STA 141C - Big Data & High Performance Statistical Computing **Spring 2022**

Week 1-2: Review of linear models and linear algebra

Lecturer: Bo Y.-C. Ning *March 31, 2022*

Disclaimer: *My notes may contain errors, distribution outside this class is allowed only with the permission of the Instructor.*

Announcement

- Lecture recording available (but not sure about its quality) and will be uploaded on canvas
- OHs updated:
  - Instructor's OH: Tuesday from 3:00 pm to 5:00pm
  - Wei's OH: Thursday from 9:00 and 11:00am
  - Wei will monitor Piazza on Monday and Wednesday, I will do Friday.
  - When there is a homework due, that week's discussion session is used as TA's office hours

Today

- Review linear models
- Review linear algebra

-1.1 Linear models

The linear model is one of the most classical models in statistics/machine learning. The models describe the response variable Y with a linear combination of predictor variables $x_1, \dots, x_p$.

Suppose we observe $y_1, \dots, y_n$, we can use the shorthand notation $y = (y_1, \dots, y_n)'$. In this way, y is a length n vector, usually we say $y \in \mathbb{R}^n$, as most of the time, we work with real values. On the other hand, we denote $x_1 = (x_{11}, x_{12}, \dots, x_{1n})'$ and similar notations are used for $x_2, \dots, x_p$. Then for each x_j, $j = 1, \dots, p$, $x_j \in \mathbb{R}^n$. Since we have p x_js, we can stack them together in this way

$$X = [x_1, x_2, \dots, x_p],$$

then X is an $n \times p$ matrix; in short, $X \in \mathbb{R}^{n \times p}$.

We can write the model as

$$Y = X\beta + \epsilon,$$

where $\epsilon \in \mathbb{R}^n$ is some random error and β is often called the regression coefficients. The goal is to estimate β.

Note: often, in practice, one want to add an intercept β_0, we can re-write

$$X = [\mathbb{1}_n, x_1, x_2, \dots, x_p],$$

where $\mathbb{1}_n = (1, \dots, 1)'$, a vector with n 1s.

-1.1.1 How to solve β?

A most popular method is the *least squares*. In this approach, we estimate β by minimizing the residual sum of squares (RSS)

$$\min_\beta \text{RSS}(\beta), \quad \text{RSS}(\beta) = (Y - X\beta)'(Y - X\beta).$$

We solve $d\text{RSS}(\beta)/d\beta = 0 \to X'(y - X\beta) = 0$. Then we have $X'X\beta = X'y$. If $X'X$ is invertible, then

$$\hat{\beta} = (X'X)^{-1}X'y.$$

In some statistical textbook, one further assume $\epsilon_i \sim N(0, \sigma^2)$ follows a normal distribution with variance σ^2 (for simplicity, we set $\sigma^2 = 1$). Then, $\epsilon \sim \text{MVN}(0, I_n)$ follows a multivariate normal distribution and $Y \sim \text{MVN}(X\beta, I_n)$ is also a multivariate normal. Then the log-likelihood function can be written as

$$\log L(\beta) = -np\log(2\pi) - \frac{1}{2}(Y - X\beta)'(Y - X\beta).$$

One wishes to obtain the maximum likelihood estimator (MLE), i.e., solving β such that $d\log L(\beta)/d\beta = 0$. You will find that, again, $\hat{\beta}$ can be obtained by solving $X'X\beta = X'y$.

-1.1.2 Other examples

1. Ridge regression

   $$\min_\beta (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta$$

   This expression looks tedious, in short, we introduce the norm notation, we denote $\beta'\beta = \|\beta\|_2^2$ (or even shorter, $\|\beta\|^2$). So the above equation is often written as

   $$\min_\beta \|Y - X\beta\|^2 + \lambda\|\beta\|^2.$$

   The solution for β is known as the ridge estimator by solving $(X'X + \lambda I_p)'\beta = X'y$.

2. Nonlinear models

   The linear assumption can be strong, one might consider a nonlinear model:

   $$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + \epsilon_i, \; i = 1, \dots, n$$

   You can redefine $X = [\mathbb{1}_n, x_1, x_1^2, x_1^3]$ and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ and write $Y = X\beta + \epsilon$.

3. Nonparametric regression - regression splines

   A more flexible approach to predict Y is to use the so-called nonparametric approach. One example is regression splines. We minimize

   $$\sum_{i=1}^n \left(y_i - \sum_{i=1}^K \beta_j g_j(x_i) \right)^2,$$

where $g_j(x_i)$ is a spline function, e.g., B-splines. K is unknown.
Eventually, one can show that we solve β from

$$G'G\beta = G'y,$$

where G is a matrix containing $g_j(x_i)$s.

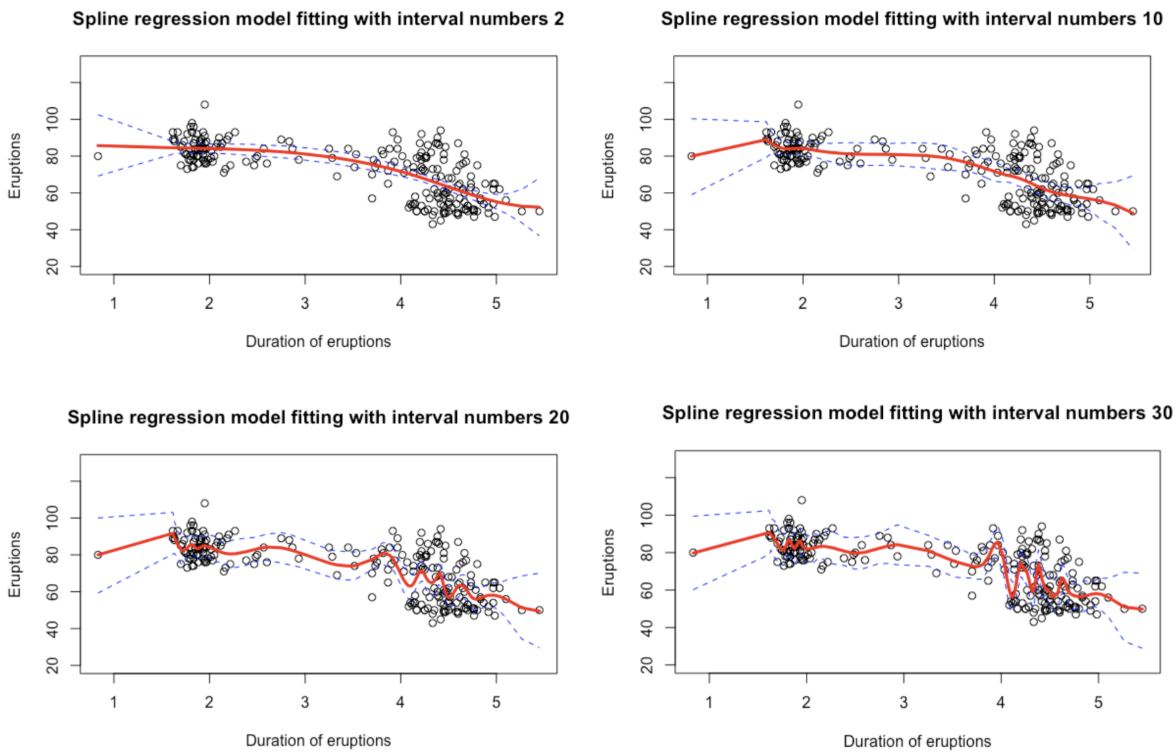


Figure -1.1: B-spline fitting of the Old Faithful Geyser data

-1.2 Why linear algebra?

As you can see from above, statistical computation often requires solving linear regressions in this form: $Ax = b$:

- Regression problem: $X'X\beta = X'y$
- Eigen-decomposition problem: $Ax = \lambda x$
- generalized eigen-decomposition problem: $Ax = \lambda Bx$
- sigular value decomposition: $A = U\Sigma V'$,
- ......

so we have to review the basic concepts in linear algebra.

-1.3 Review of linear algebra

Useful book: *The Matrix Cookbook*: `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

-1.3.1 Vector norms

- Vector norm $\|\cdot\|$: $\mathbb{R}^n \to \mathbb{R}$, for $a \in \mathbb{R}^n$:
  1. $\|a\| \geq 0$
  2. $\|a\| = 0$ if and only if $a = 0$
  3. homogeneity: $\|ca\| = c\|a\|$, $c \geq 0$
  4. triangle inequality: $\|a + b\| \leq \|a\| + \|b\|$
- ℓ_p-norm for $a = (a_1, \ldots, a_n)$, $a \in \mathbb{R}^n$,
  - ℓ_1-norm: $\|a\|_1 = \sum_{i=1}^n |a_i|$
  - ℓ_2-norm: $\|a\|_2 = \sqrt{\sum_{i=1}^n a_i^2}$
  - ℓ_∞-norm: $\|a\|_\infty = \max_i |a_i|$
  - In general, ℓ_p-norm: $\|a\|_p = (\sum_i |a_i|^p)^{1/p}$, $p \in [1, \infty]$
- Example: $a = (1, 2)'$:
  - ℓ_1-norm: $\|a\|_1 = 1 + 2 = 3$
  - ℓ_2-norm: $\|a\|_2 = \sqrt{1^2 + 2^2} = \sqrt{5}$
  - ℓ_∞-norm: $\|a\|_\infty = \max\{1, 2\} = 2$
- $\|a\|_\infty \leq \cdots \leq \|a\|_2 \leq \|a\|_1 \leq n\|a\|_\infty$
- Cauchy-Schwarz inequality: $|a'b| \leq \|a\|_2 \|b\|_2$ for $a, b \in \mathbb{R}^n$

-1.3.2 Distances between two vectors

- $a = [1, 2]$ and $b = [1.1, 2.5]$, how close is a and b?
- $b - a = [1.1 - 1, 2.5 - 2] = [0.1, 0.5]$
  - ℓ_1-norm: $d_1(b, a) = \|b - a\|_1 = 0.1 + 0.5 = 0.6$
  - ℓ_2-norm: $d_2(b, a) = \|b - a\|_2 = \sqrt{0.1^2 + 0.5^2} = \sqrt{0.26} \approx 0.51$
  - ℓ_∞-norm: $d_\infty(b, a) = \|b - a\|_\infty = \max\{0.1, 0.5\} = 0.5$
- Application: asymptotically consistent estimator: $d(\hat{\beta}, \beta_0) \to 0$ as sample size $n \to \infty$
- $d(a, b) \geq 0$; $d(a, b) = 0$ if and only if $a = b$; $d(a, b) = d(b, a)$; $d(a, b) \leq d(a, c) + d(c, b)$ (triangular inequality)

-1.3.3 Matrix norms

- Matrix norm $\|\cdot\|$: $\mathbb{R}^{m\times n} \to \mathbb{R}$, for $A \in \mathbb{R}^{m\times n}$:
  - We future require $\|AB\| \leq \|A\|\|B\|$ for $B \in \mathbb{R}^{n\times p}$
- Matrix norm
  - Maximum absolute column sum norm (Matrix-1 norm): $\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$
  - Spectra norm: $\|A\|_2 = \sqrt{\rho(A'A)}$, the square root of the maximum eigenvalue of $A'A$
  - Maximum absolute row sum norm (Matrix-∞ norm): $\|A\|_\infty = \max_i \sum_{j=1}^m |a_{ij}|$
  - Frobenius norm: $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$
- Similarly, one can also define the distances between two matrices $d(A, B)$
- Application: Matrix completion problem.
  - Observe a very sparse matrix $Y = (y_{ij})$. Want to impute all the missing entries. It is possible only when the matrix is structured, e.g., of low rank.

Figure -1.2: Netflix problem: impute the unobserved ratings for personalized recommendation. See `https://en.wikipedia.org/wiki/Netflix_Prize`

- Let $\Omega = \{(i, j) : \text{observed entries}\}$ index the observed entries and $P_\Omega(M)$ be the projection of matrix M to Ω, the problem

$$\min_{\text{rank}(X)\leq 2} \frac{1}{2}\|P_\Omega(Y) - P_\Omega(X)\|_F^2 = \frac{1}{2}\sum_{(i,j)\in\Omega}(y_{ij} - x_{ij})^2$$

See matrix completion (`https://en.wikipedia.org/wiki/Matrix_completion`)

-1.3.4 System of linear equations

The problem: $A\beta = b$, $A \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, $b \in \mathbb{R}^n$

- When is there a solution? The following statements are equivalent
  - The linear system $A\beta = b$ has a solution
  - $b \in \mathcal{C}(A)$
    - The column space of a matrix A, $\mathcal{C}(A)$, is the vector space made up of all linear combinations of the columns of A.
  - $\text{rank}((A, b)) = \text{rank}(A)$
  - $AA^-b = b$, A^- is the generalized inverse of A
- $Ax = b$ has a unique solution if and only if A has a full column rank
- If A has full row and column rank, then A is non-singular and has the unique solution $A^{-1}b$

-1.3.5 Linear independent and rank

- $x_1, \dots, x_n$ are linear independent: there exist scalars $a_1, \dots, a_n$ such that $a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0$ if and only if $a_1 = a_2 = \cdots = a_n = 0$; otherwise, $x_1, \dots, x_n$ are linear dependent
- Example: suppose we have x_1, x_2, x_3, $x_1 = a_2x_2 + a_3x_3$, then x_1 is linear dependent on x_2 and x_3.

Assume A is an $m \times n$ matrix

- $\text{rank}(A)$ is the maximum number of linearly independent rows (or columns) of a matrix.
- $\text{rank}(A) \le \min\{m, n\}$
- A matrix is full rank if $\text{rank}(A) = \min\{m, n\}$
- If $\text{rank}(A) = m$, A is full row rank; if $\text{rank}(A) = n$, A is full column rank
- If $m = n$, A is a square matrix
- For a square matrix $A \in \mathbb{R}^{m\times m}$, A is singular if $\text{rank}(A) < m$ and is non-singular if $\text{rank}(A) = m$
- $\text{rank}(AB) \le \min\{\text{rank}(A), \text{rank}(B)\}$, matrix multiplication cannot increase the rank
- A' (or A^T) is the transpose of A
- $\text{rank}(A) = \text{rank}(A') = \text{rank}(AA') = \text{rank}(A'A)$
- $\text{rank}(AB) = \text{rank}(A)$ if B has full row rank
- $\text{rank}(AB) = \text{rank}(B)$ if A has full column rank
- $\text{rank}(A + B) \le \text{rank}(A) + \text{rank}(B)$

-1.3.6 Matrix inverses

For $A \in \mathbb{R}^{m \times n}$

- The Moore-Penrose inverse of A is a matrix $A^+ \in \mathbb{R}^{n \times m}$ with the following properties:
  1. $AA^+A = A$
  2. $A^+AA^+ = A^+$
  3. A^+A and AA^+ are both symmetric
- Generalized inverse (g_1 inverse) satisfies (1): not unique
- Reflexive generalized inverse (g_2 inverse) satisfies (1) + (2): not unique
- Moore-Penrose inverse satisfies (1) + (2) + (3): unique
- Examples? (`https://en.wikipedia.org/wiki/Generalized_inverse`)

For $A \in \mathbb{R}^{n \times n}$

- A is invertible if there exist B such that $AB = BA = I_n$.
- If A is full rank (positive definite or nonsingular), then the generalized inverse is unique and denoted by A^{-1}.

For $A = X'X \in \mathbb{R}^{m \times m}$ and $X \in \mathbb{R}^{n \times m}$

- A is symmetric and positive semidefinite.
- If A is positive definite, A^{-1} is unique and $\beta = A^{-1}b$
- A is positive definite if and only if the columns of X are linearly independent (X has a full column rank)
- $\text{rank}(X) = \text{rank}(X') = \text{rank}(A) = \text{rank}(A')$
- $A = 0$ if and only if $X = 0$
- $P_X = X(X'X)^{-1}X'$ is symmetric, idempotent, P_X is known as the projection matrix

-1.3.7 Positive definite matrix

Assume $A \in \mathbb{R}^{n \times n}$ is symmetric ($A = A'$)

- A is positive definite if $x'Ax > 0$ for all x, we write $A \succ 0_{n \times n}$
- A is positive semi-definite (or nonnegative definite) if $x'Ax' \geq 0$ for all x, we write $A \succeq 0_{n \times n}$
- If A is a covariance matrix, $A \succeq 0$
- A positive definite matrix is full rank; that is, $\text{rank}(A) = n$
- For example, $X'X$ (also known as the Gramian matrix)
- $A \succeq B$ means $A - B$ is positive semi-definite

-1.3.8 Orthogonality

- v_1 is orthogonal to v_2, we write $v_1 \perp v_2$ (more often $\langle v_1, v_2 \rangle = v_1' v_2 = 0$)
- v_1 is orthonormal to v_2 if v_1 is orthogonal to v_2 and $\|v_1\| = 1$ and $\|v_2\| = 1$
- A set of nonzero, mutually orthogonal vectors are linearly independent.
- A real square matrix $A \in \mathbb{R}^{n \times n}$ is orthogonal if $A'A = I_n$
- Orthogonal matrix is of full rank, thus $A' = A^{-1}$ and $AA' = A'A = I_n$.

-1.3.9 Method of least squares

Goal: Approximate $y \in \mathbb{R}^n$ by a linear combination of column of $X = (x_1, \dots, x_p)$, $X \in \mathbb{R}^{n \times p}$

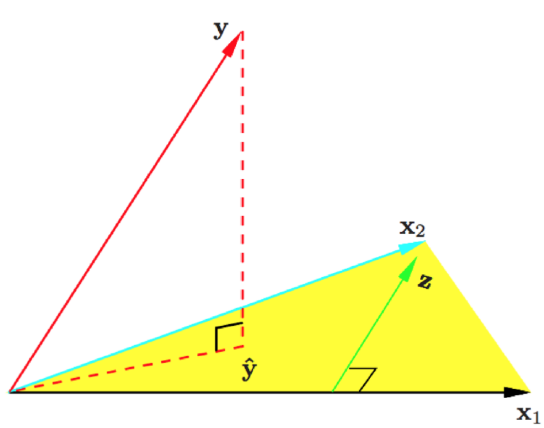
- Least square criterion: $\min Q(\beta) = \|y - X\beta\|_2^2$
- Any solution to the normal equation $X'X\beta = X'y$ is a minimizer of the least squares criterion $Q(b)$
- Solutions to the normal equation (if $X'X$ is positive definite)

  $$\hat{\beta} = (X'X)^{-1} X'y$$

  If $X'X$ is positive semidefinite

  $$\hat{\beta} = (X'X)^{-} X'y + (I_p - (X'X)^{-} X'X)q,$$

  where q is arbitrary.
- $P_X = X(X'X)^{-1}X'$ is the orthogonal projection onto $\mathcal{C}(X)$
- The fitted value from the least squares solution $\hat{y} = P_X y$ is the orthogonal projection of the response y onto the column space $\mathcal{C}(X)$.



- Decompose y

  $$y = P_X y + (I - P_X)y = \hat{y} + \hat{e}$$

  and $\|y\|_2^2 = \|\hat{y}\|_2^2 + \|\hat{e}\|_2^2$.

-1.3.10 Idempotent matrix and projection

For a matrix $P \in \mathbb{R}^{n\times n}$

- P is idempotent if and only if $P^2 = PP = P$
- A matrix P is a projection on a vector space $\mathcal{V}$ if
  - P is idempotent
  - $Px \in \mathcal{V}$ for all x
  - $Pz = z$ for all $z \in \mathcal{V}$
- A symmetric, idempotent matrix is called an orthogonal projection (P_X)
- Many books use the term "projection" in the sense of of orthogonal projection.

-1.3.11 Eigenvalue and eigenvector

Assume $A \in \mathbb{R}^{n\times n}$ is a square matrix

- Eigenvalues are defined as roots of the characteristic equation $\det(\lambda I_n - A) = 0$
- If λ is an eigenvalue of A, then there exist non-zero $x, y \in \mathbb{R}^n$ such that $Ax = \lambda x$ and $y'A = \lambda y'$, x is the (column) eigenvector and y is the row eigenvector of A associated with the eigenvalue λ
- A is singular if and only if it has at least one 0 eigenvalue.
- Eigenvectors associated with distinct eigenvalues are linearly independent
- Eigenvalues of an upper or lower triangular matrix are its diagonal entries: $\lambda = a_{ii}$
- Eigenvalues of an idempotent matrix are either 0 or 1
- In most statistical applications, we deal with eigenvalues/eigenvectors of symmetric matrices. The eigenvalues and eigenvectors of a real symmetric matrix are real.
- Eigenvectors associated with distinct eigenvalues of a symmetry matrix are orthogonal.
- Eigen-decompostion of a symmetric matrix: $A = U\Lambda U'$, where
  - $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$
  - Columns of U are the eigenvectors which are mutually orthonormal
  - A real symmetric matrix is positive semidefinite (positive definite) if and only if all eigenvalues are nonnegative (positive).
  - $\mathrm{tr}(A)$ (a square matrix not require to be symmetric), $\mathrm{tr}(A) = \mathrm{tr}(U\Lambda U') = \mathrm{tr}(U'U\Lambda) = \mathrm{tr}(\Lambda) = \sum_i \lambda_{ii}$

-1.3.12 Trace

A is a square matrix, $A \in \mathbb{R}^{n\times n}$

- $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$
- $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$
- $\mathrm{tr}(\lambda A) = \lambda \mathrm{tr}(A)$
- $\mathrm{tr}(A') = \mathrm{tr}(A)$
- In general, $\mathrm{tr}(A_1 A_2 \dots A_k) = \mathrm{tr}(A_k A_1 \dots A_{k-1}) = \mathrm{tr}(A_{j+1} \dots A_k A_1 \dots A_j)$

-1.3.13 Determinant

- If $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $\det(A) = a_{11}a_{22} - a_{12}a_{21}$.
- $\det(AB) = \det(A)\det(B)$ for $A, B \in \mathbb{R}^{n\times n}$
- $\det A^{-1} = 1/\det(A)$
- $\det(A) = \det(A')$ and, for a scalar c and $A \in \mathbb{R}^{n\times n}$, $\det(cA) = c^n \det(A)$
- The determinant of an upper or lower triangular matrix is the product of its diagonal elements.
- A is nonsingular (or positive definite) if and only if $\det(A) \neq 0$
- A is singular if and only if if $\det(A) = 0$
- If $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ is a block matrix, $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{n\times m}$, and $D \in \mathbb{R}^{n\times n}$, then

$$\det(M) = \det(A - BD^{-1}C)\det(D) = \det(A)\det(D - CA^{-1}B)$$

-1.3.14 Singular value decomposition (SVD)

For $A \in \mathbb{R}^{m\times n}$ and $p = \min\{m, n\}$

- Singular value decomposition: $A = U\Sigma V'$, where
  - $U = (u_1, \dots, u_m) \in \mathbb{R}^{m\times m}$ is orthogonal
  - $V = (v_1, \dots, v_p) \in \mathbb{R}^{n\times n}$ is orthogonal
  - $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m\times n}$, $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$
  - σ_i are called the singular values, u_i are the left singular vectors, and v_i are the right singular vectors.
- Thin (compact) SVD: assume $m \geq n$, $A = U\Sigma V'$, where
  - $U \in \mathbb{R}^{m\times n}$, $U'U = I_n$
  - $V \in \mathbb{R}^{n\times n}$, $V'V = I_n$

  - $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{m \times n}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$

- Relation to eigen-decomposition. Using thin SVD:

$$A'A = V\Sigma UU'\Sigma V' = V\Sigma^2 V'$$

$$AA' = U\Sigma VV'\Sigma U' = U\Sigma^2 U'$$

- Application: Principal component analysis (dimension deduction)
  - Principal components are eigenvectors of the covariance matrix (example: `https://arxiv.org/pdf/1708.00491.pdf`)