Another look at subset selection using linear least squares

Student: Mădălina-Alina Racoviță

Special Chapters of Artificial Intelligence Presentation
Master of Computational Optimization
Faculty of Computer Science, UAIC, Iasi

Today's Presentation Agenda

Small recap

Some numerical calculus and regression concepts: Linear Least Squares, Subset Selection, orthogonal matrixes, conditionning number of a matrix, QR factorization

Usages of QR factorization

PCA, SVD, selecting subsets of regression variables by least squares methods.

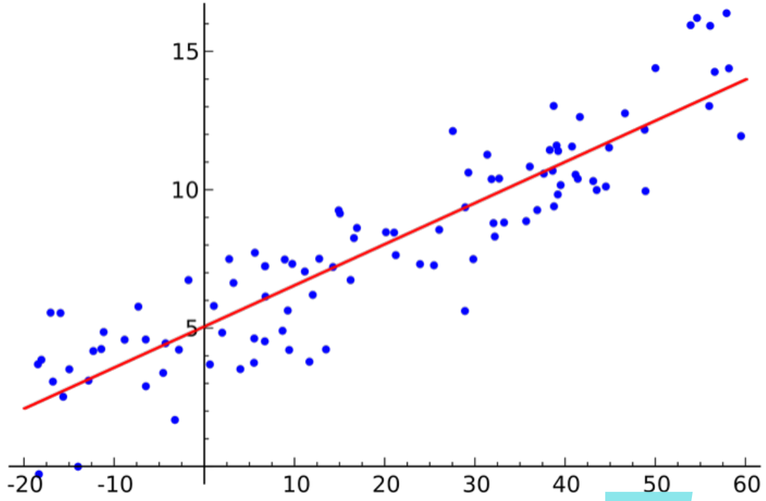
Othogonal projections and their properties

Appendix. An estimator for squared projections

Q&A

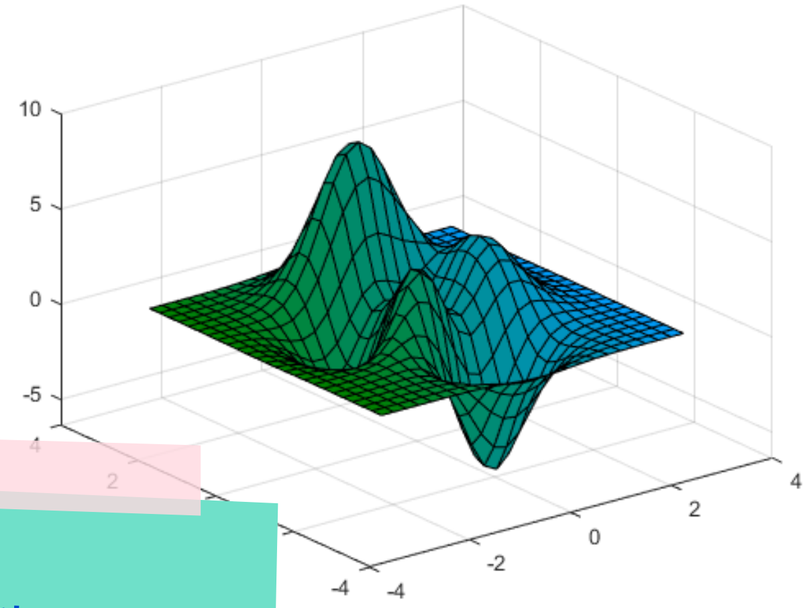
Small recap

1. **Linear Least Squares** - simple idea
2. **Subset selection** - usage
3. What is an **orthogonal matrix**?
4. **Condition number of a matrix.** What is an ill-conditioned matrix?
5. **QR factorization**



Linear Least Squares What is its simple idea?

If we have a simple simple system of equations, noted with AX = b, if the system is compatible then the equation is clear. But if it is not, then we would like to find the closest one to the real solution. This is the reason why we proceed the Least Squares method.



Subset selection

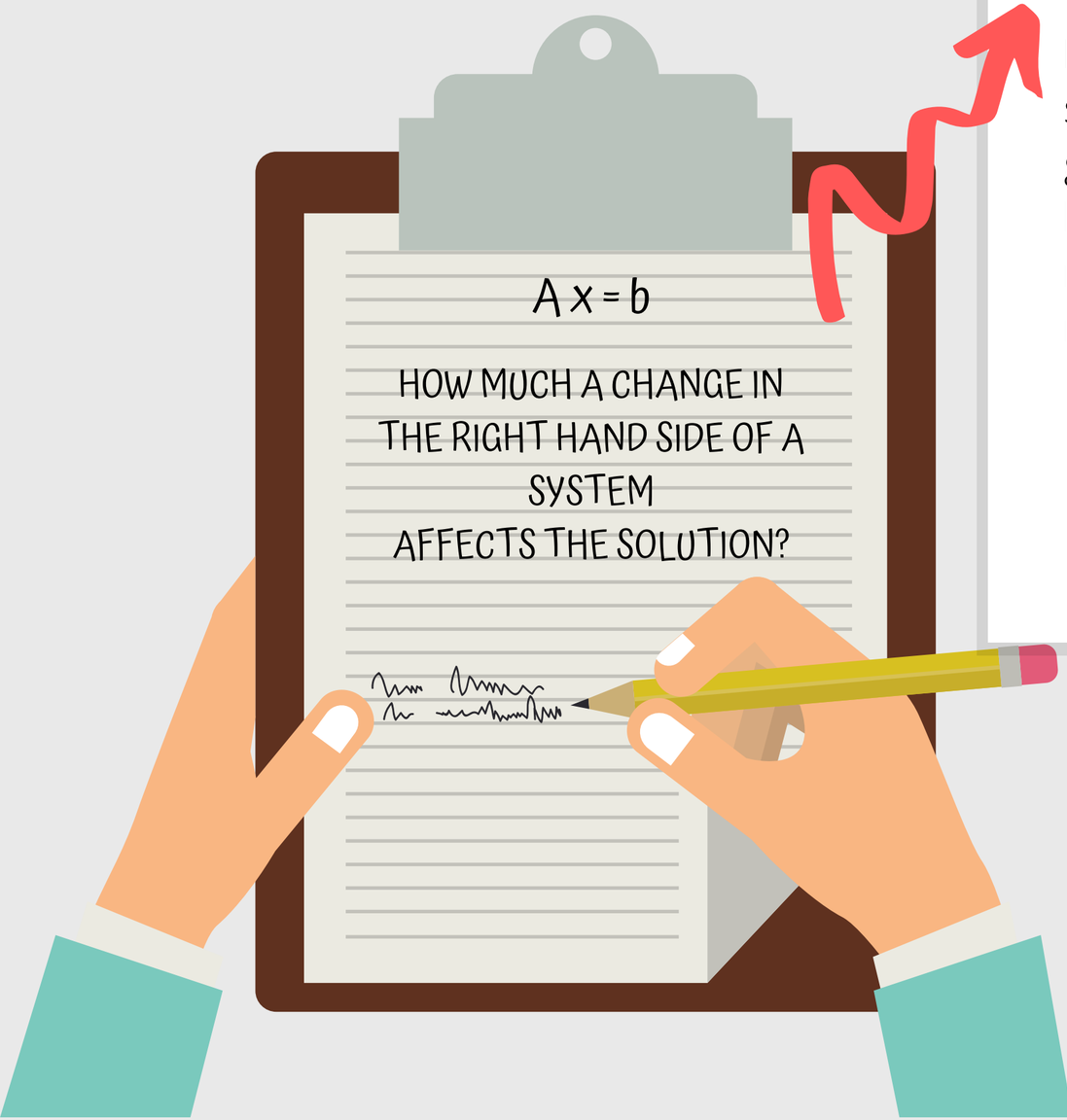
The subset selection's principal purpose is to reduce the number dimensions of a linear model.
"Curse of dimensionality"

Orthogonal matrix

$$A^{-1} = A^T$$

$$A^T A = A A^T = I$$

Condition number of a matrix, what does it mean?



$\delta A \rightarrow \delta X$ or $\delta A \rightarrow \delta b$

If the change in A implies a big change in the solution system then the condition number of the matrix A is going to have a large value. In this case the matrix will be **ill-conditioned**. If not it will be close to 1 and the matrix will be **well-conditioned**. The condition number can be computed like this:

$$\kappa(A) = \|A^{-1}\| \, \|A\| \geq \|A^{-1} A\| = 1$$

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|,$$

Algorithms that computes the QR decomposition are the following:

- **Givens algorithm**. It uses **rotation matrixes** whose columns form an orthonormal basis. This algorithm can be used for finding also the eigenvalues of a matrix.
- **Gram-Schmidt algorithm** computes the QR decomposition on columns and it is an **algorithm of orthonormalisation of a basis**
- **Householder algorithm** uses **reflexion matrixes**

$$R_y(\theta) = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix}$$

$$P = I_n - 2vv^T, \quad v \in \mathbb{R}^n, \quad \|v\|_2 = \sqrt{\sum_{j=1}^{n} |v_j|^2} = 1$$

QR decomposition

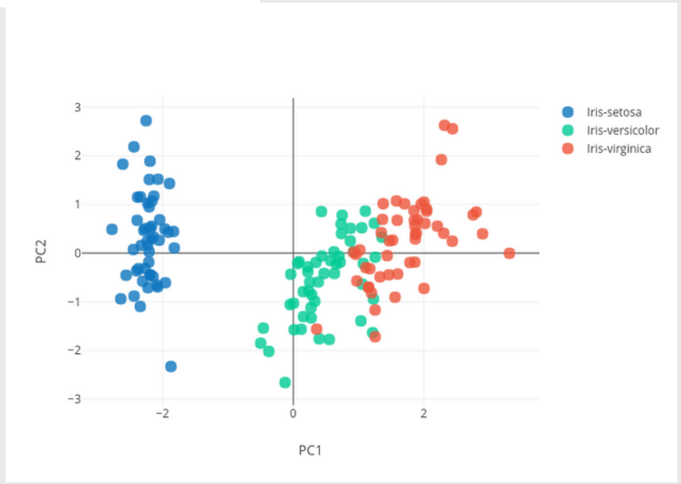
In linear algebra, a **QR decomposition**, also known as a **QR factorization** is a decomposition of a matrix A into a product **A = QR** of an **orthogonal matrix Q** and an **upper triangular matrix R**.

$$\begin{bmatrix} 2 & 3 \\ 2 & 4 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \star & \star \\ \star & \star \\ \star & \star \end{bmatrix} \begin{bmatrix} \star & \star \\ 0 & \star \end{bmatrix}$$

Usages of QR factorization:

$$A^T A\mathbf{x} = A^T\mathbf{b}$$

$$\hat{\mathbf{x}} = R^{-1}Q^T\mathbf{b}.$$



- Solving linear systems of equations
- Principal components analysis
- Singular Value Decomposition
- Finding the eigenvalues of a matrix by applying QR algorithm

The article presents the fact that **by using QR factorisation in subset selection the bias can be introduced**, with the mentioning that the OLS estimator is an unbiased estimator.

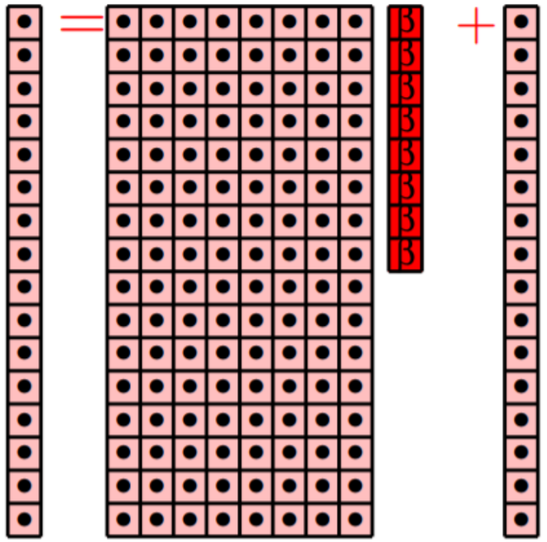
Orthogonal projections and their properties

Full regression model

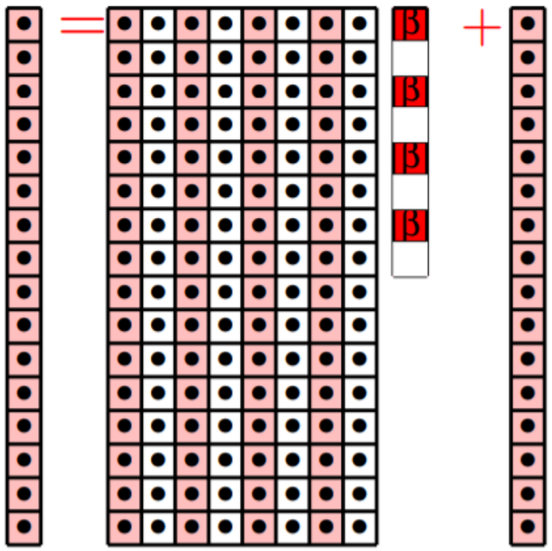
Subset selection

QR decomposition

$$y = A\beta + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I_m)$$



$$y = A_* \beta_* + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I_m)$$



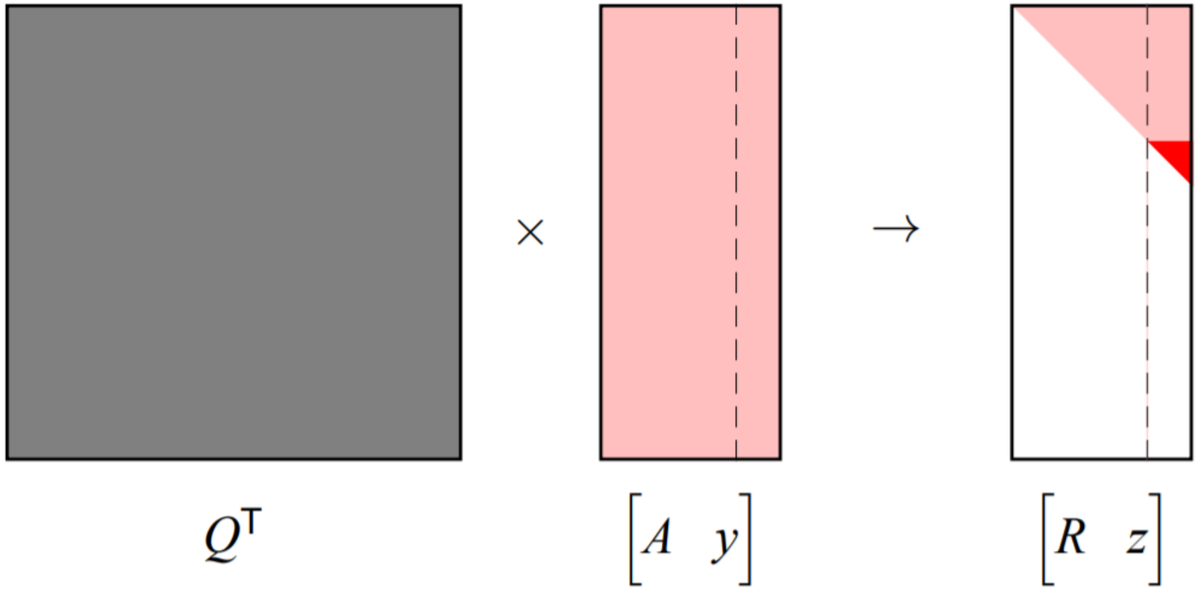
- ☑ QR decomposition of A:

$$Q^T A = \begin{pmatrix} R \\ 0 \end{pmatrix} \begin{matrix} n \\ m-n \end{matrix}$$

$$Q^T y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \begin{matrix} n \\ m-n \end{matrix}$$

- ☑ LS solution:

$$\hat{\beta} = R^{-1} y_1 \quad \text{RSS} = y_2^T y_2$$



$$\hat{\beta} = R^{-1}z$$

$$\text{RSS}(\hat{\beta}) = \rho^2$$

In the article is presented the idea that by using the QR decomposition we can solve the system of equations like this:

$$A^T A\mathbf{x} = A^T\mathbf{b}$$

where **b** is represented by the **y** vector and **x** is out vector of regression coefficients, i.e. **beta.**

$$\beta = R^{-1}Q^T y$$

$$Q^{-1}y = t$$

Properties of QR decomposition

1. The matrix **R** is the **Cholesky factor** for AA^T where the Cholesky decomposition is the decomposition of symmetric and positively defined matrixes $A = LL^T$

$$AA^T = R^T R$$

2. The first **p rows and columns of R** are the **Cholesky factor for the first p rows and columns** of AA^T
This means that for the regression of Y agains the first p columns of X, we can use the same calculations as for the full regression but just 'forget' the last p rown and columns.

3. The order of the variables in the Cholesky factor, R, can easily be changed to derive regressions of Y upon other subsets of X-variables.

4. If the data satisfies $y = A\beta + \epsilon$, then $t \sim \mathcal{N}(\mu, \sigma^2)$

5.If only the p elements of beta are non-zero, then the expected values of the last (n-p) elements of t is to be zero.

Appendix. An estimator for squared projections

In the appendix of the article is presented an estimator for μ^2 , given a $x \sim N(\mu = ?, \sigma^2)$

$0 \leq C \leq 1, C \in \mathbb{R}$ where C have to minimize the following squared error loss:

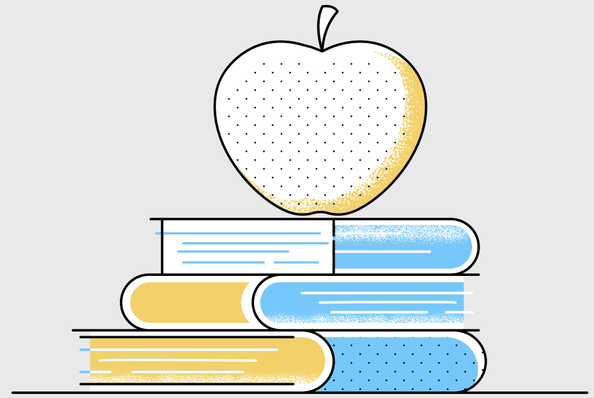
$$E(Cx^2 - \mu^2)^2 = C^2(\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4) - 2C\mu^2(\mu^2 + \sigma^2) + \mu^4$$

(this is a formula resulted from the normal distribution))

$$\frac{\partial E(Cx^2 - \mu^2)}{\partial C} = C(\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4) - 2\mu^2(\mu^2 + \sigma^2)$$

min when

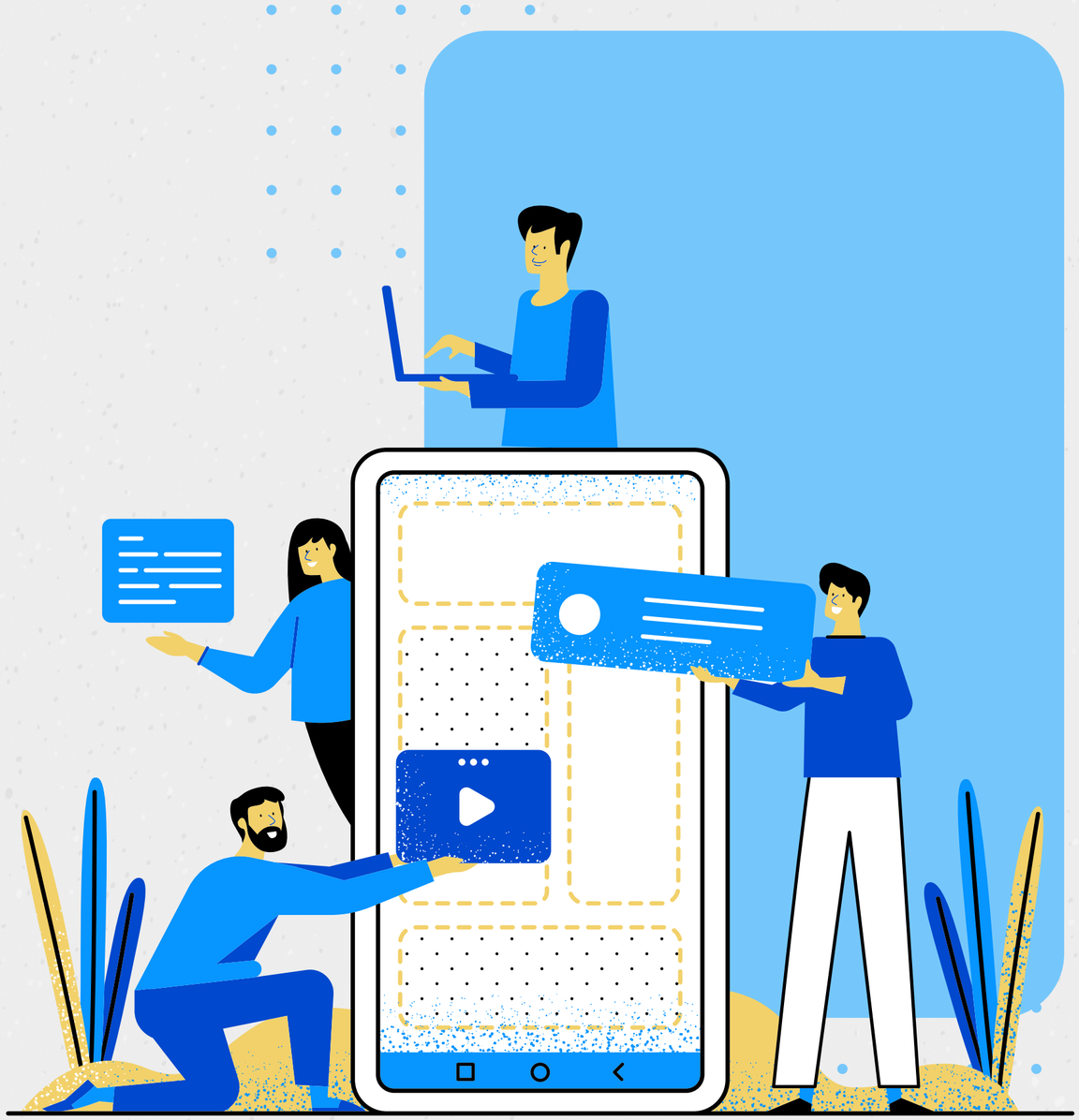
$$C = \frac{(\frac{\mu^2}{\sigma^2})^2 + \frac{\mu^2}{\sigma^2}}{(\frac{\mu^2}{\sigma^2})^2 + 6\frac{\mu^2}{\sigma^2} + 3} \xrightarrow[\mu = ?]{} \mu^2 = x^2 \frac{(\frac{x^2}{\sigma^2})^2 + \frac{x^2}{\sigma^2}}{(\frac{x^2}{\sigma^2})^2 + 6\frac{x^2}{\sigma^2} + 3}$$



BIBLIOGRAPHY

1. https://profs.info.uaic.ro/~cgatu/csia/res/lecture-07.pdf
2. https://profs.info.uaic.ro/~cgatu/csia/research/model_selection/56_Miller_CommStatTheoryMeth_00.PDF
3. https://profs.info.uaic.ro/~ancai/CN/curs/CN%20-%20curs%2005%20-%202019.pdf
4. https://en.wikipedia.org/wiki/QR_decomposition
5. https://en.wikipedia.org/wiki/Condition_number

Q&A



Thank you!