# Machine Learning for Finance (FIN 570)
## Singular Value Decomposition (SVD)
## and Principal Component Analysis (PCA)

Instructor: Jaehyuk Choi

Peking University HSBC Business School, Shenzhen, China

2018-19 Module 1 (Fall 2018)

# Eigen(spectral) decomposition

For a matrix $A$, eigenvalue $\lambda_k$ and eigenvector $v_k$ satisfy

$$Av_k = \lambda_k v_k.$$

The matrix $A$ can be decomposed into

$$A = Q\Lambda Q^{-1},$$

where $\Lambda$ is a diagonal matrix with values $\lambda_k$ and $Q = (v_1 \cdots v_n)$, i.e., $Q_{*j} = v_j$.
When $A$ is real and symmetric, $Q$ is an orthonormal matrix, $QQ^T = I$,

$$A = Q\Lambda Q^T,$$

# Singular Value Decomposition (SVD)

The single most useful practical concept in linear algebra:

- Any matrix (even rectangular) has a SVD.
- SVD tells everything on a matrix.

For any $m \times n$ matrix $A$, there is a unique decomposition:
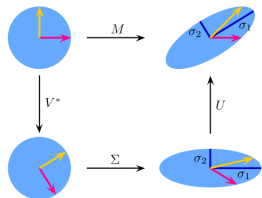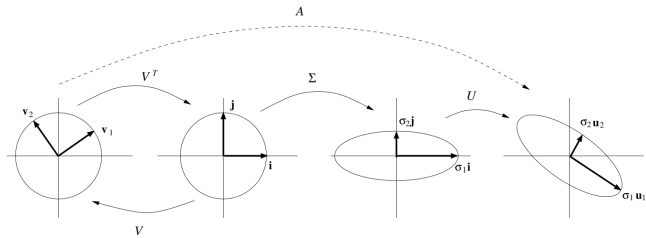
$$A = USV^T, \quad \text{where}$$

- $U$ ($m \times m$): orthonormal ($UU^T = U^TU = I$)
- $S$ ($m \times n$): diagonal. Singular values, $s_k \geq 0$, are in decreasing order for $1 \leq k \leq \min(m, n)$
- $V$ ($n \times n$): orthonormal ($VV^T = V^TV = I$)



A = U Σ V^T

m x n    m x m    m x n    n x n
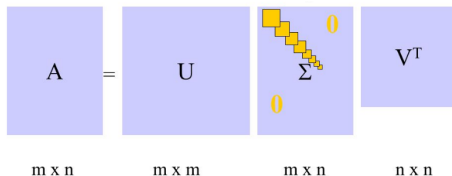
# SVD: Intuition

Linear transformation $A$ is decomposed into
- a rotation by $V^T$
- a scaling by $S$
- a rotation by $U$



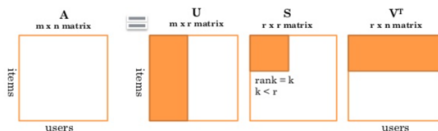$$M = U \cdot \Sigma \cdot V^*$$

# SVD: Compact Form, Low Rank Approximation



$A = U \times S \times V^T$



$A_k = U_k \times S_k \times V_k^T$

- For a non-square matrix, a compact form is enough:
  $U$ $(m \times r)$, $S$ $(r \times r)$, $V$ $(n \times r)$ where $r = \min(m, n)$.
- If the rank is $k$ $(\leq r)$, $s_{j>k} = 0$:
  $U$ $(m \times k)$, $S$ $(k \times k)$, $V$ $(n \times k)$
- Using the first $j$ $(< k)$ biggest singular values,

# SVD: Image Compression

An image file is nothing but a matrix, so the low-rank approximation of SVD works as an image compression method. The storage is reduced from $mn$ to $(m + n + 1)k$.
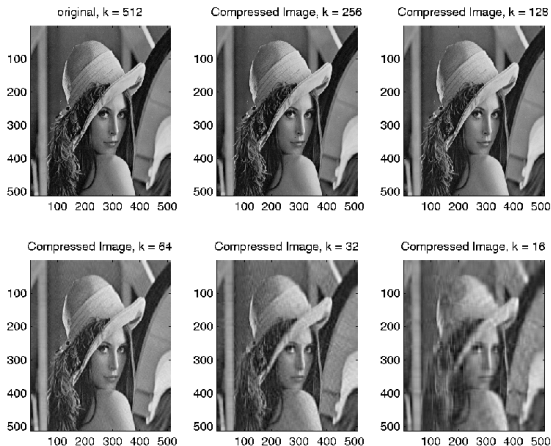
# Principal Component Analysis (PCA)

If $\boldsymbol{X}$ is a matrix of $n$ samples of $p$ features ($n \times p$), the covariance matrix is

$$\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} : (p \times p) \text{ symmetric matrix}$$

The covariance matrix of the transformed space $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{W}$ is

$$\mathsf{Cov}(\boldsymbol{Z}) = \frac{1}{n}(\boldsymbol{X}\boldsymbol{W})^T(\boldsymbol{X}\boldsymbol{W}) = \frac{1}{n}\boldsymbol{W}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{W} = \boldsymbol{W}^T\Sigma\boldsymbol{W}$$

If we pick $\boldsymbol{W}$ to be the orthogonal transformation of $SVD$, i.e., $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{S}\boldsymbol{W}^T$,

$$\mathsf{Cov}(\boldsymbol{Z}) = \boldsymbol{S} = \mathsf{diag}(S_{11}, \cdots, S_{pp}).$$

Notice that $\mathsf{Cov}(Z_i, Z_j) = \boldsymbol{W}_{*i}^T\boldsymbol{\Sigma}\boldsymbol{W}_{*j} = S_{ij}$ is zero if $i \neq j$, so the extracted features are orthogonal.

# Process of finding $W$

Let $W = (W_{*1} \ W_{*2} \ \cdots W_{*p})$.

- Find $W_{*1}$ such that $|W_{*1}| = 1$ and $|W_{*1}^T \Sigma W_{*1}|$ is maximized.
- Find $W_{*2}$ such that $|W_{*2}| = 1$, $|W_{*2}^T \Sigma W_{*2}|$ is maximized and $W_{*1}^T W_{*2} = 0$.
- $\cdots$
- Find $W_{*k}$ such that $|W_{*k}| = 1$, $|W_{*k}^T \Sigma W_{*k}|$ is maximized and $W_{*k}$ is orthogonal to $\{W_{*j}\}$ for $j < k$.
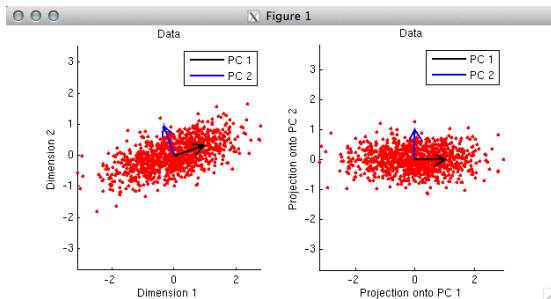
# Total and Explained Variance

The total variance is the variance of all original features. Under PCA,

$$\sum_{k=1}^{p} \mathsf{Var}(X_k) = \sum_{k=1}^{p} S_{kk}.$$

Therefore the ratio

$$\frac{\sum_{j=1}^{k} S_{jj}}{\sum_{j=1}^{p} S_{jj}}$$

indicates how much of the total variance is *explained* by the first $k$ PCA factors. Extracting features from PCA is an unsupervised learning, NOT supervised learning, because the response variable is not associated.

# PCA vs Simple Linear Regression for $(x, y)$

PCA is not same as Simple Linear regression (OLS)!

- **Linear Regression** minimize the the (squared) distance in $y$-axis.
- **PCA** (1st component) minimize the (squared) shortest distance.