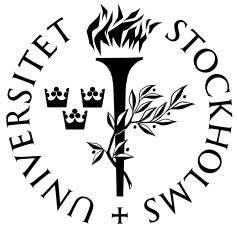

SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Practical Linear Algebra for Applied General Linear Systems

av

Jakub Olczak

2011 - No 7

Practical Linear Algebra for Applied General Linear Systems

Jakub Olczak

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Paul Vaderlind

2011

Abstract. We study the underlying theory of matrix equations, their interpretation and develop some of the practical linear algebra behind the standard tools used, in applied mathematics, to solve systems of linear equations: the LU factorization, the QR factorization and the SVD (Singular Value Decomposition.) We also extend our study to more general systems giving rise to linear least squares problems and show how the QR and SVD factorizations are used to solve overdetermined problems and can be applied to rank deficient problems.

Contents

1. Introduction

1.1. **On Solving Matrix Equations.** The most important problem in applied mathematics is equations solving. If A is our coefficient matrix, the solution to the matrix equation $A\mathbf{x} = \mathbf{y}$ is $\mathbf{x} = A^{-1}\mathbf{y}$, assuming the *inverse* A^{-1} exists, which from elementary linear algebra implies that A must be *square* and *nonsingular*. Even if it does exist, it is not certain whether it or the solution $\mathbf{x}$ is useful. What happens if we operate in finite precision arithmetic - for example doing calculations on a calculator - if the data, i.e. the coefficients of the matrix A and/or vector $\mathbf{y}$, is collected using a measuring instrument with inaccuracies? A simple illustration of a new class of problems:

Given the equations $.01x + 1.6y = 32.1$ and $x + .6y = 22$ inserting $x^* = 10$, $y^* = 20$ verifies that this is the true solution. Solving the system of equations with three digit accuracy using the standard Gaussian elimination (with reduction to row echelon form) we get the augmented system

$$\left(\begin{array}{cc|c} .01 & 1.6 & 32.1 \\ 1 & .6 & -3190 \end{array}\right) \overset{\text{Gaussian elimination}}{\underset{\text{3 digit accuracy}}{\Longrightarrow}} \left(\begin{array}{cc|c} 1 & 0 & -10 \\ 0 & 1 & 20.1 \end{array}\right).$$

We got $x = -10$ and $y = 20.1$. A small error caused catastrophic errors in the solution. If we reorder the rows (*pivot*) we instead get

$$\left(\begin{array}{cc|c} 1 & .6 & -3190 \\ .01 & 1.6 & 32.1 \end{array}\right) \overset{Gaussianelimination}{\underset{\text{3 digit accuracy}}{\Longrightarrow}} \left(\begin{array}{cc|c} 1 & 0 & 9.9 \\ 0 & 1 & 20.1 \end{array}\right).$$

This is acceptable, but how would we reorder an arbitrary 500×500 matrix to get an acceptable solution? Even worse, suppose we are given the systems

$$\left(\begin{array}{cc|c} .01 & 1.6 & 32.1 \\ 1 & .6 & -3190 \\ 3 & -2.1 & 332 \end{array}\right) \text{ or } \left(\begin{array}{ccc|c} .01 & 1.6 & 3 & 32.1 \\ 1 & .6 & -2.1 & -3190 \\ -.10004 & -16.0001 & -30.0009 & -321 \end{array}\right).$$

One is rectangular and the other is singular in three digit accuracy, and neither is solvable by matrix inversion. Both situations can and do arise in practice. Do they have solutions?

These simple examples leads us to study linear systems of equations, their solutions and their meanings, to see if we can extend our understanding and develop practical methods for equation solving in applications.

1.2. **About This Paper.** This paper will focus on studying the theory of linear systems of equations and their solutions, to see how it is expanded into practical situations. We will limit our scope and focus on the practical linear algebra, developing the framework leading up to linear least squares solutions to problems. We mostly ignore questions of implementation (algorithms, complexity, errors, perturbation theory, problem conditioning and so on) as this is the domain of numerical analysis and scientific computing.

Especially error propagation, perturbation theory and problem conditioning are crucial, but a fair study of these would far oversize this paper, and its study depends on findings in this paper (for example singular values.) Because it is so important we none the less give a brief background in the Appendix. While these considerations are important to fully understand the methods developed, the study of these would still be highly selective, heavily implementation dependent and situation specific. There are good practices for every problem, but no best implementation for all situations. Nor is this strictly necessary. The underlining linear algebra that will be developed is still sound, and we will take a general approach, focusing on

understanding the standard well-proven and generally robust methods that are used in applications.

2. Background Theory

We begin by reviewing some concepts of linear algebra and build onwards, but first some important notes on notation and matrices.

2.1. Note on Matrices and Their Representation.

It will be assumed throughout that vectors $\mathbf{v}$ are column vectors, and row vectors are represented as $\mathbf{v}^T$. Often when studying the properties of algorithms it is more enlightening, convenient, and sufficient ([Demmel]) to look at a matrix as divided into blocks of submatrices, rather than individual elements or column vectors.

Definition 1. Let A be a matrix $m \times n$. Let $1 \leq i_1 \leq \ldots \leq i_k \leq m$ and $1 \leq j_1 \leq \ldots \leq j_k \leq n$ be two sets of contiguous indexes. The $k \times l$ matrix S of entries $s_{pq} = a_{i_p j_q}$ with $p \in [1,k]$, $q = [1,l]$ is called a *submatrix* of A. If $k = l$ and $i_r = j_r$ for $r \in [1,k]$, S is called a *principal submatrix* of A. If $S = A_{1:k,1:k}$ where $k \leq \min(m,n)$ we call it a *leading principal submatrix*.

A submatrix S, of A, is A with any rows and columns removed. A leading principal submatrix is the square $k \times k$ upper left corner of A. A principal submatrix is a square matrix with corresponding rows and columns deleted. In our rectangular definition, all rows/columns$> \max(m,n)$ are first removed to make it square.

We will also use more specific notation. If A is an $m \times m$ matrix, then $A_{k:l,x:y}$ denotes the submatrix consisting of the elements in rows k to l from columns x to y. $A_{k:l,p}$ signifies the k to l elements of column p, and $A_{1,1:m}$ would denote the entire first row of A. If $A = 0$, in the previous, it will signify matrix of zeroes of that size and likewise $A = I$ will be used to signify an (always square) identity *submatrix*.

Definition 2. A $m \times n$ matrix A is called *block partitioned* or said to be *partitioned into submatrices* if

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1l} \\ \vdots & \ddots & \vdots \\ A_{k1} & \cdots & A_{kl} \end{bmatrix}$$

where A_{ij} are submatrices of A.

Provided that the size of each single block is such that any single matrix operation is well-defined, from [QuaSacSal] we gather the following useful results.

Proposition 3. *Let A and B be block matrices*

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1l} \\ \vdots & \ddots & \vdots \\ A_{k1} & \cdots & A_{kl} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & \cdots & B_{1l} \\ \vdots & \ddots & \vdots \\ B_{k1} & \cdots & B_{kl} \end{bmatrix}$$

where A_{ij} and B_{ij} are matrices $(k_i \times l_j)$ and $(m_i \times n_j)$. Then we have

1. $$\lambda A = \begin{bmatrix} \lambda A_{11} & \cdots & \lambda A_{1l} \\ \vdots & \ddots & \vdots \\ \lambda A_{k1} & \cdots & \lambda A_{kl} \end{bmatrix}, \quad \lambda \in \mathbb{C}; \quad A^T = \begin{bmatrix} A_{11}^T & \cdots & A_{k1}^T \\ \vdots & \ddots & \vdots \\ A_{1l}^T & \cdots & A_{kl}^T \end{bmatrix};$$

2. *if $k = m$, $l = n$, $m_i = k_i$ and $n_j = l_j$, then*

$$A + B = \begin{bmatrix} A_{11} + B_{11} & \cdots & A_{1l} + B_{1l} \\ \vdots & \ddots & \vdots \\ A_{k1} + B_{k1} & \cdots & A_{kl} + B_{kl} \end{bmatrix};$$

3. if $l = m$, $l_i = m_i$, $n = k_i$, then letting $C_{ij} = \sum_{s=1}^{m} A_{is}B_{sj}$,

$$C = \begin{bmatrix} C_{11} & \cdots & C_{1l} \\ \vdots & \ddots & \vdots \\ C_{k1} & \cdots & C_{kl} \end{bmatrix}.$$

This simply means that we can treat the submatrices as elements in their own right, as long as the resulting matrix operations between submatrices are defined.

Definition 4. A *permutation matrix* P is a identity matrix with permuted rows.

From [OlvShaDraft] (Chapter 1, Lemma 1.9) and [Demmel] (Section 2.3, Lemma 2.2) and we give the following without further proof.

Lemma 5. *A matrix P is a permutation matrix iff each row of P contains all* 0 *entries except for a single* 1*, and, in addition, each column of P also contains all* 0 *entries except for a single* 1*.*

Lemma 6. *Let P, P_1, and P_2 be $n \times n$ permutation matrices and X be an $n \times n$ matrix. Then*
1. PX (XP) is the same as X with its rows (columns) permuted.
2. $P^{-1} = P^T$.
3. $\det(P) = \pm 1$.
4. $P_1 \cdot P_2$ is also a permutation matrix.

Definition 7. Let S be *any* nonsingular matrix. Then A and $B = S^{-1}AS$ are called *similar* matrices, and S is a similarity transformation.

Looking forward somewhat (to Definition 16 and on) we conclude the following for similar matrices.

Lemma 8. *Similar matrices have the same rank.*

Proof. Assume $\text{rank}P = n \geq r$. If $\text{rank}A = r$ then PA cannot have larger rank and neither can AP^{-1} or $PAP^{-1} = B$, so $\text{rank}B \leq \text{rank}A$ (excessive columns end up in $\ker A$). If $\text{rank}B = k$, and reversing the argument, we find that $\text{rank}A \leq \text{rank}B$ which leaves us with $\text{rank}A = \text{rank}B = r = k$. Similar matrices have the same rank. □

It is possible to show many other important properties for similar matrices, for example that they form equivalence relationships[1], have the same eigenvalues etc.

2.2. **Vector spaces, Subspaces and Basis.**

Definition 9. A *vector space* is a set $\mathcal{V}$ equipped with two operations:
(i) *Addition*: if $\mathbf{v}$, $\mathbf{w} \in \mathcal{V}$ then $\mathbf{v} + \mathbf{w} \in \mathcal{V}$; (ii) *Scalar multiplication*: for $c \in \mathbb{R}$, $c\mathbf{v} \in \mathcal{V}$.

The operations are required to satisfy the following axioms for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$ and all $c, d \in \mathbb{R}$:
(a) *Commutativity of Addition*: $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$
(b) *Associativity of Addition*: $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$
(c) *Additive Identity*: There is a zero element $\mathbf{0} \in V$ satisfying $\mathbf{v} + \mathbf{0} = \mathbf{v} = \mathbf{0} + \mathbf{v}$
(d) *Additive Inverse*: For each $\mathbf{v} \in V$ there is an element $-\mathbf{v} \in V$ such that $\mathbf{v} + (-\mathbf{v}) = 0 = (-\mathbf{v}) + \mathbf{v}$.
(e) *Distributivity*: $(c + d)\mathbf{v} = (c\mathbf{v}) + (d\mathbf{v})$, and $c(\mathbf{v} + \mathbf{w}) = (c\mathbf{v}) + (c\mathbf{w})$.
(f) *Associativity of Scalar Multiplication*: $c(d\mathbf{v}) = (cd)\mathbf{v}$.
(g) *Unit for Scalar Multiplication*: the scalar $1 \in \mathbb{R}$ satisfies $1\mathbf{v} = \mathbf{v}$.

[1] $S = I$ gives A similar to A; $S^{-1} = M \Rightarrow M^{-1}BM = A$ show symmetry; If $N^{-1}BN = C \Leftrightarrow N^{-1}S^{-1}ASN = C$ show A similar to C via SN, show transitivity; $A \sim B$.

Though we will focus on vectors, the members of the "vector space" do not have to be vectors. They can just as well be matrices, polynomials, functions etc. In practice we often work with subsets of the vector space.

Definition 10. A *subspace* of a vector space $\mathcal{V}$ is a subset $\mathcal{W} \subset \mathcal{V}$ which is a vector space in its own right.

Proposition 11. *A subset $\mathcal{W} \subset \mathcal{V}$ of a vector space is a subspace iff (a) for every $\mathbf{v}, \mathbf{w} \in \mathcal{W}$, the sum $\mathbf{v} + \mathbf{w} \in \mathcal{W}$, and (b) for every $\mathbf{v} \in \mathcal{W}$ and every $c \in \mathbb{R}$, the scalar product $c\mathbf{v} \in \mathcal{W}$.*

Proof. We want to show that given (a) and (b) the subset is a vector space and that the operations fulfill all axioms of Definition 9. Let $c \in \mathbb{R}$, and $\mathbf{v}, \mathbf{w} \in \mathcal{W}$ which we can regard as part of $\mathcal{V}$. We know that $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$ because $\mathcal{V}$ is a vector space, but closure also implies that it is part of $\mathcal{W}$. This shows (a) of Definition 9 is fulfilled. The other properties follow from equally trivial argumentation. □

Definition 12. Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be a finite collection of elements of a vector space $\mathcal{V}$. A sum of the form

$$\sum_{i=1}^{k} c_i \mathbf{v}_i$$

where $c_1, \ldots, c_k$ are any scalars, is known as a *linear combination* of the elements $\mathbf{v}_1, \ldots, \mathbf{v}_k$ and their *span* is the subset $\mathcal{W} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset \mathcal{V}$ consisting of all possible linear combinations.

Proposition 13. *The span of a collection of vectors, $\mathcal{W} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, forms a subspace of the underlying vector space.*

Proof. Let $\mathbf{v} = \sum_{i=1}^{k} c_i \mathbf{v}_i$ and $\widehat{\mathbf{v}} = \sum_{i=1}^{k} \widehat{c}_i \mathbf{v}_i$. If there are any two linear combinations, then their sum $\mathbf{v} + \hat{\mathbf{v}} = (c_1 + \hat{c}_1)\mathbf{v}_1 + \ldots + (c_k + \hat{c}_k)\mathbf{v}_k$ and any scalar multiple $a\mathbf{v} = (ac_1)\mathbf{v}_1 + \ldots + (ac_k)\mathbf{v}_k$ are also linear combinations. □

Definition 14. The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathcal{V}$ are called *linearly dependent* if there exists a collection of scalars $c_1, \ldots, c_k$, *not all zero*, such that $\sum_{i=1}^{k} c_i \mathbf{v}_i = 0$. Vectors which are not linearly dependent are *linearly independent*.

Theorem 15. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^m$ and let $A = (\mathbf{v}_1 \cdots \mathbf{v}_n)$ be the corresponding $m \times n$ matrix.*

(a) The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^m$ are linearly dependent iff there is a non-zero solution $\mathbf{v} \neq \mathbf{0}$ to the homogenous linear system $A\mathbf{c} = \mathbf{0}$.

(b) The vectors are linearly independent iff the only solution to the homogenous system $A\mathbf{c} = \mathbf{0}$ is the trivial one $\mathbf{c} = \mathbf{0}$.

(c) A vector $\mathbf{b}$ lies in the span of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ iff the linear system $A\mathbf{c} = \mathbf{b}$ is compatible, i.e., it has at least one solution.

Proof. For $\mathbf{v_1}, \ldots, \mathbf{v}_n$ to be linearly dependent a $\mathbf{c} = (c_1, \ldots, c_n)^T \neq \mathbf{0}$ must exist such that the linear combination $A\mathbf{c} = c_1\mathbf{v}_1 + \ldots + c_k\mathbf{v}_n = \mathbf{0}$. Therefore the linear dependence requires the existence of a nontrivial solution to the homogenous linear system $A\mathbf{c} = \mathbf{0}$. This shows (a). Property (b) follows directly from (a) and Definition 14, since any other solution $\mathbf{c} \neq \mathbf{0}$ would make it linearly dependent. This also follows from (a) and (c) since two or more solutions implies a non-unique linear combination $\Leftrightarrow$ infinitely many solutions.

To show (c) we write $A = [\mathbf{v}_1 \ \ldots \ \mathbf{v}_n] = [\mathbf{v}]$. Then $A\mathbf{c} = \mathbf{b}$ can be expressed as a linear combination $c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n = \mathbf{b}$. Assume $c_i^*, c_i \in \mathbb{R}$, and $c_i\mathbf{v}_i = c_i^*\mathbf{v}_i \Leftrightarrow (c_i - c_i^*)\mathbf{v}_i = \mathbf{0}$, for any $i = [1, n]$. If this is to be valid (compatible) $\mathbf{v}_i = \mathbf{0} \vee c_i = c_i^*$. Especially, if $\mathbf{v}_i = \mathbf{0}$ for any i (i.e. is a linear combination of some of the other $\mathbf{v}_{n-1}$ vectors) then any combination of c_i and c_i^* will work and the system has infinitely

many solutions. Assuming $\mathbf{v}_i \neq \mathbf{0}$, either $c_i = c_i^*$ for all i and the solution is unique, or $c_i \neq c_i^*$ for some i and the linear combination does not make sense (incompatible) and hence $\mathbf{b}$ has no solution and cannot lie in the space spanned by $\mathbf{v}$. □

Definition 16. The *rank* r of a matrix A is the number of linearly independent columns of A.

Lemma 17. *Rank* r of A is also=#pivots=#linearly independent rows.

Proof. The number of *pivots* (diagonal entries of the row echelon form) is the same as the number of nonzero columns. If some column is a linear combination of the others, it can be zeroed (expressed as linear combination of the other columns.) This can be repeated for any linearly dependent column until no more linear combinations are left. The number of nonzero columns is then the number of linearly independent columns, the rank. These could be rearranged into row echelon form and we find that #pivots=rank. This rearrangement can be preformed by relabeling, and/or rearranging A accordingly (e.g. $\mathbf{v}_i \longleftrightarrow \mathbf{v}_k$, where the $\mathbf{v}$'s are columns of A) or applying a permutation matrix P. It is also clear that the same argument holds for the rows: #pivots=#nonzero rows.[2] □

Proposition 18. *A set of* n *vectors in* $\mathbb{R}^m$ *is linearly independent iff the corresponding* $m \times n$ *matrix* A *has rank* n*. In particular, this requires* $n \leq m$*. Any collection of* $n > m$ *vectors in* $\mathbb{R}^m$ *is linearly dependent.*

Proof. This follows from Theorem 15 and Lemma 17. The second part follows from the fact that if $m < n$ there are more free parameters than equations and hence infinitely many solutions, since any solution $\mathbf{c} = (\underbrace{0, \ldots, 0}_{m}, \underbrace{c_1, \ldots, c_{m-n}}_{n-m}) \neq \mathbf{0}$ solves the homogenous system. □

Proposition 19. *A collection of* n *vectors will span* $\mathbb{R}^m$ *iff their* $m \times n$ *matrix has rank* m*. In particular, this requires* $n \geq m$*.*

Proof. $n < m$ linearly independent vectors cannot span $\mathbb{R}^m$, because not all possible linear combinations $\mathbf{b} \in \mathbb{R}^m$ would be representable and it would not be closed under addition. This requires rank $n = m$. If $n > m$, the further $n - m$ vectors will be linear combinations of the first m and thus lie in their span. □

Definition 20. A *basis* of a vector space $\mathcal{V}$ is a finite collection of elements $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathcal{V}$ which span $\mathcal{V}$, and are linearly independent.

Proposition 21. *Every basis of* $\mathbb{R}^m$ *contains exactly* m *vectors. A set of* m *vectors* $\mathbf{v_1}, \ldots, \mathbf{v}_m \in \mathbb{R}^m$ *is a basis iff the* $m \times m$ *matrix* $A = (\mathbf{v}_1 \ \ldots \ \mathbf{v}_m)$ *is nonsingular.*

Proof. Linear independence requires that the only solution to the homogenous system $A\mathbf{x} = \mathbf{0}$ is the trivial one $\mathbf{x} = \mathbf{0}$. Also, a vector $\mathbf{b} \in \mathbb{R}^m$ will lie in the span$\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ iff the linear system $A\mathbf{x} = \mathbf{b}$ has a solution. For $\mathbf{v}_1, \ldots, \mathbf{v}_m$ to span $\mathbb{R}^m$, this must hold for all possible right hand sides $\mathbf{b}$. Both results require that rank$A = m$, meaning that it is square and full rank, i.e. *nonsingular.* □

Lemma 22. *Suppose* $\mathbf{v}_1, \ldots, \mathbf{v}_m$ *span a vector space* $\mathcal{V}$*. Then every set of* $n > m$ *elements* $\mathbf{w}_1, \ldots, \mathbf{w}_n \in \mathcal{V}$ *is linearly dependent.*

[2] We will see this again clearly in Section 3.4 when the rank is the number of singular values $\sigma_i > 0$.

Proof. Each element $\mathbf{w}_j = \sum_{i=1}^{m} a_{ij}\mathbf{v}_i$, $j = 1, \ldots, n$ can be written as a linear combination of the spanning elements.

$$c_1\mathbf{w}_1 + \ldots + c_n\mathbf{w}_n = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}c_j\mathbf{v}_i.$$

This linear combination will be zero whenever $\mathbf{c} = (c_1, \ldots, c_n)^T$ solves the homogenous linear system $\sum_{j=1}^{n} a_{ij}c_j = 0$, $i = 1, \ldots, m$, of m equations in $n > m$ unknowns. Any homogenous system with more unknowns than equations always has a non-trivial solution $\mathbf{c} \neq \mathbf{0}$, and this immediately implies that $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are linearly dependent. □

Proposition 23. *Suppose the vector space $\mathcal{V}$ has a basis $\mathbf{v}_1, \ldots, \mathbf{v}_m$. Then every other basis of $\mathcal{V}$ has the same number of elements in it. The number is called the dimension of $\mathcal{V}$ and is written as* $\dim\mathcal{V} = m$.

Proof. Suppose we have two bases containing a different number of elements. By definition, the smaller basis spans the vector space. But then Lemma 22 demands that the elements in the larger supposed basis must be linearly dependent. This contradicts our assumption that both sets are bases, and proves the proposition. □

From this we can summarize the following optimality property for bases.

Theorem 24. *Suppose $\mathcal{V}$ is a n-dimensional vector space. Then*
(1) Every set of more than n elements of $\mathcal{V}$ is linearly dependent.
(2) No set less than n elements span $\mathcal{V}$.
(3) A set of n elements forms a basis iff it spans $\mathcal{V}$.
(4) A set of n elements forms a basis iff it is linearly independent.

Lemma 25. *The elements $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form a basis of $\mathcal{V}$ iff every $\mathbf{v} \in \mathcal{V}$ can be written uniquely as a linear combination thereof:*

$$\mathbf{v} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n = \sum_{i=1}^{n} c_i\mathbf{v}_i.$$

Proof. The condition that the basis span $\mathcal{V}$ implies every $\mathbf{v} \in \mathcal{V}$ can be written as some linear combination of the basis elements. Suppose we can write an element $\mathbf{x} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n = \hat{c}_1\mathbf{v}_1 + \ldots + \hat{c}_n\mathbf{v}_n$ as two different combinations.

Subtracting one from the other we find that $(c_1 - \hat{c}_1)\mathbf{v}_1 + \ldots + (c_n - \hat{c}_n)\mathbf{v}_n = \mathbf{0}$. Linear independence of the basis elements implies that the coefficients $c_i - \hat{c}_i = 0 \Leftrightarrow c_i = \hat{c}_i$ and the linear combinations are the same. □

2.3. **Inner Products and Norms.**

Definition 26. An *inner product* on the real vector space $\mathcal{V}$ is a pairing that takes two vectors $\mathbf{v}$, $\mathbf{w} \in \mathcal{V}$ and produces a real number $\langle \mathbf{v}; \mathbf{w} \rangle \in \mathbb{R}$. The inner product is required to satisfy the following three axioms for all $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w} \in V$ and c, $d \in \mathbb{R}$
(i) Bilinearity

$$\begin{aligned} \langle c\mathbf{u} + d\mathbf{v}; \mathbf{w} \rangle &= c\,\langle \mathbf{u}; \mathbf{w} \rangle + d\,\langle \mathbf{v}; \mathbf{w} \rangle, \\ \langle \mathbf{u}; c\mathbf{v} + d\mathbf{w} \rangle &= c\,\langle \mathbf{u}; \mathbf{v} \rangle + d\,\langle \mathbf{u}; \mathbf{w} \rangle. \end{aligned}$$

(ii) Symmetry

$$\langle \mathbf{v}; \mathbf{w} \rangle = \langle \mathbf{w}; \mathbf{v} \rangle.$$

(iii) Positivity

$$\langle \mathbf{v}; \mathbf{v} \rangle > 0 \quad \text{whenever} \quad \mathbf{v} \neq \mathbf{0}, \quad \langle \mathbf{0}; \mathbf{0} \rangle = 0.$$

A vector space equipped with an inner product is called an *inner product space.*

A familiar example is the Euclidean dot product in $\mathbb{R}^n$ $\langle \mathbf{v}; \mathbf{w} \rangle = \mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T\mathbf{w} = \sum_{i=1}^{n} v_i w_i$, which we already know from previous experience satisfy (i)-(iii) and is thus an inner product norm.

Definition 27. Given an inner product, the associated *norm of a vector* $\mathbf{v} \in \mathcal{V}$ *is defined as* $||\mathbf{v}|| = \sqrt{\langle \mathbf{v}; \mathbf{v} \rangle}$.

Note. Definition 26 ensures that $\mathbb{R} \ni ||\mathbf{v}|| \geq 0$ with equality only if $\mathbf{v} = \mathbf{0}$.

Proposition 28. *Every inner product satisfies the Cauchy-Schwarz inequality* $|\langle \mathbf{v}; \mathbf{w} \rangle| \leq ||\mathbf{v}|| \, ||\mathbf{w}||$, $\mathbf{v}, \mathbf{w} \in \mathcal{V}$. *Here* $||\mathbf{v}||$ *is the associated norm, while* $|\cdot|$ *denotes the absolute value. Equality holds iff* $\mathbf{v}$ *and* $\mathbf{w}$ *are parallel vectors.*

Proof. The case when $\mathbf{w} = \mathbf{0}$ is trivial, since both sides of the inequality equal 0. Thus we may suppose $\mathbf{w} \neq \mathbf{0}$. Let $t \in \mathbb{R}$ be an arbitrary scalar. Using the three basic inner product axioms, we have

$$0 \leq ||\mathbf{v} + t\mathbf{w}||^2 = \langle \mathbf{v} + t\mathbf{w}; \mathbf{v} + t\mathbf{w} \rangle = ||\mathbf{v}||^2 + 2t \langle \mathbf{v}; \mathbf{w} \rangle + t^2 ||\mathbf{w}||^2, \tag{2.1}$$

with equality holding iff $\mathbf{v} = -t\mathbf{w}$, which requires $\mathbf{v}$ and $\mathbf{w}$ to be parallel vectors. We fix $\mathbf{v}$ and $\mathbf{w}$, and consider their right hand side of Equation (2.1) as a quadratic function, $p(t) = ||\mathbf{w}||^2 t^2 + 2 \langle \mathbf{v}; \mathbf{w} \rangle t + ||\mathbf{v}||^2$, of the scalar variable t. $p(t)$ assumes it minimum when $p'(t) = 2||\mathbf{w}||^2 t + 2 \langle \mathbf{v}; \mathbf{w} \rangle = 0$, so at $t = -\langle \mathbf{v}; \mathbf{w} \rangle / ||\mathbf{w}||^2$. Substituting this value for t into Equation (2.1) gives

$$0 \leq ||\mathbf{v}||^2 - 2\frac{\langle \mathbf{v}; \mathbf{w} \rangle^2}{||\mathbf{w}||^2} + \frac{\langle \mathbf{v}; \mathbf{w} \rangle^2}{||\mathbf{w}||^2} = ||\mathbf{v}||^2 - \frac{\langle \mathbf{v}; \mathbf{w} \rangle^2}{||\mathbf{w}||^2}.$$

which with rearranging becomes $\langle \mathbf{v}; \mathbf{w} \rangle^2 \leq ||\mathbf{v}||^2 ||\mathbf{w}||^2$. Taking the positive square root of both sides gives the desired inequality. □

Theorem 29. *The norm associated with an inner product satisfies the triangle inequality* $||\mathbf{v} + \mathbf{w}|| \leq ||\mathbf{v}|| + ||\mathbf{w}||$ *for every* $\mathbf{v}, \mathbf{w} \in \mathcal{V}$. *Equality holds iff* $\mathbf{v}$ *and* $\mathbf{w}$ *are parallel vectors.*

Proof.

$$\begin{aligned} ||\mathbf{v} + \mathbf{w}||^2 = \langle \mathbf{v} + \mathbf{w}; \mathbf{v} + \mathbf{w} \rangle &= ||\mathbf{v}||^2 + 2 \langle \mathbf{v}; \mathbf{w} \rangle + ||\mathbf{w}||^2 \\ &\leq ||\mathbf{v}||^2 + 2||\mathbf{v}|| \, ||\mathbf{w}|| + ||\mathbf{w}||^2 = (||\mathbf{v}|| + ||\mathbf{w}||)^2, \end{aligned}$$

using Cauchy-Schwartz inequality. Taking the positive square root of both sides gives the desired result. □

Definition 30. A *norm* on the vector space $\mathcal{V}$ assigns a real number $||\mathbf{v}||$ to each vector $\mathbf{v} \in \mathcal{V}$, subject to the following axioms for all $\mathbf{v}$, $\mathbf{w} \in \mathcal{V}$, and $c \in \mathbb{R}$:

(i) Positivity: $||\mathbf{v}|| \geq 0$, with $||\mathbf{v}|| = 0$ iff $\mathbf{v} = 0$.
(ii) Homogeneity: $||c\mathbf{v}|| = |c| ||\mathbf{v}||$.
(iii) Triangle inequality: $||\mathbf{v} + \mathbf{w}|| \leq ||\mathbf{v}|| + ||\mathbf{w}||$.

There are many different norms but the most common norms are the p-norms.

Definition 31. The general *p-norm* is defined as

$$||\mathbf{v}||_p = \sqrt[p]{\sum_{i=1}^{n} |v_i|^p}.$$

In Proposition 106 we show that, in some sense, all norms are equal in a finite-dimensional vector space. Properties (i) and (ii) are straightforward for the p-norm and property (iii) is known as *Minkowski's inequality,* but we will use $||\cdot||$ as the standard Euclidean ($p = 2$-norm) throughout.

Lemma 32. *If* $\mathbf{v} = \mathbf{0}$ *is any nonzero vector, then the vector* $\mathbf{u} = \mathbf{v}/||\mathbf{v}||$ *obtained by dividing* $\mathbf{v}$ *by its norm is a unit vector parallel to* $\mathbf{v}$.

Proof. Making use of the homogeneity property of the norm, $||\mathbf{u}|| = ||\,\mathbf{v}/||\mathbf{v}||\,|| = ||\mathbf{v}||/||\mathbf{v}|| = 1$. □

Definition 33. Let A be $m \times n$, $||\cdot||_{\hat{m}}$ be a vector norm on $\mathbb{R}^m$, and $||\cdot||_{\hat{n}}$ be a vector norm on $\mathbb{R}^n$. Then

$$||A||_{\hat{m}\hat{n}} \equiv \max_{\mathbb{R}^n \ni \mathbf{x} \neq \mathbf{0}} \frac{||A\mathbf{x}||_{\hat{m}}}{||\mathbf{x}||_{\hat{n}}} = \max_{\mathbb{R}^n \ni ||\mathbf{x}||_{\hat{n}}=1} ||A\mathbf{x}||_{\hat{m}}$$

is called an *operator norm* or *induced matrix norm* or *subordinate matrix norm.*

It is the smallest C such that $||A\mathbf{x}||_{\hat{m}} \leq C||\mathbf{x}||_{\hat{n}}$ i.e. the maximum factor by which A can stretch $\mathbf{x}$. Is usefulness comes from the behavior of a matrix as an operation from its (normed) domain and range spaces. We state the following without proof (see [Demmel, GolubVanLoan] and others.)

Lemma 34. *An operator norm is a matrix norm.*

2.4. **Orthogonality.**

Definition 35. Two elements $\mathbf{v}, \mathbf{w} \in \mathcal{V}$ of an inner product space $\mathcal{V}$ are called *orthogonal* if their inner product $\langle \mathbf{v}; \mathbf{w} \rangle = 0$.

Definition 36. A basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of a subspace $\mathcal{V}$ is called *orthogonal* if $\langle \mathbf{v}_i; \mathbf{v}_j \rangle = 0$ for all $i \neq j$. The basis is called *orthonormal* if, in addition, each vector has unit length: $||\mathbf{v}|| = 1$, for all $i = 1, \ldots, n$.

Lemma 37. *If $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is any orthogonal basis, then the normalized vectors $\mathbf{u}_i = \mathbf{v}_i/||\mathbf{v}_i||$ form an orthonormal basis.*

Proof. Follows from Lemma 32. Since $||\mathbf{v}_i|| = v_i \in \mathbb{R}$, and dividing by a scalar only affects the length of and not the orientation of the orthogonal vectors. □

Proposition 38. *If $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathcal{V}$ are nonzero, mutually orthogonal, so $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ for all $i \neq j$, then they are linearly independent.*

Proof. Suppose $c_1\mathbf{v}_1 + \ldots + c_k\mathbf{v}_k = \mathbf{0} = \mathbf{v}$. Taking any $\mathbf{v}_i$ and using that the elements are orthogonal and the linearity of inner product we get: $\langle \mathbf{v}; \mathbf{v}_i \rangle = c_1 \langle \mathbf{v}_1 : \mathbf{v}_i \rangle + \cdots + c_k \langle \mathbf{v}_k : \mathbf{v}_i \rangle = c_i||\mathbf{v}_i||^2 = 0$. Provided $\mathbf{v}_i \neq \mathbf{0}$, we conclude that the coefficient $c_i = 0$. Since this holds for all $i = 1, \ldots, k$, linear independence of $\mathbf{v}_1, \ldots, \mathbf{v}_k$ follows. □

Corollary 39. *Suppose $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathcal{V}$ are mutually orthogonal nonzero elements of an inner product space $\mathcal{V}$. Then $\mathbf{v}_1, \ldots, \mathbf{v}_k$ form an orthogonal basis for their span $\mathcal{W} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathcal{V}$, which is therefore a subspace of dimension $n = \dim W$. In particular, if $\dim \mathcal{V} = n$, then they form a orthogonal basis for $\mathcal{V}$.*

Theorem 40. *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an orthonormal basis for an inner product space $\mathcal{V}$. Then one can write any element $\mathbf{v} \in \mathcal{V}$ as a linear combination $\mathbf{v} = c_1\mathbf{u}_1 + \ldots + c_n\mathbf{u_n}$, in which the coordinates $c_i = \langle \mathbf{v}; \mathbf{u}_i \rangle$, $i = 1, \ldots, n$, are explicitly given as inner products. Moreover, the norm*

$$||\mathbf{v}|| = \sqrt{c_1^2 + \ldots + c_n^2} = \sqrt{\sum_{i=1}^{n} \langle \mathbf{v}; \mathbf{u}_i \rangle^2}$$

is the square root of the sum of the squares of its coordinates.

Proof. The orthonormality condition is $\langle \mathbf{u}_i; \mathbf{u}_j \rangle = 0$ if $i \neq j$ else $= 1$ if $i = j$ and because of bilinearity of the inner product

$$\langle \mathbf{v}; \mathbf{u} \rangle = \left\langle \sum_{j=1}^{n} c_j \mathbf{u}_j; \mathbf{u}_i \right\rangle = \sum_{j=1}^{n} c_j \langle \mathbf{u}_j; \mathbf{u}_i \rangle = c_i||\mathbf{u}_i||^2 = c_i.$$

Similarly using orthogonality of the basis elements we get

$$||\mathbf{v}||^2 = \langle \mathbf{v}; \mathbf{v} \rangle = \sum_{i,j=1}^{n} c_i c_j \langle \mathbf{u}_i; \mathbf{u}_j \rangle = \sum_{i=1}^{n} c_i^2.$$

□

Proposition 41. *If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form an orthogonal basis, then the corresponding coordinates of a vector $\mathbf{v} = a_1\mathbf{v}_1 + \ldots + a_n\mathbf{v}_n$ are given by $a_i = \langle \mathbf{v}; \mathbf{v}_i \rangle / ||\mathbf{v}_i||^2$. In this case the norm can be computed via*

$$||\mathbf{v}||^2 = \sum_{i=1}^{n} a_i^2 ||\mathbf{v}_i||^2 = \sum_{i=1}^{n} \left(\frac{\langle \mathbf{v}; \mathbf{v}_i \rangle}{||\mathbf{v}_i||} \right)^2.$$

Proof. This proof is practically identical to previous proof (Theorem 40.) □

Definition 42. A square matrix Q is called an *orthogonal matrix* if it satisfies $Q^T Q = I$. This implies that $Q^{-1} = Q$ for an orthogonal matrix.

Proposition 43. *A matrix Q is orthogonal*[3] *iff its columns form an orthonormal basis with respect to the Euclidean dot product on $\mathbb{R}^n$.*

Proof. Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be columns of Q and $\mathbf{u}_1^T, \ldots, \mathbf{u}_n^T$ the rows of Q^T. The $(i,j)^{\text{th}}$ entry of $Q^T Q = I$ is given as the product of the i^{th} row of Q^T times the j^{th} column of Q. Thus $\mathbf{u}_i \cdot \mathbf{u}_j = \mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$ which is the condition for $\mathbf{u}_1, \ldots, \mathbf{u}_n$ to form an orthonormal basis. □

Lemma 44. *An orthogonal matrix has determinant* $\det Q = \pm 1$.

Proof. From Definition 42 taking the determinant gives $1 = \det I = \det(Q^T Q) = \det Q^T \det Q = (\det Q)^2$. □

Proposition 45. *The product of two orthogonal matrices is also orthogonal.*

Proof. If $Q_1^T Q_1 = I = Q_2^T Q_2$, then $(Q_1 Q_2)^T (Q_1 Q_2) = Q_1^T Q_2^T Q_1 Q_2 = I$, and so $Q_1 Q_2$ is also orthogonal. □

Definition 46. A vector $\mathbf{z} \in \mathcal{V}$ is said to be orthogonal to the subspace $\mathcal{W}$ if it is orthogonal to every vector in $\mathcal{W}$, so $\langle \mathbf{z}; \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in \mathcal{W}$.

Note. $\mathbf{z}$ is orthogonal to $\mathcal{W}$ if it is orthogonal to every basis vector in $\mathcal{W}$.

Definition 47. The *orthogonal projection* of $\mathbf{v}$ onto the subspace $\mathcal{W}$ is the element $\mathbf{w} \in \mathcal{W}$ that makes the difference $\mathbf{z} = \mathbf{v} - \mathbf{w}$ orthogonal to $\mathcal{W}$.

Proposition 48. *Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an orthonormal basis for the subspace $\mathcal{W} \subset \mathcal{V}$. Then the orthogonal projection of a vector $\mathbf{v} \in \mathcal{V}$ onto $\mathcal{W}$ is $\mathbf{w} = c_1\mathbf{u}_1 + \ldots + c_n\mathbf{u}_n$ where $c_i = \langle \mathbf{v}; \mathbf{u}_i \rangle$, $i = 1, \ldots, n$.*

Proof. First, since $\mathbf{u}_1 \ldots, \mathbf{u}_n$ form a basis of the subspace, the orthogonal projection element $\mathbf{w} = c_1\mathbf{u}_1 + \ldots + c_n\mathbf{u}_n$ must be some linear combination thereof. Definition 47 requires that the difference $\mathbf{z} = \mathbf{v} - \mathbf{w}$ be orthogonal to $\mathcal{W}$. It suffices to check orthogonality to the basis vectors of $\mathcal{W}$. By our orthonormality assumption, for each $1 \leq i \leq n$,

$$\begin{aligned} 0 = \langle \mathbf{z}; \mathbf{u}_i \rangle \quad &= \langle \mathbf{v}; \mathbf{u}_i \rangle - \langle \mathbf{w}; \mathbf{u}_i \rangle = \langle \mathbf{v}; \mathbf{u}_i \rangle - \langle c_i\mathbf{u}_i + \ldots + c_n\mathbf{u}_n; \mathbf{u}_i \rangle \\ &= \langle \mathbf{v}; \mathbf{u}_i \rangle - c_1 \langle \mathbf{u}_1; \mathbf{u}_i \rangle - \ldots - c_n \langle \mathbf{u}_n; \mathbf{u}_i \rangle = \langle \mathbf{v}; \mathbf{u}_i \rangle - c_i. \end{aligned} \tag{2.2}$$

We deduce that the coefficients $c_i = \langle \mathbf{v}; \mathbf{u}_i \rangle$ of the orthogonal projection $\mathbf{w}$ are uniquely prescribed by the orthogonality requirement. □

[3] This definition is standard throughout linear algebra. Matrices with non-normalized orthogonal columns do not have a specific name.

Note. By the same reasoning, or by simply putting $\mathbf{u}_i = \mathbf{v}_i/||\mathbf{v}_i||$ (where $\mathbf{v}_i$ is not normalized) above, the orthogonal projection of $\mathbf{v}$ onto $\mathcal{W}$, having a general orthogonal basis $\mathbf{v}_1, \dots \mathbf{v}_n$, is given by $\mathbf{w} = a_1\mathbf{v}_i + \dots a_n\mathbf{v}_n$, where $a_i = \langle \mathbf{v}; \mathbf{v}_i \rangle / ||\mathbf{v}_i||^2$, $i = 1, \dots n$.

Definition 49. Two subspaces $\mathcal{W}, \mathcal{Z} \subset \mathcal{V}$ are called *orthogonal* if every vector in $\mathcal{W}$ is orthogonal to every vector in $\mathcal{Z}$.

Definition 50. The *orthogonal complement* to a subspace $\mathcal{W} \subset \mathcal{V}$, denoted $\mathcal{W}^\perp$ is defined as the set of all vectors which are orthogonal to $\mathcal{W}$, so $\mathcal{W}^\perp = \{\mathbf{v} \in \mathcal{W} \mid \langle \mathbf{v}; \mathbf{w} \rangle = 0 \text{ for all } \mathbf{w} \in \mathcal{W}\}$.

Proposition 51. *Suppose that $\mathcal{W} \subset \mathcal{V}$ is a finite-dimensional subspace of an inner product space. Then every vector $\mathbf{v} \in \mathcal{V}$ can be uniquely decomposed into $\mathbf{v} = \mathbf{w} + \mathbf{z}$ where $\mathbf{w} \in \mathcal{W}$ and $\mathbf{z} \in \mathcal{W}^\perp$.*

Proof. We let $\mathbf{w} \in \mathcal{W}$ be the orthogonal projection of $\mathbf{v}$ onto $\mathcal{W}$. Then $\mathbf{z} = \mathbf{v} - \mathbf{w}$ is by definition, orthogonal to $\mathcal{W}$ and hence belongs to $\mathcal{W}^\perp$. Note that $\mathbf{z}$ can be viewed as the orthogonal projection of $\mathbf{v}$ onto the complementary subspace $\mathcal{W}^\perp$. If we are given two such decompositions, $\mathbf{v} = \mathbf{w} + \mathbf{z} = \tilde{\mathbf{w}} + \tilde{\mathbf{z}}$, then $\tilde{\mathbf{w}} - \mathbf{w} = \tilde{\mathbf{z}} - \mathbf{z}$. The left hand side of this equation lies in $\mathcal{W}$ while the right hand side belongs to $\mathcal{W}^\perp$. But since they are orthogonal the only vector that can belong to both subspaces $\mathcal{W}$ and $\mathcal{W}^\perp$ is the zero vector and thus $\mathbf{w} = \tilde{\mathbf{w}}$ and $\mathbf{z} = \tilde{\mathbf{z}}$, which proves uniqueness. □

Corollary 52. *If $\mathcal{W}$ is a finite-dimensional subspace of an inner product space, then $(\mathcal{W}^\perp)^\perp = \mathcal{W}$.*

Proposition 53. *If $\dim\mathcal{W} = m$ and $\dim\mathcal{V} = n$, then $\dim\mathcal{W}^\perp = n - m$.*

Proof. This is a direct consequence of Proposition 51. Since n basis vectors span $\mathcal{V}$ in total, removing the m basis vectors that span the orthogonal subspace $\mathcal{W}$, leaves $n - m$ basis vectors orthogonal to $\mathcal{W}$, which form an orthogonal subspace on their own. □

2.5. **Systems of Linear Equations.** First we state the most basic and familiar results, found in any book on elementary linear algebra. They can also be directly inferred from the earlier discussion on linear independence.

Theorem 54. *A linear system $A\mathbf{x} = \mathbf{b}$ has a unique solution for every choice of right hand side $\mathbf{b}$ iff its coefficient matrix A is square and nonsingular.*

Theorem 55. *If A is invertible, then the unique solution to the linear system $A\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x} = A^{-1}\mathbf{b}$.*

Theorem 56. *A homogenous linear system $A\mathbf{x} = \mathbf{0}$ of m equations in n unknowns has a nontrivial solution $\mathbf{x} \neq 0$ iff the rank of A is $r < n$. If $m < n$, the system always has a nontrivial solution. If $m = n$, the system has a nontrivial solution iff A is nonsingular.*

For our purposes the use of the inverse A^{-1} is purely theoretical. $\mathbf{x} = A^{-1}\mathbf{b}$ should not be thought of as a matrix-vector multiplication or that A^{-1} is actually computed, but viewed as a change of basis (expressing $\mathbf{x}$ as a linear combination of $\mathbf{y}$) or alternatively as solving a linear system of equations. This will be used indirectly throughout this text so to clarify: since A^{-1} is defined it has full rank, i.e. has all linearly independent columns $\iff$ from a basis for the span of $A^{-1} \iff$ $\mathbf{x}$ is a linear combination $\mathbf{x} = \sum_{j=1}^{\text{rank}A} b_j \mathbf{a}_j^{(-1)}$, where $\mathbf{a}_j^{(-1)}$ signifies the j^{th} column of A^{-1}.

Definition 57. The *range* of an $m \times n$ matrix A is the subspace $\mathrm{rng}A \subset \mathbb{R}^m$ spanned by the columns of A. The *kernel* or *null space* of A is the subspace $\ker A \subset \mathbb{R}^n$ consisting of all vectors which are annihilated by A, so

$$\mathrm{rng}A = \{A\mathbf{x} \,|\, \mathbf{x} \in \mathbb{R}^n\} \subset \mathbb{R}^m \quad \text{and} \quad \ker A = \{\mathbf{z} \in \mathbb{R}^n \,|\, A\mathbf{z} = \mathbf{0}\} \subset \mathbb{R}^n.$$

Alternative names for the range are *image* and *column space*, as by definition a vector $\mathbb{R}^m \ni \mathbf{b} \in \mathrm{rng}A$ iff it can be written as a linear combination of the columns of $A = (\mathbf{a}_1\, \mathbf{a}_2\, \ldots\, \mathbf{a}_n)$ i.e. $\mathbf{b} = x_1\mathbf{a}_1 + \ldots x_n\mathbf{a}_n$ and so $\mathbf{b} = A\mathbf{x}$ for some $\mathbf{x}$, meaning that *a vector* $\mathbf{b}$ *lies in the range of* A *iff the linear system* $A\mathbf{x} = \mathbf{b}$ *has a solution.*

An alternatives name for the kernel is the *null space*, as $\ker A$ is the set of solutions to the homogenous system $A\mathbf{z} = \mathbf{0}$. Suppose that $\mathbf{z}, \mathbf{w} \in \ker A$ so that $A\mathbf{z} = \mathbf{0} = A\mathbf{w}$. Then for any $c, d \in \mathbb{R}$: $A(c\mathbf{z} + d\mathbf{w}) = cA\mathbf{z} + dA\mathbf{w} = \mathbf{0} \in \ker A$, and so $\ker A$ is a subspace. This is known as the *superposition principle* for solutions to homogenous linear system of equations.

Proposition 58. *If* $\mathbf{z}_1, \ldots, \mathbf{z}_k \in \ker A$ *(are solutions to* $A\mathbf{z} = \mathbf{0}$*), then so are* $c_1\mathbf{z}_1 + \ldots + c_k\mathbf{z}_k \in \ker A$.

Proof. $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \ker A \Leftrightarrow A\mathbf{z}_1 = \ldots = A\mathbf{z}_n = \mathbf{0}$ so $c_1A\mathbf{z}_1 = \ldots = c_nA\mathbf{z}_n = \mathbf{0} \in \ker A$, which also shows that $\ker A$ is a subspace ($c_kA\mathbf{z}_k + c_iA\mathbf{z}_i \in \ker A$.) □

Note. The set $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of solutions to inhomogenous $A\mathbf{x} = \mathbf{b}$, $\mathbf{b} \neq \mathbf{0}$, is *not* a subspace (it would not contain $\mathbf{x} = \mathbf{0}$.)

Theorem 59. *The linear system* $A\mathbf{x} = \mathbf{b}$ *has a solution* $\mathbf{x}^*$ *iff* $\mathbf{b} \in \mathrm{rng}A$. *If this occurs, then* $\mathbf{x}$ *is a solution to the linear system iff*

$$\mathbf{x} = \mathbf{x}^* + \mathbf{z},$$

where $\mathbf{z} \in \ker A$ *is any element in the kernel of* A.

Proof. The first part follows from Definition 57. If $A\mathbf{x} = A\mathbf{x}^* = \mathbf{b}$, their difference $\mathbf{z} = \mathbf{x} - \mathbf{x}^*$ satisfies $A\mathbf{z} = A(\mathbf{x} - \mathbf{x}^*) = A\mathbf{x} - A\mathbf{x}^* = \mathbf{b} - \mathbf{b} = \mathbf{0}$ and $\mathbf{z} \in \ker A$ and $\mathbf{x} = \mathbf{x}^* + \mathbf{z}$ follows. □

In order to find the general solution to the system one needs to find a particular solution $\mathbf{x}$ and the general solution $\mathbf{z} \in \ker A$ to homogenous equation (as in the case of linear ordinary differential equations).

Definition 60. The *adjoint* to a linear system $A\mathbf{x} = \mathbf{b}$ of m equations in n unknowns is the linear system

$$A^T\mathbf{y} = \mathbf{f}$$

of n equations in m unknowns. Here $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{f} \in \mathbb{R}^n$.

Definition 61. The *corange* (or alternatively *row space* or *coimage*) of an $m \times n$ matrix A is the range of its transpose,

$$\mathrm{corng}A = \mathrm{rng}A^T = \{A^T\mathbf{y} \,|\, \mathbf{y} \in \mathbb{R}^m\} \subset \mathbb{R}^n.$$

The *cokernel* or *left null space* of A is the kernel of its transpose,

$$\mathrm{coker}A = \ker A^T = \{\mathbf{w} \in \mathbb{R}^n \,|\, A^T\mathbf{w} = \mathbf{0}\} \subset \mathbb{R}^m,$$

That is, the set of solutions to the homogenous adjoint system.

The following ([OlvShaDraft] Theorem 2.47) is the *Fundamental Theorem of Linear Algebra*, found in any elementary linear algebra text.

Theorem 62. *Let* A *be a* $m \times n$ *matrix of rank* r. *Then*

$$\dim \mathrm{corng}A = \dim \mathrm{rng}A = \mathrm{rank}A = \mathrm{rank}A^T = r,$$
$$\dim \ker A = n - r, \qquad \dim \mathrm{coker}A = m - r.$$

Proof. Briefly. The rank r of A is the #linearly independent columns=#linearly independent rows= $\dim A$ = #pivots. The linearly independent rows of A are the linearly independent columns of A^T and so $\text{rank}A^T = r$ = #pivots. It follows that #linearly dependent columns of $A = n - r = \dim \ker A$. Since $\text{coker}A = \ker A^T$ (A^T is $n \times m$) similarly #linearly independent columns of $A^T = m - r$. □

Theorem 63. *Let A be an $m \times n$ matrix of rank r. Then its kernel and corange are orthogonal complements as subspaces of $\mathbb{R}^n$, of respective dimension $n - r$ and r, while its cokernel and range are orthogonal complements in $\mathbb{R}^m$, of respective dimensions $m - r$ and r:*

$$\ker A = (\text{corng}A)^\perp \subset \mathbb{R}^n, \quad \text{coker}A = (\text{rng}A)^\perp \subset \mathbb{R}^m. \tag{2.3}$$

Proof. A vector $\mathbf{x} \in \mathbb{R}^n$ lies in $\ker A$ iff $A\mathbf{x} = \mathbf{0}$. According to the rules of matrix multiplication, the i^{th} entry of $A\mathbf{x}$ equals the product of the i^{th} row $\mathbf{r}_i^T$ of A and $\mathbf{x}$. But this product vanishes, $\mathbf{r}_i^T\mathbf{x} = \mathbf{r}_i \cdot \mathbf{x} = 0$, iff $\mathbf{x}$ is orthogonal to $\mathbf{r}_i$. Therefore $\mathbf{x} \in \ker A$ iff $\mathbf{x}$ is orthogonal to all the rows of A. Since the rows span $\text{corng}A = \text{rng}A^T$, this is equivalent to the statement that $\mathbf{x}$ lies in the orthogonal complement $(\text{corng}A)^\perp$, which proves the first statement. The proof of range and cokernel follows the same argument applied to the transposed matrix A^T. □

A linear system $A\mathbf{x} = \mathbf{b}$ will have a solution iff the right hand side $\mathbf{b} \in \text{rng}A$ which requires $\mathbf{b} \perp \text{coker}A$, and we can write the compatibility conditions for $A\mathbf{x} = \mathbf{b}$ as $\mathbf{y} \cdot \mathbf{b} = 0$ for any $\mathbf{y}$ satisfying $A^T\mathbf{y} = \mathbf{0}$. Following [OlvShaDraft] we state the following characterization of compatible linear systems, without proof, but is actually a combination of Theorem 62 and Theorem 63.

Theorem 64. *(Fredholm* alternative*) The linear system $A\mathbf{x} = \mathbf{b}$ has a solution iff $\mathbf{b}$ is orthogonal to the cokernel of A.*

We state the following theorem without proof.

Proposition 65. *Multiplication by an $m \times n$ matrix A of rank r defines a one-to-one correspondence between the r-dimensional subspace $\text{corng}A \subset \mathbb{R}^n$ and $\text{rng}A \subset \mathbb{R}^m$. Moreover, if $\mathbf{v}_1, \ldots, \mathbf{v}_r$ forms a basis of $\text{corng}A$ then their images $A\mathbf{v}_1, \ldots, A\mathbf{v}_n$ form a basis for $\text{rng}A$.*

Proposition 66. *A compatible linear system $A\mathbf{v} = \mathbf{b}$ with $\mathbf{b} \in \text{rng}A = (\text{coker}A)^\perp$ has a unique solution $\mathbf{w} \in \text{corng}A$ with $A\mathbf{w} = \mathbf{b}$. The general solution is $\mathbf{x} = \mathbf{w} + \mathbf{z}$ where $\mathbf{z} \in \ker A$. The particular solution is distinguished by the fact that it has minimum Euclidean norm $||\mathbf{w}||$ among possible solutions.*

We will briefly return to these in Section 3.3.3 where these results will become clear.

2.6. **Positive Definite Matrices.**

Definition 67. An $n \times n$ matrix K is called *symmetric positive definite* - s.p.d - if it is symmetric, $K^T = K$, and satisfies the positivity condition $\mathbf{x}^T K\mathbf{x} > 0$ for all $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n$.[4]

Theorem 68. *Every inner product on $\mathbb{R}^n$ is given by $\langle \mathbf{x}; \mathbf{y} \rangle = \mathbf{x}^T K\mathbf{y}$, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, where K is s.p.d.*

Proof. Let $\langle \mathbf{x}; \mathbf{y} \rangle$ denote the inner product between the vectors $\mathbf{x} = (x_1 \ \ldots \ x_n)^T$ and $\mathbf{y} = (y_1 \ \ldots \ y_n)^T$, in $\mathbb{R}^n$. Writing the vectors in terms of the standard basis

[4]This is sometimes written $K > 0$, but does not imply that all entries are > 0.

$\mathbf{x} = x_1\mathbf{e}_1 + \ldots + x_n\mathbf{e}_n = \sum_{i=1}^n x_i\mathbf{e}_i$, $\mathbf{y} = \sum_{j=1}^n y_j\mathbf{e}_i$. Bilinearity of the inner product gives

$$\langle \mathbf{x}, \mathbf{y} \rangle = \left\langle \sum_{i=1}^n x_i\mathbf{e}_i; \sum_{j=1}^n y_j\mathbf{e}_j \right\rangle = \sum_{i,j=1}^n x_i x_j \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \sum_{i,j=1}^n k_{ij} x_i y_i = \mathbf{x}^T K \mathbf{y}, \tag{2.4}$$

where K is the $n \times n$ matrix of inner products of the basis vectors, with entries $k_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle$, $i, j = 1, \ldots, n$. So any inner product can/must be expressed in the general *bilinear form*.

Symmetry of the inner product implies that $k_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \langle \mathbf{e}_j, \mathbf{e}_i \rangle = k_{ji}$, $i, j = 1, \ldots, n$. Consequently, the inner product matrix K is symmetric with $K = K^T$. Since $\langle \mathbf{x}; \mathbf{y} \rangle = \mathbf{x}^T K \mathbf{y} =$ [since scalar] $= (\mathbf{x}^T K \mathbf{y})^T = \mathbf{y}^T K^T \mathbf{x} = \mathbf{y}^T K \mathbf{x} = \langle \mathbf{y}; \mathbf{x} \rangle$, symmetry of K ensures the bilinear form is also symmetric.

Finally $||\mathbf{x}||^2 = \langle \mathbf{x}; \mathbf{x} \rangle = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^n k_{ij} x_i x_j \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, and equality only iff $\mathbf{x} = \mathbf{0}$. □

Proposition 69. *All s.p.d matrices K are non-singular.*

Proof. For the $\mathbf{x}^T(K\mathbf{x}) = (K^T\mathbf{x})^T\mathbf{x} > 0$ to hold (for K to be s.p.d.) $\mathrm{rng} K \ni \mathbf{x} \neq \mathbf{0} \in \mathrm{corng} K$ and only $\{\mathbf{0}\} = \ker A$ (i.e. $\dim \ker K = 0$) and so square K has no linearly independent columns and is invertible. More succinctly: if $\mathbf{x}^T A \mathbf{x} = \mathbf{0}$ we could find a nonzero $\mathbf{x} \neq \mathbf{0}$ to satisfy the equation (we would have linear dependent columns/rows.) □

Definition 70. Let $\mathcal{V}$ be an inner product space, and let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathcal{V}$. The associated *Gram matrix*

$$K = \begin{pmatrix} \langle \mathbf{v}_1; \mathbf{v}_1 \rangle & \cdots & \langle \mathbf{v}_1; \mathbf{v}_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{v}_n; \mathbf{v}_1 \rangle & \cdots & \langle \mathbf{v}_n; \mathbf{v}_n \rangle \end{pmatrix} \tag{2.5}$$

is the the $n \times n$ matrix whose entries are the inner products between the chosen vector space elements.

Proposition 71. *All Gram matrices are positive semi-definite. A Gram matrix is s.p.d. iff the elements $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathcal{V}$ are linearly independent.*

Proof. To prove (semi-)definiteness of K, we need to examine the associated quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^n k_{ij} x_i x_j.$$

Symmetry of the inner product implies symmetry of Gram matrix so $k_{ij} = \langle \mathbf{v}_i; \mathbf{v}_j \rangle = \langle \mathbf{v}_j; \mathbf{v}_i \rangle = k_{ji}$, and hence $K^T = K$. Substituting this into the above we get

$$q(\mathbf{x}) = \sum_{i,j=1}^n \langle \mathbf{v}_i; \mathbf{v}_j \rangle x_i x_j.$$

Bilinearity of the inner product of $\mathcal{V}$ implies that we can assemble this summation into a single inner product

$$q(\mathbf{x}) = \left\langle \sum_{i=1}^n x_i \mathbf{v}_i; \sum_{j=1}^n x_j \mathbf{v}_j \right\rangle = \langle \mathbf{v}; \mathbf{v} \rangle = ||\mathbf{v}||^2 \geq 0,$$

where $\mathbf{v} = x_1\mathbf{v}_1 + \ldots + x_n\mathbf{v}_n \in \mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n)$, so K is positive semi-definite. Moreover, $q(\mathbf{x}) = ||\mathbf{v}||^2 > 0$ as long as $\mathbf{v} \neq \mathbf{0}$. If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent then $\mathbf{v} = \mathbf{0}$ iff $x_1 = \cdots = x_n = 0$, and hence, in this case, $q(\mathbf{x})$ and K are s.p.d. □

Given vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^m$ form the $m \times n$ matrix $A = (\mathbf{v}_1 \ \ldots \ \mathbf{v_n})$. The Euclidean inner product (dot product) $\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T\mathbf{w}$ gives that $k_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j = \mathbf{v}_i^T\mathbf{v}_j$ of the i^{th} row of A^T with the j^{th} column of A i.e.

$$K = A^T A, \tag{2.6}$$

which by Proposition 71 is s.p.d. iff the columns of A are linearly independent.

Theorem 72. *Given an $m \times n$ matrix A, the following are equivalent:*
(i) The $n \times n$ Gram matrix $K = A^T A$ is positive definite.
(ii) A has linearly independent columns.
(iii) $\text{rank}A = n \leq m$.
(iv) $\ker A = \{0\}$.

3. Factorizations

We now turn to the problem of matrix factorizations, one of the most important tools of linear algebra. The idea is to reduce a matrix into parts that are either easier to solve, or display some important property of the matrix (for example number of pivots, rank, invertibility, *singular values*, *eigenvalues* etc.) We will be using the notations and definitions from Section 2.1 extensively.

3.1. Gauss Reduction and LU Factorization.

Gauss transformations. If $M, A \in \mathbb{R}^{m\times m}$ we can express the matrix-matrix product MA as matrix-vector products $MA = [M\mathbf{a}_1 \cdots M\mathbf{a}_m]$, where $\mathbf{a}_k$ is the k^{th} columns of A, and $M\mathbf{a}_k$ forms the k^{th} column of MA. Now suppose $\mathbf{a}_k \in \mathbb{R}^m$ with $\mathbf{a}_k \ni a_{k,k} \neq 0$ (the k^{th} element of $\mathbf{a}_k$.) Let $\mathbf{l}_k^T = (\underbrace{0, \ldots, 0}_{k}, l_{k+1}, \ldots, l_m)$, $l_i = a_{k,i}/a_{k,k}$, $i = k+1, \ldots, m$ and $a_{k,i}$ is the i^{th} element of $\mathbf{a}_k$. We will call $\mathbf{l}_k$ a *Gauss vector* with *multipliers* l_i, and define the *Gauss transform* as $M(\mathbf{l}_k) = M_k = I - \mathbf{l}_k\mathbf{e}_k^T \in \mathbb{R}^{m\times m}$, where $\mathbf{e}_k \in \mathbb{R}^m$ is the k^{th} unit vector ($e_{i=k} = 1$, $e_{i\neq k} = 0$.)[5]

On applying a Gauss transform M_k to A, the k^{th} column of the resulting M_kA becomes

$$M_k\mathbf{a}_k = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 \\ 0 & \ddots & \vdots & & \\ & & 1 & & \vdots \\ \vdots & & -l_{k+1} & & \\ & & \vdots & \ddots & 0 \\ 0 & \cdots & l_m & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_k \\ a_{k+1} \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Note 73. If we apply M_k to a $m \times m$ matrix B we get $M_kB = (I - \mathbf{l}_k\mathbf{e}_k^T)B = B - \mathbf{l}_k(\mathbf{e}_k^T B) = B - \mathbf{l}_k B_{k,1:m} = B - \tilde{B}$, (where $B_{k,1:m}$ is the k^{th} row of B.) Since $l_{1:k} = 0$ in $\mathbf{l}_k$ (i.e. the first k elements of $\mathbf{l}_k$ are zero) we get $\tilde{B} = \begin{bmatrix} 0_{1:k,1:m} \\ \tilde{B}_{k+1:m,1:m} \end{bmatrix}$. Only the submatrix $B_{k+1:m,1:m}$ is affected, and the application will leave "subcolumn"

[5]The condition that $l_i = 0$ (for $i = 1, \ldots, k$) is required for M_k to be a Gauss transform.

$B_{k+1:m,1} = \mathbf{0}$. For example, if $m = 5$ and $k = 3$ we get

$$\mathbf{l}_3\mathbf{e}_3^T B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -l_4 \\ -l_5 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ B_{3,1} & B_{3,2} & B_{3,3} & B_{3,4} & B_{3,5} \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 0 \\ -l_4 \\ -l_5 \end{bmatrix} \begin{bmatrix} B_{3,1} & B_{3,2} & B_{3,3} & B_{3,4} & B_{3,5} \end{bmatrix} = \mathbf{l}_k\mathbf{b}^T =$$

$$\tilde{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -l_4B_{3,1} & -l_4B_{3,2} & -l_4B_{3,3} & -l_4B_{3,4} & -l_4B_{3,5} \\ -l_5B_{3,1} & -l_5B_{3,2} & -l_5B_{3,3} & -l_5B_{3,4} & -l_5B_{3,5} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \bullet & \bullet & B_{4,3}B_{3,3}/B_{3,3} & \bullet & \bullet \\ \bullet & \bullet & B_{5,3}B_{3,3}/B_{3,3} & \bullet & \bullet \end{bmatrix}$$

and, observing that only elements in $B_{4:5,1}$ (the subdiagonal column) coincide especially with those of B, we get

$$B - \tilde{B} = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix} = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & 0 & \bullet & \bullet \\ \bullet & \bullet & 0 & \bullet & \bullet \end{bmatrix}.$$

Further, if for example $B_{5,4} = 0$ then after application of $\tilde{B}$ it will be $-l_5B_{3,4}$, and we have "destroyed" a zero. Looking at the second step, $\mathbf{l}_k\mathbf{b}^T$, we see that zero elements in $\mathbf{b}^T$ will introduce zero columns in $\tilde{B}$. We use this fact in the following:

Definition 74. Let $M_k = M(\mathbf{l}_k)$ where $\mathbf{l}_k = \mathbf{l}(A_{k-1})$. Let A_k be the matrix A after k applications $M_1 \cdots M_k$. *Gauss reduction* (or *upper triangularization*) is the process of successively applying such a sequence of M_k $(k = 1, \ldots, m-1)$ to zero the subdiagonal and reduce A to row echelon form.

If A is $m \times m$, the Gauss reduction will need at most $m - 1$ steps since in the last column the pivot is all that remains. Also, returning to Note 73 we find, in step $k \neq 1$, that since the subdiagonal entries in columns $1, \ldots, k-1$ are zero, the corresponding columns of $\tilde{B}$ will be zero, and only the lower right rectangular corner of B is affected i.e. B_{22}:

$$B - \tilde{B} = \begin{bmatrix} B_{1:k,1:k-1} & B_{1:k,k:m} \\ B_{k+1:m,1:k-1} & B_{22} \end{bmatrix} - \begin{bmatrix} 0_{1:k,1:k-1} & 0_{1:k,k:m} \\ 0_{k+1:m,1:k-1} & \tilde{B}_{k+1:m,k:m} \end{bmatrix}.$$

Illustrating Gauss reduction with an arbitrary 4×4 matrix A we see that

$$
\begin{aligned}
A \underset{(1)}{\overset{[M_1]}{\rightarrow}} & \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ a & \bullet & \bullet & \bullet \\ b & \bullet & \bullet & \bullet \\ c & \bullet & \bullet & \bullet \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 \\ a & \bullet & \bullet & \bullet \\ b & \bullet & \bullet & \bullet \\ c & \bullet & \bullet & \bullet \end{bmatrix} \\
\underset{(2)}{\overset{[M_2]}{\rightarrow}} & \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & d & \bullet & \bullet \\ 0 & e & \bullet & \bullet \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & d & \bullet & \bullet \\ 0 & e & \bullet & \bullet \end{bmatrix} \\
\underset{(3)}{\overset{[M_3]}{\rightarrow}} & \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & f & \bullet \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & f & \bullet \end{bmatrix} \\
= & \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & 0 & \bullet \end{bmatrix} = U
\end{aligned}
$$

Other possibilities for reducing A to row echelon from are possible, but some properties make the Gauss transform, and the resulting Gauss reduction, particularly useful.

Proposition 75. *If M and $\tilde{M}$ are Gauss transforms, then*
(1) M is nonsingular and its inverse M^{-1} is also unit lower triangular,
(2) M^{-1} is equal to M with its subdiagonal elements of opposite sign,
(3) $M\hat{M}$ is also unit lower triangular.

Proof. (1) and (2): The pattern of mostly zeroes (*sparsity*[6]), in the general case, in $\mathbf{l}_k$ and $\mathbf{e}_k$ implies that $\mathbf{e}_k^T\mathbf{l}_k = 0$, because $e_k = 1$ and the rest zero, but $l_k = 0$ (the k^{th} entry of $\mathbf{l}_k$.) Therefore $(I-\mathbf{l}_k\mathbf{e}_k^T)(I+\mathbf{l}_k\mathbf{e}_k^T) = I-\tau_k\mathbf{e}_k^T\mathbf{l}_k\mathbf{e}_k^T = I-\mathbf{l}(\mathbf{e}_k^T\tau)\mathbf{e}_k^T = I$ and so $M_k^{-1} = I+\mathbf{l}_k\mathbf{e}_k^T(= L_k)$, and we can easily determine the inverse (which implies non-singularity.)
(3) Sparsity again gives $\mathbf{e}_k^T\mathbf{l}_{k+1} = 0$, $M_kM_{k+1} = (I+\mathbf{l}_k\mathbf{e}_k^T)(I+\mathbf{l}_{k+1}\mathbf{e}_{k+1}^T) = I+\mathbf{l}_k\mathbf{e}_k^T+\mathbf{l}_{k+1}\mathbf{e}_{k+1}^T$, a unit lower triangular matrix is obtained. Putting $M_kM_{k+1} = M$ and multiplying with $M_n = I+\mathbf{l}_n\mathbf{e}_n^T$ will give $MM_n = I+\mathbf{l}_k\mathbf{e}_k^T+\mathbf{l}_{k+1}\mathbf{e}_{k+1}^T+\mathbf{l}_n\mathbf{e}_n^T$, which is also unit lower triangular. □

We change the notation to $M_k = L_k^{-1}$ and $M_k^{-1} = L_k$ for the k^{th} Gauss transform and its inverse. So by choosing L_k^{-1} properly it is usually possible to zero the subdiagonal elements in column k of the matrix A, and under the right circumstances (for example we required $a_{k,k} \neq 0$) one can find a sequence of Gauss transforms $L_1^{-1}, \ldots, L_{m-1}^{-1}$ such that $L_{m-1}^{-1}\cdots L_1^{-1}A = U$ is upper triangular. This is the idea behind LU factorization.

LU factorization. LU factorization is the result of a complete Gauss reduction of a nonsingular matrix A, resulting in a upper triangular matrix U and unit lower triangular matrix L, such that $A = LU$.

We just saw that applying the L_1^{-1} to L_{m-1}^{-1} Gauss transformations successively will give $L_{m-1}^{-1}\ldots L_1^{-1}A = U$, where U is upper triangular. Setting $L_{m-1}^{-1}\cdots L_1^{-1} = L^{-1}$, where L^{-1} is invertible and unit lower triangular (Proposition 75) we get that

[6]Sparsity is a very important concept to making practical linear algebra practical. The opposite is "full", making no assumptions on the distribution or existence of zero entries.

$L^{-1}A = U$. Then $L = (L_{m-1}^{-1} \cdots L_1^{-1})^{-1} = L_1 \cdots L_{m-1}$ (also unit lower triangular) and we get $A = LU$. This is the LU factorization of A.

Proposition 76. *The following statements are equivalent:*

1. There exists a unique unit lower triangular L and nonsingular upper triangular U such that $A = LU$.

2. All leading principal submatrices of A are nonsingular.

Proof. (1) $\implies$ (2): $A = LU$ may also be written

$$\begin{aligned}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} &= \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} \\ &= \begin{bmatrix} L_{11}U_{11} & L_{11}U_{12} \\ L_{21}U_{11} & L_{21}U_{12} + L_{22}U_{22} \end{bmatrix},\end{aligned}$$

where A_{11} is a j-by-j leading principal submatrix, as are L_{11} and U_{11}. Therefore $\det A_{11} = \det(L_{11}U_{11}) = \det L_{11} \det U_{11} = 1 \cdot \prod_k^j (U_{11})_{kk} \neq 0$[7], since L is unit triangular and U is triangular.

(2) $\implies$ (1): Using induction on m: in the basic case, for 1×1 matrices: $a = 1 \cdot a$. To prove for $m \times m$ matrices $\tilde{A}$ we will find unique $(m-1) \times (m-1)$ triangular matrices L and U ($LU = A \in \mathbb{R}^{(m-1)\times(m-1)}$), unique $(m-1) \times 1$ vectors $\mathbf{b}$, $\mathbf{c}$, $\mathbf{l}$ and $\mathbf{u}$, and unique scalars $\delta, \eta \neq 0$ such that

$$\tilde{A} = \begin{bmatrix} A & \mathbf{b} \\ \mathbf{c}^T & \delta \end{bmatrix} = \tilde{L}\tilde{U} = \begin{bmatrix} L & 0 \\ \mathbf{l}^T & 1 \end{bmatrix} \begin{bmatrix} U & \mathbf{u} \\ 0 & \eta \end{bmatrix} = \begin{bmatrix} LU & L\mathbf{u} \\ \mathbf{l}^T U & \mathbf{l}^T\mathbf{u} + \eta \end{bmatrix}.$$

By our induction assumption, unique L and U exist such that $A = LU$. Now let $\mathbf{u} = L^{-1}\mathbf{b}$, $\mathbf{l}^T = \mathbf{c}^T U^{-1}$, and $\eta = \delta - \mathbf{l}^T\mathbf{u}$, all of which are unique. The diagonal entries of U are nonzero by induction (and those of L are 1), and $\eta \neq 0$ since $0 \neq \det(\tilde{A}) = \det(U) \cdot \eta$. If either of L and U are singular (though $\tilde{A}$ is not) the LU factorization fails. But by induction, since LU held for $(m-1)$, $\tilde{L}\tilde{U}$ holds for m. □

As we just saw, though A may be nonsingular, submatrices may not be and then LU factorization fails. Reordering (*permuting*) the components of A, if A is nonsingular, we can get nonsingular leading principal submatrices as required.

Proposition 77. *If A is nonsingular, then there exists permutations P_1 and P_2, a nonsingular unit lower triangular matrix L and nonsingular upper triangular matrix U, such that $P_1AP_2 = LU$. Only one of P_1 and P_2 are necessary.*

Proof. We use induction on dimension m. For 1×1 matrices: $P_1 = P_2 = L = 1$ and $U = A$. Assume that it is true for dimension $n-1$. If A is nonsingular, then it has a nonzero entry; choose permutations P_1' and P_2' so that the $(1,1)$ entry of $P_1'AP_2'$ is nonzero. (Only one of P_1' and P_2' is needed since nonsingularity implies that each row and column has nonzero entry.)

Now we write the desired factorization and solve for the unknown components:

$$\begin{aligned}\begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix} \\ &= \begin{bmatrix} u_{11} & U_{12} \\ L_{21}u_{11} & L_{21}U_{12} + \tilde{A}_{22} \end{bmatrix},\end{aligned}$$

where A_{22} and $\tilde{A}_{22}$ are $(n-1)\times(n-1)$, and L_{21} and U_{12}^T are $(n-1)\times 1$. Solving for the components of this 2×2 block factorization we get $u_{11} - a_{11} = 0$ ($u_{11} = a_{11} \neq 0$), $U_{12} = A_{12}$, and $L_{21}u_{11} = A_{21}$. Since $u_{11} = a_{11} \neq 0$, we can solve for $L_{21} = A_{21}/a_{11}$. Finally, $L_{21}U_{12} + \tilde{A}_{22} = A_{22}$ implies $\tilde{A}_{22} = A_{22} - L_{21}U_{12}$.

[7] This is the product of the *traces* of L and U.

We want to apply induction to $\tilde{A}_{22}$, but to do so we need to check that $\det\tilde{A}_{22} \neq 0$: Since $\det P_1' A P_2' = \pm\det A \neq 0$ and also

$$\det P_1' A P_2' = \det \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \cdot \det \begin{bmatrix} u_{11} & U_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix} = 1 \cdot (u_{11} \cdot \det\tilde{A}_{22}),$$

then $\det\tilde{A}_{22}$ must be nonzero.

Therefore, by induction there exist permutations $\tilde{P}_1$ and $\tilde{P}_2$ so that $\tilde{P}_1\tilde{A}_{22}\tilde{P}_2 = \tilde{L}\tilde{U}$, with $\tilde{L}$ unit lower triangular and $\tilde{U}$ upper triangular and nonsingular. Substituting this in the above 2×2 block factorization yields

$$\begin{aligned} P_1' A P_2' &= \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & \tilde{P}_1^T\tilde{L}\tilde{U}\tilde{P}_2^T \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_1^T\tilde{L} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & \tilde{U}\tilde{P}_2^T \end{bmatrix}, \\ &= \begin{bmatrix} 1 & 0 \\ L_{21} & \tilde{P}_1^T\tilde{L} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12}\tilde{P}_2 \\ 0 & \tilde{U} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_2^T \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_1^T \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \tilde{P}_1 L_{21} & \tilde{L} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12}\tilde{P}_2 \\ 0 & \tilde{U} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_2^T \end{bmatrix} \end{aligned}$$

so we get the desired factorization of A:

$$\begin{aligned} P_1 A P_2 &= \left(\begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_1 \end{bmatrix} P_1' \right) A \left(P_2' \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_2 \end{bmatrix} \right) \\ &= \begin{bmatrix} 1 & 0 \\ \tilde{P}_1 L_{21} & \tilde{L} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12}\tilde{P}_2 \\ 0 & \tilde{U} \end{bmatrix}. \end{aligned}$$

□

Corollary 78. *We can choose P_2' and P_1' so that a_{11} is the largest entry in absolute value in the whole matrix. More generally, at step i of Gaussian elimination, where we are computing the i^{th} column of L, we reorder the rows and columns i through n so that the largest entry in this submatrix is on the diagonal. This is called* Gaussian elimination with complete pivoting - *GECP.*

We can choose $P_2' = I$ and P_1' so that a_{11} is the largest entry in absolute value in the column. More generally, at step i of Gaussian elimination, where we are computing the i^{th} column of L, we reorder the rows so that the largest entry in the column is on the diagonal. This is called Gaussian elimination with partial pivoting - *GEPP.*[8]

Summarizing we get the following:

Theorem 79. *Let A be a $m \times m$ matrix. Then the following conditions are equivalent (i) A is nonsingular; (ii) A has m nonzero pivots; (iii) A admits a permuted LU factorization: $PA = LU$; (iv) once $B\,(= PA)$ admits a LU factorization, it is unique (for every P.)*

Just as when forming the L of the LU factorization, we do not have the resulting matrices P_{row} or P_{column}, the row or column ordering, beforehand. Rather any permutation P is a consequence of Corollary 78, where at each step we do an (hopefully) sufficient pivot P_i (the resulting leading principal submatrix is nonsingular) for that, step and P_{row} is the resulting permutation matrix after we are done with the LU factorization. To see clearly how this comes about, we illustrate if for GEPP.

[8]GECP is almost never used in practice. GEPP is (almost) always sufficient in practice because GEPP almost always works. ([Demmel])

By Corollary 78, at each step $A \in \mathbb{R}^{(m+1)\times(m+1)}$ is first permuted by P_1 so that the maximum magnitude element is found in the diagonal and then the Gauss transform L_1^{-1} is applied. In the next step this is repeated and we get a sequence $L_m^{-1}P_m \cdots L_1^{-1}P_1A = U$. If we put $L_m^{-1} = L'_m$, $L'_{m-1} = P_mL_{m-1}^{-1}P_m^{-1}$, $L'_{m-2} = P_mP_{m-1}L_{m-2}^{-1}P_{m-1}^{-1}P_m^{-1}$, with the general $L'_k = P_m \dots P_{k+1}L_k^{-1}P_{k+1}^{-1} \dots P_m^{-1}$ and the final $L'_1 = P_m \dots P_2L_1^{-1}P_2^{-1} \dots P_m^{-1}$.

We see that L'_k and L_k^{-1} are similar and will have the same structure. When we show what happens for $m = 3$ it is a trivial mater to extend this result to the general $L'_m \dots L'_1P_m \cdots P_1 = L_m^{-1}P_m \cdots L_1^{-1}P_1 = L'P$, where $L' = (L'_m \cdots L'_1)$ which we already know to be L^{-1}.

$$\begin{aligned} L'P_{row} &= L'_3L'_2L'_1P_3P_2P_1 = L_3^{-1}(P_3L_2^{-1}P_3^{-1})(P_3P_2L_1^{-1}P_2^{-1}P_3^{-1})P_3P_2P_1 \\ &= L_3^{-1}P_3L_2^{-1}P_2L_1^{-1}P_1. \end{aligned}$$

3.2. **Cholesky Factorization.**

Proposition 80. *If all leading principal submatrices of $A \in \mathbb{R}^{n\times n}$ are nonsingular, then there exists unique lower triangular matrices L and M and a unique diagonal matrix $D = \operatorname{diag}(d_1, \dots, d_n)$ such that $A = LDM^T$.*

Proof. We know that A has a LU factorization $A = LU$. Set $D = \operatorname{diag}(d_1, \dots, d_n)$ with $d_i = u_{ii}$ for $i = [1, n]$. D is nonsingular and $M^T = D^{-1}U$ is unit upper triangular. Thus $A = LU = LD(D^{-1}U) = LDM^T$. Uniqueness follows from the uniqueness of LU factorization. □

Theorem 81. *If $A = LDM^T$ is the LDM^T factorization of a nonsingular symmetric matrix A, then $L = M$.*

Proof. The matrix $M^{-1}AM^{-T} = MLD$ is symmetric and lower triangular, i.e. diagonal. Since D is nonsingular, this implies that $M^{-1}L$ is also diagonal, but $M^{-1}L$ is unit lower triangular and so $M^{-1}L = I$. □

Proposition 82. *If X is nonsingular, then A is s.p.d. iff X^TAX is s.p.d.*

Proof. X nonsingular implies $\mathbf{y} = X\mathbf{x} \neq \mathbf{0}$ for all $\mathbf{x} \neq \mathbf{0}$. So $0 < \mathbf{x}^TX^TAX\mathbf{x} = (X\mathbf{x})^TAX\mathbf{x} = \mathbf{y}^TA\mathbf{y} > 0$. Since this was true for all $\mathbf{x} \neq \mathbf{0}$ (and hence $\mathbf{y} \neq \mathbf{0}$) this implies A is s.p.d. when X^TAX is s.p.d. □

Proposition 83. *If X is nonsingular, then X^TAX is s.p.d. if A is s.p.d.*

Proof. This is simply the reversal of Lemma 82. Beginning with A s.p.d and nonsingular X. For $\mathbf{x} \neq \mathbf{0} \Leftrightarrow \mathbf{y} = X\mathbf{x} \neq \mathbf{0}$, and $0 < \mathbf{y}^TA\mathbf{y}$ gives $X^TAX = K > 0$. □

Note. We used X nonsingular to ensure that $\mathbf{x} \neq \mathbf{0}$ meant $X\mathbf{x} \neq \mathbf{0}$. If we extend the argument to rectangular $X \in \mathbb{R}^{m\times n}$ with full rank n and $\mathbf{x} \in R^n$, the argument would still hold.

Lemma 84. *If A is s.p.d. and S is any principal submatrix of A then S is s.p.d.*

Proof. Suppose $S \in \mathbb{R}^{m\times m}$ is any principal submatrix of $A \in \mathbb{R}^{n\times n}$. Let $\mathbf{y} \neq \mathbf{0} \in \mathbb{R}^m$, $\mathbf{z} = \mathbf{0} \in \mathbb{R}^{n-m}$ and let $\mathbf{x}^* = [\mathbf{y}^T, \mathbf{z}^T]^T \in \mathbb{R}^n$. Let P_S be a permutation such that $P_{\mathbf{s}}\mathbf{x}^* = \mathbf{x}_s$, where $\mathbf{0} \neq \mathbf{x}_s \ni x_k = y_i$ is located in the k^{th} row where row/column i of S comes from in A (see Definition 1.) We find that $\mathbf{y}^TS\mathbf{y} = c = \mathbf{x}_s^TA\mathbf{x}_s$, and because $c = \mathbf{x}_s^TA\mathbf{x}_s > 0$ for all such $\mathbf{x}_s \neq \mathbf{0}$, then $\mathbf{y}^TS\mathbf{y} > 0$ for all $\mathbf{y} \neq \mathbf{0}$ and S must also be s.p.d. □

Corollary 85. *The diagonal entries of a s.p.d matrix A are all positive.*

Proof. Trivially: Select $S_i = [a_{i,i}] \in A$, then for any $x \neq 0$, $ax^2 > 0$ since A is s.p.d. □

Proposition 86. *If A is s.p.d. then the factorization LDL^T exists and $D = \mathrm{diag}(d_1, \ldots, d_n)$ all positive diagonal entries.*

Proof. By Lemma 84 all leading principal submatrices of A are nonsingular so by Proposition 80 $A = LDM^T$ exist, and since L is nonsingular we can put $X = L^{-T}$ and apply Lemma 83 we have that $XAX^T = L^{-1}AL^{-T} = DM^TL^{-T} = G$ which must also be s.p.d. Since M and L^{-T} are unit upper triangular, ML^{-T} also is and G must have the same positive diagonal as D.

Finally, since A is symmetric, by Theorem 81 $M = L$ and $A = LDL^T$ exists. □

Note that for Proposition 82 up to the last line of Proposition 86 we never used symmetry, and these results hold for more general positive definite matrices. The following result is s.p.d. specific.

Theorem 87. *If $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, then there exists a unique lower triangular $G \in \mathbb{R}^{n\times n}$ with positive diagonal entries such that $A = GG^T$, and $a_{ii} > 0$. This reduction is known as the* Cholesky factorization.

Proof. Since A is s.p.d. by Proposition 86 there exists a unit lower triangular L and a diagonal $D = \mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$ such that $A = LDL^T$. Since the d_k are positive (by Lemma 85) the matrix $G = L \cdot \mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$ is real lower triangular with positive diagonal entries. It also satisfies $A = GG^T$. Uniqueness follows from uniqueness of LDL^T factorization (Theorem 81). $a_{ii} > 0$ follows because $a_{ii} = (\mathrm{diag}G)^2 > 0$ (A is nonsingular.) □

Note. Equivalently one can use a upper triangular matrix U and write $A = U^TU$.

3.2.1. *Computing the Cholesky Factor.* We want to derive an algorithm for the Cholesky factorization L. We have just proved that it exists and thus $A = LL^T$ where A is s.p.d. Deriving it again using a different approach, provides us with the means for an algorithm to compute the Cholesky factors L.

Lemma 88. *Let A be s.p.d. Then there exists a unique lower triangular nonsingular matrix L, with positive diagonal entries, such that $A = LL^T$.*

Proof. Choosing $l_{ii} > 0$ will determine L uniquely. Using induction on the dimension n. If $n = 1$, choose $l_{11} = \sqrt{a_{11}}$, which must exist since $a_{ii} > 0$ for s.p.d. A.

$$A = \begin{bmatrix} a_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix} = \begin{bmatrix} \sqrt{a_{11}} & 0 \\ \frac{A_{12}^T}{\sqrt{a_{11}}} & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{A}_{22} \end{bmatrix} \begin{bmatrix} \sqrt{a_{11}} & \frac{A_{12}^T}{\sqrt{a_{11}}} \\ 0 & I \end{bmatrix}$$
$$= \begin{bmatrix} a_{11} & A_{12} \\ A_{12}^T & \tilde{A}_{22} + \frac{A_{12}^TA_{12}}{a_{11}} \end{bmatrix}$$

so the $(n-1)\times(n-1)$ matrix $\tilde{A}_{22} = A_{22} - \frac{A_{12}^TA_{12}}{a_{11}}$ is symmetric. By Lemma 82 $D = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{A} \end{bmatrix}$ is s.p.d. and thus by Lemma 83 $\tilde{A}_{22}$ is s.p.d. By induction there exists an $\tilde{L}$ such that $\tilde{A}_{22} = \tilde{L}\tilde{L}^T$ and

$$A = \begin{bmatrix} \sqrt{a_{11}} & 0 \\ \frac{A_{12}^T}{\sqrt{a_{11}}} & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{L}\tilde{L}^T \end{bmatrix} \begin{bmatrix} \sqrt{a_{11}} & \frac{A_{12}^T}{\sqrt{a_{11}}} \\ 0 & I \end{bmatrix}$$
$$= \begin{bmatrix} \sqrt{a_{11}} & 0 \\ \frac{A_{12}^T}{\sqrt{a_{11}}} & \tilde{L} \end{bmatrix} \begin{bmatrix} \sqrt{a_{11}} & \frac{A_{12}^T}{\sqrt{a_{11}}} \\ 0 & \tilde{L}^T \end{bmatrix} = LL^T.$$

□

We see that D is successively reduced to I by a series of operations that operate on the $\tilde{L}\tilde{L}^T$ portion of D. This gives the following algorithm called the *Outer product version Cholesky*:[9] :

Let $A_1 = A$. At step i we have $A_i = \begin{bmatrix} I_{i-1} & 0 & 0 \\ 0 & a_{ii} & \mathbf{a}_i^* \\ 0 & \mathbf{a}_i & B_i \end{bmatrix}$. If we have a lower triangular $L_i = \begin{bmatrix} I_{i-1} & 0 & 0 \\ 0 & \sqrt{a_{ii}} & 0 \\ 0 & \frac{1}{\sqrt{a_{ii}}}\mathbf{a}_i & I_{n-i} \end{bmatrix}$ we can write $A_i = L_i A_{i+1} L_i^T$ where $A_{i+1} = \begin{bmatrix} I_i & 0 \\ 0 & B_i - \frac{1}{a_{ii}}\mathbf{a}_i\mathbf{a}_i^T \end{bmatrix}$. If we repeat this for $i = [1, n]$ we get $A_{n+1} = I$ and from this we can deduce $L = L_1 \cdots L_n$.

Note. From the structure of L_i (a Gauss transform) and its updating effects on A_i we see that in fact we have a symmetric Gauss elimination! We don't prove it explicitly but by Lemma 84 Cholesky factorization will never require pivoting for convergence and s.p.d. matrices are "diagonally dominant".

3.3. **Orthogonal Decompositions - QR Factorization.** The idea of QR factorization is to successively form a sequence of orthonormal vectors $\mathbf{q}_i$ to span the same space as the columns of A, i.e. to find an orthonormal basis for A. We also want an upper triangular matrix R where the diagonal entries $\neq 0$, which is easy to solve by back substitution (Section 5.)

3.3.1. *Gram-Schmidt Orthogonalization Process.* Let $\mathcal{V}$ be a n-dimensional inner product space with some known basis $\mathbf{a}_1, \ldots, \mathbf{a}_n$ (so $\mathbf{a}_i \neq \mathbf{0}$.) We will construct orthogonal elements $\mathbf{q}_i$ of the basis $\mathbf{q}_1, \ldots, \mathbf{q}_n$. The projection of $\mathbf{x}$ onto $\mathbf{y}$ is $\text{proj}_{\mathbf{y}}(\mathbf{x}) = \frac{\langle \mathbf{x};\mathbf{y} \rangle}{\langle \mathbf{y};\mathbf{y} \rangle}\mathbf{y} = \frac{\langle \mathbf{x};\mathbf{y} \rangle}{||\mathbf{y}||^2}\mathbf{y}$.

We can choose any $\mathbf{a}_i$ to start building our orthogonal basis from so we select $\mathbf{a}_1$, and set $\mathbf{q}_1 = \mathbf{a}_1/||\mathbf{a}_1||$. Now $\mathbf{q}_2$ (2^{nd} basis vector) must be orthogonal to $\mathbf{q}_1$. We can achieve this by subtracting a suitable multiple of $\mathbf{q}_1$ so $\mathbf{q}_2 = \mathbf{a}_2 - r_1\mathbf{q}_1$. Since by orthogonality $\langle \mathbf{q}_2; \mathbf{q}_1 \rangle = \langle \mathbf{a}_2; \mathbf{q}_1 \rangle - r_1 \langle \mathbf{q}_1; \mathbf{q}_1 \rangle = \langle \mathbf{a}_2; \mathbf{q}_1 \rangle - r_1 ||\mathbf{q}_1||^2 = 0$ we get that $r_1 = \langle \mathbf{a}_2; \mathbf{q}_1 \rangle / ||\mathbf{q}_1||^2$, and

$$\mathbf{q}_2 = \mathbf{a}_2 - \frac{\langle \mathbf{a}_2; \mathbf{q}_1 \rangle}{||\mathbf{q}_1||^2}\mathbf{q}_1 = \mathbf{a}_2 - \text{proj}_{\mathbf{q}_1}(\mathbf{a}_2) = \mathbf{a}_2 - r_1\mathbf{q}_1. \tag{3.1}$$

Next $\mathbf{q}_3 = \mathbf{a}_3 - r_{13}\mathbf{q}_1 - r_{23}\mathbf{q}_2$. Since $\mathbf{q}_1$ and $\mathbf{q}_2$ are orthonormal we only have to look at $\langle \mathbf{a}_3; \mathbf{q}_1 \rangle = 0$ and $\langle \mathbf{a}_3; \mathbf{q}_2 \rangle = 0$. We get $r_{12} = \langle \mathbf{a}_3; \mathbf{q}_1 \rangle ||\mathbf{q}_1||^{-1}$ and $r_{23} = \langle \mathbf{a}_3; \mathbf{q}_2 \rangle ||\mathbf{q}_2||^{-1}$ and

$$\mathbf{q}_3 = \mathbf{a}_3 - \text{proj}_{\mathbf{q}_1}(\mathbf{a}_3)\mathbf{q}_1 - \text{proj}_{\mathbf{q}_2}(\mathbf{a}_3)\mathbf{q}_2.$$

Continuing, we get orthogonal vectors $\mathbf{q}_1, \ldots, \mathbf{q}_{j-1}$ as linear combinations of $\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}$, and $\mathbf{q}_j$ can be computed from $\mathbf{q}_j = \mathbf{a}_j - r_1\mathbf{q}_1 - \ldots - r_{j-1}\mathbf{q}_{j-1}$. The orthogonality constraint $\langle \mathbf{q}_j; \mathbf{q}_i \rangle = \langle \mathbf{a}_j; \mathbf{q}_i \rangle - r_j \langle \mathbf{q}_j; \mathbf{q}_i \rangle = 0$ requires $r_i = \langle \mathbf{a}_j; \mathbf{q}_i \rangle / \langle \mathbf{q}_j; \mathbf{q}_i \rangle$ and the general *classical Gram-Schmidt formula* is

$$\mathbf{q}_j = \mathbf{a}_j - \sum_{i=1}^{k-1} \text{proj}_{\mathbf{q}_i}(\mathbf{a}_j) = \mathbf{a}_j - \sum_{i=1}^{k-1} r_{ij}\mathbf{q}_i \tag{3.2}$$

We can get an orthonormal basis by taking $\mathbf{q}_i = \mathbf{q}_i/||\mathbf{q}_i||$. Let $||\mathbf{q}_i||^{-1} = c_i$ and set $r_{ij} = r_{ij}c_i$ above (it is convention to ensure $r_{jj} > 0$ - since $c = ||\mathbf{q}|| = \pm\sqrt{q_1^2 + \ldots + q_n^2}$.) As a result the Gram-Schmidt process shows existence of an

[9]Based on [GolubVanLoan] with *notation* from [Wikipedia].

orthogonal (orthonormal) basis for a finite-dimensional inner product space. In addition, ensuring that all $r_{jj} > 0$ removes ambiguity of signs.

In finite precision, roundoff (usually) result in loss of orthogonality ([Demmel]) and therefore a mathematically equivalent but algorithmically modified approach is taken - *modified Gram-Schmidt* - given in Equation 3.3. Instead of updating one vector at a time against the others to produce $\mathbf{q}_j$, all vectors are instead updated at each step.

$$\begin{aligned} \mathbf{q}_j^{(1)} &= \mathbf{a}_j - \text{proj}_{\mathbf{q}_1}(\mathbf{a}_j) \\ \mathbf{q}_j^{(2)} &= \mathbf{q}_j^{(1)} - \text{proj}_{\mathbf{q}_2}(\mathbf{q}_j^{(1)}) \\ &\vdots \\ \mathbf{q}_j^{(j-1)} &= \mathbf{q}_j^{(j-2)} - \text{proj}_{\mathbf{q}_{j-1}}(\mathbf{q}_j^{(j-2)}) \end{aligned} \tag{3.3}$$

In the first step all components non-orthogonal to $\mathbf{q}_1$ will removed from all $\mathbf{a}_{i>1}$, making $\mathbf{q}_1$ orthogonal to them. In the next step, all components non-orthogonal to $\mathbf{q_2}$ will be removed from all $\mathbf{a}_{i>2}$ (labeled $\mathbf{q}_{i>2}^{(1)}$), but not from $\mathbf{q}_1$ since it is already orthogonal.) By correcting the remaining bases in steps, rounding errors are leveled out and it is therefore more stable when precision is finite.

If we use the normalized basis $\mathbf{q}_1, \ldots, \mathbf{q_n}$, looking at Equation (3.2) we can rearrange the terms as

$$\begin{aligned} \mathbf{a}_1 &= r_{11}\mathbf{q}_1, \\ \mathbf{a}_2 &= r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2, \\ &\vdots \\ \mathbf{a}_n &= r_{1n}\mathbf{q}_n + r_{2n}\mathbf{q}_2 + \ldots + r_{nn}\mathbf{q}_n. \end{aligned}$$

This shows that the original matrix A can be expressed as the product of an orthogonal matrix Q and an upper triangular matrix R (where we choose $r_{ii} > 0$) - the (*full*) *QR factorization*.

$$A = [\mathbf{a}_1 \ldots \mathbf{a}_n],\ Q = [\mathbf{q}_1 \ldots \mathbf{q}_n],\ R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & & r_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix} \implies A = QR.$$

Proposition 89. *Assume $\mathbb{R}^{m\times n} \ni A = QR$, $m \geq n$ and full rank n. Assume $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ and $Q = \{\mathbf{q}_1, \ldots, \mathbf{q}_m\}$ are column partitionings. Then* $\text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = \text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$, $k \in [1, n]$. *In particular if $Q_n = Q(1:m, 1:n)$ and $Q_{m-n} = Q(1:m, n+1:m)$ then* $\text{rng}A = \text{rng}Q_n$ *and* $\text{rng}A^{\perp} = \text{rng}Q_{m-n}$ *and in addition $A = Q_nR_n$, $R_n = R(1:n, 1:n)$.*

Note. $R = \begin{bmatrix} R_n \\ 0 \end{bmatrix}$, $Q = [Q_n\ Q_{m-n}]$ so $A = QR = [Q_n\ Q_{m-n}] \begin{bmatrix} R_n \\ 0 \end{bmatrix} = Q_nR_n$.
Also, we will use $\hat{Q}\hat{R}$ to denote the *reduced* (or *thin*) *QR factors* Q_nR_n.

Proof. Since we saw from the QR factorization $\mathbf{a}_k = \sum_{i=1}^{k} r_{ik}\mathbf{q}_i \in \text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$, $\supseteq \text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$. Since $\text{rank}A = n$ we get $\dim \text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\} = k$ and hence $\text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\} = \text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$. We see the rest as a trivial consequence. □

Proposition 90. *Suppose $A \in \mathbb{R}^{m\times n}$ has full column rank. Then the thin QR factorization $A = \hat{Q}\hat{R}$ is unique where $\hat{Q} \in \mathbb{R}^{m\times n}$ has orthonormal columns and $\hat{R}$ is upper triangular with positive diagonal entries. Moreover $\hat{R} = G^T$, where G is the lower triangular Cholesky factor of A^TA.*

Proof. Since $A^TA = (\hat{Q}\hat{R})^T\hat{Q}\hat{R} = \hat{R}^T\hat{R}$ we see that $G = \hat{R}^T$ is the Cholesky factor of A^TA. This factor is unique by Theorem 87 and since $\hat{Q} = A\hat{R}^{-1}$, $\hat{Q}$ is also unique. □

Note. This of course again proves that $A = QR$ is unique since it is a special case of Proposition 90, and motivates the choice of $r_{ii} > 0$.

Putting all this together we reach the following result:

Theorem 91. *Any nonsingular matrix A can be factorized, $A = QR$, into the product of an orthogonal matrix Q and an upper triangular matrix R. The factorization is unique if all the diagonal entries of R are assumed to be positive.*

3.3.2. *Projectors.* We saw the use of projections in the Gram-Schmidt process, when we computed the familiar projections of one vector onto another. We will study this a bit further and see why projectors turn out to be important in applications.

Definition 92. A *projector* is a square matrix P that satisfies $P^2 = P$.

Such a matrix is said to be *idempotent*. The definition includes both orthogonal and *oblique* (nonorthogonal) projectors. If P is a projector, $I-P$ is also a projector, because $(I-P)^2 = I - 2P + P^2 = I - 2P + P = I - P$.

Note 93. If $\mathbf{v} \in \text{rng}P$, then $\mathbf{v}$ applying P results in $\mathbf{v}$ itself, it lies in its own projection. For some $\mathbf{x}$, $\mathbf{v} = P\mathbf{x}$ and $P\mathbf{v} = P^2\mathbf{x} = P\mathbf{x} = \mathbf{v}$. If $\mathbf{v} \neq P\mathbf{v}$ then $P(P\mathbf{v} - \mathbf{v}) = P^2\mathbf{v} - P\mathbf{v} = \mathbf{0}$ and $P\mathbf{v} - \mathbf{v} \in \ker P$, or $(I-P)P\mathbf{v} = \mathbf{0}$. From this we see that $\text{rng}(I-P) = \ker P$ and the complementary fact that $P = I - (I-P)$ gives $\ker(I-P) = \text{rng}P$. Also, if $\mathbf{v} \in \ker P$ then $\mathbf{v} = \mathbf{v} - P\mathbf{v} = \mathbf{0}$ and also $\mathbf{v} \in \ker(I-P)$ then $\mathbf{v}(I-P) = \mathbf{0}$ and $\ker(I-P) \cap \ker P = \mathbf{0}$, and a projector separates $\mathbb{R}^m$ into two spaces S_1 and S_2.

We wait until Section 3.4 to show the following:

Proposition 94. *A projector P is orthogonal iff $P = P^T$.*

From Note 101 we deduce that we can write $P = \hat{Q}\hat{Q}^*$, where the columns of $\hat{Q}$ are orthonormal. A special case of this is $P_{\mathbf{q}} = \mathbf{q}\mathbf{q}^T$ which isolates the component in the $\mathbf{q}$ direction ($||\mathbf{q}|| = 1$). The complement of this is $P_{\perp\mathbf{q}} = I - \mathbf{q}\mathbf{q}^T$ or if we have non-normalized: $Q = \mathbf{q}\mathbf{q}^T/\mathbf{q}^T\mathbf{q}$ or

$$Q = I - \frac{\mathbf{q}\mathbf{q}^T}{\mathbf{q}^T\mathbf{q}} \tag{3.4}$$

which we recognize from the Gram-Schmit process.

3.3.3. *Arbitrary Basis Projectors.* We now show how to construct a projection with an arbitrary basis (i.e. possibly not orthogonal) onto a subspace $\mathcal{V}$ of $\mathbb{R}^m$.

Suppose $\mathcal{V}$ is spanned by the linearly independent vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = A$ (i.e. A is $m \times n$.) As $\mathbf{w}$ passes from $\mathbf{v}$ to its orthogonal projection $\mathbf{b} \in \text{rng}A$, $\mathbf{w} = \mathbf{b} - \mathbf{v}$ must be orthogonal to $\text{rng}A \Leftrightarrow \mathbf{a}_j^T(\mathbf{b} - \mathbf{v}) = \mathbf{0}$ for every j. Since $\mathbf{b} \in \text{rng}A$, we can set $\mathbf{b} = A\mathbf{x}$ and write $\mathbf{a}_j^T(A\mathbf{x} - \mathbf{v}) = \mathbf{0}$ for each j, or equivalently $A^T(A\mathbf{x} - \mathbf{v}) = \mathbf{0}$ or $A^TA\mathbf{v} = A^T\mathbf{v}$, where A^TA of course is non-singular by Equation (2.6), so

$$\mathbf{x} = (A^TA)^{-1}A^T\mathbf{v}. \tag{3.5}$$

The projection of $\mathbf{v}$, $\mathbf{y} = A\mathbf{x} = A(A^TA)^{-1}A^T\mathbf{v}$ and the orthogonal projector onto $\text{rng}A$ can be written as $P = A(A^TA)^{-1}A^T$, a multidimensional generalization of Equation (3.4). In Theorem 109 we will see that $\mathbf{x}$ minimizes the norm, as promised by Proposition 66.

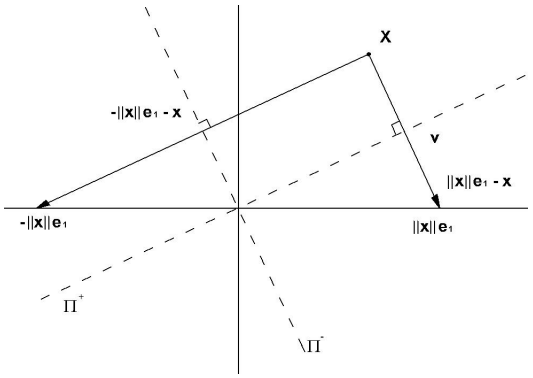

FIGURE 3.1. Orthogonal reflectors about hyperplanes Π^+ and Π^-.

3.3.4. *Householder Reflection.* Using projectors, we now show two other often used methods to construct the QR factorization.

We want to construct an orthogonal matrix $Q_k \in \mathbb{R}^{m\times n}$ that zeros out the subdiagonal elements in the k^{th} column of a matrix A, without affecting subdiagonal entries in previous columns. Each Q_k is chosen to be an orthogonal matrix

$$Q_k = \begin{bmatrix} I_{1:k-1} & 0 \\ 0 & H_{k:m} \end{bmatrix}, \tag{3.6}$$

where $I_{1:k-1}$ is the $(k-1)\times(k-1)$ identity matrix and $H_{k:m}$ is an $(m-k+1)\times(m-k+1)$ orthogonal matrix (where $I_{1:k-1}$ of course disappears when $k=1$) and multiplication by H introduces zeros in the k^{th} column (in the $k+1$ elements.) Suppose at step k, the entries k to m of the k^{th} column are given by $\mathbf{x}\in\mathbb{R}^{m-k+1}$. The desired H should

$$\mathbf{x} = \begin{bmatrix} \bullet \\ \bullet \\ \bullet \\ \bullet \end{bmatrix} \rightarrow H\mathbf{x} = \begin{bmatrix} \pm||\mathbf{x}|| \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \pm||\mathbf{x}||e_1,$$

where $\bullet$ are any elements (not all $= 0$). There are many mappings that could do this, the *Householder reflector* turns out to be easy to compute. The reflector will reflect the space $\mathbb{R}^{m-k+1}$ across the hyperplanes Π^+ och Π^- orthogonal to $\mathbf{v} = \pm||\mathbf{x}||\mathbf{e}_1 - \mathbf{x}$. From Figure 3.1 we get a geometrical picture of the situation, and why we get $\pm$, and we will want to select the reflection that moves $\mathbf{x}$ the larger distance. When applied, every point on one side of the hyperplane is mapped to the other. $\mathbf{x}$ is mapped to $\pm||\mathbf{x}||\mathbf{e}_1$. The orthogonal projection of any vector $\mathbf{y}\in\mathbb{R}^m$ onto Π is $P\mathbf{y} = \left(I - \dfrac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}\right)\mathbf{y} = \mathbf{y} - \mathbf{v}\left(\dfrac{\mathbf{v}^T\mathbf{y}}{\mathbf{v}^T\mathbf{v}}\right)$, but with H we want to go even further and reflect across H, go twice as far and thus H becomes

$$H = I - 2\frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}} = I - 2\mathbf{u}\mathbf{u}^T, \quad \mathbf{u} = \mathbf{v}/||\mathbf{v}||.$$

Mathematically either of '+' or '−' is satisfactory, but it is common practice to choose the *Householder vector* $\mathbf{v} = -\text{sign}(x_1)||\mathbf{x}||\mathbf{e}_1 - \mathbf{x}$ with $\text{sign}(0) = 1$, or more commonly $\mathbf{v} = \text{sign}(x_1)||\mathbf{x}||\mathbf{e}_1 + \mathbf{x}$, to minimize cancelation in the x_i component.

Because $H^T = I - 2(\mathbf{u}\mathbf{u}^T)^T = H$ and $HH^T = (I - 2\mathbf{u}\mathbf{u}^T) = I - 4\mathbf{u}\mathbf{u}^T - 4\mathbf{u}(\mathbf{u}^T\mathbf{u})\mathbf{u}^T = I = H^2$, H is a symmetric and orthogonal matrix and H gives an orthogonal transformation Q_k as desired.

Householder QR. First noting that the product of orthogonal matrices will be orthogonal (Proposition 45), we briefly illustrate how Householder reflections are used to produce the QR factorization of a matrix $A \in \mathbb{R}^{5\times 4}$. At each step orthogonal Q is easily constructed with H via $\mathbf{v}$ (depending on our current A at step i.)

$$A \overset{[H]}{\underset{(1)}{\rightarrow}} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \end{bmatrix} \overset{\begin{bmatrix} 1 & 0 \\ 0 & H_{2:m} \end{bmatrix}}{\underset{(2)}{\rightarrow}} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \end{bmatrix} \overset{\begin{bmatrix} I_{1:2} & 0 \\ 0 & H_{3:m} \end{bmatrix}}{\underset{(3)}{\rightarrow}}$$

$$\overset{\cdots}{\underset{(3)}{\rightarrow}} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & 0 & \bullet \\ 0 & 0 & 0 & \bullet \end{bmatrix} \overset{\begin{bmatrix} I_{1:3} & 0 \\ 0 & H_{4:5} \end{bmatrix}}{\underset{(4)}{\rightarrow}} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & 0 & \bullet \\ 0 & 0 & 0 & 0 \end{bmatrix} = Q_4 \cdots Q_1 A = QR.$$

3.3.5. *Givens Rotation.* A 2×2 *Givens rotation* $G(\theta) \equiv \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}$ is a counterclockwise rotation through an angle θ of $\mathbf{x} \in \mathbb{R}^2$. Putting $G^T G = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} c & -s \\ s & c \end{bmatrix} = \begin{bmatrix} c^2+s^2 & sc-cs \\ cs-sc & s^2+c^2 \end{bmatrix} = I$, so G is orthogonal. The orthogonal transformation $Q(G_{ij\theta}) = Q_{ij\theta}$ matrix, with the general Givens rotation $G_{ij\theta}$, is a θ radians rotation about the (i,k) coordinate plane of $\mathbf{x} \in \mathbb{R}^m$ given by

(3.7)
$$Q_{ij\theta} = \begin{bmatrix} I_{1:i-1} & & & & 0 \\ & c & & -s & \\ & & I_{i+1:j-1} & & \\ & s & & c & \\ 0 & & & & I_{j+1:m} \end{bmatrix} = \begin{bmatrix} I_{1:i-1} & & 0 \\ & G_{ij\theta} & \\ 0 & & I_{j+1:m} \end{bmatrix}.$$

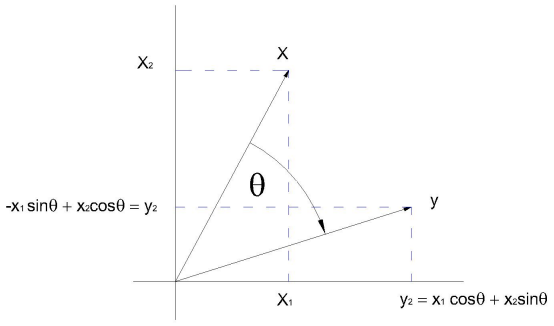

FIGURE 3.2. A θ radians Givens rotation about the axis.

Say we want to zero $y_k{}^{\text{th}}$ component of $\mathbf{y}$, then $\mathbf{y} = G_{ij\theta}\mathbf{x}$ gives

$$y_k = \begin{cases} cx_i - sx_j & k = i \\ sx_i + cx_k & k = j \\ x_k & k \neq i, j \end{cases} = 0 \text{ if } \begin{array}{l} c = x_i/\sigma, \\ s = -x_j/\sigma, \\ \sigma = \sqrt{x_i^2 + x_j^2}. \end{array} \tag{3.8}$$

The way we showed that G was orthogonal, it follows trivially that $Q_{ij\theta}$ is also orthogonal. Zeroing y_k as in Equation (3.8) is not optimal (if $\sigma \to 0$), but shows how Givens rotations can be chosen to selectively zero off-diagonal elements.

Givens QR. We briefly illustrate how Givens rotations are used to produce the QR factorization of A in the previous 5×4 example for Householder transformations, interrupting after step (2).

$$Q_{ij\theta}^{(1)} A_2 = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & c & -s \\ & & & s & c \end{bmatrix} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \end{bmatrix} \underset{(2.5)}{=} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & 0 & \bullet \end{bmatrix} \to$$

$$Q_{ij\theta}^{(2.5)} A_{2.5} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & c & -s & \\ & & s & c & \\ & & & & 1 \end{bmatrix} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & 0 & \bullet \end{bmatrix} \underset{(3)}{=} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & 0 & \bullet \\ 0 & 0 & 0 & \bullet \end{bmatrix}$$

We do not prove this explicitly, but it is clear that it is possible to construct these orthogonal transformations to zero out the rows of A as well (e.g. $QA^T \to AQ^T$, where Q is an orthogonal transformation) and is used in some applications.

Each of the three different methods construct the QR in different ways (remember that the QR is unique) and have their own problems and uses.

3.4. **Orthogonal Decompositions - Singular Value Decomposition.** The following decomposition is a very important theoretical tool that also has many practical applications. It turns out to be an extremely powerful and comprehensive description of *any* matrix. Some of the most interesting properties, for example its relation to the eigenvalue decomposition or its importance in studying conditioning of problems, will not be examined in detail.

Theorem 95. *Let $A \in \mathbb{C}^{m\times n}$ (m, n arbitrary and A possibly rank deficient). Then there exists a factorization (*Singular Value Decomposition - *SVD)*

$$A = U\Sigma V^T$$

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal, and $\Sigma = diag(\sigma_1, \ldots, \sigma_{p=\min(m,n)})$ where $\sigma_1 \geq \ldots \geq \sigma_p \geq 0 \in \mathbb{R}$. The columns $\mathbf{u}_1, \ldots, \mathbf{u}_m$ of U are called left singular vectors. *The columns $\mathbf{v}_1 \ldots, \mathbf{v}_n$ are called* right singular vectors. *The σ_i are called* singular values.

Proof. We use induction on m and n (and assume $m \geq n$, if $m < n$ we can consider A^T instead.) Assume the SVD exists for $(m-1) \times (n-1)$ and $A \neq 0$; otherwise we can take $\Sigma = 0$ and let U and V be arbitrary orthogonal matrices.

The base case is when $n = 1$ (since $m \geq n$). We write $A = U\Sigma V^T$ with $U = A/||A||_2$, $\Sigma = ||A||_2$, and $V = 1$.

For the induction step choose $\mathbf{v}$ so $||\mathbf{v}||_2 = 1$ and $||A||_2 = ||A\mathbf{v}||_2 > 0$. Such a $\mathbf{v}$ exists by the definition of $||A||_2 = \max_{||\mathbf{v}||_2=1} ||A\mathbf{v}||_2$ (Definition 33). Let $\mathbf{u} = A\mathbf{v}/||A\mathbf{v}||_2$, which is a unit vector. Choose $\tilde{U}$ and $\tilde{V}$ to that $U = [\mathbf{u}, \tilde{U}]$ is an $m \times n$ orthogonal matrix, and $V = [\mathbf{v}, \tilde{V}]$ is $n \times n$ orthogonal matrix.

$$U^T A V = \begin{bmatrix} \mathbf{u}^T \\ \tilde{U}^T \end{bmatrix} A \begin{bmatrix} \mathbf{v} & \tilde{V} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^T A\mathbf{v} & \mathbf{u}^T A\tilde{V} \\ \tilde{U}^T A\mathbf{v} & \tilde{U}^T A\tilde{V} \end{bmatrix}.$$

Then

$$\mathbf{u}^T A\mathbf{v} = \frac{(A\mathbf{v})^T(A\mathbf{v})}{||A\mathbf{v}||_2} = \frac{||A\mathbf{v}||_2^2}{||A\mathbf{v}||_2} = ||A\mathbf{v}||_2 = ||A||_2 \equiv \sigma$$

and $\tilde{U}^* A\mathbf{v} = \tilde{U}^*\mathbf{u}||A\mathbf{v}||_2 = 0$. We claim $\mathbf{u}^T A\tilde{V} = 0$ too because otherwise $\sigma = ||A||_2 = ||U^T AV||_2 \geq ||\,[1, 0, \ldots, 0]\, U^* AV||_2 = ||\left[\sigma|\mathbf{u}^T A\tilde{V}|\right]||_2 > \sigma$, a contradiction.[10]

So $U^*AV = \begin{bmatrix} \sigma & 0 \\ 0 & \tilde{U}^T A\tilde{V} \end{bmatrix} = \begin{bmatrix} \sigma & 0 \\ 0 & A \end{bmatrix}$. We may now apply the induction hypothesis to $\tilde{A}$ to get $\tilde{A} = U_1 \Sigma V_1^T$, where U_1 is $(m-1)\times(n-1)$, Σ_1 is $(n-1)\times(n-1)$, and V_1 is $(n-1)\times(n-1)$. So

$$U^*AV = \begin{bmatrix} \sigma & 0 \\ 0 & U_1\Sigma V_1^T \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix} \begin{bmatrix} \sigma & 0 \\ 0 & \Sigma \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V_1 \end{bmatrix}^T$$

or

$$A = \left(U \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix}\right) \begin{bmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{bmatrix} \left(V \begin{bmatrix} 1 & 0 \\ 0 & V_1 \end{bmatrix}\right)^T,$$

which is our desired decomposition. □

Note. We denote $\sigma_{max} = \sigma_1$ the largest singular value, the i^{th} largest σ_i, the smallest singular value as σ_{min}, and σ_p as the p^{th} ($p = \min\{m, n\}$ of a $A \in \mathbb{R}^{m\times n}$.)

Proposition 96. *Let $A \in \mathbb{R}^{m\times n}$, $r \leq p$ the number of $\sigma_i \neq 0$, and $\langle \mathbf{x}, \ldots, \mathbf{z}\rangle = \text{span}\{\mathbf{x}, \ldots, \mathbf{z}\}$. Then*

1. $\text{rank}A = r$,
2. $\text{rng}A = \langle \mathbf{u}_1, \ldots, \mathbf{u}_r\rangle$ *and* $\ker A = \langle \mathbf{v}_{r+1}, \ldots, \mathbf{v_n}\rangle$,
3. *For $A \in \mathbb{R}^{m\times m}$, $|\det A| = \prod_{i=1}^m \sigma_i$.*

Proof. (1) Choose $m \times (m-n)$ $\tilde{U}$ such that $\hat{U} = [U, \tilde{U}]$ is square and orthogonal. Since $\hat{U}$ and V are nonsingular, A and $\hat{U}^T AV = \begin{bmatrix} \Sigma^{n\times n} \\ 0^{(m-n)\times n} \end{bmatrix} = \hat{\Sigma}$ have same rank r (the number of pivots in A).

(2) Clearly $\text{rng}\Sigma = \langle \mathbf{e}_1, \ldots, \mathbf{e}_r\rangle \subseteq \mathbb{R}^m$ and $\ker \Sigma = \langle \mathbf{e}_{r+1}, \ldots, \mathbf{e}_n\rangle \subseteq \mathbb{R}^n$.

$A\mathbf{z} = \hat{U}\hat{\Sigma}V^T\mathbf{z} = \mathbf{0} \Leftrightarrow \hat{U}^T A\mathbf{z} = \hat{\Sigma}V^T\mathbf{z} = \hat{U}^T AV(V^T\mathbf{z}) = \hat{\Sigma}V^T\mathbf{z} = \mathbf{0}$. So for $\mathbf{z} \in \ker A$, $V^T\mathbf{z} \in \ker\hat{\Sigma}$, which leaves us with $\langle \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\rangle = \ker A$.

A similar argument gives that $\hat{U} \cdot \text{rng}(\hat{U}^T AV = \hat{\Sigma}) = \text{rng}A = \langle \mathbf{u}_1, \ldots, \mathbf{u}_r\rangle$.

(3) If U orthogonal, $\det U = |1|$ and $\det U^T = (\det U)^T$, so $|\det A| = |\det(U\Sigma V^T)| = |\det U| \cdot |\det\Sigma| \cdot |\det V^T| = |\det\Sigma| = \prod_{i=1}^m \sigma_i$. □

From [GolubVanLoan] we provide the following results regarding norms without proof.

Lemma 97. *Let $A \in \mathbb{R}^{m\times n}$ then $||A||_2 = \sigma_1 = \sigma_{max}$ and $\min_{\mathbf{x}\neq 0} ||A\mathbf{x}||_2/||\mathbf{x}||_2 = \sigma_{min}$.*

If we have $A = U\Sigma V^T$ then $A\mathbf{x} = \mathbf{b} \Leftrightarrow U^T A\mathbf{x} = U^T U\Sigma V^T\mathbf{x} = U^T\mathbf{b}$. Putting $U^T\mathbf{b} = \mathbf{b}'$ and $V^T\mathbf{x} = \mathbf{x}'$, we get $\mathbf{b}' = \Sigma\mathbf{x}'$. Since the SVD exists for *any* matrix this change of basis shows that *any* matrix A can be made diagonal if we use the proper bases for $\text{rng}A$ and $\ker A$. Later we will show additional useful properties of the SVD (related to the eigenvalue problem) that explain why it is such a powerful decomposition.

Definition 98. If $A = U\Sigma V^T \in \mathbb{R}^{m\times n}_{\{m\geq n\}}$ is the SVD of A, then $A = \hat{U}\hat{\Sigma}V^T$, where $\hat{U} = U_{1:m,1:n} = [\mathbf{u}_1, \ldots, \mathbf{u}_n] \in \mathbb{R}^{m\times n}$, and $\hat{\Sigma} = \Sigma_{1:n,1:n} = \text{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{m\times n}$, is the *thin SVD*.

[10] See [Demmel] p. 22-23, Lemma 1.7 part 7.

Proposition 99. *Assume we have the SVD of $A \in \mathbb{R}^{m\times n}$. If $k < r = \text{rank}A$ and*

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{3.9}$$

then $\min_{\text{rank}B=k} ||A - B||_2 = ||A - A_k|| = \sigma_{k+1}$.

Proof. $U^T A_k V = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$ it follows that $\text{rank}A_k = k$ and that $U^T(A - A_k)V = \text{diag}(0, \ldots, \sigma_{k+1}, \ldots, \sigma_p)$ and so $||A - A_k||_2 = \sigma_{k+1}$. Now suppose $\text{rank}B = k$ for $B \in \mathbb{R}^{m\times n}$. It follows that we can find orthonormal vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{n-k}$ such that $\ker B = \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_{n-k}\}$. Since $(n-k)+(k+1) > n$, $\text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_{n-k}\} \cap \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\} \neq \{\mathbf{0}\}$ (they intersect). If $\mathbf{z}$ is a unit 2-norm vector in this intersection, $B\mathbf{z} = \mathbf{0}$ and $A\mathbf{z} = \sum_{i=1}^{k+1} \sigma_i(\mathbf{v}_i^T\mathbf{z})\mathbf{u}_i$ we have that

$$||A - B||_2^2 \geq ||(A - B)\mathbf{z}||_2^2 = ||A\mathbf{z}||_2^2 = \sum_{i=1}^{k+1} \sigma_i(\mathbf{v}_i^T\mathbf{z})\mathbf{u}_i \geq \sigma_{k+1}^2.$$

□

We say that Equation (3.9) is a *low rank approximation SVD* of A. This also shows that the SVD maximizes the "energy" of A, or the "information contents" of A. Low-rank approximations A_k are used as a way to compress information with minimal information loss or extract the "dominant" information in data mining (see [Eldén] for an example). We give a final result that provides a geometrical understanding of the SVD.

Proposition 100. *Let $S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : ||\mathbf{x}||_2 = 1\}$, the unit sphere in $\mathbb{R}^n$. Let $A \cdot S^{n-1} = \{A\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \in \mathbb{R}^n \text{ and } ||\mathbf{x}||_2 = 1\}$. Then $A \cdot S^{n-1}$ is an ellipsoid centered at the origin of $\mathbb{R}^n$, with principal axes $\sigma_i\mathbf{u}_i$.*

Proof. We will multiply by one factor of $A = U\Sigma V^T$ at the time to form $A \cdot S^{n-1}$. Assume for simplicity that A nonsingular. Since V is orthogonal it maps unit vectors to unit vectors and $V^T \cdot S^{n-1} = S^{n-1}$. Then, since $\mathbf{v} \in S^{n-1}$ iff $||\mathbf{v}||_2 = 1$, $\mathbf{w} = \Sigma\mathbf{v} \in \Sigma S^{n-1}$ iff $||\Sigma^{-1}\mathbf{w}||_2 = 1$ or $\sum_{i=1}^{n}(w_i/\sigma_i)^2 = 1$. This defines an ellipsoid with principal axes $\sigma_i\mathbf{e}_i$, where $\mathbf{e}_i$ is the i^{th} column of the identity matrix. Multiplying each $\mathbf{w} = \Sigma\mathbf{v}$ by U just rotates the ellipsoid so that each $\mathbf{e}_i$ becomes $\mathbf{u}_i$, the i^{th} column of U. □

Hence the SVD describes a rotation (by V,) a scaling of the unit hypersphere axes (by the elements of Σ) and another rotation (by U) of the resulting ellipsoid (higher dimensional ellipse). In the 2-dimensional case (or the rank 2 approximation) it is a rotated ellipse.

We now present the proof of Proposition 94:

Proof. $P\mathbf{x} \in S_1$ and $(I - P)\mathbf{y} \in S_2$. If $P = P^*$ then the inner product is $\mathbf{x}^*P^*(I - P)\mathbf{y} = \mathbf{x}^*(P - P^2)\mathbf{y} = \mathbf{0}$.

Now suppose P projects S_1 along S_2 where $S_1 \perp S_2$ and $\dim S_1 = n$. Let $\{\mathbf{q}_1, \ldots, \mathbf{q}_m\}$ be an orthonormal basis for $\mathbb{R}^m$, where $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is a basis for S_1 and $\{\mathbf{q}_{n+1}, \ldots, \mathbf{q}_m\}$ is a basis for S_2. For $j \leq n$ we have $P\mathbf{q}_j = \mathbf{q}_j$, and for $j > n$ $P\mathbf{q}_j = \mathbf{0}$. Now let Q be the orthogonal matrix whose j^{th} column is $\mathbf{q}_j$. We then have $PQ = [\mathbf{q}_1, \ldots, \mathbf{q}_n, 0, \ldots]$ so that $Q^TPQ = \text{diag}(\underbrace{1, \ldots, 1}_{n}, 0, \ldots) = \Sigma$. Thus we have constructed $P = Q\Sigma Q^T$, a SVD of P. Since $P^T = Q\Sigma^T Q^T = Q\Sigma Q^T = P$, P is orthogonal. □

Note 101. From this we can deduce that $P = QQ^T = \hat{Q}\hat{Q}^T$ because P will have $m - n$ zero eigenvalues and we can drop those columns.

3.4.1. *On Computing the SVD.* The following theorem illustrates the connection between the SVD and the eigendecomposition of a matrix A.

Theorem 102. *Let $\mathbb{R}^{m\times n} \ni A = U\Sigma V^T$ be the SVD of A, with $m \geq n$.*[11]

1. Suppose $A = A^T$ with eigenvalues λ_i and orthonormal eigenvectors $\mathbf{u}_i$, i.e. $A = U\Lambda U^T$ is the eigendecomposition with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $U = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$, and $UU^T = I$. Then a SVD of A is $A = U\Sigma V^T$, where $\sigma_i = |\lambda_i|$ and $\mathbf{v}_i = \operatorname{sign}(\lambda_i)\mathbf{u}_i$, with $\operatorname{sign}(0) = 1$.

2. The eigenvalues of the symmetric A^TA are σ_i^2. The right singular vectors $\mathbf{v}_i$ are corresponding orthonormal eigenvectors.

3. The eigenvalues of the symmetric matrix AA^T are σ_i^T and $m-n$ zeroes. The left singular eigenvectors $\mathbf{u}_i$ are the corresponding orthonormal eigenvectors for the eigenvalues σ_i^2. One can take any $m-n$ other orthonormal vectors as eigenvectors for the eigenvalue 0.

4. Let $H = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$ where A is square and $A = U\Sigma V^T$, $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, $V = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ and U as before. Then the $2n$ eigenvalues of H are $\pm\sigma_i$, with corresponding unit vectors $\pm 1/\sqrt{2}\begin{bmatrix} \mathbf{v}_i \\ \pm\mathbf{u}_i \end{bmatrix}$.

Proof. (1) Is true by definition of SVD. (2) $A^TA = V\Sigma U^TU\Sigma V^T = V\Sigma^2V^T$. This is an eigendecomposition of A^TA, with the columns of V the eigenvectors and the diagonal entries of Σ^2 the eigenvalues. (3) Choose $m \times (m-n)$ matrix $\tilde{U}$ so that $[U, \tilde{U}]$ is square and orthogonal. We see that we get the eigendecomposition of AA^T if we write[12]

$$AA^T = U\Sigma V^TV\Sigma U^T = U\Sigma^2U^T = [U, \tilde{U}]\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}[U, \tilde{U}]^T.$$

(4) See [Demmel, GolubVanLoan] for more on this property. □

As we see, the SVD is closely related to the symmetric eigenvalue problem for a matrix. Computing the SVD accurately, is an intricate matter and before the modern approach[13] the SVD was not used, but from the few properties we have shown we see why it would be desirable. The solution of the symmetric eigenvalue problem is well beyond the scope of this paper, indeed a paper in itself, and we refer to in particular [Demmel, GolubVanLoan] for a detailed study of the eigenvalue problem leading up to the the stable modern SVD algorithms. For a less detailed but enlightening treatment refer to [TreBau, Björck].

A Naive Computation. None the less, for anyone familiar with basic eigenvalue theory, we give a brief description of a "naive" way, based on [TreBau]. Given a matrix A, we first form the s.p.d. $A^TA = H(= V\Sigma^T\Sigma V^T)$, and compute the *eigenvalues* (the roots λ_i of $p(\lambda) = \det(H - \lambda I) = 0$.) Then use the λ_i to form the *eigenvectors* $\mathbf{x}_i$ in $A\mathbf{x_i} = \lambda_i\mathbf{x}_i$ by $(A - \lambda_iI)\mathbf{x}_i = \mathbf{0}$. We know from Theorem 102 that all $\mathbf{x}_i$ will be orthogonal and $\lambda_i \geq 0$. It can be shown that the right and left eigenvectors are equal and arranging λ_i in descending magnitude $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with their corresponding $\mathbf{x}_i$ we can express $H = XDX^T = U\Sigma_HV^T$. Since $\sigma_{i,H} = \sigma_{i,A}^2$ we

[11]There are similar results for $m < n$.

[12]This is the Schur decomposition of *the nondefective (real) heremetian* matrix AA^T, and since Σ diagonal, the σ_i will be the eigenvalues. [Demmel]

[13]Formulated in: Golub, Gene H.; Kahan, William (1965). "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 and Golub, G. H.; Reinsch, C. (1970). "Singular value decomposition and least squares solutions". Numerische Mathematik 14 (5): 403–420.

take the $\sigma_{i,A} = |\sqrt{\sigma_{i,H}}|$. Then solve $U\Sigma = AV$ for orthogonal U (for example with the QR algorithm.)

4. Closest Point and Least Squares Problem

4.1. **Quadratic Functions.** We being studying quadratic functions, which are important in studying certain minimization problems.

We want to minimize[14] a real multivariate *quadratic function*

$$p(\mathbf{x}) = p(x_1, \ldots, x_n) = \sum_{i,j=1}^{n} k_{ij} x_i x_j - 2\sum_{i=1}^{n} f_i x_i + c = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c, \tag{4.1}$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ and $k_{ij}, f_i, c \in \mathbb{R}$, and assume that $k_{ij} = k_{ji}$ (the quadratic terms are symmetric) so that $K = (k_{ij})$ is a symmetric $n \times n$ (quadratic coefficient) matrix and $\mathbf{f}$ a constant vector. In order for this to have a minimum K must be *positive definite* (i.e. *K symmetric positive definite - s.p.d.*)

Proposition 103. *If $K > 0$ is s.p.d then the quadratic function, Equation* (4.1)*, has a unique minimizer, which is the solution to the linear system $K\mathbf{x} = \mathbf{f}$, namely $\mathbf{x}^* = K^{-1}\mathbf{f}$. The minimum value of $p(\mathbf{x})$ is equal to (any of)*

$$p(\mathbf{x}^*) = p(K^{-1}\mathbf{f}) = c - \mathbf{f}^T K^{-1}\mathbf{f} = c - \mathbf{f}^T \mathbf{x}^* = c - (\mathbf{x}^*)^T K \mathbf{x}^* \tag{4.2}$$

Proof. Suppose $\mathbf{x}^* = K^{-1}\mathbf{f}$ is the unique (the inverse is unique) solution to $K\mathbf{x} = \mathbf{f}$. Then for any $\mathbf{x} \in \mathbb{R}^n$

$$\begin{aligned} p(\mathbf{x}) =& \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T K \mathbf{x}^* + c \\ =& (\mathbf{x} - \mathbf{x}^*)^T K (\mathbf{x} - \mathbf{x}^*) + [c - (\mathbf{x}^*)^T K \mathbf{x}^*], \end{aligned}$$

where the symmetry of $K = K^T$ gives $\mathbf{x}^T K \mathbf{x}^* = (\mathbf{x}^*)^T K \mathbf{x}$. The second term in the final formula does not depend om $\mathbf{x}$. Moreover, the first term has the form $\mathbf{y}^T K \mathbf{y}$ where $\mathbf{y} = \mathbf{x} - \mathbf{x}^*$. Since we assumed that K is positive definite, $\mathbf{y}^T K \mathbf{y} \geq 0$ and vanishes iff $\mathbf{y} = \mathbf{x} - \mathbf{x}^* = \mathbf{0}$, which assumes minimum. Therefore the minimum of $p(\mathbf{x})$ occurs at $\mathbf{x} = \mathbf{x}^*$. The minimum value of $p(\mathbf{x})$ is equal to the constant term. The alternative expressions in Equation (4.2) follow from substitutions. □

Proposition 104. *If $K > 0$ is s.p.d., the the quadratic function $p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c$ has a unique global minimizer $\mathbf{x}^*$ satisfying $K\mathbf{x}^* = \mathbf{f}$. If $K \geq 0$ is positive semi-definite, and $\mathbf{f} \in \mathrm{rng}K$, then every solution $K\mathbf{x}^* = \mathbf{f}$ is a global minimum of $p(\mathbf{x})$. However, in the semi definite case, the minimum is not unique since $p(\mathbf{x}^* + \mathbf{z}) = p(\mathbf{x}^*)$ for any vector $\mathbf{z} \in \ker K$. In all other cases, there is no global minimum, and $p(\mathbf{x})$ can assume some arbitrarily large negative values.*

Proof. The first part is just Proposition 103. The second part uses that definite matrices have trivial kernels (but semi-definite do not.) If K is not semi-definite (and not positive definite, i.e. is negative definite), then one can find a vector $\mathbf{y}$ such that $a = \mathbf{y}^T K \mathbf{y} < 0$. Set $\mathbf{x} = t\mathbf{y}$ so that $p(\mathbf{x}) = p(t\mathbf{y}) = at^2 + 2bt + c$, with $b = \mathbf{y}^T \mathbf{f}$. Since $a < 0$, choosing $|t| \gg 0$ sufficiently large, one can arrange that $p(t\mathbf{y}) \ll 0$ is an arbitrarily large negative quantity. The remaining case is when K is positive semi-definite but $\mathbf{f} \notin \mathrm{rng}A$. □

[14] Maximizing a function $f(\mathbf{x})$ is the same as minimizing $-f(\mathbf{x})$.

4.2. **Closest Point or Distance to Subspace.** Given a point $\mathbf{b} \in \mathbb{R}^m$ and a subset $\mathcal{V} \subset \mathbb{R}^m$ we want to minimize the *distance* $d(\mathbf{b}, \mathbf{v}) = ||\mathbf{v} - \mathbf{b}||$ over possible $\mathbf{v}$, i.e. find the point $\mathbf{v}^* \in \mathcal{V}$ that is closest to $\mathbf{b}$. If $\mathbf{b} \in \mathcal{V}$, in the subspace, the distance is 0, and so we are left with studying the case when $\mathbf{b} \notin \mathcal{V}$.

We assume $\mathcal{V} \subset \mathbb{R}^n$ but any finite-dimensional subspace of any inner product space will do for the methods to come. It is common to assume that $||\cdot|| = ||\cdot||_2$ because the 2-norm relates it to the usual interpretation of length in Euclidian space. This will not be explicitly assumed here. Rather any norm coming from an inner product ($||\mathbf{v}|| = \sqrt{\langle \mathbf{v}; \mathbf{v} \rangle}$) will be assumed. This gives a simpler linear minimization problem in contrast to, for example the 1-norm or the ∞-norm, which give rise to nonlinear minimization problems ([OlvShaDraft]) which we will not deal with. In the end, all norms are, in some sense, analytically equivalent. From [Internet-blog] we adapt the following for completeness.

Lemma 105. *Suppose $\mathcal{X}$ is a n-dimensional normed space over $\mathbb{R}$ with basis $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. There exists a $c > 0$ such that $||\alpha_1\mathbf{x}_1 + \ldots + \alpha_n\mathbf{x}_n|| \geq c(|\alpha_1| + \ldots + |\alpha_n|)$ for any selection in $\mathcal{X}$.*

Proof. Let $s = |\alpha_1| + \ldots + |\alpha_n|$ and $\beta_i = \alpha_i/s$. The inequality $||\alpha_1\mathbf{x}_1 + \ldots + \alpha_n\mathbf{x}_n|| \geq c(|\alpha_1| + \ldots + |\alpha_n|)$ becomes $||\beta_1\mathbf{x}_1 + \ldots + \beta_n\mathbf{x}_n|| \geq c$ for all $\beta_1, \ldots, \beta_n$ satisfying $\sum_{i=1}^n |\beta_i| = 1$. If we suppose no such $c > 0$ exists, we can construct a sequence $(\mathbf{y}_m)$, where $\mathbf{y}_m = \sum_{i=1}^n \beta_i^{(m)}\mathbf{x}_i$ and $\sum_{i=1}^n |\beta_i^{(m)}| = 1$ for each m, and $||\mathbf{y}_m|| \to 0$. $\sum_{i=1}^n |\beta_i^{(m)}| = 1$ implies that the sequence $(\beta_i^{(m)})_{m\in\mathbb{N}}$ (where i is fixed) is bounded, and we must have a convergent subsequence. Apply this for $i = 1$, let the limit of that subsequence be β_1 and let $\mathbf{y}_{1,m}$ be the associated subsequence of the original $(\mathbf{y}_m)$ sequence. On that subsequence, apply again for $i = 2$, then on that subsequence again on $i = 3$ and so on until we have $\mathbf{y}_{n,m} = \sum_{i=1}^n \beta_i^{(m)}\mathbf{x}_i$.

Now note that each $\beta_i^{(m)} \to \beta$ (where β_i is the limit of the subsequence with the associated i from earlier.) This implies $\mathbf{y}_{n,m} \to \sum_{i=1}^n \beta_i\mathbf{x}_i = \mathbf{y}$. Since we required $\sum_{i=1}^n |\beta_i^{(m)}| = 1$ for each m earlier, we now have that $\sum_{i=1}^n |\beta_i| = 1$. This means that $\mathbf{y} \neq 0$, so the subsequence converged to non-zero element, which means that the original sequence can not converge to the zero element. We could only do this because we assumed no such $c > 0$ existed, so a $c > 0$ with the desired property must exist. □

Proposition 106. *We say two norms, $||\cdot||_a$ and $||\cdot||_b$, on the same vector space $\mathcal{X}$ are equivalent if there exist $c, C > 0$ such that for every $\mathbf{x} \in \mathcal{X}$, $c||\mathbf{x}||_a \leq ||\mathbf{x}||_b \leq C||\mathbf{x}||_a$. In a finite dimensional normed space all norms are equivalent.*

Proof. Suppose $\mathcal{X}$ is a n-dimensional space with basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and that $||\cdot||_a$ and $||\cdot||_b$ are any norms on $\mathcal{X}$. From Lemma 105 we know that for any $\mathbf{x} \in \mathcal{X}$ we can choose $\gamma > 0$ such that $||\alpha_1\mathbf{x}_1 + \ldots + \alpha_n\mathbf{x}_n||_a = ||\mathbf{x}||_a \geq \gamma(|\alpha_1| + \ldots + |\alpha_n|)$. If we consider the triangle inequality $||\mathbf{x}||_b \leq k\sum_{i=1}^n |\alpha_1|$, where $k = \max(||\mathbf{e}_1||_b, \ldots, ||\mathbf{e}_n||_b)$, applying the earlier inequality we get $||\mathbf{x}||_b \leq k||\mathbf{x}||_a/\gamma = C||\mathbf{x}||_a$.

If we reverse $||\cdot||_a$ and $||\cdot||_b$ and repeat the process we obtain $||\mathbf{x}||_b \geq c||\mathbf{x}||_a$. $\mathbf{x} \in \mathcal{X}$, $||\cdot||_a$ and $||\cdot||_b$ were all arbitrary, and k, γ were not depend on the choice of $\mathbf{x}$. Therefore all norms on $\mathcal{X}$ are equivalent. □

With that behind us we can comfortably proceed with another "geometric" result, which turns out to be important, a recurring theme, in binding together the theory of matrix equations and their solutions.

Theorem 107. *Let $\mathbf{v}_1, \ldots \mathbf{v}_n$ form a basis for the subspace $\mathcal{V} \subset \mathbb{R}^m$. Given $\mathbf{b} \in \mathbb{R}^m$, the closest point $\mathbf{v}^* = x_1^*\mathbf{v}_1 + \ldots + x_n^*\mathbf{v}_n \in \mathcal{V}$ is prescribed by the solution*

$\mathbf{x}^* = K^{-1}\mathbf{f}$ *to the linear system* $K\mathbf{x} = \mathbf{f}$*, where* $K(i,j) = k_{ij} = \langle \mathbf{v}_i; \mathbf{v}_j \rangle$ *and* $\mathbf{f} = \langle \mathbf{v}_i; \mathbf{b} \rangle$*. The distance between the point and the subspace is*

$$||\mathbf{v}^* - \mathbf{b}|| = \sqrt{||\mathbf{b}||^2 - \mathbf{f}^T\mathbf{v}^*}. \tag{4.3}$$

Proof. The closest point is found by minimizing the distance to the subspace i.e.

$$||\mathbf{v} - \mathbf{b}||^2 = ||\mathbf{v}||^2 - 2\langle \mathbf{v}; \mathbf{b} \rangle + ||\mathbf{b}||^2 \tag{4.4}$$

over all possible $\mathbf{v} \in \mathcal{V} \subset \mathbb{R}^m$. If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis for $\mathcal{V}$ (so that $\dim\mathcal{V} = n$) then any $\mathbf{v} \in \mathcal{V}$ is a linear combination of these. So $\mathbf{v} = x_1\mathbf{v}_1 + \ldots + x_n\mathbf{v}_n$ inserted into Equation (4.4) gives that

$$||\mathbf{v}||^2 = \langle \mathbf{v}; \mathbf{v} \rangle = \sum_{i,j=1}^{n} x_i x_j \langle \mathbf{v}_i; \mathbf{v}_j \rangle = \sum_{i,j}^{n} k_{ij} x_i x_j = \mathbf{x}^T K \mathbf{x} \tag{4.5}$$

Since $K(i,j) = k_{ij} = \langle \mathbf{v}_i; \mathbf{v}_j \rangle$ and inner products are symmetric $k_{ji} = k_{ij}$, K is the symmetric $n \times n$ Gram matrix (also see the discussion following Equation (4.1).) In addition

$$\langle \mathbf{v}; \mathbf{b} \rangle = \langle x_1\mathbf{v}_1 + \ldots + x_n\mathbf{v}_n; \mathbf{b} \rangle = \sum_{i=1}^{n} x_i \langle \mathbf{v}_i; \mathbf{b} \rangle = \sum_{i=1}^{n} x_i f_i = \mathbf{x}^T\mathbf{f}, \tag{4.6}$$

where $\mathbf{f} = (f_1 \ldots, f_n,) \in \mathbb{R}^n$ and $f_i = \langle \mathbf{v}_i; \mathbf{b} \rangle$. Putting $||\mathbf{b}||^2 = c$ and substituting this, and equations (4.5) and (4.6) into Equation (4.4) gives

$$\begin{aligned} ||\mathbf{v} - \mathbf{b}||^2 &= ||\mathbf{v}||^2 - 2\langle \mathbf{v}; \mathbf{b} \rangle + ||\mathbf{b}||^2 = \sum_{i,j}^{n} k_{ij} x_i x_j - 2\sum f_i x_i + c \\ &= \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T\mathbf{f} + c = p(\mathbf{x}). \end{aligned}$$

Since the basis of $\mathbf{v}$ is are linearly independent Proposition 71 ensures that the Gram matrix $K = A^TA$ is positive definite and we can apply Proposition 103 to solve the closest point problem (Equation (4.4)) and we get the result. □

4.3. **Theory of Least Squares.**

4.3.1. *Overdetermined Systems.* We now study the "solution" to equations $A\mathbf{x} = \mathbf{b}$ when it does not have an solution, i.e. the solution is not in the range of A, for example in the case when there are more equations than unknowns ($m > n$). Though there might not exist a solution, there should exist a "best" solution $\mathbf{x}^*$, that most closely matches a true solution. The task is to minimize the *residual* $\mathbf{r}(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$, or rather the norm of the residual $||\mathbf{r}(\mathbf{x})|| = ||A\mathbf{x} - \mathbf{b}||$. It is possible to use different norms, but if we choose the 2-norm we get the *least squares solution*, which makes sense for the same reasons as before.

Definition 108. The *least squares solution* to a linear system of equations $A\mathbf{x} = \mathbf{b}$ is the vector $\mathbf{x}^* \in \mathbb{R}^n$ that minimizes the Euclidean norm $||A\mathbf{x} - \mathbf{b}|| = ||\mathbf{r}||$.

Theorem 109. *Assume* $\ker A = \{\mathbf{0}\}$*. Set* $K = A^TA$ *and* $\mathbf{f} = A^T\mathbf{b}$*. Then the least squares solution to* $A\mathbf{x} = \mathbf{b}$ *is the unique solution to the normal equations*

$$K\mathbf{x} = \mathbf{f} \quad \text{or} \quad (A^TA)\mathbf{x} = A^T\mathbf{b}, \tag{4.7}$$

namely[15]

$$\mathbf{x}^* = (A^TA)^{-1}A^T\mathbf{b}. \tag{4.8}$$

The least squares error is $||A\mathbf{x}^* - \mathbf{b}||^2 = ||\mathbf{b}||^2 - \mathbf{f}^T\mathbf{x}^* = ||\mathbf{b}||^2 - \mathbf{b}^TA(A^TA)^{-1}A^T\mathbf{b}$.

[15] $(A^TA)^{-1}A^T$ is called the *pseudoinverse* A^+)

Proof. Let $\mathcal{V} = \text{rng}A \subset \mathbb{R}^m$, the range of the column space of A. If the columns are linearly independent (required for the solution to be unique) they form a basis $\mathbf{v}_1, \ldots, \mathbf{v_n} \in \mathbb{R}^m$ for the range $\mathcal{V}$, i.e. $A = (\mathbf{v}_1 \ \ldots \ \mathbf{v_n})$. So every element in the range can be written as $\mathbf{v} = A\mathbf{x}$ and therefore minimizing it is the same as minimizing the distance $||A\mathbf{x} - \mathbf{b}|| = ||\mathbf{v} - \mathbf{b}||$ between point and subspace. Since the minimizer $\mathbf{x}^*$ is both the least square solution and the closest point $\mathbf{v}^* = A\mathbf{x}^* \in \mathcal{V}$, the least square solution follows from Theorem 107.

For vectors $\mathbf{v}, \mathbf{w} \in$ euclidean $\mathbb{R}^n$ the dot product can be identified with the matrix product so $\langle \mathbf{v}; \mathbf{w} \rangle = \mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T\mathbf{w}$. Therefore (Equation (2.6)), under Euclidean inner product the entries of the (Gram) matrix K and the vector $\mathbf{f}$ are given by $k_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j = \mathbf{v}_i^T\mathbf{v}_j (= k_{ji})$ and $f_i = \mathbf{v}_i \cdot \mathbf{b} = \mathbf{v}_i^T\mathbf{b}$ and hence we can write $K = A^TA$, $\mathbf{f} = A^T\mathbf{b}$ to express $K\mathbf{x} = \mathbf{f}$ and Equation (4.3) of Theorem (107) as stated, and the equations follow. □

Orthogonal Least Square Connection.

Theorem 110. *Let $\mathcal{W} \subset \mathcal{V}$ be a finite-dimensional subspace of an inner product space. Given a vector $\mathbf{v} \in \mathcal{V}$, the closest point or least squares minimizer $\mathbf{w} \in \mathcal{W}$ is the same as the orthogonal projection of $\mathbf{v}$ onto $\mathcal{W}$.*

Proof. Let $\mathbf{w} \in \mathcal{W}$ be the orthogonal projection of $\mathbf{v}$ onto the subspace, which requires that the difference $\mathbf{z} = \mathbf{v} - \mathbf{w}$ be orthogonal to W. Suppose $\tilde{\mathbf{w}} \in \mathcal{W}$ is any other vector in the subspace. Then

$$||\mathbf{v} - \tilde{\mathbf{w}}||^2 = ||\mathbf{w} + \mathbf{z} - \tilde{\mathbf{w}}||^2 = ||\mathbf{w} - \tilde{\mathbf{w}}||^2 + 2\langle \mathbf{w} - \tilde{\mathbf{w}}; \mathbf{z} \rangle + ||\mathbf{z}||^2 = ||\mathbf{w} - \tilde{\mathbf{w}}||^2 + ||\mathbf{z}||^2.$$

The inner product term vanishes (= 0) because $\mathbf{z}$ is orthogonal to every vector in $\mathcal{W}$, including $\mathbf{w} - \tilde{\mathbf{w}}$. Since $\mathbf{z} = \mathbf{v} - \mathbf{w}$ is uniquely prescribed by the vector $\mathbf{v}$, the second term $||\mathbf{z}||^2$ does not change with the choice of the point $\tilde{\mathbf{w}} \in \mathcal{W}$. Therefore $||\mathbf{v} - \tilde{\mathbf{w}}||^2$ will be minimized iff $||\mathbf{w} - \tilde{\mathbf{w}}||^2$ is minimized. Since $\tilde{\mathbf{w}} \in \mathcal{W}$ is allowed to be any element of the subspace $\mathcal{W}$, the minimal value $||\mathbf{w} - \tilde{\mathbf{w}}||^2 = 0$ occurs when $\tilde{\mathbf{w}} = \mathbf{w}$. Thus, the closest point $\tilde{\mathbf{w}}$ coincides with the orthogonal projection $\mathbf{w}$. □

4.3.2. *Underdetermined (Rank Deficient) Systems.* An underdetermined problem is a rank deficient problem, when a matrix has more linearly independent columns than equations. There is clearly no unique solution to such a problem.

Lemma 111. *Let $A \in \mathbb{R}^{m \times n}$, with $m \geq n$ and* rank$A = r < n$. *Then there is a $n - r$ dimensional set of vectors that minimize $||A\mathbf{x} - \mathbf{b}||_2$.*

Proof. Let $A\mathbf{z} = \mathbf{0}$. If $\mathbf{x}$ minimizes $||A\mathbf{x} - \mathbf{b}||_2$, so does $\mathbf{x} + \mathbf{z}$. □

So there is no unique least squares solution to an underdetermined system. In practice, due to roundoff in the entries, A will often have one or more very small computed singular values and the system will be near singular, rather than singular. So despite singularity, we can often obtain a unique solution nonetheless, but it is likely to be near singular, very large and sensitive to errors in $\mathbf{b}$.

Proposition 112. *Let $\sigma_{\min}$ the smallest singular value of A. Assume $\sigma_{\min} > 0$. Then*

1. if $\mathbf{x}$ minimizes $||A\mathbf{x} - \mathbf{b}||_2$, then $||\mathbf{x}||_2 \geq |\mathbf{u}_n^T\mathbf{b}|/\sigma_{\min}$ where $\mathbf{u}_n$ is the last column of U in $A - U\Sigma V^T$.

2. changing $\mathbf{b}$ to $\mathbf{b} + \delta\mathbf{b}$ can change $\mathbf{x}$ to $\mathbf{x} + \delta\mathbf{x}$, where $||\delta\mathbf{x}||_2$ is as large as $||\delta\mathbf{b}||_2/\sigma_{\min}$.

Hence, if A is nearly rank deficient ($\sigma_{\min}$ small) the solution $\mathbf{x}$ is ill-conditioned and possibly large.

Proof. (1) $\mathbf{x} = A^+\mathbf{b} = V\Sigma^{-1}U^T\mathbf{b}$, so $||\mathbf{x}||_2 = ||\Sigma^{-1}U^T\mathbf{b}||_2 \geq |(\Sigma^{-1}U^T\mathbf{b})_n| = |\mathbf{u}_n^T\mathbf{b}|/\sigma_{min}$. (2) choose $\delta\mathbf{b}$ parallel to $\mathbf{u}_n$ in previous. □

Proposition 113. *When A is exactly singular, the $\mathbf{x}$ that minimize $||A\mathbf{x}-\mathbf{b}||_2$ can be characterized as follows. Let $A = U\Sigma V^T$ have rank $r < n$, and write the SVD of aA as*

$$A = [U_1, U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} [V_1, V_2]^T = U_1 \Sigma_1 V_1^T, \tag{4.9}$$

where Σ_1 is $r \times r$ and nonsingular and U_1 and V_1 have r columns. Let $\sigma = \sigma_{min}(\Sigma_1)$, the smallest nonzero singular value of A. Then
1. All solutions $\mathbf{x}$ can be written $\mathbf{x} = V_1\Sigma_1^{-1}U_1^T\mathbf{b} + V_2\mathbf{z}$, $\mathbf{z}$ an arbitrary vector.
2. The solution $\mathbf{x}$ has minimal norm $||\mathbf{x}||_2$ precisely when $\mathbf{z} = \mathbf{0}$, in which case $\mathbf{x} = V_1\Sigma_1^{-1}U_1^T\mathbf{b}$ and $||\mathbf{x}||_2 \le ||\mathbf{b}||_2/\sigma$.
3. Changing $\mathbf{b}$ to $\delta\mathbf{b}$ can change the minimal norm solution $\mathbf{x}$ by at most $||\delta\mathbf{b}||_2/\sigma$.

In other words, the norm and condition number of the unique minimal norm solution $\mathbf{x}$ depends on the smallest nonzero singular value of A.

Proof. Choose $\tilde{U}$ so $\left[U, \tilde{U}\right] = \left[U_1, U_2, \tilde{U}\right]$ is an $m \times m$ orthogonal matrix. Then

$$||A\mathbf{x}-\mathbf{b}||_2^2 = ||[U,\tilde{U}]^T(A\mathbf{x}-\mathbf{b})||_2^2 = \left\| \begin{bmatrix} U_1^T \\ U_2^T \\ \tilde{U}^T \end{bmatrix} (U_1\Sigma_1V_1^T\mathbf{x}-\mathbf{b}) \right\|_2^2 =$$

$$= \left\| \begin{bmatrix} \Sigma_1V_1^T\mathbf{x} - U_1^T\mathbf{b} \\ U_2^T\mathbf{b} \\ \tilde{U}^T\mathbf{b} \end{bmatrix} \right\|_2^2 = ||\Sigma_1V_1^T\mathbf{x} - U_1^T\mathbf{b}||_2^2 + ||U_2^T\mathbf{b}||_2^2 + ||\tilde{U}^T\mathbf{b}||_2^2.$$

(1) $||A\mathbf{x}-\mathbf{b}||_2$ is minimized when $\Sigma_1V_1^T\mathbf{x} = U_1^T\mathbf{b}$, or $\mathbf{x} = V_1\Sigma_1^{-1}U_1^T\mathbf{b} + V_2\mathbf{z}$ since $V_1^TV_2\mathbf{z} = \mathbf{0}$ for all $\mathbf{z}$.
(2) Since the columns of V_1 and V_2 are mutually orthogonal, the Pythagorean theorem implies that $||\mathbf{x}||_2^2 = ||V_1\Sigma_1^{-1}U_1^T\mathbf{b}||_2^2 + ||V_2\mathbf{z}||_2^2$, and this is minimized by $\mathbf{z} = \mathbf{0}$.
(3) Changing $\mathbf{b}$ by $\delta\mathbf{b}$ changes $\mathbf{x}$ by at most $||V_1\Sigma_1^{-1}U_1^T\delta\mathbf{b}||_2 \le ||\Sigma_1^{-1}||_2||\delta\mathbf{b}||_2 = ||\delta\mathbf{b}||_2/\sigma$. □

This tells us that there is at least a minimum norm solution $\mathbf{x}$, which is unique and that may be well-conditioned if the smallest singular value is not to small.

Definition 114. Let $A = U\Sigma V^T = U_1\Sigma_1V_1^T$ as in Equation (4.9). Then $A^+ \equiv V_1\Sigma_1^{-1}U_1^T$, or also written as $A^+ = V^T\Sigma^+U$, where $\Sigma^+ = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}^+ = \begin{bmatrix} \Sigma_1^{-1} & 0 \\ 0 & 0 \end{bmatrix}$, is the *Moore-Penrose pseudoinverse* for the possibly rank-deficient problem A. The Moore-Penrose pseudoinverse is also written as $A^\dagger$.

So we find that the solution to the least squares problem is always $\mathbf{x} = A^\dagger\mathbf{b}$ and, "even" when A is rank deficient, $\mathbf{x}$ has minimum norm out of the possible solutions.

But from ([Björck] p. 26, Theorem 1.4.1) we get the following:

Proposition 115. *If* $\text{rank}(A+E) \neq \text{rank}A$ *(where E is a perturbation) then* $||(A+E)^+ + A^+|| \ge 1/||E||_2$.

The point of this proposition is that $A^\dagger$ can vary discontinuously when rank$A$ changes (because from Lemma 97 this is related to $\sigma_{\min}$.) When the rank changes, the change in $A^\dagger$ may be unbounded when $||E||_2 \to 0$. This matters for rank-deficient problems, since they are ill-conditioned. Therefore it is very important to know what rank we are operating with.

Definition 116. A matrix A is said to have *numerical δ-rank* equal to k if $k = \min\{\text{rank}B \ ||A-B||_2 \le \delta\}$, where $B = A + E$ and E is a perturbation.

From Proposition 99 we see that if $k < n$ then $\inf_{\text{rank}B \leq k} ||A - B||_2 = \sigma_{k+1}$, and our low-rank approximation becomes $A = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, which generally implies that a matrix A has numerical rank (δ-rank) k iff $\sigma_1 \geq \ldots \geq \sigma_k > \delta > \sigma_{k+1} \geq \ldots \geq \sigma_n$. The selection of δ can be difficult, but a simple and reasonable way to go about it would be to say that if the coefficients of A are accurate within $\pm\varepsilon$ we put $\delta = \varepsilon$, any smaller value would be outside the accuracy.

It follows that $||A - A_k||_2 = ||A\tilde{V}||_2 \leq \delta$, $\tilde{V} = \{\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$ and $\mathcal{N}(\tilde{V}) = \text{span}(\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n)$ is the *numerical nullspace* (kernel) of A.

5. Solving Linear Systems of Equations

We now summarize how to solve systems of linear equations $A\mathbf{x} = \mathbf{b}$ using the methods we have developed.

Forward and Backward Substitution. Forward substitution is used to solve a lower triangular system $L\mathbf{x} = \mathbf{b}$, and solving for x_i is given by $x_i = {}^{(b_i - \sum_{j=1}^{i-1} l_{ij} x_j)}/_{l_{ii}}$. *Backward substitution* is used to solve the upper triangular system $U\mathbf{x} = \mathbf{b}$, and solving for x_i is given by $x_i = {}^{(b_i - \sum_{j=i+1}^{n} u_{ij} x_j)}/_{u_{ii}}$.

5.1. **Solving the Nonsingular Problems.** Theoretically this problem can be solved by computing and applying the inverse to both sides, but as we have said this is not done in practice. We refer to any of [Demmel, OlvShaDraft, TreBau, GolubVanLoan] or other texts on numerical analysis for the details but give two intuitive motivations. (1) computing the inverse is done by Gauss-Jordan elimination (using Gauss elimination and then more work to finalize the reduction) followed by the added work of forming $\mathbf{x} = A^{-1}\mathbf{b}$; (2) it is prone to numerical instability, possibly leading to rank deficiency in inexact arithmetic. If $\left[\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 10^4 & 0 & 10^{-3} \end{array}\right]$ is one step in the augmented Gauss-Jordan system, the final step would be to divide the 2^{nd} row by 10^4 to obtain $[I|A^{-1}]$, but $10^{-3} \cdot 10^{-4} = 10^{-7} = 0$ in three digit accuracy, making A^{-1} singular. This was just a simple example but we know that A^{-1} is undefined in this situation, though it is not clear how our new methods would deal with this problem better, we can see the inverse is no good.

5.1.1. *LU Solution.* In Section 3.1 we showed the existence of the partially pivoted LU factorization - GEPP.

If the $m \times m$ matrix A is nonsingular, there exists a permutation matrix P, a nonsingular lower triangular matrix L, and a nonsingular upper triangular matrix U such that $P^T A = LU \Leftrightarrow A = PLU$. To solve $A\mathbf{x} = \mathbf{b}$, we solve the equivalent system $PLU\mathbf{x} = \mathbf{b}$ as follows:

(1) $LU\mathbf{x} = P^{-1}\mathbf{b} = P^T\mathbf{b}$ (permute entries of $\mathbf{b}$)
(2) $U\mathbf{x} = L^{-1}(P^T\mathbf{b})$ (forward substitution)
(3) $\mathbf{x} = U^{-1}(L^{-1}P^T\mathbf{b})$ (backward substitution)

Putting this together the equation $A\mathbf{x} = \mathbf{b} \Leftrightarrow PA = LU\mathbf{x} = \mathbf{b}$. This is then first reduced to $L\mathbf{y} = \mathbf{b}$, giving $\mathbf{y} = U\mathbf{x} = L^{-1}\mathbf{b} = \mathbf{c}$ (where L^{-1} is only symbolic for the change of basis operation preformed by forward substitution). Then backward substitution gives that $U\mathbf{x} = \mathbf{c}$ becomes $\mathbf{x} = U^{-1}\mathbf{c}$. All this amounts to $\mathbf{x} = U^{-1}L^{-1}\mathbf{b}$.

5.1.2. *QR factorization.* Using QR factorization to solve $A\mathbf{x} = \mathbf{b}$ is straight forward.

(1) Compute QR factorization $A = QR$ so that $QR\mathbf{x} = \mathbf{b}$.
(2) Since Q is orthonormal, $R\mathbf{x} = Q^T\mathbf{b}$.
(3) Since R is upper triangular solve for $\mathbf{x}$ with back substitution.

The resulting algorithm is numerically more stable for some (ill-conditioned) matrices than LU factorization, but needs more work ([Demmel]).

5.1.3. *SVD Solution.* Forming the SVD is laborious but using it is simple.

Proposition 117. *If A has full rank, the solution of $\min_x ||A\mathbf{x} - \mathbf{b}||_2$ is $\mathbf{x} = V\Sigma^{-1}U^T\mathbf{b}$.*

Proof. This is just a special case of Proposition 113 where we put $\mathbf{z} = \mathbf{0}$. □

We know A has full rank so we get:

(1) Compute $A = U\Sigma V^T$ the SVD.
(2) Compute the vector $U^T\mathbf{b}$.
(3) Solve the diagonal system $\Sigma\mathbf{w} = U^T\mathbf{b}$ for $\mathbf{w}$.
(4) Set $\mathbf{x} = V\mathbf{w}$.

5.2. **Least Squares - Overdetermined Systems.** A very common situation in many applications are overdetermined systems, having more equations than unknowns i.e. a matrix $A \in \mathbb{R}^{m\times n}$, $m > n$, of rank$A = n$ (full rank.) Usually the problem is then to determine the coefficients of the equation that best fits the data, i.e. that minimizes the error $\mathbf{r} = A\mathbf{x} - \mathbf{b} = A(\mathbf{x} - \mathbf{x}^*)$, where $\mathbf{x}$ is the computed solution and $\mathbf{x}^*$ is the true solution.

5.2.1. *Normal Equations Solution to Least Squares.* The *normal equations least squares method* first forms the s.p.d. matrix A^TA and then computes the Cholesky factorization of this.

Assume $A \in \mathbb{R}^{m\times n}$ with rank$A = n$.

(1) Form the matrix A^TA and the vector $A^T\mathbf{b}$.
(2) Compute the Cholesky factorization $A^TA = R^TR$
(3) Solve the lower triangular system $R^T\mathbf{w} = A^T\mathbf{b}$ using forward substitution for $\mathbf{w}$.
(4) Solve the upper triangular system $R\mathbf{x} = \mathbf{w}$ using back substitution.

Note. The Cholesky factorization is effective and "inexpensive" to compute, but the formation of A^TA can lead to problems for matrices that are, for example, near rank deficient. It is possible to show that in the formation of A^TA, the accuracy will be negatively affected and how much ([Demmel, Björck]). This motivates the QR method as the standard least square solver for all but well-conditioned problems.

5.2.2. *QR Method Solution.* The usual approach to solving the least squares problem is the $\hat{Q}\hat{R}$ factorization, via Gram-Schmidt or Householder reflections. We saw that the orthogonal projector P can be written $P = \hat{Q}\hat{Q}^T$ and we have $\mathbf{y} = P\mathbf{b} = \hat{Q}\hat{Q}^T\mathbf{b}$. Since $\mathbf{y} \in$ rng$A$ the system $A\mathbf{x} = \mathbf{y}$ has an exact solution. Combining $\mathbf{y} = \hat{Q}\hat{Q}^T\mathbf{b}$ and the QR factorization gives $\hat{Q}\hat{R}\mathbf{x} = \hat{Q}\hat{Q}^T\mathbf{b} \quad \Rightarrow \quad \hat{R}\mathbf{x} = \hat{Q}^T\mathbf{b}$, with upper-triangular $\hat{R}$ which is solved via backward substitution. If A has full rank, the system is nonsingular.

Another way to derive this is $A^TA = A^T\mathbf{b}$ from the normal equations. Substituting the QR factorization into this gives $\hat{R}^T\hat{Q}^T\hat{Q}\hat{R}\mathbf{x} = \hat{R}^T\hat{Q}^T\mathbf{b}$, which implies $\hat{R}\mathbf{x} = \hat{Q}^T\mathbf{b}$.

(1) Compute the reduced QR factorization $A = \hat{Q}\hat{R}$.
(2) Compute the vector $\hat{Q}^T\mathbf{b}$.
(3) Solve upper-triangular system $\hat{R}\mathbf{x} = \hat{Q}^T\mathbf{b}$ for $\mathbf{x}$.

5.2.3. *SVD Solution.* Let $A = \hat{U}\hat{\Sigma}V^T$. Now we can write the projector $P = \hat{U}\hat{U}^T$, so that $\mathbf{y} = P\mathbf{b} = \hat{U}\hat{U}^*\mathbf{b}$, giving $A\mathbf{x} = \hat{U}\hat{\Sigma}V^T\mathbf{x} = \hat{U}\hat{U}\mathbf{b}$ and $\hat{U}\hat{\Sigma}\mathbf{x} = \hat{U}^T\mathbf{b}$.

(1) Compute the reduced SVD $A = \hat{U}\hat{\Sigma}V^T$.
(2) Compute the vector $\hat{U}^T\mathbf{b}$.
(3) Solve the diagonal system $\hat{\Sigma}\mathbf{w} = \hat{U}^T\mathbf{b}$ for $\mathbf{w}$.
(4) Set $\mathbf{x} = V\mathbf{w}$.

5.3. **Solving the Rank-Deficient Problem.** We saw that the underdetermined, rank-deficient, problem does not have a unique solution, but rather infinitely many solutions. We also saw that for rank deficient problems, due to round-off, new singular values (i.e. ranks) could be introduced but that they would be near-singular and that this would "blow up" the solution. The way to deal with this is called *regularization* (improving poor conditioning.) A meaningful discussion on regularization or the preferred non-SVD method, *column pivoted QR*, is once more beyond the scope of this paper[16] but [Björck], and to some extent [GolubVanLoan], deal extensively with the rank deficient problem. Even so, given our discussions in this paper, we briefly end with the two most obvious ways.

5.3.1. *SVD.*

Naive Approach. The naive approach to solving rank deficient $A\mathbf{x} = \mathbf{b}$ is to compute $A = U\Sigma V^T$, form $A^\dagger$ and then obtain $\mathbf{x}$ from $\mathbf{x} = A^\dagger\mathbf{b}$.

As we have seen this will often be numerically unstable and unusable, as the computed rank and original rank may differ.

Truncated SVD Solution. In Proposition 99 we studied the low rank SVD. We can think of it as a way to regularize A.

Assume that $A \in \mathbb{R}^{m\times n}$ is rank-deficient with rank$A = k(= \delta\text{-rank})$. Assume $A = U\Sigma V^T = \sum_{i=1}^{n} u_i\sigma_i v_i^T$ is the SVD of A.

(1) Given δ-rank $= k$, we put all $\sigma_{i>k} = 0$, and our best rank k approximation is $A_k = \sum_{i=1}^{k} \mathbf{u}_i\sigma_i\mathbf{v}_i^T$. The least squares problem becomes $\min_{\mathbf{x}} ||A_k\mathbf{x} - \mathbf{b}||_2$,
(2) The *truncated SVD* (TSVD) solution is $\mathbf{x} = \sum_{i=1}^{k} c_i\mathbf{v}_i/\sigma_i$, $\mathbf{c} = U^T\mathbf{b}$, where $\mathbf{v}_i$ are the right singular vectors.

[16]It would require a more extensive study of errors, stability and conditioning. Simply determining the δ-rank to use is not always straight forward.

Appendix A. On Numerical Analysis

The introduction of errors (inexactness) leads to a range of complexities. If we want to compute and use a result, usually dependent on many intermediate steps, it is necessary to examine what happens to these errors in each step, how they propagate, to know the final error. Different methods, formulations of the same problem, can lead to different results since they take different paths. Practical aspect of simply implementing the different methods, the algorithms, also need to be taken into consideration. Resource balancing (time, accuracy, computer memory and much more) forces different approaches.

The following informal discussions is based mainly on [Demmel, TreBau] but can be found in most literature on numerical analysis.

A.1. **Perturbations, Conditions Numbers and Errors.** Numerical algorithms rarely give exactly correct answers, with broadly two sources of errors. Errors in input (e.g. measurement errors) and errors from the approximations the algorithm does itself.

Let $x, \delta x, f(x) \in \mathbb{R}$, where δx is a small *perturbation* (change) in data. The expression $e_a = |f(x+\delta x) - f(x)|$ is called the *absolute error*, while $e_r = |f(x+\delta x) - f(x)|/|f(x)|$ is called the *relative error*.

It is desirable to bound the resulting error, to know how bad things can get. Using a linear approximation of the Taylor expansion $f(x+\delta x) - f(x) \approx f(x) + \delta x f'(x) \Rightarrow |f(x+\delta) - f(x)| \approx |\delta x| \cdot |f'(x)|$ we get an approximation of the bound. The term $|f'(x)|$ is called the *absolute condition number* of f at x, and bounds the *absolute error* given a error bound on the perturbation. Similarly bounding the relative error

$$\frac{|f(x+\delta x) - f(x)|}{|f(x)|} \approx \frac{|\delta x|}{|x|} \cdot \frac{|f'(x)| \cdot |x|}{|f(x)|},$$

the term $|f'(x)| \cdot |x|/|f(x)|$ is called the *relative condition number*.

Definition. If $\text{alg}(x)$ is an algorithm for $f(x)$, including the effects of roundoff, we call $\text{alg}(x)$ a *backward stable algorithm* for $f(x)$ if for all x there is a "small" δx such that $\text{alg}(x) = f(x+\delta x)$. δx is called the *backward error*.

Informally, we say that we get the exact answer $f(x+\delta x)$ for a slightly wrong problem $(x+\delta x)$. It implies that we may bound the error $|\text{alg}(x) - f(x)| = |f(x+\delta x) - f(x)| \approx |f'(x)| \cdot |\delta x|$.

We extend this discussion to multivariate functions, matrices in particular, of which our previous findings are a special case. Let $\delta\mathbf{x}$ be a small *perturbation* (change) in $\mathbf{x}$, and $\delta f = f(\mathbf{x}+\delta\mathbf{x}) - f(\mathbf{x})$. Remembering Proposition 106, we will simply assume that $||\cdot||$ is any norm and work with the corresponding induced norms.

Definition. The *absolute condition number* $\kappa_a = \kappa_a(\mathbf{x})$, of κ_a at $\mathbf{x}$, is

$$\kappa_a = \lim_{\delta \to 0} \sup_{||\delta \mathbf{x}|| \leq \delta} \frac{||\delta f||}{||\delta \mathbf{x}||},$$

and the *relative condition number* $\kappa(\mathbf{x}) = \kappa$, of κ at $\mathbf{x}$, is

$$\kappa = \lim_{\delta \to 0} \sup_{||\delta \mathbf{x}|| \leq \delta} \left(\frac{||\delta f||}{||f(\mathbf{x})||} \Big/ \frac{||\delta f||}{||\delta \mathbf{x}||} \right).$$

We call κ the *condition number*, and we say that a problem is *well-conditioned* if κ is small. A problem is *ill-conditioned* if κ is large.

Note. Usually $\lim_{\delta\to 0}\sup_{||\delta\mathbf{x}||\leq\delta}$ is simplified as $\sup_{\delta\mathbf{x}}$ when we are talking about the supremum of all $\delta\mathbf{x}\to\mathbf{0}$. We will assume this. Also if f is differentiable then we can express $\delta f=\delta f(\mathbf{x})$ in terms of the Jacobian $J(\mathbf{x})$, where $J_{ij}(\mathbf{x})=\partial f_i(\mathbf{x})/\partial x_i$, as $\lim_{||\delta\mathbf{x}||\to\mathbf{0}}\delta f=J\delta\mathbf{x}$. Then $\kappa_a=||J(\mathbf{x})||$ and $\kappa=(||J(\mathbf{x})||\cdot||\mathbf{x}||)/||f(\mathbf{x})||$.

From the definition we gather that a well-conditioned problem has the property that a small $\delta\mathbf{x}$ leads to only a "small" perturbation in $f(\mathbf{x})$, but to a "large" perturbation in $f(\mathbf{x})$ for an ill-conditioned problem. "Small" and "large" depends on the situation. $\kappa=1$ is good, since we have a one-to-one correspondence, whereas $\kappa\approx 10^9$ is bad since it roughly corresponds to a 9 digit loss of accuracy.

If $||\cdot||_2=||\cdot||$ then $||A||=\sigma_1$ and $||A^{-1}||=1/\sigma_{\min}$ and $\kappa(A)=\sigma_1/\sigma_{\min}$ [TreBau]. We set $\kappa(A)=\infty$ if A is singular (since $\sigma_{min}=0$.)

Conditioning of $A\mathbf{x}$ and $A^{-1}\mathbf{b}$. Let $A\in\mathbb{R}^{m\times n}$. From the definition we get

$$\kappa=\sup_{\delta\mathbf{x}}\left(\frac{||A(\mathbf{x}+\delta\mathbf{x})||}{||A\mathbf{x}||}\Big/\frac{||\delta\mathbf{x}||}{||\mathbf{x}||}\right)=\sup_{\delta\mathbf{x}}\left(\frac{||A\delta\mathbf{x}||}{||\delta\mathbf{x}||}\Big/\frac{||A\mathbf{x}||}{||\mathbf{x}||}\right)=||A||\frac{||\mathbf{x}||}{||A\mathbf{x}||}.$$

If A is invertible then we get that $||\mathbf{x}||/||A\mathbf{x}||\leq||A^{-1}||$ (see [TreBau].) To remove the dependence on $\mathbf{x}$, we write

$$\kappa=||A||\cdot||x||/||Ax||\leq||A||\cdot||A^{-1}||.$$

Later we will provide a result that states that $||A||\cdot||A^{-1}||=\kappa(A)$ under certain conditions. For the inverse problem, $A^{-1}\mathbf{b}$, we simply replace A by A^{-1} and the result is identical. If A is not invertible but has full rank we can replace A^{-1} with A^{+} (from Theorem 109.)

Conditioning of a System of Equations. We now perturb the coefficient matrix A by an infinitesimal δA and get $(A+\delta A)(\mathbf{x}+\delta\mathbf{x})=\mathbf{b}$. We drop the $\delta A\cdot\delta\mathbf{x}\approx 0$ for $(\delta A)\mathbf{x}+A(\delta\mathbf{x})$ which is ≈ 0 because $\delta A,\delta\mathbf{x}\approx 0$ and $\delta\mathbf{x}=-A^{-1}(\delta A)\mathbf{x}\Rightarrow||\delta\mathbf{x}||\leq||A^{-1}||\cdot||\delta A||\cdot||\mathbf{x}||$ which is also

$$\frac{||\delta\mathbf{x}||}{||\mathbf{x}||}\Big/\frac{||\delta A||}{||A||}\leq||A^{-1}||\cdot||A||=\kappa(A).$$

Equality can be shown to hold when δA is such that $||A^{-1}(\delta A)\mathbf{x}||=||A^{-1}||\cdot||\delta A||\cdot||\mathbf{x}||$, and that such a δA exists.

We conclude with two theorems from [TreBau] that summarize and generalize these findings.

Theorem. *Let $A\in\mathbb{R}^{m\times m}$ be nonsingular and consider equation $A\mathbf{x}=\mathbf{b}$. The problem of computing $\mathbf{b}$ given $\mathbf{x}$ has condition number*

$$\kappa=||A||\frac{||\mathbf{x}||}{||\mathbf{b}||}\leq||A||\cdot||A^{-1}|| \tag{A.1}$$

with respect to perturbation of $\mathbf{x}$. The problem of computing $\mathbf{x}$ given $\mathbf{b}$, has condition number

$$\kappa=||A^{-1}||\frac{||\mathbf{b}||}{||\mathbf{x}||}\leq||A||\cdot||A^{-1}|| \tag{A.2}$$

with respect to perturbations of $\mathbf{b}$. If $||\cdot||=||\cdot||_2$, then equality holds in Equation (A.1) *if $\mathbf{x}$ is a multiple of a right singular vector of A corresponding to the minimal singular value σ_{min}. Equality holds in Equation* (A.2) *if $\mathbf{b}$ is a multiple of a left singular vector of A corresponding to the maximal singular value σ_1.*

Theorem. *Let $\mathbf{b}$ be fixed and consider the problem of computing $\mathbf{x}=A^{-1}\mathbf{b}$, where A is square and nonsingular. The condition number of this problem with respect to perturbations in A is $\kappa=||A|\cdot||A^{-1}||=\kappa(A)$.*

References

[Björck] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM 1996

[Demmel] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM 1997

[Eldén] Lars Eldén: *Numerical linear algebra in data mining*, Acta Numerica 2006, pp. 327-384

[GolubVanLoan] G. H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd Ed., The Johns Hopkins University Press, 1996

[OlvShaDraft] P. J. Olver and C. Shakiban, *Applied Mathematics*, draft paper

[QuaSacSal] Quarteroni A., Sacco A., Saleri F., *Numerical Mathematics*, Springer 2000

[TreBau] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM 1997

[Wikipedia] http://en.wikipedia.org/wiki/Cholesky_decomposition

[Internet-blog] http://mathprelims.wordpress.com/2008/07/16/