Part VII

Orthogonality

Section 31

Orthogonal Diagonalization

Focus Questions

By the end of this section, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the section.

- What does it mean for a matrix to be orthogonally diagonalizable and why is this concept important?
- What is a symmetric matrix and what important property related to diagonalization does a symmetric matrix have?
- What is the spectrum of a matrix?

Application: The Multivariable Second Derivative Test

In single variable calculus, we learn that the second derivative can be used to classify a critical point of the type where the derivative of a function is 0 as a local maximum or minimum.

Theorem 31.1 (The Second Derivative Test for Single-Variable Functions)**.** *If a is a critical number of a function f so that $f'(a) = 0$ and if $f''(a)$ exists, then*

- *if $f''(a) < 0$, then $f(a)$ is a local maximum value of f,*
- *if $f''(a) > 0$, then $f(a)$ is a local minimum value of f, and*
- *if $f''(a) = 0$, this test yields no information.*

In the two-variable case we have an analogous test, which is usually seen in a multivariable calculus course.

Theorem 31.2 (The Second Derivative Test for Functions of Two Variables)**.** *Suppose (a, b) is a critical point of the function f for which $f_x(a, b) = 0$ and $f_y(a, b) = 0$. Let D be the quantity*

defined by

$$D = f_{xx}(a,b)f_{yy}(a,b) - f_{xy}(a,b)^2.$$

- *If $D > 0$ and $f_{xx}(a,b) < 0$, then f has a local maximum at (a,b).*
- *If $D > 0$ and $f_{xx}(a,b) > 0$, then f has a local minimum at (a,b).*
- *If $D < 0$, then f has a saddle point at (a,b).*
- *If $D = 0$, then this test yields no information about what happens at (a,b).*

A proof of this test for two-variable functions is based on Taylor polynomials, and relies on symmetric matrices, eigenvalues, and quadratic forms. The steps for a proof will be found later in this section.

Introduction

We have seen how to diagonalize a matrix – if we can find n linearly independent eigenvectors of an $n \times n$ matrix A and let P be the matrix whose columns are those eigenvectors, then $P^{-1}AP$ is a diagonal matrix with the eigenvalues down the diagonal in the same order corresponding to the eigenvectors placed in P. We will see that in certain cases we can take this one step further and create an orthogonal matrix with eigenvectors as columns to diagonalize a matrix. This is called orthogonal diagonalization. Orthogonal diagonalizability is useful in that it allows us to find a "convenient" coordinate system in which to interpret the results of certain matrix transformations. A set of orthonormal basis vectors for an orthogonally diagonalizable matrix A is called a set of *principal axes* for A. Orthogonal diagonalization will also play a crucial role in the singular value decomposition of a matrix, a decomposition that has been described by some as the "pinnacle" of linear algebra.

Definition 31.3. An $n \times n$ matrix A is **orthogonally diagonalizable** if there is an orthogonal matrix P such that

$$P^{\mathsf{T}}AP$$

is a diagonal matrix. We say that the matrix P *orthogonally diagonalizes* the matrix A.

Preview Activity 31.1.

(1) For each matrix A whose eigenvalues and corresponding eigenvectors are given, find a matrix P such that $P^{-1}AP$ is a diagonal matrix.

(a) $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ with eigenvalues -1 and 3 and corresponding eigenvectors $\mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

(b) $A = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$ with eigenvalues 0 and 3 and corresponding eigenvectors $\mathbf{v}_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

(c) $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ with eigenvalues 0, 1, and 3 and corresponding eigenvectors $\mathbf{v}_1 = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$, and $\mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$.

(2) Which matrices in part 1 seem to satisfy the orthogonal diagonalization requirement? Do you notice any common traits among these matrices?

Symmetric Matrices

As we saw in Preview Activity 31.1, matrices that are not symmetric need not be orthogonally diagonalizable, but the symmetric matrix examples are orthogonally diagonalizable. We explore that idea in this section.

If P is a matrix that orthogonally diagonalizes the matrix A, then $P^\mathsf{T}AP = D$, where D is a diagonal matrix. Since $D^\mathsf{T} = D$ and $A = PDP^\mathsf{T}$, we have

$$\begin{aligned} A &= PDP^\mathsf{T} \\ &= PD^\mathsf{T}P^\mathsf{T} \\ &= \left(P^\mathsf{T}\right)^\mathsf{T} D^\mathsf{T}P^\mathsf{T} \\ &= \left(PDP^\mathsf{T}\right)^\mathsf{T} \\ &= A^\mathsf{T}. \end{aligned}$$

Therefore, $A^\mathsf{T} = A$ and matrices with this property are the only matrices that can be orthogonally diagonalized. Recall that any matrix A satisfying $A^\mathsf{T} = A$ is a symmetric matrix.

While we have just shown that the only matrices that can be orthogonally diagonalized are the symmetric matrices, the amazing thing about symmetric matrices is that *every* symmetric matrix can be orthogonally diagonalized. We will prove this shortly.

Symmetric matrices have useful properties, a few of which are given in the following activity (we will use some of these properties later in this section).

Activity 31.1. Let A be a symmetric $n \times n$ matrix and let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\mathbb{R}^n$.

(a) Show that $\mathbf{x}^\mathsf{T}A\mathbf{y} = (A\mathbf{x})^\mathsf{T}\mathbf{y}$.

(b) Show that $(A\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (A\mathbf{y})$.

(c) Show that the eigenvalues of a 2×2 symmetric matrix $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ are real.

Activity 31.1 (c) shows that a 2×2 symmetric matrix has real eigenvalues. This is a general result about real symmetric matrices.

Theorem 31.4. *Let A be an $n \times n$ symmetric matrix with real entries. Then the eigenvalues of A are real.*

Proof. Let A be an $n \times n$ symmetric matrix with real entries and let λ be an eigenvalue of A with eigenvector $\mathbf{v}$. To show that λ is real, we will show that $\overline{\lambda} = \lambda$. We know

$$A\mathbf{v} = \lambda\mathbf{v}. \tag{31.1}$$

Since A has real entries, we also know that $\overline{\lambda}$ is an eigenvalue for A with eigenvector $\overline{\mathbf{v}}$. Multiply both sides of (31.1) on the left by $\overline{\mathbf{v}}^{\mathsf{T}}$ to obtain

$$\overline{\mathbf{v}}^{\mathsf{T}} A\mathbf{v} = \overline{\mathbf{v}}^{\mathsf{T}} \lambda\mathbf{v} = \lambda\left(\overline{\mathbf{v}}^{\mathsf{T}}\mathbf{v}\right). \tag{31.2}$$

Now

$$\overline{\mathbf{v}}^{\mathsf{T}} A\mathbf{v} = (A\overline{\mathbf{v}})^{\mathsf{T}}\mathbf{v} = (\overline{\lambda}\,\overline{\mathbf{v}})^{\mathsf{T}}\mathbf{v} = \overline{\lambda}\left(\overline{\mathbf{v}}^{\mathsf{T}}\mathbf{v}\right)$$

and equation (31.2) becomes

$$\overline{\lambda}\left(\overline{\mathbf{v}}^{\mathsf{T}}\mathbf{v}\right) = \lambda\left(\overline{\mathbf{v}}^{\mathsf{T}}\mathbf{v}\right).$$

Since $\mathbf{v} \neq \mathbf{0}$, this implies that $\overline{\lambda} = \lambda$ and λ is real. ■

To orthogonally diagonalize a matrix, it must be the case that eigenvectors corresponding to different eigenvalues are orthogonal. This is an important property and it would be useful to know when it happens.

Activity 31.2. Let A be a real symmetric matrix with eigenvalues λ_1 and λ_2 and corresponding eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$, respectively.

(a) Use Activity 31.1 (b) to show that $\lambda_1\mathbf{v}_1 \cdot \mathbf{v}_2 = \lambda_2\mathbf{v}_1 \cdot \mathbf{v}_2$.

(b) Explain why the result of part (a) shows that $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal if $\lambda_1 \neq \lambda_2$.

Activity 31.2 proves the following theorem.

Theorem 31.5. *If A is a real symmetric matrix, then eigenvectors corresponding to distinct eigenvalues are orthogonal.*

Recall that the only matrices than can be orthogonally diagonalized are the symmetric matrices. Now we show that every real symmetric matrix can be orthogonally diagonalized, which completely characterizes the matrices that are orthogonally diagonalizable.

Theorem 31.6. *Let A be a real symmetric matrix. Then A is orthogonally diagonalizable.*

Proof. We will assume that all matrices are real matrices. To prove that every symmetric matrix is orthogonally diagonalizable, we will proceed by contradiction and assume that there are $n \times n$ symmetric matrices that are not orthogonally diagonalizable for some values of n. Since n must be positive (greater than 1, in fact, since every 1×1 matrix is orthogonally diagonalizable), there must be a smallest value of n so that there is an $n \times n$ symmetric matrix that is not orthogonally diagonalizable. Let A be one of these smallest $n \times n$ matrices that is not orthogonally diagonalizable. By

our assumption, every $k \times k$ symmetric matrix M with $k < n$ is orthogonally diagonalizable. Since A is not orthogonally diagonalizable, there cannot exist an orthogonal basis for $\mathbb{R}^n$ that consists of eigenvectors of A. We will now show that this is impossible, which will force our assumption that A is not orthogonally diagonalizable to be false.

Since A is a symmetric matrix, we know that A has n (counting multiplicities) real eigenvalues (not necessarily distinct). Let λ_1 be one of these real eigenvalues and $\mathbf{u}_1$ a corresponding unit eigenvector. Let

$$W = \text{Span}\{\mathbf{u}_1\}^{\perp}.$$

So $\mathbf{x} \in W$ if $0 = \mathbf{u}_1 \cdot \mathbf{x} = \mathbf{u}_1^{\mathsf{T}}\mathbf{x}$. Now W is a subspace of $\mathbb{R}^n$ and $\text{Span}\{\mathbf{u}_1\}$ has dimension 1, so $\dim(W) = n - 1$. We can then construct an orthonormal basis $\mathcal{B} = \{\mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_n\}$ for W.

Note that W is invariant under left multiplication by A. To see why, let $\mathbf{w} \in W$. Then

$$\mathbf{u}_1 \cdot A\mathbf{w} = (A\mathbf{u}_1)^{\mathsf{T}}\mathbf{w} = \lambda_1 \mathbf{u}_1^{\mathsf{T}}\mathbf{w} = \lambda_1 \mathbf{u}_1 \cdot \mathbf{w} = 0,$$

so $A\mathbf{w}$ is orthogonal to $\mathbf{u}_1$ and is therefore in W.

Since W is invariant under left multiplication by A, we can think of left multiplication by A as a linear transformation from the $n-1$ dimensional vector space W to itself. We call this the *restriction* of A to W and will denote it by $A|_W$. The coordinate transformation T defined by $T(\mathbf{x}) = [\mathbf{x}]_{\mathcal{B}}$ maps the $n-1$ dimensional vector space W to $\mathbb{R}^{n-1}$ and we know that T is an invertible map. We can then view left multiplication by A on W as a composite of the coordinate map T, a matrix transformation M from $\mathbb{R}^{n-1}$ to $\mathbb{R}^{n-1}$ and then T^{-1} as shown in Figure 31.1.

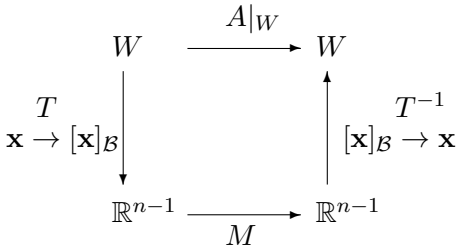

Figure 31.1: Composite diagram

Let us focus on the coordinate mapping T for a moment. This mapping has two important properties that we will need.

Claim.

(1) $T(\mathbf{x}) \cdot T(\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ *for every* $\mathbf{x}, \mathbf{y}$ *in* W.

(2) $T^{-1}([\mathbf{x}]_{\mathcal{B}}) \cdot T^{-1}([\mathbf{y}]_{\mathcal{B}}) = [\mathbf{x}]_{\mathcal{B}} \cdot [\mathbf{y}]_{\mathcal{B}}$ *for every* $[\mathbf{x}]_{\mathcal{B}}, [\mathbf{y}]_{\mathcal{B}}$ *in* $\mathbb{R}^{n-1}$.

Proof of the Claim: Let $\mathbf{x}, \mathbf{y}$ in W. Since $\mathcal{B}$ is a basis for W, we know that

$$\mathbf{x} = x_2\mathbf{u}_2 + x_3\mathbf{u}_3 + \cdots + x_n\mathbf{u}_n \text{ and } \mathbf{y} = y_2\mathbf{u}_2 + y_3\mathbf{u}_3 + \cdots + y_n\mathbf{u}_n$$

for some scalars x_2, x_3, ..., x_n and y_2, y_3, ..., y_n. So $[\mathbf{x}]_{\mathcal{B}} = [x_2 \ x_3 \ \ldots \ x_n]^{\mathsf{T}}$ and $[\mathbf{y}]_{\mathcal{B}} = [y_2 \ y_3 \ \ldots \ y_n]^{\mathsf{T}}$ and it follows that

$$T(\mathbf{x}) \cdot T(\mathbf{y}) = [\mathbf{x}]_{\mathcal{B}} \cdot [\mathbf{y}]_{\mathcal{B}} = x_2y_2 + x_3y_3 + \cdots + x_ny_n. \tag{31.3}$$

Now $\mathcal{B}$ is an orthonormal basis, so $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ for $i \neq j$, $\mathbf{u}_i \cdot \mathbf{u}_i = 1$ for all i, and

$$\begin{aligned}\mathbf{x} \cdot \mathbf{y} &= (x_2\mathbf{u}_2 + x_3\mathbf{u}_3 + \cdots + x_n\mathbf{u}_n) \cdot (y_2\mathbf{u}_2 + y_3\mathbf{u}_3 + \cdots + y_n\mathbf{u}_n) \\ &= (x_2y_2\mathbf{u}_2 \cdot \mathbf{u}_2) + (x_3y_3\mathbf{u}_3 \cdot \mathbf{u}_3) + \cdots + (x_ny_n\mathbf{u}_n \cdot \mathbf{u}_n) \\ &= x_2y_2 + x_3y_3 + \cdots + x_ny_n. \end{aligned} \tag{31.4}$$

Thus, $T(\mathbf{x}) \cdot T(\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$, proving part 1 of the claim.

For part 2, note that

$$T^{-1}\left([\mathbf{x}]_{\mathcal{B}}\right) \cdot T^{-1}\left([\mathbf{y}]_{\mathcal{B}}\right) = \mathbf{x} \cdot \mathbf{y} = [\mathbf{x}]_{\mathcal{B}} \cdot [\mathbf{y}]_{\mathcal{B}}$$

by (31.3) and (31.4), verifying part 2 of the claim. ■

We now apply this claim to show that the matrix M is a symmetric matrix. Since the coordinate transformation T is an onto mapping, every vector in $\mathbb{R}^{n-1}$ can be written as $[\mathbf{x}]_{\mathcal{B}}$ for some $\mathbf{x}$ in W. Let $[\mathbf{x}]_{\mathcal{B}}$ and $[\mathbf{y}]_{\mathcal{B}}$ be arbitrary vectors in $\mathbb{R}^{n-1}$. From Figure 31.1 we can see that $M = TA|_W T^{-1} = TAT^{-1}$, so

$$\begin{aligned} M[\mathbf{x}]_{\mathcal{B}} \cdot [\mathbf{y}]_{\mathcal{B}} &= TAT^{-1}([\mathbf{x}]_{\mathcal{B}}) \cdot [\mathbf{y}]_{\mathcal{B}} \\ &= TAT^{-1}([\mathbf{x}]_{\mathcal{B}}) \cdot TT^{-1}\left([\mathbf{y}]_{\mathcal{B}}\right) \\ &= AT^{-1}([\mathbf{x}]_{\mathcal{B}}) \cdot T^{-1}\left([\mathbf{y}]_{\mathcal{B}}\right) \\ &= T^{-1}([\mathbf{x}]_{\mathcal{B}}) \cdot AT^{-1}\left([\mathbf{y}]_{\mathcal{B}}\right) \\ &= TT^{-1}([\mathbf{x}]_{\mathcal{B}}) \cdot TAT^{-1}\left([\mathbf{y}]_{\mathcal{B}}\right) \\ &= [\mathbf{x}]_{\mathcal{B}} \cdot M[\mathbf{y}]_{\mathcal{B}}. \end{aligned}$$

Therefore, M is a symmetric matrix. Since M is $(n-1) \times (n-1)$, we can conclude that M is orthogonally diagonalizable. Thus, there is an orthonormal basis $\{[\mathbf{x}_2]_{\mathcal{B}}, [\mathbf{x}_3]_{\mathcal{B}}, \ldots, [\mathbf{x}_n]_{\mathcal{B}}\}$ for $\mathbb{R}^{n-1}$ consisting of eigenvectors of M. Let $M[\mathbf{x}_i]_{\mathcal{B}} = \lambda_i[\mathbf{x}_i]_{\mathcal{B}}$ for each $2 \leq i \leq n$.

Finally, we will show that the set $\mathcal{C} = \{\mathbf{u}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n\}$ is an orthonormal set of eigenvectors of A. This will contradict the fact that we assumed that A was *not* orthogonally diagonalizable. The set $\mathcal{B}$ is an orthonormal subset of W and so by the definition of W each vector in $\mathcal{B}$ is orthogonal to $\mathbf{u}_1$. Thus the set $\mathcal{C}$ is an orthogonal set. To complete our proof we only need show that each $\mathbf{x}_i$ is an eigenvector of A. Choose an i arbitrarily between 2 and n. Now

$$A\mathbf{x}_i = T^{-1}MT(\mathbf{x}_i) = T^{-1}M[\mathbf{x}_i]_{\mathcal{B}} = T^{-1}\left(\lambda_i[\mathbf{x}_i]_{\mathcal{B}}\right) = \lambda_i T^{-1}\left([\mathbf{x}_i]_{\mathcal{B}}\right) = \lambda_i\mathbf{x}_i.$$

Therefore, the set $\mathcal{C} = \{\mathbf{u}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n\}$ is an orthogonal set of eigenvectors of A. Since the transformations T and T^{-1} preserve dot products, they also preserve lengths. So, in fact, $\mathcal{C}$ is an orthonormal set of eigenvectors of A as desired. This completes our proof. ■

The set of eigenvalues of a matrix A is called the *spectrum* of A and we have just proved the following theorem.

Theorem 31.7 (The Spectral Theorem for Real Symmetric Matrices)**.** *Let A be an $n \times n$ symmetric matrix with real entries. Then*

(1) A has n real eigenvalues (counting multiplicities)

(2) the dimension of each eigenspace of A is the multiplicity of the corresponding eigenvalue as a root of the characteristic polynomial

(3) eigenvectors corresponding to different eigenvalues are orthogonal

(4) A is orthogonally diagonalizable.

So *any* real symmetric matrix is orthogonally diagonalizable. We have seen examples of the orthogonal diagonalization of $n \times n$ real symmetric matrices with n distinct eigenvalues, but how do we orthogonally diagonalize a symmetric matrix having eigenvalues of multiplicity greater than 1? The next activity shows us the process.

Activity 31.3. Let $A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}$. The eigenvalues of A are 2 and 8, with eigenspace of dimension 2 and dimension 1, respectively.

(a) Explain why A can be orthogonally diagonalized.

(b) Two linearly independent eigenvectors for A corresponding to the eigenvalue 2 are $\mathbf{v}_1 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$. Note that $\mathbf{v}_1, \mathbf{v}_2$ are not orthogonal, so cannot be in an orthogonal basis of $\mathbb{R}^3$ consisting of eigenvectors of A. So find a set $\{\mathbf{w}_1, \mathbf{w}_2\}$ of orthogonal eigenvectors of A so that $\text{Span}\{\mathbf{w}_1, \mathbf{w}_2\} = \text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$.

(c) The vector $\mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ is an eigenvector for A corresponding to the eigenvalue 8. What can you say about the orthogonality relationship between $\mathbf{w}_i$'s and $\mathbf{v}_3$?

(d) Find a matrix P that orthogonally diagonalizes A. Verify your work.

The Spectral Decomposition of a Symmetric Matrix A

Let A be an $n \times n$ symmetric matrix with real entries. The Spectral Theorem tells us we can find an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ of eigenvectors of A. Let $A\mathbf{u}_i = \lambda_i \mathbf{u}_i$ for each $1 \leq i \leq n$. If $P = [\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3\ \cdots\ \mathbf{u}_n]$, then we know that

$$P^{\mathsf{T}} A P = P^{-1} A P = D,$$

where D is the $n \times n$ diagonal matrix

$$\begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}.$$

Since $A = PDP^{\mathsf{T}}$ we see that

$$\begin{aligned} A &= [\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3\ \cdots\ \mathbf{u}_n] \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^{\mathsf{T}} \\ \mathbf{u}_2^{\mathsf{T}} \\ \mathbf{u}_3^{\mathsf{T}} \\ \vdots \\ \mathbf{u}_n^{\mathsf{T}} \end{bmatrix} \\ &= [\lambda_1\mathbf{u}_1\ \lambda_2\mathbf{u}_2\ \lambda_3\mathbf{u}_3\ \cdots\ \lambda_n\mathbf{u}_n] \begin{bmatrix} \mathbf{u}_1^{\mathsf{T}} \\ \mathbf{u}_2^{\mathsf{T}} \\ \mathbf{u}_3^{\mathsf{T}} \\ \vdots \\ \mathbf{u}_n^{\mathsf{T}} \end{bmatrix} \\ &= \lambda_1\mathbf{u}_1\mathbf{u}_1^{\mathsf{T}} + \lambda_2\mathbf{u}_2\mathbf{u}_2^{\mathsf{T}} + \lambda_3\mathbf{u}_3\mathbf{u}_3^{\mathsf{T}} + \cdots + \lambda_n\mathbf{u}_n\mathbf{u}_n^{\mathsf{T}}, \end{aligned} \tag{31.5}$$

where the last product follows from Exercise 4. The expression in (31.5) is called a *spectral decomposition* of the matrix A. Let $P_i = \mathbf{u}_i\mathbf{u}_i^{\mathsf{T}}$ for each i. The matrices P_i satisfy several special conditions given in the next theorem. The proofs are left to the exercises.

Theorem 31.8. *Let A be an $n \times n$ symmetric matrix with real entries, and let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be an orthonormal basis of eigenvectors of A with $A\mathbf{u}_i = \lambda_i\mathbf{u}_i$ for each i. For each i, let $P_i = \mathbf{u}_i\mathbf{u}_i^{\mathsf{T}}$. Then*

(1) $A = \lambda_1 P_1 + \lambda_2 P_2 + \cdots + \lambda_n P_n$,

(2) P_i *is a symmetric matrix for each* i,

(3) P_i *is a rank 1 matrix for each* i,

(4) $P_i^2 = P_i$ *for each* i,

(5) $P_i P_j = 0$ *if* $i \neq j$,

(6) $P_i\mathbf{u}_i = \mathbf{u}_i$ *for each* i,

(7) $P_i\mathbf{u}_j = \mathbf{0}$ *if* $i \neq j$,

(8) For any vector $\mathbf{v}$ *in* $\mathbb{R}^n$, $P_i\mathbf{v} = proj_{Span\{\mathbf{u}_i\}}\mathbf{v}$.

The consequence of Theorem 31.8 is that any symmetric matrix can be written as the sum of symmetric, rank 1 matrices. As we will see later, this kind of decomposition contains much information about the matrix product $A^{\mathsf{T}}A$ for any matrix A.

Activity 31.4. Let $A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}$. Let $\lambda_1 = 2$, $\lambda_2 = 2$, and $\lambda_3 = 8$ be the eigenvalues of A. A basis for the eigenspace E_8 of A corresponding to the eigenvalue 8 is $\{[1\ 1\ 1]^{\mathsf{T}}\}$ and a basis for the eigenspace E_2 of A corresponding to the eigenvalue 2 is $\{[1\ -1\ 0]^{\mathsf{T}}, [1\ 0\ -1]^{\mathsf{T}}\}$. (Compare to Activity 31.3.)

(a) Find orthonormal eigenvectors $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ of A corresponding to λ_1, λ_2, and λ_3, respectively.

(b) Compute $\lambda_1 \mathbf{u}_1 \mathbf{u}_1^\mathsf{T}$

(c) Compute $\lambda_2 \mathbf{u}_2 \mathbf{u}_2^\mathsf{T}$

(d) Compute $\lambda_3 \mathbf{u}_3 \mathbf{u}_3^\mathsf{T}$

(e) Verify that $A = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^\mathsf{T} + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^\mathsf{T} + \lambda_3 \mathbf{u}_3 \mathbf{u}_3^\mathsf{T}$.

Examples

What follows are worked examples that use the concepts from this section.

Example 31.9. For each of the following matrices A, determine if A is diagonalizable. If A is not diagonalizable, explain why. If A is diagonalizable, find a matrix P so that $P^{-1}AP$ is a diagonal matrix. If the matrix is diagonalizable, is it orthogonally diagonalizable? If orthogonally diagonalizable, find an orthogonal matrix that diagonalizes A. Use appropriate technology to find eigenvalues and eigenvectors.

(a) $A = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 3 & 2 \\ 1 & -1 & 0 \end{bmatrix}$ (b) $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ (c) $A = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 7 & 2 \\ 1 & 2 & 4 \end{bmatrix}$

Example Solution.

(a) Recall that an $n \times n$ matrix A is diagonalizable if and only if A has n linearly independent eigenvectors, and A is orthogonally diagonalizable if and only if A is symmetric. Since A is not symmetric, A is not orthogonally diagonalizable. Technology shows that the eigenvalues of A are 2 and 1 and bases for the corresponding eigenspaces are $\{[1\ 1\ 0]^\mathsf{T}, [2\ 0\ 1]^\mathsf{T}\}$ and $\{[0\ -1\ 1]^\mathsf{T}\}$. So A is diagonalizable and if $P = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}$, then

$$P^{-1}AP = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

(b) Since A is not symmetric, A is not orthogonally diagonalizable. Technology shows that the eigenvalues of A are 0 and 1 and bases for the corresponding eigenspaces are $\{[0\ 0\ 1]^\mathsf{T}\}$ and $\{[1\ 0\ 0]^\mathsf{T}\}$. We cannot create a basis of $\mathbb{R}^3$ consisting of eigenvectors of A, so A is not diagonalizable.

(c) Since A is symmetric, A is orthogonally diagonalizable. Technology shows that the eigenvalues of A are 3 and 9 and bases for the eigenspaces $\{[-1\ 0\ 1]^\mathsf{T}, [-2\ 1\ 0]^\mathsf{T}\}$ and $\{[1\ 2\ 1]^\mathsf{T}\}$, respectively. To find an orthogonal matrix that diagonalizes A, we must find an orthonormal basis of $\mathbb{R}^3$ consisting of eigenvectors of A. To do that, we use the Gram-Schmidt process

to obtain an orthogonal basis for the eigenspace of A corresponding to the eigenvalue 3. Doing so gives an orthogonal basis $\{\mathbf{v}_1, \mathbf{v}_2\}$, where $\mathbf{v}_1 = [-1\ 0\ 1]^\mathsf{T}$ and

$$\begin{aligned}\mathbf{v}_2 &= [-2\ 1\ 0]^\mathsf{T} - \frac{[-2\ 1\ 0]^\mathsf{T} \cdot [-1\ 0\ 1]^\mathsf{T}}{[-1\ 0\ 1]^\mathsf{T} \cdot [-1\ 0\ 1]^\mathsf{T}}[-1\ 0\ 1]^\mathsf{T} \\ &= [-2\ 1\ 0]^\mathsf{T} - [-1\ 0\ 1]^\mathsf{T} \\ &= [-1\ 1\ -1]^\mathsf{T}.\end{aligned}$$

So an orthonormal basis for $\mathbb{R}^3$ of eigenvectors of A is

$$\left\{\frac{1}{\sqrt{2}}[-1\ 0\ 1]^\mathsf{T}, \frac{1}{\sqrt{3}}[-1\ 1\ -1]^\mathsf{T}, \frac{1}{\sqrt{6}}[1\ 1\ 1]^\mathsf{T}\right\}.$$

Therefore, A is orthogonally diagonalizable and if P is the matrix $\begin{bmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \end{bmatrix}$, then

$$P^{-1}AP = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 9 \end{bmatrix}.$$

Example 31.10. Let $A = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$. Find an orthonormal basis for $\mathbb{R}^4$ consisting of eigenvectors of A.

Example Solution.

Since A is symmetric, there is an orthogonal matrix P such that $P^{-1}AP$ is diagonal. The columns of P will form an orthonormal basis for $\mathbb{R}^4$. Using a cofactor expansion along the first row shows that

$$\begin{aligned}\det(A - \lambda I_4) &= \det\left(\begin{bmatrix} -\lambda & 0 & 0 & 1 \\ 0 & -\lambda & 1 & 0 \\ 0 & 1 & -\lambda & 0 \\ 1 & 0 & 0 & -\lambda \end{bmatrix}\right) \\ &= \left(\lambda^2 - 1\right)^2 \\ &= (\lambda + 1)^2(\lambda - 1)^2.\end{aligned}$$

So the eigenvalues of A are 1 and -1. The reduced row echelon forms of $A - I_4$ and $A + I_4$ are, respectively,

$$\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Thus, a basis for the eigenspace E_1 of A is $\{[0\ 1\ 1\ 0]^\mathsf{T}, [1\ 0\ 0\ 1]^\mathsf{T}\}$ and a basis for the eigenspace E_{-1} of A is $\{[0\ 1\ -1\ 0]^\mathsf{T}, [1\ 0\ 0\ -1]^\mathsf{T}\}$. The set $\{[0\ 1\ 1\ 0]^\mathsf{T}, [1\ 0\ 0\ 1]^\mathsf{T}, [0\ 1\ -1\ 0]^\mathsf{T}, [1\ 0\ 0\ -1]^\mathsf{T}\}$

is an orthogonal set, so an orthonormal basis for $\mathbb{R}^4$ consisting of eigenvectors of A is

$$\left\{ \frac{1}{\sqrt{2}}[0\ 1\ 1\ 0]^{\mathsf{T}}, \frac{1}{\sqrt{2}}[1\ 0\ 0\ 1]^{\mathsf{T}}, \frac{1}{\sqrt{2}}[0\ 1\ -1\ 0]^{\mathsf{T}}, \frac{1}{\sqrt{2}}[1\ 0\ 0\ -1]^{\mathsf{T}} \right\}.$$

Summary

- An $n \times n$ matrix A is orthogonally diagonalizable if there is an orthogonal matrix P such that $P^{\mathsf{T}}AP$ is a diagonal matrix. Orthogonal diagonalizability is useful in that it allows us to find a "convenient" coordinate system in which to interpret the results of certain matrix transformations. Orthogonal diagonalization also a plays a crucial role in the singular value decomposition of a matrix.

- An $n \times n$ matrix A is symmetric if $A^{\mathsf{T}} = A$. The symmetric matrices are exactly the matrices that can be orthogonally diagonalized.

- The spectrum of a matrix is the set of eigenvalues of the matrix.

Exercises

(1) For each of the following matrices, find an orthogonal matrix P so that $P^{\mathsf{T}}AP$ is a diagonal matrix, or explain why no such matrix exists.

(a) $A = \begin{bmatrix} 3 & -4 \\ -4 & -3 \end{bmatrix}$ (b) $A = \begin{bmatrix} 4 & 1 & 1 \\ 1 & 1 & 4 \\ 1 & 4 & 1 \end{bmatrix}$ (c) $A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \\ 3 & 0 & 5 & 2 \end{bmatrix}$

(2) For each of the following matrices find an orthonormal basis of eigenvectors of A. Then find a spectral decomposition of A.

(a) $A = \begin{bmatrix} 3 & -4 \\ -4 & -3 \end{bmatrix}$ (b) $A = \begin{bmatrix} 4 & 1 & 1 \\ 1 & 1 & 4 \\ 1 & 4 & 1 \end{bmatrix}$

(c) $A = \begin{bmatrix} -4 & 0 & -24 \\ 0 & -8 & 0 \\ -24 & 0 & 16 \end{bmatrix}$ (d) $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & 2 & -3 \end{bmatrix}$

(3) Find a non-diagonal 4×4 matrix with eigenvalues 2, 3 and 6 which can be orthogonally diagonalized.

(4) Let $A = [a_{ij}] = [\mathbf{c}_1\ \mathbf{c}_2\ \cdots\ \mathbf{c}_m]$ be an $k \times m$ matrix with columns $\mathbf{c}_1$, $\mathbf{c}_2$, ..., $\mathbf{c}_m$, and let

$B = [b_{ij}] = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_m \end{bmatrix}$ be an $m \times n$ matrix with rows $\mathbf{r}_1$, $\mathbf{r}_2$, ..., $\mathbf{r}_m$. Show that

$$AB = [\mathbf{c}_1\ \mathbf{c}_2\ \cdots\ \mathbf{c}_m] \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_m \end{bmatrix} = \mathbf{c}_1\mathbf{r}_1 + \mathbf{c}_2\mathbf{r}_2 + \cdots + \mathbf{c}_m\mathbf{r}_m.$$

(5) Let A be an $n \times n$ symmetric matrix with real entries and let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be an orthonormal basis of eigenvectors of A. For each i, let $P_i = \mathbf{u}_i\mathbf{u}_i^\mathsf{T}$. Prove Theorem 31.8 – that is, verify each of the following statements.

(a) For each i, P_i is a symmetric matrix.

(b) For each i, P_i is a rank 1 matrix.

(c) For each i, $P_i^2 = P_i$.

(d) If $i \neq j$, then $P_iP_j = 0$.

(e) For each i, $P_i\mathbf{u}_i = \mathbf{u}_i$.

(f) If $i \neq j$, then $P_i\mathbf{u}_j = 0$.

(g) If $\mathbf{v}$ is in $\mathbb{R}^n$, show that

$$P_i\mathbf{v} = \text{proj}_{\text{Span}\{\mathbf{u}_i\}}\mathbf{v}.$$

For this reason we call P_i an *orthogonal projection matrix*.

(6) Show that if M is an $n \times n$ matrix and $(M\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (M\mathbf{y})$ for every $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^n$, then M is a symmetric matrix. (Hint: Try $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$.)

(7) Let A be an $n \times n$ symmetric matrix and assume that A has an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ of eigenvectors of A so that $A\mathbf{u}_i = \lambda_i\mathbf{u}_i$ for each i. Let $P_i = \mathbf{u}_i\mathbf{u}_i^\mathsf{T}$ for each i. It is possible that not all of the eigenvalue of A are distinct. In this case, some of the eigenvalues will be repeated in the spectral decomposition of A. If we want only distinct eigenvalues to appear, we might do the following. Let $\mu_1, \mu_2, \ldots, \mu_k$ be the distinct eigenvalues of A. For each j between 1 and k, let Q_j be the sum of all of the P_i that have μ_j as eigenvalue.

(a) The eigenvalues for the matrix $A = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 2 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 2 & 3 \end{bmatrix}$ are -1 and 4. Find a basis for each eigenspace and determine each P_i. Then find k, $\mu_1, \ldots, \mu_k$, and each Q_j.

(b) Show in general (not just for the specific example in part (a), that the Q_j satisfy the same properties as the P_i. That is, verify the following.

i. $A = \mu_1Q_1 + \mu_2Q_2 + \cdots \mu_kQ_k$

ii. Q_j is a symmetric matrix for each j

iii. $Q_j^2 = Q_j$ for each j

iv. $Q_jQ_\ell = 0$ when $j \neq \ell$

v. if E_{μ_j} is the eigenspace for A corresponding to the eigenvalue μ_j, and if $\mathbf{v}$ is in $\mathbb{R}^n$, then $Q_j\mathbf{v} = \text{proj}_{E_{\mu_j}}\mathbf{v}$.

(c) What is the rank of Q_j? Verify your answer.

(8) Label each of the following statements as True or False. Provide justification for your response.

(a) **True/False** Every real symmetric matrix is diagonalizable.

(b) **True/False** If P is a matrix whose columns are eigenvectors of a symmetric matrix, then the columns of P are orthogonal.

(c) **True/False** If A is a symmetric matrix, then eigenvectors of A corresponding to distinct eigenvalues are orthogonal.

(d) **True/False** If $\mathbf{v}_1$ and $\mathbf{v}_2$ are distinct eigenvectors of a symmetric matrix A, then $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal.

(e) **True/False** Any symmetric matrix can be written as a sum of symmetric rank 1 matrices.

(f) **True/False** If A is a matrix satisfying $A^\mathsf{T} = A$, and $\mathbf{u}$ and $\mathbf{v}$ are vectors satisfying $A\mathbf{u} = 2\mathbf{u}$ and $A\mathbf{v} = -2\mathbf{v}$, then $\mathbf{u} \cdot \mathbf{v} = 0$.

(g) **True/False** If an $n \times n$ matrix A has n orthogonal eigenvectors, then A is a symmetric matrix.

(h) **True/False** If an $n \times n$ matrix has n real eigenvalues (counted with multiplicity), then A is a symmetric matrix.

(i) **True/False** For each eigenvalue of a symmetric matrix, the algebraic multiplicity equals the geometric multiplicity.

(j) **True/False** If A is invertible and orthogonally diagonalizable, then so is A^{-1}.

(k) **True/False** If A, B are orthogonally diagonalizable $n \times n$ matrices, then so is AB.

Project: The Second Derivative Test for Functions of Two Variables

In this project we will verify the Second Derivative Test for functions of two variables.[1] This test will involve Taylor polynomials and linear algebra. As a quick review, recall that the second order Taylor polynomial for a function f of a single variable x at $x = a$ is

$$P_2(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2. \tag{31.6}$$

As with the linearization of a function, the second order Taylor polynomial is a good approximation to f around a – that is $f(x) \approx P_2(x)$ for x close to a. If a is a critical number for f with $f'(a) = 0$,

[1] Many thanks to Professor Paul Fishback for sharing his activity on this topic. Much of this project comes from his activity.

then

$$P_2(x) = f(a) + \frac{f''(a)}{2}(x-a)^2.$$

In this situation, if $f''(a) < 0$, then $\frac{f''(a)}{2}(x-a)^2 \leq 0$ for x close to a, which makes $P_2(x) \leq f(a)$. This implies that $f(x) \approx P_2(x) \leq f(a)$ for x close to a, which makes $f(a)$ a relative maximum value for f. Similarly, if $f''(a) > 0$, then $f(a)$ is a relative minimum.

We now need a Taylor polynomial for a function of two variables. The complication of the additional independent variable in the two variable case means that the Taylor polynomials will need to contain all of the possible monomials of the indicated degrees. Recall that the linearization (or tangent plane) to a function $f = f(x, y)$ at a point (a, b) is given by

$$P_1(x, y) = f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b).$$

Note that $P_1(a, b) = f(a, b)$, $\frac{\partial P_1}{\partial x}(a, b) = f_x(a, b)$, and $\frac{\partial P_1}{\partial y}(a, b) = f_y(a, b)$. This makes $P_1(x, y)$ the best linear approximation to f near the point (a, b). The polynomial $P_1(x, y)$ is the first order Taylor polynomial for f at (a, b).

Similarly, the second order Taylor polynomial $P_2(x, y)$ centered at the point (a, b) for the function f is

$$\begin{aligned} P_2(x, y) = f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b) + \frac{f_{xx}(a, b)}{2}(x - a)^2 \\ + f_{xy}(a, b)(x - a)(y - b) + \frac{f_{yy}(a, b)}{2}(y - b)^2. \end{aligned}$$

Project Activity 31.1. To see that $P_2(x, y)$ is the best approximation for f near (a, b), we need to know that the first and second order partial derivatives of P_2 agree with the corresponding partial derivatives of f at the point (a, b). Verify that this is true.

We can rewrite this second order Taylor polynomial using matrices and vectors so that we can apply techniques from linear algebra to analyze it. Note that

$$\begin{aligned} P_2(x, y) = f(a, b) + \nabla f(a, b)^{\mathsf{T}} \begin{bmatrix} x - a \\ y - b \end{bmatrix} \\ + \frac{1}{2} \begin{bmatrix} x - a \\ y - b \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} f_{xx}(a, b) & f_{xy}(a, b) \\ f_{xy}(a, b) & f_{yy}(a, b) \end{bmatrix} \begin{bmatrix} x - a \\ y - b \end{bmatrix}, \end{aligned} \tag{31.7}$$

where $\nabla f(x, y) = \begin{bmatrix} f_x(x, y) \\ f_y(x, y) \end{bmatrix}$ is the gradient of f and H is the *Hessian* of f, where $H(x, y) = \begin{bmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{bmatrix}$.[2]

Project Activity 31.2. Use Equation (31.7) to compute $P_2(x, y)$ for $f(x, y) = x^4 + y^4 - 4xy + 1$ at $(a, b) = (2, 3)$.

[2]Note that under reasonable conditions (e.g., that f has continuous second order mixed partial derivatives in some open neighborhood containing (x, y)) we have that $f_{xy}(x, y) = f_{yx}(x, y)$ and $H(x, y) = \begin{bmatrix} f_{xx}(a, b) & f_{xy}(a, b) \\ f_{xy}(a, b) & f_{yy}(a, b) \end{bmatrix}$ is a symmetric matrix. We will only consider functions that satisfy these reasonable conditions.

The important idea for us is that if (a, b) is a point at which f_x and f_y are zero, then ∇f is the zero vector and Equation (31.7) reduces to

$$P_2(x, y) = f(a, b) + \frac{1}{2} \begin{bmatrix} x - a \\ y - b \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} f_{xx}(a, b) & f_{xy}(a, b) \\ f_{xy}(a, b) & f_{yy}(a, b) \end{bmatrix} \begin{bmatrix} x - a \\ y - b \end{bmatrix}, \tag{31.8}$$

To make the connection between the multivariable second derivative test and properties of the Hessian, $H(a, b)$, at a critical point of a function f at which $\nabla f = \mathbf{0}$, we will need to connect the eigenvalues of a matrix to the determinant and the trace.

Let A be an $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ (not necessarily distinct). Exercise 2 in Section 17 shows that

$$\det(A) = \lambda_1 \lambda_2 \cdots \lambda_n. \tag{31.9}$$

In other words, the determinant of a matrix is equal to the product of the eigenvalues of the matrix. In addition, Exercise 9 in Section 18 shows that

$$\text{trace}(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n. \tag{31.10}$$

for a diagonalizable matrix, where $\text{trace}(A)$ is the sum of the diagonal entries of A. Equation (31.10) is true for any square matrix, but we don't need the more general result for this project.

The fact that the Hessian is a symmetric matrix makes it orthogonally diagonalizable. We denote the eigenvalues of $H(a, b)$ as λ_1 and λ_2. Thus there exists an orthogonal matrix P and a diagonal matrix $D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ such that $P^{\mathsf{T}} H(a, b) P = D$, or $H(a, b) = PDP^{\mathsf{T}}$. Equations 31.9 and 31.10 show that

$$\lambda_1 \lambda_2 = f_{xx}(a, b) f_{yy}(a, b) - f_{xy}(a, b)^2 \text{ and } \lambda_1 + \lambda_2 = f_{xx}(a, b) + f_{yy}(a, b).$$

Now we have the machinery to verify the Second Derivative Test for Two-Variable Functions. We assume (a, b) is a point in the domain of a function f so that $\nabla f(a, b) = \mathbf{0}$. First we consider the case where $f_{xx}(a, b) f_{yy}(a, b) - f_{xy}(a, b)^2 < 0$.

Project Activity 31.3. Explain why if $f_{xx}(a, b) f_{yy}(a, b) - f_{xy}(a, b)^2 < 0$, then

$$\begin{bmatrix} x - a \\ y - b \end{bmatrix}^{\mathsf{T}} H(a, b) \begin{bmatrix} x - a \\ y - b \end{bmatrix}$$

is indefinite. Explain why this implies that f is "saddle-shaped" near (a, b). (Hint: Substitute $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = P^{\mathsf{T}} \begin{bmatrix} x - a \\ y - b \end{bmatrix}$. What does the graph of f look like in the w_1 and w_2 directions?)

Now we examine the situation when $f_{xx}(a, b) f_{yy}(a, b) - f_{xy}(a, b)^2 > 0$.

Project Activity 31.4. Assume that $f_{xx}(a, b) f_{yy}(a, b) - f_{xy}(a, b)^2 > 0$.

(a) Explain why either both $f_{xx}(a, b)$ and $f_{yy}(a, b)$ are positive or both are negative.

(b) If $f_{xx}(a, b) > 0$ and $f_{yy}(a, b) > 0$, explain why λ_1 and λ_2 must be positive.

(c) Explain why, if $f_{xx}(a,b) > 0$ and $f_{yy}(a,b) > 0$, then $f(a,b)$ is a local minimum value for f.

When $f_{xx}(a,b)f_{yy}(a,b) - f_{xy}(a,b)^2 > 0$ and either $f_{xx}(a,b)$ or $f_{yy}(a,b)$ is negative, a slight modification of the preceding argument leads to the fact that f has a local maximum at (a,b) (the details are left to the reader). Therefore, we have proved the Second Derivative Test for functions of two variables!

Project Activity 31.5. Use the Hessian to classify the local maxima, minima, and saddle points of $f(x,y) = x^4 + y^4 - 4xy + 1$. Draw a graph of f to illustrate.

Section 32

Quadratic Forms and the Principal Axis Theorem

Focus Questions

By the end of this section, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the section.

- What is a quadratic form on $\mathbb{R}^n$?
- What does the Principal Axis Theorem tell us about quadratic forms?

Application: The Tennis Racket Effect

Try an experiment with a tennis racket (or a squash racket, or a ping pong paddle). Let us define a 3D coordinate system with the center of the racket as the origin with the head of the racket lying in the xy-plane. We let $\mathbf{u}_1$ be the vector in the direction of the handle and $\mathbf{u}_2$ the perpendicular direction (still lying in the plane defined by the head) as illustrated in Figure 32.1. We then let $\mathbf{u}_3$ be a vector perpendicular to the plane of the head. Hold the racket by the handle and spin it to make one rotation around the $\mathbf{u}_1$ axis. This is pretty easy. It is also not difficult to throw the racket so that it rotates around the $\mathbf{u}_3$. Now toss the racket into the air to make one complete rotation around the axis of the vector $\mathbf{u}_2$ and catch the handle. Repeat this several times. You should notice that in most instances, the racket will also have made a half rotation around the $\mathbf{u}_1$ axis so that the other face of the racket now points up. This is quite different than the rotations around the $\mathbf{u}_1$ and $\mathbf{u}_3$ axes. A good video that illustrates this behavior can be seen at https://www.youtube.com/watch?v=4dqCQqI-Gis.

This effect is a result in classical mechanics that describes the rotational movement of a rigid body in space, called the *tennis racket effect* (or the Dzhanibekov effect, after the Russian cosmonaut Vladimir Dzhanibekov who discovered the theorem's consequences while in zero gravity in space – you can see an illustration of this in the video at https://www.youtube.com/watch?v=L2o9eBl_Gzw). The result is simple to see in practice, but is difficult to intuitively

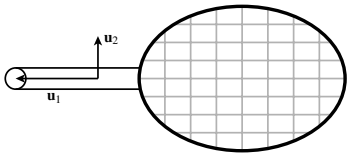

Figure 32.1: Two principal axes of a tennis racket.

understand why the behavior is different around the intermediate axis. There is a story of a student who asked the famous physicist Richard Feynman if there is any intuitive way to understand the result; Feynman supposedly went into deep thought for about 10 or 15 seconds and answered, "no." As we will see later in this section, we can understand this effect using the principal axes of a rigid body.

Introduction

We are familiar with quadratic equations in algebra. Examples of quadratic equations include $x^2 = 1$, $x^2 + y^2 = 1$, and $x^2 + xy + y^2 = 3$. We don't, however, have to restrict ourselves to two variables. A quadratic equation in n variables is any equation in which the sum of the exponents in any monomial term is 2. So a quadratic equation in the variables x_1, x_2, ..., x_n is an equation in the form

$$\begin{aligned} a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + \cdots + a_{nn}x_n^2 \\ + a_{12}x_1x_2 + a_{13}x_1x_3 + \cdots a_{1n}x_1x_n \\ + a_{23}x_2x_3 + a_{24}x_2x_4 + \cdots a_{2n}x_2x_n + \cdots \\ + a_{n-1n}x_{n-1}x_n \\ = c \end{aligned}$$

for some constant c. In matrix notation this expression on the left of this equation has the form

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j = \mathbf{x}^{\mathsf{T}}A\mathbf{x}$$

where $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ and A is the $n \times n$ matrix $A = [a_{ij}]$. For example, if $A = \begin{bmatrix} 1 & 3 & -2 \\ -1 & 1 & 2 \\ 0 & 2 & -2 \end{bmatrix}$, then we get the quadratic expression $x_1^2 + 3x_1x_2 - 2x_1x_3 - x_2x_1 + x_2^2 + 2x_2x_3 + 2x_3x_2 - 2x_3^2$. We should note here that the terms involving x_ix_j and x_jx_i are repeated in our sum, but

$$a_{ij}x_ix_j + a_{ji}x_jx_i = 2\left(\frac{a_{ij} + a_{ji}}{2}\right)x_ix_j$$

and so we could replace a_{ij} and a_{ji} both with $\left(\frac{a_{ij}+a_{ji}}{2}\right)$ without changing the quadratic form. With this alteration in mind, we can then assume that A is a symmetric matrix. So in the previous

example, the symmetric matrix $A' = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 2 \\ -1 & 2 & -2 \end{bmatrix}$ gives the same quadratic expression. This leads to the following definition.

Definition 32.1. A **quadratic form** on $\mathbb{R}^n$ is a function Q defined by

$$Q(\mathbf{x}) = \mathbf{x}^\mathsf{T} A\mathbf{x}$$

for some $n \times n$ symmetric matrix A.

As we show in Exercise 7., the symmetric matrix A is unique to the quadratic form, so we call the symmetric matrix A is the *matrix of the quadratic form*. It is these quadratic forms that we will study in this section.

Preview Activity 32.1.

(1) To get a little more comfortable with quadratic forms, write the quadratic forms in matrix form, explicitly identifying the vector $\mathbf{x}$ and the symmetric matrix A of the quadratic form.

(a) $3x_1^2 - 2x_2^2 + 4x_1x_2 + x_2x_3$

(b) $x_1x_4 + 4x_2x_3 - x_2^2 + 10x_1x_5$

(2) Some quadratic forms form equations in $\mathbb{R}^2$ that are very familiar: $x^2+y^2 = 1$ is an equation of a circle, $2x^2 + 3y^2 = 2$ is an equation of an ellipse, and $x^2 - y^2 = 1$ is an equation of a hyperbola. Of course, these do not represent all of the quadratic forms in $\mathbb{R}^2$ – some contain cross-product terms. We can recognize the equations above because they contain no cross-product terms (terms involving xy). We can more easily recognize the quadratic forms that contain cross-product terms if we can somehow rewrite the forms in a different format with no cross-product terms. We illustrate how this can be done with the quadratic form Q defined by $Q(\mathbf{x}) = x^2 - xy + y^2$.

(a) Write $Q(\mathbf{x})$ in the form $\mathbf{x}^\mathsf{T} A\mathbf{x}$, where A is a 2×2 symmetric matrix.

(b) Since A is a symmetric matrix we can orthogonally diagonalize A. Given that the eigenvalues of A are $\frac{3}{2}$ and $\frac{1}{2}$ with corresponding eigenvectors $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, respectively, find a matrix P that orthogonally diagonalizes A.

(c) Define $\mathbf{y} = \begin{bmatrix} w \\ z \end{bmatrix}$ to satisfy $\mathbf{x} = P\mathbf{y}$. Substitute for $\mathbf{x}$ in the quadratic form $Q(\mathbf{x})$ to write the quadratic form in terms of w and z. What kind of graph does the quadratic equation $Q(\mathbf{x}) = 1$ have?

Equations Involving Quadratic Forms in $\mathbb{R}^2$

When we consider equations of the form $Q(\mathbf{x}) = d$, where Q is a quadratic form in $\mathbb{R}^2$ and d is a constant, we wind up with old friends like $x^2+y^2 = 1$, $2x^2+3y^2 = 2$, or $x^2-y^2 = 1$. As we saw in Preview Activity 32.1 these equations are relatively easy to recognize. However, when we have

cross-product terms, like in $x^2 - xy + y^2 = 1$, it is not so easy to identify the curve the equation represents. If there was a way to eliminate the cross-product term xy from this form, we might be more easily able to recognize its graph. The discussion in this section will focus on quadratic forms in $\mathbb{R}^2$, but we will see later that the arguments work in any number of dimensions. While working in $\mathbb{R}^2$ we will use the standard variables x and y instead of x_1 and x_2.

In general, the equation of the form $Q(\mathbf{x}) = d$, where Q is a quadratic form in $\mathbb{R}^2$ defined by a matrix $A = \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix}$ and d is a constant looks like

$$ax^2 + bxy + cy^2 = d.$$

The graph of an equation like this is either an ellipse (a circle is a special case of an ellipse), a hyperbola, two non-intersecting lines, a point, or the empty set (see Exercise 5.). The quadratic forms do not involve linear terms, so we don't consider the cases of parabolas. One way to see into which category one of these quadratic form equations falls is to write the equation in standard form.

The standard forms for quadratic equations in $\mathbb{R}^2$ are as follows, where a and b are nonzero constants and h and k are any constants.

Lines: $ax^2 = 1$ or $ay^2 = 1$ $(a > 0)$

Ellipse: $\dfrac{(x-h)^2}{a^2} + \dfrac{(y-k)^2}{b^2} = 1$

Hyperbola: $\dfrac{(x-h)^2}{a^2} - \dfrac{(y-k)^2}{b^2} = 1$ or $\dfrac{(y-k)^2}{b^2} - \dfrac{(x-h)^2}{a^2} = 1$

Preview Activity 32.1 contains the main tool that we need to convert a quadratic form into one of these standard forms. By this we mean that if we have a quadratic form Q in the variables x_1, x_2, …, x_n, we want to find variables y_1, y_2, …, y_n in terms of x_1, x_2, …, x_n so that when written in terms of the variables y_1, y_2, …, y_n the quadratic form Q contains no cross terms. In other words, we want to find a vector $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ so that $Q(\mathbf{x}) = \mathbf{y}^\mathsf{T} D\mathbf{y}$, where D is a diagonal matrix. Since every real symmetric matrix is orthogonally diagonalizable, we will always be able to find a matrix P that orthogonally diagonalizes A. The details are as follows.

Let Q be the quadratic form defined by $Q(\mathbf{x}) = \mathbf{x}^\mathsf{T} A\mathbf{x}$, where A is an $n \times n$ symmetric matrix. As in Preview Activity 32.1, the fact that A is symmetric means that we can find an orthogonal matrix $P = [\mathbf{p}_1\ \mathbf{p}_2\ \mathbf{p}_3\ \cdots\ \mathbf{p}_n]$ whose columns are orthonormal eigenvectors of A corresponding to eigenvalues λ_1, λ_2, …, λ_n, respectively. Letting $\mathbf{y} = P^\mathsf{T}\mathbf{x}$ give us $\mathbf{x} = P\mathbf{y}$ and

$$\mathbf{x}^\mathsf{T} A\mathbf{x} = (P\mathbf{y})^\mathsf{T} A(P\mathbf{y}) = \mathbf{y}^\mathsf{T}(P^\mathsf{T} AP)\mathbf{y} = \mathbf{y}^\mathsf{T} D\mathbf{y},$$

where D is the diagonal matrix whose ith diagonal entry is λ_i.

Moreover, the set $\mathcal{B} = \{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n\}$ is an orthonormal basis for $\mathbb{R}^n$ and so defines a coordinate system for $\mathbb{R}^n$. Note that if $y = [y_1\ y_2\ \cdots\ y_n]^\mathsf{T}$, then

$$\mathbf{x} = P\mathbf{y} = y_1\mathbf{p}_1 + y_2\mathbf{p}_2 + \cdots + y_n\mathbf{p}_n.$$

Thus, the coordinate vector of $\mathbf{x}$ with respect to $\mathcal{B}$ is $\mathbf{y}$, or $[\mathbf{x}]_{\mathcal{B}} = \mathbf{y}$. We summarize in Theorem 32.2.

Theorem 32.2 (Principal Axis Theorem)**.** *Let A be an $n \times n$ symmetric matrix. There is an orthogonal change of variables $\mathbf{x} = P\mathbf{y}$ so that the quadratic form Q defined by $Q(\mathbf{x}) = \mathbf{x}^{\mathsf{T}} A\mathbf{x}$ is transformed into the quadratic form $\mathbf{y}^{\mathsf{T}} D\mathbf{y}$ where D is a diagonal matrix.*

The columns of the orthogonal matrix P in the Principal Axis Theorem form an orthogonal basis for $\mathbb{R}^n$ and are called the *principal axes* for the quadratic form Q. Also, the coordinate vector of $\mathbf{x}$ with respect to this basis is $\mathbf{y}$.

Activity 32.1. Let Q be the quadratic form defined by $Q(\mathbf{x}) = 2x^2 + 4xy + 5y^2 = \mathbf{x}^{\mathsf{T}} A\mathbf{x}$, where $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ and $A = \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix}$.

(a) The eigenvalues of A are $\lambda_1 = 6$ and $\lambda_2 = 1$ with corresponding eigenvectors $\mathbf{v}_1 = [1\ 2]^{\mathsf{T}}$ and $\mathbf{v}_2 = [-2\ 1]^{\mathsf{T}}$, respectively. Find an orthogonal matrix P with determinant 1 that diagonalizes A. Is P unique? Explain. Is there a matrix without determinant 1 that orthogonally diagonalizes A? Explain.

(b) Use the matrix P to write the quadratic form without the cross-product.

(c) We can view P as a change of basis matrix from the coordinate system defined by $\mathbf{y} = P^{\mathsf{T}}\mathbf{x}$ to the standard coordinate system. In other words, in the standard xy coordinate system, the quadratic form is written as $\mathbf{x}^{\mathsf{T}} A\mathbf{x}$, but in the new coordinate system defined by $\mathbf{y}$ the quadratic form is written as $(P\mathbf{y})^{\mathsf{T}} A(P\mathbf{y})$. As a change of basis matrix, P performs a rotation. See if you can recall what we learned about rotation matrices and determine the angle of rotation P defines. Plot the graph of the quadratic equation $Q(\mathbf{x}) = 1$ in the new coordinate system and identify this angle on the graph. Interpret the result.

Classifying Quadratic Forms

If we draw graphs of equations of the type $z = Q(\mathbf{x})$, where Q is a quadratic form, we can see that a quadratic form whose matrix does not have 0 as an eigenvalue can take on all positive values (except at $\mathbf{x} = \mathbf{0}$) as shown at left in Figure 32.2, all negative values (except at $\mathbf{x} = \mathbf{0}$) as shown in the center of Figure 32.2, or both positive and negative values as depicted at right in Figure 32.2. We can see when these cases happen by analyzing the eigenvalues of the matrix that defines the quadratic form. Let A be a 2×2 symmetric matrix with eigenvalues λ_1 and λ_2, and let P be a matrix that orthogonally diagonalizes A so that $P^{\mathsf{T}} AP = D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$. If we let $\mathbf{y} = \begin{bmatrix} w \\ z \end{bmatrix} = P^{\mathsf{T}}\mathbf{x}$, then

$$\begin{aligned} Q(\mathbf{x}) &= \mathbf{x}^{\mathsf{T}} A\mathbf{x} \\ &= \mathbf{y}^{\mathsf{T}} D\mathbf{y} \\ &= \lambda_1 w^2 + \lambda_2 z^2. \end{aligned}$$

Then $Q(\mathbf{x}) \geq 0$ if all of the eigenvalues of A are positive (with $Q(\mathbf{x}) > 0$ when $\mathbf{x} \neq \mathbf{0}$) and $Q(\mathbf{x}) \leq 0$ if all of the eigenvalues of A are negative (with $Q(\mathbf{x}) < 0$ when $\mathbf{x} \neq \mathbf{0}$). If one eigenvalue

of A is positive and the other negative, then $Q(\mathbf{x})$ will take on both positive and negative values. As a result, we classify symmetric matrices (and their corresponding quadratic forms) quadratic forms according to these behaviors.

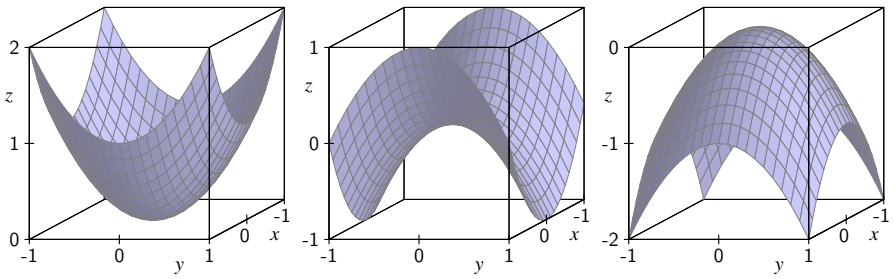

Figure 32.2: Left: Paraboloid $Q(\mathbf{x}) = x^2 + y^2$. Center: Hyperbolic Paraboloid $Q(\mathbf{x}) = x^2 - y^2$. Right: Paraboloid $Q(\mathbf{x}) = -x^2 - y^2$.

Definition 32.3. A symmetric matrix A (and its associated quadratic form Q) is

(a) **positive definite** if $\mathbf{x}^\mathsf{T} A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$,

(b) **positive semidefinite** if $\mathbf{x}^\mathsf{T} A\mathbf{x} \geq 0$ for all $\mathbf{x}$,

(c) **negative definite** if $\mathbf{x}^\mathsf{T} A\mathbf{x} < 0$ for all $\mathbf{x} \neq \mathbf{0}$,

(d) **negative semidefinite** if $\mathbf{x}^\mathsf{T} A\mathbf{x} \leq 0$ for all $\mathbf{x}$,

(e) **indefinite** if $\mathbf{x}^\mathsf{T} A\mathbf{x}$ takes on both positive and negative values.

For example, the quadratic form $Q(\mathbf{x}) = x^2 + y^2$ at left in Figure 32.2 is positive definite (with repeated eigenvalue 1), the quadratic form $Q(\mathbf{x}) = -(x^2 + y^2)$ in the center of Figure 32.2 is negative definite (repeated eigenvalue -1), and the hyperbolic paraboloid $Q(\mathbf{x}) = x^2 - y^2$ at right in Figure 32.2 is indefinite (eigenvalues 1 and -1).

So we have argued that a quadratic form $Q(\mathbf{x}) = \mathbf{x}^\mathsf{T} A\mathbf{x}$ is positive definite if A has all positive eigenvalues, negative definite if A has all negative eigenvalues, and indefinite if A has both positive and negative eigenvalues. Similarly, the quadratic form is positive semidefinite if A has all nonnegative eigenvalues and negative semidefinite if A has all nonpositive eigenvalues. Positive definite matrices are important, as the following activity illustrates.

Activity 32.2. Let A be a symmetric $n \times n$ matrix, and define $\langle\ ,\ \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\mathsf{T} A\mathbf{v}. \tag{32.1}$$

(a) Explain why it is necessary for A to be positive definite in order for (32.1) to define an inner product on $\mathbb{R}^n$.

(b) Show that (32.1) defines an inner product on $\mathbb{R}^n$ if A is positive definite.

(c) Let $\langle\ ,\ \rangle$ be the mapping from $\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right\rangle = 2x_1y_1 - x_1y_2 - x_2y_1 + x_2y_2.$$

Find a matrix A so that $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\mathsf{T} A\mathbf{y}$ and explain why $\langle\ ,\ \rangle$ defines an inner product.

Examples

What follows are worked examples that use the concepts from this section.

Example 32.4. Write the given quadratic equation in a system in which it has no cross-product terms.

(a) $8x^2 - 4xy + 5y^2 = 1$

(b) $x^2 + 4xy + y^2 = 1$

(c) $4x^2 + 4y^2 + 4z^2 + 4xy + 4xz + 4yz - 3 = 0$

Example Solution.

(a) We write the quadratic form $Q(x, y) = 8x^2 - 4xy + 5y^2$ as $\mathbf{x}^\mathsf{T} A\mathbf{x}$, where $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ and $A = \begin{bmatrix} 8 & -2 \\ -2 & 5 \end{bmatrix}$. The eigenvalues for A are 9 and 4, and bases for the corresponding eigenspaces E_9 and E_4 are $\{[-2\ 1]^\mathsf{T}\}$ and $\{[1\ 2]^\mathsf{T}\}$, respectively. An orthogonal matrix P that orthogonally diagonalizes A is

$$P = \begin{bmatrix} -\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix}.$$

If $\mathbf{y} = [u\ v]^\mathsf{T}$ and we let $\mathbf{x} = P\mathbf{y}$, then we can rewrite the quadratic equation $8x^2 - 4xy + 5y^2 = 1$ as

$$\begin{aligned} 8x^2 - 4xy + 5y^2 &= 1 \\ \mathbf{x}^\mathsf{T} A\mathbf{x} &= 1 \\ (P\mathbf{y})^\mathsf{T} A(P\mathbf{y}) &= 1 \\ \mathbf{y}^\mathsf{T} P^\mathsf{T} AP\mathbf{y} &= 1 \\ \mathbf{y}^\mathsf{T} \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix} \mathbf{y} &= 1 \\ 9u^2 + 4v^2 &= 1. \end{aligned}$$

So the quadratic equation $8x^2 - 4xy + 5y^2 = 1$ is an ellipse.

(b) We write the quadratic form $Q(x, y) = x^2 + 4xy + y^2$ as $\mathbf{x}^\mathsf{T} A\mathbf{x}$, where $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ and $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$. The eigenvalues for A are 3 and -1, and bases for the corresponding eigenspaces E_3 and E_{-1} are $\{[1\ 1]^\mathsf{T}\}$ and $\{[-1\ 1]^\mathsf{T}\}$, respectively. An orthogonal matrix P that orthogonally diagonalizes A is

$$P = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

If $\mathbf{y} = [u\ v]^\mathsf{T}$ and we let $\mathbf{x} = P\mathbf{y}$, then we can rewrite the quadratic equation $x^2 + 4xy + y^2 = 1$ as

$$\begin{aligned} x^2 + 4xy + y^2 &= 1 \\ \mathbf{x}^\mathsf{T} A\mathbf{x} &= 1 \\ (P\mathbf{y})^\mathsf{T} A(P\mathbf{y}) &= 1 \\ \mathbf{y}^\mathsf{T} P^\mathsf{T} AP\mathbf{y} &= 1 \\ \mathbf{y}^\mathsf{T} \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{y} &= 1 \\ 3u^2 - v^2 &= 1. \end{aligned}$$

So the quadratic equation $x^2 + 4xy + y^2 = 1$ is a hyperbola.

(c) We write the quadratic form $Q(x, y, z) = 4x^2 + 4y^2 + 4z^2 + 4xy + 4xz + 4yz$ as $\mathbf{x}^\mathsf{T} A\mathbf{x}$, where $\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ and $A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}$. The eigenvalues for A are 2 and 8, and bases for the corresponding eigenspaces E_2 and E_8 are $\{[-1\ 0\ 1]^\mathsf{T}, [-1\ 1\ 0]^\mathsf{T}\}$ and $\{[1\ 1\ 1]^\mathsf{T}\}$, respectively. Applying the Gram-Schmidt process to the basis for E_2 gives us an orthogonal basis $\{\mathbf{w}_1, \mathbf{w}_2\}$ of E_2, where $\mathbf{w}_1 = [-1\ 0\ 1]^\mathsf{T}$ and

$$\begin{aligned} \mathbf{w}_2 &= [-1\ 1\ ]^\mathsf{T} - \frac{[-1\ 0\ 1]^\mathsf{T} \cdot [-1\ 1\ 0]^\mathsf{T}}{[-1\ 0\ 1]^\mathsf{T} \cdot [-1\ 0\ 1]^\mathsf{T}} [-1\ 0\ 1]^\mathsf{T} \\ &= [-1\ 1\ 0]^\mathsf{T} - \frac{1}{2}[-1\ 0\ 1]^\mathsf{T} \\ &= \frac{1}{2}[-1\ 2\ -1]^\mathsf{T}. \end{aligned}$$

An orthogonal matrix P that orthogonally diagonalizes A is

$$P = \begin{bmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{bmatrix}.$$

If $\mathbf{y} = [u\ v\ w]^\mathsf{T}$ and we let $\mathbf{x} = P\mathbf{y}$, then we can rewrite the quadratic equation $4x^2 +$

$4y^2 + 4z^2 + 4xy + 4xz + 4yz = 3$ as

$$\begin{aligned} 4x^2 + 4y^2 + 4z^2 + 4xy + 4xz + 4yz &= 3 \\ \mathbf{x}^\mathsf{T} A\mathbf{x} &= 3 \\ (P\mathbf{y})^\mathsf{T} A(P\mathbf{y}) &= 3 \\ \mathbf{y}^\mathsf{T} P^\mathsf{T} AP\mathbf{y} &= 3 \\ \mathbf{y}^\mathsf{T} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 8 \end{bmatrix} \mathbf{y} &= 3 \\ 2u^2 + 2v^2 + 8w^2 &= 3. \end{aligned}$$

So the quadratic equation $4x^2 + 4y^2 + 4z^2 + 4xy + 4xz + 4yz - 3 = 0$ is a ellipsoid.

Example 32.5. Let A and B be positive definite matrices, and let $C = \begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix}$.

(a) Must A be invertible? Justify your answer.

(b) Must A^{-1} be positive definite? Justify your answer.

(c) Must A^2 be positive definite? Justify your answer.

(d) Must $A + B$ be positive definite? Justify your answer.

(e) Is C positive definite? Justify your answer.

Example Solution.

(a) Since A has all positive eigenvalues and $\det(A)$ is the product of the eigenvalues of A, then $\det(A) > 0$. Thus, A is invertible.

(b) The fact that A is positive definite means that A is also symmetric. Recall that $\left(A^{-1}\right)^\mathsf{T} = \left(A^\mathsf{T}\right)^{-1}$. Since A is symmetric, it follows that $\left(A^{-1}\right)^\mathsf{T} = A^{-1}$ and A^{-1} is symmetric. The eigenvalues of A^{-1} are the reciprocals of the eigenvalues of A. Since the eigenvalues of A are all positive, so are the eigenvalues of A^{-1}. Thus, A^{-1} is positive definite.

(c) Notice that

$$\left(A^2\right)^\mathsf{T} = (AA)^\mathsf{T} = A^\mathsf{T} A^\mathsf{T} = A^2,$$

so A^2 is symmetric. The eigenvalues of A^2 are the squares of the eigenvalues of A. Since no eigenvalue of A is 0, the eigenvalues of A^2 are all positive and A^2 is positive definite.

(d) We know that B is symmetric, and

$$(A + B)^\mathsf{T} = A^\mathsf{T} + B^\mathsf{T} = A + B,$$

so $A + B$ is symmetric. Also, the fact that $\mathbf{x}^\mathsf{T} A\mathbf{x} > 0$ and $\mathbf{x}^\mathsf{T} B\mathbf{x} > 0$ for all $\mathbf{x}$ implies that

$$\mathbf{x}^\mathsf{T}(A + B)\mathbf{x} = \mathbf{x}^\mathsf{T} A\mathbf{x} + \mathbf{x}^\mathsf{T} B\mathbf{x} > 0$$

for all $\mathbf{x}$. Thus, $A + B$ is positive definite.

(e) The matrix C is symmetric and

$$\det(C - \lambda I_2) = (5 - \lambda)(3 - \lambda) - 9 = \lambda^2 - 8\lambda + 6.$$

So the eigenvalues of C are $4 + \sqrt{10}$ and $4 - \sqrt{10} \approx 0.8$. Since the eigenvalues of C are both positive, C is positive definite.

Summary

- A quadratic form on $\mathbb{R}^n$ is a function Q defined by

$$Q(\mathbf{x}) = \mathbf{x}^\mathsf{T} A\mathbf{x}$$

for some $n \times n$ symmetric matrix A.

- The Principal Axis Theorem tells us that there is a change of variable $\mathbf{x} = P\mathbf{y}$ that will remove the cross-product terms from a quadratic form and allow us to identify the form and determine the principal axes for the form.

Exercises

(1) Find the matrix for each quadratic form.

(a) $x_1^2 - 2x_1x_2 + 4x_2^2$ if $\mathbf{x}$ is in $\mathbb{R}^2$

(b) $10x_1^2 + 4x_1x_3 + 2x_2x_3 + x_3^2$ if $\mathbf{x}$ is in $\mathbb{R}^3$

(c) $2x_1x_2 + 2x_1x_3 - x_1x_4 + 5x_2^2 + 4x_3x_4 + 8x_4^2$ if $\mathbf{x}$ is in $\mathbb{R}^4$

(2) For each quadratic form, identify the matrix A of the form, find a matrix P that orthogonally diagonalizes A, and make a change of variable that transforms the quadratic form into one with no cross-product terms.

(a) $x_1^2 + 2x_1x_2 + x_2^2$

(b) $-2x_1^2 + 2x_1x_2 + 4x_1x_3 - 2x_2^2 - 4x_2x_3 - x_3^2$

(c) $11x_1^2 - 12x_1x_2 - 12x_1x_3 - 12x_1x_4 - x_2^2 - 2x_3x_4$

(3) One topic in multivariable calculus is constrained optimization. We can use the techniques of this section to solve certain types of constrained optimization problems involving quadratic forms. As an example, we will find the maximum and minimum values of the quadratic form defined by the matrix $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ on the unit circle.

(a) First we determine some bounds on the values of a quadratic form. Let Q be the quadratic form defined by the $n \times n$ real symmetric matrix A. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of A, and let P be a matrix that orthogonally diagonalizes A, with $P^\mathsf{T}AP = D$ as the matrix with diagonal entries λ_1, λ_2, …, λ_n in order. Let $\mathbf{y} = [y_1\ y_2\ \cdots\ y_n]^\mathsf{T} = P^\mathsf{T}[x_1\ x_2\ \cdots\ x_n]^\mathsf{T}$.

i. Show that
$$Q(\mathbf{x}) = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2.$$
ii. Use the fact that $\lambda_1 \geq \lambda_i$ for each i and the fact that P (and P^{T}) is an orthogonal matrix to show that
$$Q(\mathbf{x}) \leq \lambda_1 ||\mathbf{x}||.$$
iii. Now show that $Q(\mathbf{x}) \geq \lambda_n ||\mathbf{x}||$.

(b) Use the result of part (a) to find the maximum and minimum values of the quadratic form defined by the matrix $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ on the unit circle.

(4) In this exercise we characterize the symmetric, positive definite, 2×2 matrices with real entries in terms of the entries of the matrices. Let $A = \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix}$ for some real numbers a, b, and c.

(a) Assume that A is positive definite.
i. Show that a must be positive.
ii. Use the fact that the eigenvalues of A must be positive to show that $ac > b^2$. We conclude that if A is positive definite, then $a > 0$ and $ac > b^2$.

(b) Now show that if $a > 0$ and $ac > b^2$, then A is positive definite. This will complete our classification of positive definite 2×2 matrices.

(5) In this exercise we determine the form of
$$\mathbf{x}^{\mathsf{T}} A \mathbf{x} = 1, \tag{32.2}$$
where A is a symmetric 2×2 matrix. Let P be a matrix that orthogonally diagonalizes A and let $\mathbf{y} = P^{\mathsf{T}}\mathbf{x}$.

(a) Substitute $\mathbf{y}$ for $P^{\mathsf{T}}\mathbf{x}$ in the equation $\mathbf{x}^{\mathsf{T}} A \mathbf{x} = 1$. What form does the resulting equation have (write this form in terms of the eigenvalues of A)?

(b) What kind of graph does the equation (32.2) have if A is positive definite? Why?

(c) What kind of graph does the equation (32.2) have if A has both positive and negative eigenvalues? Why?

(d) What kind of graph does the equation (32.2) have if one eigenvalue of A is zero and the other non-zero? Why?

(6) Let $A = [a_{ij}]$ be a symmetric $n \times n$ matrix.

(a) Show that $\mathbf{e}_i^{\mathsf{T}} A \mathbf{e}_j = a_{ij}$, where $\mathbf{e}_i$ is the ith standard unit vector for $\mathbb{R}^n$. (This result will be useful in Exercise 7..)

(b) Let $\mathbf{u}$ be a unit eigenvector of A with eigenvalue λ. Find $\mathbf{u}^{\mathsf{T}} A \mathbf{u}$ in terms of λ.

(7) Suppose A and B are symmetric $n \times n$ matrices, and let $Q_A(\mathbf{x}) = \mathbf{x}^{\mathsf{T}} A \mathbf{x}$ and $Q_B(\mathbf{x}) = \mathbf{x}^{\mathsf{T}} B \mathbf{x}$. If $Q_A(\mathbf{x}) = Q_B(\mathbf{x})$ for all $\mathbf{x}$ in $\mathbb{R}^n$, show that $A = B$. (Hint: Use Exercise 6 (a) to compare $Q_A(\mathbf{e}_i)$ and $Q_B(\mathbf{e}_i)$, then compare $Q_A(\mathbf{e}_i + \mathbf{e}_j)$ to $Q_B(\mathbf{e}_i + \mathbf{e}_j)$ for $i \neq j$.) Thus, quadratic forms are uniquely determined by their symmetric matrices.

(8) In this exercise we analyze all inner products on $\mathbb{R}^n$. Let $\langle\ ,\ \rangle$ be an inner product on $\mathbb{R}^n$. Let $\mathbf{x} = [x_1\ x_2\ \ldots\ x_n]^\mathsf{T}$ and $\mathbf{y} = [y_1\ y_2\ \ldots\ y_n]^\mathsf{T}$ be arbitrary vectors in $\mathbb{R}^n$. Then

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i \quad \text{and} \quad \mathbf{y} = \sum_{j=1}^{n} y_j \mathbf{e}_j,$$

where $\mathbf{e}_i$ is the ith standard vector in $\mathbb{R}^n$.

(a) Explain why

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \langle \mathbf{e}_i, \mathbf{e}_j \rangle y_j. \tag{32.3}$$

(b) Calculate the matrix product

$$\mathbf{x}^\mathsf{T} \begin{bmatrix} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle & \langle \mathbf{e}_1, \mathbf{e}_2 \rangle & \cdots & \langle \mathbf{e}_1, \mathbf{e}_n \rangle \\ \langle \mathbf{e}_2, \mathbf{e}_1 \rangle & \langle \mathbf{e}_2, \mathbf{e}_2 \rangle & \cdots & \langle \mathbf{e}_2, \mathbf{e}_n \rangle \\ \vdots & \vdots & & \vdots \\ \langle \mathbf{e}_n, \mathbf{e}_1 \rangle & \langle \mathbf{e}_n, \mathbf{e}_2 \rangle & \cdots & \langle \mathbf{e}_n, \mathbf{e}_n \rangle \end{bmatrix} \mathbf{y}$$

and compare to (32.3). What do you notice?

(c) Explain why any inner product on $\mathbb{R}^n$ is of the form $\mathbf{x}^\mathsf{T} A \mathbf{y}$ for some symmetric, positive definite matrix A.

(9) Label each of the following statements as True or False. Provide justification for your response.

(a) **True/False** If Q is a quadratic form, then there is exactly one matrix A such that $Q(\mathbf{x}) = \mathbf{x}^\mathsf{T} A \mathbf{x}$.

(b) **True/False** The matrix of a quadratic form is unique.

(c) **True/False** If the matrix of a quadratic form is a diagonal matrix, then the quadratic form has no cross-product terms.

(d) **True/False** The eigenvectors of the symmetric matrix A form the principal axes of the quadratic form $\mathbf{x}^\mathsf{T} A \mathbf{x}$.

(e) **True/False** The principal axes of a quadratic form are orthogonal.

(f) **True/False** If a and c are positive, then the quadratic equation $ax^2 + bxy + cy^2 = 1$ defines an ellipse.

(g) **True/False** If the entries of a symmetric matrix A are all positive, then the quadratic form $\mathbf{x}^\mathsf{T} A \mathbf{x}$ is positive definite.

(h) **True/False** If a quadratic form $\mathbf{x}^\mathsf{T} A \mathbf{x}$ defined by a symmetric matrix A is positive definite, then the entries of A are all non-negative.

(i) **True/False** If a quadratic form $Q(\mathbf{x})$ on $\mathbb{R}^2$ is positive definite, then the graph of $z = Q(\mathbf{x})$ is a paraboloid opening upward.

(j) **True/False** If a quadratic form $Q(\mathbf{x})$ on $\mathbb{R}^2$ is negative definite, then the graph of $z = Q(\mathbf{x})$ is a paraboloid opening downward.

(k) **True/False** If a quadratic form $Q(\mathbf{x})$ on $\mathbb{R}^2$ is indefinite, then there is a nonzero vector $\mathbf{x}$ such that $Q(\mathbf{x}) = 0$.

(l) **True/False** If $Q(\mathbf{x})$ is positive definite, then so is the quadratic form $aQ(\mathbf{x})$ for $a > 0$.

(m) **True/False** If If $Q(\mathbf{x}) = \mathbf{x}^\mathsf{T} A\mathbf{x}$ is indefinite, then at least one of the eigenvalues of A is negative and at least one positive.

(n) **True/False** If $n \times n$ symmetric matrices A and B define positive definite quadratic forms, then so does $A + B$.

(o) **True/False** If an invertible symmetric matrix A defines a positive definite quadratic form, then so does A^{-1}.

Project: The Tennis Racket Theorem

If a particle of mass m and velocity v is moving in a straight line, its kinetic energy KE is given by $KE = \frac{1}{2}mv^2$. If, instead, the particle rotates around an axis with angular velocity ω (in radians per unit of time), its linear velocity is $v = r\omega$, where r is the radius of the particle's circular path. Substituting into the kinetic energy formula shows that the kinetic energy of the rotating particle is then $KE = \frac{1}{2}\left(mr^2\right)\omega^2$. The quantity mr^2 is called the *moment of inertia* of the particle and is denoted by I. So $KE = \frac{1}{2}I\omega^2$ for a rotating particle. Notice that the larger the value of r, the larger the inertia. You can imagine this with a figure skater. When a skater spins along their major axis with their arms outstretched, the speed at which they rotate is lower than when they bring their arms into their bodies. The moment of inertia for rotational motion plays a role similar to the mass in linear motion. Essentially, the inertia tells us how resistant the particle is to rotation.

To understand the tennis racket effect, we are interested in rigid bodies as they move through space. Any rigid body in three space has three principal axes about which it likes to spin. These axes are at right angles to each other and pass through the center of mass. Think of enclosing the object in an ellipsoid – the longest axis is the *primary* axis, the middle axis is the *intermediate* axis, and the third axis is the third axis. As a rigid body moves through space, it rotates around these axes and there is inertia along each axis. Just like with a tennis racket, if you were to imagine an axle along any of the principal axes and spin the object along that axel, it will either rotate happily with no odd behavior like flipping, or it won't. The former behavior is that of a stable axis and the latter an unstable axis. The Tennis Racket Theorem is a statement about the rotation of the body. Essentially, the Tennis Racket Theorem states that the rotation of a rigid object around its primary and third principal axes is stable, while rotation around its intermediate axis is not. To understand why this is so, we need to return to moments of inertia.

Assume that we have a rigid body moving through space. Euler's (rotation) equation describes the rotation of a rigid body with respect to the body's principal axes of inertia. Assume that I_1, I_2, and I_3 are the moments of inertia around the primary, intermediate, and third principal axes with $I_1 > I_2 > I_3$. Also assume that ω_1, ω_2, and ω_3 are the components of the angular velocity along each axis. When there is no torque applied, using a principal orthogonal coordinates, Euler's

equation tells us that

$$I_1\dot{\omega}_1 = (I_2 - I_3)\omega_2\omega_3 \tag{32.4}$$
$$I_2\dot{\omega}_2 = (I_3 - I_1)\omega_3\omega_1 \tag{32.5}$$
$$I_3\dot{\omega}_3 = (I_1 - I_2)\omega_1\omega_2. \tag{32.6}$$

(The dots indicate a derivative with respect to time, which is common notation in physics.) We will use Euler's equations to understand the Tennis Racket Theorem.

Project Activity 32.1. To start, we consider rotation around the first principal axis. Our goal is to show that rotation around this axis is stable. That is, small perturbations in angular velocity will have only small effects on the rotation of the object. So we assume that ω_2 and ω_3 are small. In general, the product of two small quantities will be much smaller, so (32.4) implies that $\dot{\omega}_1$ must be very small. So we can disregard $\dot{\omega}_1$ in our calculations.

(a) Differentiate (32.5) with respect to time to explain why

$$I_2\ddot{\omega}_2 \approx (I_3 - I_1)\dot{\omega}_3\omega_1.$$

(b) Substitute for $\dot{\omega}_3$ from (32.6) to show that ω_2 is an approximate solution to

$$\ddot{\omega}_2 = -k\omega_2 \tag{32.7}$$

for some positive constant k.

(c) The equation (32.7) is a differential equation because it is an equation that involves derivatives of a function. Show by differentiating twice that, if

$$\omega_2 = A\cos\left(\sqrt{k}t + B\right) \tag{32.8}$$

(where A and B are any scalars), then ω_2 is a solution to (32.7). (In fact, ω_2 is the general solution to (32.7), which is verified in just about any course in differential equations.)

Equation 32.8 shows that ω_2 is is bounded, so that any slight perturbations in angular velocity have a limited effect on ω_2. A similar argument can be made for ω_3. This implies that the rotation around the principal axes is stable – slight changes in angular velocity have limited effects on the rotations around the other axes.

We can make a similar argument for rotation around the third principal axes.

Project Activity 32.2. In this activity, repeat the process from Project Activity to show that rotation around the third principal axis is stable. So assume that ω_1 and ω_3 are small, which implies by (32.6) implies that $\dot{\omega}_3$ must be very small and can be disregarded in calculations.

Now the issue is why is rotation around the second principal axis different.

Project Activity 32.3. Now assume that ω_1 and ω_3 are small. Thus, $\dot{\omega}_2$ is very small by (32.5), and we consider $\dot{\omega}_2$ to be negligible.

(a) Differentiate (32.4) to show that

$$I_1\ddot{\omega}_1 \approx (I_2 - I_3)\omega_2\dot{\omega}_3.$$

(b) Substitute for $\dot{\omega}_3$ from (32.6) to show that ω_1 is an approximate solution to

$$\ddot{\omega}_1 = k\omega_1 \tag{32.9}$$

for some positive scalar k.

(c) The fact that the constant multiplier in (32.9) is positive instead of negative as in (32.7) completely changes the type of solution. Show that

$$\omega_1 = Ae^{\sqrt{k}t+B} \tag{32.10}$$

(where A and B are any scalars) is a solution to (32.9) (and, in fact, is the general solution). Explain why this shows that rotation around the second principal axis is not stable.

Section 33

The Singular Value Decomposition

Focus Questions

By the end of this section, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the section.

- What is the operator norm of a matrix and what does it tell us about the matrix?
- What is a singular value decomposition of a matrix? Why is a singular value decomposition important?
- How does a singular value decomposition relate fundamental subspaces connected to a matrix?
- What is an outer product decomposition of a matrix and how is it useful?

Application: Search Engines and Semantics

Effective search engines search for more than just words. Language is complex and search engines must deal with the fact that there are often many ways to express a given concept (this is called *synonymy*, that multiple words can have the same meaning), and that a single word can have multiple meanings (*polysemy*). As a consequence, a search on a word may provide irrelevant matches (e.g., searching for *derivative* could provide pages on mathematics or financial securities) or you might search for articles on *cats* but the paper you really want uses the word *felines*. A better search engine will not necessarily try to match terms, but instead retrieve information based on concept or intent. Latent Semantic Indexing (LSI) (or *Latent Semantic Analysis*), developed in the late 1980s, helps search engines determine concept and intent in order to provide more accurate and relevant results. LSI essentially works by providing underlying (latent) relationships between words (semantics) that search engines need to provide context and understanding (indexing). LSI provides a mapping of both words and documents into a lower dimensional "concept" space, and makes the search in this new space. The mapping is provided by the singular value decomposition.

Introduction

The singular value decomposition (SVD) of a matrix is an important and useful matrix decomposition. Unlike other matrix decompositions, *every* matrix has a singular value decomposition. The SVD is used in a variety of applications including scientific computing, digital signal processing, image compression, principal component analysis, web searching through latent semantic indexing, and seismology. Recall that the eigenvector decomposition of an $n \times n$ diagonalizable matrix M has the form $P^{-1}MP$, where the columns of the matrix P are n linearly independent eigenvectors of M and the diagonal entries of the diagonal matrix $P^{-1}MP$ are the eigenvalues of M. The singular value decomposition does something similar for any matrix of any size. One of the keys to the SVD is that the matrix $A^{\mathsf{T}}A$ is symmetric for any matrix A.

The Operator Norm of a Matrix

Before we introduce the Singular Value Decomposition, let us work through some preliminaries to motivate the idea. The first is to provide an answer to the question "How 'big' is a matrix?" There are many ways to interpret and answer this question, but a substantial (and useful) answer should involve more than just the dimensions of the matrix. A good measure of the size of a matrix, which we will refer to as the norm of the matrix, should take into account the action of the linear transformation defined by the matrix on vectors. This then will lead to questions about how difficult or easy is it to solve a matrix equation $A\mathbf{x} = \mathbf{b}$.

If we want to incorporate the action of a matrix A into a calculation of the norm of A, we might think of measuring how much A can change a vector $\mathbf{x}$. This could lead us to using $||A\mathbf{x}||$ as some sort of measure of a norm of A. However, since $||A(c\mathbf{x})|| = |c|\ ||A\mathbf{x}||$ for any scalar c, scaling $\mathbf{x}$ by a large scalar will produce a large norm, so this is not a viable definition of a norm. We could instead measure the *relative* effect that A has on a vector $\mathbf{x}$ as $\dfrac{||A\mathbf{x}||}{||\mathbf{x}||}$, since this ratio does not change when $\mathbf{x}$ is multiplied by an scalar. The largest of all of these ratios would provide a good sense of how much A can change vectors. Thus, we define the operator norm of a matrix A as follows.

Definition 33.1. The **operator norm**[1] of a matrix A is

$$||A|| = \max_{||\mathbf{x}||\neq 0} \left\{ \frac{||A\mathbf{x}||}{||\mathbf{x}||} \right\}.$$

Due to the linearity of matrix multiplication, we can restrict ourselves to unit vectors for an equivalent definition of the operator norm of the matrix A as

$$||A|| = \max_{||\mathbf{x}||=1} \{||A\mathbf{x}||\}.$$

Preview Activity 33.1.

(1) Determine $||A||$ if A is the zero matrix.

[1] Technically this definition should be in terms of a supremum, but because the equivalent definition restricts the $\mathbf{x}$'s to a compact subset, the sup is achieved and we can use max.

(2) Determine $||I_n||$, where I_n is the $n \times n$ identity matrix.

(3) Let $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$. Find $||A||$. Justify your answer. (Hint: $x_1^2 + 4x_2^2 \leq 4(x_1^2 + x_2^2)$.)

(4) If P is an orthogonal matrix, what is $||P||$? Why?

The operator norm of a matrix tells us that how big the action of an $m \times n$ matrix is can be determined by its action on the unit sphere in $\mathbb{R}^n$ (the unit sphere is the set of terminal point of unit vectors). Let us consider two examples.

Example 33.2. Let $A = \begin{bmatrix} 2 & 1 \\ 2 & 5 \end{bmatrix}$. We can draw a graph to see the action of A on the unit circle. A picture of the set $\{A\mathbf{x} \ : \ ||\mathbf{x}|| = 1\}$ is shown in Figure 33.1.

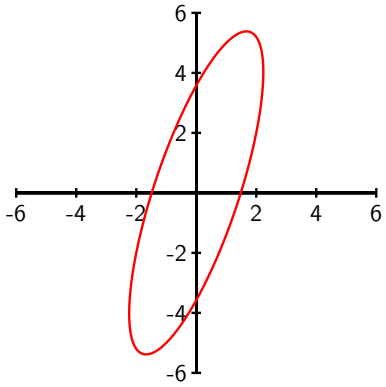

Figure 33.1: The image of the unit circle under the action of A

It appears that A transforms the unit circle into an ellipse. To find $||A||$ we want to maximize $||A\mathbf{x}||$ for $\mathbf{x}$ on the unit circle. This is the same as maximizing

$$||A\mathbf{x}||^2 = (A\mathbf{x})^\mathsf{T}(A\mathbf{x}) = \mathbf{x}^\mathsf{T}A^\mathsf{T}A\mathbf{x}.$$

Now $A^\mathsf{T}A = \begin{bmatrix} 8 & 12 \\ 12 & 26 \end{bmatrix}$ is a symmetric matrix, so we can orthogonally diagonalize $A^\mathsf{T}A$. The eigenvalues of $A^\mathsf{T}A$ are 32 and 2. Let $P = [\mathbf{u}_1 \ \mathbf{u}_2]$, where $\mathbf{u}_1 = \begin{bmatrix} \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix}^\mathsf{T}$ is a unit eigenvector of $A^\mathsf{T}A$ with eigenvalue 32 and $\mathbf{u}_2 = \begin{bmatrix} -\frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \end{bmatrix}^\mathsf{T}$ is a unit eigenvector of $A^\mathsf{T}A$ with eigenvalue 2. Then P is an orthogonal matrix such that $P^\mathsf{T}(A^\mathsf{T}A)P = \begin{bmatrix} 32 & 0 \\ 0 & 2 \end{bmatrix} = D$. It follows that

$$\mathbf{x}^\mathsf{T}(A^\mathsf{T}A)\mathbf{x} = \mathbf{x}^\mathsf{T}PDP^\mathsf{T}\mathbf{x} = (P^\mathsf{T}\mathbf{x})^\mathsf{T}D(P^\mathsf{T}\mathbf{x}).$$

Now P^T is orthogonal, so $||P^\mathsf{T}\mathbf{x}|| = ||\mathbf{x}||$ and P^T maps the unit circle to the unit circle. Moreover, if $\mathbf{x}$ is on the unit circle, then $\mathbf{y} = P\mathbf{x}$ is also on the unit circle and $P^\mathsf{T}\mathbf{y} = P^\mathsf{T}P\mathbf{x} = \mathbf{x}$. So every point $\mathbf{x}$ on the unit circle corresponds to a point $P\mathbf{x}$ on the unit circle. Thus, the forms $\mathbf{x}^\mathsf{T}(A^\mathsf{T}A)\mathbf{x}$ and $(P^\mathsf{T}\mathbf{x})^\mathsf{T}D(P^\mathsf{T}\mathbf{x})$ take on exactly the same values over all points on the unit circle. Now we

just need to find the maximum value of $(P^{\mathsf{T}}\mathbf{x})^{\mathsf{T}}D(P^{\mathsf{T}}\mathbf{x})$. This turns out to be relatively easy since D is a diagonal matrix.

Let's simplify the notation. Let $\mathbf{y} = P^{\mathsf{T}}\mathbf{x}$. Then our job is to maximize $\mathbf{y}^{\mathsf{T}}D\mathbf{y}$. If $\mathbf{y} = [y_1\ y_2]^{\mathsf{T}}$, then

$$\mathbf{y}^{\mathsf{T}}D\mathbf{y} = 32y_1^2 + 2y_2^2.$$

We want to find the maximum value of this expression for $\mathbf{y}$ on the unit circle. Note that $2y_2^2 \leq 32y_2^2$ and so

$$32y_1^2 + 2y_2^2 \leq 32y_1^2 + 32y_2^2 = 32(y_1^2 + y_2^2) = 32||\mathbf{y}||^2 = 32.$$

Since $[1\ 0]^{\mathsf{T}}$ is on the unit circle, the expression $32y_1^2 + 2y_2^2$ attains the value 32 at some point on the unit circle, so 32 is the maximum value of $\mathbf{y}^{\mathsf{T}}D\mathbf{y}$ over all $\mathbf{y}$ on the unit circle. While we are at it, we can similarly find the minimum value of $\mathbf{y}^{\mathsf{T}}D\mathbf{y}$ for $\mathbf{y}$ on the unit circle. Since $2y_1^2 \leq 32y_1^2$ we see that

$$32y_1^2 + 2y_2^2 \geq 2y_1^2 + 2y_2^2 = 2(y_1^2 + y_2^2) = 2||\mathbf{y}||^2 = 2.$$

Since the expression $\mathbf{y}^{\mathsf{T}}D\mathbf{y}$ attains the value 2 at $[0\ 1]^{\mathsf{T}}$ on the unit circle, we can see that $\mathbf{y}^{\mathsf{T}}D\mathbf{y}$ attains the minimum value of 2 on the unit circle.

Now we can return to the expression $\mathbf{x}^{\mathsf{T}}(A^{\mathsf{T}}A)\mathbf{x}$. Since $\mathbf{y}^{\mathsf{T}}D\mathbf{y}$ assumes the same values as $\mathbf{x}^{\mathsf{T}}(A^{\mathsf{T}}A)\mathbf{x}$, we can say that the maximum value of $\mathbf{x}^{\mathsf{T}}(A^{\mathsf{T}}A)\mathbf{x}$ for $\mathbf{x}$ on the unit circle is 32 (and the minimum value is 2). Moreover, the quadratic form $(P^{\mathsf{T}}\mathbf{x})^{\mathsf{T}}D(P^{\mathsf{T}}\mathbf{x})$ assumes its maximum value when $P^{\mathsf{T}}\mathbf{x} = [1\ 0]^{\mathsf{T}}$ or $[-1\ 0]^{\mathsf{T}}$. Thus, the form $\mathbf{x}^{\mathsf{T}}(A^{\mathsf{T}}A)\mathbf{x}$ assumes its maximum value at the vector $\mathbf{x} = P[1\ 0]^{\mathsf{T}} = \mathbf{u}_1$ or $-\mathbf{u}_1$. Similarly, the quadratic form $\mathbf{x}^{\mathsf{T}}(A^{\mathsf{T}}A)\mathbf{x}$ attains its minimum value at $P[0\ 1]^{\mathsf{T}} = \mathbf{u}_2$ or $-\mathbf{u}_2$. We conclude that $||A|| = \sqrt{32}$.

Figure 33.2 shows the image of the unit circle under the action of A and the images of $A\mathbf{u}_1$ and $A\mathbf{u}_2$ where $\mathbf{u}_1, \mathbf{u}_2$ are the two unit eigenvectors of $A^{\mathsf{T}}A$. The image also supports that $A\mathbf{x}$ assumes its maximum and minimum values for points on the unit circle at $\mathbf{u}_1$ and $\mathbf{u}_2$.

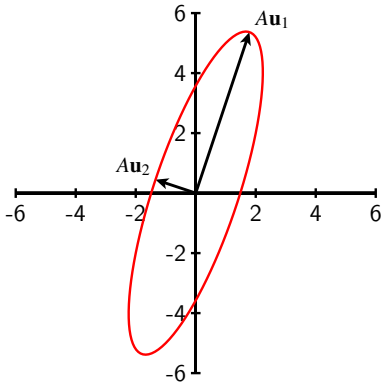

Figure 33.2: The image of the unit circle under the action of A, and the vectors $A\mathbf{u}_1$ and $A\mathbf{u}_2$

IMPORTANT NOTE 1: What we have just argued is that the maximum value of $||A\mathbf{x}||$ for $\mathbf{x}$ on the unit sphere in $\mathbb{R}^n$ is the square root of the largest eigenvalue of $A^{\mathsf{T}}A$ and occurs at a corresponding unit eigenvector.

Example 33.3. This same process works for matrices other than 2×2 ones. For example, consider $A = \begin{bmatrix} -2 & 8 & 20 \\ 14 & 19 & 10 \end{bmatrix}$. In this case A maps $\mathbb{R}^3$ to $\mathbb{R}^2$. The image of the unit sphere $\{\mathbf{x} \in \mathbb{R}^3 : ||\mathbf{x}|| = 1\}$ under left multiplication by A is a filled ellipse as shown in Figure 33.3. As with the

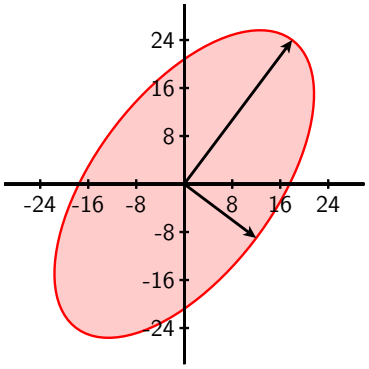

Figure 33.3: The image of the unit circle under the action of A, and the vectors $A\mathbf{u}_1$ and $A\mathbf{u}_2$

previous example, the norm of A is the square root of the maximum value of $\mathbf{x}^\mathsf{T}(A^\mathsf{T}A)\mathbf{x}$ and this maximum value is the dominant eigenvalue of $A^\mathsf{T}A = \begin{bmatrix} 200 & 250 & 100 \\ 250 & 425 & 350 \\ 100 & 350 & 500 \end{bmatrix}$. The eigenvalues of A are $\lambda_1 = 900$, $\lambda_2 = 225$, and $\lambda_3 = 0$ with corresponding unit eigenvectors $\mathbf{u}_1 = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \end{bmatrix}^\mathsf{T}$, $\mathbf{u}_1 = \begin{bmatrix} -\frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}^\mathsf{T}$, and $\mathbf{u}_3 = \begin{bmatrix} \frac{2}{3} & -\frac{2}{3} & \frac{1}{3} \end{bmatrix}^\mathsf{T}$. So in this case we have $||A|| = \sqrt{900} = 30$. The transformation defined by matrix multiplication by A from $\mathbb{R}^3$ to $\mathbb{R}^2$ has a one-dimensional kernel which is spanned by the eigenvector corresponding to λ_3. The image of the transformation is 2-dimensional and the image of the unit circle is an ellipse where $A\mathbf{u}_1$ gives the major axis of the ellipse and $A\mathbf{u}_2$ gives the minor axis. Essentially, the square roots of the eigenvalues of $A^\mathsf{T}A$ tell us how A stretches the image space in each direction.

IMPORTANT NOTE 2: We have just argued that the image of the unit n-sphere under the action of an $m \times n$ matrix is an ellipsoid in $\mathbb{R}^m$ stretched the greatest amount, $\sqrt{\lambda_1}$, in the direction of an eigenvector for the largest eigenvalue (λ_1) of $A^\mathsf{T}A$; the next greatest amount, $\sqrt{\lambda_2}$, in the direction of a unit vector for the second largest eigenvalue (λ_2) of $A^\mathsf{T}A$; and so on.

Activity 33.1. Let $A = \begin{bmatrix} 0 & 5 \\ 4 & 3 \\ -2 & 1 \end{bmatrix}$. Then $A^\mathsf{T}A = \begin{bmatrix} 20 & 10 \\ 10 & 35 \end{bmatrix}$. The eigenvalues of $A^\mathsf{T}A$ are $\lambda_1 = 40$ and $\lambda_2 = 15$ with respective eigenvectors $\mathbf{v}_1 = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$.

(a) Find $||A||$.

(b) Find a unit vector $\mathbf{x}$ at which $||A\mathbf{x}||$ assumes its maximum value.

The SVD

The Singular Value Decomposition (SVD) is essentially a concise statement of what we saw in the previous section that works for *any* matrix. We will uncover the SVD in this section.

Preview Activity 33.2. Let $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$. Since A is not square, we cannot diagonalize A. However, the matrix

$$A^{\mathsf{T}}A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

is a symmetric matrix and can be orthogonally diagonalized. The eigenvalues of $A^{\mathsf{T}}A$ are 3, 1, and 0 with corresponding eigenvectors

$$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \text{ and } \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix},$$

respectively. Use appropriate technology to do the following.

(1) Find an orthogonal matrix $V = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]$ that orthogonally diagonalizes $A^{\mathsf{T}}A$, where

$$V^{\mathsf{T}}\left(A^{\mathsf{T}}A\right)V = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

(2) For $i = 1, 2$, let $\mathbf{u}_i = \frac{A\mathbf{v}_i}{||A\mathbf{v}_i||}$. Find each $\mathbf{u}_i$. Why don't we define $\mathbf{u}_3$ in this way?

(3) Let $U = [\mathbf{u}_1\ \mathbf{u}_2]$. What kind of matrix is U? Explain.

(4) Calculate the matrix product $U^{\mathsf{T}}AV$. What do you notice? How is this similar to the eigenvector decomposition of a matrix?

Preview Activity 33.2 contains the basic ideas behind the Singular Value Decomposition. Let A be an $m \times n$ matrix with real entries. Note that $A^{\mathsf{T}}A$ is a symmetric $n \times n$ matrix and, hence, it can be orthogonally diagonalized. Let $V = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3\ \cdots\ \mathbf{v}_n]$ be an $n \times n$ orthogonal matrix whose columns form an orthonormal set of eigenvectors for $A^{\mathsf{T}}A$. For each i, let $(A^{\mathsf{T}}A)\mathbf{v}_i = \lambda_i\mathbf{v}_i$. We know

$$V^{\mathsf{T}}(A^{\mathsf{T}}A)V = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

Now notice that for each i we have

$$||A\mathbf{v}_i||^2 = (A\mathbf{v}_i)^{\mathsf{T}}(A\mathbf{v}_i) = \mathbf{v}_i^{\mathsf{T}}(A^{\mathsf{T}}A)\mathbf{v}_i = \mathbf{v}_i^{\mathsf{T}}\lambda_i\mathbf{v}_i = \lambda_i||\mathbf{v}_i||^2 = \lambda_i, \tag{33.1}$$

so $\lambda_i \geq 0$. Thus, the matrix $A^{\mathsf{T}}A$ has no negative eigenvalues. We can always arrange the eigenvectors and eigenvalues of $A^{\mathsf{T}}A$ so that

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

Also note that

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

if $i \neq j$. So the set $\{A\mathbf{v}_1, A\mathbf{v}_2, \ldots, A\mathbf{v}_n\}$ is an orthogonal set in $\mathbb{R}^m$. Each of the vectors $A\mathbf{v}_i$ is in Col A, and so $\{A\mathbf{v}_1, A\mathbf{v}_2, \ldots, A\mathbf{v}_n\}$ is an orthogonal subset of Col A. It is possible that $A\mathbf{v}_i = \mathbf{0}$ for some of the $\mathbf{v}_i$ (if $A^{\mathsf{T}}A$ has 0 as an eigenvalue). Let $\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_r$ be the eigenvectors corresponding to the nonzero eigenvalues. Then the set

$$\mathcal{B} = \{A\mathbf{v}_1, A\mathbf{v}_2, \ldots, A\mathbf{v}_r\}$$

is a linearly independent set of nonzero orthogonal vectors in Col A. Now we will show that $\mathcal{B}$ is a basis for Col A. Let $\mathbf{y}$ be a vector in Col A. Then $\mathbf{y} = A\mathbf{x}$ for some vector $\mathbf{x}$ in $\mathbb{R}^n$. Recall that the vectors $\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_n$ form an orthonormal basis of $\mathbb{R}^n$, so

$$\mathbf{x} = x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \cdots + x_n\mathbf{v}_n$$

for some scalars x_1, x_2, …, x_n. Since $A\mathbf{v}_j = \mathbf{0}$ for $r+1 \leq j \leq n$ we have

$$\begin{aligned}\mathbf{y} &= A\mathbf{x}\\ &= A(x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \cdots + x_n\mathbf{v}_n)\\ &= x_1A\mathbf{v}_1 + x_2A\mathbf{v}_2 + \cdots + x_rA\mathbf{v}_r + x_{r+1}A\mathbf{v}_{r+1} + \cdots + x_nA\mathbf{v}_n\\ &= x_1A\mathbf{v}_1 + x_2A\mathbf{v}_2 + \cdots + x_rA\mathbf{v}_r.\end{aligned}$$

So Span $\mathcal{B}$ = Col A and $\mathcal{B}$ is an orthogonal basis for Col A.

Now we are ready to find the Singular Value Decomposition of A. First we create an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ for Col A by normalizing the vectors $A\mathbf{v}_i$. So we let

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

for i from 1 to r.

Remember from (33.1) that $||A\mathbf{v}_i||^2 = \lambda_i$, so if we let $\sigma_i = \sqrt{\lambda_i}$, then we have

$$\mathbf{u}_i = \frac{A\mathbf{v}_i}{\sigma_i} \text{ and } A\mathbf{v}_i = \sigma_i\mathbf{u}_i.$$

We ordered the λ_i so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, so we also have

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

The scalars σ_1, σ_2, …, σ_n are called the *singular values* of A.

Definition 33.4. Let A be an $m \times n$ matrix. The **singular values** of A are the square roots of the eigenvalues of $A^{\mathsf{T}}A$.

The vectors $\mathbf{u}_1$, $\mathbf{u}_2$, …, $\mathbf{u}_r$ are r orthonormal vectors in $\mathbb{R}^m$. We can extend the set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ to an orthonormal basis $\mathcal{C} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}\mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$ of $\mathbb{R}^m$. Recall that $A\mathbf{v}_i = \sigma_i\mathbf{u}_i$ for $1 \leq i \leq r$ and $A\mathbf{v}_j = \mathbf{0}$ for $r+1 \leq j \leq n$, so

$$\begin{aligned}AV &= A[\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_n]\\ &= [A\mathbf{v}_1\ A\mathbf{v}_2\ \cdots\ A\mathbf{v}_n]\\ &= [\sigma_1\mathbf{u}_1\ \sigma_2\mathbf{u}_2\ \cdots\ \sigma_r\mathbf{u}_r\ \mathbf{0}\ \mathbf{0}\ \cdots\ \mathbf{0}].\end{aligned}$$

We can write the matrix $[\sigma_1\mathbf{v}_1\ \sigma_2\mathbf{v}_2\ \cdots\ \sigma_r\mathbf{v}_r\ \mathbf{0}\ \mathbf{0}\ \cdots\ \mathbf{0}]$ in another way. Let Σ be the $m \times n$ matrix defined as

$$\Sigma = \left[\begin{array}{ccccc|c} \sigma_1 & & & & 0 & \\ & \sigma_2 & & & & 0 \\ & & \sigma_3 & & & \\ & & & \ddots & & \\ 0 & & & & \sigma_r & \\ \hline & & 0 & & & 0 \end{array}\right].$$

Now

$$[\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_m]\Sigma = [\sigma_1\mathbf{u}_1\ \sigma_1\mathbf{u}_2\ \cdots\ \sigma_r\mathbf{u}_r\ \mathbf{0}\ \mathbf{0}\ \cdots\ \mathbf{0}] = AV.$$

So if $U = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_m]$, then

$$U\Sigma = AV.$$

Since V is an orthogonal matrix, we have that

$$U\Sigma V^{\mathsf{T}} = AVV^{\mathsf{T}} = A.$$

This is the Singular Value Decomposition of A.

Theorem 33.5 (The Singular Value Decomposition)**.** *Let A be an $m \times n$ matrix of rank r. There exist an $m \times m$ orthogonal matrix U, an $n \times n$ orthogonal matrix V, and an $m \times n$ matrix Σ whose first r diagonal entries are the singular values σ_1, σ_2, ..., σ_r and whose other entries are 0, such that*

$$A = U\Sigma V^{\mathsf{T}}.$$

SVD Summary: A Singular Value Decomposition of an $m \times n$ matrix A of rank r can be found as follows.

(1) Find an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n\}$ of eigenvectors of $A^{\mathsf{T}}A$ such that $(A^{\mathsf{T}}A)\mathbf{v}_i = \lambda_i\mathbf{v}_i$ for i from 1 to n with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ with the first r eigenvalues being positive. The vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, ..., $\mathbf{v}_n$ are the *right singular vectors* of A.

(2) Let

$$V = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3\ \cdots\ \mathbf{v}_n].$$

Then V orthogonally diagonalizes $A^{\mathsf{T}}A$.

(3) The singular values of A are the numbers σ_i, where $\sigma_i = \sqrt{\lambda_i} > 0$ for i from 1 to r. Let Σ be the $m \times n$ matrix

$$\Sigma = \left[\begin{array}{ccccc|c} \sigma_1 & & & & 0 & \\ & \sigma_2 & & & & 0 \\ & & \sigma_3 & & & \\ & & & \ddots & & \\ 0 & & & & \sigma_r & \\ \hline & & 0 & & & 0 \end{array}\right]$$

(4) For i from 1 to r, let $\mathbf{u}_i = \frac{A\mathbf{v}_i}{||A\mathbf{v}_i||}$. Then the set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ forms an orthonormal basis of Col A.

(5) Extend the set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ to an orthonormal basis

$$\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}\mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$$

of $\mathbb{R}^m$. Let

$$U = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_m].$$

The vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_m$ are the *left singular vectors* of A.

(6) Then $A = U\Sigma V^{\mathsf{T}}$ is a singular value decomposition of A.

Activity 33.2. Let $A = \begin{bmatrix} 0 & 5 \\ 4 & 3 \\ -2 & 1 \end{bmatrix}$. Then $A^{\mathsf{T}}A = \begin{bmatrix} 20 & 10 \\ 10 & 35 \end{bmatrix}$. The eigenvalues of $A^{\mathsf{T}}A$ are $\lambda_1 = 40$ and $\lambda_2 = 15$ with respective eigenvectors $\mathbf{w}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\mathbf{w}_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$.

(a) Find an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n\}$ of eigenvectors of $A^{\mathsf{T}}A$. What is n? Find the matrix V in a SVD for A.

(b) Find the singular values of A. What is the rank r of A? Why?

(c) What are the dimensions of the matrix Σ in the SVD of A? Find Σ.

(d) Find the vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_r$. If necessary, extend this set to an orthonormal basis

$$\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}\mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$$

of $\mathbb{R}^m$.

(e) Find the matrix U so that $A = U\Sigma V^{\mathsf{T}}$ is a SVD for A.

There is another way we can write this SVD of A. Let the $m \times n$ matrix A have a singular value decomposition $U\Sigma V^{\mathsf{T}}$, where

$$U = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_m],$$

$$\Sigma = \left[\begin{array}{cccc|c} \sigma_1 & & & 0 & \\ & \sigma_2 & & & 0 \\ & & \sigma_3 & & \\ & & & \ddots & \\ 0 & & & \sigma_r & \\ \hline & 0 & & & 0 \end{array}\right], \text{ and}$$

$$V = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3\ \cdots\ \mathbf{v}_n].$$

Since $A = U\Sigma V^{\mathsf{T}}$ we see that

$$A = [\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3\ \cdots\ \mathbf{u}_m] \left[\begin{array}{cccc|c} \sigma_1 & & & & 0 \\ & \sigma_2 & & & \\ & & \sigma_3 & & 0 \\ & & & \ddots & \\ 0 & & & \sigma_r & \\ \hline & 0 & & & 0 \end{array}\right] \begin{bmatrix} \mathbf{v}_1^{\mathsf{T}} \\ \mathbf{v}_2^{\mathsf{T}} \\ \mathbf{v}_3^{\mathsf{T}} \\ \vdots \\ \mathbf{v}_n^{\mathsf{T}} \end{bmatrix}$$

$$= [\sigma_1\mathbf{u}_1\ \sigma_2\mathbf{u}_2\ \sigma_3\mathbf{u}_3\ \cdots\ \sigma_r\mathbf{u}_r\ \mathbf{0}\ \cdots\ \mathbf{0}] \begin{bmatrix} \mathbf{v}_1^{\mathsf{T}} \\ \mathbf{v}_2^{\mathsf{T}} \\ \mathbf{v}_3^{\mathsf{T}} \\ \vdots \\ \mathbf{v}_n^{\mathsf{T}} \end{bmatrix}$$

$$= \sigma_1\mathbf{u}_1\mathbf{v}_1^{\mathsf{T}} + \sigma_2\mathbf{u}_2\mathbf{v}_2^{\mathsf{T}} + \sigma_3\mathbf{u}_3\mathbf{v}_3^{\mathsf{T}} + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^{\mathsf{T}}. \tag{33.2}$$

This is called an *outer product decomposition* of A and tells us everything we learned above about the action of the matrix A as a linear transformation. Each of the products $\mathbf{u}_i\mathbf{v}_i^{\mathsf{T}}$ is a rank 1 matrix (see Exercise 9.), and $||A\mathbf{v}_1|| = \sigma_1$ is the largest value A takes on the unit n-sphere, $||A\mathbf{v}_2|| = \sigma_2$ is the next largest dilation of the unit n-sphere, and so on. An outer product decomposition allows us to approximate A with smaller rank matrices. For example, the matrix $\sigma_1\mathbf{u}_1\mathbf{v}_1^{\mathsf{T}}$ is the best rank 1 approximation to A, $\sigma_1\mathbf{u}_1\mathbf{v}_1^{\mathsf{T}} + \sigma_2\mathbf{u}_2\mathbf{v}_2^{\mathsf{T}}$ is the best rank 2 approximation, and so on. This will be very useful in applications, as we will see in the next section.

SVD and the Null, Column, and Row Spaces of a Matrix

We conclude this section with a short discussion of how a singular value decomposition relates fundamental subspaces of a matrix. We have seen that the vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_r$ in an SVD for an $m \times n$ matrix A form a basis for Col A. Recall also that $A\mathbf{v}_j = \mathbf{0}$ for $r+1 \leq j \leq n$. Since $\dim(\text{Nul } A) + \dim(\text{Col } A) = n$, it follows that the vectors $\mathbf{v}_{r+1}$, $\mathbf{v}_{r+2}$, ..., $\mathbf{v}_n$ form a basis for Nul A. As you will show in the exercises, the set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ is a basis for Row A. Thus, an SVD for a matrix A tells us about three fundamental vector spaces related to A.

Examples

What follows are worked examples that use the concepts from this section.

Example 33.6. Let $A = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}$.

(a) Find a singular value decomposition for A. You may use technology to find eigenvalues and eigenvectors of matrices.

(b) Use the singular value decomposition to find a basis for Col A, Row A, and Nul A.

Example Solution.

(a) With A as given, we have $A^{\mathsf{T}}A = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 5 & 4 & 0 \\ 0 & 4 & 5 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$. Technology shows that the eigenvalues of $A^{\mathsf{T}}A$ are $\lambda_1 = 9$, $\lambda_2 = 4$, $\lambda_3 = 1$, and $\lambda_4 = 0$ with corresponding orthonormal eigenvectors $\mathbf{v}_1 = \frac{1}{\sqrt{2}}[0\,1\,1\,0]^{\mathsf{T}}$, $\mathbf{v}_2 = [1\,0\,0\,0]^{\mathsf{T}}$, $\mathbf{v}_3 = \frac{1}{\sqrt{2}}[0\,-1\,1\,0]^{\mathsf{T}}$, and $\mathbf{v}_4 = [0\,0\,0\,1]^{\mathsf{T}}$. This makes $V = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3\ \mathbf{v}_4]$. The singular values of A are $\sigma_1 = \sqrt{9} = 3$, $\sigma_2 = \sqrt{4} = 2$, $\sigma_3 = \sqrt{1} = 1$, and $\sigma_4 = 0$, so Σ is the 3×4 matrix with the nonzero singular values along the diagonal and zeros everywhere else. Finally, we define the vectors $\mathbf{u}_i$ as $\mathbf{u}_i = \frac{1}{||A\mathbf{v}_i||}A\mathbf{v}_i$. Again, technology gives us $\mathbf{u}_1 = \frac{1}{\sqrt{2}}[0\ 1\ 1]^{\mathsf{T}}$, $\mathbf{u}_2 = [1\ 0\ 0]^{\mathsf{T}}$, and $\mathbf{u}_3 = \frac{1}{\sqrt{2}}[0\ -1\ 1]^{\mathsf{T}}$. Thus, a singular value decomposition of A is $U\Sigma V^{\mathsf{T}}$, where

$$U = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \text{ and}$$

$$V = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

(b) Recall that the right singular vectors of an $m \times n$ matrix A of rank r form an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n\}$ of eigenvectors of $A^{\mathsf{T}}A$ such that $(A^{\mathsf{T}}A)\mathbf{v}_i = \lambda_i\mathbf{v}_i$ for i from 1 to n with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. These vectors are the columns of the matrix $V = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_n]$ in a singular value decomposition of A.

For i from 1 to r, we let $\mathbf{u}_i = \frac{A\mathbf{v}_i}{||A\mathbf{v}_i||}$. Then the set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ forms an orthonormal basis of Col A. We extend this set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ to an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}\mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$ of $\mathbb{R}^m$.

Recall also that $A\mathbf{v}_j = \mathbf{0}$ for $r + 1 \leq j \leq n$. Since $\dim(\text{Nul } A) + \dim(\text{Col } A) = n$, it follows that the vectors $\mathbf{v}_{r+1}$, $\mathbf{v}_{r+2}$, …, $\mathbf{v}_n$ form a basis for Nul A. Finally, the set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ is a basis for Row A.

So in our example, we have $m = 3$, $n = 4$, $\mathbf{v}_1 = \frac{1}{\sqrt{2}}[0\ 1\ 1\ 0]^{\mathsf{T}}$, $\mathbf{v}_2 = [1\ 0\ 0\ 0]^{\mathsf{T}}$, $\mathbf{v}_3 = \frac{1}{\sqrt{2}}[0\ -1\ 1\ 0]^{\mathsf{T}}$, and $\mathbf{v}_4 = [0\ 0\ 0\ 1]^{\mathsf{T}}$. Since the singular values of A are 3, 2, 1, and 0, it follows that $r = \text{rank}(A) = 3$. We also have $\mathbf{u}_1 = \frac{1}{\sqrt{2}}[0\ 1\ 1]^{\mathsf{T}}$, $\mathbf{u}_2 = [1\ 0\ 0]^{\mathsf{T}}$, and $\mathbf{u}_3 = \frac{1}{\sqrt{2}}[0\ -1\ 1]^{\mathsf{T}}$. So

$$\left\{\frac{1}{\sqrt{2}}[0\,1\,1\,0]^{\mathsf{T}}, [1\,0\,0\,0]^{\mathsf{T}}, \frac{1}{\sqrt{2}}[0\ -1\,1\,0]^{\mathsf{T}}\right\}$$

is a basis for Row A and

$$\{[0\ 0\ 0\ 1]^{\mathsf{T}}\}$$

is a basis for Nul A. Finally, the set

$$\left\{\frac{1}{\sqrt{2}}[0\ 1\ 1]^{\mathsf{T}}, [1\ 0\ 0]^{\mathsf{T}}, \frac{1}{\sqrt{2}}[[0\ -1\ 1]^{\mathsf{T}}\right\}$$

is a basis for Col A.

Example 33.7. Let

$$A = \begin{bmatrix} 2 & 5 & 4 \\ 6 & 3 & 0 \\ 6 & 3 & 0 \\ 2 & 5 & 4 \end{bmatrix}.$$

The eigenvalues of $A^{\mathsf{T}}A$ are $\lambda_1 = 144$, $\lambda_2 = 36$, and $\lambda_3 = 0$ with corresponding eigenvectors

$$\mathbf{w}_1 = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix},\ \mathbf{w}_1 = \begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix},\ \text{and}\ \mathbf{w}_1 = \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix}.$$

In addition,

$$A\mathbf{w}_1 = \begin{bmatrix} 18 \\ 18 \\ 18 \\ 18 \end{bmatrix} \text{ and } A\mathbf{w}_2 = \begin{bmatrix} 9 \\ -9 \\ -9 \\ 9 \end{bmatrix}.$$

(a) Find orthogonal matrices U and V, and the matrix Σ, so that $U\Sigma V^{\mathsf{T}}$ is a singular value decomposition of A.

(b) Determine the best rank 1 approximation to A.

Example Solution.

(a) Normalizing the eigenvectors $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ to normal eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, respectively, gives us an orthogonal matrix

$$V = \begin{bmatrix} \frac{2}{3} & -\frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} & -\frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \end{bmatrix}.$$

Now $A\mathbf{v}_i = A\frac{\mathbf{w}_i}{||\mathbf{w}_i||} = \frac{1}{||\mathbf{w}_i||}A\mathbf{w}_i$, so normalizing the vectors $A\mathbf{v}_1$ and $A\mathbf{v}_2$ gives us vectors

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

that are the first two columns of our matrix U. Given that U is a 4×4 matrix, we need to find two other vectors orthogonal to $\mathbf{u}_1$ and $\mathbf{u}_2$ that will combine with $\mathbf{u}_1$ and $\mathbf{u}_2$ to form an orthogonal basis for $\mathbb{R}^4$. Letting $\mathbf{z}_1 = [1\ 1\ 1\ 1]^\mathsf{T}$, $\mathbf{z}_2 = [1\ -1\ -1\ 1]^\mathsf{T}$, $\mathbf{z}_3 = [1\ 0\ 0\ 0]^\mathsf{T}$, and $\mathbf{z}_4 = [0\ 1\ 0\ 1]^\mathsf{T}$, a computer algebra system shows that the reduced row echelon form of the matrix $[\mathbf{z}_1\ \mathbf{z}_2\ \mathbf{z}_3\ \mathbf{z}_4]$ is I_4, so that vectors $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$, $\mathbf{z}_4$ are linearly independent. Letting $\mathbf{w}_1 = \mathbf{z}_1$ and $\mathbf{w}_2 = \mathbf{z}_2$, the Gram-Schmidt process shows that the set $\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4\}$ is an orthogonal basis for $\mathbb{R}^4$, where

$$\begin{aligned}
\mathbf{w}_3 &= [1\ 0\ 0\ 0]^\mathsf{T} - \frac{[1\ 0\ 0\ 0]^\mathsf{T} \cdot [1\ 1\ 1\ 1]^\mathsf{T}}{[1\ 1\ 1\ 1]^\mathsf{T} \cdot [1\ 1\ 1\ 1]^\mathsf{T}}[1\ 1\ 1\ 1]^\mathsf{T} \\
&\qquad - \frac{[1\ 0\ 0\ 0]^\mathsf{T} \cdot [1\ -1\ -1\ 1]^\mathsf{T}}{[1\ -1\ -1\ 1]^\mathsf{T} \cdot [1\ -1\ -1\ 1]^\mathsf{T}}[1\ -1\ -1\ 1]^\mathsf{T} \\
&= [1\ 0\ 0\ 0]^\mathsf{T} - \frac{1}{4}[1\ 1\ 1\ 1]^\mathsf{T} - \frac{1}{4}[1\ -1\ -1\ 1]^\mathsf{T} \\
&= \frac{1}{4}[2\ 0\ 0\ -2]^\mathsf{T}
\end{aligned}$$

and (using $[1\ 0\ 0\ -1]^\mathsf{T}$ for $\mathbf{w}_3$)

$$\begin{aligned}
\mathbf{w}_4 &= [0\ 1\ 0\ 0]^\mathsf{T} - \frac{[0\ 1\ 0\ 0]^\mathsf{T} \cdot [1\ 1\ 1\ 1]^\mathsf{T}}{[1\ 1\ 1\ 1]^\mathsf{T} \cdot [1\ 1\ 1\ 1]^\mathsf{T}}[1\ 1\ 1\ 1]^\mathsf{T} \\
&\qquad - \frac{[0\ 1\ 0\ 0]^\mathsf{T} \cdot [1\ -1\ -1\ 1]^\mathsf{T}}{[1\ -1\ -1\ 1]^\mathsf{T} \cdot [1\ -1\ -1\ 1]^\mathsf{T}}[1\ -1\ -1\ 1]^\mathsf{T} \\
&\qquad - \frac{[0\ 1\ 0\ 0]^\mathsf{T} \cdot [1\ 0\ 0\ -1]^\mathsf{T}}{[1\ 0\ 0\ -1]^\mathsf{T} \cdot [1\ 0\ 0\ -1]^\mathsf{T}}[1\ 0\ 0\ -1]^\mathsf{T} \\
&= [0\ 1\ 0\ 0]^\mathsf{T} - \frac{1}{4}[1\ 1\ 1\ 1]^\mathsf{T} + \frac{1}{4}[1\ -1\ -1\ 1]^\mathsf{T} - \mathbf{0} \\
&= \frac{1}{4}[0\ 2\ -2\ 0]^\mathsf{T}.
\end{aligned}$$

The set $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$ where $\mathbf{u}_1 = \frac{1}{2}[1\ 1\ 1\ 1]^\mathsf{T}$, $\mathbf{u}_2 = \frac{1}{2}[1\ -1\ -1\ 1]^\mathsf{T}$, $\mathbf{u}_3 = \frac{1}{\sqrt{2}}[1\ 0\ 0\ -1]^\mathsf{T}$ and $\mathbf{u}_4 = \frac{1}{\sqrt{2}}[0\ 1\ -1\ 0]^\mathsf{T}$ is an orthonormal basis for $\mathbb{R}^4$ and we can let

$$U = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}.$$

The singular values of A are $\sigma_1 = \sqrt{\lambda_1} = 12$ and $\sigma_2 = \sqrt{\lambda_2} = 6$, and so

$$\Sigma = \begin{bmatrix} 12 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Therefore, a singular value decomposition of A is $U\Sigma V^{\mathsf{T}}$ of

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} \begin{bmatrix} 12 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ -\frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \end{bmatrix}.$$

(b) Determine the best rank 1 approximation to A. The outer product decomposition of A is

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{T}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{T}}.$$

So the rank one approximation to A is

$$\sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{T}} = 12 \left(\frac{1}{2}\right) \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 4 & 4 & 2 \\ 4 & 4 & 2 \\ 4 & 4 & 2 \\ 4 & 4 & 2 \end{bmatrix}.$$

Note that the rows in this rank one approximation are the averages of the two distinct rows in the matrix A, which makes sense considering that this is the closest rank one matrix to A.

Summary

We learned about the singular value decomposition of a matrix.

- The operator norm of an $m \times n$ matrix A is

$$||A|| = \max_{||\mathbf{x}|| \neq 0} \frac{||A\mathbf{x}||}{||\mathbf{x}||} = \max_{||\mathbf{x}||=1} ||A\mathbf{x}||.$$

  The operator norm of a matrix tells us that how big the action of an $m \times n$ matrix is can be determined by its action on the unit sphere in $\mathbb{R}^n$.

- A singular value decomposition of an $m \times n$ matrix is of the form $A = U\Sigma V^{\mathsf{T}}$, where
  - $V = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3\ \cdots\ \mathbf{v}_n]$ where $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n\}$ is an orthonormal basis of eigenvectors of $A^{\mathsf{T}}A$ such that $(A^{\mathsf{T}}A)\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for i from 1 to n with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$,
  - $U = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_m]$ where $\mathbf{u}_i = \frac{A\mathbf{v}_i}{||A\mathbf{v}_i||}$ for i from 1 to r, and this orthonormal basis of Col A is extended to an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}\mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$ of $\mathbb{R}^m$,
  - $\Sigma = \left[\begin{array}{cccc|c} \sigma_1 & & & 0 & \\ & \sigma_2 & & & 0 \\ & & \sigma_3 & & \\ & & & \ddots & \\ 0 & & & \sigma_r & \\ \hline & & 0 & & 0 \end{array}\right]$, where $\sigma_i = \sqrt{\lambda_i} > 0$ for i from 1 to r.

A singular value decomposition is important in that every matrix has a singular value decomposition, and a singular value decomposition has a variety of applications including scientific computing and digital signal processing, image compression, principal component analysis, web searching through latent semantic indexing, and seismology.

- The vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_r$ in an SVD for an $m \times n$ matrix A form a basis for Col A while the vectors $\mathbf{v}_{r+1}$, $\mathbf{v}_{r+2}$, ..., $\mathbf{v}_n$ form a basis for Nul A. Also, the set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ is a basis for Row A.

- Let A have an SVD as in the second bullet. Decomposing A as

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\mathsf{T} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\mathsf{T} + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^\mathsf{T} + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^\mathsf{T}$$

is an outer product decomposition of A. An outer product decomposition allows us to approximate A with smaller rank matrices. For example, the matrix $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^\mathsf{T}$ is the best rank 1 approximation to A, $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^\mathsf{T} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\mathsf{T}$ is the best rank 2 approximation, and so on.

Exercises

(1) Find a singular value decomposition of the following matrices.

(a) $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 1 \\ 1 & 3 \end{bmatrix}$

(e) $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}$

(2) Let A be an $m \times n$ matrix of rank r with singular value decomposition $U\Sigma V^\mathsf{T}$, where $U = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_m]$ and $V = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_n]$. We have seen that the set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ is a basis for Col A, and the vectors $\mathbf{v}_{r+1}$, $\mathbf{v}_{r+2}$, ..., $\mathbf{v}_n$ form a basis for Nul A. In this exercise we examine the set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ and determine what this set tells us about Row A.

(a) Find a singular value decomposition for A^T. (Hint: Use the singular value decomposition $U\Sigma V^\mathsf{T}$ for A.)

(b) Explain why the result of (a) shows that the set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ is a basis for Row A.

(3) Let $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{bmatrix}$.

(a) Find the singular values of A.

(b) Find a singular value decomposition of A.

(c) Use a singular value decomposition to find orthonormal bases for the following:

i. Nul A
ii. Col A
iii. Row A

(4) Let A have the singular value decomposition as in (33.2).

(a) Show, using (33.2), that $||A\mathbf{v}_j|| = \sigma_j$.

(b) Explain why $||A|| = \sigma_1$. (Hint: The set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is an orthonormal basis of $\mathbb{R}^n$. Use this to show that $||A\mathbf{x}||^2 \leq \sigma_1^2$ for any unit vector $\mathbf{x}$ in $\mathbb{R}^n$.)

(5) Show that A and A^T have the same nonzero singular values. How are their singular value decompositions related?

(6) The vectors $\mathbf{v}_i$ that form the columns of the matrix V in a singular value decomposition of a matrix A are eigenvectors of $A^\mathsf{T}A$. In this exercise we investigate the vectors $\mathbf{u}_i$ that make up the columns of the matrix U in a singular value decomposition of a matrix A for each i between 1 and the rank of A, and their connection to the matrix AA^T.

(a) Let $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$. A singular value decomposition of A is $U\Sigma V^\mathsf{T}$, where

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix},$$
$$\Sigma = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$
$$V = \begin{bmatrix} -\frac{1}{\sqrt{6}} & -\frac{1}{3\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}.$$

i. Determine the rank r of $A^\mathsf{T}A$ and identify the vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_r$.
ii. Calculate $AA^\mathsf{T}\mathbf{u}_i$ for each i between 1 and r. How is $AA^\mathsf{T}\mathbf{u}_i$ related to $\mathbf{u}_i$?

(b) Now we examine the result of part (a) in general. Let A be an arbitrary matrix. Calculate $AA^\mathsf{T}\mathbf{u}_i$ for $1 \leq i \leq \text{rank}(A)$ and determine specifically how $AA^\mathsf{T}\mathbf{u}_i$ is related to $\mathbf{u}_i$. What does this tell us about the vectors $\mathbf{u}_i$ and the matrix AA^T?

(c) Now show in general that the columns of U are orthonormal eigenvectors for AA^T. (That is, what can we say about the vectors $\mathbf{u}_i$ if $i > \text{rank}(A)$?)

(7) If A is a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, what is $||A||$? Justify your answer.

(8) Let A be a $n \times n$ symmetric matrix.

(a) Show that if $\mathbf{v}$ is an eigenvector of A with eigenvalue λ, then $\mathbf{v}$ is an eigenvector for $A^\mathsf{T}A$. What is the corresponding eigenvalue?

(b) Show that if $\mathbf{v}$ is an eigenvector of $A^\mathsf{T}A$ with non-negative eigenvalue λ, then $A\mathbf{v}$ is an eigenvector of $A^\mathsf{T}A$. What is the corresponding eigenvalue?

(c) Suppose $U\Sigma V^{\mathsf{T}}$ is a singular value decomposition of A. Explain why $V\Sigma V^{\mathsf{T}}$ is also a singular value decomposition of A.

(9) Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ be the vectors found in a singular value decomposition of a matrix A, where r is the rank of A. Show that $\mathbf{u}_i\mathbf{v}_i^{\mathsf{T}}$ is a rank 1 matrix for each i. (Hint: Compare to Exercise 5. in Section 31.)

(10) Is it possible for a matrix A to have a singular value decomposition $U\Sigma V^{\mathsf{T}}$ in which $U = V$? If no, explain why. If yes, determine for which matrices we can have $U = V$.

(11) Label each of the following statements as True or False. Provide justification for your response.

(a) **True/False** If σ is a singular value of a matrix A, then σ is an eigenvalue of $A^{\mathsf{T}}A$.

(b) **True/False** A set of right singular vectors of a matrix A is also a set of left singular vectors of A^{T}.

(c) **True/False** The transpose of a singular value decomposition of a matrix A is a singular value decomposition for A^{T}.

(d) **True/False** Similar matrices have the same singular values.

(e) **True/False** If A is an $n \times n$ matrix, then the singular values of A^2 are the squares of the singular values of A.

(f) **True/False** The Σ matrix in an SVD of A is unique.

(g) **True/False** The matrices U, V in an SVD of A are unique.

(h) **True/False** If A is a positive definite matrix, then an orthogonal diagonalization of A is an SVD of A.

Project: Latent Semantic Indexing

As an elementary example to illustrate the idea behind Latent Semantic Indexing (LSI), consider the problem of creating a program to search a collection of documents for words, or words related to a given word. Document collections are usually very large, but we use a small example for illustrative purposes. A standard example that is given in several publications[2] is the following. Suppose we have nine documents c_1 through c_5 (titles of documents about human-computer interaction) and m_1 through m_4 (titles of graph theory papers) that make up our library:

- c_1: *Human* machine *interface* for ABC *computer* applications
- c_2: A *survey* of *user* opinion of *computer system response time*
- c_3: The *EPS user interface* management *system*
- c_4: *System* and *human system* engineering testing of *EPS*

[2] e.g., Deerwester, S., Dumais, S. T., Fumas, G. W., Landauer, T. K. and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990, 41: 391?407, and Landauer, T. and Dutnais, S. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 1997. Vol. 1M. No. 2, 211-240.

- c_5: Relation of *user* perceived *response time* to error measurement
- m_1: The generation of random, binary, ordered *trees*
- m_2: The intersection *graph* of paths in *trees*
- m_3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
- m_4: *Graph minors*: A *survey*

To make a searchable database, one might start by creating a list of key terms that appear in the documents (generally removing common words such as "a", "the", "of", etc., called *stop words* – these words contribute little, if any, context). In our documents we identify the key words that are shown in italics. (Note that we are just selecting key words to make our example manageable, not necessarily identifying the most important words.) Using the key words we create a *term-document* matrix. The term-document matrix is the matrix in which the terms form the rows and the documents the columns. If $A = [a_{ij}]$ is the term-document matrix, then a_{ij} counts the number of times word i appears in document j. The term-document matrix A for our library is

$$\begin{array}{l} \\ \text{human} \\ \text{interface} \\ \text{computer} \\ \text{user} \\ \text{system} \\ \text{response} \\ \text{time} \\ \text{EPS} \\ \text{survey} \\ \text{trees} \\ \text{graph} \\ \text{minors} \end{array} \begin{array}{c} \begin{array}{ccccccccc} c_1 & c_2 & c_3 & c_4 & c_5 & m_1 & m_2 & m_3 & m_4 \end{array} \\ \left[\begin{array}{ccccccccc} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right] \end{array}. \tag{33.3}$$

One of our goals is to rate the pages in our library for relevance if we search for a query. For example, suppose we want to rate the pages for the query *survey, computer*. This query can be represented by the vector $\mathbf{q} = [0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0]^{\mathsf{T}}$.

Project Activity 33.1. In a standard term-matching search with $m \times n$ term-document matrix A, a query vector $\mathbf{q}$ would be matched with the terms to determine the number of matches. The matching counts the number of times each document agrees with the query.

(a) Explain why this matching is accomplished by the matrix-vector product $A^{\mathsf{T}}\mathbf{q}$.

(b) Let $\mathbf{y} = [y_1\ y_2\ \ldots\ y_n]^{\mathsf{T}} = A^{\mathsf{T}}\mathbf{q}$. Explain why $y_i = \cos(\theta_i)||\mathbf{a}_i||||\mathbf{q}||$, where $\mathbf{a}_i$ is the ith column of A and θ_i is the angle between $\mathbf{a}_i$ and $\mathbf{q}$.

(c) We can use the cosine calculation from part (b) to compare matches to our query – the closer the cosine is to 1, the better the match (dividing by the product of the norms is essentially converting all vectors to unit vectors for comparison purposes). This is often referred to as the cosine distance. Calculate the cosines of the θ_i for our example of the query $\mathbf{q} = [0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0]^{\mathsf{T}}$. Order the documents from best to worst match for this query.

Though we were able to rate the documents in Project Activity 33.1 using the cosine distance, the result is less than satisfying. Documents c_3, c_4, and c_5 are all related to computers but do not appear at all in or results. This is a problem with language searches – we don't want to compare just words, but we also need to compare the concepts the words represent. The fact that words can represent different things implies that a random choice of word by different authors can introduce noise into the word-concept relationship. To filter out this noise, we can apply the singular value decomposition to find a smaller set of concepts to better represent the relationships. Before we do so, we examine some useful properties of the term-document matrix.

Project Activity 33.2. Let $A = [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_9]$, where $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_9$ are the columns of A.

(a) In Project Activity 33.1 you should have seen that $b_{ij} = \mathbf{a}_i^\mathsf{T}\mathbf{a}_j = \mathbf{a}_i \cdot \mathbf{a}_j$. Assume for the moment that all of the entries in A are either 0 or 1. Explain why in this case the dot product $\mathbf{a}_i \cdot \mathbf{a}_j$ tells us how many terms documents i and j have in common. Also, the matrix $A^\mathsf{T}A$ takes dot products of the columns of A, which refer to what's happening in each document and so is looking at document-document interactions. For these reasons, we call $A^\mathsf{T}A$ the document-document matrix.

(b) Use appropriate technology to calculate the entries of the matrix $C = [c_{ij}] = AA^\mathsf{T}$. This matrix is the term-term matrix. Assume for the moment that all of the entries in A are either 0 or 1. Explain why if terms i and j occur together in k documents, then $c_{ij} = k$.

The nature of the term-term and document-document matrices makes it realistic to think about a SVD.

Project Activity 33.3. To see why a singular value decomposition might be useful, suppose our term-document matrix A has singular value decomposition $A = U\Sigma V^\mathsf{T}$. (Don't actually calculate the SVD yet).

(a) Show that the document-document matrix $A^\mathsf{T}A$ satisfies $A^\mathsf{T}A = \left(V\Sigma^\mathsf{T}\right)\left(V\Sigma^\mathsf{T}\right)^\mathsf{T}$. This means that we can compare document i and document j using the dot product of row i and column j of the matrix product $V\Sigma^\mathsf{T}$.

(b) Show that the term-term matrix AA^T satisfies $AA^\mathsf{T} = (U\Sigma)\,(U\Sigma)^\mathsf{T}$. Thus we can compare term i and term j using the dot product of row i and column j of the matrix product $U\Sigma$. (Exercise 6 shows that the columns of U are orthogonal eigenvectors of AA^T.)

As we will see, the connection of the matrices U and V to documents and terms that we saw in Project Activity 33.3 will be very useful when we use the SVD of the term-document matrix to reduce dimensions to a "concept" space. We will be able to interpret the rows of the matrices U and V as providing coordinates for terms and documents in this space.

Project Activity 33.4. The singular value decomposition (SVD) allows us to produce new, improved term-document matrices. For this activity, use the term-document matrix A in (33.3).

(a) Use appropriate technology to find a singular value decomposition of A so that $A = U\Sigma V^\mathsf{T}$. Print your entries to two decimal places (but keep as many as possible for computational purposes).

(b) Each singular value tells us how important its semantic dimension is. If we remove the smaller singular values (the less important dimensions), we retain the important information but eliminate minor details and noise. We produce a new term-document matrix A_k by keeping the largest k of the singular values and discarding the rest. This gives us an approximation

$$A_k = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\mathsf{T} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\mathsf{T} + \cdots + \sigma_k \mathbf{u}_k \mathbf{v}_k^\mathsf{T}$$

using the outer product decomposition, where σ_1, σ_2, ..., σ_k are the k largest singular values of A. Note that if A is an $m \times n$ matrix, letting $U_k = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_k]$ (an $m \times k$ matrix), Σ_k the $k \times k$ matrix with the first k singular values along the diagonal, and $V^\mathsf{T} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_k]^\mathsf{T}$ (a $k \times n$ matrix), then we can also write $A_k = U_k \Sigma_k V_k^\mathsf{T}$. This is sometimes referred to as a reduced SVD. Find U_2, Σ_2, and V_2^T, and find the new term-document matrix A_2.

Once we have our term-document matrix, there are three basic comparisons to make: comparing terms, comparing documents, and comparing terms and documents. Term-document matrices are usually very large, with dimension being the number of terms. By using a reduced SVD we can create a much smaller approximation. In our example, the matrix A_k in Project Activity 33.4 reduces our problem to a k-dimensional space. Intuitively, we can think of LSI as representing terms as averages of all of the documents in which they appear and documents as averages of all of the terms they contain. Through this process, LSI attempts to combine the surface information in our library into a deeper abstraction (the "concept" space) that captures the mutual relationships between terms and documents.

We now need to understand how we can represent documents and terms in this smaller space where $A_k = U_k \Sigma_k V_k^\mathsf{T}$. Informally, we can consider the rows of U_k as representing the coordinates of each term in the lower dimensional concept space and the columns of V_k^T as the coordinates of the documents, while the entries of Σ_k tell us how important each semantic dimension is. The dot product of two row vectors of A_k indicates how terms compare across documents. This product is $A_k A_k^\mathsf{T}$. Just as in Project Activity 33.3, we have $A_k A_k^\mathsf{T} = (U_k \Sigma_k)\,(U_k \Sigma_k)^\mathsf{T}$. In other words, if we consider the rows of $U_k \Sigma_k$ as coordinates for terms, then the dot products of these rows give us term to term comparisons. (Note that multiplying U by Σ just stretches the rows of U by the singular values according to the importance of the concept represented by that singular value.) Similarly, the dot product between columns of A provide a comparison of documents. This comparison is given by $A_k^\mathsf{T} A_k = \left(V_k \Sigma_k^\mathsf{T}\right)\left(V_k \Sigma_k^\mathsf{T}\right)^\mathsf{T}$ (again by Project Activity 33.3). So we can consider the rows of $V\Sigma^\mathsf{T}$ as providing coordinates for documents.

Project Activity 33.5. We have seen how to compare terms to terms and documents to documents. The matrix A_k itself compares terms to documents. Show that $A_k = \left(U_k \Sigma_k^{1/2}\right)\left(V_k \Sigma_k^{1/2}\right)^\mathsf{T}$, where $\Sigma_k^{1/2}$ is the diagonal matrix of the same size as Σ_k whose diagonal entries are the square roots of the corresponding diagonal entries in Σ_k. Thus, all useful comparisons of terms and documents can be made using the rows of the matrices U and V, scaled in some way by the singular values in Σ.

To work in this smaller concept space, it is important to be able to find appropriate comparisons to objects that appeared in the original search. For example, to complete the latent structure view of the system, we must also convert the original query to a representation within the new term-document system represented by A_k. This new representation is called a *pseudo-document*.

$$\begin{array}{l} \text{human} \\ \text{interface} \\ \text{computer} \\ \text{user} \\ \text{system} \\ \text{response} \\ \text{time} \\ \text{EPS} \\ \text{survey} \\ \text{trees} \\ \text{graph} \\ \text{minors} \end{array} \begin{bmatrix} -0.22 & -0.11 \\ -0.20 & -0.07 \\ -0.24 & 0.04 \\ -0.40 & 0.06 \\ -0.64 & -0.17 \\ -0.27 & 0.11 \\ -0.27 & 0.11 \\ -0.30 & -0.14 \\ -0.21 & 0.27 \\ -0.01 & 0.49 \\ -0.04 & 0.62 \\ -0.03 & 0.45 \end{bmatrix} \tag{33.4}$$

Terms in the reduced concept space.

$$\begin{array}{cccccccccc} c_1 & c_2 & c_3 & c_4 & c_5 & m_1 & m_2 & m_3 & m_4 \\ \begin{bmatrix} -0.20 & -0.61 & -0.46 & -0.54 & -0.28 & -0.00 & -0.01 & -0.02 & -0.08 \\ -0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53 \end{bmatrix} \end{array} \tag{33.5}$$

Documents in the reduced concept space.

Project Activity 33.6. For an original query $\mathbf{q}$, we start with its term vector $\mathbf{a_q}$ (a vector in the coordinate system determined by the columns of A) and find a representation $\mathbf{v_q}$ that we can use as a column of V^T in the document-document comparison matrix. If this representation was perfect, then it would take a real document in the original system given by A and produce the corresponding column of U if we used the full SVD. In other words, we would have $\mathbf{a_q} = U\Sigma\mathbf{v_q}^\mathsf{T}$.

(a) Use the fact that $A_k = U_k\Sigma_k V_k^\mathsf{T}$, to show that $V_k = A_k^\mathsf{T} U_k \Sigma_k^{-1}$. It follows that $\mathbf{q}$ is transformed into the query $\mathbf{q}_k = \mathbf{q}^\mathsf{T} U_k \Sigma_k^{-1}$.

(b) In our example, using $k = 2$, the terms can now be represented as 2-dimensional vectors (the rows of U_2, see (33.4)), or as points in the plane. More specifically, *human* is represented by the vector (to two decimal places) $[-0.22\ -0.11]^\mathsf{T}$, *interface* by $[-0.20\ -0.07]^\mathsf{T}$, etc. Similarly, the documents are represented by columns of V_2 (see (33.5)), so that the document c_1 is represented by $[-0.20\ -0.06]^\mathsf{T}$, c_2 by $[-0.61\ 0.17]^\mathsf{T}$, etc. From this perspective we can visualize these documents in the plane. Plot the documents and the query in the 2-dimensional concept space. Then calculate the cosine distances from the query to the documents in this space. Which documents now give the best three matches to the query? Compare the matches to your plot.

As we can see from Project Activity 33.6, the original query had no match at all with any documents except c_1, c_2, and m_4. In the new concept space, the query now has some connection to every document,. So LSI has made semantic connections between the terms and documents that were not present in the original term-document matrix, which gives us better results for our search.

Section 34

Approximations Using the Singular Value Decomposition

Focus Questions

By the end of this section, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the section.

- What is the condition number of a matrix and what does it tell us about the matrix?
- What is the pseudoinverse of a matrix?
- Why are pseudoinverses useful?
- How can we find a least squares solution to an equation $A\mathbf{x} = \mathbf{b}$?

Application: Global Positioning System

You are probably familiar with the Global Positioning System (GPS). The system allows anyone with the appropriate software to accurately determine their location at any time. The applications are almost endless, including getting real-time driving directions while in your car, guiding missiles, and providing distances on golf courses.

The GPS is a worldwide radio-navigation system owned by the US government and operated by the US Air Force. GPS is one of four global navigation satellite systems. At least twenty four GPS satellites orbit the Earth at an altitude of approximately 11,000 nautical miles. The satellites are placed so that at any time at least four of them can be accessed by a GPS receiver. Each satellite carries an atomic clock to relay a time stamp along with its position in space. There are five ground stations to coordinate and ensure that the system is working properly.

The system works by triangulation, but there is also error involved in the measurements that go into determining position. Later in this section we will see how the method of least squares can be

used to determine the receiver's position.

Introduction

A singular value decomposition has many applications, and in this section we discuss how a singular value decomposition can be used in image compression, to determine how sensitive a matrix can be to rounding errors in the process of row reduction, and to solve least squares problems.

Image Compression

The digital age has brought many new opportunities for the collection, analysis, and dissemination of information. Along with these opportunities come new difficulties as well. All of this digital information must be stored in some way and be retrievable in an efficient manner. A singular value decomposition of digitally stored information can be used to compress the information or clean up corrupted information. In this section we will see how a singular value decomposition can be used in image compression. While a singular value decomposition is normally used with very large matrices, we will restrict ourselves to small examples so that we can more clearly see how a singular value decomposition is applied.

Preview Activity 34.1. Let $A = \frac{1}{4}\begin{bmatrix} 67 & 29 & -31 & -73 \\ 29 & 67 & -73 & -31 \\ 31 & 73 & -67 & -29 \\ 73 & 31 & -29 & -67 \end{bmatrix}$. A singular value decomposition for A is $U\Sigma V^{\mathsf{T}}$, where

$$U = [\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3\ \mathbf{u}_4] = \frac{1}{2}\begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} 50 & 0 & 0 & 0 \\ 0 & 20 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$V = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3\ \mathbf{v}_4] = \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix}.$$

(1) Write the summands in the corresponding outer product decomposition of A.

(2) The outer product decomposition of A writes A as a sum of rank 1 matrices (the summands $\sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}$).Each summand contains some information about the matrix A. Since σ_1 is the largest of the singular values, it is reasonable to expect that the summand $A_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{T}}$ contains the most information about A among all of the summands. To get a measure of how much information A_1 contains of A, we can think of A as simply a long vector in $\mathbb{R}^{mn}$

where we have folded the data into a rectangular array (we will see later why taking the norm as the norm of the vector in $\mathbb{R}^{nm}$ makes sense, but for now, just use this definition). If we are interested in determining the error in approximating an image by a compressed image, it makes sense to use the standard norm in $\mathbb{R}^{mn}$ to determine length and distance, which is really just the Frobenius norm that comes from the Frobenius inner product defined by

$$\langle U, V \rangle = \sum u_{ij} v_{ij},$$

where $U = [u_{ij}]$ and $V = [v_{ij}]$ are $m \times n$ matrices. So in this section all the norms for matrices will refer to the Frobenius norm. Rather than computing the distance between A_1 and A to measure the error, we are more interested in the relative error

$$\frac{||A - A_1||}{||A||}.$$

(a) Calculate the relative error in approximating A by A_1. What does this tell us about how much information A_1 contains about A?

(b) Let $A_2 = \sum_{k=1}^{2} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}}$. Calculate the relative error in approximating A by A_2. What does this tell us about how much information A_2 contains about A?

(c) Let $A_3 = \sum_{k=1}^{3} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}}$. Calculate the relative error in approximating A by A_3. What does this tell us about how much information A_3 contains about A?

(d) Let $A_4 = \sum_{k=1}^{4} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}}$. Calculate the relative error in approximating A by A_4. What does this tell us about how much information A_4 contains about A? Why?

The first step in compressing an image is to digitize the image. (If you completed the image compression application project in Section 23, then this will seem familiar to you.) There are many ways to do this and we will consider one of the simplest ways and only work with gray-scale images, with the scale from 0 (black) to 255 (white). A digital image can be created by taking a small grid of squares (called pixels) and coloring each pixel with some shade of gray. The resolution of this grid is a measure of how many pixels are used per square inch. As an example, consider the 16 by 16 pixel picture of a flower shown in Figure 34.1.

To store this image pixel by pixel would require $16 \times 16 = 256$ units of storage space (1 for each pixel). If we let M be the matrix whose i, jth entry is the scale of the i, jth pixel, then M is the matrix

$$\begin{bmatrix}
240 & 240 & 240 & 240 & 130 & 130 & 240 & 130 & 130 & 240 & 240 & 240 & 240 & 240 & 240 & 240 \\
240 & 240 & 240 & 130 & 175 & 175 & 130 & 175 & 175 & 130 & 240 & 240 & 240 & 240 & 240 & 240 \\
240 & 240 & 130 & 130 & 175 & 175 & 130 & 175 & 175 & 130 & 130 & 240 & 240 & 240 & 240 & 240 \\
240 & 130 & 175 & 175 & 130 & 175 & 175 & 175 & 130 & 175 & 175 & 130 & 240 & 240 & 240 & 240 \\
240 & 240 & 130 & 175 & 175 & 130 & 175 & 130 & 175 & 175 & 130 & 240 & 240 & 240 & 240 & 240 \\
255 & 240 & 240 & 130 & 130 & 175 & 175 & 175 & 130 & 130 & 240 & 240 & 225 & 240 & 240 & 240 \\
240 & 240 & 130 & 175 & 175 & 130 & 130 & 130 & 175 & 175 & 130 & 240 & 225 & 255 & 240 & 240 \\
240 & 240 & 130 & 175 & 130 & 240 & 130 & 240 & 130 & 175 & 130 & 240 & 255 & 255 & 255 & 240 \\
240 & 240 & 240 & 130 & 240 & 240 & 75 & 240 & 240 & 130 & 240 & 255 & 255 & 255 & 255 & 255 \\
240 & 240 & 240 & 240 & 240 & 240 & 75 & 240 & 240 & 240 & 240 & 240 & 240 & 240 & 240 & 240 \\
240 & 240 & 240 & 75 & 75 & 240 & 75 & 240 & 75 & 75 & 240 & 240 & 240 & 240 & 240 & 240 \\
50 & 240 & 240 & 240 & 75 & 240 & 75 & 240 & 75 & 240 & 240 & 240 & 240 & 50 & 240 & 240 \\
240 & 75 & 240 & 240 & 240 & 75 & 75 & 75 & 240 & 240 & 50 & 240 & 50 & 240 & 240 & 50 \\
240 & 240 & 75 & 240 & 240 & 240 & 75 & 240 & 240 & 50 & 240 & 50 & 240 & 240 & 50 & 240 \\
75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 \\
75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75 & 75
\end{bmatrix}.$$

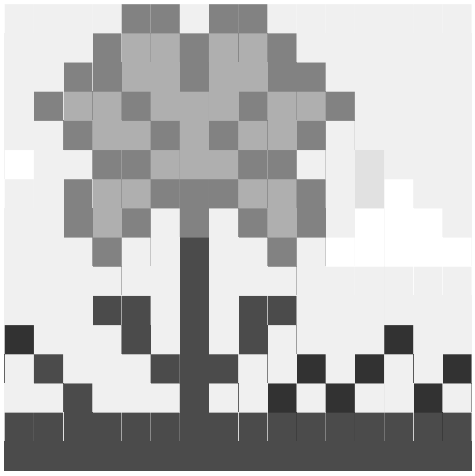

Figure 34.1: A 16 by 16 pixel image

Recall that if $U\Sigma V^{\mathsf{T}}$ is a singular value decomposition for M, then we can also write M in the form

$$M = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{T}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{T}} + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^{\mathsf{T}} + \cdots + \sigma_{16} \mathbf{u}_{16} \mathbf{v}_{16}^{\mathsf{T}}.$$

given in (33.2). For this M, the singular values are approximately

$$\begin{bmatrix} 3006.770088367795 \\ 439.13109000200205 \\ 382.1756550649652 \\ 312.1181752764884 \\ 254.45105800344953 \\ 203.36470770057494 \\ 152.8696215072527 \\ 101.29084240890717 \\ 63.80803769229468 \\ 39.6189181773536 \\ 17.091891798245463 \\ 12.304589419140656 \\ 4.729898943556077 \\ 2.828719409809012 \\ 6.94442317024232 \times 10^{-15} \\ 2.19689952047833 \times 10^{-15} \end{bmatrix}. \tag{34.1}$$

Notice that some of these singular values are very small compared to others. As in Preview Activity 34.1, the terms with the largest singular values contain most of the information about the matrix. Thus, we shouldn't lose much information if we eliminate the small singular values. In this particular example, the last 4 singular values are significantly smaller than the rest. If we let

$$M_{12} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{T}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{T}} + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^{\mathsf{T}} + \cdots + \sigma_{12} \mathbf{u}_{12} \mathbf{v}_{12}^{\mathsf{T}},$$

then we should expect the image determined by M_{12} to be close to the image made by M. The two images are presented side by side in Figure 34.2.

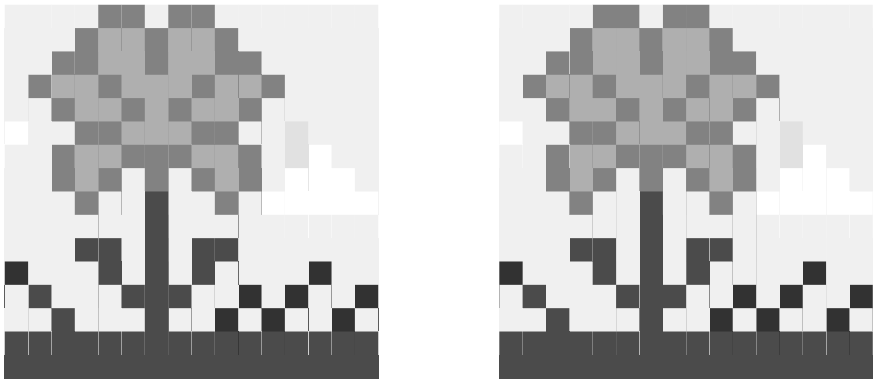

Figure 34.2: A 16 by 16 pixel image and a compressed image using a singular value decomposition.

This small example illustrates the general idea. Suppose we had a satellite image that was 1000×1000 pixels and we let M represent this image. If we have a singular value decomposition of this image M, say

$$M = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{T}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{T}} + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^{\mathsf{T}} + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^{\mathsf{T}},$$

if the rank of M is large, it is likely that many of the singular values will be very small. If we only keep s of the singular values, we can approximate M by

$$M_s = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{T}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{T}} + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^{\mathsf{T}} + \cdots + \sigma_s \mathbf{u}_s \mathbf{v}_s^{\mathsf{T}}$$

and store the image with only the vectors $\sigma_1 \mathbf{u}_1$, $\sigma_2 \mathbf{u}_2$, ..., $\sigma_s \mathbf{u}_s$, $\mathbf{v}_1$, $\mathbf{v}_1$, ..., $\mathbf{v}_s$. For example, if we only need 10 of the singular values of a satellite image ($s = 10$), then we can store the satellite image with only 20 vectors in $\mathbb{R}^{1000}$ or with $20 \times 1000 = 20,000$ numbers instead of $1000 \times 1000 = 1,000,000$ numbers.

A similar process can be used to denoise data.[1]

Calculating the Error in Approximating an Image

In the context where a matrix represents an image, the operator aspect of the matrix is irrelevant – we are only interested in the matrix as a holder of information. In this situation, we think of an $m \times n$ matrix as simply a long vector in $\mathbb{R}^{mn}$ where we have folded the data into a rectangular array. If we are interested in determining the error in approximating an image by a compressed image, it makes sense to use the standard norm in $\mathbb{R}^{mn}$ to determine length and distance. This leads to what is called the *Frobenius* norm of a matrix. The Frobenius norm $||M||_F$ of an $m \times n$ matrix $M = [m_{ij}]$ is

$$||M||_F = \sqrt{\sum m_{ij}^2}.$$

[1] For example, as stated in http://www2.imm.dtu.dk/~pch/Projekter/tsvd.html, "The SVD [singular value decomposition] has also applications in digital signal processing, e.g., as a method for noise reduction. The central idea is to let a matrix A represent the noisy signal, compute the SVD, and then discard small singular values of A. It can be shown that the small singular values mainly represent the noise, and thus the rank-k matrix A_k represents a filtered signal with less noise."

There is a natural corresponding inner product on the set of $m \times n$ matrices (called the *Frobenius product*) defined by

$$\langle A, B \rangle = \sum a_{ij} b_{ij},$$

where $A = [a_{ij}]$ and $B = [b_{ij}]$ are $m \times n$ matrices.[2] Note that

$$||A||_F = \sqrt{\langle A, A \rangle}.$$

If an $m \times n$ matrix M of rank r has a singular value decomposition $M = U\Sigma V^{\mathsf{T}}$, we have seen that we can write M as an outer product

$$M = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{T}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{T}} + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^{\mathsf{T}} + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^{\mathsf{T}}, \tag{34.2}$$

where the $\mathbf{u}_i$ are the columns of U and the $\mathbf{v}_j$ the columns of V. Each of the products $\mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}$ is an $m \times n$ matrix. Since the columns of $\mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}$ are all scalar multiples of $\mathbf{u}_i$, the matrix $\mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}$ is a rank 1 matrix. So (34.2) expresses M as a sum of rank 1 matrices. Moreover, if we let $\mathbf{x}$ and $\mathbf{w}$ be $m \times 1$ vectors and let $\mathbf{y}$ and $\mathbf{z}$ be $n \times 1$ vectors with $\mathbf{y} = [y_1 \ y_2 \ \ldots \ y_n]^{\mathsf{T}}$ and $\mathbf{z} = [z_1 \ z_2 \ \ldots \ z_n]^{\mathsf{T}}$, then

$$\begin{aligned}
\langle \mathbf{x}\mathbf{y}^{\mathsf{T}}, \mathbf{w}\mathbf{z}^{\mathsf{T}} \rangle &= \langle [y_1\mathbf{x} \ y_2\mathbf{x} \ \cdots \ y_n\mathbf{x}], [z_1\mathbf{w} \ z_2\mathbf{w} \ \cdots \ z_n\mathbf{w}] \rangle \\
&= \sum (y_i \mathbf{x}) \cdot (z_i \mathbf{w}) \\
&= \sum (y_i z_i)(\mathbf{x} \cdot \mathbf{w}) \\
&= (\mathbf{x} \cdot \mathbf{w}) \sum (y_i z_i) \\
&= (\mathbf{x} \cdot \mathbf{w})(\mathbf{y} \cdot \mathbf{z}).
\end{aligned}$$

Using the vectors from the singular value decomposition of M as in (34.2) we see that

$$\langle \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}, \mathbf{u}_j \mathbf{v}_j^{\mathsf{T}} \rangle = (\mathbf{u}_i \cdot \mathbf{u}_j)(\mathbf{v}_i \cdot \mathbf{v}_j) = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

It follows that

$$||M||_F^2 = \sum \sigma_i^2 (\mathbf{u}_i \cdot \mathbf{u}_i)(\mathbf{v}_i \cdot \mathbf{v}_i) = \sum \sigma_i^2. \tag{34.3}$$

Activity 34.1. Verify (34.3) that $||M||_F^2 = \sum \sigma_i^2$.

When we used the singular value decomposition to approximate the image defined by M, we replaced M with a matrix of the form

$$M_k = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{T}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{T}} + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^{\mathsf{T}} + \cdots + \sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}}. \tag{34.4}$$

We call M_k the rank k approximation to M. Notice that the outer product expansion in (34.4) is in fact a singular value decomposition for M_k. The error E_k in approximating M with M_k is

$$E_k = M - M_k = \sigma_{k+1} \mathbf{u}_{k+1} \mathbf{v}_{k+1}^{\mathsf{T}} + \sigma_{k+2} \mathbf{u}_{k+2} \mathbf{v}_{k+2}^{\mathsf{T}} + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^{\mathsf{T}}. \tag{34.5}$$

[2] This is the same inner product that we defined as the Frobenius inner product in Section 29, where $\langle A, B \rangle = \text{trace}\left(AB^{\mathsf{T}}\right)$.

Once again, notice that (34.5) is a singular value decomposition for E_k. We define the relative error in approximating M with M_k as

$$\frac{||E_k||}{||M||}.$$

Now (34.3) shows that

$$\frac{||E_k||}{||M||} = \sqrt{\frac{\sum_{i=k+1}^{r} \sigma_i^2}{\sum_{i=1}^{r} \sigma_i^2}}.$$

In applications, we often want to retain a certain degree of accuracy in our approximations and this error term can help us accomplish that.

In our flower example, the singular values of M are given in (34.1). The relative error in approximating M with M_{12} is

$$\sqrt{\frac{\sum_{i=13}^{16} \sigma_i^2}{\sum_{i=1}^{16} \sigma_i^2}} \approx 0.03890987666.$$

Errors (rounded to 4 decimal places) for approximating M with some of the M_k are shown in Table 34.1

k	10	9	8	7	6
$\frac{\|E_k\|}{\|M\|}$	0.0860	0.1238	0.1677	0.2200	0.2811
k	5	4	3	2	1
$\frac{\|E_k\|}{\|M\|}$	0.3461	0.4132	0.4830	0.5566	0.6307

Table 34.1: Errors in approximating M by M_k

Activity 34.2. Let M represent the flower image.

(a) Find the relative errors in approximating M by M_{13} and M_{14}. You can use the fact that $\sum_{i=1}^{16} \sigma_i^2 \approx 4992.553293$.

(b) About how much of the information in the image is contained in the rank 1 approximation? Explain.

The Condition Number of a Matrix

A singular value decomposition for a matrix A can tell us a lot about how difficult it is to accurately solve a system $A\mathbf{x} = \mathbf{b}$. Solutions to systems of linear equations can be very sensitive to rounding as the next exercise demonstrates.

Activity 34.3. Find the solution to each of the systems.

(a) $\begin{bmatrix} 1.0000 & 1.0000 \\ 1.0000 & 1.0005 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2.0000 \\ 2.0050 \end{bmatrix}$

(b) $\begin{bmatrix} 1.000 & 1.000 \\ 1.000 & 1.001 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2.000 \\ 2.005 \end{bmatrix}$

Notice that a simple rounding in the $(2, 2)$ entry of the coefficient matrix led to a significantly different solution. If there are rounding errors at any stage of the Gaussian elimination process, they can be compounded by further row operations. This is an important problem since computers can only approximate irrational numbers with rational numbers and so rounding can be critical. Finding ways of dealing with these kinds of errors is an area of on-going research in numerical linear algebra. This problem is given a name.

Definition 34.1. A matrix A is **ill-conditioned** if relatively small changes in any entries of A can produce significant changes in solutions to the system $A\mathbf{x} = \mathbf{b}$.

A matrix that is not ill-conditioned is said to be *well-conditioned*. Since small changes in entries of ill-conditioned matrices can lead to large errors in computations, it is an important problem in linear algebra to have a way to measure how ill-conditioned a matrix is. This idea will ultimately lead us to the condition number of a matrix.

Suppose we want to solve the system $A\mathbf{x} = \mathbf{b}$, where A is an invertible matrix. Activity 34.3 illustrates that if A is really close to being singular, then small changes in the entries of A can have significant effects on the solution to the system. So the system can be very hard to solve accurately if A is close to singular. It is important to have a sense of how "good" we can expect any calculated solution to be. Suppose we think we solve the system $A\mathbf{x} = \mathbf{b}$ but, through rounding error in our calculation of A, get a solution $\mathbf{x}'$ so that $A\mathbf{x}' = \mathbf{b}'$, where $\mathbf{b}'$ is not exactly $\mathbf{b}$. Let $\Delta\mathbf{x}$ be the error in our calculated solution and $\Delta\mathbf{b}$ the difference between $\mathbf{b}'$ and $\mathbf{b}$. We would like to know how large the error $||\Delta\mathbf{x}||$ can be. But this isn't exactly the right question. We could scale everything to make $||\Delta\mathbf{x}||$ as large as we want. What we really need is a measure of the *relative error* $\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||}$, or how big the error is compared to $||\mathbf{x}||$ itself. More specifically, we want to know how large the relative error in $\Delta\mathbf{x}$ is compared to the relative error in $\Delta\mathbf{b}$. In other words, we want to know how good the relative error in $\Delta\mathbf{b}$ is as a predictor of the relative error in $\Delta\mathbf{x}$ (we may have some control over the relative error in $\Delta\mathbf{b}$, perhaps by keeping more significant digits). So we want know if there is a best constant C such that

$$\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||} \leq C\frac{||\Delta\mathbf{b}||}{||\mathbf{b}||}.$$

This best constant C is the condition number – a measure of how well the relative error in $\Delta\mathbf{b}$ predicts the relative error in $\Delta\mathbf{x}$. How can we find C?

Since $A\mathbf{x}' = \mathbf{b}'$ we have

$$A(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}.$$

Distributing on the left and using the fact that $A\mathbf{x} = \mathbf{b}$ gives us

$$A\Delta\mathbf{x} = \Delta\mathbf{b}.$$

We return for a moment to the operator norm of a matrix. This is an appropriate norm to use here since we are considering A to be a transformation. Recall that if A is an $m \times n$ matrix, we defined the operator norm of A to be

$$||A|| = \max_{||\mathbf{x}||\neq\mathbf{0}} \left\{ \frac{||A\mathbf{x}||}{||\mathbf{x}||} \right\} = \max_{||\mathbf{x}||=1} \{||A\mathbf{x}||\}.$$

One important property that the norm has is that if the product AB is defined, then

$$||AB|| \leq ||A||\,||B||.$$

To see why, notice that

$$\frac{||AB\mathbf{x}||}{||\mathbf{x}||} = \frac{||A(B\mathbf{x})||}{||B\mathbf{x}||}\,\frac{||B\mathbf{x}||}{||\mathbf{x}||}.$$

Now $\frac{||A(B\mathbf{x})||}{||B\mathbf{x}||} \leq ||A||$ and $\frac{||B\mathbf{x}||}{||\mathbf{x}||} \leq ||B||$ by the definition of the norm, so we conclude that

$$\frac{||AB\mathbf{x}||}{||\mathbf{x}||} \leq ||A||\,||B||$$

for every $\mathbf{x}$. Thus,

$$||AB|| \leq ||A||\,||B||.$$

Now we can find the condition number. From $A\Delta\mathbf{x} = \Delta\mathbf{b}$ we have

$$\Delta\mathbf{x} = A^{-1}\Delta\mathbf{b},$$

so

$$||\Delta\mathbf{x}|| \leq ||A^{-1}||\,||\Delta\mathbf{b}||. \tag{34.6}$$

Similarly, $\mathbf{b} = A\mathbf{x}$ implies that $||\mathbf{b}|| \leq ||A||\,||\mathbf{x}||$ or

$$\frac{1}{||\mathbf{x}||} \leq \frac{||A||}{||\mathbf{b}||}. \tag{34.7}$$

Combining (34.6) and (34.7) gives

$$\begin{aligned}\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||} &\leq \frac{||A^{-1}||\,||\Delta\mathbf{b}||}{||\mathbf{x}||}\\ &= ||A^{-1}||\,||\Delta\mathbf{b}||\left(\frac{1}{||\mathbf{x}||}\right)\\ &\leq ||A^{-1}||\,||\Delta\mathbf{b}||\frac{||A||}{||\mathbf{b}||}\\ &= ||A^{-1}||\,||A||\,\frac{||\Delta\mathbf{b}||}{||\mathbf{b}||}.\end{aligned}$$

This constant $||A^{-1}||\,||A||$ is the best bound and so is called the condition number of A.

Definition 34.2. The **condition number** of an invertible matrix A is the number $||A^{-1}||\,||A||$.

How does a singular value decomposition tell us about the condition number of a matrix? Recall that the maximum value of $||A\mathbf{x}||$ for $\mathbf{x}$ on the unit n-sphere is σ_1. So $||A|| = \sigma_1$. If A is an invertible matrix and $A = U\Sigma V^{\mathsf{T}}$ is a singular value decomposition for A, then

$$A^{-1} = (U\Sigma V^{\mathsf{T}})^{-1} = (V^{\mathsf{T}})^{-1}\Sigma^{-1}U^{-1} = V\Sigma^{-1}U^{\mathsf{T}},$$

where

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1} & & & & 0 \\ & \frac{1}{\sigma_2} & & & \\ & & \frac{1}{\sigma_3} & & \\ & & & \ddots & \\ 0 & & & & \frac{1}{\sigma_n} \end{bmatrix}.$$

Now $V\Sigma^{-1}U^{\mathsf{T}}$ is a singular value decomposition for A^{-1} with the diagonal entries in reverse order, so

$$||A^{-1}|| = \frac{1}{\sigma_n}.$$

Therefore, the condition number of A is

$$||A^{-1}||\,||A|| = \frac{\sigma_1}{\sigma_n}.$$

Activity 34.4. Let $A = \begin{bmatrix} 1.0000 & 1.0000 \\ 1.0000 & 1.0005 \end{bmatrix}$. A computer algebra system gives the singular values of A as 2.00025003124999934 and 0.000249968750000509660. What is the condition number of A. What does that tell us about A? Does this seem reasonable given the result of Activity 34.3?

Activity 34.5.

(a) What is the smallest the condition number of a matrix can be? Find an entire class of matrices with this smallest condition number.

(b) What is the condition number of an orthogonal matrix? Why does this make sense? (Hint: If P is an orthogonal matrix, what is $||P\mathbf{x}||$ for any vector $\mathbf{x}$? What does this make $||P||$?)

(c) What is the condition number of an invertible symmetric matrix in terms of its eigenvalues?

(d) Why do we not define the condition number of a non-invertible matrix? If we did, what would the condition number have to be? Why?

Pseudoinverses

Not every matrix is invertible, so we cannot always solve a matrix equation $A\mathbf{x} = \mathbf{b}$. However, every matrix has a pseudoinverse A^+ that acts something like an inverse. Even when we can't solve a matrix equation $A\mathbf{x} = \mathbf{b}$ because $\mathbf{b}$ isn't in Col A, we can use the pseudoinverse of A to "solve" the equation $A\mathbf{x} = \mathbf{b}$ with the "solution" $A^+\mathbf{b}$. While not an exact solution, $A^+\mathbf{b}$ turns out to be the best approximation to a solution in the least squares sense. In this use the singular value decomposition to find the pseudoinverse of a matrix.

Preview Activity 34.2. Let $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$. The singular value decomposition of A is $U\Sigma V^{\mathsf{T}}$

where

$$U = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix},$$
$$\Sigma = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$
$$V = \frac{1}{6}\begin{bmatrix} \sqrt{6} & -3\sqrt{2} & 2\sqrt{3} \\ 2\sqrt{6} & 0 & -2\sqrt{3} \\ \sqrt{6} & 3\sqrt{2} & 2\sqrt{3} \end{bmatrix}.$$

(1) Explain why A is not an invertible matrix.

(2) Explain why the matrices U and V are invertible. How are U^{-1} and V^{-1} related to U^{T} and V^{T}?

(3) Recall that one property of invertible matrices is that the inverse of a product of invertible matrices is the product of the inverses in the reverse order. If A were invertible, then A^{-1} would be $\left(U\Sigma V^{\mathsf{T}}\right)^{-1} = V\Sigma^{-1}U^{\mathsf{T}}$. Even though U and V are invertible, the matrix Σ is not. But Σ does contain non-zero eigenvalues that have reciprocals, so consider the matrix $\Sigma^{+} = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$. Calculate the products $\Sigma\Sigma^{+}$ and $\Sigma^{+}\Sigma$. How are the results similar to that obtained with a matrix inverse?

(4) The only matrix in the singular value decomposition of A that is not invertible is Σ. But the matrix Σ^{+} acts somewhat like an inverse of Σ, so let us define A^{+} as $V\Sigma^{+}U^{\mathsf{T}}$. Now we explore a few properties of the matrix A^{+}.

   (a) Calculate AA^{+} and $A^{+}A$ for $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$. What do you notice?

   (b) Calculate $A^{+}AA^{+}$ and $AA^{+}A$ for $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$. What do you notice?

Only some square matrices have inverses. However, every matrix has a pseudoinverse. A pseudoinverse A^{+} of a matrix A provides something like an inverse when a matrix doesn't have an inverse. Pseudoinverses are useful to approximate solutions to linear systems. If A is invertible, then the equation $A\mathbf{x} = \mathbf{b}$ has the solution $\mathbf{x} = A^{-1}\mathbf{b}$, but when A is not invertible and $\mathbf{b}$ is not in Col A, then the equation $A\mathbf{x} = \mathbf{b}$ has no solution. In the invertible case of an $n \times n$ matrix A, there is a matrix B so that $AB = BA = I_n$. This also implies that $BAB = B$ and $ABA = A$. To mimic this situation when A is not invertible, we search for a matrix A^{+} (a pseudoinverse of A) so that $AA^{+}A = A$ and $A^{+}AA^{+} = A^{+}$, as we saw in Preview Activity 34.2. Then it turns out that A^{+} acts something like an inverse for A. In this case, we approximate the solution to $A\mathbf{x} = \mathbf{b}$ by $\mathbf{x}^{*} = A^{+}\mathbf{b}$, and we will see that the vector $A\mathbf{x}^{*} = AA^{+}\mathbf{b}$ turns out to be the vector in Col A that is closest to $\mathbf{b}$ in the least squares sense.

A reasonable question to ask is how we can find a pseudoinverse of a matrix A. A singular value decomposition provides an answer to this question. If A is an invertible $n \times n$ matrix, then 0

is not an eigenvalue of A. As a result, in the singular value decomposition $U\Sigma V^{\mathsf{T}}$ of A, the matrix Σ is an invertible matrix (note that U, Σ, and V are all $n \times n$ matrices in this case). So

$$A^{-1} = \left(U\Sigma V^{\mathsf{T}}\right)^{-1} = V\Sigma^{-1}U^{\mathsf{T}},$$

where

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & 0 & \\ & & \frac{1}{\sigma_3} & & \\ & 0 & & \ddots & \\ & & & & \frac{1}{\sigma_n} \end{bmatrix}.$$

In this case, $V\Sigma^{-1}U^{\mathsf{T}}$ is a singular value decomposition for A^{-1}.

To understand in general how a pseudoinverse is found, let A be an $m \times n$ matrix with $m \neq n$, or an $n \times n$ with rank less than n. In these cases A does not have an inverse. But as in Preview Activity 34.2, a singular value decomposition provides a pseudoinverse A^+ for A. Let $U\Sigma V^{\mathsf{T}}$ be a singular value decomposition of an $m \times n$ matrix A of rank r, with

$$\Sigma = \left[\begin{array}{cccc|c} \sigma_1 & & & & \\ & \sigma_2 & & 0 & 0 \\ & & \sigma_3 & & \\ & 0 & & \ddots & \\ & & & \sigma_r & \\ \hline & & 0 & & 0 \end{array}\right]$$

The matrices U and V are invertible, but the matrix Σ is not if A is not invertible. If we let Σ^+ be the $n \times m$ matrix defined by

$$\Sigma^+ = \left[\begin{array}{cccc|c} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & 0 & 0 \\ & & \frac{1}{\sigma_3} & & \\ & 0 & & \ddots & \\ & & & \frac{1}{\sigma_r} & \\ \hline & & 0 & & 0 \end{array}\right],$$

then Σ^+ will act much like an inverse of Σ might. In fact, it is not difficult to see that

$$\Sigma\Sigma^+ = \left[\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right] \text{ and } \Sigma^+\Sigma = \left[\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right],$$

where $\Sigma\Sigma^+$ is an $m \times m$ matrix and $\Sigma^+\Sigma$ is an $n \times n$ matrix.

The matrix

$$A^+ = V\Sigma^+U^{\mathsf{T}} \tag{34.8}$$

is a *pseudoinverse* of A.

Activity 34.6.

(a) Find the pseudoinverse A^+ of $A = \begin{bmatrix} 0 & 5 \\ 4 & 3 \\ -2 & 1 \end{bmatrix}$. Use the singular value decomposition $U\Sigma V^\mathsf{T}$ of A, where

$$U = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} & 1 \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{3}}{3} & 0 \\ 0 & \frac{\sqrt{3}}{3} & 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{15} \\ 0 & 0 \end{bmatrix}, V = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}.$$

(b) The vector $\mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ is not in Col A. The vector $\mathbf{x}^* = A^+\mathbf{b}$ is an approximation to a solution of $A\mathbf{x} = \mathbf{b}$, and $AA^+\mathbf{b}$ is in Col A. Find $A\mathbf{x}^*$ and determine how far $A\mathbf{x}^*$ is from $\mathbf{b}$.

Pseudoinverses satisfy several properties that are similar to those of inverses. For example, we had an example in Preview Activity 34.2 where $AA^+A = A$ and $A^+AA^+ = A^+$. That A^+ always satisfies these properties is the subject of the next activity.

Activity 34.7. Let A be an $m \times n$ matrix with singular value decomposition $U\Sigma V^\mathsf{T}$. Let A^+ be defined as in (34.8).

(a) Show that $AA^+A = A$.

(b) Show that $A^+AA^+ = A^+$.

Activity 34.7 shows that A^+ satisfies properties that are similar to those of an inverse of A. In fact, A^+ satisfies several other properties (that together can be used as defining properties) as stated in the next theorem. The conditions of Theorem 34.3 are called the *Penrose* or *Moore-Penrose* conditions.[3] Verification of the remaining parts of this theorem are left for the exercises.

Theorem 34.3 (The Moore-Penrose Conditions.)**.** *A pseudoinverse of a matrix A is a matrix A^+ that satisfies the following properties.*

(1) $AA^+A = A$

(2) $A^+AA^+ = A^+$

(3) $(AA^+)^\mathsf{T} = AA^+$

(4) $(A^+A)^\mathsf{T} = A^+A$

Also, there is a unique matrix A^+ that satisfies these properties. The verification of this property is left to the exercises.

[3] Theorem 34.3 is often given as the definition of a pseudoinverse.

Least Squares Approximations

In many situations we want to be able to make predictions from data. However, data is rarely well-behaved and so we need to use approximation techniques to estimate from the data.

Preview Activity 34.3. NBC was awarded the U.S. television broadcast rights to the 2016 and 2020 summer Olympic games. Table 34.2 lists the amounts paid (in millions of dollars) by NBC sports for the 2008 through 2012 summer Olympics plus the recently concluded bidding for the 2016 and 2020 Olympics, where year 0 is the year 2008. Figure 34.3 shows a plot of the data. Our goal in this activity is to find a linear function f defined by $f(x) = a_0 + a_1 x$ that fits the data well.

Year	Amount
0	894
4	1180
8	1226
12	1418

Table 34.2: Olympics data.

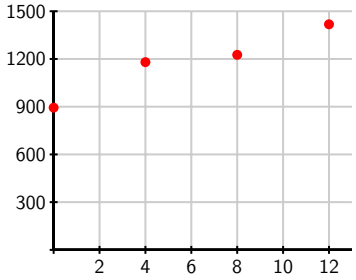

Figure 34.3: A plot of the data.

If the data were actually linear, then the data would satisfy the system

$$\begin{aligned} a_0 + \ 0a_1 &= \ 894 \\ a_0 + \ 4a_1 &= 1180 \\ a_0 + \ 8a_1 &= 1226 \\ a_0 + 12a_1 &= 1418. \end{aligned}$$

In matrix form this system is $A\mathbf{x} = \mathbf{b}$, where $A = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 12 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$, and $\mathbf{b} = \begin{bmatrix} 894 \\ 1180 \\ 1226 \\ 1418 \end{bmatrix}$.

Assume that a singular value decomposition of A (with entries rounded to 4 decimal places) is $U\Sigma V^{\mathsf{T}}$, where

$$U \approx \begin{bmatrix} 0.0071 & -0.8366 & 0.2236 & 0.5000 \\ 0.2713 & -0.4758 & -0.6708 & -0.5000 \\ 0.5355 & -0.1150 & 0.6708 & -0.5000 \\ 0.7997 & 0.2459 & -0.2236 & 0.5000 \end{bmatrix},$$

$$\Sigma \approx \begin{bmatrix} 15.0528 & 0 \\ 0 & 1.1884 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$V \approx \begin{bmatrix} 0.1072 & -0.9942 \\ 0.9942 & 0.1072 \end{bmatrix}.$$

(1) Explain why A is not an invertible matrix. Find a pseudoinverse A^+ of A and calculate $A^+\mathbf{b}$. Round calculations to four decimal places.

(2) Use $A^+\mathbf{b}$ to find the linear approximation $f(x) = a_0 + a_1x$. Plot your linear approximation against the data in Figure 34.3.

Preview Activity 34.3 illustrates how $A^+\mathbf{b}$ can be used as an approximation to a solution to the equation $A\mathbf{x} = \mathbf{b}$ when $\mathbf{b}$ is not in Col A. Now we will examine what kind of approximation $A^+\mathbf{b}$ actually is.

Let $U\Sigma V^\mathsf{T}$ be a singular value decomposition for an $m \times n$ matrix A of rank r. Then the columns of

$$U = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_m]$$

form an orthonormal basis for $\mathbb{R}^m$ and $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ is a basis for Col A. Remember from Section 30 that if $\mathbf{b}$ is any vector in $\mathbb{R}^m$, then

$$\text{proj}_{\text{Col } A}\mathbf{b} = (\mathbf{b}\cdot\mathbf{u}_1)\mathbf{u}_1 + (\mathbf{b}\cdot\mathbf{u}_2)\mathbf{u}_2 + \cdots + (\mathbf{b}\cdot\mathbf{u}_r)\mathbf{u}_r$$

is the least squares approximation of the vector $\mathbf{b}$ by a vector in Col A. We can extend this sum to all of columns of U as

$$\text{proj}_{\text{Col } A}\mathbf{b} = (\mathbf{b}\cdot\mathbf{u}_1)\mathbf{u}_1 + (\mathbf{b}\cdot\mathbf{u}_2)\mathbf{u}_2 + \cdots + (\mathbf{b}\cdot\mathbf{u}_r)\mathbf{u}_r + 0\mathbf{u}_{r+1} + 0\mathbf{u}_{r+2} + \cdots + 0\mathbf{u}_m.$$

It follows that

$$\begin{aligned}
\text{proj}_{\text{Col } A}\mathbf{b} &= \sum_{i=1}^{r} \mathbf{u}_i(\mathbf{u}_i\cdot\mathbf{b}) \\
&= \sum_{i=1}^{r} \mathbf{u}_i(\mathbf{u}_i^\mathsf{T}\mathbf{b}) \\
&= \sum_{i=1}^{r} (\mathbf{u}_i\mathbf{u}_i^\mathsf{T})\mathbf{b} \\
&= \left(\sum_{i=1}^{r}(1)(\mathbf{u}_i\mathbf{u}_i^\mathsf{T})\right)\mathbf{b} + \left(\sum_{i=r+1}^{m} 0(\mathbf{u}_i\mathbf{u}_i^\mathsf{T})\right)\mathbf{b} \\
&= (UDU^\mathsf{T})\mathbf{b},
\end{aligned}$$

where

$$D = \left[\begin{array}{c|c} I_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right].$$

Now, if $\mathbf{z} = A^+\mathbf{b}$, then

$$A\mathbf{z} = (U\Sigma V^\mathsf{T})(V\Sigma^+U^\mathsf{T}\mathbf{b}) = (U\Sigma\Sigma^+U^\mathsf{T})\mathbf{b} = (UDU^\mathsf{T})\mathbf{b} = \text{proj}_{\text{Col } A}\mathbf{b},$$

and hence the vector $A\mathbf{z} = AA^+\mathbf{b}$ is the vector $A\mathbf{x}$ in Col A that minimizes $||A\mathbf{x} - \mathbf{b}||$. Thus, $A\mathbf{z}$ is in actuality the least squares approximation to $\mathbf{b}$. So a singular value decomposition allows us to construct the pseudoinverse of a matrix A and then directly solve the least squares problem.

Activity 34.8. Having to calculate eigenvalues and eigenvectors for a matrix to produce a singular value decomposition to find pseudoinverse can be computationally intense. As we demonstrate in this activity, the process is easier if the columns of A are linearly independent. More specifically, we will prove the following theorem.

Theorem 34.4. *If the columns of a matrix A are linearly independent, then* $A^+ = \left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}$.

To see how, suppose that A is an $m \times n$ matrix with linearly independent columns.

(a) Given that the columns of A are linearly independent, what must be the relationship between n and m?

(b) Since the columns of A are linearly independent, it follows that $A^\mathsf{T}A$ is invertible (see Exercise 13.). So the eigenvalues of A are all non-zero. Let σ_1, σ_2, ..., σ_r be the singular values of A. How is r related to n, and what do Σ and Σ^+ look like?

(c) Let us now investigate the form of the invertible matrix $A^\mathsf{T}A$ (note that neither A nor A^T is necessarily invertible). If a singular value decomposition of A is $U\Sigma V^\mathsf{T}$, show that

$$A^\mathsf{T}A = V\Sigma^\mathsf{T}\Sigma V^\mathsf{T}.$$

(d) Let $\lambda_i = \sigma_i^2$ for i from 1 to n. It is straightforward to see that $\Sigma^\mathsf{T}\Sigma$ is an $n \times n$ diagonal matrix D, where

$$D = \Sigma^\mathsf{T}\Sigma = \begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & 0 & \\ & & \lambda_3 & & \\ & & & \ddots & \\ & 0 & & & \lambda_n \end{bmatrix}.$$

Then $(A^\mathsf{T}A)^{-1} = VD^{-1}V^\mathsf{T}$. Recall that $A^+ = V\Sigma^+U^\mathsf{T}$, so to relate $A^\mathsf{T}A$ to A^+ we need a product that is equal to Σ^+. Explain why

$$D^{-1}\Sigma^\mathsf{T} = \Sigma^+.$$

(e) Complete the activity by showing that

$$\left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T} = A^+.$$

Therefore, to calculate A^+ and solve a least squares problem, Theorem 34.4 shows that as long as the columns of A are linearly independent, we can avoid using a singular value decomposition of A in finding A^+. As an example, if $A = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 12 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 894 \\ 1180 \\ 1226 \\ 1418 \end{bmatrix}$ as in Preview Activity 34.3, then

$$\left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T} \approx \begin{bmatrix} 0.7000 & 0.4000 & 0.1000 & -0.2000 \\ -0.0750 & -0.0250 & 0.0250 & 0.0750 \end{bmatrix}$$

and

$$\left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}\mathbf{b} \approx \begin{bmatrix} 936.8000 \\ 40.4500 \end{bmatrix}.$$

So $f(x) = 936.8 + 40.45x$ is (to 4 decimal places) the least squares linear approximation to the data. A graph of f versus the data is shown in Figure 34.4. By least squares we mean that $f(x)$ approximates the data so that the sum of the squares of the vertical distances between the data and corresponding values of f (as illustrated in Figure 34.4) is as small as possible.

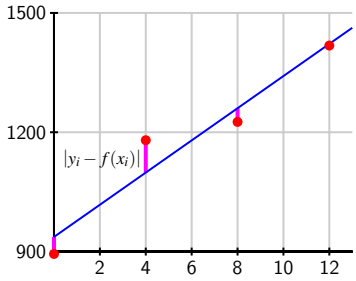

Figure 34.4: Linear approximation plotted against the data set.

Activity 34.9. Suppose we wanted to fit a quadratic or a cubic polynomial to a set of data. Preview Activity 34.3 showed us how to fit a line to data and we can extend that idea to any degree polynomial. To fit a polynomial $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 + a_0$ of degree n to m data points (x_1, y_1), (x_2, y_2), …, (x_m, y_m), no two of which have the same x coordinate in the least squares sense, we want to find the least squares approximate solution to the system .

$$\begin{aligned}
y_1 &= a_0 + x_1 a_1 + x_1^2 a_2 + \cdots + x_1^{n-1} a_{n-1} + x_1^n a_n \\
y_2 &= a_0 + x_2 a_1 + x_2^2 a_2 + \cdots + x_2^{n-1} a_{n-1} + x_2^n a_n \\
y_3 &= a_0 + x_3 a_1 + x_3^2 a_2 + \cdots + x_3^{n-1} a_{n-1} + x_3^n a_n \\
&\vdots \\
y_m &= a_0 + x_m a_1 + x_m^2 a_2 + \cdots + x_m^{n-1} a_{n-1} + x_m^n a_n.
\end{aligned}$$

of m equations in the $n+1$ unknowns a_0, a_1, …, a_{n-1}, and a_n. In matrix form we can write this system as $M\mathbf{a} = \mathbf{y}$, where

$$M = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} & x_2^n \\ 1 & x_3 & x_3^2 & \cdots & x_3^{n-1} & x_3^n \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^{n-1} & x_m^n \end{bmatrix}$$

while the vectors $\mathbf{a}$ and $\mathbf{y}$ are

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{m-1} \\ y_m \end{bmatrix}.$$

(a) Set up the matrix equation to fit a quadratic to the Olympics data in Table Table 34.2.

(b) Use Theorem 34.4 and appropriate technology to find the least squares quadratic polynomial for the Olympics data. Draw your approximation against the data as shown in Figure 34.3. Round approximations to 4 decimal places.

Examples

What follows are worked examples that use the concepts from this section.

Example 34.5. Let

$$A = \begin{bmatrix} 2 & 5 & 4 \\ 6 & 3 & 0 \\ 6 & 3 & 0 \\ 2 & 5 & 4 \end{bmatrix}.$$

The eigenvalues of $A^{\mathsf{T}}A$ are $\lambda_1 = 144$, $\lambda_2 = 36$, and $\lambda_3 = 0$ with corresponding eigenvectors

$$\mathbf{w}_1 = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}, \ \mathbf{w}_1 = \begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix}, \text{ and } \mathbf{w}_1 = \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix}.$$

In addition,

$$A\mathbf{w}_1 = \begin{bmatrix} 18 \\ 18 \\ 18 \\ 18 \end{bmatrix} \text{ and } A\mathbf{w}_2 = \begin{bmatrix} 9 \\ -9 \\ -9 \\ 9 \end{bmatrix}.$$

(a) Find orthogonal matrices U and V, and the matrix Σ, so that $U\Sigma V^{\mathsf{T}}$ is a singular value decomposition of A.

(b) Determine the best rank 1 approximation to A. Give an appropriate numerical estimate as to how good this approximation is to A.

(c) Find the pseudoinverse A^{+} of A.

(d) Let $\mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}^{\mathsf{T}}$. Does the matrix equation

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

have a solution? If so, find the solution. If not, find the best approximation you can to a solution to this matrix equation.

(e) Use the orthogonal basis $\{\frac{1}{2}[1\ 1\ 1\ 1]^{\mathsf{T}}, \frac{1}{2}[1\ -1\ -1\ 1]^{\mathsf{T}}\}$ of Col A to find the projection of $\mathbf{b}$ onto Col A. Compare to your solution in part (c).

Example Solution.

(a) Normalizing the eigenvectors $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ to normal eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, respectively, gives us an orthogonal matrix

$$V = \begin{bmatrix} \frac{2}{3} & -\frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} & -\frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \end{bmatrix}.$$

Now $A\mathbf{v}_i = A\frac{\mathbf{w}_i}{||\mathbf{w}_i||} = \frac{1}{||\mathbf{w}_i}A\mathbf{w}_i$, so normalizing the vectors $A\mathbf{w}_1$ and $A\mathbf{w}_2$ gives us vectors

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

that are the first two columns of our matrix U. Given that U is a 4×4 matrix, we need to find two other vectors orthogonal to $\mathbf{u}_1$ and $\mathbf{u}_2$ that will combine with $\mathbf{u}_1$ and $\mathbf{u}_2$ to form an orthogonal basis for $\mathbb{R}^4$. Letting $\mathbf{z}_1 = [1\ 1\ 1\ 1]^\mathsf{T}$, $\mathbf{z}_2 = [1\ -1\ -1\ 1]^\mathsf{T}$, $\mathbf{z}_3 = [1\ 0\ 0\ 0]^\mathsf{T}$, and $\mathbf{z}_4 = [0\ 1\ 0\ 1]^\mathsf{T}$, a computer algebra system shows that the reduced row echelon form of the matrix $[\mathbf{z}_1\ \mathbf{z}_2\ \mathbf{z}_3\ \mathbf{z}_4]$ is I_4, so that vectors $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$, $\mathbf{z}_4$ are linearly independent. Letting $\mathbf{w}_1 = \mathbf{z}_1$ and $\mathbf{w}_2 = \mathbf{z}_2$, the Gram-Schmidt process shows that the set $\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4\}$ is an orthogonal basis for $\mathbb{R}^4$, where $\mathbf{w}_3 = \frac{1}{4}[2\ 0\ 0\ -2]^\mathsf{T}$ and (using $[1\ 0\ 0\ -1]^\mathsf{T}$ for $\mathbf{w}_3$) $\mathbf{w}_4 = \frac{1}{4}[0\ 2\ -2\ 0]^\mathsf{T}$.

The set $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$ where $\mathbf{u}_1 = \frac{1}{2}[1\ 1\ 1\ 1]^\mathsf{T}$, $\mathbf{u}_2 = \frac{1}{2}[1\ -1\ -1\ 1]^\mathsf{T}$, $\mathbf{u}_3 = \frac{1}{\sqrt{2}}[1\ 0\ 0\ -1]^\mathsf{T}$ and $\mathbf{u}_4 = \frac{1}{\sqrt{2}}[0\ 1\ -1\ 0]^\mathsf{T}$ is an orthonormal basis for $\mathbb{R}^4$ and we can let

$$U = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}.$$

The singular values of A are $\sigma_1 = \sqrt{\lambda_1} = 12$ and $\sigma_2 = \sqrt{\lambda_2} = 6$, and so

$$\Sigma = \begin{bmatrix} 12 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Therefore, a singular value decomposition of A is $U\Sigma V^\mathsf{T}$ of

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} \begin{bmatrix} 12 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ -\frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \end{bmatrix}.$$

(b) The outer product decomposition of A is

$$A = \sigma_1\mathbf{u}_1\mathbf{v}_1^\mathsf{T} + \sigma_2\mathbf{u}_2\mathbf{v}_2^\mathsf{T}.$$

So the rank one approximation to A is

$$\sigma_1\mathbf{u}_1\mathbf{v}_1^\mathsf{T} = 12\left(\frac{1}{2}\right)\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}\begin{bmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 4 & 4 & 2 \\ 4 & 4 & 2 \\ 4 & 4 & 2 \\ 4 & 4 & 2 \end{bmatrix}.$$

The error in approximating A with this rank one approximation is

$$\sqrt{\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}} = \sqrt{\frac{36}{180}} = \sqrt{\frac{1}{5}} \approx 0.447.$$

(c) Given that $A = U\Sigma V^{\mathsf{T}}$, we use the pseudoinverse Σ^+ of Σ to find the pseudoinverse A^+ of A by

$$A^+ = V\Sigma^+ U^{\mathsf{T}}.$$

Now

$$\Sigma^+ = \begin{bmatrix} \frac{1}{12} & 0 & 0 \\ 0 & \frac{1}{6} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

so

$$\begin{aligned} A^+ &= \begin{bmatrix} \frac{2}{3} & -\frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} & -\frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} \frac{1}{12} & 0 & 0 \\ 0 & \frac{1}{6} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}^{\mathsf{T}} \\ &= \frac{1}{72} \begin{bmatrix} -2 & 6 & 6 & -2 \\ 4 & 0 & 0 & 4 \\ 5 & -3 & -3 & 5 \end{bmatrix}. \end{aligned}$$

(d) Augmenting A with $\mathbf{b}$ and row reducing shows that

$$[A\ \mathbf{b}] \sim \begin{bmatrix} 2 & 5 & 4 & 1 \\ 0 & -12 & -12 & -3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

so $\mathbf{b}$ is not in Col A and the equation $A\mathbf{x} = \mathbf{b}$ has no solution. However, the best approximation to a solution to $A\mathbf{x} = \mathbf{b}$ is found using the pseudoinverse A^+ of A. That best solution is

$$\begin{aligned} \mathbf{x}^* &= AA^+\mathbf{b} \\ &= \begin{bmatrix} 2 & 5 & 4 \\ 6 & 3 & 0 \\ 6 & 3 & 0 \\ 2 & 5 & 4 \end{bmatrix} \frac{1}{72} \begin{bmatrix} -2 & 6 & 6 & -2 \\ 4 & 0 & 0 & 4 \\ 5 & -3 & -3 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} 2 \\ 1 \\ 1 \\ 2 \end{bmatrix}. \end{aligned}$$

(e) The rank of A is 2 and an orthonormal basis for Col A is $\{\mathbf{u}_1, \mathbf{u}_2\}$, where $\mathbf{u}_1 = \frac{1}{2}[1\ 1\ 1\ 1]^\mathsf{T}$ and $\mathbf{u}_2 = \frac{1}{2}[1\ -1\ -1\ 1]^\mathsf{T}$. So

$$\begin{aligned}\text{proj}_{\text{Col } A}\mathbf{b} &= (\mathbf{b}\cdot\mathbf{u}_1)\mathbf{u}_1 + (\mathbf{b}\cdot\mathbf{u}_2)\mathbf{u}_2 \\ &= \left(\frac{3}{2}\right)\left(\frac{1}{2}\right)[1\ 1\ 1\ 1]^\mathsf{T} + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)[1\ -1\ -1\ 1]^\mathsf{T} \\ &= \frac{1}{2}[2\ 1\ 1\ 2]^\mathsf{T}\end{aligned}$$

as expected from part (c).

Example 34.6. According to the Centers for Disease Control and Prevention[4], the average length of a male infant (in centimeters) in the US as it ages (with time in months from 1.5 to 8.5) is given in Table 34.3. In this problem we will find the line and the quadratic of best fit in the least squares

Age (months)	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5
Average Length (cm)	56.6	59.6	62.1	64.2	66.1	67.9	69.5	70.9

Table 34.3: Average lengths of male infants.

sense to this data. We treat time in months as the independent variable and length in centimeters as the dependent variable.

(a) Find a line that is the best fit to the data in the least squares sense. Draw a picture of your least squares solution against a scatterplot of the data.

(b) Now find the least squares quadratic of the form $q(x) = a_2x^2 + a_1x + a_0$ to the data. Draw a picture of your least squares solution against a scatterplot of the data.

Example Solution.

(a) We assume that a line of the form $f(x) = a_1x + a_0$ contains all of the data points. The first date point would satisfy $1.5a_1 + a_0 = 56.6$, the second $2.5a_1 + a_0 = 59.6$, and so on, giving us the linear system

$$\begin{aligned}1.5a_1 + a_0 &= 56.6 \\ 2.5a_1 + a_0 &= 59.6 \\ 3.5a_1 + a_0 &= 62.1 \\ 4.5a_1 + a_0 &= 64.2 \\ 5.5a_1 + a_0 &= 66.1 \\ 6.5a_1 + a_0 &= 67.9 \\ 7.5a_1 + a_0 &= 69.5 \\ 8.5a_1 + a_0 &= 70.9.\end{aligned}$$

[4] https://www.cdc.gov/growthcharts/html_charts/lenageinf.htm

Letting

$$A = \begin{bmatrix} 1.5 & 1 \\ 2.5 & 1 \\ 3.5 & 1 \\ 4.5 & 1 \\ 5.5 & 1 \\ 6.5 & 1 \\ 7.5 & 1 \\ 8.5 & 1 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} a_1 \\ a_0 \end{bmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 56.6 \\ 59.6 \\ 62.1 \\ 64.2 \\ 66.1 \\ 67.9 \\ 69.5 \\ 70.9 \end{bmatrix},$$

we can write this system in the matrix form $A\mathbf{x} = \mathbf{b}$. Neither column of A is a multiple of the other, so the columns of A are linearly independent. The least squares solution to the system is then found by

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

Technology shows that (with entries rounded to 3 decimal places), A^{+} is

$$\begin{bmatrix} -0.083 & -0.060 & -0.036 & -0.012 & 0.012 & 0.036 & 0.060 & 0.083 \\ 0.542 & 0.423 & 0.304 & 0.185 & 0.065 & -0.054 & -0.173 & -0.292 \end{bmatrix},$$

and

$$A^{+}\mathbf{b} \approx \begin{bmatrix} 2.011 \\ 54.559 \end{bmatrix}.$$

So the least squares linear solution to $A\mathbf{x} = \mathbf{b}$ is f defined by $f(x) \approx 2.011x + 54.559$. A graph of f against the data points is shown at left in Figure 34.5.

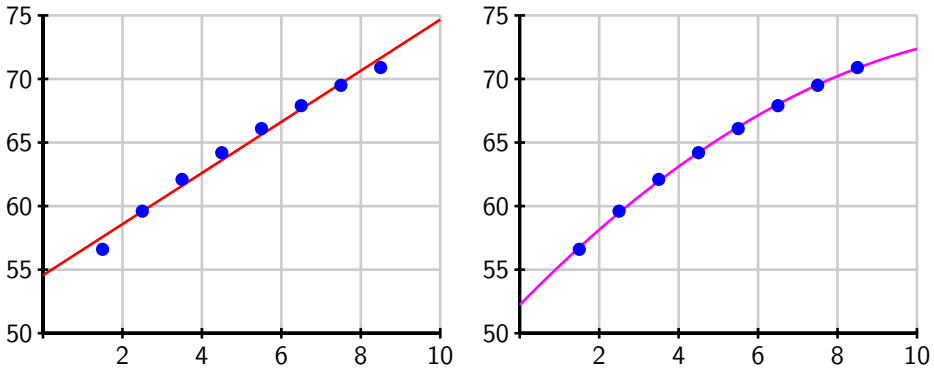

Figure 34.5: Left: Least squares line. Right: Least squares quadratic.

(b) The first date point would satisfy $(1.5^2)a_2 + 1.5a_1 + a_0 = 56.6$, the second $(2.5)^2a_2 +$

$2.5a_1 + a_0 = 59.6$, and so on, giving us the linear system

$$\begin{aligned}
1.5^2a_2 + 1.5a_1 + a_0 &= 56.6\\
2.5^2a_2 + 2.5a_1 + a_0 &= 59.6\\
3.5^2a_2 + 3.5a_1 + a_0 &= 62.1\\
4.5^2a_2 + 4.5a_1 + a_0 &= 64.2\\
5.5^2a_2 + 5.5a_1 + a_0 &= 66.1\\
6.5^2a_2 + 6.5a_1 + a_0 &= 67.9\\
7.5^2a_2 + 7.5a_1 + a_0 &= 69.5\\
8.5^2a_2 + 8.5a_1 + a_0 &= 70.9.
\end{aligned}$$

Letting

$$A = \begin{bmatrix} 1.5^2 & 1.5 & 1\\ 2.5^2 & 2.5 & 1\\ 3.5^2 & 3.5 & 1\\ 4.5^2 & 4.5 & 1\\ 5.5^2 & 5.5 & 1\\ 6.5^2 & 6.5 & 1\\ 7.5^2 & 7.5 & 1\\ 8.5^2 & 8.5 & 1 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} a_2\\ a_1\\ a_0 \end{bmatrix}, \text{ and } \mathbf{b} = \begin{bmatrix} 56.6\\ 59.6\\ 62.1\\ 64.2\\ 66.1\\ 67.9\\ 69.5\\ 70.9 \end{bmatrix},$$

we can write this system in the matrix form $A\mathbf{x} = \mathbf{b}$.

Technology shows that every column of the reduced row echelon form of A contains a pivot, so the columns of A are linearly independent. The least squares solution to the system is then found by

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

Technology shows that (with entries rounded to 3 decimal places) A^+ is

$$\begin{bmatrix} 0.042 & 0.006 & -0.018 & -0.030 & -0.030 & -0.018 & 0.006 & 0.042\\ -0.500 & -0.119 & 0.143 & 0.286 & 0.310 & 0.214 & 0.000 & -0.333\\ 1.365 & 0.540 & -0.049 & -0.403 & -0.522 & -0.406 & -0.055 & 0.531 \end{bmatrix},$$

and

$$A^+\mathbf{b} \approx \begin{bmatrix} -0.118\\ 3.195\\ 52.219 \end{bmatrix}.$$

So the least squares quadratic solution to $A\mathbf{x} = \mathbf{b}$ is q defined by $q(x) \approx -0.118x^2 + 3.195x + 52.219$. A graph of q against the data points is shown at right in Figure 34.5.

Summary

- The condition number of an $m \times n$ matrix A is the number $||A^{-1}||\ ||A||$. The condition number provides a measure of how well the relative error in a calculated value $\Delta\mathbf{b}$ predicts the relative error in $\Delta\mathbf{x}$ when we are trying to solve a system $A\mathbf{x} = \mathbf{b}$.

- A pseudoinverse A^+ of a matrix A can be found through a singular value decomposition. Let $U\Sigma V^\mathsf{T}$ be a singular value decomposition of an $m \times n$ matrix A of rank r, with

$$\Sigma = \left[\begin{array}{cccc|c} \sigma_1 & & & & \\ & \sigma_2 & & 0 & \\ & & \sigma_3 & & 0 \\ & 0 & & \ddots & \\ & & & & \sigma_r \\ \hline & & 0 & & 0 \end{array}\right]$$

  If Σ^+ is the $n \times m$ matrix defined by

$$\Sigma^+ = \left[\begin{array}{cccc|c} \frac{1}{\sigma_1} & & & & \\ & \frac{1}{\sigma_2} & & 0 & \\ & & \frac{1}{\sigma_3} & & 0 \\ & 0 & & \ddots & \\ & & & & \frac{1}{\sigma_r} \\ \hline & & 0 & & 0 \end{array}\right],$$

  then $A^+ = V\Sigma^+U^\mathsf{T}$.

- A pseudoinverse A^+ of a matrix A acts like an inverse for A. So if we can't solve a matrix equation $A\mathbf{x} = \mathbf{b}$ because $\mathbf{b}$ isn't in Col A, we can use the pseudoinverse of A to "solve" the equation $A\mathbf{x} = \mathbf{b}$ with the "solution" $A^+\mathbf{b}$. While not an exact solution, $A^+\mathbf{b}$ turns out to be the best approximation to a solution in the least squares sense.

- If the columns of A are linearly independent, then we can alternatively find the least squares solution as $\left(A^\mathsf{T}A\right)^{-1}A^\mathsf{T}\mathbf{b}$.

Exercises

(1) Let $A = \begin{bmatrix} 20 & 4 & 32 \\ -4 & 4 & 2 \\ 35 & 22 & 26 \end{bmatrix}$. Then A has singular value decomposition $U\Sigma V^\mathsf{T}$, where

$$U = \frac{1}{5}\begin{bmatrix} 3 & 4 & 0 \\ 0 & 0 & 5 \\ 4 & -3 & 0 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 60 & 0 & 0 \\ 0 & 15 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

$$V = \frac{1}{3}\begin{bmatrix} 2 & -1 & -2 \\ 1 & -2 & 2 \\ 2 & 2 & 1 \end{bmatrix}.$$

(a) What are the singular values of A?

(b) Write the outer product decomposition of A.

(c) Find the best rank 1 approximation to A. What is the relative error in approximating A by this rank 1 matrix?

(d) Find the best rank 2 approximation to A. What is the relative error in approximating A by this rank 2 matrix?

(2) Let $A = \begin{bmatrix} 861 & 3969 & 70 & 140 \\ 3969 & 861 & 70 & 140 \\ 3969 & 861 & -70 & -140 \\ 861 & 3969 & -70 & -140 \end{bmatrix}$.

(a) Find a singular value decomposition for A.

(b) What are the singular values of A?

(c) Write the outer product decomposition of A.

(d) Find the best rank 1, 2, and 3 approximations to A. How much information about A does each of these approximations contain?

(3) The University of Denver Infant Study Center investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months. The study sought a relationship between babies' first crawling age and the average temperature during the month they first try to crawl (about 6 months after birth). Some of the data from the study is in Table 34.4. Let x represent the temperature in degrees Fahrenheit and $C(x)$ the average crawling age in months.

x	33	37	48	57
$C(x)$	33.83	33.35	33.38	32.32

Table 34.4: Crawling age.

(a) Use Theorem 34.4 to find the least squares line to fit this data. Plot the data and your line on the same set of axes.

(b) Use your least squares line to predict the average crawling age when the temperature is 65.

(4) The cost, in cents, of a first class postage stamp in years from 1981 to 1995 is shown in Table 34.5.

Year	1981	1985	1988	1991	1995
Cost	20	22	25	29	32

Table 34.5: Cost of postage.

(a) Use Theorem 34.4 to find the least squares line to fit this data. Plot the data and your line on the same set of axes.

(b) Now find the least squares quadratic approximation to this data. Plot the quadratic function on same axes as your linear function.

(c) Use your least squares line and quadratic to predict the cost of a postage stamp in this year. Look up the cost of a stamp today and determine how accurate your prediction is. Which function gives a better approximation? Provide reasons for any discrepancies.

(5) Assume that the number of feet traveled by a batted baseball at various angles in degrees (all hit at the same bat speed) is given in Table 34.6.

Angle	10°	20°	30°	40°	50°	60°
Distance	116	190	254	285	270	230

Table 34.6: Distance traveled by batted ball.

(a) Plot the data and explain why a quadratic function is likely a better fit to the data than a linear function.

(b) Find the least squares quadratic approximation to this data. Plot the quadratic function on same axes as your data.

(c) At what angle (or angles), to the nearest degree, must a player bat the ball in order for the ball to travel a distance of 220 feet?

(6) Not all data is well modeled with polynomials – populations tend to grow at rates proportional to the population, which implies exponential growth. For example, Table 34.7 shows the approximate population of the United States in years between 1920 and 2000, with the population measured in millions. If we assume the population grows exponentially, we would

Year	1920	1930	1940	1950	1960	1970	1980	1990	2000
Population	106	123	142	161	189	213	237	259	291

Table 34.7: U.S. population.

want to find the best fit function f of the form $f(t) = ae^{kt}$, where a and k are constants. To apply the methods we have developed, we could instead apply the natural logarithm to both sides of $y = ae^{kt}$ to obtain the equation $\ln(y) = \ln(a) + kt$. We can then find the best fit line to the data in the form $(t, \ln(y))$ to determine the values of $\ln(a)$ and k. Use this approach to find the best fit exponential function in the least squares sense to the U.S. population data.

(7) How close can a matrix be to being non-invertible? We explore that idea in this exercise. Let $A = [a_{ij}]$ be the $n \times n$ upper triangular matrix with 1s along the diagonal and with every other entry being -1.

(a) What is $\det(A)$? What are the eigenvalues of A? Is A invertible?

(b) Let $B = [b_{ij}]$ be the $n \times n$ matrix so that $b_{n1} = -\frac{1}{2^{n-2}}$ and $b_{ij} = a_{ij}$ for all other i and j.

i. For the matrix B with $n = 3$, show that the equation $B\mathbf{x} = \mathbf{0}$ has a non-trivial solution. Find one non-trivial solution.

ii. For the matrix B with $n = 4$, show that the equation $B\mathbf{x} = \mathbf{0}$ has a non-trivial solution. Find one non-trivial solution.

iii. Use the pattern established in parts (i.) and (ii.) to find a non-trivial solution to the equation $B\mathbf{x} = \mathbf{0}$ for an arbitrary value of n. Be sure to verify that you have a solution. has a non-trivial solution. Is B invertible? (Hint: For any positive integer m, the sum $1 + \sum_{k=0}^{m-1} 2^k$ is the partial sum of a geometric series with ratio 2 and so $1 + \sum_{k=0}^{m-1} 2^k = 1 + \frac{1-2^m}{1-2} = 2^m$.)

iv. Explain why B is not an invertible matrix. Notice that A and B differ by a single entry, and that A is invertible and B is not. Let us examine how close A is to B. Calculate $||A - B||_F$? What happens to $||A - B||_F$ as n goes to infinity? How close can an invertible matrix be to becoming non-invertible?

(8) Let $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}$. In this exercise we find a matrix B so that $B^2 = A$, that is, find a square root of the matrix A.

(a) Find the eigenvalues and corresponding eigenvectors for A and $A^\mathsf{T}A$. Explain what you see.

(b) Find a matrix V that orthogonally diagonalizes $A^\mathsf{T}A$.

(c) Exercise 8 in Section 33 shows if $U\Sigma V^\mathsf{T}$ is a singular value decomposition for a symmetric matrix A, then so is $V\Sigma V^\mathsf{T}$. Recall that $A^n = \left(V\Sigma V^\mathsf{T}\right)^n = V\Sigma^n V^\mathsf{T}$ for any positive integer n. We can exploit this idea to define $\sqrt{A}$ to be the matrix

$$V\Sigma^{1/2}V^\mathsf{T},$$

where $\Sigma^{1/2}$ is the matrix whose diagonal entries are the square roots of the corresponding entries of Σ. Let $B = \sqrt{A}$. Calculate B and show that $B^2 = A$.

(d) Why was it important that A b ea symmetric matrix for this process to work, and what had to be true about the eigenvalues of A for this to work?

(e) Can you extend the process in this exercise to find a cube root of A?

(9) Let A be an $m \times n$ matrix with singular value decomposition $U\Sigma V^\mathsf{T}$. Let A^+ be defined as in (34.8). In this exercise we prove the remaining parts of Theorem 34.3.

(a) Show that $(AA^+)^\mathsf{T} = AA^+$. (Hint: $\Sigma\Sigma^+$ is a symmetric matrix.)

(b) Show that $(A^+A)^\mathsf{T} = A^+A$.

(10) In this exercise we show that the pseudoinverse of a matrix is the unique matrix that satisfies the Moore-Penrose conditions. Let A be an $m \times n$ matrix with singular value decomposition $U\Sigma V^\mathsf{T}$ and pseudoinverse $X = V\Sigma^+U^\mathsf{T}$. To show that A^+ is the unique matrix that satisfies the Moore-Penrose conditions, suppose that there is another matrix Y that also satisfies the Moore-Penrose conditions.

(a) Show that $X = YAX$.

(b) Show that $Y = YAX$.

(c) How do the results of parts (a) and (b) show that A^+ is the unique matrix satisfying the Moore-Penrose conditions?

(11) Find the pseudo-inverse of the $m \times n$ zero matrix $A = 0$. Explain the conclusion.

(12) In all of the examples that we have done finding a singular value decomposition of a matrix, it has been the case (though we haven't mentioned it), that if A is an $m \times n$ matrix, then $\text{rank}(A) = \text{rank}\left(A^\mathsf{T}A\right)$. Prove this result.

(13) Show that if the columns of a matrix A are linearly independent, then $A^\mathsf{T}A$ is invertible. (Hint: If $A^\mathsf{T}A\mathbf{x} = \mathbf{0}$, what is $\mathbf{x}^\mathsf{T}A^\mathsf{T}A\mathbf{x}$?)

(14) Label each of the following statements as True or False. Provide justification for your response.

(a) **True/False** A matrix has a pseudo-inverse if and only if the matrix is singular.

(b) **True/False** The pseudoinverse of an invertible matrix A is the matrix A^{-1}.

(c) **True/False** If the columns of A are linearly dependent, then there is no least squares solution to $A\mathbf{x} = \mathbf{b}$.

(d) **True/False** If the columns of A are linearly independent, then there is a unique least squares solution to $A\mathbf{x} = \mathbf{b}$.

(e) **True/False** If T is the matrix transformation defined by a matrix A and S is the matrix transformation defined by A^+, then T and S are inverse transformations.

Project: GPS and Least Squares

In this project we discuss some of the details about how the GPS works. The idea is based on intersections of spheres. To build a basic understanding of the system, we begin with a 2-dimensional example.

Project Activity 34.1. Suppose that there are three base stations A, B, and C in $\mathbb{R}^2$ that can send and receive signals from your mobile phone. Assume that A is located at point $(-1, -2)$, B at point $(36, 5)$, and C at point $(16, 35)$. Also assume that your mobile phone location is point (x, y). Based on the time that it takes to receive the signals from the three base stations, it can be determined that your distance to base station A is 28 km, your distance to base station B is 26 km, and your distance to base station C is 14 km using a coordinate system with measurements in kilometers based on a reference point chosen to be $(0, 0)$. Due to limitations on the measurement equipment, these measurements all contain some unknown error which we will denote as z. The goal is to determine your location in $\mathbb{R}^2$ based on this information.

If the distance readings were accurate, then the point (x, y) would lie on the circle centered at A of radius 29. The distance from (x, y) to base station A can be represented in two different ways:

28 km and $\sqrt{(x+1)^2+(y+2)^2}$. However, there is some error in the measurements (due to the receiver clock and satellite clocks not being snychronized), so we really have

$$\sqrt{(x+1)^2+(y+2)^2}+z=28,$$

where z is the error. Similarly, (x,y) must also satisfy

$$\sqrt{(x-36)^2+(y-5)^2}+z=26$$

and

$$\sqrt{(x-16)^2+(y-35)^2}+z=14.$$

(a) Explain how these three equations can be written in the equivalent form

$$(x+1)^2+(y+2)^2=(28-z)^2 \tag{34.9}$$
$$(x-36)^2+(y-5)^2=(26-z)^2 \tag{34.10}$$
$$(x-16)^2+(y-35)^2=(14-z)^2. \tag{34.11}$$

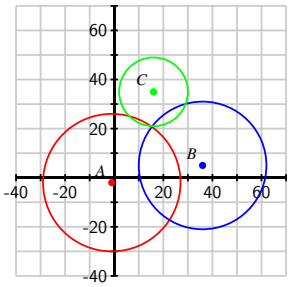

Figure 34.6: Intersections of circles.

(b) If all measurements were accurate, your position would be at the intersection of the circles centered at A with radius 28 km, centered at B with radius 26 km, and centered at C with radius 14 km as shown in Figure 34.6. Even though the figure might seem to imply it, because of the error in the measurements the three circles do not intersect in one point. So instead, we want to find the best estimate of a point of intersection that we can. The system of equations 34.9, 34.10, and 34.11 is non-linear and can be difficult to solve, if it even has a solution. To approximate a solution, we can linearize the system. To do this, show that if we subtract corresponding sides of equation (34.9) from (34.10) and expand both sides, we can obtain the linear equation

$$37x+7y+2z=712$$

in the unknowns x, y, and z.

(c) Repeat the process in part (b), subtracting (34.9) from (34.11) and show that we can obtain the linear equation

$$17x+37y+14z=1032$$

in x, y, and z.

(d) We have reduced our system of three non-linear equations to the system

$$\begin{aligned} 37x + \ \ 7y + \ \ 2z &= 712 \\ 17x + 37y + 14z &= 1032 \end{aligned}$$

of two linear equations in the unknowns x, y, and z. Use technology to find a pseudoinverse of the coefficient matrix of this system. Use the pseudoinverse to find the least squares solution to this system. Does your solution correspond to an approximate point of intersection of the three circles?

Project Activity 34.1 provides the basic idea behind GPS. Suppose you receive a signal from a GPS satellite. The transmission from satellite i provides four pieces of information – a location (x_i, y_i, z_i) and a time stamp t_i according to the satellite's atomic clock. The time stamp allows the calculation of the distance between you and the ith satellite. The transmission travel time is calculated by subtracting the current time on the GPS receiver from the satellite's time stamp. Distance is then found by multiplying the transmission travel time by the rate, which is the speed of light $c = 299792.458$ km/s.[5] So distance is found as $c(t_i - d)$, where d is the time at the receiver. This signal places your location within in a sphere of that radius from the center of the satellite. If you receive a signal at the same time from two satellites, then your position is at the intersection of two spheres. As can be seen at left in Figure 34.7, that intersection is a circle. So your position has been narrowed quite a bit. Now if you receive simultaneous signals from three spheres, your position is narrowed to the intersection of three spheres, or two points as shown at right in Figure 34.7. So if we could receive perfect information from three satellites, then your location would be exactly determined.

Figure 34.7: Intersections of spheres.

There is a problem with the above analysis – calculating the distances. These distances are determined by the time it takes for the signal to travel from the satellite to the GPS receiver. The times are measured by the clocks in the satellites and the clocks in the receivers. Since the GPS receiver clock is unlikely to be perfectly synchronized with the satellite clock, the distance calculations are not perfect. In addition, the rate at which the signal travels can change as the signal moves through

[5]The signals travel in radio waves, which are electromagnetic waves, and travel at the speed of light. Also, c is the speed of light in a vacuum, but atmosphere is not too dense so we assume this value of c

the ionosphere and the troposphere. As a result, the calculated distance measurements are not exact, and are referred to as *pseudoranges*. In our calculations we need to factor in the error related to the time discrepancy and other factors. We will incorporate these errors into our measure of d and treat d as an unknown. (Of course, this is all more complicated that is presented here, but this provides the general idea.)

To ensure accuracy, the GPS uses signals from four satellites. Assume a satellite is positioned at point (x_1, y_1, z_1) at a distance d_1 from the GPS receiver located at point (x, y, z). The distance can also be measured in two ways: as

$$\sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2}.$$

and as $c(t_1 - d)$. So

$$c(t_1 - d) = \sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2}.$$

Again, we are treating d as an unknown, so this equation has the four unknowns x, y, z, and d. Using signals from four satellites produces the system of equations

$$\sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2} = c(t_1 - d) \tag{34.12}$$

$$\sqrt{(x-x_2)^2+(y-y_2)^2+(z-z_2)^2} = c(t_2 - d) \tag{34.13}$$

$$\sqrt{(x-x_3)^2+(y-y_3)^2+(z-z_3)^2} = c(t_3 - d) \tag{34.14}$$

$$\sqrt{(x-x_4)^2+(y-y_4)^2+(z-z_4)^2} = c(t_4 - d). \tag{34.15}$$

Project Activity 34.2. The system of equations (34.12), (34.13), (34.14), and (34.15) is a nonlinear system and is difficult to solve, if it even has a solution. We want a method that will provide at least an approximate solution as well as apply if we use more than four satellites. We choose a reference node (say (x_1, y_1, z_1)) and make calculations relative to that node as we did in Project Activity 34.1.

(a) First square both sides of the equations (34.12), (34.13), (34.14), and (34.15) to remove the roots. Then subtract corresponding sides of the new first equation (involving (x_1, y_1, z_1)) from the new second equation (involving (x_2, y_2, z_2)) to show that we can obtain the linear equation

$$2(x_2 - x_1)x + 2(y_2 - y_1)y + 2(z_2 - z_1)z + 2c^2(t_1 - t_2)d = c^2(t_1^2 - t_2^2) - h_1 + h_2,$$

where $h_i = x_i^2 + y_i^2 + z_i^2$. (Note that the unknowns are x, y, z, and d – all other quantities are known.)

(b) Use the result of part (a) to write a linear system that can be obtained by subtracting the first equation from the third and fourth equations as well.

(c) The linearizations from part (b) determine a system $A\mathbf{x} = \mathbf{b}$ of linear equations. Identify A, $\mathbf{x}$, and $\mathbf{b}$. Then explain how we can approximate a best solution to this system in the least squares sense.

We conclude this project with a final note. At times a GPS receiver may only be able to receive signals from three satellites. In these situations, the receiver can substitute the surface of the Earth as a fourth sphere and continue the computation.