Drew University
College of Liberal Arts

CUR: An Alternative to SVD-based Methods

A Thesis in Mathematics

by
Gregory J. Hunt

Submitted in Partial Fulfillment of the Requirements
for the Degree of Bachelor of Arts
With Specialized Honors in Mathematics

May 2013

Advisor: Dr. Jon Kettenring
Thesis Committee Members:
Christopher Andrews
Christopher Apelian
Peter Likarish

Contents

Chapter 1

The Singular Value Decomposition

The purpose of this chapter is an introduction to the singular value decomposition. The singular value decomposition (SVD) is a beautiful and unifying matrix factorization. It is one of the most useful matrix factorizations in applied linear algebra and is fundamental to the work and hand. While it is important to understand the form and mechanics of the SVD it is, in my opinion, equally as important to understand the linear algebraic geometry surrounding the SVD. In this light then this chapter will not simply be matrix algebraic proofs. Indeed it will be a discussion about linear operators and vector spaces as much as about matrices. While this may seem somewhat detached from applied statistical research progressing through the SVD in this manner is not purely a product of my own affinity for such topics. Understanding the SVD and its relation to the fundamental subspaces has direct importance to the research at hand and is thus the route this chapter will take.

1.1 Diagonalization

Our first step towards the singular value decomposition starts with the diagonalization of matrices. This topic is at the heart of the SVD.

1.1.1 Diagonal And Diagonalizable Matrices

Diagonal Matrices

An important class of matrices is the class of diagonal matrices. A diagonal matrix A is a square $p \times p$ matrix for which all of the entries off the main diagonal are zero.

Example 1

The following matrices A and B are diagonal,

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & \pi \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

We will often omit the zeros are write these as

$$A = \begin{bmatrix} 1 & & \\ & -5 & \\ & & \pi \end{bmatrix} \text{ and } \quad B = \begin{bmatrix} 1 & \\ & \end{bmatrix}.$$

◀

The only parameters needed to specify a diagonal matrix are its size and its diagonal entries. Thus a diagonal matrix D with diagonal entries $d_1, \ldots, d_n$ may be denoted

$diag(d_1, \ldots, d_n)$. Here we assume that all of the diagonal entries are specified and so $diag(d_1, \ldots, d_n)$ is order n, that is, it is of size $n \times n$.

Example 2

Using the matrices A and B from Example 1,

$$A = diag(1, -5, \pi)$$

$$B = diag(1, 0).$$

◀

There are many reasons why diagonal matrices are important in linear algebra. Indeed we will see many applications of diagonal matrices throughout this paper. For now let us consider a centrally important one: the linear transformation of a diagonal matrix. Let us think of any real $n \times p$ matrix A as describing a linear transformation $f : \mathbb{R}^p \to \mathbb{R}^n$ where

$$f(x) = A\mathbf{x} \text{ for every } \mathbf{x} \in \mathbb{R}^p.$$

Since diagonal matrices are square then they are a transformation on $\mathbb{R}^p$. That is, they map vectors from $\mathbb{R}^p$ into $\mathbb{R}^p$. Thus we return to the same space from whence we came. More importantly, diagonal linear transformations have very simple behavior on $\mathbb{R}^p$. They just scale the components of vectors.

For an arbitrary $p \times p$ diagonal matrix $D = diag(d_1, \ldots, d_p)$ the the associated linear transformation $g : \mathbb{R}^p \to \mathbb{R}^p$ gives us a scaling. That is,

$$g(\mathbf{u}) = g\left(\begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix}\right) = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix} = \begin{bmatrix} d_1 u_1 \\ \vdots \\ d_p u_p \end{bmatrix} \text{ for any } \mathbf{u} \in \mathbb{R}^p.$$

Thus each of the components u_i of $\mathbf{u}$ is scaled by the i^{th} diagonal entry d_i. So we can think of a diagonal matrix as leaving the underlying space untouched and scaling each vector in the space by stretching or shrinking it along each direction. Alternatively we can think of such a transformation as leaving the vectors in place and stretching or shrinking the underlying space of the vectors.

Example 3

As an example consider the diagonal matrix $M = diag(3, 3)$ and its associated linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$. For any $\mathbf{v} = (x, y)^T \in \mathbb{R}^2$ we have that

$$T(\mathbf{v}) = M\mathbf{v} = \begin{bmatrix} 3 & \\ & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3x \\ 3y \end{bmatrix} = 3 \begin{bmatrix} x \\ y \end{bmatrix} = 3\mathbf{v}.$$

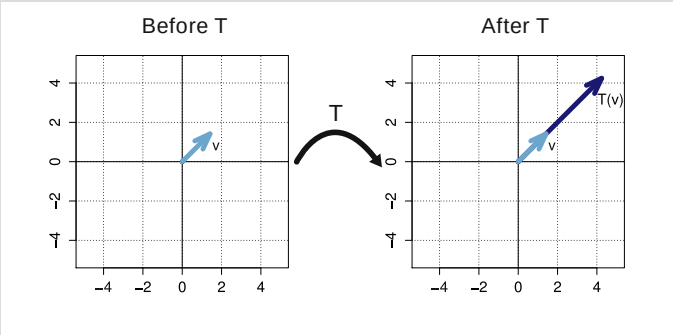

Figure 1.1: The map defined in Example 3 stretches vectors by a factor of 3.

◀

In Example 3 the transformation T just scales the vectors in $\mathbb{R}^2$ by 3. The matrix of the transformation is $3I_2$, $I_2 = diag(1, 1)$ being the 2-dimensional identity matrix. Generally

$$I_n = diag(\overbrace{1, \ldots, 1}^{n})$$

with n $1's$. We will omit the subscript n when the dimension is clear. The identity matrix I_n is in many ways the n-dimensional generalization of the number 1. Thus $3I_2$ is the two dimensional generalization of the real number 3. So T is the 2-dimensional generalization of the map $x \mapsto 3x$ for $x \in \mathbb{R}$. Generally in the case where all of the diagonal entries are equal then a diagonal matrix represents a homogeneous scaling of the space. Here we use homogeneous to mean that each of the components is scaled by the same factor (3 in the above example).

Example 4

Using the matrix A from above and a vector $\mathbf{v} = (x, y, z)^T$ then the linear transformation $\mathbf{v} \mapsto A\mathbf{v}$ maps

$$\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \text{ into } A\mathbf{v} = \begin{bmatrix} x \\ -5y \\ \pi z \end{bmatrix}.$$

◀

This example serves to show that not all diagonal transformations are homogeneous. One might call the transformation of Example 4 an *in*homogeneous scaling because it doesn't scale the components by the same factor. In this case it scales the components by $1, -5$ and π respectively. Nonetheless the behavior of diagonal transformations are quite easy to understand. They simply scale each component by some value.

Diagonalizing Matrices

Obviously not all square matrices are diagonal. Yet this is not to say that we can't view their associated linear transformations as scalings. Indeed the linear transformations of many $p \times p$ square matrices may be viewed as scalings of $\mathbb{R}^p$ if we look at their action through the lens of a suitably chosen basis.

Consider a linear transformation T on $\mathbb{R}^p$ defined, with respect to the standard basis $\mathcal{E}$, by a $p \times p$ matrix A. What we would like to do is find a basis $\mathcal{B}$ such that the matrix defining T under the basis $\mathcal{B}$ is diagonal. Let $\mathbf{x}$ be a vector in $\mathbb{R}^p$ and assume that such a basis $\mathcal{B}$ exists. Let us define $[\mathbf{x}]_\mathcal{E}$ and $[\mathbf{x}]_\mathcal{B}$ to be the standard basis and $\mathcal{B}$ basis representations of $\mathbf{x}$ respectively. Then T is defined in the standard $\mathcal{E}$ basis by

$$[T(\mathbf{x})]_\mathcal{E} = A[\mathbf{x}]_\mathcal{E}$$

or in the $\mathcal{B}$ basis by

$$[T(\mathbf{x})]_\mathcal{B} = D[\mathbf{x}]_\mathcal{B}.$$

where D is diagonal. This second definition allows us to view T as a scaling. Instead of scaling the standard basis components of $\mathbf{x}$ the map T scales the $\mathcal{B}$ basis components of $\mathbf{x}$.

Assuming such a basis $\mathcal{B}$ exists then given a standard basis representation of a vector $\mathbf{x}$, $[\mathbf{x}]_\mathcal{E}$, we can compute $[T(\mathbf{x})]_\mathcal{E}$ by the following steps.

1. Convert $[\mathbf{x}]_\mathcal{E}$ to a $\mathcal{B}$ basis representation $[\mathbf{x}]_\mathcal{B}$.
2. Apply the $\mathcal{B}$ representation of T by left multiplying by D to get $[T(\mathbf{x})]_\mathcal{B}$.
3. Convert $[T(\mathbf{x})]_\mathcal{B}$ back to the $\mathcal{E}$ basis representation $[T(\mathbf{x})]_\mathcal{E}$.

The only thing to be explained in this process is how switch between bases. If B is the matrix whose columns are the standard basis representation of the $\mathcal{B}$ basis vectors then surely $[\mathbf{x}]_\mathcal{E} = B[\mathbf{x}]_\mathcal{B}$ since $[\mathbf{x}]_\mathcal{B}$ is just the vector of $\mathcal{B}$ basis coordinates to $\mathbf{x}$. Thus B maps $[\mathbf{x}]_\mathcal{B}$ to $[\mathbf{x}]_\mathcal{E}$. Then B^{-1} defines the inverse linear transformation $[\mathbf{x}]_\mathcal{E} \mapsto [\mathbf{x}]_\mathcal{B}$. We can see this because if $[\mathbf{x}]_\mathcal{E} = B[\mathbf{x}]_\mathcal{B}$ then $B^{-1}[\mathbf{x}]_\mathcal{E} = B^{-1}B[\mathbf{x}]_\mathcal{B} = [\mathbf{x}]_\mathcal{B}$.

Then the three steps outlined above are

1. Compute $[\mathbf{x}]_{\mathcal{B}}$ as $[\mathbf{x}]_{\mathcal{B}} = B^{-1}[\mathbf{x}]_{\mathcal{E}}$.
2. Compute $[T(\mathbf{x})]_{\mathcal{B}}$ as $[T(\mathbf{x})]_{\mathcal{B}} = D[\mathbf{x}]_{\mathcal{B}}$.
3. Compute $[T(\mathbf{x})]_{\mathcal{E}}$ as $[T(\mathbf{x})]_{\mathcal{E}} = B[T(\mathbf{x})]_{\mathcal{B}}$.

Thus we have,

$$[T(\mathbf{x})]_{\mathcal{E}} = B[T(\mathbf{x})]_{\mathcal{B}} = BD[\mathbf{x}]_{\mathcal{B}} = BDB^{-1}[\mathbf{x}]_{\mathcal{E}}.$$

or, dropping the basis subscript,

$$T(\mathbf{x}) = BDB^{-1}\mathbf{x}$$

where $\mathbf{x}$ is a standard basis representation of a vector in $\mathbb{R}^p$.

Now we know that in the standard basis

$$T(\mathbf{x}) = A\mathbf{x}$$

and so it must be that

$$A = BDB^{-1}$$

or

$$AB = BD$$

and since $B = \begin{bmatrix}\mathbf{b_1} & \cdots & \mathbf{b_p}\end{bmatrix}$ then

$$A\begin{bmatrix}\mathbf{b_1} & \cdots & \mathbf{b_p}\end{bmatrix} = \begin{bmatrix}\mathbf{b_1} & \cdots & \mathbf{b_p}\end{bmatrix} D$$

or

$$\begin{bmatrix}A\mathbf{b_1} & \cdots & A\mathbf{b_p}\end{bmatrix} = \begin{bmatrix}d_1\mathbf{b_1} & \cdots & d_p\mathbf{b_p}\end{bmatrix}.$$

This means that for $i = 1, \ldots, p$

$$A\mathbf{b_i} = d_i\mathbf{b_i}$$

which is precisely an eigensystem problem. The $\mathbf{b_i}$ are eigenvectors and the d_i are the associated eigenvalues.

Thus we have solved our problem completely. Consider a linear transformation T defined in the standard basis by a matrix A. Then if there is a basis $\mathcal{B}$ under which T is defined by a diagonal matrix then the ordered basis $\mathcal{B}$ is comprised of eigenvectors $\mathbf{v_1}, \ldots, \mathbf{v_p}$ of A and the diagonal representation of T in this basis is $D = diag(\lambda_1, \ldots, \lambda_p)$ where λ_i is the eigenvalue associated with the eigenvector $\mathbf{v_i}$. When such a basis exists then we say that A and D are similar. They are similar in the sense that they represent the same linear transformation under different bases. Furthermore we know that they are related by the relation

$$A = BDB^{-1}$$

which is called a similarity transformation of D.

Solving for D we find that

$$D = B^{-1}AB.$$

If such an eigenbasis for $\mathbb{R}^p$ exists then we say that A is diagonalizable since we may diagonalize it via a similarity transformation. We may state this entire discussion as a theorem.

Theorem 1

A $p \times p$ matrix A is diagonalizable if and only if it has an eigenbasis of p linearly independent eigenvectors. In this case the diagonal matrix D to which A is similar is comprised of the associated eigenvalues of A.

Proof.
$\Rightarrow$ The forward direction is a direct corollary of our above discussion. If A is diagonalizable then we need p linearly independent eigenvectors to form the matrix B. Furthermore the diagonal matrix is one consisting of the associated eigenvalues of A. By associated we mean that if A has p eigenpairs $\{(\mathbf{v_i}, \lambda_i)\}_{i=1}^{p}$ then the i^{th} diagonal element of D is the eigenvalue corresponding to the i^{th} column of B, the i^{th} eigenvector.

$\Leftarrow$ The reverse direction is not harder. If A has p linearly independent eigenvectors then construct B by making the i^{th} column of B the i^{th} eigenvector $\mathbf{v_i}$. Similarly construct $D = diag(\lambda_1, \ldots, \lambda_p)$. Then we know that $A\mathbf{v_i} = \lambda\mathbf{v_i}$ for $i = 1, \ldots, p$ and so as above

$$AB = BD$$

or (since B is a square matrix with linearly independent columns),

$$B^{-1}AB = D$$

and thus A is diagonalizable.

■

While it is possible to lift this theorem out of the linear transformation lingo and state it simply as a theorem of matrix algebra it is important to understand the former. What the above theorem states is that if we can find an eigenbasis for the space using the eigenvectors of the matrix then the linear transformation defined by this matrix is quite simple, its just a scaling.

A general theme to keep in mind in this chapter is the following. We would like to understand matrices however there are a lot of them. The vector space of all matrices of size $n \times p$ is a np–dimensional space. If we could break down matrices (or certain classes of matrices) into the product of several very simple matrices then they would be easier to understand. If we can do this then the behavior of matrices would simply be the combined behavior of some very simple matrices. The purpose of this chapter is to see how we may do this. More on this later. For now let us close with an example.

Example 5

Consider the matrix

$$A = \begin{bmatrix} 5 & -5 \\ 0 & 4 \end{bmatrix}.$$

The eigenvalues of this matrix are $\lambda_1 = 5$ and $\lambda_2 = 4$ and we can find two associated eigenvectors of

$$\mathbf{v_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \mathbf{v_2} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}.$$

Since $\mathbf{v_1}$ and $\mathbf{v_2}$ aren't multiples of each other then they form a basis for $\mathbb{R}^2$. Then according to our above theorem we may diagonalize A by a similarity transformation. Indeed if

$$V = \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix} \text{ then } V^{-1} = \begin{bmatrix} 1 & -5 \\ 0 & 1 \end{bmatrix}$$

and

$$V^{-1}AV = \begin{bmatrix} 1 & -5 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & -5 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 4 \end{bmatrix}.$$

Figure 1.1 plots two unit eigenvectors $\mathbf{u_1} = \mathbf{v_1}$ and $\mathbf{u_2} = \frac{1}{\sqrt{26}}\mathbf{v_2}$. What we can see is that while the eigenvectors are not colinear they are also not orthogonal. Thus while the

eigenvectors of A form a basis for $\mathbb{R}^2$ they don't form an orthonormal basis. We would like to determine for which matrices the eigenvectors can form not only a basis but an orthonormal one.

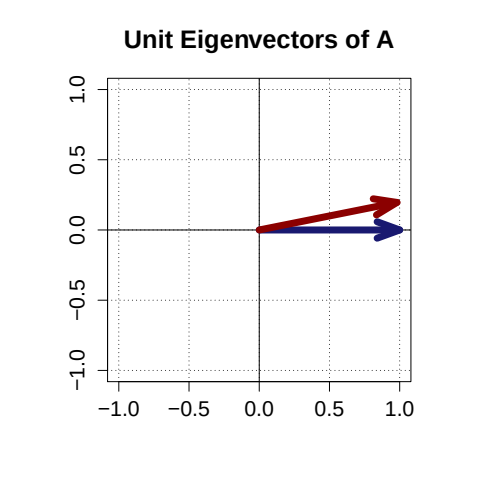

Figure 1.2: Blue arrow corresponds to the unit eigenvector $\mathbf{u_1}$ and the red line corresponds to the unit eigenvector $\mathbf{u_2}$.

◀

1.1.2 Orthogonal Diagonalization

We previously discovered that many linear operators defined by a square $p \times p$ matrix may be seen as a diagonal operator on $\mathbb{R}^p$ if we use the correct basis for the space. What would be nice is if we could find an orthonormal basis for $\mathbb{R}^p$ under which a linear transformation is defined by a diagonal matrix. That will the topic of this section.

Orthogonal Bases and Matrices

An orthonormal basis for a space is a basis $\mathcal{B} = \{\mathbf{b_1}, \ldots, \mathbf{b_p}\}$ for which the basis vectors $\{\mathbf{b_i}\}$ are

1. unit vectors, meaning
$$||\mathbf{b_i}|| = \sqrt{\mathbf{b_i}^T \mathbf{b_i}} = 1,$$
2. mutually orthogonal such that
$$\mathbf{b_i}^T \mathbf{b_j} = 0 \text{ for } i \neq j.$$

An orthogonal matrix U (which would probably better be called an orthonormal matrix) is a $p \times p$ matrix whose columns form an orthonormal basis for $\mathbb{R}^p$. Orthogonal matrices are always full rank since their columns are linearly independent. Their inverses are quite easy to find as they are simply their transpose. Indeed the property of a matrix U that
$$U^{-1} = U^T$$
is an alternative definition for orthogonal matrices.

Theorem 2

For a real matrix U, $U^{-1} = U^T$ if and only if U is orthogonal.

Proof.

$\Leftarrow$ For the reverse direction we need to show that if U is orthogonal then U^T is its inverse. Consider the matrix product U^TU. Since

$$(U^TU)_{i,j} = row(i, U^T)\, col(j, U) = col(i, U)^T\, col(j, U)$$

then since the columns of U form an orthonormal basis distinct columns are orthogonal and so

$$col(i, U)^T\, col(j, U) = 0 \text{ if } i \neq j$$

and

$$col(i, U)^T col(i, U) = ||col(i, U)||^2 = 1$$

because the columns of U are unit vectors. All together then,

$$(U^TU)_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

and so $U^TU = I$. Here we are using the notation $row(i, A)$ to denote the $1 \times p$ vector that is the i^{th} row of A and $col(j, A)$ to denote the $p \times 1$ j^{th} column of A.

Since we can play the same game and find that $UU^T = I$ then we have that

$$UU^T = U^TU = I$$

and so U^T is the inverse of U.

$\Rightarrow$ For the forward direction we need to show that if $U^{-1} = U^T$ then U is orthogonal. This is not much different from the reverse direction. If $U^{-1} = U^T$ then

$$U^TU = U^{-1}U = I$$

and so since

$$(U^TU)_{i,j} = row(i, U^T)\, col(j, U) = col(i, U)^T col(j, U)$$

then

$$col(i, U)^T col(j, U) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

and so all distinct pairs of the columns are orthogonal since their inner product is zero. Furthermore all of the columns are unit vectors since

$$col(i, U)^T col(i, U) = ||col(i, U)||^2 = 1$$

meaning that

$$||col(i, U)|| = 1.$$

These two conditions give us that U is then orthogonal because the columns of U are mutually orthogonal and unit vectors.

■

A direct corollary from this discussion is that the transpose (or equivalently inverse) of an orthogonal matrix is orthogonal. If U is orthogonal then the inverse of U^T is $U = (U^T)^T$ and so the inverse of U^T is its transpose and hence U^T is orthogonal.

Furthermore for an orthogonal matrix U, $det(U) = \pm 1$.

Corollary

For U, an orthogonal matrix, $det(U) = \pm 1$.

Proof. Clearly

$$det(U^{-1}) = det(U^T) = \frac{1}{det(U)}$$

but surely $det(A) = det(A^T)$ for any real matrix A and so

$$det(U) = \frac{1}{det(U)}$$

meaning that $det(U)$ is its own multiplicative inverse and hence $det(U) = \pm 1$.

■

Now since the columns of a $p \times p$ orthogonal matrix U form a basis for $\mathbb{R}^p$ then the linear transformation of an orthogonal matrix is a change of basis. However because the basis defined by the columns is orthonormal an orthogonal matrix is a particularly nice change of basis. An orthogonal matrix is simply a rotation, reflection or some roto-inversion of the standard basis vectors. While we won't go into a formal proof of this we would like to point out the more important point idea which is that an orthogonal linear transformation preserves inner product.

Theorem 3

Orthogonal linear transformations preserve inner product

Proof. Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\mathbb{R}^p$ and for some orthogonal matrix U let

$$\mathbf{x}_U = U\mathbf{x}$$

and

$$\mathbf{y}_U = U\mathbf{y}$$

be the images of $\mathbf{x}$ and $\mathbf{y}$ under the linear transformation defined by U. Then

$$\mathbf{x}_U^T\mathbf{y}_U = (U\mathbf{x})^T(U\mathbf{y}) = \mathbf{x}^T U^T U\mathbf{y} = \mathbf{x}^T\mathbf{y}$$

since $UU^T = I$.

■

Since inner product is preserved then vector norms are preserved. Thus the length of vectors is preserved under orthogonal matrices as well as the distance between vectors. After all the distance between vectors is simply

$$dist(u, v) = ||u - v||$$

for $u, v \in \mathbb{R}^p$. Thus the distances of all vectors from the origin (their norms) and the distance between all vectors are unchanged under orthogonal linear transformations. So the geometry of the space is really unchanged under the transformation. What has changed is that the vectors are expressed in different basis. While non-orthogonal linear transformations will stretch or collapse the space orthogonal transformations do not do this. If we imagine a vector space as some kind of Euclidean space with points representing the vectors then an orthogonal transformation simply removes one coordinate system and replaces it with a new (yet still orthonormal) system. Alternatively we can think of leaving the underlying space alone and rotating (really roto-inverting) all of the vectors in the space.

Example 6

The $p \times p$ identity matrix I_p is an orthogonal matrix because $I^{-1} = I^T = I$. Furthermore the 2×2 rotation matrix of the plane $\mathbb{R}^2$ through an angle θ,

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

is an orthogonal matrix. We can see this because the columns are orthogonal since

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) \end{bmatrix} \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix} = -\cos(\theta)\sin(\theta) + \sin(\theta)\cos(\theta) = 0$$

and the columns are unit vectors because

$$||col(1,R)|| = ||col(2,R)|| = \sqrt{cos^2(\theta) + \sin^2(\theta)} = 1.$$

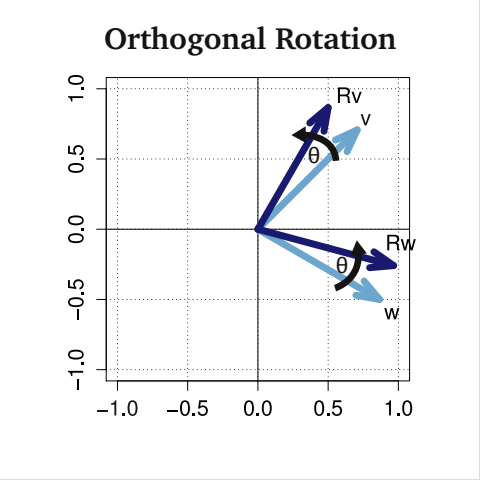

Figure 1.3: The rotation matrix in Example 6 rotates all of the vectors in the space through some angle θ. Notice that the distances between all vectors remain unchanged under the transformation by the matrix R.

◀

Orthogonal Diagonalization

Now Theorem 1 told us that if a $p \times p$ matrix A had an eigenbasis for $\mathbb{R}^p$ then we could diagonalize A via a similarity transformation $B^{-1}AB$ such that

$$D = B^{-1}AB$$

where the columns of B are an eigenbasis for $\mathbb{R}^p$ and the diagonal entries of D are the corresponding eigenvalues. What we would like to find out is when we can diagonalize a matrix A via an orthogonal similarity transformation such that

$$D = U^{-1}AU = U^TAU$$

where U is an orthogonal matrix and so $U^{-1} = U^T$. If we can do this then the linear transformation of the matrix A may be seen as a scaling under the appropriate roto-inversion of the space.

The first thing that we must notice is that if a matrix A is orthogonally diagonalizable, that is

$$D = U^T AU$$

then

$$A = UDU^T$$

and so

$$A^T = (UDU^T)^T = (U^T)^T D^T U^T = UDU^T = A$$

and so A must be symmetric (meaning $A = A^T$). It turns out that this is enough, a matrix is orthogonally diagonalizable if and only if it is symmetric.

To prove this we will need the following fact. While only symmetric matrices are orthogonally diagonalizable by a similarity transformation a wider class of real matrices are orthogonally triagulatable by an orthogonal similarity transformation.

Theorem 4

A $p \times p$ real matrix A is similar (via an orthogonal similarity transformation) to a upper triangular matrix $\mathcal{T}$ if A has p (counting multiplicities) real eigenvalues.

Proof. This factorization is know as the Schur factorization. We will prove it by induction on the order of the matrix.

Vacuously this is true for a 1×1 matrix (really just a scalar). Now assume it is true for all matrices of order $n-1$, i.e. square $(n-1) \times (n-1)$ matrices, such that all order $n-1$ matrices with $n-1$ real eigenvalues are similar via an orthogonal similarity transformation to an upper triangular matrix.

Consider an $n \times n$ matrix A with n real eigenvalues. Let μ be an eigenvalue and pick an associated unit eigenvector $\mathbf{u}$. Now find vectors $\mathbf{w_2}, \ldots, \mathbf{w_n}$ (via Gram-Schmidt or the like) such that $\{\mathbf{u}, \mathbf{w_2}, \ldots, \mathbf{w_n}\}$ is an ordered orthonormal basis for $\mathbb{R}^n$. Let U be the matrix which has $\mathbf{u}, \mathbf{w_2}, \ldots, \mathbf{w_n}$ as its columns. Then

$$\begin{aligned} U^T AU &= U^T \begin{bmatrix} A\mathbf{u} & A\mathbf{w_2} & \cdots & A\mathbf{w_n} \end{bmatrix} \\ &= U^T \begin{bmatrix} \mu\mathbf{u} & * & \cdots & * \end{bmatrix} \\ &= \left[\begin{array}{c|c} \mu\mathbf{u}^T\mathbf{u} & * \cdots * \\ \hline \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & C \end{array}\right] = \left[\begin{array}{c|c} \mu & * \\ \hline 0 & C \end{array}\right] \end{aligned}$$

where C is the corner $(n-1) \times (n-1)$ block of the matrix and the $*'s$ indicate entries whose values we do not particularly care about. Now $det(U) = det(U^T) = \pm 1$ since U is orthogonal and so for some scalar λ,

$$\begin{aligned} det(A - \lambda I) &= det(U^T) det(A - \lambda I) det(U) \\ &= det(U^T (A - \lambda I) U) \\ &= det(U^T AU - U^T U \lambda I) \\ &= det(U^T AU - \lambda I) \end{aligned}$$

and so the characteristic equation for $U^T AU$ is the same as that for A meaning they share the same eigenvalues.

We can use the row operations of row interchange or row addition to manipulate the $C - \lambda I$ block of $U^T AU - \lambda I$ into an upper triangular matrix $(C - \lambda I)'$. That is, since

$$U^T AU - \lambda I = \left[\begin{array}{c|c} \mu - \lambda & * \\ \hline 0 & C - \lambda I \end{array}\right]$$

then via row operations

$$U^T AU - \lambda I \rightsquigarrow M$$

where

$$M = \left[\begin{array}{c|c} \mu - \lambda & * \\ \hline 0 & (C - \lambda I)' \end{array}\right]$$

and $(C - \lambda I)'$ is upper triangular.

Notice that we can do this without touching the first row of $U^T AU - \lambda I$. Furthermore since row addition doesn't change the determinant of a matrix and row interchange only flips the determinant's sign (and these are the only operations we need use) then

$$det(U^T AU - \lambda I) = \pm det(M)$$

and so the characteristic equation for the two are the same since if

$$det(U^T AU - \lambda I) = 0$$

then

$$\pm det(M) = 0$$

but we can simply drop the $\pm$ and so $det(M) = 0$.

However M is an upper triangular matrix and so its determinant is simply the product of its main diagonal. Since the only diagonal entry of M not in $(C - \lambda I)'$ is $\mu - \lambda$ then

$$det(M) = (\mu - \lambda) det((C - \lambda I)') = \pm(\mu - \lambda) det(C - \lambda I)$$

since when row reducing $U^T AU - \lambda I$ into M we use only use row addition and row interchange to manipulate the submatrices

$$C - \lambda I \rightsquigarrow (C - \lambda I)'.$$

Thus the characteristic polynomial defined by M is

$$\pm(\mu - \lambda) det(C - \lambda I)$$

which has the same roots of the characteristic polynomial for $U^T AU$ and hence A and so if $\mu = \mu_1, \mu_2 \ldots, \mu_n$ are the n real eigenvalues of A then $\mu_2, \ldots, \mu_n$ must be roots of

$$det(C - \lambda I) = 0$$

and hence C has $n - 1$ real eigenvalues.

Thus we have discovered that C is a real $(n - 1) \times (n - 1)$ matrix with $n - 1$ real eigenvalues and hence by our induction hypothesis there is some orthogonal matrix $\mathcal{O}$ such that

$$\mathcal{O}^T C \mathcal{O} = R$$

where R is an upper triangular matrix. Then if

$$S = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \mathcal{O} \end{array}\right]$$

it is easy to verify that

$$\begin{aligned} S^T U^T AUS &= \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \mathcal{O}^T \end{array}\right] \left[\begin{array}{c|c} \mu & * \\ \hline 0 & C \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \mathcal{O} \end{array}\right] \\ &= \left[\begin{array}{c|c} \mu & * \\ \hline 0 & \mathcal{O}^T C \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \mathcal{O} \end{array}\right] \\ &= \left[\begin{array}{c|c} \mu & * \\ \hline 0 & \mathcal{O}^T C \mathcal{O} \end{array}\right] \\ &= \left[\begin{array}{c|c} \mu & * \\ \hline 0 & R \end{array}\right]. \end{aligned}$$

Thus if $V = US$ then $V^{-1} = (US)^{-1} = S^{-1} U^{-1} = S^T U^T = (US)^T$ and so V is orthogonal and

$$V^T AV = S^T U^T ASU = \left[\begin{array}{c|c} \mu & * \\ \hline 0 & R \end{array}\right]$$

which is upper triangular since R is upper triangular. Thus we have found an orthogonal matrix V such that V^TAV is upper triangular.

■

This factorization is more useful in theory than in practice. For our purposes we will prove that every $p \times p$ symmetric matrix has p real eigenvalues and hence permits such a factorization.

Theorem 5

The eigenvalues of a real symmetric matrix are real.

Proof. Consider a real symmetric matrix A and an eigenpair $(\mathbf{v}, \lambda)$. Now we know that

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

Consider the operation of conjugate transpose for a matrix B over the complex field $\mathbb{C}$. The conjugate transpose of B, denoted $B^\dagger$, is precisely as it sounds,

$$B^\dagger = \overline{B}^T$$

where $\overline{B}$ is the element-wise complex conjugate of B. Let us take the conjugate transpose of our first equation. Then

$$(A\mathbf{v})^\dagger = (\lambda\mathbf{v})^\dagger$$

implies

$$\mathbf{v}^\dagger A = \overline{\lambda}\mathbf{v}^\dagger$$

since A is real and symmetric so

$$A^\dagger = \overline{A}^T = \overline{A} = A$$

and λ is a scalar so

$$\lambda^\dagger = \overline{\lambda}^T = \overline{\lambda}.$$

Now consider multiplying on the right by $\mathbf{v}$. Then

$$\mathbf{v}^\dagger A\mathbf{v} = \overline{\lambda}\mathbf{v}^\dagger\mathbf{v}$$

and since $A\mathbf{v} = \lambda\mathbf{v}$ then

$$\mathbf{v}^\dagger\lambda\mathbf{v} = \overline{\lambda}\mathbf{v}^\dagger\mathbf{v}$$

or because λ is a scalar and hence commutes,

$$\lambda\mathbf{v}^\dagger\mathbf{v} = \overline{\lambda}\mathbf{v}^\dagger\mathbf{v}$$

meaning that

$$\lambda = \overline{\lambda}$$

since $\mathbf{v} \neq 0$ because it is an eigenvector. Hence it must be that λ is real.

■

Since every $p \times p$ matrix has (counting multiplicities) p eigenvalues over the complex field then a symmetric $p \times p$ matrix not only has all real eigenvalues it must have p real eigenvalues and thus permits a Schur Factorization. This is enough to prove the main theorem of this section.

Theorem 6

A matrix A is orthogonally diagonalizable if and only if A is symmetric.

Proof.

$\Rightarrow$ We have already shown that if a matrix A is orthogonally diagonalizable then it must be symmetric since

$$A^T = (UDU^T)^T = (U^T)^T D^T U^T = UDU^T = A.$$

$\Leftarrow$ Now if a $p \times p$ matrix A is symmetric then A has p real eigenvalues and so it has a Schur factorization such that A is similar via an orthogonal similarity transformation, $\mathcal{O}$, to an upper triangular $\mathcal{T}$ as

$$\mathcal{O}^T A \mathcal{O} = \mathcal{T}.$$

Then $\mathcal{T}^T = (\mathcal{O}^T A \mathcal{O})^T = \mathcal{O}^T A^T (\mathcal{O}^T)^T = \mathcal{O}^T A \mathcal{O} = \mathcal{T}$ and so $\mathcal{T}$ is a symmetric triangular matrix and so it must be diagonal.

■

Thus we have completely characterized those matrices permitting orthogonal diagonalizations. The action of any real symmetric matrix may be seen as the product of orthogonal and diagonal matrices. Such a view of symmetric matrices is quite nice. The linear transformation defined by these matrices may be viewed as simply a scaling. We can see this if we abandon the standard basis and literally rotate of our viewpoint. On the other hand diagonal and orthogonal matrices are quite simple to manipulate with matrix algebra and so expressing symmetric matrices as such can simplify matrix algebraic manipulations.

However we also see that only a very narrow family, those symmetric matrices, permit orthogonal diagonalizations. It would be nice if we could extend such ideas to a wider class of matrices. The singular value decomposition is precisely this idea. It is a matrix factorization which, as close as possible, allows orthogonal diagonalization for any arbitrary $n \times p$ matrix.

To close this section and transition into the singular value decomposition let us consider the following example.

Example 7

While any real $n \times p$ matrix A is not necessarily symmetric the associated matrices $A^T A$ and AA^T are symmetric since

$$(A^T A)^T = A^T (A^T)^T = A^T A$$

and

$$(AA^T)^T = (A^T)^T A^T = AA^T.$$

For example consider the real matrix

$$A = \begin{bmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{bmatrix}.$$

Then

$$A^T A = \begin{bmatrix} 333 & 81 \\ 81 & 117 \end{bmatrix}$$

and

$$AA^T = \begin{bmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{bmatrix}.$$

Now both of these matrices are symmetric and so they are orthogonally diagonalizable by our previous theorem.

For A^TA we find that the eigenvalues are $\lambda_1 = 360$ and $\lambda_2 = 90$ with associated eigenvectors of

$$v_1 = \begin{bmatrix} 81 \\ 27 \end{bmatrix} \text{ and } v_2 = \begin{bmatrix} -27 \\ 81 \end{bmatrix}$$

and so if

$$V = \begin{bmatrix} 81 & -27 \\ 27 & 81 \end{bmatrix}$$

then

$$V^TA^TAV = diag(360, 90).$$

For AA^T we find that the eigenvalues are $\mu_1 = 360, \mu_2 = 90$ and $\mu_3 = 0$ with associated eigenvectors of

$$u_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, u_2 = \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix} \text{ and } u_3 = \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix}$$

and so if

$$U = \begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ 2 & -2 & -1 \end{bmatrix}$$

then

$$U^TAA^TU = diag(360, 90, 0).$$

Notice that both U and V are orthogonal and so we have indeed orthogonally diagonalized these matrices.

◀

1.2 The Singular Value Decomposition

An interesting fact about Example 7 is that the nonzero eigenvalues of A^TA and AA^T are equal. This is not contrived but always true. We will see that the symmetric forms A^TA and AA^T tell us very important information about the matrix A and will thus be fundamental in developing the singular value decomposition.

The singular value decomposition is a factorization of a real $n \times p$ matrix A as

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

where U is an $n \times n$ orthogonal matrix, V is a $p \times p$ orthogonal matrix and Σ is an $n \times p$ matrix whose elements off the main diagonal are zero.

This should look very reminiscent of the orthogonal diagonalizations of the previous section. While only symmetric matrices permit an orthogonal diagonalization *any* $n \times p$ real matrix A has a singular value decomposition. With symmetric matrices we may diagonalize them by a similarity transformation of a single orthogonal matrix. With any real $n \times p$ matrix we may transform it into a "basically" diagonal matrix Σ with two orthogonal matrices. By basically diagonal we mean that Σ is zero off the main diagonal but not necessarily square. We will call such matrices "rectangular diagonal matrices". Thus we may rectangularly diagonalize any real $n \times p$ matrix via two orthogonal matrices U and V as

$$U^TAV = \Sigma.$$

We notice that since $A = U\Sigma V^T$ then

$$A^TA = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^TU^TU\Sigma V^T = V(\Sigma^T\Sigma)V^T$$

and

$$AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma V^TV\Sigma^TU^T = U(\Sigma\Sigma^T)U^T.$$

Then the matrix V is the matrix which orthogonally diagonalizes the symmetric form A^TA and the matrix U is the matrix orthogonally diagonalizing AA^T. Furthermore the Σ matrix is related to the diagonalized forms of A^TA and AA^T and thus their respective eigenvalues.

1.2.1 The Symmetric Forms A^TA and AA^T

Singular Values

While the above is not a proof of the existence of the SVD it shows us some properties that must be true should such a factorization exist. In this light then it seems appropriate to investigate the symmetric forms A^TA and AA^T. We will start by proving as a theorem the curiosity we noticed in Example 7.

Theorem 7

For a real matrix A the non-zero eigenvalues of A^TA and AA^T are the same.

Proof. Consider an eigenpair $(\mathbf{v}, \lambda)$ of A^TA such that $\lambda \neq 0$. Then

$$A^TA\mathbf{v} = \lambda\mathbf{v}$$

and so left-multiplying by A we get

$$AA^TA\mathbf{v} = \lambda A\mathbf{v}$$

or, adding parentheses for emphasis,

$$AA^T(A\mathbf{v}) = \lambda(A\mathbf{v})$$

and so $A\mathbf{v}$ is an eigenvector of AA^T and hence λ is an eigenvalue of AA^T. We can do the same thing for AA^T and show that all non-zero eigenvalues of AA^T are eigenvalues of A^TA. Thus the set of non-zero eigenvalues is shared between the two symmetric forms.

The thing to check, however, is that these symmetric forms not only have the same set of nonzero eigenvalues but have the same number of repetitions of nonzero eigenvalues. That is, we want to check the algebraic and geometric multiplicities of the shared non-zero eigenvalues.

Since A^TA and AA^T are symmetric matrices then we know that if an eigenvalue has an algebraic multiplicity of k then we can find k orthogonal (not just linearly independent) eigenvectors. This follows from the fact that these matrices are orthogonally diagonalizable and thus there must be an orthonormal eigenbasis. A corollary is that the geometric multiplicity equals the algebraic. Thus let us assume that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal eigenvectors corresponding to the eigenspace of an eigenvalue λ of A^TA. Then

$$A^TA\mathbf{u} = \lambda\mathbf{u} \text{ and } A^TA\mathbf{v} = \lambda\mathbf{v}$$

and as above

$$AA^TA\mathbf{u} = \lambda A\mathbf{u} \text{ and } AA^TA\mathbf{v} = \lambda A\mathbf{v}$$

meaning

$$A\mathbf{u} \text{ and } A\mathbf{v}$$

are eigenvectors of AA^T corresponding to the eigenvalue λ. Then

$$(A\mathbf{u})^T(A\mathbf{v}) = \mathbf{u}^TA^TA\mathbf{v} = \lambda\mathbf{u}^T\mathbf{v}$$

since $\mathbf{v}$ is an eigenvector of A^TA. However since $\mathbf{u}$ and $\mathbf{v}$ are orthogonal then $\mathbf{u}^T\mathbf{v} = 0$ and hence

$$\lambda\mathbf{u}^T\mathbf{v} = 0$$

and so $A\mathbf{v}$ and $A\mathbf{u}$ are orthogonal eigenvectors of AA^T corresponding to the eigenvalue λ. Thus if the eigenspace for an eigenvalue λ of A^TA is k-dimensional then we can find k orthogonal eigenvectors for the eigenvalue λ of AA^T and so the eigenvalue λ of AA^T has a geometric multiplicity of at least k. Thus the geometric multiplicity of any λ of

AA^T is at least that of the corresponding λ to A^TA. We can play the other side and show that the eigenspace of any eigenvalue of A^TA is at least as big as the associated space of AA^T and so they must be equal.

Thus the nonzero eigenvalues of A^TA and AA^T are the same and they have the same algebraic and geometric multiplicities.

■

Notice that we are ignoring the zero eigenvalues because they are (of course) the same for both matrices in terms of value. However they are not necessarily equal in terms of geometric or algebraic multiplicity. Indeed since A^TA and AA^T need not be the same size then since they have the same number of nonzero eigenpairs surely, in general, they will have a different number of zero eigenvalues. This follows since if A is $n \times p$ then A^TA is $p \times p$, AA^T is $n \times n$ yet they are both symmetric and hence have precisely p and n real eigenvalues respectively. However since they both have the same number of nonzero eigenvalues, say m, then respectively they have zero as an eigenvalue $p - m$ and $n - m$ times. Since $p \neq n$ in general then generally $p - m \neq n - m$. Thus they will have zero as a repeated eigenvalues a different number of times.

Furthermore keep in mind the following corollary from the above proof.

Corollary

If $(\mathbf{u}, \lambda)$ and $(\mathbf{v}, \mu)$ are eigenpairs of A^TA where u and v are unit vectors that are mutually orthogonal then $A\mathbf{u}$ and $A\mathbf{v}$ are orthogonal eigenvectors of AA^T corresponding to the same eigenvalues. Alternatively if $(\mathbf{w}, \omega)$ and $(\mathbf{x}, \xi)$ are eigenpairs of AA^T where w and x are unit vectors and mutually orthogonal then $A^T\mathbf{w}$ and $A^T\mathbf{x}$ are orthogonal eigenvectors of A^TA corresponding to the same eigenvalues.

■

Now previously we had noted that

$$A^TA = V(\Sigma^T\Sigma)V^T \text{ and } AA^T = U(\Sigma\Sigma^T)U^T$$

with orthogonal matrices U and V. Similarly we had specified that Σ was an $n \times p$ matrix with zeros off the main diagonal. That is, for some nonzero σ_i,

$$\Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_k & \\ \hline & 0 & & 0 \end{array}\right] = \left[\begin{array}{c|c} D & 0_{1,2} \\ \hline 0_{2,1} & 0_{2,2} \end{array}\right]$$

where $D = diag(\sigma_1, \ldots, \sigma_k)$ is a $k \times k$ diagonal matrix of the non-zero elements of Σ, $0_{1,2}$ is a $k \times (p-k)$ matrix of zeros, $0_{2,1}$ is a $(n-k) \times k$ matrix of zeros and $0_{2,2}$ is a $(n-k) \times (p-k)$ matrix of zeros with $1 \leq k \leq min\{n, p\}$. Schematically

$$[\Sigma] = \left[\begin{array}{c|c} [D] & [0_{1,2}] \\ \hline [0_{2,1}] & [0_{2,2}] \end{array}\right] = \left[\begin{array}{c|c} k \times k & k \times (p-k) \\ \hline (n-k) \times k & (n-k) \times (p-k) \end{array}\right].$$

Thus

$$\Sigma^T\Sigma = \left[\begin{array}{c|c} D^2 & 0 \\ \hline 0 & 0 \end{array}\right] = diag(\sigma_1^2, \ldots, \sigma_k^2, \overbrace{0, \ldots, 0}^{p-k \text{ times}})$$

a $p \times p$ diagonal matrix while,

$$\Sigma\Sigma^T = \left[\begin{array}{c|c} D^2 & 0 \\ \hline 0 & 0 \end{array}\right] = diag(\sigma_1^2, \ldots, \sigma_k^2, \overbrace{0, \ldots, 0}^{n-k \text{ times}})$$

is a $n \times n$ diagonal matrix. The unspecified matrices of zeros are assumed to be the correct size to complete the matrices. Then since

$$U^T AA^T U = \Sigma\Sigma^T = diag(\sigma_1^2, \ldots, \sigma_k^2, 0, \ldots, 0)$$

with orthogonal U then

$$diag(\sigma_1^2, \ldots, \sigma_k^2, 0, \ldots, 0) = diag(\lambda_1, \ldots, \lambda_n)$$

where the λ_i are the eigenvalues of AA^T. Similarly since

$$V^T A^T AV = \Sigma^T\Sigma = diag(\sigma_1^2, \ldots, \sigma_k^2, 0, \ldots, 0)$$

with orthogonal V then

$$diag(\sigma_1^2, \ldots, \sigma_k^2, 0, \ldots, 0) = diag(\mu_1, \ldots, \mu_p).$$

where the μ_i are the eigenvalues of $A^T A$. Both of these follow because U and V orthogonally diagonalize the symmetric forms and so the diagonal matrices to which they are similar must have their eigenvalues on the diagonal.

Then since precisely k of the diagonal elements of $\Sigma^T\Sigma$ or $\Sigma\Sigma^T$ are non-zero, precisely the σ_i^2's, these must be precisely the non-zero eigenvalues $\{\mu_i\}$ and $\{\lambda_i\}$ of $A^T A$ and AA^T. Hence

$$\sigma_i^2 = \lambda_i = \mu_i$$

and so

$$\sigma_i = \sqrt{\lambda_i} = \sqrt{\mu_i}$$

the square roots of the non-zero eigenvalues of $A^T A$ or AA^T. This is okay because we know that there are the same number of non-zero eigenvalues of these two symmetric forms. Furthermore the eigenvalues are non-negative so a real square root exists. We can see this because if $(\mathbf{v}, \lambda)$ is an eigenpair for $A^T A$ and $\mathbf{v}$ is a unit vector then

$$||A\mathbf{v}||^2 = (A\mathbf{v})^T(A\mathbf{v}) = \mathbf{v}^T A^T A\mathbf{v} = \lambda\mathbf{v}^T\mathbf{v} = \lambda$$

and so since $||A\mathbf{v}||^2 \geq 0$ then $\lambda \geq 0$ and so there is a real square root of the eigenvalues of $A^T A$ or AA^T.

Thus we have discovered that if the singular value decomposition exists then the non-zero elements along the diagonal of the Σ matrix are the square roots of the eigenvalues of $A^T A$ or AA^T. Let us order the σ_i for $i = 1, \ldots, k$ in decreasing order and define $\sigma_i = 0$ for $i = k+1, \ldots, s$ where $s = min\{n, p\}$. Then the σ_i for $i = 1, \ldots, s$ are called the singular values of the matrix A. These are the square roots of the eigenvalues of $A^T A$ or AA^T. Now instead of talking about eigenvectors of $A^T A$ and AA^T corresponding to non-zero eigenvalues we can talk about the eigenvectors corresponding to non-zero singular values. We are simply using our new definition of the non-zero singular values being the square roots of the non-zero eigenvalues of $A^T A$ and AA^T.

Fundamental Subspaces

Besides allowing us to define the singular values of a matrix the symmetric forms $A^T A$ and AA^T are intimately related to the four fundamental subspaces of the linear transformation defined by the matrix A. The following theorem will not only display this relationship but will give us the last theorem we need before we can prove the existence of the singular value decomposition.

Theorem 8

The unit-eigenvectors of AA^T corresponding to non-zero singular values of A can form an orthonormal basis for the column space of A.

Proof. Let $\mathcal{B} = \{\mathbf{v_1}, \ldots, \mathbf{v_p}\}$ be an orthonormal eigenbasis for $\mathbb{R}^p$ formed from the eigenvectors of $A^T A$. We know we can do this since $A^T A$ is $p \times p$ and a symmetric matrix and so it has p orthogonal eigenvectors. Then $A\mathbf{v_i}$ and $A\mathbf{v_j}$ are orthogonal for $i \neq j$ since

$$(A\mathbf{v_i})^T(A\mathbf{v_j}) = \mathbf{v_i}^T A^T A\mathbf{v_j} = \lambda\mathbf{v_i}^T\mathbf{v_j} = 0$$

since distinct eigenvectors are orthogonal. Furthermore we previously established that the non-zero singular values are the lengths of the $A\mathbf{v_i}$ where the $\mathbf{v_i}$ are the unit-eigenvectors of A^TA corresponding to non-zero singular values. This follows because

$$||A\mathbf{v_i}||^2 = (A\mathbf{v_i})^T(A\mathbf{v_i}) = \mathbf{v_i}^T A^T A\mathbf{v_i} = \lambda_i \mathbf{v_i}^T\mathbf{v_i} = \lambda_i = \sigma_i^2$$

and so

$$||A\mathbf{v_i}|| = \sigma_i.$$

Furthermore if $\mathbf{v_i}$ is a unit-eigenvector of A^TA corresponding to an eigenvalue of zero then by the same logic

$$||A\mathbf{v_i}|| = 0$$

and hence $A\mathbf{v_i} = \mathbf{0}$.

If there are k non-zero singular values then let us order our basis in decreasing order such that $\{\mathbf{v_1}, \ldots, \mathbf{v_k}\}$ are the eigenvectors of A^TA corresponding to non-zero singular values. Then $\{A\mathbf{v_1}, \ldots, A\mathbf{v_k}\}$ is a set of orthogonal (and non-zero) vectors in the columns space of A. We want to show that $\{A\mathbf{v_i}\}_{i=1}^k$ is a basis for the column space of A. To see this consider a vector $\mathbf{x}$ in the column space of A then

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

for some $\mathbf{y} \in \mathbb{R}^p$ and using our eigenbasis $\{\mathbf{v_i}\}$ then

$$\mathbf{y} = c_1\mathbf{v_1} + \cdots + c_p\mathbf{v_p}$$

and so

$$\mathbf{x} = A\mathbf{y} = c_1A\mathbf{v_1} + \cdots + c_pA\mathbf{v_p} = c_1A\mathbf{v_1} + \cdots + c_kA\mathbf{v_k} + \mathbf{0} + \cdots + \mathbf{0}$$

meaning that if $\mathbf{x}$ is in the column space of A then it is in the span of $\{\mathbf{Av_i}\}_{i=1}^k$ and hence $\{\mathbf{Av_i}\}_{i=1}^k$ is a basis for the column space of A since it is a linearly independent set spanning the space. Moreover it is an orthogonal basis since the $A\mathbf{v_i}$ are mutually orthogonal.

Now we already discovered that if $\mathbf{v_i}$ is an eigenvector of A^TA then $A\mathbf{v_i}$ is an eigenvector of AA^T. Thus the basis

$$\{\mathbf{Av_i}\}_{i=1}^k$$

is a set of orthogonal eigenvectors of AA^T. Hence if we divide these vectors by their norms (remember $||A\mathbf{v_i}|| = \sigma_i$) to obtain the set

$$\{\mathbf{u_i}\}_{i=1}^k = \left\{\frac{A\mathbf{v_i}}{\sigma_i}\right\}_{i=1}^k$$

then these unit-eigenvectors of AA^T are an orthonormal basis for the column space of A.

■

Now if the unit-eigenvectors of AA^T can form a basis for the column space of A then the unit-eigenvectors of $A^TA = (A^T)(A^T)^T$ are a basis for the column space of A^T, i.e. the row space of A.

Thus the two symmetric forms A^TA and AA^T give us nice bases for the four fundamental subspaces of A. If $\{\mathbf{u_i}\}_{i=1}^n$ are the unit eigenvectors of AA^T and A has k non-zero singular values then $\{\mathbf{u_i}\}_{i=1}^k$ is an orthonormal basis for the column space of A and $\{\mathbf{u_i}\}_{i=k+1}^n$ is an orthonormal basis for the null space of A. On the other hand if $\{\mathbf{v_i}\}_{i=1}^p$ are the unit eigenvectors of A^TA then $\{\mathbf{v_i}\}_{i=1}^k$ is an ortho-normal basis for the row space of A and $\{\mathbf{v_i}\}_{i=k+1}^p$ is an ortho-normal basis for the null space of A^T.

Notice that the previous discussion implies that the rank of the matrix A is the number of non-zero singular values k since the bases of the row and column spaces have k vectors and hence the dimensions of the row and column space is k.

1.2.2 Existence and Uniqueness

We now have the requisite groundwork laid necessary to give a proof of the singular value decomposition.

Theorem 9

Let A be an $n \times p$ rank k real matrix. Then there is an $n \times p$ matrix Σ that is rectangular diagonal (zeros off the main diagonal), an $n \times n$ orthogonal matrix U and $p \times p$ orthogonal matrix p such that $A = U\Sigma V^T$.

Proof. Let $\{\mathbf{v_i}\}_{i=1}^{p}$ be an orthonormal eigenbasis (eigenvectors of A^TA) for $\mathbb{R}^p$ such that the vectors $\{\mathbf{v_i}\}_{i=1}^{k}$ are associated with nonzero singular values $\sigma_1, \ldots, \sigma_k$ where

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0$$

and $\{\mathbf{v_i}\}_{i=k+1}^{p}$ are associated with eigenvalues of zero. Then as we noticed before the set $\{\mathbf{u_i}\}_{i=1}^{k}$ where

$$\mathbf{u_i} = \frac{A\mathbf{v_i}}{\sigma_i} \text{ for } i = 1, \ldots, k$$

is an orthonormal basis for the column space of A where each $\mathbf{u_i}$ is an eigenvector of AA^T. This relationship implies that for $i = 1, \ldots, k$

$$A\mathbf{v_i} = \sigma_i \mathbf{u_i}$$

and

$$A\mathbf{v_i} = \mathbf{0}$$

for $i = k+1, \ldots, p$ as we noticed previously.

Now extend the $\mathbf{u_i}$ to be an orthonormal basis for $\mathbb{R}^n$ then if

$$U = [\mathbf{u_1} \cdots \mathbf{u_n}] \text{ and } V = [\mathbf{v_1} \cdots \mathbf{v_p}]$$

then U and V are orthogonal matrices and

$$AV = [A\mathbf{v_1} \quad \cdots \quad A\mathbf{v_p}] = [\sigma_1\mathbf{u_1} \quad \cdots \quad \sigma_k\mathbf{u_k} \quad \mathbf{0} \quad \cdots \quad \mathbf{0}].$$

Furthermore if

$$\Sigma = \left[\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right] \text{ where } D = diag(\sigma_1, \ldots, \sigma_k)$$

and D is $k \times k$ then

$$\begin{aligned} U\Sigma &= [\mathbf{u_1} \quad \cdots \quad \mathbf{u_n}] \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_k & \\ \hline & 0 & & 0 \end{array}\right] \\ &= [\sigma_1\mathbf{u_1} \quad \cdots \quad \sigma_k\mathbf{u_k} \quad \mathbf{0} \quad \cdots \quad \mathbf{0}] = AV \end{aligned}$$

Thus $AV = U\Sigma$ and since V is orthogonal then $A = U\Sigma V^T$.

■

We should note that while we have talked about *the* singular value decomposition there is no one unique singular value decomposition. Remember that if $s = min\{n, p\}$ then there are s singular values and some subset of those singular values are nonzero. Normally we list the singular values in decreasing order down the main diagonal of Σ such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_s \geq 0$. So while technically we could reorder the singular values and corresponding vectors of U and V we will not do so and consider them the be non-increasing down the main diagonal of Σ. Note however that if $\sigma_i = \sigma_j$ for some $i \neq j$ then we could

switch the right singular vectors $\mathbf{v_i}$ and $\mathbf{v_j}$ as long as we switched the corresponding left singular vectors $\mathbf{u_i}$ and $\mathbf{v_i}$. Furthermore we may multiply both its right singular vector $\mathbf{v_i}$ and its left singular vector $\mathbf{u_i}$ by -1 and still have a valid decomposition.

Note that the right and left singular vectors associated with degenerate singular values $\sigma_i = 0$ (or those not associated with singular values at all) there is not much unique about these. These vectors were chosen only entirely to complete the orthonormal eigenbases of A^TA and AA^T and thus there is quite a lot of freedom in their specification. Thus the the right and left singular vectors associated with non-degenerate singular values are specified up to sign and possible permutation with other vectors associated with an equal singular value. However those singular vectors associated with singular values of zero (or not associated with a singular value at all) are quite arbitrary.

The singular value decomposition is a quite beautiful decomposition. It is the closest we can come to orthogonally diagonalizing an arbitrary matrix. Furthermore it displays the rank of the matrix while giving bases for the four fundamental subspaces. The decomposition allows us to view the action of the matrix A in very simple terms. In terms of a linear transformation the action of any arbitrary matrix A can be seen as the product of an orthogonal, rectangular diagonal and then orthogonal transformation. Orthogonal transformations are nice because they are roto-inversions and diagonal (or basically diagonal) transformations are nice because they are scalings or collapsings of the space. Finally, as we noted for the orthogonal diagonalization of symmetric matrices, the factorization makes matrix algebra easy because manipulating orthogonal and diagonal matrices is easy.

Example 8

Using the matrix

$$A = \begin{bmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{bmatrix}$$

from Example 7 we may compute the SVD of A quickly.

Σ is the rectangularly diagonal matrix with singular values decreasing on its main diagonal so since the non-zero eigenvalues of A^TA and AA^T are 360 and 90 then

$$\Sigma = \begin{bmatrix} 6\sqrt{10} & 0 \\ 0 & 3\sqrt{10} \\ 0 & 0 \end{bmatrix}$$

Let us take the matrix V as the previous one except with unit length columns. Then

$$V = \frac{1}{27\sqrt{10}} \begin{bmatrix} 81 & -27 \\ 27 & 81 \end{bmatrix}$$

and

$$AV = \sqrt{10} \begin{bmatrix} 2 & 2 \\ 4 & 1 \\ 4 & -2 \end{bmatrix} = U\Sigma$$

so

$$\sqrt{10} \begin{bmatrix} 2 & 2 \\ 4 & 1 \\ 4 & -2 \end{bmatrix} = \begin{bmatrix} \mathbf{u_1} & \mathbf{u_2} & \mathbf{u_3} \end{bmatrix} \begin{bmatrix} 6\sqrt{10} & 0 \\ 0 & 3\sqrt{10} \\ 0 & 0 \end{bmatrix}$$

meaning

$$\mathbf{u_1} = \frac{1}{3} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \text{ and } \mathbf{u_2} = \frac{1}{3} \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}$$

and choosing $\mathbf{u_3}$ to complete a basis for $\mathbb{R}^n$ as

$$\mathbf{u_3} = \frac{1}{3} \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}$$

then

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

It is easy to verify that with these matrices

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

with U and V orthogonal.

◀

Before we move on to the next topics let us take a minute to discuss what the singular value decomposition says about the geometry of a linear transformation. We know that if an $n \times p$ matrix A is rank k and has a singular value decomposition of $A = U\Sigma V^T$ then $\{\mathbf{v_1}, \ldots, \mathbf{v_k}\}$ is a basis for the row space of A, $\{\mathbf{v_{k+1}}, \ldots, \mathbf{v_p}\}$ is a basis for the null space of A, $\{\mathbf{u_1}, \ldots, \mathbf{u_k}\}$ is a basis for the column space of A and $\{\mathbf{u_{k+1}}, \ldots, \mathbf{u_n}\}$ is a basis for the null space of A^T.

Let us consider the transformation $T : \mathbb{R}^p \to \mathbb{R}^n$ defined by

$$T(\mathbf{x}) = A\mathbf{x} = U\Sigma V^T\mathbf{x}.$$

Remember that the map A knocks the orthonormal $\{\mathbf{v_i}\}_{i=1}^k$ basis for the row space into the orthogonal basis $\{\sigma_i\mathbf{u_i}\}_{i=1}^k$ for the column space, that is, $A\mathbf{u_i} = \sigma_i\mathbf{v_i}$ for $i = 1, \ldots, k$. This makes since because the dimension of the column space and row space are equal to each other and to the rank. Thus for a vector $\mathbf{x} \in \mathbb{R}^p$,

$$\mathbf{x} = x_1\mathbf{v_1} + \cdots + x_k\mathbf{v_k}$$

under the $\{\mathbf{v_i}\}_{i=1}^k$ basis and so

$$\begin{aligned} A\mathbf{x} &= x_1A\mathbf{v_1} + \cdots + x_kA\mathbf{v_k} \\ &= x_1\sigma_1\mathbf{u_1} + \cdots + x_k\sigma_k\mathbf{u_k} \end{aligned}$$

under the $\{\mathbf{u_i}\}_{i=1}^n$ basis.

The action of any real $n \times p$ matrix A may be viewed as follows. First we left multiply by V^T which is equivalent to transforming $\mathbf{x}$ from its standard basis representation to a representation in the basis $\{\mathbf{v_i}\}_{i=1}^p$. Let

$$x = (x_1, \ldots, x_k, \overbrace{0, \ldots, 0}^{p-k})^T$$

be the V basis representation of $\mathbf{x}$. Then Σ maps

$$(x_1, \ldots, x_k, \overbrace{0, \ldots, 0}^{p-k})^T \text{ to } (\sigma_1x_1, \ldots, \sigma_kx_k, \overbrace{0, \ldots, 0}^{n-k})^T$$

which is precisely a U basis representation of $A\mathbf{x}$. Thus the final left multiplication by U simply takes this U basis representation of $A\mathbf{x}$ are transforms it back into the standard basis representation of $\mathbf{x}$ for $\mathbb{R}^n$. The idea here is that if we restrict ourselves to thinking about the vectors mapped between the row and column spaces then the behavior of any matrix (through the lens of the "correct" bases for these two spaces) is a very simple scaling of the space. This is a very powerful idea and we will do our best to get good mileage out of it.

1.3 Important Applications

The remainder of this chapter will look at two applications of the singular value decomposition. Using the hammer of the SVD makes quick work of what might otherwise have been tedious. The power of the SVD is that it cuts directly to the heart of these applications and allows succinct, intuitive definitions and theorems.

1.3.1 Pseudoinversion and Projection

The SVD makes defining an "inverse" for singular matrices quite simple.

Inverting Rectangular Diagonal Matrices

Consider a diagonal $n \times n$ matrix $D = diag(d_1, \ldots, d_n)$ where $d_i \neq 0$ for all $i = 1, \ldots, n$. Then D is a full rank square matrix and so it must be invertible. Indeed the inverse if easy to find, $D^{-1} = diag(1/d_1, \ldots, 1/d_n)$.

Now consider a rectangularly diagonal $n \times p$ matrix Σ such that

$$\Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_k & \\ \hline & 0 & & 0 \end{array}\right]$$

where $\sigma_i \neq 0$. Now clearly Σ is not generally invertible since it is generally not even square. However consider left multiplication by Σ. For $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$

$$\Sigma\mathbf{x} = \begin{bmatrix} \sigma_1 x_1 \\ \vdots \\ \sigma_k x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Define the matrix E to be the $p \times n$ matrix

$$E = \left[\begin{array}{ccc|c} 1/\sigma_1 & & & \\ & \ddots & & 0 \\ & & 1/\sigma_k & \\ \hline & 0 & & 0 \end{array}\right]$$

then

$$E\Sigma\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Let $\mathbf{y} \in Span\{\mathbf{e_1}, \ldots, \mathbf{e_k}\} = Row(\Sigma)$ where $\mathbf{e_i}$ is the i^{th} column of I_p. Then

$$E\Sigma\mathbf{y} = \mathbf{y}.$$

Let us call E the pesudoinverse of Σ and denote it as Σ^+. Thus for vectors in the row space of Σ then Σ^+ is the inverse map to Σ. That is,

$$\Sigma^+\Sigma\mathbf{y} = \mathbf{y} \text{ for } \mathbf{y} \in Row(\Sigma).$$

However if $x_1 = \cdots = x_k = 0$ and so $\mathbf{x} \notin Row(\Sigma)$ then $\mathbf{x} \in Nul(\Sigma)$ and so $\Sigma\mathbf{x} = \mathbf{0}$ and hence

$$\Sigma^+\Sigma\mathbf{x} = \mathbf{0} \text{ for } \mathbf{x} \in Nul(\Sigma).$$

Similarly $(\Sigma^+)^T = (\Sigma^T)^+$ is the inverse map for Σ^T. Thus if $\mathbf{z}$ is in the column space of Σ then $(\Sigma^+)^T\Sigma^T\mathbf{z} = \mathbf{z}$ or $(\Sigma\Sigma^+)^T\mathbf{z} = \mathbf{z}$. However surely $\Sigma\Sigma^+$ is symmetric (it is a $n \times n$ diagonal matrix) and so

$$\Sigma\Sigma^+\mathbf{z} = \mathbf{z} \text{ for } \mathbf{z} \in Col(\Sigma).$$

On the other hand

$$\Sigma\Sigma^+\mathbf{z} = \mathbf{0} \text{ for } \mathbf{z} \in Nul(\Sigma^T).$$

Now $Row(\Sigma) \subseteq \mathbb{R}^p$ and so consider a vector $\mathbf{x} \in \mathbb{R}^p$. We know that

$$Nul(\Sigma) \oplus Row(\Sigma) = \mathbb{R}^p$$

and that $Nul(\Sigma) = Row(\Sigma)^{\perp}$ meaning

$$\mathbf{x} = \mathbf{x}_{\parallel} + \mathbf{x}_{\perp}$$

where $\mathbf{x}_{\parallel} \in Row(\Sigma)$ and $\mathbf{x}_{\perp} \in Nul(\Sigma)$. Thus

$$\Sigma^{+}\Sigma\mathbf{x} = \Sigma^{+}\Sigma(\mathbf{x}_{\parallel} + \mathbf{x}_{\perp}) = \Sigma^{+}\Sigma\mathbf{x}_{\parallel} + \Sigma^{+}\Sigma\mathbf{x}_{\perp} = \mathbf{x}_{\parallel}$$

as per our previous discussion. Similarly if $\mathbf{y} \in \mathbb{R}^n$ then

$$\Sigma\Sigma^{+}\mathbf{y} = \mathbf{y}_{\parallel}$$

where $\mathbf{y}_{\parallel} \in Col(\Sigma)$.

Thus what is going on is that the map $P_R : \mathbb{R}^p \to \mathbb{R}^p$ defined by

$$P_R(\mathbf{x}) = \Sigma^{+}\Sigma\mathbf{x} \text{ for } \mathbf{x} \in \mathbb{R}^p$$

is the map projecting vectors from $\mathbb{R}^p$ onto the row space of Σ. Similarly the map $P_C : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$P_C(\mathbf{y}) = \Sigma\Sigma^{+}\mathbf{y} \text{ for } \mathbf{y} \in \mathbb{R}^n$$

is the map projecting vectors from $\mathbb{R}^n$ onto the column space of Σ.

It should be clear why we call Σ^{+} the pseudoinverse of Σ. A map Σ is singular either because it is square but not full rank or not square at all. In this case Σ or Σ^T is not injective. This is because $Nul(\Sigma)$ or $Nul(\Sigma^T)$ is nontrivial. Consider the first to be true. If this isn't the case then we can really just transpose the following argument and have basically the same discussion. Then assume the null space of Σ is nontrivial. Thus when we map a vector via left multiplication by Σ we necessarily lose any part of the vector in the null space because all such vectors are mapped to zero. Normally an inverse map would recover the vector originally mapped. However under this singular mapping we can't recover anything mapped from the null space. There are multiple vectors mapped to zero and after the fact we don't know which one of them was originally mapped in any particular case. The pseudoinverse recovers as much of the vector as possible, i.e. the part in the row space. A similar story may be told about Σ^T and $(\Sigma^T)^{+}$.

The idea is that we can't find an inverse map for $\Sigma : \mathbb{R}^p \to \mathbb{R}^n$ however using the first isomorphism theorem for vector spaces we know that

$$\mathbb{R}^p / Ker(\Sigma) \approx Im(\Sigma)$$

or, using the language of vector spaces, $Row(\Sigma)$ is isomorphic to $Col(\Sigma)$ and so if $f : Row(\Sigma) \to Col(\Sigma)$ is the map defined by

$$f(\mathbf{x}) = \Sigma\mathbf{x} \text{ for } \mathbf{x} \in Row(\Sigma)$$

then f^{-1} is the map $f^{-1} : Col(\Sigma) \to Row(\Sigma)$ defined by

$$f^{-1}(\mathbf{x}) = \Sigma^{+}\mathbf{x} \text{ for } \mathbf{x} \in Col(\Sigma).$$

Inverting Arbitrary Matrices

It may seem strange that up to this point we have only defined the pseudoinverse (properly called the Moore-Penrose pseudoinverse) for rectangularly diagonal matrices. However we know that, via the singular value decomposition, every matrix is transformable with the proper orthogonal transformations to a rectangularly diagonal matrix. Thus if A is a $n \times p$ matrix and if the SVD of A is

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

then we defined the pseudoinverse of A to be

$$A^{+} = (U\Sigma V^T)^{+} = V\Sigma^{+}U^T$$

Note that if $A = U\Sigma V^T$ is an invertible matrix then

$$A^{-1} = V\Sigma^{-1}U^T$$

and so our above definition makes some sense. To show that this definition is really what we want we will show that $A^{+}A$ is the projection matrix onto the row space of A and AA^{+}

is the projection matrix onto the column space of A. To see this first note that if A is rank k then

$$\Sigma^+\Sigma = diag(\overbrace{1,\dots,1}^{k},\overbrace{0,\dots,0}^{p-k}) = \left[\begin{array}{c|c} I_k & 0 \\ \hline 0 & 0 \end{array}\right]$$

and

$$\Sigma\Sigma^+ = diag(\overbrace{1,\dots,1}^{k},\overbrace{0,\dots,0}^{n-k}) = \left[\begin{array}{c|c} I_k & 0 \\ \hline 0 & 0 \end{array}\right].$$

Thus we have that

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

and since the first k columns of V are a basis for the row space of A whereas the last $p-k$ are a basis for the null space of A and so if $\mathbf{x} \in \mathbb{R}^p$ then

$$A^+A\mathbf{x} = V\Sigma^+\Sigma V^T\mathbf{x} = V\left[\begin{array}{c|c} I_k & 0 \\ \hline 0 & 0 \end{array}\right]V^T\mathbf{x}$$

which simply makes a change of basis to the V basis with $V^T\mathbf{x}$ while $\Sigma^+\Sigma$ does nothing to the first k components of these new vectors (the parts in $Row(A)$), kills last $p-k$ components (those in $Nul(A)$), and then changes back to the standard basis with left multiplication by V. Thus A^+A is the projection matrix onto the row space of A because it only keeps the parts in the row space.

Similarly we can show that

$$AA^+\mathbf{x} = U\Sigma\Sigma^+U^T\mathbf{x}$$

which changes to the U basis for $\mathbb{R}^n$, kills the components of vectors in the null space of A^T, doesn't touch the vectors in $Col(A)$, and then changes back to the standard basis. Thus AA^+ keeps only the parts in $Col(A)$ and is thus the projection from $\mathbb{R}^n$ onto $Col(A)$.

Then as we discussed with the pseudoinverse of rectangular diagonal matrices, the pseudoinverse of any $n \times p$ real matrix A is the inverse map when we restrict ourselves to look at vectors mapped between $Row(A)$ and $Col(A)$. Otherwise it is the projection down onto the appropriate subspace in an attempt to recover as much of the originally mapped vector as possible.

Example 9

Consider the points $(4,0),(0,2)$ and $(1,1)$ in the xy-plane. We would like to fit a linear function

$$\hat{y} = \beta_1 x + \beta_0$$

to the points such that given a pair (x,y) then $\hat{y}$ is a good prediction of y. Then our equation should, if it predicts perfectly, give us that

$$\begin{cases} 0 &= \beta_1 * 4 + \beta_0 \\ 2 &= \beta_1 * 0 + \beta_0 \\ 1 &= \beta_1 + \beta_0 \end{cases}$$

or, in matrix form,

$$\begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \mathbf{y} = X\beta = \begin{bmatrix} 4 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix}.$$

Thus to solve for $(\beta_1,\beta_0)^T$ we should find

$$\beta = X^{-1}\mathbf{y}.$$

However X isn't invertible however we can use the pseduoinverse of X and so

$$\beta = X^+\mathbf{y} = \frac{1}{13}\begin{bmatrix} -6 \\ 23 \end{bmatrix}$$

meaning our function is

$$\hat{y} = \frac{-6}{13}x + \frac{23}{13}.$$

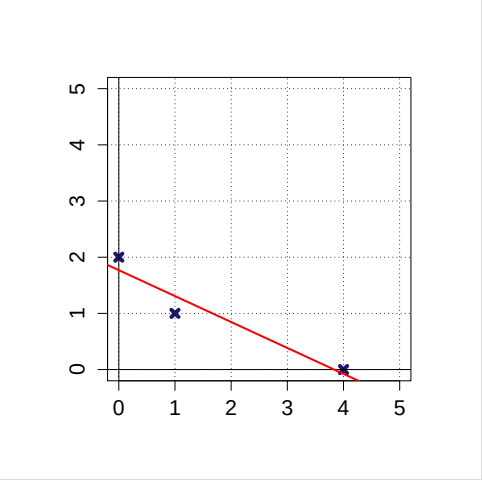

Figure 1.4: The best fit for the points (blue) line is in red.

◀

1.3.2 Quadratic Forms and Principal Axes

The final application of the material in this chapter is finding the principal axes of ellipses. We will see that this subject is the heart of principal components analysis and will hopefully then be a nice transition into our next chapter on precisely this analysis.

Ellipses as Quadratic Forms

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

$$Q(\mathbf{x}) = Q((x_1, \ldots, x_n)) = \sum_{i=1}^{n} \sum_{j \leq i} a_{i,j} x_i x_j$$

for any $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. It is a polynomial $Q((x_1, \ldots, x_n))$ in the variables $x_1, \ldots, x_n$ that is homogeneous of degree 2 such that

$$Q(t\mathbf{x}) = t^2 Q(\mathbf{x})$$

for some scalar $t \in \mathbb{R}$. This means that the degrees of each of its constituent monomials is two.

Example 10

The quadratic form $S : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$S(\mathbf{x}) = S((x, y)) = (x + y)^2 = x^2 + 2xy + y^2$$

defines a prototypical quadratic form.

Notice that generally

$$(x_1 + \cdots + x_n)^2 = \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} \sum_{j<i} x_i x_j = \sum_{i=1}^{n} \sum_{j \leq i} (2 - \delta_{i,j}) x_i x_j$$

where $\delta_{i,j}$ is Kronecker's Delta. Thus if $P : \mathbb{R}^n \to \mathbb{R}$ is a function defined by

$$P(\mathbf{x}) = P((x_1 + \cdots + x_n)) = (x_1 + \cdots + x_n)^2$$

then P is a quadratic form where $a_{i,j} = 2 - \delta_{i,j}$.

◀

An equivalent way to define a quadratic form $Q : \mathbb{R}^n \to \mathbb{R}$ is

$$Q(\mathbf{x}) = Q((x_1, \ldots, x_n)) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j} x_i x_j$$

where transforming our first definition we have that

$$c_{i,j} = \begin{cases} a_{i,j} & i = j \\ \frac{1}{2} a_{i,j} \text{ or } \frac{1}{2} a_{j,i} & i \neq j. \end{cases}$$

We halve the $a_{i,j}$ when $i \neq j$ because this second form counts both $x_i x_j$ and $x_j x_i$ in comparison to our first definition which counts only one of these.

Consider a matrix C such that $C_{i,j} = c_{i,j}$ then for $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$

$$C\mathbf{x} = \begin{bmatrix} c_{1,1}x_1 + & \cdots & +c_{1,n}x_n \\ \vdots & & \vdots \\ c_{n,1}x_1 + & \cdots & +c_{n,n}x_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} c_{1,j}x_j \\ \vdots \\ \sum_{j=1}^{n} c_{n,j}x_j \end{bmatrix}$$

and so

$$\begin{aligned} \mathbf{x}^T C\mathbf{x} &= (x_1, \ldots, x_n) \begin{bmatrix} \sum_{j=1}^{n} c_{1,j}x_j \\ \vdots \\ \sum_{j=1}^{n} c_{n,j}x_j \end{bmatrix} \\ &= x_1 \sum_{j=1}^{n} c_{1,j}x_j + \cdots + x_n \sum_{j=1}^{n} c_{n,j}x_j \\ &= \sum_{j=1}^{n} x_1 c_{1,j}x_j + \cdots + \sum_{j=1}^{n} x_n c_{n,j}x_j \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j}x_i x_i = Q(\mathbf{x}). \end{aligned}$$

Thus if C is a matrix of coefficients as defined above then any quadratic form Q may be defined as $Q(\mathbf{x}) = \mathbf{x}^T C\mathbf{x}$. Furthermore since $C_{i,j} = c_{i,j} = c_{j,i} = C_{j,i}$ then C is symmetric. Hence any quadratic form may be represented as

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

where C is a symmetric matrix.

Example 11

Using the quadratic form P from Example 10,

$$P(\mathbf{x}) = \mathbf{x}^T \mathbf{1}\mathbf{1}^T \mathbf{x} = \mathbf{x}^T \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \mathbf{x} \text{ for } \mathbf{x} \in \mathbb{R}^n$$

where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$ is a vector of $1's$.

◀

For any quadratic form Q defined by

$$Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} \text{ for } \mathbf{x} \in \mathbb{R}^n$$

and some $n \times n$ symmetric matrix A then the locus of points defined, for some scalar c, by

$$Q(\mathbf{x}) = c^2$$

is either an ellipsoid or hyperboloid. These are the n-dimensional generalizations of hyperbolas and ellipses. For convenience we assume that $c > 0$ such that $c = \sqrt{c^2}$.

Example 12

Consider the quadratic form $R : \mathbb{R}^2 \to \mathbb{R}$ defined for $\mathbf{x} = (x, y)^T \in \mathbb{R}^2$ by

$$R(\mathbf{x}) = R((x, y)) = \frac{1}{4}x^2 + y^2.$$

We plot $R(\mathbf{x}) = \frac{1}{4}x^2 + y^2 = 1$ below in Figure 1.4.

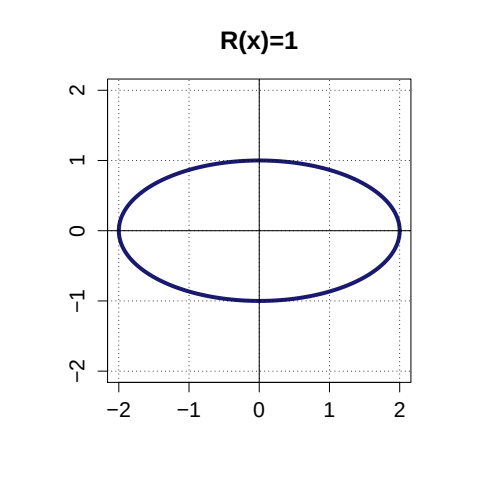

Figure 1.5: The quadratic form $R(\mathbf{x})$ defines an ellipse in $\mathbb{R}^2$.

◀

Principal Axes of Ellipsoids

For our purposes we are interested in finding the principal axes of ellipsoids. We define the principal axes of an ellipsoid to be its axes of symmetry. We will see that this definition will need some clarification later but for now it will suffice.

The principal axes of the quadratic form in Example 12 are easy to see. They are simply the x and y axes themselves. This is true because there are no mixed xy terms in the defining polynoimal,

$$R((x, y)) = \frac{1}{4}x^2 + 0xy + y^2.$$

This will generally be true. Consider a quadratic form $Q : \mathbb{R}^n \to \mathbb{R}$ and define an ellipsoid as the locus of points $\mathbf{x} \in \mathbb{R}^n$ where

$$Q(\mathbf{x}) = \mathbf{x}^T C\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j}x_i x_j = a^2$$

for some constant $a > 0$ and symmetric matrix C where $C_{i,j} = c_{i,j}$. If there are no mixed terms $x_i x_j$ where $i \neq j$, i.e. $c_{i,j} = 0$ when $i \neq j$, in the defining formula for Q then the principal axes of the ellipsoid $Q(\mathbf{x}) = a^2$ align with the axes of the space.

Theorem 10

The principal axes of a quadratic form Q align with the spaces axes if and only if there are no mixed terms in the defining polynomial.

Proof.
$\Rightarrow$ To see this notice that if the principal axes align with the space axes then definitionally if we reflect any point $\mathbf{y}$ on the ellipsoid over any of the axes we should end up back on the ellipsoid. If this were not true then the ellipse wouldn't have symmetry over the space's axes. Thus if $\mathbf{y} = (y_1, \ldots, y_k, \ldots, y_n)^T \in \mathbb{R}^n$ with $1 \leq k \leq n$ is a point on the ellipsoid (i.e. $Q(\mathbf{y}) = a^2$) then define the point $\hat{\mathbf{y}}$ to be

$$\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T = (y_1, \ldots, -y_k, \ldots, y_n)^T \in \mathbb{R}^n.$$

This point better be on the ellipsoid because it is the reflection of $\mathbf{y}$ over one of the axes. Thus $Q(\hat{\mathbf{y}}) = a^2$ and so

$$Q(\mathbf{y}) = Q(\hat{\mathbf{y}})$$

and hence

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j} y_i y_j = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j} \hat{y}_i \hat{y}_j.$$

for any $\mathbf{y}$ on the ellipsoid. Now surely the constant a does not change the principal axes of the quadratic form. It only changes the length of the axes. Thus let us choose a constant b so to make finding the principal axes nice. Consider some $\mathbf{x} \in \mathbb{R}^n$ such that the only non-zero elements of $\mathbf{x}$ are x_k and x_s for $s \neq k$ meaning

$$\mathbf{x} = (0, \ldots, 0, x_s, 0, \ldots, 0, x_k, 0, \ldots, 0)$$

and

$$\hat{\mathbf{x}} = (0, \ldots, 0, x_s, 0, \ldots, 0, -x_k, 0, \ldots, 0).$$

and so if $Q(\mathbf{x}) = b^2$ for some scalar $b > 0$ then $Q(\hat{\mathbf{x}}) = b^2$. Thus

$$\begin{aligned} Q(\mathbf{x}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j} x_i x_j \\ &= c_{s,s} x_s^2 + c_{s,k} x_s x_k + c_{k,s} x_k x_s + c_{k,k} x_k^2 \\ &= c_{s,s} x_s^2 + 2c_{s,k} x_s x_k + c_{k,k} x_k^2 \end{aligned}$$

and

$$\begin{aligned} Q(\hat{\mathbf{x}}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j} \hat{x}_i \hat{x}_j \\ &= c_{s,s} \hat{x}_s^2 + c_{s,k} \hat{x}_s \hat{x}_k + c_{k,s} \hat{x}_k \hat{x}_s + c_{k,k} \hat{x}_k^2 \\ &= c_{s,s} x_s^2 - c_{s,k} x_s x_k - c_{k,s} x_k x_s + c_{k,k} x_k^2 \\ &= c_{s,s} x_s^2 - 2c_{s,k} x_s x_k + c_{k,k} x_k^2 \end{aligned}$$

and so since $Q(\mathbf{x}) = Q(\hat{\mathbf{x}})$ then

$$c_{s,s} x_s^2 + 2c_{s,k} x_s x_k + c_{k,k} x_k^2 = c_{s,s} x_s^2 - 2c_{s,k} x_s x_k + c_{k,k} x_k^2$$

or

$$4c_{s,k} x_k x_s = 0.$$

It follows that since x_k and x_s are nonzero then $c_{s,k} = 0$. Since we can do this for any s and k where $s \neq k$ then

$$c_{i,j} = 0 \text{ for } i \neq j.$$

This means that $C = diag((c_{i,i}))$ because the off diagonal elements are zero. Thus if the principal axes are the spaces axes then there are no mixed terms.

$\Leftarrow$ Notice that this argument can be run backwards. Clearly if there are no mixed terms then

$$Q(\mathbf{x}) = c_{1,1} x_1^2 + \cdots + c_{n,n} x_n^2$$

and

$$Q(\hat{\mathbf{x}}) = Q(\mathbf{x})$$

and so the quadratic form has symmetry over the space axes. Thus the principal axes are the axes of the space

■

So we have determined the principal axes for quadratic forms determined by diagonal matrices (they are simply the axes of the space). Now to determine the principal axes for other ellipses we will operate as is the theme of this chapter and leverage the fact that symmetric matrices are orthogonally diagonalizable.

Consider an arbitrary quadratic form $P : \mathbb{R}^n \to \mathbb{R}$ defined by

$$P(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} \text{ for } \mathbf{x} \in \mathbb{R}^n$$

and some symmetric A. Then A is diagonalizable by an orthogonal similarity transformation

$$D = U^T AU$$

where U is an orthogonal matrix whose columns are eigenvectors of A and $D = diag(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix with the eigenvalues of A on its diagonal. Let us make a change of basis (change of variable)

$$\mathbf{y} = U^T\mathbf{x}$$

such that $\mathbf{y}$ is the representation of $\mathbf{x}$ in the eigenbasis of A defined by the columns of U. Then

$$\mathbf{x} = U\mathbf{y}$$

and so

$$\mathbf{x}^T A\mathbf{x} = (U\mathbf{y})^T A(U\mathbf{y}) = \mathbf{y}^T U^T AU\mathbf{y} = \mathbf{y}^T D\mathbf{y}.$$

So since D is a diagonal matrix defining the quadratic form under the U basis then the principal axes of the ellipsoid $P(\mathbf{y}) = c^2$ under this basis are the axes of the space. However the axes of this space are not the standard basis axes they are the axes defined by the eigenbasis of A. Thus the principal axes are those axes defined by the orthonormal eigenvectors of A.

Furthermore under the eigenbasis U then

$$P(\mathbf{y}) = \mathbf{y}^T D\mathbf{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2$$

and so if we call the i^{th} principal axis the i^{th} eigenvector of A (orderd descending according to eigenvalues) then the length of the i^{th} principal axis for $P(\mathbf{x}) = P(\mathbf{y}) = c^2$ is

$$\frac{c}{\sqrt{\lambda_i}}.$$

This follows because the length of the i^{th} principal axis is simply the distance from the origin to the ellipse along the i^{th} principal axis. It is simply the value y_i such that

$$P((0, \ldots, y_i, \ldots, 0)) = c^2.$$

This means

$$\lambda_i y_i^2 = c^2$$

or

$$y_i = \frac{c}{\sqrt{\lambda_i}}.$$

Example 13

Consider the matrix

$$A = \begin{bmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{bmatrix}$$

from Example 8 and its associated symmetric form

$$A^T A = \begin{bmatrix} 333 & 81 \\ 81 & 117 \end{bmatrix}$$

Then we already determined the unit-eigenvectors of $A^T A$ to be the columns of the matrix

$$V = \frac{1}{27\sqrt{10}} \begin{bmatrix} 81 & -27 \\ 27 & 81 \end{bmatrix}$$

with associated eigenvalues of $\lambda_1 = 360$ and $\lambda_2 = 90$. Thus if $R : \mathbb{R}^2 \to \mathbb{R}$ is a quadratic form defined by

$$R(\mathbf{x}) = \mathbf{x}^T A^T A \mathbf{x} \text{ for } \mathbf{x} \in \mathbb{R}^2$$

and E is the ellipse defined by $R(\mathbf{x}) = 1$ for $\mathbf{x} \in \mathbb{R}^2$. Then the principal axes of this ellipse are determined by unit-eigenvectors and the lengths of these eigenvectors are

$$\frac{1}{6\sqrt{10}} \text{ and } \frac{1}{3\sqrt{10}}$$

notice that these lengths are

$$\frac{1}{\sqrt{\lambda_i}} = \frac{1}{\sigma_i}$$

where σ_i is the i^{th} singular value of A.

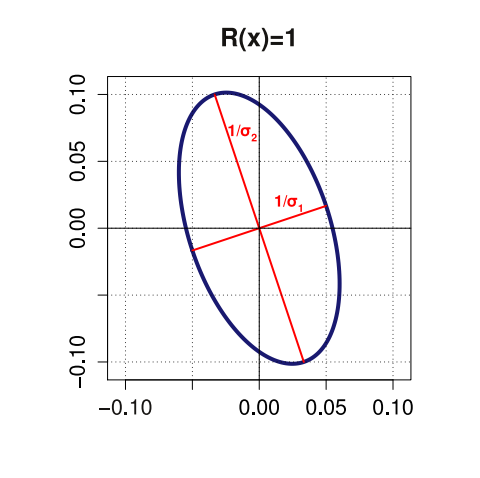

Figure 1.6: The quadratic form defined in Example 13. The lines are the axes of the ellipse.

◀

Although we defined *the* principal axes to be the axes of symmetry there may be more than one set of axes for the space over which the ellipsoid is symmetric. The ellipse defined by

$$\mathbf{x}^T I \mathbf{x} = 1$$

is one for which *any* orthogonal basis of the space would define a set of principal axes. Thus as we saw with the singular value decomposition if there are repeated eigenvalues then there is some degree of choice in the eigenvectors chosen. The important point is that we can always find a new orthonormal basis for the space (a new set of axes) with respect to which the ellipse is symmetric. We call any such set a set of principal axes of the ellipse.

The last fact we need cover is that a quadratic form can define an ellipse if and only if the eigenvalues of its defining matrix are all non-negative. We call matrices with non-negative eigenvalues positive semi-definite.

Theorem 11

A quadratic form Q defined by $Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ for some $n \times n$ symmetric matrix A and $\mathbf{x} \in \mathbb{R}^n$ defines an ellipse $Q(\mathbf{x}) = c^2$ for $\mathbf{x} \in \mathbb{R}^n$ and $c > 0$ if and only if the eigenvalues of A are non-negative.

From the previous discussion we know that there is some basis under which the quadratic form Q is defined by a diagonal matrix $D = diag(\lambda_1, \ldots, \lambda_n)$ where the λ_i are the eigenvalues of A and so there is some basis where

$$Q(\mathbf{x}) = \mathbf{x}^T D\mathbf{x} = \lambda_1 x_1^2 + \cdots + \lambda_n x_n^2.$$

The equation $Q(\mathbf{x}) = c^2$ will define an ellipsoid if and only if the λ_i are non-negative which is true if and only if the eigenvalues of A are non-negative.

■

Notice that for a quadratic form Q we have $Q(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$ if and only if the matrix defining Q is positive semi-definite. Surely if all of the eigenvalues λ_i are non-negative then, looking at our above formula, $Q(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. Furthermore if there is some k such that λ_k that is negative then we can pick a $\mathbf{x}$ such that $Q(\mathbf{x}) < 0$ by making the x_k the only non-zero element of $\mathbf{x}$.

Chapter 2

Principal Components Analysis

This chapter serves to introduce principal components analysis. Principal components analysis (PCA) is widely popular multivariate method often used in exploratory data analysis. With the information in Chapter 1 understood the development of PCA is not a very laborious ordeal. This chapter will introduce some of the statistical notions needed to understand PCA, introduce the method itself, and give examples of the method both in theory and through case studies.

2.1 Multivariate Generalizations

The first order of business is to extend some of the familiar univariate statistical notions into a multivariate setting. Thus instead of considering for some sample space Ω a random scalar $\mathcal{X} : \Omega \to \mathbb{R}$ we consider a random vector $\boldsymbol{\mathcal{X}} : \Omega \to \mathbb{R}^p$ such that

$$\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \cdots, \mathcal{X}_p)^T$$

where $\{\mathcal{X}_i\}_{i=1}^p$ are random scalars. Similarly to random vectors we may consider a random matrix

$$Y : \Omega \to \mathbb{R}^{n \times p} \text{ where } Y_{i,j} = \mathcal{Y}_{i,j}$$

for some set of random scalars $\{\mathcal{Y}_{i,j}\}_{i=1\ j=1}^{n\ p}$.

2.1.1 Covariance and Correlation

Covariance and Correlation Matrices

One of the most important concepts in multivariate statistics is the variance-covariance matrix. Also known as the covariance or dispersion matrix it is the multivariate generalization of variance. As with the univariate variance we may define it both for the theoretical probability distribution of a population as well as an estimate obtained by sampling from the population.

Consider a random vector $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)^T$ comprised of p random scalars $\{\mathcal{X}_i\}_{i=1}^p$. We define the variance-covariance matrix $\text{var}[\boldsymbol{\mathcal{X}}] = \Sigma$ to be the matrix such that

$$\Sigma_{i,j} = \text{cov}[\mathcal{X}_i, \mathcal{X}_j].$$

We can write this in another form as

$$\Sigma = \text{E}[(\boldsymbol{\mathcal{X}} - \text{E}[\boldsymbol{\mathcal{X}}])(\boldsymbol{\mathcal{X}} - \text{E}[\boldsymbol{\mathcal{X}}])^T]$$

where the expectation of a random matrix (or vector) is the element-wise expectation and, furthermore, multiplication of random matrices (or vectors) is defined in precise analogy to that of real matrices (or vectors).

The equality of the two definitions may be seen because

$$\Sigma_{i,j} = \text{E}\left[(\boldsymbol{\mathcal{X}} - \text{E}\left[\boldsymbol{\mathcal{X}}\right])_i(\boldsymbol{\mathcal{X}} - \text{E}\left[\boldsymbol{\mathcal{X}}\right])_j\right] = \text{E}\left[(\mathcal{X}_i - \text{E}\left[\mathcal{X}_i\right])(\mathcal{X}_j - \text{E}\left[\mathcal{X}_j\right])\right] = \text{cov}\left[\mathcal{X}_i, \mathcal{X}_j\right].$$

In either case the covariance of $\mathcal{X}_i$ and $\mathcal{X}_j$, denoted σ_{x_i,x_j}, is the $(i,j)^{th}$ entry of the covariance matrix Σ.

Let us have taken n observations of $\mathfrak{X}$. This is tantamount to n simultaneous observations of each of the p random variables $\mathcal{X}_i$. We may similarly define an unbiased estimator of Σ as the sample covariance matrix S such that

$$S_{i,j} = s_{\mathcal{X}_i,\mathcal{X}_j}$$

where $s_{\mathcal{X}_i,\mathcal{X}_j}$ is the univariate sample covariance between $\mathcal{X}_i$ and $\mathcal{X}_j$. Then leveraging the univariate definition for sample covariance we have that

$$S_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{k,i} - m_i)(x_{k,j} - m_j)$$

where $x_{s,t}$ is the s^{th} observation of $\mathcal{X}_t$ and $m_i = \frac{1}{n}\sum_{k=1}^{n} x_{k,i}$ is the arithmetic mean of the sample of $\mathcal{X}_i$ and similarly for m_j.

This notation is very messy so let us simplify it with vector notation. Define $\mathbf{x_j} \in \mathbb{R}^n$ for $j = 1, \ldots, p$ to be a vector containing the observations of $\mathcal{X}_j$. That is

$$\mathbf{x_j} = (x_{1,j}, \ldots, x_{n,j})^T$$

where $x_{i,j}$ is the i^{th} observation of variable $\mathcal{X}_j$. Then if

$$\mathbf{m_j} = (m_j, \ldots, m_j)^T = m_j(\overbrace{1, \ldots, 1}^{n})^T \in \mathbb{R}^n$$

is the vector containing m_j in all n entries then

$$S_{i,j} = \frac{1}{n-1}(\mathbf{x_i} - \mathbf{m_i})^T(\mathbf{x_j} - \mathbf{m_j}).$$

Since $S_{i,j}$ is simply the univariate covariance among two random scalars $\mathcal{X}_i$ and $\mathcal{X}_j$ then the above gives a new way of writing the univariate covariance or indeed univariate variance. For a univariate random variable $\mathcal{Y}$ with observation vector $\mathbf{y} \in \mathbb{R}^n$ its univariate variance estimation may be written as

$$s_{\mathcal{Y}}^2 = \frac{1}{n-1}(\mathbf{y} - \overline{\mathbf{y}})^T(\mathbf{y} - \overline{\mathbf{y}})$$

where $\overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean vector for $\mathbf{y}$. This should look very familiar to our definition for the covariance matrix Σ. Remember that we said Σ and S are the generalization of univariate variance. Thus it should not be surprising that we can write the sample covariance matrix S in a very analogous form.

Define the data matrix X as

$$X = \begin{bmatrix}\mathbf{x_1} & \cdots & \mathbf{x_p}\end{bmatrix}$$

having as its j^{th} column the vector $\mathbf{x_j}$. We call this the data matrix because the $(i,j)^{th}$ entry of X is the i^{th} observation of the random variable $\mathcal{X}_j$. Alternatively the $(i,j)^{th}$ entry is the j^{th} variable $\mathcal{X}_j$ measured in the i^{th} observation. Each *row* of X is one of the n observations of $\mathfrak{X}$. Alternatively we could define the data matrix in a row major manner such that

$$X = \begin{bmatrix}\mathcal{O}_1 \\ \vdots \\ \mathcal{O}_n\end{bmatrix}$$

where $\mathcal{O}_i$ for $i = 1, \ldots, n$ is the i^{th} observation of $\mathfrak{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)^T$. (Admittedly since $\mathfrak{X}$ is a random *column* vector it would have been better to define X as having as each of its *columns* one of the observations of $\mathfrak{X}$ however this is not the usual convention. Thus we will not buck the trend and use the customary definition for a data matrix.)

Let us also define the mean matrix as follows. Let $\mathbf{1}_p \in \mathbb{R}^p$ to be the vector of all 1's such that $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^p$ and define $D = diag(m_1, m_2, \ldots, m_p)$ to an $p \times p$ diagonal matrix with i^{th} diagonal entry as the arithmetic mean of the sample of $\mathcal{X}_i$. Then the matrix $\mathbf{1}_n\mathbf{1}_p^T$ is a $n \times p$ matrix of all 1's and so the matrix M defined by

$$M = \mathbf{1}_n\mathbf{1}_p^T D = \begin{bmatrix}1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1\end{bmatrix}\begin{bmatrix}m_p & & \\ & \ddots & \\ & & m_p\end{bmatrix} = \begin{bmatrix}m_1 & \cdots & m_p \\ \vdots & \ddots & \vdots \\ m_1 & \cdots & m_p\end{bmatrix}$$

is a matrix whose columns are the mean vectors $\mathbf{m_i}$.

Thus we have that

$$X = \begin{bmatrix} \mathbf{x_1} & \cdots & \mathbf{x_p} \end{bmatrix} \text{ and } M = \begin{bmatrix} \mathbf{m_1} & \cdots & \mathbf{m_p} \end{bmatrix}$$

and so

$$X - M = \begin{bmatrix} \mathbf{x_1} - \mathbf{m_1} & \cdots & \mathbf{x_p} - \mathbf{m_p} \end{bmatrix}.$$

Note the similarity of

$$\mathfrak{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)^T \text{ to } X = \begin{bmatrix} \mathbf{x_1} & \cdots & \mathbf{x_p} \end{bmatrix}$$

and

$$\mathrm{E}[\mathfrak{X}] = (\mathrm{E}[\mathcal{X}_1], \ldots, \mathrm{E}[\mathcal{X}_p])^T \text{ to } M = \begin{bmatrix} \mathbf{m_1} & \cdots & \mathbf{m_p} \end{bmatrix}.$$

Then analogously to the definition of the variance for the univariate random variable $\mathcal{Y}$ and similarly to how we defined Σ we can define the sample covariance matrix S of the data matrix X as

$$S = \frac{1}{n-1}(X - M)^T(X - M).$$

We can see this because

$$\begin{aligned} \left(\frac{1}{n-1}(X - M)^T(X - M)\right)_{i,j} &= \frac{1}{n-1} row\left(i, (X - M)^T\right) col(j, X - M) \\ &= \frac{1}{n-1} col(i, X - M)^T col(j, X - M) \end{aligned}$$

which is precisely

$$\frac{1}{n-1}(\mathbf{x_i} - \mathbf{m_i})^T(\mathbf{x_j} - \mathbf{m_j}) = s_{\mathcal{X}_i, \mathcal{X}_j}.$$

As with the univariate case we can define correlation from a definition of covariance. Given the same random vector $\mathfrak{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)^T$ we can define the correlation matrix P as

$$P_{i.j} = \rho_{\mathcal{X}_i, \mathcal{X}_j} = \frac{\sigma_{\mathcal{X}_i, \mathcal{X}_j}}{\sigma_{\mathcal{X}_i} \sigma_{\mathcal{X}_j}}$$

so that the $(i, j)^{th}$ entry of P is the correlation between $\mathcal{X}_i$ and $\mathcal{X}_j$.

Similarly if we have made n observations then we can define a sample correlation matrix R as

$$R_{i,j} = r_{\mathcal{X}_i, \mathcal{X}_j} = \frac{s_{\mathcal{X}_i, \mathcal{X}_j}}{s_{\mathcal{X}_i} s_{\mathcal{X}_j}}.$$

If we take our data matrix X from previously and define a scaling matrix A such that $A = diag(1/s_{\mathcal{X}_1}, \ldots, 1/s_{\mathcal{X}_p})$ then right multiplication by A gives us

$$(X - M)A = \begin{bmatrix} (\mathbf{x_1} - \mathbf{m_1})/s_{\mathcal{X}_1} & \cdots & (\mathbf{x_p} - \mathbf{m_p})/s_{\mathcal{X}_p} \end{bmatrix}.$$

which standardizes the columns of $X - M$ and so

$$R = \frac{1}{n-1}((X - M)A)^T(X - M)A = \frac{1}{n-1}A(X - M)^T(X - M)A.$$

This is easy to see since $S = \frac{1}{n-1}(X-M)^T(X-M)$ is our covariance matrix and multiplying by A on the right scales the j^{th} column of S by $1/s_{\mathcal{X}_j}$ and multiplying by A on the left scales the i^{th} row by $1/s_{\mathcal{X}_i}$ giving us that

$$\left(A\left(\frac{1}{n-1}(X - M)^T(X - M)\right)A\right)_{i,j} = \frac{S_{i,j}}{s_{\mathcal{X}_i} s_{\mathcal{X}_j}} = \frac{s_{\mathcal{X}_i, \mathcal{X}_j}}{s_{\mathcal{X}_i} s_{\mathcal{X}_j}} = r_{\mathcal{X}_i, \mathcal{X}_j}.$$

Note that a nearly identical argument can be made in forming the correlation matrix P by scaling component random scalars $\mathcal{X}_i$ of $\mathfrak{X}$ by their respective variances and finding the covariance matrix of the standardized random vector $B\mathfrak{X}$ where $B = diag(1/\sigma_1, \ldots, 1/\sigma_p)$.

Example 1

Consider the 5×3 data matrix X of 5 measurements of 3 variables

$$X = \begin{bmatrix} 4.0 & 2.0 & .6 \\ 4.2 & 2.1 & .59 \\ 3.9 & 2.0 & .58 \\ 4.3 & 2.1 & .62 \\ 4.1 & 2.2 & .63 \end{bmatrix}.$$

Then the means of the columns are $(4.1, 2.08, .604)$ and so the mean matrix is

$$M = \begin{bmatrix} 4.1 & 2.08 & .604 \\ 4.1 & 2.08 & .604 \\ 4.1 & 2.08 & .604 \\ 4.1 & 2.08 & .604 \\ 4.1 & 2.08 & .604 \end{bmatrix} \text{ meaning } X - M = \begin{bmatrix} -.1 & -.08 & -.004 \\ .1 & .02 & -.014 \\ -.2 & -.08 & -.024 \\ .2 & .02 & .016 \\ 0 & .12 & .026 \end{bmatrix}$$

Then the sample covariance matrix is

$$S = \frac{1}{5-1}(X-M)^T(X-M) = \begin{bmatrix} .025 & .0075 & .00175 \\ .0075 & .007 & .00135 \\ .00175 & .00135 & .0043 \end{bmatrix}.$$

The diagonals of this matrix are the variances and so their square roots are the standard deviations meaning

$$s_1 = .1581, s_2 = .0836 \text{ and } s_3 = .0207$$

and so if $A = diag(1/s_1, 1/s_2, 1/s_3)$ then $(X-M)A$ standardizes the columns of $X-M$ and so

$$(X-M)A = \begin{bmatrix} -.6324 & -.9561 & -.1928 \\ .0632 & .2390 & -.6751 \\ -1.264 & -.9561 & -1.157 \\ 1.264 & .2390 & .7715 \\ 0 & 1.434 & 1.253 \end{bmatrix}$$

meaning the sample correlation matrix R is

$$R = \frac{1}{4}A(X-M)^T(X-M)A = \begin{bmatrix} 1 & .5669 & .5337 \\ .5669 & 1 & .7781 \\ .5337 & .7781 & 1 \end{bmatrix}.$$

◀

Properties and Linear Combinations

The manner in which we defined the sample correlation matrix (via right multiplication by a matrix A) can be generalized into an important fact that will be our first theorem of this chapter. Consider an $n \times p$ data matrix X and a $p \times m$ real matrix L. Then let Y be the product XL such that

$$\begin{aligned} Y = XL &= \begin{bmatrix} \mathbf{x_1} & \cdots & \mathbf{x_p} \end{bmatrix} \begin{bmatrix} \mathbf{L_1} & \cdots & \mathbf{L_m} \end{bmatrix} \\ &= \begin{bmatrix} L_{1,1}\mathbf{x_1} + \cdots + L_{p,1}\mathbf{x_p} & \cdots & L_{1,m}\mathbf{x_1} + \cdots + L_{p,m}\mathbf{x_p} \end{bmatrix} \end{aligned}$$

such that each column of $Y = XL$ is a linear combination of the observation vectors $\{\mathbf{x_i}\}_{i=1}^p$. That is,

$$col(j, Y) = col(j, XL) = L_{1,j}\mathbf{x_1} + \cdots + L_{p,j}\mathbf{x_p}$$

and so whereas the j^{th} column of X represented the observation vector of $\mathcal{X}_j$ the j^{th} column of XL represents the observation vector of a new random scalar

$$\mathcal{Y}_j = L_{1,j}\mathcal{X}_1 + \cdots + L_{p,j}\mathcal{X}_p.$$

Thus while the matrix X is a data matrix for samples of the random vector $\boldsymbol{\mathcal{X}} : \Omega \to \mathbb{R}^p$ the matrix $Y = XL$ is a data matrix for samples of the random vector $\boldsymbol{\mathcal{Y}} = L^T\boldsymbol{\mathcal{X}}$ which maps $\Omega \mapsto \mathbb{R}^m$ where

$$\begin{aligned}\boldsymbol{\mathcal{Y}} = L^T\boldsymbol{\mathcal{X}} &= \begin{bmatrix} \text{—} \mathbf{L_1} \text{—} \\ \vdots \\ \text{—} \mathbf{L_m} \text{—} \end{bmatrix} \begin{bmatrix} \mathcal{X}_1 \\ \vdots \\ \mathcal{X}_p \end{bmatrix} \\ &= (L_{1,1}\mathcal{X}_1 + \cdots + L_{p,1}\mathcal{X}_p,\ \ldots,\ L_{1,m}\mathcal{X}_1 + \cdots + L_{p,m}\mathcal{X}_p) \\ &= (\mathcal{Y}_1, \ldots, \mathcal{Y}_m).\end{aligned}$$

Let us prove an important theorem about such linear combinations.

Theorem 1

Consider a random vector $\boldsymbol{\mathcal{X}}$ with associated covariance matrix Σ and $n \times p$ sample data matrix X with sample covariance matrix S. If L is a $p \times m$ matrix then the covariance matrix of the random vector $L^T\boldsymbol{\mathcal{X}}$ is $L^T\Sigma L$ and the sample covariance matrix associated with the data matrix XL of $L^T\boldsymbol{\mathcal{X}}$ is L^TSL.

We already established that the data matrix for $L^T\boldsymbol{\mathcal{X}}$ is XL and it is easy to show that if the sample mean matrix of $\boldsymbol{\mathcal{X}}$ is M then the sample mean matrix of $L^T\boldsymbol{\mathcal{X}}$ is ML since average plays nicely with linear operators such as matrix multiplication. Thus the sample covariance matrix of $L^T\boldsymbol{\mathcal{X}}$ is

$$\begin{aligned}S_L &= \frac{1}{n-1}(XL - ML)^T(XL - ML) \\ &= \frac{1}{n-1}L^T(X - M)^T(X - M)L \\ &= L^T\left(\frac{1}{n-1}(X - M)^T(X - M)\right)L \\ &= L^TSL.\end{aligned}$$

This is also true for the covariance matrix Σ of $\boldsymbol{\mathcal{X}}$ because of its analogous definition. The covariance matrix of $L^T\boldsymbol{\mathcal{X}}$ is

$$\begin{aligned}\Sigma_L &= \mathrm{E}\left[(L^T\boldsymbol{\mathcal{X}} - \mathrm{E}\left[L^T\boldsymbol{\mathcal{X}}\right])(L^T\boldsymbol{\mathcal{X}} - \mathrm{E}\left[L^T\boldsymbol{\mathcal{X}}\right])^T\right] \\ &= \mathrm{E}\left[(L^T\boldsymbol{\mathcal{X}} - L^T\mathrm{E}\left[\boldsymbol{\mathcal{X}}\right])(L^T\boldsymbol{\mathcal{X}} - L^T\mathrm{E}\left[\boldsymbol{\mathcal{X}}\right])^T\right] \\ &= \mathrm{E}\left[(L^T(\boldsymbol{\mathcal{X}} - \mathrm{E}\left[\boldsymbol{\mathcal{X}}\right]))(L^T(\boldsymbol{\mathcal{X}} - \mathrm{E}\left[\boldsymbol{\mathcal{X}}\right]))^T\right] \\ &= \mathrm{E}\left[L^T(\boldsymbol{\mathcal{X}} - \mathrm{E}\left[\boldsymbol{\mathcal{X}}\right])(\boldsymbol{\mathcal{X}} - \mathrm{E}\left[\boldsymbol{\mathcal{X}}\right])^TL\right] \\ &= L^T\mathrm{E}\left[(\boldsymbol{\mathcal{X}} - \mathrm{E}\left[\boldsymbol{\mathcal{X}}\right])(\boldsymbol{\mathcal{X}} - \mathrm{E}\left[\boldsymbol{\mathcal{X}}\right])^T\right]L \\ &= L^T\Sigma L.\end{aligned}$$

■

Consider a random vector $\boldsymbol{\mathcal{X}} : \Omega \to \mathbb{R}^p$ with covariance matrix Σ. If we let B be the $p \times p$ diagonal matrix $B = diag(1/\sigma_1, \ldots, 1/\sigma_p)$ then the covariance matrix of $B\boldsymbol{\mathcal{X}}$ is

$$\Sigma_B = B\Sigma B$$

which is precisely the correlation matrix P. Similarly if we have a $p \times p$ sample covariance matrix S of $\boldsymbol{\mathcal{X}}$ and $A = diag(1/s_1, \ldots, 1/s_p)$ then the sample covariance matrix of $A\boldsymbol{\mathcal{X}}$ is

$$S_A = ASA$$

which is the sample correlation matrix R.

Thus in general the correlation matrix P is simply the covariance matrix of the standardized random vector $\mathcal{Y} = \left(\frac{\mathcal{X}_1}{\sigma_1}, \ldots, \frac{\mathcal{X}_p}{\sigma_p}\right)^T$. Similarly the sample correlation matrix R is the sample covariance matrix of the associated standardized data matrix with columns $Y = \left[\frac{\mathbf{x_1}}{s_1} \cdots \frac{\mathbf{x_1}}{s_1}\right]$. Thus all correlation matrices are covariance matrices. If not stated otherwise we may now assume that all results on the covariance matrix extends to the correlation matrix.

Our definitions of the covariance and sample covariance matrices Σ and S hint at another property. We define these matrices multiplying a suitable matrix (or vector) by its transpose. Consider the sample covariance and correlation matrices S and R. These (and also Σ and P) are symmetric. We can see this because in the case of S,

$$S_{i,j} = s_{\mathcal{X}_i,\mathcal{X}_j} = s_{\mathcal{X}_j,\mathcal{X}_i} = S_{j,i}.$$

and similarly for R (and Σ or P). More richly however notice that for $B = \frac{1}{\sqrt{n-1}}(X - M)$

$$S = B^T B$$

which is a symmetric form that should bring to mind much of the last chapter. A similar story may be told for R since for $C = \frac{1}{\sqrt{n-1}}(X - M)A$ we have $R = C^T C$ and, recalling the definitions for Σ and P, it is true for these matrices also. However we know much more from last chapter about these forms than that they are simply symmetric. The next theorem will be the first of many applications of last chapter's material given what we now know about these matrices.

Theorem 2

A matrix M is a covariance matrix if and only if it positive semi-definite.

$\Rightarrow$ We already showed in the previous chapter that the symmetric form $A^T A$ has non-negative eigenvalues for a real A and so since the sample covariance matrix is

$$S = \left(\frac{1}{\sqrt{n-1}}X\right)^T \left(\frac{1}{\sqrt{n-1}}X\right)$$

where $\frac{1}{\sqrt{n-1}}X$ is real then S has non-negative eigenvalues and is thus positive semi-definite.

On the other hand for a random variable $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)^T$ with covariance matrix Σ and $p \times 1$ vector

$$L = (l_1, \ldots, l_p)^T$$

then the variance of the univariate random variable defined by the linear combination

$$l_1\mathcal{X}_1 + \cdots + l_p\mathcal{X}_p$$

is the (1×1) covariance matrix of $L^T\boldsymbol{\mathcal{X}}$. Theorem 1 tells us that this is

$$\Sigma_{L^T\boldsymbol{\mathcal{X}}} = L^T \Sigma L.$$

Since this is the covariance of a univariate random variable then $L^T\Sigma L \geq 0$ however $L^T\Sigma L$ is a quadratic form in L. Thus, as we discovered in the previous chapter, Σ must be positive semi-definite since its associated quadratic form is non-negative.

$\Leftarrow$ Now for the reverse direction consider a symmetric positive semi-definite M. Then M is orthogonally diagonalizable as

$$M = UDU^T$$

for some orthogonal matrix U and diagonal matrix D. Then define the square root of M to be

$$\sqrt{M} = U\sqrt{D}U^T$$

where if $D = diag(\lambda_1, \ldots, \lambda_p)$ then $\sqrt{D} = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$. This okay because M is positive semi-definite and so its eigenvalues are non-negative since so a real square root exists for λ_i. With this definition then

$$\left(\sqrt{M}\right)^2 = U\sqrt{M}U^T U\sqrt{M}U^T = U\left(\sqrt{D}\right)^2 U^T = UDU^T = M$$

and

$$\left(\sqrt{M}\right)^T = (U\sqrt{D}U^T)^T = (U^T)^T\left(\sqrt{D}\right)U^T = U\sqrt{D}U^T = \sqrt{M}.$$

Now if $\mathfrak{X}$ is a random vector with identity covariance structure ($\Sigma = I$) then we know from Theorem 1 that

$$\text{var}\left[\sqrt{M}\mathfrak{X}\right] = \left(\sqrt{M}\right)^T \text{var}[\mathfrak{X}]\sqrt{M} = \sqrt{M}I\sqrt{M} = \left(\sqrt{M}\right)^2 = M.$$

Thus M is the covariance matrix of $\sqrt{M}\mathfrak{X}$ and so it is the covariance matrix of some random vector. A similar result is available for the sample covariance matrix S.

■

Since all correlation matrices are also covariance matrices then the above theorem is true for them also.

The Frobenius Norm

In the spirit of generalizing we would like to define a norm for matrices. Normally we deal with norms of vectors. The norm of a vector

$$\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$$

is defined as

$$||\mathbf{x}|| = \sqrt{\sum_{i=1}^{p} x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}.$$

For a real $n \times p$ matrix A we may define a norm, called the Frobenius norm, as

$$||A||_F = \sqrt{\sum_{j=1}^{p}\sum_{i=1}^{n} A_{i,j}^2} = \sqrt{tr(A^TA)} = \sqrt{tr(AA^T)}$$

where $tr(B)$ is the trace of a square $p \times p$ matrix B defined by $tr(B) = \sum_{i=1}^{p} B_{i,i}$ the sum of its diagonal elements. Notice how analogously we define the matrix norm. Just as with a vector it is the square root of the sum of the constituent components. Indeed when A is a vector ($1 \times p$ or $p \times 1$ matrix) then the Frobenius norm reduces to the familiar Euclidean norm for vectors.

We can see that $\sqrt{\sum_{j=1}^{p}\sum_{i=1}^{n} A_{i,j}^2} = \sqrt{tr(A^TA)}$ because

$$(A^TA)_{i,j} = row(i, A^T)col(j, A) = col(i, A)^T col(j, A)$$

and so

$$(A^TA)_{j,j} = col(j, A)^T col(j, A) = ||col(j, A)||^2 = \sum_{i=1}^{n} A_{i,j}^2.$$

This means that

$$tr(A^TA) = \sum_{j=1}^{p}\sum_{i=1}^{n} A_{i,j}^2$$

and so

$$\sqrt{tr(A^TA)} = \sqrt{\sum_{j=1}^{p}\sum_{i=1}^{n} A_{i,j}^2}.$$

Now the trace of A^TA is equal to the sum of the eigenvalues of A^TA. To see this note that A^TA is a $p \times p$ square matrix and we proved last chapter that it had p real eigenvalues. Furthermore since it is symmetric it is orthogonally diagonalizable as

$$D = U^TA^TAU$$

and so

$$tr(D) = tr(U^TA^TAU)$$

and so since $tr(YZ) = tr(ZY)$ for same sized matrices Y and Z then

$$tr(D) = tr(U^TA^TAU) = tr(UU^TA^TA) = tr(A^TA).$$

Thus since $D = diag(\lambda_1, \ldots, \lambda_p)$ is a diagonal with matrix of eigenvalues

$$tr(A^TA) = tr(D) = \sum_{i=1}^{p} \lambda_i.$$

Hence

$$||A||_F = \sqrt{tr(A^TA)} = \sqrt{\sum_{i=1}^{p} \lambda_i}.$$

We can play the same game with AA^T and get that $||A||_F = \sqrt{tr(A^TA)} = \sqrt{tr(AA^T)} = \sqrt{\sum_{i=1}^{p} \lambda_i}$ since the non-zero eigenvalues of A^TA and AA^T are the same. However we defined the non-zero singular values of A to be the square roots of the eigenvalues of A^TA (or AA^T) and so $\lambda_i = \sigma_i^2$ meaning

$$||A||_F = \sqrt{\sum_{i=1}^{s} \sigma_i^2}$$

where $s \leq min\{n, p\}$ is the number of non-zero singular values.

There is another important way we can see this. The trick of switching the order of the trace of a product can be used to show that the Frobenius norm is invariant under multiplication by orthogonal matrices. For some orthogonal matrix U

$$||AU||_F = \sqrt{tr(U^TA^TAU)} = \sqrt{tr(UU^TA^TA)} = \sqrt{tr(A^TA)} = ||A||_F$$

and similarly

$$||UA||_F = ||A||_F.$$

Thus if A has a singular value decomposition $A = U\Sigma V^T$ for orthogonal U and V then

$$||A||_F = ||U\Sigma V^T||_F = ||\Sigma||_F = \sqrt{tr(\Sigma^T\Sigma)} = \sqrt{\sum_{i=1}^{s} \sigma_i^2}.$$

This fact should not be surprising as the we already know that for a matrix A the singular values σ_i tell us how much vectors in the row space are stretched when sent into the column space. As with vector norms matrix norms are trying to tell us something about the size of a matrix. Thus it seems sensible that the Frobenius norm tells us something about how much the matrix stretches the space (and the vectors in it).

Total Variance

We would like to define a summary statistic for the covariance matrix that can give us an overall picture. After all the covariance matrix has $\binom{p+1}{2}$ unique entries and so this is quite a lot of information to take in at once. Let us define the total variance of a random vector $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)^T$ with associated covariance matrix Σ to be

$$\text{total variance} = tr(\Sigma) = \sum_{i=1}^{p} \sigma_{\mathcal{X}_i}^2.$$

Notice that if we define the Frobenius norm of a random matrix Y to be the square root of the trace of its symmetric form YY^T then

$$\begin{aligned}||(\mathfrak{X} - \mathrm{E}[\mathfrak{X}])||_F^2 &= tr\left((\mathfrak{X} - \mathrm{E}[\mathfrak{X}])(\mathfrak{X} - \mathrm{E}[\mathfrak{X}])^T\right) \\ &= \sum_{i=1}^{p}(\mathfrak{X}_i - \mathrm{E}[\mathfrak{X}_i])^2\end{aligned}$$

and so

$$\begin{aligned}\mathrm{E}\left[||(\mathfrak{X} - \mathrm{E}[\mathfrak{X}])||_F^2\right] &= \mathrm{E}\left[\sum_{i=1}^{p}(\mathfrak{X}_i - \mathrm{E}[\mathfrak{X}_i])^2\right] \\ &= \sum_{i=1}^{p}\mathrm{E}\left[(\mathfrak{X}_i - \mathrm{E}[\mathfrak{X}_i])^2\right] \\ &= \sum_{i=1}^{p}\sigma_i^2.\end{aligned}$$

Thus the total variance of a random vector $\mathfrak{X}$ is the expectation of the squared Frobenius norm of the centered random vector $\mathfrak{X} - \mathrm{E}[\mathfrak{X}]$.

Similarly if the associated sample covariance matrix of n observations of $\mathfrak{X}$ is S then define

$$\text{total sample variance} = tr(S) = \sum_{i=1}^{p} s_{\mathfrak{X}_i}^2.$$

Remember that if X is the data matrix of the n observations of $\mathfrak{X}$ then

$$S = \frac{1}{n-1}(X - M)^T(X - M)$$

for associated mean matrix M and so

$$tr(S) = tr\left(\frac{1}{n-1}(X - M)^T(X - M)\right)$$

or if $Y = \frac{1}{\sqrt{N-1}}(X - M)$ then

$$tr(S) = tr\left(Y^T Y\right) = ||Y||_F^2 = \frac{1}{n-1}||X - M||_F^2.$$

Thus the total variance is proportional to the squared Frobenius norm of the centered data matrix $X - M$. Notice that this implies that the total variance is proportional to the sum of the squares of the singular values of $X - M$. Put another way, it is proportional to the sum of the eigenvalues of the sample covariance matrix.

Example 2

Considering the sample covariance matrix from Example 1 let us compute the total sample variance. One definition is that

$$\text{total sample variance } = tr(S) = .03243 \approx .025 + .007 + .0043$$

or taking the Frobenius norm of X we see that

$$||X - M||_F = .3601$$

and so

$$\frac{1}{4}||X - M||_F^2 = .03243.$$

◀

Note that the total variance is only one way to define a summary statistic for the covariance matrix. We introduce this notion in particular because it is the most natural such statistic when doing principal components analysis.

2.1.2 The Normal Distribution

The last generalization we want to make is that of the normal distribution. In the first part of this section we will develop a couple important features of the univariate normal distribution. While the univariate normal distribution is likely quite familiar developing its properties will be helpful in seeing the development of analogous properties in the more general multivariate normal distribution. Discussing the multivariate normal distribution is the last step necessary to discussing principal components analysis.

The Univariate Case

A univariate normally distributed random variable $\mathcal{X}$ is a simple model of a phenomena in which samples tend to be clustered around some center. The distribution is parameterized by two parameters $\mu = \mathrm{E}[\mathcal{X}]$ and $\sigma^2 = \mathrm{var}[\mathcal{X}]$. The goal of this first section is to show how these parameters control for location and spread of the data respectively.

For a positive spread $\sigma^2 > 0$ the random variable $\mathcal{X} \sim N(\mu, \sigma^2)$ has a probability density function $\phi : \mathbb{R} \to [0, 1]$ given by

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

for any $x \in \mathbb{R}$.

Consider the exponent of the density function $-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$. First notice that since the $\frac{x-\mu}{\sigma}$ term is squared then for a constant c all points x such that

$$\left|\frac{x-\mu}{\sigma}\right| = \sqrt{\left(\frac{x-\mu}{\sigma}\right)^2} = c$$

have an equal density. For the univariate normal distribution all this amounts to the fact that points symmetric around the mean have equal density.

Now we claimed that the mean parameter $\mu = \mathrm{E}[\mathcal{X}]$ is a point around which values tend to cluster. This seems plausible since the mean is the expected value and thus purports to capture a typical value. We can state this a little more rigorously by noting that the probability density is highest around the mean. As the standardized distance from the mean of a point $x \in \mathbb{R}$, $\left|\frac{x-\mu}{\sigma}\right|$, increases the probability density $\phi(x)$ decreases. This occurs because $\left|\frac{x-\mu}{\sigma}\right|$ increasing entails $\left|\frac{x-\mu}{\sigma}\right|^2 = \left(\frac{x-\mu}{\sigma}\right)^2$ increasing and thus the exponent of the density function becoming more negative. Alternatively we can look at the derivaties of ϕ

$$\phi'(x) = -\left(\frac{x-\mu}{\sigma}\right)\phi(x) \text{ and } \phi''(x) = \left(-\frac{1}{\sigma} + \left(\frac{x-\mu}{\sigma}\right)^2\right)\phi(x)$$

and note that since $\phi(x) > 0$ for all $x \in \mathbb{R}$ then $\phi'(x) = 0$ if and only if $x = \mu$ in which case $\phi''(x) < 0$ and so we have a maximum only at $x = \mu$. Thus if we sample from the distribution we will tend to see samples closer rather than further from the mean since this is where the probability density is greatest.

This discussion begins to highlight the importance of the term

$$\frac{x-\mu}{\sigma}$$

in the exponent of the density function in determining the random variable's behavior. Let us transform $\mathcal{X}$ into a new random variable $\mathcal{Z}$ by

$$\mathcal{Z} = \frac{\mathcal{X}-\mu}{\sigma}.$$

By doing this transformation we have that

$$\mathrm{E}[\mathcal{Z}] = \mathrm{E}\left[\frac{\mathcal{X}-\mu}{\sigma}\right] = 0$$

and

$$\mathrm{var}[\mathcal{Z}] = \mathrm{var}\left[\frac{\mathcal{X}-\mu}{\sigma}\right] = \frac{1}{\sigma^2}\mathrm{var}[\mathcal{X}] = 1.$$

Thus for any normally distributed random variable $\mathcal{X} \sim N(\mu, \sigma^2)$ the random variable $\mathcal{Z} = (\mathcal{X} - \mu)/\sigma$ has a distribution of $\mathcal{Z} \sim N(0, 1)$.

Consider two normally distributed random variables $\mathcal{W} \sim N(\mu_{\mathcal{W}}, \sigma^2_{\mathcal{W}})$ and $\mathcal{Y} \sim N(\mu_{\mathcal{Y}}, \sigma^2_{\mathcal{Y}})$ with standardized version of $\mathcal{Z}_{\mathcal{W}}$ and $\mathcal{Z}_{\mathcal{Y}}$ respectively. Note that $\mathcal{Z}_{\mathcal{W}} = \mathcal{Z}_{\mathcal{Y}}$ and so the standardized versions of $\mathcal{W}$ and $\mathcal{Y}$ are the same random variable. The standardization process of subtracting off the mean and dividing through by the standard deviation washes away any difference between the random variables. It zeroes out the mean and unitizes the variance and thus makes equal the only parameters of the distribution. Clearly then for any constant c,

$$P(|\mathcal{Z}_{\mathcal{W}}| < c) = P(|\mathcal{Z}_{\mathcal{Y}}| < c)$$

because $\mathcal{Z}_{\mathcal{W}} = \mathcal{Z}_{\mathcal{Y}} \sim N(0, 1)$. However using the definition of $\mathcal{Z}_{\mathcal{W}}$ and $\mathcal{Z}_{\mathcal{Y}}$ we may write this as

$$P\left(\left|\frac{\mathcal{W} - \mu_{\mathcal{W}}}{\sigma_{\mathcal{W}}}\right| < c\right) = P\left(\left|\frac{\mathcal{Y} - \mu_{\mathcal{Y}}}{\sigma_{\mathcal{Y}}}\right| < c\right)$$

meaning

$$P\left(\mu_{\mathcal{W}} - c\sigma_{\mathcal{W}} < \mathcal{W} < \mu_{\mathcal{W}} + c\sigma_{\mathcal{W}}\right) = P\left(\mu_{\mathcal{Y}} - c\sigma_{\mathcal{Y}} < \mathcal{Y} < \mu_{\mathcal{Y}} + a\sigma_{\mathcal{Y}}\right).$$

This is a quite powerful statement. It tells us that the probability of being within c standard deviations from the mean is equal for all normal random variables. Notice that while the intervals

$$(\mu_{\mathcal{W}} - c\sigma_{\mathcal{W}}, \mu_{\mathcal{W}} + c\sigma_{\mathcal{W}}) \text{ and } (\mu_{\mathcal{Y}} - c\sigma_{\mathcal{Y}}, \mu_{\mathcal{Y}} + c\sigma_{\mathcal{Y}})$$

capture the same probability under the respective density functions to $\mathcal{W}$ and $\mathcal{Y}$ the intervals themselves are not generally equal in (Euclidean) size. Indeed they are sized $2\sigma_{\mathcal{W}}$ and $2\sigma_{\mathcal{Y}}$ respectively where $\sigma_{\mathcal{W}} \neq \sigma_{\mathcal{Y}}$ generally. Thus the same probability density is spread out over intervals of differing sizes depending on the variances (or standard deviations). All this is to say is that the variance parameter of the normal distribution controls the spread with which the density is allocated. This in turn determines the spread (sample variance) of the samples we take from the random variable.

Our discussion so far should not be earth-shattering. We have described the univariate normal distribution $N(\mu, \sigma^2)$ as being controlled by two parameters μ and σ which control the general location and spread of the distribution (and samples thereof). Thus if we sample from a normal distribution we expect to see points generally clustered around the mean μ. The degree of tightness with which they cluster around the mean is dependent upon the variance paramter σ^2. While it is likely that this much about the univariate normal was already understood we now want to generalize to the multivariate normal distribution.

The Multivariate Case

The multivariate normal distribution is one parameterized by a mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. These are, respectively, the generalization of the mean μ and variance σ^2. We already discussed the covariance matrix Σ at length and the mean vector for a p–variate normal random vector $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)^T \sim N_p(\boldsymbol{\mu}, \Sigma)$ is precisely what you think it is,

$$\boldsymbol{\mu} = (\mathrm{E}[\mathcal{X}_1], \ldots, \mathrm{E}[\mathcal{X}_p])^T.$$

Now for the univariate normal distribution if $\sigma^2 > 0$ we were able to define a density function ϕ. A similar condition holds in the multivariate case. We already established that the covariance matrix Σ is positive semi-definite meaning that its eigenvalues are non-negative. We say that Σ is positive definite if all of its eigenvalues are positive and denote this by writing $\Sigma > 0$. In the case that Σ is positive definite we may define a probability density function $\boldsymbol{\phi} : \mathbb{R}^p \to [0, 1]$ by

$$\boldsymbol{\phi}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}\left((\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)\right)$$

for $\mathbf{x} \in \mathbb{R}^p$. Now since Σ is a real symmetric matrix then its eigenvalues are real and we may orthogonally diagonalize it as

$$\Sigma = UDU^T$$

meaning

$$|\Sigma| = det(\Sigma) = det(UDU^T) = det(U)det(D)det(U^T) = det(D).$$

However $det(D)$ is simply the product of the diagonal elements of D and is thus the product of the eigenvalues of Σ. Thus if Σ were positive semi-definite and not positive definite then 0 would be an eigenvalue and we would have $|\Sigma| = 0$. Thus we see the reason that Σ must be positive definite to define a density function. Otherwise $|\Sigma| = 0$ meaning Σ^{-1} wouldn't exist and we could not divide by $\sqrt{|\Sigma|}$. Either of these problems breaks the definition of the probability density function given above. Thus for the above definition to work we really need $\Sigma > 0$.

Now we had a similar condition in the univariate case that $\sigma^2 > 0$ in order to define the univariate density function so it is not too surprising that we have this condition here. The univariate case where $\sigma^2 = 0$ is a very boring case and so we don't lose too much sleep over the fact that we can't define a density function properly here. However the multivariate case where Σ has 0 as an eigenvalue is a more interesting case and thus it is more troubling that we can't properly define a density function in such cases. There is a fix we can apply here but since it slightly complicates our discussion we will save it to the end. First we will discuss the so called non-degenerate case where $\Sigma > 0$ and the probability density function is well defined. After that is well understood we can come back and make some adjustments so that we can deal with degenerate covariance matrices.

Let us begin as we did in the univariate case and focus on the exponent of the density function. In the multivariate case this exponent is

$$-\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Previously for univariate normal distributions we looked at all points $x \in \mathbb{R}$ such that

$$\left|\frac{x-\mu}{\sigma}\right| = c$$

and noted that the distribution's density had symmetry about the mean. For the multivariate case we note that

$$||\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})|| = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

and so all points such that for some constant c

$$||\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})|| = c$$

have the same density determined by $\boldsymbol{\phi}$. Here $\Sigma^{-1/2} = \left(\sqrt{\Sigma}\right)^{-1} = \sqrt{\Sigma^{-1}}$ for symmetric $\Sigma > 0$. However if

$$||\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})|| = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} = c$$

then

$$||\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})||^2 = (\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

which looks a lot like a quadratic form from last chapter. Indeed we will show that this *is* a quadratic because the defining matrix Σ^{-1} is symmetric.

Notice that the variable for our quadratic form is a difference of vectors $\mathbf{x}-\boldsymbol{\mu}$ which is something we did not explicitly deal with previously. If we look at the quadratic form through a shifted origin of the coordinate system where we consider the coordinates in terms of the displacement $\mathbf{y}$ of $\mathbf{x}$ from $\boldsymbol{\mu}$ as $\mathbf{y} = \mathbf{x}-\boldsymbol{\mu}$ then the equation is

$$\mathbf{y}^T\Sigma^{-1}\mathbf{y} = c^2$$

and we are back to a form we recognize.

The behavior of the equation

$$(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

depends much upon the matrix Σ^{-1} as we discovered in the previous chapter. Thus the next theorem will greatly aid us in understanding this equation.

Theorem 3

The eigenpair $(\lambda, \mathbf{v})$ belongs to an invertible covariance matrix Σ if and only if $(\frac{1}{\lambda}, \mathbf{v})$ is an eigenpair of Σ^{-1}.

$\Rightarrow$ If $(\lambda, \mathbf{v})$ is an eigenpair of Σ then we already established that $\lambda \in \mathbb{R}$ and $\lambda > 0$ (otherwise it wouldn't be invertible). Furthermore if this is an eigenpair then

$$\Sigma\mathbf{v} = \lambda\mathbf{v}$$

and so

$$\Sigma^{-1}\Sigma\mathbf{v} = \lambda\Sigma^{-1}\mathbf{v}$$

which means that

$$\Sigma^{-1}\mathbf{v} = \frac{1}{\lambda}\mathbf{v}.$$

This means $(\frac{1}{\lambda}, \mathbf{v})$ is an eigenpair of Σ^{-1}.

$\Leftarrow$ Since this process is reversible then we have that $(\lambda, \mathbf{v})$ is an eigenpair of Σ if and only if $(\frac{1}{\lambda}, \mathbf{v})$ is an eigenpair of Σ^{-1}.

■

The first implication of this theorem is that if $\Sigma > 0$ then $\Sigma^{-1} > 0$ since if $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of Σ then $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_p}$ are the eigenvalues of Σ^{-1}. Thus if $\lambda_i > 0$ for $i = 1, \ldots, p$ then $\frac{1}{\lambda_i} > 0$ for $i = 1, \ldots, p$ and so all the eigenvalues of Σ^{-1} are positive. Now surely if Σ is symmetric then so is Σ^{-1} because if $\Sigma = \Sigma^T$ then $\Sigma^{-1} = (\Sigma^T)^{-1}$ and

$$I = I^T = (\Sigma^{-1}\Sigma)^T = \Sigma^T(\Sigma^{-1})^T$$

meaning $(\Sigma^T)^{-1} = (\Sigma^{-1})^T$ or $\Sigma^{-1} = (\Sigma^{-1})^T$ and hence it is symmetric.

Then the equation defined by

$$(\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$$

is a quadratic form because Σ^{-1} is symmetric. Futhermore it is an ellipsoid because Σ^{-1} is positive definite.

Now if we make the translating change of variable $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$ then we get an equation, as before, of $\mathbf{y}^T\Sigma^{-1}\mathbf{y} = c^2$. Since the center of $\mathbf{y}^T\Sigma^{-1}\mathbf{y} = c^2$ is the point $\mathbf{y} = \mathbf{0}$ as discussed in the previous chapter then the center of

$$(\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$$

is $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} = \mathbf{0}$ or $\mathbf{x} = \boldsymbol{\mu}$. Thus far we have found that our equation describes an ellipsoid centered at $\boldsymbol{\mu}$.

Now the last thing we need to discuss to fully characterize the ellipse is to discuss its principal axes. We know from chapter 1 that the principal axes are the eigenvectors of the defining matrix and the lengths of the principal axes are proportional to the eigenvalues. However we know that for a positive definite Σ the eigenvectors of Σ^{-1} are those of Σ and the eigenvalues of Σ^{-1} are the reciprocals of those of Σ. Thus the principal axes of the ellipse defined by our equation are the eigenvectors of Σ. Furthermore the lengths of the principal axes are

$$\frac{c}{\sqrt{\gamma_i}}$$

where γ_i is an eigenvalue of Σ^{-1}. However we already established that $\gamma_i = \frac{1}{\lambda_i}$ where λ_i is an eigenvalue of Σ. So the length of the principal axes are

$$c\sqrt{\lambda_i}$$

where λ_i is the i^{th} eigenvalue of Σ ordered in non-increasing order. Now if $\boldsymbol{\mathcal{X}}$ is our multivariate normal random variable and $\mathbf{v_i}$ is a unit eigenvector of $\text{var}[\boldsymbol{\mathcal{X}}] = \Sigma$ corresponding to an eigenvalue of λ_i then by Theorem 1

$$\text{var}\left[\mathbf{v_i}^T(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})\right] = \text{var}\left[\mathbf{v_i}^T\boldsymbol{\mathcal{X}}\right] = \mathbf{v_i}^T\text{var}[\boldsymbol{\mathcal{X}}]\,\mathbf{v_i} = \mathbf{v_i}^T\Sigma\mathbf{v_i}.$$

However $\mathbf{v_i}$ is a unit eigenvector of Σ so

$$\text{var}\left[\mathbf{v_i}^T(\boldsymbol{\mathcal{X}}-\boldsymbol{\mu})\right] = \mathbf{v_i}^T\Sigma\mathbf{v_i} = \mathbf{v_i}^T(\lambda_i\mathbf{v_i}) = \lambda_i\mathbf{v_i}^T\mathbf{v_i} = \lambda_i.$$

Thus the i^{th} principal axis is in the direction of $\mathbf{v_i}$, the i^{th} eigenvector of Σ, and has a length proportional to $\sqrt{\lambda_i} = \sqrt{\text{var}\left[\mathbf{v_i}^T(\boldsymbol{\mathcal{X}}-\boldsymbol{\mu})\right]}$ which is the standard deviation of the random variable $\mathbf{v_i}^T(\boldsymbol{\mathcal{X}}-\boldsymbol{\mu})$.

Now this whole conversation began because we wanted to determine all points $\mathbf{x} \in \mathbb{R}^p$ having equal probability density. The density function went something like

$$\exp\left(||\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})||\right)$$

so that all points having the same standardized distance $||\Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})|| = c$ had the same density.

In the univariate case the parameter σ^2 changed how stretched or spread out the probability density was along the axis. The matrix Σ does the same thing for the multivariate case. However the multivariate case is much richer because we not only have to decide how stretched out the density is in any particular direction (the eigenvalues of Σ) we have to determine the directions in which this stretching is occurring (the eigenvectors of Σ). The function of Σ is quite similar to σ^2 however. They both just determine how the probability density is spread out over the space.

We have now discovered that all points satisfying this equation lie on an ellipsoid with center $\boldsymbol{\mu}$ principal axes in the directions of the eigenvectors of Σ and corresponding lengths of those principal axes being c times the standard deviation of the random variable in that direction. Thus as c changes the only thing which varies with respect to the ellipsoids are the lengths of the principal axes. All of the axes get bigger as c increases and they all get smaller as c decreases. Every value of $c > 0$ defines a different ellipsoid representing a different density. The ellipsoids are telescoping like Russian dolls such that they sit neatly inside each other with no intersections. They are the level curves of $\boldsymbol{\phi}$ if we think of $\boldsymbol{\phi}(\mathbf{x})$ being a surface over $\mathbb{R}^p$.

Consider sitting at the centroid $\boldsymbol{\mu}$. If we move outward from the centroid we will be sitting on one of many telescoping ellipsoids defined by the covariance matrix. Depending on the direction we moved and the distance we are from the centroid will determine the ellipsoid on which we sit. Note that moving the same distance in any direction doesn't guarantee that we are on the same ellipsoid since these ellipsoids are flattened or elongated depending upon the direction in which we move. If our point is sitting on an ellipsoid defined by a larger constant c then it has larger principal axes and is, in some sense, further from the centroid. The normal euclidean distance

$$d(\mathbf{x},\boldsymbol{\mu}) = ||\mathbf{x}-\boldsymbol{\mu}|| = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})}$$

defines concentric a p–spheres around the centroid. We say that a point $\mathbf{a}$ is further from $\boldsymbol{\mu}$ than $\mathbf{b}$ if $\mathbf{a}$ sits on a sphere with larger radius. In the same way our quadratic form defines concentric p–dimensional ellipsoids and we can say that $\mathbf{a}$ if further from the centroid than $\mathbf{b}$ if it sits on an ellipse with larger principal axes.

Notice that the "further" the ellipsoid upon which a point $\mathbf{x}$ sits the larger the quadratic form in the exponent of the density function and thus more negative the exponent and the smaller the density. Thus we expect to see points lying in ellipsoids closer to the centroid than ones further away because the density dies off as we move away from $\boldsymbol{\mu}$. Hence samples form the multivariate normal distribution will be clustered around the mean in a somewhat ellipsoidal arrangement mimicking Σ.

We can show that the probability density function has a maximum at $\mathbf{x} = \boldsymbol{\mu}$. Consider V to be an orthonormal eigenbasis for $\mathbb{R}^p$ formed from the eigenvectors of Σ. Then we can make a change of variable and let $\boldsymbol{\mathcal{Y}} = V^T(\boldsymbol{\mathcal{X}}-\boldsymbol{\mu})$ so that

$$\text{var}\left[\boldsymbol{\mathcal{Y}}\right] = V^T\Sigma V = D = diag(1/\lambda_1,\ldots,1/\lambda_p)$$

for λ_i an eigenvalue of Σ and $\text{E}\left[\mathcal{Y}\right] = 0$. Then we have that

$$\boldsymbol{\phi}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^p|D|}}\exp\left(-\frac{1}{2}\left(\mathbf{y}^T D\mathbf{y}\right)\right)$$

for $\mathbf{y} \in \mathbb{R}^p$. Then

$$(\nabla\boldsymbol{\phi})(\mathbf{y}) = -D\mathbf{y}\boldsymbol{\phi}(\mathbf{y}) = \begin{bmatrix} -\boldsymbol{\phi}(\mathbf{y})\lambda_1 y_1 \\ \vdots \\ -\boldsymbol{\phi}(\mathbf{y})\lambda_p y_p \end{bmatrix}$$

meaning

$$\begin{aligned} \text{Hess}(\boldsymbol{\phi})(\mathbf{y}) &= \begin{bmatrix} \frac{\partial}{\partial y_1}(-\boldsymbol{\phi}(\mathbf{y})\lambda_1 y_1) & \cdots & \frac{\partial}{\partial y_1}(-\boldsymbol{\phi}(\mathbf{y})\lambda_p y_p) \\ \vdots & & \vdots \\ \frac{\partial}{\partial y_p}(-\boldsymbol{\phi}(\mathbf{y})\lambda_1 y_1) & \cdots & \frac{\partial}{\partial y_p}(-\boldsymbol{\phi}(\mathbf{y})\lambda_p y_p) \end{bmatrix} \\ &= \begin{bmatrix} -\lambda_1\boldsymbol{\phi}(\mathbf{y}) + -\lambda_1 y_1(-\lambda_1 y_1\boldsymbol{\phi}(\mathbf{y})) & \cdots & -\lambda_p y_p(-\lambda_1 y_1\boldsymbol{\phi}(\mathbf{y})) \\ \vdots & & \vdots \\ -\lambda_1 y_1(-\lambda_p y_p\boldsymbol{\phi}(\mathbf{y})) & \cdots & -\lambda_p\boldsymbol{\phi}(\mathbf{y}) + -\lambda_p y_p(-\lambda_p y_p\boldsymbol{\phi}(\mathbf{y})) \end{bmatrix} \end{aligned}$$

and when $\mathbf{x} = \boldsymbol{\mu}$ we correspondingly have $\mathbf{y} = \mathbf{0}$ in which case

$$(\nabla\boldsymbol{\phi})(\mathbf{0}) = \mathbf{0}$$

and

$$\text{Hess}(\boldsymbol{\phi})(\mathbf{0}) = -D\boldsymbol{\phi}(\mathbf{0})$$

and so since the gradient is zero and the Hessian is negative definite (since D is positive definite and $\boldsymbol{\phi}(\mathbf{0}) > 0$) then we have a minimum at $\mathbf{x} = \boldsymbol{\mu}$. Thus we really have the highest density at $\mathbf{x} = \boldsymbol{\mu}$ and decreasing density as we move away from the mean (as determined by Σ).

Notice that for any $\boldsymbol{\mathcal{X}} \sim N(\boldsymbol{\mu}, \Sigma)$ we can transform $\boldsymbol{\mathcal{X}}$ into $\boldsymbol{\mathcal{Z}}$ such that

$$\boldsymbol{\mathcal{Z}} = \Sigma^{-1/2}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})$$

giving us that

$$\text{E}[\boldsymbol{\mathcal{Z}}] = \text{E}\left[\Sigma^{-1/2}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})\right] = \Sigma^{-1}(\text{E}[\boldsymbol{\mathcal{X}}] - \boldsymbol{\mu}) = \mathbf{0}$$

and

$$\text{var}[\boldsymbol{\mathcal{Z}}] = \text{var}\left[\Sigma^{-1/2}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})\right] = \Sigma^{-1/2}\text{var}[\boldsymbol{\mathcal{X}}]\,\Sigma^{-1/2} = I.$$

Thus our $\boldsymbol{\mathcal{Z}}$ transformation transforms any p–variate normal random vector into a standard $N_p(\mathbf{0}, I)$ random vector. Hence similar to the univariate case the probability

$$P(||\boldsymbol{\mathcal{Z}}_{\boldsymbol{\mathcal{X}}}|| < c)$$

for some $c \in \mathbb{R}$ is the same for all p–variate normal random vectors $\boldsymbol{\mathcal{X}}$. However precisely what kind of region in $\mathbb{R}^p$ this event corresponds to (what kind of region captures the constant amount of probability) depends on the shape of the ellipses of concentration defined by $\Sigma_{\boldsymbol{\mathcal{X}}}$.

Example 3

Consider a bivariate normal random vector $\boldsymbol{\mathcal{X}} \sim N_2(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu} = (0, 0)^T$ and

$$\Sigma = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}.$$

Then unit eigenvectors of Σ are

$$\mathbf{v_1} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } \mathbf{v_2} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

with corresponding eigenvalues of $\lambda_1 = 1$ and $\lambda_2 = 5$. The following figures give a geometrical interpretation to our entire discussion thus far.

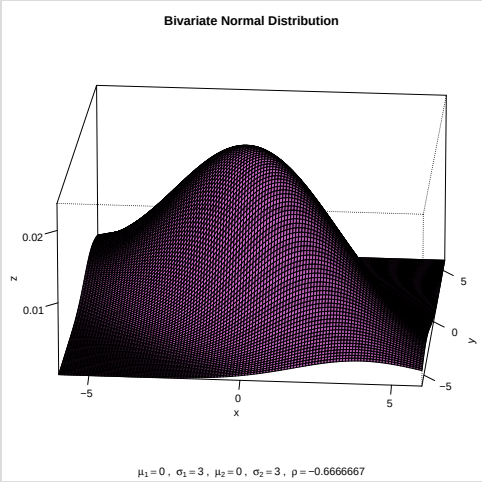

Figure 2.1: The density function of $\mathfrak{X}$ as a surface over $\mathbb{R}^2$.

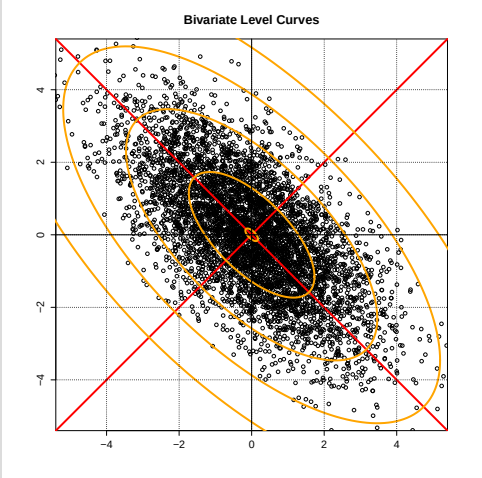

Figure 2.2: A sample of $n = 5000$ from $\mathfrak{X}$ overlain with the principal axes (red) and level curves of ϕ (orange).

◀

2.1.3 The Degenerate Σ

Now thus far we have assume that $\Sigma > 0$ and so that 0 wasn't an eigenvalue of Σ. However singular covariance matrices come up in practice as well as in theory so it is worth briefly addressing them.

The idea behind a degenerate $p \times p$ covariance structure is that all of the action is happening in some subspace of $\mathbb{R}^p$. A singular covariance matrix means there is some linear relationship among the columns and thus there is some linear relationship among the variables. This means that some linear combination has a a correlation of 1. Thus we aren't really in a p dimensional space because we don't really have p independent variables. Instead if $rank(\Sigma) = k < p$ then we only have k variables giving us different information and thus anything we sample from this distribution is going to lie in some subspace. Furthermore if Σ is close to being degenerate (some of the eigenvalues are quite small) then this will be approximately true and *most* of the action will be contained in some subspace. This will be an important concept to keep in mind.

If 0 is an eigenvalue corresponding to some eigenvector u of Σ then the variance $\mathfrak{X}$ in that direction, $u^T\mathfrak{X}$, is zero and thus the length of the principal axis in that direction is zero. Thus instead of having true p–dimensional ellipsoids of concentration we have

degenerate ones that line entirely in some proper subspace. The problem for defining a density function as before is that a distribution $\mathfrak{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ has zero density when viewed as a p–dimensional random vector because it lies in some subspace.

We can remedy this problem by ignoring the null space of Σ. We aren't really loosing anything because $Nul(\Sigma)$ has no variance. Thus if $rank(\Sigma) = k < p$ we can view $\mathfrak{X}$ as a k–dimensional random vector with support in the affine subspace

$$\boldsymbol{\mu} + Row(\Sigma) = \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{x} = \boldsymbol{\mu} + \mathbf{y} \text{ for } \mathbf{y} \in Row(\Sigma)\}.$$

By considering just those vectors in $\boldsymbol{\mu} + Row(\Sigma) = \boldsymbol{\mu} + Col(\Sigma)$ then we can define the density function for $\mathfrak{X}$ as

$$\boldsymbol{\phi}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|_+}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^+ (\mathbf{x} - \boldsymbol{\mu}))\right)$$

for $\mathbf{x} \in \boldsymbol{\mu} + Row(\Sigma)$. Note that $|\Sigma|_+$ is the pseudo-determinant of Σ defined by the product of its non-zero eigenvalues and Σ^+ is the pseudoinverse of Σ as previously discussed and $rank(\Sigma) = k$.

We aren't going to go into great detail about such degenerate normal distributions. All we need really note is that our entire previous conversation essentially maps directly to these distributions when considering them under the restricted domain of support. This claim is justified more or less *entirely* because if $\mathbf{x} \in \boldsymbol{\mu} + Row(\Sigma)$ then $\mathbf{x} - \boldsymbol{\mu} \in Row(\Sigma)$ and so $\Sigma^+ = \Sigma^{-1}$ for these $\mathbf{x} - \boldsymbol{\mu} \in Row(\Sigma)$.

2.2 Principal Components Analysis

2.2.1 Basic Idea

2.2.2 Correlation Among Components

2.2.3 Variance Maximization

2.2.4 Best Low-Rank Approximation

2.3 Examples

2.3.1 Classical PCA

2.3.2 Finding Interesting Projections

Chapter 3

The CUR Algorithm

3.1 The Definition In Literature

3.2 PCA-like uses of CUR

3.3 Leverage Scores

3.4 The k parameter

3.5 The c parameter

Chapter 4

Empirical Testing

4.1 Data and Metrics

4.2 Single Populations

4.2.1 Synthetic Data

4.2.2 Case Studies

4.3 Data with Groups

4.3.1 Synthetic Data

4.3.2 Case Studies