Matrix Decomposition and Applications

Jun Lu
jun.lu.locky@gmail.com

Abstract

In 1954, Alston S. Householder published *Principles of Numerical Analysis*, one of the first modern treatments on matrix decomposition that favored a (block) LU decomposition-the factorization of a matrix into the product of lower and upper triangular matrices. And now, matrix decomposition has become a core technology in machine learning, largely due to the development of the back propagation algorithm in fitting a neural network. The sole aim of this survey is to give a self-contained introduction to concepts and mathematical tools in numerical linear algebra and matrix analysis in order to seamlessly introduce matrix decomposition techniques and their applications in subsequent sections. However, we clearly realize our inability to cover all the useful and interesting results concerning matrix decomposition and given the paucity of scope to present this discussion, e.g., the separated analysis of the Euclidean space, Hermitian space, Hilbert space, and things in the complex domain. We refer the reader to literature in the field of linear algebra for a more detailed introduction to the related fields.

This survey is primarily a summary of purpose, significance of important matrix decomposition methods, e.g., LU, QR, and SVD, and the origin and complexity of the methods which shed light on their modern applications. Most importantly, this article presents improved procedures for most of the calculations of the decomposition algorithms which potentially reduce the complexity they induce. Again, this is a decomposition-based context, thus we will introduce the related background when it is needed and necessary. In many other textbooks on linear algebra, the principal ideas are discussed and the matrix decomposition methods serve as "byproduct". However, we focus on the decomposition methods instead and the principal ideas serve as fundamental tools for them. The mathematical prerequisite is a first course in linear algebra. Other than this modest background, the development is self-contained, with rigorous proof provided throughout.

Keywords: Existence and computing of matrix decompositions, Complexity, Floating point operations (flops), Low-rank approximation, Pivot, LU decomposition for nonzero leading principal minors, Data distillation, CR decomposition, CUR/Skeleton decomposition, Interpolative decomposition, Biconjugate decomposition, Coordinate transformation, Hessenberg decomposition, ULV decomposition, URV decomposition, Rank decomposition, Gram-Schmidt process, Householder reflector, Givens rotation, Rank revealing decomposition, Cholesky decomposition and update/downdate, Eigenvalue problems, Alternating least squares, Randomized algorithm.

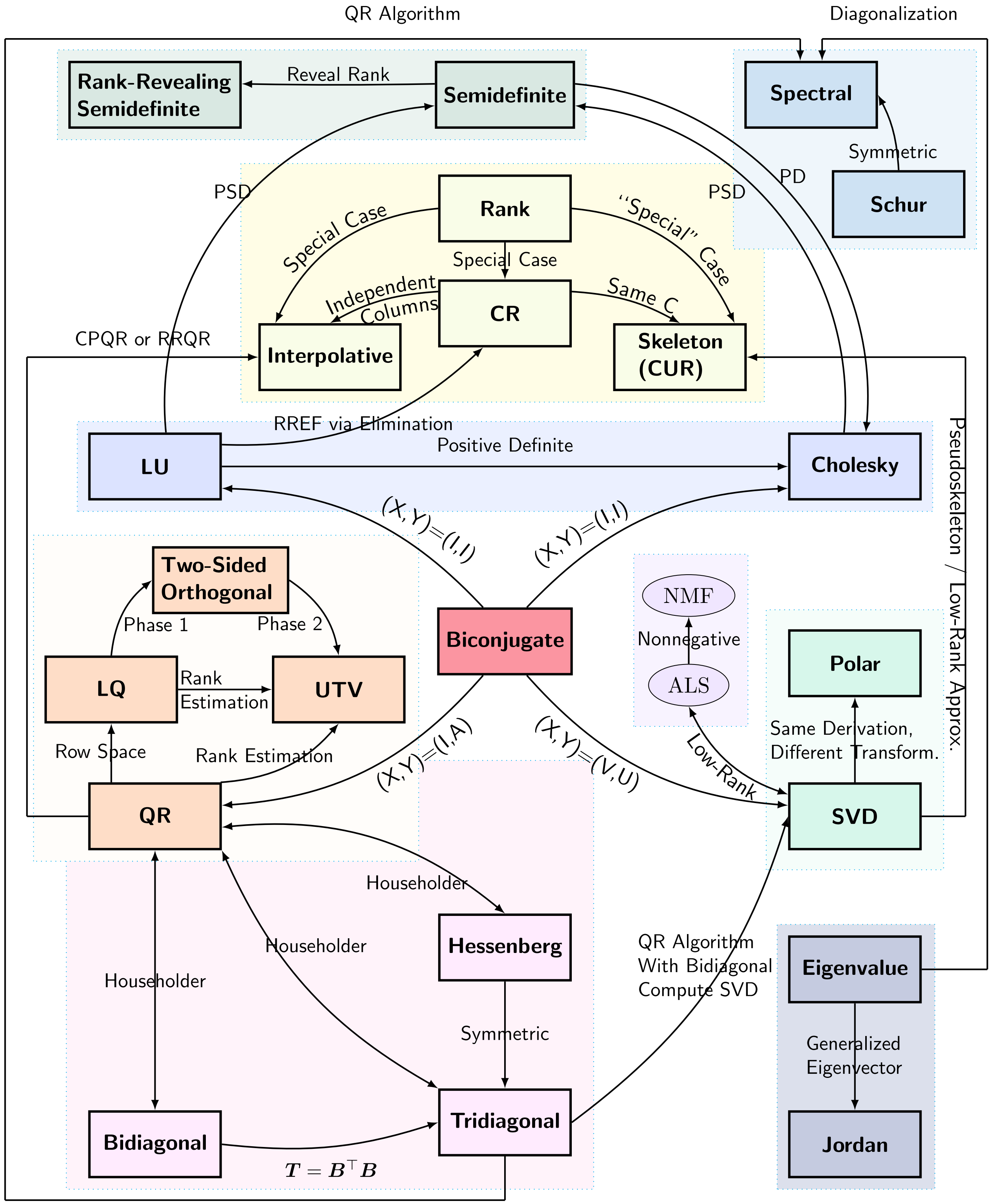


Figure 1: Matrix Decomposition World Map.

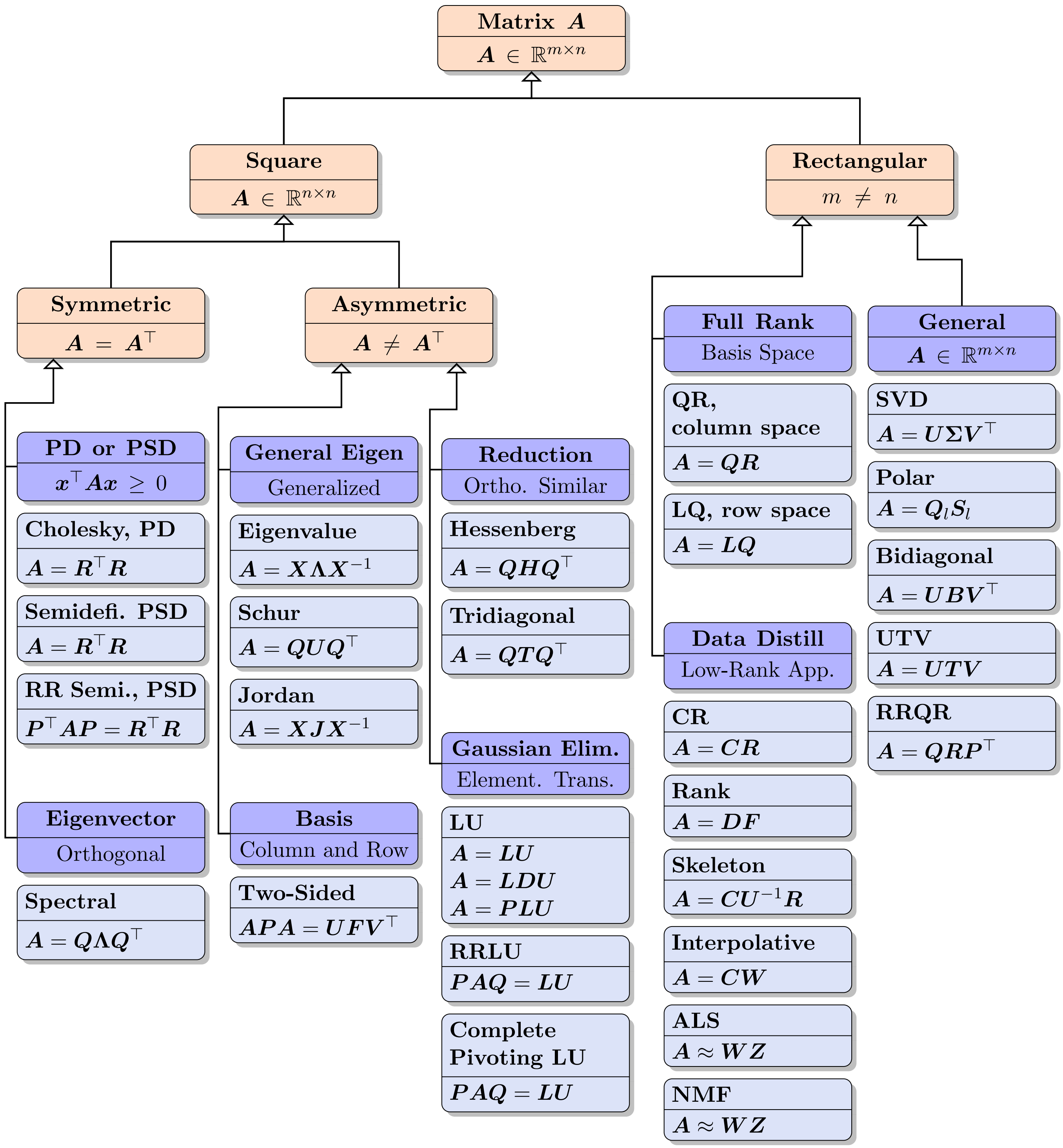


Figure 2: Matrix Decomposition World Map Under Conditions.

Contents

Introduction and Background

Matrix decomposition has become a core technology in statistics (Banerjee and Roy, 2014; Gentle, 1998), optimization (Gill et al., 2021), machine learning (Goodfellow et al., 2016; Bishop, 2006), and deep learning largely due to the development of back propagation algorithm in fitting a neural network and the low-rank neural networks in efficient deep learning. The sole aim of this survey is to give a self-contained introduction to concepts and mathematical tools in numerical linear algebra and matrix analysis in order to seamlessly introduce matrix decomposition techniques and their applications in subsequent sections. However, we clearly realize our inability to cover all the useful and interesting results concerning matrix decomposition and given the paucity of scope to present this discussion, e.g., the separated analysis of the Euclidean space, Hermitian space, and Hilbert space. We refer the reader to literature in the field of linear algebra for a more detailed introduction to the related fields. Some excellent examples include (Householder, 2006; Trefethen and Bau III, 1997; Strang, 2009; Stewart, 2000; Gentle, 2007; Higham, 2002; Quarteroni et al., 2010; Golub and Van Loan, 2013; Beck, 2017; Gallier and Quaintance, 2017; Boyd and Vandenberghe, 2018; Strang, 2019; van de Geijn and Myers, 2020; Strang, 2021). Most importantly, this survey will only cover the compact proofs of the existence of the matrix decomposition methods. For more details on how to reduce the calculation complexity, rigorous discussion in various applications and examples, why each matrix decomposition method is important in practice, and preliminaries on tensor decomposition, one can refer to (Lu, 2021c).

A matrix decomposition is a way of reducing a complex matrix into its constituent parts which are in simpler forms. The underlying principle of the decompositional approach to matrix computation is that it is not the business of the matrix algorithmists to solve particular problems, but it is an approach that can simplify more complex matrix operations which can be performed on the decomposed parts rather than on the original matrix itself. At a general level, a matrix decomposition task on matrix $\boldsymbol{A}$ can be cast as

- $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{U}$: where $\boldsymbol{Q}$ is an orthogonal matrix that contains the same column space as $\boldsymbol{A}$ and $\boldsymbol{U}$ is a relatively simple and sparse matrix to reconstruct $\boldsymbol{A}$.
- $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^\top$: where $\boldsymbol{Q}$ is orthogonal such that $\boldsymbol{A}$ and $\boldsymbol{T}$ are *similar matrices* that share the same properties such as same eigenvalues, sparsity. Moreover, working on $\boldsymbol{T}$ is an easier task compared to that of $\boldsymbol{A}$.
- $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{T}\boldsymbol{V}$: where $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal matrices such that the columns of $\boldsymbol{U}$ and the rows of $\boldsymbol{V}$ constitute an orthonormal basis of the column space and row space of $\boldsymbol{A}$ respectively.
- $\underset{m\times n}{\boldsymbol{A}} = \underset{m\times r}{\boldsymbol{B}} \quad \underset{r\times n}{\boldsymbol{C}}$: where $\boldsymbol{B}, \boldsymbol{C}$ are full rank matrices that can reduce the memory storage of $\boldsymbol{A}$. In practice, a low-rank approximation $\underset{m\times n}{\boldsymbol{A}} \approx \underset{m\times k}{\boldsymbol{D}} \quad \underset{k\times n}{\boldsymbol{F}}$ can be employed where $k < r$ is called the *numerical rank* of the matrix such that the matrix can be stored much more inexpensively and can be multiplied rapidly with vectors or other matrices. An approximation of the form $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{F}$ is useful for storing the matrix $\boldsymbol{A}$ more frugally (we can store $\boldsymbol{D}$ and $\boldsymbol{F}$ using $k(m+n)$ floats, as opposed to mn numbers for storing $\boldsymbol{A}$), for efficiently computing a matrix-vector product $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ (via $\boldsymbol{c} = \boldsymbol{F}\boldsymbol{x}$ and $\boldsymbol{b} = \boldsymbol{D}\boldsymbol{c}$), for data interpretation, and much more.

- A matrix decomposition, which though is usually expensive to compute, can be reused to solve new problems involving the original matrix in different scenarios, e.g., as long as the factorization of $\boldsymbol{A}$ is obtained, it can be reused to solve the set of linear systems $\{\boldsymbol{b}_1 = \boldsymbol{A}\boldsymbol{x}_1, \boldsymbol{b}_2 = \boldsymbol{A}\boldsymbol{x}_2, \ldots, \boldsymbol{b}_k = \boldsymbol{A}\boldsymbol{x}_k\}$.
- More generally, a matrix decomposition can help to understand the internal meaning of what happens when multiplied by the matrix such that each constituent has a geometrical transformation (see Section 15, p. 148).

The matrix decomposition algorithms can fall into many categories. Nonetheless, six categories hold the center and we sketch it here:

1. Factorizations arise from Gaussian elimination including the LU decomposition and its positive definite alternative - Cholesky decomposition;
2. Factorizations obtained when orthogonalizing the columns or the rows of a matrix such that the data can be explained well in an orthonormal basis;
3. Factorizations where the matrices are skeletoned such that a subset of the columns or the rows can represent the whole data in a small reconstruction error, whilst, the sparsity and nonnegativity of the matrices are kept as they are;
4. Reduction to Hessenberg, tridiagonal, or bidiagonal form, as a result, the properties of the matrices can be explored in these reduced matrices such as rank, eigenvalues, and so on;
5. Factorizations result from the computation of the eigenvalues of matrices;
6. In particular, the rest can be cast as a special kind of decompositions that involve optimization methods, high-level ideas where the category may not be straightforward to determine.

The world pictures for decomposition in Figure 1 and 2 connect each decomposition method by their internal relations and also separate different methods by the criteria or prerequisites of them. Readers will get more information about the two pictures after reading the text.

Notation and preliminaries In the rest of this section we will introduce and recap some basic knowledge about linear algebra. For the rest of the important concepts, we define and discuss them as per need for clarity. The readers with enough background in matrix analysis can skip this section. In the text, we simplify matters by considering only matrices that are real. Without special consideration, the eigenvalues of the discussed matrices are also real. We also assume throughout that $||\cdot|| = ||\cdot||_2$.

In all cases, scalars will be denoted in a non-bold font possibly with subscripts (e.g., a, α, α_i). We will use **boldface** lower case letters possibly with subscripts to denote vectors (e.g., $\boldsymbol{\mu}$, $\boldsymbol{x}$, $\boldsymbol{x}_n$, $\boldsymbol{z}$) and **boldface** upper case letters possibly with subscripts to denote matrices (e.g., $\boldsymbol{A}$, $\boldsymbol{L}_j$). The i-th element of a vector $\boldsymbol{z}$ will be denoted by $\boldsymbol{z}_i$ in bold font (or z_i in the non-bold font). *The i-th row and j-th column value of matrix $\boldsymbol{A}$ will be denoted by $\boldsymbol{A}_{ij}$ if block submatrices are involved, or by a_{ij} if block submatrices are not involved.* Furthermore, it will be helpful to utilize the **Matlab-style notation**, the i-th row to j-th row and k-th column to m-th column submatrix of matrix $\boldsymbol{A}$ will be denoted by $\boldsymbol{A}_{i:j,k:m}$. When the index is not continuous, given ordered subindex sets I and J, $\boldsymbol{A}[I, J]$ denotes the submatrix of $\boldsymbol{A}$

obtained by extracting the rows and columns of $\boldsymbol{A}$ indexed by I and J, respectively; and $\boldsymbol{A}[:,J]$ denotes the submatrix of $\boldsymbol{A}$ obtained by extracting the columns of $\boldsymbol{A}$ indexed by J.

And in all cases, vectors are formulated in a column rather than in a row. A row vector will be denoted by a transpose of a column vector such as $\boldsymbol{a}^\top$. A specific column vector with values is split by the symbol "$;$", e.g., $\boldsymbol{x}=[1;2;3]$ is a column vector in $\mathbb{R}^3$. Similarly, a specific row vector with values is split by the symbol "$,$", e.g., $\boldsymbol{y}=[1,2,3]$ is a row vector with 3 values. Further, a column vector can be denoted by the transpose of a row vector e.g., $\boldsymbol{y}=[1,2,3]^\top$ is a column vector.

The transpose of a matrix $\boldsymbol{A}$ will be denoted by $\boldsymbol{A}^\top$ and its inverse will be denoted by $\boldsymbol{A}^{-1}$. We will denote the $p\times p$ identity matrix by $\boldsymbol{I}_p$. A vector or matrix of all zeros will be denoted by a **boldface** zero $\mathbf{0}$ whose size should be clear from context, or we denote $\mathbf{0}_p$ to be the vector of all zeros with p entries.

Definition 0.1: (Eigenvalue)

Given any vector space E and any linear map $A: E\rightarrow E$, a scalar $\lambda\in K$ is called an eigenvalue, or proper value, or characteristic value of $\boldsymbol{A}$ if there is some nonzero vector $\boldsymbol{u}\in E$ such that

$$\boldsymbol{A}\boldsymbol{u}=\lambda\boldsymbol{u}.$$

Definition 0.2: (Spectrum and Spectral Radius)

The set of all eigenvalues of $\boldsymbol{A}$ is called the spectrum of $\boldsymbol{A}$ and denoted by $\Lambda(\boldsymbol{A})$. The largest magnitude of the eigenvalues is known as the spectral radius $\rho(\boldsymbol{A})$:

$$\rho(\boldsymbol{A})=\max_{\lambda\in\Lambda(\boldsymbol{A})}|\lambda|.$$

Definition 0.3: (Eigenvector)

A vector $\boldsymbol{u}\in E$ is called an eigenvector, or proper vector, or characteristic vector of $\boldsymbol{A}$ if $\boldsymbol{u}\neq 0$ and if there is some $\lambda\in K$ such that

$$\boldsymbol{A}\boldsymbol{u}=\lambda\boldsymbol{u},$$

where the scalar λ is then an eigenvalue. And we say that $\boldsymbol{u}$ is an eigenvector associated with λ.

Moreover, the tuple $(\lambda,\boldsymbol{u})$ above is said to be an **eigenpair**. Intuitively, the above definitions mean that multiplying matrix $\boldsymbol{A}$ by the vector $\boldsymbol{u}$ results in a new vector that is in the same direction as $\boldsymbol{u}$, but only scaled by a factor λ. For any eigenvector $\boldsymbol{u}$, we can scale it by a scalar s such that $s\boldsymbol{u}$ is still an eigenvector of $\boldsymbol{A}$. That's why we call the eigenvector as an eigenvector of $\boldsymbol{A}$ associated with eigenvalue λ. To avoid ambiguity, we usually assume that the eigenvector is normalized to have length 1 and the first entry is positive (or negative) since both $\boldsymbol{u}$ and $-\boldsymbol{u}$ are eigenvectors.

In this context, we will highly use the idea about the linear independence of a set of vectors. Two equivalent definitions are given as follows.

Definition 0.4: (Linearly Independent)

A set of vectors $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\}$ is called linearly independent if there is no combination can get $x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \ldots + x_m\boldsymbol{a}_m = 0$ except all x_i's are zero. An equivalent definition is that $\boldsymbol{a}_1 \neq \mathbf{0}$, and for every $k > 1$, the vector $\boldsymbol{a}_k$ does not belong to the span of $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{k-1}\}$.

In the study of linear algebra, every vector space has a basis and every vector is a linear combination of members of the basis. We then define the span and dimension of a subspace via the basis.

Definition 0.5: (Span)

If every vector $\boldsymbol{v}$ in subspace $\mathcal{V}$ can be expressed as a linear combination of $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\}$, then $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\}$ is said to span $\mathcal{V}$.

Definition 0.6: (Subspace)

A nonempty subset $\mathcal{V}$ of $\mathbb{R}^n$ is called a subspace if $x\boldsymbol{a} + y\boldsymbol{a} \in \mathcal{V}$ for every $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{V}$ and every $x, y \in \mathbb{R}$.

Definition 0.7: (Basis and Dimension)

A set of vectors $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m\}$ is called a basis of $\mathcal{V}$ if they are linearly independent and span $\mathcal{V}$. Every basis of a given subspace has the same number of vectors, and the number of vectors in any basis is called the dimension of the subspace $\mathcal{V}$. By convention, the subspace $\{\mathbf{0}\}$ is said to have dimension zero. Furthermore, every subspace of nonzero dimension has a basis that is orthogonal, i.e., the basis of a subspace can be chosen orthogonal.

Definition 0.8: (Column Space (Range))

If $\boldsymbol{A}$ is an $m \times n$ real matrix, we define the column space (or range) of $\boldsymbol{A}$ to be the set spanned by its columns:

$$\mathcal{C}(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^m : \exists \boldsymbol{x} \in \mathbb{R}^n, \, \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}\}.$$

And the row space of $\boldsymbol{A}$ is the set spanned by its rows, which is equal to the column space of $\boldsymbol{A}^\top$:

$$\mathcal{C}(\boldsymbol{A}^\top) = \{\boldsymbol{x} \in \mathbb{R}^n : \exists \boldsymbol{y} \in \mathbb{R}^m, \, \boldsymbol{x} = \boldsymbol{A}^\top\boldsymbol{y}\}.$$

Definition 0.9: (Null Space (Nullspace, Kernel))

If $\boldsymbol{A}$ is an $m \times n$ real matrix, we define the null space (or kernel, or nullspace) of $\boldsymbol{A}$ to be the set:

$$\mathcal{N}(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^n : \, \boldsymbol{A}\boldsymbol{y} = \mathbf{0}\}.$$

And the null space of $\boldsymbol{A}^\top$ is defined as

$$\mathcal{N}(\boldsymbol{A}^\top) = \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{A}^\top \boldsymbol{x} = \boldsymbol{0}\}.$$

Both the column space of $\boldsymbol{A}$ and the null space of $\boldsymbol{A}^\top$ are subspaces of $\mathbb{R}^n$. In fact, every vector in $\mathcal{N}(\boldsymbol{A}^\top)$ is perpendicular to $\mathcal{C}(\boldsymbol{A})$ and vice versa.[1]

Definition 0.10: (Rank)

The *rank* of a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the dimension of the column space of $\boldsymbol{A}$. That is, the rank of $\boldsymbol{A}$ is equal to the maximal number of linearly independent columns of $\boldsymbol{A}$, and is also the maximal number of linearly independent rows of $\boldsymbol{A}$. The matrix $\boldsymbol{A}$ and its transpose $\boldsymbol{A}^\top$ have the same rank. We say that $\boldsymbol{A}$ has full rank, if its rank is equal to $min\{m, n\}$. In another word, this is true if and only if either all the columns of $\boldsymbol{A}$ are linearly independent, or all the rows of $\boldsymbol{A}$ are linearly independent. Specifically, given a vector $\boldsymbol{u} \in \mathbb{R}^m$ and a vector $\boldsymbol{v} \in \mathbb{R}^n$, then the $m \times n$ matrix $\boldsymbol{u}\boldsymbol{v}^\top$ is of rank 1. In short, the rank of a matrix is equal to:

- number of linearly independent columns;
- number of linearly independent rows;
- and remarkably, these are always the same (see Corollary 0.13, p. 12).

Definition 0.11: (Orthogonal Complement in General)

The orthogonal complement $\mathcal{V}^\perp$ of a subspace $\mathcal{V}$ contains every vector that is perpendicular to $\mathcal{V}$. That is,

$$\mathcal{V}^\perp = \{\boldsymbol{v} | \boldsymbol{v}^\top \boldsymbol{u} = 0, \ \ \forall \boldsymbol{u} \in \mathcal{V}\}.$$

The two subspaces are disjoint that span the entire space. The dimensions of $\mathcal{V}$ and $\mathcal{V}^\perp$ add to the dimension of the whole space. Furthermore, $(\mathcal{V}^\perp)^\perp = \mathcal{V}$.

Definition 0.12: (Orthogonal Complement of Column Space)

If $\boldsymbol{A}$ is an $m \times n$ real matrix, the orthogonal complement of $\mathcal{C}(\boldsymbol{A})$, $\mathcal{C}^\perp(\boldsymbol{A})$ is the subspace defined as:

$$\begin{aligned}\mathcal{C}^\perp(\boldsymbol{A}) &= \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}, \forall \boldsymbol{x} \in \mathbb{R}^n\} \\ &= \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{y}^\top \boldsymbol{v} = \boldsymbol{0}, \forall \boldsymbol{v} \in \mathcal{C}(\boldsymbol{A})\}.\end{aligned}$$

Then we have the four fundamental spaces for any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with rank r as shown in Theorem 0.15.

Theorem 0.13: (Row Rank Equals Column Rank)

1. Every vector in $\mathcal{N}(\boldsymbol{A})$ is also perpendicular to $\mathcal{C}(\boldsymbol{A}^\top)$ and vice versa.

The dimension of the column space of a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is equal to the dimension of its row space, i.e., the row rank and the column rank of a matrix $\boldsymbol{A}$ are equal.

Proof [**of Theorem 0.13, A First Way**] We first notice that the null space of $\boldsymbol{A}$ is orthogonal complementary to the row space of $\boldsymbol{A}$: $\mathcal{N}(\boldsymbol{A})\bot\mathcal{C}(\boldsymbol{A}^\top)$ (where the row space of $\boldsymbol{A}$ is exactly the column space of $\boldsymbol{A}^\top$), that is, vectors in the null space of $\boldsymbol{A}$ are orthogonal to vectors in the row space of $\boldsymbol{A}$. To see this, suppose $\boldsymbol{A}$ has rows $\boldsymbol{a}_1^\top, \boldsymbol{a}_2^\top, \ldots, \boldsymbol{a}_m^\top$ and $\boldsymbol{A} = [\boldsymbol{a}_1^\top; \boldsymbol{a}_2^\top; \ldots; \boldsymbol{a}_m^\top]$. For any vector $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{A})$, we have $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$, that is, $[\boldsymbol{a}_1^\top\boldsymbol{x}; \boldsymbol{a}_2^\top\boldsymbol{x}; \ldots; \boldsymbol{a}_m^\top\boldsymbol{x}] = \boldsymbol{0}$. And since the row space of $\boldsymbol{A}$ is spanned by $\boldsymbol{a}_1^\top, \boldsymbol{a}_2^\top, \ldots, \boldsymbol{a}_m^\top$. Then $\boldsymbol{x}$ is perpendicular to any vectors from $\mathcal{C}(\boldsymbol{A}^\top)$ which means $\mathcal{N}(\boldsymbol{A})\bot\mathcal{C}(\boldsymbol{A}^\top)$.

Now suppose, the dimension of row space of $\boldsymbol{A}$ is r. Let $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r$ be a set of vectors in $\mathbb{R}^n$ and form a basis for the row space. Then the r vectors $\boldsymbol{A}\boldsymbol{r}_1, \boldsymbol{A}\boldsymbol{r}_2, \ldots, \boldsymbol{A}\boldsymbol{r}_r$ are in the column space of $\boldsymbol{A}$, which are linearly independent. To see this, suppose we have a linear combination of the r vectors: $x_1\boldsymbol{A}\boldsymbol{r}_1 + x_2\boldsymbol{A}\boldsymbol{r}_2 + \ldots + x_r\boldsymbol{A}\boldsymbol{r}_r = \boldsymbol{0}$, that is, $\boldsymbol{A}(x_1\boldsymbol{r}_1 + x_2\boldsymbol{r}_2 + \ldots + x_r\boldsymbol{r}_r) = \boldsymbol{0}$ and the vector $\boldsymbol{v} = x_1\boldsymbol{r}_1 + x_2\boldsymbol{r}_2 + \ldots + x_r\boldsymbol{r}_r$ is in null space of $\boldsymbol{A}$. But since $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r$ is a basis for the row space of $\boldsymbol{A}$, $\boldsymbol{v}$ is thus also in the row space of $\boldsymbol{A}$. We have shown that vectors from null space of $\boldsymbol{A}$ is perpendicular to vectors from row space of $\boldsymbol{A}$, thus $\boldsymbol{v}^\top\boldsymbol{v} = 0$ and $x_1 = x_2 = \ldots = x_r = 0$. Then $\boldsymbol{A}\boldsymbol{r}_1, \boldsymbol{A}\boldsymbol{r}_2, \ldots, \boldsymbol{A}\boldsymbol{r}_r$ are in the column space of $\boldsymbol{A}$ and they are linearly independent which means the dimension of the column space of $\boldsymbol{A}$ is larger than r. This result shows that **row rank of $\boldsymbol{A}$ $\leq$ column rank of $\boldsymbol{A}$**.

If we apply this process again for $\boldsymbol{A}^\top$. We will have **column rank of $\boldsymbol{A}$ $\leq$ row rank of $\boldsymbol{A}$**. This completes the proof. ■

Further information can be drawn from this proof is that if $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r$ is a set of vectors in $\mathbb{R}^n$ that forms a basis for the row space, then $\boldsymbol{A}\boldsymbol{r}_1, \boldsymbol{A}\boldsymbol{r}_2, \ldots, \boldsymbol{A}\boldsymbol{r}_r$ is a basis for the column space of $\boldsymbol{A}$. We formulate this finding into the following lemma.

Lemma 0.14: (Column Basis from Row Basis)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, suppose that $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r\}$ is a set of vectors in $\mathbb{R}^n$ which forms a basis for the row space, then $\{\boldsymbol{A}\boldsymbol{r}_1, \boldsymbol{A}\boldsymbol{r}_2, \ldots, \boldsymbol{A}\boldsymbol{r}_r\}$ is a basis for the column space of $\boldsymbol{A}$.

For any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, it can be easily verified that any vector in the row space of $\boldsymbol{A}$ is perpendicular to any vector in the null space of $\boldsymbol{A}$. Suppose $\boldsymbol{x}_n \in \mathcal{N}(\boldsymbol{A})$, then $\boldsymbol{A}\boldsymbol{x}_n = \boldsymbol{0}$ such that $\boldsymbol{x}_n$ is perpendicular to every row of $\boldsymbol{A}$ which agrees with our claim.

Similarly, we can also show that any vector in the column space of $\boldsymbol{A}$ is perpendicular to any vector in the null space of $\boldsymbol{A}^\top$. Further, the column space of $\boldsymbol{A}$ together with the null space of $\boldsymbol{A}^\top$ span the whole $\mathbb{R}^m$ which is known as the fundamental theorem of linear algebra.

The fundamental theorem contains two parts, the dimension of the subspaces and the orthogonality of the subspaces. The orthogonality can be easily verified as shown above. Moreover, when the row space has dimension r, the null space has dimension $n - r$. This cannot be easily stated and we prove in the following theorem.

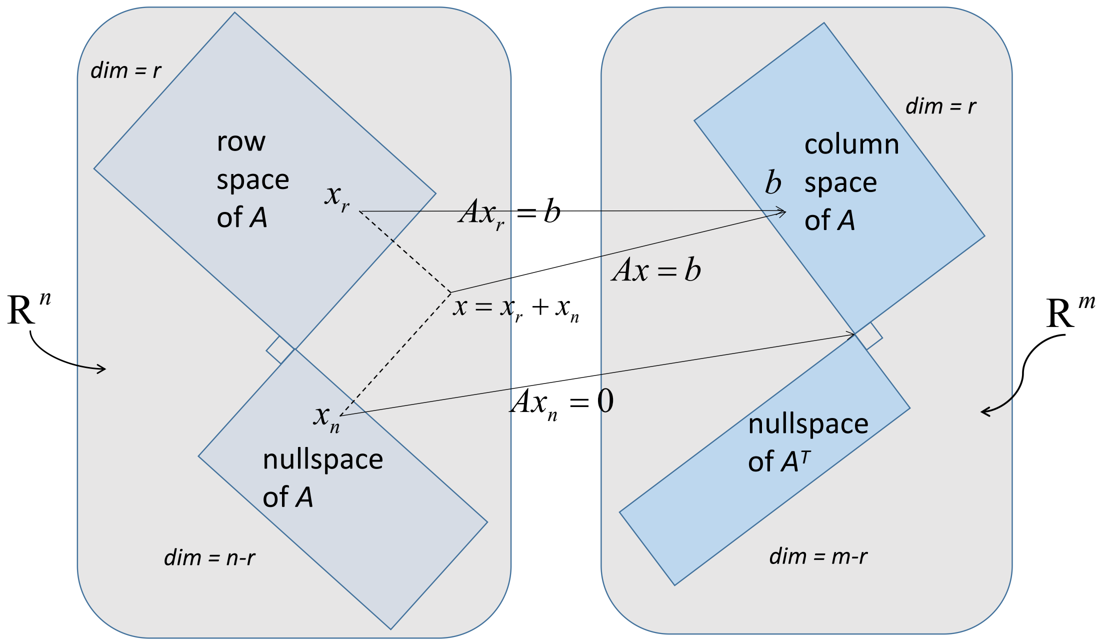


Figure 3: Two pairs of orthogonal subspaces in $\mathbb{R}^n$ and $\mathbb{R}^m$. $\dim(\mathcal{C}(\boldsymbol{A}^\top))+\dim(\mathcal{N}(\boldsymbol{A})) = n$ and $\dim(\mathcal{N}(\boldsymbol{A}^\top)) + \dim(\mathcal{C}(\boldsymbol{A})) = m$. The null space component goes to zero as $\boldsymbol{A}\boldsymbol{x}_n = \boldsymbol{0} \in \mathbb{R}^m$. The row space component goes to column space as $\boldsymbol{A}\boldsymbol{x}_r = \boldsymbol{A}(\boldsymbol{x}_r + \boldsymbol{x}_n) = \boldsymbol{b} \in \mathcal{C}(\boldsymbol{A})$.

Theorem 0.15: (The Fundamental Theorem of Linear Algebra)

Orthogonal Complement and Rank-Nullity Theorem: for any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we have

- $\mathcal{N}(\boldsymbol{A})$ is orthogonal complement to the row space $\mathcal{C}(\boldsymbol{A}^\top)$ in $\mathbb{R}^n$: $dim(\mathcal{N}(\boldsymbol{A})) + dim(\mathcal{C}(\boldsymbol{A}^\top)) = n$;
- $\mathcal{N}(\boldsymbol{A}^\top)$ is orthogonal complement to the column space $\mathcal{C}(\boldsymbol{A})$ in $\mathbb{R}^m$: $dim(\mathcal{N}(\boldsymbol{A}^\top)) + dim(\mathcal{C}(\boldsymbol{A})) = m$;
- For rank-r matrix $\boldsymbol{A}$, $dim(\mathcal{C}(\boldsymbol{A}^\top)) = dim(\mathcal{C}(\boldsymbol{A})) = r$, that is, $dim(\mathcal{N}(\boldsymbol{A})) = n-r$ and $dim(\mathcal{N}(\boldsymbol{A}^\top)) = m-r$.

Proof [of Theorem 0.15] Following from the proof of Theorem 0.13. Let $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r$ be a set of vectors in $\mathbb{R}^n$ that form a basis for the row space, then $\boldsymbol{A}\boldsymbol{r}_1, \boldsymbol{A}\boldsymbol{r}_2, \ldots, \boldsymbol{A}\boldsymbol{r}_r$ is a basis for the column space of $\boldsymbol{A}$. Let $\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_k \in \mathbb{R}^n$ form a basis for the null space of $\boldsymbol{A}$. Following again from the proof of Theorem 0.13, $\mathcal{N}(\boldsymbol{A}) \perp \mathcal{C}(\boldsymbol{A}^\top)$, thus, $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r$ are perpendicular to $\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_k$. Then, $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r, \boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_k\}$ is linearly independent in $\mathbb{R}^n$.

For any vector $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{A}\boldsymbol{x}$ is in the column space of $\boldsymbol{A}$. Then it can be expressed by a combination of $\boldsymbol{A}\boldsymbol{r}_1, \boldsymbol{A}\boldsymbol{r}_2, \ldots, \boldsymbol{A}\boldsymbol{r}_r$: $\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^r a_i \boldsymbol{A}\boldsymbol{r}_i$ which states that $\boldsymbol{A}(\boldsymbol{x} - \sum_{i=1}^r a_i \boldsymbol{r}_i) = \boldsymbol{0}$ and $\boldsymbol{x} - \sum_{i=1}^r a_i \boldsymbol{r}_i$ is thus in $\mathcal{N}(\boldsymbol{A})$. Since $\{\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_k\}$ is a basis for the null space of $\boldsymbol{A}$, $\boldsymbol{x}-\sum_{i=1}^{r} a_i\boldsymbol{r}_i$ can be expressed by a combination of $\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_k$: $\boldsymbol{x}-\sum_{i=1}^{r} a_i\boldsymbol{r}_i = \sum_{j=1}^{k} b_j\boldsymbol{n}_j$, i.e., $\boldsymbol{x} = \sum_{i=1}^{r} a_i\boldsymbol{r}_i + \sum_{j=1}^{k} b_j\boldsymbol{n}_j$. That is, any vector $\boldsymbol{x} \in \mathbb{R}^n$ can be expressed by $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r, \boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_k\}$ and the set forms a basis for $\mathbb{R}^n$. Thus the dimension sum to n: $r+k=n$, i.e., $dim(\mathcal{N}(\boldsymbol{A})) + dim(\mathcal{C}(\boldsymbol{A}^\top)) = n$. Similarly, we can prove $dim(\mathcal{N}(\boldsymbol{A}^\top)) + dim(\mathcal{C}(\boldsymbol{A})) = m$. ■

Figure 3 demonstrates two pairs of such orthogonal subspaces and shows how $\boldsymbol{A}$ takes $\boldsymbol{x}$ into the column space. The dimensions of the row space of $\boldsymbol{A}$ and the null space of $\boldsymbol{A}$ add to n. And the dimensions of the column space of $\boldsymbol{A}$ and the null space of $\boldsymbol{A}^\top$ add to m. The null space component goes to zero as $\boldsymbol{A}\boldsymbol{x_n} = \boldsymbol{0} \in \mathbb{R}^m$ which is the intersection of column space of $\boldsymbol{A}$ and null space of $\boldsymbol{A}^\top$. The row space component goes to column space as $\boldsymbol{A}\boldsymbol{x_r} = \boldsymbol{A}(\boldsymbol{x_r}+\boldsymbol{x_n}) = \boldsymbol{b} \in \mathbb{R}^m$.

Definition 0.16: (Orthogonal Matrix)

A real square matrix $\boldsymbol{Q}$ is an orthogonal matrix if the inverse of $\boldsymbol{Q}$ equals the transpose of $\boldsymbol{Q}$, that is $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\top$ and $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$. In another word, suppose $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n]$ where $\boldsymbol{q}_i \in \mathbb{R}^n$ for all $i \in \{1, 2, \ldots, n\}$, then $\boldsymbol{q}_i^\top\boldsymbol{q}_j = \delta(i,j)$ with $\delta(i,j)$ being the Kronecker delta function. If $\boldsymbol{Q}$ contains only γ of these columns with $\gamma < n$, then $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_\gamma$ stills holds with $\boldsymbol{I}_\gamma$ being the $\gamma\times\gamma$ identity matrix. But $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$ will not be true. For any vector $\boldsymbol{x}$, the orthogonal matrix will preserve the length: $||\boldsymbol{Q}\boldsymbol{x}|| = ||\boldsymbol{x}||$.

Definition 0.17: (Permutation Matrix)

A permutation matrix $\boldsymbol{P}$ is a square binary matrix that has exactly one entry of 1 in each row and each column and 0's elsewhere.

Row Point That is, the permutation matrix $\boldsymbol{P}$ has the rows of the identity $\boldsymbol{I}$ in any order and the order decides the sequence of the row permutation. Suppose we want to permute the rows of matrix $\boldsymbol{A}$, we just multiply on the left by $\boldsymbol{P}\boldsymbol{A}$.

Column Point Or, equivalently, the permutation matrix $\boldsymbol{P}$ has the columns of the identity $\boldsymbol{I}$ in any order and the order decides the sequence of the column permutation. And now, the column permutation of $\boldsymbol{A}$ is to multiply on the right by $\boldsymbol{A}\boldsymbol{P}$.

The permutation matrix $\boldsymbol{P}$ can be more efficiently represented via a vector $J \in \mathbb{Z}_+^n$ of indices such that $\boldsymbol{P} = \boldsymbol{I}[:, J]$ where $\boldsymbol{I}$ is the $n\times n$ identity matrix and notably, the elements in vector J sum to $1+2+\ldots+n = \frac{n^2+n}{2}$.

Part I
Gaussian Elimination

1. LU Decomposition

Perhaps the best known and the first matrix decomposition we should know about is the LU decomposition. We now illustrate the results in the following theorem and the proof of the existence of which will be delayed in the next sections.

Theorem 1.1: (LU Decomposition with Permutation)

Every nonsingular $n \times n$ square matrix $\boldsymbol{A}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{PLU},$$

where $\boldsymbol{P}$ is a permutation matrix, $\boldsymbol{L}$ is a unit lower triangular matrix (i.e., lower triangular matrix with all 1's on the diagonal), and $\boldsymbol{U}$ is a nonsingular upper triangular matrix.

Note that, in the remainder of this text, we will put the decomposition-related results in the blue box. And other claims will be in a gray box. This rule will be applied for the rest of the survey without special mention.

Remark 1.2: (Decomposition Notation)

The above decomposition applies to any nonsingular matrix $\boldsymbol{A}$. We will see that this decomposition arises from the elimination steps in which case row operations of subtraction and exchange of two rows are allowed where the subtractions are recorded in matrix $\boldsymbol{L}$ and the row exchanges are recorded in matrix $\boldsymbol{P}$. To make this row exchange explicit, the common form for the above decomposition is $\boldsymbol{QA} = \boldsymbol{LU}$ where $\boldsymbol{Q} = \boldsymbol{P}^\top$ that records the exact row exchanges of the rows of $\boldsymbol{A}$. Otherwise, the $\boldsymbol{P}$ would record the row exchanges of $\boldsymbol{LU}$. In our case, we will make the decomposition to be clear for matrix $\boldsymbol{A}$ rather than for $\boldsymbol{QA}$. For this reason, we will put the permutation matrix on the right-hand side of the equation for the remainder of the text without special mention.

Specifically, in some cases, we will not need the permutation matrix. This decomposition relies on the leading principal minors. We provide the definition which is important for the illustration.

Definition 1.3: (Leading Principal Minors)

Let $\boldsymbol{A}$ be an $n \times n$ square matrix. A $k \times k$ submatrix of $\boldsymbol{A}$ obtained by deleting the **last** $n-k$ columns and the **last** $n-k$ rows from $\boldsymbol{A}$ is called a k-th order **leading principal submatrix** of $\boldsymbol{A}$, that is, the $k \times k$ submatrix taken from the top left corner of $\boldsymbol{A}$. The determinant of the $k \times k$ leading principal submatrix is called a k-th order **leading principal minor** of $\boldsymbol{A}$.

Under mild conditions on the leading principal minors of matrix $\boldsymbol{A}$, the LU decomposition will not involve the permutation matrix.

Theorem 1.4: (LU Decomposition without Permutation)

For any $n \times n$ square matrix $\boldsymbol{A}$, if all the leading principal minors are nonzero, i.e., $\det(\boldsymbol{A}_{1:k,1:k}) \neq 0$, for all $k \in \{1, 2, \ldots, n\}$, then $\boldsymbol{A}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U},$$

where $\boldsymbol{L}$ is a unit lower triangular matrix (i.e., lower triangular matrix with all 1's on the diagonal), and $\boldsymbol{U}$ is a **nonsingular** upper triangular matrix.

Specifically, this decomposition is **unique**. See Corollary 1.8.

Remark 1.5: (Other Forms of the LU Decomposition without Permutation)

The leading principal minors are nonzero, in another word, means the leading principal submatrices are nonsingular.

Singular $\boldsymbol{A}$ In the above theorem, we assume $\boldsymbol{A}$ is nonsingular as well. The LU decomposition also exists for singular matrix $\boldsymbol{A}$. However, the matrix $\boldsymbol{U}$ will be singular as well in this case. This can be shown in the following section that, if matrix $\boldsymbol{A}$ is singular, some pivots will be zero, and the corresponding diagonal values of $\boldsymbol{U}$ will be zero.

Singular leading principal submatrices Even if we assume matrix $\boldsymbol{A}$ is nonsingular, the leading principal submatrices might be singular. Suppose further that some of the leading principal minors are zero, the LU decomposition also exists, but if so, it is again not unique.

We will discuss where this decomposition comes from in the next section. There are also generalizations of LU decomposition to non-square or singular matrices, such as rank-revealing LU decomposition. Please refer to (Pan, 2000; Miranian and Gu, 2003; Dopico et al., 2006) or we will have a short discussion in Section 1.10.

1.1 Relation to Gaussian Elimination

Solving linear system equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is the basic problem in linear algebra. Gaussian elimination transforms a linear system into an upper triangular one by applying simple *elementary row transformations* on the left of the linear system in $n-1$ stages if $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. As a result, it is much easier to solve by a backward substitution. The elementary transformation is defined rigorously as follows.

Definition 1.6: (Elementary Transformation)

For square matrix $\boldsymbol{A}$, the following three transformations are referred as **elementary row/column transformations**:

1. Interchanging two rows (or columns) of $\boldsymbol{A}$;
2. Multiplying all elements of a row (or a column) of $\boldsymbol{A}$ by some nonzero number;

3. Adding any row (or column) of $\boldsymbol{A}$ multiplied by a nonzero number to any other row (or column);

Specifically, the elementary row transformations of $\boldsymbol{A}$ are unit lower triangular to multiply $\boldsymbol{A}$ on the left, and the elementary column transformations of $\boldsymbol{A}$ are unit upper triangular to multiply $\boldsymbol{A}$ on the right.

The Gaussian elimination is described by the third type - elementary row transformation above. Suppose the upper triangular matrix obtained by Gaussian elimination is given by $\boldsymbol{U} = \boldsymbol{E}_{n-1}\boldsymbol{E}_{n-2}\ldots\boldsymbol{E}_1\boldsymbol{A}$, and in the k-th stage, the k-th column of $\boldsymbol{E}_{k-1}\boldsymbol{E}_{k-2}\ldots\boldsymbol{E}_1\boldsymbol{A}$ is $\boldsymbol{x} \in \mathbb{R}^n$. Gaussian elimination will introduce zeros below the diagonal of $\boldsymbol{x}$ by

$$\boldsymbol{E}_k = \boldsymbol{I} - \boldsymbol{z}_k\boldsymbol{e}_k^\top,$$

where $\boldsymbol{e}_k \in \mathbb{R}^n$ is the k-th unit basis vector, and $\boldsymbol{z}_k \in \mathbb{R}^n$ is given by

$$\boldsymbol{z}_k = [0, \ldots, 0, z_{k+1}, \ldots, z_n]^\top, \qquad z_i = \frac{x_i}{x_k}, \quad \forall i \in \{k+1, \ldots, n\}.$$

We realize that $\boldsymbol{E}_k$ is a unit lower triangular matrix (with 1's on the diagonal) with only the k-th column of the lower submatrix being nonzero,

$$\boldsymbol{E}_k = \begin{bmatrix} 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & -z_{k+1} & 1 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & -z_n & 0 & \ldots & 1 \end{bmatrix},$$

and multiplying on the left by $\boldsymbol{E}_k$ will introduce zeros below the diagonal:

$$\boldsymbol{E}_k\boldsymbol{x} = \begin{bmatrix} 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & -z_{k+1} & 1 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & -z_n & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

For example, we write out the Gaussian elimination steps for a 4×4 matrix. For simplicity, we assume there are no row permutations. And in the following matrix, $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

A Trivial Gaussian Elimination For a 4×4 Matrix:

$$\underset{\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}} \xrightarrow{\boldsymbol{E}_1} \underset{\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}} \xrightarrow{\boldsymbol{E}_2} \underset{\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \boxtimes & \boxtimes & \boxtimes \\ 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}} \xrightarrow{\boldsymbol{E}_3} \underset{\boldsymbol{E}_3\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \boxtimes & \boxtimes & \boxtimes \\ 0 & 0 & \boxtimes & \boxtimes \\ 0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} \end{bmatrix}}, \quad (1.1)$$

where $\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3$ are lower triangular matrices. Specifically, as discussed above, Gaussian transformation matrices $\boldsymbol{E}_i$'s are unit lower triangular matrices with 1's on the diagonal. This can be explained that for the k-th transformation $\boldsymbol{E}_k$, working on the matrix $\boldsymbol{E}_{k-1} \ldots \boldsymbol{E}_1 \boldsymbol{A}$, the transformation subtracts multiples of the k-th row from rows $\{k+1, k+2, \ldots, n\}$ to get zeros below the diagonal in the k-th column of the matrix. And never use rows $\{1, 2, \ldots, k-1\}$.

For the transformation example above, at step 1, we multiply on the left by $\boldsymbol{E}_1$ so that multiples of the 1-st row are subtracted from rows $2, 3, 4$ and the first entries of rows $2, 3, 4$ are set to zero. Similar situations for step 2 and step 3. By setting $\boldsymbol{L} = \boldsymbol{E}_1^{-1}\boldsymbol{E}_2^{-1}\boldsymbol{E}_3^{-1}$ and letting the matrix after elimination be $\boldsymbol{U}$, [2] we get $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$. Thus we obtain an LU decomposition for this 4×4 matrix $\boldsymbol{A}$.

Definition 1.7: (Pivot)

First nonzero entry in the row after each elimination step is called a **pivot**. For example, the blue crosses in Equation (1.1) are pivots.

But sometimes, it can happen that the value of $\boldsymbol{A}_{11}$ is zero. No $\boldsymbol{E}_1$ can make the next elimination step successful. So we need to interchange the first row and the second row via a permutation matrix $\boldsymbol{P}_1$. This is known as the **pivoting**, or **permutation**.

Gaussian Elimination With a Permutation In the Beginning:

$$
\underset{\boldsymbol{A}}{\begin{bmatrix} 0 & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}}
\xrightarrow{\boldsymbol{P}_1}
\underset{\boldsymbol{P}_1\boldsymbol{A}}{\begin{bmatrix} \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}}
\xrightarrow{\boldsymbol{E}_1}
\underset{\boldsymbol{E}_1\boldsymbol{P}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \mathbf{0} & {\color{blue}\boldsymbol{\boxtimes}} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}}
$$

$$
\xrightarrow{\boldsymbol{E}_2}
\underset{\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{P}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & {\color{blue}\boxtimes} & \boxtimes & \boxtimes \\ 0 & \mathbf{0} & {\color{blue}\boldsymbol{\boxtimes}} & \boldsymbol{\boxtimes} \\ 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}}
\xrightarrow{\boldsymbol{E}_3}
\underset{\boldsymbol{E}_3\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{P}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & {\color{blue}\boxtimes} & \boxtimes & \boxtimes \\ 0 & 0 & {\color{blue}\boxtimes} & \boxtimes \\ 0 & 0 & \mathbf{0} & {\color{blue}\boldsymbol{\boxtimes}} \end{bmatrix}}.
$$

By setting $\boldsymbol{L} = \boldsymbol{E}_1^{-1}\boldsymbol{E}_2^{-1}\boldsymbol{E}_3^{-1}$ and $\boldsymbol{P} = \boldsymbol{P}_1^{-1}$, we get $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U}$. Therefore we obtain a full LU decomposition with permutation for this 4×4 matrix $\boldsymbol{A}$.

In some situations, other permutation matrices $\boldsymbol{P}_2, \boldsymbol{P}_3, \ldots$ will appear in between the lower triangular $\boldsymbol{E}_i$'s. An example is shown as follows.

Gaussian Elimination With a Permutation In Between:

$$
\underset{\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}}
\xrightarrow{\boldsymbol{E}_1}
\underset{\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}}
\xrightarrow{\boldsymbol{P}_1}
\underset{\boldsymbol{P}_1\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \mathbf{0} & {\color{blue}\boldsymbol{\boxtimes}} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \mathbf{0} & {\color{blue}\boldsymbol{\boxtimes}} & \boldsymbol{\boxtimes} \\ 0 & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}}
\xrightarrow{\boldsymbol{E}_2}
\underset{\boldsymbol{E}_2\boldsymbol{P}_1\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & {\color{blue}\boxtimes} & \boxtimes & \boxtimes \\ 0 & 0 & {\color{blue}\boxtimes} & \boxtimes \\ 0 & \mathbf{0} & \mathbf{0} & {\color{blue}\boldsymbol{\boxtimes}} \end{bmatrix}}.
$$

2. The inverses of unit lower triangular matrices are also unit lower triangular matrices. And the products of unit lower triangular matrices are also unit lower triangular matrices.

In this case, we find $\boldsymbol{U} = \boldsymbol{E}_2\boldsymbol{P}_1\boldsymbol{E}_1\boldsymbol{A}$. In Section 1.3 or Section 1.9.1, we will show that the permutations in-between will result in the same form $\boldsymbol{A} = \boldsymbol{PLU}$ where $\boldsymbol{P}$ takes account of all the permutations.

The above examples can be easily extended to any $n \times n$ matrix if we assume there are no row permutations in the process. And we will have $n-1$ such lower triangular transformations. The k-th transformation $\boldsymbol{E}_k$ introduces zeros below the diagonal in the k-th column of $\boldsymbol{A}$ by subtracting multiples of the k-th row from rows $\{k+1, k+2, \ldots, n\}$. Finally, by setting $\boldsymbol{L} = \boldsymbol{E}_1^{-1}\boldsymbol{E}_2^{-1}\ldots\boldsymbol{E}_{n-1}^{-1}$ we obtain the LU decomposition $\boldsymbol{A} = \boldsymbol{LU}$ (without permutation).

1.2 Existence of the LU Decomposition without Permutation

The Gaussian elimination or Gaussian transformation shows the origin of the LU decomposition. We then prove Theorem 1.4 rigorously, i.e., the existence of the LU decomposition without permutation by induction.

Proof [**of Theorem 1.4: LU Decomposition without Permutation**] We will prove by induction that every $n \times n$ square matrix $\boldsymbol{A}$ with nonzero leading principal minors has a decomposition $\boldsymbol{A} = \boldsymbol{LU}$. The 1×1 case is trivial by setting $L = 1, U = A$, thus, $A = LU$.

Suppose for any $k \times k$ matrix $\boldsymbol{A}_k$ with all the leading principal minors being nonzero has an LU decomposition without permutation. If we prove any $(k+1) \times (k+1)$ matrix $\boldsymbol{A}_{k+1}$ can also be factored as this LU decomposition without permutation, then we complete the proof.

For any $(k+1)\times(k+1)$ matrix $\boldsymbol{A}_{k+1}$, suppose the k-th order leading principal submatrix of $\boldsymbol{A}_{k+1}$ is $\boldsymbol{A}_k$ with the size of $k \times k$. Then $\boldsymbol{A}_k$ can be factored as $\boldsymbol{A}_k = \boldsymbol{L}_k\boldsymbol{U}_k$ with $\boldsymbol{L}_k$ being a unit lower triangular matrix and $\boldsymbol{U}_k$ being a nonsingular upper triangular matrix from the assumption. Write out $\boldsymbol{A}_{k+1}$ as $\boldsymbol{A}_{k+1} = \begin{bmatrix} \boldsymbol{A}_k & \boldsymbol{b} \\ \boldsymbol{c}^\top & d \end{bmatrix}$. Then it admits the factorization:

$$\boldsymbol{A}_{k+1} = \begin{bmatrix} \boldsymbol{A}_k & \boldsymbol{b} \\ \boldsymbol{c}^\top & d \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_k & \boldsymbol{0} \\ \boldsymbol{x}^\top & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_k & \boldsymbol{y} \\ \boldsymbol{0} & z \end{bmatrix} = \boldsymbol{L}_{k+1}\boldsymbol{U}_{k+1},$$

where $\boldsymbol{b} = \boldsymbol{L}_k\boldsymbol{y}$, $\boldsymbol{c}^\top = \boldsymbol{x}^\top\boldsymbol{U}_k$, $d = \boldsymbol{x}^\top\boldsymbol{y} + z$, $\boldsymbol{L}_{k+1} = \begin{bmatrix} \boldsymbol{L}_k & \boldsymbol{0} \\ \boldsymbol{x}^\top & 1 \end{bmatrix}$, and $\boldsymbol{U}_{k+1} = \begin{bmatrix} \boldsymbol{U}_k & \boldsymbol{y} \\ \boldsymbol{0} & z \end{bmatrix}$. From the assumption, $\boldsymbol{L}_k$ and $\boldsymbol{U}_k$ are nonsingular. Therefore

$$\boldsymbol{y} = \boldsymbol{L}_k^{-1}\boldsymbol{b}, \qquad \boldsymbol{x}^\top = \boldsymbol{c}^\top\boldsymbol{U}_k^{-1}, \qquad z = d - \boldsymbol{x}^\top\boldsymbol{y}.$$

If further, we could prove z is nonzero such that $\boldsymbol{U}_{k+1}$ is nonsingular, we complete the proof.

Since all the leading principal minors of $\boldsymbol{A}_{k+1}$ are nonzero, we have $\det(\boldsymbol{A}_{k+1}) =$ [3] $\det(\boldsymbol{A}_k)\cdot \det(d - \boldsymbol{c}^\top\boldsymbol{A}_k^{-1}\boldsymbol{b}) = \det(\boldsymbol{A}_k)\cdot(d - \boldsymbol{c}^\top\boldsymbol{A}_k^{-1}\boldsymbol{b}) \neq 0$, where $d - \boldsymbol{c}^\top\boldsymbol{A}_k^{-1}\boldsymbol{b}$ is a scalar. As $\det(\boldsymbol{A}_k) \neq 0$ from the assumption, we obtain $d - \boldsymbol{c}^\top\boldsymbol{A}_k^{-1}\boldsymbol{b} \neq 0$. Substitute $\boldsymbol{b} = \boldsymbol{L}_k\boldsymbol{y}$ and $\boldsymbol{c}^\top = \boldsymbol{x}^\top\boldsymbol{U}_k$ into the formula, we have $d - \boldsymbol{x}^\top\boldsymbol{U}_k\boldsymbol{A}_k^{-1}\boldsymbol{L}_k\boldsymbol{y} = d - \boldsymbol{x}^\top\boldsymbol{U}_k(\boldsymbol{L}_k\boldsymbol{U}_k)^{-1}\boldsymbol{L}_k\boldsymbol{y} = d - \boldsymbol{x}^\top\boldsymbol{y} \neq 0$ which is exactly the form of $z \neq 0$. Thus we find $\boldsymbol{L}_{k+1}$ with all the values

3. By the fact that if matrix $\boldsymbol{M}$ has a block formulation: $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$, then $\det(\boldsymbol{M}) = \det(\boldsymbol{A})\det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})$.

on the diagonal being 1, and $\boldsymbol{U}_{k+1}$ with all the values on the diagonal being nonzero which means $\boldsymbol{L}_{k+1}$ and $\boldsymbol{U}_{k+1}$ are nonsingular, [4] from which the result follows. ■

We further prove that if no permutation involves, the LU decomposition is unique.

Corollary 1.8: (Uniqueness of the LU Decomposition without Permutation)

Suppose the $n \times n$ square matrix $\boldsymbol{A}$ has nonzero leaning principal minors. Then, the LU decomposition is unique.

Proof [of Corollary 1.8] Suppose the LU decomposition is not unique, then we can find two decompositions such that $\boldsymbol{A} = \boldsymbol{L}_1\boldsymbol{U}_1 = \boldsymbol{L}_2\boldsymbol{U}_2$ which implies $\boldsymbol{L}_2^{-1}\boldsymbol{L}_1 = \boldsymbol{U}_2\boldsymbol{U}_1^{-1}$. The left of the equation is a unit lower triangular matrix and the right of the equation is an upper triangular matrix. This implies both sides of the above equation are diagonal matrices. Since the inverse of a unit lower triangular matrix is also a unit lower triangular matrix, and the product of unit lower triangular matrices is also a unit lower triangular matrix, this results in that $\boldsymbol{L}_2^{-1}\boldsymbol{L}_1 = \boldsymbol{I}$. The equality implies that both sides are identity such that $\boldsymbol{L}_1 = \boldsymbol{L}_2$ and $\boldsymbol{U}_1 = \boldsymbol{U}_2$ and leads to a contradiction. ■

In the proof of Theorem 1.4, we have shown that the diagonal values of the upper triangular matrix are all nonzero if the leading principal minors of $\boldsymbol{A}$ are all nonzero. We then can formulate this decomposition in another form if we divide each row of $\boldsymbol{U}$ by each diagonal value of $\boldsymbol{U}$. This is called the *LDU decomposition*.

Corollary 1.9: (LDU Decomposition)

For any $n \times n$ square matrix $\boldsymbol{A}$, if all the leading principal minors are nonzero, i.e., $\det(\boldsymbol{A}_{1:k,1:k}) \neq 0$, for all $k \in \{1, 2, \ldots, n\}$, then $\boldsymbol{A}$ can be **uniquely** factored as

$$\boldsymbol{A} = \boldsymbol{LDU},$$

where $\boldsymbol{L}$ is a unit lower triangular matrix, $\boldsymbol{U}$ is a **unit** upper triangular matrix, and $\boldsymbol{D}$ is a diagonal matrix.

The proof is trivial that from the LU decomposition of $\boldsymbol{A} = \boldsymbol{LR}$, we can find a diagonal matrix $\boldsymbol{D} = diag(\boldsymbol{R}_{11}, \boldsymbol{R}_{22}, \ldots, \boldsymbol{R}_{nn})$ such that $\boldsymbol{D}^{-1}\boldsymbol{R} = \boldsymbol{U}$ is a unit upper triangular matrix. And the uniqueness comes from the uniqueness of the LU decomposition.

1.3 Existence of the LU Decomposition with Permutation

In Theorem 1.4, we require that matrix $\boldsymbol{A}$ has nonzero leading principal minors. However, this is not necessarily. Even when the leading principal minors are zero, nonsingular matrices still have an LU decomposition, but with an additional permutation. The proof is still from induction.

4. A triangular matrix (upper or lower) is nonsingular if and only if all the entries on its main diagonal are nonzero.

Proof [**of Theorem 1.1: LU Decomposition with Permutation**] We note that any 1×1 nonsingular matrix has a full LU decomposition $A = PLU$ by simply setting $P = 1$, $L = 1$, $U = A$. We will show that if every $(n-1) \times (n-1)$ nonsingular matrix has a full LU decomposition, then this is also true for every $n \times n$ nonsingular matrix. By induction, we prove that every nonsingular matrix has a full LU decomposition.

We will formulate the proof in the following order. If $\boldsymbol{A}$ is nonsingular, then its row permuted matrix $\boldsymbol{B}$ is also nonsingular. And Schur complement of $\boldsymbol{B}_{11}$ in $\boldsymbol{B}$ is also nonsingular. Finally, we formulate the decomposition of $\boldsymbol{A}$ by $\boldsymbol{B}$ from this property.

We notice that at least one element in the first column of $\boldsymbol{A}$ must be nonzero otherwise $\boldsymbol{A}$ will be singular. We can then apply a row permutation that makes the element in entry $(1, 1)$ to be nonzero. That is, there exists a permutation $\boldsymbol{P}_1$ such that $\boldsymbol{B} = \boldsymbol{P}_1\boldsymbol{A}$ in which case $\boldsymbol{B}_{11} \neq 0$. Since $\boldsymbol{A}$ and $\boldsymbol{P}_1$ are both nonsingular and the product of nonsingular matrices is also nonsingular, then $\boldsymbol{B}$ is also nonsingular.

Schur complement of $\boldsymbol{B}$ is also nonsingular:

Now consider the Schur complement of $\boldsymbol{B}_{11}$ in $\boldsymbol{B}$ with size $(n-1) \times (n-1)$

$$\bar{\boldsymbol{B}} = \boldsymbol{B}_{2:n,2:n} - \frac{1}{\boldsymbol{B}_{11}} \boldsymbol{B}_{2:n,1} \boldsymbol{B}_{1,2:n}.$$

Suppose there is an $(n-1)$-vector $\boldsymbol{x}$ satisfies

$$\bar{\boldsymbol{B}}\boldsymbol{x} = 0. \tag{1.2}$$

Then $\boldsymbol{x}$ and $y = -\frac{1}{\boldsymbol{B}_{11}} \boldsymbol{B}_{1,2:n}\boldsymbol{x}$ satisfy

$$\boldsymbol{B} \begin{bmatrix} \boldsymbol{x} \\ y \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{1,2:n} \\ \boldsymbol{B}_{2:n,1} & \boldsymbol{B}_{2:n,2:n} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ y \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ 0 \end{bmatrix}.$$

Since $\boldsymbol{B}$ is nonsingular, $\boldsymbol{x}$ and y must be zero. Hence, Equation (1.2) holds only if $\boldsymbol{x} = \boldsymbol{0}$ which means that the null space of $\bar{\boldsymbol{B}}$ is of dimension 0 and thus $\bar{\boldsymbol{B}}$ is nonsingular with size $(n-1) \times (n-1)$.

By the induction assumption that any $(n-1) \times (n-1)$ nonsingular matrix can be factorized as the full LU decomposition form

$$\bar{\boldsymbol{B}} = \boldsymbol{P}_2\boldsymbol{L}_2\boldsymbol{U}_2.$$

We then factor $\boldsymbol{A}$ as

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{P}_1^\top \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{1,2:n} \\ \boldsymbol{B}_{2:n,1} & \boldsymbol{B}_{2:n,2:n} \end{bmatrix} \\
&= \boldsymbol{P}_1^\top \begin{bmatrix} 1 & 0 \\ 0 & \boldsymbol{P}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{1,2:n} \\ \boldsymbol{P}_2^\top \boldsymbol{B}_{2:n,1} & \boldsymbol{P}_2^\top \boldsymbol{B}_{2:n,2:n} \end{bmatrix} \\
&= \boldsymbol{P}_1^\top \begin{bmatrix} 1 & 0 \\ 0 & \boldsymbol{P}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{1,2:n} \\ \boldsymbol{P}_2^\top \boldsymbol{B}_{2:n,1} & \boldsymbol{L}_2\boldsymbol{U}_2 + \boldsymbol{P}_2^\top \frac{1}{\boldsymbol{B}_{11}} \boldsymbol{B}_{2:n,1}\boldsymbol{B}_{1,2:n} \end{bmatrix} \\
&= \boldsymbol{P}_1^\top \begin{bmatrix} 1 & 0 \\ 0 & \boldsymbol{P}_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{1}{\boldsymbol{B}_{11}} \boldsymbol{P}_2^\top \boldsymbol{B}_{2:n,1} & \boldsymbol{L}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{1,2:n} \\ \boldsymbol{0} & \boldsymbol{U}_2 \end{bmatrix}.
\end{aligned}$$

Therefore, we find the full LU decomposition of $\boldsymbol{A} = \boldsymbol{PLU}$ by defining

$$\boldsymbol{P} = \boldsymbol{P}_1^\top \begin{bmatrix} 1 & 0 \\ 0 & \boldsymbol{P}_2 \end{bmatrix}, \qquad \boldsymbol{L} = \begin{bmatrix} 1 & 0 \\ \frac{1}{\boldsymbol{B}_{11}} \boldsymbol{P}_2^\top \boldsymbol{B}_{2:n,1} & \boldsymbol{L}_2 \end{bmatrix}, \qquad \boldsymbol{U} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{1,2:n} \\ \boldsymbol{0} & \boldsymbol{U}_2 \end{bmatrix},$$

from which the result follows. ■

1.4 Bandwidth Preserving in the LU Decomposition without Permutation

For any matrix, the bandwidth of it can be defined as follows.

Definition 1.10: (Matrix Bandwidth)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with entry (i, j) element denoted by $\boldsymbol{A}_{ij}$. Then $\boldsymbol{A}$ has **upper bandwidth** q if $\boldsymbol{A}_{ij} = 0$ for $j > i + q$, and **lower bandwidth** p if $\boldsymbol{A}_{ij} = 0$ for $i > j + p$.

An example of a 6×6 matrix with upper bandwidth 2 and lower bandwidth 3 is shown as follows:

$$\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & 0 & 0 & 0 \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & 0 & 0 \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & 0 \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & 0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}.$$

Then, we prove that the bandwidth after the LU decomposition without permutation is preserved.

Lemma 1.11: (Bandwidth Preserving)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with upper bandwidth q and lower bandwidth p. If $\boldsymbol{A}$ has an LU decomposition $\boldsymbol{A} = \boldsymbol{LU}$, then $\boldsymbol{U}$ has upper bandwidth q and $\boldsymbol{L}$ has lower bandwidth p.

Proof [of Lemma 1.11] The LU decomposition without permutation can be obtained as follows:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{1,2:n} \\ \boldsymbol{A}_{2:n,1} & \boldsymbol{A}_{2:n,2:n} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{1}{\boldsymbol{A}_{11}} \boldsymbol{A}_{2:n,1} & \boldsymbol{I}_{n-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{1,2:n} \\ 0 & \boldsymbol{S} \end{bmatrix} = \boldsymbol{L}_1 \boldsymbol{U}_1,$$

where $\boldsymbol{S} = \boldsymbol{A}_{2:n,2:n} - \frac{1}{\boldsymbol{A}_{11}} \boldsymbol{A}_{2:n,1} \boldsymbol{A}_{1,2:n}$ is the Schur complement of $\boldsymbol{A}_{11}$ in $\boldsymbol{A}$. We can name this decomposition of $\boldsymbol{A}$ as the s-decomposition of $\boldsymbol{A}$. The first column of $\boldsymbol{L}_1$ and the first row of $\boldsymbol{U}_1$ have the required structure (bandwidth p and q respectively), and the Schur complement $\boldsymbol{S}$ of $\boldsymbol{A}_{11}$ has upper bandwidth $q - 1$ and lower bandwidth $p - 1$ respectively. The result follows by induction on the s-decomposition of $\boldsymbol{S}$. ■

1.5 Block LU Decomposition

Another form of the LU decomposition is to factor the matrix into block triangular matrices.

Theorem 1.12: (Block LU Decomposition without Permutation)

For any $n \times n$ square matrix $\boldsymbol{A}$, if the first m leading principal block submatrices are nonsingular, then $\boldsymbol{A}$ can be factored as

$$
\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{I} & & & \\ \boldsymbol{L}_{21} & \boldsymbol{I} & & \\ \vdots & & \ddots & \\ \boldsymbol{L}_{m1} & \ldots & \boldsymbol{L}_{m,m-1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{11} & \boldsymbol{U}_{12} & \ldots & \boldsymbol{U}_{1m} \\ & \boldsymbol{U}_{22} & & \vdots \\ & & \ddots & \boldsymbol{U}_{m-1,m} \\ & & & \boldsymbol{U}_{mm} \end{bmatrix},
$$

where $\boldsymbol{L}_{i,j}$'s and $\boldsymbol{U}_{ij}$'s are some block matrices.

Specifically, this decomposition is unique.

Note that the $\boldsymbol{U}$ in the above theorem is not necessarily upper triangular. An example can be shown as follows:

$$
\boldsymbol{A} = \left[\begin{array}{cc:cc} 0 & 1 & 1 & 1 \\ -1 & 2 & -1 & 2 \\ \hdashline 2 & 1 & 4 & 2 \\ 1 & 2 & 3 & 3 \end{array}\right] = \left[\begin{array}{cc:cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hdashline 5 & -2 & 1 & 0 \\ 4 & -1 & 0 & 1 \end{array}\right] \left[\begin{array}{cc:cc} 0 & 1 & 1 & 1 \\ -1 & 2 & -1 & 2 \\ \hdashline 0 & 0 & -3 & 1 \\ 0 & 0 & -2 & 1 \end{array}\right].
$$

The trivial non-block LU decomposition fails on $\boldsymbol{A}$ since the entry $(1, 1)$ is zero. However, the block LU decomposition exists.

1.6 Application: Linear System via the LU Decomposition

Consider the well-determined linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ with $\boldsymbol{A}$ of size $n \times n$ and nonsingular. Avoid solving the system by computing the inverse of $\boldsymbol{A}$, we solving linear equation by the LU decomposition. Suppose $\boldsymbol{A}$ admits the LU decomposition $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U}$. The solution is given by the following algorithm.

Algorithm 1 Solving Linear Equations by LU Decomposition

Require: matrix $\boldsymbol{A}$ is nonsingular and square with size $n \times n$, solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$;

1: LU Decomposition: factor $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U}$; ▷ $(2/3)n^3$ flops
2: Permutation: $\boldsymbol{w} = \boldsymbol{P}^\top \boldsymbol{b}$; ▷ 0 flops
3: Forward substitution: solve $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{w}$; ▷ $1 + 3 + ... + (2n-1) = n^2$ flops
4: Backward substitution: solve $\boldsymbol{U}\boldsymbol{x} = \boldsymbol{v}$; ▷ $1 + 3 + ... + (2n-1) = n^2$ flops

The complexity of the decomposition step is $(2/3)n^3$ flops (Lu, 2021c), the backward and forward substitution steps both cost $1 + 3 + ... + (2n-1) = n^2$ flops. Therefore, the total cost for computing the linear system via the LU factorization is $(2/3)n^3 + 2n^2$ flops. If we keep only the leading term, the Algorithm 1 costs $(2/3)n^3$ flops where the most cost comes from the LU decomposition.

Linear system via the block LU decomposition For a block LU decomposition of $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$, we need to solve $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{w}$ and $\boldsymbol{U}\boldsymbol{x} = \boldsymbol{v}$. But the latter system is not triangular and requires some extra computations.

1.7 Application: Computing the Inverse of Nonsingular Matrices

By Theorem 1.1, for any nonsingular matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, we have a full LU factorization $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U}$. Then the inverse can be obtained by solving the matrix equation

$$\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I},$$

which contains n linear systems computation: $\boldsymbol{A}\boldsymbol{x}_i = \boldsymbol{e}_i$ for all $i \in \{1, 2, \ldots, n\}$ where $\boldsymbol{x}_i$ is the i-the column of $\boldsymbol{X}$ and $\boldsymbol{e}_i$ is the i-th column of $\boldsymbol{I}$ (i.e., the i-th unit vector).

Theorem 1.13: (Inverse of Nonsingular Matrix by Linear System)

Computing the inverse of a nonsingular matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ by n linear systems needs $\sim (2/3)n^3 + n(2n^2) = (8/3)n^3$ flops where $(2/3)n^3$ comes from the computation of the LU decomposition of $\boldsymbol{A}$.

The proof is trivial by using Algorithm 1. However, the complexity can be reduced by taking the advantage of the structures of $\boldsymbol{U}, \boldsymbol{L}$. We find that the inverse of the nonsingular matrix is $\boldsymbol{A}^{-1} = \boldsymbol{U}^{-1}\boldsymbol{L}^{-1}\boldsymbol{P}^{-1} = \boldsymbol{U}^{-1}\boldsymbol{L}^{-1}\boldsymbol{P}^{T}$. By taking this advantage, the complexity is reduced from $(8/3)n^3$ to $2n^3$ flops.

1.8 Application: Computing the Determinant via the LU Decomposition

We can find the determinant easily by using the LU decomposition. If $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$, then $\det(\boldsymbol{A}) = \det(\boldsymbol{L}\boldsymbol{U}) = \det(\boldsymbol{L})\det(\boldsymbol{U}) = \boldsymbol{U}_{11}\boldsymbol{U}_{22}\ldots\boldsymbol{U}_{nn}$ where $\boldsymbol{U}_{ii}$ is the i-th diagonal of $\boldsymbol{U}$ for $i \in \{1, 2, \ldots, n\}$. [5]

Further, for the LU decomposition with permutation $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U}$, $\det(\boldsymbol{A}) = \det(\boldsymbol{P}\boldsymbol{L}\boldsymbol{U}) = \det(\boldsymbol{P})\boldsymbol{U}_{11}\boldsymbol{U}_{22}\ldots\boldsymbol{U}_{nn}$. The determinant of a permutation matrix is either 1 or –1 because after changing rows around (which changes the sign of the determinant [6]) a permutation matrix becomes identity matrix $\boldsymbol{I}$, whose determinant is one.

1.9 Pivoting

1.9.1 Partial Pivoting

In practice, it is desirable to pivot even when it is not necessary. When dealing with a linear system via the LU decomposition as shown in Algorithm 1, if the diagonal entries of $\boldsymbol{U}$ are small, it can lead to inaccurate solutions for the linear solution. Thus, it is common to pick the largest entry to be the pivot. This is known as the *partial pivoting*. For example,

5. The determinant of a lower triangular matrix (or an upper triangular matrix) is the product of the diagonal entries.
6. The determinant changes sign when two rows are exchanged (sign reversal).

Partial Pivoting For a 4×4 Matrix:

$$
\underset{\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}}
\xrightarrow{\boldsymbol{E}_1}
\underset{\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \mathbf{2} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ 0 & \mathbf{5} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ 0 & \mathbf{7} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}}
\xrightarrow{\boldsymbol{P}_1}
\underset{\boldsymbol{P}_1\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & 7 & \boxtimes & \boxtimes \\ 0 & 5 & \boxtimes & \boxtimes \\ 0 & 2 & \boxtimes & \boxtimes \end{bmatrix}}
\xrightarrow{\boldsymbol{E}_2}
\underset{\boldsymbol{E}_2\boldsymbol{P}_1\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & 7 & \boxtimes & \boxtimes \\ 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ 0 & \mathbf{0} & \mathbf{0} & \boldsymbol{\boxtimes} \end{bmatrix}},
\tag{1.3}
$$

in which case, we pick 7 as the pivot after transformation by $\boldsymbol{E}_1$ even when it is not necessary. This interchange permutation can guarantee that no multiplier is greater than 1 in absolute value during the Gaussian elimination. More generally, the procedure for computing the LU decomposition with partial pivoting of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is given in Algorithm 2.

Algorithm 2 LU Decomposition with Partial Pivoting

Require: Matrix $\boldsymbol{A}$ with size $n \times n$;

```
Let U = A;
for k = 1 to n−1 do                                  ▷ i.e., get the k-th column of U
    Find a row permutation P_k that swaps U_kk with the largest element in |U_{k:n,k}|;
    U = P_k U;
    Determine the Gaussian transformation E_k to introduce zeros below the diagonal of
    the k-th column of U;
    U = E_k U;
end for
Output U;
```

The algorithm requires $2/3(n^3)$ flops and $(n-1)+(n-2)+\ldots+1 \sim O(n^2)$ comparisons resulting from the pivoting procedure. Upon completion, the upper triangular matrix $\boldsymbol{U}$ is given by

$$
\boldsymbol{U} = \boldsymbol{E}_{n-1}\boldsymbol{P}_{n-1}\ldots\boldsymbol{E}_2\boldsymbol{P}_2\boldsymbol{E}_1\boldsymbol{P}_1\boldsymbol{A}.
$$

Computing the final $\boldsymbol{L}$ And we here show that Algorithm 2 computes the LU decomposition in the following form

$$
\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U},
$$

where $\boldsymbol{P} = \boldsymbol{P}_1\boldsymbol{P}_2\ldots\boldsymbol{P}_{n-1}$, $\boldsymbol{U}$ is the upper triangular matrix results directly from the algorithm, $\boldsymbol{L}$ is unit lower triangular with $|\boldsymbol{L}_{ij}| \leq 1$ for all $1 \leq i, j \leq n$. $\boldsymbol{L}_{k+1:n,k}$ is a permuted version of $\boldsymbol{E}_k$'s multipliers. To see this, we notice that the permutation matrices used in the algorithm fall into a special kind of permutation matrix since we only interchange two rows of the matrix. *This implies the $\boldsymbol{P}_k$'s are symmetric and $\boldsymbol{P}_k^2 = \boldsymbol{I}$ for $k \in \{1, 2, \ldots, n-1\}$.* Suppose

$$
\boldsymbol{M}_k = (\boldsymbol{P}_{n-1}\ldots\boldsymbol{P}_{k+1})\boldsymbol{E}_k(\boldsymbol{P}_{k+1}\ldots\boldsymbol{P}_{n-1}).
$$

Then, $\boldsymbol{U}$ can be written as

$$
\boldsymbol{U} = \boldsymbol{M}_{n-1}\ldots\boldsymbol{M}_2\boldsymbol{M}_1\boldsymbol{P}^\top\boldsymbol{A}.
$$

To see what $\boldsymbol{M}_k$ is, we realize that $\boldsymbol{P}_{k+1}$ is a permutation with the upper left $k \times k$ block being an identity matrix. And thus we have

$$
\begin{aligned}
\boldsymbol{M}_k &= (\boldsymbol{P}_{n-1} \ldots \boldsymbol{P}_{k+1})(\boldsymbol{I}_n - \boldsymbol{z}_k \boldsymbol{e}_k^\top)(\boldsymbol{P}_{k+1} \ldots \boldsymbol{P}_{n-1}) \\
&= \boldsymbol{I}_n - (\boldsymbol{P}_{n-1} \ldots \boldsymbol{P}_{k+1})(\boldsymbol{z}_k \boldsymbol{e}_k^\top)(\boldsymbol{P}_{k+1} \ldots \boldsymbol{P}_{n-1}) \\
&= \boldsymbol{I}_n - (\boldsymbol{P}_{n-1} \ldots \boldsymbol{P}_{k+1}\boldsymbol{z}_k)(\boldsymbol{e}_k^\top \boldsymbol{P}_{k+1} \ldots \boldsymbol{P}_{n-1}) \\
&= \boldsymbol{I}_n - (\boldsymbol{P}_{n-1} \ldots \boldsymbol{P}_{k+1}\boldsymbol{z}_k)\boldsymbol{e}_k^\top. \qquad\qquad (\text{since } \boldsymbol{e}_k^\top \boldsymbol{P}_{k+1} \ldots \boldsymbol{P}_{n-1} = \boldsymbol{e}_k^\top)
\end{aligned}
$$

This implies that $\boldsymbol{M}_k$ is unit lower triangular with the k-th column being the permuted version of $\boldsymbol{E}_k$. And the final lower triangular $\boldsymbol{L}$ is thus given by

$$
\boldsymbol{L} = \boldsymbol{M}_1^{-1}\boldsymbol{M}_2^{-1} \ldots \boldsymbol{M}_{n-1}^{-1}.
$$

1.9.2 Complete Pivoting

In partial pivoting, when introducing zeros below the diagonal of the k-th column of $\boldsymbol{U}$, the k-th pivot is determined by scanning the current subcolumn $\boldsymbol{U}_{k:n,k}$. In complete pivoting, the largest absolute entry in the current submatrix $\boldsymbol{U}_{k:n,k:n}$ is interchanged into the entry (k,k) of $\boldsymbol{U}$. Therefore, an additional *column permutation* $\boldsymbol{Q}_k$ is needed in each step. The final upper triangular matrix $\boldsymbol{U}$ is obtained by

$$
\boldsymbol{U} = \boldsymbol{E}_{n-1}\boldsymbol{P}_{n-1} \ldots (\boldsymbol{E}_2\boldsymbol{P}_2(\boldsymbol{E}_1\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{Q}_1)\boldsymbol{Q}_2) \ldots \boldsymbol{Q}_{n-1}.
$$

Similarly, the complete pivoting algorithm is formulated in Algorithm 3.

Algorithm 3 LU Decomposition with Complete Pivoting

Require: Matrix $\boldsymbol{A}$ with size $n \times n$;
1: Let $\boldsymbol{U} = \boldsymbol{A}$;
2: **for** $k = 1$ to $n-1$ **do** $\quad \triangleright$ the value k is to get the k-th column of $\boldsymbol{U}$
3: Find a row permutation matrix $\boldsymbol{P}_k$, and a column permutation $\boldsymbol{Q}_k$ that swaps $\boldsymbol{U}_{kk}$ with the largest element in $|\boldsymbol{U}_{k:n,k:n}|$, say $\boldsymbol{U}_{u,v} = \max |\boldsymbol{U}_{k:n,k:n}|$;
4: $\boldsymbol{U} = \boldsymbol{P}_k\boldsymbol{U}\boldsymbol{Q}_k$;
5: Determine the Gaussian transformation $\boldsymbol{E}_k$ to introduce zeros below the diagonal of the k-th column of $\boldsymbol{U}$;
6: $\boldsymbol{U} = \boldsymbol{E}_k\boldsymbol{U}$;
7: **end for**
8: Output $\boldsymbol{U}$;

The algorithm requires $2/3(n^3)$ flops and $(n^2+(n-1)^2+\ldots+1^2) \sim O(n^3)$ comparisons resulting from the pivoting procedure. Again, let $\boldsymbol{P} = \boldsymbol{P}_1\boldsymbol{P}_2 \ldots \boldsymbol{P}_{n-1}$, $\boldsymbol{Q} = \boldsymbol{Q}_1\boldsymbol{Q}_2 \ldots \boldsymbol{Q}_{n-1}$,

$$
\boldsymbol{M}_k = (\boldsymbol{P}_{n-1} \ldots \boldsymbol{P}_{k+1})\boldsymbol{E}_k(\boldsymbol{P}_{k+1} \ldots \boldsymbol{P}_{n-1}), \qquad \text{for all } k \in \{1, 2, \ldots, n-1\}
$$

and

$$
\boldsymbol{L} = \boldsymbol{M}_1^{-1}\boldsymbol{M}_2^{-1} \ldots \boldsymbol{M}_{n-1}^{-1}.
$$

We have $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{L}\boldsymbol{U}\boldsymbol{Q}^\top$ or $\boldsymbol{P}^\top\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{U}$ as the final decomposition.

1.9.3 Rook Pivoting

The *rook pivoting* provides an alternative to the partial and complete pivoting. Instead of choosing the larges value in $|\boldsymbol{U}_{k:n,k:n}|$ in the k-th step, it searches for an element of $\boldsymbol{U}_{k:n,k:n}$ that is maximal in both its row and column. Apparently, the rook pivoting is not unique such that we could find many entries that satisfy the criteria. For example, for a submatrix $\boldsymbol{U}_{k:n,k:n}$ as follows

$$\boldsymbol{U}_{k:n,k:n} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 7 & 3 \\ 5 & 2 & 1 & 2 \\ 2 & 1 & 2 & 1 \end{bmatrix},$$

where the 7 will be chosen by complete pivoting. And one of $5, 4, 7$ will be identified as a rook pivot.

1.10 Rank-Revealing LU Decomposition

In many applications, a factorization produced by Gaussian elimination with pivoting when $\boldsymbol{A}$ has rank r will reveal rank in the following form

$$\boldsymbol{PAQ} = \begin{bmatrix} \boldsymbol{L}_{11} & \boldsymbol{0} \\ \boldsymbol{L}_{21}^\top & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{11} & \boldsymbol{U}_{12} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

where $\boldsymbol{L}_{11} \in \mathbb{R}^{r\times r}$ and $\boldsymbol{U}_{11} \in \mathbb{R}^{r\times r}$ are nonsingular, $\boldsymbol{L}_{21}, \boldsymbol{U}_{21} \in \mathbb{R}^{r\times(n-r)}$, and $\boldsymbol{P}, \boldsymbol{Q}$ are permutations. Gaussian elimination with rook pivoting or complete pivoting can result in such decomposition (Hwang et al., 1992; Higham, 2002).

2. Cholesky Decomposition

Theorem 2.1: (Cholesky Decomposition)

Every positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{R},$$

where $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ is an upper triangular matrix **with positive diagonal elements**. This decomposition is known as the **Cholesky decomposition** of $\boldsymbol{A}$. $\boldsymbol{R}$ is known as the **Cholesky factor** or **Cholesky triangle** of $\boldsymbol{A}$.

Alternatively, $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$ where $\boldsymbol{L} = \boldsymbol{R}^\top$ is a lower triangular matrix **with positive diagonals**.

Specifically, the Cholesky decomposition is unique (Corollary 2.8, p. 38).

The Cholesky decomposition is named after a French military officer and mathematician, André-Louis Cholesky (1875-1918), who developed the Cholesky decomposition in his surveying work. Similar to the LU decomposition for solving linear systems, the Cholesky decomposition is further used primarily to solve positive definite linear systems. The development of the solution is similar to that of the LU decomposition in Section 1.6 (p. 24), and we shall not repeat the details.

2.1 Existence of the Cholesky Decomposition via Recursive Calculation

In this section, we will prove the existence of the Cholesky decomposition via recursive calculation. In Section 13.5.4 (p. 133), we will also prove the existence of the Cholesky decomposition via the QR decomposition and spectral decomposition. Before showing the existence of Cholesky decomposition, we need the following definitions and lemmas.

Definition 2.2: (Positive Definite and Positive Semidefinite)

A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is positive definite (PD) if $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$ for all nonzero $\boldsymbol{x} \in \mathbb{R}^n$. And a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is positive semidefinite (PSD) if $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

One of the prerequisites for the Cholesky decomposition is the definition of the above positive definiteness of a matrix. We sketch several properties of this PD matrix as follows:

Positive Definite Matrix Property 1 of 5

We will show the equivalent definition on the positive definiteness of a matrix $\boldsymbol{A}$ is that $\boldsymbol{A}$ only has positive eigenvalues, or on the positive semidefiniteness of a matrix $\boldsymbol{A}$ is that $\boldsymbol{A}$ only has nonnegative eigenvalues. The proof is provided in Section 13.5.2 (p. 131) as a consequence of the spectral theorem.

Positive Definite Matrix Property 2 of 5

Lemma 2.3: (Positive Diagonals of Positive Definite Matrices)

The diagonal elements of a positive definite matrix $\boldsymbol{A}$ are all **positive**. And similarly, the diagonal elements of a positive semidefinite matrix $\boldsymbol{B}$ are all **nonnegative**.

Proof [of Lemma 2.3] From the definition of positive definite matrices, we have $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$ for all nonzero $\boldsymbol{x}$. In particular, let $\boldsymbol{x} = \boldsymbol{e}_i$ where $\boldsymbol{e}_i$ is the i-th unit vector with the i-th entry being equal to 1 and other entries being equal to 0. Then,

$$\boldsymbol{e}_i^\top \boldsymbol{A}\boldsymbol{e}_i = a_{ii} > 0, \qquad \forall i \in \{1, 2, \ldots, n\},$$

where a_{ii} is the i-th diagonal component. The proof for the second part follows similarly. This completes the proof. ■

Positive Definite Matrix Property 3 of 5

Lemma 2.4: (Schur Complement of Positive Definite Matrices)

For any positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, its Schur complement of $\boldsymbol{A}_{11}$ is given by $\boldsymbol{S}_{n-1} = \boldsymbol{A}_{2:n,2:n} - \frac{1}{\boldsymbol{A}_{11}}\boldsymbol{A}_{2:n,1}\boldsymbol{A}_{2:n,1}^\top$ which is also positive definite.

A word on the notation Note that the subscript $n-1$ of $\boldsymbol{S}_{n-1}$ means it is of size $(n-1)\times(n-1)$ and it is a Schur complement of an $n\times n$ positive definite matrix. We will use this notation in the following sections.

Proof [of Lemma 2.4] For any nonzero vector $\boldsymbol{v} \in \mathbb{R}^{n-1}$, we can construct a vector $\boldsymbol{x} \in \mathbb{R}^n$ by the following equation:

$$\boldsymbol{x} = \begin{bmatrix} -\frac{1}{\boldsymbol{A}_{11}}\boldsymbol{A}_{2:n,1}^\top\boldsymbol{v} \\ \boldsymbol{v} \end{bmatrix},$$

which is nonzero. Then

$$\begin{aligned}
\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} &= \begin{bmatrix} -\dfrac{1}{\boldsymbol{A}_{11}}\boldsymbol{v}^\top\boldsymbol{A}_{2:n,1} & \boldsymbol{v}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{2:n,1}^\top \\ \boldsymbol{A}_{2:n,1} & \boldsymbol{A}_{2:n,2:n} \end{bmatrix} \begin{bmatrix} -\frac{1}{\boldsymbol{A}_{11}}\boldsymbol{A}_{2:n,1}^\top\boldsymbol{v} \\ \boldsymbol{v} \end{bmatrix} \\
&= \begin{bmatrix} -\dfrac{1}{\boldsymbol{A}_{11}}\boldsymbol{v}^\top\boldsymbol{A}_{2:n,1} & \boldsymbol{v}^\top \end{bmatrix} \begin{bmatrix} 0 \\ \boldsymbol{S}_{n-1}\boldsymbol{v} \end{bmatrix} \\
&= \boldsymbol{v}^\top\boldsymbol{S}_{n-1}\boldsymbol{v}.
\end{aligned}$$

Since $\boldsymbol{A}$ is positive definite, we have $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} = \boldsymbol{v}^\top\boldsymbol{S}_{n-1}\boldsymbol{v} > 0$ for all nonzero $\boldsymbol{v}$. Thus, the Schur complement $\boldsymbol{S}_{n-1}$ is positive definite as well. ■

The above argument can be extended to PSD matrices as well. If $\boldsymbol{A}$ is PSD, then the Schur complement $\boldsymbol{S}_{n-1}$ is also PSD.

A word on the Schur complement In the proof of Theorem 1.1, we have shown this Schur complement $\boldsymbol{S}_{n-1} = \boldsymbol{A}_{2:n,2:n} - \frac{1}{\boldsymbol{A}_{11}}\boldsymbol{A}_{2:n,1}\boldsymbol{A}_{2:n,1}^\top$ is also nonsingular if $\boldsymbol{A}$ is nonsngular and $\boldsymbol{A}_{11} \neq 0$. Similarly, the Schur complement of $\boldsymbol{A}_{nn}$ in $\boldsymbol{A}$ is $\bar{\boldsymbol{S}}_{n-1} = \boldsymbol{A}_{1:n-1,1:n-1} - \frac{1}{\boldsymbol{A}_{nn}}\boldsymbol{A}_{1:n-1,n}\boldsymbol{A}_{1:n-1,n}^\top$ which is also positive definite if $\boldsymbol{A}$ is positive definite. This property can help prove the fact that the leading principal minors of positive definite matrices are all positive. See Section 2.2 for more details.

We then prove the existence of Cholesky decomposition using these lemmas.

Proof [**of Theorem 2.1: Existence of Cholesky Decomposition Recursively**] For any positive definite matrix $\boldsymbol{A}$, we can write out (since $\boldsymbol{A}_{11}$ is positive by Lemma 2.3)

$$
\begin{aligned}
\boldsymbol{A} &= \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{2:n,1}^\top \\ \boldsymbol{A}_{2:n,1} & \boldsymbol{A}_{2:n,2:n} \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{\boldsymbol{A}_{11}} & \mathbf{0} \\ \frac{1}{\sqrt{\boldsymbol{A}_{11}}}\boldsymbol{A}_{2:n,1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \sqrt{\boldsymbol{A}_{11}} & \frac{1}{\sqrt{\boldsymbol{A}_{11}}}\boldsymbol{A}_{2:n,1}^\top \\ \mathbf{0} & \boldsymbol{A}_{2:n,2:n} - \frac{1}{\boldsymbol{A}_{11}}\boldsymbol{A}_{2:n,1}\boldsymbol{A}_{2:n,1}^\top \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{\boldsymbol{A}_{11}} & \mathbf{0} \\ \frac{1}{\sqrt{\boldsymbol{A}_{11}}}\boldsymbol{A}_{2:n,1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_{2:n,2:n} - \frac{1}{\boldsymbol{A}_{11}}\boldsymbol{A}_{2:n,1}\boldsymbol{A}_{2:n,1}^\top \end{bmatrix} \begin{bmatrix} \sqrt{\boldsymbol{A}_{11}} & \frac{1}{\sqrt{\boldsymbol{A}_{11}}}\boldsymbol{A}_{2:n,1}^\top \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \\
&= \boldsymbol{R}_1^\top \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{n-1} \end{bmatrix} \boldsymbol{R}_1.
\end{aligned}
$$

where

$$
\boldsymbol{R}_1 = \begin{bmatrix} \sqrt{\boldsymbol{A}_{11}} & \frac{1}{\sqrt{\boldsymbol{A}_{11}}}\boldsymbol{A}_{2:n,1}^\top \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix}.
$$

Since we proved the Schur complement $\boldsymbol{S}_{n-1}$ is positive definite in Lemma 2.4, then we can factor it in the same way as

$$
\boldsymbol{S}_{n-1} = \hat{\boldsymbol{R}}_2^\top \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{n-2} \end{bmatrix} \hat{\boldsymbol{R}}_2.
$$

Therefore, we have

$$
\begin{aligned}
\boldsymbol{A} &= \boldsymbol{R}_1^\top \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{R}}_2^\top \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{n-2} \end{bmatrix} \hat{\boldsymbol{R}}_2. \end{bmatrix} \boldsymbol{R}_1 \\
&= \boldsymbol{R}_1^\top \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{R}}_2^\top \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{n-2} \end{bmatrix} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{R}}_2 \end{bmatrix} \boldsymbol{R}_1 \\
&= \boldsymbol{R}_1^\top \boldsymbol{R}_2^\top \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{n-2} \end{bmatrix} \end{bmatrix} \boldsymbol{R}_2 \boldsymbol{R}_1.
\end{aligned}
$$

The same formula can be recursively applied. This process gradually continues down to the bottom-right corner giving us the decomposition

$$
\begin{aligned}
\boldsymbol{A} &= \boldsymbol{R}_1^\top \boldsymbol{R}_2^\top \ldots \boldsymbol{R}_n^\top \boldsymbol{R}_n \ldots \boldsymbol{R}_2 \boldsymbol{R}_1 \\
&= \boldsymbol{R}^\top \boldsymbol{R},
\end{aligned}
$$

where $\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_n$ are upper triangular matrices with positive diagonal elements and $\boldsymbol{R} = \boldsymbol{R}_1 \boldsymbol{R}_2 \ldots \boldsymbol{R}_n$ is also an upper triangular matrix with positive diagonal elements from which the result follows. ■

The process in the proof can also be used to calculate the Cholesky decomposition and compute the complexity of the algorithm.

Lemma 2.5: ($\boldsymbol{R}^\top \boldsymbol{R}$ is PD)

For any upper triangular matrix $\boldsymbol{R}$ with positive diagonal elements, then $\boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{R}$ is positive definite.

Proof [of Lemma 2.5] If an upper triangular matrix $\boldsymbol{R}$ has positive diagonals, it has full column rank, and the null space of $\boldsymbol{R}$ is of dimension 0 by the fundamental theorem of linear algebra (Theorem 0.15, p. 14). As a result, $\boldsymbol{R}\boldsymbol{x} \neq \boldsymbol{0}$ for any nonzero vector $\boldsymbol{x}$. Thus $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = ||\boldsymbol{R}\boldsymbol{x}||^2 > 0$ for any nonzero vector $\boldsymbol{x}$. ■

This corollary above works not only for the upper triangular matrices $\boldsymbol{R}$, but can be extended to any $\boldsymbol{R}$ with linearly independent columns.

A word on the two clams Combine Theorem 2.1 and Lemma 2.5, we can claim that matrix $\boldsymbol{A}$ is positive definite if and only if $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{R}$ where $\boldsymbol{R}$ is an upper triangular matrix with positive diagonals.

2.2 Sylvester's Criterion: Leading Principal Minors of PD Matrices

In Lemma 2.4, we proved for any positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, its Schur complement of $\boldsymbol{A}_{11}$ is $\boldsymbol{S}_{n-1} = \boldsymbol{A}_{2:n,2:n} - \frac{1}{\boldsymbol{A}_{11}}\boldsymbol{A}_{2:n,1}\boldsymbol{A}_{2:n,1}^\top$ and it is also positive definite. This is also true for its Schur complement of $\boldsymbol{A}_{nn}$, i.e., $\boldsymbol{S}'_{n-1} = \boldsymbol{A}_{1:n-1,1:n-1} - \frac{1}{\boldsymbol{A}_{nn}}\boldsymbol{A}_{1:n-1,n}\boldsymbol{A}_{1:n-1,n}^\top$ is also positive definite.

We then claim that all the leading principal minors (Definition 1.3, p. 16) of a positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ are positive. This is also known as the Sylvester's criterion (Swamy, 1973; Gilbert, 1991). Recall that these positive leading principal minors imply the existence of the LU decomposition for positive definite matrix $\boldsymbol{A}$ by Theorem 1.4 (p. 17).

To show the Sylvester's criterion, we need the following lemma.

Positive Definite Matrix Property 4 of 5

Lemma 2.6: (Quadratic PD)

Let $\boldsymbol{E}$ be any invertible matrix. Then $\boldsymbol{A}$ is positive definite if and only if $\boldsymbol{E}^\top \boldsymbol{A}\boldsymbol{E}$ is also positive definite.

Proof [of Lemma 2.6] We will prove by forward implication and reverse implication separately as follows.

Forward implication Suppose $\boldsymbol{A}$ is positive definite, then for any nonzero vector $\boldsymbol{x}$, we have $\boldsymbol{x}^\top \boldsymbol{E}^\top \boldsymbol{A}\boldsymbol{E}\boldsymbol{x} = \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y} > 0$, since $\boldsymbol{E}$ is invertible such that $\boldsymbol{E}\boldsymbol{x}$ is nonzero. [7] This implies $\boldsymbol{E}^\top \boldsymbol{A}\boldsymbol{E}$ is PD.

Reverse implication Conversely, suppose $\boldsymbol{E}^\top \boldsymbol{A}\boldsymbol{E}$ is positive definite, for any nonzero $\boldsymbol{x}$, we have $\boldsymbol{x}^\top \boldsymbol{E}^\top \boldsymbol{A}\boldsymbol{E}\boldsymbol{x} > 0$. For any nonzero $\boldsymbol{y}$, there exists a nonzero $\boldsymbol{x}$ such that $\boldsymbol{y} = \boldsymbol{E}\boldsymbol{x}$

7. Since the null space of $\boldsymbol{E}$ is of dimension 0 and the only solution for $\boldsymbol{E}\boldsymbol{x} = \boldsymbol{0}$ is the trivial solution $\boldsymbol{x} = 0$.

since $\boldsymbol{E}$ is invertible. This implies $\boldsymbol{A}$ is PD as well. ■

We then provide the rigorous proof for Sylvester's criterion.

Positive Definite Matrix Property 5 of 5

Theorem 2.7: (Sylvester's Criterion)

The real symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is positive definite if and only if all the leading principal minors of $\boldsymbol{A}$ are positive.

Proof [of Theorem 2.7] We will prove by forward implication and reverse implication separately as follows.

Forward implication: We will prove by induction for the forward implication. Suppose $\boldsymbol{A}$ is positive definite. Since all the components on the diagonal of positive definite matrices are all positive (Lemma 2.3, p. 29). The case for $n = 1$ is trivial that $\det(\boldsymbol{A}) > 0$ if $\boldsymbol{A}$ is a scalar.

Suppose all the leading principal minors for $k \times k$ matrices are all positive. If we could prove this is also true for $(k+1)\times(k+1)$ PD matrices, then we complete the proof.

For a $(k+1)\times(k+1)$ matrix with the block form $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^\top & d \end{bmatrix}$, where $\boldsymbol{A}$ is a $k \times k$ submatrix. Then its Schur complement of d, $\boldsymbol{S}_k = \boldsymbol{A} - \frac{1}{d}\boldsymbol{b}\boldsymbol{b}^\top$ is also positive definite and its determinant is positive from the assumption. Therefore, $\det(\boldsymbol{M}) = \det(d)\det(\boldsymbol{A} - \frac{1}{d}\boldsymbol{b}\boldsymbol{b}^\top) =$ [8] $d \cdot \det(\boldsymbol{A} - \frac{1}{d}\boldsymbol{b}\boldsymbol{b}^\top) > 0$, which completes the proof.

Reverse implication: Conversely, suppose all the leading principal minors of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ are positive, i.e., leading principal submatrices are nonsingular. Suppose further the (i,j)-th entry of $\boldsymbol{A}$ is denoted by a_{ij}, we realize that $a_{11} > 0$ by the assumption. Subtract multiples of the first row of $\boldsymbol{A}$ to zero out the entries in the first column of $\boldsymbol{A}$ below the first diagonal a_{11}. That is,

$$
\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \stackrel{\boldsymbol{E}_1\boldsymbol{A}}{\longrightarrow} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} & \ldots & a_{nn} \end{bmatrix}.
$$

This operation preserves the values of the principal minors of $\boldsymbol{A}$. The $\boldsymbol{E}_1$ might be mysterious to the readers. Actually, the $\boldsymbol{E}_1$ contains two steps $\boldsymbol{E}_1 = \boldsymbol{Z}_{12}\boldsymbol{Z}_{11}$. The first step $\boldsymbol{Z}_{11}$

8. By the fact that if matrix $\boldsymbol{M}$ has a block formulation: $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$, then $\det(\boldsymbol{M}) = \det(\boldsymbol{D})\det(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})$.

is to subtract the 2-nd row to the n-th row by a multiple of the first row, that is

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \xrightarrow{\boldsymbol{Z}_{11}\boldsymbol{A}} \begin{bmatrix} 1 & 0 & \ldots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{11}} & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & (a_{22} - \frac{a_{21}}{a_{11}}a_{12}) & \ldots & (a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & (a_{n2} - \frac{a_{n1}}{a_{11}}a_{12}) & \ldots & (a_{nn} - \frac{a_{n1}}{a_{11}}a_{1n}) \end{bmatrix},$$

where we subtract the bottom-right $(n-1)\times(n-1)$ by some terms additionally. $\boldsymbol{Z}_{12}$ is to add back these terms.

$$\boldsymbol{Z}_{11}\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & (a_{22} - \frac{a_{21}}{a_{11}}a_{12}) & \ldots & (a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & (a_{n2} - \frac{a_{n1}}{a_{11}}a_{12}) & \ldots & (a_{nn} - \frac{a_{n1}}{a_{11}}a_{1n}) \end{bmatrix}$$

$$\xrightarrow{\boldsymbol{Z}_{12}(\boldsymbol{Z}_{11}\boldsymbol{A})} \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \frac{a_{21}}{a_{11}} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{a_{n1}}{a_{11}} & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & (a_{22} - \frac{a_{21}}{a_{11}}a_{12}) & \ldots & (a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & (a_{n2} - \frac{a_{n1}}{a_{11}}a_{12}) & \ldots & (a_{nn} - \frac{a_{n1}}{a_{11}}a_{1n}) \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} & \ldots & a_{nn} \end{bmatrix} = \boldsymbol{E}_1\boldsymbol{A}.$$

Now subtract multiples of the first column of $\boldsymbol{E}_1\boldsymbol{A}$, from the other columns of $\boldsymbol{E}_1\boldsymbol{A}$ to zero out the entries in the first row of $\boldsymbol{E}_1\boldsymbol{A}$ to the right of the first column. Since $\boldsymbol{A}$ is symmetric, we can multiply on the right by $\boldsymbol{E}_1^\top$ to get what we want. We then have

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \xrightarrow{\boldsymbol{E}_1\boldsymbol{A}} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} & \ldots & a_{nn} \end{bmatrix} \xrightarrow{\boldsymbol{E}_1\boldsymbol{A}\boldsymbol{E}_1^\top} \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} & \ldots & a_{nn} \end{bmatrix}.$$

This operation also preserves the values of the principal minors of $\boldsymbol{A}$. The leading principal minors of $\boldsymbol{E}_1\boldsymbol{A}\boldsymbol{E}_1^\top$ are exactly the same as those of $\boldsymbol{A}$.

Continue this process, we will transform $\boldsymbol{A}$ into a diagonal matrix $\boldsymbol{E}_n \ldots \boldsymbol{E}_1\boldsymbol{A}\boldsymbol{E}_1^\top \ldots \boldsymbol{E}_n^\top$ whose diagonal values are exactly the same as the diagonals of $\boldsymbol{A}$ and are positive. Let $\boldsymbol{E} = \boldsymbol{E}_n \ldots \boldsymbol{E}_1$, which is an invertible matrix. Apparently, $\boldsymbol{E}\boldsymbol{A}\boldsymbol{E}^\top$ is PD, which implies $\boldsymbol{A}$ is PD as well from Lemma 2.6. ■

2.3 Existence of the Cholesky Decomposition via the LU Decomposition without Permutation

By Theorem 2.7 on Sylvester's criterion and the existence of LU decomposition without permutation in Theorem 1.4 (p. 17), there is a unique LU decomposition for positive definite matrix $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}_0$ where $\boldsymbol{L}$ is a unit lower triangular matrix and $\boldsymbol{U}_0$ is an upper triangular matrix. Since *the signs of the pivots of a symmetric matrix are the same as the signs of the eigenvalues* (Strang, 2009):

$$\text{number of positive pivots} = \text{number of positive eigenvalues}.$$

And $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}_0$ has the following form

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}_0 = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ l_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \ldots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \ldots & u_{1n} \\ 0 & u_{22} & \ldots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & u_{nn} \end{bmatrix}.$$

This implies that the diagonals of $\boldsymbol{U}_0$ contain the pivots of $\boldsymbol{A}$. And all the eigenvalues of PD matrices are positive (see Lemma 13.32, p. 131, which is a consequence of spectral decomposition). Thus the diagonals of $\boldsymbol{U}_0$ are positive.

Taking the diagonal of $\boldsymbol{U}_0$ out into a diagonal matrix $\boldsymbol{D}$, we can rewrite $\boldsymbol{U}_0 = \boldsymbol{D}\boldsymbol{U}$ as shown in the following equation

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}_0 = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ l_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \ldots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & 0 & \ldots & 0 \\ 0 & u_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & u_{nn} \end{bmatrix} \begin{bmatrix} 1 & u_{12}/u_{11} & \ldots & u_{1n}/u_{11} \\ 0 & 1 & \ldots & u_{2n}/u_{22} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{U},$$

where $\boldsymbol{U}$ is a *unit* upper triangular matrix. By the uniqueness of the LU decomposition without permutation in Corollary 1.8 (p. 21) and the symmetry of $\boldsymbol{A}$, it follows that $\boldsymbol{U} = \boldsymbol{L}^\top$, and $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\top$. Since the diagonals of $\boldsymbol{D}$ are positive, we can set $\boldsymbol{R} = \boldsymbol{D}^{1/2}\boldsymbol{L}^\top$ where $\boldsymbol{D}^{1/2} = \text{diag}(\sqrt{u_{11}}, \sqrt{u_{22}}, \ldots, \sqrt{u_{nn}})$ such that $\boldsymbol{A} = \boldsymbol{R}^\top\boldsymbol{R}$ is the Cholesky decomposition of $\boldsymbol{A}$, and $\boldsymbol{R}$ is upper triangular with positive diagonals.

2.3.1 Diagonal Values of the Upper Triangular Matrix

Suppose $\boldsymbol{A}$ is a PD matrix, take $\boldsymbol{A}$ as a block matrix $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$ where $\boldsymbol{A}_{11} \in \mathbb{R}^{k\times k}$, and its block LU decomposition is given by

$$\begin{aligned} \boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} = \boldsymbol{L}\boldsymbol{U}_0 &= \begin{bmatrix} \boldsymbol{L}_{11} & \boldsymbol{0} \\ \boldsymbol{L}_{21} & \boldsymbol{L}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{11} & \boldsymbol{U}_{12} \\ \boldsymbol{0} & \boldsymbol{U}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{L}_{11}\boldsymbol{U}_{11} & \boldsymbol{L}_{11}\boldsymbol{U}_{12} \\ \boldsymbol{L}_{21}\boldsymbol{U}_{11} & \boldsymbol{L}_{21}\boldsymbol{U}_{12} + \boldsymbol{L}_{22}\boldsymbol{U}_{22} \end{bmatrix}. \end{aligned}$$

Then the leading principal minor (Definition 1.3, p. 16), $\Delta_k = \det(\boldsymbol{A}_{1:k,1:k}) = \det(\boldsymbol{A}_{11})$ is given by

$$\Delta_k = \det(\boldsymbol{A}_{11}) = \det(\boldsymbol{L}_{11}\boldsymbol{U}_{11}) = \det(\boldsymbol{L}_{11})\det(\boldsymbol{U}_{11}).$$

We notice that $\boldsymbol{L}_{11}$ is a unit lower triangular matrix and $\boldsymbol{U}_{11}$ is an upper triangular matrix. By the fact that the determinant of a lower triangular matrix (or an upper triangular matrix) is the product of the diagonal entries, we obtain

$$\Delta_k = \det(\boldsymbol{U}_{11}) = u_{11}u_{22}\ldots u_{kk},$$

i.e., the k-th leading principal minor of $\boldsymbol{A}$ is the determinant of the $k \times k$ submatrix of $\boldsymbol{U}_0$. That is also the product of the first k diagonals of $\boldsymbol{D}$ ($\boldsymbol{D}$ is the matrix from $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\top$). Let $\boldsymbol{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, therefore, we have

$$\Delta_k = d_1 d_2 \ldots d_k = \Delta_{k-1} d_k.$$

This gives us an alternative form of $\boldsymbol{D}$, i.e., the **squared** diagonal values of $\boldsymbol{R}$ ($\boldsymbol{R}$ is the Cholesky factor from $\boldsymbol{A} = \boldsymbol{R}^\top\boldsymbol{R}$), and it is given by

$$\boldsymbol{D} = \mathrm{diag}\left(\Delta_1, \frac{\Delta_2}{\Delta_1}, \ldots, \frac{\Delta_n}{\Delta_{n-1}}\right),$$

where Δ_k is the k-th leading principal minor of $\boldsymbol{A}$, for all $k \in \{1, 2, \ldots, n\}$. That is, the diagonal values of $\boldsymbol{R}$ are given by

$$\mathrm{diag}\left(\sqrt{\Delta_1}, \sqrt{\frac{\Delta_2}{\Delta_1}}, \ldots, \sqrt{\frac{\Delta_n}{\Delta_{n-1}}}\right).$$

2.3.2 Block Cholesky Decomposition

Following from the last section, suppose $\boldsymbol{A}$ is a PD matrix, take $\boldsymbol{A}$ as a block matrix $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_k & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{bmatrix}$ where $\boldsymbol{A}_k \in \mathbb{R}^{k\times k}$, and its block LU decomposition is given by

$$\begin{aligned}\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_k & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{bmatrix} = \boldsymbol{L}\boldsymbol{U}_0 &= \begin{bmatrix}\boldsymbol{L}_k & \mathbf{0}\\ \boldsymbol{L}_{21} & \boldsymbol{L}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_k & \boldsymbol{U}_{12}\\ \mathbf{0} & \boldsymbol{U}_{22}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{L}_k\boldsymbol{U}_k & \boldsymbol{L}_{11}\boldsymbol{U}_{12}\\ \boldsymbol{L}_{21}\boldsymbol{U}_{11} & \boldsymbol{L}_{21}\boldsymbol{U}_{12} + \boldsymbol{L}_{22}\boldsymbol{U}_{22}\end{bmatrix}.\end{aligned}$$

where the k-th leading principal submatrix $\boldsymbol{A}_k$ of $\boldsymbol{A}$ also has its LU decomposition $\boldsymbol{A}_k = \boldsymbol{L}_k\boldsymbol{U}_k$. Then, it is trivial that the Cholesky decomposition of an $n\times n$ matrix contains $n-1$ other Cholesky decompositions within it: $\boldsymbol{A}_k = \boldsymbol{R}_k^\top\boldsymbol{R}_k$, for all $k \in \{1, 2, \ldots, n-1\}$. This is particularly true that any leading principal submatrix $\boldsymbol{A}_k$ of the positive definite matrix $\boldsymbol{A}$ is also positive definite. This can be shown that for positive definite matrix $\boldsymbol{A}_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$, and any nonzero vector $\boldsymbol{x}_k \in \mathbb{R}^k$ appended by a zero element $\boldsymbol{x}_{k+1} = \begin{bmatrix}\boldsymbol{x}_k\\ 0\end{bmatrix}$. It follows that

$$\boldsymbol{x}_k^\top\boldsymbol{A}_k\boldsymbol{x}_k = \boldsymbol{x}_{k+1}^\top\boldsymbol{A}_{k+1}\boldsymbol{x}_{k+1} > 0,$$

and $\boldsymbol{A}_k$ is positive definite. If we start from $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, we will recursively get that $\boldsymbol{A}_{n-1}$ is PD, $\boldsymbol{A}_{n-2}$ is PD, ... And all of them admit a Cholesky decomposition.

2.4 Existence of the Cholesky Decomposition via Induction

In the last section, we proved the existence of the Cholesky decomposition via the LU decomposition without permutation. Following from the proof of the LU decomposition in Section 1.2, we realize that the existence of Cholesky decomposition can be a direct consequence of induction as well.

Proof [**of Theorem 2.1: Existence of Cholesky Decomposition by Induction**] We will prove by induction that every $n \times n$ positive definite matrix $\boldsymbol{A}$ has a decomposition $\boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{R}$. The 1×1 case is trivial by setting $R = \sqrt{A}$, thus, $A = R^2$.

Suppose for any $k \times k$ PD matrix $\boldsymbol{A}_k$ has a Cholesky decomposition. If we prove any $(k+1) \times (k+1)$ PD matrix $\boldsymbol{A}_{k+1}$ can also be factored as this Cholesky decomposition, then we complete the proof.

For any $(k+1) \times (k+1)$ PD matrix $\boldsymbol{A}_{k+1}$, Write out $\boldsymbol{A}_{k+1}$ as

$$\boldsymbol{A}_{k+1} = \begin{bmatrix} \boldsymbol{A}_k & \boldsymbol{b} \\ \boldsymbol{b}^\top & d \end{bmatrix}.$$

We note that $\boldsymbol{A}_k$ is PD from the last section. By the inductive hypothesis, it admits a Cholesky decomposition $\boldsymbol{A}_k = \boldsymbol{R}_k^\top \boldsymbol{R}_k$. We can construct the upper triangular matrix

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_k & \boldsymbol{r} \\ 0 & s \end{bmatrix},$$

such that it follows that

$$\boldsymbol{R}_{k+1}^\top \boldsymbol{R}_{k+1} = \begin{bmatrix} \boldsymbol{R}_k^\top \boldsymbol{R}_k & \boldsymbol{R}_k^\top \boldsymbol{r} \\ \boldsymbol{r}^\top \boldsymbol{R}_k & \boldsymbol{r}^\top \boldsymbol{r} + s^2 \end{bmatrix}.$$

Therefore, if we can prove $\boldsymbol{R}_{k+1}^\top \boldsymbol{R}_{k+1} = \boldsymbol{A}_{k+1}$ is the Cholesky decomposition of $\boldsymbol{A}_{k+1}$ (which requires the value s to be positive), then we complete the proof. That is, we need to prove

$$\begin{aligned} \boldsymbol{b} &= \boldsymbol{R}_k^\top \boldsymbol{r}, \\ d &= \boldsymbol{r}^\top \boldsymbol{r} + s^2. \end{aligned}$$

Since $\boldsymbol{R}_k$ is nonsingular, we have a unique solution for $\boldsymbol{r}$ and s that

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{R}_k^{-\top} \boldsymbol{b}, \\ s &= \sqrt{d - \boldsymbol{r}^\top \boldsymbol{r}} = \sqrt{d - \boldsymbol{b}^\top \boldsymbol{A}_k^{-1} \boldsymbol{b}}, \end{aligned}$$

since we assume s is nonnegative. However, we need to further prove that s is not only nonnegative, but also positive. Since $\boldsymbol{A}_k$ is PD, from Sylvester's criterion, and the fact that if matrix $\boldsymbol{M}$ has a block formulation: $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$, then $\det(\boldsymbol{M}) = \det(\boldsymbol{A}) \det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})$. We have

$$\det(\boldsymbol{A}_{k+1}) = \det(\boldsymbol{A}_k) \det(d - \boldsymbol{b}^\top \boldsymbol{A}_k^{-1} \boldsymbol{b}) = \det(\boldsymbol{A}_k)(d - \boldsymbol{b}^\top \boldsymbol{A}_k^{-1} \boldsymbol{b}) > 0.$$

Since $\det(\boldsymbol{A}_k) > 0$, we then obtain that $(d - \boldsymbol{b}^\top \boldsymbol{A}_k^{-1} \boldsymbol{b}) > 0$ and this implies $s > 0$. We complete the proof. ■

2.5 Uniqueness of the Cholesky Decomposition

Corollary 2.8: (Uniqueness of Cholesky Decomposition)

The Cholesky decomposition $\boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{R}$ for any positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is unique.

The uniqueness of the Cholesky decomposition can be an immediate consequence of the uniqueness of the LU decomposition without permutation. Or, an alternative rigorous proof is provided as follows.

Proof [of Corollary 2.8] Suppose the Cholesky decomposition is not unique, then we can find two decompositions such that $\boldsymbol{A} = \boldsymbol{R}_1^\top \boldsymbol{R}_1 = \boldsymbol{R}_2^\top \boldsymbol{R}_2$ which implies

$$\boldsymbol{R}_1 \boldsymbol{R}_2^{-1} = \boldsymbol{R}_1^{-\top} \boldsymbol{R}_2^\top.$$

From the fact that the inverse of an upper triangular matrix is also an upper triangular matrix, and the product of two upper triangular matrices is also an upper triangular matrix, [9] we realize that the left-side of the above equation is an upper triangular matrix and the right-side of it is a lower triangular matrix. This implies $\boldsymbol{R}_1 \boldsymbol{R}_2^{-1} = \boldsymbol{R}_1^{-\top} \boldsymbol{R}_2^\top$ is a diagonal matrix, and $\boldsymbol{R}_1^{-\top} \boldsymbol{R}_2^\top = (\boldsymbol{R}_1^{-\top} \boldsymbol{R}_2^\top)^\top = \boldsymbol{R}_2 \boldsymbol{R}_1^{-1}$. Let $\boldsymbol{\Lambda} = \boldsymbol{R}_1 \boldsymbol{R}_2^{-1} = \boldsymbol{R}_2 \boldsymbol{R}_1^{-1}$ be the diagonal matrix. We notice that the diagonal value of $\boldsymbol{\Lambda}$ is the product of the corresponding diagonal values of $\boldsymbol{R}_1$ and $\boldsymbol{R}_2^{-1}$ (or $\boldsymbol{R}_2$ and $\boldsymbol{R}_1^{-1}$). That is, for

$$\boldsymbol{R}_1 = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ 0 & r_{22} & \ldots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & r_{nn} \end{bmatrix}, \qquad \boldsymbol{R}_2 = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1n} \\ 0 & s_{22} & \ldots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & s_{nn} \end{bmatrix},$$

we have,

$$\boldsymbol{R}_1 \boldsymbol{R}_2^{-1} = \begin{bmatrix} \frac{r_{11}}{s_{11}} & 0 & \ldots & 0 \\ 0 & \frac{r_{22}}{s_{22}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{r_{nn}}{s_{nn}} \end{bmatrix} = \begin{bmatrix} \frac{s_{11}}{r_{11}} & 0 & \ldots & 0 \\ 0 & \frac{s_{22}}{r_{22}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{s_{nn}}{r_{nn}} \end{bmatrix} = \boldsymbol{R}_2 \boldsymbol{R}_1^{-1}.$$

Since both $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ have positive diagonals, this implies $r_{11} = s_{11}, r_{22} = s_{22}, \ldots, r_{nn} = s_{nn}$. And $\boldsymbol{\Lambda} = \boldsymbol{R}_1 \boldsymbol{R}_2^{-1} = \boldsymbol{R}_2 \boldsymbol{R}_1^{-1} = \boldsymbol{I}$. That is, $\boldsymbol{R}_1 = \boldsymbol{R}_2$ and this leads to a contradiction. The Cholesky decomposition is thus unique. ■

2.6 Last Words on Positive Definite Matrices

In Section 13.5.2 (p. 131), we will prove that a matrix $\boldsymbol{A}$ is PD if and only if $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{P}^\top \boldsymbol{P}$ where $\boldsymbol{P}$ is nonsingular. And in Section 13.5.5 (p. 133), we will

9. Same for lower triangular matrices: the inverse of a lower triangular matrix is also a lower triangular matrix, and the product of two lower triangular matrices is also a lower triangular matrix.

prove that PD matrix $\boldsymbol{A}$ can be uniquely factored as $\boldsymbol{A} = \boldsymbol{B}^2$ where $\boldsymbol{B}$ is also PD. The two results are both consequences of the spectral decomposition of PD matrices.

To conclude, for PD matrix $\boldsymbol{A}$, we can factor it into $\boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{R}$ where $\boldsymbol{R}$ is an upper triangular matrix with positive diagonals as shown in Theorem 2.1 by Cholesky decomposition, $\boldsymbol{A} = \boldsymbol{P}^\top \boldsymbol{P}$ where $\boldsymbol{P}$ is nonsingular in Theorem 13.33 (p. 131), and $\boldsymbol{A} = \boldsymbol{B}^2$ where $\boldsymbol{B}$ is PD in Theorem 13.34 (p. 133). For clarity, the different factorizations of positive definite matrix $\boldsymbol{A}$ are summarized in Figure 4.

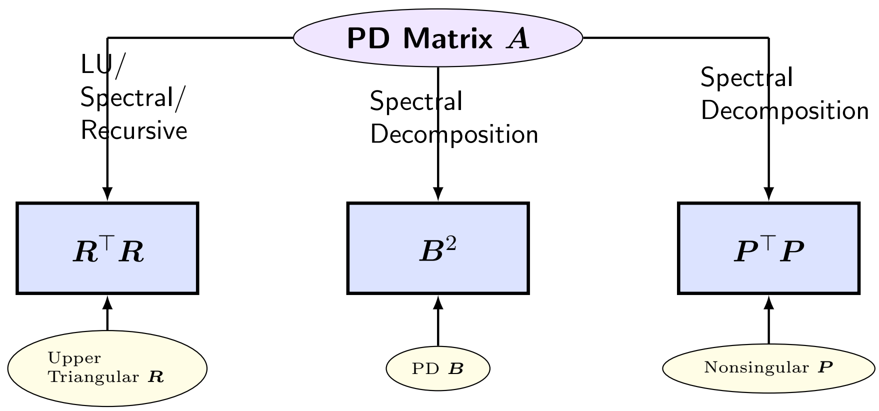


Figure 4: Demonstration of different factorizations on positive definite matrix $\boldsymbol{A}$.

2.7 Decomposition for Semidefinite Matrices

For positive semidefinite matrices, the Cholesky decomposition also exists with slight modification.

Theorem 2.9: (Semidefinite Decomposition)

Every positive semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{R},$$

where $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ is an upper triangular matrix with possible **zero** diagonal elements.

For such decomposition, the diagonal of $\boldsymbol{R}$ may not display the rank of $\boldsymbol{A}$ (Higham, 2009). More generally, a rank-revealing decomposition for semidefinite decomposition is provided as follows.

Theorem 2.10: (Semidefinite Rank-Revealing Decomposition)

Every positive semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with rank r can be factored as

$$\boldsymbol{P}^\top \boldsymbol{A} \boldsymbol{P} = \boldsymbol{R}^\top \boldsymbol{R}, \qquad \text{with} \qquad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{n\times n},$$

where $\boldsymbol{R}_{11} \in \mathbb{R}^{r\times r}$ is an upper triangular matrix with positive diagonal elements, and $\boldsymbol{R}_{12} \in \mathbb{R}^{r\times(n-r)}$.

The proof for the existence of the above rank-revealing decomposition for semidefinite matrices is delayed in Section 13.5.3 (p. 132) as a consequence of the spectral decomposition (Theorem 13.1, p. 113) and the column-pivoted QR decomposition (Theorem 3.2, p. 53).

2.8 Application: Rank-One Update/Downdate

Updating linear systems after low-rank modifications of the system matrix is widespread in machine learning, statistics, and many other fields. However, it is well known that this update can lead to serious instabilities in the presence of round-off error (Seeger, 2004). If the system matrix is positive definite, it is almost always possible to use a representation based on the Cholesky decomposition which is much more numerically stable. We will shortly provide the proof for this rank one update/downdate via Cholesky decomposition in this section.

2.8.1 Rank-One Update

A rank-one update $\boldsymbol{A}'$ of matrix $\boldsymbol{A}$ by vector $\boldsymbol{x}$ is of the form:

$$
\begin{aligned}
\boldsymbol{A}' &= \boldsymbol{A} + \boldsymbol{v}\boldsymbol{v}^\top \\
\boldsymbol{R}'^\top \boldsymbol{R}' &= \boldsymbol{R}^\top \boldsymbol{R} + \boldsymbol{v}\boldsymbol{v}^\top.
\end{aligned}
$$

If we have already calculated the Cholesky factor $\boldsymbol{R}$ of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, then the Cholesky factor $\boldsymbol{R}'$ of $\boldsymbol{A}'$ can be calculated efficiently. Note that $\boldsymbol{A}'$ differs from $\boldsymbol{A}$ only via the symmetric rank-one matrix. Hence we can compute $\boldsymbol{R}'$ from $\boldsymbol{R}$ using the rank-one Cholesky update, which takes $O(n^2)$ operations each saving from $O(n^3)$ if we do know $\boldsymbol{R}$, the Cholesky decomposition of $\boldsymbol{A}$ up front, i.e., we want to compute the Cholesky decomposition of $\boldsymbol{A}'$ via that of $\boldsymbol{A}$. To see this, suppose there is a set of orthogonal matrices $\boldsymbol{Q}_n\boldsymbol{Q}_{n-1}\ldots\boldsymbol{Q}_1$ such that that

$$
\boldsymbol{Q}_n\boldsymbol{Q}_{n-1}\ldots\boldsymbol{Q}_1 \begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{R} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{R}' \end{bmatrix}.
$$

Then we find out the expression for the Cholesky factor of $\boldsymbol{A}'$ by $\boldsymbol{R}'$. Specifically, multiply the left-hand side (l.h.s.,) of above equation by its transpose,

$$
\begin{bmatrix} \boldsymbol{v} & \boldsymbol{R}^\top \end{bmatrix} \boldsymbol{Q}_1\ldots\boldsymbol{Q}_{n-1}\boldsymbol{Q}_n\boldsymbol{Q}_n\boldsymbol{Q}_{n-1}\ldots\boldsymbol{Q}_1 \begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{R} \end{bmatrix} = \boldsymbol{R}^\top\boldsymbol{R} + \boldsymbol{v}\boldsymbol{v}^\top.
$$

And multiply the right-hand side (r.h.s.,) by its transpose,

$$
\begin{bmatrix} \mathbf{0} & \boldsymbol{R}'^\top \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \boldsymbol{R}' \end{bmatrix} = \boldsymbol{R}'^\top\boldsymbol{R}',
$$

which agrees with the l.h.s., equation. Givens rotations are such orthogonal matrices that can transfer $\boldsymbol{R}, \boldsymbol{v}$ into $\boldsymbol{R}'$. We will discuss the intrinsic meaning of Givens rotation shortly to prove the existence of QR decomposition in Section 3.12 (p. 60). Here, we only introduce the definition of it and write out the results directly. Feel free to skip this section for a first reading.

Definition 2.11: (n-th Order Givens Rotation)

A Givens rotation is represented by a matrix of the following form

$$
\boldsymbol{G}_{kl} = \begin{bmatrix}
1 & & & & & & & & & \\
 & \ddots & & & & & & & & \\
 & & 1 & & & & & & & \\
 & & & c & & & & s & & \\
 & & & & 1 & & & & & \\
 & & & & & \ddots & & & & \\
 & & & & & & 1 & & & \\
 & & & -s & & & & c & & \\
 & & & & & & & & 1 & \\
 & & & & & & & & & \ddots
\end{bmatrix}_{n\times n},
$$

where the $(k,k),(k,l),(l,k),(l,l)$ entries are $c,s,-s,c$ respectively, and $s = \cos\theta$ and $c = \cos\theta$ for some θ.

Let $\boldsymbol{\delta}_k \in \mathbb{R}^n$ be the zero vector except that the k-th entry is 1. Then mathematically, the Givens rotation defined above can be denoted by

$$
\boldsymbol{G}_{kl} = \boldsymbol{I} + (c-1)(\boldsymbol{\delta}_k\boldsymbol{\delta}_k^\top + \boldsymbol{\delta}_l\boldsymbol{\delta}_l^\top) + s(\boldsymbol{\delta}_k\boldsymbol{\delta}_l^\top - \boldsymbol{\delta}_l\boldsymbol{\delta}_k^\top).
$$

where the subscripts k,l indicate the **rotation is in plane** k **and** l. Specifically, one can also define the n-th order Givens rotation where $(k,k),(k,l),(l,k),(l,l)$ entries are $c, -s, s, c$ respectively. The ideas are the same.

It can be easily verified that the n-th order Givens rotation is an orthogonal matrix and its determinant is 1. For any vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$, we have $\boldsymbol{y} = \boldsymbol{G}_{kl}\boldsymbol{x}$, where

$$
\begin{cases}
y_k = c\cdot x_k + s\cdot x_l, \\
y_l = -s\cdot x_k + c\cdot x_l, \\
y_j = x_j, & (j \neq k, l)
\end{cases}
$$

That is, a Givens rotation applied to $\boldsymbol{x}$ rotates two components of $\boldsymbol{x}$ by some angle θ and leaves all other components the same.

Now, suppose we have an $(n+1)$-th order Givens rotation indexed from 0 to n, and it is given by

$$
\boldsymbol{G}_k = \boldsymbol{I} + (c_k - 1)(\boldsymbol{\delta}_0\boldsymbol{\delta}_0^\top + \boldsymbol{\delta}_k\boldsymbol{\delta}_k^\top) + s_k(\boldsymbol{\delta}_0\boldsymbol{\delta}_k^\top - \boldsymbol{\delta}_k\boldsymbol{\delta}_0^\top),
$$

where $c_k = \cos\theta_k, s_k = \sin\theta_k$ for some θ_k, $\boldsymbol{G}_k \in \mathbb{R}^{(n+1)\times(n+1)}$, $\delta_k \in \mathbb{R}^{n+1}$ is a zero vector except that the $(k+1)$-th entry is 1.

Taking out the k-th column of the following equation

$$
\begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{R}' \end{bmatrix},
$$

where we let the k-th element of $\boldsymbol{v}$ be v_k, and the k-th diagonal of $\boldsymbol{R}$ be r_{kk}. We realize that $\sqrt{v_k^2 + r_{kk}^2} \neq 0$, let $c_k = \frac{r_{kk}}{\sqrt{v_k^2+r_{kk}^2}}$, $s_k = -\frac{v_k}{\sqrt{v_k^2+r_{kk}^2}}$. Then,

$$\begin{cases} v_k \rightarrow c_k v_k + s_k r_{kk} = 0; \\ r_{kk} \rightarrow -s_k v_k + c_k r_{kk} = \sqrt{v_k^2 + r_{kk}^2} = r'_{kk}. \end{cases}$$

That is, G_k will introduce zero value to the k-th element to $\boldsymbol{v}$ and nonzero value to r_{kk}.

This finding above is essential for the rank-one update. And we obtain

$$\boldsymbol{G}_n \boldsymbol{G}_{n-1} \ldots \boldsymbol{G}_1 \begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{R}' \end{bmatrix}.$$

For each Givens rotation, it takes $6n$ flops. And there are n such rotations, which requires $6n^2$ flops if keeping only the leading term. The complexity to calculate the Cholesky factor of $\boldsymbol{A}'$ is thus reduced from $\frac{1}{3}n^3$ to $6n^2$ flops if we already know the Cholesky factor of $\boldsymbol{A}$ by the rank-one update. The above algorithm is essential to reduce the complexity of the posterior calculation in the Bayesian inference for Gaussian mixture model (Lu, 2021a).

2.8.2 Rank-One Downdate

Now suppose we have calculated the Cholesky factor of $\boldsymbol{A}$, and the $\boldsymbol{A}'$ is the downdate of $\boldsymbol{A}$ as follows:

$$\boldsymbol{A}' = \boldsymbol{A} - \boldsymbol{v}\boldsymbol{v}^\top$$
$$\boldsymbol{R}'^\top \boldsymbol{R}' = \boldsymbol{R}^\top \boldsymbol{R} - \boldsymbol{v}\boldsymbol{v}^\top.$$

The algorithm is similar by proceeding as follows:

$$\boldsymbol{G}_1 \boldsymbol{G}_2 \ldots \boldsymbol{G}_n \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{R}' \end{bmatrix}. \tag{2.1}$$

Again, $\boldsymbol{G}_k = \boldsymbol{I} + (c_k - 1)(\boldsymbol{\delta}_0\boldsymbol{\delta}_0^\top + \boldsymbol{\delta}_k\boldsymbol{\delta}_k^\top) + s_k(\boldsymbol{\delta}_0\boldsymbol{\delta}_k^\top - \boldsymbol{\delta}_k\boldsymbol{\delta}_0^\top)$, can be constructed as follows:

Taking out the k-th column of the following equation

$$\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{R} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{R}' \end{bmatrix}.$$

We realize that $r_{kk} \neq 0$, let $c_k = \frac{\sqrt{r_{kk}^2 - v_k^2}}{r_{kk}}$, $s_k = \frac{v_k}{r_{kk}}$. Then,

$$\begin{cases} 0 \rightarrow s_k r_{kk} = v_k; \\ r_{kk} \rightarrow c_k r_{kk} = \sqrt{r_{kk}^2 - v_k^2} = r'_{kk}. \end{cases}$$

This requires $r_{kk}^2 > v_k^2$ to make $\boldsymbol{A}'$ to be positive definite. Otherwise, c_k above will not exist.

Again, one can check that, multiply the l.h.s., of Equation (2.1) by its transpose, we have

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{R}^\top \end{bmatrix} \boldsymbol{G}_n \ldots \boldsymbol{G}_2 \boldsymbol{G}_1 \boldsymbol{G}_1 \boldsymbol{G}_2 \ldots \boldsymbol{G}_n \begin{bmatrix} \mathbf{0} \\ \boldsymbol{R} \end{bmatrix} = \boldsymbol{R}^\top \boldsymbol{R}.$$

And multiply the r.h.s., by its transpose, we have

$$\begin{bmatrix} \boldsymbol{v} & \boldsymbol{R}'^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{R}' \end{bmatrix} = \boldsymbol{v}\boldsymbol{v}^\top + \boldsymbol{R}'^\top \boldsymbol{R}'.$$

This results in $\boldsymbol{R}'^\top \boldsymbol{R}' = \boldsymbol{R}^\top \boldsymbol{R} - \boldsymbol{v}\boldsymbol{v}^\top$.

2.9 Application: Indefinite Rank Two Update

Let $\boldsymbol{A} = \boldsymbol{R}^\top \boldsymbol{R}$ be the Cholesky decomposition of $\boldsymbol{A}$, (Goldfarb, 1976; Seeger, 2004) give a stable method for the indefinite rank-two update of the form

$$\boldsymbol{A}' = (\boldsymbol{I} + \boldsymbol{v}\boldsymbol{u}^\top)\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^\top).$$

Let

$$\begin{cases} \boldsymbol{z} = \boldsymbol{R}^{-\top}\boldsymbol{v}, \\ \boldsymbol{w} = \boldsymbol{R}\boldsymbol{u}, \end{cases} \qquad \rightarrow \qquad \begin{cases} \boldsymbol{v} = \boldsymbol{R}^\top \boldsymbol{z}, \\ \boldsymbol{u} = \boldsymbol{R}^{-1}\boldsymbol{w}. \end{cases}$$

And suppose the LQ decomposition [10] of $\boldsymbol{I} + \boldsymbol{z}\boldsymbol{w}^\top$ is given by $\boldsymbol{I} + \boldsymbol{z}\boldsymbol{w}^\top = \boldsymbol{L}\boldsymbol{Q}$, where $\boldsymbol{L}$ is lower triangular and $\boldsymbol{Q}$ is orthogonal. Thus, we have

$$\begin{aligned} \boldsymbol{A}' &= (\boldsymbol{I} + \boldsymbol{v}\boldsymbol{u}^\top)\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^\top) \\ &= (\boldsymbol{I} + \boldsymbol{R}^\top \boldsymbol{z}\boldsymbol{w}^\top \boldsymbol{R}^{-\top})\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{R}^{-1}\boldsymbol{w}\boldsymbol{z}^\top \boldsymbol{R}) \\ &= \boldsymbol{R}^\top(\boldsymbol{I} + \boldsymbol{z}\boldsymbol{w}^\top)(\boldsymbol{I} + \boldsymbol{w}\boldsymbol{z}^\top)\boldsymbol{R} \\ &= \boldsymbol{R}^\top \boldsymbol{L}\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{L}^\top \boldsymbol{R} \\ &= \boldsymbol{R}^\top \boldsymbol{L}\boldsymbol{L}^\top \boldsymbol{R}. \end{aligned}$$

Let $\boldsymbol{R}' = \boldsymbol{R}^\top \boldsymbol{L}$ which is lower triangular, we find the Cholesky decomposition of $\boldsymbol{A}'$.

Part II
Triangularization, Orthogonalization and Gram-Schmidt Process

3. QR Decomposition

In many applications, we are interested in the column space of a matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}$. The successive spaces spanned by the columns $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots$ of $\boldsymbol{A}$ are

$$\mathcal{C}([\boldsymbol{a}_1]) \quad \subseteq \quad \mathcal{C}([\boldsymbol{a}_1, \boldsymbol{a}_2]) \quad \subseteq \quad \mathcal{C}([\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3]) \quad \subseteq \quad \ldots,$$

10. We will shortly introduce in Theorem 3.11 (p. 65).

where $\mathcal{C}([\ldots])$ is the subspace spanned by the vectors included in the brackets. The idea of QR decomposition is the construction of a sequence of orthonormal vectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots$ that span the same successive subspaces.

$$
\left\{\mathcal{C}([\boldsymbol{q}_1]) = \mathcal{C}([\boldsymbol{a}_1])\right\} \subseteq \left\{\mathcal{C}([\boldsymbol{q}_1, \boldsymbol{q}_2]) = \mathcal{C}([\boldsymbol{a}_1, \boldsymbol{a}_2])\right\} \subseteq \left\{\mathcal{C}([\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3]) = \mathcal{C}([\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3])\right\} \subseteq \ldots,
$$

We provide the result of QR decomposition in the following theorem and we delay the discussion of its existence in the next sections.

Theorem 3.1: (QR Decomposition)

Every $m \times n$ matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n]$ (whether linearly independent or dependent columns) with $m \geq n$ can be factored as

$$
\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R},
$$

where

1. **Reduced**: $\boldsymbol{Q}$ is $m \times n$ with orthonormal columns and $\boldsymbol{R}$ is an $n \times n$ upper triangular matrix which is known as the **reduced QR decomposition**;
2. **Full**: $\boldsymbol{Q}$ is $m \times m$ with orthonormal columns and $\boldsymbol{R}$ is an $m \times n$ upper triangular matrix which is known as the **full QR decomposition**. If further restrict the upper triangular matrix to be a square matrix, the full QR decomposition can be denoted as
$$
\boldsymbol{A} = \boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}_0 \\ \boldsymbol{0}\end{bmatrix},
$$
where $\boldsymbol{R}_0$ is an $m \times m$ upper triangular matrix.

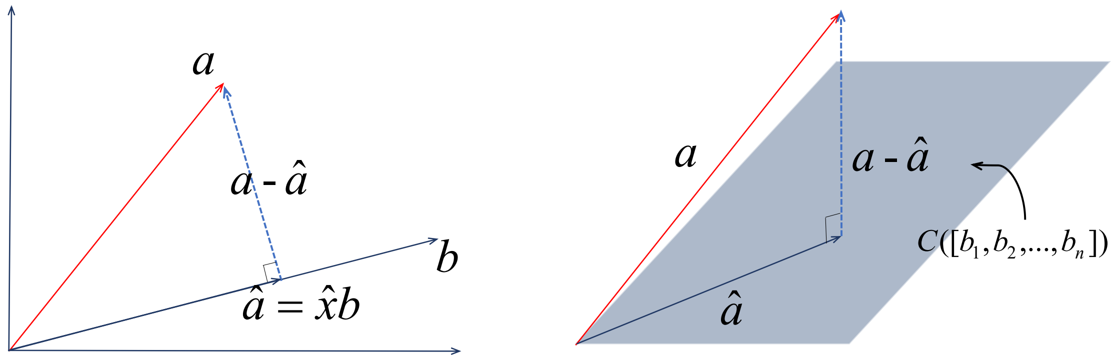
Specifically, when $\boldsymbol{A}$ has full rank, i.e., has linearly independent columns, $\boldsymbol{R}$ also has linearly independent columns, and $\boldsymbol{R}$ is nonsingular for the *reduced* case. This implies diagonals of $\boldsymbol{R}$ are nonzero. Under this condition, when we further restrict elements on the diagonal of $\boldsymbol{R}$ are positive, the *reduced* QR decomposition is **unique**. The *full* QR decomposition is normally not unique since the right-most $(m-n)$ columns in $\boldsymbol{Q}$ can be in any order.

3.1 Project a Vector Onto Another Vector

Project a vector $\boldsymbol{a}$ to a vector $\boldsymbol{b}$ is to find the vector closest to $\boldsymbol{a}$ on the line of $\boldsymbol{b}$. The projection vector $\widehat{\boldsymbol{a}}$ is some multiple of $\boldsymbol{b}$. Let $\widehat{\boldsymbol{a}} = \widehat{x}\boldsymbol{b}$ and $\boldsymbol{a} - \widehat{\boldsymbol{a}}$ is perpendicular to $\boldsymbol{b}$ as shown in Figure 5(a). We then get the following result:

Project Vector $\boldsymbol{a}$ Onto Vector $\boldsymbol{b}$

$\boldsymbol{a}^\perp = \boldsymbol{a} - \widehat{\boldsymbol{a}}$ is perpendicular to $\boldsymbol{b}$, so $(\boldsymbol{a} - \widehat{x}\boldsymbol{b})^\top \boldsymbol{b} = 0$: $\widehat{x} = \frac{\boldsymbol{a}^\top\boldsymbol{b}}{\boldsymbol{b}^\top\boldsymbol{b}}$ and $\widehat{\boldsymbol{a}} = \frac{\boldsymbol{a}^\top\boldsymbol{b}}{\boldsymbol{b}^\top\boldsymbol{b}}\boldsymbol{b} = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\boldsymbol{b}^\top\boldsymbol{b}}\boldsymbol{a}$.



(a) Project onto a line (b) Project onto a space

Figure 5: Project a vector onto a line and a space.

3.2 Project a Vector Onto a Plane

Project a vector $\boldsymbol{a}$ to a space spanned by $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$ is to find the vector closest to $\boldsymbol{a}$ on the column space of $[\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n]$. The projection vector $\widehat{\boldsymbol{a}}$ is a combination of $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$: $\widehat{\boldsymbol{a}} = \widehat{x}_1\boldsymbol{b}_1 + \widehat{x}_2\boldsymbol{b}_2 + \ldots + \widehat{x}_n\boldsymbol{b}_n$. This is actually a least squares problem. To find the projection, we just solve the normal equation $\boldsymbol{B}^\top\boldsymbol{B}\widehat{\boldsymbol{x}} = \boldsymbol{B}^\top\boldsymbol{a}$ where $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n]$ and $\widehat{\boldsymbol{x}} = [\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n]$. We refer the details of this projection view in the least squares to (Strang, 2009; Trefethen and Bau III, 1997; Yang, 2000; Golub and Van Loan, 2013; Lu, 2021e) as it is not the main interest of this survey. For each vector $\boldsymbol{b}_i$, the projection of $\boldsymbol{a}$ in the direction of $\boldsymbol{b}_i$ can be analogously obtained by

$$\widehat{\boldsymbol{a}}_i = \frac{\boldsymbol{b}_i\boldsymbol{b}_i^\top}{\boldsymbol{b}_i^\top\boldsymbol{b}_i}\boldsymbol{a}, \qquad \forall i \in \{1, 2, \ldots, n\}.$$

Let $\widehat{\boldsymbol{a}} = \sum_{i=1}^n \widehat{\boldsymbol{a}}_i$, this results in

$$\boldsymbol{a}^\perp = (\boldsymbol{a} - \widehat{\boldsymbol{a}}) \perp \mathcal{C}(\boldsymbol{B}),$$

i.e., $(\boldsymbol{a} - \widehat{\boldsymbol{a}})$ is perpendicular to the column space of $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n]$ as shown in Figure 5(b).

3.3 Existence of the QR Decomposition via the Gram-Schmidt Process

For three linearly independent vectors $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3\}$ and the space spanned by the three linearly independent vectors $\mathcal{C}([\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3])$, i.e., the column space of the matrix $[\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3]$. We intend to construct three orthogonal vectors $\{\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3\}$ in which case $\mathcal{C}([\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3]) = \mathcal{C}([\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3])$. Then we divide the orthogonal vectors by their length to normalize. This process produces three mutually orthonormal vectors $\boldsymbol{q}_1 = \frac{\boldsymbol{b}_1}{||\boldsymbol{b}_1||}$, $\boldsymbol{q}_2 = \frac{\boldsymbol{b}_2}{||\boldsymbol{b}_2||}$, $\boldsymbol{q}_2 = \frac{\boldsymbol{b}_2}{||\boldsymbol{b}_2||}$.

For the first vector, we choose $\boldsymbol{b}_1 = \boldsymbol{a}_1$ directly. The second vector $\boldsymbol{b}_2$ must be perpendicular to the first one. This is actually the vector $\boldsymbol{a}_2$ subtracting its projection along

$\boldsymbol{b}_1$:

$$
\begin{aligned}
\boldsymbol{b}_2 = \boldsymbol{a}_2 - \frac{\boldsymbol{b}_1\boldsymbol{b}_1^\top}{\boldsymbol{b}_1^\top\boldsymbol{b}_1}\boldsymbol{a}_2 &= (\boldsymbol{I} - \frac{\boldsymbol{b}_1\boldsymbol{b}_1^\top}{\boldsymbol{b}_1^\top\boldsymbol{b}_1})\boldsymbol{a}_2 && \text{(Projection view)}\\
&= \boldsymbol{a}_2 - \underbrace{\frac{\boldsymbol{b}_1^\top\boldsymbol{a}_2}{\boldsymbol{b}_1^\top\boldsymbol{b}_1}\boldsymbol{b}_1}_{\widehat{\boldsymbol{a}}_2}, && \text{(Combination view)}
\end{aligned}
$$

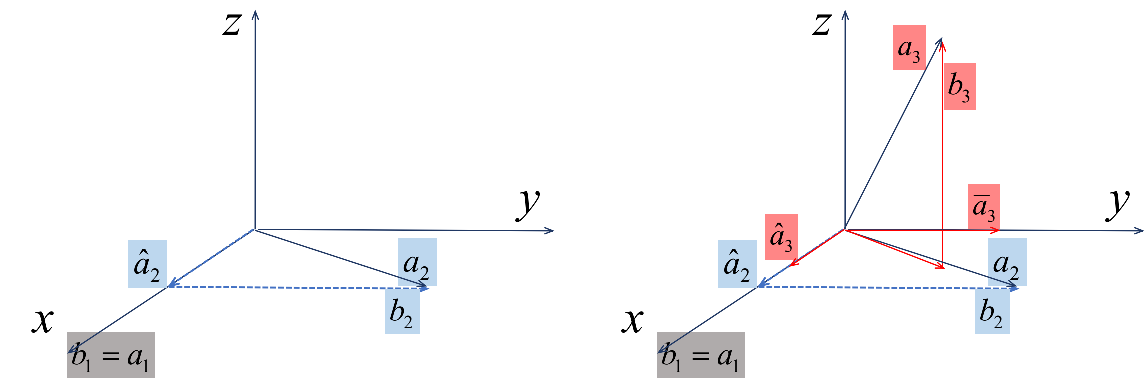
where the first equation shows $\boldsymbol{b}_2$ is a multiplication of the matrix $(\boldsymbol{I} - \frac{\boldsymbol{b}_1\boldsymbol{b}_1^\top}{\boldsymbol{b}_1^\top\boldsymbol{b}_1})$ and the vector $\boldsymbol{a}_2$, i.e., project $\boldsymbol{a}_2$ onto the orthogonal complement space of $\mathcal{C}([\boldsymbol{b}_1])$. The second equality in the above equation shows $\boldsymbol{a}_2$ is a combination of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$. Clearly, the space spanned by $\boldsymbol{b}_1, \boldsymbol{b}_2$ is the same space spanned by $\boldsymbol{a}_1, \boldsymbol{a}_2$. The situation is shown in Figure 6(a) in which we choose **the direction of $\boldsymbol{b}_1$ as the x-axis in the Cartesian coordinate system**. $\widehat{\boldsymbol{a}}_2$ is the projection of $\boldsymbol{a}_2$ onto line $\boldsymbol{b}_1$. It can be clearly shown that the part of $\boldsymbol{a}_2$ perpendicular to $\boldsymbol{b}_1$ is $\boldsymbol{b}_2 = \boldsymbol{a}_2 - \widehat{\boldsymbol{a}}_2$ from the figure.

For the third vector $\boldsymbol{b}_3$, it must be perpendicular to both the $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ which is actually the vector $\boldsymbol{a}_3$ subtracting its projection along the plane spanned by $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$

$$
\begin{aligned}
\boldsymbol{b}_3 = \boldsymbol{a}_3 - \frac{\boldsymbol{b}_1\boldsymbol{b}_1^\top}{\boldsymbol{b}_1^\top\boldsymbol{b}_1}\boldsymbol{a}_3 - \frac{\boldsymbol{b}_2\boldsymbol{b}_2^\top}{\boldsymbol{b}_2^\top\boldsymbol{b}_2}\boldsymbol{a}_3 &= (\boldsymbol{I} - \frac{\boldsymbol{b}_1\boldsymbol{b}_1^\top}{\boldsymbol{b}_1^\top\boldsymbol{b}_1} - \frac{\boldsymbol{b}_2\boldsymbol{b}_2^\top}{\boldsymbol{b}_2^\top\boldsymbol{b}_2})\boldsymbol{a}_3 && \text{(Projection view)}\\
&= \boldsymbol{a}_3 - \underbrace{\frac{\boldsymbol{b}_1^\top\boldsymbol{a}_3}{\boldsymbol{b}_1^\top\boldsymbol{b}_1}\boldsymbol{b}_1}_{\widehat{\boldsymbol{a}}_3} - \underbrace{\frac{\boldsymbol{b}_2^\top\boldsymbol{a}_3}{\boldsymbol{b}_2^\top\boldsymbol{b}_2}\boldsymbol{b}_2}_{\bar{\boldsymbol{a}}_3}, && \text{(Combination view)}
\end{aligned}
\tag{3.1}
$$

where the first equation shows $\boldsymbol{b}_3$ is a multiplication of the matrix $(\boldsymbol{I} - \frac{\boldsymbol{b}_1\boldsymbol{b}_1^\top}{\boldsymbol{b}_1^\top\boldsymbol{b}_1} - \frac{\boldsymbol{b}_2\boldsymbol{b}_2^\top}{\boldsymbol{b}_2^\top\boldsymbol{b}_2})$ and the vector $\boldsymbol{a}_3$, i.e., project $\boldsymbol{a}_3$ onto the orthogonal complement space of $\mathcal{C}([\boldsymbol{b}_1, \boldsymbol{b}_2])$. The second equality in the above equation shows $\boldsymbol{a}_3$ is a combination of $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$. We will see this property is essential in the idea of the QR decomposition. Again, it can be shown that the space spanned by $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ is the same space spanned by $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$. The situation is shown in Figure 6(b), in which we choose **the direction of $\boldsymbol{b}_2$ as the y-axis of the Cartesian coordinate system**. $\widehat{\boldsymbol{a}}_3$ is the projection of $\boldsymbol{a}_3$ onto line $\boldsymbol{b}_1$, $\bar{\boldsymbol{a}}_3$ is the projection of $\boldsymbol{a}_3$ onto line $\boldsymbol{b}_2$. It can be shown that the part of $\boldsymbol{a}_3$ perpendicular to both $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ is $\boldsymbol{b}_3 = \boldsymbol{a}_3 - \widehat{\boldsymbol{a}}_3 - \bar{\boldsymbol{a}}_3$ from the figure.

Finally, we normalize each vector by dividing their length which produces three orthonormal vectors $\boldsymbol{q}_1 = \frac{\boldsymbol{b}_1}{||\boldsymbol{b}_1||}$, $\boldsymbol{q}_2 = \frac{\boldsymbol{b}_2}{||\boldsymbol{b}_2||}$, $\boldsymbol{q}_2 = \frac{\boldsymbol{b}_2}{||\boldsymbol{b}_2||}$.



(a) Project $\boldsymbol{a}_2$ onto the space perpendicular to $\boldsymbol{b}_1$.

(b) Project $\boldsymbol{a}_3$ onto the space perpendicular to $\boldsymbol{b}_1, \boldsymbol{b}_2$.

Figure 6: The Gram-Schmidt process.

This idea can be extended to a set of vectors rather than only three. And we call this process as *Gram-Schmidt process*. After this process, matrix $\boldsymbol{A}$ will be triangularized. The method is named after Jørgen Pedersen Gram and Erhard Schmidt, but it appeared earlier in the work of Pierre-Simon Laplace in the theory of Lie group decomposition.

As we mentioned previously, the idea of the QR decomposition is the construction of a sequence of orthonormal vectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots$ that span the same successive subspaces.

$$\left\{\mathcal{C}([\boldsymbol{q}_1]) = \mathcal{C}([\boldsymbol{a}_1])\right\} \subseteq \left\{\mathcal{C}([\boldsymbol{q}_1, \boldsymbol{q}_2]) = \mathcal{C}([\boldsymbol{a}_1, \boldsymbol{a}_2])\right\} \subseteq \left\{\mathcal{C}([\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3]) = \mathcal{C}([\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3])\right\} \subseteq \ldots,$$

This implies any $\boldsymbol{a}_k$ is in the space spanned by $\mathcal{C}([\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_k])$. [11] As long as we have found these orthonormal vectors, to reconstruct $\boldsymbol{a}_i$'s from the orthonormal matrix $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n]$, an upper triangular matrix $\boldsymbol{R}$ is needed such that $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$.

The Gram–Schmidt process is not the only algorithm for finding the QR decomposition. Several other QR decomposition algorithms exist such as Householder reflections and Givens rotations which are more reliable in the presence of round-off errors. These QR decomposition methods may also change the order in which the columns of $\boldsymbol{A}$ are processed.

3.4 Orthogonal vs Orthonormal

The vectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n \in \mathbb{R}^m$ are mutually orthogonal when their dot products $\boldsymbol{q}_i^\top \boldsymbol{q}_j$ are zero whenever $i \neq j$. When each vector is divided by its length, the vectors become orthogonal unit vectors. Then the vectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$ are mutually orthonormal. We put the orthonormal vectors into a matrix $\boldsymbol{Q}$.

When $m \neq n$: the matrix $\boldsymbol{Q}$ is easy to work with because $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I} \in \mathbb{R}^{n \times n}$. Such $\boldsymbol{Q}$ with $m \neq n$ is sometimes referred to as a **semi-orthogonal** matrix.

When $m = n$: the matrix $\boldsymbol{Q}$ is square, $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}$ means that $\boldsymbol{Q}^\top = \boldsymbol{Q}^{-1}$, i.e., the transpose of $\boldsymbol{Q}$ is also the inverse of $\boldsymbol{Q}$. Then we also have $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$, i.e., $\boldsymbol{Q}^\top$ is the

11. And also, any $\boldsymbol{q}_k$ is in the space spanned by $\mathcal{C}([\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k])$.

two-sided inverse of $\boldsymbol{Q}$. We call this $\boldsymbol{Q}$ an **orthogonal matrix.** [12] To see this, we have

$$\begin{bmatrix}\boldsymbol{q}_1^\top\\ \boldsymbol{q}_2^\top\\ \vdots\\ \boldsymbol{q}_n^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{q}_1 & \boldsymbol{q}_2 & \ldots & \boldsymbol{q}_n\end{bmatrix}=\begin{bmatrix}1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}.$$

In other words, $\boldsymbol{q}_i^\top\boldsymbol{q}_j=\delta_{ij}$ where δ_{ij} is the Kronecker delta. The columns of an orthogonal matrix $\boldsymbol{Q}\in\mathbb{R}^{n\times n}$ form an **orthonormal basis** of $\mathbb{R}^n$.

Orthogonal matrices can be viewed as matrices that change the basis of other matrices. Hence they preserve the angle (inner product) between the vectors: $\boldsymbol{u}^\top\boldsymbol{v}=(\boldsymbol{Q}\boldsymbol{u})^\top(\boldsymbol{Q}\boldsymbol{v})$. This invariance of the inner products of angles between the vectors is preserved, which also relies on the invariance of their lengths: $||\boldsymbol{Q}\boldsymbol{u}||=||\boldsymbol{u}||$. In real cases, multiplied by a orthogonal matrix $\boldsymbol{Q}$ will rotate (if $\det(\boldsymbol{Q})=1$) or reflect (if $\det(\boldsymbol{Q})=-1$) the original vector space. Many decomposition algorithms will result in two orthogonal matrices, thus such rotations or reflections will happen twice.

3.5 Computing the Reduced QR Decomposition via CGS and MGS

We write out this form of the reduced QR Decomposition such that $\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{Q}\in\mathbb{R}^{m\times n}$ and $\boldsymbol{R}\in\mathbb{R}^{n\times n}$:

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{a}_1 & \boldsymbol{a}_2 & ... & \boldsymbol{a}_n\end{bmatrix}=\begin{bmatrix}\boldsymbol{q}_1 & \boldsymbol{q}_2 & ... & \boldsymbol{q}_n\end{bmatrix}\begin{bmatrix}r_{11} & r_{12} & \ldots & r_{1n}\\ & r_{22} & \ldots & r_{2n}\\ \text{\huge 0} & & \ddots & \vdots\\ & & & r_{nn}\end{bmatrix}.$$

The orthonormal matrix $\boldsymbol{Q}$ can be easily calculated by the Gram-Schmidt process. To see why we have the upper triangular matrix $\boldsymbol{R}$, we write out these equations

$$\begin{aligned}
\boldsymbol{a}_1 &= r_{11}\boldsymbol{q}_1 &&= \sum_{i=1}^{1} r_{i1}\boldsymbol{q}_1,\\
&\vdots\\
\boldsymbol{a}_k &= r_{1k}\boldsymbol{q}_1+r_{2k}\boldsymbol{q}_2+\ldots+r_{kk}\boldsymbol{q}_k &&= \sum_{i=1}^{k} r_{ik}\boldsymbol{q}_k,\\
&\vdots
\end{aligned}$$

which coincides with the second equation of Equation (3.1) and conforms to the form of an upper triangular matrix $\boldsymbol{R}$. And if we extend the idea of Equation (3.1) into the k-th term, we will get

$$\boldsymbol{a}_k=\sum_{i=1}^{k-1}(\boldsymbol{q}_i^\top\boldsymbol{a}_k)\boldsymbol{q}_i+\boldsymbol{a}_k^\perp=\sum_{i=1}^{k-1}(\boldsymbol{q}_i^\top\boldsymbol{a}_k)\boldsymbol{q}_i+||\boldsymbol{a}_k^\perp||\cdot\boldsymbol{q}_k,$$

12. Note here we use the term *orthogonal matrix* to mean the matrix $\boldsymbol{Q}$ has orthonormal columns. The term *orthonormal matrix* is **not** used for historical reasons.

where $\boldsymbol{a}_k^\perp$ is such $\boldsymbol{b}_k$ in Equation (3.1) that we emphasize the "perpendicular" property here. This implies we can gradually orthonormalize $\boldsymbol{A}$ to an orthonormal set $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n]$ by

$$
\begin{cases}
r_{ik} = \boldsymbol{q}_i^\top \boldsymbol{a}_k, \quad \forall i \in \{1, 2, \ldots, k-1\}; \\
\boldsymbol{a}_k^\perp = \boldsymbol{a}_k - \sum_{i=1}^{k-1} r_{ik}\boldsymbol{q}_i; \\
r_{kk} = ||\boldsymbol{a}_k^\perp||; \\
\boldsymbol{q}_k = \boldsymbol{a}_k^\perp / r_{kk}.
\end{cases}
\tag{3.2}
$$

Orthogonal Projection

We notice again from Equation (3.2), the first two equality imply that

$$
\left.
\begin{aligned}
r_{ik} &= \boldsymbol{q}_i^\top \boldsymbol{a}_k, \quad \forall i \in \{1, 2, \ldots, k-1\} \\
\boldsymbol{a}_k^\perp &= \boldsymbol{a}_k - \sum_{i=1}^{k-1} r_{ik}\boldsymbol{q}_i
\end{aligned}
\right\} \rightarrow \boldsymbol{a}_k^\perp = \boldsymbol{a}_k - \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top \boldsymbol{a}_k = (\boldsymbol{I} - \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top)\boldsymbol{a}_k,
\tag{3.3}
$$

where $\boldsymbol{Q}_{k-1} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{k-1}]$. This implies $\boldsymbol{q}_k$ can be obtained by

$$
\boldsymbol{q}_k = \frac{\boldsymbol{a}_k^\perp}{||\boldsymbol{a}_k^\perp||} = \frac{(\boldsymbol{I} - \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top)\boldsymbol{a}_k}{||(\boldsymbol{I} - \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top)\boldsymbol{a}_k||}.
$$

The matrix $(\boldsymbol{I} - \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top)$ in above equation is known as an *orthogonal projection matrix* that will project $\boldsymbol{a}_k$ along the column space of $\boldsymbol{Q}_{k-1}$, i.e., project a vector so that the vector is perpendicular to the column space of $\boldsymbol{Q}_{k-1}$. The net result is that the $\boldsymbol{a}_k^\perp$ or $\boldsymbol{q}_k$ calculated in this way will be orthogonal to the $\mathcal{C}(\boldsymbol{Q}_{k-1})$, i.e., in the null space of $\boldsymbol{Q}_{k-1}^\top$: $\mathcal{N}(\boldsymbol{Q}_{k-1}^\top)$ by the fundamental theorem of linear algebra (Theorem 0.15, p. 14).

Let $\boldsymbol{P}_1 = (\boldsymbol{I} - \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top)$ and we claimed above $\boldsymbol{P}_1 = (\boldsymbol{I} - \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top)$ is an orthogonal projection matrix such that $\boldsymbol{P}_1\boldsymbol{v}$ will project the $\boldsymbol{v}$ onto the null space of $\boldsymbol{Q}_{k-1}$. And actually, let $\boldsymbol{P}_2 = \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top$, then $\boldsymbol{P}_2$ is also an orthogonal projection matrix such that $\boldsymbol{P}_2\boldsymbol{v}$ will project the $\boldsymbol{v}$ onto the column space of $\boldsymbol{Q}_{k-1}$.

But why do the matrix $\boldsymbol{P}_1, \boldsymbol{P}_2$ can magically project a vector onto the corresponding subspaces? It can be easily shown that the column space of $\boldsymbol{Q}_{k-1}$ is equal to the column space of $\boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top$:

$$
\mathcal{C}(\boldsymbol{Q}_{k-1}) = \mathcal{C}(\boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top) = \mathcal{C}(\boldsymbol{P}_2).
$$

Therefore, the result of $\boldsymbol{P}_2\boldsymbol{v}$ is a linear combination of the columns of $\boldsymbol{P}_2$, which is in the column space of $\boldsymbol{P}_2$ or the column space of $\boldsymbol{Q}_{k-1}$. The formal definition of a *projection matrix* $\boldsymbol{P}$ is that it is idempotent $\boldsymbol{P}^2 = \boldsymbol{P}$ such that projecting twice is equal to projecting once. What makes the above $\boldsymbol{P}_2 = \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top$ different is that the projection $\widehat{\boldsymbol{v}}$ of any vector $\boldsymbol{v}$ is perpendicular to $\boldsymbol{v} - \widehat{\boldsymbol{v}}$:

$$
(\widehat{\boldsymbol{v}} = \boldsymbol{P}_2\boldsymbol{v}) \perp (\boldsymbol{v} - \widehat{\boldsymbol{v}}).
$$

This goes to the original definition we gave above: the *orthogonal projection matrix*. To avoid confusion, one may use the term *oblique projection matrix* in the nonorthogonal case.

When $\boldsymbol{P}_2$ is an orthogonal projection matrix, $\boldsymbol{P}_1 = \boldsymbol{I} - \boldsymbol{P}_2$ is also an orthogonal projection matrix that will project any vector onto the space perpendicular to the $\mathcal{C}(\boldsymbol{Q}_{k-1})$, i.e., $\mathcal{N}(\boldsymbol{Q}_{k-1}^\top)$. Therefore, we conclude the two orthogonal projections:

$$
\begin{cases}
\boldsymbol{P}_1: & \text{project onto } \mathcal{N}(\boldsymbol{Q}_{k-1}^\top);\\
\boldsymbol{P}_2: & \text{project onto } \mathcal{C}(\boldsymbol{Q}_{k-1}).
\end{cases}
$$

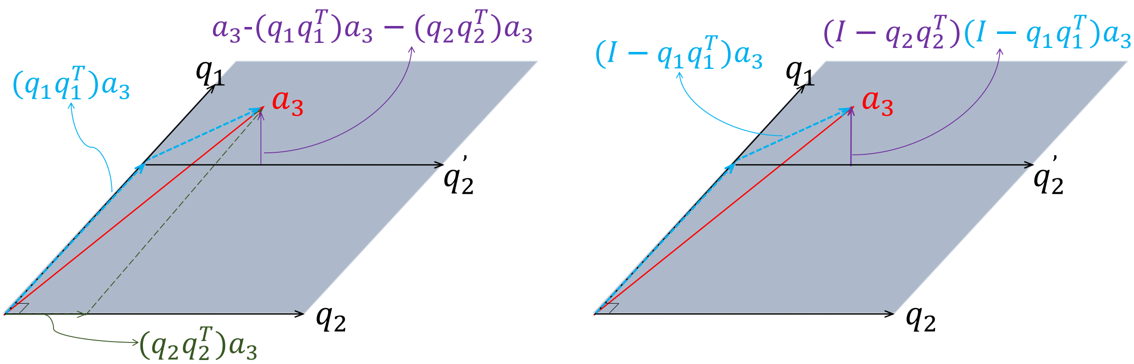
The further result that is important to notice is when the columns of $\boldsymbol{Q}_{k-1}$ are mutually orthonormal, we have the following decomposition:

$$
\boxed{\boldsymbol{P}_1 = \boldsymbol{I} - \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top = (\boldsymbol{I} - \boldsymbol{q}_1\boldsymbol{q}_1^\top)(\boldsymbol{I} - \boldsymbol{q}_2\boldsymbol{q}_2^\top)\ldots(\boldsymbol{I} - \boldsymbol{q}_{k-1}\boldsymbol{q}_{k-1}^\top),} \tag{3.4}
$$

where $\boldsymbol{Q}_{k-1} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{k-1}]$ and each $(\boldsymbol{I} - \boldsymbol{q}_i\boldsymbol{q}_i^\top)$ is to project a vector into the perpendicular space of $\boldsymbol{q}_i$. This finding is important to make a step further to a *modified Gram-Schmidt process (MGS)* where we project and subtract on the fly. To avoid confusion, the previous Gram-Schmidt is called the *classical Gram-Schmidt process (CGS)*. The difference between the CGS and MGS is, in the CGS, we project the same vector onto the orthogonormal ones and subtract afterwards. However, in the MGS, the projection and subtraction are done in an interleaved manner. A three-column example $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3]$ is shown in Figure 7 where each step is denoted in a different color. We summarize the difference between the CGS and MGS processes for obtaining $\boldsymbol{q}_k$ via the k-th column $\boldsymbol{a}_k$ of $\boldsymbol{A}$ and the orthonormalized vectors $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{k-1}\}$:

$$
\begin{aligned}
&\text{(CGS)}: \text{obtain } \boldsymbol{q}_k \text{ by normalizing } \boldsymbol{a}_k^\perp = (\boldsymbol{I} - \boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^\top)\boldsymbol{a}_k;\\
&\text{(MGS)}: \text{obtain } \boldsymbol{q}_k \text{ by normalizing } \boldsymbol{a}_k^\perp = \left\{(\boldsymbol{I} - \boldsymbol{q}_{k-1}\boldsymbol{q}_{k-1}^\top)\ldots\left[(\boldsymbol{I} - \boldsymbol{q}_2\boldsymbol{q}_2^\top)\left((\boldsymbol{I} - \boldsymbol{q}_1\boldsymbol{q}_1^\top)\boldsymbol{a}_k\right)\right]\right\},
\end{aligned}
$$

where the parentheses of the MGS indicate the order of the computation.



(a) CGS, step 1: blue vector; step 2: green vector; step 3: purple vector.

(b) MGS, step 1: blue vector; step 2: purple vector.

Figure 7: CGS vs MGS in 3-dimensional space where $\boldsymbol{q}_2'$ is parallel to $\boldsymbol{q}_2$ so that projecting on $\boldsymbol{q}_2$ is equivalent to projecting on $\boldsymbol{q}_2'$.

What's the difference? Taking the three-column matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3]$ as an example. Suppose we have computed $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ such that $span\{\boldsymbol{q}_1, \boldsymbol{q}_2\} = span\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$. And we want to proceed to compute the $\boldsymbol{q}_3$.

In the CGS, the orthogonalization of column $\boldsymbol{a}_n$ against column $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ is performed by projecting the original column $\boldsymbol{a}_3$ of $\boldsymbol{A}$ onto $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ respectively and subtracting at once:

$$
\begin{cases}
\boldsymbol{a}_3^\perp = \boldsymbol{a}_3 - (\boldsymbol{q}_1^\top \boldsymbol{a}_3)\boldsymbol{q}_1 - (\boldsymbol{q}_2^\top \boldsymbol{a}_3)\boldsymbol{q}_2 \\
\qquad = \boldsymbol{a}_3 - (\boldsymbol{q}_1\boldsymbol{q}_1^\top)\boldsymbol{a}_3 - \boxed{(\boldsymbol{q}_2\boldsymbol{q}_2^\top)\boldsymbol{a}_3} \\
\boldsymbol{q}_3 = \dfrac{\boldsymbol{a}_3^\perp}{||\boldsymbol{a}_3^\perp||},
\end{cases}
\tag{3.5}
$$

as shown in Figure 7(a).

In the MGS, on the other hand, the components along each $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ are immediately subtracted out of rest of the column $\boldsymbol{a}_3$ as soon as the $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ are computed. Therefore the orthogonalization of column $\boldsymbol{a}_3$ against $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ is not performed by projecting the original column $\boldsymbol{a}_3$ against $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ as it is in the CGS, but rather against a vector obtained by subtracting from that column $\boldsymbol{a}_3$ of $\boldsymbol{A}$ the components in the direction of $\boldsymbol{q}_1, \boldsymbol{q}_2$ successively. This is important because the error components of $\boldsymbol{q}_i$ in *span*$\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ will be smaller (we will further discuss in the next paragraphs).

More precisely, in the MGS the orthogonalization of column $\boldsymbol{a}_3$ against $\boldsymbol{q}_1$ is performed by subtracting the component of $\boldsymbol{q}_1$ from the vector $\boldsymbol{a}_3$:

$$
\boldsymbol{a}_3^{(1)} = (\boldsymbol{I} - \boldsymbol{q}_1\boldsymbol{q}_1^\top)\boldsymbol{a}_3 = \boldsymbol{a}_3 - (\boldsymbol{q}_1\boldsymbol{q}_1^\top)\boldsymbol{a}_3,
$$

where $\boldsymbol{a}_3^{(1)}$ is the component of $\boldsymbol{a}_3$ lies in a space perpendicular to $\boldsymbol{q}_1$. And further step is performed by

$$
\begin{aligned}
\boldsymbol{a}_3^{(2)} = (\boldsymbol{I} - \boldsymbol{q}_2\boldsymbol{q}_2^\top)\boldsymbol{a}_3^{(1)} = \boldsymbol{a}_3^{(1)} - (\boldsymbol{q}_2\boldsymbol{q}_2^\top)\boldsymbol{a}_3^{(1)} \\
= \boldsymbol{a}_3 - (\boldsymbol{q}_1\boldsymbol{q}_1^\top)\boldsymbol{a}_3 - \boxed{(\boldsymbol{q}_2\boldsymbol{q}_2^\top)\boldsymbol{a}_3^{(1)}}
\end{aligned}
\tag{3.6}
$$

where $\boldsymbol{a}_3^{(2)}$ is the component of $\boldsymbol{a}_3^{(1)}$ lies in a space perpendicular to $\boldsymbol{q}_2$ and we highlight the difference to the CGS in Equation (3.5) by blue text. This net result is that $\boldsymbol{a}_3^{(2)}$ is the component of $\boldsymbol{a}_3$ lies in the space perpendicular to $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ as shown in Figure 7(b).

Main difference and catastrophic cancellation The key difference is that the $\boldsymbol{a}_3$ can in general have large components in *span*$\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ in which case one starts with large values and ends up with small values with large relative errors in them. This is known as the problem of *catastrophic cancellation*. Whereas $\boldsymbol{a}_3^{(1)}$ is in the direction perpendicular to $\boldsymbol{q}_1$ and has only a small "error" component in the direction of $\boldsymbol{q}_1$. Compare the $\boxed{\text{boxed}}$ text in Equation (3.5) and (3.6), it is not hard to see $(\boldsymbol{q}_2\boldsymbol{q}_2^\top)\boldsymbol{a}_3^{(1)}$ in Equation (3.6) is more accurate by the above argument. And thus, because of the much smaller error in this projection factor, the MGS introduces less orthogonalization error at each subtraction step than that is in the CGS. In fact, it can be shown that the final $\boldsymbol{Q}$ obtained in the CGS satisfies

$$
||\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\top|| \leq O(\epsilon\kappa^2(\boldsymbol{A})),
$$

where $\kappa(\boldsymbol{A})$ is a value larger than 1 determined by $\boldsymbol{A}$. Whereas, in the MGS, the error satisfies

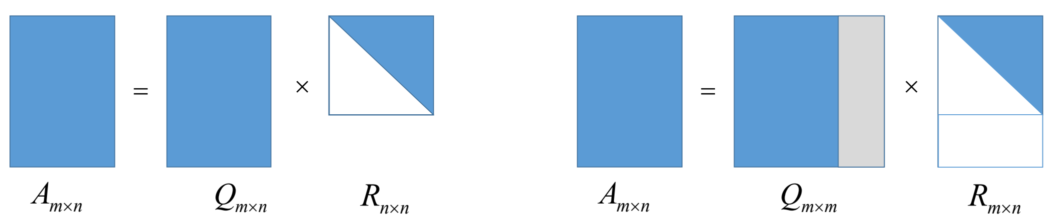
$$
||\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\top|| \leq O(\epsilon\kappa(\boldsymbol{A})).
$$

That is, the $\boldsymbol{Q}$ obtained in the MGS is more orthogonal.

More to go, preliminaries for Householder and Givens methods Although, we claimed here the MGS usually works better than the CGS in practice. The MGS can still fall victim to the *catastrophic cancellation* problem. Suppose in iteration k of the MGS algorithm, $\boldsymbol{a}_k$ is almost in the span of $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{k-1}\}$. This will result in that $\boldsymbol{a}_k^\perp$ has only a small component that is perpendicular to $span\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{k-1}\}$, whereas the "error" component in the $span\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{k-1}\}$ will be amplified and the net result is $\boldsymbol{Q}$ will be less orthonormal. In this case, if we can find a successive set of orthogonal matrices $\{\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_l\}$ such that $\boldsymbol{Q}_l \ldots \boldsymbol{Q}_2\boldsymbol{Q}_1\boldsymbol{A}$ is triangularized, then $\boldsymbol{Q} = (\boldsymbol{Q}_l \ldots \boldsymbol{Q}_2\boldsymbol{Q}_1)^\top$ will be "more" orthogonal than the CGS or the MGS. We will discuss this method in Section 3.11 and 3.12 via Householder reflectors and Givens rotations.

3.6 Computing the Full QR Decomposition via the Gram-Schmidt Process

A full QR decomposition of an $m \times n$ matrix with linearly independent columns goes further by appending additional $m - n$ orthonormal columns to $\boldsymbol{Q}$ so that it becomes an $m \times m$ orthogonal matrix. In addition, rows of zeros are appended to $\boldsymbol{R}$ so that it becomes an $m \times n$ upper triangular matrix. We call the additional columns in $\boldsymbol{Q}$ **silent columns** and additional rows in $\boldsymbol{R}$ **silent rows**. The comparison between the reduced QR decomposition and the full QR decomposition is shown in Figure 8 where silent columns in $\boldsymbol{Q}$ are denoted in gray, blank entries are zero and blue entries are elements that are not necessarily zero.



(a) Reduced QR decomposition

(b) Full QR decomposition

Figure 8: Comparison between the reduced and full QR decompositions.

3.7 Dependent Columns

Previously, we assumed matrix $\boldsymbol{A}$ has linearly independent columns. However, this is not always necessary. Suppose in step k of CGS or MGS, $\boldsymbol{a}_k$ is in the plane spanned by $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{k-1}$ which is equivalent to the space spanned by $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{k-1}$, i.e., vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k$ are dependent. Then r_{kk} will be zero and $\boldsymbol{q}_k$ does not exist because of the zero division. At this moment, we simply pick $\boldsymbol{q}_k$ arbitrarily to be any normalized vector that is orthogonal to $\mathcal{C}([\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{k-1}])$ and continue the Gram-Schmidt process. Again, for matrix $\boldsymbol{A}$ with dependent columns, we have both reduced and full QR decomposition algorithms.

This idea can be further extended that when $\boldsymbol{q}_k$ does not exist, we just skip the current steps. And add the silent columns in the end. In this sense, QR decomposition for a matrix with dependent columns is not unique. However, as long as you stick to a systematic process, QR decomposition for any matrix is unique. This finding can also help to decide whether a set of vectors are linearly independent or not. Whenever r_{kk} in CGS or MGS is zero, we report the vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k$ are dependent and stop the algorithm for "independent checking".

3.8 QR with Column Pivoting: Column-Pivoted QR (CPQR)

Suppose $\boldsymbol{A}$ has dependent columns, a column-pivoted QR (CPQR) decomposition can be found as follows.

Theorem 3.2: (Column-Pivoted QR Decomposition)

Every $m \times n$ matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n]$ with $m \geq n$ and rank r can be factored as

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

where $\boldsymbol{R}_{11} \in \mathbb{R}^{r\times r}$ is upper triangular, $\boldsymbol{R}_{12} \in \mathbb{R}^{r\times(n-r)}$, $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ is an orthogonal matrix, and $\boldsymbol{P}$ is a permutation matrix. This is also known as the **full** CPQR decomposition. Similarly, the **reduced** version is given by

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{Q}_r \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \end{bmatrix},$$

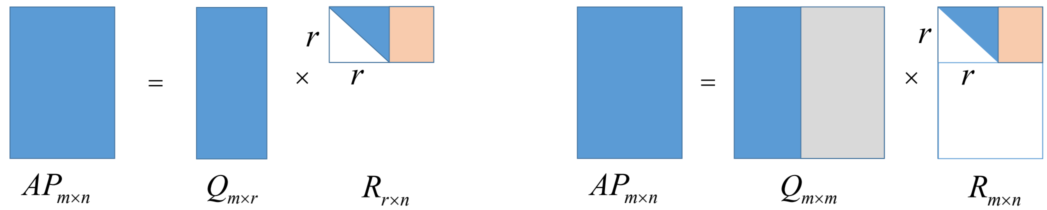
where $\boldsymbol{R}_{11} \in \mathbb{R}^{r\times r}$ is upper triangular, $\boldsymbol{R}_{12} \in \mathbb{R}^{r\times(n-r)}$, $\boldsymbol{Q}_r \in \mathbb{R}^{m\times r}$ contains orthonormal columns, and $\boldsymbol{P}$ is a permutation matrix.

3.8.1 A Simple CPQR via CGS

A Simple CPQR via CGS The classical Gram-Schmidt process can compute this CPQR decomposition. Following from the QR decomposition for dependent columns that when $r_{kk} = 0$, the column k of $\boldsymbol{A}$ is dependent on the previous $k-1$ columns. Whenever this happens, we permute this column into the last column and continue the Gram-Schmidt process. We notice that $\boldsymbol{P}$ is the permutation matrix that interchanges the dependent columns into the last $n-r$ columns. Suppose the first r columns of $\boldsymbol{A}\boldsymbol{P}$ are $[\widehat{\boldsymbol{a}}_1, \widehat{\boldsymbol{a}}_2, \ldots, \widehat{\boldsymbol{a}}_r]$, and the span of them is just the same as the span of $\boldsymbol{Q}_r$ (in the reduced version), or the span of $\boldsymbol{Q}_{:,:r}$ (in the full version)

$$\mathcal{C}([\widehat{\boldsymbol{a}}_1, \widehat{\boldsymbol{a}}_2, \ldots, \widehat{\boldsymbol{a}}_r]) = \mathcal{C}(\boldsymbol{Q}_r) = \mathcal{C}(\boldsymbol{Q}_{:,:r}).$$

And $\boldsymbol{R}_{12}$ is a matrix that recovers the dependent $n-r$ columns from the column space of $\boldsymbol{Q}_r$ or column space of $\boldsymbol{Q}_{:,:r}$. The comparison of reduced and full CPQR decomposition is shown in Figure 9 where silent columns in $\boldsymbol{Q}$ are denoted in grey, blank entries are zero and blue/orange entries are elements that are not necessarily zero.



(a) Reduced CPQR decomposition (b) Full CPQR decomposition

Figure 9: Comparison between the reduced and full CPQR decompositions.

3.8.2 A Practical CPQR via CGS

A Practical CPQR via CGS We notice that the simple CPQR algorithm pivot the first r independent columns into the first r columns of $\boldsymbol{AP}$. Let $\boldsymbol{A}_1$ be the first r columns of $\boldsymbol{AP}$, and $\boldsymbol{A}_2$ be the rest. Then, from the full CPQR, we have

$$[\boldsymbol{A}_1, \boldsymbol{A}_2] = \boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}_{11} & \boldsymbol{R}_{12}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} = \left[\boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}_{11}\\ \mathbf{0}\end{bmatrix}, \boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}_{12}\\ \mathbf{0}\end{bmatrix}\right].$$

It is not easy to see that

$$||\boldsymbol{A}_2|| = \left\|\boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}_{12}\\ \mathbf{0}\end{bmatrix}\right\| = \left\|\begin{bmatrix}\boldsymbol{R}_{12}\\ \mathbf{0}\end{bmatrix}\right\| = \|\boldsymbol{R}_{12}\|,$$

where the penultimate equality comes from the orthogonal equivalence under the matrix norm. Therefore, the norm of $\boldsymbol{R}_{12}$ is decided by the norm of $\boldsymbol{A}_2$. When favoring well-conditioned CPQR, $\boldsymbol{R}_{12}$ should be small in norm. And a practical CPQR decomposition is to permute columns of the matrix $\boldsymbol{A}$ firstly such that the columns are ordered decreasingly in vector norm:

$$\widetilde{\boldsymbol{A}} = \boldsymbol{AP}_0 = [\boldsymbol{a}_{j_1}, \boldsymbol{a}_{j_2}, \ldots, \boldsymbol{a}_{j_n}],$$

where $\{j_1, j_2, \ldots, j_n\}$ is a permuted index set of $\{1, 2, \ldots, n\}$ and

$$||\boldsymbol{a}_{j_1}|| \geq ||\boldsymbol{a}_{j_2}|| \geq \ldots \geq ||\boldsymbol{a}_{j_n}||.$$

Then apply the "simple" reduced CPQR decomposition on $\widetilde{\boldsymbol{A}}$ such that $\widetilde{\boldsymbol{A}}\boldsymbol{P}_1 = \boldsymbol{Q}_r[\boldsymbol{R}_{11}, \boldsymbol{R}_{12}]$. The "practical" reduced CPQR of $\boldsymbol{A}$ is then recovered as

$$\boldsymbol{A}\underbrace{\boldsymbol{P}_0\boldsymbol{P}_1}_{\boldsymbol{P}} = \boldsymbol{Q}_r[\boldsymbol{R}_{11}, \boldsymbol{R}_{12}].$$

The further optimization on the CPQR algorithm is via the MGS where the extra bonus is to stop at a point when the factorization works on a rank deficient submatrix and the CPQR via this MGS can find the numerical rank (Lu, 2021c). This is known as the *partial factorization* and we shall not give the details.

3.9 QR with Column Pivoting: Revealing Rank One Deficiency

We notice that column-pivoted QR is just one method to find the column permutation where $\boldsymbol{A}$ is rank deficient and we interchange the first linearly independent r columns of $\boldsymbol{A}$ into the first r columns of the $\boldsymbol{AP}$. If $\boldsymbol{A}$ is nearly rank-one deficient and we would like to find a column permutation of $\boldsymbol{A}$ such that the resulting pivotal element r_{nn} of the QR decomposition is small. This is known as the *revealing rank-one deficiency* problem.

Theorem 3.3: (Revealing Rank One Deficiency, (Chan, 1987))

If $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{v} \in \mathbb{R}^n$ is a unit 2-norm vector (i.e., $||\boldsymbol{v}|| = 1$), then there exists a permutation $\boldsymbol{P}$ such that the reduced QR decomposition

$$\boldsymbol{AP} = \boldsymbol{QR}$$

satisfies $r_{nn} \leq \sqrt{n}\epsilon$ where $\epsilon = ||\boldsymbol{Av}||$ and r_{nn} is the n-th diagonal of $\boldsymbol{R}$. Note that $\boldsymbol{Q} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ in the reduced QR decomposition.

Proof [of Theorem 3.3] Suppose $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is a permutation matrix such that if $\boldsymbol{w} = \boldsymbol{P}^\top \boldsymbol{v}$ where

$$|w_n| = \max |v_i|, \quad \forall i \in \{1, 2, \ldots, n\}.$$

That is, the last component of $\boldsymbol{w}$ is equal to the max component of $\boldsymbol{v}$ in absolute value. Then we have $|w_n| \geq 1/\sqrt{n}$. Suppose the QR decomposition of $\boldsymbol{AP}$ is $\boldsymbol{AP} = \boldsymbol{QR}$, then

$$\epsilon = ||\boldsymbol{Av}|| = ||(\boldsymbol{Q}^\top \boldsymbol{AP})(\boldsymbol{P}^\top \boldsymbol{v})|| = ||\boldsymbol{Rw}|| = \begin{bmatrix} \vdots \\ r_{nn}w_n \end{bmatrix} \geq |r_{nn}w_n| \geq |r_{nn}|/\sqrt{n},$$

This completes the proof. ■

The following discussion is based on the existence of the singular value decomposition (SVD) which will be introduced in Section 14 (p. 134). Feel free to skip at a first reading. Suppose the SVD of $\boldsymbol{A}$ is given by $\boldsymbol{A} = \sum_{i=1}^n \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where σ_i's are singular values with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$, i.e., σ_n is the smallest singular value, and $\boldsymbol{u}_i$'s, $\boldsymbol{v}_i$'s are the left and right singular vectors respectively. Then, if we let $\boldsymbol{v} = \boldsymbol{v}_n$ such that $\boldsymbol{Av}_n = \sigma_n \boldsymbol{u}_n$, [13] we have

$$||\boldsymbol{Av}|| = \sigma_n.$$

By constructing a permutation matrix $\boldsymbol{P}$ such that

$$|\boldsymbol{P}^\top \boldsymbol{v}|_n = \max |\boldsymbol{v}_i|, \quad \forall i \in \{1, 2, \ldots, n\},$$

we will find a QR decomposition of $\boldsymbol{A} = \boldsymbol{QR}$ with a pivot r_{nn} smaller than $\sqrt{n}\sigma_n$. If $\boldsymbol{A}$ is rank-one deficient, then σ_n will be close to 0 and r_{nn} is thus bounded to a small value in magnitude which is close to 0.

13. We will prove that the right singular vector of $\boldsymbol{A}$ is equal to the right singular vector of $\boldsymbol{R}$ if the $\boldsymbol{A}$ has QR decomposition $\boldsymbol{A} = \boldsymbol{QR}$ in Lemma 14.11 (p. 140). The claim can also be applied to the singular values. So $\boldsymbol{v}_n$ here is also the right singular vector of $\boldsymbol{R}$.

3.10 QR with Column Pivoting: Revealing Rank r Deficiency*

Following from the last section, suppose now we want to compute the reduced QR decomposition where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is nearly rank r deficient with $r > 1$. Our goal now is to find a permutation $\boldsymbol{P}$ such that

$$\boldsymbol{AP} = \boldsymbol{QR} = \boldsymbol{Q}\begin{bmatrix} \boldsymbol{L} & \boldsymbol{M} \\ \boldsymbol{0} & \boldsymbol{N} \end{bmatrix}, \tag{3.7}$$

where $\boldsymbol{N} \in \mathbb{R}^{r\times r}$ and $||\boldsymbol{N}||$ is small in some norm. A recursive algorithm can be applied to do so. Suppose we have already isolated a small $k \times k$ block $\boldsymbol{N}_k$, based on which, if we can isolate a small $(k+1)\times(k+1)$ block $\boldsymbol{N}_{k+1}$, then we can find the permutation matrix recursively. To repeat, suppose we have the permutation $\boldsymbol{P}_k$ such that the $\boldsymbol{N}_k \in \mathbb{R}^{k\times k}$ has a small norm,

$$\boldsymbol{AP}_k = \boldsymbol{Q}_k\boldsymbol{R}_k = \boldsymbol{Q}_k\begin{bmatrix} \boldsymbol{L}_k & \boldsymbol{M}_k \\ \boldsymbol{0} & \boldsymbol{N}_k \end{bmatrix}.$$

We want to find a permutation $\boldsymbol{P}_{k+1}$, such that $\boldsymbol{N}_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$ also has a small norm,

$$\boxed{\boldsymbol{AP}_{k+1} = \boldsymbol{Q}_{k+1}\boldsymbol{R}_{k+1} = \boldsymbol{Q}_{k+1}\begin{bmatrix} \boldsymbol{L}_{k+1} & \boldsymbol{M}_{k+1} \\ \boldsymbol{0} & \boldsymbol{N}_{k+1} \end{bmatrix}}.$$

From the algorithm introduced in the last section, there is an $(n-k)\times(n-k)$ permutation matrix $\widetilde{\boldsymbol{P}}_{k+1}$ such that $\boldsymbol{L}_k \in \mathbb{R}^{(n-k)\times(n-k)}$ has the QR decomposition $\boldsymbol{L}_k\widetilde{\boldsymbol{P}}_{k+1} = \widetilde{\boldsymbol{Q}}_{k+1}\widetilde{\boldsymbol{L}}_k$ such that the entry $(n-k, n-k)$ of $\widetilde{\boldsymbol{L}}_k$ is small. By constructing

$$\boldsymbol{P}_{k+1} = \boldsymbol{P}_k\begin{bmatrix} \widetilde{\boldsymbol{P}}_{k+1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}, \qquad \boldsymbol{Q}_{k+1} = \boldsymbol{Q}_k\begin{bmatrix} \widetilde{\boldsymbol{Q}}_{k+1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix},$$

we have

$$\boxed{\boldsymbol{AP}_{k+1} = \boldsymbol{Q}_{k+1}\begin{bmatrix} \widetilde{\boldsymbol{L}}_k & \widetilde{\boldsymbol{Q}}_{k+1}^\top\boldsymbol{M}_k \\ \boldsymbol{0} & \boldsymbol{N}_k \end{bmatrix}}.$$

We know that entry $(n-k, n-k)$ of $\widetilde{\boldsymbol{L}}_k$ is small, if we can prove the last row of $\widetilde{\boldsymbol{Q}}_{k+1}^\top\boldsymbol{M}_k$ is small in norm, then we find the QR decomposition revealing rank $k+1$ deficiency (see (Chan, 1987) for a proof).

3.11 Existence of the QR Decomposition via the Householder Reflector

We first give the formal definition of a Householder reflector and we will take a look at its properties.

Definition 3.4: (Householder Reflector)

Let $\boldsymbol{u} \in \mathbb{R}^n$ be a vector of unit length (i.e., $||\boldsymbol{u}|| = 1$). Then $\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^\top$ is said to be a *Householder reflector*, a.k.a., a *Householder transformation*. We call this $\boldsymbol{H}$ the Householder reflector associated with the unit vector $\boldsymbol{u}$ where the unit vector $\boldsymbol{u}$ is also known as *Householder vector*. If a vector $\boldsymbol{x}$ is multiplied by $\boldsymbol{H}$, then it is reflected in the hyperplane $span\{\boldsymbol{u}\}^\perp$.

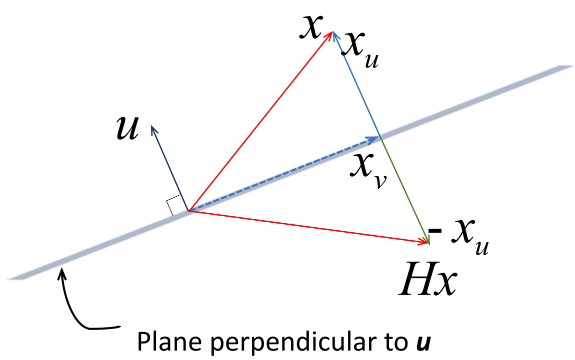


Figure 10: Demonstration of the Householder reflector. The Householder reflector obtained by $\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^\top$ where $||\boldsymbol{u}|| = 1$ will reflect vector $\boldsymbol{x}$ along the plane perpendicular to $\boldsymbol{u}$: $\boldsymbol{x} = \boldsymbol{x_v} + \boldsymbol{x_u} \rightarrow \boldsymbol{x_v} - \boldsymbol{x_u}$.

Note that if $||\boldsymbol{u}|| \neq 1$, we can define $\boldsymbol{H} = \boldsymbol{I} - 2\frac{\boldsymbol{u}\boldsymbol{u}^\top}{\boldsymbol{u}^\top\boldsymbol{u}}$ as the Householder reflector.

Then we have the following corollary from this definition.

Corollary 3.5: (Unreflected by Householder)

Any vector $\boldsymbol{v}$ that is perpendicular to $\boldsymbol{u}$ is left unchanged by the Householder transformation, that is, $\boldsymbol{H}\boldsymbol{v} = \boldsymbol{v}$ if $\boldsymbol{u}^\top\boldsymbol{v} = 0$.

The proof is trivial that $(\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{v} = \boldsymbol{v} - 2\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{v} = \boldsymbol{v}$.

Suppose $\boldsymbol{u}$ is a unit vector with $||\boldsymbol{u}|| = 1$, and a vector $\boldsymbol{v}$ is perpendicular to $\boldsymbol{u}$. Then any vector $\boldsymbol{x}$ on the plane can be decomposed into two parts $\boldsymbol{x} = \boldsymbol{x_v} + \boldsymbol{x_u}$: the first one $\boldsymbol{x_u}$ is parallel to $\boldsymbol{u}$ and the second one $\boldsymbol{x_v}$ is perpendicular to $\boldsymbol{u}$ (i.e., parallel to $\boldsymbol{v}$). From Section 3.1 on the projection of a vector onto another one, $\boldsymbol{x_u}$ can be computed by $\boldsymbol{x_u} = \frac{\boldsymbol{u}\boldsymbol{u}^\top}{\boldsymbol{u}^\top\boldsymbol{u}}\boldsymbol{x} = \boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{x}$. We then transform this $\boldsymbol{x}$ by the Householder reflector associated with $\boldsymbol{u}$, $\boldsymbol{H}\boldsymbol{x} = (\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^\top)(\boldsymbol{x_v} + \boldsymbol{x_u}) = \boldsymbol{x_v} - \boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{x} = \boldsymbol{x_v} - \boldsymbol{x_u}$. That is, the space perpendicular to $\boldsymbol{u}$ acts as a mirror and any vector $\boldsymbol{x}$ is reflected by the Householder reflector associated with $\boldsymbol{u}$ (i.e., reflected in the hyperplane $span\{\boldsymbol{u}\}^\perp$). The situation is shown in Figure 10.

If we know two vectors are reflected to each other, the next corollary tells us how to find the corresponding Householder reflector.

Corollary 3.6: (Finding the Householder Reflector)

Suppose $\boldsymbol{x}$ is reflected to $\boldsymbol{y}$ by a Householder reflector with $||\boldsymbol{x}|| = ||\boldsymbol{y}||$, then the Householder reflector is obtained by

$$\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^\top, \text{ where } \boldsymbol{u} = \frac{\boldsymbol{x} - \boldsymbol{y}}{||\boldsymbol{x} - \boldsymbol{y}||}.$$

Proof [of Corollary 3.6] Write out the equation, we have

$$\begin{aligned}
\boldsymbol{H}\boldsymbol{x} = \boldsymbol{x} - 2\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{x} &= \boldsymbol{x} - 2\frac{(\boldsymbol{x} - \boldsymbol{y})(\boldsymbol{x}^\top - \boldsymbol{y}^\top)}{(\boldsymbol{x} - \boldsymbol{y})^\top(\boldsymbol{x} - \boldsymbol{y})}\boldsymbol{x} \\
&= \boldsymbol{x} - (\boldsymbol{x} - \boldsymbol{y}) = \boldsymbol{y}.
\end{aligned}$$

Note that the condition $||\boldsymbol{x}|| = ||\boldsymbol{y}||$ is required to prove the result. ■

The Householder reflectors are useful to set a block of components of a given vector to zero. Particularly, we usually would like to set the vector $\boldsymbol{a} \in \mathbb{R}^n$ to be zero except the i-th element. Then the Householder vector can be chosen to be

$$\boldsymbol{u} = \frac{\boldsymbol{a} - r\boldsymbol{e}_i}{||\boldsymbol{a} - r\boldsymbol{e}_i||}, \qquad \text{where } r = \pm||\boldsymbol{a}||$$

which is a reasonable Householder vector since $||\boldsymbol{a}|| = ||r\boldsymbol{e}_i|| = |r|$. We carefully notice that when $r = ||\boldsymbol{a}||$, $\boldsymbol{a}$ is reflected to $||\boldsymbol{a}||\boldsymbol{e}_i$ via the Householder reflector $\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^\top$; otherwise when $r = -||\boldsymbol{a}||$, $||\boldsymbol{a}||$ is reflected to $-||\boldsymbol{a}||\boldsymbol{e}_i$ via the Householder reflector.

Remark 3.7: (Householder Properties)

If $\boldsymbol{H}$ is a Householder reflector, then it has the following properties:

- $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{I}$;
- $\boldsymbol{H} = \boldsymbol{H}^\top$;
- $\boldsymbol{H}^\top\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}^\top = \boldsymbol{I}$ such that Householder reflector is an orthogonal matrix;
- $\boldsymbol{H}\boldsymbol{u} = -\boldsymbol{u}$, if $\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^\top$.

We see in the Gram-Schmidt section that QR decomposition is to use a triangular matrix to orthogonalize a matrix $\boldsymbol{A}$. The further idea is that, if we have a set of orthogonal matrices that can make $\boldsymbol{A}$ to be triangular step by step, then we can also recover the QR decomposition. Specifically, if we have an orthogonal matrix $\boldsymbol{Q}_1$ that can introduce zeros to the 1-st column of $\boldsymbol{A}$ except the entry (1,1); and an orthogonal matrix $\boldsymbol{Q}_2$ that can introduce zeros to the 2-nd column except the entries (2,1), (2,2); .... Then, we can also find the QR decomposition. For the way to introduce zeros, we could reflect the columns of the matrix to a basis vector $\boldsymbol{e}_1$ whose entries are all zero except the first entry.

Let $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}$ be the column partition of $\boldsymbol{A}$, and let

$$r_1 = ||\boldsymbol{a}_1||, \qquad \boldsymbol{u}_1 = \frac{\boldsymbol{a}_1 - r_1\boldsymbol{e}_1}{||\boldsymbol{a}_1 - r_1\boldsymbol{e}_1||}, \qquad \text{and} \qquad \boldsymbol{H}_1 = \boldsymbol{I} - 2\boldsymbol{u}_1\boldsymbol{u}_1^\top, \tag{3.8}$$

where $\boldsymbol{e}_1$ here is the first basis for $\mathbb{R}^m$, i.e., $\boldsymbol{e}_1 = [1; 0; 0; \ldots; 0] \in \mathbb{R}^m$. Then

$$\boldsymbol{H}_1\boldsymbol{A} = [\boldsymbol{H}_1\boldsymbol{a}_1, \boldsymbol{H}_1\boldsymbol{a}_2, \ldots, \boldsymbol{H}_1\boldsymbol{a}_n] = \begin{bmatrix} r_1 & \boldsymbol{R}_{1,2:n} \\ \boldsymbol{0} & \boldsymbol{B}_2 \end{bmatrix}, \tag{3.9}$$

which reflects $\boldsymbol{a}_1$ to $r_1\boldsymbol{e}_1$ and introduces zeros below the diagonal in the 1-st column. We observe that the entries below r_1 are all zero now under this specific reflection. Notice that we reflect $\boldsymbol{a}_1$ to $||\boldsymbol{a}_1||\boldsymbol{e}_1$ the two of which have same length, rather than reflect $\boldsymbol{a}_1$ to $\boldsymbol{e}_1$ directly. This is for the purpose of **numerical stability**.

Choice of r_1: moreover, the choice of r_1 is **not unique**. For **numerical stability**, it is also desirable to choose $r_1 = -\text{sign}(a_{11})||\boldsymbol{a}_1||$, where a_{11} is the first component of $\boldsymbol{a}_1$. Or even, $r_1 = \text{sign}(a_{11})||\boldsymbol{a}_1||$ is also possible as long as $||\boldsymbol{a}_1||$ is equal to $||r_1\boldsymbol{e}_1||$. However, we will not cover this topic here.

We can then apply this process to $\boldsymbol{B}_2$ in Equation (3.9) to make the entries below the entry (2,2) to be all zeros. Note that, we do not apply this process to the entire $\boldsymbol{H}_1\boldsymbol{A}$ but rather the submatrix $\boldsymbol{B}_2$ in it because we have already introduced zeros in the first column, and reflecting again will introduce nonzero values back.

Suppose $\boldsymbol{B}_2 = [\boldsymbol{b}_2, \boldsymbol{b}_3, \ldots, \boldsymbol{b}_n]$ is the column partition of $\boldsymbol{B}_2$, and let

$$r_2 = ||\boldsymbol{b}_2||, \qquad \boldsymbol{u}_2 = \frac{\boldsymbol{b}_2 - r_2\boldsymbol{e}_1}{||\boldsymbol{b}_2 - r_2\boldsymbol{e}_1||}, \qquad \widetilde{\boldsymbol{H}}_2 = \boldsymbol{I} - 2\boldsymbol{u}_2\boldsymbol{u}_2^\top, \qquad \text{and} \qquad \boldsymbol{H}_2 = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{H}}_2 \end{bmatrix},$$

where now $\boldsymbol{e}_1$ here is the first basis for $\mathbb{R}^{m-1}$ and $\boldsymbol{H}_2$ is also an orthogonal matrix since $\widetilde{\boldsymbol{H}}_2$ is an orthogonal matrix. Then it follows that

$$\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A} = [\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{a}_1, \boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{a}_2, \ldots, \boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{a}_n] = \begin{bmatrix} r_1 & \boldsymbol{R}_{12} & \boldsymbol{R}_{1,3:n} \\ 0 & r_2 & \boldsymbol{R}_{2,3:n} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{C}_3 \end{bmatrix}.$$

The same process can go on, and we will finally triangularize $\boldsymbol{A} = (\boldsymbol{H}_n\boldsymbol{H}_{n-1}\ldots\boldsymbol{H}_1)^{-1}\boldsymbol{R} = \boldsymbol{Q}\boldsymbol{R}$. And since the $\boldsymbol{H}_i$'s are symmetric and orthogonal, we have $\boldsymbol{Q} = (\boldsymbol{H}_n\boldsymbol{H}_{n-1}\ldots\boldsymbol{H}_1)^{-1} = \boldsymbol{H}_1\boldsymbol{H}_2\ldots\boldsymbol{H}_n$.

An example of a 5×4 matrix is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

$$\underset{\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}} \xrightarrow{\boldsymbol{H}_1} \underset{\boldsymbol{H}_1\boldsymbol{A}}{\begin{bmatrix} \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}} \xrightarrow{\boldsymbol{H}_2} \underset{\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}}$$

$$\xrightarrow{\boldsymbol{H}_3} \underset{\boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \boxtimes & \boxtimes & \boxtimes \\ 0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ 0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} \\ 0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} \end{bmatrix}} \xrightarrow{\boldsymbol{H}_4} \underset{\boldsymbol{H}_4\boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \boxtimes & \boxtimes & \boxtimes \\ 0 & 0 & \boxtimes & \boxtimes \\ 0 & 0 & 0 & \boldsymbol{\boxtimes} \\ 0 & 0 & 0 & \mathbf{0} \end{bmatrix}}$$

A closer look at the QR factorization The Householder algorithm is a process that makes a matrix triangular by a sequence of orthogonal matrix operations. In the Gram-Schmidt process (both CGS and MGS), we use a triangular matrix to orthogonalize the matrix. However, in the Householder algorithm, we use orthogonal matrices to triangularize. The difference between the two approaches is summarized as follows:

- Gram-Schmidt: triangular orthogonalization;
- Householder: orthogonal triangularization.

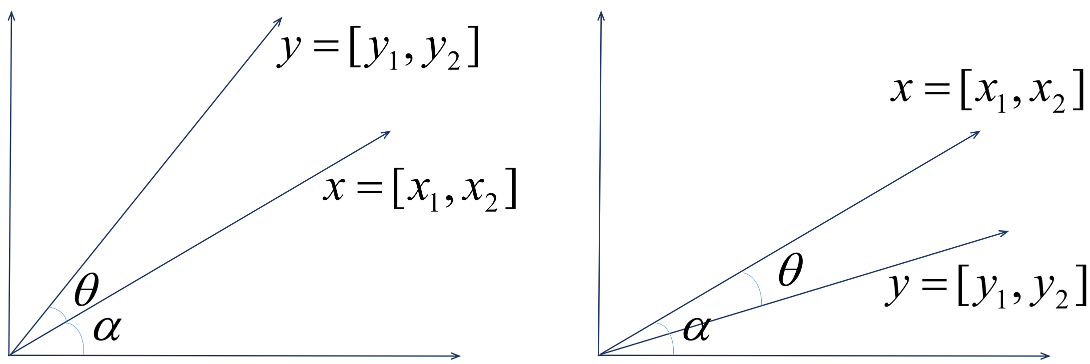
We further notice that, in the Householder algorithm or the Givens algorithm that we will shortly see, a set of orthogonal matrices are applied so that the QR decomposition obtained is a *full* QR decomposition. Whereas, the direct QR decomposition obtained by CGS or MGS is a *reduced* one (although the silent columns or rows can be further added to find the full one).

3.12 Existence of the QR Decomposition via the Givens Rotation

We have defined the Givens rotation in Definition 2.11 (p. 41) to find the rank-one update/downdate of the Cholesky decomposition. Now consider the following 2×2 orthogonal matrices

$$\boldsymbol{F}=\begin{bmatrix}-c & s\\ s & c\end{bmatrix}, \qquad \boldsymbol{J}=\begin{bmatrix}c & -s\\ s & c\end{bmatrix}, \qquad \boldsymbol{G}=\begin{bmatrix}c & s\\ -s & c\end{bmatrix},$$

where $s=\sin\theta$ and $c=\cos\theta$ for some θ. The first matrix has $\det(\boldsymbol{F})=-1$ and is a special case of a Householder reflector in dimension 2 such that $\boldsymbol{F}=\boldsymbol{I}-2\boldsymbol{u}\boldsymbol{u}^\top$ where $\boldsymbol{u}=\left[\sqrt{\frac{1+c}{2}},\ \sqrt{\frac{1-c}{2}}\right]^\top$ or $\boldsymbol{u}=\left[-\sqrt{\frac{1+c}{2}},\ -\sqrt{\frac{1-c}{2}}\right]^\top$. The latter two matrices have $\det(\boldsymbol{J})=\det(\boldsymbol{G})=1$ and effects a rotation instead of a reflection. Such a matrix is called a ***Givens rotation***.



(a) $\boldsymbol{y}=\boldsymbol{J}\boldsymbol{x}$, counter-clockwise rotation. (b) $\boldsymbol{y}=\boldsymbol{G}\boldsymbol{x}$, clockwise rotation.

Figure 11: Demonstration of two Givens rotations.

Figure 11 demonstrate a rotation of $\boldsymbol{x}$ under $\boldsymbol{J}$, where $\boldsymbol{y}=\boldsymbol{J}\boldsymbol{x}$ such that

$$\begin{cases}y_1=c\cdot x_1-s\cdot x_2,\\ y_2=s\cdot x_1+c\cdot x_2.\end{cases}$$

We want to verify the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$ is actually θ (and counter-clockwise rotation) after the Givens rotation $\boldsymbol{J}$ as shown in Figure 11(a). Firstly, we have

$$\begin{cases}\cos(\alpha)=\dfrac{x_1}{\sqrt{x_1^2+x_2^2}},\\ \sin(\alpha)=\dfrac{x_2}{\sqrt{x_1^2+x_2^2}}.\end{cases} \qquad \text{and} \qquad \begin{cases}\cos(\theta)=c,\\ \sin(\theta)=s.\end{cases}$$

This implies $\cos(\theta+\alpha)=\cos(\theta)\cos(\alpha)-\sin(\theta)\sin(\alpha)$. If we can show $\cos(\theta+\alpha)=\cos(\theta)\cos(\alpha)-\sin(\theta)\sin(\alpha)$ is equal to $\frac{y_1}{\sqrt{y_1^2+y_2^2}}$, then we complete the proof.

For the former one, $\cos(\theta+\alpha)=\cos(\theta)\cos(\alpha)-\sin(\theta)\sin(\alpha)=\frac{c\cdot x_1-s\cdot x_2}{\sqrt{x_1^2+x_2^2}}$. For the latter one, it can be verified that $\sqrt{y_1^2+y_2^2}=\sqrt{x_1^2+x_2^2}$, and $\frac{y_1}{\sqrt{y_1^2+y_2^2}}=\frac{c\cdot x_1-s\cdot x_2}{\sqrt{x_1^2+x_2^2}}$. This completes the proof. Similarly, we can also show that the angle between $\boldsymbol{y}=\boldsymbol{G}\boldsymbol{x}$ and $\boldsymbol{x}$ is also θ in Figure 11(b) and the rotation is clockwise.

It can be easily verified that the n-th order Givens rotation (Definition 2.11, p. 41) is an orthogonal matrix and its determinant is 1. For any vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$, we have $\boldsymbol{y} = \boldsymbol{G}_{kl}\boldsymbol{x}$, where

$$\begin{cases} y_k = c \cdot x_k + s \cdot x_l, \\ y_l = -s \cdot x_k + c \cdot x_l, \\ y_j = x_j, & (j \neq k, l) \end{cases}$$

That is, a Givens rotation applied to $\boldsymbol{x}$ rotates two components of $\boldsymbol{x}$ by some angle θ and leaves all other components the same. When $\sqrt{x_k^2 + x_l^2} \neq 0$, let $c = \frac{x_k}{\sqrt{x_k^2+x_l^2}}$, $s = \frac{x_l}{\sqrt{x_k^2+x_l^2}}$. Then,

$$\begin{cases} y_k = \sqrt{x_k^2 + x_l^2}, \\ y_l = 0, \\ y_j = x_j. & (j \neq k, l) \end{cases}$$

This finding above is essential for the QR decomposition via the Givens rotation.

Corollary 3.8: (Basis From Givens Rotations Forwards)

For any vector $\boldsymbol{x} \in \mathbb{R}^n$, there exists a set of Givens rotations $\{\boldsymbol{G}_{12}, \boldsymbol{G}_{13}, \ldots, \boldsymbol{G}_{1n}\}$ such that $\boldsymbol{G}_{1n} \ldots \boldsymbol{G}_{13}\boldsymbol{G}_{12}\boldsymbol{x} = ||\boldsymbol{x}||\boldsymbol{e}_1$ where $\boldsymbol{e}_1 \in \mathbb{R}^n$ is the first unit basis in $\mathbb{R}^n$.

Proof [of Corollary 3.8] From the finding above, we can find a $\boldsymbol{G}_{12}, \boldsymbol{G}_{13}, \boldsymbol{G}_{14}$ such that

$$\boldsymbol{G}_{12}\boldsymbol{x} = \left[\sqrt{x_1^2 + x_2^2}, 0, x_3, \ldots, x_n\right]^\top,$$

$$\boldsymbol{G}_{13}\boldsymbol{G}_{12}\boldsymbol{x} = \left[\sqrt{x_1^2 + x_2^2 + x_3^2}, 0, 0, x_4, \ldots, x_n\right]^\top,$$

and

$$\boldsymbol{G}_{14}\boldsymbol{G}_{13}\boldsymbol{G}_{12}\boldsymbol{x} = \left[\sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2}, 0, 0, 0, x_5, \ldots, x_n\right]^\top.$$

Continue this process, we will obtain $\boldsymbol{G}_{1n} \ldots \boldsymbol{G}_{13}\boldsymbol{G}_{12} = ||\boldsymbol{x}||\boldsymbol{e}_1$. ■

Remark 3.9: (Basis From Givens Rotations Backwards)

In Corollary 3.8, we find the Givens rotation that introduces zeros from the 2-nd entry to th n-th entry (i.e., forward). Sometimes we want the reverse order, i.e., introduce zeros from the n-th entry to the 2-nd entrysuch that $\boldsymbol{G}_{12}\boldsymbol{G}_{13} \ldots \boldsymbol{G}_{1n}\boldsymbol{x} = ||\boldsymbol{x}||\boldsymbol{e}_1$ where $\boldsymbol{e}_1 \in \mathbb{R}^n$ is the first unit basis in $\mathbb{R}^n$.

The procedure is similar, we can find a $\boldsymbol{G}_{1n}, \boldsymbol{G}_{1,(n-1)}, \boldsymbol{G}_{1,(n-2)}$ such that

$$\boldsymbol{G}_{1n}\boldsymbol{x} = \left[\sqrt{x_1^2 + x_n^2}, x_2, x_3, \ldots, x_{n-1}, 0\right]^\top,$$

$$\boldsymbol{G}_{1,(n-1)}\boldsymbol{G}_{1n}\boldsymbol{x} = \left[\sqrt{x_1^2+x_{n-1}^2+x_n^2}, x_2, x_3, \ldots, x_{n-2}, 0, 0\right]^\top,$$

and

$$\boldsymbol{G}_{1,(n-2)}\boldsymbol{G}_{1,(n-1)}\boldsymbol{G}_{1n}\boldsymbol{x} = \left[\sqrt{x_1^2+x_{n-2}^2+x_{n-1}^2+x_n^2}, x_2, x_3, \ldots, x_{n-3}, 0, 0, 0\right]^\top.$$

Continue this process, we will obtain $\boldsymbol{G}_{12}\boldsymbol{G}_{13}\ldots\boldsymbol{G}_{1n}\boldsymbol{x} = ||\boldsymbol{x}||\boldsymbol{e}_1$.

An alternative form Alternatively, there are rotations $\{\boldsymbol{G}_{12}, \boldsymbol{G}_{23}, \ldots, \boldsymbol{G}_{(n-1),n}\}$ such that $\boldsymbol{G}_{12}\boldsymbol{G}_{23}\ldots\boldsymbol{G}_{(n-1),n}\boldsymbol{x} = ||\boldsymbol{x}||\boldsymbol{e}_1$ where $\boldsymbol{e}_1 \in \mathbb{R}^n$ is the first unit basis in $\mathbb{R}^n$ with

$$\boldsymbol{G}_{(n-1),n}\boldsymbol{x} = \left[x_1, x_2, \ldots, x_{n-2}, \sqrt{x_{n-1}^2+x_n^2}, 0\right]^\top,$$

$$\boldsymbol{G}_{(n-2),(n-1)}\boldsymbol{G}_{(n-1),n}\boldsymbol{x} = \left[x_1, x_2, \ldots, x_{n-3}, \sqrt{x_{n-2}^2+x_{n-1}^2+x_n^2}, 0, 0\right]^\top,$$

and

$$\boldsymbol{G}_{(n-3),(n-2)}\boldsymbol{G}_{(n-2),(n-1)}\boldsymbol{G}_{(n-1),n}\boldsymbol{x} = \left[x_1, x_2, \ldots, x_{n-4}, \sqrt{x_{n-3}^2+x_{n-2}^2+x_{n-1}^2+x_n^2}, 0, 0, 0\right]^\top.$$

Continue this process, we will obtain $\boldsymbol{G}_{12}\boldsymbol{G}_{23}\ldots\boldsymbol{G}_{(n-1),n}\boldsymbol{x} = ||\boldsymbol{x}||\boldsymbol{e}_1$.

From the Corollary 3.8 above, for the way to introduce zeros, we could **rotate** the columns of the matrix to a basis vector $\boldsymbol{e}_1$ whose entries are all zero except the first entry. Let $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}$ be the column partition of $\boldsymbol{A}$, and let

$$\boldsymbol{G}_1 = \boldsymbol{G}_{1m}\ldots\boldsymbol{G}_{13}\boldsymbol{G}_{12}, \tag{3.10}$$

where $\boldsymbol{e}_1$ here is the first basis for $\mathbb{R}^m$, i.e., $\boldsymbol{e}_1 = [1; 0; 0; \ldots; 0] \in \mathbb{R}^m$. Then

$$\boldsymbol{G}_1\boldsymbol{A} = [\boldsymbol{G}_1\boldsymbol{a}_1, \boldsymbol{G}_1\boldsymbol{a}_2, \ldots, \boldsymbol{G}_1\boldsymbol{a}_n] = \begin{bmatrix} ||\boldsymbol{a}_1|| & \boldsymbol{R}_{1,2:n} \\ \boldsymbol{0} & \boldsymbol{B}_2 \end{bmatrix}, \tag{3.11}$$

which rotates $\boldsymbol{a}_1$ to $||\boldsymbol{a}_1||\boldsymbol{e}_1$ and introduces zeros below the diagonal in the 1-st column.

We can then apply this process to $\boldsymbol{B}_2$ in Equation (3.11) to make the entries below the (2,2)-th entry to be all zeros. Suppose $\boldsymbol{B}_2 = [\boldsymbol{b}_2, \boldsymbol{b}_3, \ldots, \boldsymbol{b}_n]$, and let

$$\boldsymbol{G}_2 = \boldsymbol{G}_{2m}\ldots\boldsymbol{G}_{24}\boldsymbol{G}_{23},$$

where $\boldsymbol{G}_{2n}, \ldots, \boldsymbol{G}_{24}, \boldsymbol{G}_{23}$ can be implied from context. Then

$$\boldsymbol{G}_2\boldsymbol{G}_1\boldsymbol{A} = [\boldsymbol{G}_2\boldsymbol{G}_1\boldsymbol{a}_1, \boldsymbol{G}_2\boldsymbol{G}_1\boldsymbol{a}_2, \ldots, \boldsymbol{G}_2\boldsymbol{G}_1\boldsymbol{a}_n] = \begin{bmatrix} ||\boldsymbol{a}_1|| & \boldsymbol{R}_{12} & \boldsymbol{R}_{1,3:n} \\ 0 & ||\boldsymbol{b}_2|| & \boldsymbol{R}_{2,3:n} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{C}_3 \end{bmatrix}.$$

The same process can go on, and we will finally triangularize $\boldsymbol{A} = (\boldsymbol{G}_n\boldsymbol{G}_{n-1}\ldots\boldsymbol{G}_1)^{-1}\boldsymbol{R} = \boldsymbol{QR}$. And since $\boldsymbol{G}_i$'s are orthogonal, we have $\boldsymbol{Q} = (\boldsymbol{G}_n\boldsymbol{G}_{n-1}\ldots\boldsymbol{G}_1)^{-1} = \boldsymbol{G}_1^\top\boldsymbol{G}_2^\top\ldots\boldsymbol{G}_n^\top$, and

$$
\begin{aligned}
\boldsymbol{G}_1^\top\boldsymbol{G}_2^\top\ldots\boldsymbol{G}_n^\top &= (\boldsymbol{G}_n\ldots\boldsymbol{G}_2\boldsymbol{G}_1)^\top \\
&= \left\{(\boldsymbol{G}_{nm}\ldots\boldsymbol{G}_{n,(n+1)})\ldots(\boldsymbol{G}_{2m}\ldots\boldsymbol{G}_{23})(\boldsymbol{G}_{1m}\ldots\boldsymbol{G}_{12})\right\}^\top.
\end{aligned}
\tag{3.12}
$$

The Givens rotation algorithm works better when $\boldsymbol{A}$ already has a lot of zeros below the main diagonal. An example of a 5×4 matrix is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

Givens rotations in $\boldsymbol{G}_1$ For a 5×4 example, we realize that $\boldsymbol{G}_1 = \boldsymbol{G}_{15}\boldsymbol{G}_{14}\boldsymbol{G}_{13}\boldsymbol{G}_{12}$. And the process is shown as follows:

$$
\underset{\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\xrightarrow{\boldsymbol{G}_{12}}
\underset{\boldsymbol{G}_{12}\boldsymbol{A}}{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\xrightarrow{\boldsymbol{G}_{13}}
\underset{\boldsymbol{G}_{13}\boldsymbol{G}_{12}\boldsymbol{A}}{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boxtimes & \boxtimes & \boxtimes \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
$$

$$
\xrightarrow{\boldsymbol{G}_{14}}
\underset{\boldsymbol{G}_{14}\boldsymbol{G}_{13}\boldsymbol{G}_{12}\boldsymbol{A}}{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\xrightarrow{\boldsymbol{G}_{15}}
\underset{\boldsymbol{G}_{15}\boldsymbol{G}_{14}\boldsymbol{G}_{13}\boldsymbol{G}_{12}\boldsymbol{A}}{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
$$

Givens rotation as a big picture Take $\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3, \boldsymbol{G}_4$ as a single matrix, we have

$$
\underset{\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\xrightarrow{\boldsymbol{G}_1}
\underset{\boldsymbol{G}_1\boldsymbol{A}}{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\xrightarrow{\boldsymbol{G}_2}
\underset{\boldsymbol{G}_2\boldsymbol{G}_1\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
$$

$$
\xrightarrow{\boldsymbol{G}_3}
\underset{\boldsymbol{G}_3\boldsymbol{G}_2\boldsymbol{G}_1\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\xrightarrow{\boldsymbol{G}_4}
\underset{\boldsymbol{G}_4\boldsymbol{G}_3\boldsymbol{G}_2\boldsymbol{G}_1\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \mathbf{0}
\end{bmatrix}}
$$

Orders to introduce the zeros With the Givens rotations for the QR decomposition, it is flexible to choose different orders to introduce the zeros of $\boldsymbol{R}$. In our case, we introduce zeros column by column. It is also possible to introduce zeros row by row.

3.13 Uniqueness of the QR Decomposition

The results of the QR decomposition from the Gram-Schmidt process , the Householder algorithm, and the Givens algorithms are different. Even in the Householder algorithm, we have different methods to choose the sign of r_1 in Equation (3.8). Thus, from this point, QR decomposition is not unique.

However, if we use just the procedure described in the Gram-Schmidt process, or systematically choose the sign in the Householder algorithm, then the decomposition is unique. The uniqueness of the *reduced* QR decomposition for full column rank matrix $\boldsymbol{A}$ is assured when $\boldsymbol{R}$ has positive diagonals by inductive analysis (Lu, 2021c). We here provide another proof for the uniqueness of the *reduced* QR decomposition for matrices if the diagonal values of $\boldsymbol{R}$ are positive which will shed light on the implicit Q theorem in Hessenberg decomposition (Section 8.3, p. 94) or tridiagonal decomposition (Theorem 9.1, p. 97).

Corollary 3.10: (Uniqueness of the reduced QR Decomposition)

Suppose matrix $\boldsymbol{A}$ is an $m \times n$ matrix with full column rank n and $m \geq n$. Then, the *reduced* QR decomposition is unique if the main diagonal values of $\boldsymbol{R}$ are positive.

Proof [of Corollary 3.10] Suppose the *reduced* QR decomposition is not unique, we can complete it into a *full* QR decomposition, then we can find two such full decompositions so that $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1 = \boldsymbol{Q}_2\boldsymbol{R}_2$ which implies $\boldsymbol{R}_1 = \boldsymbol{Q}_1^{-1}\boldsymbol{Q}_2\boldsymbol{R}_2 = \boldsymbol{V}\boldsymbol{R}_2$ where $\boldsymbol{V} = \boldsymbol{Q}_1^{-1}\boldsymbol{Q}_2$ is an orthogonal matrix. Write out the equation, we have

$$
\boldsymbol{R}_1 = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ & r_{22} & \ldots & r_{2n} \\ \Large 0 & & \ddots & \vdots \\ & & & r_{nn} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1m} \\ v_{21} & v_{22} & \ldots & v_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m1} & v_{m2} & \ldots & v_{mm} \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1n} \\ & s_{22} & \ldots & s_{2n} \\ \Large 0 & & \ddots & \vdots \\ & & & s_{nn} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \end{bmatrix} = \boldsymbol{V}\boldsymbol{R}_2,
$$

This implies

$$
r_{11} = v_{11}s_{11}, \qquad v_{21} = v_{31} = v_{41} = \ldots = v_{m1} = 0.
$$

Since $\boldsymbol{V}$ contains mutually orthonormal columns and the first column of $\boldsymbol{V}$ is of norm 1. Thus, $v_{11} = \pm 1$. We notice that $r_{ii} > 0$ and $s_{ii} > 0$ for $i \in \{1, 2, \ldots, n\}$ by assumption such that $r_{11} > 0$ and $s_{11} > 0$ and v_{11} can only be positive 1. Since $\boldsymbol{V}$ is an orthogonal matrix, we also have

$$
v_{12} = v_{13} = v_{14} = \ldots = v_{1m} = 0.
$$

Applying this process to the submatrices of $\boldsymbol{R}_1, \boldsymbol{V}, \boldsymbol{R}_2$, we will find the upper-left submatrix of $\boldsymbol{V}$ is an identity: $\boldsymbol{V}[1:n, 1:n] = \boldsymbol{I}_n$ such that $\boldsymbol{R}_1 = \boldsymbol{R}_2$. This implies $\boldsymbol{Q}_1[:, 1:n] = \boldsymbol{Q}_2[:, 1:n]$ and leads to a contradiction such that the reduced QR decomposition is unique. ■

3.14 LQ Decomposition

We previously proved the existence of the QR decomposition via the Gram-Schmidt process in which case we are interested in the column space of a matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}$. However, in many applications (see (Schilders, 2009)), we are also interested in the row space of a matrix $\boldsymbol{B} = [\boldsymbol{b}_1^\top; \boldsymbol{b}_2^\top; ...; \boldsymbol{b}_m^\top] \in \mathbb{R}^{m\times n}$, where $\boldsymbol{b}_i$ is the i-th row of $\boldsymbol{B}$. The successive spaces spanned by the rows $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots$ of $\boldsymbol{B}$ are

$$\mathcal{C}([\boldsymbol{b}_1]) \quad \subseteq \quad \mathcal{C}([\boldsymbol{b}_1, \boldsymbol{b}_2]) \quad \subseteq \quad \mathcal{C}([\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3]) \quad \subseteq \quad \ldots.$$

The QR decomposition thus has its sibling which finds the orthogonal row space. By applying QR decomposition on $\boldsymbol{B}^\top = \boldsymbol{Q}_0\boldsymbol{R}$, we recover the LQ decomposition of the matrix $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{Q}$ where $\boldsymbol{Q} = \boldsymbol{Q}_0^\top$ and $\boldsymbol{L} = \boldsymbol{R}^\top$.

Theorem 3.11: (LQ Decomposition)

Every $m \times n$ matrix $\boldsymbol{B}$ (whether linearly independent or dependent rows) with $n \geq m$ can be factored as

$$\boldsymbol{B} = \boldsymbol{L}\boldsymbol{Q},$$

where

1. **Reduced**: $\boldsymbol{L}$ is an $m\times m$ lower triangular matrix and $\boldsymbol{Q}$ is $m\times n$ with orthonormal rows which is known as the **reduced LQ decomposition**;
2. **Full**: $\boldsymbol{L}$ is an $m \times n$ lower triangular matrix and $\boldsymbol{Q}$ is $n \times n$ with orthonormal rows which is known as the **full LQ decomposition**. If further restrict the lower triangular matrix to be a square matrix, the full LQ decomposition can be denoted as

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{L}_0 & \boldsymbol{0} \end{bmatrix} \boldsymbol{Q},$$

where $\boldsymbol{L}_0$ is an $m \times m$ square lower triangular matrix.

Row-pivoted LQ (RPLQ) Similar to the column-pivoted QR in Section 3.8, there exists a row-pivoted LQ decomposition:

$$\begin{cases} \text{Reduced RPLQ:} \qquad \boldsymbol{P}\boldsymbol{B} &= \underbrace{\begin{bmatrix} \boldsymbol{L}_{11} \\ \boldsymbol{L}_{21} \end{bmatrix}}_{m\times r} \underbrace{\boldsymbol{Q}_r}_{r\times n}; \\ \text{Full RPLQ:} \qquad \boldsymbol{P}\boldsymbol{B} &= \underbrace{\begin{bmatrix} \boldsymbol{L}_{11} & \boldsymbol{0} \\ \boldsymbol{L}_{21} & \boldsymbol{0} \end{bmatrix}}_{m\times m} \underbrace{\boldsymbol{Q}}_{m\times n}, \end{cases}$$

where $\boldsymbol{L}_{11} \in \mathbb{R}^{r\times r}$ is lower triangular, $\boldsymbol{Q}_r$ or $\boldsymbol{Q}_{1:r,:}$ spans the same row space as $\boldsymbol{B}$, and $\boldsymbol{P}$ is a permutation matrix that interchange independent rows into the upper-most rows.

3.15 Two-Sided Orthogonal Decomposition

Theorem 3.12: (Two-Sided Orthogonal Decomposition)

When square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with rank r, the full CPQR, RPLQ of $\boldsymbol{A}$ are given by

$$\boldsymbol{A}\boldsymbol{P}_1 = \boldsymbol{Q}_1 \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad \boldsymbol{P}_2\boldsymbol{A} = \begin{bmatrix} \boldsymbol{L}_{11} & \mathbf{0} \\ \boldsymbol{L}_{21} & \mathbf{0} \end{bmatrix} \boldsymbol{Q}_2$$

respectively. Then we would find out

$$\boldsymbol{A}\boldsymbol{P}\boldsymbol{A} = \boldsymbol{Q}_1 \underbrace{\begin{bmatrix} \boldsymbol{R}_{11}\boldsymbol{L}_{11} + \boldsymbol{R}_{12}\boldsymbol{L}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\text{rank } r} \boldsymbol{Q}_2,$$

where the first r columns of $\boldsymbol{Q}_1$ span the same column space of $\boldsymbol{A}$, first r rows of $\boldsymbol{Q}_2$ span the same row space of $\boldsymbol{A}$, and $\boldsymbol{P}$ is a permutation matrix. We name this decomposition as **two-sided orthogonal decomposition**.

This decomposition is very similar to the property of SVD: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ that the first r columns of $\boldsymbol{U}$ span the column space of $\boldsymbol{A}$ and the first r columns of $\boldsymbol{V}$ span the row space of $\boldsymbol{A}$ (we shall see in Lemma 14.8, p. 138). Therefore, the two-sided orthogonal decomposition can be regarded as an inexpensive alternative in this sense.

Lemma 3.13: (Four Orthonormal Basis)

Given the two-sided orthogonal decomposition of matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with rank r: $\boldsymbol{A}\boldsymbol{P}\boldsymbol{A} = \boldsymbol{U}\boldsymbol{F}\boldsymbol{V}^\top$, where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ are the column partitions of $\boldsymbol{U}$ and $\boldsymbol{V}$. Then, we have the following property:

- $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{A}^\top)$;
- $\{\boldsymbol{v}_{r+1}, \boldsymbol{v}_{r+2}, \ldots, \boldsymbol{v}_n\}$ is an orthonormal basis of $\mathcal{N}(\boldsymbol{A})$;
- $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r\}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{A})$;
- $\{\boldsymbol{u}_{r+1}, \boldsymbol{u}_{r+2}, \ldots, \boldsymbol{u}_n\}$ is an orthonormal basis of $\mathcal{N}(\boldsymbol{A}^\top)$.

3.16 Rank-One Changes

We previously discussed the rank-one update/downdate of the Cholesky decomposition in Section 2.8 (p. 40). The rank-one change $\boldsymbol{A}'$ of matrix $\boldsymbol{A}$ in the QR decomposition is defined in a similar form:

$$\begin{aligned} \boldsymbol{A}' &= \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^\top, \\ \downarrow \;\; & \quad \downarrow \\ \boldsymbol{Q}'\boldsymbol{R}' &= \boldsymbol{Q}\boldsymbol{R} + \boldsymbol{u}\boldsymbol{v}^\top, \end{aligned}$$

where if we set $\boldsymbol{A}' = \boldsymbol{A} - (-\boldsymbol{u})\boldsymbol{v}^\top$, we recover the downdate form such that the update or downdate in the QR decomposition are the same. Let $\boldsymbol{w} = \boldsymbol{Q}^\top\boldsymbol{u}$, we have

$$\boldsymbol{A}' = \boldsymbol{Q}(\boldsymbol{R} + \boldsymbol{w}\boldsymbol{v}^\top).$$

From the second form in Remark 3.9 on introducing zeros backwards, there exists a set of Givens rotations $\boldsymbol{G}_{12}\boldsymbol{G}_{23}\ldots\boldsymbol{G}_{(n-1),n}$ such that

$$
\boldsymbol{G}_{12}\boldsymbol{G}_{23}\ldots\boldsymbol{G}_{(n-1),n}\boldsymbol{w} = \pm||\boldsymbol{w}||\boldsymbol{e}_1,
$$

where $\boldsymbol{G}_{(k-1),k}$ is the Givens rotation in plane $k-1$ and k that introduces zero in the k-th entry of $\boldsymbol{w}$. Apply this rotation to $\boldsymbol{R}$, we have

$$
\boldsymbol{G}_{12}\boldsymbol{G}_{23}\ldots\boldsymbol{G}_{(n-1),n}\boldsymbol{R} = \boldsymbol{H}_0,
$$

where the Givens rotations in this *reverse order* are useful to transform the upper triangular $\boldsymbol{R}$ into a "simple" upper Hessenberg which is close to upper triangular matrices (see Definition 8.1 that we will introduce in the Hessenberg decomposition). If the rotations are transforming $\boldsymbol{w}$ into $\pm||\boldsymbol{w}||\boldsymbol{e}_1$ from *forward order* as in Corollary 3.8, we will not have this upper Hessenberg $\boldsymbol{H}_0$. To see this, suppose $\boldsymbol{R} \in \mathbb{R}^{5\times 5}$, an example is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed. The backwards rotations result in the upper Hessenberg $\boldsymbol{H}_0$ which is relatively simple to handle:

$$
\text{Backwards:}\;
\underset{\boldsymbol{R}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & 0 & 0 & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{G}_{45}}{\rightarrow}
\underset{\boldsymbol{G}_{45}\boldsymbol{R}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\stackrel{\boldsymbol{G}_{34}}{\rightarrow}
\underset{\boldsymbol{G}_{34}\boldsymbol{G}_{45}\boldsymbol{R}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \boxtimes & \boxtimes
\end{bmatrix}}
$$

$$
\stackrel{\boldsymbol{G}_{23}}{\rightarrow}
\underset{\boldsymbol{G}_{23}\boldsymbol{G}_{34}\boldsymbol{G}_{45}\boldsymbol{R}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{G}_{12}}{\rightarrow}
\underset{\boldsymbol{G}_{12}\boldsymbol{G}_{23}\boldsymbol{G}_{34}\boldsymbol{G}_{45}\boldsymbol{R}}{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes
\end{bmatrix}}.
$$

And the forward rotations result in a full matrix:

$$
\text{Forwards:}\;
\underset{\boldsymbol{R}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & 0 & 0 & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{G}_{12}}{\rightarrow}
\underset{\boldsymbol{G}_{12}\boldsymbol{R}}{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & 0 & 0 & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{G}_{23}}{\rightarrow}
\underset{\boldsymbol{G}_{23}\boldsymbol{G}_{12}\boldsymbol{R}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & 0 & 0 & \boxtimes
\end{bmatrix}}
$$

$$
\stackrel{\boldsymbol{G}_{34}}{\rightarrow}
\underset{\boldsymbol{G}_{34}\boldsymbol{G}_{23}\boldsymbol{G}_{12}\boldsymbol{R}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & 0 & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{G}_{45}}{\rightarrow}
\underset{\boldsymbol{G}_{45}\boldsymbol{G}_{34}\boldsymbol{G}_{23}\boldsymbol{G}_{12}\boldsymbol{R}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}.
$$

Generally, the backward rotation results in,

$$
\boldsymbol{G}_{12}\boldsymbol{G}_{23}\ldots\boldsymbol{G}_{(n-1),n}(\boldsymbol{R}+\boldsymbol{w}\boldsymbol{v}^\top) = \boldsymbol{H}_0 \pm ||\boldsymbol{w}||\boldsymbol{e}_1\boldsymbol{v}^\top = \boldsymbol{H},
$$

which is also upper Hessenberg. Similar to triangularization via the Givens rotation in Section 3.12, there exists a set of rotations $\boldsymbol{J}_{12}, \boldsymbol{J}_{23}, \ldots, \boldsymbol{J}_{(n-1),n}$ such that

$$
\boldsymbol{J}_{(n-1),n}\ldots\boldsymbol{J}_{23}\boldsymbol{J}_{12}\boldsymbol{H} = \boldsymbol{R}',
$$

is upper triangular. Following from the 5×5 example, the triangularization is shown as follows

$$
\underbrace{\boldsymbol{H}_0 \pm ||\boldsymbol{w}||\boldsymbol{e}_1\boldsymbol{v}^\top}_{\boldsymbol{H}} = \mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes
\end{bmatrix}}_{\boldsymbol{H}}
\stackrel{\boldsymbol{J}_{12}}{\rightarrow}
\mathop{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes
\end{bmatrix}}_{\boldsymbol{J}_{12}\boldsymbol{H}}
\stackrel{\boldsymbol{J}_{23}}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes
\end{bmatrix}}_{\boldsymbol{J}_{23}\boldsymbol{J}_{12}\boldsymbol{H}}
$$

$$
\stackrel{\boldsymbol{J}_{34}}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \boxtimes & \boxtimes
\end{bmatrix}}_{\boldsymbol{J}_{34}\boldsymbol{J}_{23}\boldsymbol{J}_{12}\boldsymbol{H}}
\stackrel{\boldsymbol{J}_{45}}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes}
\end{bmatrix}}_{\boldsymbol{J}_{45}\boldsymbol{J}_{34}\boldsymbol{J}_{23}\boldsymbol{J}_{12}\boldsymbol{H}}.
$$

And the QR decomposition of $\boldsymbol{A}'$ is thus given by

$$
\boldsymbol{A}' = \boldsymbol{Q}'\boldsymbol{R}',
$$

where

$$
\begin{cases}
\boldsymbol{R}' = (\boldsymbol{J}_{(n-1),n}\ldots\boldsymbol{J}_{23}\boldsymbol{J}_{12})(\boldsymbol{G}_{12}\boldsymbol{G}_{23}\ldots\boldsymbol{G}_{(n-1),n})(\boldsymbol{R}+\boldsymbol{w}\boldsymbol{v}^\top); \\
\boldsymbol{Q}' = \boldsymbol{Q}\left\{(\boldsymbol{J}_{(n-1),n}\ldots\boldsymbol{J}_{23}\boldsymbol{J}_{12})(\boldsymbol{G}_{12}\boldsymbol{G}_{23}\ldots\boldsymbol{G}_{(n-1),n})\right\}^\top.
\end{cases}
\tag{3.13}
$$

3.17 Appending or Deleting a Column

Deleting a column Suppose the QR decomposition of $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ is given by $\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{R}$ where the column partition of $\boldsymbol{A}$ is $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\ldots,\boldsymbol{a}_n]$. Now, if we delete the k-th column of $\boldsymbol{A}$ such that $\boldsymbol{A}'=[\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{k-1},\boldsymbol{a}_{k+1},\ldots,\boldsymbol{a}_n]\in\mathbb{R}^{m\times(n-1)}$. We want to find the QR decomposition of $\boldsymbol{A}'$ efficiently. Suppose further $\boldsymbol{R}$ has the following form

$$
\boldsymbol{R} = \begin{array}{c}
\begin{bmatrix}
\boldsymbol{R}_{11} & \boldsymbol{a} & \boldsymbol{R}_{12} \\
\mathbf{0} & r_{kk} & \boldsymbol{b}^\top \\
\mathbf{0} & \mathbf{0} & \boldsymbol{R}_{22}
\end{bmatrix}
\begin{matrix} k-1 \\ 1 \\ m-k \end{matrix} \\
\begin{matrix} k-1 & \quad 1 \quad & n-k \end{matrix}
\end{array}.
$$

Apparently,

$$
\boldsymbol{Q}^\top\boldsymbol{A}' = \begin{bmatrix}
\boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\
\mathbf{0} & \boldsymbol{b}^\top \\
\mathbf{0} & \boldsymbol{R}_{22}
\end{bmatrix} = \boldsymbol{H}
$$

is upper Hessenberg. A 6×5 example is shown as follows where $k=3$:

$$
\underset{\boldsymbol{R}=\boldsymbol{Q}^\top\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & 0 & 0 & \boxtimes \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}}
\longrightarrow
\underset{\boldsymbol{H}=\boldsymbol{Q}^\top\boldsymbol{A}'}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes \\
0 & 0 & 0 & 0
\end{bmatrix}}.
$$

Again, for columns k to $n-1$ of $\boldsymbol{H}$, there exists a set of rotations $\boldsymbol{G}_{k,k+1}$, $\boldsymbol{G}_{k+1,k+2}$, $\ldots$, $\boldsymbol{G}_{n-1,n}$ that could introduce zeros for the elements $h_{k+1,k}$, $h_{k+2,k+1}$, $\ldots$, $h_{n,n-1}$ of $\boldsymbol{H}$. The the triangular matrix $\boldsymbol{R}'$ is given by

$$
\boldsymbol{R}' = \boldsymbol{G}_{n-1,n}\ldots\boldsymbol{G}_{k+1,k+2}\boldsymbol{G}_{k,k+1}\boldsymbol{Q}^\top\boldsymbol{A}'.
$$

And the orthogonal matrix

$$
\boldsymbol{Q}' = (\boldsymbol{G}_{n-1,n}\ldots\boldsymbol{G}_{k+1,k+2}\boldsymbol{G}_{k,k+1}\boldsymbol{Q}^\top)^\top = \boldsymbol{Q}\boldsymbol{G}_{k,k+1}^\top\boldsymbol{G}_{k+1,k+2}^\top\ldots\boldsymbol{G}_{n-1,n}^\top, \tag{3.14}
$$

such that $\boldsymbol{A}'=\boldsymbol{Q}'\boldsymbol{R}'$.

Appending a column Similarly, suppose $\widetilde{\boldsymbol{A}} = [\boldsymbol{a}_1, \boldsymbol{a}_k, \boldsymbol{w}, \boldsymbol{a}_{k+1}, \ldots, \boldsymbol{a}_n]$ where we append $\boldsymbol{w}$ into the $(k+1)$-th column of $\boldsymbol{A}$. We can obtain

$$
\boldsymbol{Q}^\top\widetilde{\boldsymbol{A}} = [\boldsymbol{Q}^\top\boldsymbol{a}_1, \ldots, \boldsymbol{Q}^\top\boldsymbol{a}_k, \boldsymbol{Q}^\top\boldsymbol{w}, \boldsymbol{Q}^\top\boldsymbol{a}_{k+1}, \ldots, \boldsymbol{Q}^\top\boldsymbol{a}_n] = \widetilde{\boldsymbol{H}}.
$$

A set of Givens rotations $\boldsymbol{J}_{m-1,m}, \boldsymbol{J}_{m-2,m-1}, \ldots, \boldsymbol{J}_{k+1,k+2}$ can introduce zeros for the $\widetilde{h}_{m,k+1}$, $\widetilde{h}_{m-1,k+1}$, $\ldots$, $\widetilde{h}_{k+2,k+1}$ elements of $\widetilde{\boldsymbol{H}}$ such that

$$
\widetilde{\boldsymbol{R}} = \boldsymbol{J}_{k+1,k+2}\ldots\boldsymbol{J}_{m-2,m-1}\boldsymbol{J}_{m-1,m}\boldsymbol{Q}^\top\widetilde{\boldsymbol{A}}.
$$

Suppose $\widetilde{\boldsymbol{H}}$ is of size 6×5 and $k=2$, an example is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

$$
\underset{\widetilde{\boldsymbol{H}}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & 0 & \boxtimes \\
0 & 0 & \boxtimes & 0 & 0 \\
0 & 0 & \boxtimes & 0 & 0
\end{bmatrix}}
\stackrel{\boldsymbol{J}_{56}}{\rightarrow}
\underset{\boldsymbol{J}_{56}\widetilde{\boldsymbol{H}}\rightarrow \widetilde{h}_{63}=0}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & 0 & \boxtimes \\
0 & 0 & \boldsymbol{\boxtimes} & 0 & 0 \\
0 & 0 & \mathbf{0} & 0 & 0
\end{bmatrix}}
\stackrel{\boldsymbol{J}_{45}}{\rightarrow}
\underset{\boldsymbol{J}_{45}\boldsymbol{J}_{56}\widetilde{\boldsymbol{H}}\rightarrow \widetilde{h}_{53}=0}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boldsymbol{\boxtimes} & 0 & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & 0 & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}}
\stackrel{\boldsymbol{J}_{34}}{\rightarrow}
\underset{\boldsymbol{J}_{34}\boldsymbol{J}_{45}\boldsymbol{J}_{56}\widetilde{\boldsymbol{H}}\rightarrow \widetilde{h}_{43}=0}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & 0 & \boxtimes \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}}.
$$

And finally, the orthogonal matrix

$$
\widetilde{\boldsymbol{Q}} = (\boldsymbol{J}_{k+1,k+2}\ldots\boldsymbol{J}_{m-2,m-1}\boldsymbol{J}_{m-1,m}\boldsymbol{Q}^\top)^\top = \boldsymbol{Q}\boldsymbol{J}_{m-1,m}^\top\boldsymbol{J}_{m-2,m-1}^\top\ldots\boldsymbol{J}_{k+1,k+2}^\top, \tag{3.15}
$$

such that $\widetilde{\boldsymbol{A}} = \widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{R}}$.

Real world application The method introduced above is useful for the efficient variable selection in the least squares problem via the QR decomposition. At each time we delete a column of the data matrix $\boldsymbol{A}$, and apply an F-test to see if the variable is significant or not. If not, we will delete the variable and favor a simpler model (Lu, 2021e).

3.18 Appending or Deleting a Row

Appending a row Suppose the full QR decomposition of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is given by $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} = \boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{A}_1 \in \mathbb{R}^{k\times n}$ and $\boldsymbol{A}_2 \in \mathbb{R}^{(m-k)\times n}$. Now, if we add a row such that $\boldsymbol{A}' = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{w}^\top \\ \boldsymbol{A}_2 \end{bmatrix} \in \mathbb{R}^{(m+1)\times n}$. We want to find the full QR decomposition of $\boldsymbol{A}'$ efficiently. Construct a permutation matrix

$$
\boldsymbol{P} = \begin{bmatrix} \mathbf{0} & 1 & \mathbf{0} \\ \boldsymbol{I}_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{m-k} \end{bmatrix} \longrightarrow \boldsymbol{P} \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{w}^\top \\ \boldsymbol{A}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}^\top \\ \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix}.
$$

Then,

$$
\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^\top \end{bmatrix} \boldsymbol{P}\boldsymbol{A}' = \begin{bmatrix} \boldsymbol{w}^\top \\ \boldsymbol{R} \end{bmatrix} = \boldsymbol{H}
$$

is upper Hessenberg. Similarly, a set of rotations $\boldsymbol{G}_{12}, \boldsymbol{G}_{23}, \ldots, \boldsymbol{G}_{n,n+1}$ can be applied to introduce zeros for the elements h_{21}, h_{32}, $\ldots$, $h_{n+1,n}$ of $\boldsymbol{H}$. The triangular matrix $\boldsymbol{R}'$ is given by

$$
\boldsymbol{R}' = \boldsymbol{G}_{n,n+1} \ldots \boldsymbol{G}_{23}\boldsymbol{G}_{12} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^\top \end{bmatrix} \boldsymbol{P}\boldsymbol{A}'.
$$

And the orthogonal matrix

$$
\boldsymbol{Q}' = \left(\boldsymbol{G}_{n,n+1} \ldots \boldsymbol{G}_{23}\boldsymbol{G}_{12} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^\top \end{bmatrix} \boldsymbol{P} \right)^\top = \boldsymbol{P}^\top \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q} \end{bmatrix} \boldsymbol{G}_{12}^\top \boldsymbol{G}_{23}^\top \ldots \boldsymbol{G}_{n,n+1}^\top,
$$

such that $\boldsymbol{A}' = \boldsymbol{Q}'\boldsymbol{R}'$.

Deleting a row Suppose $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{w}^\top \\ \boldsymbol{A}_2 \end{bmatrix} \in \mathbb{R}^{m\times n}$ where $\boldsymbol{A}_1 \in \mathbb{R}^{k\times n}$, $\boldsymbol{A}_2 \in \mathbb{R}^{(m-k-1)\times n}$ with the full QR decomposition given by $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{Q} \in \mathbb{R}^{m\times m}, \boldsymbol{R} \in \mathbb{R}^{m\times n}$. We want to compute the full QR decomposition of $\widetilde{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix}$ efficiently (assume $m-1 \geq n$). Analogously, we can construct a permutation matrix

$$
\boldsymbol{P} = \begin{bmatrix} \mathbf{0} & 1 & \mathbf{0} \\ \boldsymbol{I}_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{m-k-1} \end{bmatrix}
$$

such that

$$
\boldsymbol{P}\boldsymbol{A} = \begin{bmatrix} \mathbf{0} & 1 & \mathbf{0} \\ \boldsymbol{I}_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{m-k-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{w}^\top \\ \boldsymbol{A}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}^\top \\ \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} = \boldsymbol{P}\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{M}\boldsymbol{R},
$$

where $\boldsymbol{M} = \boldsymbol{P}\boldsymbol{Q}$ is an orthogonal matrix. Let $\boldsymbol{m}^\top$ be the first row of $\boldsymbol{M}$, and a set of givens rotations $\boldsymbol{G}_{m-1,m}, \boldsymbol{G}_{m-2,m-1}, \ldots, \boldsymbol{G}_{1,2}$ introducing zeros for elements $m_m, m_{m-1}, \ldots, m_2$ of $\boldsymbol{m}$ respectively such that $\boldsymbol{G}_{1,2} \ldots \boldsymbol{G}_{m-2,m-1}\boldsymbol{G}_{m-1,m}\boldsymbol{m} = \alpha \boldsymbol{e}_1$ where $\alpha = \pm 1$. Therefore, we have

$$
\boldsymbol{G}_{1,2} \ldots \boldsymbol{G}_{m-2,m-1}\boldsymbol{G}_{m-1,m}\boldsymbol{R} = \begin{array}{cc} \begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{R}_1 \end{bmatrix} & \begin{array}{c} 1 \\ m-1 \end{array} \end{array},
$$

which is upper Hessenberg with $\boldsymbol{R}_1 \in \mathbb{R}^{(m-1)\times n}$ being upper triangular. And

$$
\boldsymbol{M}\boldsymbol{G}_{m-1,m}^\top \boldsymbol{G}_{m-2,m-1}^\top \ldots \boldsymbol{G}_{1,2}^\top = \begin{bmatrix} \alpha & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q}_1 \end{bmatrix},
$$

where $\boldsymbol{Q}_1 \in \mathbb{R}^{(m-1)\times(m-1)}$ is an orthogonal matrix. The bottom-left block of the above matrix is a zero vector since $\alpha = \pm 1$ and $\boldsymbol{M}$ is orthogonal. To see this, let $\boldsymbol{G} = \boldsymbol{G}_{m-1,m}^\top \boldsymbol{G}_{m-2,m-1}^\top \ldots \boldsymbol{G}_{1,2}^\top$ with the first column being $\boldsymbol{g}$ and $\boldsymbol{M} = [\boldsymbol{m}^\top; \boldsymbol{m}_2^\top; \boldsymbol{m}_3^\top; \ldots, \boldsymbol{m}_m^\top]$ being the row partition of $\boldsymbol{M}$. We have

$$
\begin{aligned}
\boldsymbol{m}^\top \boldsymbol{g} = \pm 1 \qquad &\rightarrow \qquad \boldsymbol{g} = \pm \boldsymbol{m}, \\
\boldsymbol{m}_i^\top \boldsymbol{m} = 0, \qquad &\forall i \in \{2, 3, \ldots, m\}.
\end{aligned}
$$

This results in

$$
\begin{aligned}
\boldsymbol{P}\boldsymbol{A} &= \boldsymbol{M}\boldsymbol{R} \\
&= (\boldsymbol{M}\boldsymbol{G}_{m-1,m}^\top \boldsymbol{G}_{m-2,m-1}^\top \ldots \boldsymbol{G}_{1,2}\top)(\boldsymbol{G}_{1,2} \ldots \boldsymbol{G}_{m-2,m-1}\boldsymbol{G}_{m-1,m}\boldsymbol{R}) \\
&= \begin{bmatrix} \alpha & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q}_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}^\top \\ \boldsymbol{R}_1 \end{bmatrix} = \begin{bmatrix} \alpha \boldsymbol{v}^\top \\ \boldsymbol{Q}_1\boldsymbol{R}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}^\top \\ \widetilde{\boldsymbol{A}} \end{bmatrix}.
\end{aligned}
$$

This implies $\boldsymbol{Q}_1\boldsymbol{R}_1$ is the full QR decomposition of $\widetilde{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix}$.

4. UTV Decomposition: ULV and URV Decomposition

4.1 UTV Decomposition

The UTV decomposition goes further by factoring the matrix into two orthogonal matrices $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{T}\boldsymbol{V}$, where $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal, whilst $\boldsymbol{T}$ is (upper/lower) triangular.[14] The resulting $\boldsymbol{T}$ supports rank estimation. The matrix $\boldsymbol{T}$ can be lower triangular which results in the ULV decomposition, or it can be upper triangular which results in the URV decomposition. The UTV framework shares a similar form as the singular value decomposition (SVD, see Section 14, p. 134) and can be regarded as inexpensive alternatives to the SVD.

Theorem 4.1: (Full ULV Decomposition)

14. These decompositions fall into a category known as the *double-sided orthogonal decomposition*. We will see, the UTV decomposition, complete orthogonal decomposition, and singular value decomposition are all in this notion.

Every $m \times n$ matrix $\boldsymbol{A}$ with rank r can be factored as

$$\boldsymbol{A} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{L} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{V},$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are two orthogonal matrices, and $\boldsymbol{L} \in \mathbb{R}^{r\times r}$ is a lower triangular matrix.

The existence of the ULV decomposition is from the QR and LQ decomposition.

Proof [of Theorem 4.1] For any rank r matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n]$, we can use a column permutation matrix $\boldsymbol{P}$ (Definition 0.17, p. 15) such that the linearly independent columns of $\boldsymbol{A}$ appear in the first r columns of $\boldsymbol{AP}$. Without loss of generality, we assume $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_r$ are the r linearly independent columns of $\boldsymbol{A}$ and

$$\boldsymbol{AP} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_r, \boldsymbol{b}_{r+1}, \ldots, \boldsymbol{b}_n].$$

Let $\boldsymbol{Z} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_r] \in \mathbb{R}^{m\times r}$. Since any $\boldsymbol{b}_i$ is in the column space of $\boldsymbol{Z}$, we can find a $\boldsymbol{E} \in \mathbb{R}^{r\times(n-r)}$ such that

$$[\boldsymbol{b}_{r+1}, \boldsymbol{b}_{r+2}, \ldots, \boldsymbol{b}_n] = \boldsymbol{ZE}.$$

That is,

$$\boldsymbol{AP} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_r, \boldsymbol{b}_{r+1}, \ldots, \boldsymbol{b}_n] = \boldsymbol{Z} \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{E} \end{bmatrix},$$

where $\boldsymbol{I}_r$ is an $r \times r$ identity matrix. Moreover, $\boldsymbol{Z} \in \mathbb{R}^{m\times r}$ has full column rank such that its full QR decomposition is given by $\boldsymbol{Z} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix}$, where $\boldsymbol{R} \in \mathbb{R}^{r\times r}$ is an upper triangular matrix with full rank and $\boldsymbol{U}$ is an orthogonal matrix. This implies

$$\boldsymbol{AP} = \boldsymbol{Z} \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{E} \end{bmatrix} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{E} \end{bmatrix} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{RE} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \tag{4.1}$$

Since $\boldsymbol{R}$ has full rank, this means $\begin{bmatrix} \boldsymbol{R} & \boldsymbol{RE} \end{bmatrix}$ also has full rank such that its full LQ decomposition is given by $\begin{bmatrix} \boldsymbol{L} & \boldsymbol{0} \end{bmatrix} \boldsymbol{V}_0$ where $\boldsymbol{L} \in \mathbb{R}^{r\times r}$ is a lower triangular matrix and $\boldsymbol{V}_0$ is an orthogonal matrix. Substitute into Equation (4.1), we have

$$\boldsymbol{A} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{L} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{V}_0 \boldsymbol{P}^{-1}.$$

Let $\boldsymbol{V} = \boldsymbol{V}_0 \boldsymbol{P}^{-1}$ which is a product of two orthogonal matrices, and is also an orthogonal matrix. This completes the proof. ∎

A second way to see the proof of the ULV decomposition will be discussed in the proof of Theorem 4.3 shortly via the rank-revealing QR decomposition and trivial QR decomposition. Now suppose the ULV decomposition of matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{L} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{V}.$$

Let $\boldsymbol{U}_0 = \boldsymbol{U}_{:,1:r}$ and $\boldsymbol{V}_0 = \boldsymbol{V}_{1:r,:}$, i.e., $\boldsymbol{U}_0$ contains only the first r columns of $\boldsymbol{U}$, and $\boldsymbol{V}_0$ contains only the first r rows of $\boldsymbol{V}$. Then, we still have $\boldsymbol{A} = \boldsymbol{U}_0 \boldsymbol{L} \boldsymbol{V}_0$. This is known as the **reduced ULV decomposition**. Similarly, we can also claim the URV decomposition as follows.

Theorem 4.2: (Full URV Decomposition)
Every $m \times n$ matrix $\boldsymbol{A}$ with rank r can be factored as

$$\boldsymbol{A} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \boldsymbol{V},$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are two orthogonal matrices, and $\boldsymbol{R} \in \mathbb{R}^{r\times r}$ is an upper triangular matrix.

The proof is just similar to that of ULV decomposition and we shall not give the details. Again, there is a version of reduced URV decomposition and the difference between the full and reduced URV can be implied from the context. The ULV and URV sometimes are referred to as the UTV decomposition framework (Fierro and Hansen, 1997; Golub and Van Loan, 2013).

We will shortly see that the forms of ULV and URV are very close to the singular value decomposition (SVD). All of the three factor the matrix $\boldsymbol{A}$ into two orthogonal matrices. Specially, there exists a set of basis for the four subspaces of $\boldsymbol{A}$ in the fundamental theorem of linear algebra via the ULV and the URV. Taking ULV as an example, the first r columns of $\boldsymbol{U}$ form an orthonormal basis of $\mathcal{C}(\boldsymbol{A})$, and the last $(m-r)$ columns of $\boldsymbol{U}$ form an orthonormal basis of $\mathcal{N}(\boldsymbol{A}^\top)$. The first r rows of $\boldsymbol{V}$ form an orthonormal basis for the row space $\mathcal{C}(\boldsymbol{A}^\top)$, and the last $(n-r)$ rows form an orthonormal basis for $\mathcal{N}(\boldsymbol{A})$ (similar to the two-sided orthogonal decomposition):

$$\begin{cases} \mathcal{C}(\boldsymbol{A}) = span\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r\}, \\ \mathcal{N}(\boldsymbol{A}) = span\{\boldsymbol{v}_{r+1}, \boldsymbol{v}_{r+2}, \ldots, \boldsymbol{v}_n\}, \\ \mathcal{C}(\boldsymbol{A}^\top) = span\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}, \\ \mathcal{N}(\boldsymbol{A}^\top) = span\{\boldsymbol{u}_{r+1}, \boldsymbol{u}_{r+2}, \ldots, \boldsymbol{u}_m\}. \end{cases}$$

The SVD goes further that there is a connection between the two pairs of orthonormal basis, i.e., transforming from column basis into row basis, or left null space basis into right null space basis. We will get more details in the SVD section.

4.2 Complete Orthogonal Decomposition

What is related to the UTV decomposition is called the *complete orthogonal decomposition* which factors into two orthogonal matrices as well.

Theorem 4.3: (Complete Orthogonal Decomposition)
Every $m \times n$ matrix $\boldsymbol{A}$ with rank r can be factored as

$$\boldsymbol{A} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \boldsymbol{V},$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are two orthogonal matrices, and $\boldsymbol{T} \in \mathbb{R}^{r\times r}$ is an rank-r matrix.

Proof [of Theorem 4.3] By rank-revealing QR decomposition (Theorem 3.2, p. 53), $\boldsymbol{A}$ can be factored as

$$\boldsymbol{Q}_1^\top \boldsymbol{A}\boldsymbol{P} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\boldsymbol{R}_{11} \in \mathbb{R}^{r\times r}$ is upper triangular, $\boldsymbol{R}_{12} \in \mathbb{R}^{r\times(n-r)}$, $\boldsymbol{Q}_1 \in \mathbb{R}^{m\times m}$ is an orthogonal matrix, and $\boldsymbol{P}$ is a permutation matrix.

Then, it is not hard to find a decomposition such that

$$\begin{bmatrix} \boldsymbol{R}_{11}^\top \\ \boldsymbol{R}_{12}^\top \end{bmatrix} = \boldsymbol{Q}_2 \begin{bmatrix} \boldsymbol{S} \\ \mathbf{0} \end{bmatrix}, \tag{4.2}$$

where $\boldsymbol{Q}_2$ is an orthogonal matrix, $\boldsymbol{S}$ is an rank-r matrix. The decomposition is reasonable in the sense the matrix $\begin{bmatrix} \boldsymbol{R}_{11}^\top \\ \boldsymbol{R}_{12}^\top \end{bmatrix} \in \mathbb{R}^{n\times r}$ has rank r of which the columns stay in a subspace of $\mathbb{R}^n$. Nevertheless, the columns of $\boldsymbol{Q}_2$ span the whole space of $\mathbb{R}^n$, where we can assume the first r columns of $\boldsymbol{Q}_2$ span the same space as that of $\begin{bmatrix} \boldsymbol{R}_{11}^\top \\ \boldsymbol{R}_{12}^\top \end{bmatrix}$. The matrix $\begin{bmatrix} \boldsymbol{S} \\ \mathbf{0} \end{bmatrix}$ is to transform $\boldsymbol{Q}_2$ into $\begin{bmatrix} \boldsymbol{R}_{11}^\top \\ \boldsymbol{R}_{12}^\top \end{bmatrix}$.

Then, it follows that

$$\boldsymbol{Q}_1^\top \boldsymbol{A}\boldsymbol{P}\boldsymbol{Q}_2 = \begin{bmatrix} \boldsymbol{S}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Let $\boldsymbol{U} = \boldsymbol{Q}_1$, $\boldsymbol{V} = \boldsymbol{Q}_2^\top \boldsymbol{P}^\top$ and $\boldsymbol{T} = \boldsymbol{S}^\top$, we complete the proof. ■

We can find that when Equation (4.2) is taken to be the reduced QR decomposition of $\begin{bmatrix} \boldsymbol{R}_{11}^\top \\ \boldsymbol{R}_{12}^\top \end{bmatrix}$, then the complete orthogonal decomposition reduces to the ULV decomposition.

4.3 Application: Row Rank equals Column Rank Again via UTV

As mentioned above, the UTV framework can prove the important theorem in linear algebra that the row rank and column rank of a matrix are equal. Notice that to apply the UTV in the proof, a slight modification on the claim of the existence of the UTV decomposition needs to be taken care of. For example, in Theorem 4.1, the assumption of the matrix $\boldsymbol{A}$ is to have rank r. Since rank r already admits the fact that row rank equals column rank. A better claim here to this aim is to say matrix $\boldsymbol{A}$ has column rank r in Theorem 4.1. See (Lu, 2021b) for a detailed discussion.

Proof [of Theorem 0.13, p. 12, A Second Way] Any $m \times n$ matrix $\boldsymbol{A}$ with rank r can be factored as

$$\boldsymbol{A} = \boldsymbol{U}_0 \begin{bmatrix} \boldsymbol{L} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \boldsymbol{V}_0,$$

where $\boldsymbol{U}_0 \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V}_0 \in \mathbb{R}^{n\times n}$ are two orthogonal matrices, and $\boldsymbol{L} \in \mathbb{R}^{r\times r}$ is a lower triangular matrix [15]. Let $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{L} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$, the row rank and column rank of $\boldsymbol{D}$ are apparently the same. If we could prove the column rank of $\boldsymbol{A}$ equals the column rank of $\boldsymbol{D}$, and the row rank of $\boldsymbol{A}$ equals the row rank of $\boldsymbol{D}$, then we complete the proof.

Let $\boldsymbol{U} = \boldsymbol{U}_0^\top$, $\boldsymbol{V} = \boldsymbol{V}_0^\top$, then $\boldsymbol{D} = \boldsymbol{UAV}$. Decompose the above idea into two steps, a moment of reflexion reveals that, if we could first prove the row rank and column rank of $\boldsymbol{A}$ are equal to those of $\boldsymbol{UA}$, and then, if we further prove the row rank and column rank of $\boldsymbol{UA}$ are equal to those of $\boldsymbol{UAV}$, we could also complete the proof.

Row rank and column rank of $\boldsymbol{A}$ are equal to those of $\boldsymbol{UA}$ Let $\boldsymbol{B} = \boldsymbol{UA}$, and let further $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n]$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n]$ be the column partitions of $\boldsymbol{A}$ and $\boldsymbol{B}$. Therefore, $[\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n] = [\boldsymbol{U}\boldsymbol{a}_1, \boldsymbol{U}\boldsymbol{a}_2, \ldots, \boldsymbol{U}\boldsymbol{a}_n]$. If $x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \ldots + x_n\boldsymbol{a}_n = 0$, then we also have

$$\boldsymbol{U}(x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \ldots + x_n\boldsymbol{a}_n) = x_1\boldsymbol{b}_1 + x_2\boldsymbol{b}_2 + \ldots + x_n\boldsymbol{b}_n = 0.$$

Let $j_1, j_2, \ldots, j_r$ be distinct indices between 1 and n, if the set $\{\boldsymbol{a}_{j_1}, \boldsymbol{a}_{j_2}, \ldots, \boldsymbol{a}_{j_r}\}$ is independent, the set $\{\boldsymbol{b}_{j_1}, \boldsymbol{b}_{j_2}, \ldots, \boldsymbol{b}_{j_r}\}$ must also be linearly independent. This implies

$$dim(\mathcal{C}(\boldsymbol{B})) \leq dim(\mathcal{C}(\boldsymbol{A})).$$

Similarly, by $\boldsymbol{A} = \boldsymbol{U}^\top\boldsymbol{B}$, it follows that

$$dim(\mathcal{C}(\boldsymbol{A})) \leq dim(\mathcal{C}(\boldsymbol{B})).$$

This implies

$$dim(\mathcal{C}(\boldsymbol{B})) = dim(\mathcal{C}(\boldsymbol{A})).$$

Apply the process onto $\boldsymbol{B}^\top$ and $\boldsymbol{A}^\top$, we have

$$dim(\mathcal{C}(\boldsymbol{B}^\top)) = dim(\mathcal{C}(\boldsymbol{A}^\top)).$$

This implies the row rank and column rank of $\boldsymbol{A}$ and $\boldsymbol{B} = \boldsymbol{UA}$ are the same. Similarly, we can also show that the row rank and column rank of $\boldsymbol{UA}$ and $\boldsymbol{UAV}$ are the same. This completes the proof. ■

15. Instead of using the ULV decomposition, in some texts, the authors use elementary transformations $\boldsymbol{E}_1, \boldsymbol{E}_2$ such that $\boldsymbol{A} = \boldsymbol{E}_1 \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{E}_2$, to prove the result.

Part III
Data Interpretation and Information Distillation

5. CR Decomposition

CR decomposition is proposed in (Strang, 2021; Strang and Moler, 2021). As usual, we firstly give the result and we will discuss the existence and the origin of this decomposition in the following sections.

Theorem 5.1: (CR Decomposition)

Any rank-r matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ can be factored as

$$\underset{m\times n}{\boldsymbol{A}} = \underset{m\times r}{\boldsymbol{C}} \quad \underset{r\times n}{\boldsymbol{R}}$$

where $\boldsymbol{C}$ is the first r linearly independent columns of $\boldsymbol{A}$, and $\boldsymbol{R}$ is an $r \times n$ matrix to reconstruct the columns of $\boldsymbol{A}$ from columns of $\boldsymbol{C}$. In particular, $\boldsymbol{R}$ is the row reduced echelon form (RREF) of $\boldsymbol{A}$ without the zero rows.

The storage for the decomposition is then reduced or potentially increased from mn to $r(m+n)$.

5.1 Existence of the CR Decomposition

Since matrix $\boldsymbol{A}$ is of rank r, there are some r linearly independent columns in $\boldsymbol{A}$. We then choose linearly independent columns from $\boldsymbol{A}$ and put them into $\boldsymbol{C}$:

Find r linearly Independent Columns From $\boldsymbol{A}$

1. If column 1 of $\boldsymbol{A}$ is not zero, put it into the column of $\boldsymbol{C}$;
2. If column 2 of $\boldsymbol{A}$ is not a multiple of column 1, put it into the column of $\boldsymbol{C}$;
3. If column 3 of $\boldsymbol{A}$ is not a combination of columns 1 and 2, put it into the column of $\boldsymbol{C}$;
4. Continue this process until we find r linearly independent columns (or all the linearly independent columns if we do not know the rank r beforehand).

When we have the r linearly independent columns from $\boldsymbol{A}$, we can prove the existence of CR decomposition by the column space view of matrix multiplication.

Column space view of matrix multiplication A multiplication of two matrices $\boldsymbol{D} \in \mathbb{R}^{m\times k}, \boldsymbol{E} \in \mathbb{R}^{k\times n}$ is $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{E} = \boldsymbol{D}[\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n] = [\boldsymbol{D}\boldsymbol{e}_1, \boldsymbol{D}\boldsymbol{e}_2, \ldots, \boldsymbol{D}\boldsymbol{e}_n]$, i.e., each column of $\boldsymbol{A}$ is a combination of columns from $\boldsymbol{D}$.

Proof [of Theorem 5.1] As the rank of matrix $\boldsymbol{A}$ is r and $\boldsymbol{C}$ contains r linearly independent columns from $\boldsymbol{A}$, the column space of $\boldsymbol{C}$ is equivalent to the column space of $\boldsymbol{A}$. If we take any other column $\boldsymbol{a}_i$ of $\boldsymbol{A}$, $\boldsymbol{a}_i$ can be represented as a linear combination of the columns of

$\boldsymbol{C}$, i.e., there exists a vector $\boldsymbol{r}_i$ such that $\boldsymbol{a}_i = \boldsymbol{C}\boldsymbol{r}_i$, $\forall i \in \{1, 2, \ldots, n\}$. Put these $\boldsymbol{r}_i$'s into the columns of matrix $\boldsymbol{R}$, we obtain

$$\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n] = [\boldsymbol{C}\boldsymbol{r}_1, \boldsymbol{C}\boldsymbol{r}_2, \ldots, \boldsymbol{C}\boldsymbol{r}_n] = \boldsymbol{C}\boldsymbol{R},$$

from which the result follows. ■

5.2 Reduced Row Echelon Form (RREF)

In Gaussian elimination Section 1.1, we introduced the elimination matrix (a lower triangular matrix) and permutation matrix to transform $\boldsymbol{A}$ into an upper triangular form. We rewrite the Gaussian elimination for a 4×4 square matrix, where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed:

Gaussian Elimination for a Square Matrix

$$\underset{\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}} \xrightarrow{\boldsymbol{E}_1} \underset{\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}} \xrightarrow{\boldsymbol{P}_1} \underset{\boldsymbol{P}_1\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \mathbf{0} & {\color{blue}\boldsymbol{\boxtimes}} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \mathbf{0} & {\color{blue}\boldsymbol{\boxtimes}} & \boldsymbol{\boxtimes} \\ 0 & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}} \xrightarrow{\boldsymbol{E}_2} \underset{\boldsymbol{E}_2\boldsymbol{P}_1\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & {\color{blue}\boxtimes} & \boxtimes & \boxtimes \\ 0 & 0 & {\color{blue}\boxtimes} & \boxtimes \\ 0 & \mathbf{0} & \mathbf{0} & {\color{blue}\boldsymbol{\boxtimes}} \end{bmatrix}}.$$

Furthermore, the Gaussian elimination can also be applied on a rectangular matrix, we give an example for a 4×5 matrix as follows:

Gaussian Elimination for a Rectangular Matrix

$$\underset{\boldsymbol{A}}{\begin{bmatrix} {\color{blue}2} & \boxtimes & 10 & 9 & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \end{bmatrix}} \xrightarrow{\boldsymbol{E}_1} \underset{\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} {\color{blue}2} & \boxtimes & 10 & 9 & \boxtimes \\ \mathbf{0} & \mathbf{0} & {\color{blue}\mathbf{5}} & \mathbf{6} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \mathbf{0} & \mathbf{2} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \end{bmatrix}} \xrightarrow{\boldsymbol{E}_2} \underset{\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} {\color{blue}2} & \boxtimes & 10 & 9 & \boxtimes \\ 0 & 0 & {\color{blue}5} & 6 & \boxtimes \\ 0 & 0 & \mathbf{0} & {\color{blue}\mathbf{3}} & \boldsymbol{\boxtimes} \\ 0 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}},$$

where the blue-colored numbers are **pivots** as we defined previously and we call the last matrix above **row echelon form**. Note that we get the 4-th row as a zero row in this specific example. Going further, if we subtract each row by a multiple of the next row to make the entries above the pivots to be zero:

Reduced Row Echelon Form: Get Zero Above Pivots

$$\underset{\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} {\color{blue}2} & \boxtimes & 10 & 9 & \boxtimes \\ 0 & 0 & {\color{blue}5} & 6 & \boxtimes \\ 0 & 0 & 0 & {\color{blue}3} & \boxtimes \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}} \xrightarrow{\boldsymbol{E}_3} \underset{\boldsymbol{E}_3\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} {\color{blue}2} & \boxtimes & \mathbf{0} & \mathbf{-3} & \boldsymbol{\boxtimes} \\ 0 & 0 & {\color{blue}5} & 6 & \boxtimes \\ 0 & 0 & 0 & {\color{blue}3} & \boxtimes \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}} \xrightarrow{\boldsymbol{E}_4} \underset{\boldsymbol{E}_4\boldsymbol{E}_3\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix} {\color{blue}2} & \boxtimes & 0 & \mathbf{0} & \boldsymbol{\boxtimes} \\ 0 & 0 & {\color{blue}5} & \mathbf{0} & \boldsymbol{\boxtimes} \\ 0 & 0 & 0 & {\color{blue}3} & \boxtimes \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}},$$

where $\boldsymbol{E}_3$ subtracts 2 times the 2-nd row from the 1-st row, and $\boldsymbol{E}_4$ adds the 3-rd row to the 1-st row and subtracts 2 times the 3-rd row from the 2-nd row. Finally, we get the full row reduced echelon form by making the pivots to be 1:

Reduced Row Echelon Form: Make The Pivots To Be 1

$$
\underset{\boldsymbol{E}_4\boldsymbol{E}_3\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix}2 & \boxtimes & 0 & 0 & \boxtimes\\ 0 & 0 & 5 & 0 & \boxtimes\\ 0 & 0 & 0 & 3 & \boxtimes\\ 0 & 0 & 0 & 0 & 0\end{bmatrix}} \xrightarrow{\boldsymbol{E}_5} \underset{\boldsymbol{E}_5\boldsymbol{E}_4\boldsymbol{E}_3\boldsymbol{E}_2\boldsymbol{E}_1\boldsymbol{A}}{\begin{bmatrix}\mathbf{1} & \boxtimes & \mathbf{0} & \mathbf{0} & \boxtimes\\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \boxtimes\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \boxtimes\\ 0 & 0 & 0 & 0 & 0\end{bmatrix}},
$$

where $\boldsymbol{E}_5$ makes the pivots to be 1. Note here, the transformation matrix $\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_5$ are not necessarily to be lower triangular matrices as they are in LU decomposition. They can also be permutation matrices or other matrices. We call this final matrix the **reduced row echelon form** of $\boldsymbol{A}$ where it has 1's as pivots and zeros above the pivots.

Lemma 5.2: (Rank and Pivots)

The rank of $\boldsymbol{A}$ is equal to the number of pivots.

Lemma 5.3: (RREF in CR)

The reduced row echelon form of the matrix $\boldsymbol{A}$ without zero rows is the matrix $\boldsymbol{R}$ in the CR decomposition.

In short, we first compute the reduced row echelon form of matrix $\boldsymbol{A}$ by $rref(\boldsymbol{A})$, Then $\boldsymbol{C}$ is obtained by removing from $\boldsymbol{A}$ all the non-pivot columns (which can be determined by looking for columns in $rref(\boldsymbol{A})$ which do not contain a pivot). And $\boldsymbol{R}$ is obtained by eliminating zero rows of $rref(\boldsymbol{A})$. And this is actually a special case of **rank decomposition** of matrix $\boldsymbol{A}$. However, CR decomposition is so special that it involves the reduced row echelon form so that we introduce it here particularly.

$\boldsymbol{R}$ has a remarkable form whose r columns containing the pivots form an $r \times r$ identity matrix. Note again that we can just remove the zero rows from the row reduced echelon form to obtain this matrix $\boldsymbol{R}$. In (Strang, 2021), the authors give a specific notation for the row reduced echelon form without removing the zero rows as $\boldsymbol{R}_0$:

$$
\boldsymbol{R}_0 = rref(\boldsymbol{A}) = \begin{bmatrix}\boldsymbol{R}\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{F}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\boldsymbol{P}, \text{[16]}
$$

where the $n \times n$ permutation matrix $\boldsymbol{P}$ puts the columns of $r \times r$ identity matrix $\boldsymbol{I}_r$ into the correct positions, matching the first r linearly independent columns of the original matrix $\boldsymbol{A}$.

The CR decomposition reveals a great theorem of linear algebra that the row rank equals the column rank of any matrix.

16. Permutation matrix $\boldsymbol{P}$ in the right side of a matrix is to permute the column of that matrix.

Proof [**of Theorem 0.13, A Third Way**] For CR decomposition of matrix $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{R}$, we have $\boldsymbol{R} = [\boldsymbol{I}_r, \boldsymbol{F}]\boldsymbol{P}$, where $\boldsymbol{P}$ is an $n \times n$ permutation to put the columns of the $r \times r$ identity matrix $\boldsymbol{I}_r$ into the correct positions as shown above. It can be easily verified that the r rows of $\boldsymbol{R}$ are linearly independent of the submatrix of $\boldsymbol{I}_r$ (since $\boldsymbol{I}_r$ is nonsingular) such that the row rank of $\boldsymbol{R}$ is r.

Firstly, from the definition of the CR decomposition, the r columns of $\boldsymbol{C}$ are from r linearly independent columns of $\boldsymbol{A}$, the column rank of $\boldsymbol{A}$ is r. Further,

• Since $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{R}$, all rows of $\boldsymbol{A}$ are combinations of the rows of $\boldsymbol{R}$. That is, the row rank of $\boldsymbol{A}$ is no larger than the row rank of $\boldsymbol{R}$;

• From $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{R}$, we also have $(\boldsymbol{C}^\top\boldsymbol{C})^{-1}\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{R} = (\boldsymbol{C}^\top\boldsymbol{C})^{-1}\boldsymbol{C}^\top\boldsymbol{A}$, that is $\boldsymbol{R} = (\boldsymbol{C}^\top\boldsymbol{C})^{-1}\boldsymbol{C}^\top\boldsymbol{A}$. $\boldsymbol{C}^\top\boldsymbol{C}$ is nonsingular since it has full column rank r. Then all rows of $\boldsymbol{R}$ are also combinations of the rows of $\boldsymbol{A}$. That is, the row rank of $\boldsymbol{R}$ is no larger than the row rank of $\boldsymbol{A}$;

• By "sandwiching", the row rank of $\boldsymbol{A}$ is equal to the row rank of $\boldsymbol{R}$ which is r.

Therefore, both the row rank and column rank of $\boldsymbol{A}$ are equal to r from which the result follows. ■

5.3 Rank Decomposition

We previously mentioned that the CR decomposition is a special case of rank decomposition. Formally, we prove the existence of the rank decomposition rigorously in the following theorem.

Theorem 5.4: (Rank Decomposition)

Any rank-r matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ can be factored as

$$\underset{m\times n}{\boldsymbol{A}} = \underset{m\times r}{\boldsymbol{D}} \quad \underset{r\times n}{\boldsymbol{F}},$$

where $\boldsymbol{D} \in \mathbb{R}^{m\times r}$ has rank r, and $\boldsymbol{F} \in \mathbb{R}^{r\times n}$ also has rank r, i.e., $\boldsymbol{D}, \boldsymbol{F}$ have full rank r.

The storage for the decomposition is then reduced or potentially increased from mn to $r(m+n)$.

Proof [of Theorem 5.4] By ULV decomposition in Theorem 4.1 (p. 71), we can decompose $\boldsymbol{A}$ by

$$\boldsymbol{A} = \boldsymbol{U}\begin{bmatrix}\boldsymbol{L} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{V}.$$

Let $\boldsymbol{U}_0 = \boldsymbol{U}_{:,1:r}$ and $\boldsymbol{V}_0 = \boldsymbol{V}_{1:r,:}$, i.e., $\boldsymbol{U}_0$ contains only the first r columns of $\boldsymbol{U}$, and $\boldsymbol{V}_0$ contains only the first r rows of $\boldsymbol{V}$. Then, we still have $\boldsymbol{A} = \boldsymbol{U}_0\boldsymbol{L}\boldsymbol{V}_0$ where $\boldsymbol{U}_0 \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V}_0 \in \mathbb{R}^{r\times n}$. This is also known as the reduced ULV decomposition. Let $\{\boldsymbol{D} = \boldsymbol{U}_0\boldsymbol{L}$ and $\boldsymbol{F} = \boldsymbol{V}_0\}$, or $\{\boldsymbol{D} = \boldsymbol{U}_0$ and $\boldsymbol{F} = \boldsymbol{L}\boldsymbol{V}_0\}$, we find such rank decomposition. ■

The rank decomposition is not unique. Even by elementary transformations, we have

$$\boldsymbol{A} = \boldsymbol{E}_1\begin{bmatrix}\boldsymbol{Z} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{E}_2,$$

where $\boldsymbol{E}_1 \in \mathbb{R}^{m\times m}$, $\boldsymbol{E}_2 \in \mathbb{R}^{n\times n}$ represent elementary row and column operations, $\boldsymbol{Z} \in \mathbb{R}^{r\times r}$. The transformation is rather general, and there are dozens of these $\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{Z}$. Similar construction on this decomposition as shown in the above proof, we can recover another rank decomposition.

Analogously, we can find such $\boldsymbol{D}, \boldsymbol{F}$ by SVD, URV, CR, CUR, and many other decompositional algorithms. However, we may connect the different rank decompositions by the following lemma.

Lemma 5.5: (Connection Between Rank Decompositions)

For any two rank decompositions of $\boldsymbol{A} = \boldsymbol{D}_1\boldsymbol{F}_1 = \boldsymbol{D}_2\boldsymbol{F}_2$, there exists a nonsingular matrix $\boldsymbol{P}$ such that

$$\boldsymbol{D}_1 = \boldsymbol{D}_2\boldsymbol{P} \qquad \text{and} \qquad \boldsymbol{F}_1 = \boldsymbol{P}^{-1}\boldsymbol{F}_2.$$

Proof [of Lemma 5.5] Since $\boldsymbol{D}_1\boldsymbol{F}_1 = \boldsymbol{D}_2\boldsymbol{F}_2$, we have $\boldsymbol{D}_1\boldsymbol{F}_1\boldsymbol{F}_1^\top = \boldsymbol{D}_2\boldsymbol{F}_2\boldsymbol{F}_1^\top$. It is trivial that $rank(\boldsymbol{F}_1\boldsymbol{F}_1^\top) = rank(\boldsymbol{F}_1) = r$ such that $\boldsymbol{F}_1\boldsymbol{F}_1^\top$ is a square matrix with full rank and thus is nonsingular. This implies $\boldsymbol{D}_1 = \boldsymbol{D}_2\boldsymbol{F}_2\boldsymbol{F}_1^\top(\boldsymbol{F}_1\boldsymbol{F}_1^\top)^{-1}$. Let $\boldsymbol{P} = \boldsymbol{F}_2\boldsymbol{F}_1^\top(\boldsymbol{F}_1\boldsymbol{F}_1^\top)^{-1}$, we have $\boldsymbol{D}_1 = \boldsymbol{D}_2\boldsymbol{P}$ and $\boldsymbol{F}_1 = \boldsymbol{P}^{-1}\boldsymbol{F}_2$. ■

5.4 Application: Rank and Trace of an Idempotent Matrix

The CR decomposition is quite useful to prove the rank of an idempotent matrix. See also how it works in the orthogonal projection in (Lu, 2021c,e).

Lemma 5.6: (Rank and Trace of an Idempotent Matrix)

For any $n \times n$ idempotent matrix $\boldsymbol{A}$ (i.e., $\boldsymbol{A}^2 = \boldsymbol{A}$), the rank of $\boldsymbol{A}$ equals the trace of $\boldsymbol{A}$.

Proof [of Lemma 5.6] Any $n \times n$ rank-r matrix $\boldsymbol{A}$ has CR decomposition $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{R}$, where $\boldsymbol{C} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{R} \in \mathbb{R}^{r\times n}$ with $\boldsymbol{C}, \boldsymbol{R}$ having full rank r. Then,

$$\begin{aligned}
\boldsymbol{A}^2 &= \boldsymbol{A}, \\
\boldsymbol{C}\boldsymbol{R}\boldsymbol{C}\boldsymbol{R} &= \boldsymbol{C}\boldsymbol{R}, \\
\boldsymbol{R}\boldsymbol{C}\boldsymbol{R} &= \boldsymbol{R}, \\
\boldsymbol{R}\boldsymbol{C} &= \boldsymbol{I}_r,
\end{aligned}$$

where $\boldsymbol{I}_r$ is an $r \times r$ identity matrix. Thus

$$trace(\boldsymbol{A}) = trace(\boldsymbol{C}\boldsymbol{R}) = trace(\boldsymbol{R}\boldsymbol{C}) = trace(\boldsymbol{I}_r) = r,$$

which equals the rank of $\boldsymbol{A}$. The equality above is from the invariant of cyclic permutation of trace. ■

6. Skeleton/CUR Decomposition

Theorem 6.1: (Skeleton Decomposition)

Any rank-r matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ can be factored as

$$\underset{m\times n}{\boldsymbol{A}} = \underset{m\times r}{\boldsymbol{C}} \quad \underset{r\times r}{\boldsymbol{U}^{-1}} \quad \underset{r\times n}{\boldsymbol{R}},$$

where $\boldsymbol{C}$ is some r linearly independent columns of $\boldsymbol{A}$, $\boldsymbol{R}$ is some r linearly independent rows of $\boldsymbol{A}$ and $\boldsymbol{U}$ is the nonsingular submatrix on the intersection.

- The storage for the decomposition is then reduced or potentially increased from mn floats to $r(m+n)+r^2$ floats.
- Or further, if we only record the position of the indices, it requires mr, nr floats for storing $\boldsymbol{C}, \boldsymbol{R}$ respectively and extra $2r$ integers to remember the position of each column of $\boldsymbol{C}$ in that of $\boldsymbol{A}$ and each row of $\boldsymbol{R}$ in that of $\boldsymbol{A}$ (i.e., construct $\boldsymbol{U}$ from $\boldsymbol{C}, \boldsymbol{R}$).

Skeleton decomposition is also known as the *CUR decomposition* follows from the notation in the decomposition. The illustration of skeleton decomposition is shown in Figure 12 where the yellow vectors denote the linearly independent columns of $\boldsymbol{A}$ and green vectors denote the linearly independent rows of $\boldsymbol{A}$. In case $\boldsymbol{A}$ is square and invertible, we have skeleton decomposition $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{U}^{-1}\boldsymbol{R}$ where $\boldsymbol{C} = \boldsymbol{R} = \boldsymbol{U} = \boldsymbol{A}$ such that the decomposition reduces to $\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{A}$. Specifically, if I, J index vectors both with size r that contain the indices of rows and columns selected from $\boldsymbol{A}$ into $\boldsymbol{R}$ and $\boldsymbol{C}$ respectively, $\boldsymbol{U}$ can be denoted as $\boldsymbol{U} = \boldsymbol{A}[I, J]$.

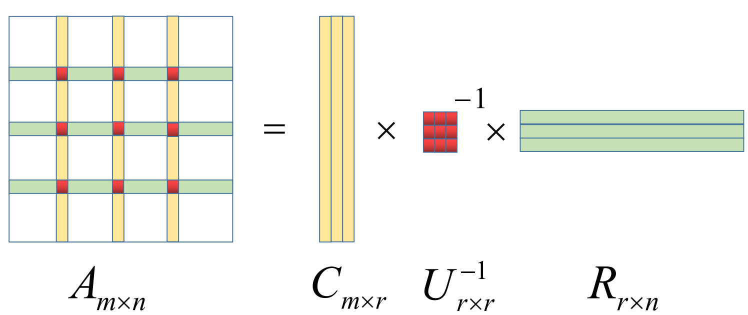


Figure 12: Demonstration of skeleton decomposition of a matrix.

6.1 Existence of the Skeleton Decomposition

In Corollary 0.13, we proved the row rank and the column rank of a matrix are equal. In another word, we can also claim that the dimension of the column space and the dimension of the row space are equal. This property is essential for the existence of the skeleton decomposition.

We are then ready to prove the existence of the skeleton decomposition. The proof is rather elementary.

Proof [of Theorem 6.1] The proof relies on the existence of such nonsingular matrix $\boldsymbol{U}$ which is central to this decomposition method.

Existence of such nonsingular matrix $\boldsymbol{U}$ Since matrix $\boldsymbol{A}$ is rank-r, we can pick r columns from $\boldsymbol{A}$ so that they are linearly independent. Suppose we put the specific r independent columns $\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2}, \ldots, \boldsymbol{a}_{ir}$ into the columns of an $m\times r$ matrix $\boldsymbol{N} = [\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2}, \ldots, \boldsymbol{a}_{ir}] \in \mathbb{R}^{m\times r}$. The dimension of the column space of $\boldsymbol{N}$ is r so that the dimension of the row space of $\boldsymbol{N}$ is also r by Corollary 0.13. Again, we can pick r linearly independent rows $\boldsymbol{n}_{j1}^\top, \boldsymbol{n}_{j2}^\top, \ldots, \boldsymbol{n}_{jr}^\top$ from $\boldsymbol{N}$ and put the specific r rows into rows of an $r \times r$ matrix $\boldsymbol{U} = [\boldsymbol{n}_{j1}^\top; \boldsymbol{n}_{j2}^\top; \ldots; \boldsymbol{n}_{jr}^\top] \in \mathbb{R}^{r\times r}$. Using Corollary 0.13 again, the dimension of the column space of $\boldsymbol{U}$ is also r which means there are the r linearly independent columns from $\boldsymbol{U}$. So $\boldsymbol{U}$ is such a nonsingular matrix with size $r \times r$.

Main proof As long as we find the nonsingular $r \times r$ matrix $\boldsymbol{U}$ inside $\boldsymbol{A}$, we can find the existence of the skeleton decomposition as follows.

Suppose $\boldsymbol{U} = \boldsymbol{A}[I, J]$ where I, J are index vectors of size r. Since $\boldsymbol{U}$ is a nonsingular matrix, the columns of $\boldsymbol{U}$ are linearly independent. Thus the columns of matrix $\boldsymbol{C}$ based on the columns of $\boldsymbol{U}$ are also linearly independent (i.e., select the r columns of $\boldsymbol{A}$ with the same entries of the matrix $\boldsymbol{U}$. Here $\boldsymbol{C}$ is equal to the $\boldsymbol{N}$ we construct above and $\boldsymbol{C} = \boldsymbol{A}[:, J]$).

As the rank of the matrix $\boldsymbol{A}$ is r, if we take any other column $\boldsymbol{a}_i$ of $\boldsymbol{A}$, $\boldsymbol{a}_i$ can be represented as a linear combination of the columns of $\boldsymbol{C}$, i.e., there exists a vector $\boldsymbol{x}$ such that $\boldsymbol{a}_i = \boldsymbol{C}\boldsymbol{x}$, for all $i \in \{1, 2, \ldots, n\}$. Let r rows of $\boldsymbol{a}_i$ corresponding to the row entries of $\boldsymbol{U}$ be $\boldsymbol{r}_i \in \mathbb{R}^r$ for all $i \in \{1, 2, \ldots, n\}$ (i.e., $\boldsymbol{r}_i$ contains r entries of $\boldsymbol{a}_i$). That is, select the r entries of $\boldsymbol{a}_i$'s corresponding to the entries of $\boldsymbol{U}$ as follows:

$$\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m\times n} \qquad \longrightarrow \qquad \boldsymbol{A}[I, :] = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n] \in \mathbb{R}^{r\times n}.$$

Since $\boldsymbol{a}_i = \boldsymbol{C}\boldsymbol{x}$, $\boldsymbol{U}$ is a submatrix inside $\boldsymbol{C}$, and $\boldsymbol{r}_i$ is a subvector inside $\boldsymbol{a}_i$, we have $\boldsymbol{r}_i = \boldsymbol{U}\boldsymbol{x}$ which is equivalent to $\boldsymbol{x} = \boldsymbol{U}^{-1}\boldsymbol{r}_i$. Thus for every i, we have $\boldsymbol{a}_i = \boldsymbol{C}\boldsymbol{U}^{-1}\boldsymbol{r}_i$. Combine the n columns of such $\boldsymbol{r}_i$ into $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n]$, we obtain

$$\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n] = \boldsymbol{C}\boldsymbol{U}^{-1}\boldsymbol{R},$$

from which the result follows.

In short, we first find r linearly independent columns of $\boldsymbol{A}$ into $\boldsymbol{C} \in \mathbb{R}^{m\times r}$. From $\boldsymbol{C}$, we find an $r \times r$ nonsingular submatrix $\boldsymbol{U}$. The r rows of $\boldsymbol{A}$ corresponding to entries of $\boldsymbol{U}$ can help to reconstruct the columns of $\boldsymbol{A}$. Again, the situation is shown in Figure 12. ■

CR decomposition vs skeleton decomposition We note that CR decomposition and skeleton decomposition share a similar form. Even for the symbols used $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{R}$ for the CR decomposition and $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{U}^{-1}\boldsymbol{R}$ for the skeleton decomposition.

Both in the CR decomposition and the skeleton decomposition, we **can** select the first r independent columns to obtain the matrix $\boldsymbol{C}$ (the symbol for both the CR decomposition and the skeleton decomposition). So $\boldsymbol{C}$'s in the CR decomposition and the skeleton decomposition are exactly the same. On the contrary, $\boldsymbol{R}$ in the CR decomposition is the reduced row echelon form without the zero rows, whereas $\boldsymbol{R}$ in the skeleton decomposition is exactly some rows from $\boldsymbol{A}$ so that $\boldsymbol{R}$'s have different meanings in the two decompositional methods.

A word on the uniqueness of CR decomposition and skeleton decomposition As mentioned above, both in the CR decomposition and the skeleton decomposition, we select the first r linearly independent columns to obtain the matrix $\boldsymbol{C}$. In this sense, the CR and skeleton decompositions have a unique form. However, if we select the last r linearly independent columns, we will get a different CR decomposition or skeleton decomposition. We will not discuss this situation here as it is not the main interest of this text.

To repeat, in the above proof for the existence of the skeleton decomposition, we first find the r linearly independent columns of $\boldsymbol{A}$ into the matrix $\boldsymbol{C}$. From $\boldsymbol{C}$, we find an $r \times r$ nonsingular submatrix $\boldsymbol{U}$. From the submatrix $\boldsymbol{U}$, we finally find the final row submatrix $\boldsymbol{R} \in \mathbb{R}^{r\times n}$. A further question can be posed that if matrix $\boldsymbol{A}$ has rank r, matrix $\boldsymbol{C}$ contains r linearly independent columns, and matrix $\boldsymbol{R}$ contains r linearly independent rows, then whether the $r \times r$ "intersection" of $\boldsymbol{C}$ and $\boldsymbol{R}$ is invertible or not [17].

Corollary 6.2: (Nonsingular Intersection)

If matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has rank r, matrix $\boldsymbol{C}$ contains r linearly independent columns, and matrix $\boldsymbol{R}$ contains r linearly independent rows, then the $r \times r$ "intersection" matrix $\boldsymbol{U}$ of $\boldsymbol{C}$ and $\boldsymbol{R}$ is invertible.

Proof [of Corollary 6.2] If I, J are the indices of rows and columns selected from $\boldsymbol{A}$ into $\boldsymbol{R}$ and $\boldsymbol{C}$ respectively, then, $\boldsymbol{R}$ can be denoted as $\boldsymbol{R} = \boldsymbol{A}[I,:]$, $\boldsymbol{C}$ can be represented as $\boldsymbol{C} = \boldsymbol{A}[:,J]$, and $\boldsymbol{U}$ can be denoted as $\boldsymbol{U} = \boldsymbol{A}[I,J]$.

Since $\boldsymbol{C}$ contains r linearly independent columns of $\boldsymbol{A}$, any column $\boldsymbol{a}_i$ of $\boldsymbol{A}$ can be represented as $\boldsymbol{a}_i = \boldsymbol{C}\boldsymbol{x}_i = \boldsymbol{A}[:,J]\boldsymbol{x}_i$ for all $i \in \{1, 2, \ldots, n\}$. This implies the r entries of $\boldsymbol{a}_i$ corresponding to the I indices can be represented by the columns of $\boldsymbol{U}$ such that $\boldsymbol{a}_i[I] = \boldsymbol{U}\boldsymbol{x}_i \in \mathbb{R}^r$ for all $i \in \{1, 2, \ldots, n\}$, i.e.,

$$\boldsymbol{a}_i = \boldsymbol{C}\boldsymbol{x}_i = \boldsymbol{A}[:,J]\boldsymbol{x}_i \in \mathbb{R}^m \qquad \longrightarrow \qquad \boldsymbol{a}_i[I] = \boldsymbol{A}[I,J]\boldsymbol{x}_i = \boldsymbol{U}\boldsymbol{x}_i \in \mathbb{R}^r.$$

Since $\boldsymbol{R}$ contains r linearly independent rows of $\boldsymbol{A}$, the row rank and column rank of $\boldsymbol{R}$ are equal to r. Combining the facts above, the r columns of $\boldsymbol{R}$ corresponding to indices J (i.e., the r columns of $\boldsymbol{U}$) are linearly independent.

Again, by applying Corollary 0.13, the dimension of the row space of $\boldsymbol{U}$ is also equal to r which means there are the r linearly independent rows from $\boldsymbol{U}$, and $\boldsymbol{U}$ is invertible. ■

7. Interpolative Decomposition (ID)

Column interpolative decomposition (ID) factors a matrix as the product of two matrices, one of which contains selected columns from the original matrix, and the other of which has a subset of columns consisting of the identity matrix and all its values are no greater than 1 in absolute value. Formally, we have the following theorem describing the details of the column ID.

17. We thank Gilbert Strang for raising this interesting question.

Theorem 7.1: (Column Interpolative Decomposition)
Any rank-r matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ can be factored as

$$\underset{m\times n}{\boldsymbol{A}} = \underset{m\times r}{\boldsymbol{C}} \quad \underset{r\times n}{\boldsymbol{W}},$$

where $\boldsymbol{C} \in \mathbb{R}^{m\times r}$ is some r linearly independent columns of $\boldsymbol{A}$, $\boldsymbol{W} \in \mathbb{R}^{r\times n}$ is the matrix to reconstruct $\boldsymbol{A}$ which contains an $r \times r$ identity submatrix (under a mild column permutation). Specifically entries in $\boldsymbol{W}$ have values no larger than 1 in magnitude:

$$\max |w_{ij}| \leq 1, \ \forall \ i \in [1, r], j \in [1, n].$$

The storage for the decomposition is then reduced or potentially increased from mn floats to mr, $(n-r)r$ floats for storing $\boldsymbol{C}, \boldsymbol{W}$ respectively and extra r integers are required to remember the position of each column of $\boldsymbol{C}$ in that of $\boldsymbol{A}$.

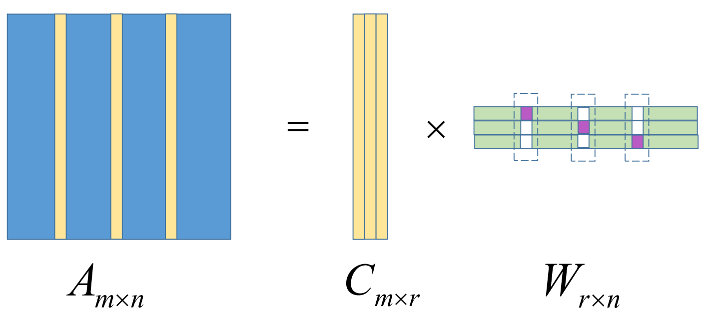


Figure 13: Demonstration of the column ID of a matrix where the yellow vector denotes the linearly independent columns of $\boldsymbol{A}$, white entries denote zero, and purple entries denote one.

The illustration of the column ID is shown in Figure 13 where the yellow vectors denote the linearly independent columns of $\boldsymbol{A}$ and the purple vectors in $\boldsymbol{W}$ form an $r \times r$ identity submatrix. The positions of the purple vectors inside $\boldsymbol{W}$ are exactly the same as the positions of the corresponding yellow vectors inside $\boldsymbol{A}$. The column ID is very similar to the CR decomposition (Theorem 5.1, p. 76), both select r linearly independent columns into the first factor and the second factor contains an $r \times r$ identity submatrix. The difference is in that the CR decomposition will exactly choose the first r linearly independent columns into the first factor and the identity submatrix appears in the pivots (Definition 1.7, p. 19). And more importantly, the second factor in the CR decomposition comes from the RREF (Lemma 5.3, p. 78). Therefore, the column ID can also be utilized in the applications of the CR decomposition, say proving the fact of rank equals trace in idempotent matrices (Lemma 5.6, p. 80), and proving the elementary theorem in linear algebra that column rank equals row rank of a matrix (Corollary 0.13, p. 12). Moreover, the column ID is also a special case of rank decomposition (Theorem 5.4, p. 79) and is apparently not unique. The connection between different column IDs is given by Lemma 5.5 (p. 80).

Notations that will be extensively used in the sequel Following again the Matlab-style notation, if J_s is an index vector with size r that contains the indices of columns selected from $\boldsymbol{A}$ into $\boldsymbol{C}$, then $\boldsymbol{C}$ can be denoted as $\boldsymbol{C} = \boldsymbol{A}[:, J_s]$. The matrix $\boldsymbol{C}$ contains "skeleton" columns of $\boldsymbol{A}$, hence the subscript s in J_s. From the "skeleton" index vector J_s, the $r \times r$ identity matrix inside $\boldsymbol{W}$ can be recovered by

$$\boldsymbol{W}[:, J_s] = \boldsymbol{I}_r \in \mathbb{R}^{r\times r}.$$

Suppose further we put the remaining indices of $\boldsymbol{A}$ into an index vector J_r where

$$J_s \cap J_r = \varnothing \qquad \text{and} \qquad J_s \cup J_r = \{1, 2, \ldots, n\}.$$

The remaining $n-r$ columns in $\boldsymbol{W}$ consists of an $r \times (n-r)$ *expansion matrix* since the matrix contains *expansion coefficients* to reconstruct the columns of $\boldsymbol{A}$ from $\boldsymbol{C}$:

$$\boldsymbol{E} = \boldsymbol{W}[:, J_r] \in \mathbb{R}^{r\times(n-r)},$$

where the entries of $\boldsymbol{E}$ are known as the *expansion coefficients*. Moreover, let $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ be a (column) permutation matrix (Definition 0.17, p. 15) defined by $\boldsymbol{P} = \boldsymbol{I}_n[:, (J_s, J_r)]$ so that

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{A}[:, (J_s, J_r)] = [\boldsymbol{C}, \boldsymbol{A}[:, J_r]],$$

and

$$\boldsymbol{W}\boldsymbol{P} = \boldsymbol{W}[:, (J_s, J_r)] = [\boldsymbol{I}_r, \boldsymbol{E}] \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{W} = [\boldsymbol{I}_r, \boldsymbol{E}]\, \boldsymbol{P}^\top. \tag{7.1}$$

7.1 Existence of the Column Interpolative Decomposition

Cramer's rule The proof of the existence of the column ID relies on the Cramer's rule that we shall shortly discuss here. Consider a system of n linear equations for n unknowns, represented in matrix multiplication form as follows :

$$\boldsymbol{M}\boldsymbol{x} = \boldsymbol{l},$$

where $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is nonsingular and $\boldsymbol{x}, \boldsymbol{l} \in \mathbb{R}^n$. Then the theorem states that in this case, the system has a unique solution, whose individual values for the unknowns are given by:

$$x_i = \frac{\det(\boldsymbol{M}_i)}{\det(\boldsymbol{M})}, \qquad \text{for all} \quad i \in \{1, 2, \ldots, n\},$$

where $\boldsymbol{M}_i$ is the matrix formed by replacing the i-th column of $\boldsymbol{M}$ with the column vector $\boldsymbol{l}$. In full generality, the Cramer's rule considers the matrix equation

$$\boldsymbol{M}\boldsymbol{X} = \boldsymbol{L},$$

where $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is nonsingular and $\boldsymbol{X}, \boldsymbol{L} \in \mathbb{R}^{n\times m}$. Let $I = [i_1, i_2, \ldots, i_k]$ and $J = [j_1, j_2, \ldots, j_k]$ be two index vectors where $1 \leq i_1 \leq i_2 \leq \ldots \leq i_k \leq n$ and $1 \leq j_1 \leq j_2 \leq \ldots \leq j_k \leq n$. Then $\boldsymbol{X}[I, J]$ is a $k \times k$ submatrix of $\boldsymbol{X}$. Let further $\boldsymbol{M}_{\boldsymbol{L}}(I, J)$ be the $n \times n$ matrix formed by replacing the i_s column of $\boldsymbol{M}$ by j_s column of $\boldsymbol{L}$ for all $s \in \{1, 2, \ldots, k\}$. Then

$$\det(\boldsymbol{X}[I, J]) = \frac{\det\left(\boldsymbol{M}_{\boldsymbol{L}}(I, J)\right)}{\det(\boldsymbol{M})}.$$

When I, J are of size 1, it follows that

$$x_{ij} = \frac{\det(\boldsymbol{M_L}(i,j))}{\det(\boldsymbol{M})}. \tag{7.2}$$

Now we are ready to prove the existence of the column ID.

Proof [of Theorem 7.1] We have mentioned above the proof relies on the Cramer's rule. If we can show the entries of $\boldsymbol{W}$ can be denoted by the Cramer's rule equality in Equation (7.2) and the numerator is smaller than the denominator, then we can complete the proof. However, we notice that the matrix in the denominator of Equation (7.2) is a square matrix. Here comes the trick.

Step 1: column ID for full row rank matrix For a start, we first consider the full row rank matrix $\boldsymbol{A}$ (which implies $r = m$, $m \leq n$, and $\boldsymbol{A} \in \mathbb{R}^{r\times n}$ such that the matrix $\boldsymbol{C} \in \mathbb{R}^{r\times r}$ is a square matrix in the column ID $\boldsymbol{A} = \boldsymbol{CW}$ that we want). Determine the "skeleton" index vector J_s by

$$\boxed{J_s = \arg\max_J \{|\det(\boldsymbol{A}[:,J])| : J \text{ is a subset of } \{1,2,\ldots,n\} \text{ with size } r=m\},} \tag{7.3}$$

i.e., J_s is the index vector that is determined by maximizing the magnitude of the determinant of $\boldsymbol{A}[:,J]$. As we have discussed in the last section, there exists a (column) permutation matrix such that

$$\boldsymbol{AP} = \begin{bmatrix}\boldsymbol{A}[:,J_s] & \boldsymbol{A}[:,J_r]\end{bmatrix}.$$

Since $\boldsymbol{C} = \boldsymbol{A}[:,J_s]$ has full column rank $r = m$, it is then nonsingular. The above equation can be rewritten as

$$\begin{aligned}
\boldsymbol{A} &= \begin{bmatrix}\boldsymbol{A}[:,J_s] & \boldsymbol{A}[:,J_r]\end{bmatrix}\boldsymbol{P}^\top \\
&= \boldsymbol{A}[:,J_s]\begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{A}[:,J_s]^{-1}\boldsymbol{A}[:,J_r]\end{bmatrix}\boldsymbol{P}^\top \\
&= \boldsymbol{C}\underbrace{\begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{C}^{-1}\boldsymbol{A}[:,J_r]\end{bmatrix}\boldsymbol{P}^\top}_{\boldsymbol{W}}
\end{aligned},$$

where the matrix $\boldsymbol{W}$ is given by $\begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{C}^{-1}\boldsymbol{A}[:,J_r]\end{bmatrix}\boldsymbol{P}^\top = \begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{E}\end{bmatrix}\boldsymbol{P}^\top$ by Equation (7.1). To prove the claim that the magnitude of $\boldsymbol{W}$ is no larger than 1 is equivalent to proving that entries in $\boldsymbol{E} = \boldsymbol{C}^{-1}\boldsymbol{A}[:,J_r] \in \mathbb{R}^{r\times(n-r)}$ are no greater than 1 in absolute value.

Define the index vector $[j_1, j_2, \ldots, j_n]$ as a permutation of $[1, 2, \ldots, n]$ such that

$$[j_1, j_2, \ldots, j_n] = [1, 2, \ldots, n]\boldsymbol{P} = [J_s, J_r].^{18}$$

Thus, it follows from $\boldsymbol{CE} = \boldsymbol{A}[:,J_r]$ that

$$\underbrace{[\boldsymbol{a}_{j_1}, \boldsymbol{a}_{j_2}, \ldots, \boldsymbol{a}_{j_r}]}_{=\boldsymbol{C}=\boldsymbol{A}[:,J_s]}\boldsymbol{E} = \underbrace{[\boldsymbol{a}_{j_{r+1}}, \boldsymbol{a}_{j_{r+2}}, \ldots, \boldsymbol{a}_{j_n}]}_{=\boldsymbol{A}[:,J_r]:=\boldsymbol{B}},$$

where $\boldsymbol{a}_i$ is the i-th column of $\boldsymbol{A}$ and let $\boldsymbol{B} = \boldsymbol{A}[:,J_r]$. Therefore, by Cramer's rule in Equation (7.2), we have

$$\boldsymbol{E}_{kl} = \frac{\det(\boldsymbol{C_B}(k,l))}{\det(\boldsymbol{C})}, \tag{7.4}$$

18. Note here $[j_1, j_2, \ldots, j_n]$, $[1, 2, \ldots, n]$, J_s, and J_r are row vectors.

where $\boldsymbol{E}_{kl}$ is the entry (k,l) of $\boldsymbol{E}$ and $\boldsymbol{C}_{\boldsymbol{B}}(k,l)$ is the $r\times r$ matrix formed by replacing the k-th column of $\boldsymbol{C}$ by the l-th column of $\boldsymbol{B}$. For example,

$$
\boldsymbol{E}_{11}=\frac{\det\left([\boldsymbol{a}_{j_{r+1}},\boldsymbol{a}_{j_2},\ldots,\boldsymbol{a}_{j_r}]\right)}{\det\left([\boldsymbol{a}_{j_1},\boldsymbol{a}_{j_2},\ldots,\boldsymbol{a}_{j_r}]\right)},\qquad \boldsymbol{E}_{12}=\frac{\det\left([\boldsymbol{a}_{j_{r+2}},\boldsymbol{a}_{j_2},\ldots,\boldsymbol{a}_{j_r}]\right)}{\det\left([\boldsymbol{a}_{j_1},\boldsymbol{a}_{j_2},\ldots,\boldsymbol{a}_{j_r}]\right)},
$$
$$
\boldsymbol{E}_{21}=\frac{\det\left([\boldsymbol{a}_{j_1},\boldsymbol{a}_{j_{r+1}},\ldots,\boldsymbol{a}_{j_r}]\right)}{\det\left([\boldsymbol{a}_{j_1},\boldsymbol{a}_{j_2},\ldots,\boldsymbol{a}_{j_r}]\right)},\qquad \boldsymbol{E}_{22}=\frac{\det\left([\boldsymbol{a}_{j_1},\boldsymbol{a}_{j_{r+2}},\ldots,\boldsymbol{a}_{j_r}]\right)}{\det\left([\boldsymbol{a}_{j_1},\boldsymbol{a}_{j_2},\ldots,\boldsymbol{a}_{j_r}]\right)}.
$$

Since J_s is chosen to maximize the magnitude of $\det(\boldsymbol{C})$ in Equation (7.3), it follows that

$$
|\boldsymbol{E}_{kl}|\leq 1,\qquad \text{for all}\quad k\in\{1,2,\ldots,r\}, l\in\{1,2,\ldots,n-r\}.
$$

Step 2: apply to general matrices To summarize what we have proved above and to abuse the notation. For any matrix $\boldsymbol{F}\in\mathbb{R}^{r\times n}$ with **full** rank $r\leq n$, the column ID exists that $\boldsymbol{F}=\boldsymbol{C}_0\boldsymbol{W}$ where the values in $\boldsymbol{W}$ are no greater than 1 in absolute value.

Apply the finding to the full general matrix $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ with rank $r\leq\{m,n\}$, it is trivial that the matrix $\boldsymbol{A}$ admits a rank decomposition (Theorem 5.4, p. 79):

$$
\underset{m\times n}{\boldsymbol{A}}=\underset{m\times r}{\boldsymbol{D}}\quad\underset{r\times n}{\boldsymbol{F}},
$$

where $\boldsymbol{D},\boldsymbol{F}$ have full column rank r and full row rank r respectively. For the column ID of $\boldsymbol{F}=\boldsymbol{C}_0\boldsymbol{W}$ where $\boldsymbol{C}_0=\boldsymbol{F}[:,J_s]$ contains r linearly independent columns of $\boldsymbol{F}$. We notice by $\boldsymbol{A}=\boldsymbol{D}\boldsymbol{F}$ such that

$$
\boldsymbol{A}[:,J_s]=\boldsymbol{D}\boldsymbol{F}[:,J_s],
$$

i.e., the columns indexed by J_s of $(\boldsymbol{D}\boldsymbol{F})$ can be obtained by $\boldsymbol{D}\boldsymbol{F}[:,J_s]$ which in turn are the columns of $\boldsymbol{A}$ indexed by J_s. This makes

$$
\underbrace{\boldsymbol{A}[:,J_s]}_{\boldsymbol{C}}=\underbrace{\boldsymbol{D}\boldsymbol{F}[:,J_s]}_{\boldsymbol{D}\boldsymbol{C}_0},
$$

And

$$
\boldsymbol{A}=\boldsymbol{D}\boldsymbol{F}=\boldsymbol{D}\boldsymbol{C}_0\boldsymbol{W}=\underbrace{\boldsymbol{D}\boldsymbol{F}[:,J_s]}_{\boldsymbol{C}}\boldsymbol{W}=\boldsymbol{C}\boldsymbol{W}.
$$

This completes the proof. ■

The above proof reveals an intuitive way to compute the optimal column ID of matrix $\boldsymbol{A}$. However, any algorithm that is guaranteed to find such an optimally-conditioned factorization must have combinatorial complexity (Martinsson, 2019). Therefore, randomized algorithms, approximation by column-pivoted QR (Section 3.8, p. 53) and rank-revealing QR (Section 3.10, p. 56) are applied to find a relatively well-conditioned decomposition for the column ID where $\boldsymbol{W}$ is small in norm rather than having entries all smaller than 1 in magnitude. See (Lu, 2021c) for more details.

7.2 Row ID and Two-Sided ID

We term the decomposition above as column ID. This is no coincidence since it has its siblings:

Theorem 7.2: (The Whole Interpolative Decomposition)

Any rank-r matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ can be factored as

$$
\begin{aligned}
\text{Column ID:} \qquad \underset{m\times n}{\boldsymbol{A}} &= \boxed{\underset{m\times r}{\boldsymbol{C}}} \quad \underset{r\times n}{\boldsymbol{W}};\\
\text{Row ID:} \qquad &= \underset{m\times r}{\boldsymbol{Z}} \quad \boxed{\underset{r\times n}{\boldsymbol{R}}};\\
\text{Two-Sided ID:} \qquad &= \underset{m\times r}{\boldsymbol{Z}} \quad \boxed{\underset{r\times r}{\boldsymbol{U}}} \quad \underset{r\times n}{\boldsymbol{W}},
\end{aligned}
$$

where

- $\boldsymbol{C} = \boldsymbol{A}[:, J_s] \in \mathbb{R}^{m\times r}$ is some r linearly independent columns of $\boldsymbol{A}$, $\boldsymbol{W} \in \mathbb{R}^{r\times n}$ is the matrix to reconstruct $\boldsymbol{A}$ which contains an $r \times r$ identity submatrix (under a mild column permutation): $\boldsymbol{W}[:, J_s] = \boldsymbol{I}_r$;
- $\boldsymbol{R} = \boldsymbol{A}[I_s, :] \in \mathbb{R}^{r\times n}$ is some r linearly independent rows of $\boldsymbol{R}$, $\boldsymbol{Z} \in \mathbb{R}^{m\times r}$ is the matrix to reconstruct $\boldsymbol{A}$ which contains an $r \times r$ identity submatrix (under a mild row permutation): $\boldsymbol{Z}[I_s, :] = \boldsymbol{I}_r$;
- Entries in $\boldsymbol{W}, \boldsymbol{Z}$ have values no larger than 1 in magnitude: $\max |w_{ij}| \leq 1$ and $\max |z_{ij}| \leq 1$;
- $\boldsymbol{U} = \boldsymbol{A}[I_s, J_s] \in \mathbb{R}^{r\times r}$ is the nonsingular submatrix on the intersection of $\boldsymbol{C}, \boldsymbol{R}$;
- The three matrices $\boldsymbol{C}, \boldsymbol{R}, \boldsymbol{U}$ in the $\boxed{\text{boxed}}$ texts share same notation as the skeleton decomposition (Theorem 6.1, p. 81) where they even have same meanings such that the three matrices make the skeleton decomposition of $\boldsymbol{A}$: $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{U}^{-1}\boldsymbol{R}$.

The proof of the row ID is just similar to that of the column ID. Suppose the column ID of $\boldsymbol{A}^\top$ is given by $\boldsymbol{A}^\top = \boldsymbol{C}_0\boldsymbol{W}_0$ where $\boldsymbol{C}_0$ contains r linearly independent columns of $\boldsymbol{A}^\top$ (i.e., r linearly independent rows of $\boldsymbol{A}$). Let $\boldsymbol{R} = \boldsymbol{C}_0, \boldsymbol{Z} = \boldsymbol{W}_0$, the row ID is obtained by $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{R}$.

For the two-sided ID, recall from the skeleton decomposition (Theorem 6.1, p. 81). When $\boldsymbol{U}$ is the intersection of $\boldsymbol{C}, \boldsymbol{R}$, it follows that $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{U}^{-1}\boldsymbol{R}$. Thus $\boldsymbol{C}\boldsymbol{U}^{-1} = \boldsymbol{Z}$ by the row ID. And this implies $\boldsymbol{C} = \boldsymbol{Z}\boldsymbol{U}$. By column ID, it follows that $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{U}\boldsymbol{W}$ which proves the existence of the two-sided ID.

Data storage For the data storage of each ID, we summarize as follows

- *Column ID.* It requires mr and $(n-r)r$ floats to store $\boldsymbol{C}$ and $\boldsymbol{W}$ respectively , and r integers to store the indices of the selected columns in $\boldsymbol{A}$;
- *Row ID.* It requires nr and $(m-r)r$ floats to store $\boldsymbol{R}$ and $\boldsymbol{Z}$ respectively, and r integers to store the indices of the selected rows in $\boldsymbol{A}$;
- *Two-Sided ID.* It requires $(m-r)r$, $(n-r)r$, and r^2 floats to store $\boldsymbol{Z}, \boldsymbol{W}$, and $\boldsymbol{U}$ respectively. And extra $2r$ integers are required to store the indices of the selected rows and columns in $\boldsymbol{A}$.

Further reduction on the storage for two-sided ID for sparse matrix $\boldsymbol{A}$ Suppose the column ID of $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{W}$ where $\boldsymbol{C} = \boldsymbol{A}[:, J_s]$ and a good spanning rows index I_s set of $\boldsymbol{C}$ could be found:

$$\boldsymbol{A}[I_s, :] = \boldsymbol{C}[I_s, :]\boldsymbol{W}.$$

We observe that $\boldsymbol{C}[I_s, :] = \boldsymbol{A}[I_s, J_s] \in \mathbb{R}^{r\times r}$ which is nonsingular (since full rank r in the sense of both row rank and column rank). It follows that

$$\boldsymbol{W} = (\boldsymbol{A}[I_s, J_s])^{-1}\boldsymbol{A}[I_s, :].$$

Therefore, there is no need to store the matrix $\boldsymbol{W}$ explicitly. We only need to store $\boldsymbol{A}[I_s, :]$ and $(\boldsymbol{A}[I_s, J_s])^{-1}$. Or when we can compute the inverse of $\boldsymbol{A}[I_s, J_s]$ on the fly, it only requires r integers to store J_s and recover $\boldsymbol{A}[I_s, J_s]$ from $\boldsymbol{A}[I_s, :]$. The storage of $\boldsymbol{A}[I_s, :]$ is cheap if $\boldsymbol{A}$ is sparse.

Part IV
Reduction to Hessenberg, Tridiagonal, and Bidiagonal Form

8. Hessenberg Decomposition

We firstly give the rigorous definition of the upper Hessenberg matrix.

Definition 8.1: (Upper Hessenberg Matrix)

An *upper Hessenberg matrix* is a square matrix where all the entries below the first diagonal (i.e., the ones below the *main diagonal*) (a.k.a., *lower subdiagonal*) are zeros. Similarly, a lower Hessenberg matrix is a square matrix where all the entries above the first diagonal (i.e., the ones above the main diagonal) are zeros.

The definition of the upper Hessenberg can also be extended to rectangular matrices, and the form can be implied from the context.

In matrix language, for any matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$, and the entry (i, j) denoted by h_{ij} for all $i, j \in \{1, 2, \ldots, n\}$. Then $\boldsymbol{H}$ with $h_{ij} = 0$ for all $i \geq j + 2$ is known as an Hessenberg matrix.

Let i denote the smallest positive integer for which $h_{i+1,i} = 0$ where $i \in \{1, 2, \ldots, n-1\}$, then $\boldsymbol{H}$ is *unreduced* if $i = n - 1$.

Take a 5×5 matrix as an example, the lower triangular below the lower sub-diagonal are zero in the upper Hessenberg matrix:

$$\underset{\textit{possibly unreduced}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & 0 & \boxtimes & \boxtimes & \boxtimes \\ 0 & 0 & 0 & \boxtimes & \boxtimes \end{bmatrix}} \qquad \text{or} \qquad \underset{\textit{reduced}}{\begin{bmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & 0 & {\color{blue}0} & \boxtimes & \boxtimes \\ 0 & 0 & 0 & \boxtimes & \boxtimes \end{bmatrix}}.$$

Then we have the following Hessenberg decomposition:

Theorem 8.2: (Hessenberg Decomposition)

Every $n \times n$ square matrix $\boldsymbol{A}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}^\top \qquad \text{or} \qquad \boldsymbol{H} = \boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{Q},$$

where $\boldsymbol{H}$ is an upper Hessenberg matrix, and $\boldsymbol{Q}$ is an orthogonal matrix.

It's not hard to find that a lower Hessenberg decomposition of $\boldsymbol{A}^\top$ is given by $\boldsymbol{A}^\top = \boldsymbol{Q}\boldsymbol{H}^\top\boldsymbol{Q}^\top$ if $\boldsymbol{A}$ has the Hessenberg decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}^\top$. The Hessenberg decomposition shares a similar form as the QR decomposition in that they both reduce a matrix into a sparse form where the lower parts of both are zero.

Remark 8.3: (Why Hessenberg Decomposition)

We will see that the zeros introduced into $\boldsymbol{H}$ from $\boldsymbol{A}$ is accomplished by the left orthogonal matrix $\boldsymbol{Q}$ (same as the QR decomposition) and the right orthogonal matrix $\boldsymbol{Q}^\top$ here does not transform the matrix into any better or simple form. Then why do we want the Hessenberg decomposition rather than just a QR decomposition which has a simpler structure in that it even has zeros in the lower sub-diagonal? The answer is that the Hessenberg decomposition is usually used by other algorithms as a phase 1 step to find a decomposition that factor the matrix into two orthogonal matrices, e.g., SVD, UTV, and so on. And if we employ an aggressive algorithm that even favors zeros in the lower sub-diagonal (again, as in the QR decomposition), the right orthogonal transform $\boldsymbol{Q}^\top$ will destroy the zeros that can be seen very shortly.

On the other hand, the form $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}^\top$ on $\boldsymbol{H}$ is known as the *orthogonal similarity transformation* (Definition 8.4, p. 90) on $\boldsymbol{A}$ such that the eigenvalues, rank and trace of $\boldsymbol{A}$ and $\boldsymbol{H}$ are the same (Lemma 8.5, p. 91). Then if we want to study the properties of $\boldsymbol{A}$, exploration on $\boldsymbol{H}$ can be a relatively simpler task.

8.1 Similarity Transformation and Orthogonal Similarity Transformation

As mentioned previously, the Hessenberg decomposition introduced in this section, the tridiagonal decomposition in the next section, the Schur decomposition (Theorem 12.1, p. 110), and the spectral decomposition (Theorem 13.1, p. 113) share a similar form that transforms the matrix into a similar matrix. We now give the rigorous definition of similar matrices and similarity transformations.

Definition 8.4: (Similar Matrices and Similarity Transformation)

$\boldsymbol{A}$ and $\boldsymbol{B}$ are called *similar matrices* if there exists a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$.

In words, for any nonsingular matrix $\boldsymbol{P}$, the matrices $\boldsymbol{A}$ and $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$ are similar matrices. And in this sense, given the nonsingular matrix $\boldsymbol{P}$, $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$ is called a *similarity transformation* applied to matrix $\boldsymbol{A}$.

Moreover, when $\boldsymbol{P}$ is orthogonal, then $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^\top$ is also known as the *orthogonal similarity transformation* of $\boldsymbol{A}$.

The difference between the similarity transformation and orthogonal similarity transformation is partly explained in the sense of coordinate transformation (Section 15, p. 148). Now we prove the important properties of similar matrices that will be proved very useful in the sequel.

Lemma 8.5: (Eigenvalue, Trace and Rank of Similar Matrices)

Any eigenvalue of $\boldsymbol{A}$ is also an eigenvalue of $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$. The converse is also true that any eigenvalue of $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$ is also an eigenvalue of $\boldsymbol{A}$. I.e., $\Lambda(\boldsymbol{A}) = \Lambda(\boldsymbol{B})$, where $\Lambda(\boldsymbol{X})$ is the spectrum of matrix $\boldsymbol{X}$ (Definition 0.2, p. 10).

And also the trace and rank of $\boldsymbol{A}$ are equal to those of matrix $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$ for any nonsingular matrix $\boldsymbol{P}$.

Proof [of Lemma 8.5] For any eigenvalue λ of $\boldsymbol{A}$, we have $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$. Then $\lambda\boldsymbol{P}\boldsymbol{x} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{x}$ such that $\boldsymbol{P}\boldsymbol{x}$ is an eigenvector of $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$ corresponding to λ.

Similarly, for any eigenvalue λ of $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$, we have $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{x} = \lambda\boldsymbol{x}$. Then $\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{x} = \lambda\boldsymbol{P}^{-1}\boldsymbol{x}$ such that $\boldsymbol{P}^{-1}\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$ corresponding to λ.

For the trace of $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$, we have $trace(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}) = trace(\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{P}) = trace(\boldsymbol{A})$, where the first equality comes from the fact that trace of a product is invariant under cyclical permutations of the factors:

$$
trace(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = trace(\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}) = trace(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}),
$$

if all $\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}$, $\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}$, and $\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}$ exist.

For the rank of $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$, we separate it into two claims as follows.

Rank claim 1: $rank(\boldsymbol{Z}\boldsymbol{A}) = rank(\boldsymbol{A})$ if $\boldsymbol{Z}$ is nonsingular We will first show that $rank(\boldsymbol{Z}\boldsymbol{A}) = rank(\boldsymbol{A})$ if $\boldsymbol{Z}$ is nonsingular. For any vector $\boldsymbol{n}$ in the null space of $\boldsymbol{A}$, that is $\boldsymbol{A}\boldsymbol{n} = \boldsymbol{0}$. Thus, $\boldsymbol{Z}\boldsymbol{A}\boldsymbol{n} = \boldsymbol{0}$, that is, $\boldsymbol{n}$ is also in the null space of $\boldsymbol{Z}\boldsymbol{A}$. And this implies $\mathcal{N}(\boldsymbol{A}) \subseteq \mathcal{N}(\boldsymbol{Z}\boldsymbol{A})$.

Conversely, for any vector $\boldsymbol{m}$ in the null space of $\boldsymbol{Z}\boldsymbol{A}$, that is $\boldsymbol{Z}\boldsymbol{A}\boldsymbol{m} = \boldsymbol{0}$, we have $\boldsymbol{A}\boldsymbol{m} = \boldsymbol{Z}^{-1}\boldsymbol{0} = \boldsymbol{0}$. That is, $\boldsymbol{m}$ is also in the null space of $\boldsymbol{A}$. And this indicates $\mathcal{N}(\boldsymbol{Z}\boldsymbol{A}) \subseteq \mathcal{N}(\boldsymbol{A})$.

By "sandwiching", the above two arguments imply

$$
\mathcal{N}(\boldsymbol{A}) = \mathcal{N}(\boldsymbol{Z}\boldsymbol{A}) \quad \longrightarrow \quad rank(\boldsymbol{Z}\boldsymbol{A}) = rank(\boldsymbol{A}).
$$

Rank claim 2: $rank(\boldsymbol{A}\boldsymbol{Z}) = rank(\boldsymbol{A})$ if $\boldsymbol{Z}$ is nonsingular We notice that the row rank is equal to the column rank of any matrix (Corollary 0.13, p. 12). Then $rank(\boldsymbol{A}\boldsymbol{Z}) = rank(\boldsymbol{Z}^\top\boldsymbol{A}^\top)$. Since $\boldsymbol{Z}^\top$ is nonsingular, by claim 1, we have $rank(\boldsymbol{Z}^\top\boldsymbol{A}^\top) = rank(\boldsymbol{A}^\top) = rank(\boldsymbol{A})$ where the last equality is again from the fact that the row rank is equal to the column rank of any matrix. This results in $rank(\boldsymbol{A}\boldsymbol{Z}) = rank(\boldsymbol{A})$ as claimed.

Since $\boldsymbol{P}, \boldsymbol{P}^{-1}$ are nonsingular, we then have $rank(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}) = rank(\boldsymbol{A}\boldsymbol{P}^{-1}) = rank(\boldsymbol{A})$ where the first equality is from claim 1 and the second equality is from claim 2. We complete the proof. ■

8.2 Existence of the Hessenberg Decomposition

We will prove that any $n \times n$ matrix can be reduced to Hessenberg form via a sequence of Householder transformations that are applied from the left and the right to the matrix. Previously, we utilized a Householder reflector to triangularize matrices and introduce zeros below the diagonal to obtain the QR decomposition. A similar approach can be applied to introduce zeros below the subdiagonal.

Before introducing the mathematical construction of such decomposition, we emphasize the following remark which will be very useful in the finding of the decomposition.

Remark 8.6: (Left and Right Multiplied by a Matrix with Block Identity)
For square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, and a matrix

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_{n-k} \end{bmatrix},$$

where $\boldsymbol{I}_k$ is a $k \times k$ identity matrix. Then $\boldsymbol{BA}$ will not change the first k rows of $\boldsymbol{A}$, and $\boldsymbol{AB}$ will not change the first k columns of $\boldsymbol{A}$.

The proof of this remark is trivial.

First Step: Introduce Zeros for the First Column

Let $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n]$ be the column partitions of $\boldsymbol{A}$, and each $\boldsymbol{a}_i \in \mathbb{R}^n$. Suppose $\bar{\boldsymbol{a}}_1, \bar{\boldsymbol{a}}_2, \ldots, \bar{\boldsymbol{a}}_n \in \mathbb{R}^{n-1}$ are vectors removing the first component in $\boldsymbol{a}_i$'s. Let

$$r_1 = ||\bar{\boldsymbol{a}}_1||, \qquad \boldsymbol{u}_1 = \frac{\bar{\boldsymbol{a}}_1 - r_1\boldsymbol{e}_1}{||\bar{\boldsymbol{a}}_1 - r_1\boldsymbol{e}_1||}, \qquad \text{and} \qquad \widetilde{\boldsymbol{H}}_1 = \boldsymbol{I} - 2\boldsymbol{u}_1\boldsymbol{u}_1^\top \in \mathbb{R}^{(n-1)\times(n-1)},$$

where $\boldsymbol{e}_1$ here is the first basis for $\mathbb{R}^{n-1}$, i.e., $\boldsymbol{e}_1 = [1; 0; 0; \ldots; 0] \in \mathbb{R}^{n-1}$. To introduce zeros below the sub-diagonal and operate on the submatrix $\boldsymbol{A}_{2:n,1:n}$, we append the Householder reflector into

$$\boldsymbol{H}_1 = \begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{\boldsymbol{H}}_1 \end{bmatrix},$$

in which case, $\boldsymbol{H}_1\boldsymbol{A}$ will introduce zeros in the first column of $\boldsymbol{A}$ below entry (2,1). The first row of $\boldsymbol{A}$ will not be affected at all and kept unchanged by Remark 8.6. And we can easily verify that both $\boldsymbol{H}_1$ and $\widetilde{\boldsymbol{H}}_1$ are orthogonal matrices and they are symmetric (from the definition of Householder reflector). To have the form in Theorem 8.2, we multiply $\boldsymbol{H}_1\boldsymbol{A}$ on the right by $\boldsymbol{H}_1^\top$ which results in $\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top$. The $\boldsymbol{H}_1^\top$ on the right will not change the first column of $\boldsymbol{H}_1\boldsymbol{A}$ and thus keep the zeros introduced in the first column.

An example of a 5×5 matrix is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

$$
\underset{\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{H}_1\times}{\rightarrow}
\underset{\boldsymbol{H}_1\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\stackrel{\times\boldsymbol{H}_1^\top}{\rightarrow}
\underset{\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top}{\begin{bmatrix}
\boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
$$

Second Step: Introduce Zeros for the Second Column

Let $\boldsymbol{B} = \boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top$, where the entries in the first column below entry (2,1) are all zeros. And the goal is to introduce zeros in the second column below entry (3,2). Let $\boldsymbol{B}_2 = \boldsymbol{B}_{2:n,2:n} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{n-1}]$. Suppose again $\bar{\boldsymbol{b}}_1, \bar{\boldsymbol{b}}_2, \ldots, \bar{\boldsymbol{b}}_{n-1} \in \mathbb{R}^{n-2}$ are vectors removing the first component in $\boldsymbol{b}_i$'s. We can again construct a Householder reflector

$$
r_1 = ||\bar{\boldsymbol{b}}_1||, \qquad \boldsymbol{u}_2 = \frac{\bar{\boldsymbol{b}}_1 - r_1\boldsymbol{e}_1}{||\bar{\boldsymbol{b}}_1 - r_1\boldsymbol{e}_1||}, \qquad \text{and} \qquad \widetilde{\boldsymbol{H}}_2 = \boldsymbol{I} - 2\boldsymbol{u}_2\boldsymbol{u}_2^\top \in \mathbb{R}^{(n-2)\times(n-2)}, \tag{8.1}
$$

where $\boldsymbol{e}_1$ now is the first basis for $\mathbb{R}^{n-2}$. To introduce zeros below the sub-diagonal and operate on the submatrix $\boldsymbol{B}_{3:n,1:n}$, we append the Householder reflector into

$$
\boldsymbol{H}_2 = \begin{bmatrix} \boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{\boldsymbol{H}}_2 \end{bmatrix},
$$

where $\boldsymbol{I}_2$ is a 2×2 identity matrix. We can see that $\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top$ will not change the first two rows of $\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top$, and as the Householder cannot reflect a zero vector such that the zeros in the first column will be kept. Again, putting $\boldsymbol{H}_2^\top$ on the right of $\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top$ will not change the first 2 columns so that the zeros will be kept.

Following the example of a 5×5 matrix, the second step is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

$$
\underset{\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{H}_2\times}{\rightarrow}
\underset{\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\stackrel{\times\boldsymbol{H}_2^\top}{\rightarrow}
\underset{\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boxtimes & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
$$

The same process can go on, and there are $n-2$ such steps. We will finally triangularize by

$$
\boldsymbol{H} = \boldsymbol{H}_{n-2}\boldsymbol{H}_{n-3}\ldots\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top\ldots\boldsymbol{H}_{n-2}^\top.
$$

And since $\boldsymbol{H}_i$'s are symmetric and orthogonal, the above equation can be simply reduced to

$$
\boldsymbol{H} = \boldsymbol{H}_{n-2}\boldsymbol{H}_{n-3}\ldots\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1\boldsymbol{H}_2\ldots\boldsymbol{H}_{n-2}.
$$

Note here only $n-2$ such stages exist rather than $n-1$ or n. We will verify this number of steps by the example below. The example of a 5×5 matrix as a whole is shown as follows where again $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

A Complete Example of Hessenberg Decomposition

$$
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}_{\boldsymbol{A}}
\stackrel{\boldsymbol{H}_1\times}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}_{\boldsymbol{H}_1\boldsymbol{A}}
\stackrel{\times\boldsymbol{H}_1^\top}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}_{\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top}
$$

$$
\stackrel{\boldsymbol{H}_2\times}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}_{\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top}
\stackrel{\times\boldsymbol{H}_2^\top}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boxtimes & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}_{\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top}
$$

$$
\stackrel{\boldsymbol{H}_3\times}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}_{\boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top}
\stackrel{\times\boldsymbol{H}_3^\top}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boxtimes & \boxtimes & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \boxtimes & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}_{\boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top\boldsymbol{H}_3^\top},
$$

8.3 Properties of the Hessenberg Decomposition

The Hessenberg decomposition is not unique in the different ways to construct the Householder reflectors (say Equation (8.1), p. 93). However, under mild conditions, we can claim a similar structure in different decompositions.

Theorem 8.7: (Implicit Q Theorem for Hessenberg Decomposition)

Suppose two Hessenberg decompositions of matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ are given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{H}\boldsymbol{U}^\top = \boldsymbol{V}\boldsymbol{G}\boldsymbol{V}^\top$ where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ are the column partitions of $\boldsymbol{U}, \boldsymbol{V}$. Suppose further that k is the smallest positive integer for which $h_{k+1,k} = 0$ where h_{ij} is the entry (i, j) of $\boldsymbol{H}$. Then

- If $\boldsymbol{u}_1 = \boldsymbol{v}_1$, then $\boldsymbol{u}_i = \pm\boldsymbol{v}_i$ and $|h_{i,i-1}| = |g_{i,i-1}|$ for $i \in \{2, 3, \ldots, k\}$.
- When $k = n-1$, the Hessenberg matrix $\boldsymbol{H}$ is known as *unreduced*. However, if $k < n-1$, then $g_{k+1,k} = 0$.

Proof [of Theorem 8.7] Define the orthogonal matrix $\boldsymbol{Q}=\boldsymbol{V}^\top\boldsymbol{U}$ and we have

$$\left.\begin{aligned}\boldsymbol{GQ}=\boldsymbol{V}^\top\boldsymbol{A}\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{U}=\boldsymbol{V}^\top\boldsymbol{A}\boldsymbol{U}\\ \boldsymbol{QH}=\boldsymbol{V}^\top\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{U}=\boldsymbol{V}^\top\boldsymbol{A}\boldsymbol{U}\end{aligned}\right\} \qquad \xrightarrow{\text{leads to}} \qquad \boldsymbol{GQ}=\boldsymbol{QH},$$

the $(i-1)$-th column of each can be represented as

$$\boldsymbol{G}\boldsymbol{q}_{i-1}=\boldsymbol{Q}\boldsymbol{h}_{i-1},$$

where $\boldsymbol{q}_{i-1}$ and $\boldsymbol{h}_{i-1}$ are the $(i-1)$-th column of $\boldsymbol{Q}$ and $\boldsymbol{H}$ respectively. Since $h_{l,i-1}=0$ for $l\geq i+1$ (by the definition of upper Hessenberg matrices), $\boldsymbol{Q}\boldsymbol{h}_{i-1}$ can be represented as

$$\boldsymbol{Q}\boldsymbol{h}_{i-1}=\sum_{j=1}^{i}h_{j,i-1}\boldsymbol{q}_j=h_{i,i-1}\boldsymbol{q}_i+\sum_{j=1}^{i-1}h_{j,i-1}\boldsymbol{q}_j.$$

Combine the two findings above, it follows that

$$h_{i,i-1}\boldsymbol{q}_i=\boldsymbol{G}\boldsymbol{q}_{i-1}-\sum_{j=1}^{i-1}h_{j,i-1}\boldsymbol{q}_j.$$

A moment of reflexion reveals that $[\boldsymbol{q}_1,\boldsymbol{q}_2,\ldots,\boldsymbol{q}_k]$ is upper triangular. And since $\boldsymbol{Q}$ is orthogonal, it must be diagonal and each value on the diagonal is in $\{-1,1\}$ for $i\in\{2,\ldots,k\}$. Then, $\boldsymbol{q}_1=\boldsymbol{e}_1$ and $\boldsymbol{q}_i=\pm\boldsymbol{e}_i$ for $i\in\{2,\ldots,k\}$. Further, since $\boldsymbol{q}_i=\boldsymbol{V}^\top\boldsymbol{u}_i$ and $h_{i,i-1}=\boldsymbol{q}_i^\top(\boldsymbol{G}\boldsymbol{q}_{i-1}-\sum_{j=1}^{i-1}h_{j,i-1}\boldsymbol{q}_j)=\boldsymbol{q}_i^\top\boldsymbol{G}\boldsymbol{q}_{i-1}$. For $i\in\{2,\ldots,k\}$, $\boldsymbol{q}_i^\top\boldsymbol{G}\boldsymbol{q}_{i-1}$ is just $\pm g_{i,i-1}$. It follows that

$$\begin{aligned}|h_{i,i-1}|&=|g_{i,i-1}|, &&\forall i\in\{2,\ldots,k\},\\ \boldsymbol{u}_i&=\pm\boldsymbol{v}_i, &&\forall i\in\{2,\ldots,k\}.\end{aligned}$$

This proves the first part. For the second part, if $k<n-1$,

$$\begin{aligned}g_{k+1,k}=\boldsymbol{e}_{k+1}^\top\boldsymbol{G}\boldsymbol{e}_k=\pm\boldsymbol{e}_{k+1}^\top\underbrace{\boldsymbol{GQ}}_{\boldsymbol{QH}}\boldsymbol{e}_k=\pm\boldsymbol{e}_{k+1}^\top\underbrace{\boldsymbol{QH}\boldsymbol{e}_k}_{k\text{-th column of }\boldsymbol{QH}}\\ =\pm\boldsymbol{e}_{k+1}^\top\boldsymbol{Q}\boldsymbol{h}_k=\pm\boldsymbol{e}_{k+1}^\top\sum_{j}^{k+1}h_{jk}\boldsymbol{q}_j=\pm\boldsymbol{e}_{k+1}^\top\sum_{j}^{k}h_{jk}\boldsymbol{q}_j=0,\end{aligned}$$

where the penultimate equality is from the assumption that $h_{k+1,k}=0$. This completes the proof. ■

We observe from the above theorem, when two Hessenberg decompositions of matrix $\boldsymbol{A}$ are both unreduced and have the same first column in the orthogonal matrices, then the Hessenberg matrices $\boldsymbol{H},\boldsymbol{G}$ are similar matrices such that $\boldsymbol{H}=\boldsymbol{DGD}^{-1}$ where $\boldsymbol{D}=\mathrm{diag}(\pm1,\pm1,\ldots,\pm1)$. *Moreover, and most importantly, if we restrict the elements in the lower sub-diagonal of the Hessenberg matrix $\boldsymbol{H}$ to be positive (if possible), then the Hessenberg decomposition $\boldsymbol{A}=\boldsymbol{QHQ}^\top$ is uniquely determined by $\boldsymbol{A}$ and the first column of $\boldsymbol{Q}$.* This is similar to what we have claimed on the uniqueness of the QR decomposition (Corollary 3.10, p. 64) and it is important to reduce the complexity of the QR algorithm for computing the singular value decomposition or eigenvalues of a matrix in general (Lu, 2021c).

The next finding involves a Krylov matrix defined as follows:

Definition 8.8: (Krylov Matrix)

Given matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, a vector $\boldsymbol{q} \in \mathbb{R}^n$, and a scalar k, the *Krylov matrix* is defined to be

$$\boldsymbol{K}(\boldsymbol{A}, \boldsymbol{q}, k) = \begin{bmatrix}\boldsymbol{q} & \boldsymbol{A}\boldsymbol{q} & \ldots & \boldsymbol{A}^{k-1}\boldsymbol{q}\end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Theorem 8.9: (Reduced Hessenberg)

Suppose there exists an orthogonal matrix $\boldsymbol{Q}$ such that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factored as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}^\top$. Then $\boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{H}$ is an unreduced upper Hessenberg matrix if and only if $\boldsymbol{R} = \boldsymbol{Q}^\top\boldsymbol{K}(\boldsymbol{A}, \boldsymbol{q}_1, n)$ is nonsingular and upper triangular where $\boldsymbol{q}_1$ is the first column of $\boldsymbol{Q}$.

If $\boldsymbol{R}$ is singular and k is the smallest index so that $r_{kk} = 0$, then k is also the smallest index that $h_{k,k-1} = 0$.

Proof [of Theorem 8.9] We prove by forward implication and converse implication separately as follows:

Forward implication Suppose $\boldsymbol{H}$ is unreduced, write out the following matrix

$$\boldsymbol{R} = \boldsymbol{Q}^\top\boldsymbol{K}(\boldsymbol{A}, \boldsymbol{q}_1, n) = [\boldsymbol{e}_1, \boldsymbol{H}\boldsymbol{e}_1, \ldots, \boldsymbol{H}^{n-1}\boldsymbol{e}_1],$$

where $\boldsymbol{R}$ is upper triangular with $r_{11} = 1$ obviously. Observe that $r_{ii} = h_{21}h_{32}\ldots h_{i,i-1}$ for $i \in \{2, 3, \ldots, n\}$. When $\boldsymbol{H}$ is unreduced, $\boldsymbol{R}$ is nonsingular as well.

Converse implication Now suppose $\boldsymbol{R}$ is upper triangular and nonsingular, we observe that $\boldsymbol{r}_{k+1} = \boldsymbol{H}\boldsymbol{r}_k$ such that the $(k+2:n)$-th rows of $\boldsymbol{H}$ are zero and $h_{k+1,k} \neq 0$ for $k \in \{1, 2, \ldots, n-1\}$. Then $\boldsymbol{H}$ is unreduced.

If $\boldsymbol{R}$ is singular and k is the smallest index so that $r_{kk} = 0$, then

$$\left.\begin{aligned} r_{k-1,k-1} &= h_{21}h_{32}\ldots h_{k-1,k-2} && \neq 0 \\ r_{kk} &= h_{21}h_{32}\ldots h_{k-1,k-2}h_{k,k-1} && = 0 \end{aligned}\right\} \qquad \underrightarrow{\text{leads to}} \qquad h_{k,k-1} = 0,$$

from which the result follows. ■

9. Tridiagonal Decomposition: Hessenberg in Symmetric Matrices

We firstly give the formal definition of the tridiagonal matrix.

Definition 9.1: (Tridiagonal Matrix)

A tridiagonal matrix is a square matrix where all the entries below the lower sub-diagonal and the entries above the upper sub-diagonal are zeros. I.e., the tridiagonal matrix is a *band matrix*.

The definition of the tridiagonal matrix can also be extended to rectangular matrices, and the form can be implied from the context.

In matrix language, for any matrix $\boldsymbol{T} \in \mathbb{R}^{n\times n}$, and the entry (i,j) denoted by t_{ij} for all $i,j \in \{1,2,\ldots,n\}$. Then $\boldsymbol{T}$ with $t_{ij}=0$ for all $i \geq j+2$ and $i \leq j-2$ is known as a tridiagonal matrix.

Let i denote the smallest positive integer for which $h_{i+1,i}=0$ where $i \in \{1,2,\ldots,n-1\}$, then $\boldsymbol{T}$ is *unreduced* if $i=n-1$.

Take a 5×5 matrix as an example, the lower triangular below the lower sub-diagonal and upper triangular above the upper sub-diagonal are zero in the tridiagonal matrix:

$$
\underset{\textit{possibly unreduced}}{\begin{bmatrix}
\boxtimes & \boxtimes & 0 & 0 & 0\\
\boxtimes & \boxtimes & \boxtimes & 0 & 0\\
0 & \boxtimes & \boxtimes & \boxtimes & 0\\
0 & 0 & \boxtimes & \boxtimes & \boxtimes\\
0 & 0 & 0 & \boxtimes & \boxtimes
\end{bmatrix}}
\qquad
\underset{\textit{reduced}}{\begin{bmatrix}
\boxtimes & \boxtimes & 0 & 0 & 0\\
\boxtimes & \boxtimes & \boxtimes & 0 & 0\\
0 & \boxtimes & \boxtimes & \boxtimes & 0\\
0 & 0 & \textcolor{blue}{0} & \boxtimes & \boxtimes\\
0 & 0 & 0 & \boxtimes & \boxtimes
\end{bmatrix}}.
$$

Obviously, a tridiagonal matrix is a special case of an upper Hessenberg matrix. Then we have the following tridiagonal decomposition:

Theorem 9.2: (Tridiagonal Decomposition)

Every $n\times n$ symmetric matrix $\boldsymbol{A}$ can be factored as

$$
\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^\top \qquad \text{or} \qquad \boldsymbol{T} = \boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{Q},
$$

where $\boldsymbol{T}$ is a *symmetric* tridiagonal matrix, and $\boldsymbol{Q}$ is an orthogonal matrix.

The existence of the tridiagonal matrix is trivial by applying the Hessenberg decomposition to symmetric matrix $\boldsymbol{A}$.

9.1 Properties of the Tridiagonal Decomposition

Similarly, the tridiagonal decomposition is not unique. However, and most importantly, if we restrict the elements in the lower sub-diagonal of the tridiagonal matrix $\boldsymbol{T}$ to be positive (if possible), then the tridiagonal decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^\top$ is uniquely determined by $\boldsymbol{A}$ and the first column of $\boldsymbol{Q}$.

Theorem 9.3: (Implicit Q Theorem for Tridiagonal)

Suppose two Tridiagonal decompositions of symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ are given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^\top = \boldsymbol{V}\boldsymbol{G}\boldsymbol{V}^\top$ where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ are the column partitions of $\boldsymbol{U}, \boldsymbol{V}$. Suppose further that k is the smallest positive integer for which $t_{k+1,k}=0$ where t_{ij} is the entry (i,j) of $\boldsymbol{T}$. Then

- If $\boldsymbol{u}_1 = \boldsymbol{v}_1$, then $\boldsymbol{u}_i = \pm\boldsymbol{v}_i$ and $|t_{i,i-1}| = |g_{i,i-1}|$ for $i \in \{2,3,\ldots,k\}$.
- When $k=n-1$, the tridiagonal matrix $\boldsymbol{T}$ is known as unreduced. However, if $k<n-1$, then $g_{k+1,k}=0$.

From the above theorem, we observe that if we restrict the elements in the lower sub-diagonal of the tridiagonal matrix $\boldsymbol{T}$ to be positive (if possible), i.e., *unreduced*, then the tridiagonal decomposition $\boldsymbol{A} = \boldsymbol{QTQ}^\top$ is uniquely determined by $\boldsymbol{A}$ and the first column of $\boldsymbol{Q}$. This again is similar to what we have claimed on the uniqueness of the QR decomposition (Corollary 3.10, p. 64).

Similarly, a reduced tridiagonal decomposition can be obtained from the implication of the Krylov matrix (Definition 8.8, p. 96).

Theorem 9.4: (Reduced Tridiagonal)

Suppose there exists an orthogonal matrix $\boldsymbol{Q}$ such that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factored as $\boldsymbol{A} = \boldsymbol{QTQ}^\top$. Then $\boldsymbol{Q}^\top \boldsymbol{AQ} = \boldsymbol{T}$ is an unreduced tridiagonal matrix if and only if $\boldsymbol{R} = \boldsymbol{Q}^\top \boldsymbol{K}(\boldsymbol{A}, \boldsymbol{q}_1, n)$ is nonsingular and upper triangular where $\boldsymbol{q}_1$ is the first column of $\boldsymbol{Q}$.

If $\boldsymbol{R}$ is singular and k is the smallest index so that $r_{kk} = 0$, then k is also the smallest index that $t_{k,k-1} = 0$.

The proofs of the above two theorems are the same as those in Theorem 8.7 and 8.9.

10. Bidiagonal Decomposition

We firstly give the rigorous definition of the upper Bidiagonal matrix as follows:

Definition 10.1: (Upper Bidiagonal Matrix)

An upper bidiagonal matrix is a square matrix which is a banded matrix with non-zero entries along the *main diagonal* and the *upper subdiagonal* (i.e., the ones above the main diagonal). This means there are exactly two nonzero diagonals in the matrix.

Furthermore, when the diagonal below the main diagonal has the non-zero entries, the matrix is lower bidiagonal.

The definition of bidigonal matrices can also be extended to rectangular matrices, and the form can be implied from the context.

Take a 7×5 matrix as an example, the lower triangular below the main diagonal and the upper triangular above the upper subdiagonal are zero in the upper bidiagonal matrix:

$$
\begin{bmatrix}
\boxtimes & \boxtimes & 0 & 0 & 0 \\
0 & \boxtimes & \boxtimes & 0 & 0 \\
0 & 0 & \boxtimes & \boxtimes & 0 \\
0 & 0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & 0 & 0 & \boxtimes \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}.
$$

Then we have the following bidiagonal decomposition:

Theorem 10.2: (Bidiagonal Decomposition)
Every $m \times n$ matrix $\boldsymbol{A}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^\top \qquad \text{or} \qquad \boldsymbol{B} = \boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{V},$$

where $\boldsymbol{B}$ is an upper bidiagonal matrix, and $\boldsymbol{U}, \boldsymbol{V}$ are orthogonal matrices.

We will see the bidiagonalization resembles the form of a singular value decomposition where the only difference is the values of $\boldsymbol{B}$ in bidiagonal form has nonzero entries on the upper sub-diagonal such that it will be shown to play an important role in the calculation of the singular value decomposition.

10.1 Existence of the Bidiagonal Decomposition: Golub-Kahan Bidiagonalization

Previously, we utilized a Householder reflector to triangularize matrices and introduce zeros below the diagonal to obtain the QR decomposition, and introduce zeros below the sub-diagonal to obtain the Hessenberg decomposition. A similar approach can be employed to find the bidiagonal decomposition.

First Step 1.1: Introduce Zeros for the First Column

Let $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n]$ be the column partitions of $\boldsymbol{A}$, and each $\boldsymbol{a}_i \in \mathbb{R}^m$. We can construct the Householder reflector as follows:

$$r_1 = ||\boldsymbol{a}_1||, \qquad \boldsymbol{u}_1 = \frac{\boldsymbol{a}_1 - r_1\boldsymbol{e}_1}{||\boldsymbol{a}_1 - r_1\boldsymbol{e}_1||}, \qquad \text{and} \qquad \boldsymbol{H}_1 = \boldsymbol{I} - 2\boldsymbol{u}_1\boldsymbol{u}_1^\top \in \mathbb{R}^{m\times m},$$

where $\boldsymbol{e}_1$ here is the first basis for $\mathbb{R}^m$, i.e., $\boldsymbol{e}_1 = [1; 0; 0; \ldots; 0] \in \mathbb{R}^m$. In this case, $\boldsymbol{H}_1\boldsymbol{A}$ will introduce zeros in the first column of $\boldsymbol{A}$ below entry (1,1), i.e., reflect $\boldsymbol{a}_1$ to $r_1\boldsymbol{e}_1$. We can easily verify that both $\boldsymbol{H}_1$ is a symmetric and orthogonal matrix (from the definition of Householder reflector).

An example of a 7×5 matrix is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

$$\underset{\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{H}_1\times}{\rightarrow}
\underset{\boldsymbol{H}_1\boldsymbol{A}}{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
\boldsymbol{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}.$$

Till now, this is exactly what we have done in the QR decomposition via the Householder reflector in Section 3.11 (p. 56). Going further, to introduce zeros above the upper sub-diagonal of $\boldsymbol{H}_1\boldsymbol{A}$ is equivalent to introducing zeros below the lower subdiagonal of $(\boldsymbol{H}_1\boldsymbol{A})^\top$.

First Step 1.2: Introduce Zeros for the First Row

Now suppose we are looking at the *transpose* of $\boldsymbol{H}_1\boldsymbol{A}$, that is $(\boldsymbol{H}_1\boldsymbol{A})^\top = \boldsymbol{A}^\top\boldsymbol{H}_1^\top \in \mathbb{R}^{n\times m}$ and the column partition is given by $\boldsymbol{A}^\top\boldsymbol{H}_1^\top = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m]$ where each $\boldsymbol{z}_i \in \mathbb{R}^n$. Suppose $\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2, \ldots, \bar{\boldsymbol{z}}_m \in \mathbb{R}^{n-1}$ are vectors removing the first component in $\boldsymbol{z}_i$'s. Let

$$
r_1 = ||\bar{\boldsymbol{z}}_1||, \qquad \boldsymbol{v}_1 = \frac{\bar{\boldsymbol{z}}_1 - r_1\boldsymbol{e}_1}{||\bar{\boldsymbol{z}}_1 - r_1\boldsymbol{e}_1||}, \qquad \text{and} \qquad \widetilde{\boldsymbol{L}}_1 = \boldsymbol{I} - 2\boldsymbol{v}_1\boldsymbol{v}_1^\top \in \mathbb{R}^{(n-1)\times(n-1)},
$$

where $\boldsymbol{e}_1$ now is the first basis for $\mathbb{R}^{n-1}$, i.e., $\boldsymbol{e}_1 = [1; 0; 0; \ldots; 0] \in \mathbb{R}^{n-1}$. To introduce zeros below the sub-diagonal and operate on the submatrix $(\boldsymbol{A}^\top\boldsymbol{H}_1^\top)_{2:n,1:m}$, we append the Householder reflector into

$$
\boldsymbol{L}_1 = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{L}}_1 \end{bmatrix},
$$

in which case, $\boldsymbol{L}_1(\boldsymbol{A}^\top\boldsymbol{H}_1^\top)$ will introduce zeros in the first column of $(\boldsymbol{A}^\top\boldsymbol{H}_1^\top)$ below entry (2,1), i.e., reflect $\bar{\boldsymbol{z}}_1$ to $r_1\boldsymbol{e}_1$. The first row of $(\boldsymbol{A}^\top\boldsymbol{H}_1^\top)$ will not be affected at all and kept unchanged by Remark 8.6 (p. 92) such that the zeros introduced in Step 1.1) will be kept. And we can easily verify that both $\boldsymbol{L}_1$ and $\widetilde{\boldsymbol{L}}_1$ are orthogonal matrices and they are symmetric (from the definition of Householder reflector).

Come back to the original *untransposed* matrix $\boldsymbol{H}_1\boldsymbol{A}$, multiply on the right by $\boldsymbol{L}_1^\top$ is to introduce zeros in the first row to the right of entry (1,2). Again, following the example above, a 7×5 matrix is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

$$
\underset{\boldsymbol{A}^\top\boldsymbol{H}_1^\top}{\begin{bmatrix}
\boxtimes & 0 & 0 & 0 & 0 & 0 & 0\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{L}_1\times}{\rightarrow}
\underset{\boldsymbol{L}_1\boldsymbol{A}^\top\boldsymbol{H}_1^\top}{\begin{bmatrix}
\boxtimes & 0 & 0 & 0 & 0 & 0 & 0\\
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\stackrel{(\cdot)^\top}{\rightarrow}
\underset{\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top}{\begin{bmatrix}
\boxtimes & \boldsymbol{\boxtimes} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}.
$$

In short, $\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top$ finishes the first step to introduce zeros for the first column and the first row of $\boldsymbol{A}$.

Second Step 2.1: Introduce Zeros for the Second Column

Let $\boldsymbol{B} = \boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top$, where the entries in the first column below entry (1,1) are all zeros and the entries in the first row to the right of entry (1,2) are all zeros as well. And the goal is to introduce zeros in the second column below entry (2,2). Let $\boldsymbol{B}_2 = \boldsymbol{B}_{2:m,2:n} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{n-1}] \in \mathbb{R}^{(m-1)\times(n-1)}$. We can again construct a Householder reflector

$$
r_1 = ||\boldsymbol{b}_1||, \qquad \boldsymbol{u}_2 = \frac{\boldsymbol{b}_1 - r_1\boldsymbol{e}_1}{||\boldsymbol{b}_1 - r_1\boldsymbol{e}_1||}, \qquad \text{and} \qquad \widetilde{\boldsymbol{H}}_2 = \boldsymbol{I} - 2\boldsymbol{u}_2\boldsymbol{u}_2^\top \in \mathbb{R}^{(m-1)\times(m-1)},
$$

where $\boldsymbol{e}_1$ now is the first basis for $\mathbb{R}^{m-1}$ i.e., $\boldsymbol{e}_1 = [1; 0; 0; \ldots; 0] \in \mathbb{R}^{m-1}$. To introduce zeros below the main diagonal and operate on the submatrix $\boldsymbol{B}_{2:m,2:n}$, we append the Householder reflector into

$$\boldsymbol{H}_2 = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{H}}_2 \end{bmatrix},$$

in which case, we can see that $\boldsymbol{H}_2(\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top)$ will not change the first row of $(\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top)$ by Remark 8.6 (p. 92), and as the Householder cannot reflect a zero vector such that the zeros in the first column will be kept as well.

Following the example above, a 7×5 matrix is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

$$\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & 0 & 0 & 0 \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}_{\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top}
\stackrel{\boldsymbol{H}_2\times}{\rightarrow}
\mathop{\begin{bmatrix}
\boxtimes & \boxtimes & 0 & 0 & 0 \\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}_{\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top}.$$

Second Step 2.2: Introduce Zeros for the Second Row

Same as step 1.2), now suppose we are looking at the *transpose* of $\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top$, that is $(\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top)^\top = \boldsymbol{L}_1\boldsymbol{A}^\top\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top \in \mathbb{R}^{n\times m}$ and the column partition is given by $\boldsymbol{L}_1\boldsymbol{A}^\top\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$ where each $\boldsymbol{x}_i \in \mathbb{R}^n$. Suppose $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \ldots, \bar{\boldsymbol{x}}_m \in \mathbb{R}^{n-2}$ are vectors removing the first two components in $\boldsymbol{x}_i$'s. Construct the Householder reflector as follows:

$$r_1 = ||\bar{\boldsymbol{x}}_1||, \qquad \boldsymbol{v}_2 = \frac{\bar{\boldsymbol{x}}_1 - r_1\boldsymbol{e}_1}{||\bar{\boldsymbol{x}}_1 - r_1\boldsymbol{e}_1||}, \qquad \text{and} \qquad \widetilde{\boldsymbol{L}}_2 = \boldsymbol{I} - 2\boldsymbol{v}_2\boldsymbol{v}_2^\top \in \mathbb{R}^{(n-2)\times(n-2)},$$

where $\boldsymbol{e}_1$ now is the first basis for $\mathbb{R}^{n-2}$, i.e., $\boldsymbol{e}_1 = [1; 0; 0; \ldots; 0] \in \mathbb{R}^{n-2}$. To introduce zeros below the sub-diagonal and operate on the submatrix $(\boldsymbol{L}_1\boldsymbol{A}^\top\boldsymbol{H}_1\boldsymbol{H}_2)_{3:n,1:m}$, we append the Householder reflector into

$$\boldsymbol{L}_1 = \begin{bmatrix} \boldsymbol{I}_2 & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{L}}_2 \end{bmatrix},$$

where $\boldsymbol{I}_2$ is a 2×2 identity matrix. In this case, $\boldsymbol{L}_2(\boldsymbol{L}_1\boldsymbol{A}^\top\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top)$ will introduce zeros in the second column of $(\boldsymbol{L}_1\boldsymbol{A}^\top\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top)$ below entry (3,2). The first two rows of $(\boldsymbol{L}_1\boldsymbol{A}^\top\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top)$ will not be affected at all and kept unchanged by Remark 8.6 (p. 92). **Further, the first column of it will be kept unchanged as well**. And we can easily verify that both $\boldsymbol{L}_1$ and $\widetilde{\boldsymbol{L}}_1$ are orthogonal matrices and they are symmetric (from the definition of Householder reflector).

Come back to the original *untransposed* matrix $\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top$, multiply on the right by $\boldsymbol{L}_2^\top$ is to introduce zeros in the second row to the right of entry (2,3). Following the example above, a 7×5 matrix is shown as follows where $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

$$
\underset{\boldsymbol{L}_1\boldsymbol{A}^\top\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top}{\begin{bmatrix}
\boxtimes & 0 & 0 & 0 & 0 & 0 & 0\\
\boxtimes & \boxtimes & 0 & 0 & 0 & 0 & 0\\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{L}_2\times}{\rightarrow}
\underset{\boldsymbol{L}_2\boldsymbol{L}_1\boldsymbol{A}^\top\boldsymbol{H}_1^\top\boldsymbol{H}_2^\top}{\begin{bmatrix}
\boxtimes & 0 & 0 & 0 & 0 & 0 & 0\\
\boxtimes & \boxtimes & 0 & 0 & 0 & 0 & 0\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\stackrel{(\cdot)^\top}{\rightarrow}
\underset{\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top\boldsymbol{L}_2^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & 0 & 0 & 0\\
0 & \boxtimes & \boldsymbol{\boxtimes} & \mathbf{0} & \mathbf{0}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}.
$$

In short, $\boldsymbol{H}_2(\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top)\boldsymbol{L}_2^\top$ finish the second step to introduce zeros for the second column and the second row of $\boldsymbol{A}$. The same process can go on, and we shall notice that there are n such $\boldsymbol{H}_i$ Householder reflectors on the left and $n-2$ such $\boldsymbol{L}_i$ Householder reflectors on the right (suppose $m > n$ for simplicity). The interleaved Householder factorization is known as the *Golub-Kahan Bidiagonalization* (Golub and Kahan, 1965). We will finally bidiagonalize

$$
\boldsymbol{B} = \boldsymbol{H}_n\boldsymbol{H}_{n-1}\ldots\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top\boldsymbol{L}_2^\top\ldots\boldsymbol{L}_{n-2}^\top.
$$

And since the $\boldsymbol{H}_i$'s and $\boldsymbol{L}_i$'s are symmetric and orthogonal, we have

$$
\boldsymbol{B} = \boldsymbol{H}_n\boldsymbol{H}_{n-1}\ldots\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1\boldsymbol{L}_2\ldots\boldsymbol{L}_{n-2}.
$$

A full example of a 7×5 matrix is shown as follows where again $\boxtimes$ represents a value that is not necessarily zero, and **boldface** indicates the value has just been changed.

A Complete Example of Golub-Kahan Bidiagonalization

$$
\underset{\boldsymbol{A}}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes
\end{bmatrix}}
\stackrel{\boldsymbol{H}_1\times}{\rightarrow}
\underset{\boldsymbol{H}_1\boldsymbol{A}}{\begin{bmatrix}
\boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
\mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\stackrel{\times\boldsymbol{L}_1^\top}{\rightarrow}
\underset{\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top}{\begin{bmatrix}
\boxtimes & \boldsymbol{\boxtimes} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
$$

$$
\stackrel{\boldsymbol{H}_2\times}{\rightarrow}
\underset{\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes\\
0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\stackrel{\times\boldsymbol{L}_2^\top}{\rightarrow}
\underset{\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top\boldsymbol{L}_2^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & 0 & 0 & 0\\
0 & \boxtimes & \boldsymbol{\boxtimes} & \mathbf{0} & \mathbf{0}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}\\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
$$

$$
\stackrel{\boldsymbol{H}_3\times}{\rightarrow}
\underset{\boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top\boldsymbol{L}_2^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\stackrel{\times\boldsymbol{L}_3^\top}{\rightarrow}
\underset{\boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top\boldsymbol{L}_2^\top\boldsymbol{L}_3^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & 0 & 0 & 0 \\
\boxtimes & \boxtimes & \boxtimes & 0 & 0 \\
0 & \boxtimes & \boxtimes & \boldsymbol{\boxtimes} & \mathbf{0} \\
0 & 0 & \boxtimes & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes}
\end{bmatrix}}
$$

$$
\stackrel{\boldsymbol{H}_4\times}{\rightarrow}
\underset{\boldsymbol{H}_4\boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top\boldsymbol{L}_2\boldsymbol{L}_3^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boldsymbol{\boxtimes} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & \mathbf{0} & \boldsymbol{\boxtimes}
\end{bmatrix}}
\stackrel{\boldsymbol{H}_5\times}{\rightarrow}
\underset{\boldsymbol{H}_5\boldsymbol{H}_4\boldsymbol{H}_3\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{L}_1^\top\boldsymbol{L}_2\boldsymbol{L}_3^\top}{\begin{bmatrix}
\boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & \boxtimes & \boxtimes & \boxtimes \\
0 & 0 & 0 & \boxtimes & \boxtimes \\
0 & 0 & 0 & 0 & \boldsymbol{\boxtimes} \\
0 & 0 & 0 & 0 & \mathbf{0} \\
0 & 0 & 0 & 0 & \mathbf{0}
\end{bmatrix}}.
$$

We present in a way where a right Householder reflector $\boldsymbol{L}_i$ follows from a left one $\boldsymbol{H}_i$. However, a trivial error that might be employed is that we do the left ones altogether, and the right ones follow. That is, a bidiagonal decomposition is a combination of a QR decomposition and a Hessenberg decomposition. Nevertheless, this is problematic, the right Householder reflector $\boldsymbol{L}_1$ will destroy the zeros introduced by the left ones. Therefore, the left and right reflectors need to be employed in an interleaved manner to introduce back the zeros.

The Golub-Kahan bidiagonalization is not the most efficient way to calculate the bidiagonal decomposition. It requires $\sim 4mn^2 - \frac{4}{3}n^3$ flops to compute a bidiagonal decomposition of an $m \times n$ matrix with $m > n$. Further, if $\boldsymbol{U}, \boldsymbol{V}$ are needed explicitly, additional $\sim 4m^2n - 2mn^2 + 2n^3$ flops are required.

LHC Bidiagonalization Nevertheless, when $m \gg n$, we can extract the square triangular matrix (i.e., the QR decomposition) and apply the Golub-Kahan diagonalization on the square $n \times n$ matrix. This is known as the *Lawson-Hanson-Chan (LHC) bidiagonalization* (Lawson and Hanson, 1995; Chan, 1982) and the procedure is shown in Figure 14. The LHC bidiagonalization starts by computing the QR decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$. Then follows by applying the Golub-Kahan process such that $\widetilde{\boldsymbol{R}} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{B}}\boldsymbol{V}^\top$ where $\widetilde{\boldsymbol{R}}$ is the square $n \times n$ triangular submatrix inside $\boldsymbol{R}$. Append $\widetilde{\boldsymbol{U}}$ into

$$
\boldsymbol{U}_0 = \begin{bmatrix} \widetilde{\boldsymbol{U}} & \\ & \boldsymbol{I}_{m-n} \end{bmatrix},
$$

which results in $\boldsymbol{R} = \boldsymbol{U}_0\boldsymbol{B}\boldsymbol{V}^\top$ and $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{U}_0\boldsymbol{B}\boldsymbol{V}^\top$. Let $\boldsymbol{U} = \boldsymbol{Q}\boldsymbol{U}_0$, we obtain the bidiagonal decomposition. The QR decomposition requires $2mn^2 - \frac{2}{3}n^3$ flops and the Golub-Kahan process now requires $\frac{8}{3}n^3$ (operating on an $n \times n$ submatrix). Thus the total complexity to obtain the bidiagonal matrix $\boldsymbol{B}$ is then reduced to

$$
\text{LHC bidiagonalization: } \sim 2mn^2 + 2n^3 \text{ flops.}
$$

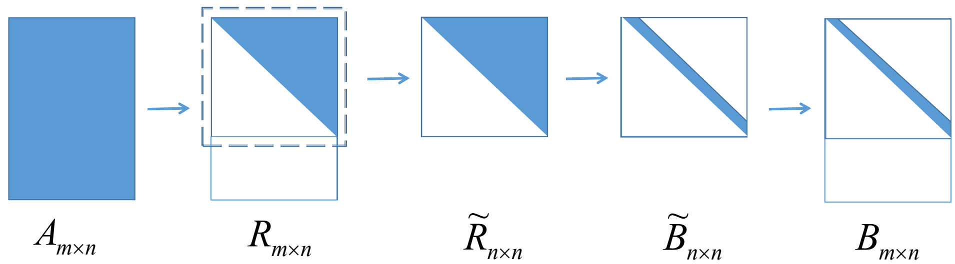


Figure 14: Demonstration of LHC-bidiagonalization of a matrix

The LHC process creates zeros and then destroys them again in the lower triangle of the upper $n \times n$ square of $\boldsymbol{R}$, but the zeros in the lower $(m-n) \times n$ rectangular matrix of $\boldsymbol{R}$ will be kept. Thus when $m-n$ is large enough (or $m \gg n$), there is a net gain. Simple calculations will show the LHC bidiagonalization costs less when $m > \frac{5}{3}n$ compared to the Golub-Kahan bidiagonalization.

Three-Step Bidiagonalization The LHC procedure is advantageous only when $m > \frac{5}{3}n$. A further trick is to apply the QR decomposition not at the beginning of the computation, but at a suitable point in the middle (Trefethen and Bau III, 1997). In particular, the procedure is shown in Figure 15 where we apply the first k steps of left and right Householder reflectors as in the Golub-Kahan process leaving the bottom-right $(m-k)\times(n-k)$ submatrix "unreflected". Then following the same LHC process on the submatrix to obtain the final bidiagonal decomposition. By doing so, the complexity reduces when $n < m < 2n$.

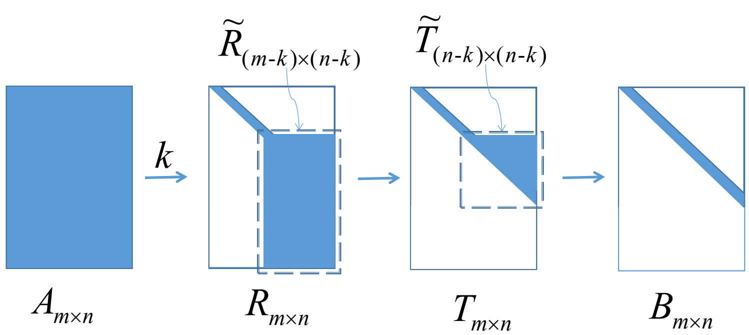


Figure 15: Demonstration of Three-Step bidiagonalization of a matrix

To conclude, the costs of the three methods are shown as follows:

$$
\begin{cases}
\text{Golub-Kahan: } \sim 4mn^2 - \dfrac{4}{3}n^3 \text{ flops,} \\
\text{LHC: } \sim 2mn^2 + 2n^3 \text{ flops,} \\
\text{Three-Step: } \sim 2mn^2 + 2m^2n - \dfrac{2}{3}m^3 - \dfrac{2}{3}n^3 \text{ flops.}
\end{cases}
$$

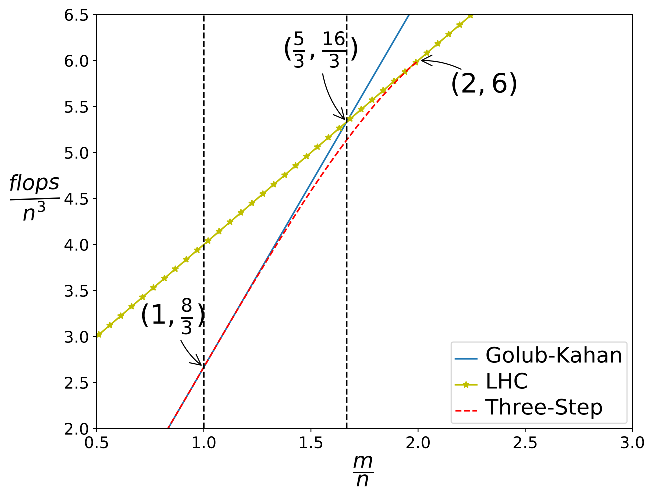


Figure 16: Comparison of the complexity among the three bidiagonal methods. When $m > 2n$, LHC is preferred; when $n < m < 2n$, the Three-Step method is preferred though the improvement is small enough.

When $m > 2n$, LHC is preferred; when $n < m < 2n$, the Three-Step method is preferred though the improvement is small enough as shown in Figure 16 where the operation counts for the three methods are plotted as a function of $\frac{m}{n}$. Notice that the complexity discussed here does not involve the extra computation of $\boldsymbol{U}, \boldsymbol{V}$. We shall not discuss the issue for simplicity.

10.2 Connection to Tridiagonal Decomposition

We fist illustrate the connection by the following lemma that reveals how to construct a tridiagonal matrix from a bidiagonal one.

Lemma 10.3: (Construct Tridiagonal From Bidiagonal)

Suppose $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ is upper bidiagonal, then $\boldsymbol{T}_1 = \boldsymbol{B}^\top\boldsymbol{B}$ and $\boldsymbol{T}_2 = \boldsymbol{B}\boldsymbol{B}^\top$ are *symmetric* triangular matrices.

The lemma above reveals an important property. Suppose $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^\top$ is the bidiagonal decomposition of $\boldsymbol{A}$, then the symmetric matrix $\boldsymbol{A}\boldsymbol{A}^\top$ has a tridiagonal decomposition

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{B}^\top\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{U}^\top.$$

And the symmetric matrix $\boldsymbol{A}^\top\boldsymbol{A}$ has a tridiagonal decomposition

$$\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{B}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{V}^\top.$$

As a final result in this section, we state a theorem giving the tridiagonal decomposition of a symmetric matrix with special eigenvalues.

Theorem 10.4: (Tridiagonal Decomposition for Nonnegative Eigenvalues)

Suppose $n \times n$ symmetric matrix $\boldsymbol{A}$ has nonnegative eigenvalues, then there exists a matrix $\boldsymbol{Z}$ such that

$$\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}^\top.$$

Moreover, the tridiagonal decomposition of $\boldsymbol{A}$ can be reduced to a problem to find the bidiagonal decomposition of $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^\top$ such that the tridiagonal decomposition of $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}^\top = \boldsymbol{U}\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{U}^\top.$$

Proof [of Theorem 10.4] The eigenvectors of symmetric matrices can be chosen to be orthogonal (Lemma 13.3, p. 114) such that symmetric matrix $\boldsymbol{A}$ can be decomposed into $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ (spectral theorem 13.1, p. 113) where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{A}$. When eigenvalues are nonnegative, $\boldsymbol{\Lambda}$ can be factored as $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}$. Let $\boldsymbol{Z} = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}$, $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}^\top$. Thus, combining our findings yields the result. ■

Part V
Eigenvalue Problem

11. Eigenvalue and Jordan Decomposition

Theorem 11.1: (Eigenvalue Decomposition)

Any square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with linearly independent eigenvectors can be factored as

$$\boldsymbol{A} = \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^{-1},$$

where $\boldsymbol{X}$ contains the eigenvectors of $\boldsymbol{A}$ as columns, and $\boldsymbol{\Lambda}$ is a diagonal matrix $diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are eigenvalues of $\boldsymbol{A}$.

Eigenvalue decomposition is also known as to diagonalize the matrix $\boldsymbol{A}$. When no eigenvalues of $\boldsymbol{A}$ are repeated, the eigenvectors are sure to be linearly independent. Then $\boldsymbol{A}$ can be diagonalized. Note here without n linearly independent eigenvectors, we cannot diagonalize. In Section 13.3 (p. 119), we will further discuss conditions under which the matrix has linearly independent eigenvectors.

11.1 Existence of the Eigenvalue Decomposition

Proof [of Theorem 11.1] Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ as the linearly independent eigenvectors of $\boldsymbol{A}$. Clearly, we have

$$\boldsymbol{A}\boldsymbol{x}_1 = \lambda_1\boldsymbol{x}_1, \qquad \boldsymbol{A}\boldsymbol{x}_2 = \lambda_2\boldsymbol{x}_2, \qquad \ldots, \qquad \boldsymbol{A}\boldsymbol{x}_n = \lambda_n\boldsymbol{x}_n.$$

In the matrix form,

$$\boldsymbol{A}\boldsymbol{X} = [\boldsymbol{A}\boldsymbol{x}_1, \boldsymbol{A}\boldsymbol{x}_2, \ldots, \boldsymbol{A}\boldsymbol{x}_n] = [\lambda_1\boldsymbol{x}_1, \lambda_2\boldsymbol{x}_2, \ldots, \lambda_n\boldsymbol{x}_n] = \boldsymbol{X}\boldsymbol{\Lambda}.$$

Since we assume the eigenvectors are linearly independent, then $\boldsymbol{X}$ has full rank and is invertible. We obtain

$$\boldsymbol{A} = \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^{-1}.$$

This completes the proof. ■

We will discuss some similar forms of eigenvalue decomposition in the spectral decomposition section, where the matrix $\boldsymbol{A}$ is required to be symmetric, and the $\boldsymbol{X}$ is not only nonsingular but also orthogonal. Or, the matrix $\boldsymbol{A}$ is required to be a *simple matrix*, that is, the algebraic multiplicity and geometric multiplicity are the same for $\boldsymbol{A}$, and $\boldsymbol{X}$ will be a trivial nonsingular matrix that may not contain the eigenvectors of $\boldsymbol{A}$. The decomposition also has a geometric meaning, which we will discuss in Section 15 (p. 148).

A matrix decomposition in the form of $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^{-1}$ has a nice property that we can compute the m-th power efficiently.

Remark 11.2: (m-th Power)

The m-th power of $\boldsymbol{A}$ is $\boldsymbol{A}^m = \boldsymbol{X}\boldsymbol{\Lambda}^m\boldsymbol{X}^{-1}$ if the matrix $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^{-1}$.

We notice that we require $\boldsymbol{A}$ have linearly independent eigenvectors to prove the existence of the eigenvalue decomposition. Under specific conditions, the requirement is intrinsically satisfied.

Lemma 11.3: (Different Eigenvalues)

Suppose the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ are all different. Then the corresponding eigenvectors are automatically independent. In another word, any square matrix with different eigenvalues can be diagonalized.

Proof [of Lemma 11.3] Suppose the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are all different, and the eigenvectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are dependent. That is, there exists a nonzero vector $\boldsymbol{c} = [c_1, c_2, \ldots, c_{n-1}]^\top$ such that

$$\boldsymbol{x}_n = \sum_{i=1}^{n-1} c_i \boldsymbol{x}_i.$$

Then we have

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{x}_n &= \boldsymbol{A}(\sum_{i=1}^{n-1} c_i \boldsymbol{x}_i) \\
&= c_1\lambda_1\boldsymbol{x}_1 + c_2\lambda_2\boldsymbol{x}_2 + \ldots + c_{n-1}\lambda_{n-1}\boldsymbol{x}_{n-1}.
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{x}_n &= \lambda_n \boldsymbol{x}_n \\
&= \lambda_n(c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 + \ldots + c_{n-1}\boldsymbol{x}_{n-1}).
\end{aligned}$$

Combine above two equations, we have

$$\sum_{i=1}^{n-1} (\lambda_n - \lambda_i) c_i \boldsymbol{x}_i = \boldsymbol{0}.$$

This leads to a contradiction since $\lambda_n \neq \lambda_i$ for all $i \in \{1, 2, \ldots, n-1\}$, from which the result follows. ■

Remark 11.4: (Limitation of Eigenvalue Decomposition)

The limitation of eigenvalue decomposition is that:

• The eigenvectors in $\boldsymbol{X}$ are usually not orthogonal and there are not always enough eigenvectors (i.e., some eigenvalues are equal).

• To compute the eigenvalues and eigenvectors $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ requires $\boldsymbol{A}$ to be square. Rectangular matrices cannot be diagonalized by eigenvalue decomposition.

11.2 Jordan Decomposition

In eigenvalue decomposition, we suppose matrix $\boldsymbol{A}$ has n linearly independent eigenvectors. However, this is not necessarily true for all square matrices. We introduce further a generalized version of eigenvalue decomposition which is called the Jordan decomposition named after Camille Jordan (Jordan, 1870).

We first introduce the definitin of Jordan blocks and Jordan form for the further description of Jordan decomposition.

Definition 11.5: (Jordan Block)

An $m \times m$ upper triangular matrix $B(\lambda, m)$ is called a Jordan block provided all m diagonal elements are the same eigenvalue λ and all upper sub-digonal elements are all ones:

$$B(\lambda, m) = \begin{bmatrix} \lambda & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & \lambda & 1 & \ldots & 0 & 0 & 0 \\ 0 & 0 & \lambda & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \ldots & \lambda & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & \lambda & 1 \\ 0 & 0 & 0 & \ldots & 0 & 0 & \lambda \end{bmatrix}_{m \times m}$$

Definition 11.6: (Jordan Form)

Given an $n \times n$ matrix $\boldsymbol{A}$, a Jordan form $\boldsymbol{J}$ for $\boldsymbol{A}$ is a block diagonal matrix defined as

$$\boldsymbol{J} = diag(B(\lambda_1, m_1), B(\lambda_2, m_2), \ldots B(\lambda_k, m_k))$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are eigenvalues of $\boldsymbol{A}$ (duplicates possible) and $m_1 + m_2 + \ldots + m_k = n$.

Then, the Jordan decomposition follows:

Theorem 11.7: (Jordan Decomposition)

Any square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{X}\boldsymbol{J}\boldsymbol{X}^{-1},$$

where $\boldsymbol{X}$ is a nonsingular matrix containing the generalized eigenvectors of $\boldsymbol{A}$ as columns, and $\boldsymbol{J}$ is a Jordan form matrix $diag(\boldsymbol{J}_1, \boldsymbol{J}_2, \ldots, \boldsymbol{J}_k)$ where

$$\boldsymbol{J}_i = \begin{bmatrix} \lambda_i & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & \lambda_i & 1 & \ldots & 0 & 0 & 0 \\ 0 & 0 & \lambda_i & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \ldots & \lambda_i & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & \lambda_i & 1 \\ 0 & 0 & 0 & \ldots & 0 & 0 & \lambda_i \end{bmatrix}_{m_i\times m_i}$$

is an $m_i \times m_i$ square matrix with m_i being the number of repetitions of eigenvalue λ_i and $m_1 + m_2 + \ldots + m_k = n$. $\boldsymbol{J}_i$'s are referred to as Jordan blocks.

Further, nonsingular matrix $\boldsymbol{X}$ is called the **matrix of generalized eigenvectors** of $\boldsymbol{A}$.

As an example, a Jordan form can have the following structure:

$$\begin{aligned} \boldsymbol{J} &= diag(B(\lambda_1, m_1), B(\lambda_2, m_2), \ldots, B(\lambda_k, m_k)) \\ &= \begin{bmatrix} \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 1 \\ 0 & 0 & \lambda_1 \end{bmatrix} & & & & \\ & \begin{bmatrix} \lambda_2 \end{bmatrix} & & & \\ & & \begin{bmatrix} \lambda_3 & 1 \\ 0 & \lambda_3 \end{bmatrix} & & \\ & & & \ddots & \\ & & & & \begin{bmatrix} \lambda_k & 1 \\ 0 & \lambda_k \end{bmatrix} \end{bmatrix}. \end{aligned}$$

Decoding a Jordan Decomposition: Note that zeros can appear on the upper sub-diagonal of $\boldsymbol{J}$ and in each block, the first column is always a diagonal containing only eigenvalues of $\boldsymbol{A}$. Take out one block to decode, without loss of generality, we take out the first block $\boldsymbol{J}_1$. We shall show the columns $1, 2, \ldots, m_1$ of $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{J}$ with $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x}_1 &= \lambda_1\boldsymbol{x}_1 \\ \boldsymbol{A}\boldsymbol{x}_2 &= \lambda_1\boldsymbol{x}_2 + \boldsymbol{x}_1 \\ \vdots &= \vdots \\ \boldsymbol{A}\boldsymbol{x}_{m_1} &= \lambda_1\boldsymbol{x}_{m_1} + \boldsymbol{x}_{m_1-1}. \end{aligned}$$

For more details about Jordan decomposition, please refer to (Gohberg and Goldberg, 1996; Hales and Passi, 1999).

The Jordan decomposition is not particularly interesting in practice as it is extremely sensitive to perturbation. Even with the smallest random change to a matrix , the matrix can be made diagonalizable (van de Geijn and Myers, 2020). As a result, there is no practical mathematical software library or tool that computes it. And the proof takes dozens of pages to discuss. For this reason, we leave the proof to interesting readers.

12. Schur Decomposition

Theorem 12.1: (Schur Decomposition)

Any square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with real eigenvalues can be factored as

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{U}\boldsymbol{Q}^\top,$$

where $\boldsymbol{Q}$ is an orthogonal matrix, and $\boldsymbol{U}$ is an upper triangular matrix. That is, all square matrix $\boldsymbol{A}$ with real eigenvalues can be triangularized.

A close look at Schur decomposition The first column of $\boldsymbol{AQ}$ and $\boldsymbol{QU}$ are $\boldsymbol{A}\boldsymbol{q}_1$ and $\boldsymbol{U}_{11}\boldsymbol{q}_1$. Then, $\boldsymbol{U}_{11}, \boldsymbol{q}_1$ are eigenvalue and eigenvector of $\boldsymbol{A}$. But other columns of $\boldsymbol{Q}$ need not be eigenvectors of $\boldsymbol{A}$.

Schur decomposition for symmetric matrices Symmetric matrix $\boldsymbol{A} = \boldsymbol{A}^\top$ leads to $\boldsymbol{QUQ}^\top = \boldsymbol{QU}^\top\boldsymbol{Q}^\top$. Then $\boldsymbol{U}$ is a diagonal matrix. And this diagonal matrix actually contains eigenvalues of $\boldsymbol{A}$. All the columns of $\boldsymbol{Q}$ are eigenvectors of $\boldsymbol{A}$. We conclude that all symmetric matrices are diagonalizable even with repeated eigenvalues.

12.1 Existence of the Schur Decomposition

To prove Theorem 12.1, we need to use the following lemmas.

Lemma 12.2: (Determinant Intermezzo)

We have the following properties for determinant of matrices:
- The determinant of multiplication of two matrices is $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$;
- The determinant of the transpose is $\det(\boldsymbol{A}^\top) = \det(\boldsymbol{A})$;
- Suppose matrix $\boldsymbol{A}$ has eigenvalue λ, then $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$;
- Determinant of any identity matrix is 1;
- Determinant of an orthogonal matrix $\boldsymbol{Q}$:

$$\det(\boldsymbol{Q}) = \det(\boldsymbol{Q}^\top) = \pm 1, \qquad \text{since } \det(\boldsymbol{Q}^\top)\det(\boldsymbol{Q}) = \det(\boldsymbol{Q}^\top\boldsymbol{Q}) = \det(\boldsymbol{I}) = 1;$$

- Any square matrix $\boldsymbol{A}$, we then have an orthogonal matrix $\boldsymbol{Q}$:

$$\det(\boldsymbol{A}) = \det(\boldsymbol{Q}^\top)\det(\boldsymbol{A})\det(\boldsymbol{Q}) = \det(\boldsymbol{Q}^\top\boldsymbol{AQ});$$

Lemma 12.3: (Submatrix with Same Eigenvalue)

Suppose square matrix $\boldsymbol{A}_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$ has real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{k+1}$. Then we can construct a $k \times k$ matrix $\boldsymbol{A}_k$ with eigenvalues $\lambda_2, \lambda_3, \ldots, \lambda_{k+1}$ by

$$\boldsymbol{A}_k = \begin{bmatrix} -\boldsymbol{p}_2^\top - \\ -\boldsymbol{p}_3^\top - \\ \vdots \\ -\boldsymbol{p}_{k+1}^\top - \end{bmatrix} \boldsymbol{A}_{k+1} \begin{bmatrix} \boldsymbol{p}_2 & \boldsymbol{p}_3 & \ldots & \boldsymbol{p}_{k+1} \end{bmatrix},$$

where $\boldsymbol{p}_1$ is a eigenvector of $\boldsymbol{A}_{k+1}$ with norm 1 corresponding to eigenvalue λ_1, and $\boldsymbol{p}_2, \boldsymbol{p}_3, \ldots, \boldsymbol{p}_{k+1}$ are any orthonormal vectors orthogonal to $\boldsymbol{p}_1$.

Proof [of Lemma 12.3] Let $\boldsymbol{P}_{k+1} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{k+1}]$. Then $\boldsymbol{P}_{k+1}^\top \boldsymbol{P}_{k+1} = \boldsymbol{I}$, and

$$\boldsymbol{P}_{k+1}^\top \boldsymbol{A}_{k+1} \boldsymbol{P}_{k+1} = \begin{bmatrix} \lambda_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_k \end{bmatrix}.$$

For any eigenvalue $\lambda = \{\lambda_2, \lambda_3, \ldots, \lambda_{k+1}\}$, by Lemma 12.2, we have

$$\begin{aligned} \det(\boldsymbol{A}_{k+1} - \lambda \boldsymbol{I}) &= \det(\boldsymbol{P}_{k+1}^\top (\boldsymbol{A}_{k+1} - \lambda \boldsymbol{I}) \boldsymbol{P}_{k+1}) \\ &= \det(\boldsymbol{P}_{k+1}^\top \boldsymbol{A}_{k+1} \boldsymbol{P}_{k+1} - \lambda \boldsymbol{P}_{k+1}^\top \boldsymbol{P}_{k+1}) \\ &= \det\left(\begin{bmatrix} \lambda_1 - \lambda & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_k - \lambda \boldsymbol{I} \end{bmatrix} \right) \\ &= (\lambda_1 - \lambda) \det(\boldsymbol{A}_k - \lambda \boldsymbol{I}). \end{aligned}$$

Where the last equality is from the fact that if matrix $\boldsymbol{M}$ has a block formulation: $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{E} & \boldsymbol{F} \\ \boldsymbol{G} & \boldsymbol{H} \end{bmatrix}$, then $\det(\boldsymbol{M}) = \det(\boldsymbol{E}) \det(\boldsymbol{H} - \boldsymbol{G}\boldsymbol{E}^{-1}\boldsymbol{F})$. Since λ is an eigenvalue of $\boldsymbol{A}$ and $\lambda \neq \lambda_1$, then $\det(\boldsymbol{A}_{k+1} - \lambda \boldsymbol{I}) = (\lambda_1 - \lambda) \det(\boldsymbol{A}_k - \lambda \boldsymbol{I}) = 0$ means λ is also an eigenvalue of $\boldsymbol{A}_k$. ■

We then prove the existence of the Schur decomposition by induction.

Proof [**of Theorem 12.1: Existence of Schur Decomposition**] We note that the theorem is trivial when $n = 1$ by setting $Q = 1$ and $U = A$. Suppose the theorem is true for $n = k$ for some $k \geq 1$. If we prove the theorem is also true for $n = k + 1$, then we complete the proof.

Suppose for $n = k$, the theorem is true for $\boldsymbol{A}_k = \boldsymbol{Q}_k \boldsymbol{U}_k \boldsymbol{Q}_k^\top$.

Suppose further $\boldsymbol{P}_{k+1}$ *contains orthogonal vectors* $\boldsymbol{P}_{k+1} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{k+1}]$ *as constructed in Lemma 12.3 where* $\boldsymbol{p}_1$ *is an eigenvector of* $\boldsymbol{A}_{k+1}$ *corresponding to eigenvalue* λ_1 *and its norm is 1,* $\boldsymbol{p}_2, \ldots, \boldsymbol{p}_{k+1}$ *are orthonormal to* $\boldsymbol{p}_1$. Let the other k eigenvalues of $\boldsymbol{A}_{k+1}$ be $\lambda_2, \lambda_3, \ldots, \lambda_{k+1}$. Since we suppose for $n = k$, the theorem is true, we can find a matrix $\boldsymbol{A}_k$ with eigenvalues $\lambda_2, \lambda_3, \ldots, \lambda_{k+1}$. So we have the following property by Lemma 12.3:

$$\boldsymbol{P}_{k+1}^\top \boldsymbol{A}_{k+1} \boldsymbol{P}_{k+1} = \begin{bmatrix} \lambda & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_k \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{A}_{k+1} \boldsymbol{P}_{k+1} = \boldsymbol{P}_{k+1} \begin{bmatrix} \lambda_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_k \end{bmatrix}.$$

Let $\boldsymbol{Q}_{k+1} = \boldsymbol{P}_{k+1} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_k \end{bmatrix}$. Then, it follows that

$$
\begin{aligned}
\boldsymbol{A}_{k+1}\boldsymbol{Q}_{k+1} &= \boldsymbol{A}_{k+1}\boldsymbol{P}_{k+1} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_k \end{bmatrix} \\
&= \boldsymbol{P}_{k+1} \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_k \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_k \end{bmatrix} \\
&= \boldsymbol{P}_{k+1} \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_k\boldsymbol{Q}_k \end{bmatrix} \\
&= \boldsymbol{P}_{k+1} \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_k\boldsymbol{U}_k \end{bmatrix} & \text{(By the assumption for } n=k) \\
&= \boldsymbol{P}_{k+1} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_k \end{bmatrix} \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{U}_k \end{bmatrix} \\
&= \boldsymbol{Q}_{k+1}\boldsymbol{U}_{k+1}. & \left(\boldsymbol{U}_{k+1} = \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{U}_k \end{bmatrix}\right)
\end{aligned}
$$

We then have $\boldsymbol{A}_{k+1} = \boldsymbol{Q}_{k+1}\boldsymbol{U}_{k+1}\boldsymbol{Q}_{k+1}^\top$, where $\boldsymbol{U}_{k+1}$ is an upper triangular matrix, and $\boldsymbol{Q}_{k+1}$ is an orthogonal matrix since $\boldsymbol{P}_{k+1}$ and $\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_k \end{bmatrix}$ are both orthogonal matrices. ■

12.2 Other Forms of the Schur Decomposition

From the proof of the Schur decomposition, we obtain the upper triangular matrix $\boldsymbol{U}_{k+1}$ by appending the eigenvalue λ_1 to $\boldsymbol{U}_k$. From this process, the values on the diagonal are always eigenvalues. Therefore, we can decompose the upper triangular into two parts.

Corollary 12.4: (Form 2 of Schur Decomposition)

Any square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with real eigenvalues can be factored as

$$
\boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{\Lambda} + \boldsymbol{T}, \qquad \textbf{or} \qquad \boldsymbol{A} = \boldsymbol{Q}(\boldsymbol{\Lambda}+\boldsymbol{T})\boldsymbol{Q}^\top,
$$

where $\boldsymbol{Q}$ is an orthogonal matrix, $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{T}$ is a *strictly upper triangular* matrix.

A strictly upper triangular matrix is an upper triangular matrix having 0's along the diagonal as well as the lower portion. Another proof for this decomposition is that $\boldsymbol{A}$ and $\boldsymbol{U}$ (where $\boldsymbol{U} = \boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{Q}$) are similar matrices so that they have the same eigenvalues (Lemma 8.5, p. 91). And the eigenvalues of any upper triangular matrices are on the diagonal. To see this, for any upper triangular matrix $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ where the diagonal values are r_{ii} for all $i \in \{1, 2, \ldots, n\}$. We have

$$
\boldsymbol{R}\boldsymbol{e}_i = r_{ii}\boldsymbol{e}_i,
$$

where $\boldsymbol{e}_i$ is the i-th basis vector in $\mathbb{R}^n$, i.e., $\boldsymbol{e}_i$ is the i-th column of the $n \times n$ identity matrix $\boldsymbol{I}_n$. So we can decompose $\boldsymbol{U}$ into $\boldsymbol{\Lambda}$ and $\boldsymbol{T}$.

A final observation on the second form of the Schur decomposition is shown as follows. From $\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\Lambda} + \boldsymbol{T})$, it follows that

$$\boldsymbol{A}\boldsymbol{q}_k = \lambda_k \boldsymbol{q}_k + \sum_{i=1}^{k-1} t_{ik}\boldsymbol{q}_i,$$

where t_{ik} is the (i,k)-th entry of $\boldsymbol{T}$. The form is quite close to the eigenvalue decomposition. Nevertheless, the columns become orthonormal bases and the orthonormal bases are correlated.

13. Spectral Decomposition (Theorem)

Theorem 13.1: (Spectral Decomposition)

A real matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric if and only if there exists an orthogonal matrix $\boldsymbol{Q}$ and a diagonal matrix $\boldsymbol{\Lambda}$ such that

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top,$$

where the columns of $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n]$ are eigenvectors of $\boldsymbol{A}$ and are mutually orthonormal, and the entries of $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ are the corresponding eigenvalues of $\boldsymbol{A}$, which are real. And the rank of $\boldsymbol{A}$ is the number of nonzero eigenvalues. This is known as the **spectral decomposition** or **spectral theorem** of real symmetric matrix $\boldsymbol{A}$. Specifically, we have the following properties:

1. A symmetric matrix has only **real eigenvalues**;
2. The eigenvectors are orthogonal such that they can be chosen **orthonormal** by normalization;
3. The rank of $\boldsymbol{A}$ is the number of nonzero eigenvalues;
4. If the eigenvalues are distinct, the eigenvectors are unique as well.

The above decomposition is called the spectral decomposition for real symmetric matrices and is often known as the *spectral theorem*.

Spectral theorem vs eigenvalue decomposition In the eigenvalue decomposition, we require the matrix $\boldsymbol{A}$ to be square and the eigenvectors to be linearly independent. Whereas in the spectral theorem, any symmetric matrix can be diagonalized, and the eigenvectors are chosen to be orthogonormal.

A word on the spectral decomposition In Lemma 8.5 (p. 91), we proved that the eigenvalues of similar matrices are the same. From the spectral decomposition, we notice that $\boldsymbol{A}$ and $\boldsymbol{\Lambda}$ are similar matrices such that their eigenvalues are the same. For any diagonal matrices, the eigenvalues are the diagonal components. [19] To see this, we realize that

$$\boldsymbol{\Lambda}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$ is the i-th basis vector. Therefore, the matrix Λ contains the eigenvalues of $\boldsymbol{A}$.

19. Actually, we have shown in the last section that the diagonal values for triangular matrices are the eigenvalues of it.

13.1 Existence of the Spectral Decomposition

We prove the theorem in several steps.

Symmetric Matrix Property 1 of 4

Lemma 13.2: (Real Eigenvalues)

The eigenvalues of any symmetric matrix are all real.

Proof [of Lemma 13.2] Suppose eigenvalue λ is a complex number $\lambda = a + ib$ where a, b are real. Its complex conjugate is $\bar{\lambda} = a - ib$. Same for complex eigenvector $\boldsymbol{x} = \boldsymbol{c} + i\boldsymbol{d}$ and its complex conjugate $\bar{\boldsymbol{x}} = \boldsymbol{c} - i\boldsymbol{d}$ where $\boldsymbol{c}, \boldsymbol{d}$ are real vectors. We then have the following property

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x} \qquad \underrightarrow{\text{ leads to }} \qquad \boldsymbol{A}\bar{\boldsymbol{x}} = \bar{\lambda}\bar{\boldsymbol{x}} \qquad \underrightarrow{\text{ transpose to }} \qquad \bar{\boldsymbol{x}}^\top \boldsymbol{A} = \bar{\lambda}\bar{\boldsymbol{x}}^\top.$$

We take the dot product of the first equation with $\bar{\boldsymbol{x}}$ and the last equation with $\boldsymbol{x}$:

$$\bar{\boldsymbol{x}}^\top \boldsymbol{A}\boldsymbol{x} = \lambda \bar{\boldsymbol{x}}^\top \boldsymbol{x}, \qquad \text{and} \qquad \bar{\boldsymbol{x}}^\top \boldsymbol{A}\boldsymbol{x} = \bar{\lambda}\bar{\boldsymbol{x}}^\top \boldsymbol{x}.$$

Then we have the equality $\lambda \bar{\boldsymbol{x}}^\top \boldsymbol{x} = \bar{\lambda}\bar{\boldsymbol{x}}^\top \boldsymbol{x}$. Since $\bar{\boldsymbol{x}}^\top \boldsymbol{x} = (\boldsymbol{c} - i\boldsymbol{d})^\top(\boldsymbol{c} + i\boldsymbol{d}) = \boldsymbol{c}^\top\boldsymbol{c} + \boldsymbol{d}^\top\boldsymbol{d}$ is a real number. Therefore the imaginary part of λ is zero and λ is real. ■

Symmetric Matrix Property 2 of 4

Lemma 13.3: (Orthogonal Eigenvectors)

The eigenvectors corresponding to distinct eigenvalues of any symmetric matrix are orthogonal so that we can normalize eigenvectors to make them orthonormal since $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x} \underrightarrow{\text{ leads to }} \boldsymbol{A}\frac{\boldsymbol{x}}{||\boldsymbol{x}||} = \lambda \frac{\boldsymbol{x}}{||\boldsymbol{x}||}$ which corresponds to the same eigenvalue.

Proof [of Lemma 13.3] Suppose eigenvalues λ_1, λ_2 correspond to eigenvectors $\boldsymbol{x}_1, \boldsymbol{x}_2$ so that $\boldsymbol{A}\boldsymbol{x}_1 = \lambda \boldsymbol{x}_1$ and $\boldsymbol{A}\boldsymbol{x}_2 = \lambda_2 \boldsymbol{x}_2$. We have the following equality:

$$\boldsymbol{A}\boldsymbol{x}_1 = \lambda_1 \boldsymbol{x}_1 \qquad \underrightarrow{\text{ leads to }} \qquad \boldsymbol{x}_1^\top \boldsymbol{A} = \lambda_1 \boldsymbol{x}_1^\top \qquad \underrightarrow{\text{ leads to }} \qquad \boldsymbol{x}_1^\top \boldsymbol{A}\boldsymbol{x}_2 = \lambda_1 \boldsymbol{x}_1^\top \boldsymbol{x}_2,$$

and

$$\boldsymbol{A}\boldsymbol{x}_2 = \lambda_2 \boldsymbol{x}_2 \qquad \underrightarrow{\text{ leads to }} \qquad \boldsymbol{x}_1^\top \boldsymbol{A}\boldsymbol{x}_2 = \lambda_2 \boldsymbol{x}_1^\top \boldsymbol{x}_2,$$

which implies $\lambda_1 \boldsymbol{x}_1^\top \boldsymbol{x}_2 = \lambda_2 \boldsymbol{x}_1^\top \boldsymbol{x}_2$. Since eigenvalues $\lambda_1 \neq \lambda_2$, the eigenvectors are orthogonal. ■

In the above Lemma 13.3, we prove that the eigenvectors corresponding to distinct eigenvalues of symmetric matrices are orthogonal. More generally, we prove the important theorem that eigenvectors corresponding to distinct eigenvalues of any matrix are linearly independent.

Theorem 13.4: (Independent Eigenvector Theorem)

If a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has k distinct eigenvalues, then any set of k corresponding eigenvectors are linearly independent.

Proof [of Theorem 13.4] We will prove by induction. Firstly, we will prove that any two eigenvectors corresponding to distinct eigenvalues are linearly independent. Suppose $\boldsymbol{v}_1, \boldsymbol{v}_2$ correspond to distinct eigenvalues λ_1 and λ_2 respectively. Suppose further there exists a nonzero vector $\boldsymbol{x} = [x_1, x_2] \neq \mathbf{0}$ that

$$x_1\boldsymbol{v}_1 + x_2\boldsymbol{v}_2 = \mathbf{0}. \tag{13.1}$$

That is, $\boldsymbol{v}_1, \boldsymbol{v}_2$ are linearly independent. Multiply Equation (13.1) on the left by $\boldsymbol{A}$, we get

$$x_1\lambda_1\boldsymbol{v}_1 + x_2\lambda_2\boldsymbol{v}_2 = \mathbf{0}. \tag{13.2}$$

Multiply Equation (13.1) on the left by λ_2, we get

$$x_1\lambda_2\boldsymbol{v}_1 + x_2\lambda_2\boldsymbol{v}_2 = \mathbf{0}. \tag{13.3}$$

Subtract Equation (13.2) from Equation (13.3) to find

$$x_1(\lambda_2 - \lambda_1)\boldsymbol{v}_1 = \mathbf{0}.$$

Since $\lambda_2 \neq \lambda_1$, $\boldsymbol{v}_1 \neq \mathbf{0}$, we must have $x_1 = 0$. From Equation (13.1), $\boldsymbol{v}_2 \neq \mathbf{0}$, we must also have $x_2 = 0$ which arrives at a contradiction. Thus $\boldsymbol{v}_1, \boldsymbol{v}_2$ are linearly independent.

Now, suppose any $j < k$ eigenvectors are linearly independent, if we could prove that any $j+1$ eigenvectors are also linearly independent, we finish the proof. Suppose $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_j$ are linearly independent and $\boldsymbol{v}_{j+1}$ is dependent on the first j eigenvectors. That is, there exists a nonzero vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_j] \neq \mathbf{0}$ that

$$\boldsymbol{v}_{j+1} = x_1\boldsymbol{v}_1 + x_2\boldsymbol{v}_2 + \ldots + x_j\boldsymbol{v}_j. \tag{13.4}$$

Suppose the $j+1$ eigenvectors correspond to distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_j, \lambda_{j+1}$. Multiply Equation (13.4) on the left by $\boldsymbol{A}$, we get

$$\lambda_{j+1}\boldsymbol{v}_{j+1} = x_1\lambda_1\boldsymbol{v}_1 + x_2\lambda_2\boldsymbol{v}_2 + \ldots + x_j\lambda_j\boldsymbol{v}_j. \tag{13.5}$$

Multiply Equation (13.4) on the left by λ_{j+1}, we get

$$\lambda_{j+1}\boldsymbol{v}_{j+1} = x_1\lambda_{j+1}\boldsymbol{v}_1 + x_2\lambda_{j+1}\boldsymbol{v}_2 + \ldots + x_j\lambda_{j+1}\boldsymbol{v}_j. \tag{13.6}$$

Subtract Equation (13.6) from Equation (13.5), we find

$$x_1(\lambda_{j+1} - \lambda_1)\boldsymbol{v}_1 + x_2(\lambda_{j+1} - \lambda_2)\boldsymbol{v}_2 + \ldots + x_j(\lambda_{j+1} - \lambda_j)\boldsymbol{v}_j = \mathbf{0}.$$

From assumption, $\lambda_{j+1} \neq \lambda_i$ for all $i \in \{1, 2, \ldots, j\}$, and $\boldsymbol{v}_i \neq \mathbf{0}$ for all $i \in \{1, 2, \ldots, j\}$. We must have $x_1 = x_2 = \ldots = x_j = 0$ which leads to a contradiction. Then $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_j, \boldsymbol{v}_{j+1}$ are linearly independent. This completes the proof. ■

Corollary 13.5: (Independent Eigenvector Theorem, CNT.)

If a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has n distinct eigenvalues, then any set of n corresponding eigenvectors form a basis for $\mathbb{R}^n$.

Symmetric Matrix Property 3 of 4

Lemma 13.6: (Orthonormal Eigenvectors for Duplicate Eigenvalue)

If $\boldsymbol{A}$ has a duplicate eigenvalue λ_i with multiplicity $k \geq 2$, then there exist k orthonormal eigenvectors corresponding to λ_i.

Proof [of Lemma 13.6] We note that there is at least one eigenvector $\boldsymbol{x}_{i1}$ corresponding to λ_i. And for such eigenvector $\boldsymbol{x}_{i1}$, we can always find additional $n-1$ orthonormal vectors $\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_n$ so that $\{\boldsymbol{x}_{i1}, \boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_n\}$ forms an orthonormal basis in $\mathbb{R}^n$. Put the $\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_n$ into matrix $\boldsymbol{Y}_1$ and $\{\boldsymbol{x}_{i1}, \boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_n\}$ into matrix $\boldsymbol{P}_1$

$$\boldsymbol{Y}_1 = [\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_n] \qquad \text{and} \qquad \boldsymbol{P}_1 = [\boldsymbol{x}_{i1}, \boldsymbol{Y}_1].$$

We then have

$$\boldsymbol{P}_1^\top \boldsymbol{A} \boldsymbol{P}_1 = \begin{bmatrix} \lambda_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Y}_1^\top \boldsymbol{A} \boldsymbol{Y}_1 \end{bmatrix}.$$

As a result, $\boldsymbol{A}$ and $\boldsymbol{P}_1^\top \boldsymbol{A} \boldsymbol{P}_1$ are similar matrices such that they have the same eigenvalues since $\boldsymbol{P}_1$ is nonsingular (even orthogonal here, see Lemma 8.5, p. 91). We obtain

$$\det(\boldsymbol{P}_1^\top \boldsymbol{A} \boldsymbol{P}_1 - \lambda \boldsymbol{I}_n) \stackrel{20}{=} (\lambda_i - \lambda) \det(\boldsymbol{Y}_1^\top \boldsymbol{A} \boldsymbol{Y}_1 - \lambda \boldsymbol{I}_{n-1}).$$

If λ_i has multiplicity $k \geq 2$, then the term $(\lambda_i - \lambda)$ occurs k times in the polynomial from the determinant $\det(\boldsymbol{P}_1^\top \boldsymbol{A} \boldsymbol{P}_1 - \lambda \boldsymbol{I}_n)$, i.e., the term occurs $k-1$ times in the polynomial from $\det(\boldsymbol{Y}_1^\top \boldsymbol{A} \boldsymbol{Y}_1 - \lambda \boldsymbol{I}_{n-1})$. In another word, $\det(\boldsymbol{Y}_1^\top \boldsymbol{A} \boldsymbol{Y}_1 - \lambda_i \boldsymbol{I}_{n-1}) = 0$ and λ_i is an eigenvalue of $\boldsymbol{Y}_1^\top \boldsymbol{A} \boldsymbol{Y}_1$.

Let $\boldsymbol{B} = \boldsymbol{Y}_1^\top \boldsymbol{A} \boldsymbol{Y}_1$. Since $\det(\boldsymbol{B} - \lambda_i \boldsymbol{I}_{n-1}) = 0$, the null space of $\boldsymbol{B} - \lambda_i \boldsymbol{I}_{n-1}$ is not none. Suppose $(\boldsymbol{B} - \lambda_i \boldsymbol{I}_{n-1})\boldsymbol{n} = \boldsymbol{0}$, i.e., $\boldsymbol{B}\boldsymbol{n} = \lambda_i \boldsymbol{n}$ and $\boldsymbol{n}$ is an eigenvector of $\boldsymbol{B}$.

From $\boldsymbol{P}_1^\top \boldsymbol{A} \boldsymbol{P}_1 = \begin{bmatrix} \lambda_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{bmatrix}$, we have $\boldsymbol{A}\boldsymbol{P}_1 \begin{bmatrix} z \\ \boldsymbol{n} \end{bmatrix} = \boldsymbol{P}_1 \begin{bmatrix} \lambda_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} z \\ \boldsymbol{n} \end{bmatrix}$, where z is any scalar. From the left side of this equation, we have

$$\begin{aligned} \boldsymbol{A}\boldsymbol{P}_1 \begin{bmatrix} z \\ \boldsymbol{n} \end{bmatrix} &= \begin{bmatrix} \lambda_i \boldsymbol{x}_{i1}, \boldsymbol{A}\boldsymbol{Y}_1 \end{bmatrix} \begin{bmatrix} z \\ \boldsymbol{n} \end{bmatrix} \\ &= \lambda_i z \boldsymbol{x}_{i1} + \boldsymbol{A}\boldsymbol{Y}_1 \boldsymbol{n}. \end{aligned} \tag{13.7}$$

20. By the fact that if matrix $\boldsymbol{M}$ has a block formulation: $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$, then $\det(\boldsymbol{M}) = \det(\boldsymbol{A}) \det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})$.

And from the right side of the equation, we have

$$
\begin{aligned}
\boldsymbol{P}_1 \begin{bmatrix} \lambda_i & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} z \\ \boldsymbol{n} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{x}_{i1} & \boldsymbol{Y}_1 \end{bmatrix} \begin{bmatrix} \lambda_i & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} z \\ \boldsymbol{n} \end{bmatrix} \\
&= \begin{bmatrix} \lambda_i \boldsymbol{x}_{i1} & \boldsymbol{Y}_1 \boldsymbol{B} \end{bmatrix} \begin{bmatrix} z \\ \boldsymbol{n} \end{bmatrix} \\
&= \lambda_i z \boldsymbol{x}_{i1} + \boldsymbol{Y}_1 \boldsymbol{B} \boldsymbol{n} \\
&= \lambda_i z \boldsymbol{x}_{i1} + \lambda_i \boldsymbol{Y}_1 \boldsymbol{n}. \qquad (\text{Since } \boldsymbol{B}\boldsymbol{n} = \lambda_i \boldsymbol{n})
\end{aligned} \tag{13.8}
$$

Combine Equation (13.8) and Equation (13.7), we obtain

$$
\boldsymbol{A}\boldsymbol{Y}_1\boldsymbol{n} = \lambda_i \boldsymbol{Y}_1 \boldsymbol{n},
$$

which means $\boldsymbol{Y}_1\boldsymbol{n}$ is an eigenvector of $\boldsymbol{A}$ corresponding to the eigenvalue λ_i (same eigenvalue corresponding to $\boldsymbol{x}_{i1}$). Since $\boldsymbol{Y}_1\boldsymbol{n}$ is a combination of $\boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_n$ which are orthonormal to $\boldsymbol{x}_{i1}$, the $\boldsymbol{Y}_1\boldsymbol{n}$ can be chosen to be orthonormal to $\boldsymbol{x}_{i1}$.

To conclude, if we have one eigenvector $\boldsymbol{x}_{i1}$ corresponding to λ_i whose multiplicity is $k \geq 2$, we could construct the second eigenvector by choosing one vector from the null space of $(\boldsymbol{B} - \lambda_i \boldsymbol{I}_{n-1})$ constructed above. Suppose now, we have constructed the second eigenvector $\boldsymbol{x}_{i2}$ which is orthonormal to $\boldsymbol{x}_{i1}$. For such eigenvectors $\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}$, we can always find additional $n-2$ orthonormal vectors $\boldsymbol{y}_3, \boldsymbol{y}_4, \ldots, \boldsymbol{y}_n$ so that $\{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{y}_3, \boldsymbol{y}_4, \ldots, \boldsymbol{y}_n\}$ forms an orthonormal basis in $\mathbb{R}^n$. Put the $\boldsymbol{y}_3, \boldsymbol{y}_4, \ldots, \boldsymbol{y}_n$ into matrix $\boldsymbol{Y}_2$ and $\{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{y}_3, \boldsymbol{y}_4, \ldots, \boldsymbol{y}_n\}$ into matrix $\boldsymbol{P}_2$:

$$
\boldsymbol{Y}_2 = [\boldsymbol{y}_3, \boldsymbol{y}_4, \ldots, \boldsymbol{y}_n] \qquad \text{and} \qquad \boldsymbol{P}_2 = [\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{Y}_1].
$$

We then have

$$
\boldsymbol{P}_2^\top \boldsymbol{A} \boldsymbol{P}_2 = \begin{bmatrix} \lambda_i & 0 & \mathbf{0} \\ 0 & \lambda_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Y}_2^\top \boldsymbol{A} \boldsymbol{Y}_2 \end{bmatrix} = \begin{bmatrix} \lambda_i & 0 & \mathbf{0} \\ 0 & \lambda_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{C} \end{bmatrix},
$$

where $\boldsymbol{C} = \boldsymbol{Y}_2^\top \boldsymbol{A} \boldsymbol{Y}_2$ such that $\det(\boldsymbol{P}_2^\top \boldsymbol{A} \boldsymbol{P}_2 - \lambda \boldsymbol{I}_n) = (\lambda_i - \lambda)^2 \det(\boldsymbol{C} - \lambda \boldsymbol{I}_{n-2})$. If the multiplicity of λ_i is $k \geq 3$, $\det(\boldsymbol{C} - \lambda_i \boldsymbol{I}_{n-2}) = 0$ and the null space of $\boldsymbol{C} - \lambda_i \boldsymbol{I}_{n-2}$ is not none so that we can still find a vector from null space of $\boldsymbol{C} - \lambda_i \boldsymbol{I}_{n-2}$ and $\boldsymbol{C}\boldsymbol{n} = \lambda_i \boldsymbol{n}$. Now we can construct a vector $\begin{bmatrix} z_1 \\ z_2 \\ \boldsymbol{n} \end{bmatrix} \in \mathbb{R}^n$, where z_1, z_2 are any scalar values, such that

$$
\boldsymbol{A}\boldsymbol{P}_2 \begin{bmatrix} z_1 \\ z_2 \\ \boldsymbol{n} \end{bmatrix} = \boldsymbol{P}_2 \begin{bmatrix} \lambda_i & 0 & \mathbf{0} \\ 0 & \lambda_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{C} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \boldsymbol{n} \end{bmatrix}.
$$

Similarly, from the left side of the above equation, we will get $\lambda_i z_1 \boldsymbol{x}_{i1} + \lambda_i z_2 \boldsymbol{x}_{i2} + \boldsymbol{A}\boldsymbol{Y}_2\boldsymbol{n}$. From the right side of the above equation, we will get $\lambda_i z_1 \boldsymbol{x}_{i1} + \lambda_i z_2 \boldsymbol{x}_{i2} + \lambda_i \boldsymbol{Y}_2 \boldsymbol{n}$. As a result,

$$
\boldsymbol{A}\boldsymbol{Y}_2\boldsymbol{n} = \lambda_i \boldsymbol{Y}_2 \boldsymbol{n},
$$

where $\boldsymbol{Y}_2\boldsymbol{n}$ is an eigenvector of $\boldsymbol{A}$ and orthogonal to $\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}$. And it is easy to construct the eigenvector to be orthonormal to the first two.

The process can go on, and finally, we will find k orthonormal eigenvectors corresponding to λ_i.

Actually, the dimension of the null space of $\boldsymbol{P}_1^\top \boldsymbol{A}\boldsymbol{P}_1 - \lambda_i \boldsymbol{I}_n$ is equal to the multiplicity k. It also follows that if the multiplicity of λ_i is k, there cannot be more than k orthogonal eigenvectors corresponding to λ_i. Otherwise, it will come to the conclusion that we could find more than n orthogonal eigenvectors which leads to a contradiction. ■

The proof of the existence of the spectral decomposition is trivial from the lemmas above. Also, we can use Schur decomposition to prove the existence of it.

Proof [**of Theorem 13.1: Existence of Spectral Decomposition**] From the Schur decomposition in Theorem 12.1 (p. 110), symmetric matrix $\boldsymbol{A} = \boldsymbol{A}^\top$ leads to $\boldsymbol{Q}\boldsymbol{U}\boldsymbol{Q}^\top = \boldsymbol{Q}\boldsymbol{U}^\top\boldsymbol{Q}^\top$. Then $\boldsymbol{U}$ is a diagonal matrix. And this diagonal matrix actually contains eigenvalues of $\boldsymbol{A}$. All the columns of $\boldsymbol{Q}$ are eigenvectors of $\boldsymbol{A}$. We conclude that all symmetric matrices are diagonalizable even with repeated eigenvalues. ■

For any matrix multiplication, we have the rank of the multiplication result no larger than the rank of the inputs. However, the symmetric matrix $\boldsymbol{A}^\top\boldsymbol{A}$ is rather special in that the rank of $\boldsymbol{A}^\top\boldsymbol{A}$ is equal to that of $\boldsymbol{A}$ which will be used in the proof of singular value decomposition in the next section.

Lemma 13.7: (Rank of $\boldsymbol{AB}$)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times k}$, then the matrix multiplication $\boldsymbol{AB} \in \mathbb{R}^{m\times k}$ has $rank(\boldsymbol{AB}) \leq \min(rank(\boldsymbol{A}), rank(\boldsymbol{B}))$.

Proof [of Lemma 13.7] For matrix multiplication $\boldsymbol{AB}$, we have

- All rows of $\boldsymbol{AB}$ are combinations of rows of $\boldsymbol{B}$, the row space of $\boldsymbol{AB}$ is a subset of the row space of $\boldsymbol{B}$. Thus $rank(\boldsymbol{AB}) \leq rank(\boldsymbol{B})$.
- All columns of $\boldsymbol{AB}$ are combinations of columns of $\boldsymbol{A}$, the column space of $\boldsymbol{AB}$ is a subset of the column space of $\boldsymbol{A}$. Thus $rank(\boldsymbol{AB}) \leq rank(\boldsymbol{A})$.

Therefore, $rank(\boldsymbol{AB}) \leq \min(rank(\boldsymbol{A}), rank(\boldsymbol{B}))$. ■

Symmetric Matrix Property 4 of 4

Lemma 13.8: (Rank of Symmetric Matrices)

If $\boldsymbol{A}$ is an $n\times n$ real symmetric matrix, then rank$(\boldsymbol{A})$ = the total number of nonzero eigenvalues of $\boldsymbol{A}$. In particular, $\boldsymbol{A}$ has full rank if and only if $\boldsymbol{A}$ is nonsingular. Further, $\mathcal{C}(\boldsymbol{A})$ is the linear space spanned by the eigenvectors of $\boldsymbol{A}$ that correspond to nonzero eigenvalues.

Proof [of Lemma 13.8] For any symmetric matrix $\boldsymbol{A}$, we have $\boldsymbol{A}$, in spectral form, as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ and also $\boldsymbol{\Lambda} = \boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{Q}$. Since we have shown in Lemma 13.7 that the rank of the multiplication $rank(\boldsymbol{AB})\leq\min(rank(\boldsymbol{A}),\, rank(\boldsymbol{B}))$.

- From $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, we have $rank(\boldsymbol{A}) \leq rank(\boldsymbol{Q}\boldsymbol{\Lambda}) \leq rank(\boldsymbol{\Lambda})$;
- From $\boldsymbol{\Lambda} = \boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{Q}$, we have $rank(\boldsymbol{\Lambda}) \leq rank(\boldsymbol{Q}^\top\boldsymbol{A}) \leq rank(\boldsymbol{A})$,

The inequalities above give us a contradiction. And thus $rank(\boldsymbol{A}) = rank(\boldsymbol{\Lambda})$ which is the total number of nonzero eigenvalues.

Since $\boldsymbol{A}$ is nonsingular if and only if all of its eigenvalues are nonzero, $\boldsymbol{A}$ has full rank if and only if $\boldsymbol{A}$ is nonsingular. ■

Similar to the eigenvalue decomposition, we can compute the m-th power of matrix $\boldsymbol{A}$ via the spectral decomposition more efficiently.

Remark 13.9: (m-th Power)

The m-th power of $\boldsymbol{A}$ is $\boldsymbol{A}^m = \boldsymbol{Q}\boldsymbol{\Lambda}^m\boldsymbol{Q}^\top$ if the matrix $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$.

13.2 Uniqueness of Spectral Decomposition

Clearly, the spectral decomposition is not unique essentially because of the multiplicity of eigenvalues. One can imagine that eigenvalue λ_i and λ_j are the same for some $1 \leq i, j \leq n$, and interchange the corresponding eigenvectors in $\boldsymbol{Q}$ will have the same results but the decompositions are different. But the *eigenspaces* (i.e., the null space $\mathcal{N}(\boldsymbol{A} - \lambda_i\boldsymbol{I})$ for eigenvalue λ_i) corresponding to each eigenvalue are fixed. So there is a unique decomposition in terms of eigenspaces and then any orthonormal basis of these eigenspaces can be chosen.

13.3 Other Forms, Connecting Eigenvalue Decomposition*

In this section, we discuss other forms of the spectral decomposition under different conditions.

Definition 13.10: (Characteristic Polynomial)

For any square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the **characteristic polynomial** $\det(\boldsymbol{A} - \lambda\boldsymbol{I})$ is given by

$$
\begin{aligned}
\det(\lambda\boldsymbol{I} - \boldsymbol{A}) &= \lambda^n - \gamma_{n-1}\lambda^{n-1} + \ldots + \gamma_1\lambda + \gamma_0 \\
&= (\lambda - \lambda_1)^{k_1}(\lambda - \lambda_2)^{k_2}\ldots(\lambda - \lambda_m)^{k_m},
\end{aligned}
$$

where $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the distinct roots of $\det(\lambda\boldsymbol{I} - \boldsymbol{A})$ and also the eigenvalues of $\boldsymbol{A}$, and $k_1 + k_2 + \ldots + k_m = n$, i.e., $\det(\lambda\boldsymbol{I} - \boldsymbol{A})$ is a polynomial of degree n for any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ (see proof of Lemma 13.6, p. 116).

An important multiplicity arises from the characteristic polynomial of a matrix is then defined as follows:

Definition 13.11: (Algebraic Multiplicity and Geometric Multiplicity)

Given the characteristic polynomial of matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$:

$$\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = (\lambda - \lambda_1)^{k_1}(\lambda - \lambda_2)^{k_2} \ldots (\lambda - \lambda_m)^{k_m}.$$

The integer k_i is called the **algebraic multiplicity** of the eigenvalue λ_i, i.e., the algebraic multiplicity of eigenvalue λ_i is equal to the multiplicity of the corresponding root of the characteristic polynomial.

The **eigenspace associated to eigenvalue** λ_i is defined by the null space of $(\boldsymbol{A} - \lambda_i \boldsymbol{I})$, i.e., $\mathcal{N}(\boldsymbol{A} - \lambda_i \boldsymbol{I})$.

And the dimension of the eigenspace associated to λ_i, $\mathcal{N}(\boldsymbol{A} - \lambda_i \boldsymbol{I})$, is called the **geometric multiplicity** of λ_i.

In short, we denote the algebraic multiplicity of λ_i by $alg(\lambda_i)$, and its geometric multiplicity by $geo(\lambda_i)$.

Remark 13.12: (Geometric Multiplicity)

Note that for matrix $\boldsymbol{A}$ and the eigenspace $\mathcal{N}(\boldsymbol{A} - \lambda_i \boldsymbol{I})$, the dimension of the eigenspace is also the number of linearly independent eigenvectors of $\boldsymbol{A}$ associated to λ_i, namely a basis for the eigenspace. This implies that while there are an infinite number of eigenvectors associated with each eigenvalue λ_i, the fact that they form a subspace (provided the zero vector is added) means that they can be described by a finite number of vectors.

By definition, the sum of the algebraic multiplicities is equal to n, but the sum of the geometric multiplicities can be strictly smaller.

Corollary 13.13: (Multiplicity in Similar Matrices)

Similar matrices have same algebraic multiplicities and geometric multiplicities.

Proof [of Corollary 13.13] In Lemma 8.5 (p. 91), we proved that the eigenvalues of similar matrices are the same, therefore, the algebraic multiplicities of similar matrices are the same as well.

Suppose $\boldsymbol{A}$ and $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$ are similar matrices where $\boldsymbol{P}$ is nonsingular. And the geometric multiplicity of an eigenvalue of $\boldsymbol{A}$, say λ, is k. Then there exists a set of orthogonal vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ that are the basis for the eigenspace $\mathcal{N}(\boldsymbol{A} - \lambda \boldsymbol{I})$ such that $\boldsymbol{A}\boldsymbol{v}_i = \lambda \boldsymbol{v}_i$ for all $i \in \{1, 2, \ldots, k\}$. Then, $\boldsymbol{w}_i = \boldsymbol{P}\boldsymbol{v}_i$'s are the eigenvectors of $\boldsymbol{B}$ associated with eigenvalue λ. Further, $\boldsymbol{w}_i$'s are linearly independent since $\boldsymbol{P}$ is nonsingular. Thus, the dimension of the eigenspace $\mathcal{N}(\boldsymbol{B} - \lambda_i \boldsymbol{I})$ is at least k, that is, $dim(\mathcal{N}(\boldsymbol{A} - \lambda \boldsymbol{I})) \leq dim(\mathcal{N}(\boldsymbol{B} - \lambda \boldsymbol{I}))$.

Similarly, there exists a set of orthogonal vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k$ that are the bases for the eigenspace $\mathcal{N}(\boldsymbol{B} - \lambda \boldsymbol{I})$, then $\boldsymbol{v}_i = \boldsymbol{P}^{-1}\boldsymbol{w}_i$ for all $i \in \{1, 2, \ldots, k\}$ are the eigenvectors of $\boldsymbol{A}$ associated to λ. This will result in $dim(\mathcal{N}(\boldsymbol{B} - \lambda \boldsymbol{I})) \leq dim(\mathcal{N}(\boldsymbol{A} - \lambda \boldsymbol{I}))$.

Therefore, by "sandwiching", we get $dim(\mathcal{N}(\boldsymbol{A} - \lambda \boldsymbol{I})) = dim(\mathcal{N}(\boldsymbol{B} - \lambda \boldsymbol{I}))$, which is the equality of the geometric multiplicities, and the claim follows. ■

Lemma 13.14: (Bounded Geometric Multiplicity)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, its geometric multiplicity is bounded by algebraic multiplicity for any eigenvalue λ_i:

$$geo(\lambda_i) \leq alg(\lambda_i).$$

Proof [of Lemma 13.14] If we can find a similar matrix $\boldsymbol{B}$ of $\boldsymbol{A}$ that has a specific form of the characteristic polynomial, then we complete the proof.

Suppose $\boldsymbol{P}_1 = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k]$ contains the eigenvectors of $\boldsymbol{A}$ associated with λ_i which are linearly independent. That is, the k vectors are bases for the eigenspace $\mathcal{N}(\boldsymbol{A} - \lambda\boldsymbol{I})$ and the geometric multiplicity associated with λ_i is k. We can expand it to n linearly independent vectors such that

$$\boldsymbol{P} = [\boldsymbol{P}_1, \boldsymbol{P}_2] = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k, \boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_n],$$

where $\boldsymbol{P}$ is nonsingular. Then $\boldsymbol{A}\boldsymbol{P} = [\lambda_i \boldsymbol{P}_1, \boldsymbol{A}\boldsymbol{P}_2]$.

Construct a matrix $\boldsymbol{B} = \begin{bmatrix} \lambda_i \boldsymbol{I}_k & \boldsymbol{C} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix}$ where $\boldsymbol{A}\boldsymbol{P}_2 = \boldsymbol{P}_1\boldsymbol{C} + \boldsymbol{P}_2\boldsymbol{D}$, then $\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{B}$ such that $\boldsymbol{A}$ and $\boldsymbol{B}$ are similar matrices. We can always find such $\boldsymbol{C}, \boldsymbol{D}$ that satisfy the above condition, since $\boldsymbol{v}_i$'s are linearly independent with spanning the whole space $\mathbb{R}^n$, and any column of $\boldsymbol{A}\boldsymbol{P}_2$ is in the column space of $\boldsymbol{P} = [\boldsymbol{P}_1, \boldsymbol{P}_2]$. Therefore,

$$\begin{aligned}
\det(\boldsymbol{A} - \lambda\boldsymbol{I}) &= \det(\boldsymbol{P}^{-1})\det(\boldsymbol{A} - \lambda\boldsymbol{I})\det(\boldsymbol{P}) && (\det(\boldsymbol{P}^{-1}) = 1/\det(\boldsymbol{P}))\\
&= \det(\boldsymbol{P}^{-1}(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{P}) && (\det(\boldsymbol{A})\det(\boldsymbol{B}) = \det(\boldsymbol{A}\boldsymbol{B}))\\
&= \det(\boldsymbol{B} - \lambda\boldsymbol{I})\\
&= \det(\begin{bmatrix} (\lambda_i - \lambda)\boldsymbol{I}_k & \boldsymbol{C} \\ \boldsymbol{0} & \boldsymbol{D} - \lambda\boldsymbol{I} \end{bmatrix})\\
&= (\lambda_i - \lambda)^k \det(\boldsymbol{D} - \lambda\boldsymbol{I}),
\end{aligned}$$

where the last equality is from the fact that if matrix $\boldsymbol{M}$ has a block formulation: $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$, then $\det(\boldsymbol{M}) = \det(\boldsymbol{A})\det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})$. This implies

$$geo(\lambda_i) \leq alg(\lambda_i).$$

And we complete the proof. ■

Following from the proof of Lemma 13.6, we notice that the algebraic multiplicity and geometric multiplicity are the same for symmetric matrices. We call these matrices simple matrices.

Definition 13.15: (Simple Matrix)

When the algebraic multiplicity and geometric multiplicity are the same for a matrix, we call it a simple matrix.

Definition 13.16: (Diagonalizable)

A matrix $\boldsymbol{A}$ is diagonalizable if there exists a nonsingular matrix $\boldsymbol{P}$ and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$.

Eigenvalue decomposition in Theorem 11.1 and spectral decomposition in Theorem 13.1 are such kinds of matrices that are diagonalizable.

Lemma 13.17: (Simple Matrices are Diagonalizable)

A matrix is a simple matrix if and only if it is diagonalizable.

Proof [of Lemma 13.17] We will show by forward implication and backward implication separately as follows.

Forward implication Suppose that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a simple matrix, such that the algebraic and geometric multiplicities for each eigenvalue are equal. For a specific eigenvalue λ_i, let $\{\boldsymbol{v}_1^i, \boldsymbol{v}_2^i, \ldots, \boldsymbol{v}_{k_i}^i\}$ be a basis for the eigenspace $\mathcal{N}(\boldsymbol{A} - \lambda_i \boldsymbol{I})$, that is, $\{\boldsymbol{v}_1^i, \boldsymbol{v}_2^i, \ldots, \boldsymbol{v}_{k_i}^i\}$ is a set of linearly independent eigenvectors of $\boldsymbol{A}$ associated to λ_i, where k_i is the algebraic or geometric multiplicity associated to λ_i: $alg(\lambda_i) = geo(\lambda_i) = k_i$. Suppose there are m distinct eigenvalues, since $k_1 + k_2 + \ldots + k_m = n$, the set of eigenvectors consists of the union of n vectors. Suppose there is a set of x_j's such that

$$\boldsymbol{z} = \sum_{j=1}^{k_1} x_j^1 \boldsymbol{v}_j^1 + \sum_{j=1}^{k_2} x_j^2 \boldsymbol{v}_j^2 + \ldots \sum_{j=1}^{k_m} x_j^m \boldsymbol{v}_j^m = \boldsymbol{0}. \tag{13.9}$$

Let $\boldsymbol{w}^i = \sum_{j=1}^{k_i} x_j^i \boldsymbol{v}_j^i$. Then $\boldsymbol{w}^i$ is either an eigenvector associated to λ_i, or it is a zero vector. That is $\boldsymbol{z} = \sum_{i=1}^{m} \boldsymbol{w}^i$ is a sum of either zero vector or an eigenvector associated with different eigenvalues of $\boldsymbol{A}$. Since eigenvectors associated with different eigenvalues are linearly independent. We must have $\boldsymbol{w}^i = 0$ for all $i \in \{1, 2, \ldots, m\}$. That is

$$\boldsymbol{w}^i = \sum_{j=1}^{k_i} x_j^i \boldsymbol{v}_j^i = \boldsymbol{0}, \qquad \text{for all } i \in \{1, 2, \ldots, m\}.$$

Since we assume the eigenvectors $\boldsymbol{v}_j^i$'s associated to λ_i are linearly independent, we must have $x_j^i = 0$ for all $i \in \{1, 2, \ldots, m\}, j \in \{1, 2, \ldots, k_i\}$. Thus, the n vectors are linearly independent:

$$\{\boldsymbol{v}_1^1, \boldsymbol{v}_2^1, \ldots, \boldsymbol{v}_{k_i}^1\}, \{\boldsymbol{v}_1^2, \boldsymbol{v}_2^2, \ldots, \boldsymbol{v}_{k_i}^2\}, \ldots, \{\boldsymbol{v}_1^m, \boldsymbol{v}_2^m, \ldots, \boldsymbol{v}_{k_i}^m\}.$$

By eigenvalue decomposition in Theorem 11.1, matrix $\boldsymbol{A}$ can be diagonalizable.

Backward implication Suppose $\boldsymbol{A}$ is diagonalizable. That is, there exists a nonsingular matrix $\boldsymbol{P}$ and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$. $\boldsymbol{A}$ and $\boldsymbol{D}$ are similar matrices such that they have the same eigenvalues (Lemma 8.5, p. 91), same algebraic multiplicities, and geometric multiplicities (Corollary 13.13, p. 120). It can be easily verified that a diagonal matrix has equal algebraic multiplicity and geometric multiplicity such that $\boldsymbol{A}$ is a simple matrix. ■

Remark 13.18: (Equivalence on Diagonalization)

From Theorem 13.4 that any eigenvectors corresponding to different eigenvalues are linearly independent, and Remark 13.12 that the geometric multiplicity is the dimension of the eigenspace. We realize, if the geometric multiplicity is equal to the algebraic multiplicity, the eigenspace can span the whole space $\mathbb{R}^n$ if matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. So the above Lemma is equivalent to claim that if the eigenspace can span the whole space $\mathbb{R}^n$, then $\boldsymbol{A}$ can be diagonalizable.

Corollary 13.19

A square matrix $\boldsymbol{A}$ with linearly independent eigenvectors is a simple matrix. Or if $\boldsymbol{A}$ is symmetric, it is also a simple matrix.

From the eigenvalue decomposition in Theorem 11.1 and the spectral decomposition in Theorem 13.1, the proof is trivial for the corollary.

Now we are ready to show the second form of the spectral decomposition.

Theorem 13.20: (Spectral Decomposition: The Second Form)

A **simple matrix** $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factored as a sum of a set of idempotent matrices

$$\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{A}_i,$$

where λ_i for all $i \in \{1, 2, \ldots, n\}$ are eigenvalues of $\boldsymbol{A}$ (duplicate possible), and also known as the **spectral values** of $\boldsymbol{A}$. Specifically, we have the following properties:

1. Idempotent: $\boldsymbol{A}_i^2 = \boldsymbol{A}_i$ for all $i \in \{1, 2, \ldots, n\}$;
2. Orthogonal: $\boldsymbol{A}_i\boldsymbol{A}_j = \boldsymbol{0}$ for all $i \neq j$;
3. Additivity: $\sum_{i=1}^{n} \boldsymbol{A}_i = \boldsymbol{I}_n$;
4. Rank-Additivity: $rank(\boldsymbol{A}_1) + rank(\boldsymbol{A}_2) + \ldots + rank(\boldsymbol{A}_n) = n$.

Proof [of Theorem 13.20] Since $\boldsymbol{A}$ is a simple matrix, from Lemma 13.17, there exists a nonsingular matrix $\boldsymbol{P}$ and a diagonal matrix $\boldsymbol{\Lambda}$ such that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$ where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, and λ_i's are eigenvalues of $\boldsymbol{A}$ and columns of $\boldsymbol{P}$ are eigenvectors of $\boldsymbol{A}$. Suppose

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_n \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{P}^{-1} = \begin{bmatrix} \boldsymbol{w}_1^\top \\ \boldsymbol{w}_2^\top \\ \vdots \\ \boldsymbol{w}_n^\top \end{bmatrix}$$

are the column and row partitions of $\boldsymbol{P}$ and $\boldsymbol{P}^{-1}$ respectively. Then, we have

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_n \end{bmatrix} \boldsymbol{\Lambda} \begin{bmatrix} \boldsymbol{w}_1^\top \\ \boldsymbol{w}_2^\top \\ \vdots \\ \boldsymbol{w}_n^\top \end{bmatrix} = \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i \boldsymbol{w}_i^\top.$$

Let $\boldsymbol{A}_i = \boldsymbol{v}_i \boldsymbol{w}_i^\top$, we have $\boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{A}_i$. We realize that $\boldsymbol{P}^{-1}\boldsymbol{P} = \boldsymbol{I}$ such that

$$\begin{cases} \boldsymbol{w}_i^\top \boldsymbol{v}_j = 1, & \text{if } i = j. \\ \boldsymbol{w}_i^\top \boldsymbol{v}_j = 0, & \text{if } i \neq j. \end{cases}$$

Therefore,

$$\boldsymbol{A}_i \boldsymbol{A}_j = \boldsymbol{v}_i \boldsymbol{w}_i^\top \boldsymbol{v}_j \boldsymbol{w}_j^\top = \begin{cases} \boldsymbol{v}_i \boldsymbol{w}_i^\top = \boldsymbol{A}_i, & \text{if } i = j. \\ \boldsymbol{0}, & \text{if } i \neq j. \end{cases}$$

This implies the idempotency and orthogonality of $\boldsymbol{A}_i$'s. We also notice that $\sum_{i=1}^n \boldsymbol{A}_i = \boldsymbol{P}\boldsymbol{P}^{-1} = \boldsymbol{I}$, that is the additivity of $\boldsymbol{A}_i$'s. The rank-additivity of the $\boldsymbol{A}_i$'s is trivial since $rank(\boldsymbol{A}_i) = 1$ for all $i \in \{1, 2, \ldots, n\}$. ■

The decomposition is highly related to the Cochran's theorem and its application in the distribution theory of linear models (Lu, 2021c,e).

Theorem 13.21: (Spectral Decomposition: The Third Form)

A **simple matrix** $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with k distinct eigenvalues can be factored as a sum of a set of idempotent matrices

$$\boldsymbol{A} = \sum_{i=1}^{k} \lambda_i \boldsymbol{A}_i,$$

where λ_i for all $i \in \{1, 2, \ldots, k\}$ are the distinct eigenvalues of $\boldsymbol{A}$, and also known as the **spectral values** of $\boldsymbol{A}$. Specifically, we have the following properties:

1. Idempotent: $\boldsymbol{A}_i^2 = \boldsymbol{A}_i$ for all $i \in \{1, 2, \ldots, k\}$;
2. Orthogonal: $\boldsymbol{A}_i \boldsymbol{A}_j = \boldsymbol{0}$ for all $i \neq j$;
3. Additivity: $\sum_{i=1}^k \boldsymbol{A}_i = \boldsymbol{I}_n$;
4. Rank-Additivity: $rank(\boldsymbol{A}_1) + rank(\boldsymbol{A}_2) + \ldots + rank(\boldsymbol{A}_k) = n$.

Proof [of Theorem 13.21] From Theorem 13.20, we can decompose $\boldsymbol{A}$ by $\boldsymbol{A} = \sum_{j=1}^n \beta_i \boldsymbol{B}_j$. Without loss of generality, the eigenvalues β_i's are ordered such that $\beta_1 \leq \beta_2 \leq \ldots \beta_n$ where duplicate is possible. Let λ_i's be the distinct eigenvalues, and $\boldsymbol{A}_i$ be the sum of the $\boldsymbol{B}_j$'s associated with λ_i. Suppose the multiplicity of λ_i is m_i, and the $\boldsymbol{B}_j$'s associated to λ_i can be denoted as $\{\boldsymbol{B}_1^i, \boldsymbol{B}_2^i, \ldots, \boldsymbol{B}_{m_i}^i\}$. Then $\boldsymbol{A}_i$ can be denoted as $\boldsymbol{A}_i = \sum_{j=1}^{m_i} \boldsymbol{B}_j^i$. Apparently $\boldsymbol{A} = \sum_{i=1}^k \lambda_i \boldsymbol{A}_i$.

Idempotency $\boldsymbol{A}_i^2 = (\boldsymbol{B}_1^i + \boldsymbol{B}_2^i + \ldots \boldsymbol{B}_{m_i}^i)(\boldsymbol{B}_1^i + \boldsymbol{B}_2^i + \ldots \boldsymbol{B}_{m_i}^i) = \boldsymbol{B}_1^i + \boldsymbol{B}_2^i + \ldots \boldsymbol{B}_{m_i}^i = \boldsymbol{A}_i$ from the idempotency and orthogonality of $\boldsymbol{B}_j^i$'s.

Ortogonality $\boldsymbol{A}_i \boldsymbol{A}_j = (\boldsymbol{B}_1^i + \boldsymbol{B}_2^i + \ldots \boldsymbol{B}_{m_i}^i)(\boldsymbol{B}_1^j + \boldsymbol{B}_2^j + \ldots \boldsymbol{B}_{m_j}^j) = \boldsymbol{0}$ from the orthogonality of the $\boldsymbol{B}_j^i$'s.

Additivity It is trivial that $\sum_{i=1}^k \boldsymbol{A}_i = \boldsymbol{I}_n$.

Rank-Additivity $rank(\boldsymbol{A}_i) = rank(\sum_{j=1}^{m_i} \boldsymbol{B}_j^i) = m_i$ such that $rank(\boldsymbol{A}_1) + rank(\boldsymbol{A}_2) + \ldots + rank(\boldsymbol{A}_k) = m_1 + m_2 + \ldots + m_k = n$. ■

Theorem 13.22: (Spectral Decomposition: Backward Implication)

If a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with k distinct eigenvalues can be factored as a sum of a set of idempotent matrices

$$\boldsymbol{A} = \sum_{i=1}^{k} \lambda_i \boldsymbol{A}_i,$$

where λ_i for all $i \in \{1,2,\ldots,k\}$ are the distinct eigenvalues of $\boldsymbol{A}$, and

1. Idempotent: $\boldsymbol{A}_i^2 = \boldsymbol{A}_i$ for all $i \in \{1,2,\ldots,k\}$;
2. Orthogonal: $\boldsymbol{A}_i\boldsymbol{A}_j = \boldsymbol{0}$ for all $i \neq j$;
3. Additivity: $\sum_{i=1}^{k} \boldsymbol{A}_i = \boldsymbol{I}_n$;
4. Rank-Additivity: $rank(\boldsymbol{A}_1) + rank(\boldsymbol{A}_2) + \ldots + rank(\boldsymbol{A}_k) = n$.

Then, the matrix $\boldsymbol{A}$ is a simple matrix.

Proof [of Corollary 13.22] Suppose $rank(\boldsymbol{A}_i) = r_i$ for all $i \in \{1,2,\ldots,k\}$. By ULV decomposition in Theorem 4.1, $\boldsymbol{A}_i$ can be factored as

$$\boldsymbol{A}_i = \boldsymbol{U}_i \begin{bmatrix} \boldsymbol{L}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{V}_i,$$

where $\boldsymbol{L}_i \in \mathbb{R}^{r_i\times r_i}$, $\boldsymbol{U}_i \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V}_i \in \mathbb{R}^{n\times n}$ are orthogonal matrices. Let

$$\boldsymbol{X}_i = \boldsymbol{U}_i \begin{bmatrix} \boldsymbol{L}_i \\ \boldsymbol{0} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{V}_i = \begin{bmatrix} \boldsymbol{Y}_i \\ \boldsymbol{Z}_i \end{bmatrix},$$

where $\boldsymbol{X}_i \in \mathbb{R}^{n\times r_i}$, and $\boldsymbol{Y}_i \in \mathbb{R}^{r_i\times n}$ is the first r_i rows of $\boldsymbol{V}_i$. Then, we have

$$\boldsymbol{A}_i = \boldsymbol{X}_i\boldsymbol{Y}_i.$$

This can be seen as a **reduced** ULV decomposition of $\boldsymbol{A}_i$. Appending the $\boldsymbol{X}_i$'s and $\boldsymbol{Y}_i$'s into $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k], \qquad \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_k \end{bmatrix},$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ (from rank-additivity). By block matrix multiplication and the additivity of $\boldsymbol{A}_i$'s, we have

$$\boldsymbol{X}\boldsymbol{Y} = \sum_{i=1}^{k} \boldsymbol{X}_i\boldsymbol{Y}_i = \sum_{i=1}^{k} \boldsymbol{A}_i = \boldsymbol{I}.$$

Therefore $\boldsymbol{Y}$ is the inverse of $\boldsymbol{X}$, and

$$\boldsymbol{Y}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_k \end{bmatrix} [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k] = \begin{bmatrix} \boldsymbol{Y}_1\boldsymbol{X}_1 & \boldsymbol{Y}_1\boldsymbol{X}_2 & \ldots & \boldsymbol{Y}_1\boldsymbol{X}_k \\ \boldsymbol{Y}_2\boldsymbol{X}_1 & \boldsymbol{Y}_2\boldsymbol{X}_2 & \ldots & \boldsymbol{Y}_2\boldsymbol{X}_k \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Y}_k\boldsymbol{X}_1 & \boldsymbol{Y}_k\boldsymbol{X}_2 & \ldots & \boldsymbol{Y}_k\boldsymbol{X}_k \end{bmatrix} = \boldsymbol{I},$$

such that

$$\boldsymbol{Y}_i\boldsymbol{X}_j = \begin{cases} \boldsymbol{I}_{r_i}, & \text{if } i=j; \\ \boldsymbol{0}, & \text{if } i \neq j. \end{cases}$$

This implies

$$\boldsymbol{A}_i\boldsymbol{X}_j = \begin{cases} \boldsymbol{X}_i, & \text{if } i=j; \\ \boldsymbol{0}, & \text{if } i \neq j, \end{cases} \qquad \text{and} \qquad \boldsymbol{A}\boldsymbol{X}_i = \lambda_i\boldsymbol{X}_i.$$

Finally, we have

$$\boldsymbol{A}\boldsymbol{X} = \boldsymbol{A}[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k] = [\lambda_1\boldsymbol{X}_1, \lambda_2\boldsymbol{X}_2, \ldots, \lambda_k\boldsymbol{X}_k] = \boldsymbol{X}\boldsymbol{\Lambda},$$

where

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1\boldsymbol{I}_{r_1} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \lambda_2\boldsymbol{I}_{r_2} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \lambda_k\boldsymbol{I}_{r_k} \end{bmatrix}$$

is a diagonal matrix. This implies $\boldsymbol{A}$ can be diagonalized and from Lemma 13.17, $\boldsymbol{A}$ is a simple matrix. ■

Corollary 13.23: (Forward and Backward Spectral)

Combine Theorem 13.21 and Theorem 13.22, we can claim that matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a simple matrix with k distinct eigenvalues if and only if it can be factored as a sum of a set of idempotent matrices

$$\boldsymbol{A} = \sum_{i=1}^{k} \lambda_i\boldsymbol{A}_i,$$

where λ_i for all $i \in \{1, 2, \ldots, k\}$ are the distinct eigenvalues of $\boldsymbol{A}$, and

1. Idempotent: $\boldsymbol{A}_i^2 = \boldsymbol{A}_i$ for all $i \in \{1, 2, \ldots, k\}$;
2. Orthogonal: $\boldsymbol{A}_i\boldsymbol{A}_j = \boldsymbol{0}$ for all $i \neq j$;
3. Additivity: $\sum_{i=1}^{k} \boldsymbol{A}_i = \boldsymbol{I}_n$;
4. Rank-Additivity: $rank(\boldsymbol{A}_1) + rank(\boldsymbol{A}_2) + \ldots + rank(\boldsymbol{A}_k) = n$.

13.4 Skew-Symmetric Matrices and its Properties*

We have introduced the spectral decomposition for symmetric matrices. A special kind of matrices that's related to symmetric is called the skew-symmetric matrices.

Definition 13.24: (Skew-Symmetric Matrix)

If matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ have the following property, then it is known as a **skew-symmetric matrix**:

$$\boldsymbol{A}^\top = -\boldsymbol{A}.$$

Note that under this definition, for the diagonal values a_{ii} for all $i \in \{1, 2, \ldots, n\}$, we have $a_{ii} = -a_{ii}$ which implies all the diagonal components are 0.

We have proved in Lemma 13.2 that all the eigenvalues of symmetric matrices are real. Similarly, we could show that all the eigenvalues of skew-symmetric matrices are imaginary.

Lemma 13.25: (Imaginary Eigenvalues)

The eigenvalues of any skew-symmetric matrix are all imaginary or zero.

Proof [of Lemma 13.25] Suppose eigenvalue λ is a complex number $\lambda = a + ib$ where a, b are real. Its complex conjugate is $\bar{\lambda} = a - ib$. Same for complex eigenvector $\boldsymbol{x} = \boldsymbol{c} + i\boldsymbol{d}$ and its complex conjugate $\bar{\boldsymbol{x}} = \boldsymbol{c} - i\boldsymbol{d}$ where $\boldsymbol{c}, \boldsymbol{d}$ are real vectors. We then have the following property

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \qquad \xrightarrow{\text{leads to}} \qquad \boldsymbol{A}\bar{\boldsymbol{x}} = \bar{\lambda}\bar{\boldsymbol{x}} \qquad \xrightarrow{\text{transpose to}} \qquad \bar{\boldsymbol{x}}^\top\boldsymbol{A}^\top = \bar{\lambda}\bar{\boldsymbol{x}}^\top.$$

We take the dot product of the first equation with $\bar{\boldsymbol{x}}$ and the last equation with $\boldsymbol{x}$:

$$\bar{\boldsymbol{x}}^\top\boldsymbol{A}\boldsymbol{x} = \lambda\bar{\boldsymbol{x}}^\top\boldsymbol{x}, \qquad \text{and} \qquad \bar{\boldsymbol{x}}^\top\boldsymbol{A}^\top\boldsymbol{x} = \bar{\lambda}\bar{\boldsymbol{x}}^\top\boldsymbol{x}.$$

Then we have the equality $-\lambda\bar{\boldsymbol{x}}^\top\boldsymbol{x} = \bar{\lambda}\bar{\boldsymbol{x}}^\top\boldsymbol{x}$ (since $\boldsymbol{A}^\top = -\boldsymbol{A}$). Since $\bar{\boldsymbol{x}}^\top\boldsymbol{x} = (\boldsymbol{c} - i\boldsymbol{d})^\top(\boldsymbol{c} + i\boldsymbol{d}) = \boldsymbol{c}^\top\boldsymbol{c} + \boldsymbol{d}^\top\boldsymbol{d}$ is a real number. Therefore the real part of λ is zero and λ is either imaginary or zero. ■

Lemma 13.26: (Odd Skew-Symmetric Determinant)

For skew-symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, if n is odd, then $\det(\boldsymbol{A}) = 0$.

Proof [of Lemma 13.26] When n is odd, we have

$$\det(\boldsymbol{A}) = \det(\boldsymbol{A}^\top) = \det(-\boldsymbol{A}) = (-1)^n\det(\boldsymbol{A}) = -\det(\boldsymbol{A}).$$

This implies $\det(\boldsymbol{A}) = 0$. ■

Theorem 13.27: (Block-Diagonalization of Skew-Symmetric Matrices)

A real skew-symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^\top,$$

where $\boldsymbol{Z}$ is an $n \times n$ nonsingular matrix, and $\boldsymbol{D}$ is a block-diagonal matrix with the following form

$$\boldsymbol{D} = \mathrm{diag}\left(\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, 0, \ldots, 0\right).$$

Proof [of Theorem 13.27] We will prove by recursive calculation. As usual, we will denote the entry (i, j) of matrix $\boldsymbol{A}$ by $\boldsymbol{A}_{ij}$.

Case 1). Suppose the first row of $\boldsymbol{A}$ is nonzero, we notice that $\boldsymbol{EAE}^\top$ is skew-symmetric if $\boldsymbol{A}$ is skew-symmetric for any matrix $\boldsymbol{E}$. This will make both the diagonals of $\boldsymbol{A}$ and $\boldsymbol{EAE}^\top$ are zeros, and the upper-left 2×2 submatrix of $\boldsymbol{EAE}^\top$ has the following form

$$
(\boldsymbol{EAE}^\top)_{1:2,1:2} = \begin{bmatrix} 0 & x \\ -x & 0 \end{bmatrix}.
$$

Since we suppose the first row of $\boldsymbol{A}$ is nonzero, there exists a permutation matrix $\boldsymbol{P}$ (Definition 0.17, p. 15), such that we will exchange the nonzero value, say a, in the first row to the second column of $\boldsymbol{PAP}^\top$. And as discussed above, the upper-left 2×2 submatrix of $\boldsymbol{PAP}^\top$ has the following form

$$
(\boldsymbol{PAP}^\top)_{1:2,1:2} = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix}.
$$

Construct a nonsingular matrix $\boldsymbol{M} = \begin{bmatrix} 1/a & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{n-1} \end{bmatrix}$ such that the upper left 2×2 submatrix of $\boldsymbol{MPAP}^\top\boldsymbol{M}^\top$ has the following form

$$
(\boldsymbol{MPAP}^\top\boldsymbol{M}^\top)_{1:2,1:2} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.
$$

Now we finish diagonalizing the upper-left 2×2 block. Suppose now $(\boldsymbol{MPAP}^\top\boldsymbol{M}^\top)$ above has a nonzero value, say b, in the first row with entry $(1, j)$ for some $j > 2$, we can construct a nonsingular matrix $\boldsymbol{L} = \boldsymbol{I} - b \cdot \boldsymbol{E}_{j2}$ where $\boldsymbol{E}_{2j}$ is an all-zero matrix except the entry $(2, j)$ is 1, such that $(\boldsymbol{LMPAP}^\top\boldsymbol{M}^\top\boldsymbol{L}^\top)$ will introduce 0 for the entry with value b.

A Trivial Example

For example, suppose $\boldsymbol{MPAP}^\top\boldsymbol{M}^\top$ is a 3×3 matrix with the following value

$$
\boldsymbol{MPAP}^\top\boldsymbol{M}^\top = \begin{bmatrix} 0 & 1 & b \\ -1 & 0 & \times \\ \times & \times & 0 \end{bmatrix}, \qquad \text{and} \qquad \boldsymbol{L} = \boldsymbol{I} - b \cdot \boldsymbol{E}_{j2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -b & 1 \end{bmatrix},
$$

where $j = 3$ for this specific example. This results in

$$
\boldsymbol{LMPAP}^\top\boldsymbol{M}^\top\boldsymbol{L}^\top = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -b & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & b \\ -1 & 0 & \times \\ \times & \times & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -b \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & \times \\ \times & \times & 0 \end{bmatrix}.
$$

Similarly, if the second row of $\boldsymbol{LMPAP}^\top\boldsymbol{M}^\top\boldsymbol{L}^\top$ contains a nonzero value, say c, we could construct a nonsingular matrix $\boldsymbol{K} = \boldsymbol{I} + c \cdot \boldsymbol{E}_{j1}$ such that $\boldsymbol{KLMPAP}^\top\boldsymbol{M}^\top\boldsymbol{L}^\top\boldsymbol{K}^\top$ will introduce 0 for the entry with value c.

A Trivial Example

For example, suppose $\boldsymbol{LMPAP}^\top \boldsymbol{M}^\top \boldsymbol{L}^\top$ is a 3×3 matrix with the following value

$$\boldsymbol{LMPAP}^\top \boldsymbol{M}^\top \boldsymbol{L}^\top = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & c \\ \times & \times & 0 \end{bmatrix}, \qquad \text{and} \qquad \boldsymbol{K} = \boldsymbol{I} + c\cdot \boldsymbol{E}_{j1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ c & 0 & 1 \end{bmatrix},$$

where $j=3$ for this specific example. This results in

$$\boldsymbol{KLMPAP}^\top \boldsymbol{M}^\top \boldsymbol{L}^\top \boldsymbol{K}^\top = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ c & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & c \\ \times & \times & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & c \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ \times & \times & 0 \end{bmatrix}.$$

Since we have shown that $\boldsymbol{KLMPAP}^\top \boldsymbol{M}^\top \boldsymbol{L}^\top \boldsymbol{K}^\top$ is also skew-symmetric, then, it is actually

$$\boldsymbol{KLMPAP}^\top \boldsymbol{M}^\top \boldsymbol{L}^\top \boldsymbol{K}^\top = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

so that we do not need to tackle the first 2 columns of the above equation.

Apply this process for the bottom-right $(n-2)\times(n-2)$ submatrix, we will complete the proof.

Case 2). Suppose the first row of $\boldsymbol{A}$ is zero, a permutation matrix to put the first row into the last row and apply the process in case 1 to finish the proof. ■

From the block-diagonalization of skew-symmetric matrices above, we could easily find that the rank of a skew-symmetric matrix is even. And we could prove the determinant of skew-symmetric with even order is nonnegative as follows.

Lemma 13.28: (Even Skew-Symmetric Determinant)

For skew-symmetric matrix $\boldsymbol{A}\in\mathbb{R}^{n\times n}$, if n is even, then $\det(\boldsymbol{A})\geq 0$.

Proof [of Lemma 13.28] By Theorem 13.27, we could block-diagonalize $\boldsymbol{A}=\boldsymbol{ZDZ}^\top$ such that

$$\det(\boldsymbol{A}) = \det(\boldsymbol{ZDZ}^\top) = \det(\boldsymbol{Z})^2\det(\boldsymbol{D}) \geq 0.$$

This completes the proof. ■

13.5 Applications

13.5.1 Application: Eigenvalue of Projection Matrix

In Section 14.4 (p. 141), we will introduce the QR decomposition can be applied to solve the least squares problem, where we consider the overdetermined system $\boldsymbol{Ax}=\boldsymbol{b}$ with $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ being the data matrix, $\boldsymbol{b}\in\mathbb{R}^m$ with $m>n$ being the observation matrix. Normally $\boldsymbol{A}$

will have full column rank since the data from real work has a large chance to be unrelated. And the least squares solution is given by $\boldsymbol{x}_{LS} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{b}$ for minimizing $||\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}||^2$, where $\boldsymbol{A}^\top\boldsymbol{A}$ is invertible since $\boldsymbol{A}$ has full column rank and $rank(\boldsymbol{A}^T\boldsymbol{A}) = rank(\boldsymbol{A})$. The recovered observation matrix is then $\hat{\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{x}_{LS} = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{b}$. $\boldsymbol{b}$ may not be in the column space of $\boldsymbol{A}$, but the recovered $\hat{\boldsymbol{b}}$ is in this column space. We then define such matrix $\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$ to be a projection matrix, i.e., projecting $\boldsymbol{b}$ onto the column space of $\boldsymbol{A}$. Or, it is also known as hat matrix, since we put a hat on $\boldsymbol{b}$. It can be easily verified the projection matrix is symmetric and idempotent (i.e., $\boldsymbol{H}^2 = \boldsymbol{H}$).

Remark 13.29: (Column Space of Projection Matrix)

We notice that the hat matrix $\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$ is to project any vector in $\mathbb{R}^m$ into the column space of $\boldsymbol{A}$. That is, $\boldsymbol{H}\boldsymbol{y} \in \mathcal{C}(\boldsymbol{A})$. Notice again $\boldsymbol{H}\boldsymbol{y}$ is the nothing but a combination of the columns of $\boldsymbol{H}$, thus $\mathcal{C}(\boldsymbol{H}) = \mathcal{C}(\boldsymbol{A})$.

In general, for any projection matrix $\boldsymbol{H}$ to project vector onto subspace $\mathcal{V}$, then $\mathcal{C}(\boldsymbol{H}) = \mathcal{V}$. More formally, in a mathematical language, this property can be proved by SVD.

We now show that for any projection matrix, it has specific eigenvalues.

Proposition 13.30: (Eigenvalue of Projection Matrix)

The only possible eigenvalues of a projection matrix are 0 and 1.

Proof [of Proposition 13.30] Since $\boldsymbol{H}$ is symmetric, we have spectral decomposition $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$. From the idempotent property, we have

$$\begin{aligned}
(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top)^2 &= \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top \\
\boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top &= \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top \\
\boldsymbol{\Lambda}^2 &= \boldsymbol{\Lambda} \\
\lambda_i^2 &= \lambda_i,
\end{aligned}$$

Therefore, the only possible eigenvalues for $\boldsymbol{H}$ are 0 and 1. ■

This property of the projection matrix is important for the analysis of distribution theory for linear models. See (Lu, 2021e) for more details. Following from the eigenvalue of the projection matrix, it can also give rise to the perpendicular projection $\boldsymbol{I}-\boldsymbol{H}$.

Proposition 13.31: (Project onto $\mathcal{V}^\perp$)

Let $\mathcal{V}$ be a subspace and $\boldsymbol{H}$ be a projection onto $\mathcal{V}$. Then $\boldsymbol{I}-\boldsymbol{H}$ is the projection matrix onto $\mathcal{V}^\perp$.

Proof [of Proposition 13.31] First, $(\boldsymbol{I}-\boldsymbol{H})$ is symmetric, $(\boldsymbol{I}-\boldsymbol{H})^\top = \boldsymbol{I}-\boldsymbol{H}^\top = \boldsymbol{I}-\boldsymbol{H}$ since $\boldsymbol{H}$ is symmatrix. And

$$(\boldsymbol{I}-\boldsymbol{H})^2 = \boldsymbol{I}^2 - \boldsymbol{I}\boldsymbol{H} - \boldsymbol{H}\boldsymbol{I} + \boldsymbol{H}^2 = \boldsymbol{I}-\boldsymbol{H}.$$

Thus $\boldsymbol{I}-\boldsymbol{H}$ is a projection matrix. By spectral theorem again, let $\boldsymbol{H}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$. Then $\boldsymbol{I}-\boldsymbol{H}=\boldsymbol{Q}\boldsymbol{Q}^\top-\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top=\boldsymbol{Q}(\boldsymbol{I}-\boldsymbol{\Lambda})\boldsymbol{Q}^\top$. Hence the column space of $\boldsymbol{I}-\boldsymbol{H}$ is spanned by the eigenvectors of $\boldsymbol{H}$ corresponding to the zero eigenvalues of $\boldsymbol{H}$ (by Proposition 13.30, p. 130), which coincides with $\mathcal{V}^\perp$. ■

Again, for a detailed analysis of the origin of the projection matrix and results behind the projection matrix, we highly recommend the readers refer to (Lu, 2021c) although it is not the main interest of matrix decomposition results.

13.5.2 Application: An Alternative Definition on PD and PSD of Matrices

In Definition 2.2 (p. 29), we defined the positive definite matrices and positive semidefinite matrices by the quadratic form of the matrices. We here prove that a symmetric matrix is positive definite if and only if all eigenvalues are positive.

Lemma 13.32: (Eigenvalues of PD and PSD Matrices)

A matrix $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ is positive definite (PD) if and only if $\boldsymbol{A}$ has only positive eigenvalues. And a matrix $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ is positive semidefinite (PSD) if and only if $\boldsymbol{A}$ has only nonnegative eigenvalues.

Proof [of Lemma 13.32] We will prove by forward implication and reverse implication separately as follows.

Forward implication: Suppose $\boldsymbol{A}$ is PD, then for any eigenvalue λ and its corresponding eigenvector $\boldsymbol{v}$ of $\boldsymbol{A}$, we have $\boldsymbol{A}\boldsymbol{v}=\lambda\boldsymbol{v}$. Thus

$$\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v}=\lambda||\boldsymbol{v}||^2>0.$$

This implies $\lambda>0$.

Reverse implication: Conversely, suppose the eigenvalues are positive. By spectral decomposition of $\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$. If $\boldsymbol{x}$ is a nonzero vector, let $\boldsymbol{y}=\boldsymbol{Q}^\top\boldsymbol{x}$, we have

$$\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}=\boldsymbol{x}^\top(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top)\boldsymbol{x}=(\boldsymbol{x}^\top\boldsymbol{Q})\boldsymbol{\Lambda}(\boldsymbol{Q}^\top\boldsymbol{x})=\boldsymbol{y}^\top\boldsymbol{\Lambda}\boldsymbol{y}=\sum_{i=1}^n\lambda_i y_i^2>0.$$

That is, $\boldsymbol{A}$ is PD.

Analogously, we can prove the second part of the claim. ■

Theorem 13.33: (Nonsingular Factor of PSD and PD Matrices)

A real symmetric matrix $\boldsymbol{A}$ is PSD if and only if $\boldsymbol{A}$ can be factored as $\boldsymbol{A}=\boldsymbol{P}^\top\boldsymbol{P}$, and is PD if and only if $\boldsymbol{P}$ is nonsingular.

Proof [of Theorem 13.33] For the first part, we will prove by forward implication and reverse implication separately as follows.

Forward implication: Suppose $\boldsymbol{A}$ is PSD, its spectral decomposition is given by $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$. Since eigenvalues of PSD matrices are nonnegative, we can decompose $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}$. Let $\boldsymbol{P} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\top$, we can decompose $\boldsymbol{A}$ by $\boldsymbol{A} = \boldsymbol{P}^\top\boldsymbol{P}$.

Reverse implication: If $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{P}^\top\boldsymbol{P}$, then all eigenvalues of $\boldsymbol{A}$ are nonnegative since for any eigenvalues λ and its corresponding eigenvector $\boldsymbol{v}$ of $\boldsymbol{A}$, we have

$$\lambda = \frac{\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{v}^\top\boldsymbol{v}} = \frac{\boldsymbol{v}^\top\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{v}}{\boldsymbol{v}^\top\boldsymbol{v}} = \frac{||\boldsymbol{P}\boldsymbol{v}||^2}{||\boldsymbol{v}||^2} \geq 0.$$

This implies $\boldsymbol{A}$ is PSD by Lemma 13.32.

Similarly, we can prove the second part for PD matrices where the positive definiteness will result in the nonsingular $\boldsymbol{P}$ and the nonsingular $\boldsymbol{P}$ will result in the positiveness of the eigenvalues. [21] ■

13.5.3 Proof for Semidefinite Rank-Revealing Decomposition

In this section, we provide a proof for Theorem 2.10 (p. 39), the existence of the rank-revealing decomposition for positive semidefinite matrix.

Proof [of Theorem 2.10] The proof is a consequence of the nonsingular factor of PSD matrices (Theorem 13.33, p. 131) and the existence of column-pivoted QR decomposition (Theorem 3.2, p. 53).

By Theorem 13.33, the nonsingular factor of PSD matrix $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{Z}^\top\boldsymbol{Z}$, where $\boldsymbol{Z} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\top$ and $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ is the spectral decomposition of $\boldsymbol{A}$.

By Lemma 13.8, the rank of matrix $\boldsymbol{A}$ is the number of nonzero eigenvalues (here the number of positive eigenvalues since $\boldsymbol{A}$ is PSD). Therefore only r components in the diagonal of $\boldsymbol{\Lambda}^{1/2}$ are nonzero, and $\boldsymbol{Z} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\top$ contains only r independent columns, i.e., $\boldsymbol{Z}$ is of rank r. By column-pivoted QR decomposition, we have

$$\boldsymbol{Z}\boldsymbol{P} = \boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}_{11} & \boldsymbol{R}_{12}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix},$$

where $\boldsymbol{P}$ is a permutation matrix, $\boldsymbol{R}_{11} \in \mathbb{R}^{r\times r}$ is upper triangular with positive diagonals, and $\boldsymbol{R}_{12} \in \mathbb{R}^{r\times(n-r)}$. Therefore

$$\boldsymbol{P}^\top\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}^\top\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{P} = \begin{bmatrix}\boldsymbol{R}_{11}^\top & \boldsymbol{0}\\ \boldsymbol{R}_{12}^\top & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_{11} & \boldsymbol{R}_{12}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}.$$

Let

$$\boldsymbol{R} = \begin{bmatrix}\boldsymbol{R}_{11} & \boldsymbol{R}_{12}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix},$$

we find the rank-revealing decomposition for semidefinite matrix $\boldsymbol{P}^\top\boldsymbol{A}\boldsymbol{P} = \boldsymbol{R}^\top\boldsymbol{R}$. ■

This decomposition is produced by using complete pivoting, which at each stage permutes the largest diagonal element in the active submatrix into the pivot position. The procedure is similar to the partial pivoting discussed in Section 1.9.1 (p. 25).

21. See also wiki page: https://en.wikipedia.org/wiki/Sylvester's_criterion.

13.5.4 Application: Cholesky Decomposition via the QR Decomposition and the Spectral Decomposition

In this section, we provide another proof for the existence of Cholesky decomposition.

Proof [of Theorem 2.1] From Theorem 13.33, the PD matrix $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{P}^\top \boldsymbol{P}$ where $\boldsymbol{P}$ is a nonsingular matrix. Then, the QR decomposition of $\boldsymbol{P}$ is given by $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{R}$. This implies

$$\boldsymbol{A} = \boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{R} = \boldsymbol{R}^\top \boldsymbol{R},$$

where we notice that the form is very similar to the Cholesky decomposition except that we do not claim the $\boldsymbol{R}$ has only positive diagonal values. From the CGS algorithm to compute the QR decomposition, we realize that the diagonals of $\boldsymbol{R}$ are nonnegative, and if $\boldsymbol{P}$ is nonsingular, the diagonals of $\boldsymbol{R}$ are also positive. ■

The proof for the above theorem is a consequence of the existence of both the QR decomposition and the spectral decomposition. Thus, the existence of Cholesky decomposition can be proved via the QR decomposition and the spectral decomposition in this sense.

13.5.5 Application: Unique Power Decomposition of Positive Definite Matrices

Theorem 13.34: (Unique Power Decomposition of PD Matrices)

Any $n \times n$ positive matrix $\boldsymbol{A}$ can be **uniquely** factored as a product of a positive definite matrix $\boldsymbol{B}$ such that $\boldsymbol{A} = \boldsymbol{B}^2$.

Proof [of Theorem 13.34] We first prove that there exists such positive definite matrix $\boldsymbol{B}$ so that $\boldsymbol{A} = \boldsymbol{B}^2$.

Existence Since $\boldsymbol{A}$ is PD which is also symmetric, the spectral decomposition of $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$. Since eigenvalues of PD matrices are positive by Lemma 13.32, the square root of $\boldsymbol{\Lambda}$ exists. We can define $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\top$ such that $\boldsymbol{A} = \boldsymbol{B}^2$ where $\boldsymbol{B}$ is apparently PD.

Uniqueness Suppose such factorization is not unique, then there exist two of this decomposition such that

$$\boldsymbol{A} = \boldsymbol{B}_1^2 = \boldsymbol{B}_2^2,$$

where $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are both PD. The spectral decompositions of them are given by

$$\boldsymbol{B}_1 = \boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^\top, \qquad \text{and} \qquad \boldsymbol{B}_2 = \boldsymbol{Q}_2\boldsymbol{\Lambda}_2\boldsymbol{Q}_2^\top.$$

We notice that $\boldsymbol{\Lambda}_1^2$ and $\boldsymbol{\Lambda}_2^2$ contains the eigenvalues of $\boldsymbol{A}$, and both eigenvalues of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ contained in $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are positive (since $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are both PD). Without loss of generality, we suppose $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2 = \boldsymbol{\Lambda}^{1/2}$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. By $\boldsymbol{B}_1^2 = \boldsymbol{B}_2^2$, we have

$$\boldsymbol{Q}_1\boldsymbol{\Lambda}\boldsymbol{Q}_1^\top = \boldsymbol{Q}_2\boldsymbol{\Lambda}\boldsymbol{Q}_2^\top \qquad \xrightarrow{\text{leads to}} \qquad \boldsymbol{Q}_2^\top\boldsymbol{Q}_1\boldsymbol{\Lambda} = \boldsymbol{\Lambda}\boldsymbol{Q}_2^\top\boldsymbol{Q}_1.$$

Let $\boldsymbol{Z} = \boldsymbol{Q}_2^\top \boldsymbol{Q}_1$, this implies $\boldsymbol{\Lambda}$ and $\boldsymbol{Z}$ commute, and $\boldsymbol{Z}$ must be a block diagonal matrix whose partitioning conforms to the block structure of $\boldsymbol{\Lambda}$. This results in $\boldsymbol{\Lambda}^{1/2} = \boldsymbol{Z}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z}^\top$ and

$$\boldsymbol{B}_2 = \boldsymbol{Q}_2\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}_2^\top = \boldsymbol{Q}_2\boldsymbol{Q}_2^\top\boldsymbol{Q}_1\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}_1^\top\boldsymbol{Q}_2\boldsymbol{Q}_2^\top = \boldsymbol{B}_1.$$

This completes the proof. ■

Similarly, we could prove the unique decomposition of PSD matrix $\boldsymbol{A} = \boldsymbol{B}^2$ where $\boldsymbol{B}$ is PSD. A more detailed discussion on this topic can be referred to (Koeber and Schäfer, 2006).

Decomposition for PD matrices To conclude, for PD matrix $\boldsymbol{A}$, we can factor it into $\boldsymbol{A} = \boldsymbol{R}^\top\boldsymbol{R}$ where $\boldsymbol{R}$ is an upper triangular matrix with positive diagonals as shown in Theorem 2.1 by Cholesky decomposition, $\boldsymbol{A} = \boldsymbol{P}^\top\boldsymbol{P}$ where $\boldsymbol{P}$ is nonsingular in Theorem 13.33, and $\boldsymbol{A} = \boldsymbol{B}^2$ where $\boldsymbol{B}$ is PD in Theorem 13.34.

14. Singular Value Decomposition (SVD)

In the eigenvalue decomposition, we factor the matrix into a diagonal matrix. However, this is not always true. If $\boldsymbol{A}$ does not have linearly independent eigenvectors, such diagonalization does not exist. The singular value decomposition (SVD) fills this gap. Instead of factoring the matrix into an eigenvector matrix, SVD gives rise to two orthogonal matrices. We provide the result of SVD in the following theorem and we will discuss the existence of SVD in the next sections.

Theorem 14.1: (Reduced SVD for Rectangular Matrices)

For every real $m \times n$ matrix $\boldsymbol{A}$ with rank r, then matrix $\boldsymbol{A}$ can be factored as

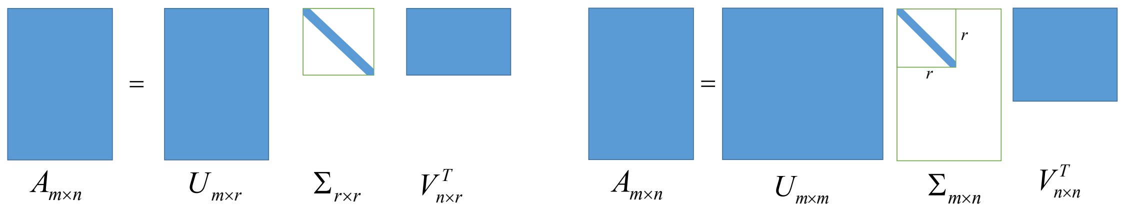
$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ is a diagonal matrix $\boldsymbol{\Sigma} = diag(\sigma_1, \sigma_2 \ldots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ and

- σ_i's are the nonzero **singular values** of $\boldsymbol{A}$, in the meantime, they are the (positive) square roots of the nonzero **eigenvalues** of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$.
- Columns of $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ contain the r eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$ corresponding to the r nonzero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$.
- Columns of $\boldsymbol{V} \in \mathbb{R}^{n\times r}$ contain the r eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$ corresponding to the r nonzero eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$.
- Moreover, the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are called the **left and right singular vectors** of $\boldsymbol{A}$, respectively.
- Further, the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal (by Spectral Theorem 13.1, p. 113).

In particular, we can write out the matrix decomposition by the sum of outer products of vectors $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, which is a sum of r rank-one matrices.

If we append additional $m - r$ silent columns that are orthonormal to the r eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$, just like the silent columns in the QR decomposition, we will have an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{m\times m}$. Similar situation for the columns of $\boldsymbol{V}$. The comparison between the reduced and the full SVD is shown in Figure 17 where white entries are zero and blue entries are not necessarily zero.



(a) Reduced SVD decomposition

(b) Full SVD decomposition

Figure 17: Comparison between the reduced and full SVD.

14.1 Existence of the SVD

To prove the existence of the SVD, we need to use the following lemmas. We mentioned that the singular values are the square roots of the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$. While, negative values do not have square roots such that the eigenvalues must be nonnegative.

Lemma 14.2: (Nonnegative Eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{A}^\top\boldsymbol{A}$ has nonnegative eigenvalues.

Proof [of Lemma 14.2] For eigenvalue and its corresponding eigenvector $\lambda, \boldsymbol{x}$ of $\boldsymbol{A}^\top\boldsymbol{A}$, we have

$$\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}^\top\boldsymbol{x}.$$

Since $\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = ||\boldsymbol{A}\boldsymbol{x}||^2 \geq 0$ and $\boldsymbol{x}^\top\boldsymbol{x} \geq 0$. We then have $\lambda \geq 0$. ■

Since $\boldsymbol{A}^\top\boldsymbol{A}$ has nonnegative eigenvalues, we then can define the singular value $\sigma \geq 0$ of $\boldsymbol{A}$ such that σ^2 is the eigenvalue of $\boldsymbol{A}^\top\boldsymbol{A}$, i.e., $\boxed{\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v} = \sigma^2\boldsymbol{v}}$. This is essential to the existence of the SVD.

We have shown in Lemma 13.7 (p. 118) that $rank(\boldsymbol{A}\boldsymbol{B}) \leq \min\{rank(\boldsymbol{A}), rank(\boldsymbol{B})\}$. However, the symmetric matrix $\boldsymbol{A}^\top\boldsymbol{A}$ is rather special in that the rank of $\boldsymbol{A}^\top\boldsymbol{A}$ is equal to $rank(\boldsymbol{A})$. And we now prove it.

Lemma 14.3: (Rank of $\boldsymbol{A}^\top\boldsymbol{A}$)

$\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}$ have same rank.

Proof [of Lemma 14.3] Let $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{A})$, we have

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0},$$

i.e., $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{A})$ $\underrightarrow{\text{leads to}}$ $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{A}^\top\boldsymbol{A})$, therefore $\mathcal{N}(\boldsymbol{A}) \subseteq \mathcal{N}(\boldsymbol{A}^\top\boldsymbol{A})$.

Further, let $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{A}^\top \boldsymbol{A})$, we have

$$\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \underrightarrow{\text{leads to}} \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = 0 \underrightarrow{\text{leads to}} ||\boldsymbol{A}\boldsymbol{x}||^2 = 0 \underrightarrow{\text{leads to}} \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0},$$

i.e., $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{A}^\top \boldsymbol{A}) \underrightarrow{\text{leads to}} \boldsymbol{x} \in \mathcal{N}(\boldsymbol{A})$, therefore $\mathcal{N}(\boldsymbol{A}^\top \boldsymbol{A}) \subseteq \mathcal{N}(\boldsymbol{A})$.

As a result, by "sandwiching", it follows that

$$\mathcal{N}(\boldsymbol{A}) = \mathcal{N}(\boldsymbol{A}^\top \boldsymbol{A}) \qquad \text{and} \qquad dim(\mathcal{N}(\boldsymbol{A})) = dim(\mathcal{N}(\boldsymbol{A}^\top \boldsymbol{A})).$$

By the fundamental theorem of linear algebra (Theorem 0.15, p. 14), $\boldsymbol{A}^\top \boldsymbol{A}$ and $\boldsymbol{A}$ have the same rank. ■

Apply the observation to $\boldsymbol{A}^\top$, we can also prove that $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}$ have the same rank:

$$rank(\boldsymbol{A}) = rank(\boldsymbol{A}^\top \boldsymbol{A}) = rank(\boldsymbol{A}\boldsymbol{A}^\top).$$

In the form of the SVD, we claimed the matrix $\boldsymbol{A}$ is a sum of r rank-one matrices where r is the number of nonzero singular values. And the number of nonzero singular values is actually the rank of the matrix.

Lemma 14.4: (The Number of Nonzero Singular Values Equals the Rank)

The number of nonzero singular values of matrix $\boldsymbol{A}$ equals the rank of $\boldsymbol{A}$.

Proof [of Lemma 14.4] The rank of any symmetric matrix (here $\boldsymbol{A}^\top \boldsymbol{A}$) equals the number of nonzero eigenvalues (with repetitions) by Lemma 13.8. So the number of nonzero singular values equals the rank of $\boldsymbol{A}^\top \boldsymbol{A}$. By Lemma 14.3, the number of nonzero singular values equals the rank of $\boldsymbol{A}$. ■

We are now ready to prove the existence of the SVD.

Proof [**of Theorem 14.1: Existence of the SVD**] Since $\boldsymbol{A}^\top \boldsymbol{A}$ is a symmetric matrix, by Spectral Theorem 13.1 (p. 113) and Lemma 14.2, there exists an orthogonal matrix $\boldsymbol{V}$ such that

$$\boxed{\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^2 \boldsymbol{V}^\top},$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix containing the singular values of $\boldsymbol{A}$, i.e., $\boldsymbol{\Sigma}^2$ contains the eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$. Specifically, $\boldsymbol{\Sigma} = diag(\sigma_1, \sigma_2, \ldots, \sigma_r)$ and $\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2\}$ are the nonzero eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$ with r being the rank of $\boldsymbol{A}$. I.e., $\{\sigma_1, \ldots, \sigma_r\}$ are the singular values of $\boldsymbol{A}$. In this case, $\boldsymbol{V} \in \mathbb{R}^{n\times r}$. Now we are into the central part.

Start from $\boxed{\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{v}_i = \sigma_i^2 \boldsymbol{v}_i}$, $\forall i \in \{1, 2, \ldots, r\}$, i.e., the eigenvector $\boldsymbol{v}_i$ of $\boldsymbol{A}^\top \boldsymbol{A}$ corresponding to σ_i^2:

1. Multiply both sides by $\boldsymbol{v}_i^\top$:

$$\boldsymbol{v}_i^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{v}_i = \sigma_i^2 \boldsymbol{v}_i^\top \boldsymbol{v}_i \qquad \underrightarrow{\text{leads to}} \qquad ||\boldsymbol{A}\boldsymbol{v}_i||^2 = \sigma_i^2 \qquad \underrightarrow{\text{leads to}} \qquad ||\boldsymbol{A}\boldsymbol{v}_i|| = \sigma_i$$

2. Multiply both sides by $\boldsymbol{A}$:

$$\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{v}_i = \sigma_i^2 \boldsymbol{A}\boldsymbol{v}_i \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{A}\boldsymbol{A}^\top \frac{\boldsymbol{A}\boldsymbol{v}_i}{\sigma_i} = \sigma_i^2 \frac{\boldsymbol{A}\boldsymbol{v}_i}{\sigma_i} \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{u}_i = \sigma_i^2 \boldsymbol{u}_i$$

where we notice that this form can find the eigenvector of $\boldsymbol{A}\boldsymbol{A}^\top$ corresponding to σ_i^2 which is $\boldsymbol{A}\boldsymbol{v}_i$. Since the length of $\boldsymbol{A}\boldsymbol{v}_i$ is σ_i, we then define $\boldsymbol{u}_i = \frac{\boldsymbol{A}\boldsymbol{v}_i}{\sigma_i}$ with norm 1.

These $\boldsymbol{u}_i$'s are orthogonal because $(\boldsymbol{A}\boldsymbol{v}_i)^\top(\boldsymbol{A}\boldsymbol{v}_j) = \boldsymbol{v}_i^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}_j = \sigma_j^2\boldsymbol{v}_i^\top\boldsymbol{v}_j = 0$. That is

$$\boxed{\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top}.$$

Since $\boxed{\boldsymbol{A}\boldsymbol{v}_i = \sigma_i\boldsymbol{u}_i}$, we have

$$[\boldsymbol{A}\boldsymbol{v}_1, \boldsymbol{A}\boldsymbol{v}_2, \ldots, \boldsymbol{A}\boldsymbol{v}_r] = [\sigma_1\boldsymbol{u}_1, \sigma_2\boldsymbol{u}_2, \ldots, \sigma_r\boldsymbol{u}_r] \qquad \xrightarrow{\text{leads to}} \qquad \boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma},$$

which completes the proof. ■

By appending silent columns in $\boldsymbol{U}$ and $\boldsymbol{V}$, we can easily find the full SVD. A byproduct of the above proof is that the spectral decomposition of $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top$ will result in the spectral decomposition of $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top$ with the same eigenvalues.

Corollary 14.5: (Eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$)

The nonzero eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ are the same.

We have shown in Lemma 14.2 that the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ are nonnegative, such that the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$ are nonnegative as well.

Corollary 14.6: (Nonnegative Eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$)

The eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ are nonnegative.

The existence of the SVD is important for defining the effective rank of a matrix.

Definition 14.7: (Effective Rank vs Exact Rank)

Effective rank, or also known as the *numerical rank*. Following from Lemma 14.4, the number of nonzero singular values is equal to the rank of a matrix. Assume the i-th largest singular value of $\boldsymbol{A}$ is denoted as $\sigma_i(\boldsymbol{A})$. Then if $\sigma_r(\boldsymbol{A}) \gg \sigma_{r+1}(\boldsymbol{A}) \approx 0$, r is known as the numerical rank of $\boldsymbol{A}$. Whereas, when $\sigma_i(\boldsymbol{A}) > \sigma_{r+1}(\boldsymbol{A}) = 0$, it is known as having *exact rank* r as we have used in most of our discussions.

14.2 Properties of the SVD

14.2.1 Four Subspaces in SVD

For any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we have the following property:

• $\mathcal{N}(\boldsymbol{A})$ is the orthogonal complement of the row space $\mathcal{C}(\boldsymbol{A}^\top)$ in $\mathbb{R}^n$: $dim(\mathcal{N}(\boldsymbol{A})) + dim(\mathcal{C}(\boldsymbol{A}^\top)) = n$;

• $\mathcal{N}(\boldsymbol{A}^\top)$ is the orthogonal complement of the column space $\mathcal{C}(\boldsymbol{A})$ in $\mathbb{R}^m$: $dim(\mathcal{N}(\boldsymbol{A}^\top)) + dim(\mathcal{C}(\boldsymbol{A})) = m$;

This is called the fundamental theorem of linear algebra and is also known as the rank-nullity theorem. From the SVD, we can find an orthonormal basis for each subspace.

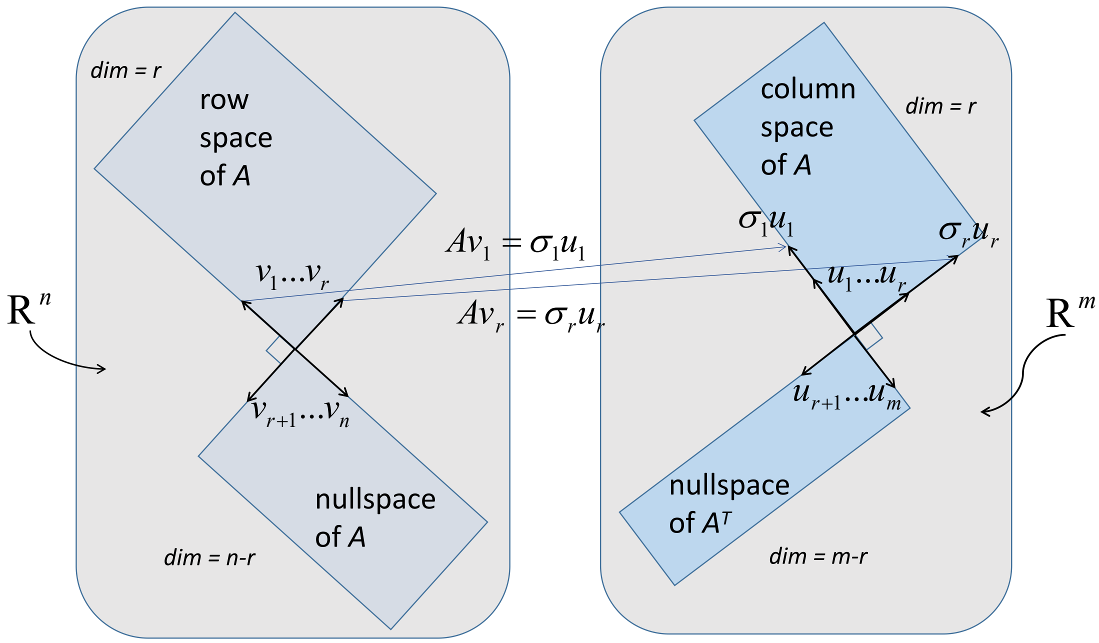


Figure 18: Orthonormal bases that diagonalize $\boldsymbol{A}$ from SVD.

Lemma 14.8: (Four Orthonormal Basis)

Given the full SVD of matrix $\boldsymbol{A} = \boldsymbol{U\Sigma V}^\top$, where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ are the column partitions of $\boldsymbol{U}$ and $\boldsymbol{V}$. Then, we have the following property:

- $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{A}^\top)$;
- $\{\boldsymbol{v}_{r+1}, \boldsymbol{v}_{r+2}, \ldots, \boldsymbol{v}_n\}$ is an orthonormal basis of $\mathcal{N}(\boldsymbol{A})$;
- $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r\}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{A})$;
- $\{\boldsymbol{u}_{r+1}, \boldsymbol{u}_{r+2}, \ldots, \boldsymbol{u}_m\}$ is an orthonormal basis of $\mathcal{N}(\boldsymbol{A}^\top)$.

The relationship of the four subspaces is demonstrated in Figure 18 where $\boldsymbol{A}$ transfer the row basis $\boldsymbol{v}_i$ into column basis $\boldsymbol{u}_i$ by $\sigma_i \boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{v}_i$ for all $i \in \{1, 2, \ldots, r\}$.

Proof [of Lemma 14.8] From Lemma 13.8, for symmetric matrix $\boldsymbol{A}^\top\boldsymbol{A}$, $\mathcal{C}(\boldsymbol{A}^\top\boldsymbol{A})$ is spanned by the eigenvectors, thus $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{A}^\top\boldsymbol{A})$.

Since,

1. $\boldsymbol{A}^\top\boldsymbol{A}$ is symmetric, then the row space of $\boldsymbol{A}^\top\boldsymbol{A}$ equals the column space of $\boldsymbol{A}^\top\boldsymbol{A}$.

2. All rows of $\boldsymbol{A}^\top\boldsymbol{A}$ are the combinations of the rows of $\boldsymbol{A}$, so the row space of $\boldsymbol{A}^\top\boldsymbol{A} \subseteq$ the row space of $\boldsymbol{A}$, i.e., $\mathcal{C}(\boldsymbol{A}^\top\boldsymbol{A}) \subseteq \mathcal{C}(\boldsymbol{A}^\top)$.

3. Since $rank(\boldsymbol{A}^\top\boldsymbol{A}) = rank(\boldsymbol{A})$ by Lemma 14.3, we then have

The row space of $\boldsymbol{A}^\top\boldsymbol{A}$ = the column space of $\boldsymbol{A}^\top\boldsymbol{A}$ = the row space of $\boldsymbol{A}$, i.e., $\mathcal{C}(\boldsymbol{A}^\top\boldsymbol{A}) = \mathcal{C}(\boldsymbol{A}^\top)$. Thus $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{A}^\top)$.

Further, the space spanned by $\{\boldsymbol{v}_{r+1}, \boldsymbol{v}_{r+2}, \ldots, \boldsymbol{v}_n\}$ is an orthogonal complement to the space spanned by $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$, so $\{\boldsymbol{v}_{r+1}, \boldsymbol{v}_{r+2}, \ldots, \boldsymbol{v}_n\}$ is an orthonormal basis of $\mathcal{N}(\boldsymbol{A})$.

If we apply this process to $\boldsymbol{A}\boldsymbol{A}^\top$, we will prove the rest claims in the lemma. Also, we can see that $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r\}$ is a basis for the column space of $\boldsymbol{A}$ by Lemma 0.14 [22], since $\boldsymbol{u}_i = \frac{\boldsymbol{A}\boldsymbol{v}_i}{\sigma_i}$, $\forall i \in \{1, 2, \ldots, r\}$. ■

14.2.2 Relationship between Singular Values and Determinant

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a square matrix and the singular value decomposition of $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, it follows that

$$|\det(\boldsymbol{A})| = |\det(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)| = |\det(\boldsymbol{\Sigma})| = \sigma_1\sigma_2\ldots\sigma_n.$$

If all the singular values σ_i are nonzero, then $\det(\boldsymbol{A}) \neq 0$. That is, $\boldsymbol{A}$ is **nonsingular**. If there is at least one singular value such that $\sigma_i = 0$, then $\det(\boldsymbol{A}) = 0$, and $\boldsymbol{A}$ does not have full rank, and is not invertible. Then the matrix is called **singular**. This is why σ_i's are known as the singular values.

14.2.3 Orthogonal Equivalence

We have defined in Definition 8.4 (p. 90) that $\boldsymbol{A}$ and $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$ are similar matrices for any nonsingular matrix $\boldsymbol{P}$. The orthogonal equivalence is defined in a similar way.

Definition 14.9: (Orthogonal Equivalent Matrices)

For any orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, the matrices $\boldsymbol{A}$ and $\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}$ are called *orthogonal equivalent matrices*. Or *unitary equivalent* in complex domain when $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary.

Then, we have the following property for orthogonal equivalent matrices.

Lemma 14.10: (Orthogonal Equivalent Matrices)

For any orthogonal equivalent matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, then singular values are the same.

Proof [of Lemma 14.10] Since $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal equivalent, there exist orthogonal matrices that $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{A}\boldsymbol{V}$. We then have

$$\boldsymbol{B}\boldsymbol{B}^\top = (\boldsymbol{U}\boldsymbol{A}\boldsymbol{V})(\boldsymbol{V}^\top\boldsymbol{A}^\top\boldsymbol{U}^\top) = \boldsymbol{U}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{U}^\top.$$

This implies $\boldsymbol{B}\boldsymbol{B}^\top$ and $\boldsymbol{A}\boldsymbol{A}^\top$ are similar matrices. By Lemma 8.5 (p. 91), the eigenvalues of similar matrices are the same, which proves the singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$ are the same. ■

14.2.4 SVD for QR

22. For any matrix $\boldsymbol{A}$, let $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r\}$ be a set of vectors in $\mathbb{R}^n$ which forms a basis for the row space, then $\{\boldsymbol{A}\boldsymbol{r}_1, \boldsymbol{A}\boldsymbol{r}_2, \ldots, \boldsymbol{A}\boldsymbol{r}_r\}$ is a basis for the column space of $\boldsymbol{A}$.

Lemma 14.11: (Orthogonal Equivalent Matrices)

Suppose the full QR decomposition for matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$ is given by $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ is orthogonal and $\boldsymbol{R} \in \mathbb{R}^{m\times n}$ is upper triangular. Then $\boldsymbol{A}$ and $\boldsymbol{R}$ have the same singular values and right singular vectors.

Proof [of Lemma 14.11] We notice that $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{R}^\top\boldsymbol{R}$ such that $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{R}^\top\boldsymbol{R}$ have the same eigenvalues and eigenvectors, i.e., $\boldsymbol{A}$ and $\boldsymbol{R}$ have the same singular values and right singular vectors (i.e., the eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$ or $\boldsymbol{R}^\top\boldsymbol{R}$). ∎

The above lemma implies that an SVD of a matrix can be constructed by the QR decomposition of itself. Suppose the QR decomposition of $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ and the SVD of $\boldsymbol{R}$ is given by $\boldsymbol{R} = \boldsymbol{U}_0\boldsymbol{\Sigma}\boldsymbol{V}^\top$. Therefore, the SVD of $\boldsymbol{A}$ can be obtained by

$$\boldsymbol{A} = \underbrace{\boldsymbol{Q}\boldsymbol{U}_0}_{\boldsymbol{U}}\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

14.3 Polar Decomposition

Theorem 14.12: (Polar Decomposition)

For every real $n \times n$ square matrix $\boldsymbol{A}$ with rank r, then matrix $\boldsymbol{A}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{Q}_l\boldsymbol{S},$$

where $\boldsymbol{Q}_l$ is an orthogonal matrix, and $\boldsymbol{S}$ is a positive semidefinite matrix. And this form is called the **left polar decomposition**. Also matrix $\boldsymbol{A}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{S}\boldsymbol{Q}_r,$$

where $\boldsymbol{Q}_r$ is an orthogonal matrix, and $\boldsymbol{S}$ is a positive semidefinite matrix. And this form is called the **right polar decomposition**.

Specially, the left and right polar decomposition of a square matrix $\boldsymbol{A}$ is **unique**.

Since every $n \times n$ square matrix $\boldsymbol{A}$ has full SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where both $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n \times n$ orthogonal matrix. We then have $\boldsymbol{A} = (\boldsymbol{U}\boldsymbol{V}^\top)(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{Q}_l\boldsymbol{S}$ where it can be easily verified that $\boldsymbol{Q}_l = \boldsymbol{U}\boldsymbol{V}^\top$ is an orthogonal matrix and $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is a symmetric matrix. We notice that the singular values in $\boldsymbol{\Sigma}$ are nonnegative, such that $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2\top}\boldsymbol{V}^\top$ showing that $\boldsymbol{S}$ is PSD.

Similarly, we have $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{V}^\top = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top)(\boldsymbol{U}\boldsymbol{V}^\top) = \boldsymbol{S}\boldsymbol{Q}_r$. And $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2\top}\boldsymbol{U}^\top$ such that $\boldsymbol{S}$ is PSD as well.

For the uniqueness of the right polar decomposition, we suppose the decomposition is not unique, and two of the decompositions are given by

$$\boldsymbol{A} = \boldsymbol{S}_1\boldsymbol{Q}_1 = \boldsymbol{S}_2\boldsymbol{Q}_2,$$

such that

$$\boldsymbol{S}_1 = \boldsymbol{S}_2\boldsymbol{Q}_2\boldsymbol{Q}_1^\top.$$

Since $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are symmetric, we have

$$\boldsymbol{S}_1^2 = \boldsymbol{S}_1\boldsymbol{S}_1^\top = \boldsymbol{S}_2\boldsymbol{Q}_2\boldsymbol{Q}_1^\top\boldsymbol{Q}_1\boldsymbol{Q}_2^\top\boldsymbol{S}_2 = \boldsymbol{S}_2^2.$$

This implies $\boldsymbol{S}_1 = \boldsymbol{S}_2$, and the decomposition is unique (Theorem 13.34, p. 133). Similarly, the uniqueness of the left polar decomposition can be implied from the context.

Corollary 14.13: (Full Rank Polar Decomposition)

When $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has full rank, then the $\boldsymbol{S}$ in both the left and right polar decomposition above is a symmetric positive definite matrix.

14.4 Application: Least Squares via the Full QR Decomposition, UTV, SVD

Least Squares via the Full QR Decomposition

Let's consider the overdetermined system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the data matrix, $\boldsymbol{b} \in \mathbb{R}^m$ with $m > n$ is the observation matrix. Normally $\boldsymbol{A}$ will have full column rank since the data from real work has a large chance to be unrelated. And the least squares (LS) solution is given by $\boldsymbol{x}_{LS} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{b}$ for minimizing $||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||^2$, where $\boldsymbol{A}^\top\boldsymbol{A}$ is invertible since $\boldsymbol{A}$ has full column rank and $rank(\boldsymbol{A}^T\boldsymbol{A}) = rank(\boldsymbol{A})$.

However, the inverse of a matrix is not easy to compute, we can then use QR decomposition to find the least squares solution as illustrated in the following theorem.

Theorem 14.14: (LS via QR for Full Column Rank Matrix)

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ is its full QR decomposition with $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ being an orthogonal matrix, $\boldsymbol{R} \in \mathbb{R}^{m\times n}$ being an upper triangular matrix appended by additional $m-n$ zero rows, and $\boldsymbol{A}$ has full column rank with $m \geq n$. Suppose $\boldsymbol{R} = \begin{bmatrix}\boldsymbol{R}_1 \\ \boldsymbol{0}\end{bmatrix}$, where $\boldsymbol{R}_1 \in \mathbb{R}^{n\times n}$ is the square upper triangular in $\boldsymbol{R}$, $\boldsymbol{b} \in \mathbb{R}^m$, then the LS solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is given by

$$\boldsymbol{x}_{LS} = \boldsymbol{R}_1^{-1}\boldsymbol{c},$$

where $\boldsymbol{c}$ is the first n components of $\boldsymbol{Q}^\top\boldsymbol{b}$.

Proof [of Theorem 14.14] Since $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ is the full QR decomposition of $\boldsymbol{A}$ and $m \geq n$, the last $m-n$ rows of $\boldsymbol{R}$ are zero as shown in Figure 8. Then $\boldsymbol{R}_1 \in \mathbb{R}^{n\times n}$ is the square upper triangular in $\boldsymbol{R}$ and $\boldsymbol{Q}^\top\boldsymbol{A} = \boldsymbol{R} = \begin{bmatrix}\boldsymbol{R}_1 \\ \boldsymbol{0}\end{bmatrix}$. Thus,

$$\begin{aligned}
||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||^2 &= (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\
&= (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\top\boldsymbol{Q}\boldsymbol{Q}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \qquad &\text{(Since } \boldsymbol{Q} \text{ is an orthogonal matrix)} \\
&= ||\boldsymbol{Q}^\top\boldsymbol{A}\boldsymbol{x} - \boldsymbol{Q}^\top\boldsymbol{b}||^2 \qquad &\text{(Invariant under orthogonal)} \\
&= \left\|\begin{bmatrix}\boldsymbol{R}_1 \\ \boldsymbol{0}\end{bmatrix}\boldsymbol{x} - \boldsymbol{Q}^\top\boldsymbol{b}\right\|^2 \\
&= ||\boldsymbol{R}_1\boldsymbol{x} - \boldsymbol{c}||^2 + ||\boldsymbol{d}||^2,
\end{aligned}$$

where $\boldsymbol{c}$ is the first n components of $\boldsymbol{Q}^\top\boldsymbol{b}$ and $\boldsymbol{d}$ is the last $m-n$ components of $\boldsymbol{Q}^\top\boldsymbol{b}$. And the LS solution can be calculated by back substitution of the upper triangular system $\boldsymbol{R}_1\boldsymbol{x}=\boldsymbol{c}$, i.e., $\boldsymbol{x}_{LS}=\boldsymbol{R}_1^{-1}\boldsymbol{c}$. ■

To verify Theorem 14.14, for the full QR decomposition of $\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{Q}\in\mathbb{R}^{m\times m}$ and $\boldsymbol{R}\in\mathbb{R}^{m\times n}$. Together with the LS solution $\boldsymbol{x}_{LS}=(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{b}$, we obtain

$$
\begin{aligned}
\boldsymbol{x}_{LS} &= (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{b}\\
&= (\boldsymbol{R}^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{Q}^\top\boldsymbol{b}\\
&= (\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{Q}^\top\boldsymbol{b}\\
&= (\boldsymbol{R}_1^\top\boldsymbol{R}_1)^{-1}\boldsymbol{R}^\top\boldsymbol{Q}^\top\boldsymbol{b}\\
&= \boldsymbol{R}_1^{-1}\boldsymbol{R}_1^{-\top}\boldsymbol{R}^\top\boldsymbol{Q}^\top\boldsymbol{b}\\
&= \boldsymbol{R}_1^{-1}\boldsymbol{R}_1^{-\top}\boldsymbol{R}_1^\top\boldsymbol{Q}_1^\top\boldsymbol{b}\\
&= \boldsymbol{R}_1^{-1}\boldsymbol{Q}_1^\top\boldsymbol{b},
\end{aligned}
\tag{14.1}
$$

where $\boldsymbol{R}=\begin{bmatrix}\boldsymbol{R}_1\\ \boldsymbol{0}\end{bmatrix}$ and $\boldsymbol{R}_1\in\mathbb{R}^{n\times n}$ is an upper triangular matrix, and $\boldsymbol{Q}_1=\boldsymbol{Q}_{1:m,1:n}\in\mathbb{R}^{m\times n}$ is the first n columns of $\boldsymbol{Q}$ (i.e., $\boldsymbol{Q}_1\boldsymbol{R}_1$ is the reduced QR decomposition of $\boldsymbol{A}$). Then the result of Equation (14.1) agrees with Theorem 14.14.

To conclude, using the QR decomposition, we first derive directly the least squares result which results in the argument in Theorem 14.14. Moreover, we verify the result of LS from calculus indirectly by the QR decomposition as well. The two results coincide with each other. For those who are interested in LS in linear algebra, a pictorial view of least squares for full column rank $\boldsymbol{A}$ in the fundamental theorem of linear algebra is provided in (Lu, 2021d).

Least Squares via ULV/URV for Rank Deficient Matrices

In the above section, we introduced the LS via the full QR decomposition for full rank matrices. However, if often happens that the matrix may be rank-deficient. If $\boldsymbol{A}$ does not have full column rank, $\boldsymbol{A}^\top\boldsymbol{A}$ is not invertible. We can then use the ULV/URV decomposition to find the least squares solution as illustrated in the following theorem.

Theorem 14.15: (LS via ULV/URV for Rank Definient Matrix)

Let $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ with rank r and $m\geq n$. Suppose $\boldsymbol{A}=\boldsymbol{U}\boldsymbol{T}\boldsymbol{V}$ is its full ULV/URV decomposition with $\boldsymbol{U}\in\mathbb{R}^{m\times m},\boldsymbol{V}\in\mathbb{R}^{n\times n}$ being orthogonal matrix matrices, and

$$
\boldsymbol{T}=\begin{bmatrix}\boldsymbol{T}_{11} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}
$$

where $\boldsymbol{T}_{11} \in \mathbb{R}^{r\times r}$ is a lower triangular matrix or an upper triangular matrix. Suppose $\boldsymbol{b} \in \mathbb{R}^m$, then the LS solution with the minimal 2-norm to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is given by

$$\boldsymbol{x}_{LS} = \boldsymbol{V}^\top \begin{bmatrix} \boldsymbol{T}_{11}^{-1}\boldsymbol{c} \\ \boldsymbol{0} \end{bmatrix},$$

where $\boldsymbol{c}$ is the first r components of $\boldsymbol{U}^\top\boldsymbol{b}$.

Proof [of Theorem 14.15] Since $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ is the full QR decomposition of $\boldsymbol{A}$ and $m > n$, the last $m - n$ rows of $\boldsymbol{R}$ are zero as shown in Figure 8. Then $\boldsymbol{R}_1 \in \mathbb{R}^{n\times n}$ is the square upper triangular in $\boldsymbol{R}$ and $\boldsymbol{Q}^\top\boldsymbol{A} = \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix}$. Thus,

$$\begin{aligned}
||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||^2 &= (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\
&= (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\top\boldsymbol{U}\boldsymbol{U}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) && \text{(Since } \boldsymbol{U} \text{ is an orthogonal matrix)} \\
&= ||\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{x} - \boldsymbol{U}^\top\boldsymbol{b}||^2 && \text{(Invariant under orthogonal)} \\
&= ||\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{T}\boldsymbol{V}\boldsymbol{x} - \boldsymbol{U}^\top\boldsymbol{b}||^2 \\
&= ||\boldsymbol{T}\boldsymbol{V}\boldsymbol{x} - \boldsymbol{U}^\top\boldsymbol{b}||^2 \\
&= ||\boldsymbol{T}_{11}\boldsymbol{e} - \boldsymbol{c}||^2 + ||\boldsymbol{d}||^2,
\end{aligned}$$

where $\boldsymbol{c}$ is the first r components of $\boldsymbol{U}^\top\boldsymbol{b}$ and $\boldsymbol{d}$ is the last $m - r$ components of $\boldsymbol{U}^\top\boldsymbol{b}$, $\boldsymbol{e}$ is the first r components of $\boldsymbol{V}\boldsymbol{x}$ and $\boldsymbol{f}$ is the last $n - r$ components of $\boldsymbol{V}\boldsymbol{x}$:

$$\boldsymbol{U}^\top\boldsymbol{b} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{bmatrix}, \qquad \boldsymbol{V}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{e} \\ \boldsymbol{f} \end{bmatrix}$$

And the LS solution can be calculated by back/forward substitution of the upper/lower triangular system $\boldsymbol{T}_{11}\boldsymbol{e} = \boldsymbol{c}$, i.e., $\boldsymbol{e} = \boldsymbol{T}_{11}^{-1}\boldsymbol{c}$. For $\boldsymbol{x}$ to have a minimal 2-norm, $\boldsymbol{f}$ must be zero. That is

$$\boldsymbol{x}_{LS} = \boldsymbol{V}^\top \begin{bmatrix} \boldsymbol{T}_{11}^{-1}\boldsymbol{c} \\ \boldsymbol{0} \end{bmatrix}.$$

This completes the proof. ■

A word on the minimal 2-norm LS solution For the least squares problem, the set of all minimizers

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n : ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}|| = \min\}$$

is convex (Golub and Van Loan, 2013). And if $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$, then

$$||\boldsymbol{A}(\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) - \boldsymbol{b}|| \leq \lambda||\boldsymbol{A}\boldsymbol{x}_1 - \boldsymbol{b}|| + (1-\lambda)||\boldsymbol{A}\boldsymbol{x}_2 - \boldsymbol{b}|| = \min_{\boldsymbol{x}\in\mathbb{R}^n} ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||.$$

Thus $\lambda\boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2 \in \mathcal{X}$. In above proof, if we do not set $\boldsymbol{f} = \boldsymbol{0}$, we will find more least squares solutions. However, the minimal 2-norm least squares solution is unique. For full-rank case in the previous section, the least squares solution is unique and it must have a minimal 2-norm. See also (Foster, 2003; Golub and Van Loan, 2013) for a more detailed discussion on this topic.

Least Squares via SVD for Rank Deficient Matrices

Apart form the UTV decomposition for rank-deficient least squares solution, SVD serves as an alternative.

Theorem 14.16: (LS via SVD for Rank Deficient Matrix)

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is its full SVD decomposition with $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ being orthogonal matrices and $rank(\boldsymbol{A}) = r$. Suppose $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m]$, $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ and $\boldsymbol{b} \in \mathbb{R}^m$, then the LS solution with minimal 2-norm to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is given by

$$\boldsymbol{x}_{LS} = \sum_{i=1}^{r} \frac{\boldsymbol{u}_i^\top \boldsymbol{b}}{\sigma_i}\boldsymbol{v}_i = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^\top\boldsymbol{b}, \tag{14.2}$$

where the upper-left side of $\boldsymbol{\Sigma}^+ \in \mathbb{R}^{n\times m}$ is a diagonal matrix, $\boldsymbol{\Sigma}^+ = \begin{bmatrix}\boldsymbol{\Sigma}_1^+ & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}$ where $\boldsymbol{\Sigma}_1^+ = diag(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r})$.

Proof [of Theorem 14.16] Write out the loss to be minimized

$$\begin{aligned}
||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||^2 &= (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\
&= (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\top\boldsymbol{U}\boldsymbol{U}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) && \text{(Since } \boldsymbol{U} \text{ is an orthogonal matrix)} \\
&= ||\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{x} - \boldsymbol{U}^\top\boldsymbol{b}||^2 && \text{(Invariant under orthogonal)} \\
&= ||\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{x} - \boldsymbol{U}^\top\boldsymbol{b}||^2 && \text{(Since } \boldsymbol{V} \text{ is an orthogonal matrix)} \\
&= ||\boldsymbol{\Sigma}\boldsymbol{\alpha} - \boldsymbol{U}^\top\boldsymbol{b}||^2 && \text{(Let } \boldsymbol{\alpha} = \boldsymbol{V}^\top\boldsymbol{x}) \\
&= \sum_{i=1}^{r}(\sigma_i\boldsymbol{\alpha}_i - \boldsymbol{u}_i^\top\boldsymbol{b})^2 + \sum_{i=r+1}^{m}(\boldsymbol{u}_i^\top\boldsymbol{b})^2. && \text{(Since } \sigma_{r+1} = \sigma_{r+2} = \ldots = \sigma_m = 0)
\end{aligned}$$

Since $\boldsymbol{x}$ only appears in $\boldsymbol{\alpha}$, we just need to set $\boldsymbol{\alpha}_i = \frac{\boldsymbol{u}_i^\top\boldsymbol{b}}{\sigma_i}$ for all $i \in \{1, 2, \ldots, r\}$ to minimize the above equation. For any value of $\boldsymbol{\alpha}_{r+1}, \boldsymbol{\alpha}_{r+2}, \ldots, \boldsymbol{\alpha}_n$, it won't change the result. From the regularization point of view (or here, we want the minimal 2-norm) we can set them to be 0. This gives us the LS solution via SVD:

$$\boldsymbol{x}_{LS} = \sum_{i=1}^{r} \frac{\boldsymbol{u}_i^\top \boldsymbol{b}}{\sigma_i}\boldsymbol{v}_i = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^\top\boldsymbol{b} = \boldsymbol{A}^+\boldsymbol{b},$$

where $\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^\top \in \mathbb{R}^{n\times m}$ is known as the **pseudo-inverse** of $\boldsymbol{A}$. ■

14.5 Application: Principal Component Analysis (PCA) via the Spectral Decomposition and the SVD

Given a data set of n observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ for all $i \in \{1, 2, \ldots, n\}$. Our goal is to project the data onto a low-dimensional space, say $m < p$. Define the sample mean vector and sample covariance matrix

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \qquad \text{and} \qquad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top.$$

where the $n-1$ term in the covariance matrix is to make it to be an unbiased consistent estimator of the covariance (Lu, 2021e). Or the covariance matrix can also be defined as $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top$ which is also a consistent estimator of covariance matrix [23].

Each data point $\boldsymbol{x}_i$ is then projected onto a scalar value by $\boldsymbol{u}_1$ such that $\boldsymbol{u}_1^\top\boldsymbol{x}_i$. The mean of the projected data is obtained by $\mathrm{E}[\boldsymbol{u}_1^\top\boldsymbol{x}_i] = \boldsymbol{u}_1^\top\overline{\boldsymbol{x}}$, and the variance of the projected data is given by

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{u}_1^\top\boldsymbol{x}_i] &= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{u}_1^\top\boldsymbol{x}_i - \boldsymbol{u}_1^\top\overline{\boldsymbol{x}})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{u}_1^\top(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top\boldsymbol{u}_1\\ &= \boldsymbol{u}_1^\top\boldsymbol{S}\boldsymbol{u}_1.\end{aligned}$$

We want to maximize the projected variance $\boldsymbol{u}_1^\top\boldsymbol{S}\boldsymbol{u}_1$ with respect to $\boldsymbol{u}_1$ where we must constrain $||\boldsymbol{u}_1||$ to prevent $||\boldsymbol{u}_1|| \to \infty$ by setting it to be $\boldsymbol{u}_1^\top\boldsymbol{u}_1 = 1$. By Lagrange multiplier (see (Bishop, 2006; Boyd et al., 2004)), we have

$$\boldsymbol{u}_1^\top\boldsymbol{S}\boldsymbol{u}_1 + \lambda_1(1 - \boldsymbol{u}_1^\top\boldsymbol{u}_1).$$

Trivial calculation will lead to

$$\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1 \qquad \xrightarrow{\text{leads to}} \qquad \boldsymbol{u}_1^\top\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1.$$

That is, $\boldsymbol{u}_1$ is an eigenvector of $\boldsymbol{S}$ corresponding to eigenvalue λ_1. And the maximum variance projection $\boldsymbol{u}_1$ is corresponding to the largest eigenvalues of $\boldsymbol{S}$. The eigenvector is known as the *first principal axis*.

Define the other principal axes by decremental eigenvalues until we have m such principal components bring about the dimension reduction. This is known as the *maximum variance formulation* of PCA (Hotelling, 1933; Bishop, 2006; Shlens, 2014). A *minimum-error formulation* of PCA is discussed in (Pearson, 1901; Bishop, 2006).

PCA via the spectral decomposition Now let's assume the data are centered such that $\overline{\boldsymbol{x}}$ is zero, or we can set $\boldsymbol{x}_i = \boldsymbol{x}_i - \overline{\boldsymbol{x}}$ to centralize the data. Let the data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ contain the data observations as rows. The covariance matrix is given by

$$\boldsymbol{S} = \frac{\boldsymbol{X}^\top\boldsymbol{X}}{n-1},$$

which is a symmetric matrix, and its spectral decomposition is given by

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top, \tag{14.3}$$

23. Consistency: An estimator θ_n of θ constructed on the basis of a sample of size n is said to be consistent if $\theta_n \xrightarrow{p} \theta$ as $n \to \infty$.

where $\boldsymbol{U}$ is an orthogonal matrix of eigenvectors (columns of $\boldsymbol{U}$ are eigenvectors of $\boldsymbol{S}$), and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is a diagonal matrix with eigenvalues (ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$). The eigenvectors are called *principal axes* of the data, and they *decorrelate* the the covariance matrix. Projections of the data on the principal axes are called the *principal components*. The i-th principal component is given by the i-th column of $\boldsymbol{XU}$. If we want to reduce the dimension from p to m, we just select the first m columns of $\boldsymbol{XU}$.

PCA via the SVD If the SVD of $\boldsymbol{X}$ is given by $\boldsymbol{X} = \boldsymbol{P\Sigma Q}^\top$, then the covariance matrix can be written as

$$\boldsymbol{S} = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{n-1} = \boldsymbol{Q}\frac{\boldsymbol{\Sigma}^2}{n-1}\boldsymbol{Q}^\top, \tag{14.4}$$

where $\boldsymbol{Q} \in \mathbb{R}^{p\times p}$ is an orthogonal matrix and contains the right singular vectors of $\boldsymbol{X}$, and the upper left of $\boldsymbol{\Sigma}$ is a diagonal matrix containing the singular values $\text{diag}(\sigma_1, \sigma_2, \ldots)$ with $\sigma_1 \geq \sigma_2 \geq \ldots$. The number of singular values is equal to $\min\{n, p\}$ which will not be larger than p and some of which are zeros.

The above Equation (14.4) compared with Equation (14.3) implies Equation (14.4) is also a spectral decomposition of $\boldsymbol{S}$, since the eigenvalues in $\boldsymbol{\Lambda}$ and singular values in $\boldsymbol{\Sigma}$ are ordered in a descending way and the uniqueness of the spectral decomposition in terms of the eigenspaces (Section 13.2, p. 119).

This results in the right singular vectors $\boldsymbol{Q}$ are also the principal axes which decorrelate the covariance matrix, and the singular values are related to the eigenvalues of the covariance matrix via $\lambda_i = \frac{\sigma_i^2}{n-1}$. To reduce the dimensionality of the data from p to m, we should select largest m singular values and the corresponding right singular vectors. This is also related to the truncated SVD (TSVD) $\boldsymbol{X}_m = \sum_{i=1}^m \sigma_i \boldsymbol{p}_i \boldsymbol{q}_i^\top$ as will be shown in the next section, where $\boldsymbol{p}_i$'s and $\boldsymbol{q}_i$'s are the columns of $\boldsymbol{P}$ and $\boldsymbol{Q}$.

A byproduct of PCA via the SVD for high-dimensional data For a principle axis $\boldsymbol{u}_i$ of $\boldsymbol{S} = \frac{\boldsymbol{X}^\top\boldsymbol{X}}{n-1}$, we have

$$\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n-1}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i.$$

Left multiply by $\boldsymbol{X}$, we obtain

$$\frac{\boldsymbol{X}\boldsymbol{X}^\top}{n-1}(\boldsymbol{X}\boldsymbol{u}_i) = \lambda_i(\boldsymbol{X}\boldsymbol{u}_i),$$

which implies λ_i is also an eigenvalue of $\frac{\boldsymbol{X}\boldsymbol{X}^\top}{n-1} \in \mathbb{R}^{n\times n}$, and the corresponding eigenvector is $\boldsymbol{X}\boldsymbol{u}_i$. This is also stated in the proof of Theorem 14.1, the existence of the SVD. If $p \gg n$, instead of finding the eigenvector of $\boldsymbol{S}$, i.e., the principle axes of $\boldsymbol{S}$, we can find the eigenvector of $\frac{\boldsymbol{X}\boldsymbol{X}^\top}{n-1}$. This reduces the complexity from $O(p^3)$ to $O(n^3)$. Suppose now, the eigenvector of $\frac{\boldsymbol{X}\boldsymbol{X}^\top}{n-1}$ is $\boldsymbol{v}_i$ corresponding to nonzero eigenvalue λ_i,

$$\frac{\boldsymbol{X}\boldsymbol{X}^\top}{n-1}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i.$$

Left multiply by $\boldsymbol{X}^\top$, we obtain

$$\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n-1}(\boldsymbol{X}^\top \boldsymbol{v}_i) = \boldsymbol{S}(\boldsymbol{X}^\top \boldsymbol{v}_i) = \lambda_i(\boldsymbol{X}^\top \boldsymbol{v}_i),$$

i.e., the eigenvector $\boldsymbol{u}_i$ of $\boldsymbol{S}$, is proportional to $\boldsymbol{X}^\top \boldsymbol{v}_i$, where $\boldsymbol{v}_i$ is the eigenvector of $\frac{\boldsymbol{X}\boldsymbol{X}^\top}{n-1}$ corresponding to the same eigenvalue λ_i. A further normalization step is needed to make $||\boldsymbol{u}_i|| = 1$.

14.6 Application: Low-Rank Approximation

For a low-rank approximation problem, there are basically two types related due to the interplay of rank and error: *fixed-precision approximation problem* and *fixed-rank approximation problem*. In the fixed-precision approximation problem, for a given matrix $\boldsymbol{A}$ and a given tolerance ϵ, one wants to find a matrix $\boldsymbol{B}$ with rank $r = r(\epsilon)$ such that $||\boldsymbol{A} - \boldsymbol{B}|| \leq \epsilon$ in an appropriate matrix norm. On the contrary, in the fixed-rank approximation problem, one looks for a matrix $\boldsymbol{B}$ with fixed rank k and an error $||\boldsymbol{A} - \boldsymbol{B}||$ as small as possible. In this section, we will consider the latter. Some excellent examples can also be found in (Kishore Kumar and Schneider, 2017; Martinsson, 2019).

Suppose we want to approximate matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with rank r by a rank $k < r$ matrix $\boldsymbol{B}$. The approximation is measured by spectral norm:

$$\boldsymbol{B} = \arg\min_{\boldsymbol{B}} \quad ||\boldsymbol{A} - \boldsymbol{B}||_2,$$

where the spectral norm is defined as follows:

Definition 14.17: (Spectral Norm)

The spectral norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is defined as

$$||\boldsymbol{A}||_2 = \max_{\boldsymbol{x}\neq \boldsymbol{0}} \frac{||\boldsymbol{A}\boldsymbol{x}||_2}{||\boldsymbol{x}||_2} = \max_{\boldsymbol{u}\in\mathbb{R}^n : ||\boldsymbol{u}||_2 = 1} ||\boldsymbol{A}\boldsymbol{x}||_2,$$

which is also the maximal singular value of $\boldsymbol{A}$, i.e., $||\boldsymbol{A}||_2 = \sigma_1(\boldsymbol{A})$.

Then, we can recover the best rank-k approximation by the following theorem.

Theorem 14.18: (Eckart-Young-Misky Theorem w.r.t. Spectral Norm)

Given matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $1 \leq k \leq rank(\boldsymbol{A}) = r$, and let $\boldsymbol{A}_k$ be the truncated SVD (TSVD) of $\boldsymbol{A}$ with the largest k terms, i.e., $\boldsymbol{A}_k = \sum_{i=1}^k \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ from SVD of $\boldsymbol{A} = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ by zeroing out the $r - k$ trailing singular values of $\boldsymbol{A}$. Then $\boldsymbol{A}_k$ is the best rank-k approximation to $\boldsymbol{A}$ in terms of the spectral norm.

Proof [of Theorem 14.18] We need to show for any matrix $\boldsymbol{B}$, if $rank(\boldsymbol{B}) = k$, then $||\boldsymbol{A} - \boldsymbol{B}||_2 \geq ||\boldsymbol{A} - \boldsymbol{A}_k||_2$.

Since $rank(\boldsymbol{B}) = k$, then $dim(\mathcal{N}(\boldsymbol{B})) = n - k$. As a result, any $k + 1$ basis in $\mathbb{R}^n$ intersects $\mathcal{N}(\boldsymbol{B})$. As shown in Lemma 14.8, $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{A}^\top) \subset \mathbb{R}^n$, so that we can choose the first $k + 1$ $\boldsymbol{v}_i$'s as a $k + 1$ basis for $\mathbb{R}^n$. Let $\boldsymbol{V}_{k+1} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{k+1}]$, then there is a vector $\boldsymbol{x}$ that

$$\boldsymbol{x} \in \mathcal{N}(\boldsymbol{B}) \cap \mathcal{C}(\boldsymbol{V}_{k+1}), \qquad s.t. \;\; ||\boldsymbol{x}||_2 = 1.$$

That is $\boldsymbol{x} = \sum_{i=1}^{k+1} a_i \boldsymbol{v}_i$, and $|| \sum_{i=1}^{k+1} a_i \boldsymbol{v}_i||_2 = \sum_{i=1}^{k+1} a_i^2 = 1$. Thus,

$$
\begin{aligned}
||\boldsymbol{A}-\boldsymbol{B}||_2^2 \geq ||(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{x}||_2^2 \cdot ||\boldsymbol{x}||_2^2, & \qquad \text{(From defintion of spectral norm)}\\
= ||\boldsymbol{A}\boldsymbol{x}||_2^2, & \qquad (\boldsymbol{x} \text{ in null space of } \boldsymbol{B})\\
= \sum_{i=1}^{k+1} \sigma_i^2 (\boldsymbol{v}_i^\top \boldsymbol{x})^2, & \qquad (\boldsymbol{x} \text{ orthogonal to } \boldsymbol{v}_{k+2}, \ldots, \boldsymbol{v}_r)\\
\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} (\boldsymbol{v}_i^\top \boldsymbol{x})^2, & \qquad (\sigma_{k+1} \leq \sigma_k \leq \ldots \leq \sigma_1)\\
\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} a_i^2, & \qquad (\boldsymbol{v}_i^\top \boldsymbol{x} = a_i)\\
= \sigma_{k+1}^2. &
\end{aligned}
$$

It is trivial that $||\boldsymbol{A}-\boldsymbol{A}_k||_2^2 = || \sum_{i=k+1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top ||_2^2 = \sigma_{k+1}^2$. Thus, $||\boldsymbol{A}-\boldsymbol{A}_k||_2 \leq ||\boldsymbol{A}-\boldsymbol{B}||_2$, which completes the proof. ■

Moreover, readers can prove that $\boldsymbol{A}_k$ is the best rank-k approximation to $\boldsymbol{A}$ in terms of the Frobenius norm. The minimal error is given by the Euclidean norm of the singular values that have been zeroed out in the process $||\boldsymbol{A}-\boldsymbol{A}_k||_F = \sqrt{\sigma_{k+1}^2 + \sigma_{k+2}^2 + \ldots + \sigma_r^2}$.

SVD gives the best approximation of a matrix. As mentioned in (Stewart, 1998; Kishore Kumar and Schneider, 2017), *the singular value decomposition is the creme de la creme of rank-reducing decompositions — the decomposition that all others try to beat.* And also *The SVD is the climax of this linear algebra course* in (Strang, 2009).

Part VI
Special Topics

15. Coordinate Transformation in Matrix Decomposition

Suppose a vector $\boldsymbol{v} \in \mathbb{R}^3$ and it has elements $\boldsymbol{v} = [3; 7; 2]$. But what do these values 3, 7, and 2 mean? In the Cartesian coordinate system, it means it has a component of 3 on the x-axis, a component of 7 on the y-axis, and a component of 2 on the z-axis.

15.1 An Overview of Matrix Multiplication

Coordinate defined by a nonsingular matrix Suppose further a 3×3 nonsingular matrix $\boldsymbol{B}$ which means $\boldsymbol{B}$ is invertible and columns of $\boldsymbol{B}$ are linearly independent. Thus the 3 columns of $\boldsymbol{B}$ form a basis for the space $\mathbb{R}^3$. One step forward, we can take the 3 columns of $\boldsymbol{B}$ as a basis for a new coordinate system, which we call the B coordinate system. Going back to the Cartesian coordinate system, we also have three vectors as a basis, $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$. If we put the three vectors into columns of a matrix, the matrix will be an identity matrix. So $\boldsymbol{I}\boldsymbol{v} = \boldsymbol{v}$ means transfer $\boldsymbol{v}$ from the Cartesian coordinate system into the Cartesian coordinate system, the same coordinate. Similarly, $\boldsymbol{B}\boldsymbol{v} = \boldsymbol{u}$ is to transfer $\boldsymbol{v}$ from the Cartesian coordinate system into the B system. Specifically, for $\boldsymbol{v} = [3; 7; 2]$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3]$, we have $\boldsymbol{u} = \boldsymbol{B}\boldsymbol{v} = 3\boldsymbol{b}_1 + 7\boldsymbol{b}_2 + 2\boldsymbol{b}_3$, i.e., $\boldsymbol{u}$ contains 3 of the first basis $\boldsymbol{b}_1$ of $\boldsymbol{B}$, 7 of the second basis $\boldsymbol{b}_2$ of $\boldsymbol{B}$, and 2 of the third basis $\boldsymbol{b}_3$ of $\boldsymbol{B}$. If again, we want to transfer the vector $\boldsymbol{u}$ from B coordinate system back to the Cartesian coordinate system, we just need to multiply by $\boldsymbol{B}^{-1}\boldsymbol{u} = \boldsymbol{v}$.

Coordinate defined by an orthogonal matrix A 3×3 orthogonal matrix $\boldsymbol{Q}$ defines a "better" coordinate system since the three columns (i.e., basis) are orthonormal to each other. $\boldsymbol{Q}\boldsymbol{v}$ is to transfer $\boldsymbol{v}$ from the Cartesian to the coordinate system defined by the orthogonal matrix. Since the basis vectors from the orthogonal matrix are orthonormal, just like the three vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ in the Cartesian coordinate system, the transformation defined by the orthogonal matrix just rotates or reflects the Cartesian system. $\boldsymbol{Q}^\top$ can help transfer back to the Cartesian coordinate system.

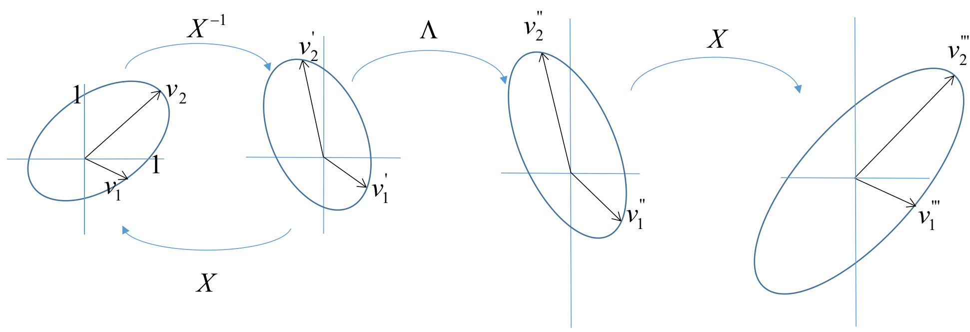


Figure 19: Eigenvalue Decomposition: $\boldsymbol{X}^{-1}$ transforms to a different coordinate system. $\boldsymbol{\Lambda}$ stretches and $\boldsymbol{X}$ transforms back. $\boldsymbol{X}^{-1}$ and $\boldsymbol{X}$ are nonsingular, which will change the basis of the system, and the angle between the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ will **not** be preserved, that is, the angle between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is **different** from the angle between $\boldsymbol{v}_1'$ and $\boldsymbol{v}_2'$. The length of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are also **not** preserved, that is, $||\boldsymbol{v}_1|| \neq ||\boldsymbol{v}_1'||$ and $||\boldsymbol{v}_2|| \neq ||\boldsymbol{v}_2'||$.

15.2 Eigenvalue Decomposition

A square matrix $\boldsymbol{A}$ with linearly independent eigenvectors can be factored as $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^{-1}$ where $\boldsymbol{X}$ and $\boldsymbol{X}^{-1}$ are nonsingular so that they define a system transformation intrinsically. $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^{-1}\boldsymbol{u}$ firstly transfers $\boldsymbol{u}$ into the system defined by $\boldsymbol{X}^{-1}$. Let's call this system the **eigen coordinate system**. $\boldsymbol{\Lambda}$ is to stretch each component of the vector in the eigen system by the length of the eigenvalue. And then $\boldsymbol{X}$ helps to transfer the resulting vector back to the Cartesian coordinate system. A demonstration of how the eigenvalue decomposition transforms between coordinate systems is shown in Figure 19 where $\boldsymbol{v}_1, \boldsymbol{v}_2$ are two linearly independent eigenvectors of $\boldsymbol{A}$ such that they form a basis for $\mathbb{R}^2$.

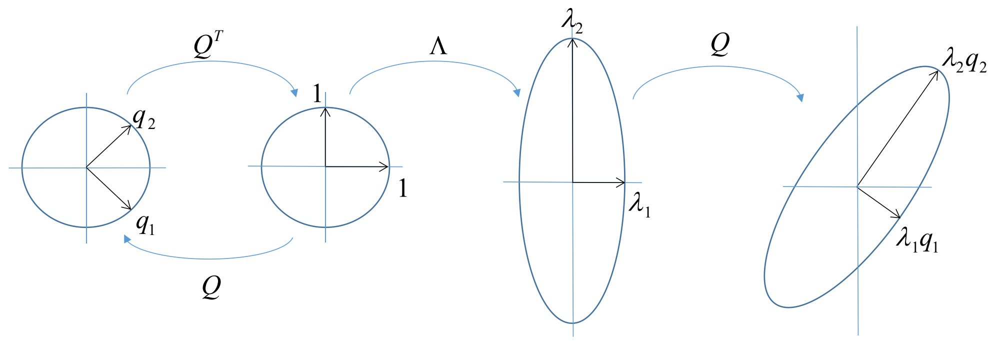


Figure 20: Spectral Decomposition $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$: $\boldsymbol{Q}^\top$ rotates or reflects, $\boldsymbol{\Lambda}$ stretches cycle to ellipse, and $\boldsymbol{Q}$ rotates or reflects back. Orthogonal matrices $\boldsymbol{Q}^\top$ and $\boldsymbol{Q}$ only change the basis of the system. However, they preserve the angle between the vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, and the lengths of them.

15.3 Spectral Decomposition

A symmetric matrix $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ where $\boldsymbol{Q}$ and $\boldsymbol{Q}^\top$ are orthogonal so that they define a system transformation intrinsically. $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top\boldsymbol{u}$ firstly rotates or reflects $\boldsymbol{u}$ into the system defined by $\boldsymbol{Q}^\top$. Let's call this system the **spectral coordinate system**. $\boldsymbol{\Lambda}$ is to stretch each component of the vector in the spectral system by the length of eigenvalue. And then $\boldsymbol{Q}$ helps to rotate or reflect the resulting vector back to the original coordinate system. A demonstration of how the spectral decomposition transforms between coordinate systems is shown in Figure 20 where $\boldsymbol{q}_1, \boldsymbol{q}_2$ are two linearly independent eigenvectors of $\boldsymbol{A}$ such that they form a basis for $\mathbb{R}^2$. The coordinate transformation in the spectral decomposition is similar to that of the eigenvalue decomposition. Except that in the spectral decomposition, the orthogonal vectors transferred by $\boldsymbol{Q}^\top$ are still orthogonal. This is also a property of orthogonal matrices. That is, orthogonal matrices can be viewed as matrices which change the basis of other matrices. Hence they preserve the angle (inner product) between the vectors

$$\boldsymbol{u}^\top\boldsymbol{v} = (\boldsymbol{Q}\boldsymbol{u})^\top(\boldsymbol{Q}\boldsymbol{v}).$$

The above invariance of the inner products of angles between the vectors are preserved, which also relies on the invariance of their lengths:

$$||\boldsymbol{Q}\boldsymbol{u}|| = ||\boldsymbol{u}||.$$

15.4 SVD

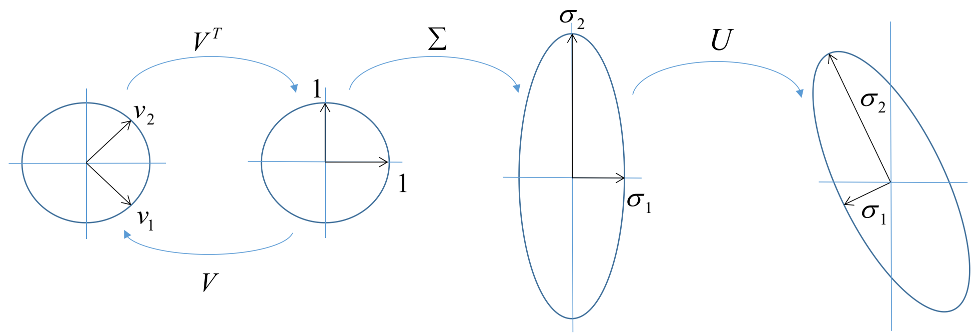


Figure 21: SVD: $\boldsymbol{V}^\top$ and $\boldsymbol{U}$ rotate or reflect, $\boldsymbol{\Sigma}$ stretches the circle to an ellipse. Orthogonal matrices $\boldsymbol{V}^\top$ and $\boldsymbol{U}$ only change the basis of the system. However, they preserve the angle between the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, and the lengths of them.

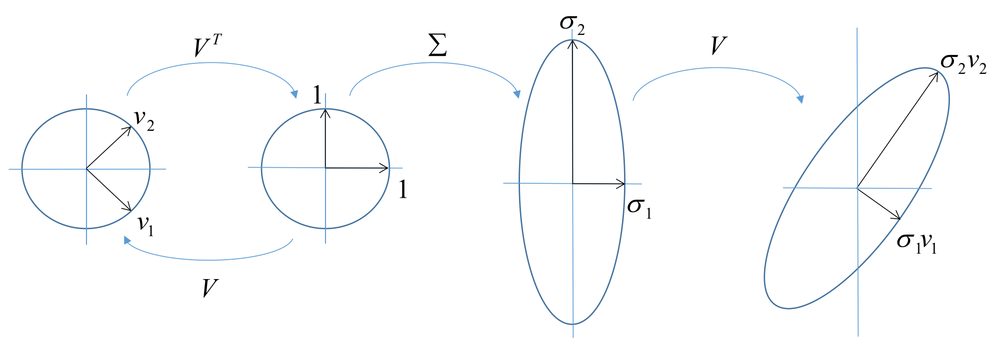


Figure 22: $\boldsymbol{V\Sigma V}^\top$ from SVD or Polar decomposition: $\boldsymbol{V}^\top$ rotates or reflects, $\boldsymbol{\Sigma}$ stretches cycle to ellipse, and $\boldsymbol{V}$ rotates or reflects back. Orthogonal matrices $\boldsymbol{V}^\top$ and $\boldsymbol{V}$ only change the basis of the system. However, they preserve the angle between the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, and the lengths of them.

Any $m \times n$ matrix can be factored as $\boldsymbol{A} = \boldsymbol{U\Sigma V}^\top$. $\boldsymbol{Au} = \boldsymbol{U\Sigma V}^\top\boldsymbol{u}$ then firstly rotates or reflects $\boldsymbol{u}$ into the system defined by $\boldsymbol{V}^\top$, which we call the V **coordinate system**. $\boldsymbol{\Sigma}$ stretches the first r components of the resulted vector in the V system by the lengths of the singular values. If $n \geq m$, then $\boldsymbol{\Sigma}$ only keeps additional $m - r$ components which are stretched to zero while removing the final $n - m$ components. If $m > n$, the $\boldsymbol{\Sigma}$ stretches $n - r$ components to zero, and also adds additional $m - n$ zero components. Finally, $\boldsymbol{U}$ rotates or reflects the resulting vector into the U **coordinate system** defined by $\boldsymbol{U}$. A demonstration of how the SVD transforms in a 2×2 example is shown in Figure 21. Further, Figure 22 demonstrates the transformation of $\boldsymbol{V\Sigma V}^\top$ in a 2×2 example. Similar to the spectral decomposition, orthogonal matrices $\boldsymbol{V}^\top$ and $\boldsymbol{U}$ only change the basis of the system. However, they preserve the angle between the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

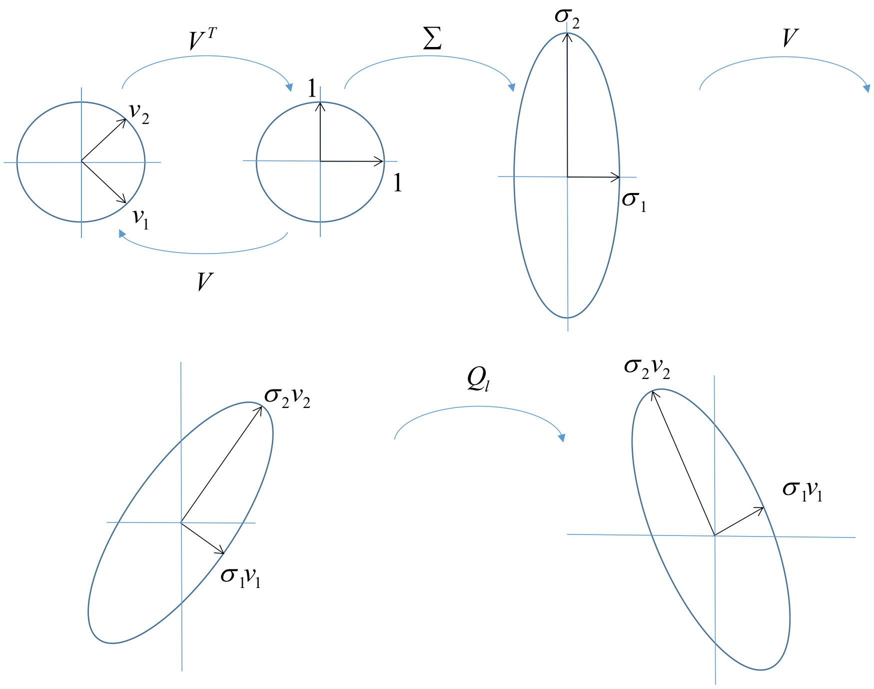


Figure 23: Polar decomposition: $\boldsymbol{V}^\top$ rotates or reflects, $\boldsymbol{\Sigma}$ stretches cycle to ellipse, and $\boldsymbol{V}$ rotates or reflects back. Orthogonal matrices $\boldsymbol{V}^\top$, $\boldsymbol{V}$, $\boldsymbol{Q}_l$ only change the basis of the system. However, they preserve the angle between the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, and the lengths of them.

15.5 Polar Decomposition

Any $n \times n$ square matrix $\boldsymbol{A}$ can be factored as left polar decomposition $\boldsymbol{A} = (\boldsymbol{U}\boldsymbol{V}^\top)(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{Q}_l\boldsymbol{S}$. Similarly, $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{Q}_l(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top)\boldsymbol{u}$ is to transfer $\boldsymbol{u}$ into the system defined by $\boldsymbol{V}^\top$ and stretch each component by the lengths of the singular values. Then the resulted vector is transferred back into the Cartesian coordinate system by $\boldsymbol{V}$. Finally, $\boldsymbol{Q}_l$ will rotate or reflect the resulting vector from the Cartesian coordinate system into the Q system defined by $\boldsymbol{Q}_l$. The meaning of right polar decomposition shares a similar description. Similar to the spectral decomposition, orthogonal matrices $\boldsymbol{V}^\top$ and $\boldsymbol{V}$ only change the basis of the system. However, they preserve the angle between the vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

16. Alternating Least Squares

16.1 Netflix Recommender and Matrix Factorization

In the Netflix prize (Bennett et al., 2007), the goal was to predict ratings of users for different movies, given the existing ratings of those users for other movies. We index M movies with $m = 1, 2, \ldots, M$ and N users by $n = 1, 2, \ldots, N$. We denote the rating of the n-th user for the m-th movie by a_{mn}. Define $\boldsymbol{A}$ to be an $M \times N$ rating matrix with columns $\boldsymbol{a}_n \in \mathbb{R}^M$ containing ratings of the n-th user. Note that many ratings a_{mn} are missing and our goal is to predict those missing ratings accurately.

We formally consider algorithms for solving the following problem: The matrix $\boldsymbol{A}$ is approximately factorized into an $M \times K$ matrix $\boldsymbol{W}$ and a $K \times N$ matrix $\boldsymbol{Z}$. Usually K is chosen to be smaller than M or N, so that $\boldsymbol{W}$ and $\boldsymbol{Z}$ are smaller than the original matrix $\boldsymbol{A}$. This results in a compressed version of the original data matrix. An appropriate decision on the value of K is critical in practice, but the choice of K is very often problem dependent. The factorization is significant in the sense, suppose $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N]$ and $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N]$ are the column partitions of $\boldsymbol{A}, \boldsymbol{Z}$ respectively, then $\boldsymbol{a}_n = \boldsymbol{W}\boldsymbol{z}_n$, i.e., each column $\boldsymbol{a}_n$ is approximated by a linear combination of the columns of $\boldsymbol{W}$ weighted by the components in $\boldsymbol{z}_n$. Therefore, columns of $\boldsymbol{W}$ can be thought of containing column basis of $\boldsymbol{A}$. This is similar to the factorization in the data interpretation part (Part III, p. 76). What's different is that we are not restricting $\boldsymbol{W}$ to be exact columns from $\boldsymbol{A}$.

To find the approximation $\boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{Z}$, we need to define a loss function such that the distance between $\boldsymbol{A}$ and $\boldsymbol{W}\boldsymbol{Z}$ can be measured. The loss function is selected to be the Frobenius norm between two matrices which vanishes to zero if $\boldsymbol{A} = \boldsymbol{W}\boldsymbol{Z}$ where the advantage will be seen shortly.

To simplify the problem, let us assume that there are no missing ratings firstly. Project data vectors $\boldsymbol{a}_n$ to a smaller dimension $\boldsymbol{z}_n \in \mathbb{R}^K$ with $K < M$, such that the *reconstruction error* measured by Frobenius norm is minimized (assume K is known):

$$\min_{\boldsymbol{W},\boldsymbol{Z}} \sum_{n=1}^{N} \sum_{m=1}^{M} \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)^2, \tag{16.1}$$

where $\boldsymbol{W} = [\boldsymbol{w}_1^\top; \boldsymbol{w}_2^\top; \ldots; \boldsymbol{w}_M^\top] \in \mathbb{R}^{M\times K}$ and $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N] \in \mathbb{R}^{K\times N}$ containing $\boldsymbol{w}_m$'s and $\boldsymbol{z}_n$'s as rows and columns respectively. The loss form in Equation (16.1) is known as the *per-example loss*. It can be equivalently written as

$$L(\boldsymbol{W}, \boldsymbol{Z}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)^2 = ||\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}||^2.$$

Moreover, the loss $L(\boldsymbol{W}, \boldsymbol{Z}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)$ is convex with respect to $\boldsymbol{Z}$ given $\boldsymbol{W}$ and vice versa. Therefore, we can first minimize with respect to $\boldsymbol{Z}$ given $\boldsymbol{W}$ and then minimize with respect to $\boldsymbol{W}$ given $\boldsymbol{Z}$:

$$\begin{cases} \boldsymbol{Z} \leftarrow \arg\min_{\boldsymbol{Z}} L(\boldsymbol{W}, \boldsymbol{Z}); & \text{(ALS1)} \\ \boldsymbol{W} \leftarrow \arg\min_{\boldsymbol{W}} L(\boldsymbol{W}, \boldsymbol{Z}). & \text{(ALS2)} \end{cases}$$

This is known as the *coordinate descent algorithm* in which case we employ the least squares, it is also called the *alternating least squares (ALS)* (Comon et al., 2009; Takács and Tikk, 2012; Giampouras et al., 2018). The convergence is guaranteed if the loss function $L(\boldsymbol{W}, \boldsymbol{Z})$ decreases at each iteration and we shall discuss more on this in the sequel.

Remark 16.1: (Convexity and Global Minimum)

Although the loss function defined by Frobenius norm $||\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A}||^2$ is convex in $\boldsymbol{W}$ given $\boldsymbol{Z}$ or vice versa, it is not convex in both variables together. Therefore we are not able to find the global minimum. However, the convergence is assured to find local minima.

Given $\boldsymbol{W}$, Optimizing $\boldsymbol{Z}$ Now, let's see what is in the problem of $\boldsymbol{Z} \leftarrow \arg\min_{\boldsymbol{Z}} L(\boldsymbol{W},\boldsymbol{Z})$. When there exists a unique minimum of the loss function $L(\boldsymbol{W},\boldsymbol{Z})$ with respect to $\boldsymbol{Z}$, we speak of the *least squares* minimizer of $\arg\min_{\boldsymbol{Z}} L(\boldsymbol{W},\boldsymbol{Z})$. Given $\boldsymbol{W}$, $L(\boldsymbol{W},\boldsymbol{Z})$ can be written as $L(\boldsymbol{Z}|\boldsymbol{W})$ to emphasize on the variable of $\boldsymbol{Z}$:

$$L(\boldsymbol{Z}|\boldsymbol{W}) = ||\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A}||^2 = \|\boldsymbol{W}[\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_N]-[\boldsymbol{a}_1,\boldsymbol{a}_2,\ldots,\boldsymbol{a}_N]\|^2 = \left\| \begin{bmatrix} \boldsymbol{W}\boldsymbol{z}_1-\boldsymbol{a}_1 \\ \boldsymbol{W}\boldsymbol{z}_2-\boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{W}\boldsymbol{z}_N-\boldsymbol{a}_N \end{bmatrix} \right\|^2 .$$ [24]

Now, if we define

$$\widetilde{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{W} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{W} \end{bmatrix} \in \mathbb{R}^{MN\times KN}, \qquad \widetilde{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \\ \boldsymbol{z}_N \end{bmatrix} \in \mathbb{R}^{KN}, \qquad \widetilde{\boldsymbol{a}} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_N \end{bmatrix} \in \mathbb{R}^{MN},$$

then the (ALS1) problem can be reduced to the normal least squares for minimizing $||\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{z}}-\widetilde{\boldsymbol{a}}||^2$ with respect to $\widetilde{\boldsymbol{z}}$. And the solution is given by

$$\widetilde{\boldsymbol{z}} = (\widetilde{\boldsymbol{W}}^\top\widetilde{\boldsymbol{W}})^{-1}\widetilde{\boldsymbol{W}}^\top\widetilde{\boldsymbol{a}}.$$

The construction may seem reasonable at first glance. But since $rank(\widetilde{\boldsymbol{W}}) = \min\{M,K\}$, $(\widetilde{\boldsymbol{W}}^\top\widetilde{\boldsymbol{W}})$ is not invertible. A direct way to solve (ALS1) is to find the differential of $L(\boldsymbol{Z}|\boldsymbol{W})$ with respect to $\boldsymbol{Z}$:

$$\begin{aligned} \frac{\partial L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}} &= \frac{\partial\, tr\left((\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A})(\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A})^\top\right)}{\partial \boldsymbol{Z}} \\ &= \frac{\partial\, tr\left((\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A})(\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A})^\top\right)}{\partial(\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A})} \frac{\partial(\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A})}{\partial \boldsymbol{Z}} \\ &\stackrel{\star}{=} 2\boldsymbol{W}^\top(\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A}) \in \mathbb{R}^{K\times N}, \end{aligned} \tag{16.2}$$

where the first equality is from the definition of Frobenius such that $||\boldsymbol{A}|| = \sqrt{\sum_{i=1,j=1}^{m,n}(\boldsymbol{A}_{ij})^2} = \sqrt{tr(\boldsymbol{A}\boldsymbol{A}^\top)}$, and equality $(\star)$ comes from the fact that $\frac{\partial tr(\boldsymbol{A}\boldsymbol{A}^\top)}{\partial \boldsymbol{A}} = 2\boldsymbol{A}$. When the loss function is a differentiable function of $\boldsymbol{Z}$, we may determine the least squares solution by differential calculus, and a minimum of the function $L(\boldsymbol{Z}|\boldsymbol{W})$ must be a root of the equation:

$$\frac{\partial L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}} = \boldsymbol{0}.$$

24. The matrix norm used here is the Frobenius norm such that $||\boldsymbol{A}|| = \sqrt{\sum_{i=1,j=1}^{m,n}(\boldsymbol{A}_{ij})^2}$ if $\boldsymbol{A}\in\mathbb{R}^{m\times n}$. And the vector norm used here is the l_2 norm such that $||\boldsymbol{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$ if $\boldsymbol{x}\in\mathbb{R}^n$.

By finding the root of the above equation, we have the "candidate" update on $\boldsymbol{Z}$ that find the minimizer of $L(\boldsymbol{Z}|\boldsymbol{W})$

$$\boldsymbol{Z} = (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{A} \leftarrow \arg\min_{\boldsymbol{Z}} L(\boldsymbol{Z}|\boldsymbol{W}). \tag{16.3}$$

Before we declare a root of the above equation is actually a minimizer rather than a maximizer (that's why we call the update a "candidate" update above), we need to verify the function is convex such that if the function is twice differentiable, this can be equivalently done by verifying

$$\frac{\partial^2 L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}^2} > 0,$$

i.e., the Hessian matrix is positive definite (recall the definition of positive definiteness, Definition 2.2, p. 29). To see this, we write out the twice differential

$$\frac{\partial^2 L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}^2} = 2\boldsymbol{W}^\top\boldsymbol{W} \in \mathbb{R}^{K\times K},$$

which has full rank if $\boldsymbol{W} \in \mathbb{R}^{M\times K}$ has full rank (Lemma 14.3, p. 135) and $K < M$. We here claim that if $\boldsymbol{W}$ has full rank, then $\frac{\partial^2 L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}^2}$ is positive definite. This can be done by checking that when $\boldsymbol{W}$ has full rank, $\boldsymbol{W}\boldsymbol{x} = \boldsymbol{0}$ only when $\boldsymbol{x} = \boldsymbol{0}$ since the null space of $\boldsymbol{W}$ is of dimension 0. Therefore,

$$\boldsymbol{x}^\top(2\boldsymbol{W}^\top\boldsymbol{W})\boldsymbol{x} > 0, \qquad \text{for any nonzero vector } \boldsymbol{x} \in \mathbb{R}^K.$$

Now, the thing is that we need to check if $\boldsymbol{W}$ has full rank so that the Hessian of $L(\boldsymbol{Z}|\boldsymbol{W})$ is positive definiteness, otherwise, we cannot claim the update of $\boldsymbol{Z}$ in Equation (16.3) decreases the loss so that the matrix decomposition is going into the right way to better approximate the original matrix $\boldsymbol{A}$ by $\boldsymbol{W}\boldsymbol{Z}$. We will shortly come back to the positive definiteness of the Hessian matrix in the sequel which relies on the following lemma

Lemma 16.2: (Rank of $\boldsymbol{Z}$ after Updating)

Suppose $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ has full rank with $M \leq N$ and $\boldsymbol{W} \in \mathbb{R}^{M\times K}$ has full rank with $K < M$, then the update of $\boldsymbol{Z} = (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{A} \in \mathbb{R}^{K\times N}$ in Equation (16.3) has full rank.

Proof [of Lemma 16.2] Since $\boldsymbol{W}^\top\boldsymbol{W} \in \mathbb{R}^{K\times K}$ has full rank if $\boldsymbol{W}$ has full rank (Lemma 14.3, p. 135) such that $(\boldsymbol{W}^\top\boldsymbol{W})^{-1}$ has full rank.

Suppose $\boldsymbol{W}^\top\boldsymbol{x} = \boldsymbol{0}$, this implies $(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{x} = \boldsymbol{0}$. Thus

$$\mathcal{N}(\boldsymbol{W}^\top) \subseteq \mathcal{N}\left((\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right).$$

Moreover, suppose $(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{x} = \boldsymbol{0}$, and since $(\boldsymbol{W}^\top\boldsymbol{W})^{-1}$ is invertible. This implies $\boldsymbol{W}^\top\boldsymbol{x} = (\boldsymbol{W}^\top\boldsymbol{W})\boldsymbol{0} = \boldsymbol{0}$, and

$$\mathcal{N}\left((\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right) \subseteq \mathcal{N}(\boldsymbol{W}^\top).$$

As a result, by "sandwiching", it follows that

$$\mathcal{N}(\boldsymbol{W}^\top) = \mathcal{N}\left((\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\right). \tag{16.4}$$

Therefore, $(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$ has full rank K. Let $\boldsymbol{T} = (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top \in \mathbb{R}^{K\times M}$, and suppose $\boldsymbol{T}^\top\boldsymbol{x} = \mathbf{0}$. This implies $\boldsymbol{A}^\top\boldsymbol{T}^\top\boldsymbol{x} = \mathbf{0}$, and

$$\mathcal{N}(\boldsymbol{T}^\top) \subseteq \mathcal{N}(\boldsymbol{A}^\top\boldsymbol{T}^\top).$$

Similarly, suppose $\boldsymbol{A}^\top(\boldsymbol{T}^\top\boldsymbol{x}) = \mathbf{0}$. Since $\boldsymbol{A}$ has full rank with the dimension of the null space being 0: $dim\left(\mathcal{N}(\boldsymbol{A}^\top)\right) = 0$, $(\boldsymbol{T}^\top\boldsymbol{x})$ must be zero. The claim follows from that since $\boldsymbol{A}$ has full rank M with the row space of $\boldsymbol{A}^\top$ being equal to the column space of $\boldsymbol{A}$ where $dim\left(\mathcal{C}(\boldsymbol{A})\right) = M$ and the $dim\left(\mathcal{N}(\boldsymbol{A}^\top)\right) = M - dim\left(\mathcal{C}(\boldsymbol{A})\right) = 0$. Therefore, $\boldsymbol{x}$ is in the null space of $\boldsymbol{T}^\top$ if $\boldsymbol{x}$ is in the null space of $\boldsymbol{A}^\top\boldsymbol{T}^\top$:

$$\mathcal{N}(\boldsymbol{A}^\top\boldsymbol{T}^\top) \subseteq \mathcal{N}(\boldsymbol{T}^\top).$$

By "sandwiching" again,

$$\mathcal{N}(\boldsymbol{T}^\top) = \mathcal{N}(\boldsymbol{A}^\top\boldsymbol{T}^\top). \tag{16.5}$$

Since $\boldsymbol{T}^\top$ has full rank $K < M < N$, $dim\left(\mathcal{N}(\boldsymbol{T}^\top)\right) = dim\left(\mathcal{N}(\boldsymbol{A}^\top\boldsymbol{T}^\top)\right) = 0$. Therefore, $\boldsymbol{Z}^\top = \boldsymbol{A}^\top\boldsymbol{T}^\top$ has full rank K. We complete the proof. ■

Given $\boldsymbol{Z}$, Optimizing $\boldsymbol{W}$ Given $\boldsymbol{Z}$, $L(\boldsymbol{W},\boldsymbol{Z})$ can be written as $L(\boldsymbol{W}|\boldsymbol{Z})$ to emphasize on the variable of $\boldsymbol{W}$:

$$L(\boldsymbol{W}|\boldsymbol{Z}) = ||\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}||^2.$$

A direct way to solve (ALS2) is to find the differential of $L(\boldsymbol{W}|\boldsymbol{Z})$ with respect to $\boldsymbol{W}$:

$$\begin{aligned}
\frac{\partial L(\boldsymbol{W}|\boldsymbol{Z})}{\partial \boldsymbol{W}} &= \frac{\partial\, tr\left((\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A})(\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A})^\top\right)}{\partial \boldsymbol{W}} \\
&= \frac{\partial\, tr\left((\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A})(\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A})^\top\right)}{\partial(\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A})}\frac{\partial(\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A})}{\partial \boldsymbol{W}} \\
&= 2(\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A})\boldsymbol{Z}^\top \in \mathbb{R}^{M\times K}.
\end{aligned}$$

The "candidate" update on $\boldsymbol{W}$ is similarly to find the root of the differential $\frac{\partial L(\boldsymbol{W}|\boldsymbol{Z})}{\partial \boldsymbol{W}}$:

$$\boldsymbol{W}^\top = (\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1}\boldsymbol{Z}\boldsymbol{A}^\top \leftarrow \mathop{\arg\min}_{\boldsymbol{W}} L(\boldsymbol{W}|\boldsymbol{Z}). \tag{16.6}$$

Again, we emphasize that the update is only a "candidate" update. We need to further check whether the Hessian is positive definite or not. The Hessian matrix is given by

$$\frac{\partial^2 L(\boldsymbol{W}|\boldsymbol{Z})}{\partial \boldsymbol{W}^2} = 2\boldsymbol{Z}\boldsymbol{Z}^\top \in \mathbb{R}^{K\times K}.$$

Therefore, by analogous analysis, if $\boldsymbol{Z}$ has full rank with $K < N$, the Hessian matrix is positive definite.

Lemma 16.3: (Rank of $\boldsymbol{W}$ after Updating)

Suppose $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ has full rank with $M \leq N$ and $\boldsymbol{Z} \in \mathbb{R}^{K\times N}$ has full rank with $K < N$, then the update of $\boldsymbol{W}^\top = (\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1}\boldsymbol{Z}\boldsymbol{A}^\top$ in Equation (16.6) has full rank.

Proof [of Lemma 16.3] The proof is slightly different to that of Lemma 16.2. Since $\boldsymbol{Z} \in \mathbb{R}^{K\times N}$ and $\boldsymbol{A}^\top \in \mathbb{R}^{N\times M}$ have full rank, i.e., $\det(\boldsymbol{Z}) > 0$ and $\det(\boldsymbol{A}^\top) > 0$. The determinant of their product $\det(\boldsymbol{Z}\boldsymbol{A}^\top) = \det(\boldsymbol{Z})\det(\boldsymbol{A}^\top) > 0$ such that $\boldsymbol{Z}\boldsymbol{A}^\top$ has full rank (rank K). Similarly argument can find $\boldsymbol{W}^\top$ also has full rank. ■

Combine the observations in Lemma 16.2 and Lemma 16.3, as long as we initialize $\boldsymbol{Z}, \boldsymbol{W}$ to have full rank, the updates in Equation (16.3) and Equation (16.6) are reasonable. The requirement on the $M \leq N$ is reasonable in that there are always more users than the number of movies. We conclude the process in Algorithm 4.

Algorithm 4 Alternating Least Squares

Require: $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ with $M \leq N$;
1: initialize $\boldsymbol{W} \in \mathbb{R}^{M\times K}$, $\boldsymbol{Z} \in \mathbb{R}^{K\times N}$ with full rank and $K < M \leq N$;
2: choose a stop criterion on the approximation error δ;
3: choose maximal number of iterations C;
4: $iter = 0$;
5: **while** $||\boldsymbol{A} - \boldsymbol{W}\boldsymbol{Z}|| > \delta$ and $iter < C$ **do**
6: $\quad iter = iter + 1$;
7: $\quad \boldsymbol{Z} = (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{A} \leftarrow \arg\min_{\boldsymbol{Z}} L(\boldsymbol{Z}|\boldsymbol{W})$;
8: $\quad \boldsymbol{W}^\top = (\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1}\boldsymbol{Z}\boldsymbol{A}^\top \leftarrow \arg\min_{\boldsymbol{W}} L(\boldsymbol{W}|\boldsymbol{Z})$;
9: **end while**
10: Output $\boldsymbol{W}, \boldsymbol{Z}$;

16.2 Regularization: Extension to General Matrices

We can add a regularization to minimize the following loss:

$$L(\boldsymbol{W}, \boldsymbol{Z}) = ||\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}||^2 + \lambda_w||\boldsymbol{W}||^2 + \lambda_z||\boldsymbol{Z}||^2, \qquad \lambda_w > 0, \lambda_z > 0, \tag{16.7}$$

where the differential with respect to $\boldsymbol{Z}, \boldsymbol{W}$ are given respectively by

$$\begin{cases} \dfrac{\partial L(\boldsymbol{W}, \boldsymbol{Z})}{\partial \boldsymbol{Z}} = 2\boldsymbol{W}^\top(\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}) + 2\lambda_z\boldsymbol{Z} \in \mathbb{R}^{K\times N}; \\ \dfrac{\partial L(\boldsymbol{W}, \boldsymbol{Z})}{\partial \boldsymbol{W}} = 2(\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A})\boldsymbol{Z}^\top + 2\lambda_w\boldsymbol{W} \in \mathbb{R}^{M\times K}. \end{cases} \tag{16.8}$$

The Hessian matrices are given respectively by

$$\begin{cases} \dfrac{\partial^2 L(\boldsymbol{W}, \boldsymbol{Z})}{\partial \boldsymbol{Z}^2} = 2\boldsymbol{W}^\top\boldsymbol{W} + 2\lambda_z\boldsymbol{I} \in \mathbb{R}^{K\times K}; \\ \dfrac{\partial^2 L(\boldsymbol{W}, \boldsymbol{Z})}{\partial \boldsymbol{W}^2} = 2\boldsymbol{Z}\boldsymbol{Z}^\top + 2\lambda_w\boldsymbol{I} \in \mathbb{R}^{K\times K}, \end{cases}$$

which are positive definite due to the perturbation by the regularization. To see this,

$$\begin{cases} \boldsymbol{x}^\top(2\boldsymbol{W}^\top\boldsymbol{W}+2\lambda_z\boldsymbol{I})\boldsymbol{x} = \underbrace{2\boldsymbol{x}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{x}}_{\geq 0} + 2\lambda_z||\boldsymbol{x}||^2 > 0, & \text{for nonzero } \boldsymbol{x}; \\ \boldsymbol{x}^\top(2\boldsymbol{Z}\boldsymbol{Z}^\top+2\lambda_w\boldsymbol{I})\boldsymbol{x} = \underbrace{2\boldsymbol{x}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{x}}_{\geq 0} + 2\lambda_w||\boldsymbol{x}||^2 > 0, & \text{for nonzero } \boldsymbol{x}. \end{cases}$$

The regularization makes the Hessian matrices positive definite even if $\boldsymbol{W}, \boldsymbol{Z}$ are rank deficient. And now the matrix decomposition can be extended to any matrix even when $M > N$. In rare cases, K can be chosen as $K > \max\{M, N\}$ such that a high-rank approximation of $\boldsymbol{A}$ is obtained. However, in most scenarios, we want to find the low-rank approximation of $\boldsymbol{A}$ such that $K < \min\{M, N\}$. For example, the ALS can be utilized to find the low-rank neural networks to reduce the memory of the neural networks whilst increase the performance (Lu, 2021c).

Therefore, the minimizers are given by finding the roots of the differential:

$$\begin{cases} \boldsymbol{Z} = (\boldsymbol{W}^\top\boldsymbol{W}+\lambda_z\boldsymbol{I})^{-1}\boldsymbol{W}^\top\boldsymbol{A}; \\ \boldsymbol{W}^\top = (\boldsymbol{Z}\boldsymbol{Z}^\top+\lambda_w\boldsymbol{I})^{-1}\boldsymbol{Z}\boldsymbol{A}^\top. \end{cases} \tag{16.9}$$

The regularization parameters $\lambda_z, \lambda_2 \in \mathbb{R}$ are used to balance the trade-off between the accuracy of the approximation and the smoothness of the computed solution. The choice on the selection of the parameters is typically problem dependent and can be obtained by *cross-validation*.

16.3 Missing Entries

Since the matrix decomposition via the ALS is extensively used in the Netflix recommender data, where many entries are missing since many users have not watched some movies or they will not rate the movies for some reasons. We can make an additional mask matrix $\boldsymbol{M} \in \mathbb{R}^{M\times N}$ where $\boldsymbol{M}_{mn} \in \{0, 1\}$ means if the user n has rated the movie m or not. Therefore, the loss function can be defined as

$$L(\boldsymbol{W}, \boldsymbol{Z}) = ||\boldsymbol{M} \odot \boldsymbol{A} - \boldsymbol{M} \odot (\boldsymbol{W}\boldsymbol{Z})||^2,$$

where $\odot$ is the *Hadamard product* between matrices. For example, the Hadamard product for a 3×3 matrix $\boldsymbol{A}$ with a 3×3 matrix $\boldsymbol{B}$ is

$$\boldsymbol{A} \odot \boldsymbol{B} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \odot \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & a_{13}b_{13} \\ a_{21}b_{21} & a_{22}b_{22} & a_{23}b_{23} \\ a_{31}b_{31} & a_{32}b_{32} & a_{33}b_{33} \end{bmatrix}.$$

To find the solution of the problem, let's decompose the updates in Equation (16.9) into:

$$\begin{cases} \boldsymbol{z}_n = (\boldsymbol{W}^\top\boldsymbol{W}+\lambda_z\boldsymbol{I})^{-1}\boldsymbol{W}^\top\boldsymbol{a}_n, & \text{for } n \in \{1, 2, \ldots, N\}; \\ \boldsymbol{w}_m = (\boldsymbol{Z}\boldsymbol{Z}^\top+\lambda_w\boldsymbol{I})^{-1}\boldsymbol{Z}\boldsymbol{b}_m, & \text{for } m \in \{1, 2, \ldots, M\}, \end{cases} \tag{16.10}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N], \boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N]$ are the column partitions of $\boldsymbol{Z}, \boldsymbol{A}$ respectively. And $\boldsymbol{W}^\top = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M], \boldsymbol{A}^\top = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_M]$ are the column partitions of $\boldsymbol{W}^\top, \boldsymbol{A}^\top$ respectively. The factorization of the updates indicates the update can be done via a column by column fashion.

Given $\boldsymbol{W}$ Let $\boldsymbol{o}_n \in \mathbb{R}^M$ denote the movies rated by user n where $o_{nm} = 1$ if user n has rated movie m, and $o_{nm} = 1$ otherwise. Then the n-th column of $\boldsymbol{A}$ without missing entries can be denoted as the matlab style notation $\boldsymbol{a}_n[\boldsymbol{o}_n]$. And we want to approximate the existing n-th column by $\boldsymbol{a}_n[\boldsymbol{o}_n] \approx \boldsymbol{W}[\boldsymbol{o}_n,:]\boldsymbol{z}_n$ which is actually a rank-one least squares problem:

$$\boldsymbol{z}_n = \left(\boldsymbol{W}[\boldsymbol{o}_n,:]^\top \boldsymbol{W}[\boldsymbol{o}_n,:] + \lambda_z \boldsymbol{I}\right)^{-1} \boldsymbol{W}[\boldsymbol{o}_n,:]^\top \boldsymbol{a}_n[\boldsymbol{o}_n], \qquad \text{for } n \in \{1,2,\ldots,N\}. \tag{16.11}$$

Moreover, the loss function with respect to $\boldsymbol{z}_n$:

$$L(\boldsymbol{z}_n|\boldsymbol{W}) = \sum_{m\in \boldsymbol{o}_n} \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)^2$$

and if we are concerned about the loss for all users:

$$L(\boldsymbol{Z}|\boldsymbol{W}) = \sum_{n=1}^{N} \sum_{m\in \boldsymbol{o}_n} \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)^2$$

Given $\boldsymbol{Z}$ Similarly, if $\boldsymbol{p}_m \in \mathbb{R}^N$ denotes the users that have rated the movie m with $\boldsymbol{p}_{dn} = 1$ if the movie m has been rated by user n. Then the m-th row of $\boldsymbol{A}$ without missing entries can be denoted as the matlab style notation $\boldsymbol{b}_m[\boldsymbol{p}_m]$. And we want to approximate the existing m-th row by $\boldsymbol{b}_m[\boldsymbol{p}_m] \approx \boldsymbol{Z}[:,\boldsymbol{p}_m]^\top \boldsymbol{w}_m$, [25] which again is a rank-one least squares problem:

$$\boldsymbol{w}_m = (\boldsymbol{Z}[:,\boldsymbol{p}_m]\boldsymbol{Z}[:,\boldsymbol{p}_m]^\top + \lambda_w \boldsymbol{I})^{-1}\boldsymbol{Z}[:,\boldsymbol{p}_m]\boldsymbol{b}_m[\boldsymbol{p}_m], \qquad \text{for } m \in \{1,2,\ldots,M\}. \tag{16.12}$$

Moreover, the loss function with respect to $\boldsymbol{w}_n$:

$$L(\boldsymbol{w}_n|\boldsymbol{Z}) = \sum_{n\in \boldsymbol{p}_m} \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)^2$$

and if we are concerned about the loss for all users:

$$L(\boldsymbol{W}|\boldsymbol{Z}) = \sum_{d=1}^{M} \sum_{n\in \boldsymbol{p}_m} \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)^2$$

16.4 Vector Inner Product

We have seen the ALS is to find matrices $\boldsymbol{W}, \boldsymbol{Z}$ such that $\boldsymbol{W}\boldsymbol{Z}$ can approximate $\boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{Z}$ in terms of minimum least squared loss:

$$\min_{\boldsymbol{W},\boldsymbol{Z}} \sum_{n=1}^{N}\sum_{d=1}^{M} \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)^2,$$

25. Note that $\boldsymbol{Z}[:,\boldsymbol{p}_m]^\top$ is the transpose of $\boldsymbol{Z}[:,\boldsymbol{p}_m]$, which is equal to $\boldsymbol{Z}^\top[\boldsymbol{p}_m,:]$, i.e., transposing first and then selecting.

that is, each entry a_{mn} in $\boldsymbol{A}$ can be approximated by the inner product between the two vectors $\boldsymbol{w}_m^\top \boldsymbol{z}_n$. The geometric definition of vector inner product is given by

$$\boldsymbol{w}_m^\top \boldsymbol{z}_n = ||\boldsymbol{w}|| \cdot ||\boldsymbol{z}|| \cos\theta,$$

where θ is the angle between $\boldsymbol{w}$ and $\boldsymbol{z}$. So if the vector norms of $\boldsymbol{w}, \boldsymbol{z}$ are determined, the smaller the angle, the larger the inner product.

Come back to the Netflix data, where the rating are ranging from 0 to 5 and the larger the better. If $\boldsymbol{w}_m$ and $\boldsymbol{z}_n$ fall "close" enough, then $\boldsymbol{w}^\top \boldsymbol{z}$ will have a larger value. This reveals the meaning behind the ALS where $\boldsymbol{w}_m$ represents the features of movie m, whilst $\boldsymbol{z}_n$ contains the features of user n. And each element in $\boldsymbol{w}_m$ and $\boldsymbol{z}_n$ represents a same feature. For example, it could be that the second feature $\boldsymbol{w}_{m2}$ [26] represents if the movie is an action movie or not, and $\boldsymbol{z}_{n2}$ denotes if the user n likes action movies or not. If it happens the case, then $\boldsymbol{w}_m^\top \boldsymbol{z}_n$ will be large and approximates a_{mn} well.

Note that, in the decomposition $\boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{Z}$, we know the rows of $\boldsymbol{W}$ contain the hidden features of the movies and the columns of $\boldsymbol{Z}$ contain the hidden features of the users. However, we cannot identify what are the meanings of the rows of $\boldsymbol{W}$ or the columns of $\boldsymbol{Z}$. We know they could be something like categories or genres of the movies, that provide some underlying connections between the users and the movies, but we cannot be sure what exactly they are. This is where the terminology "hidden" comes from.

16.5 Gradient Descent

In Equation (16.10), we obtain the column-by-column update directly from the full matrix way in Equation (16.9) (with regularization considered). Now let's see what's behind the idea. Following from Equation (16.7), the loss under the regularization:

$$L(\boldsymbol{W}, \boldsymbol{Z}) = ||\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}||^2 + \lambda_w ||\boldsymbol{W}||^2 + \lambda_z ||\boldsymbol{Z}||^2, \qquad \lambda_w > 0, \lambda_z > 0, \tag{16.13}$$

Since we are now considering the minimization of above loss with respect to $\boldsymbol{z}_n$, we can decompose the loss into

$$\begin{aligned} L(\boldsymbol{z}_n) &= ||\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}||^2 + \lambda_w ||\boldsymbol{W}||^2 + \lambda_z ||\boldsymbol{Z}||^2 \\ &= ||\boldsymbol{W}\boldsymbol{z}_n - \boldsymbol{a}_n||^2 + \lambda_z ||\boldsymbol{z}_n||^2 + \underbrace{\sum_{i \neq n} ||\boldsymbol{W}\boldsymbol{z}_i - \boldsymbol{a}_i||^2 + \lambda_z \sum_{i \neq n} ||\boldsymbol{z}_i||^2 + \lambda_w ||\boldsymbol{W}||^2}_{C_{z_n}}, \end{aligned} \tag{16.14}$$

where C_{z_n} is a constant with respect to $\boldsymbol{z}_n$, and $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N], \boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N]$ are the column partitions of $\boldsymbol{Z}, \boldsymbol{A}$ respectively. Taking the differential

$$\frac{\partial L(\boldsymbol{z}_n)}{\partial \boldsymbol{z}_n} = 2\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{z}_n - 2\boldsymbol{W}^\top \boldsymbol{a}_n + 2\lambda_z \boldsymbol{z}_n,$$

under which the root is exactly the first update of the column fashion in Equation (16.10):

$$\boldsymbol{z}_n = (\boldsymbol{W}^\top \boldsymbol{W} + \lambda_z \boldsymbol{I})^{-1} \boldsymbol{W}^\top \boldsymbol{a}_n, \qquad \text{for } n \in \{1, 2, \ldots, N\}.$$

26. $\boldsymbol{w}_{m2}$ is the second element of vector $\boldsymbol{w}_{m2}$

Similarly, we can decompose the loss with respect to $\boldsymbol{w}_m$,

$$\begin{aligned} L(\boldsymbol{w}_m) &= ||\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{A}||^2+\lambda_w||\boldsymbol{W}||^2+\lambda_z||\boldsymbol{Z}||^2\\ &= ||\boldsymbol{Z}^\top\boldsymbol{W}-\boldsymbol{A}^\top||^2+\lambda_w||\boldsymbol{W}^\top||^2+\lambda_z||\boldsymbol{Z}||^2\\ &= ||\boldsymbol{Z}^\top\boldsymbol{w}_m-\boldsymbol{b}_n||^2+\lambda_w||\boldsymbol{w}_m||^2+\underbrace{\sum_{i\neq m}||\boldsymbol{Z}^\top\boldsymbol{w}_i-\boldsymbol{b}_i||^2+\lambda_w\sum_{i\neq m}||\boldsymbol{w}_i||^2+\lambda_z||\boldsymbol{Z}||^2}_{C_{w_m}}, \end{aligned} \tag{16.15}$$

where C_{w_m} is a constant with respect to $\boldsymbol{w}_m$, and $\boldsymbol{W}^\top=[\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_M]$, $\boldsymbol{A}^\top=[\boldsymbol{b}_1,\boldsymbol{b}_2,\ldots,\boldsymbol{b}_M]$ are the column partitions of $\boldsymbol{W}^\top,\boldsymbol{A}^\top$ respectively. Analogously, taking the differential with respect to $\boldsymbol{w}_m$, it follows that

$$\frac{\partial L(\boldsymbol{w}_m)}{\partial \boldsymbol{w}_m}=2\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{w}_m-2\boldsymbol{Z}\boldsymbol{b}_n+2\lambda_w\boldsymbol{w}_m,$$

under which the root is exactly the second update of the column fashion in Equation (16.10):

$$\boldsymbol{w}_m=(\boldsymbol{Z}\boldsymbol{Z}^\top+\lambda_w\boldsymbol{I})^{-1}\boldsymbol{Z}\boldsymbol{b}_m, \qquad \text{for } m\in\{1,2,\ldots,M\}.$$

Now suppose we write out the iteration as the superscript and we want to find the updates $\{\boldsymbol{z}_n^{(k+1)},\boldsymbol{w}_m^{(k+1)}\}$ base on $\{\boldsymbol{Z}^{(k)},\boldsymbol{W}^{(k)}\}$:

$$\begin{cases} \boldsymbol{z}_n^{(k+1)} \leftarrow \mathop{\arg\min}\limits_{\boldsymbol{z}_n^{(k)}} L(\boldsymbol{z}_n^{(k)});\\ \boldsymbol{w}_m^{(k+1)} \leftarrow \mathop{\arg\min}\limits_{\boldsymbol{w}_m^{(k)}} L(\boldsymbol{w}_m^{(k)}). \end{cases}$$

For simplicity, we will be looking at $\boldsymbol{z}_n^{(k+1)}\leftarrow \arg\min_{\boldsymbol{z}_n^{(k)}} L(\boldsymbol{z}_n^{(k)}|-)$, and the derivation for the update on $\boldsymbol{w}_m^{(k+1)}$ will be the same. Suppose we want to approximate $\boldsymbol{z}_n^{(k+1)}$ by a linear update on $\boldsymbol{z}_n^{(k)}$:

$$\boldsymbol{z}_n^{(k+1)}=\boldsymbol{z}_n^{(k)}+\eta\boldsymbol{v}.$$

The problem now turns to the solution of $\boldsymbol{v}$ such that

$$\boldsymbol{v}=\mathop{\arg\min}_{\boldsymbol{v}} L(\boldsymbol{z}_n^{(k)}+\eta\boldsymbol{v}).$$

By Taylor's formula, $L(\boldsymbol{z}_n^{(k)}+\eta\boldsymbol{v})$ can be approximated by

$$L(\boldsymbol{z}_n^{(k)}+\eta\boldsymbol{v})\approx L(\boldsymbol{z}_n^{(k)})+\eta\boldsymbol{v}^\top\nabla L(\boldsymbol{z}_n^{(k)}),$$

when η is small enough. Then an search under the condition $||\boldsymbol{v}||=1$ given positive η is as follows:

$$\boldsymbol{v}=\mathop{\arg\min}_{||\boldsymbol{v}||=1} L(\boldsymbol{z}_n^{(k)}+\eta\boldsymbol{v})\approx \mathop{\arg\min}_{||\boldsymbol{v}||=1}\left\{L(\boldsymbol{z}_n^{(k)})+\eta\boldsymbol{v}^\top\nabla L(\boldsymbol{z}_n^{(k)})\right\}.$$

This is known as the *greedy search*. The optimal $\boldsymbol{v}$ can be obtained by

$$\boldsymbol{v}=-\frac{\nabla L(\boldsymbol{z}_n^{(k)})}{||\nabla L(\boldsymbol{z}_n^{(k)})||},$$

i.e., $\boldsymbol{v}$ is in the opposite direction of $\nabla L(\boldsymbol{z}_n^{(k)})$. Therefore, the update of $\boldsymbol{z}_n^{(k+1)}$ is reasonable to be taken as

$$\boldsymbol{z}_n^{(k+1)} = \boldsymbol{z}_n^{(k)} + \eta \boldsymbol{v} = \boldsymbol{z}_n^{(k)} - \eta \frac{\nabla L(\boldsymbol{z}_n^{(k)})}{||\nabla L(\boldsymbol{z}_n^{(k)})||},$$

which usually called the *gradient descent*. Similarly, the gradient descent of $\boldsymbol{w}_m^{(k+1)}$ is given by

$$\boldsymbol{w}_m^{(k+1)} = \boldsymbol{w}_m^{(k)} + \eta \boldsymbol{v} = \boldsymbol{w}_m^{(k)} - \eta \frac{\nabla L(\boldsymbol{w}_m^{(k)})}{||\nabla L(\boldsymbol{w}_m^{(k)})||}.$$

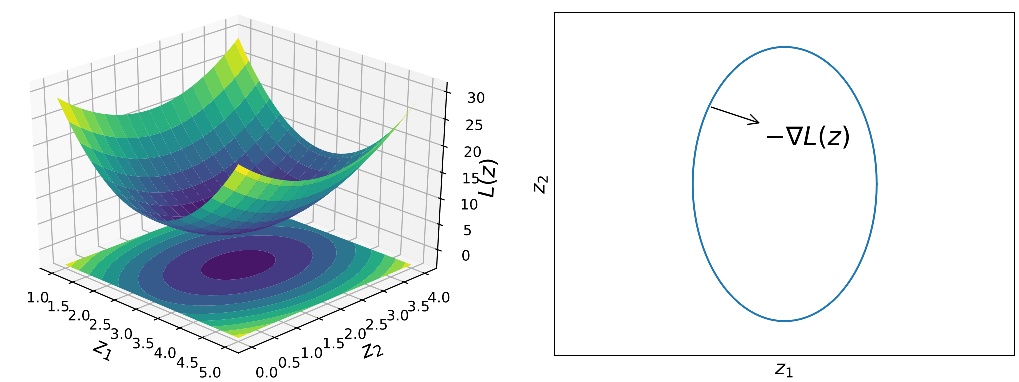
Geometrical Interpretation of Gradient Descent

Lemma 16.4: (Direction of Gradients)

An important fact is that gradients are orthogonal to level curves (a.k.a., level surface).

See (Lu, 2021c) for a proof.

The lemma above reveals the geometrical interpretation of gradient descent. For finding a solution to minimize a convex function $L(\boldsymbol{z})$, gradient descent goes to the negative gradient direction that can decrease the loss. Figure 24 depicts a 2-dimensional case, where $-\nabla L(\boldsymbol{z})$ pushes the loss to decrease for the convex function $L(\boldsymbol{z})$.



(a) A 2-dimensional convex function $L(\boldsymbol{z})$

(b) $L(\boldsymbol{z}) = c$ is a constant

Figure 24: Figure 24(a) shows a function "density" and a contour plot (blue=low, yellow=high) where the upper graph is the "density", and the lower one is the projection of it (i.e., contour). Figure 24(b): $-\nabla L(\boldsymbol{z})$ pushes the loss to decrease for the convex function $L(\boldsymbol{z})$.

16.6 Regularization: A Geometrical Interpretation

We have seen in Section 16.2 that the regularization can extend the ALS to general matrices. The gradient descent can reveal the geometric meaning of the regularization. To avoid confusion, we denote the loss function without regularization by $l(\boldsymbol{z})$ and the loss with regularization by $L(\boldsymbol{z}) = l(\boldsymbol{z}) + \lambda_z ||\boldsymbol{z}||^2$ where $l(\boldsymbol{z}) : \mathbb{R}^n \to \mathbb{R}$. When minimizing $l(\boldsymbol{z})$,

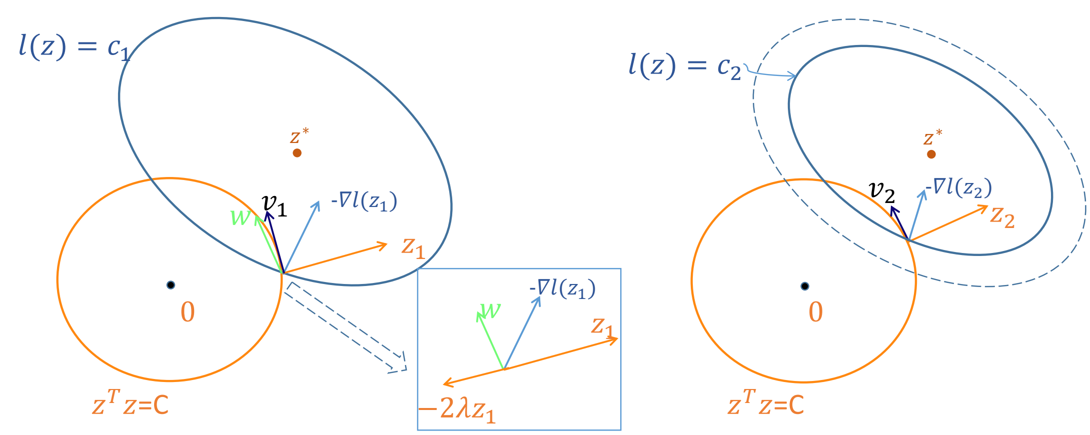


Figure 25: Constrained gradient descent with $\boldsymbol{z}^\top \boldsymbol{z} \leq C$. The green vector $\boldsymbol{w}$ is the projection of $\boldsymbol{v}_1$ into $\boldsymbol{z}^\top \boldsymbol{z} \leq C$ where $\boldsymbol{v}_1$ is the component of $-\nabla l(\boldsymbol{z})$ perpendicular to $\boldsymbol{z}_1$. The right picture is the next step after the update in the left picture. $\boldsymbol{z}^\star$ denotes the optimal solution of $\{\min l(\boldsymbol{z})\}$.

descent method will search in $\mathbb{R}^n$ for a solution. However, in machine learning, searching in the whole space can cause overfitting. A partial solution is to search in a subset of the vector space, e.g., searching in $\boldsymbol{z}^\top \boldsymbol{z} < C$ for some constant C. That is

$$\arg\min_{\boldsymbol{z}} \quad l(\boldsymbol{z}), \qquad s.t., \qquad \boldsymbol{z}^\top \boldsymbol{z} \leq C.$$

As shown above, a trivial gradient descent method will go further in the direction of $-\nabla l(\boldsymbol{z})$, i.e., update $\boldsymbol{z}$ by $\boldsymbol{z} \leftarrow \boldsymbol{z} - \eta \nabla l(\boldsymbol{z})$ for small step size η. When the level curve is $l(\boldsymbol{z}) = c_1$ and the current position of $\boldsymbol{z} = \boldsymbol{z}_1$ where $\boldsymbol{z}_1$ is the intersection of $\boldsymbol{z}^\top \boldsymbol{z} = C$ and $l(\boldsymbol{z}) = c_1$, the descent direction $-\nabla l(\boldsymbol{z}_1)$ will be perpendicular to the level curve of $l(\boldsymbol{z}_1) = c_1$ as shown in the left picture of Figure 25. However, if we further restrict the optimal value can only be in $\boldsymbol{z}^\top \boldsymbol{z} \leq C$, the trivial descent direction $-\nabla l(\boldsymbol{z}_1)$ will lead $\boldsymbol{z}_2 = \boldsymbol{z}_1 - \eta \nabla l(\boldsymbol{z}_1)$ outside of $\boldsymbol{z}^\top \boldsymbol{z} \leq C$. A solution is to decompose the step $-\nabla l(\boldsymbol{z}_1)$ into

$$-\nabla l(\boldsymbol{z}_1) = a\boldsymbol{z}_1 + \boldsymbol{v}_1,$$

where $a\boldsymbol{z}_1$ is the component perpendicular to the curve of $\boldsymbol{z}^\top \boldsymbol{z} = C$, and $\boldsymbol{v}_1$ is the component parallel to the curve of $\boldsymbol{z}^\top \boldsymbol{z} = C$. Keep only the step $\boldsymbol{v}_1$, then the update

$$\boldsymbol{z}_2 = \text{project}(\boldsymbol{z}_1 + \eta \boldsymbol{v}_1) = \text{project}\left(\boldsymbol{z}_1 + \eta \underbrace{(-\nabla l(\boldsymbol{z}_1) - a\boldsymbol{z}_1)}_{\boldsymbol{v}_1}\right)$$ [27]

will lead to a smaller loss from $l(\boldsymbol{z}_1)$ to $l(\boldsymbol{z}_2)$ and still match $\boldsymbol{z}^\top \boldsymbol{z} \leq C$. This is known as the *projection gradient descent*. It is not hard to see that the update $\boldsymbol{z}_2 = \text{project}(\boldsymbol{z}_1 + \eta \boldsymbol{v}_1)$ is equivalent to finding a vector $\boldsymbol{w}$ (shown by the green vector in the left picture of Figure 25)

27. where the project($\boldsymbol{x}$) will project the vector $\boldsymbol{x}$ to the closest point inside $\boldsymbol{z}^\top \boldsymbol{z} \leq C$. Notice here the direct update $\boldsymbol{z}_2 = \boldsymbol{z}_1 + \eta \boldsymbol{v}_1$ can still make $\boldsymbol{z}_2$ outside the curve of $\boldsymbol{z}^\top \boldsymbol{z} \leq C$.

such that $\boldsymbol{z}_2 = \boldsymbol{z}_1 + \boldsymbol{w}$ is inside the curve of $\boldsymbol{z}^\top \boldsymbol{z} \leq C$. Mathematically, the $\boldsymbol{w}$ can be obtained by $-\nabla l(\boldsymbol{z}_1) - 2\lambda \boldsymbol{z}_1$ for some λ as shown in the middle picture of Figure 25. This is exactly the negative gradient of $L(\boldsymbol{z}) = l(\boldsymbol{z}) + \lambda||\boldsymbol{z}||^2$ such that

$$\nabla L(\boldsymbol{z}) = \nabla l(\boldsymbol{z}) + 2\lambda \boldsymbol{z},$$

and

$$\boldsymbol{w} = -\nabla L(\boldsymbol{z}) \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{z}_2 \quad = \boldsymbol{z}_1 + \boldsymbol{w} = \boldsymbol{z}_1 - \nabla L(\boldsymbol{z}).$$

And in practice, a small step size η can avoid going outside the curve of $\boldsymbol{z}^\top \boldsymbol{z} \leq C$:

$$\boldsymbol{z}_2 = \boldsymbol{z}_1 - \eta \nabla L(\boldsymbol{z}),$$

which is exactly what we have discussed in Section 16.2, the regularization term.

Sparsity In rare cases, we want to find sparse solution $\boldsymbol{z}$ such that $l(\boldsymbol{z})$ is minimized. Constrained in $||\boldsymbol{z}||_1 \leq C$ exists to this purpose where $|| \cdot ||_1$ is the l_1 norm of a vector or a matrix. Similar to the previous case, the l_1 constrained optimization pushes the gradient descent towards the border of the level of $||\boldsymbol{z}||_1 = C$. The situation in the 2-dimensional case is shown in Figure 26. In a high-dimensional case, many elements in $\boldsymbol{z}$ will be pushed into the breakpoint of $||\boldsymbol{z}||_1 = C$ as shown in the right picture of Figure 26.

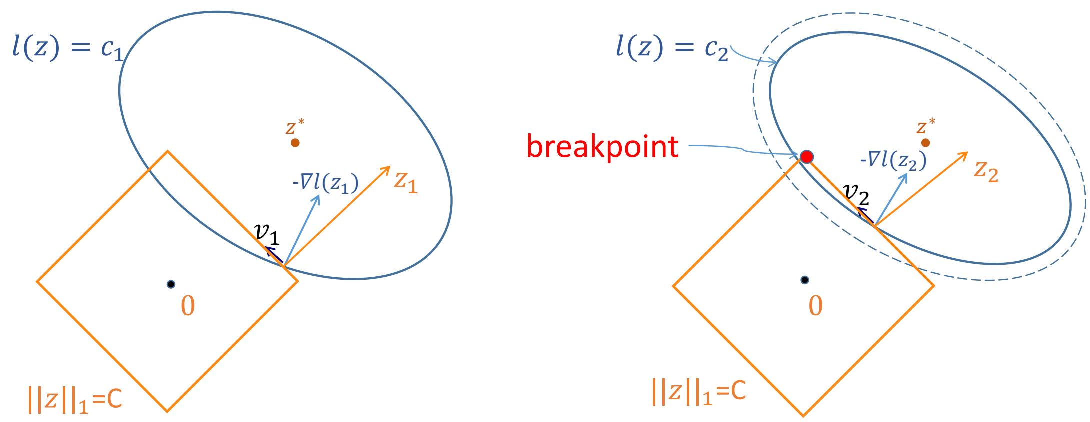


Figure 26: Constrained gradient descent with $||\boldsymbol{z}||_1 \leq C$, where the red dot denotes the breakpoint in l_1 norm. The right picture is the next step after the update in the left picture. $\boldsymbol{z}^\star$ denotes the optimal solution of $\{\min l(\boldsymbol{z})\}$.

16.7 Stochastic Gradient Descent

Now suppose we come back to the per-example loss:

$$L(\boldsymbol{W}, \boldsymbol{Z}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)^2 + \lambda_w ||\boldsymbol{w}_m||^2 + \lambda_z ||\boldsymbol{z}_n||^2.$$

And when we iteratively decrease the per-example loss term $l(\boldsymbol{w}_m, \boldsymbol{z}_n) = \left(a_{mn} - \boldsymbol{w}_m^\top \boldsymbol{z}_n\right)^2$ for all $m \in [0, M], n \in [1, N]$, the full loss $L(\boldsymbol{W}, \boldsymbol{Z})$ can also be decreased. This is known as the *stochastic coordinate descent*. The differentials with respect to $\boldsymbol{w}_m, \boldsymbol{z}_n$, and their roots are given by

$$\begin{cases} \nabla l(\boldsymbol{z}_n) = \dfrac{\partial l(\boldsymbol{w}_m, \boldsymbol{z}_n)}{\partial \boldsymbol{z}_n} = 2\boldsymbol{w}_m\boldsymbol{w}_m^\top \boldsymbol{z}_n + 2\lambda_w \boldsymbol{w}_m - 2a_{mn}\boldsymbol{w}_m \\ \qquad\qquad\qquad \underrightarrow{\text{leads to}}\ \boldsymbol{z}_n = a_{mn}(\boldsymbol{w}_m\boldsymbol{w}_m^\top + \lambda_z \boldsymbol{I})^{-1}\boldsymbol{w}_m; \\ \nabla l(\boldsymbol{w}_m) = \dfrac{\partial l(\boldsymbol{w}_m, \boldsymbol{z}_n)}{\partial \boldsymbol{w}_m} = 2\boldsymbol{z}_n\boldsymbol{z}_n^\top \boldsymbol{w}_m + 2\lambda_z \boldsymbol{z}_n - 2a_{mn}\boldsymbol{z}_n \\ \qquad\qquad\qquad \underrightarrow{\text{leads to}}\ \boldsymbol{w}_m = a_{mn}(\boldsymbol{z}_n\boldsymbol{z}_n^\top + \lambda_w \boldsymbol{I})^{-1}\boldsymbol{w}_n. \end{cases}$$

or analogously, the update can be done by gradient descent, and since we update by per-example loss, it is also known as the *stochastic gradient descent*

$$\begin{cases} \boldsymbol{z}_n = \boldsymbol{z}_n - \eta_z \dfrac{\nabla l(\boldsymbol{z}_n)}{||\nabla l(\boldsymbol{z}_n)||}; \\ \boldsymbol{w}_m = \boldsymbol{w}_m - \eta_w \dfrac{\nabla l(\boldsymbol{w}_m)}{||\nabla l(\boldsymbol{w}_m)||}. \end{cases}$$

In practice, the update for each m, n in the algorithm can be randomly produced, that's where the name *"stochastic"* comes from.

16.8 Bias Term

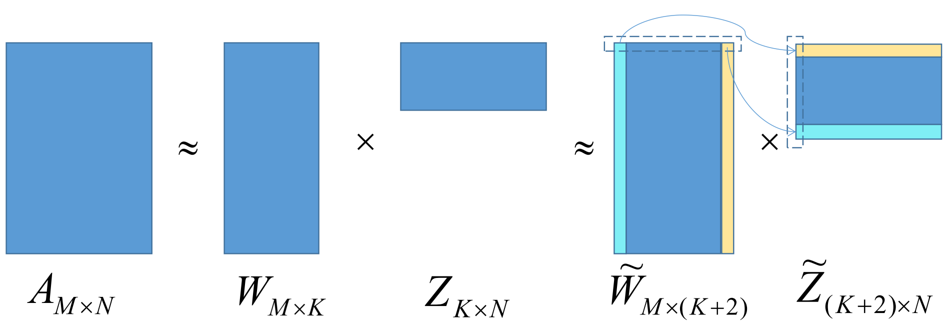


Figure 27: Bias terms in alternating least squares where the yellow entries denote ones (which are fixed) and cyan entries denote the added features to fit the bias terms. The dotted boxes give an example on how the bias terms work.

In ordinary least squares, a bias term is added to the raw matrix. A similar idea can be applied to the ALS problem. We can add a fixed column with all 1's to the last column of $\boldsymbol{W}$, thus an extra row should be added to the last row of $\boldsymbol{Z}$ to fit the features introduced by the bias term in $\boldsymbol{W}$. Analogously, a fixed row with all 1's can be added to the first row of $\boldsymbol{Z}$, and an extra column in the first column of $\boldsymbol{W}$ to fit the features. The situation is shown in Figure 27.

Following from the loss with respect to the columns of $\boldsymbol{Z}$ in Equation (16.14), suppose $\widetilde{\boldsymbol{z}}_n = \begin{bmatrix} 1 \\ \boldsymbol{z}_n \end{bmatrix}$ is the n-th column of $\widetilde{\boldsymbol{Z}}$, we have

$$
\begin{aligned}
L(\boldsymbol{z}_n) &= ||\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{Z}} - \boldsymbol{A}||^2 + \lambda_w||\widetilde{\boldsymbol{W}}||^2 + \lambda_z||\widetilde{\boldsymbol{Z}}||^2 \\
&= \left\|\widetilde{\boldsymbol{W}}\begin{bmatrix} 1 \\ \boldsymbol{z}_n \end{bmatrix} - \boldsymbol{a}_n\right\|^2 + \underbrace{\lambda_z||\widetilde{\boldsymbol{z}}_n||^2}_{=\lambda_z||\boldsymbol{z}_n||^2+\lambda_z} + \sum_{i\neq n}||\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{z}}_i - \boldsymbol{a}_i||^2 + \lambda_z\sum_{i\neq n}||\widetilde{\boldsymbol{z}}_i||^2 + \lambda_w||\widetilde{\boldsymbol{W}}||^2 \\
&= \left\|\begin{bmatrix} \overline{\boldsymbol{w}}_0 & \overline{\boldsymbol{W}} \end{bmatrix}\begin{bmatrix} 1 \\ \boldsymbol{z}_n \end{bmatrix} - \boldsymbol{a}_n\right\|^2 + \lambda_z||\boldsymbol{z}_n||^2 + C_{z_n} = \left\|\overline{\boldsymbol{W}}\boldsymbol{z}_n - \underbrace{(\boldsymbol{a}_n - \overline{\boldsymbol{w}}_0)}_{\overline{\boldsymbol{a}}_n}\right\|^2 + \lambda_z||\boldsymbol{z}_n||^2 + C_{z_n},
\end{aligned}
\tag{16.16}
$$

where $\overline{\boldsymbol{w}}_0$ is the first column of $\widetilde{\boldsymbol{W}}$ and C_{z_n} is a constant with respect to $\boldsymbol{z}_n$. Let $\overline{\boldsymbol{a}}_n = \boldsymbol{a}_n - \overline{\boldsymbol{w}}_0$, the update of $\boldsymbol{z}_n$ is just like the one in Equation (16.14) where the differential is given by:

$$
\frac{\partial L(\boldsymbol{z}_n)}{\partial \boldsymbol{z}_n} = 2\overline{\boldsymbol{W}}^\top\overline{\boldsymbol{W}}\boldsymbol{z}_n - 2\overline{\boldsymbol{W}}^\top\overline{\boldsymbol{a}}_n + 2\lambda_z\boldsymbol{z}_n.
$$

Therefore the update on $\boldsymbol{z}_n$ is given by the root of the above differential:

$$
\text{update on } \widetilde{\boldsymbol{z}}_n = \begin{cases} \boldsymbol{z}_n = (\overline{\boldsymbol{W}}^\top\overline{\boldsymbol{W}} + \lambda_z\boldsymbol{I})^{-1}\overline{\boldsymbol{W}}^\top\overline{\boldsymbol{a}}_n, & \text{for } n \in \{1,2,\ldots,N\}; \\ \widetilde{\boldsymbol{z}}_n = \begin{bmatrix} 1 \\ \boldsymbol{z}_n \end{bmatrix}. \end{cases}
$$

Similarly, follow from the loss with respect to each row of $\boldsymbol{W}$ in Equation (16.15), suppose $\widetilde{\boldsymbol{w}}_m = \begin{bmatrix} \boldsymbol{w}_m \\ 1 \end{bmatrix}$ is the m-th row of $\widetilde{\boldsymbol{W}}$ (or m-th column of $\widetilde{\boldsymbol{W}}^\top$), we have

$$
\begin{aligned}
L(\boldsymbol{w}_m) &= ||\widetilde{\boldsymbol{Z}}^\top\widetilde{\boldsymbol{W}} - \boldsymbol{A}^\top||^2 + \lambda_w||\widetilde{\boldsymbol{W}}^\top||^2 + \lambda_z||\widetilde{\boldsymbol{Z}}||^2 \\
&= ||\widetilde{\boldsymbol{Z}}^\top\widetilde{\boldsymbol{w}}_m - \boldsymbol{b}_m||^2 + \underbrace{\lambda_w||\widetilde{\boldsymbol{w}}_m||^2}_{=\lambda_w||\boldsymbol{w}_m||^2+\lambda_w} + \sum_{i\neq m}||\widetilde{\boldsymbol{Z}}^\top\widetilde{\boldsymbol{w}}_i - \boldsymbol{b}_i||^2 + \lambda_w\sum_{i\neq m}||\widetilde{\boldsymbol{w}}_i||^2 + \lambda_z||\widetilde{\boldsymbol{Z}}||^2 \\
&= \left\|\begin{bmatrix} \overline{\boldsymbol{Z}}^\top & \overline{\boldsymbol{z}}_0 \end{bmatrix}\begin{bmatrix} \boldsymbol{w}_m \\ 1 \end{bmatrix} - \boldsymbol{b}_m\right\|^2 + \lambda_w||\boldsymbol{w}_m||^2 + C_{w_m} \\
&= \left\|\overline{\boldsymbol{Z}}^\top\boldsymbol{w}_m - (\boldsymbol{b}_m - \overline{\boldsymbol{z}}_0)\right\|^2 + \lambda_w||\boldsymbol{w}_m||^2 + C_{w_m},
\end{aligned}
\tag{16.17}
$$

where $\overline{\boldsymbol{z}}_0$ is the last column of $\widetilde{\boldsymbol{Z}}^\top$ and $\overline{\boldsymbol{Z}}^\top$ is the left columns of it: $\widetilde{\boldsymbol{Z}}^\top = \begin{bmatrix} \boldsymbol{w}_m \\ 1 \end{bmatrix}$, C_{w_m} is a constant with respect to $\boldsymbol{w}_m$, and $\boldsymbol{W}^\top = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M]$, $\boldsymbol{A}^\top = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_M]$ are the column partitions of $\boldsymbol{W}^\top, \boldsymbol{A}^\top$ respectively. Let $\overline{\boldsymbol{b}}_m = \boldsymbol{b}_m - \overline{\boldsymbol{z}}_0$, the update of $\boldsymbol{w}_m$ is again just like the on in Equation (16.15) where the differential is given by:

$$
\frac{\partial L(\boldsymbol{w}_m d)}{\partial \boldsymbol{w}_m} = 2\overline{\boldsymbol{Z}}\cdot\overline{\boldsymbol{Z}}^\top\boldsymbol{w}_m - 2\overline{\boldsymbol{Z}}\cdot\overline{\boldsymbol{b}}_m + 2\lambda_w\boldsymbol{w}_m.
$$

Therefore the update on $\boldsymbol{w}_m$ is given by the root of the above differential

$$
\text{update on } \widetilde{\boldsymbol{w}}_m = \begin{cases} \boldsymbol{w}_m = (\overline{\boldsymbol{Z}} \cdot \overline{\boldsymbol{Z}}^\top + \lambda_w \boldsymbol{I})^{-1} \overline{\boldsymbol{Z}} \cdot \overline{\boldsymbol{b}}_m, & \text{for } m \in \{1, 2, \ldots, M\}; \\ \widetilde{\boldsymbol{w}}_m = \begin{bmatrix} \boldsymbol{w}_m \\ 1 \end{bmatrix}. \end{cases}
$$

Similar updates by gradient descent under the bias terms or treatment on missing entries can be deduced and we shall not repeat the details (see Section 16.5 and 16.3 for a reference).

17. Nonnegative Matrix Factorization (NMF)

Following from the matrix factorization via the ALS, we now consider algorithms for solving the nonnegative matrix factorization (NMF) problem:

- Given a nonnegative matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, find nonnegative matrix factors $\boldsymbol{W} \in \mathbb{R}^{M \times K}$ and $\boldsymbol{Z} \in \mathbb{R}^{K \times N}$ such that:
$$
\boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{Z}.
$$

To measure the approximation, the loss to evaluate is still from the Frobenius norm of the difference between the two matrices:

$$
L(\boldsymbol{W}, \boldsymbol{Z}) = ||\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}||^2.
$$

17.1 NMF via Multiplicative Update

Following from Section 16.1, given $\boldsymbol{W} \in \mathbb{R}^{M \times K}$, we want to update $\boldsymbol{Z} \in \mathbb{R}^{K \times N}$, the gradient with respect to $\boldsymbol{Z}$ is given by Equation (16.2):

$$
\frac{\partial L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}} = 2\boldsymbol{W}^\top(\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}) \in \mathbb{R}^{K \times N}.
$$

Applying the gradient descent idea in Section 16.5, the trivial update on $\boldsymbol{Z}$ can be done by

$$
(\text{GD on } \boldsymbol{Z}) \qquad \boldsymbol{Z} \leftarrow \boldsymbol{Z} - \eta \left(\frac{\partial L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}} \right) = \boldsymbol{Z} - \eta \left(2\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Z} - 2\boldsymbol{W}^\top \boldsymbol{A} \right),
$$

where η is a small positive step size. Now if we suppose a different step size for each entry of $\boldsymbol{Z}$ and incorporate the constant 2 into the step size, the update can be obtained by

$$
(\text{GD}' \text{ on } \boldsymbol{Z}) \qquad \begin{aligned} \boldsymbol{Z}_{kn} &\leftarrow \boldsymbol{Z}_{kn} - \frac{\eta_{kn}}{2} \left(\frac{\partial L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}} \right)_{kn} \\ &= \boldsymbol{Z}_{kn} - \eta_{kn} (\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Z} - \boldsymbol{W}^\top \boldsymbol{A})_{kn}, \qquad k \in [1, K], n \in [1, N], \end{aligned}
$$

where $\boldsymbol{Z}_{kn}$ is the (k, n)-th entry of $\boldsymbol{Z}$. Now if we rescale the step size:

$$
\eta_{kn} = \frac{\boldsymbol{Z}_{kn}}{(\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Z})_{kn}},
$$

then we obtain the update rule:

$$
(\text{Multiplicative update on } \boldsymbol{Z}) \qquad \boldsymbol{Z}_{kn} \leftarrow \boldsymbol{Z}_{kn} \frac{(\boldsymbol{W}^\top \boldsymbol{A})_{kn}}{(\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Z})_{kn}}, \qquad k \in [1, K], n \in [1, N],
$$

which is known as the *multiplicative update* and is first developed in (Lee and Seung, 2001) and further discussed in (Pauca et al., 2006). Analogously, the multiplicative update on $\boldsymbol{W}$ can be obtained by

$$\text{(Multiplicative update on } \boldsymbol{W}) \qquad \boldsymbol{W}_{mk} \leftarrow \boldsymbol{W}_{mk} \frac{(\boldsymbol{A}\boldsymbol{Z}^\top)_{mk}}{(\boldsymbol{W}\boldsymbol{Z}\boldsymbol{Z}^\top)_{mk}}, \qquad m \in [1, M], k \in [1, K].$$

Theorem 17.1: (Convergence of Multiplicative Update)

The loss $L(\boldsymbol{W}, \boldsymbol{Z}) = ||\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}||^2$ is non-increasing under the multiplicative update rules:

$$\begin{cases} \boldsymbol{Z}_{kn} \leftarrow \boldsymbol{Z}_{kn} \dfrac{(\boldsymbol{W}^\top \boldsymbol{A})_{kn}}{(\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Z})_{kn}}, & k \in [1, K], n \in [1, N]; \\ \boldsymbol{W}_{mk} \leftarrow \boldsymbol{W}_{mk} \dfrac{(\boldsymbol{A}\boldsymbol{Z}^\top)_{mk}}{(\boldsymbol{W}\boldsymbol{Z}\boldsymbol{Z}^\top)_{mk}}, & m \in [1, M], k \in [1, K]. \end{cases}$$

We refer the proof of the above theorem to (Lee and Seung, 2001). Clearly the approximations $\boldsymbol{W}$ and $\boldsymbol{Z}$ remain nonnegative during the updates. It is generally best to update $\boldsymbol{W}$ and $\boldsymbol{Z}$ "simultaneously", instead of updating each matrix fully before the other. In this case, after updating a row of $\boldsymbol{Z}$, we update the corresponding column of $\boldsymbol{W}$. In the implementation, a small positive quantity, say the square root of the machine precision, should be added to the denominators in the approximations of $\boldsymbol{W}$ and $\boldsymbol{Z}$ at each iteration step. And a trivial $\epsilon = 10^{-9}$ can do the job. The full procedure is shown in Algorithm 5.

Algorithm 5 NMF via Multiplicative Updates

```
Require: A ∈ ℝ^{M×N};
 initialize W ∈ ℝ^{M×K}, Z ∈ ℝ^{K×N} randomly with nonnegative entries.
 choose a stop criterion on the approximation error δ;
 choose maximal number of iterations C;
 iter = 0;
 while ||A − (WZ)||^2 > δ and iter < C do
     iter = iter + 1;
     for k = 1 to K do
         for n = 1 to N do                          ▷ udate k-th row of Z
             Z_{kn} ← Z_{kn} (W^⊤A)_{kn} / ((W^⊤WZ)_{kn} + ϵ);
         end for
         for m = 1 to M do                          ▷ udate k-th column of W
             W_{mk} ← W_{mk} (AZ^⊤)_{mk} / ((WZZ^⊤)_{mk} + ϵ);
         end for
     end for
 end while
 Output W, Z;
```

17.2 Regularization

Similar to the ALS with regularization in Section 16.2, recall the regularization helps employ the ALS into general matrices. We can also add a regularization in the context of NMF:

$$L(\boldsymbol{W}, \boldsymbol{Z}) = ||\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}||^2 + \lambda_w ||\boldsymbol{W}||^2 + \lambda_z ||\boldsymbol{Z}||^2, \qquad \lambda_w > 0, \lambda_z > 0,$$

where the induced matrix norm is still the Frobenius norm. The gradient with respective to $\boldsymbol{Z}$ given $\boldsymbol{W}$ is the same as that in Equation (16.8):

$$\frac{\partial L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}} = 2\boldsymbol{W}^\top(\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{A}) + 2\lambda_z \boldsymbol{Z} \in \mathbb{R}^{K\times N}.$$

The trivial gradient descent update can be obtained by

$$\text{(GD on } \boldsymbol{Z}) \qquad \boldsymbol{Z} \leftarrow \boldsymbol{Z} - \eta \left(\frac{\partial L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}} \right) = \boldsymbol{Z} - \eta \left(2\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Z} - 2\boldsymbol{W}^\top \boldsymbol{A} + 2\lambda_z \boldsymbol{Z} \right),$$

Analogously, if we suppose a different step size for each entry of $\boldsymbol{Z}$ and incorporate the constant 2 into the step size, the update can be obtained by

$$\text{(GD}' \text{ on } \boldsymbol{Z}) \qquad \begin{aligned} \boldsymbol{Z}_{kn} &\leftarrow \boldsymbol{Z}_{kn} - \frac{\eta_{kn}}{2} \left(\frac{\partial L(\boldsymbol{Z}|\boldsymbol{W})}{\partial \boldsymbol{Z}} \right)_{kn} \\ &= \boldsymbol{Z}_{kn} - \eta_{kn} (\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Z} - \boldsymbol{W}^\top \boldsymbol{A} + \lambda_z \boldsymbol{Z})_{kn}, \qquad k \in [1, K], n \in [1, N], \end{aligned}$$

Now if we rescale the step size:

$$\eta_{kn} = \frac{\boldsymbol{Z}_{kn}}{(\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Z})_{kn}},$$

then we obtain the update rule:

$$\text{(Multiplicative update on } \boldsymbol{Z}) \qquad \boldsymbol{Z}_{kn} \leftarrow \boldsymbol{Z}_{kn} \frac{(\boldsymbol{W}^\top \boldsymbol{A})_{kn} - \lambda_z \boldsymbol{Z}_{kn}}{(\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{Z})_{kn}}, \qquad k \in [1, K], n \in [1, N].$$

Similarly, the multiplicative update on $\boldsymbol{W}$ can be obtained by

$$\text{(Multiplicative update on } \boldsymbol{W}) \qquad \boldsymbol{W}_{mk} \leftarrow \boldsymbol{W}_{mk} \frac{(\boldsymbol{A}\boldsymbol{Z}^\top)_{mk} - \lambda_w \boldsymbol{W}_{mk}}{(\boldsymbol{W}\boldsymbol{Z}\boldsymbol{Z}^\top)_{mk}}, \qquad m \in [1, M], k \in [1, K].$$

17.3 Initialization

In the above discussion, we initialize $\boldsymbol{W}$ and $\boldsymbol{Z}$ randomly. Whereas, there are also alternative strategies designed to obtain better initial estimates in the hope of converging more rapidly to a good solution (Boutsidis and Gallopoulos, 2008; Gillis, 2014). We sketch the methods as follows:

- *Clustering techniques.* Use some clustering methods on the columns of $\boldsymbol{A}$, and make the cluster means of the top K clusters as the columns of $\boldsymbol{W}$, and initialize $\boldsymbol{Z}$ as a proper scaling of the cluster indicator matrix (that is, $\boldsymbol{Z}_{kn} \neq 0$ indicates $\boldsymbol{a}_n$ belongs to the k-th cluster);

- *Subset selection.* Pick K columns of $\boldsymbol{A}$ and set those as the initial columns for $\boldsymbol{W}$, and analogously, K rows of $\boldsymbol{A}$ are selected to form the rows of $\boldsymbol{Z}$;
- *SVD-based.* Suppose the SVD of $\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ where each factor $\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ is a rank-one matrix with possible negative values in $\boldsymbol{u}_i, \boldsymbol{v}_i$, and nonnegative σ_i. Denote $[x]_+ = \max(x, 0)$, we notice

$$
\boldsymbol{u}_i \boldsymbol{v}_i^\top = [\boldsymbol{u}_i]_+ [\boldsymbol{v}_i]_+^\top + [-\boldsymbol{u}_i]_+ [-\boldsymbol{v}_i]_+^\top - [-\boldsymbol{u}_i]_+ [\boldsymbol{v}_i]_+^\top - [\boldsymbol{u}_i]_+ [-\boldsymbol{v}_i]_+^\top .
$$

  Either $[\boldsymbol{u}_i]_+ [\boldsymbol{v}_i]_+^\top$ or $[-\boldsymbol{u}_i]_+ [-\boldsymbol{v}_i]_+^\top$ can be selected as a column and a row in $\boldsymbol{W}, \boldsymbol{Z}$.

18. Biconjugate Decomposition

18.1 Existence of the Biconjugate Decomposition

The biconjugate decomposition was proposed in (Chu et al., 1995) and discussed in (Yang, 2000). The existence of the biconjugate decomposition relies on the rank-one reduction theorem shown below. And a variety of matrix decomposition methods can be unified via this biconjugate decomposition.

Theorem 18.1: (Rank-One Reduction)

Any $m \times n$ matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with rank r, a pair of vectors $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$ such that $w = \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x} \neq 0$, then the matrix $\boldsymbol{B} = \boldsymbol{A} - w^{-1}\boldsymbol{x}\boldsymbol{y}^\top \boldsymbol{A}$ has rank $r-1$ which has exactly one less than the rank of $\boldsymbol{A}$.

Proof [of Theorem 18.1] If we can show that the dimension of $\mathcal{N}(\boldsymbol{B})$ is one larger than that of $\boldsymbol{A}$. Then this implicitly shows $\boldsymbol{B}$ has rank exactly one less than the rank of $\boldsymbol{A}$.

For any vector $\boldsymbol{n} \in \mathcal{N}(\boldsymbol{A})$, i.e., $\boldsymbol{A}\boldsymbol{n} = \boldsymbol{0}$, we then have $\boldsymbol{B}\boldsymbol{n} = \boldsymbol{A}\boldsymbol{n} - w^{-1}\boldsymbol{x}\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{n} = \boldsymbol{0}$ which means $\mathcal{N}(\boldsymbol{A}) \subseteq \mathcal{N}(\boldsymbol{B})$.

Now for any vector $\boldsymbol{m} \in \mathcal{N}(\boldsymbol{B})$, then $\boldsymbol{B}\boldsymbol{m} = \boldsymbol{A}\boldsymbol{m} - w^{-1}\boldsymbol{x}\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{m} = \boldsymbol{0}$.

Let $k = w^{-1}\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{m}$, which is a scalar, thus $\boldsymbol{A}(\boldsymbol{m} - k\boldsymbol{x}) = \boldsymbol{0}$, i.e., for any vector $\boldsymbol{n} \in \mathcal{N}(\boldsymbol{A})$, we could find a vector $\boldsymbol{m} \in \mathcal{N}(\boldsymbol{B})$ such that $\boldsymbol{n} = (\boldsymbol{m} - k\boldsymbol{x}) \in \mathcal{N}(\boldsymbol{A})$. Note that $\boldsymbol{A}\boldsymbol{x} \neq \boldsymbol{0}$ from the definition of w. Thus, the null space of $\boldsymbol{B}$ is therefore obtained from the null space of $\boldsymbol{A}$ by adding $\boldsymbol{x}$ to its basis which will increase the order of the space by 1. Thus the dimension of $\mathcal{N}(\boldsymbol{A})$ is smaller than the dimension of $\mathcal{N}(\boldsymbol{B})$ by 1 which completes the proof. ■

Suppose matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has rank r, we can define a rank reducing process to generate a sequence of Wedderburn matrices $\{\boldsymbol{A}_k\}$:

$$
\boldsymbol{A}_1 = \boldsymbol{A}, \qquad \text{and} \qquad \boldsymbol{A}_{k+1} = \boldsymbol{A}_k - w_k^{-1}\boldsymbol{A}_k \boldsymbol{x}_k \boldsymbol{y}_k^\top \boldsymbol{A}_k,
$$

where $\boldsymbol{x}_k \in \mathbb{R}^n$ and $\boldsymbol{y}_k \in \mathbb{R}^m$ are any vectors satisfying $w_k = \boldsymbol{y}_k^\top \boldsymbol{A}\boldsymbol{x}_k \neq 0$. The sequence will terminate in r steps since the rank of $\boldsymbol{A}_k$ decreases by exactly one at each step. Write out the sequence:

$$\begin{aligned}
\boldsymbol{A}_1 &= \boldsymbol{A}, \\
\boldsymbol{A}_1 - \boldsymbol{A}_2 &= w_1^{-1}\boldsymbol{A}_1\boldsymbol{x}_1\boldsymbol{y}_1^\top\boldsymbol{A}_1, \\
\boldsymbol{A}_2 - \boldsymbol{A}_3 &= w_2^{-1}\boldsymbol{A}_2\boldsymbol{x}_2\boldsymbol{y}_2^\top\boldsymbol{A}_2, \\
\boldsymbol{A}_3 - \boldsymbol{A}_4 &= w_3^{-1}\boldsymbol{A}_3\boldsymbol{x}_3\boldsymbol{y}_3^\top\boldsymbol{A}_3, \\
\vdots &= \vdots \\
\boldsymbol{A}_{r-1} - \boldsymbol{A}_r &= w_{r-1}^{-1}\boldsymbol{A}_{r-1}\boldsymbol{x}_{r-1}\boldsymbol{y}_{r-1}^\top\boldsymbol{A}_{r-1}, \\
\boldsymbol{A}_r - \boldsymbol{0} &= w_r^{-1}\boldsymbol{A}_r\boldsymbol{x}_r\boldsymbol{y}_r^\top\boldsymbol{A}_r.
\end{aligned}$$

By adding the sequence we will get

$$(\boldsymbol{A}_1 - \boldsymbol{A}_2) + (\boldsymbol{A}_2 - \boldsymbol{A}_3) + \ldots + (\boldsymbol{A}_{r-1} - \boldsymbol{A}_r) + (\boldsymbol{A}_r - \boldsymbol{0}) = \boldsymbol{A} = \sum_{i=1}^{r} w_i^{-1}\boldsymbol{A}_i\boldsymbol{x}_i\boldsymbol{y}_i^\top\boldsymbol{A}_i.$$

Theorem 18.2: (Biconjugate Decomposition: Form 1)

This equality from rank-reducing process implies the following matrix decomposition

$$\boldsymbol{A} = \boldsymbol{\Phi}\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}^\top,$$

where $\boldsymbol{\Omega} = diag(w_1, w_2, \ldots, w_r)$, $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_r] \in \mathbb{R}^{m\times r}$ and $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_r]$ with

$$\boldsymbol{\phi}_k = \boldsymbol{A}_k\boldsymbol{x}_k, \qquad \text{and} \qquad \boldsymbol{\psi}_k = \boldsymbol{A}_k^\top\boldsymbol{y}_k.$$

Obviously, different choices of $\boldsymbol{x}_k$'s and $\boldsymbol{y}_k$'s will result in different factorizations. So this factorization is rather general and we will show its connection to some well-known decomposition methods.

Remark 18.3

For the vectors $\boldsymbol{x}_k, \boldsymbol{y}_k$ in the Wedderburn sequence, we have the following property

$$\begin{aligned}
\boldsymbol{x}_k &\in \mathcal{N}(\boldsymbol{A}_{k+1}) \perp \mathcal{C}(\boldsymbol{A}_{k+1}^\top), \\
\boldsymbol{y}_k &\in \mathcal{N}(\boldsymbol{A}_{k+1}^\top) \perp \mathcal{C}(\boldsymbol{A}_{k+1}).
\end{aligned}$$

Lemma 18.4: (General Term Formula of Wedderburn Sequence: V1)

For each matrix with $\boldsymbol{A}_{k+1} = \boldsymbol{A}_k - w_k^{-1}\boldsymbol{A}_k\boldsymbol{x}_k\boldsymbol{y}_k^\top\boldsymbol{A}_k$, then $\boldsymbol{A}_{k+1}$ can be written as

$$\boldsymbol{A}_{k+1} = \boldsymbol{A} - \sum_{i=1}^{k} w_i^{-1}\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top\boldsymbol{A},$$

where

$$\boldsymbol{u}_k = \boldsymbol{x}_k - \sum_{i=1}^{k-1} \frac{\boldsymbol{v}_i^\top \boldsymbol{A}\boldsymbol{x}_k}{w_i} \boldsymbol{u}_i, \qquad \text{and} \qquad \boldsymbol{v}_k = \boldsymbol{y}_k - \sum_{i=1}^{k-1} \frac{\boldsymbol{y}_k^\top \boldsymbol{A}\boldsymbol{u}_i}{w_i} \boldsymbol{v}_i.$$

The proof of this lemma is delayed in Section 18.4. We notice that $w_i = \boldsymbol{y}_k^\top \boldsymbol{A}_k \boldsymbol{x}_k$ in the general term formula is related to $\boldsymbol{A}_k$. So it's not the true general term formula. We will write w_i to be related to $\boldsymbol{A}$ rather than $\boldsymbol{A}_k$ later. From the general term formula of Wedderburn sequence, we have

$$\begin{aligned} \boldsymbol{A}_{k+1} &= \boldsymbol{A} - \sum_{i=1}^{k} w_i^{-1} \boldsymbol{A}\boldsymbol{u}_i \boldsymbol{v}_i^\top \boldsymbol{A} \\ \boldsymbol{A}_{k} &= \boldsymbol{A} - \sum_{i=1}^{k-1} w_i^{-1} \boldsymbol{A}\boldsymbol{u}_i \boldsymbol{v}_i^\top \boldsymbol{A}. \end{aligned}$$

Thus, $\boldsymbol{A}_{k+1} - \boldsymbol{A}_k = -w_k^{-1}\boldsymbol{A}\boldsymbol{u}_k\boldsymbol{v}_k^\top \boldsymbol{A}$. Since we define the sequence by $\boldsymbol{A}_{k+1} = \boldsymbol{A}_k - w_k^{-1}\boldsymbol{A}_k\boldsymbol{x}_k\boldsymbol{y}_k^\top\boldsymbol{A}_k$. We then find $w_k^{-1}\boldsymbol{A}\boldsymbol{u}_k\boldsymbol{v}_k^\top\boldsymbol{A} = w_k^{-1}\boldsymbol{A}_k\boldsymbol{x}_k\boldsymbol{y}_k^\top\boldsymbol{A}_k$. It is trivial to see

$$\begin{aligned} \boldsymbol{A}\boldsymbol{u}_k &= \boldsymbol{A}_k\boldsymbol{x}_k, \\ \boldsymbol{v}_k^\top \boldsymbol{A} &= \boldsymbol{y}_k^\top \boldsymbol{A}_k. \end{aligned} \tag{18.1}$$

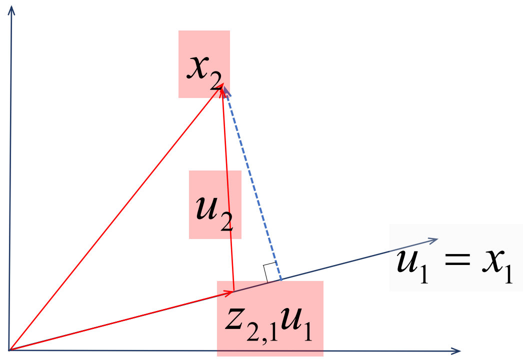
Let $z_{k,i} = \frac{\boldsymbol{v}_i^\top \boldsymbol{A}\boldsymbol{x}_k}{w_i}$ which is a scalar. From the definition of $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ in the above lemma, then

- $\boldsymbol{u}_1 = \boldsymbol{x}_1$;
- $\boldsymbol{u}_2 = \boldsymbol{x}_2 - z_{2,1}\boldsymbol{u}_1$;
- $\boldsymbol{u}_3 = \boldsymbol{x}_3 - z_{3,1}\boldsymbol{u}_1 - z_{3,2}\boldsymbol{u}_2$;
- ….

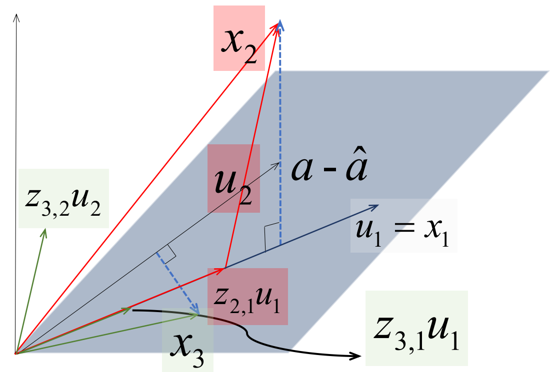
This process is just similar to the Gram-Schmidt process. But now, we do not project $\boldsymbol{x}_2$ onto $\boldsymbol{x}_1$ with the smallest distance. The vector of $\boldsymbol{x}_2$ along $\boldsymbol{x}_1$ is now defined by $z_{2,1}$. This process is shown in Figure 28. In Figure 28(a), $\boldsymbol{u}_2$ is not perpendicular to $\boldsymbol{u}_1$. But $\boldsymbol{u}_2$ does not lie on the same line of $\boldsymbol{u}_1$ so that $\boldsymbol{u}_1, \boldsymbol{u}_2$ still could span a $\mathbb{R}^2$ subspace. Similarly, in Figure 28(b), $\boldsymbol{u}_3 = \boldsymbol{x}_3 - z_{3,1}\boldsymbol{u}_1 - z_{3,2}\boldsymbol{u}_2$ does not lie in the space spanned by $\boldsymbol{u}_1, \boldsymbol{u}_2$ so that $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3$ could still span a $\mathbb{R}^3$ subspace.

A moment of reflexion would reveal that the span of $\boldsymbol{x}_2, \boldsymbol{x}_1$ is the same as the span of $\boldsymbol{u}_2, \boldsymbol{u}_1$. Similarly for $\boldsymbol{v}_i$'s. We have the following property:

$$\begin{cases} \text{span}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_j\} = \text{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_j\}; \\ \text{span}\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_j\} = \text{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_j\}. \end{cases} \tag{18.2}$$



(a) "Project" onto a line



(b) "Project" onto a space

Figure 28: "Project" a vector onto a line and onto a space.

Further, from the rank-reducing property in the Wedderburn sequence, we have

$$\begin{cases} \mathcal{C}(\boldsymbol{A}_1) \supset \mathcal{C}(\boldsymbol{A}_2) \supset \mathcal{C}(\boldsymbol{A}_3) \supset \ldots; \\ \mathcal{N}(\boldsymbol{A}_1^\top) \subset \mathcal{N}(\boldsymbol{A}_2^\top) \subset \mathcal{N}(\boldsymbol{A}_3^\top) \subset \ldots. \end{cases}$$

Since $\boldsymbol{y}_k \in \mathcal{N}(\boldsymbol{A}_{k+1}^\top)$, it then follows that $\boldsymbol{y}_j \in \mathcal{N}(\boldsymbol{A}_{k+1}^\top)$ for all $j < k+1$, i.e., $\boldsymbol{A}_{k+1}^\top \boldsymbol{y}_j = \boldsymbol{0}$ for all $j < k+1$. Which also holds true for $\boldsymbol{x}_{k+1}^\top \boldsymbol{A}_{k+1}^\top \boldsymbol{y}_j = 0$ for all $j < k+1$. From Equation (18.1), we also have $\boldsymbol{u}_{k+1}^\top \boldsymbol{A}^\top \boldsymbol{y}_j = 0$ for all $j < k+1$. Following from Equation (18.2), we obtain

$$\boldsymbol{v}_j^\top \boldsymbol{A}\boldsymbol{u}_{k+1} = 0 \text{ for all } j < k+1.$$

Similarly, we can prove

$$\boldsymbol{v}_{k+1}^\top \boldsymbol{A}\boldsymbol{u}_j = 0 \text{ for all } j < k+1.$$

Moreover, we defined $w_k = \boldsymbol{y}_k^\top \boldsymbol{A}_k \boldsymbol{x}_k$. By Equation (18.1), we can write the w_k as:

$$\begin{aligned} w_k &= \boldsymbol{y}_k^\top \boldsymbol{A}_k \boldsymbol{x}_k \\ &= \boldsymbol{v}_k^\top \boldsymbol{A}\boldsymbol{x}_k \\ &= \boldsymbol{v}_k^\top \boldsymbol{A}(\boldsymbol{u}_k + \sum_{i=1}^{k-1} \frac{\boldsymbol{v}_i^\top \boldsymbol{A}\boldsymbol{x}_k}{w_i}\boldsymbol{u}_i) && \text{(by the definition of } \boldsymbol{u}_k \text{ in Lemma 18.4)} \\ &= \boldsymbol{v}_k^\top \boldsymbol{A}\boldsymbol{u}_k, && \text{(by } \boldsymbol{v}_k^\top \boldsymbol{A}\boldsymbol{u}_j = 0 \text{ for all } j < k) \end{aligned}$$

which can be used to substitute the w_k in Lemma 18.4 and we then have the full version of the general term formula of the Wedderburn sequence such that the formula does not depend on $\boldsymbol{A}_k$'s (in the form of w_k's) with

$$\boldsymbol{u}_k = \boldsymbol{x}_k - \sum_{i=1}^{k-1} \frac{\boldsymbol{v}_i^\top \boldsymbol{A}\boldsymbol{x}_k}{\boldsymbol{v}_i^\top \boldsymbol{A}\boldsymbol{u}_i}\boldsymbol{u}_i, \qquad \text{and} \qquad \boldsymbol{v}_k = \boldsymbol{y}_k - \sum_{i=1}^{k-1} \frac{\boldsymbol{y}_k^\top \boldsymbol{A}\boldsymbol{u}_i}{\boldsymbol{v}_i^\top \boldsymbol{A}\boldsymbol{u}_i}\boldsymbol{v}_i. \tag{18.3}$$

Gram-Schmidt Process from Wedderburn Sequence: If $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r] \in \mathbb{R}^{n\times r}$, $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_r] \in \mathbb{R}^{m\times r}$ effects a rank-reducing process for $\boldsymbol{A}$. Let $\boldsymbol{A}$ be the identity matrix and $(\boldsymbol{X}, \boldsymbol{Y})$ are identical and contain the vectors for which an orthogonal basis is desired, then $(\boldsymbol{U} = \boldsymbol{V})$ give the resultant orthogonal basis.

This form of $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ in Equation (18.3) is very close to the projection to the perpendicular space of the Gram-Schmidt process in Equation (3.1). We then define $\boxed{< \boldsymbol{x}, \boldsymbol{y} >:= \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x}}$ to explicitly mimic the form of projection in Equation (3.1). We formulate the results so far in the following lemma which can help us have a clear vision about what we have been working on and we will use these results extensively in the sequel:

Lemma 18.5: (Properties of Wedderburn Sequence)

For each matrix with $\boldsymbol{A}_{k+1} = \boldsymbol{A}_k - w_k^{-1}\boldsymbol{A}_k\boldsymbol{x}_k\boldsymbol{y}_k^\top\boldsymbol{A}_k$, then $\boldsymbol{A}_{k+1}$ can be written as

$$\boldsymbol{A}_{k+1} = \boldsymbol{A} - \sum_{i=1}^{k} w_i^{-1}\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top\boldsymbol{A},$$

where

$$\boldsymbol{u}_k = \boldsymbol{x}_k - \sum_{i=1}^{k-1} \frac{< \boldsymbol{x}_k, \boldsymbol{v}_i >}{< \boldsymbol{u}_i, \boldsymbol{v}_i >}\boldsymbol{u}_i, \qquad \text{and} \qquad \boldsymbol{v}_k = \boldsymbol{y}_k - \sum_{i=1}^{k-1} \frac{< \boldsymbol{u}_i, \boldsymbol{y}_k >}{< \boldsymbol{u}_i, \boldsymbol{v}_i >}\boldsymbol{v}_i. \tag{18.4}$$

Further, we have the following properties:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{u}_k &= \boldsymbol{A}_k\boldsymbol{x}_k, \\ \boldsymbol{v}_k^\top\boldsymbol{A} &= \boldsymbol{y}_k^\top\boldsymbol{A}_k. \end{aligned} \tag{18.5}$$

$$< \boldsymbol{u}_k, \boldsymbol{v}_j > = < \boldsymbol{u}_j, \boldsymbol{v}_k > = 0 \text{ for all } j < k. \tag{18.6}$$

$$w_k = \boldsymbol{y}_k^\top\boldsymbol{A}_k\boldsymbol{x}_k = < \boldsymbol{u}_k, \boldsymbol{v}_k > \tag{18.7}$$

By substituting Equation (18.5) into Form 1 of biconjugate decomposition, and using Equation (18.7) which implies $w_k = \boldsymbol{v}_k^\top\boldsymbol{A}\boldsymbol{u}_k$, we have the Form 2 and Form 3 of this decomposition:

Theorem 18.6: (Biconjugate Decomposition: Form 2 and Form 3)

The equality from rank-reducing process implies the following matrix decomposition

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{U}_r\boldsymbol{\Omega}_r^{-1}\boldsymbol{V}_r^\top\boldsymbol{A},$$

where $\boldsymbol{\Omega}_r = diag(w_1, w_2, \ldots, w_r)$, $\boldsymbol{U}_r = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r] \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V}_r = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r] \in \mathbb{R}^{n\times r}$ with

$$\boldsymbol{u}_k = \boldsymbol{x}_k - \sum_{i=1}^{k-1} \frac{< \boldsymbol{x}_k, \boldsymbol{v}_i >}{< \boldsymbol{u}_i, \boldsymbol{v}_i >}\boldsymbol{u}_i, \qquad \text{and} \qquad \boldsymbol{v}_k = \boldsymbol{y}_k - \sum_{i=1}^{k-1} \frac{< \boldsymbol{u}_i, \boldsymbol{y}_k >}{< \boldsymbol{u}_i, \boldsymbol{v}_i >}\boldsymbol{v}_i. \tag{18.8}$$

And also the following decomposition

$$\boldsymbol{V}_\gamma^\top \boldsymbol{A}\boldsymbol{U}_\gamma = \boldsymbol{\Omega}_\gamma, \tag{18.9}$$

where $\boldsymbol{\Omega}_\gamma = diag(w_1, w_2, \ldots, w_\gamma)$, $\boldsymbol{U}_\gamma = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\gamma] \in \mathbb{R}^{m\times\gamma}$ and $\boldsymbol{V}_\gamma = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_\gamma] \in \mathbb{R}^{n\times\gamma}$. Note the difference between the subscripts r and γ we used here with $\gamma \leq r$.

We notice that, in these two forms of biconjugate decomposition, they are independent of the Wedderburn matrices $\{\boldsymbol{A}_k\}$.

A word on the notation: we will use the subscript to indicate the dimension of the matrix avoiding confusion in the sequel, e.g., the r, γ in the above theorem.

18.2 Properties of the Biconjugate Decomposition

Corollary 18.7: (Connection of $\boldsymbol{U}_\gamma$ and $\boldsymbol{X}_\gamma$)

If $(\boldsymbol{X}_\gamma, \boldsymbol{Y}_\gamma) \in \mathbb{R}^{n\times\gamma} \times \mathbb{R}^{m\times\gamma}$ effects a rank-reducing process for $\boldsymbol{A}$, then there are unique unit upper triangular matrices $\boldsymbol{R}_\gamma^{(x)} \in \mathbb{R}^{\gamma\times\gamma}$ and $\boldsymbol{R}_\gamma^{(y)} \in \mathbb{R}^{\gamma\times\gamma}$ such that

$$\boldsymbol{X}_\gamma = \boldsymbol{U}_\gamma \boldsymbol{R}_\gamma^{(x)}, \qquad \text{and} \qquad \boldsymbol{Y}_\gamma = \boldsymbol{V}_\gamma \boldsymbol{R}_\gamma^{(y)},$$

where $\boldsymbol{U}_\gamma$ and $\boldsymbol{V}_\gamma$ are matrices with columns resulting from the Wedderburn sequence as in Equation (18.9).

Proof [of Corollary 18.7] The proof is trivial from the definition of $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ in Equation (18.4) or Equation (18.8) by setting the j-th column of $\boldsymbol{R}_\gamma^{(x)}$ and $\boldsymbol{R}_\gamma^{(y)}$ as

$$\left[\frac{<\boldsymbol{x}_j, \boldsymbol{v}_1>}{<\boldsymbol{u}_1, \boldsymbol{v}_1>}, \frac{<\boldsymbol{x}_j, \boldsymbol{v}_2>}{<\boldsymbol{u}_2, \boldsymbol{v}_2>}, \ldots, \frac{<\boldsymbol{x}_j, \boldsymbol{v}_{j-1}>}{<\boldsymbol{u}_{j-1}, \boldsymbol{v}_{j-1}>}, 1, 0, 0, \ldots, 0\right]^\top,$$

and

$$\left[\frac{<\boldsymbol{u}_1, \boldsymbol{y}_j>}{<\boldsymbol{u}_1, \boldsymbol{v}_1>}, \frac{<\boldsymbol{u}_2, \boldsymbol{y}_j>}{<\boldsymbol{u}_2, \boldsymbol{v}_2>}, \ldots, \frac{<\boldsymbol{u}_{j-1}, \boldsymbol{y}_j>}{<\boldsymbol{u}_{j-1}, \boldsymbol{v}_{j-1}>}, 1, 0, 0, \ldots, 0\right]^\top.$$

This completes the proof. ■

The $(\boldsymbol{U}_\gamma, \boldsymbol{V}_\gamma) \in \mathbb{R}^{m\times\gamma} \times \mathbb{R}^{n\times\gamma}$ in Theorem 18.6 is called a **biconjugate pair** with respect to $\boldsymbol{A}$ if Ω_γ is nonsingular and diagonal. And let $(\boldsymbol{X}_\gamma, \boldsymbol{Y}_\gamma) \in \mathbb{R}^{n\times\gamma} \times \mathbb{R}^{m\times\gamma}$ effect a rank-reducing process for $\boldsymbol{A}$, then $(\boldsymbol{X}_\gamma, \boldsymbol{Y}_\gamma)$ is said to be **biconjugatable** and **biconjugated into a biconjugate pair** of matrices $(\boldsymbol{U}_\gamma, \boldsymbol{V}_\gamma)$, if there exist unit upper triangular matrices $\boldsymbol{R}_\gamma^{(x)}, \boldsymbol{R}_\gamma^{(y)}$ such that $\boldsymbol{X}_\gamma = \boldsymbol{U}_\gamma \boldsymbol{R}_\gamma^{(x)}$ and $\boldsymbol{Y}_\gamma = \boldsymbol{V}_\gamma \boldsymbol{R}_\gamma^{(y)}$.

18.3 Connection to Well-Known Decomposition Methods

18.3.1 LDU Decomposition

Theorem 18.8: (LDU, Chu et al. (1995) Theorem 2.4)

If $(\boldsymbol{X}_\gamma, \boldsymbol{Y}_\gamma) \in \mathbb{R}^{n\times\gamma} \times \mathbb{R}^{m\times\gamma}$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with γ in $\{1, 2, \ldots, r\}$. Then $(\boldsymbol{X}_\gamma, \boldsymbol{Y}_\gamma)$ can be biconjugated if and only if $\boldsymbol{Y}_\gamma^\top \boldsymbol{A}\boldsymbol{X}_\gamma$ has an LDU decomposition.

Proof [of Theorem 18.8] Suppose $\boldsymbol{X}_\gamma$ and $\boldsymbol{Y}_\gamma$ are biconjugatable, then, there exists a unit upper triangular matrices $\boldsymbol{R}_\gamma^{(x)}$ and $\boldsymbol{R}_\gamma^{(y)}$ such that $\boldsymbol{X}_\gamma = \boldsymbol{U}_\gamma \boldsymbol{R}_\gamma^{(x)}$, $\boldsymbol{Y}_\gamma = \boldsymbol{V}_\gamma \boldsymbol{R}_\gamma^{(y)}$ and $\boldsymbol{V}_\gamma^\top \boldsymbol{A}\boldsymbol{U}_\gamma = \boldsymbol{\Omega}_\gamma$ is a nonsingular diagonal matrix. Then, it follows that

$$\boldsymbol{Y}_\gamma^\top \boldsymbol{A}\boldsymbol{X}_\gamma = \boldsymbol{R}_\gamma^{(y)\top} \boldsymbol{V}_\gamma^\top \boldsymbol{A}\boldsymbol{U}_\gamma \boldsymbol{R}_\gamma^{(x)} = \boldsymbol{R}_\gamma^{(y)\top} \boldsymbol{\Omega}_\gamma \boldsymbol{R}_\gamma^{(x)}$$

is the unique unit triangular LDU decomposition of $\boldsymbol{Y}_\gamma^\top \boldsymbol{A}\boldsymbol{X}_\gamma$. This form above can be seen as the **fourth form of biconjugate decomposition**, thus we put the proof into a graybox.

Conversely, suppose $\boldsymbol{Y}_\gamma^\top \boldsymbol{A}\boldsymbol{X}_\gamma = \boldsymbol{R}_2^\top \boldsymbol{D}\boldsymbol{R}_1$ is an LDU decomposition with both $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ being unit upper triangular matrices. Then since $\boldsymbol{R}_1^{-1}$ and $\boldsymbol{R}_2^{-1}$ are also unit upper triangular matrices, and $(\boldsymbol{X}_\gamma, \boldsymbol{Y}_\gamma)$ biconjugates into $(\boldsymbol{X}_\gamma \boldsymbol{R}_1^{-1}, \boldsymbol{Y}_\gamma \boldsymbol{R}_2^{-1})$. ■

Corollary 18.9: (Determinant)

Suppose $(\boldsymbol{X}_\gamma, \boldsymbol{Y}_\gamma) \in \mathbb{R}^{n\times\gamma} \times \mathbb{R}^{m\times\gamma}$ are biconjugatable. Then

$$\det(\boldsymbol{Y}_\gamma^\top \boldsymbol{A}\boldsymbol{X}_\gamma) = \prod_{i=1}^{\gamma} w_i.$$

Proof [of Corollary 18.9] By Theorem 18.8, since $(\boldsymbol{X}_\gamma, \boldsymbol{Y}_\gamma)$ are biconjugatable, then there are unit upper triangular matrices $\boldsymbol{R}_\gamma^{(x)}$ and $\boldsymbol{R}_\gamma^{(y)}$ such that $\boldsymbol{Y}_\gamma^\top \boldsymbol{A}\boldsymbol{X}_\gamma = \boldsymbol{R}_\gamma^{(y)\top} \boldsymbol{\Omega}_\gamma \boldsymbol{R}_\gamma^{(x)}$. The determinant is just product of the trace. ■

Lemma 18.10: (Biconjugatable in Principal Minors)

Let $r = rank(\boldsymbol{A}) \geq \gamma$ with $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. In the Wedderburn sequence, take $\boldsymbol{x}_i$ as the i-th basis in $\mathbb{R}^n$ for $i \in \{1, 2, \ldots, \gamma\}$ (i.e., $\boldsymbol{x}_i = \boldsymbol{e}_i \in \mathbb{R}^n$) and $\boldsymbol{y}_i$ as the i-th basis in $\mathbb{R}^m$ for $i \in \{1, 2, \ldots, \gamma\}$ (i.e., $\boldsymbol{y}_i = \boldsymbol{e}_i \in \mathbb{R}^m$). That is $\boldsymbol{Y}_\gamma^\top \boldsymbol{A}\boldsymbol{X}_\gamma$ is the leading principal submatrix of $\boldsymbol{A}$, i.e., $\boldsymbol{Y}_\gamma^\top \boldsymbol{A}\boldsymbol{X}_\gamma = \boldsymbol{A}_{1:\gamma,1:\gamma}$. Then, $(\boldsymbol{X}_\gamma, \boldsymbol{Y}_\gamma)$ is biconjugatable if and only if the γ-th leading principal minor of $\boldsymbol{A}$ is nonzero. In this case, the γ-th leading principal minor of $\boldsymbol{A}$ is given by $\prod_{i=1}^{\gamma} w_i$.

Proof [of Lemma 18.10] The proof is trivial that the γ-th leading principal minor of $\boldsymbol{A}$ is nonzero will imply that $w_i \neq 0$ for all $i \leq \gamma$. Thus the Wedderburn sequence can be successfully obtained. The converse holds since Corollary 18.9 implies $\det(\boldsymbol{Y}_\gamma^\top \boldsymbol{A}\boldsymbol{X}_\gamma)$ is nonzero. ■

We thus finally come to the LDU decomposition for square matrices.

Theorem 18.11: (LDU: Biconjugate Decomposition for Square Matrices)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $(\boldsymbol{I}_n, \boldsymbol{I}_n)$ is biconjugatable if and only if all the leading principal minors of $\boldsymbol{A}$ are nonzero. In this case, $\boldsymbol{A}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{V}_n^{-\top}\boldsymbol{\Omega}_n\boldsymbol{U}_n^{-1} = \boldsymbol{LDU},$$

where $\boldsymbol{\Omega}_n = \boldsymbol{D}$ is a diagonal matrix with nonzero values on the diagonal, $\boldsymbol{V}_n^{-\top} = \boldsymbol{L}$ is a unit lower triangular matrix and $\boldsymbol{U}_n^{-1} = \boldsymbol{U}$ is a unit upper triangular matrix.

Proof [of Theorem 18.11] From Lemma 18.10, it is trivial that $(\boldsymbol{I}_n, \boldsymbol{I}_n)$ is biconjugatable. From Corollary 18.7, we have $\boldsymbol{U}_n\boldsymbol{R}_n^{(x)} = \boldsymbol{I}_n$ and $\boldsymbol{I}_n = \boldsymbol{V}_n\boldsymbol{R}_n^{(y)}$, thus $\boldsymbol{R}_n^{(x)} = \boldsymbol{U}_n^{-1}$ and $\boldsymbol{R}_n^{(y)} = \boldsymbol{V}_n^{-1}$ are well defined and we complete the proof. ■

18.3.2 Cholesky Decomposition

For symmetric and positive definite, the leading principal minors are positive for sure. The proof is provided in Section 2.2.

Theorem 18.12: (Cholesky: Biconjugate Decomposition for PD Matrices)

For any symmetric and positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the Cholesky decomposition of $\boldsymbol{A}$ can be obtained from the Wedderburn sequence applied to $(\boldsymbol{I}_n, \boldsymbol{I}_n)$ as $(\boldsymbol{X}_n, \boldsymbol{Y}_n)$. In this case, $\boldsymbol{A}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{U}_n^{-\top}\boldsymbol{\Omega}_n\boldsymbol{U}_n^{-1} = (\boldsymbol{U}_n^{-\top}\boldsymbol{\Omega}_n^{1/2})(\boldsymbol{\Omega}_n^{1/2}\boldsymbol{U}_n^{-1}) = \boldsymbol{R}^\top\boldsymbol{R},$$

where $\boldsymbol{\Omega}_n$ is a diagonal matrix with positive values on the diagonal, and $\boldsymbol{U}^{-1}$ is a unit upper triangular matrix.

Proof [of Theorem 18.12] Since the leading principal minors of positive definite matrices are positive, $w_i > 0$ for all $i \in \{1, 2, \ldots, n\}$. It can be easily verified via the LDU from biconjugation decomposition and the symmetric property of $\boldsymbol{A}$ that $\boldsymbol{A} = \boldsymbol{U}_n^{-\top}\boldsymbol{\Omega}_n\boldsymbol{U}_n^{-1}$. And since w_i's are positive, thus $\boldsymbol{\Omega}_n$ is positive definite and can be factored as $\boldsymbol{\Omega}_n = \boldsymbol{\Omega}_n^{1/2}\boldsymbol{\Omega}_n^{1/2}$ which implies $\boldsymbol{\Omega}_n^{1/2}\boldsymbol{U}_n^{-1}$ is the Cholesky factor. ■

18.3.3 QR Decomposition

Without loss of generality, we shall assume that $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ has full column rank so that the columns of $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{QR}$ with $\boldsymbol{Q}, \boldsymbol{R} \in \mathbb{R}^{n\times n}$

Theorem 18.13: (QR: Biconjugate Decomposition for Nonsingular Matrices)

For any nonsingular matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the QR decomposition of $\boldsymbol{A}$ can be obtained from the Wedderburn sequence applied to $(\boldsymbol{I}_n, \boldsymbol{A})$ as $(\boldsymbol{X}_n, \boldsymbol{Y}_n)$. In this case, $\boldsymbol{A}$ can be factored as

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R},$$

where $\boldsymbol{Q} = \boldsymbol{V}_n\boldsymbol{\Omega}_n^{-1/2}$ is an orthogonal matrix and $\boldsymbol{R} = \boldsymbol{\Omega}_n^{1/2}\boldsymbol{R}_n^{(x)}$ is an upper triangular matrix with Form 4 in Theorem 18.8 and let $\gamma = n$

$$\boldsymbol{Y}_n^\top \boldsymbol{A}\boldsymbol{X}_n = \boldsymbol{R}_n^{(y)\top}\boldsymbol{V}_n^\top \boldsymbol{A}\boldsymbol{U}_n\boldsymbol{R}_n^{(x)} = \boldsymbol{R}_n^{(y)\top}\boldsymbol{\Omega}_n\boldsymbol{R}_n^{(x)}.$$

where we set $\gamma = n$ since γ is any value that $\gamma \leq r$ and the rank $r = n$.

Proof [of Theorem 18.13] Since $(\boldsymbol{X}_n, \boldsymbol{Y}_n) = (\boldsymbol{I}_n, \boldsymbol{A})$. Then By Theorem 18.8, we have the decomposition

$$\boldsymbol{Y}_n^\top \boldsymbol{A}\boldsymbol{X}_n = \boldsymbol{R}_n^{(y)\top}\boldsymbol{V}_n^\top \boldsymbol{A}\boldsymbol{U}_n\boldsymbol{R}_n^{(x)} = \boldsymbol{R}_n^{(y)\top}\boldsymbol{\Omega}_n\boldsymbol{R}_n^{(x)}.$$

Substitute $(\boldsymbol{I}_n, \boldsymbol{A})$ into above decomposition, we have

$$\begin{aligned}
\boldsymbol{Y}_n^\top \boldsymbol{A}\boldsymbol{X}_n &= \boldsymbol{R}_n^{(y)\top}\boldsymbol{V}_n^\top \boldsymbol{A}\boldsymbol{U}_n\boldsymbol{R}_n^{(x)} = \boldsymbol{R}_n^{(y)\top}\boldsymbol{\Omega}_n\boldsymbol{R}_n^{(x)}\\
\boldsymbol{A}^\top\boldsymbol{A} &= \boldsymbol{R}_n^{(y)\top}\boldsymbol{\Omega}_n\boldsymbol{R}_n^{(x)}\\
\boldsymbol{A}^\top\boldsymbol{A} &= \boldsymbol{R}_1^\top\boldsymbol{\Omega}_n\boldsymbol{R}_1 \qquad (\boldsymbol{A}^\top\boldsymbol{A} \text{ is symmetric and let } \boldsymbol{R}_1 = \boldsymbol{R}_n^{(x)} = \boldsymbol{R}_n^{(y)}) \qquad (18.10)\\
\boldsymbol{A}^\top\boldsymbol{A} &= (\boldsymbol{R}_1^\top\boldsymbol{\Omega}_n^{1/2\top})(\boldsymbol{\Omega}_n^{1/2}\boldsymbol{R}_1)\\
\boldsymbol{A}^\top\boldsymbol{A} &= \boldsymbol{R}^\top\boldsymbol{R}. \qquad (\text{Let } \boldsymbol{R} = \boldsymbol{\Omega}_n^{1/2}\boldsymbol{R}_1)
\end{aligned}$$

To see why $\boldsymbol{\Omega}_n$ can be factored as $\boldsymbol{\Omega}_n = \boldsymbol{\Omega}_n^{1/2\top}\boldsymbol{\Omega}_n^{1/2}$. Suppose $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n]$. We obtain $w_i = \boldsymbol{a}_i^\top\boldsymbol{a}_i > 0$ since $\boldsymbol{A}$ is nonsingular. Thus $\boldsymbol{\Omega}_n = diag(w_1, w_2, \ldots, w_n)$ is positive definite and can be factored as

$$\boldsymbol{\Omega}_n = \boldsymbol{\Omega}_n^{1/2}\boldsymbol{\Omega}_n^{1/2} = \boldsymbol{\Omega}_n^{1/2\top}\boldsymbol{\Omega}_n^{1/2}. \qquad (18.11)$$

By $\boldsymbol{X}_\gamma = \boldsymbol{U}_\gamma\boldsymbol{R}_\gamma^{(x)}$ in Theorem 18.8 for all $\gamma \in \{1, 2, \ldots, n\}$, we have

$$\begin{aligned}
\boldsymbol{X}_n &= \boldsymbol{U}_n\boldsymbol{R}_1\\
\boldsymbol{I}_n &= \boldsymbol{U}_n\boldsymbol{R}_1, \qquad (\text{Since } \boldsymbol{X}_n = \boldsymbol{I}_n)\\
\boldsymbol{U}_n &= \boldsymbol{R}_1^{-1}
\end{aligned}$$

By $\boldsymbol{Y}_\gamma = \boldsymbol{V}_\gamma\boldsymbol{R}_\gamma^{(y)}$ in Theorem 18.8 for all $\gamma \in \{1, 2, \ldots, n\}$, we have

$$\begin{aligned}
\boldsymbol{Y}_n &= \boldsymbol{V}_n\boldsymbol{R}_1\\
\boldsymbol{A} &= \boldsymbol{V}_n\boldsymbol{R}_1, && (\boldsymbol{A} = \boldsymbol{Y}_n)\\
\boldsymbol{A}^\top\boldsymbol{A} &= \boldsymbol{R}_1^\top\boldsymbol{V}_n^\top\boldsymbol{V}_n\boldsymbol{R}_1\\
\boldsymbol{R}_1^\top\boldsymbol{\Omega}_n\boldsymbol{R}_1 &= \boldsymbol{R}_1^\top\boldsymbol{V}_n^\top\boldsymbol{V}_n\boldsymbol{R}_1, && (\text{From Equation (18.10)})\\
(\boldsymbol{R}_1^\top\boldsymbol{\Omega}_n^{1/2\top})(\boldsymbol{\Omega}_n^{1/2}\boldsymbol{R}_1) &= (\boldsymbol{R}_1^\top\boldsymbol{\Omega}_n^{1/2\top}\boldsymbol{\Omega}_n^{-1/2\top})\boldsymbol{V}_n^\top\boldsymbol{V}_n(\boldsymbol{\Omega}_n^{-1/2}\boldsymbol{\Omega}_n^{1/2}\boldsymbol{R}_1), && (\text{From Equation (18.11)})\\
\boldsymbol{R}^\top\boldsymbol{R} &= \boldsymbol{R}^\top(\boldsymbol{\Omega}_n^{-1/2\top}\boldsymbol{V}_n^\top)(\boldsymbol{V}_n\boldsymbol{\Omega}_n^{-1/2})\boldsymbol{R}
\end{aligned}$$

(18.12)

Thus, $\boldsymbol{Q} = \boldsymbol{V}_n\boldsymbol{\Omega}_n^{-1/2}$ is an orthogonal matrix. ■

18.3.4 SVD

To differentiate the notation, let $\boldsymbol{A} = \boldsymbol{U}^{\text{svd}}\boldsymbol{\Sigma}^{\text{svd}}\boldsymbol{V}^{\text{svd}\top}$ be the SVD of $\boldsymbol{A}$ where $\boldsymbol{U}^{\text{svd}} = [\boldsymbol{u}_1^{\text{svd}}, \boldsymbol{u}_2^{\text{svd}}, \ldots, \boldsymbol{u}_n^{\text{svd}}]$, $\boldsymbol{V}^{\text{svd}} = [\boldsymbol{v}_1^{\text{svd}}, \boldsymbol{v}_2^{\text{svd}}, \ldots, \boldsymbol{v}_n^{\text{svd}}]$ and $\boldsymbol{\Sigma}^{\text{svd}} = diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$. Without loss of generality, we assume $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $rank(\boldsymbol{A}) = n$. Readers can prove the equivalence for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.

If $\boldsymbol{X}_n = \boldsymbol{V}^{\text{svd}}$, $\boldsymbol{Y}_n = \boldsymbol{U}^{\text{svd}}$ effects a rank-reducing process for $\boldsymbol{A}$. From the definition of $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ in Equation (18.4) or Equation (18.8), we have

$$\boldsymbol{u}_k = \boldsymbol{v}_k^{\text{svd}} \qquad \text{and} \qquad \boldsymbol{v}_k = \boldsymbol{u}_k^{\text{svd}} \qquad \text{and} \qquad w_k = \boldsymbol{y}_k^\top \boldsymbol{A}\boldsymbol{x}_k = \sigma_k.$$

That is $\boldsymbol{V}_n = \boldsymbol{U}^{\text{svd}}$, $\boldsymbol{U}_n = \boldsymbol{V}^{\text{svd}}$, and $\boldsymbol{\Omega}_n = \boldsymbol{\Sigma}^{\text{svd}}$, where we set $\gamma = n$ since γ is any value that $\gamma \leq r$ and the rank $r = n$.

By $\boldsymbol{X}_n = \boldsymbol{U}_n\boldsymbol{R}_n^{(x)}$ in Theorem 18.8, we have

$$\boldsymbol{X}_n = \boldsymbol{U}_n\boldsymbol{R}_n^{(x)} \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{V}^{\text{svd}} = \boldsymbol{V}^{\text{svd}}\boldsymbol{R}_n^{(x)} \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{I}_n = \boldsymbol{R}_n^{(x)}$$

By $\boldsymbol{Y}_n = \boldsymbol{V}_n\boldsymbol{R}_n^{(y)}$ in Theorem 18.8, we have

$$\boldsymbol{Y}_n = \boldsymbol{V}_n\boldsymbol{R}_n^{(y)} \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{U}^{\text{svd}} = \boldsymbol{U}^{\text{svd}}\boldsymbol{R}_n^{(y)} \qquad \underrightarrow{\text{leads to}} \qquad \boldsymbol{I}_n = \boldsymbol{R}_n^{(y)}$$

Again, from Theorem 18.8 and let $\gamma = n$, we have

$$\boldsymbol{Y}_n^\top \boldsymbol{A}\boldsymbol{X}_n = \boldsymbol{R}_n^{(y)\top}\boldsymbol{V}_n^\top \boldsymbol{A}\boldsymbol{U}_n\boldsymbol{R}_n^{(x)} = \boldsymbol{R}_n^{(y)\top}\boldsymbol{\Omega}_n\boldsymbol{R}_n^{(x)}.$$

That is

$$\boldsymbol{U}^{\text{svd}\top}\boldsymbol{A}\boldsymbol{V}^{\text{svd}} = \boldsymbol{\Sigma}^{\text{svd}},$$

which is exactly the form of SVD and we prove the equivalence of SVD and biconjugate decomposition when the Wedderburn sequence is applied to $(\boldsymbol{V}^{\text{svd}}, \boldsymbol{U}^{\text{svd}})$ as $(\boldsymbol{X}_n, \boldsymbol{Y}_n)$.

18.4 Proof General Term Formula of Wedderburn Sequence

We define the Wedderburn sequence of $\boldsymbol{A}$ by $\boldsymbol{A}_{k+1} = \boldsymbol{A}_k - w_k^{-1}\boldsymbol{A}_k\boldsymbol{x}_k\boldsymbol{y}_k^\top\boldsymbol{A}_k$ and $\boldsymbol{A}_1 = \boldsymbol{A}$. The proof of the general term formula of this sequence is then:

Proof [of Lemma 18.4] For $\boldsymbol{A}_2$, we have:

$$\begin{aligned}
\boldsymbol{A}_2 &= \boldsymbol{A}_1 - w_1^{-1}\boldsymbol{A}_1\boldsymbol{x}_1\boldsymbol{y}_1^\top\boldsymbol{A}_1\\
&= \boldsymbol{A} - w_1^{-1}\boldsymbol{A}\boldsymbol{u}_1\boldsymbol{v}_1^\top\boldsymbol{A}, \text{ where } \boldsymbol{u}_1 = \boldsymbol{x}_1,\ \boldsymbol{v}_1 = \boldsymbol{y}_1.
\end{aligned}$$

For $\boldsymbol{A}_3$, we can write out the equation:

$$\begin{aligned}
\boldsymbol{A}_3 &= \boldsymbol{A}_2 - w_2^{-1}\boldsymbol{A}_2\boldsymbol{x}_2\boldsymbol{y}_2^\top\boldsymbol{A}_2\\
&= (\boldsymbol{A} - w_1^{-1}\boldsymbol{A}\boldsymbol{u}_1\boldsymbol{v}_1^\top\boldsymbol{A})\\
&\quad - w_2^{-1}(\boldsymbol{A} - w_1^{-1}\boldsymbol{A}\boldsymbol{u}_1\boldsymbol{v}_1^\top\boldsymbol{A})\boldsymbol{x}_2\boldsymbol{y}_2^\top(\boldsymbol{A} - w_1^{-1}\boldsymbol{A}\boldsymbol{u}_1\boldsymbol{v}_1^\top\boldsymbol{A}) && \text{(substitute } \boldsymbol{A}_2)\\
&= (\boldsymbol{A} - w_1^{-1}\boldsymbol{A}\boldsymbol{u}_1\boldsymbol{v}_1^\top\boldsymbol{A})\\
&\quad - w_2^{-1}\boldsymbol{A}(\boldsymbol{x}_2 - w_1^{-1}\boldsymbol{u}_1\boldsymbol{v}_1^\top\boldsymbol{A}\boldsymbol{x}_2)(\boldsymbol{y}_2^\top - w_1^{-1}\boldsymbol{y}_2^\top\boldsymbol{A}\boldsymbol{u}_1\boldsymbol{v}_1^\top)\boldsymbol{A} && \text{(take out } \boldsymbol{A})\\
&= \boldsymbol{A} - w_1^{-1}\boldsymbol{A}\boldsymbol{u}_1\boldsymbol{v}_1^\top\boldsymbol{A} - w_2^{-1}\boldsymbol{A}\boldsymbol{u}_2\boldsymbol{v}_2^\top\boldsymbol{A}\\
&= \boldsymbol{A} - \sum_{i=1}^{2} w_i^{-1}\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top\boldsymbol{A},
\end{aligned}$$

where $\boldsymbol{u}_2 = \boldsymbol{x}_2 - w_1^{-1}\boldsymbol{u}_1\boldsymbol{v}_1^\top \boldsymbol{A}\boldsymbol{x}_2 = \boldsymbol{x}_2 - \frac{\boldsymbol{v}_1^\top \boldsymbol{A}\boldsymbol{x}_2}{w_1}\boldsymbol{u}_1$, $\boldsymbol{v}_2 = \boldsymbol{y}_2 - w_1^{-1}\boldsymbol{y}_2^\top \boldsymbol{A}\boldsymbol{u}_1\boldsymbol{v}_1 = \boldsymbol{y}_2 - \frac{\boldsymbol{y}_2^\top \boldsymbol{A}\boldsymbol{u}_1}{w_1}\boldsymbol{v}_1$. Similarly, we can find the expression of $\boldsymbol{A}_4$ by $\boldsymbol{A}$:

$$
\begin{aligned}
\boldsymbol{A}_4 &= \boldsymbol{A}_3 - w_3^{-1}\boldsymbol{A}_3\boldsymbol{x}_3\boldsymbol{y}_3^\top \boldsymbol{A}_3 \\
&= \boldsymbol{A} - \sum_{i=1}^{2} w_i^{-1}\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top \boldsymbol{A} \\
&\quad - w_3^{-1}(\boldsymbol{A} - \sum_{i=1}^{2} w_i^{-1}\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top \boldsymbol{A})\boldsymbol{x}_3\boldsymbol{y}_3^\top(\boldsymbol{A} - \sum_{i=1}^{2} w_i^{-1}\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top \boldsymbol{A}) && \text{(substitute } \boldsymbol{A}_3\text{)} \\
&= \boldsymbol{A} - \sum_{i=1}^{2} w_i^{-1}\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top \boldsymbol{A} \\
&\quad - w_3^{-1}\boldsymbol{A}(\boldsymbol{x}_3 - \sum_{i=1}^{2} w_i^{-1}\boldsymbol{x}_3\boldsymbol{u}_i\boldsymbol{v}_i^\top \boldsymbol{A})(\boldsymbol{y}_3^\top - \sum_{i=1}^{2} w_i^{-1}\boldsymbol{y}_3^\top \boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top)\boldsymbol{A} && \text{(take out } \boldsymbol{A}\text{)} \\
&= \boldsymbol{A} - \sum_{i=1}^{2} w_i^{-1}\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top \boldsymbol{A} - w_3^{-1}\boldsymbol{A}\boldsymbol{u}_3\boldsymbol{v}_3^\top \boldsymbol{A} \\
&= \boldsymbol{A} - \sum_{i=1}^{3} w_i^{-1}\boldsymbol{A}\boldsymbol{u}_i\boldsymbol{v}_i^\top \boldsymbol{A},
\end{aligned}
$$

where $\boldsymbol{u}_3 = \boldsymbol{x}_3 - \sum_{i=1}^{2}\frac{\boldsymbol{v}_i^\top \boldsymbol{A}\boldsymbol{x}_3}{w_i}\boldsymbol{u}_i$, $\boldsymbol{v}_3 = \boldsymbol{y}_3 - \sum_{i=1}^{2}\frac{\boldsymbol{y}_2^\top \boldsymbol{A}\boldsymbol{u}_i}{w_i}\boldsymbol{v}_i$.

Continue this process, we can define

$$
\boldsymbol{u}_k = \boldsymbol{x}_k - \sum_{i=1}^{k-1}\frac{\boldsymbol{v}_i^\top \boldsymbol{A}\boldsymbol{x}_k}{w_i}\boldsymbol{u}_i, \qquad \text{and} \qquad \boldsymbol{v}_k = \boldsymbol{y}_k - \sum_{i=1}^{k-1}\frac{\boldsymbol{y}_k^\top \boldsymbol{A}\boldsymbol{u}_i}{w_i}\boldsymbol{v}_i,
$$

and find the general term of Wedderburn sequence. ■

19. Acknowledgments

We thank Gilbert Strang for raising the question formulated in Corollary 6.2, checking the writing of the survey, for a stream of ideas and references about the three factorizations from the steps of elimination, and for the generous sharing of the manuscript of (Strang and Drucker, 2021).

References

Sudipto Banerjee and Anindya Roy. *Linear algebra and matrix analysis for statistics*, volume 181. Crc Press Boca Raton, FL, USA:, 2014.

Amir Beck. *First-Order Methods in Optimization*, volume 25. SIAM, 2017.

James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, NY, USA., 2007.

Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.

Christos Boutsidis and Efstratios Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, 41(4):1350–1362, 2008.

Stephen Boyd and Lieven Vandenberghe. *Introduction to applied linear algebra: vectors, matrices, and least squares*. Cambridge university press, 2018.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Tony F Chan. An improved algorithm for computing the singular value decomposition. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):72–83, 1982.

Tony F Chan. Rank revealing QR factorizations. *Linear algebra and its applications*, 88: 67–82, 1987.

Moody T Chu, Robert E Funderlic, and Gene H Golub. A rank–one reduction formula and its applications to matrix factorizations. *SIAM review*, 37(4):512–530, 1995.

Pierre Comon, Xavier Luciani, and André LF De Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 23(7-8):393–405, 2009.

Froilán M Dopico, Charles R Johnson, and Juan M Molera. Multiple LU factorizations of a singular matrix. *Linear algebra and its applications*, 419(1):24–36, 2006.

Ricardo D Fierro and Per Christian Hansen. Low-rank revealing UTV decompositions. *Numerical Algorithms*, 15(1):37–55, 1997.

Leslie V Foster. Solving rank-deficient and ill-posed problems using UTV and QR factorizations. *SIAM journal on matrix analysis and applications*, 25(2):582–600, 2003.

Jean Gallier and Jocelyn Quaintance. *Fundamentals of Linear Algebra and Optimization*. Department of Computer and Information Science, University of Pennsylvania, 2017.

James E Gentle. *Numerical linear algebra for applications in statistics*. Springer Science & Business Media, 1998.

James E Gentle. Matrix algebra. *Springer texts in statistics, Springer, New York, NY, doi*, 10:978–0, 2007.

Paris V Giampouras, Athanasios A Rontogiannis, and Konstantinos D Koutroumbas. Alternating iteratively reweighted least squares minimization for low-rank matrix factorization. *IEEE Transactions on Signal Processing*, 67(2):490–503, 2018.

George T Gilbert. Positive definite matrices and Sylvester's criterion. *The American Mathematical Monthly*, 98(1):44–46, 1991.

Philip E Gill, Walter Murray, and Margaret H Wright. *Numerical linear algebra and optimization*. SIAM, 2021.

Nicolas Gillis. The why and how of nonnegative matrix factorization. *Connections*, 12:2–2, 2014.

Israel Gohberg and Seymour Goldberg. A simple proof of the jordan decomposition theorem for matrices. *The American Mathematical Monthly*, 103(2):157–159, 1996.

Donald Goldfarb. Factorized variable metric methods for unconstrained optimization. *Mathematics of Computation*, 30(136):796–811, 1976.

Gene Golub and William Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224, 1965.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2013.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

AW Hales and IBS Passi. Jordan decomposition. In *Algebra*, pages 75–87. Springer, 1999.

Nicholas J Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002.

Nicholas J Higham. Cholesky factorization. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(2):251–254, 2009.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Alston S Householder. *Principles of numerical analysis*. Courier Corporation, 2006.

Tsung-Min Hwang, Wen-Wei Lin, and Eugene K Yang. Rank revealing LU factorizations. *Linear algebra and its applications*, 175:115–141, 1992.

Camille Jordan. *Traité des substitutions et des équations algébriques*. Gauthier-Villars, 1870.

N Kishore Kumar and Jan Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017.

Martin Koeber and Uwe Schäfer. The unique square root of a positive semidefinite matrix. *International Journal of Mathematical Education in Science and Technology*, 37(8):990–992, 2006.

Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995.

Daniel D Lee and Hyunjune Sebastian Seung. Algorithms for non-negative matrix factorization. In *14th Annual Neural Information Processing Systems Conference, NIPS 2000*. Neural information processing systems foundation, 2001.

Jun Lu. A survey on Bayesian inference for Gaussian mixture model. *arXiv preprint arXiv:2108.11753*, 2021a.

Jun Lu. On the column and row ranks of a matrix. *arXiv preprint arXiv:2112.06638*, 2021b.

Jun Lu. Numerical matrix decomposition and its modern applications: A rigorous first course. *arXiv preprint arXiv:2107.02579*, 2021c.

Jun Lu. Revisit the fundamental theorem of linear algebra. *arXiv preprint arXiv:2108.04432*, 2021d.

Jun Lu. A rigorous introduction for linear models. *arXiv preprint arXiv:2105.04240*, 2021e.

Per-Gunnar Martinsson. Randomized methods for matrix computations. *The Mathematics of Data*, 25(4):187–231, 2019.

L Miranian and Ming Gu. Strong rank revealing LU factorizations. *Linear algebra and its applications*, 367:1–16, 2003.

C-T Pan. On the existence and computation of rank-revealing LU factorizations. *Linear Algebra and its Applications*, 316(1-3):199–222, 2000.

V Paul Pauca, Jon Piper, and Robert J Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear algebra and its applications*, 416(1):29–47, 2006.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*, volume 37. Springer Science & Business Media, 2010.

Wil HA Schilders. Solution of indefinite linear systems using an LQ decomposition for the linear constraints. *Linear algebra and its applications*, 431(3-4):381–395, 2009.

Matthias Seeger. Low rank updates for the Cholesky decomposition. Technical report, 2004.

Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

Gilbert W Stewart. *Matrix Algorithms: Volume 1: Basic Decompositions*. SIAM, 1998.

GW Stewart. The decompositional approach to matrix computation. *Computing in Science & Engineering*, 2(1):50–59, 2000.

Gilbert Strang. *Introduction to linear algebra*. Wellesley-Cambridge Press Wellesley, 4th edition, 2009.

Gilbert Strang. *Linear algebra and learning from data*. Wellesley-Cambridge Press Cambridge, 2019.

Gilbert Strang. *Linear algebra for everyone*. Wellesley-Cambridge Press Wellesley, 2021.

Gilbert Strang and Daniel Drucker. Three matrix factorizations from the steps of elimination. 2021.

Gilbert Strang and Cleve Moler. LU and CR elimination. 2021.

Kuduvally Swamy. On Sylvester's criterion for positive-semidefinite matrices. *IEEE Transactions on Automatic Control*, 18(3):306–306, 1973.

Gábor Takács and Domonkos Tikk. Alternating least squares for personalized ranking. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 83–90, 2012.

Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

Robert van de Geijn and Margaret Myers. Advanced linear algebra: Foundations to frontiers. *Creative Commons NonCommercial (CC BY-NC)*, 2020.

Ming Yang. Matrix decomposition. *Northwestern University, Class Notes*, 2000.