

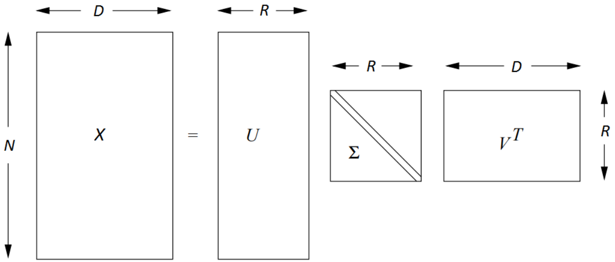
Table of Contents

SVD decomosition

Every matrix $X \in \mathbb{R}^{NxD}$ of rank R can be decomposed into the product of three matrices:

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

where $U \in \mathbb{R}^{NxR}$, $\Sigma \in \mathbb{R}^{RxR}$, $V^T \in \mathbb{R}^{RxD}$, and $\Sigma = diag\{\sigma_1, \sigma_2, ...\sigma_R\}$, $\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_R \ge 0$, $U^T U = I$, $V^T V = I$. I denotes identity matrix.

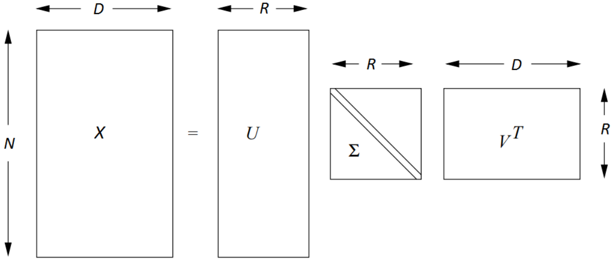


Applications of SVD

For square matrix X:

- U, V^T represent rotations, Σ represents scaling (with projection and reflection),
  every square matrix may be represented as superposition of rotation, scaling and another rotation.
- For full rank X:
  $$X^{-1} = V\Sigma^{-1}U^T,$$
  since $XX^{-1} = U\Sigma V^T V\Sigma^{-1}U^T = I$.
- For not-full-rank X: generalized inverse matrix equals
  $$X^{+} = V\Sigma^{+}U^T$$
  (A^+ is such matrix that $AA^+A = A$)

Interpretation of SVD



For X_{ij} let i denote objects and j denote properties.

- U represents standardized coordinates of concepts
- V^T represents standardized concepts representations
- Σ shows the magnitudes of presence of standardized concepts in X.

Example

	The lord of the rings	Harry Potter	Avatar	Titanic	Love story	A walk to remember
Andrew	4	5	5	0	0	0
John	4	4	5	0	0	0
Matthew	5	5	4	0	0	0
Anna	0	0	0	5	5	5
Maria	0	0	0	5	5	4
Jessika	0	0	0	4	5	4

Example

$$U = \begin{pmatrix} 0. & 0.6 & -0.3 & 0. & 0. & -0.8 \\ 0. & 0.5 & -0.5 & 0. & 0. & 0.6 \\ 0. & 0.6 & 0.8 & 0. & 0. & 0.2 \\ 0.6 & 0. & 0. & -0.8 & -0.2 & 0. \\ 0.6 & 0. & 0. & 0.2 & 0.8 & 0. \\ 0.5 & 0. & 0. & 0.6 & -0.6 & 0. \end{pmatrix}$$

$$\Sigma = \text{diag}\{\begin{pmatrix} 14. & 13.7 & 1.2 & 0.6 & 0.6 & 0.5 \end{pmatrix}\}$$

$$V^T = \begin{pmatrix} 0. & 0. & 0. & 0.6 & 0.6 & 0.5 \\ 0.5 & 0.6 & 0.6 & 0. & 0. & 0. \\ 0.5 & 0.3 & -0.8 & 0. & 0. & 0. \\ 0. & 0. & 0. & -0.2 & 0.8 & -0.6 \\ -0. & -0. & -0. & 0.8 & -0.2 & -0.6 \\ 0.6 & -0.8 & 0.2 & 0. & 0. & 0. \end{pmatrix}$$

Example (excluded insignificant concepts)

$$U_2 = \begin{pmatrix} 0. & 0.6 \\ 0. & 0.5 \\ 0. & 0.6 \\ 0.6 & 0. \\ 0.6 & 0. \\ 0.5 & 0. \end{pmatrix}$$

$$\Sigma_2 = \text{diag}\{\begin{pmatrix} 14. & 13.7 \end{pmatrix}\}$$

$$V_2^T = \begin{pmatrix} 0. & 0. & 0. & 0.6 & 0.6 & 0.5 \\ 0.5 & 0.6 & 0.6 & 0. & 0. & 0. \end{pmatrix}$$

Concepts may be

- patterns among movies (along j) - fantasy/romance
- patterns among people (along i) - boys/girls

Dimensionality reduction case: patterns along j axis.

Applications

- Example: new movie rating by new person

$$x = \begin{pmatrix} 5 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

- **Dimensionality reduction:** map x into concept space:

$$y = V_2^T x = \begin{pmatrix} 0 & 2.7 \end{pmatrix}$$

- **Recommendation system:** map y back to original movies space:

$$\widehat{x} = y V_2^T = \begin{pmatrix} 1.5 & 1.6 & 1.6 & 0 & 0 & 0 \end{pmatrix}$$

Fronebius norm

- Fronebius norm of matrix X is $\|X\|_F \stackrel{df}{=} \sqrt{\sum_{n=1}^{N} \sum_{d=1}^{D} x_{nd}^2}$
- Using properties $||X||_F = \operatorname{tr} XX^T$ and $\operatorname{tr} AB = \operatorname{tr} BA$, we obtain:

$$\begin{aligned} \|X\|_F &= \operatorname{tr}[U\Sigma V^T V\Sigma U^T] = \operatorname{tr}[U\Sigma^2 U^T] = \\ &= \operatorname{tr}[\Sigma^2 U^T U] = \operatorname{tr}[\Sigma^2] = \sum_{r=1}^{R} \sigma_r^2 \end{aligned} \qquad (2)$$

Matrix approximation

Consider approximation $X_k = U\Sigma_k V^T$, where
$\Sigma_k = \text{diag}\{\sigma_1, \sigma_2, ...\sigma_k, 0, 0, ..., 0\} \in \mathbb{R}^{RxR}$.

Theorem 2

X_k is the best approximation of X retaining k concepts.

Proof: consider matrix $Y_k = U\Sigma' V^T$, where Σ' is equal to Σ except some $R-k$ elements set to zero:
$\sigma'_{i_1} = \sigma'_{i_2} = ... = \sigma'_{i_{R-k}} = 0$. Then, using (2)

$$\|X - Y_k\|_F = \left\| U(\Sigma - \Sigma')V^T \right\|_F = \sum_{p=1}^{R-k} \sigma_{i_p}^2 \le \sum_{p=1}^{R-k} \sigma_p^2 = \|X - X_k\|_F$$

since $\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_R \ge 0$.

Matrix approximation

How many components to retain?

General case: Since

$$\|X - X_k\|_F = \left\|U(\Sigma - \Sigma_k)V^T\right\|_F = \sum_{i=k+1}^{R} \sigma_i^2$$

a reasonable choice is k^* such that

$$\frac{\|X - X_{k^*}\|_F}{\|X\|_F} = \frac{\sum_{i=k^*+1}^{R} \sigma_i^2}{\sum_{i=1}^{R} \sigma_i^2} \geq threshold$$

Visualization: 2 or 3 components.

Theorem 3

For any matrix Y_k with rank $Y_k = k$: $\|X - X_k\|_F \leq \|X - Y_k\|_F$

Finding U and V

- **Finding** V
  $X^TX=\left(U\Sigma V^T\right)^T U\Sigma V^T=(V\Sigma U^T)U\Sigma V^T=V\Sigma^2V^T$. It follows that
  $$X^TXV=V\Sigma^2V^TV=V\Sigma^2$$
  So V consists of eigenvectors of X^TX with corresponding eignvalues $\sigma_1^2,\sigma_2^2,...\sigma_R^2$.

- **Finding** U**:**
  $XX^T=U\Sigma V^T\left(U\Sigma V^T\right)^T=U\Sigma V^TV\Sigma U^T=U\Sigma^2U^T$. So
  $$XX^TU=U\Sigma^2U^TU=U\Sigma^2.$$
  So U consists of eigenvectors of $XX^T$with corresponding eigenvalues $\sigma_1^2,\sigma_2^2,...\sigma_R^2$.

Comments

- Denote the average $\bar{X} \in \mathbb{R}^D : \bar{X}_j = \sum_{i=1}^N x_{ij}$
- Denote the n-th row of X be $X_n \in \mathbb{R}^D : \; X_{nj} = x_{nj}$
- For centered X sample covariance matrix $\widehat{\Sigma}$ equals:

$$\begin{aligned} \widehat{\Sigma} &= \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X})(X_n - \bar{X})^T = \frac{1}{N}\sum_{n=1}^{N} X_n X_n^T \\ &= \frac{1}{N} X^T X \end{aligned}$$

- V **consists of principal components** since
  - V consists of eigenvectors of $X^T X$,
  - principal components are eignevectors of $\widehat{\Sigma}$ and
  - $\widehat{\Sigma} \propto X^T X$.