CS 596: Lecture Notes Take 2

16:198:596

In this set of notes, we continue our discussion of linear algebra and its use in the representation of data and models.

Basis Sets, Change of Basis, Gram Schmidt, QR Factorization

Basis Sets

Recall from the previous set of notes the idea of a *basis set*: A basis of a vector space V is a set of linearly independent vectors $\{\underline{v}_1, \ldots, \underline{v}_N\}$ such that any vector in V can be represented as a linear combination of the $\{\underline{v}_i\}$. Effectively, the set $\{\underline{v}_i\}$ gives us a basis for representation in V. If we have that some vector $\underline{v}$ satisfies

$$\underline{v} = \alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \ldots + \alpha_N \underline{v}_N, \tag{1}$$

we can represent $\underline{v}$ with respect to this basis in coordinate form, by $\underline{v} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$. The properties of a basis guarantee that this representation is unique.

You can show for instance that the vectors $\underline{v}_1 = (1, 1)$, $\underline{v}_2 = (1, -1)$ are a basis for $\mathbb{R}^2$, and could use them in place of the *standard basis* $\underline{e}_1 = (1, 0), \underline{e}_2 = (0, 1)$, utilizing the relationship that

$$\begin{aligned} \underline{e}_1 &= \frac{1}{2}\left(\underline{v}_1 + \underline{v}_2\right) \\ \underline{e}_2 &= \frac{1}{2}\left(\underline{v}_1 - \underline{v}_2\right). \end{aligned} \tag{2}$$

Change of Basis

Typically we think of vectors as being expressed relative to the standard basis, where each basis vector is all 0s except for a single 1 in a primary direction in $\mathbb{R}^N$. However, it can and will be useful to be able to *change basis* of a vector. That is, given a vector $\underline{x}$ and a basis set $\underline{v}_1, \ldots, \underline{v}_N$, we'd like to express $\underline{x}$ relative to its coordinates in this other basis set, call it $\underline{x}_\alpha$.

If $\underline{x}_\alpha = (\alpha_1, \ldots, \alpha_N)$ is the vector of coordinates with respect to this basis set, we have the relationship that

$$\underline{x} = \alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \ldots \alpha_N \underline{v}_N, \tag{3}$$

or, letting V be the matrix where the i-th column is given by the vector $\underline{v}_i$, we have the relationship that

$$\underline{x} = V \underline{x}_\alpha. \tag{4}$$

This suggests that if V is invertible (as it should be if it is square and the columns are linearly independent, as basis vectors must be), we have the simple **change of basis formula** given by

$$\underline{x}_\alpha = V^{-1} \underline{x}, \tag{5}$$

allowing us to go from a coordinate representation in the standard basis to a coordinate representation in the V-basis.

The result in Eq. (2) can be thought of as giving a change of basis matrix

$$V^{-1} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix} \tag{6}$$

It the basis vectors $\underline{v}_1, \ldots, \underline{v}_N$ are additionally *orthonormal vectors* - that is the norm of each vector is 1 and the dot product of each vector with the other vectors is 0, then you get immediately that $V^{-1} = V^{\mathrm{T}}$ - that is, the inverse or change of basis matrix is nothing more than the transpose of the original basis matrix. This highlights one of the computational simplicities associated with *orthonormal basis sets.*

Gram-Schmidt Algorithm for Orthonormal Bases

Given a vector space V, we have many reasons to want an orthonormal basis set for that vector space. Let's suppose that we have that V is spanned by the linearly independent vectors $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_N$. The **Gram-Schmidt** algorithm gives a procedure for sequentially producing an orthonormal basis set $\underline{u}_1, \ldots, \underline{u}_N$ for this vector space, based on these vectors.

The procedure begins by simply normalizing $\underline{a}_1$, to give

$$\underline{u}_1 = \frac{1}{||\underline{a}_1||_2}\underline{a}_1, \tag{7}$$

so that $\underline{u}_1$ is a unit vector in the direction of $\underline{a}_1$.

The next step is to modify $\underline{a}_2$ by *removing its projection in the direction of* $\underline{u}_1$. That is, $\underline{a}_2$ can be thought of as some amount of displacement in the direction of $\underline{u}_1$, and some amount of displacement in an orthogonal direction. Fortunately, the projection in the $\underline{u}_1$ direction is easy to compute, based on the geometric interpretation of the dot product. This gives us

$$\underline{q}_2 = \underline{a}_2 - (\underline{a}_2.\underline{u}_1)\underline{u}_1 \tag{8}$$

as a vector orthogonal to $\underline{u}_1$. To see this note that the above cannot be zero (since it is just a non-trivial linear combination of $\underline{a}_1$ and $\underline{a}_2$), and we have that $\underline{q}_2.\underline{u}_1 = 0$, that they are orthogonal. To see this, observe that

$$\underline{q}_2.\underline{u}_1 = (\underline{a}_2 - (\underline{a}_2.\underline{u}_1)\underline{u}_1).\underline{u}_1 = \underline{a}_2.\underline{u}_1 - \underline{a}_2.\underline{u}_1||\underline{u}_1||_2^2 = \underline{a}_2.\underline{u}_1 - \underline{a}_2.\underline{u}_1 = 0. \tag{9}$$

Having produced a vector $\underline{q}_2$ that is in the span of $\underline{a}_1$ and $\underline{a}_2$, but orthogonal to $\underline{u}_1$, we can produce a unit vector in that direction simply by normalizing:

$$\underline{u}_2 = \frac{1}{||\underline{q}_2||_2}\underline{q}_2. \tag{10}$$

Iteratively, we can continue removing the projections of the orthogonal directions we've produced to get the 'remainder' of $\underline{a}_k$, which must be orthogonal to everything else so far:

$$\begin{aligned} \underline{q}_k &= \underline{a}_k - \sum_{i=1}^{k-1}(\underline{a}_k.\underline{u}_i)\underline{u}_i \\ \underline{u}_k &= \frac{1}{||\underline{q}_k||_2}\underline{q}_k \end{aligned} \tag{11}$$

It follows then that because each $\underline{u}_i$ is a linear combination of $\underline{a}_1, \ldots, \underline{a}_i$, that the collection of the $\underline{u}_i$ spans the same space as the original vectors, but are all unit vectors (norm 1) and mutually orthogonal; that is, $\underline{u}_1, \ldots, \underline{u}_N$ forms an orthonormal basis of the space V.

As a question to think about: What would happen if you applied Gram-Schmidt to a set of vectors that wasn't linearly independent?

QR Factorization

The above idea of Gram-Schmidt orthonormalization gives rise to a matrix representation known as QR-Factorization. Suppose that A is a matrix, columns given by $\underline{a}_1, \ldots, \underline{a}_N$. The orthonormal basis produced by applying Gram-Schmidt can be thought of as producing a matrix Q of orthonormal columns and a matrix R such that

$$A = QR \tag{12}$$

Taking the columns of Q to correspond to the sequentially generated orthonormal vectors after processing the columns of A, we would have that the first column of R should be have a value of $||\underline{a}_1||_2$ in the upper entry and 0s everywhere else. Working backwards in this way, the second column of A should be a linear combination of the first two columns of Q - just as in the Gram-Schmidt procedure where the second orthonormal vector is a linear combination of the first two $\underline{a}_1, \underline{a}_2$, the coefficients being easy to work out from the Gram-Schmidt procedure. In this way, R is revealed to be an upper triangular matrix.

This factorization of A as QR, Q the columns of Q forming an orthonormal basis and R being an upper triangular matrix has useful implications for various applications, such as solving linear systems. Given a system of the form

$$A\underline{x} = \underline{b}, \tag{13}$$

if we have the QR factorization of A available we can write it as

$$QR\underline{x} = \underline{b} \tag{14}$$

or, noting that $Q^{-1} = Q^{\mathrm{T}}$,

$$R\underline{x} = Q^{\mathrm{T}}\underline{b}, \tag{15}$$

which is an upper-triangular system of linear equations, which can be solved straightforwardly using the Gaussian elimination procedure without any row pivoting.

Question to Think About: How can you derive the matrix R from the Gram-Schmidt Procedure?

Eigenvectors and Eigenvalues of Linear Transforms

Recalling our discussion of linear transforms: linear transforms of vector spaces amount to common manipulations like stretching, rotating, or reflecting in space. In general, these can be modeled in terms of multiplication by a matrix, $F(\underline{x}) = A\underline{x}$ for some matrix A.

For a given transformation or matrix, of particular interest are vector $\underline{v}$ and scalars λ such that:

$$A\underline{x} = \lambda\underline{x}. \tag{16}$$

Recalling the idea that scaling by a constant only stretches or shrinks a vector, the above indicates that $\underline{x}$ is special in the sense that the 'effect' of A is simply to stretch $\underline{x}$ by a factor of λ. This is the **eigenvector** $\underline{x}$ with corresponding **eigenvalue** λ. Briefly, the eigenvectors of a matrix are those for which the action of the transformation A is *particularly simple*.

Examples, Markov Chains, Significance

A classic example of this kind of thing is in analyzing Markov chains or transition matrices P, where row $\underline{P}_i$ of P represents the probabilities of transitioning from a given state j to a given state i in one timestep. In this case, the

'stationary' distribution is defined to be the likelihood that the system is in a given state 'after a long period of time'. Let $\underline{\pi}$ be the column vector of probabilities where π_j is the probability of being found in state j. A distribution is 'stationary' if it satisfies

$$\underline{\pi} = P\underline{\pi}, \tag{17}$$

or the likelihood of being in a certain state after one time step doesn't change.

But we can additionally interpret the above as saying that the stationary distribution of a transition matrix is an eigenvector of that matrix with eigenvalue 1: $P\underline{\pi} = 1\underline{\pi}$.

Author's Note: I'm not entirely sure where the tradition of referring to stationary distributions with the symbol π is from, b

One of the significances of eigenvectors, and one which will be practically useful shortly when we discuss Principal Component Analysis and dimensionality reduction, is when we can express other vectors in terms of those eigenvalues. For instance, suppose A has a collection of eigenvector/value pairs $\{(\underline{v}_i, \lambda_i)\}$, and in particular assume that they form a basis. Suppose that some vector $\underline{v}$ has the form

$$\underline{v} = \alpha_1\underline{v}_1 + \alpha_2\underline{v}_2 + \ldots + \alpha_N\underline{v}_N. \tag{18}$$

In this case, multiplication by A is easy to compute and understand:

$$\begin{aligned} A\underline{v} &= A(\alpha_1\underline{v}_1 + \alpha_2\underline{v}_2 + \ldots + \alpha_N\underline{v}_N) \\ &= \alpha_1(A\underline{v}_1) + \alpha_2(A\underline{v}_2) + \ldots + \alpha_N(A\underline{v}_N) \\ &= \alpha_1\lambda_1\underline{v}_1 + \alpha_2\lambda_2\underline{v}_2 + \ldots + \alpha_N\lambda_N\underline{v}_N. \end{aligned} \tag{19}$$

In short, the effect of A is simply to stretch the various components of $\underline{v}$ with respect to this basis of eigenvectors. This is very straightforward to implement.

When we apply this idea to principal component analysis, etc, we'll essentially be utilizing the eigenvectors of an operator to express our dataset in convenient, representative forms that makes analysis easy.

Computing Eigenvalues and Eigenvectors

Having discussed the what of eigenvalues and some of the why, this leads us to the how: given a matrix A as input, how can we find the eigenvalues and eigenvectors of that matrix?

The classical observation to make here is this: if $A\underline{x} = \lambda\underline{x}$, and we essentially want to solve for λ and $\underline{x}$, we can rearrange this to

$$A\underline{x} - \lambda\underline{x} = 0, \tag{20}$$

or equivalently that (introducing a factor of the identity matrix I for convenience),

$$[A - \lambda I]\,\underline{x} = 0. \tag{21}$$

The above suggests that some non-trivial linear combination of the columns of $A - \lambda I$ is the zero vector, i.e., the columns of $A - \lambda I$ are *linearly dependent*. As discussed previously, this connects directly to the concept of invertibility of a matrix: if a matrix's columns are linearly dependent, it is not invertible, and vice versa. One step further, a matrix is not invertible if and only if the determinant of that matrix is zero. Hence we arrive at our condition for eigenvalues: λ is an eigenvalue of A if and only if it satisfies

$$\det(A - \lambda I) = 0. \tag{22}$$

Recall that the determinant is a scalar function, a product and sum of the various terms of the specified matrix, which means that the above represents a polynomial in terms of the unknown variable λ.

The eigenvalues of a matrix A are specified by the roots of the **characteristic polynomial** $\det(A - \lambda I) = 0$.

Computing the characteristic of a polynomial, we can solve for the roots, and in doing so recover the eigenvalues of A.

As an example of this, considering the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \tag{23}$$

we can show (utilizing the formula for 2x2 matrix determinants) that the characteristic polynomial is given by

$$\det(A - \lambda I) = (1 - \lambda)^2 - 4 = \lambda^2 - 2\lambda - 3, \tag{24}$$

which has roots at $\lambda = -1, \lambda = 3$. These are the eigenvalues of this matrix.

It's worth noting here that while the characteristic polynomial does give a specification of the eigenvalues, it is not the only or necessarily best way to compute them - I will post some additional algorithms on this. In particular, calculating the roots of a polynomial (especially extremely high degree) can be quite difficult, and in practice it is frequently done by finding a matrix that gives that polynomial as its characteristic polynomial, and finding the eigenvalues of that matrix!

Given the eigenvalue λ, how do we solve for the eigenvector? One observation here is that given λ, the matrix $A - \lambda I$ is well defined, and we could consider solving the linear system $[A - \lambda I]\underline{x} = 0$ through any of our usual approaches. In particular here, thinking about the eigenvector corresponding to $\lambda = -1$, we get

$$\begin{pmatrix} 1 - (-1) & 2 \\ 2 & 1 - (-1) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0, \tag{25}$$

or

$$\begin{aligned} 2x + 2y &= 0 \\ 2x + 2y &= 0 \end{aligned} \tag{26}$$

the fact that we get redundancy in our equations is not surprising since the matrix we are multiplying by is not invertible (by construction!) and therefore the system is underspecified. But the above indicates that any vector of the form $\underline{x} = (t, -t)$ will be an eigenvector corresponding to the eigenvalue $\lambda = -1$.

Note the lack of specification here - any vector of that form will do. While the eigenvalue is generally highly specified, the corresponding eigenvector is underspecified, because any scalar multiple of the eigenvector will also be an eigenvector. If $A\underline{x} = \lambda\underline{x}$, we additionally have that

$$A(\alpha\underline{x}) = \alpha A\underline{x} = \alpha\lambda\underline{x} = \lambda(\alpha\underline{x}). \tag{27}$$

In this case, we see that any scalar multiple of the vector $(1, -1)$ suffices as an eigenvector of $\lambda = -1$. Standard practice is frequently to normalize the eigenvectors to unit vectors, to get in this case $\underline{x} = (1/\sqrt{2}, -1/\sqrt{2})$.

The other eigenvector could be found similarly.

Properties of Eigenvalues, Eigenvectors, and Symmetric Matrices

One observation to make here is that because eigenvalues are the roots of a polynomial, eigenvalues can potentially be complex. This should not come as a great surprise, as if we consider R as a rotation matrix in two dimensions

for instance, $R\underline{x}$ does not simply 'stretch' any particular $\underline{x}$ or even leave it unchanged with an eigenvalue of 1 - you can show that the eigenvalues of rotation matrices for instance are complex. But the math of eigenvalues and eigenvectors extends quite naturally in that regard.

We'll see in a moment that a common special case is to consider A as a symmetric matrix, and under this assumption the eigenvalues are all real. But that warrants its own subsection.

An important general result is this: **the eigenvectors corresponding to distinct eigenvalues are linearly independent.**

To see this, we start by assuming eigenvectors $\underline{v}_1, \ldots, \underline{v}_k$ that are have distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ but are linearly dependent. We will derive a contradiction. If they are linearly dependent, then we have some non-zero linear combination such that

$$0 = \alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \ldots + \alpha_{k-1}\underline{v}_{k-1} + \alpha_k \underline{v}_k. \tag{28}$$

Consider multiplying this by the matrix $A - \lambda_k I$. In this case, we get

$$\begin{aligned}
[A - \lambda_k I]\, 0 &= [A - \lambda_k I]\, \alpha_1 \underline{v}_1 + [A - \lambda_k I]\, \alpha_2 \underline{v}_2 + \ldots + [A - \lambda_k I]\, \alpha_{k-1}\underline{v}_{k-1} + [A - \lambda_k I]\, \alpha_k \underline{v}_k \\
&= \alpha_1 \left[A\underline{v}_1 - \lambda_k I \underline{v}_1\right] + \alpha_2 \left[A\underline{v}_2 - \lambda_k I \underline{v}_2\right] + \ldots + \alpha_{k-1}\left[A\underline{v}_{k-1} - \lambda_k I \underline{v}_{k-1}\right] + \alpha_k \left[A\underline{v}_k - \lambda_k I \underline{v}_k\right] \\
&= \alpha_1 \left[\lambda_1 \underline{v}_1 - \lambda_k \underline{v}_1\right] + \alpha_2 \left[\lambda_2 \underline{v}_2 - \lambda_k \underline{v}_2\right] + \ldots + \alpha_{k-1}\left[\lambda_{k-1}\underline{v}_{k-1} - \lambda_k \underline{v}_{k-1}\right] + \alpha_k \left[\lambda_k \underline{v}_k - \lambda_k \underline{v}_k\right] \\
&= \alpha_1(\lambda_1 - \lambda_k)\underline{v}_1 + \alpha_2(\lambda_2 - \lambda_k)\underline{v}_2 + \ldots + \alpha_{k-1}(\lambda_{k-1} - \lambda_k)\underline{v}_{k-1} + \alpha_k * 0 * \underline{v}_k,
\end{aligned} \tag{29}$$

or, simplifying

$$0 = \alpha_1(\lambda_1 - \lambda_k)\underline{v}_1 + \alpha_2(\lambda_2 - \lambda_k)\underline{v}_2 + \ldots + \alpha_{k-1}(\lambda_{k-1} - \lambda_k)\underline{v}_{k-1}. \tag{30}$$

We see that multiplying by $A - \lambda_k I$ was able to 'kill' the $\underline{v}_k$ term, and introduce a lot of non-zero factors on the other terms. If we continue in this way, multiplying by all the $A - \lambda_i I$ for $i = k-1, k-2, \ldots, 3, 2$, we are left with

$$0 = \alpha_1 \prod_{i=2}^{k} (\lambda_1 - \lambda_k)\underline{v}_1, \tag{31}$$

which requires, since $\underline{v}_1$ is non-zero and the eigenvalues are distinct, that $\alpha_1 = 0$.

This argument generalizes to give that all the α_i must be zero, in which case the $\underline{v}_1, \ldots, \underline{v}_k$ are linearly independent.

This points in the direction we are heading - if the eigenvectors are linearly independent, and there are enough of them, they may span the entire vector space, and in doing so form a basis. If we have a basis set of eigenvectors, then computations with respect to A in terms of this basis are going to be very easy.

Symmetric Matrices and their Eigenvector/Values

Recall the idea that a symmetric matrix is one where the i-th column and the i-th row are identical to one another. We can summarize this by saying $A = A^{\mathrm{T}}$, A is identical to its transpose. These matrices occur frequently in machine learning and data science, so the properties described here will be very useful in our applications (again, one of the places we are heading being principal component analysis and dimensionality reduction).

If a matrix is symmetric, it has the property that the eigenvalues of that matrix are going to be **real**. We saw an example of this previously in the eigenvalue computation of a previous section. The proof of this requires some knowledge of complex numbers and analysis (it is skippable without):

Suppose that $\underline{v}$ is an eigenvector with eigenvalue λ. Let $\underline{v}^*$ be the *complex conjugate* of $\underline{v}$. In this case, we have that

$$(\underline{v}^*)^{\mathrm{T}} A\underline{v} = (\underline{v}^{\mathrm{T}})\lambda\underline{v} = \lambda(\underline{v}^*)^{\mathrm{T}}\underline{v} = \lambda||\underline{v}||_2^2. \tag{32}$$

Taking the conjugate transpose of the above (and noting that the conjugate transpose of A is A, being symmetric and real!)

$$\lambda^*||\underline{v}||_2^2 = \left[(\underline{v}^*)^{\mathrm{T}} A\underline{v}\right]^{\mathrm{T}^*} = (\underline{v}^*)^{\mathrm{T}} A^{\mathrm{T}^*}\underline{v} = (\underline{v}^*)^{\mathrm{T}} A\underline{v} = (\underline{v}^*)^{\mathrm{T}}\lambda\underline{v} = \lambda||\underline{v}||_2^2. \tag{33}$$

Dividing out the norm we get that $\lambda^* = \lambda$, which is only possible if the eigenvalue λ is itself a real number.

Hence we see that the eigenvalues of a symmetric real matrix are themselves all real. Additionally we have that, from the previous observations, the eigenvectors of distinct eigenvalues are linearly independent. We can extend this slightly in the case of symmetric matrices to the fact that *all* the eigenvectors are linearly independent, and that a symmetric matrix will in fact have a full complement of eigenvectors, N for an $N \times N$ matrix. *The algebra of this is more advanced I think than I want to include here, but it involves a decomposition of the vector space into eigen-subspaces, and finding basis vectors in each of these sub-spaces.*

In this way, the eigenvectors of a symmetric matrix actually form a basis of the underlying vector space. And we can go one step further here: **for symmetric real matrices, there exists an orthonormal basis of eigenvectors**. We've shown already that eigenvectors (with distinct eigenvalues) are linearly independent, it remains to show that they are orthogonal to each other.

Suppose that $\underline{v}_i$ and $\underline{v}_j$ are eigenvectors of A with corresponding eigenvalues λ_i, λ_j. In this case, we have that

$$\begin{aligned} \underline{v}_i^{\mathrm{T}} A\underline{v}_j = \underline{v}_i^{\mathrm{T}}\lambda_j\underline{v}_j = \lambda_j\left(\underline{v}_i.\underline{v}_j\right) \\ \underline{v}_j^{\mathrm{T}} A\underline{v}_i = \underline{v}_j^{\mathrm{T}}\lambda_i\underline{v}_i = \lambda_i\left(\underline{v}_j.\underline{v}_i\right) \end{aligned} \tag{34}$$

However, since we have that $\left[\underline{v}_i^{\mathrm{T}} A\underline{v}_j\right]^{\mathrm{T}} = \underline{v}_j^{\mathrm{T}} A\underline{v}_i$ *(why?)*, we have that

$$\lambda_j\left(\underline{v}_i.\underline{v}_j\right) = \lambda_i\left(\underline{v}_j.\underline{v}_i\right) \tag{35}$$

or

$$(\lambda_j - \lambda_i)(\underline{v}_i.\underline{v}_j) = 0. \tag{36}$$

Since the eigenvalues are distinct, that difference is non-zero, and we are left with the fact that $\underline{v}_i.\underline{v}_j = 0$, and the vectors are orthogonal.

So tying this all together: any symmetric real matrix as an orthonormal basis of eigenvectors associated with it. We will explore this more deeply in the next section.

Similar Matrices

As a closing point for this section, I want to introduce the idea of **similar matrices**. Two matrices A, B are similar if they are related in the following way: for some invertible matrix V, we have that

$$A = VBV^{-1}. \tag{37}$$

One way to interpret this, taking V^{-1} as a change-of-basis matrix, is that the 'effect' of A with respect to the standard basis is the same as doing the following: shift to a basis in terms of V, apply B in this new basis, then

shift back to the standard basis. That is, the effect of multiplying by A is the same as multiplying by B when in a different basis.

We can put this in more concrete terms thinking about the eigenvalues and eigenvector of A and B: while they may not share eigenvectors, they will share eigenvalues. To see this, note that if $\underline{x}$ is an eigenvector of A with eigenvalues λ, so that $A\underline{x} = \lambda\underline{x}$, we have that

$$VBV^{-1}\underline{x} = \lambda\underline{x} \tag{38}$$

or

$$B\left[V^{-1}\underline{x}\right] = \lambda\left[V^{-1}\underline{x}\right]. \tag{39}$$

Hence, $V^{-1}\underline{x}$ is an eigenvalue of B with the same eigenvalue λ.

If we think about the eigenvectors of A as the 'directions where what A is doing is stretching', then B is accomplishing the exact same stretching as A (by the same factors λ), simply in different directions. This motivates to some extent the term 'similar' - A and B are 'doing the same thing', simply in different directions.

Similar matrices will come up a lot in data analysis, as it is frequently not the data itself that matters but the relative position, information, and transformations between the data points that matters. Two 'similar' data sets will frequently have the same underlying relationships.

A final observation on similar matrices: note from the previous discussions on determinants that

$$\det(A) = \det(VBV^{-1}) = \det(V)\det(B)\det(V^{-1}). \tag{40}$$

Since V is invertible, we have that $\det(V^{-1}) = 1/\det(V)$, which means that

$$\det(A) = \det(B). \tag{41}$$

This connects to our previous idea of determinants in terms of transformations: the determinant of a matrix M is the amount of stretch or distortion on the volume of the unit cube in terms of the standard basis, when that cube is multiplied or transformed by M. In the case of similar matrices, A and B are doing the *same* transformation, the same stretch and distortion of that cube, just from different perspectives/bases. And as a result, the volume is preserved.

Diagonalizable Operators

Among matrices (a more general argument would discuss the idea of Operators, transformations of the underlying data), *diagonal* matrices are particularly nice. Both in terms of computation (a power of a diagonal matrix is computed by powering the diagonal elements) and in terms of effect - multiplication by a diagonal matrix simply scales the corresponding elements of the vector.

Obviously, most matrices *aren't* diagonal, but a special case of matrices are *similar* to diagonal matrices. A matrix A is **diagonalizable** if for some invertible matrix V,

$$A = V\Delta V^{-1}, \tag{42}$$

where Δ is a real, diagonal matrix (there is a place for complex matrix entries but we largely ignore it here).

Note that in this case, the image of the standard basis vectors ($V\underline{e}_k$) are the eigenvectors of A, with eigenvalues corresponding to the k-th entry λ_k on the diagonal of Δ. That is, Δ is the diagonal matrix of the eigenvalues of A. Based on the previous description of similarity, we see that a matrix A is diagonalizable if the effect of

transformation by A is effectively scaling (by the eigenvalues λ_i) in *some* set of linearly independent directions. *Why linearly independent directions?* In this way, V represents the basis matrix where the columns are the eigenvectors of A.

If the eigenvectors of A are orthonormal, then this becomes particular simple - the inverse of V is simply V-transpose, and we have that for such a diagonalizable A

$$A = V\Delta V^{\mathrm{T}}. \tag{43}$$

This will be particularly relevant to symmetric matrices, in a moment.

Computations with diagonalizable matrices are particularly nice: note that $A^k = V\Delta^k V^{-1}$, and in this case we can use the fast exponentiation of Δ to compute the powers of A very quickly. Additionally, based on the properties of similar matrices from the previous section, we have that

$$\det A = \det \Delta, \tag{44}$$

and it's relatively straightforward to show that $\det \Delta$, for any diagonal matrix Δ, is the product of the entries. In this way, we get the following interesting result:

> For any diagonalizable matrix A, the determinant of A is given by the product of the eigenvalues.

This can be seen from either the classical formula for the determinant (noting that the only way to select D distinct rows and columns to give a non-zero product is along the diagonal) or from the perspective of transformation - the transformation of Δ is to stretch the standard basis by the appropriate factors of λ_i, multiplying the total volume of the standard unit cube by $\prod \lambda_i$, hence the effect of the similar matrix A must be the same.

At this point, there are three questions of practical and theoretical interest regarding diagonalizable matrices:

- What matrices are diagonalizable?
- Is the diagonalization unique?
- How can we compute the diagonalization of a matrix?

To answer these questions in reverse order: the problem of computing the diagonalization is subsumed somewhat by computing the eigenvectors and eigenvalues, allowing you to construct V and Δ respectively (and therefore V^{-1}). Those problems have been discussed somewhat already but will be discussed further in a section to follow.

To the question of uniqueness: very frequently a diagonalizable matrix will not have a unique diagonalization. Note, for instance, the identity matrix I. For any invertible matrix V, we have as a valid diagonalization of I,

$$I = V\Delta V^{-1}, \tag{45}$$

where $\Delta = I$. Because this holds for any V we can hardly call the diagonalization unique.

The problem here is somewhat threefold.

First, note that for any diagonalization matrix V and diagonal matrix Δ, we can construct additional diagonalization pairs V', Δ' by re-arranging the columns of V and Δ (though we must also re-arrange the rows of Δ to maintain it as a diagonal matrix). Letting P be some permutation matrix, so that $V' = VP$ and $\Delta' = P^{\mathrm{T}}\Delta P$, we have that

$$V'\Delta'(V')^{-1} = VP(P^T\Delta P)(VP)^{-1} = VPP^{\mathrm{T}}\Delta PP^{-1}V^{-1} = V\Delta V^{-1}, \tag{46}$$

utilizing the fact that $P^{\mathrm{T}} = P^{-1}$. Hence any re-ordering of the columns of V can also yield a diagonalization.

But the diagonalization is not unique up to choice of column order either - note that we see a similar problem here to one previously, that for a given eigenvalue the eigenvector is not unique. In particular, if $A\underline{x} = \lambda\underline{x}$, any scalar multiple of A will also be an eigenvector with the same eigenvalue. We could circumvent this somewhat with the restriction that all the eigenvectors of A be taken to be of norm 1, in which case we have uniqueness of eigenvectors (up to reflection).

Recalling the observation from the previous section, that eigenvectors corresponding to distinct eigenvalues are linearly independent, this suggests that if A has D-many distinct eigenvalues, then it will have D-many normalized eigenvectors (unique up to reflection) that could be arranged as the columns (unique up to ordering) of a diagonalization. This is a partial answer to the problem of uniqueness for diagonalizations.

A problem remains, however, if there are linearly independent eigenvectors corresponding to the same eigenvalue. In particular, suppose that $\underline{v}_1, \underline{v}_2$ are eigenvectors of A, linearly independent, corresponding to some common eigenvalue λ. This can be thought of as a situation where the characteristic polynomial $\det(A - \lambda I) = 0$ has a repeated root at a given λ. In this case, note that any linear combination of the two, $\alpha\underline{v}_1 + \alpha\underline{v}_2$ is also an eigenvector with eigenvalue λ, and we again lose uniqueness of choice of eigenvectors corresponding to this eigenvalue.

Note that if there are linearly independent eigenvectors corresponding to the same eigenvalue, then the span of these corresponds to an *eigenspace*, where anything in that space is an eigenvector with the same eigenvalue.

Based on these observations, we can consider the following cases:

- In the case that A has D distinct eigenvalues, or $\det(A - \lambda I) = 0$ has D distinct roots, then the eigenvectors of A are naturally linearly independent, and can be taken as the basis of a diagonalization (unique up to order and scale).
- In the case that A has eigenvalues of repeated multiplicity: suppose for each eigenvalue λ_i, it has multiplicity m_i in the characteristic polynomial; if for each such i, the eigenspace associated with λ_i has dimension m_i, then A can be diagonalized (unique up to order and choice of basis for each eigenspace).
- In the case that A has eigenvalues of repeated multiplicity: suppose for some eigenvalue λ_i with multiplicity $m_i > 1$, the eigenspace associated with λ_i has dimension less than m_i; in this case, A is not diagonalizable.

The first two cases essentially correspond to 'nice' matrices that we hope to encounter - matrices of the last form are typically said to be defective.

As an example of a defective, non-diagonalizable matrix, consider

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \tag{47}$$

In this case, we have that $\det(A - \lambda I) = (1 - \lambda)^2$, giving a repeated root of $\lambda = 1$ of multiplicity 2. However, the eigenspace associated with this eigenvalue must satisfy $A.\underline{x} = \underline{x}$, or

$$\begin{aligned} x_1 &= x_1 \\ x_1 + x_2 &= x_2 \end{aligned} \tag{48}$$

which is satisfied by anything of the form $\underline{x} = (0, y)$, an eigenspace of dimension 1. In this case, no diagonalization of A exists.

The algebra involved with these defective matrices is well studied and interesting but largely irrelevant for our own purposes. Typically, matrices we are interested in will be diagonalizable, because of the following results:

Symmetric Matrices and Diagonalization

We have the following result:

For any symmetric real matrix A, A is diagonalizable as

$$A = V\Delta V^{\mathrm{T}} \tag{49}$$

where the matrix V is the matrix composed of the orthonormal basis of eigenvectors of A, and Δ is a diagonal real matrix of the eigenvalues of A.

This follows from the results discussed on symmetric matrices in the previous section. Note, based on the above discussion of diagonalizability, we get from this that the eigenspaces associated with the eigenvalues of symmetric matrices must all have dimension equal to their multiplicities.

One convenient result of this so called spectral theorem is that for any symmetric matrix A, with eigenvector/value pairs $(\lambda_i, \underline{v}_i)$, we have the following *representation* of A:

$$A = \sum_{i=1}^{D} \lambda_i \underline{v}_i \underline{v}_i^{\mathrm{T}} \tag{50}$$

Symmetric matrices occur in a number of areas in machine learning (as discussed in the next section) so they warrant some brief discussion of their properties.

A symmetric matrix A is said to be **positive semi-definite** if for any vector $\underline{x}$, we have that

$$\underline{x}^{\mathrm{T}} A \underline{x} \geq 0. \tag{51}$$

This is sometimes tightened to **positive definite** if for any non-zero vector $\underline{x}$, we have strict positivity, $\underline{x}^{\mathrm{T}} A \underline{x} > 0$.

We can connect this to eigenvalues via the spectral theorem. Since the eigenvectors of A form an orthonormal basis, we can represent $\underline{x}$ with respect to this basis as

$$\underline{x} = \alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \ldots + \alpha_D \underline{v}_D, \tag{52}$$

in which case

$$\begin{aligned}
\underline{x}^{\mathrm{T}} A \underline{x} &= \left(\sum_{i=1}^{D} \alpha_i \underline{v}_i \right)^{\mathrm{T}} \left(\sum_{i=1}^{D} \lambda_i \underline{v}_i \underline{v}_i^{\mathrm{T}} \right) \left(\sum_{i=1}^{D} \alpha_i \underline{v}_i \right) \\
&= \left(\sum_{i=1}^{D} \alpha_i \underline{v}_i^{\mathrm{T}} \right) \left(\sum_{i=1}^{D} \sum_{j=1}^{D} \lambda_i \alpha_j \underline{v}_i \underline{v}_i^{\mathrm{T}} \underline{v}_j \right) \\
&= \left(\sum_{i=1}^{D} \alpha_i \underline{v}_i^{\mathrm{T}} \right) \left(\sum_{i=1}^{D} \lambda_i \alpha_i \underline{v}_i \right) \\
&= \sum_{i=1}^{D} \sum_{j=1}^{D} \lambda_j \alpha_j \alpha_i \underline{v}_i^{\mathrm{T}} \underline{v}_j \\
&= \lambda_1 \alpha_1^2 + \lambda_2 \alpha_2^2 + \ldots + \lambda_D \alpha_D^2.
\end{aligned} \tag{53}$$

From the above, we can see that if A is positive definite, it must be that all the λ_i are greater than 0 *(Why?)*. Similarly, if all the λ_i are greater than 0, then A must be positive definite *(Why?)*. Both together, a symmetric matrix A is positive definite if and only if its eigenvalues are strictly positive.

We can define negative definiteness similarly, and get a similar result for negative eigenvalues of A.

One nice result of this is the following: consider the possible values of $\underline{x}^{\mathrm{T}} A \underline{x}$ as a function of $\underline{x}$. If A is positive definite, these could go anywhere from 0 to ∞ *(Why?)*. But if we consider the following constrained optimization problem:

$$\text{maximize } \underline{x}^{\mathrm{T}} A \underline{x} \text{ subject to } ||\underline{x}||_2 = 1, \tag{54}$$

This has a solution given by the eigenvector (or eigenvectors) corresponding to the largest eigenvalue of A. The minimization problem is solved similarly by the eigenvector or vectors with the smallest eigenvalue of A. *Why? How does this follow from the previous results? What do you know, relative to A, about $\underline{x}$ if $||\underline{x}||_2 = 1$?*

This result is significant both in the general problem of computing eigenvectors and eigenvalues, as well as the more specific problem of principal component analysis, as outlined in the next section.

Symmetric Matrices in Machine Learning

Given a collection of N many D-dimensional data points, we typically organize them into an N-row, D-column matrix X, where each row $\underline{x}_i$ represents a data point in our data set.

We typically consider two important derived matrices from this data matrix:

- The Similarity Matrix: $S = XX^{\mathrm{T}}$. (Note, S is an $N \times N$ dimensional matrix.)
- The Covariance Matrix: $C = X^{\mathrm{T}}X$. (Note, C is a $D \times D$ dimensional matrix.)

Note that both of these matrices are symmetric - we have that $S^{\mathrm{T}} = [XX^{\mathrm{T}}]^{\mathrm{T}} = (X^{\mathrm{T}})^{\mathrm{T}}X^{\mathrm{T}} = XX^{\mathrm{T}} = S$, and similar for C. As such, both these matrices are guaranteed an orthonormal basis of eigenvectors and real valued eigenvalues. The spectra (eigenvalues and vectors) of these matrices have particular significance for the data. We can further guarantee that both these matrices are positive semi-definite. Note that in the case of C, we have that

$$\underline{v}^{\mathrm{T}} C \underline{v} = \underline{v}^{\mathrm{T}} X^{\mathrm{T}} X \underline{v} = (X\underline{v})^{\mathrm{T}}(X\underline{v}) = ||X\underline{v}||_2^2 \geq 0. \tag{55}$$

As such, the eigenvalues of both of these matrices will be strictly non-negative.

The Similarity Matrix

Notice that S is an $N \times N$ matrix, where entry $S_{i,j} = \underline{x}_i^{\mathrm{T}} . \underline{x}_j$, the dot product of the i, j-data points. Thinking of the dot product of two vectors as measuring their relative similarity to each other (identical matrices have a maximized dot product, completely orthogonal matrices have a dot product of 0), the matrix S captures all the relative similarities and therefore relative relationships between each of the data points.

As there are many sets of data that might capture the same 'relative' information, a common trick is to 'standardize' the data set in the following way: observe that being symmetric, we have that

$$S = V \Delta V^{\mathrm{T}} \tag{56}$$

where V is the orthonormal basis matrix and Δ is the diagonal matrix of corresponding eigenvalues. Note that since the eigenvalues are non-negative, we have that $\Delta = \Sigma^2$ for some diagonal matrix Σ of square roots. Hence,

$$S = V \Delta V^{\mathrm{T}} = V \Sigma^2 V^{\mathrm{T}} = V \Sigma \Sigma^{\mathrm{T}} V^{\mathrm{T}} = [V\Sigma]\,[V\Sigma]^{\mathrm{T}}. \tag{57}$$

From the above, we can interpret S as the similarity matrix arising from the pseudo-data matrix $X' = V\Sigma$. The rows of X' can be thought of as expressions of each data point in terms of an underlying linear featurespace of the original data set, that preserves the relative relationships.

One nice thing about similarity matrices is that we can generalize this: for a given data set X, imagine defining a similarity matrix by some kind of *non-linear kernel*, such as taking $S_{i,j} = \exp(-||\underline{x}_i - \underline{x}_j||_2^2/\sigma^2)$, for some parameter σ. Notice that the more similar $\underline{x}_i, \underline{x}_j$ are, the larger $S_{i,j}$ is, and the more distinct they are the smaller $S_{i,j}$ is. Being symmetric, S can again be factored into the form $S = [V\Sigma][V\Sigma]^{\mathrm{T}}$, which taking $X' = V\Sigma$ re-expresses each data point in terms of an underlying non-linear feature space of the original data set! This can be a valuable transformation of your data if the underlying relationships you're looking for are non-linear.

The Covariance Matrix

Taking $C = X^{\mathrm{T}}X$, note that $C_{i,j} = \sum_{t=1}^{N} x_t^i x_t^j$, i.e., if we look at $C_{i,j}/N$, this should approximate the expected value of $X_i * X_j$ over our underlying distribution of data. If the data is additionally *mean centered*, i.e., every row of X is in fact $\underline{x}_i - \frac{1}{N}(\underline{x}_1 + \ldots + \underline{x}_N)$, then $C_{i,j}$ approximates the covariance between the i, j features: $C_{i,j}/N \approx E[(X_i - E[X_i])(X_j - E[X_j])]$.

Further, it's interesting to note the significance of $\underline{v}^{\mathrm{T}}C\underline{v}$: from the above, this is given by $||X\underline{v}||_2^2$, or

$$\underline{v}^{\mathrm{T}}C\underline{v} = \sum_{i=1}^{N}(\underline{x}_i.\underline{v})^2, \tag{58}$$

or in other words, $\underline{v}^{\mathrm{T}}C\underline{v}$ gives the *total square projection or variance of the data set in the direction* $\underline{v}$. Taking $\underline{v}$ to be a unit vector, and asking what direction *maximizes* this variance, we see from the previous section that the answer is given by the eigenvector corresponding to the largest eigenvalue of C.

This is the *first principal component* - the direction that captures the maximal variance in the data set. Taking $\underline{v}$ to be the unit eigenvector corresponding to the eigenvalue $\lambda_{\max}$ (which must be non-negative, given the positive semi-definiteness of C), we have that

$$\sum_{i=1}^{N}(\underline{x}_i.\underline{v})^2 = \underline{v}^{\mathrm{T}}C\underline{v} = \underline{v}^{\mathrm{T}}(\lambda_{\max}\underline{v}) = \lambda_{\max}||\underline{v}||_2^2 = \lambda_{\max}, \tag{59}$$

and the total variance or projection of the data in the indicated direction is given by $\lambda_{\max}$.

Similarly, we can consider the diagonalization of C as

$$C = V\Delta V^{\mathrm{T}}, \tag{60}$$

typically ordering V and Δ with decreasing eigenvalues along the diagonal. In this case, the first element of the diagonal will be $\lambda_{\max}$, and the first column of V corresponds to this first principal component of the data. The second largest eigenvalue in Δ corresponds to the *direction of largest variance in the data, orthogonal to* $\underline{v}_1$, and each eigenvalue/vector after measures the direction of maximum remaining variance and amount of variance, until $\lambda_{\min}$ is reached, corresponding to the direction of minimal variance in the data.

Note that in this case, we get (via re-arrangement) that

$$\Delta = V^{\mathrm{T}}CV = V^{\mathrm{T}}X^{\mathrm{T}}XV = [XV]^{\mathrm{T}}[XV]. \tag{61}$$

Interpreting the above matrix Δ as the covariance matrix of the modified data matrix $X' = XV$ *(Is this data matrix still mean centered? Why or why not?)*, we see that the data features in this new system are effectively *decoupled* - distinct features in X' have a covariance of 0.

Note that the i-th row of $X' = XV$ is given by the dot product of $\underline{x}_i$ with each of the columns or eigenvectors of V, each eigenvector corresponding to an orthogonal 'direction of maximum variance' as shown above. Frequently in data sets, unless the data is effectively spherically distributed, there will be some directions with high variance (large spread to your data cloud), some directions with very low variance (directions in which your data cloud is very thin or flat). These directions of low variance contain, practically speaking, very little information about the whole data set - in the case where the variance in some direction is 0, this corresponds to the data actually being flat in that direction, and the high dimensional data set can be reduced to a lower dimensional data set.

Even if the data is not perfectly flat in these directions (corresponding to the eigenvectors of C with the lowest eigenvalues), it can be useful to essentially discard information in these directions. Letting V_k be the first k columns of V (corresponding to the directions of the k largest eigenvalues), the modified data set

$$X' = XV_k \tag{62}$$

represents the projection of the data into this smaller k-dimensional subspace, in a way that preserves the maximum variance of the data (and therefore the maximum information about the data). This is a common starting point to the idea of dimensionality reduction for data sets.

Algorithms for Computing Eigenvalues and Eigenvectors

As discussed previously, the eigenvalues of A are given by the roots of the polynomial $\det(A - \lambda I) = 0$. If you have a good mechanism for computing determinants as well as a good mechanism for computing roots of polynomials, this is a very valuable formulation of the problem.

However, computing roots of polynomials can in fact be quite hard (and computing determinants is itself an interesting problem). It can in fact be beneficial to, when solving for the roots of a polynomial, construct a matrix A with that polynomial as $\det(A - \lambda I)$, and try to solve for the eigenvalues of that matrix A. But this leads again to the question of how to compute the eigenvalues of A.

Power Iteration

Recall the idea that for a given matrix A, the 'action' of A is to stretch along the directions of its eigenvectors, an amount given by the eigenvalue. That is, if $\underline{x} = \sum \alpha_i \underline{v}_i$, then $A\underline{x} = \sum \alpha_i \lambda_i \underline{v}_i$. If all the eigenvectors start out normalized, and we imagine repeatedly applying A, i.e., computing $A^k \underline{x}$, we have

$$A^k \underline{x} = \alpha_1 \lambda_1^k \underline{v}_1 + \alpha_2 \lambda_2^k \underline{v}_2 + \ldots + \alpha_D \lambda_D^k \underline{v}_D, \tag{63}$$

and the resulting vector is very quickly going to be dominated by the eigenvector corresponding to the largest eigenvalue of A. Assume that $\underline{v}_1$ corresponds to the largest eigenvalue of A - in this case, we have that

$$A^k \underline{x} = \lambda_1^k \left[\alpha_1 \underline{v}_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1} \right)^2 \underline{v}_2 + \ldots + \alpha_D \left(\frac{\lambda_D}{\lambda_1} \right)^k \underline{v}_D \right]. \tag{64}$$

Hence, as long as α_1 is non-zero, we have that the limit of $A^k \underline{x}$ converges to $\lambda_1^k \alpha_1 \underline{v}_1$.

The method of **Power Iteration** utilizes this idea in the following way: letting $\underline{x}_0$ be some initial guess, construct the sequence of iterates given by

$$\underline{x}_{k+1} = \frac{A\underline{x}_k}{||A\underline{x}_k||_2}. \tag{65}$$

Each iterate $\underline{x}_k$ will point in the direction of $A^k\underline{x}_0$, but be rescaled to be a unit vector. The above argument shows that (if $\underline{x}_0$ has a component in the $\underline{v}_1$ direction, and the largest eigenvalue has a unique eigenvector associated with it) the iterates $\underline{x}_k$ will converge to $\underline{v}_1$. This is very useful in the case of symmetric matrices A, where an orthonormal basis of eigenvectors is guaranteed - the only potential issues being that of convergence (if the second largest eigenvalue is close to or the same as the largest eigenvalue).

If A has an eigenspace of dimension greater than 1 corresponding to its largest eigenvalue, what happens to the iterates $\underline{x}_k$?

Note that the above scheme is convenient for computing the dominant eigenvector, but not the dominant eigenvalue - this is easy to rectify however, as if the dominant eigenvector $\underline{v}_1$ computed, the corresponding eigenvalue can be recovered from dividing the elements of $A\underline{v}_1$ by the elements of $\underline{v}_1$ to recover the scale factor λ_1.

While the above gives a good scheme for determining the dominant eigenvalue and eigenvector of A, what of the remaining eigenvalues and eigenvectors?

Suppose, to simplify things, A is given as a symmetric matrix. In this case, via the spectral theorem, we have that

$$A = \sum_{i=1}^{D} \lambda_i \underline{v}_i \underline{v}_i^{\mathrm{T}}. \tag{66}$$

If we have computed $\lambda_1, \underline{v}_1$, however, we can essentially subtract this off to create a new symmetric matrix:

$$A' = A - \lambda_1 \underline{v}_1 \underline{v}_1^{\mathrm{T}}, \tag{67}$$

where the eigenspace corresponding to $\underline{v}_1$ for A has been effectively replaced by an eigenspace with eigenvalue 0.

Note then, that the second largest eigenvalue of A is now the first largest eigenvalue of A', and we can iteratively repeat the process.