کاهش ابعادی

Dimensionality Reduction

a.fahim@ut.ac.ir

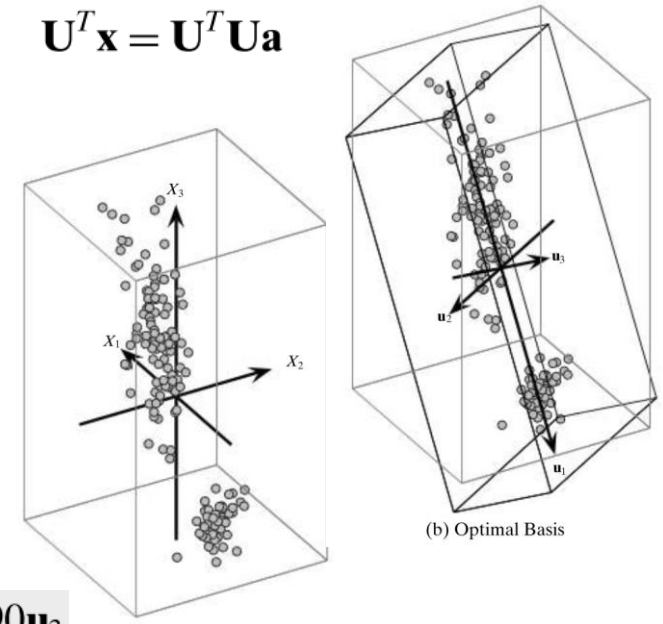
$$\mathbf{x} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_d\mathbf{u}_d \qquad \mathbf{x} = \mathbf{U}\mathbf{a}$$

Orthonormal

$$\mathbf{u}_i^T\mathbf{u}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \qquad \mathbf{U} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix}$$

$$\mathbf{U}^{-1} = \mathbf{U}^T \qquad \mathbf{U}^T\mathbf{U} = \mathbf{I} \qquad \mathbf{U}^T\mathbf{x} = \mathbf{U}^T\mathbf{U}\mathbf{a}$$

$$\mathbf{a} = \mathbf{U}^T\mathbf{x}$$



(a) Original Basis (b) Optimal Basis

Example 7.1.

$$\mathbf{x} = (-0.343, -0.754, 0.241)^T$$

$$\mathbf{x} = -0.154\mathbf{u}_1 + 0.828\mathbf{u}_2 - 0.190\mathbf{u}_3$$

$$\mathbf{a} = \mathbf{U}^T\mathbf{x} = \begin{pmatrix} -0.390 & 0.089 & -0.916 \\ -0.639 & -0.742 & 0.200 \\ -0.663 & 0.664 & 0.346 \end{pmatrix} \begin{pmatrix} -0.343 \\ -0.754 \\ 0.241 \end{pmatrix} = \begin{pmatrix} -0.154 \\ 0.828 \\ -0.190 \end{pmatrix}$$

It is natural to ask whether we can find a **reduced dimensionality** subspace that still **preserves the essential characteristics** of the data. That is, we are interested in finding the optimal r-dimensional representation of D, with r ≪ d.

$$\mathbf{x}' = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_r\mathbf{u}_r = \sum_{i=1}^{r} a_i\mathbf{u}_i$$

$$\mathbf{x}' = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \\ | & | & & | \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_r \end{pmatrix} = \mathbf{U}_r\mathbf{a}_r \qquad \mathbf{a}_r = \mathbf{U}_r^T\mathbf{x}$$

$$\mathbf{x}' = \mathbf{U}_r\mathbf{U}_r^T\mathbf{x} = \mathbf{P}_r\mathbf{x}$$

$$\boldsymbol{\epsilon} = \sum_{i=r+1}^{d} a_i\mathbf{u}_i = \mathbf{x} - \mathbf{x}'$$

$$S_r = span(\mathbf{u}_1, \ldots, \mathbf{u}_r) \qquad S_{d-r} = span(\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d)$$

are *orthogonal subspaces*, that is, all pairs of vectors $\mathbf{x} \in S_r$ and $\mathbf{y} \in S_{d-r}$ must be orthogonal. The subspace S_{d-r} is also called the *orthogonal complement* of S_r.

Principal Component Analysis (PCA) is a technique that seeks a r-dimensional basis **that best captures the variance** in the data. The direction with **the largest projected variance** is called **the first principal component**. The **orthogonal direction** that captures the **second largest** projected **variance** is called the second principal component, and so on. As we shall see, the direction that maximizes the variance is also the one that **minimizes the mean squared error**.

7.2.3 Best *r*-dimensional Approximation

The projected variance along v is given as $\sigma_{\mathbf{v}}^2 = \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}$

$$\max_{\mathbf{u}} J(\mathbf{u}) = \mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} - \alpha(\mathbf{u}^T \mathbf{u} - 1)$$

$$\boldsymbol{\Sigma}\mathbf{u} = \alpha\mathbf{u}$$

$$\sigma_{\mathbf{u}}^2 = \alpha = \lambda_1$$

$$MSE(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\epsilon}_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|\bar{\mathbf{x}}_i - \mathbf{x}_i'\|^2$$

$$MSE = \sum_{i=1}^{n} \frac{\|\bar{\mathbf{x}}_i\|^2}{n} - \mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u}$$

$$MSE(\mathbf{u}) = \text{var}(\mathbf{D}) - \mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} = \sum_{i=1}^{d} \sigma_i^2 - \mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u}$$

Thus, the principal component u_1, which is the direction that **maximizes the projected variance**, is also the direction that **minimizes the mean squared error**.

The already computed eigenvectors, $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{j-1}$, corresponding to the j − 1 largest eigenvalues of Σ, for 1 ≤ j ≤ r

$$\max_{\mathbf{v}} J(\mathbf{v}) = \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} - \alpha(\mathbf{v}^T \mathbf{v} - 1) - \sum_{i=1}^{j-1} \beta_i (\mathbf{u}_i^T \mathbf{v} - 0)$$

We find that $\beta_i = 0$ for all $i < j$ which implies that $\boldsymbol{\Sigma}\boldsymbol{v} = \alpha\boldsymbol{v}$
To maximize the variance along $\boldsymbol{v}$, we set $\alpha = \lambda_j$, the jth largest eigenvalue of $\boldsymbol{\Sigma}$, with $\boldsymbol{v} = \boldsymbol{u}_j$ giving the jth principal component.

Because $\boldsymbol{\Sigma}$ is positive semidefinite, its eigenvalues must all be non-negative, and we can thus sort them in decreasing order as follows:

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r \geq \lambda_{r+1} \cdots \geq \lambda_d \geq 0$$

We then select the r largest eigenvalues, and their corresponding eigenvectors to form the best r-dimensional approximation.

Total Projected Variance

$$\mathbf{U}_r = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \\ | & | & & | \end{pmatrix} \qquad \mathbf{P}_r = \mathbf{U}_r \mathbf{U}_r^T = \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i^T$$

$$\mathbf{x}'_i = \mathbf{P}_r \bar{\mathbf{x}}_i \qquad \mathbf{a}_i = \mathbf{U}_r^T \bar{\mathbf{x}}_i$$

The new Data Matrix

$$\mathbf{A} \in \mathbb{R}^{n \times r}$$

$$\text{var}(\mathbf{A}) = \frac{1}{n} \sum_{i=1}^{n} \bar{\mathbf{x}}_i^T \mathbf{P}_r \bar{\mathbf{x}}_i = \sum_{i=1}^{r} \mathbf{u}_i^T \mathbf{\Sigma} \mathbf{u}_i = \sum_{i=1}^{r} \lambda_i$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \|\bar{\mathbf{x}}_i - \mathbf{x}'_i\|^2 = \text{var}(\mathbf{D}) - \text{var}(\mathbf{A})$$

$$= \text{var}(\mathbf{D}) - \sum_{i=1}^{r} \mathbf{u}_i^T \mathbf{\Sigma} \mathbf{u}_i = \text{var}(\mathbf{D}) - \sum_{i=1}^{r} \lambda_i$$

$$\text{var}(\mathbf{D}) = \sum_{i=1}^{d} \sigma_i^2 = \sum_{i=1}^{d} \lambda_i$$

Choosing the Dimensionality

$$f(r) = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_r}{\lambda_1 + \lambda_2 + \cdots + \lambda_d} = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i} = \frac{\sum_{i=1}^{r} \lambda_i}{\text{var}(\mathbf{D})}$$

$$r = \min\{r' | f(r') \geq \alpha\}$$

In practice, α is usually set to 0.9 or higher, so that the reduced dataset captures at least 90% of the total variance.

Algorithm 7.1: Principal Component Analysis

PCA $(\mathbf{D}, \alpha)$:

1 $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ // compute mean

2 $\overline{\mathbf{D}} = \mathbf{D} - \mathbf{1}\cdot\boldsymbol{\mu}^T$ // center the data

3 $\boldsymbol{\Sigma} = \frac{1}{n}\left(\overline{\mathbf{D}}^T\,\overline{\mathbf{D}}\right)$ // compute covariance matrix

4 $(\lambda_1, \lambda_2, \ldots, \lambda_d) = \text{eigenvalues}(\boldsymbol{\Sigma})$ // compute eigenvalues

5 $\mathbf{U} = \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d\end{pmatrix} = \text{eigenvectors}(\boldsymbol{\Sigma})$ // compute eigenvectors

6 $f(r) = \frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{d}\lambda_i}$, for all $r = 1, 2, \ldots, d$ // fraction of total variance

7 Choose smallest r so that $f(r) \geq \alpha$ // choose dimensionality

8 $\mathbf{U}_r = \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r\end{pmatrix}$ // reduced basis

9 $\mathbf{A} = \{\mathbf{a}_i \mid \mathbf{a}_i = \mathbf{U}_r^T\bar{\mathbf{x}}_i, \text{for } i = 1, \ldots, n\}$ // reduced dimensionality data

Example 7.5. Given the 3-dimensional Iris dataset in Figure 7.1(a), its covariance matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.681 & -0.039 & 1.265 \\ -0.039 & 0.187 & -0.320 \\ 1.265 & -0.32 & 3.092 \end{pmatrix}$$

The eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ are given as

$$\lambda_1 = 3.662 \qquad \lambda_2 = 0.239 \qquad \lambda_3 = 0.059$$

$$\mathbf{u}_1 = \begin{pmatrix} -0.390 \\ 0.089 \\ -0.916 \end{pmatrix} \qquad \mathbf{u}_2 = \begin{pmatrix} -0.639 \\ -0.742 \\ 0.200 \end{pmatrix} \qquad \mathbf{u}_3 = \begin{pmatrix} -0.663 \\ 0.664 \\ 0.346 \end{pmatrix}$$

The total variance is therefore $\lambda_1 + \lambda_2 + \lambda_3 = 3.662 + 0.239 + 0.059 = 3.96$. The optimal 3-dimensional basis is shown in Figure 7.1(b).

To find a lower dimensional approximation, let $\alpha = 0.95$. The fraction of total variance for different values of r is given as

r	1	2	3
$f(r)$	0.925	0.985	1.0

For example, for $r = 1$, the fraction of total variance is given as $f(1) = \frac{3.662}{3.96} = 0.925$. Thus, we need at least $r = 2$ dimensions to capture 95% of the total variance. This optimal 2-dimensional subspace is shown as the shaded plane in Figure 7.3(a). The reduced dimensionality dataset $\mathbf{A}$ is shown in Figure 7.4. It consists of the point coordinates $\mathbf{a}_i = \mathbf{U}_2^T\bar{\mathbf{x}}_i$ in the new 2-dimensional principal components basis comprising $\mathbf{u}_1$ and $\mathbf{u}_2$.

Let φ correspond to a mapping from the **input space** to the **feature space**.

$$\boldsymbol{\Sigma}_\phi = \frac{1}{n}\sum_{i=1}^{n}\left(\phi(\mathbf{x}_i)-\boldsymbol{\mu}_\phi\right)\left(\phi(\mathbf{x}_i)-\boldsymbol{\mu}_\phi\right)^T = \frac{1}{n}\sum_{i=1}^{n}\bar{\phi}(\mathbf{x}_i)\bar{\phi}(\mathbf{x}_i)^T$$

$$\left(\frac{1}{n}\sum_{i=1}^{n}\bar{\phi}(\mathbf{x}_i)\bar{\phi}(\mathbf{x}_i)^T\right)\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \qquad \sum_{i=1}^{n} c_i\,\bar{\phi}(\mathbf{x}_i) = \mathbf{u}_1$$

where $c_i = \frac{\bar{\phi}(\mathbf{x}_i)^T\mathbf{u}_1}{n\lambda_1}$ is a scalar value.

$$\begin{aligned}
\bar{K}(\mathbf{x}_i,\mathbf{x}_j) &= \bar{\phi}(\mathbf{x}_i)^T\bar{\phi}(\mathbf{x}_j)\\
&= (\phi(\mathbf{x}_i)-\boldsymbol{\mu}_\phi)^T(\phi(\mathbf{x}_j)-\boldsymbol{\mu}_\phi)\\
&= \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) - \phi(\mathbf{x}_i)^T\boldsymbol{\mu}_\phi - \phi(\mathbf{x}_j)^T\boldsymbol{\mu}_\phi + \boldsymbol{\mu}_\phi^T\boldsymbol{\mu}_\phi\\
&= K(\mathbf{x}_i,\mathbf{x}_j) - \frac{1}{n}\sum_{k=1}^{n}\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_k) - \frac{1}{n}\sum_{k=1}^{n}\phi(\mathbf{x}_j)^T\phi(\mathbf{x}_k) + \|\boldsymbol{\mu}_\phi\|^2\\
&= K(\mathbf{x}_i,\mathbf{x}_j) - \frac{1}{n}\sum_{k=1}^{n}K(\mathbf{x}_i,\mathbf{x}_k) - \frac{1}{n}\sum_{k=1}^{n}K(\mathbf{x}_j,\mathbf{x}_k) + \frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n}K(\mathbf{x}_a,\mathbf{x}_b)
\end{aligned}$$

$$\bar{\mathbf{K}} = \mathbf{K} - \frac{1}{n}\mathbf{1}_{n\times n}\mathbf{K} - \frac{1}{n}\mathbf{K}\mathbf{1}_{n\times n} + \frac{1}{n^2}\mathbf{1}_{n\times n}\mathbf{K}\mathbf{1}_{n\times n} = \left(\mathbf{I}-\frac{1}{n}\mathbf{1}_{n\times n}\right)\mathbf{K}\left(\mathbf{I}-\frac{1}{n}\mathbf{1}_{n\times n}\right)$$

$$\left(\frac{1}{n}\sum_{i=1}^{n}\bar{\phi}(\mathbf{x}_i)\bar{\phi}(\mathbf{x}_i)^T\right)\left(\sum_{j=1}^{n}c_j\,\bar{\phi}(\mathbf{x}_j)\right) = \lambda_1\sum_{i=1}^{n}c_i\,\bar{\phi}(\mathbf{x}_i)$$

$$\sum_{i=1}^{n}\left(\bar{\phi}(\mathbf{x}_i)\sum_{j=1}^{n}c_j\,\bar{K}(\mathbf{x}_i,\mathbf{x}_j)\right) = n\lambda_1\sum_{i=1}^{n}c_i\,\bar{\phi}(\mathbf{x}_i)$$

Algorithm 7.2: Kernel Principal Component Analysis

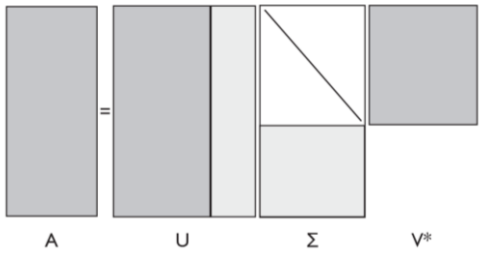
KERNELPCA $(\mathbf{D}, K, \alpha)$**:**

1. $\mathbf{K} = \left\{K(\mathbf{x}_i,\mathbf{x}_j)\right\}_{i,j=1,\ldots,n}$ // compute $n\times n$ kernel matrix
2. $\bar{\mathbf{K}} = (\mathbf{I}-\frac{1}{n}\mathbf{1}_{n\times n})\mathbf{K}(\mathbf{I}-\frac{1}{n}\mathbf{1}_{n\times n})$ // center the kernel matrix
3. $(\eta_1,\eta_2,\ldots,\eta_n) = \text{eigenvalues}(\bar{\mathbf{K}})$ // compute eigenvalues
4. $\begin{pmatrix}\mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n\end{pmatrix} = \text{eigenvectors}(\bar{\mathbf{K}})$ // compute eigenvectors
5. $\lambda_i = \frac{\eta_i}{n}$ for all $i = 1,\ldots,n$ // compute variance for each component
6. $\mathbf{c}_i = \sqrt{\frac{1}{\eta_i}}\cdot\mathbf{c}_i$ for all $i = 1,\ldots,n$ // ensure that $\mathbf{u}_i^T\mathbf{u}_i = 1$
7. $f(r) = \frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{d}\lambda_i}$, for all $r = 1,2,\ldots,d$ // fraction of total variance
8. Choose smallest r so that $f(r)\geq\alpha$ // choose dimensionality
9. $\mathbf{C}_r = \begin{pmatrix}\mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_r\end{pmatrix}$ // reduced basis
10. $\mathbf{A} = \{\mathbf{a}_i \mid \mathbf{a}_i = \mathbf{C}_r^T\,\bar{\mathbf{K}}_i, \text{for } i = 1,\ldots,n\}$ // reduced dimensionality data

7.4 Singular Value Decomposition

Principal components analysis is a special case of a more general matrix decomposition method called Singular Value Decomposition (SVD).

$$\mathbf{D} = \mathbf{L}\boldsymbol{\Delta}\mathbf{R}^T$$



where **L** is a orthogonal $n \times n$ matrix, **R** is an orthogonal $d \times d$ matrix, and **Δ** is an $n \times d$ "diagonal" matrix. The columns of **L** are called the **left singular vectors**, and the columns of **R** (or rows of $\boldsymbol{R}^T$) are called the **right singular** vectors. The matrix **Δ** is defined as

$$\boldsymbol{\Delta}(i,j) = \begin{cases} \delta_i & \text{If } i = j \\ 0 & \text{If } i \neq j \end{cases}$$

If the rank of D is $r \leq \min(n, d)$, then there will be only r nonzero singular values, which we assume are ordered as follows:

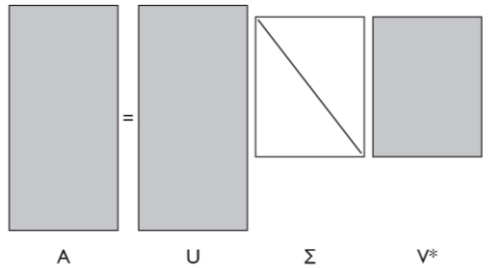
$$\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r > 0$$

One can **discard** those left and right singular vectors that **correspond to zero singular values**, to obtain the reduced SVD as

$$\mathbf{D} = \mathbf{L}_r \boldsymbol{\Delta}_r \mathbf{R}_r^T$$

$$= \begin{pmatrix} | & | & & | \\ \boldsymbol{l}_1 & \boldsymbol{l}_2 & \cdots & \boldsymbol{l}_r \\ | & | & & | \end{pmatrix} \begin{pmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_r \end{pmatrix} \begin{pmatrix} — & \mathbf{r}_1^T & — \\ — & \mathbf{r}_2^T & — \\ & \vdots & \\ — & \mathbf{r}_r^T & — \end{pmatrix}$$

$$= \delta_1 \boldsymbol{l}_1 \mathbf{r}_1^T + \delta_2 \boldsymbol{l}_2 \mathbf{r}_2^T + \cdots + \delta_r \boldsymbol{l}_r \mathbf{r}_r^T$$



In general, any $n \times d$ matrix **D** represents a *linear transformation*, $\mathbf{D}: \mathbb{R}^d \rightarrow \mathbb{R}^n$, from the space of d-dimensional vectors to the space of n-dimensional vectors because for any $\mathbf{x} \in \mathbb{R}^d$ there exists $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{Dx} = \mathbf{y}$

The set of all vectors $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{Dx} = \mathbf{y}$ over all possible $\mathbf{x} \in \mathbb{R}^d$ is called the *column space* of **D**, and the set of all vectors $\mathbf{x} \in \mathbb{R}^d$, such that $\mathbf{D}^T\mathbf{y} = \mathbf{x}$ over all $\mathbf{y} \in \mathbb{R}^n$, is called the *row space* of **D**, which is equivalent to the column space of $\mathbf{D}^T$. In other words, the column space of **D** is the set of all vectors that can be obtained as linear combinations of columns of **D**, and the row space of **D** is the set of all vectors that can be obtained as linear combinations of the rows of **D** (or columns of $\mathbf{D}^T$). Also note that the set of all vectors $\mathbf{x} \in \mathbb{R}^d$, such that $\mathbf{Dx} = \mathbf{0}$ is called the *null space* of **D**, and finally, the set of all vectors $\mathbf{y} \in \mathbb{R}^n$, such that $\mathbf{D}^T\mathbf{y} = \mathbf{0}$ is called the *left null space* of **D**.

If D has **rank r**, it means that it has **only r independent columns**, and also only **r independent rows**. Thus, the r left singular vectors $l_1, l_2, \ldots, l_r$ corresponding to the **r nonzero singular values** of D. represent a basis for the column space of D. The **remaining n−r** left singular vectors $l_{r+1}, \ldots, l_n$ represent a basis for **the left null space** of D. For the row space, the r right singular vectors $r_1, r_2, \ldots, r_r$ corresponding to the r non-zero singular values, represent a basis for the row space of D, and the remaining $d - r$ right singular vectors r_j $(j = r + 1, \ldots, d)$, represent a basis for the null space of D.

$$\mathbf{D}\mathbf{R}_r = \mathbf{L}_r \boldsymbol{\Delta}_r \mathbf{R}_r^T \mathbf{R}_r$$

$$\mathbf{D}\mathbf{R}_r = \mathbf{L}_r \boldsymbol{\Delta}_r$$

$$\mathbf{D}\mathbf{R}_r = \mathbf{L}_r \begin{pmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_r \end{pmatrix}$$

$$\mathbf{D}\begin{pmatrix} | & | & & | \\ \mathbf{r}_1 & \mathbf{r}_2 & \cdots & \mathbf{r}_r \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \delta_1 \boldsymbol{l}_1 & \delta_2 \boldsymbol{l}_2 & \cdots & \delta_r \boldsymbol{l}_r \\ | & | & & | \end{pmatrix}$$

As such, we can think of the SVD as a mapping from an orthonormal basis $(r_1, r_2, \ldots, r_r)$ in $\mathbb{R}^d$ (the row space) to an orthonormal basis $(l_1, l_2, \ldots, l_r)$ in $\mathbb{R}^n$ (the column space), with the corresponding axes scaled according to the singular values $\delta_1, \delta_2, \ldots, \delta_r$.

Connection between SVD and PCA

The scatter matrix for $\bar{\boldsymbol{D}}$

$$\begin{aligned} \bar{\mathbf{D}}^T\bar{\mathbf{D}} &= \left(\mathbf{L}\boldsymbol{\Delta}\mathbf{R}^T\right)^T\left(\mathbf{L}\boldsymbol{\Delta}\mathbf{R}^T\right) \\ &= \mathbf{R}\boldsymbol{\Delta}^T\mathbf{L}^T\mathbf{L}\boldsymbol{\Delta}\mathbf{R}^T \\ &= \mathbf{R}(\boldsymbol{\Delta}^T\boldsymbol{\Delta})\mathbf{R}^T \\ &= \mathbf{R}\boldsymbol{\Delta}_d^2\mathbf{R}^T \qquad \boldsymbol{\Delta}_d^2(i,i) = \delta_i^2 \end{aligned}$$

$$\begin{aligned} \bar{\mathbf{D}}^T\bar{\mathbf{D}} &= n\boldsymbol{\Sigma} \\ &= n\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}(n\boldsymbol{\Lambda})\mathbf{U}^T \\ \lambda_i &= \frac{\delta_i^2}{n} \end{aligned}$$

Each $\boldsymbol{r}_i$ is an eigenvector of $\boldsymbol{D}^T\boldsymbol{D}$, with the corresponding eigenvalue δ_i^2. Only $r \leq \min(d, n)$ of these eigenvalues are positive, whereas the rest are all zeros.

$$\begin{aligned} \bar{\mathbf{D}}\bar{\mathbf{D}}^T &= (\mathbf{L}\boldsymbol{\Delta}\mathbf{R}^T)(\mathbf{L}\boldsymbol{\Delta}\mathbf{R}^T)^T \\ &= \mathbf{L}\boldsymbol{\Delta}\mathbf{R}^T\mathbf{R}\boldsymbol{\Delta}^T\mathbf{L}^T \\ &= \mathbf{L}(\boldsymbol{\Delta}\boldsymbol{\Delta}^T)\mathbf{L}^T \\ &= \mathbf{L}\boldsymbol{\Delta}_n^2\mathbf{L}^T \qquad \boldsymbol{\Delta}_n^2(i,i) = \delta_i^2 \end{aligned}$$

where $\boldsymbol{\Delta}_n^2$ is the $n \times n$ diagonal matrix given as $\boldsymbol{\Delta}_n^2(i,i) = \delta_i^2$, for $i = 1, \ldots, n$. Only r of these singular values are positive, whereas the rest are all zeros. Thus, the left singular vectors in $\mathbf{L}$ are the eigenvectors of the $n \times n$ matrix $\bar{\mathbf{D}}\bar{\mathbf{D}}^T$, and the corresponding eigenvalues are given as δ_i^2.