CHAPTER 1

Linear algebra

Contents

1.1 Introduction

Linear algebra is a field of mathematics that is widely used in various disciplines. Linear algebra plays an important role in data science and machine learning. A solid understanding of linear algebra concepts can enhance the understanding of many data science and machine learning algorithms. This chapter introduces basic concepts of linear algebra that need data science including vector spaces, orthogonality, eigenvalues, matrix decomposition, and is further expanded to include linear regression and principal component analysis where linear algebra plays a central role for solving data science problems. More advanced concepts and applications of linear algebra can be found in many references [1–4,118].

Mathematical Methods in Data Science
https://doi.org/10.1016/B978-0-44-318679-0.00007-7

1.2 Elements of linear algebra

We begin with a brief review of basic concepts for linear algebra. We confine our discussions to $V = \mathbb{R}^n$.

1.2.1 Linear spaces

1.2.1.1 Linear combinations

A linear combination in linear algebra is a new vector constructed from a subset by multiplying each vector by a constant and adding the results. A linear subspace is a subset of all linear combination.

Definition 1.2.1 (Linear subspace). A linear subspace of V is a subset $U \subseteq V$ that is closed under vector addition and scalar multiplication. That is, for all $\mathbf{u}_1, \mathbf{u}_2 \in U$ and $\alpha \in \mathbb{R}$, it holds that

$$\mathbf{u}_1 + \mathbf{u}_2 \in U, \quad \text{and} \quad \alpha\, \mathbf{u}_1 \in U. \tag{1.2.1}$$

In particular, $\mathbf{0}$ is always in a linear subspace. As we can see below, a span of a set of vectors is a linear subspace.

Definition 1.2.2 (Span). Let $\mathbf{w}_1, \ldots, \mathbf{w}_m \in V$. The span of $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$, denoted $\operatorname{span}(\mathbf{w}_1, \ldots, \mathbf{w}_m)$, is the set of all linear combinations of the $\mathbf{w}_j$'s. That is,

$$\operatorname{span}(\mathbf{w}_1, \ldots, \mathbf{w}_m) = \left\{ \sum_{j=1}^{m} \alpha_j \mathbf{w}_j \,:\, \alpha_1, \ldots, \alpha_m \in \mathbb{R} \right\}. \tag{1.2.2}$$

A list of vectors that span a linear subspace U is also referred to as a spanning set of U. We can verify that a span is a linear subspace.

Lemma 1.2.3 (Every span is a linear subspace). *Let $W = \operatorname{span}(\mathbf{w}_1, \ldots, \mathbf{w}_m)$. Then W is a linear subspace.*

Proof. Let $\mathbf{u}_1, \mathbf{u}_2 \in W$, and $\alpha \in \mathbb{R}$. Then for $i = 1, 2$,

$$\mathbf{u}_i = \sum_{j=1}^{m} \beta_{i,j} \mathbf{w}_j,$$

and

$$\alpha \mathbf{u}_1 + \mathbf{u}_2 = \alpha \sum_{j=1}^{m} \beta_{1,j} \mathbf{w}_j + \sum_{j=1}^{m} \beta_{2,j} \mathbf{w}_j = \sum_{j=1}^{m} (\alpha \beta_{1,j} + \beta_{2,j}) \mathbf{w}_j.$$

We conclude that $\alpha\, \mathbf{u}_1 + \mathbf{u}_2 \in W$. □

Often it is useful to study the column space of a matrix.

Definition 1.2.4 (Column space). Let $A \in \mathbb{R}^{n\times m}$ be an $n \times m$ matrix with columns $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^n$. The column space of A, denoted $\mathrm{col}(A)$, is the span of the columns of A, that is, $\mathrm{col}(A) = \mathrm{span}(\mathbf{a}_1, \dots, \mathbf{a}_m) \in \mathbb{R}^n$.

1.2.1.2 Linear independence and dimension

For many problems in applications including data science, it is desirable to avoid redundancy in the description of a linear subspace. The concept is central to the definition of dimension of a linear space.

Definition 1.2.5 (Linear independence). A list of vectors $\mathbf{u}_1, \dots, \mathbf{u}_m$ is linearly independent if none of them can be written as a linear combination of the others, that is,

$$\forall i, \ \ \mathbf{u}_i \notin \mathrm{span}(\{\mathbf{u}_j : j \neq i\}).$$

A list of vectors is called linearly dependent if it is not linearly independent.

Lemma 1.2.6. *The vectors* $\mathbf{u}_1$, $\dots$, $\mathbf{u}_m$ *are linearly independent if and only if*

$$\sum_{j=1}^{m} \alpha_j \mathbf{u}_j = \mathbf{0} \implies \alpha_j = 0, \ \forall j.$$

Equivalently, $\mathbf{u}_1, \dots, \mathbf{u}_m$ *are linearly dependent if and only if there exist* α_j*'s, not all zero, such that* $\sum_{j=1}^{m} \alpha_j \mathbf{u}_j = \mathbf{0}$.

Proof. The equivalence follows by contradiction. We prove the second statement. Assume $\mathbf{u}_1, \dots, \mathbf{u}_m$ are linearly dependent. Then $\mathbf{u}_i = \sum_{j\neq i} \alpha_j \mathbf{u}_j$ for some i. Taking $\alpha_i = -1$ gives $\sum_{j=1}^{m} \alpha_j \mathbf{u}_j = \mathbf{0}$. On the other hand, assume $\sum_{j=1}^{m} \alpha_j \mathbf{u}_j = \mathbf{0}$ with α_j's not all zero. In particular, $\alpha_i \neq 0$ for some i. Then $\mathbf{u}_i = \frac{1}{\alpha_i} \sum_{j\neq i} \alpha_j \mathbf{u}_j$. □

For matrix form, let $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^n$ and

$$A = \begin{pmatrix} | & & | \\ \mathbf{a}_1 & \dots & \mathbf{a}_m \\ | & & | \end{pmatrix}.$$

Then linearly independence can be formulated as if there is a non-trivial solution of a linear system. It is clear that $A\mathbf{x}$ is the following linear combination of the columns of A: $\sum_{j=1}^{m} x_j \mathbf{a}_j$. Then $\mathbf{a}_1, \dots, \mathbf{a}_m$ are linearly independent if and only if $A\mathbf{x} = \mathbf{0} \implies \mathbf{x} = \mathbf{0}$. Equivalently, $\mathbf{a}_1, \ldots, \mathbf{a}_m$ are linearly dependent if and only if $\exists \mathbf{x} \neq \mathbf{0}$ such that $A\mathbf{x} = \mathbf{0}$.

We now look at the concept of bases, which give a minimal representation of a subspace. A basis is a set of vectors that generates all elements of the vector space and the vectors in the set are linearly independent.

Definition 1.2.7 (Basis of a space). Let U be a linear subspace of V. A basis of U is a list of vectors $\mathbf{u}_1, \ldots, \mathbf{u}_m$ in U that: (1) span U, that is, $U = \mathrm{span}(\mathbf{u}_1, \ldots, \mathbf{u}_m)$; and (2) are linearly independent.

We denote by $\mathbf{e}_1, \ldots, \mathbf{e}_n$ the standard basis of $\mathbb{R}^n$, where $\mathbf{e}_i$ has a one in coordinate i and zeros in all other coordinates. One of the first key properties of a basis is that it provides a unique representation of the vectors in the subspace. Indeed, let U be a linear subspace and $\mathbf{u}_1, \ldots, \mathbf{u}_m$ be a basis of U. Suppose that $\mathbf{w} \in U$ can be written as $\mathbf{w} = \sum_{j=1}^m \alpha_j \mathbf{u}_j$ and $\mathbf{w} = \sum_{j=1}^m \alpha_j' \mathbf{u}_j$. Then subtracting one equation from the other we arrive at $\sum_{j=1}^m (\alpha_j - \alpha_j')\, \mathbf{u}_j = \mathbf{0}$. By linear independence, we have $\alpha_j - \alpha_j' = 0$ for each j.

A vector space can have several bases; however, all of the bases have the same number of elements, called the dimension of the vector space. When applied to a matrix A, the dimension of the column space of A is called the (column) rank of A. We state it as the following theorem.

Theorem 1.2.8 (Dimension theorem). *Let U be a linear subspace of V. Any basis of U always has the same number of elements. All bases of U have the same length, that is, the same number of elements. We call this number the dimension of U and denote it* $\dim(U)$.

The following lemma further describes the property of a linearly dependent set, which is used to prove the dimension theorem. It states that, given a linearly dependent list of vectors, one of the vectors is in the span of the previous ones and we can remove it without changing the span.

Lemma 1.2.9 (Characterization of linearly dependent sets). *Let $\mathbf{u}_1, \ldots, \mathbf{u}_m$ be a linearly dependent list of vectors with a linearly independent subset, $\mathbf{u}_i, i \in \{1, \ldots, k\}$, $k < m$. Then there is an $i > k$ such that:*

1. $\mathbf{u}_i \in \mathrm{span}(\mathbf{u}_1, \ldots, \mathbf{u}_{i-1})$;
2. $\mathrm{span}(\{\mathbf{u}_j : j \in \{1, \ldots, m\}\}) = \mathrm{span}(\{\mathbf{u}_j : j \in \{1, \ldots, m\},\ j \neq i\})$.

Proof (Characterization of linearly dependent sets). For 1, by linear dependence, $\sum_{j=1}^m \alpha_j \mathbf{u}_j = \mathbf{0}$ with not all α_j's zero. Further, because $\mathbf{u}_i, i \in \{1, \ldots, k\}$,

$k < m$ is independent, and not all $\alpha_{k+1}, \dots, \alpha_m$ are zero. Take the largest index among the α_j's that are non-zero, say i. Then rearranging gives

$$\mathbf{u}_i = -\sum_{j=1}^{i-1} \frac{\alpha_j}{\alpha_i} \mathbf{u}_j.$$

For 2, we note that for any $\mathbf{w} \in \operatorname{span}(\{\mathbf{u}_j : j \in \{1, \dots, m\}\})$ we can write it as $\mathbf{w} = \sum_{j=1}^{m} \beta_j \mathbf{u}_j$ and we can replace $\mathbf{u}_i$ by the equation above, producing a representation of $\mathbf{w}$ in terms of $\{\mathbf{u}_j : j \in \{1, \dots, m\}, j \neq i\}$. □

We are now able to prove the dimension theorem.

Proof (Dimension theorem). Suppose we have two bases $\{\mathbf{w}_i : i \in \{1, \dots, n\}\}$ and $\{\mathbf{u}_j : j \in \{1, \dots, m\}\}$ of U. It suffices to show that $n \geq m$. First, we consider the list $\{\mathbf{u}_1, \mathbf{w}_1, \dots, \mathbf{w}_n\}$. Because the $\mathbf{w}_i$'s are spanning, adding $\mathbf{u_1} \neq \mathbf{0}$ to them necessarily produces a linearly dependent list. By the lemma for the characterization of linearly dependent sets, we can remove one of the $\mathbf{w}_i$'s without changing the span. The new list B has length n again. Then we add $\mathbf{u}_2$ to B immediately after $\mathbf{u}_1$. By the lemma for the characterization of linearly dependent sets, one of the vectors in this list is in the span of the previous ones. It cannot be $\mathbf{u}_2$ as $\{\mathbf{u}_1, \mathbf{u}_2\}$ are linearly independent by assumption. So it must be one of the remaining $\mathbf{w}_i$'s. We remove that one, without changing the span by the linear dependence lemma again. This process can be continued until we have added all the $\mathbf{u}_j$'s, as otherwise a subset of $\{\mathbf{u}_j : j \in \{1, \dots, m\}\}$ would span U, which is a contradiction. Hence $n \geq m$. □

1.2.2 Orthogonality

In many applications, the use of orthonormal bases can greatly simplify mathematical representations and reveal more insights of the underlying problems. We begin with the following definitions and lemmas.

1.2.2.1 Orthonormal bases

Definition 1.2.10 (Norm and inner product). $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u} \cdot \mathbf{v} = \sum_i^n u_i v_i$ and $\|\mathbf{u}\| = \sqrt{\sum_1^n u_i^2}$.

Definition 1.2.11. A list of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ is orthonormal if the $\mathbf{u}_i$'s are pairwise orthogonal and each has norm 1, that is, for all i and all $j \neq i$, $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$, and $\|\mathbf{u}_i\| = 1$.

We generalize the Pythagorean theorem with orthogonal vectors.

Lemma 1.2.12 (Pythagorean theorem). *Let* $\mathbf{u}, \mathbf{v} \in V$ *be orthogonal. Then* $\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$.

Proof (Pythagorean theorem). Using $\|\mathbf{w}\|^2 = \langle \mathbf{w}, \mathbf{w} \rangle$, we get

$$\|\mathbf{u}+\mathbf{v}\|^2 = \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + 2\langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2. \quad \square$$

Many useful results can be derived from Pythagorean theorem, for example, we have the following.

Lemma 1.2.13 (Cauchy–Schwarz). *For any* $\mathbf{u}, \mathbf{v} \in V$, $|\langle \mathbf{u}, \mathbf{v} \rangle| \le \|\mathbf{u}\| \|\mathbf{v}\|$.

Proof (Cauchy–Schwarz). Let $\mathbf{q} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$ be the unit vector in the direction of $\mathbf{v}$. We want to show $|\langle \mathbf{u}, \mathbf{q} \rangle| \le \|\mathbf{u}\|$. Decompose $\mathbf{u}$ into its projection onto $\mathbf{q}$ and what is left is the following:

$$\mathbf{u} = \langle \mathbf{u}, \mathbf{q} \rangle \mathbf{q} + \{\mathbf{u} - \langle \mathbf{u}, \mathbf{q} \rangle \mathbf{q}\}.$$

The two terms on the right-hand side are orthogonal, so the Pythagorean theorem gives

$$\|\mathbf{u}\|^2 = \|\langle \mathbf{u}, \mathbf{q} \rangle \mathbf{q}\|^2 + \|\mathbf{u} - \langle \mathbf{u}, \mathbf{q} \rangle \mathbf{q}\|^2 \ge \|\langle \mathbf{u}, \mathbf{q} \rangle \mathbf{q}\|^2 = \langle \mathbf{u}, \mathbf{q} \rangle^2.$$

Taking a square root gives the claim. □

Now we have the following properties for orthonormal lists.

Lemma 1.2.14. *Let* $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ *be an orthonormal list of vectors.*
1. $\|\sum_{j=1}^m \alpha_j \mathbf{u}_j\|^2 = \sum_{j=1}^m \alpha_j^2$ *for any* $\alpha_j \in \mathbb{R}$, $j \in \{1, \ldots, m\}$;
2. $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ *are linearly independent.*

Proof. For 1, noting that $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$ and $\langle \beta \mathbf{x}_1 + \mathbf{x}_2, \mathbf{x}_3 \rangle = \beta \langle \mathbf{x}_1, \mathbf{x}_3 \rangle + \langle \mathbf{x}_2, \mathbf{x}_3 \rangle$, we have

$$\begin{aligned}\left\|\sum_{j=1}^m \alpha_j \mathbf{u}_j\right\|^2 &= \left\langle \sum_{i=1}^m \alpha_i \mathbf{u}_i, \sum_{j=1}^m \alpha_j \mathbf{u}_j \right\rangle = \sum_{i=1}^m \alpha_i \left\langle \mathbf{u}_i, \sum_{j=1}^m \alpha_j \mathbf{u}_j \right\rangle \\ &= \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j \langle \mathbf{u}_i, \mathbf{u}_j \rangle,\end{aligned}$$

which $\sum_{i=1}^m \alpha_i^2$, where we used orthonormality in the rightmost equation, that is, $\langle \mathbf{u}_i, \mathbf{u}_j \rangle$ is 1 if $i = j$ and 0 otherwise.

For 2, suppose $\sum_{i=1}^{m} \beta_i \mathbf{u}_i = \mathbf{0}$. Then we must have by 1 that $\sum_{i=1}^{m} \beta_i^2 = 0$. That implies $\beta_i = 0$ for all i. Hence the $\mathbf{u}_i$'s are linearly independent. □

Given a basis $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ of a subspace $\mathscr{U}$, we know that: for any $\mathbf{w} \in \mathscr{U}$, $\mathbf{w} = \sum_{i=1}^{m} \alpha_i \mathbf{u}_i$ for some α_i's. It is not immediately obvious in general how to find these α_i's. In the orthonormal case, it is straightforward.

Theorem 1.2.15 (Orthonormal basis expansion). *Let $\mathbf{q}_1, \dots, \mathbf{q}_m$ be an orthonormal basis of $\mathscr{U}$ and let $\mathbf{u} \in \mathscr{U}$. Then*

$$\mathbf{u} = \sum_{j=1}^{m} \langle \mathbf{u}, \mathbf{q}_j \rangle \, \mathbf{q}_j.$$

Proof. Because $\mathbf{u} \in \mathscr{U}$, $\mathbf{u} = \sum_{i=1}^{m} \alpha_i \mathbf{q}_i$ for some α_i. Take the inner product with $\mathbf{q}_j$ and use orthonormality:

$$\langle \mathbf{u}, \mathbf{q}_j \rangle = \left\langle \sum_{i=1}^{m} \alpha_i \mathbf{q}_i, \mathbf{q}_j \right\rangle = \sum_{i=1}^{m} \alpha_i \langle \mathbf{q}_i, \mathbf{q}_j \rangle = \alpha_j. \qquad \square$$

1.2.2.2 Best approximation theorem

Many optimization applications can be converted to the following best approximation problem. We have a linear subspace $\mathscr{U} \subseteq V$ and a vector $\mathbf{v} \notin \mathscr{U}$. We want to find the vector $\mathbf{v}^*$ in $\mathscr{U}$ that is closest to $\mathbf{v}$ in the norm as in Fig. 1.1, that is, we want to solve

$$\min_{\mathbf{v}^* \in \mathscr{U}} \|\mathbf{v}^* - \mathbf{v}\|.$$

Example 1.2.16. Consider the two-dimensional case with a one-dimensional subspace, say $\mathscr{U} = \mathrm{span}(\mathbf{u}_1)$ with $\|\mathbf{u}_1\| = 1$. The geometrical intuition of the best approximation theorem is demonstrated in Fig. 1.1. The solution $\mathbf{v}^*$ has the property that the difference $\mathbf{v} - \mathbf{v}^*$ makes a right angle with $\mathbf{u}_1$, that is, it is orthogonal to it.

Letting $\mathbf{v}^* = \alpha^* \mathbf{u}_1$, the geometrical condition above translates into

$$0 = \langle \mathbf{u}_1, \mathbf{v} - \mathbf{v}^* \rangle = \langle \mathbf{u}_1, \mathbf{v} - \alpha^* \mathbf{u}_1 \rangle = \langle \mathbf{u}_1, \mathbf{v} \rangle - \alpha^* \langle \mathbf{u}_1, \mathbf{u}_1 \rangle = \langle \mathbf{u}_1, \mathbf{v} \rangle - \alpha^*.$$

which implies that

$$\mathbf{v}^* = \langle \mathbf{u}_1, \mathbf{v} \rangle \, \mathbf{u}_1.$$

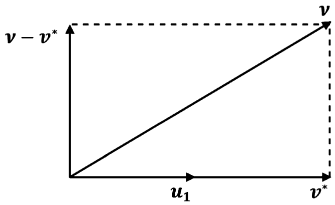

Figure 1.1

By the Pythagorean theorem, we then have for any $\alpha \in \mathbb{R}$,

$$\begin{aligned}\|\mathbf{v} - \alpha\,\mathbf{u}_1\|^2 &= \|\mathbf{v} - \mathbf{v}^* + \mathbf{v}^* - \alpha\,\mathbf{u}_1\|^2 = \|\mathbf{v} - \mathbf{v}^* + (\alpha^* - \alpha)\,\mathbf{u}_1\|^2 \\ &= \|\mathbf{v} - \mathbf{v}^*\|^2 + \|(\alpha^* - \alpha)\,\mathbf{u}_1\|^2\end{aligned}$$

and, therefore,

$$\|\mathbf{v} - \alpha\,\mathbf{u}_1\|^2 \geq \|\mathbf{v} - \mathbf{v}^*\|^2.$$

This confirms the optimality of $\mathbf{v}^*$.

The argument in the example above carries through in higher dimension, leading to the following fundamental result.

Definition 1.2.17 (Orthogonal projection). Let $\mathscr{U} \subseteq V$ be a linear subspace with orthonormal basis $\mathbf{q}_1, \ldots, \mathbf{q}_m$. The orthogonal projection of $\mathbf{v} \in V$ on $\mathscr{U}$ is defined as

$$\mathscr{P}_{\mathscr{U}}\mathbf{v} = \sum_{j=1}^{m} \langle \mathbf{v}, \mathbf{q}_j \rangle\, \mathbf{q}_j.$$

Theorem 1.2.18 (Best approximation theorem). *Let $\mathscr{U} \subseteq V$ be a linear subspace with orthonormal basis $\mathbf{q}_1, \ldots, \mathbf{q}_m$ and let $\mathbf{v} \in V$. For any $\mathbf{u} \in \mathscr{U}$,*

$$\|\mathbf{v} - \mathscr{P}_{\mathscr{U}}\mathbf{v}\| \leq \|\mathbf{v} - \mathbf{u}\|.$$

Furthermore, if $\mathbf{u} \in \mathscr{U}$ and the inequality above is an equality, then $\mathbf{u} = \mathscr{P}_{\mathscr{U}}\mathbf{v}$.

The visualization of the theorem is shown in Fig. 1.2. In addition, we note the following.

Lemma 1.2.19 (Orthogonal decomposition). *Let $\mathscr{U} \subseteq V$ be a linear subspace with orthonormal basis $\mathbf{q}_1, \ldots, \mathbf{q}_m$ and let $\mathbf{v} \in V$. For any $\mathbf{u} \in \mathscr{U}$, $\langle \mathbf{v} - \mathscr{P}_{\mathscr{U}}\mathbf{v}, \mathbf{u} \rangle = 0$. In particular, $\mathbf{v}$ can be decomposed as $(\mathbf{v} - \mathscr{P}_{\mathscr{U}}\mathbf{v}) + \mathscr{P}_{\mathscr{U}}\mathbf{v}$ where the two terms are orthogonal.*

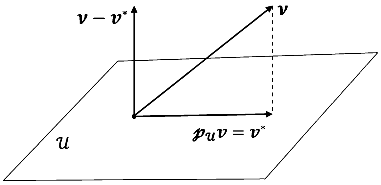

Figure 1.2 Best approximation theorem.

Proof (Orthogonal decomposition). We can write any $\mathbf{u} \in \mathcal{U}$ as $\sum_{j=1}^m \alpha_j \mathbf{q}_j$ for some α_j's. Then

$$\begin{aligned}\langle \mathbf{v} - \mathcal{P}_\mathcal{U}\mathbf{v}, \mathbf{u}\rangle &= \left\langle \mathbf{v} - \sum_{j=1}^m \langle \mathbf{v}, \mathbf{q}_j\rangle\, \mathbf{q}_j, \sum_{j=1}^m \alpha_j' \mathbf{q}_j \right\rangle \\ &= \sum_{j=1}^m \langle \mathbf{v}, \mathbf{q}_j\rangle\, \alpha_j' - \sum_{j=1}^m \alpha_j' \langle \mathbf{v}, \mathbf{q}_j\rangle = 0\end{aligned}$$

where we used the orthonormality of the $\mathbf{q}_j$'s in the rightmost equality. The second claim follows from $\mathcal{P}_\mathcal{U}\mathbf{v} \in \mathcal{U}$. □

We return to the proof of our main theorem.

Proof (Best approximation theorem). For any $\mathbf{u} \in \mathcal{U}$, the vector $\mathbf{u}' = \mathcal{P}_\mathcal{U}\mathbf{v} - \mathbf{u}$ is also in $\mathcal{U}$. By the orthogonal decomposition lemma and Pythagoras,

$$\begin{aligned}\|\mathbf{v} - \mathbf{u}\|^2 &= \|\mathbf{v} - \mathcal{P}_\mathcal{U}\mathbf{v} + \mathcal{P}_\mathcal{U}\mathbf{v} - \mathbf{u}\|^2 \\ &= \|\mathbf{v} - \mathcal{P}_\mathcal{U}\mathbf{v}\|^2 + \|\mathcal{P}_\mathcal{U}\mathbf{v} - \mathbf{u}\|^2 \geq \|\mathbf{v} - \mathcal{P}_\mathcal{U}\mathbf{v}\|^2.\end{aligned}$$

Furthermore, equality holds only if $\|\mathcal{P}_\mathcal{U}\mathbf{v} - \mathbf{u}\|^2 = 0$, which holds only if $\mathbf{u} = \mathcal{P}_\mathcal{U}\mathbf{v}$ by the point-separating property of the norm. □

The map $\mathcal{P}_\mathcal{U}$ is linear, that is, $\mathcal{P}_\mathcal{U}(\alpha\,\mathbf{x} + \mathbf{y}) = \alpha\,\mathcal{P}_\mathcal{U}\mathbf{x} + \mathcal{P}_\mathcal{U}\mathbf{y}$ for all $\alpha \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Indeed,

$$\begin{aligned}\mathcal{P}_\mathcal{U}(\alpha\,\mathbf{x} + \mathbf{y}) &= \sum_{j=1}^m \langle \alpha\,\mathbf{x} + \mathbf{y}, \mathbf{q}_j\rangle\, \mathbf{q}_j = \sum_{j=1}^m \left\{\alpha\,\langle \mathbf{x}, \mathbf{q}_j\rangle + \langle \mathbf{y}, \mathbf{q}_j\rangle\right\} \mathbf{q}_j \\ &= \alpha\,\mathcal{P}_\mathcal{U}\mathbf{x} + \mathcal{P}_\mathcal{U}\mathbf{y}.\end{aligned}$$

As a result, it can be encoded as an $n \times m$ matrix Q. Let

$$Q = \begin{pmatrix} | & & | \\ \mathbf{q}_1 & \cdots & \mathbf{q}_m \\ | & & | \end{pmatrix}$$

and note that computing

$$Q^T \mathbf{v} = \begin{pmatrix} \langle \mathbf{v}, \mathbf{q}_1 \rangle \\ \cdots \\ \langle \mathbf{v}, \mathbf{q}_m \rangle \end{pmatrix}$$

lists the coefficients in the expansion of $\mathscr{P}_{\mathscr{U}} \mathbf{v}$ over the basis $\mathbf{q}_1, \ldots, \mathbf{q}_m$. Hence we see that

$$\mathscr{P} = QQ^T.$$

On the other hand,

$$Q^T Q = \begin{pmatrix} \langle \mathbf{q}_1, \mathbf{q}_1 \rangle & \cdots & \langle \mathbf{q}_1, \mathbf{q}_m \rangle \\ \langle \mathbf{q}_2, \mathbf{q}_1 \rangle & \cdots & \langle \mathbf{q}_2, \mathbf{q}_m \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{q}_m, \mathbf{q}_1 \rangle & \cdots & \langle \mathbf{q}_m, \mathbf{q}_m \rangle \end{pmatrix} = I_{m \times m}$$

where $I_{m \times m}$ denotes the $m \times m$ identity matrix.

1.2.3 Gram–Schmidt process

The Gram–Schmidt algorithm is used to obtain an orthonormal basis. Let $\mathbf{a}_1, \ldots, \mathbf{a}_m$ be linearly independent. We intend to find an orthonormal basis of $\mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_m)$. The process takes advantage of the properties of the orthogonal projection derived above. In essence, we add the vectors $\mathbf{a}_i$ one by one, but only after taking out their orthogonal projection on the previously included vectors. The outcome spans the same subspace and orthogonal decomposition ensures orthogonality.

Theorem 1.2.20 (Gram–Schmidt). *Let* $\mathbf{a}_1, \ldots, \mathbf{a}_m$ *in* R^n *be linearly independent. Then there exist an orthonormal basis* $\mathbf{q}_1, \ldots, \mathbf{q}_m$ *of* $\mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_m)$.

Proof. The inductive step is the following. Assume that we have constructed orthonormal vectors $\mathbf{q}_1, \ldots, \mathbf{q}_{i-1}$ such that

$$U_{i-1} := \mathrm{span}(\mathbf{q}_1, \ldots, \mathbf{q}_{i-1}) = \mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}).$$

Since we have an orthonormal basis for U_{i-1}, we can compute the orthogonal projection of $\mathbf{a}_i$,

$$\mathscr{P}_{U_{i-1}}\mathbf{a}_i = \sum_{j=1}^{i-1} \langle \mathbf{a}_i, \mathbf{q}_j \rangle\, \mathbf{q}_j.$$

And we set

$$\mathbf{b}_i = \mathbf{a}_i - \mathscr{P}_{U_{i-1}}\mathbf{a}_i \quad \text{and} \quad \mathbf{q}_i = \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|}.$$

Here, we used that $\|\mathbf{b}_i\| > 0$; otherwise, $\mathbf{a}_i$ would be equal to its projection $\mathscr{P}_{U_{i-1}}\mathbf{a}_i \in \mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_{i-1})$, which would contradict linear independence of the $\mathbf{a}_j$'s. By the orthogonal decomposition result, $\mathbf{q}_i$ is orthogonal to $\mathrm{span}(\mathbf{q}_1, \ldots, \mathbf{q}_{i-1})$ and, unrolling the calculations above, $\mathbf{a}_i$ is the following linear combination of $\mathbf{q}_1, \ldots, \mathbf{q}_i$:

$$\mathbf{a}_i = \sum_{j=1}^{i-1} \langle \mathbf{a}_i, \mathbf{q}_j \rangle\, \mathbf{q}_j + \left\| \mathbf{a}_i - \sum_{j=1}^{i-1} \langle \mathbf{a}_i, \mathbf{q}_j \rangle\, \mathbf{q}_j \right\| \mathbf{q}_i.$$

Hence $\mathbf{q}_1, \ldots, \mathbf{q}_i$ forms an orthonormal list with $\mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_i) \subseteq \mathrm{span}(\mathbf{q}_1, \ldots, \mathbf{q}_i)$. The opposite inclusion holds by construction. Moreover, because $\mathbf{q}_1, \ldots, \mathbf{q}_i$ are orthonormal, they are linearly independent so we must form a basis of their span so that induction goes through. □

1.2.4 Eigenvalues and eigenvectors

Eigenvalues and eigenvectors are key concepts in many applications. As before, we work on $\mathbb{R}^d$.

Definition 1.2.21 (Eigenvalues and eigenvectors). Let $A \in \mathbb{R}^{d \times d}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an eigenvalue of A if there exists a non-zero vector $\mathbf{x} \neq \mathbf{0}$ such that

$$A\mathbf{x} = \lambda\mathbf{x}. \tag{1.2.3}$$

The vector $\mathbf{x}$ is referred to as an eigenvector.

As the next example shows, not every matrix has an eigenvalue.

Example 1.2.22 (No real eigenvalues). Set $d = 2$ and let

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

For λ to be an eigenvalue, there must be an non-zero eigenvector $\mathbf{x} = (x_1, x_2)^T$ such that

$$A\mathbf{x} = \lambda\mathbf{x}$$

or put differently

$$-x_2 = \lambda x_1 \quad \text{and} \quad x_1 = \lambda x_2.$$

Replacing these equations into each other, it must be that

$$-x_2 = \lambda^2 x_2 \quad \text{and} \quad x_1 = -\lambda^2 x_1.$$

Because x_1, x_2 cannot both be 0, λ must satisfy the equation

$$\lambda^2 = -1$$

for which there is no real solution.

As we can see from below, $A \in \mathbb{R}^{d\times d}$ has at most d distinct eigenvalues.

Lemma 1.2.23 (Number of eigenvalues). *Let $A \in \mathbb{R}^{d\times d}$ and let $\lambda_1, \ldots, \lambda_m$ be distinct eigenvalues of A with corresponding non-zero eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Then $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are linearly independent. As a result, $m \leq d$.*

Proof. Assume by contradiction that $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are linearly dependent. By linear dependence, there is $k \leq m$ such that

$$\mathbf{x}_k \in \mathrm{span}(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1})$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}$ are linearly independent. In particular, there are $a_1, \ldots, a_{k-1}$ such that

$$\mathbf{x}_k = a_1\mathbf{x}_1 + \cdots + a_{k-1}\mathbf{x}_{k-1}.$$

Transform the equation above in two ways: (1) multiply both sides by λ_k and (2) apply A. Then subtract the resulting equations. That leads to

$$\mathbf{0} = a_1(\lambda_k - \lambda_1)\mathbf{x}_1 + a_{k-1}(\lambda_k - \lambda_{k-1})\mathbf{x}_{k-1}.$$

Because the λ_i's are distinct and $\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}$ are linearly independent, we must have $a_1 = \cdots = a_{k-1} = 0$. But that implies that $\mathbf{x}_k = \mathbf{0}$, a contradiction. For the second claim, if there were more than d distinct eigenvalues, then there would be more than d corresponding linearly independent eigenvectors by the first claim, a contradiction. □

1.2.4.1 Diagonalization of symmetric matrices

We will use the notation $\operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ for the diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_d$.

Example 1.2.24 (Diagonal (and similar) matrices). Let A be similar to a matrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ with distinct diagonal entries, that is, there exists a non-singular matrix P such that

$$A = PDP^{-1}.$$

Let $\mathbf{p}_1, \ldots, \mathbf{p}_d$ be the columns of P. Then

$$AP = PD,$$

which implies that

$$A\mathbf{p}_i = \lambda_i \mathbf{p}_i.$$

Theorem 1.2.25. *If A is symmetric, then any two eigenvectors from different eigenspaces are orthogonal.*

Proof. Let $\mathbf{u}_1$ and $\mathbf{u}_2$ be eigenvectors that correspond to distinct eigenvalues, say, λ_1 and λ_2. To show that $\mathbf{u}_1 \cdot \mathbf{u}_2 = 0$, we compute

$$\begin{aligned}
\lambda_1 \mathbf{u}_1 \cdot \mathbf{u}_2 &= (\lambda_1 \mathbf{u}_1)^T \mathbf{u}_2 = (A\mathbf{u}_1)^T \mathbf{u}_2 && \text{Since } \mathbf{u}_1 \text{ is an eigenvector} \\
&= \left(\mathbf{u}_1^T A^T\right) \mathbf{u}_2 = \mathbf{u}_1^T (A\mathbf{u}_2) && \text{Since } A^T = A \\
&= \mathbf{u}_1^T (\lambda_2 \mathbf{u}_2) && \text{Since } \mathbf{u}_2 \text{ is an eigenvector} \\
&= \lambda_2 \mathbf{u}_1^T \mathbf{u}_2 = \lambda_2 \mathbf{u}_1 \cdot \mathbf{u}_2
\end{aligned}$$

Hence $(\lambda_1 - \lambda_2)\, \mathbf{u}_1 \cdot \ \mathbf{u}_2 = 0$. But $\lambda_1 - \lambda_2 \neq 0$, so $\mathbf{u}_1 \cdot \ \mathbf{u}_2 = 0$. □

A matrix A is said to be orthogonally diagonalizable if there are an orthogonal matrix P (with $P^{-1} = P^T$) and a diagonal matrix D such that

$$A = PDP^T = PDP^{-1} \tag{1.2.4}$$

To orthogonally diagonalize an $n \times n$ matrix, we must be able to find n linearly independent and orthonormal eigenvectors. If A is orthogonally diagonalizable, then

$$A^T = \left(PDP^T\right)^T = P^{TT} D^T P^T = PDP^T = A$$

Thus A is symmetric. The following results reveal more properties of a symmetric matrix, which implies that every symmetric matrix is orthogonally diagonalizable.

Theorem 1.2.26 (The spectral theorem for symmetric matrices). *An $n \times n$ symmetric matrix A has the following properties:*

- *A has n real eigenvalues, counting multiplicities.*
- *If λ is an eigenvalue of A with multiplicity k, then the eigenspace for λ is k-dimensional.*
- *The eigenspaces are mutually orthogonal, in the sense that eigenvectors corresponding to different eigenvalues are orthogonal.*
- *A is orthogonally diagonalizable.*

Proof. We only give the idea of proof. If $\mathbf{u}_1$ is a unit eigenvector corresponding to λ_1, now start with $\mathbf{u}_1$ and find $[\mathbf{u}_1, ..., \mathbf{u}_n]$ to be an orthonormal basis by the Gram–Schmidt process and let $U = [\mathbf{u}_1, ..., \mathbf{u}_n]$. Then

$$U^T A U = \begin{pmatrix} \lambda_1 & * \\ 0 & A_1 \end{pmatrix}$$

Note that A is symmetric and we must have

$$U^T A U = \begin{pmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{pmatrix}$$

Now A_1 must be symmetric and have the remaining eigenvalues and continuing this process, we arrive at

$$U^T A U = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$$

This observation will clearly lead to the conclusions of the theorem. □

Suppose that $A = PDP^{-1}$, where the columns of P are orthonormal eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$ of A and the corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$ are in the diagonal matrix D. Since $P^{-1} = P^T$,

$$A = PDP^T = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{pmatrix}$$

$$= \begin{pmatrix} \lambda_1 \mathbf{v}_1 & \cdots & \lambda_n \mathbf{v}_n \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{pmatrix}$$

Using the column-row expansion of a product, we can write

$$A = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T + \cdots + \lambda_n \mathbf{v}_n \mathbf{v}_n^T. \tag{1.2.5}$$

This representation of A is called a spectral decomposition of A because it breaks up A into pieces determined by the spectrum (eigenvalues) of A. Each $\mathbf{v}_i \mathbf{v}_i^T$ is an $n \times n$ matrix of rank 1. For example, every column of $\lambda_1 \mathbf{v}_1 \mathbf{v}_1^T$ is a multiple of $\mathbf{v}_1$. Furthermore, each matrix $\mathbf{v}_j \mathbf{v}_j^T$ is a **projection matrix** in the sense that for each $\mathbf{x}$ in $\mathbb{R}^n$, the vector $(\mathbf{v}_j \mathbf{v}_j^T)\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto the subspace spanned by $\mathbf{v}_j$.

1.2.4.2 Constrained optimization

The following result is useful for many optimization problems.

Theorem 1.2.27. *Let A be $n \times n$ symmetric matrix A with an orthogonal diagonalization $A = PDP^{-1}$. The columns of P are orthonormal eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$ of A. Assume that the diagonals of D are arranged so that $\lambda_1 \leq \lambda_2, ... \leq \lambda_n$. Then*

$$min_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \lambda_1$$

is achieved when $\mathbf{x} = \mathbf{v_1}$ and

$$max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \lambda_n$$

is achieved when $\mathbf{x} = \mathbf{v_n}$.

Proof. From the assumption, we have

$$A = P \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} P^T$$

and

$$P = \begin{bmatrix} \mathbf{v_1} & \cdots & \mathbf{v_n} \end{bmatrix},$$

Rearranging the terms gives

$$P^T A P = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

In addition, note that

$$A\mathbf{v_i} = \lambda_i \mathbf{v_i},$$

$$\mathbf{x} = P\mathbf{y},$$

and

$$\sum x_i^2 = \sum y_i^2.$$

It is easy to see that

$$\begin{aligned} \frac{\mathbf{x}^T A \mathbf{x}}{\sum x_i^2} = \frac{\mathbf{y}^T P^T A P \mathbf{y}}{\sum y_i^2} &= \frac{\lambda_1 y_1^2 + \cdots + \lambda_n y_n^2}{\sum y_i^2} \\ &\geq \lambda_1 \text{ (equality holds when } \mathbf{y} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}) \\ &\leq \lambda_n \text{ (equality holds when } \mathbf{y} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}) \end{aligned}$$

Note that

$$\mathbf{v_1} = P \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and

$$\mathbf{v_n} = P \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}. \qquad \square$$

1.3 Linear regression

Linear regression is used frequently in practical applications because of its simplicity. The models depend linearly on their unknown parameters and, therefore, are easier to fit than models which are non-linearly related to their parameters. As a result, the statistical properties of the resulting estimators are easier to determine. In this section, we first discuss QR decomposition, the least-squares problem, and return to linear regression.

1.3.1 QR decomposition

QR decomposition is a useful procedure to solve the linear least squares problem. First, we use the Gram–Schmidt algorithm to obtain an orthonormal basis span($\mathbf{a}_1, \ldots, \mathbf{a}_m$) from a linearly independent set of span($\mathbf{a}_1, \ldots, \mathbf{a}_m$). In order to derive QR decomposition, let

$$A = \begin{pmatrix} | & & | \\ \mathbf{a}_1 & \ldots & \mathbf{a}_m \\ | & & | \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} | & & | \\ \mathbf{q}_1 & \ldots & \mathbf{q}_m \\ | & & | \end{pmatrix},$$

where A, Q are $n \times m$ matrices. The output of the Gram–Schmidt algorithm above can then be written in the following compact form, known as a QR decomposition in Fig. 1.3,

$$A = QR,$$

where column i of the $m \times m$ matrix R contains the coefficients of the linear combination of $\mathbf{q}_j$'s that produces $\mathbf{a}_i$. Q is a $\mathbb{R}^{n\times m}$ matrix with $Q^T Q = I_{m\times m}$. It may be easier to verify $A = QR$ by

$$A^T = R^T Q^T.$$

By the proof of Gram–Schmidt, $\mathbf{a}_i \in \text{span}(\mathbf{q}_1, \ldots, \mathbf{q}_i)$. So column i of R has only zeros below the diagonal. Hence R has a special structure; it is upper triangular.

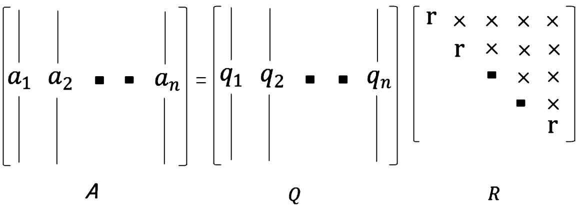

Figure 1.3 QR decomposition.

1.3.2 Least-squares problems

Let $A \in \mathbb{R}^{n\times m}$ be an $n \times m$ matrix and $\mathbf{b} \in \mathbb{R}^n$ be a vector. We try to solve the system $A\mathbf{x} = \mathbf{b}$, which is often inconsistent. We are looking to use the $A\mathbf{x}$ to approximate $\mathbf{b}$. It is reasonable to assume that matrix A has linearly independent columns. If $n = m$, that is, if A is a square matrix, we can use the matrix inverse to solve the system. But we are particularly interested in the over-determined case where $n > m$. We cannot use the matrix inverse then. One possibility to make sense of the problem in that case is to cast it as the least-squares problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \|A\mathbf{x} - \mathbf{b}\|.$$

In order to use the orthogonal decomposition result, we write

$$A = \begin{pmatrix} | & & | \\ \mathbf{a}_1 & \dots & \mathbf{a}_m \\ | & & | \end{pmatrix} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,m} \\ a_{2,1} & \cdots & a_{2,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,m} \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

Now we seek a linear combination of the columns of A that minimizes the objective

$$\left\| \sum_{j=1}^{m} x_j \mathbf{a}_j - \mathbf{b} \right\|^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} x_j a_{i,j} - b_i \right)^2 = \sum_{i=1}^{n} (\hat{y}_i - b_i)^2,$$

where

$$\hat{y}_i = \sum_{j=1}^{m} x_j a_{i,j}.$$

Now apply our characterization of the orthogonal projection on the column space of A. Let

$$\hat{\mathbf{b}} = \mathscr{P}_{\mathrm{col}(A)}\mathbf{b}.$$

Because $\hat{\mathbf{b}}$ is in the column space of A, the equation $A\mathbf{x} = \hat{\mathbf{b}}$ is consistent and there is an $\hat{x}$ such that

$$A\hat{\mathbf{x}} = \hat{\mathbf{b}}. \tag{1.3.1}$$

Since $\hat{\mathbf{b}}$ is the closed point in $\mathrm{col}(A)$ to $\mathbf{b}$ a vector $\hat{\mathbf{x}}$ is a least-square solution of $A\mathbf{x} = \mathbf{b}$ if and only (1.3.1) holds. The following theorem provides an alternative description of the solution.

Theorem 1.3.1 (Normal equations). *Let $A \in \mathbb{R}^{n\times m}$ be an $n \times m$ matrix with linearly independent columns and let $\mathbf{b} \in \mathbb{R}^n$ be a vector. The solution to the least-squares problem*

$$\min_{\mathbf{x}\in\mathbb{R}^m} \|A\mathbf{x} - \mathbf{b}\|,$$

satisfies

$$A^T A\mathbf{x} = A^T\mathbf{b},$$

which is known as the normal equations.

Proof. Let $U = \mathrm{col}(A) = \mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_m)$. By the best approximation theorem, the orthogonal projection $\hat{\mathbf{b}} = A\hat{\mathbf{x}}$ of $\mathbf{b}$ on U is the unique solution to the least-squares problem. By the orthogonal decomposition, it must satisfy $\langle \mathbf{b} - \hat{\mathbf{b}}, \mathbf{u}\rangle = 0$ for all $\mathbf{u} \in U$. Because the $\mathbf{a}_i$'s are a basis of U, it suffices that $\langle \mathbf{b} - A\hat{\mathbf{x}}, \mathbf{a}_i\rangle = 0$ for all $i \in \{1, \ldots, m\}$. In matrix form,

$$A^T(A\hat{\mathbf{x}} - \mathbf{b}) = \mathbf{0},$$

as claimed, after rearranging. □

When A has linearly independent columns, in view of the QR decomposition, it can be shown that $A^T A$ is invertible and the solution of the normal equation is

$$(A^T A)^{-1}A^T\mathbf{b}.$$

However, that approach has numerical issues, we solve the problem via QR decomposition:

- Construct an orthonormal basis of $\mathrm{col}(A)$ through a QR decomposition:

$$A = QR.$$

- Form the orthogonal projection matrix,

$$\mathscr{P}_{\mathrm{col}(A)} = QQ^T.$$

- Apply the projection to $\mathbf{b}$ and observe that $\mathbf{x}^*$ satisfies

$$A\mathbf{x}^* = QQ^T\mathbf{b}.$$

- Use the QR decomposition for A to get

$$QR\mathbf{x}^* = QQ^T\mathbf{b}.$$

- Note $Q^TQ = I_{m\times m}$ and multiply both sides by Q^T to get

$$R\mathbf{x}^* = Q^T\mathbf{b}.$$

- Because R is upper triangular, solving this system for $\mathbf{x}^*$ is straightforward. This is done via back substitution.

Theorem 1.3.2 (Least squares via QR). *Let $A \in \mathbb{R}^{n\times m}$ be an $n \times m$ matrix with linearly independent columns, let $\mathbf{b} \in \mathbb{R}^n$ be a vector, and let $A = QR$ be a QR decomposition of A, where Q is a $\mathbb{R}^{n\times m}$ matrix with $Q^TQ = I_{m\times m}$ and R is upper triangular. The solution to the least-squares problem*

$$\min_{\mathbf{x}\in\mathbb{R}^m} \|A\mathbf{x} - \mathbf{b}\|,$$

satisfies

$$R\mathbf{x}^* = Q^T\mathbf{b}.$$

1.3.3 Linear regression

Given input data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with each $\mathbf{x}_i = (x_{i1}, ..., x_{id})^T$, we seek an affine function to fit the data. The common approach involves finding coefficients β_j's that minimize the criterion

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

where

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}$$

can be viewed as the predicted values of the linear model with coefficients β_j. The minimization problem can be formulated in matrix form. Let

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad A = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ 1 & \mathbf{x}_2^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^T \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}.$$

Then the problem is transformed to

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - A\boldsymbol{\beta}\|^2.$$

This is exactly the least-squares problem that we discuss in the last section.

1.4 Principal component analysis

Principal component analysis is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. Its underlying mathematics can be explained with singular value decomposition.

1.4.1 Singular value decomposition

Let A be an $m \times n$ matrix. Then $A^T A$ is symmetric and can be orthogonally diagonalized. Let $\mathbf{v}_1, ..., \mathbf{v}_n$ be an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of $A^T A$, and let $\lambda_1, ..., \lambda_n$ be the associated eigenvalues of $A^T A$. Then, for $1 \leq i \leq n$,

$$\begin{aligned} \|A\mathbf{v}_i\|^2 = (A\mathbf{v}_i)^T A\mathbf{v}_i &= \mathbf{v}_i^T A^T A\mathbf{v}_i \\ &= \mathbf{v}_i^T(\lambda_i \mathbf{v}_i) && \text{since } \mathbf{v}_i \text{ is an eigenvectors of } A^T A \\ &= \lambda_i && \text{since } \mathbf{v}_i \text{ is a unit vector} \end{aligned} \tag{1.4.1}$$

So the eigenvalues of A are all non-negative. By renumbering, if necessary, we may assume that the eigenvalues are arranged so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n \geq 0.$$

The singular values of A are the square roots of the eigenvalues of $A^T A$, denoted by $\sigma_1, ..., \sigma_n$, and they are arranged in decreasing order. That is, $\sigma_i = \sqrt{\lambda_i}$ for $1 \leq i \leq n$. The singular values of A are the lengths of the vectors $A\mathbf{v}_1, ..., A\mathbf{v}_n$.

Theorem 1.4.1. *If an $m \times n$ matrix A has r non-zero singular values, $\sigma_1, ..., \sigma_r \geq 0$ with $\sigma_{r+1} = \cdots = \sigma_n = 0$, then the dimension of* $\mathrm{col}(A) = r$.

Proof. Let $\mathbf{v}_1, ..., \mathbf{v}_n$ be an orthonormal basis of $\mathbb{R}^n$ of $A^T A$, ordered so that the corresponding eigenvalues of $A^T A$ satisfy $\lambda_1 \geq \cdots \lambda_n$. Then for $i \neq j$,

$$(A\mathbf{v}_i)^T (A\mathbf{v}_j) = \mathbf{v}_i^T A^T A\mathbf{v}_j = \mathbf{v}_i^T (\lambda_j \mathbf{v}_j) = 0$$

since $\mathbf{v}_i$ and $\lambda_j \mathbf{v}_j$ are orthogonal. Thus $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_n\}$ is an orthogonal set. Let r be the number of non-zero singular values of A; that is, r is the number of non-zero eigenvalues of $A^T A$. We see that $A\mathbf{v}_i \neq \mathbf{0}$ if and only if $1 \leq i \leq r$. Then $[A\mathbf{v}_1, ..., A\mathbf{v}_r]$ is linearly independent and clearly is in $\mathrm{col}(A)$. Furthermore, for any $\mathbf{y}$ in $\mathrm{col}(A)$—say, $\mathbf{y} = A\mathbf{x}$—we may write $\mathbf{x} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$, and

$$\begin{aligned}\mathbf{y} = A\mathbf{x} &= c_1 A\mathbf{v}_1 + \cdots + c_r A\mathbf{v}_r + c_{r+1} A\mathbf{v}_{r+1} + \cdots + c_n A\mathbf{v}_n \\ &= c_1 A\mathbf{v}_1 + \cdots + c_r A\mathbf{v}_r + \mathbf{0} + \cdots + 0.\end{aligned}$$

Thus y is in span of $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$, which shows that $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ is an (orthogonal) basis for $\mathrm{col}(A)$. Hence the dimension of $\mathrm{col}(A) = r$. □

The decomposition of A involves an $m \times n$ diagonal matrix $\sum$ of the form

$$\sum = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$

where D is an $r \times r$ diagonal matrix for some r not exceeding the smaller of m and n. (If r equals m or n or both, some or all of the zero matrices will not appear.)

Theorem 1.4.2 (The singular value decomposition). *Let A be an $m \times n$ matrix with the dimension of* $\mathrm{col}(A) = r$. *Then there exists an $m \times n$ matrix $\sum$,* *where the diagonal entries in D are the first r singular values of* A, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$, *and there exist an* $m \times m$ *orthogonal matrix* U *and an* $n \times n$ *orthogonal matrix* V *such that*

$$A = U \sum V^T.$$

Any factorization $A = U \sum V^T$, with U and V orthogonal and $\sum$, is called a singular value decomposition **SVD** of A. The matrices U and V are not unique, but the diagonal entries of $\sum$ are necessarily the singular values of A. The column of U in such a decomposition are called **left singular vectors** of A, and the column of V are called **right singular vectors** of A. This type of matrix factorization is illustrated in Fig. 1.4.

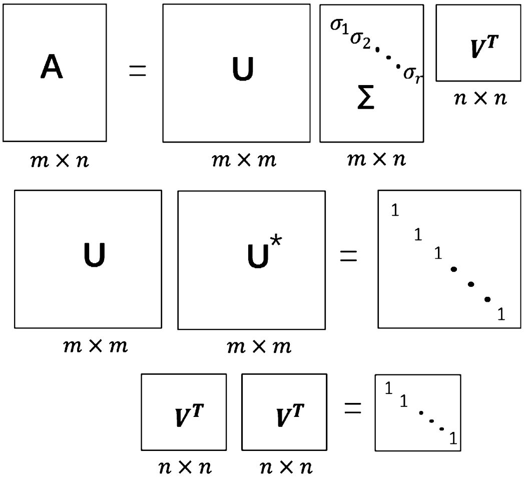

Figure 1.4 Singular value decomposition visualization.

Proof. Let λ_i be and $\mathbf{v}_i$ be as in the proof of Theorem 1.4.1. Then $\sigma_i = \sqrt{\lambda_i} = \|A\mathbf{v}_i\| \geq 0$ for $1 \leq i \leq r$, and $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ is an orthogonal basis for col(A). For $1 \leq i \leq r$, define

$$\mathbf{u}_i = \frac{1}{\|A\mathbf{v}_i\|} A\mathbf{v}_i = \frac{1}{\sigma_i} A\mathbf{v}_i$$

so that

$$A\mathbf{v}_i = \sigma_i \mathbf{u_i} \quad (1 \leq i \leq r). \tag{1.4.2}$$

Then $\mathbf{u}_1, ...\mathbf{u}_r$ is an orthonormal basis of lcol(A). Extend this set to an orthonormal basis $\mathbf{u}_1, ...\mathbf{u}_m$ of $\mathbb{R}^m$, and let

$$U = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_m \end{pmatrix} \quad \text{and} \quad V = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix}.$$

Then U and V are orthogonal matrices and

$$\begin{aligned} AV &= \begin{bmatrix} A\mathbf{v}_1 & \cdots & A\mathbf{v}_r & 0 & \cdots & 0 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1\mathbf{u}_1 & \cdots & \sigma_r\mathbf{u}_r & 0 & \cdots & 0 \end{bmatrix}. \end{aligned}$$

Let D be the diagonal matrix with diagonal entries $\sigma_1, ..., \sigma_r$. Then

$$\begin{aligned} U\sum &= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_m \end{pmatrix} \left(\begin{array}{cccc|c} \sigma_1 & & & 0 & \\ & \sigma_2 & & & 0 \\ & & \ddots & & \\ 0 & & & \sigma_r & \\ \hline & & 0 & & 0 \end{array} \right) \\ &= \begin{pmatrix} \sigma_1\mathbf{u}_1 & \cdots & \sigma_r\mathbf{u}_r & 0\cdots & 0 \end{pmatrix} = AV. \end{aligned}$$

Since V is an orthogonal matrices, $U\sum V^T = AVV^T = A$. □

1.4.2 Low-rank matrix approximations

In this section, we discuss low-rank approximations of matrices. We first introduce matrix norms, which allow us in particular to talk about the distance between two matrices.

Definition 1.4.3 (Induced norm). The 2-norm of a matrix $A \in \mathbb{R}^{n \times m}$ is

$$\|A\|_2 = \max_{\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^m} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\mathbf{x} \neq 0, \|\mathbf{x}\|=1} \|A\mathbf{x}\| = \max_{\mathbf{x} \neq 0, \|\mathbf{x}\|=1} \mathbf{x}^T A^T A\mathbf{x} \qquad (1.4.3)$$

Let $A \in \mathbb{R}^{n \times m}$ be a matrix with SVD,

$$A = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T. \qquad (1.4.4)$$

For $k < r$, truncate the sum at the kth term

$$A_k = \sum_{j=1}^{k} \sigma_j \mathbf{u}_j \mathbf{v}_j^T. \qquad (1.4.5)$$

The rank of A_k is exactly k. Indeed, by construction:

1. the vectors $\{\mathbf{u}_j : j = 1, \ldots, k\}$ are orthonormal, and
2. since $\sigma_j > 0$ for $j = 1, \ldots, k$ and the vectors $\{\mathbf{v}_j : j = 1, \ldots, k\}$ are orthonormal, $\{\mathbf{u}_j : j = 1, \ldots, k\}$ spans the column space of A_k.

Lemma 1.4.4 (Matrix norms and singular values). *Let $A \in \mathbb{R}^{n\times m}$ be a matrix with SVD,*

$$A = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T,$$

where recalling that $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > 0$ and letting A_k be the truncation defined above. Then

$$\|A - A_k\|_2^2 = \sigma_{k+1}^2.$$

Proof. For any $\mathbf{x} \neq 0$ and $\|\mathbf{x}\| = 1$,

$$\begin{aligned}\|(A - A_k)\mathbf{x}\|^2 &= \left\| \sum_{j=k+1}^{r} \sigma_j \mathbf{u}_j (\mathbf{v}_j^T \mathbf{x}) \right\|^2 = \sum_{j=k+1}^{r} \sigma_j^2 \langle \mathbf{v}_j, \mathbf{x}\rangle^2 \\ &= \mathbf{x}^T (A - A_k)^T (A - A_k)\mathbf{x}.\end{aligned}$$

Because the σ_j's are in decreasing order, this is maximized when $\langle \mathbf{v}_j, \mathbf{x}\rangle = 1$ if $j = k + 1$ and 0 otherwise. In view of Theorem 1.2.27, that is, we take $\mathbf{x} = \mathbf{v}_{k+1}$ and the norm is then σ_{k+1}^2, as claimed. □

With additional effort, we can prove the following theorem [5].

Theorem 1.4.5 (Eckart–Young–Mirsky theorem; Low-rank approximation in the induced norm). *Let $A \in \mathbb{R}^{n\times m}$ be a matrix with SVD,*

$$A = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T,$$

and let A_k be the truncation defined above with $k < r$. For any matrix $B \in \mathbb{R}^{n\times m}$ of rank at most k,

$$\|A - A_k\|_2 \leq \|A - B\|_2. \tag{1.4.6}$$

1.4.3 Principal component analysis

1.4.3.1 Covariance matrix

To prepare for principal component analysis, let $[\mathbf{X}_1 \cdots \mathbf{X}_N]$ be a $p \times N$ matrix of observation, such as described above. The **sample mean** M of the observation vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is given by

$$\mathbf{M} = \frac{1}{N}\left(\mathbf{X}_1 + \cdots + \mathbf{X}_N\right).$$

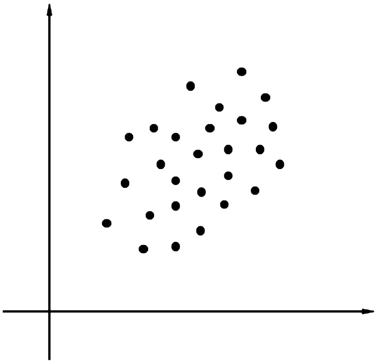

Figure 1.5 A scatter plot of observation vectors $\mathbf{X}_1, \ldots, \mathbf{X}_N$.

For the data in Fig. 1.5, the sample mean is the point in the "center" of the scatter plot. For $k = 1, \ldots, N$, let

$$\hat{\mathbf{X}}_k = \mathbf{X}_k - \mathbf{M}.$$

The columns of the $p \times N$ matrix

$$B = \left[\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \cdots \hat{\mathbf{X}}_N\right]$$

have a zero sample mean, and B is said to be in **mean-deviation form**. When the sample mean is subtracted from the data in Fig. 1.5, the resulting scatter plot has the form in Fig. 1.6.

The **(sample) covariance matrix** is the $p \times p$ matrix S defined by

$$S = \frac{1}{N-1} BB^T.$$

Since any matrix of the form BB^T is positive semidefinite, so is S.

1.4.3.2 Principal component analysis

Now assume that the columns of the $p \times N$ data matrix

$$X = [\mathbf{X}_1, \mathbf{X}_2, \cdots \mathbf{X}_N]$$

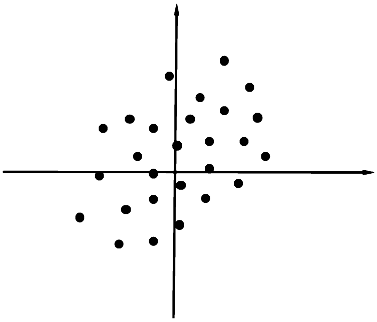

Figure 1.6 Weight-height data in mean-deviation form.

is already in mean-deviation form. The goal of principal component analysis (PCA) is to find k, ($k \leq p$) orthonormal vectors $\mathbf{v_1}, ..., \mathbf{v_k}$, (top k principal components) that maximize the objective function,

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}\langle \mathbf{X}_i \cdot \mathbf{v_j}\rangle^2, \tag{1.4.7}$$

$\langle \mathbf{X_i} \cdot \mathbf{v_j}\rangle$ is the length of projection of $\mathbf{X}_i$ on $\mathbf{v_j}$.

On the other hand, for each j, it is easy to see that

$$\mathbf{v_j}^T XX^T \mathbf{v_j} = (X^T\mathbf{v_j})^T(X^T\mathbf{v_j}) = \sum_{i=1}^{N}\langle \mathbf{X_i} \cdot \mathbf{v_j}\rangle^2 \tag{1.4.8}$$

where XX^T is a $p \times p$ matrix. As a result, for each $j \leq k$, the variance-maximization problem can be rephrased as

$$\mathrm{argmax}_{\mathbf{v}:||\mathbf{v}||=1}\mathbf{v_j}^T XX^T \mathbf{v_j}. \tag{1.4.9}$$

Assume that

$$XX^T = V\mathrm{diag}(\lambda_1, ..., \lambda_p)V^T, \text{or } \; V^T XX^T V = (\lambda_1, ..., \lambda_p).$$

In view of Theorem 1.4.5, we conclude that the optimal choice of the first k eigenvectors of XX^T corresponding to the first k largest eigenvalues, which are also the first k columns of $V = [\mathbf{v}_1, \cdots, \mathbf{v}_p]$ of the covariance matrix XX^T, which are called the principal components of the data (in the matrices of observations). The first principal component is the eigenvector corresponding to the largest eigenvalue of XX^T, the second principal component is the eigenvector corresponding to the second largest eigenvalue, and so on.

The orthogonal $p \times p$ matrix $V = \left[\mathbf{v}_1, \cdots, \mathbf{v}_p\right]$ that determines a change of variable, $\mathbf{x} = V\mathbf{y}$, or

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_p} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$$

with the property that the new variables $y_1, \ldots, y_p$ are uncorrelated and are arranged in order of decreasing variance. Indeed, we have

$$\mathbf{x}^T XX^T \mathbf{x} = \mathbf{y}^T V^T XX^T V\mathbf{y} = \mathbf{y}^T \mathrm{diag}(\lambda_1, ..., \lambda_p)\mathbf{y} = \sum_1^p \lambda_i y_i^2.$$

The orthogonal change of variable $\mathbf{x} = V\mathbf{y}$ that each observation vector $\mathbf{x}$ receives a "new name" $\mathbf{y}$, such that $\mathbf{x} = V\mathbf{y}$. Notice that $\mathbf{y} = V^{-1}\mathbf{x} = V^T\mathbf{x}$. Let $v_{1i}, \ldots, v_{pi}$ be the entries in $\mathbf{v}_i$. Since $\mathbf{v}_i^T$ the ith row of V^T the equation $\mathbf{y} = V^T\mathbf{x}$ shows that

$$y_i = \mathbf{v}_i^T \mathbf{x} = v_{1i}x_1 + v_{2i}x_2 + \cdots + v_{pi}x_p.$$

Thus y_i is a linear combination of the original variables $x_1, \ldots, x_p$, using the entries the eigenvector $\mathbf{v}_i$ as weights, which are called loadings.

1.4.3.3 Total variance

Given the columns of the $p \times N$ data matrix and assume it is already in mean-deviation form,

$$X = [\mathbf{X}_1, \mathbf{X}_2, \cdots \mathbf{X}_N],$$

and let covariance matrix S,

$$S = \frac{1}{N-1} XX^T.$$

The entries in $S = \left[S_{ij}\right]$, for $j = 1, \ldots, p$, the diagonal entry s_{jj} in S is called the **variance** of x_j, which is the first jth row of X. The variance of x_j measures the spread of the values of x_j. The total variance of the data is the sum of the variances on the diagonal of S. In general, the sum of the diagonal entries of a square matrix S is called the trace of the matrix, written $tr\,(S)$. Thus

$$\text{Total Variance} = tr\,(S)$$

Note that if

$$XX^T = V\text{diag}(\lambda_1, ..., \lambda_p)\,V^T, \;\text{ or }\; V^T XX^T V = \text{diag}(\lambda_1, ..., \lambda_p),$$

then

$$\text{tr}\,(S) = \frac{1}{N-1}\sum_1^p \lambda_j.$$

because $\text{tr}\left(VSV^T\right) = \text{tr}\,(S)$. Thus the fraction of the variances of the first k term truncation is

$$\frac{\sum_1^k \lambda_j}{\sum_1^p \lambda_j}.$$