Linear Algebra Review

Vectors To begin, let us describe an element of the state space as a point with numerical coordinates, that is

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Vectors of up to three dimensions are easy to diagram. For example,

$$\boldsymbol{x} = \begin{pmatrix} 3 \\ 2 \\ 5 \end{pmatrix}$$

can be drawn as follows.

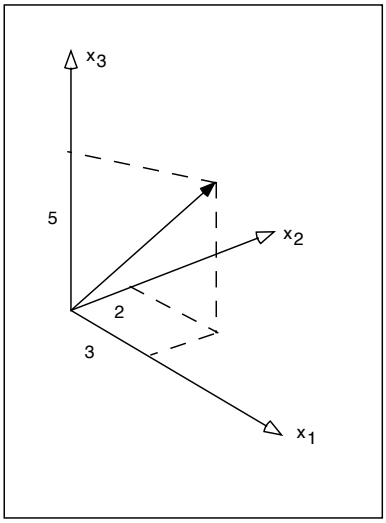


Figure 1: An example of a vector.

Addition To add two vectors together simply add their components. To multiply a vector by a scalar (number), multiply each of the components by the scalar. For example, if $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$, then if

$$\boldsymbol{x} = \begin{pmatrix} 3 \\ 5 \\ 2 \end{pmatrix} \text{ and } \boldsymbol{y} = \begin{pmatrix} -4 \\ 1 \\ 3 \end{pmatrix} \text{ then } \boldsymbol{z} = \begin{pmatrix} -1 \\ 6 \\ 5 \end{pmatrix}$$

And if $\boldsymbol{z} = \alpha\boldsymbol{x}$ for a scalar α, then

$$\boldsymbol{z} = \alpha \begin{pmatrix} 3 \\ 5 \\ 2 \end{pmatrix} = \begin{pmatrix} 3\alpha \\ 5\alpha \\ 2\alpha \end{pmatrix}$$

Dot Product The *dot product* of two vectors, denoted $\boldsymbol{x}\cdot\boldsymbol{y}$, is defined as the sum of the product of their pairwise components; that is,

$$\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

For our example, $\boldsymbol{x} \cdot \boldsymbol{y} = (3)(-4) + (5)(1) + (2)(3) = -1$.

Two vectors are said to be *orthogonal* if their dot product is zero; that is, $\boldsymbol{x} \cdot \boldsymbol{y} = 0$.

The *length* of a vector, denoted $||\boldsymbol{x}||$, is simply $\sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}$. The angle θ between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$\cos\theta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||\,||\boldsymbol{y}||}$$

From this equation it is seen that if vectors are orthogonal, then $\cos\theta = 0$ or $\theta = 90^o$.

The vector $(\boldsymbol{x} \cdot \boldsymbol{y})\frac{\boldsymbol{y}}{||\boldsymbol{y}||}$ is termed the *projection* of $\boldsymbol{x}$ onto $\boldsymbol{y}$.

The *transpose* of a matrix, denoted A^T, is simply described using the element notation as $\{a_{ji}\}$. In other words, the elements are "flipped" about the diagonal. A square $n \times n$ matrix is *symmetric* if $A^T = A$.

Linear Transformations For any function $f(x)$, a linear transformation is such that

$$f(ax + by) = af(x) + bf(y)$$

An important linear transformation is *matrix multiplication*. Matrix multiplication $A = BC$ is defined by

$$a_{ij} = \sum_{k=1}^{N} b_{ik}c_{kj}, i = 1, \ldots, P, j = 1, \ldots, Q$$

From this formula it is seen that the number of columns of B has to be the same as the number of rows of C for multiplication to be defined.

Determinant To define the *determinant* of a matrix first requires defining the number of inversions in a number sequence. Consider the sequence $\{1, 3, 4, 2\}$. The number of inversions in this sequence is 2 because 3 and 4 come after 2. Similarly the number of inversions in $\{4, 2, 1, 3\}$ is 3. Denote the number of inversions of a sequence as n. The determinant of a matrix A, denoted $|A|$, is the sum of all $n!$ possible different products that compose elements from columns of the matrix with a term that depends on the number of inversions in the row indices; that is,

$$|A| = \sum(-1)^{n(i_1,i_2,\ldots,i_n)} a_{i_1,1}a_{i_2,2} \cdots a_{i_n,n}$$

Like the inverse of a matrix, the determinant is expensive to calculate for large matrices, and a standard text

should be referred to for an algorithm. For practice calculations, however, it is useful to remember that the determinant of the 2×2 matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is given by

$$|A| = ad - bc$$

Inverse For square matrices where $N = M$, an important matrix is the *inverse* matrix A^{-1}, which is defined by

$$AA^{-1} = I$$

where I is the *identity matrix*

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & & 1 \end{bmatrix}$$

In general, like the determinant, inverses take some work to calculate, and you should find a numerical routine. For practice, however, it is useful to remember that the inverse of the 2×2 matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is given by

$$A^{-1} = \frac{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}{|A|}$$

Trace The *trace* of a matrix A is the sum of its diagonal elements; that is,

$$Tr(A) = \sum_{i=1}^{N} a_{ii}$$

Positive Definite A matrix A is *positive definite* if for every $\boldsymbol{x}$,

$$\boldsymbol{x}^T A \boldsymbol{x} > 0$$

and *positive semidefinite* if

$$\boldsymbol{x}^T A \boldsymbol{x} \geq 0$$

Orthonormal Transformation A transformation matrix A is orthonormal when

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

As a consequence

$$AA^T = I$$

1 Coordinate Systems

Let us start by considering how coordinate systems represent multi-dimensional data. Multiplying a matrix by a vector is a special case of matrix multiplication where

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

This can be written as $y_i = \sum_{k=1}^{N} a_{kj}x_j, i = 1, \ldots, M$. Alternatively we can see the transformation as a *linear combination* of the columns of A:

$$\boldsymbol{y} = \boldsymbol{a}_1 x_1 + \boldsymbol{a}_2 x_2 + \cdots + \boldsymbol{a}_N x_N$$

Often in manipulating vectors it is implicitly assumed that they are described with respect to an orthogonal coordinate system. Hence the actual coordinate vectors are not discussed. In the general case, however, the right coordinate system for data might not be orthogonal. To develop this point, consider first that the vectors $\boldsymbol{a}_i$ have a special interpretation as a coordinate system or *basis* for a multidimensional space. For example, in the traditional basis in three dimensions,

$$\boldsymbol{a}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{a}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \text{ and } \boldsymbol{a}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

allows $\boldsymbol{y}$ to be written as

$$\boldsymbol{y} = \boldsymbol{a}_1 y_1 + \boldsymbol{a}_2 y_2 + \boldsymbol{a}_3 y_3$$

A fundamentally important property of coordinate systems is that they are only describable *relative* to one another. For example, $\boldsymbol{y}$ is described in terms of the basis vectors $\boldsymbol{a}_i$.

This basis is orthogonal, since

$$\boldsymbol{a}_i \cdot \boldsymbol{a}_j = 0$$

for all i and j such that $i \neq j$, but it turns out that a nonorthogonal basis would also work. For example,

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

would still allow $\boldsymbol{y}$ to be represented (although the coefficients would of course be different). However, the matrix

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

would not work. In this case the reason is easy to see: there is no way of representing the third component. In general, to represent n-dimensional vectors, the basis must *span* the space. A general condition for this is that the columns of A must be *linearly independent*. Formally this means that the only way you could write

$$\boldsymbol{a}_1 x_1 + \boldsymbol{a}_2 x_2 + \cdots + \boldsymbol{a}_N x_N = \boldsymbol{0}$$

would be if $x_i = 0$ for all i.

What happens when the columns of an n-dimensional matrix do not span the space? The dimension of the matrix is equal to the number of linearly independent

vectors, which is known as the *rank* of the matrix. When the rank r is less than the dimension N, the vectors are said to span an r-dimensional *subspace*.

In some cases it is desirable for a matrix to have less than full rank. For example, for the equation

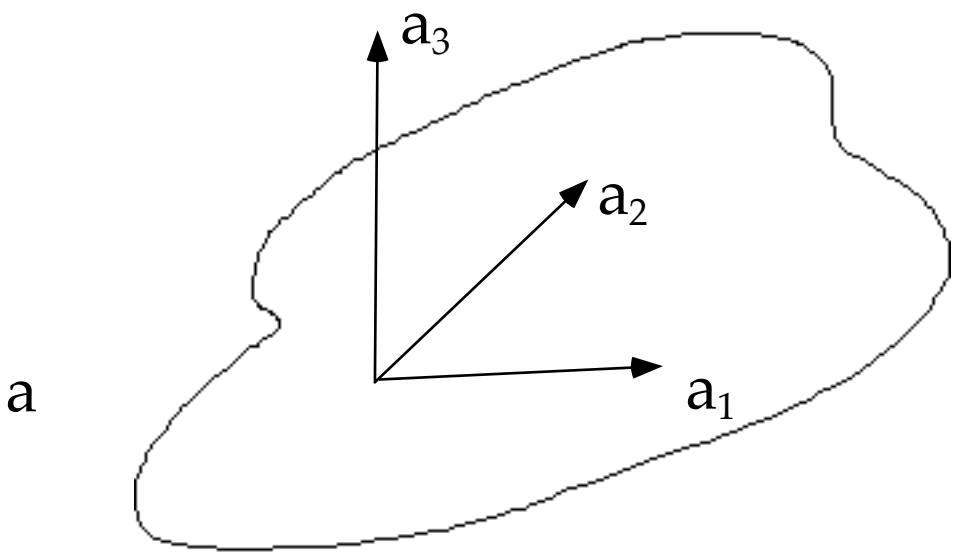
$$A\boldsymbol{x} = \mathbf{0}$$

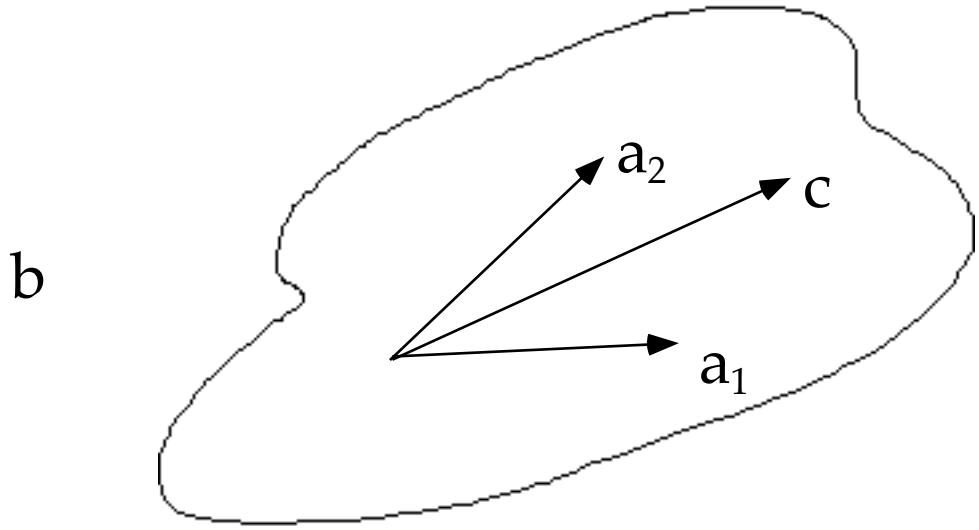
to have a nontrivial solution, the columns of A must be linearly dependent. Why? This equation is just a rewritten version of the previous equation. If the columns were linearly independent, then the only way the equations could be satisfied would be to have $x_i = 0$ for all i. But for a nontrivial solution the x_i should be nonzero. Hence for this to happen the columns must be linearly *de*pendent. For example, in three dimensions this can happen when all three of the vectors are in a plane.

In contrast, for the equation

$$A\boldsymbol{x} = \boldsymbol{c}$$

to have a unique solution the converse is true, because in order to have a solution, now the vector $\boldsymbol{c}$ must be expressed in terms of a linear combination of the columns $\boldsymbol{a}_i$. For this statement to be generally true the columns must span the space of $\boldsymbol{c}$; hence, together with the vector $\boldsymbol{c}$, they must be linearly dependent. These two cases are illustrated in Figure 2.





.

Figure 2: (a) The columns of A are linearly independent in three dimensions if the three column vectors are not coplanar. (b) For the case $A\boldsymbol{x} = \boldsymbol{c}$ the vector

Eigenvalues and Eigenvectors

At this point you should be used to the idea that any matrix can be thought of as representing a coordinate system. When a vector is multiplied by such a matrix, the general result is that the magnitude and direction of the resultant vector are different from the original. However, there is a very important special case. For any matrix, there are vector directions such that the matrix multiplication only changes the magnitude of the vector, leaving the direction unchanged. For these special directions, matrix multiplication reduces to scalar multiplication. The following example shows a case for the matrix

$$\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$$

where

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

is a special direction for the matrix, since multiplying it by the matrix just results in scaling the vector by a factor $\lambda = 4$; that is,

$$\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 4 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

In the general case, if a vector $\boldsymbol{v}$ lies along one of these directions,

$$W\boldsymbol{v} = \lambda\boldsymbol{v}$$

where λ is a scalar. Vectors that lie along these special directions are known as *eigenvectors*, and the scalars associated with a transformation matrix are known as

eigenvalues. Finding the eigenvalues of an $n \times n$ matrix for arbitrary n requires a trip to the recipe book, starting with the solution of an n^{th}-order polynomial to find the eigenvalues, but it is useful to work it out for an easy two-dimensional case, as follows:

$$\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

or, in other words,

$$\begin{bmatrix} 3-\lambda & 1 \\ 2 & 2-\lambda \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

From the previous section we know that for this equation to have a solution, the columns of the matrix must be linearly dependent, and thus $|W| = 0$. Thus

$$(3-\lambda)(2-\lambda) - 2 = 0$$

which can be solved to find the two eigenvalues $\lambda_1 = 4$ and $\lambda_2 = 1$. Now for the eigenvectors. Substituting $\lambda_1 = 4$ into the equation results in

$$\begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Now this set of equations is degenerate, meaning that there is only one useful equation in two unknowns. As a consequence there is an infinity of solutions, and you must pick one arbitrarily. Pick $v_1 = 1$. Then $v_2 = 1$. Thus the eigenvector associated with $\lambda_1 = 4$ is

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

As an easy exercise you can find the eigenvector associated with $\lambda_2 = 1$.

Now, for any particular matrix, why not pick the eigenvectors as the basis? It turns out that this is a good thing to do, since the effect is to transform the matrix into another matrix whose only nonzero elements are on the diagonal. Furthermore, these diagonal elements are the eigenvalues. The effect is to reduce matrix multiplication in the old basis to scalar multiplication in the new basis.

Changing Coordinates

What happens to transformations when the coordinate basis is changed? Suppose that the coordinate transformation is given by

$$\boldsymbol{x}^* = A\boldsymbol{x}$$

$$\boldsymbol{y}^* = A\boldsymbol{y}$$

Given the transformation

$$\boldsymbol{y} = W\boldsymbol{x}$$

what happens to W when the coordinate system is changed to the starred system? That is, for some W^* it will be true that

$$\boldsymbol{y}^* = W^*\boldsymbol{x}^*$$

What is the relation between W and W^*? One way to find out is to change back to the original system, transform by W, and then transform back to the starred system; that is,

$$\boldsymbol{x} = A^{-1}\boldsymbol{x}^*$$

$$\boldsymbol{y} = W\boldsymbol{x}$$

$$\boldsymbol{y}^* = A\boldsymbol{y}$$

Putting these transformations together:

$$\boldsymbol{y}^* = AWA^{-1}\boldsymbol{x}^*$$

Since the vector transformation taken by the two different routes should be the same, it must be true that

$$W^* = AWA^{-1}$$

Matrices related in this way are called *similar*.

Eigenvalue Transformations

Now let's relate this discussion to eigenvectors. Suppose that the eigenvectors have been chosen as the basis set. Then for a given eigenvector $\boldsymbol{y}_i$,

$$W\boldsymbol{y}_i = \lambda \boldsymbol{y}_i$$

and if Y is a matrix whose columns are the eigenvectors $\boldsymbol{y}_i$, then

$$WY = Y\Lambda$$

Here Λ is a matrix whose only nonzero components are the diagonal elements λ_i. Premultiplying both sides by Y^{-1},

$$Y^{-1}WY = \Lambda$$

What this equation means is that given a matrix W, the transformation it defines can always be simplified to that of a matrix whose only nonzero elements are diagonal by transforming to coordinates that use its eigenvectors as a basis. Furthermore, those elements are the eigenvalues.

Example

To check this result, let us use the earlier example where

$$W = \begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$$

and

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

and

$$\Lambda = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

Let us pick a particular vector as $\boldsymbol{x}$

$$\boldsymbol{x} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

so that $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = \begin{pmatrix} 13 \\ 14 \end{pmatrix}$$

First note that

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

so that $\boldsymbol{x}^* = Y^{-1}\boldsymbol{x}$, which is

$$\begin{pmatrix} \frac{10}{3} \\ -\frac{1}{3} \end{pmatrix}$$

and $\boldsymbol{y}^* = Y^{-1}\boldsymbol{y}$, which is

$$\begin{pmatrix} \frac{40}{3} \\ -\frac{1}{3} \end{pmatrix}$$

But from this you can see that $\boldsymbol{y}^* = \Lambda\boldsymbol{x}^*$.

Properties of Eigenvalues

1. An eigenvalue matrix Λ is invariant under any orthogonal transformation.
2. If all its eigenvalues are positive, a matrix A is positive definite.
3. The trace of A is the sum of all its eigenvalues and is invariant under any orthogonal transformation.
4. The trace of A^m is the sum of all its eigenvalues and is invariant under any orthogonal transformation.
5. The determinant of A is equal to the product of all its eigenvalues and is invariant under any orthogonal transformation.

Random Vectors

In this case vectors are drawn from some random distribution that captures the natural variations in the world. A random vector $\boldsymbol{X}$ is specified by a probability density function $p(\boldsymbol{X})$, where formally

$$p(\boldsymbol{X}) = \lim_{\Delta \boldsymbol{x}_i \to 0} \frac{P(\boldsymbol{X} \in \boldsymbol{I})}{\prod_i \Delta x_i}$$

where

$$\boldsymbol{I} = \{\boldsymbol{X} : x_i < X_i \leq x_i + \Delta x_i, \forall i\}$$

Although a random vector is fully characterized by its density function, such functions are often difficult to determine or mathematically complex to use. These limitations motivate modeling distributions with functions that can be described by a low number of parameters. The most important of such parameters are the *mean vector* and *covariance matrix*, which are just generalizations of the mean and variance of scalar random variables to vector random variables.

The mean vector is defined by

$$\boldsymbol{M} = E\{\boldsymbol{X}\} = \int \boldsymbol{X} p(\boldsymbol{X}) d\boldsymbol{X}$$

and the covariance matrix by

$$\Sigma = E\{(\boldsymbol{X} - M)(\boldsymbol{X} - M)^T\}$$

In practice, with real data you will use the sample mean vector and sample covariance matrix. Where $\boldsymbol{X}^k, k = 1, N$ are the samples,

$$M = \frac{1}{N}\sum_{k=1}^{N} X^k$$

$$\Sigma = \frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{X}^k - M)(\boldsymbol{X}^k - M)^T$$

You will need more than three samples in practice, but to illustrate the calculations, suppose

$$\boldsymbol{X}^1 = \begin{pmatrix} -1 \\ 3 \\ 1 \end{pmatrix},\ \boldsymbol{X}^2 = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix},\ \boldsymbol{X}^3 = \begin{pmatrix} 2 \\ 2 \\ 3 \end{pmatrix}$$

Then the mean value is

$$M = \frac{1}{3}\begin{pmatrix} 3 \\ 6 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

So that

$$\boldsymbol{X}^1 - M = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}, \boldsymbol{X}^2 - M = \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}, \boldsymbol{X}^3 - M = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$$

and the covariance matrix is given by

$$\Sigma = \frac{1}{3}\left\{\begin{bmatrix} 4 & -2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & -1 & -2 \\ -1 & 1 & 2 \\ -2 & 2 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 4 \end{bmatrix}\right\}$$

$$= \frac{1}{3}\begin{bmatrix} 6 & -3 & 0 \\ -3 & 2 & 2 \\ -2 & 2 & 8 \end{bmatrix}$$

High-Dimensional Spaces

Suppose, as is often the case, that the dimension of the space is extremely large. Now the standard way to proceed would be to choose eigenvectors u_k and eigenvalues λ_k of the sample covariance matrix Σ where

$$\begin{aligned}\Sigma &= \frac{1}{M}\sum_{n=1}^{M} \boldsymbol{X}_n \boldsymbol{X}_n^T \\ &= AA^T\end{aligned}$$

where

$$A = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M]$$

an $M \times N$ matrix of M data samples. The problem with this tack is that it is infeasible owing to the high dimensionality of the matrix Σ. Since for an image the dimension of $\boldsymbol{X}$ is n^2, then the dimension of Σ is $n^2 \times n^2$. For typical values of n, say 256, this is impossibly large. Salvation comes from the fact that the matrix Σ may be approximated by a matrix of lower rank. That is, most of the variation can be captured by projecting the data onto a subspace whose dimension is much less than the dimension of the space.

Rather than finding the eigenvectors of the larger system, consider finding the eigenvectors of the $M \times M$ system

$$A^T A\boldsymbol{v} = \mu\boldsymbol{v} \tag{1}$$

Premultiplying both sides by A,

$$AA^T A\boldsymbol{v} = \mu A\boldsymbol{v}$$

What this equation shows is that if $\boldsymbol{v}$ is an eigenvector of $A^T A$, then $A\boldsymbol{v}$ is an eigenvector of Σ. Furthermore, the eigenvalues of the smaller system are the same as those of the much larger system. It turns out also that these are the M largest eigenvalues. So to find the eigenvectors of the larger system, first find the eigenvalues and eigenvectors of the smaller system, and then multiply the eigenvectors by A.

Example: Face Recognition A lovely example of representing data is from face recognition. The task is to recognize the identity of an image of a face. The face image is described by an $N \times N$ array of brightness values. Given M exemplars of images with known identities, the objective is to take a new image and identify it with the training image to which it is most similar. The key element is in the similarity metric. It should be chosen to score the essential variations in the data. From the last section, the way to discover the essential variations is with principal components analysis, which identifies the eigenvalues of the covariance matrix of all the data.

To begin, identify the training set of images as $\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_M$. These are shown in Figure 3. To work with these it is useful to subtract the bias introduced, as all the brightness levels are positive. Thus first identify the "average face" (shown in Figure **??**)

$$\boldsymbol{I}_{ave} = \frac{1}{M} \sum_{n=1}^{M} \boldsymbol{I}_n$$

and then convert the training set by subtracting the average,

$$\boldsymbol{X}_i = \boldsymbol{I}_i - \boldsymbol{I}_{ave}, i = 1, \ldots, M$$

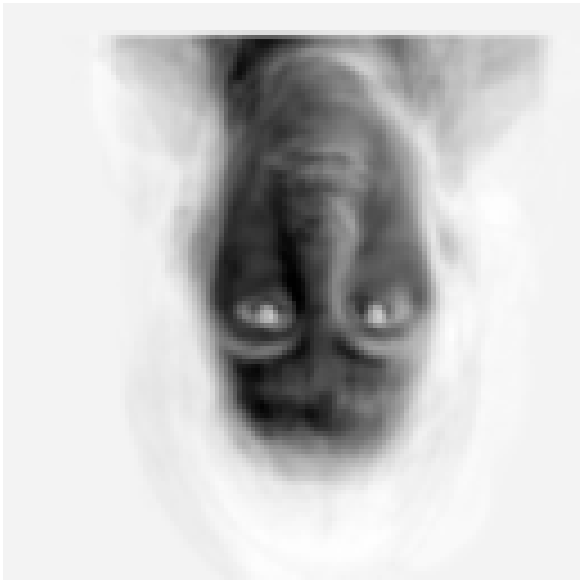
Now use Equation 1 to find M eigenvectors v_k and eigenvalues λ_k.

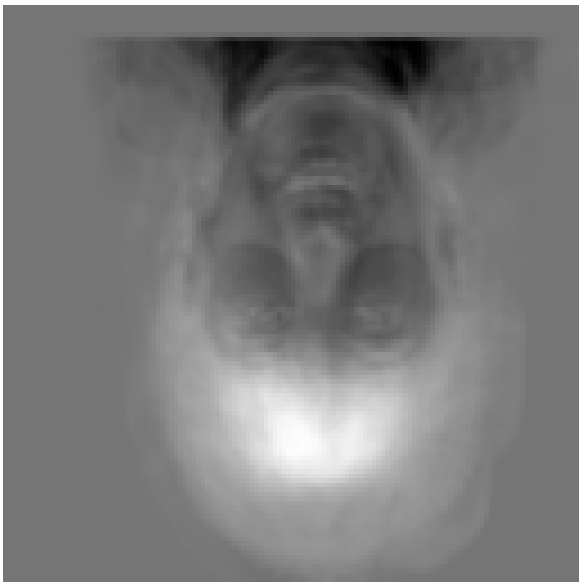
From the "short" eigenvectors (of length M), the larger eigenvectors $\boldsymbol{u}_k$, termed eigenfaces, can be constructed using $\boldsymbol{v}_k$, as follows:

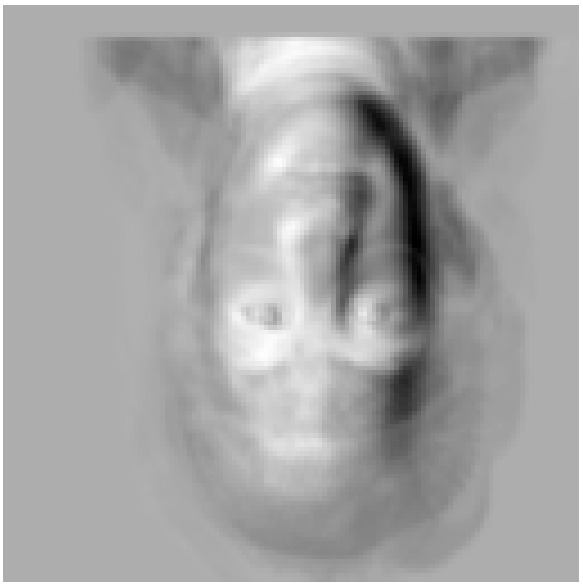
$$\boldsymbol{u}_i = \sum_{k=1}^{M} v_{ik} \boldsymbol{X}_k$$

Figure 3: A database of 12 figures used to calculate eigenvectors.

Figure **??** shows the first seven eigenvectors calculated from Equation 1.







Now that the direction principal variations have been calculated in the space of faces, this information can be used to classify a new face in terms of the faces in the data set. To do so, compute the coordinates of the new image in the eigenvector space $\boldsymbol{\Omega} = (\omega_1, \omega_2, \cdots, \omega_M)$ as follows:

$$\omega_k = u_k^T(\boldsymbol{I} - \boldsymbol{I}_{ave}), k = 1, \cdots, M$$

Next compare $\boldsymbol{\Omega}$ to the $\boldsymbol{\Omega}$s for each of the classes to pick the closest; that is, pick the class k that minimizes

$$||\boldsymbol{\Omega} - \boldsymbol{\Omega}_k||$$