

Chapter 4

Dimensionality Reduction

Last Modified: 2008-09-16 19:44

4.1 Principal Component Analysis

Principal Component Analysis or PCA is a technique used for reducing high dimensional datasets into lower dimensions. The general idea behind PCA is to transform the data to a new coordinate system that best captures the variance in the data. The direction of the largest variance is called the first principal component, the orthogonal direction that captures the second largest variance is called the second principal component, and so on.

Let the data sample consist of n points over d attributes, i.e., it is a $n \times d$ matrix, given as

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^d \\ x_2^1 & x_2^2 & \cdots & x_2^d \\ \cdots & \cdots & \cdots & \cdots \\ x_n^1 & x_n^2 & \cdots & x_n^d \end{pmatrix} \quad (4.1)$$

Each point (row) is thus given as a d -dimensional vector $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^d)^T$, whereas each attribute or random variable (column) can also be thought of as the vector $\mathbf{x}^j = (x_1^j, x_2^j, \dots, x_n^j)^T$.

We are interested in finding the best r -dimensional representation of \mathbf{X} , where $r < d$. We will start with $r = 0$, i.e., the best point approximation of \mathbf{X} . We will then consider $r = 1$, i.e., the best line that approximates \mathbf{X} , which will lead to the general PCA technique for the best $1 < r < d$ dimensional representation of \mathbf{X} .

4.1.1 Best Point Approximation

The point \mathbf{m} that best approximates \mathbf{X} is the $r = 0$ dimensional projection of \mathbf{X} , that minimizes the average sum of squared errors (SSE) from each point \mathbf{x}_i . That is, the optimization condition is given as:

$$SSE(\mathbf{m}) = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{m}\|^2 \quad (4.2)$$

To minimize $SSE(\mathbf{m})$ we differentiate it with respect to \mathbf{m} , and set the derivative to zero,

$$\frac{d}{d\mathbf{m}}SSE(\mathbf{m}) = \frac{d}{d\mathbf{m}} \frac{\sum_{i=1}^n \|\mathbf{x}_i - \mathbf{m}\|^2}{n} = 0 \quad (4.3)$$

$$-\frac{2}{n} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m}) = 0 \quad (4.4)$$

$$\sum_{i=1}^n \mathbf{x}_i - \sum_{i=1}^n \mathbf{m} = 0 \quad (4.5)$$

$$n\mathbf{m} = \sum_{i=1}^n \mathbf{x}_i \quad (4.6)$$

$$\mathbf{m} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \boldsymbol{\mu} \quad (4.7)$$

We conclude that the mean vector $\boldsymbol{\mu} = \frac{\sum_{i=1}^n \mathbf{x}_i}{n}$ is the d -dimensional point that best approximates the data \mathbf{X} .

4.1.2 Best Line Approximation

We now want to find the line \mathbf{u} that best approximates \mathbf{X} . Without loss of generality, we assume that \mathbf{u} has magnitude $\|\mathbf{u}\|^2 = \mathbf{u}^T \mathbf{u} = 1$. In other words, we want to find the direction of the unit vector, which represents the best line approximation to \mathbf{X} .

Note that each point \mathbf{x}_i can be projected onto the vector \mathbf{u} as follows:

$$\left(\frac{\mathbf{u}^T \mathbf{x}_i}{\mathbf{u}^T \mathbf{u}} \right) \mathbf{u} = (\mathbf{u}^T \mathbf{x}_i) \mathbf{u} = (\mathbf{x}'_i) \mathbf{u} \quad (4.8)$$

where the scalar

$$\boxed{\mathbf{x}'_i = \mathbf{u}^T \mathbf{x}_i} \quad (4.9)$$

gives the coordinate of the projected point in the space spanned by \mathbf{u} .

The first principal component corresponds to the direction \mathbf{u} such that the variance of the projected points \mathbf{x}'_i is maximized. That is we want to maximize

$$\sum_{i=1}^n \frac{(\mathbf{x}'_i - \boldsymbol{\mu}')^2}{n} = \frac{1}{n} \sum_{i=1}^n (\mathbf{u}^T \mathbf{x}_i - \mathbf{u}^T \boldsymbol{\mu})^2 \quad (4.10)$$

$$= \frac{1}{n} \sum_{i=1}^n (\mathbf{u}^T (\mathbf{x}_i - \boldsymbol{\mu}))^2 \quad (4.11)$$

$$= \frac{1}{n} \sum_{i=1}^n (\mathbf{u}^T (\mathbf{x}_i - \boldsymbol{\mu})) ((\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{u}) \quad (4.12)$$

$$= \mathbf{u}^T \left(\sum_{i=1}^n \frac{(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T}{n} \right) \mathbf{u} \quad (4.13)$$

$$= \mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} \quad (4.14)$$

We can now maximize the projected variance $\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u}$ with respect to \mathbf{u} , subject to the condition that $\mathbf{u}^T \mathbf{u} = 1$. This is done by introducing a Lagrangian multiplier α , and maximizing the following:

$$\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} - \alpha(\mathbf{u}^T \mathbf{u} - 1) \quad (4.15)$$

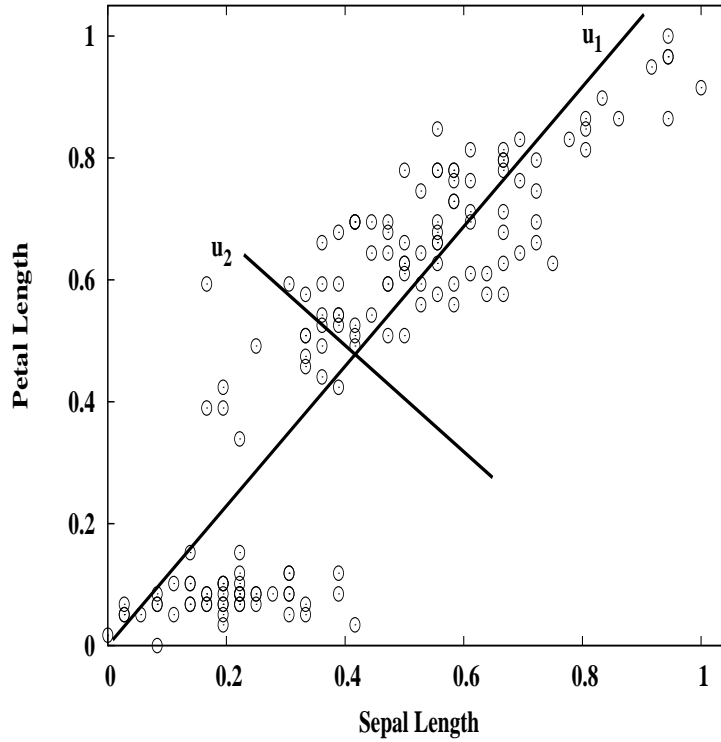


Figure 4.1: Principal Components: \mathbf{u}_1 and \mathbf{u}_2

Setting the derivative of (4.15) with respect to \mathbf{u} to zero, gives us:

$$\frac{\partial}{\partial \mathbf{u}} \mathbf{u}^T \mathbf{\Sigma} \mathbf{u} - \alpha (\mathbf{u}^T \mathbf{u} - 1) = 0 \quad (4.16)$$

$$2\mathbf{\Sigma} \mathbf{u} - 2\alpha \mathbf{u} = 0 \quad (4.17)$$

$$\boxed{\mathbf{\Sigma} \mathbf{u} = \alpha \mathbf{u}} \quad (4.18)$$

This implies that α is an eigenvalue of the covariance matrix $\mathbf{\Sigma}$, with the associated eigenvector \mathbf{u} . Furthermore,

$$\boxed{\mathbf{u}^T \mathbf{\Sigma} \mathbf{u} = \mathbf{u}^T \alpha \mathbf{u} = \alpha \mathbf{u}^T \mathbf{u} = \alpha} \quad (4.19)$$

This means that to achieve the largest projected variance $\mathbf{u}^T \mathbf{\Sigma} \mathbf{u}$, we need to choose the largest eigenvalue of $\mathbf{\Sigma}$. In other words if λ_1 is the largest eigenvalue of $\mathbf{\Sigma}$, then the choice of $\alpha = \lambda_1$ maximizes the projected variance. We conclude that the first principal component is given by the eigenvector \mathbf{u} corresponding to the largest eigenvalue λ_1 of $\mathbf{\Sigma}$.

Minimum Squared Error Approach

Here we will show that \mathbf{u} is also the direction that leads to the least sum of squared-errors between the original and projected points. First, let us assume that all points \mathbf{x}_i in the original space have been centered at the origin by subtracting the mean, as follow

$$\mathbf{z}_i = \mathbf{x}_i - \boldsymbol{\mu}$$

We will then project these mean subtracted points \mathbf{z}_i onto the line \mathbf{u} . Note that this translation of the points will not affect the perpendicular distances between the points in the original and projected spaces. Define the average sum of squared-errors optimization condition as follows:

$$SSE(\mathbf{u}) = \sum_{i=1}^n \frac{\|\mathbf{z}'_i - \mathbf{z}_i\|^2}{n} \quad (4.20)$$

$$= \frac{1}{n} \sum_{i=1}^n \|[\mathbf{u}^T(\mathbf{x}_i - \boldsymbol{\mu})]\mathbf{u} - (\mathbf{x}_i - \boldsymbol{\mu})\|^2 \quad (4.21)$$

$$= \frac{1}{n} \left(\sum_{i=1}^n [\mathbf{u}^T(\mathbf{x}_i - \boldsymbol{\mu})]^2 \|\mathbf{u}\|^2 - 2 \sum_{i=1}^n [\mathbf{u}^T(\mathbf{x}_i - \boldsymbol{\mu})][\mathbf{u}^T(\mathbf{x}_i - \boldsymbol{\mu})] + \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \right) \quad (4.22)$$

$$= -\frac{1}{n} \sum_{i=1}^n [\mathbf{u}^T(\mathbf{x}_i - \boldsymbol{\mu})]^2 + \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \quad (4.23)$$

$$= -\frac{1}{n} \sum_{i=1}^n \mathbf{u}^T(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{u} + \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \quad (4.24)$$

$$= -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} + \sum_{i=1}^n \frac{\|\mathbf{x}_i - \boldsymbol{\mu}\|^2}{n} \quad (4.25)$$

The second term above does not depend on \mathbf{u} . Thus the vector \mathbf{u} that minimizes $SSE(\mathbf{u})$ is in fact the same one that maximizes the projected variance $\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u}$.

4.1.3 Best r -dimensional Approximation

We are now interested in the best r -dimensional approximation to \mathbf{X} , where $r \leq d$.

Let us first consider the case where $r = 2$. We already computed the direction with the most variance, namely \mathbf{u}_1 which is the eigenvector corresponding to the largest eigenvalue λ_1 of $\boldsymbol{\Sigma}$. We now want to find another direction \mathbf{u}_2 , which also maximizes the projected variance, but which is orthogonal to \mathbf{u}_1 , i.e.,

$$\mathbf{u}_2^T \mathbf{u}_1 = 0$$

We also require \mathbf{u}_2 to be of unit length, i.e.,

$$\mathbf{u}_2^T \mathbf{u}_2 = 1$$

The optimization condition then becomes to maximize

$$\mathbf{u}_2^T \boldsymbol{\Sigma} \mathbf{u}_2 - \alpha(\mathbf{u}_2^T \mathbf{u}_2 - 1) - \beta(\mathbf{u}_2^T \mathbf{u}_1 - 0) \quad (4.26)$$

Taking the derivative with respect to \mathbf{u}_2 , and setting it to zero, gives

$$2\boldsymbol{\Sigma} \mathbf{u}_2 - 2\alpha \mathbf{u}_2 - \beta \mathbf{u}_1 = 0 \quad (4.27)$$

If we multiply on the left by \mathbf{u}_1^T we get

$$2\mathbf{u}_1^T \boldsymbol{\Sigma} \mathbf{u}_2 - 2\alpha \mathbf{u}_1^T \mathbf{u}_2 - \beta \mathbf{u}_1^T \mathbf{u}_1 = 0 \quad (4.28)$$

$$2\mathbf{u}_2^T \boldsymbol{\Sigma} \mathbf{u}_1 - \beta = 0 \quad (4.29)$$

$$\beta = 2\mathbf{u}_2^T \lambda_1 \mathbf{u}_1 \quad (4.30)$$

$$\beta = 2\lambda_1 \mathbf{u}_2^T \mathbf{u}_1 \quad (4.31)$$

$$\beta = 0 \quad (4.32)$$

$$(4.33)$$

Note that in the derivation above we used the fact that $\mathbf{u}_1^T \Sigma \mathbf{u}_2$ is a scalar and thus equals its transpose $\mathbf{u}_2^T \Sigma \mathbf{u}_1$. Plugging $\beta = 0$ into (4.27) gives us:

$$2\Sigma \mathbf{u}_2 - 2\alpha \mathbf{u}_2 = 0 \quad (4.34)$$

$$\boxed{\Sigma \mathbf{u}_2 = \alpha \mathbf{u}_2} \quad (4.35)$$

This means that \mathbf{u}_2 is another eigenvector of Σ , and to maximize the variance along \mathbf{u}_2 , we should choose $\alpha = \lambda_2$, the second largest eigenvalue of Σ .

Higher Dimensional Approximation

Based on the above analysis, the general approach to finding an additional principal components \mathbf{u}_i is to make sure that it is normalized to unit length, and that it is orthogonal to all previous components \mathbf{u}_j , with $1 \leq j < i$. Setting up a maximization formulation with Lagrange multipliers will give:

$$\mathbf{u}_i^T \Sigma \mathbf{u}_i - \alpha(\mathbf{u}_i^T \mathbf{u}_i - 1) - \left(\sum_{j=1}^{i-1} \beta_j (\mathbf{u}_i^T \mathbf{u}_j - 0) \right) \quad (4.36)$$

Taking the derivative with respect to \mathbf{u}_i and setting it to zero, will eventually lead to $\beta_j = 0$ for all $j < i$, and we will find that

$$\boxed{\Sigma \mathbf{u}_i = \alpha \mathbf{u}_i} \quad (4.37)$$

which means that to maximize the variance along \mathbf{u}_i , we should set $\alpha = \lambda_i$, the i -th largest eigenvalue of Σ .

To summarize, to find the best r dimensional approximation to \mathbf{X} , we compute the eigenvalues of Σ , and sort them in decreasing order

$$\lambda_1 \geq \lambda_2 \geq \dots \lambda_r \geq \lambda_{r+1} \dots \geq \lambda_d \geq 0 \quad (4.38)$$

then we select the r largest eigenvalues, and their corresponding eigenvectors to form the best r -dimensional approximation. Note above that since Σ is *positive semi-definite*, its eigenvalues must all non-negative.

Geometry of PCA

It is worth emphasizing the geometry of the PCA method. When $r = d$, PCA corresponds to a rotation of the axes, so that along the new principal directions $(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d)$, all covariances vanish. This can be seen by looking at the collective action of the full set of principal components, which can be arranged in a matrix as follows:

$$\mathbf{U} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_d \\ | & | & & | \end{pmatrix} \quad (4.39)$$

This $d \times d$ matrix is an *orthogonal* matrix, whose columns, the principal components, are pairwise orthogonal and are of unit length, i.e., the principal components are *orthonormal*

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (4.40)$$

Since \mathbf{U} is orthogonal, this means that its inverse equals its transpose

$$\mathbf{U}^{-1} = \mathbf{U}^T \quad (4.41)$$

As we derived above, each principal component \mathbf{u}_i corresponds to an eigenvector of the covariance matrix Σ , i.e.,

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i \text{ for all } 1 \leq i \leq d \quad (4.42)$$

which can be written compactly in matrix notation as follows:

$$\Sigma \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \lambda_1 \mathbf{u}_1 & \lambda_2 \mathbf{u}_2 & \cdots & \lambda_d \mathbf{u}_d \\ | & | & & | \end{pmatrix} \quad (4.43)$$

$$\Sigma \mathbf{U} = \mathbf{U} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix} \quad (4.44)$$

$$\Sigma \mathbf{U} = \mathbf{U} \Lambda \quad (4.45)$$

If we multiply (4.45) on the left by \mathbf{U}^T we obtain:

$$\mathbf{U}^T \Sigma \mathbf{U} = \mathbf{U}^T \mathbf{U} \Lambda = \Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix} \quad (4.46)$$

In other words, after transforming the data into the new set of axes given by the principal components, all of the covariances have vanished, and we are left with the diagonal matrix Λ of the variances along each of the principal components. Furthermore, the variance along each new direction is captured by the corresponding eigenvalue, that is

$$\boxed{\sigma_{\mathbf{u}_i}^2 = \lambda_i} \quad (4.47)$$

Choosing the Dimensionality

Often we may not know how many dimensions, r , to use to get a good approximation of the original data \mathbf{X} . One criteria of choosing r is to compute the fraction of the total variability for different values of r , given as follows:

$$f(r) = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_r}{\lambda_1 + \lambda_2 + \cdots + \lambda_d} = \frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^d \lambda_i} \quad (4.48)$$

This way, given a certain desired fraction, say α , starting from the first principal component, we keep on adding additional components, and stop at the lowest value r , for which $f(r) \geq \alpha$.

4.2 Singular Value Decomposition

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

For example, if we multiply (4.45) on the right by \mathbf{U}^T we obtain:

$$\Sigma \mathbf{U} \mathbf{U}^T = \mathbf{U} \Lambda \mathbf{U}^T \implies \Sigma = \mathbf{U} \Lambda \mathbf{U}^T \quad (4.49)$$

which says that the initial covariance matrix has been factorized into the orthogonal matrix \mathbf{U} containing its eigenvectors, and a diagonal matrix $\mathbf{\Lambda}$ containing its eigenvalues (sorted in decreasing order).

SVD generalizes the above factorization for any matrix. In particular for an $n \times d$ data matrix \mathbf{X} with n points and d columns, with rank r , SVD factorizes \mathbf{X} as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \quad (4.50)$$

where \mathbf{U} is a orthogonal $n \times n$ matrix, \mathbf{V} is an orthogonal $d \times d$ matrix, and $\mathbf{\Lambda}$ is a $n \times d$ matrix. The columns of \mathbf{U} are called the *left singular vectors*, the columns of \mathbf{V} (or rows of \mathbf{V}^T) are called the *right singular vectors*, and the diagonal entries $\delta_i = \mathbf{\Lambda}(i, i)$ along the main diagonal of $\mathbf{\Lambda}$ are called the *singular values* of \mathbf{X} . In fact, since the rank of the matrix is only $r \leq \min(n, d)$, there are only r non-zero singular values, which we assume are ordered as follows

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

The rest of the diagonal entries, and of course all non-diagonal entries are all zeros.

Since we are mainly interested in the non-zero singular values, we can discard those left and right singular vectors that correspond to zero singular values, to obtain the SVD as:

$$\mathbf{X} = \left(\begin{array}{c|c|c|c} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_r \\ | & | & & | \end{array} \right) \begin{pmatrix} \delta_1 & 0 & \dots & 0 \\ 0 & \delta_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \delta_r \end{pmatrix} \begin{pmatrix} - & \mathbf{v}_1^T & - \\ - & \mathbf{v}_2^T & - \\ - & \vdots & - \\ - & \mathbf{v}_r^T & - \end{pmatrix} \quad (4.51)$$

$$\mathbf{X} = \sum_{i=1}^r \delta_i \mathbf{u}_i \mathbf{v}_i^T \quad (4.52)$$

In other words, we can think of \mathbf{U} as an orthogonal $n \times r$ matrix, $\mathbf{\Lambda}$ as a $r \times r$ diagonal matrix, and \mathbf{V} as a $r \times d$ orthogonal matrix.

Equation 4.52 gives the so-called *spectral decomposition* of \mathbf{X} into rank one matrices of the form $\delta_i \mathbf{u}_i \mathbf{v}_i^T$. By selecting the q largest singular-values $\delta_1 \dots \delta_q$ and the corresponding left/right singular-vectors, we obtain the best rank q approximation to the original matrix \mathbf{X} . That is, if \mathbf{X}_q is the matrix defined as:

$$\mathbf{X}_q = \sum_{i=1}^q \delta_i \mathbf{u}_i \mathbf{v}_i^T \quad (4.53)$$

then it can be shown that \mathbf{X}_q is the rank q matrix that minimizes the expression

$$\|\mathbf{X} - \mathbf{X}_q\|_F \quad (4.54)$$

where $\|\mathbf{A}\|_F$ is called the *Frobenius Norm* of the matrix \mathbf{A} , given as:

$$\|\mathbf{A}\|_F = \sqrt{\sum_i \sum_j \mathbf{A}(i, j)^2} \quad (4.55)$$

4.2.1 Geometry of SVD

In general, any $n \times d$ matrix \mathbf{X} represents a *linear transformation*, $\mathbf{X} : \mathbb{R}^d \rightarrow \mathbb{R}^n$, from the space of d -dimensional vectors to the space of n -dimensional vectors, since for any $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^d$, we have:

$$\mathbf{X}\mathbf{v} = \mathbf{u} \quad (4.56)$$

The set of all vectors $\mathbf{u} \in \mathbb{R}^n$ such that $\mathbf{X}\mathbf{v} = \mathbf{u}$ for some $\mathbf{v} \in \mathbb{R}^d$, is called the *column space* of \mathbf{X} , and the set of all vectors $\mathbf{v} \in \mathbb{R}^d$, such that $\mathbf{X}^T\mathbf{u} = \mathbf{v}$ for some $\mathbf{u} \in \mathbb{R}^n$, is called the *row space* of \mathbf{X} . In other words, the column space of \mathbf{X} is the set of all vectors that can be obtained as the linear combinations of columns of \mathbf{X} , and the row space of \mathbf{X} is the set of all vectors that can be obtained as the linear combinations of the rows of \mathbf{X} . Also note that the set of all vectors $\mathbf{v} \in \mathbb{R}^d$, such that $\mathbf{X}\mathbf{v} = \mathbf{0}$ is called the *null space* of \mathbf{X} , and finally, the set of all vectors $\mathbf{u} \in \mathbb{R}^n$, such that $\mathbf{X}^T\mathbf{u} = \mathbf{0}$ is called the *left null space* of \mathbf{X} .

One of the main properties of SVD is that it gives a basis for each of the four fundamental spaces associated with matrix \mathbf{X} . If \mathbf{X} has rank r , it means that it has only r independent columns, and also only r independent rows! Thus the r orthonormal vectors \mathbf{u}_i corresponding to the r non-zero singular values of \mathbf{X} in (4.50), in fact, represent a basis for the column space of \mathbf{X} . The remaining $n - r$ orthonormal vectors \mathbf{u}_j , represent a basis for the left null space of \mathbf{X} . For the row space, the r vectors \mathbf{v}_i , corresponding to the r non-zero singular values, represent a basis for the row space of \mathbf{X} , and the remaining $d - r$ orthonormal vectors \mathbf{v}_j , represent a basis for the null space of \mathbf{X} .

For the moment we are only interested in the column and row spaces of \mathbf{X} , given by the left and right singular vectors, \mathbf{u}_i and \mathbf{v}_i , corresponding to the r non-zero singular values δ_i . We can think of the SVD as essentially a mapping from an orthonormal basis set of vectors $(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r)$ in \mathbb{R}^d (the row space), to an orthonormal basis set of vectors $(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r)$ in \mathbb{R}^n (the column space). Note that in general, any orthonormal basis for the row space need not map to an orthonormal basis for the column space. The goal of SVD is to find such a transformation that satisfies this requirement. In particular, we require that for any basis vector \mathbf{v}_i in the row space, we have:

$$\mathbf{X}\mathbf{v}_i = \delta_i\mathbf{u}_i \quad (4.57)$$

In other words, \mathbf{X} transforms \mathbf{v}_i in the same direction as \mathbf{u}_i , and scales it by the singular value δ_i . Over all the basis vectors, in matrix notation, we require:

$$\mathbf{X} \begin{pmatrix} | & | & \cdots & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_r \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & \cdots & | \\ \delta_1\mathbf{u}_1 & \delta_2\mathbf{u}_2 & \cdots & \delta_r\mathbf{u}_r \\ | & | & & | \end{pmatrix} \quad (4.58)$$

$$\mathbf{X}\mathbf{V} = \mathbf{U} \begin{pmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_r \end{pmatrix} \quad (4.59)$$

$$\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Delta} \quad (4.60)$$

Multiplying by \mathbf{V}^T on the right, and noting that $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, since \mathbf{V} is orthogonal (which implies that $\mathbf{V}^{-1} = \mathbf{V}^T$), we get the SVD factorization given in (4.50)

$$\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T \quad (4.61)$$

4.2.2 Computing the SVD

Let us look at the matrix $\mathbf{X}^T\mathbf{X}$; plugging in (4.50) for \mathbf{X} , we get

$$\mathbf{X}^T\mathbf{X} = (\mathbf{U}\mathbf{\Delta}\mathbf{V}^T)^T(\mathbf{U}\mathbf{\Delta}\mathbf{V}^T) \quad (4.62)$$

$$= \mathbf{V}\mathbf{\Delta}^T\mathbf{U}^T\mathbf{U}\mathbf{\Delta}\mathbf{V}^T \quad (4.63)$$

$$= \mathbf{V}(\mathbf{\Delta}^T\mathbf{\Delta})\mathbf{V}^T \quad (4.64)$$

$$= \mathbf{V}\mathbf{\Delta}_d^2\mathbf{V}^T \quad (4.65)$$

Note that since $\mathbf{\Delta}$ is a $n \times d$ “diagonal” matrix, $\mathbf{\Delta}^T \mathbf{\Delta}$ yields a $d \times d$ diagonal matrix, $\mathbf{\Delta}_d^2$. Noting the similarity with (4.49), we immediately notice that \mathbf{V} are the eigenvectors of $\mathbf{X}^T \mathbf{X}$, with the corresponding eigenvalues δ_i^2 . In other words, the singular values of \mathbf{X} are the square roots of the eigenvalues of $\mathbf{X}^T \mathbf{X}$.

Likewise, when we consider the matrix $\mathbf{X} \mathbf{X}^T$, we get:

$$\mathbf{X} \mathbf{X}^T = (\mathbf{U} \mathbf{\Delta} \mathbf{V}^T)(\mathbf{U} \mathbf{\Delta} \mathbf{V}^T)^T \quad (4.66)$$

$$= \mathbf{U} \mathbf{\Delta} \mathbf{V}^T \mathbf{V}^T \mathbf{\Delta}^T \mathbf{U}^T \quad (4.67)$$

$$= \mathbf{U}(\mathbf{\Delta} \mathbf{\Delta}^T) \mathbf{U}^T \quad (4.68)$$

$$= \mathbf{U} \mathbf{\Delta}_n^2 \mathbf{U}^T \quad (4.69)$$

where $\mathbf{\Delta}_n^2$ is a $n \times n$ diagonal matrix. This shows that \mathbf{U} are the eigenvectors of the matrix $\mathbf{X} \mathbf{X}^T$, again with the eigenvalues δ_i^2 . Note that only r of these eigenvalues are positive, whereas the rest are all zeros. Taking the positive square roots of δ_i^2 , gives us the r non-zero, positive, singular values of \mathbf{X} .

Let us consider the example matrix:

$$\mathbf{X} = \begin{pmatrix} 0 & 1 & 2 \\ -2 & -1 & 0 \end{pmatrix}$$

We have

$$\mathbf{X} \mathbf{X}^T = \begin{pmatrix} 5 & -1 \\ -1 & 5 \end{pmatrix}$$

which has eigenvalues $\delta_1^2 = 6$ and $\delta_2^2 = 4$. Solving for the corresponding eigenvectors we obtain:

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \text{ and } \mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Next, we have

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 4 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 2 & 4 \end{pmatrix}$$

which has the same positive eigenvalues $\delta_1^2 = 6$ and $\delta_2^2 = 4$. To find the eigenvector for $\delta_1^2 = 6$, we solve the equation:

$$\begin{pmatrix} 4 - \delta_1^2 & 2 & 0 \\ 2 & 2 - \delta_1^2 & 2 \\ 0 & 2 & 4 - \delta_1^2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{0}$$

$$\begin{pmatrix} -2 & 2 & 0 \\ 2 & -4 & 2 \\ 0 & 2 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{0}$$

This gives us two equations to solve:

$$\begin{aligned} -x + y &= 0 \\ x - 2y + z &= 0 \end{aligned}$$

setting the free variable $z = 1$, we obtain $y = 1$ and $x = 1$, giving us

$$\mathbf{u}_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

In a similar manner we can obtain the remaining eigenvectors:

$$\mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \text{ and } \mathbf{u}_3 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

where \mathbf{u}_2 corresponds to the eigenvalue $\delta_2^2 = 4$, and \mathbf{u}_3 corresponds to the eigenvalue $\delta_3^2 = 0$.

Putting it all together, we obtain the SVD as follows:

$$\mathbf{X} = \begin{pmatrix} 0 & 1 & 2 \\ -2 & -1 & 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{6} & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/\sqrt{6} & -2/\sqrt{6} & 1/\sqrt{6} \end{pmatrix}$$

In terms of the spectral decomposition of \mathbf{X} , we have:

$$\begin{aligned} \mathbf{X} &= \frac{\sqrt{6}}{\sqrt{2}\sqrt{3}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} + \frac{2}{\sqrt{2}\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 & 1 \\ -1 & -1 & -1 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 & 2 \\ -2 & -1 & 0 \end{pmatrix} \end{aligned}$$