We covered these notes in the tutorial sessions. I strongly recommend that you further read the presented materials in classical books on linear algebra. Please make sure that you understand the proofs and that you can regenerate them. Should you have any questions, please feel free to contact the TAs.

Matrix Vector Products

Let A be a matrix in $R^{n\times n}$ whose (i,j)-th element is denoted by a_{ij}, and v be a vector in R^n whose i-th element is denoted by v_i. Note that the elements of A and v are real numbers. The length of v, denoted by $\|v\|$, is defined as

$$\|v\| = \sqrt{\sum_{i=1}^{n} v_i^2}$$

The matrix vector product Av is a vector in R^n defined as follows:

$$[Av]_i = \sum_{j=1}^{n} a_{ij}v_j$$

The relationship between v and Av is the focus of our discussion in this section.
There are two basic questions that one needs to deal with when A is multiplied by v: what happens to v; and what happens to A? The former question is that of deforming a vector whereas the latter is that of projecting several points. The answer to each question provides us with a different interpretation of Av.
When a vector is multiplied by a matrix, its orientation changes and its length gets scaled as the following example illustrates:
Example 1: Let A be given by

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \tag{1}$$

and take

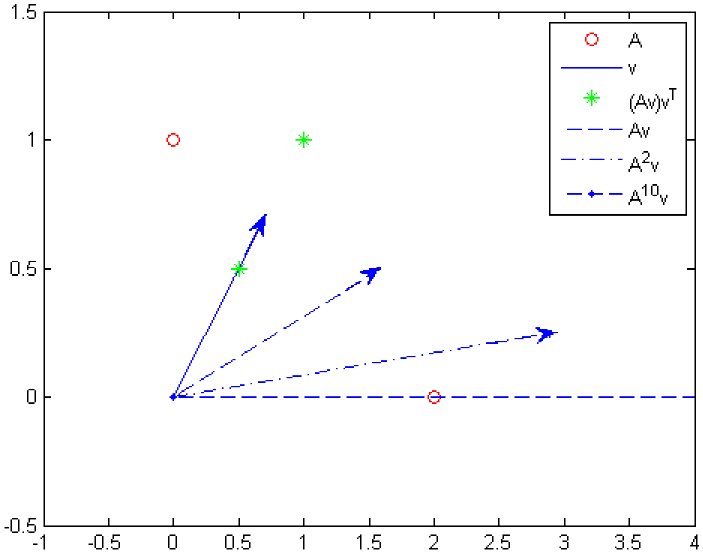
$$v = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}.$$

Note that v has unit length and makes an angle of $\frac{\pi}{4}$ with the x axis. The vector Av, however, has length $\|v\| = \sqrt{5}$ and orientation $atan(1/2) < \frac{\pi}{4}$. The amount by which the length and orientation of Ax differ from those of x are related to properties of A. Hence, Av is sometimes called the image of v under A.
Let A represent a set of m observations (rows of A) made over n features (column of A). Then, Ax represents a projection of m samples from A onto x. In other words, it is a one dimensional representation of A. To retrieve the coordinates of Ax in R^n, one needs to convert each element of the image into a vector along the direction of x. Note that the element $[Ax]_i$ is the length of the i-th row of A projected onto x. The coordinates of the projected data in the original coordinate are given by Axx^T. See Fig.1 for more details.

Eigenvalues and Eigenvectors

A scalar λ is called an eigenvalue of A if there exists $v \in R^n$ with $\|v\| \neq= 0$ such that

$$Av = \lambda v,$$



(a)

Figure 1: Different interpretations of Av. The red circles are two observations contained in A. Their projection on the unit length vector $v = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ is given by the green stars Avv^T. Therefore, Av gives us the length of the projected points on v. Furthermore, Av can also be viewed as a vector, which is called the image of v under A. Av is in general a vector with a different length and orientation than that of Av. In fact, if v does not rotate under A, then it must be an eigenvector (see the section on eigenvalues). If one images Av under A, the new image A^2v is further stretched and rotated. In this particular case, the length of the Av is $\sqrt{5/2}$ and the length of A^2v is $\sqrt{17/2}$. A^2v is more oriented toward the x-axis. If we keep imaging v under A, we obtain the largest eigenvector of A. One can see that $A^{10}v$ is almost aligned with the x-axis.

which implies $(A - \lambda I)v = 0$. The vector v is called a right eigenvector for A. Similarly, a vector satisfying $v^T A = \lambda v^T$ is a left eigenvector for A. From a geometrical point of view, an eigenvector is one that does not change its orientation when imaged under A (see Fig.1); it only gets scaled by a factor called the eigenvalue. Note that Av and λv have the same orientation and differ only by a constant factor. Furthermore, note that any λ at which $(A - \lambda I)$ becomes singular is an eigenvalue for A. If A has n distinct eigenvalues $\{\lambda_1, ..., \lambda_n\}$, we have

$$det(\lambda I - A) = (\lambda - \lambda_1)(\lambda - \lambda_2)...(\lambda - \lambda_n).$$

The right hand side of the above equation is called the characteristic polynomial of A. In fact, we can find the eigenvalues of A by finding the roots of the characteristic polynomial. This gives us an analytical approach to find the eigenvalues of either small matrices or matrices with special structures– this is not, however, computationally efficient and it is not how numerical solvers find eigenvalues. An eigenvector v_i of A belongs to the null space of $A - \lambda_i I$, which means that it satisfies the equality $(A - \lambda_i)v_i = 0$. If the eigenvalues are distinct, their corresponding eigenvectors are unique up to a scaling factor, i.e., there is a unique direction along which $(A - \lambda_i)v_i = 0$ occurs.

Example 2: Let A be given by Eq.1. The eigenvalues of A are the roots of

$$det(A - \lambda I) = (\lambda - 2)(\lambda - 1),$$

which are $\lambda_1 = 2$ and $\lambda_2 = 1$. The eigenvectors of A are given by

$$(A - \lambda_1 I) \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 0.$$

Then,

$$\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 0 \Longrightarrow$$

$$v_{11} = 1, v_{12} = 0$$

Note that $A^m v = \begin{bmatrix} 2^m & 0 \\ 0 & 1 \end{bmatrix} v$. One can see that $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is the unit length limit of $A^m v$ as $m \longrightarrow \infty$. This concept is illustrated in Fig.1.

We now consider cases where an eigenvalue λ_r is repeated. This happens when the characteristic polynomial has a repeated root:

$$det(\lambda I - A) = (\lambda - \lambda_r)^m (\lambda - \lambda_1)...(\lambda - \lambda_{n-m}), \;\; m > 1.$$

m is called the algebraic multiplicity of λ_r. In this case, the eigenvector satisfying $(A - \lambda_r I)v = 0$ may no longer be unique. More precisely, if $A - \lambda_r I$ has rank $n - m$ then there are m independent vectors that satisfy $(\lambda_r I - A)v = 0$. If this occurs, the number of independent (and, in fact, orthogonal) eigenvectors associated with λ_r, also called geometric multiplicity, is equal to m. However, if $A - \lambda_r I$ has rank $k > n - m$, we cannot find m independent vectors v that satisfy $(A - \lambda_r I)v = 0$. In this case, we need to introduce generalized eigenvectors to complete a basis for A. In what follows, let v_1 denote the regular eigenvector that satisfied $(A - \lambda_r I)v = 0$.

Example 3: Let A be the identity matrix.

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

The eigenvalues of A are given by $\lambda_1 = \lambda_2 = 1$. Therefore, 1 is a repeated eigenvalue with multiplicity 2. The eigenvectors of A must satisfy

$$(A - 1I)v = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} v = 0.$$

One can see that any vector in R^2 is an eigenvector for A. We can choose $v_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and $v_2 = \begin{bmatrix} 0 & 1 \end{bmatrix}$ to form an orthogonal basis.

Example 4: Let A be given by

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix}.$$

Again, A has a repeated eigenvalue at 1 with multiplicity 2. The associated eigenvectors must satisfy:

$$(A - 1I)v = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}.$$

In this case, there is a unique eigenvector (up to a scaling factor) $v_1 = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$ for $\lambda = 1$. This means that A is a defective matrix and we need to use generalized eigenvectors to complete a basis for A. One can show that $(A - 1I)v_2 = v_1$ yields,

$$v_2 = \begin{bmatrix} -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix}.$$

It can be verified that $(I - A)v_2 = v_1$ and that $Q = [v_1 v_2]$ forms a basis for A. Now we can prove an important property of symmetric matrices.

Lemma 1: There are no generalized eigenvectors for symmetric matrices.

Proof: Assume otherwise. Then, there exist vectors v_1 and v_2 for some symmetric matrix A, such that

$$(A - \lambda_r I)v_2 = v_1,$$

for some repeated eigenvalue λ_r. However, this implies that

$$v_1^T(A - \lambda_r I)v_2 = v_1^T v_1,$$

or

$$v_1^T A v_2 - \lambda_r v_1^T v_2^T = v_1^T v_1. \tag{2}$$

Note that for a general matrix A, if $Av = \lambda v$, we have $v^T A^T = \lambda v^T$. This means that a right eigenvector v is always a left eigenvector for A^T. However, if A is symmetric, we have

$$(Av)^T = v^T A^T = v^T A = \lambda_r v^T.$$

Replacing $v_1^T A$ in the first term of Eq.2 by $\lambda_r v^T$ yields

$$v_1^T v_1 = 0,$$

which implies $v_1 = 0$. However, this is a contradiction because $v_1 = 0$ cannot be an eigenvector. Therefore, A cannot have generalized eigenvectors. Q.E.D

Diagonalizing Symmetric Matrices

In the first homework, you were asked to prove that any matrix with a set of n linearly independent eigenvectors $Q = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}$ is diagonalizable, i.e., can be decomposed into $A = Q\Sigma Q^{-1}$. In general, defective matrices are not diagonalizable (A similar decomposition exists for defective matrices of the form $A = QJQ^{-1}$ where J contains the eigenvalues of A and is in the normal Jordan form). Therefore, using $Lemma 1$, one can conclude that symmetric matrices are always diagonalizable. The following Lemma proves a stronger property of symmetric matrices.

Lemma 2: Symmetric matrices are unitarily diagonalizable, i.e., they can be decomposed into $A = Q\Sigma Q^T$.

proof: Lemma 1 implies that Q contains n linearly independent eigenvectors. This means the decomposition $A = Q\Sigma Q^{-1}$ exists (homework 1). Therefore, it suffices to prove that Q is orthogonal, i.e., $Q^{-1} = Q^T$. To do so, one needs to prove that the columns of Q are orthogonal, i.e., $v_i^T v_j = 0$ for $i \neq j$ with $\|v_i\| = 1$ (why does this imply $Q^{-1} = Q^T$?). Two cases are to be considered: $\lambda_j \neq \lambda_i$ and $\lambda_j = \lambda_i$.
Case 1: $\lambda_i \neq \lambda_j$ It follows from the definition of eigenvectors that

$$\left.\begin{array}{c} Av_i = \lambda_i v_i \\ Av_j = \lambda_j v_j \end{array}\right\} \Longrightarrow \begin{array}{c} v_j^T A v_i = \lambda_i v_j^T v_i \\ v_i^T A v_j = \lambda_j v_i^T v_j \end{array}$$

The left hand sides are transposes of each other. Therefore,

$$(v_i^T A v_j)^T = \lambda_j v_j^T v_i = \lambda_i v_j^T v_i.$$

This implies that $(\lambda_i - \lambda_j) v_j^T v_i = 0$. Given that $\lambda_i \neq \lambda_j$, it must hold that $v_i^T v_j = 0$. Case 2: $\lambda_i = \lambda_j$ The proof follows directly from Lemma 1. Recall that since A is not defective, $A - \lambda_i I$ has rank $n - m$ with m being the multiplicity of λ_i. Therefore, there is an m dimensional subspace $\mathcal{N}$ (see Example 3) such that $(A - \lambda_i I)v = 0 \; \forall v \in \mathcal{N}$. One can always choose orthogonal vectors v_i and v_j from $\mathcal{N}$ such that $v_i^T v_j = 0$. Q.E.D

Principal Component Analysis

Let A be $m \times n$ containing m zero mean observations in n dimensions. The interest in this section is to find a 1 dimensional vector v that maximizes the variance of Av (the variance of the projection of A on v). It is proven in your textbook (see the chapter on PCA) that maximizing the variance amounts to minimizing the loss in projection . In other words,

$$argmin_{v \in R^m} \sum_{i=1}^{N} (a_i^T - a_i^T v v^T) = argmax_{v \in R^m} (v^T A^T A v).$$

Note that a_i^T refers to the i-th row of A. Let $\phi : R^m \longmapsto R$ denote the variance $\phi(y) = y^T A^T A y$. Maximizing $\phi(y)$ amounts to maximizing the norm of A. The following Lemma proves that the maximum is equal to the largest singular value of A and occurs at the principal eigenvector of $A^T A$.
Lemma 3: Let A be $m \times n$ containing m zero mean observations in n dimensions and define $\phi(y) = y^T A^T A y$. Let σ_{max} denote the largest eigenvalue of $A^T A$ and v_{max} be its associated eigenvector. It holds that:

$$v_{max} = argmax_{\|y\|=1} \phi(y).$$

In other words, the largest eigenvector of $A^T A$ is a maximizer of the variance ϕ over the set of unit length vectors.
proof 1: Since $A^T A$ is symmetric, it can be decomposed to $A^T A = Q \Sigma Q^T$ with $Q = [\; v_1 \;\; \cdots \;\; v_n \;]$. Given that Q forms a basis for $A^T A$, any vector y can be written as a linear combination of the columns of Q

$$y = \sum_{i=1}^{n} \alpha_i v_i,$$

with $\sum_{i=1}^{n} \alpha_i = 1$. The variance ϕ can be written as a function of α

$$\phi(\alpha_1, ..., \alpha_n) = (\sum_{i=1}^{n} \alpha_i v_i^T) Q^T \Sigma Q (\sum_{i=1}^{n} \alpha_i v_i^T).$$

Furthermore, note that $Q^T v_j = v_j$. Therefore,

$$\phi(\alpha_1, ..., \alpha_n) = (\sum_{i=1}^{n} \alpha_i v_i^T) \Sigma (\sum_{i=1}^{n} \alpha_i v_i^T)$$

Given that Σ is diagonal, and that $v_i^T v_j^T = 0$ for $i \neq j$, one can write

$$\phi(\alpha_1, ..., \alpha_n) = (\sum_{i=1}^{n} \alpha_i^2 \sigma_i).$$

It is clear that the maximum of $\phi(\alpha) = \sigma_{max}$ occurs at $\alpha_{max} = 1$, where α_{max} is the weight of the eigenvector for associated with σ_{max}. Thus, the maximizing y occurs at y_{max}. Q.E.D
proof 2: Consider the following convex optimization problem:

$$\begin{array}{ll} \text{maximize} & y^T A^T A y \\ \text{s.t.} & y^T y = 1, \end{array}$$

Define the Lagrangian L:

$$L(\lambda, y) = y^T A^T A y - \lambda(y^T y - 1).$$

The necessary optimality conditions require that

$$\frac{\partial L}{\partial y} = 0,$$

which means

$$A^T A y = \lambda y.$$

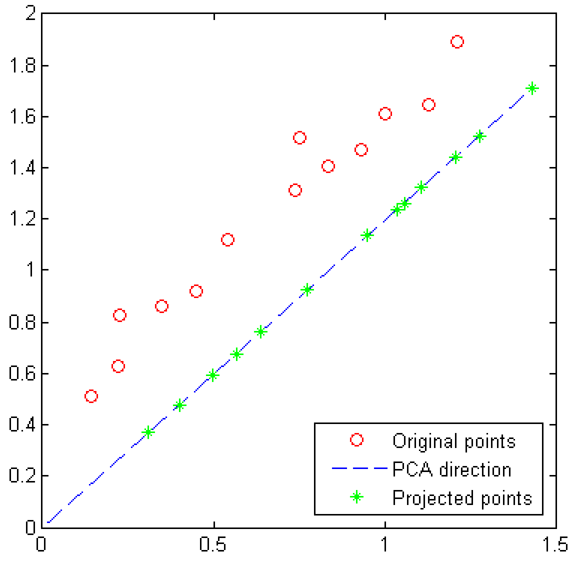
This implies that the optimal Lagrange multiplier λ is an eigenvalue of $A^T A$ and that the optimal y is an eigenvector of $A^T A$. Furthermore, note that $A^T A$ can be decomposed into $A^T A = Q \Sigma Q^T$. Imposing the constraint $y^T y = 1$ yields

$$L(y) = y^T Q^T \Sigma Q y, \ \ y \in \{v_1, ..., v_n\},$$

where v_i's are eigenvectors of $A^T A$. Let $y = v_k$ be k-th eigenvector. We have:

$$L(y, \lambda) = v_k^T Q^T \Sigma Q v_k = \sigma_k.$$

Therefore, the maximum value of L is σ_{max}, which occurs at $y = v_{max}$. Q.E.D
Example 5: Consider the set of points shown in Fig.2. The goal is to find the PCA components of the set. This requires us to find a line on which the projected points are as diverse as as possible. For instance, if the points are projected on the x-axis, the differences along the y-axis is lost. Note that if we change the y-coordinates of the points the x-axis does not change! Similarly, projecting points along the y-axis does not preserve the difference along the x-axis. However, as can be seen in the figure, the PCA line maximizes the projected variance by taking into account the differences in both coordinates. The code for generating the PCA direction is provided below.



(a)

Figure 2: A sample data set and its first principal components.

```
% choose arbitrary points and press enter to see the principal components
close all
[x,y]=getpts
plot(x,y,'ro');hold on
A=[x y];
Ab=A-repmat(mean(A),size(A,1),1);
[v,d]=eig(Ab'*Ab);
z=Ab*v(:,end)*v(:,end)';
z=z+repmat(mean(A)*v(:,end)*v(:,end)',size(z,1),1);
plot([min(z(:,1)) max(z(:,1))],...
[z(z(:,1)==min(z(:,1)),2) v(2,end)/v(1,end)*max(z(:,1))],'b--');
plot(z(:,1),z(:,2),'g*');
legend('Original points','PCA direction','Projected points')
axis equal
```