

Singular Value Decomposition (SVD) and Principal Value Analysis (PCA)

Jaehyuk Choi

March 13, 2017

Eigen(spectral) decomposition

For a matrix \mathbf{A} , eigenvalue λ_k and eigenvector \mathbf{v}_k satisfy

$$\mathbf{A}\mathbf{v}_k = \lambda_k \mathbf{v}_k.$$

The matrix \mathbf{A} can be decomposed into

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

where $\mathbf{\Lambda}$ is a diagonal matrix with values λ_k and $\mathbf{Q} = (\mathbf{v}_1 \cdots \mathbf{v}_n)$, i.e., $\mathbf{Q}_{*j} = \mathbf{v}_j$.
When \mathbf{A} is real and symmetric, \mathbf{Q} is an orthogonal matrix, $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$,

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T,$$

Singular Value Decomposition (SVD)

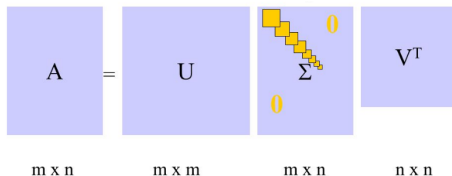
The single most useful practical concept in linear algebra:

- 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:

$$A = USV^T, \quad \text{where}$$

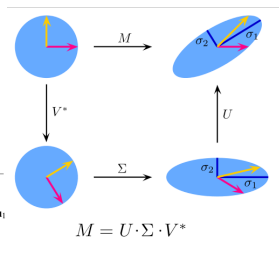
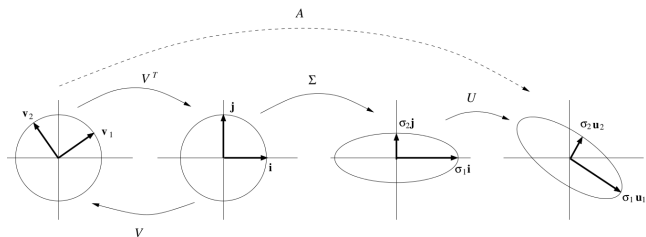
- U ($m \times m$): orthogonal ($UU^T = U^T U = I$)
- S ($m \times n$): diagonal. Singular values, $s_k \geq 0$, are in decreasing order for $1 \leq k \leq \min(m, n)$
- V ($n \times n$): orthogonal ($VV^T = V^T V = I$)



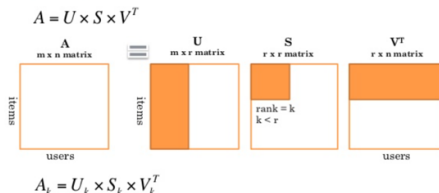
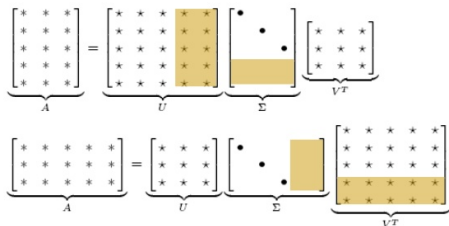
SVD: Intuition

Linear transformation A is decomposed into

- a rotation by V^T
- a scaling by S
- a rotation by U



SVD: Compact Form, Low Rank Approximation



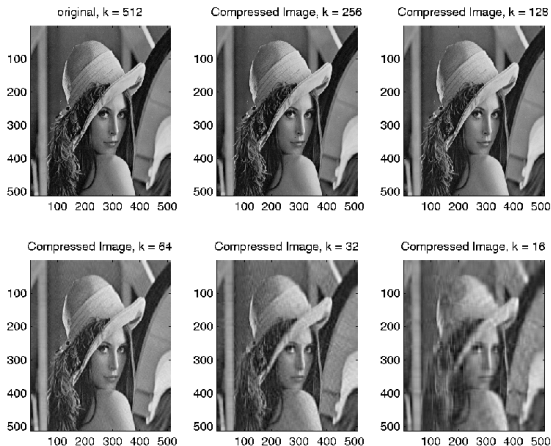
- For a non-square matrix, a compact form is enough:
 $U (m \times r)$, $S (r \times r)$, $V (n \times r)$ where $r = \min(m, n)$.
- If the rank is $k (\leq r)$, $s_{j>k} = 0$:
 $U (m \times k)$, $S (k \times k)$, $V (n \times k)$
- Using the first $j (\leq k)$ biggest singular values,

$$A_j = U_j S_j V_j^T = \sum_{i=1}^j \mathbf{u}_i s_i \mathbf{v}_i^T, \quad U_j (m \times j), \quad S_j (j \times j), \quad V_j (n \times j)$$

is the best approximation with rank j minimizing the norm $\|A - A_j\|_F$

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

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

The covariance matrix of the transformed space $\mathbf{Z} = \mathbf{XW}$ is

$$\text{Cov}(\mathbf{Z}) = \frac{1}{n} (\mathbf{XW})^T (\mathbf{XW}) = \frac{1}{n} \mathbf{W}^T (\mathbf{X}^T \mathbf{X}) \mathbf{W} = \mathbf{W}^T \mathbf{\Sigma} \mathbf{W}$$

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

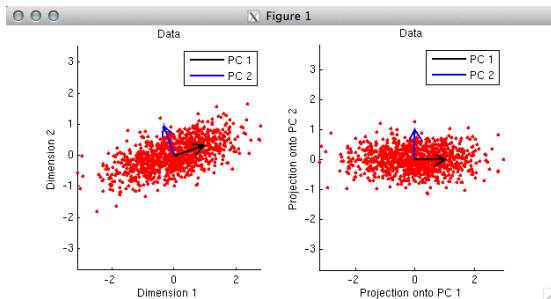
$$\text{Cov}(\mathbf{Z}) = \mathbf{S} = \text{diag}(S_{11}, \dots, S_{pp}).$$

Notice that $\text{Cov}(Z_i, Z_j) = \mathbf{W}_{*i}^T \mathbf{\Sigma} \mathbf{W}_{*j} = S_{ij}$ 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 $W_{*1}^T W_{*2} = 0$.
- ...
- 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,

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

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.