Section 09: PCA and SVD

1. Singular Value Decomposition - Proofs

Recall that if we have a symmetric, square matrix A $\in \mathbb{R}^{n\times n}$, we can eigen-decompose it in the form of $A = USU^T$, where the columns of U are eigenvectors of A with lengths of 1, and the diagonal of S is the list of eigenvalues corresponding to those eigenvectors.
Now, for a more general case, where A is a data matrix with the dimension of $\mathbb{R}^{n\times d}$, there is still a way to decompose it: $A = USV^T$, where $U \in \mathbb{R}^{n\times n}$, S is a rectangular diagonal matrix and $S \in \mathbb{R}^{n\times d}$, and $V \in \mathbb{R}^{d\times d}$.
It is called Singular Value Decomposition (SVD).

(a) Let A have SVD USV^T. Show AA^T has the columns of U as eigenvectors with associated eigenvalues S^2.
Solution:

We have $A = USV^T$ then:

$$\begin{aligned} AA^T &= USV^T(USV^T)^T \\ &= USV^T((V^T)^T S^T U^T) \\ &= USV^T VSU^T \\ &= USV^T VSU^T \\ &= USISU^T \\ &= US^2U^T \end{aligned}$$

Since we can diagonalize AA^T into US^2U^T, it has eigenvectors that are columns of U and associated eigenvalues S^2.

(b) Let A have SVD USV^T. Show A^TA has the columns of V as eigenvectors with associated eigenvalues S^2.
Solution:

We have $A = USV^T$ then:

$$\begin{aligned} A^TA &= (USV^T)^T USV^T \\ &= VSU^T USV^T \\ &= VSISV^T \\ &= VS^2V^T \end{aligned}$$

Since we can diagonalize A^TA into VS^2V^T, it has eigenvectors that are columns of V and associated eigenvalues S^2.

(c) For the matrix A, suppose we are given that $AA^T = US^2U^T$ and $A^TA = VS^2V^T$. Show that $A = USV^T$. I.e., show that for any vector $x \in \mathbb{R}^d$, we have $Ax = USV^Tx$
Solution:

Let $\{v_1, v_2, \ldots, v_n\}$ be the rows of V^T. They are orthogonal to each other and unit norm. For any $x \in \mathbb{R}^d$

we can write $x = \sum_{i=1}^{d} \alpha_i v_i$. Then we have:

$$\begin{aligned} USV^T x &= USV^T \sum_{i=1}^{d} \alpha_i v_i \\ &= \sum_{i=1}^{d} \alpha_i USV^T \ v_i \\ &= \sum_{i=1}^{d} \alpha_i USe_i \\ &= \sum_{i=1}^{d} \alpha_i U\lambda_i e_i \\ &= \sum_{i=1}^{d} \alpha_i \lambda_i u_i \end{aligned}$$

In the meantime, since v_i is an eigenvector of $A^T A$, we have $A^T A v_i = \lambda_i^2 v_i$ (1)
Multiply A on both sides of (1), we get $(AA^T)Av_i = \lambda_i^2 Av_i$
Therefore, Av_i is an eigenvector of AA^T
If we multiply v_i^T on both sides of (1), we get $v_i^T A^T A v_i = \lambda_i^2 v_i^T v_i$, which is equivalent to $||Av_i||^2 = \lambda_i^2 ||v_i||^2$
Therefore, we know that the length of vector Av_i is λ_i
Normalize the vector: $\frac{Av_i}{\lambda_i} = u_i$
Hence, $\lambda_i u_i = Av_i$
Plug it back into the formula:

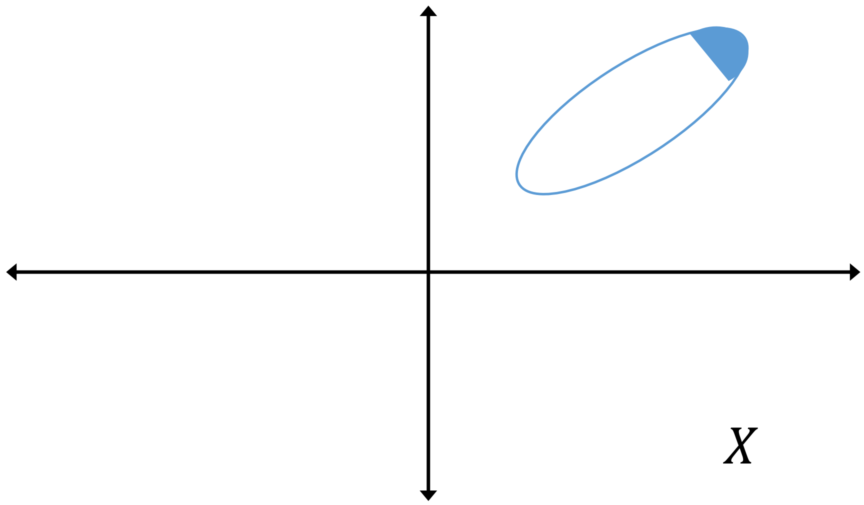
$$\begin{aligned} USV^T x &= \sum_{i=1}^{d} \alpha_i \lambda_i u_i \\ &= \sum_{i=1}^{d} \alpha_i A v_i \\ &= A \sum_{i=1}^{d} \alpha_i v_i \\ &= Ax \end{aligned}$$

2. Singular Value Decomposition - Whitening

We've seen before that demeaning our data makes it easier to work with. There's a more general operation called "whitening" where we also normalize the important directions of our data. In this problem, we'll do the operations corresponding to one version of whitening as a way to get better intuition on how SVD works.

Let $X \in \mathbb{R}^{n\times d}$ be a matrix of data points, and J be $\mathbf{I} - \mathbf{1}\mathbf{1}^T/n$. Let JX have a singular value decomposition of $JX = USV^T$.

Suppose we know that our points were drawn from a Gaussian distribution with covariance Σ. We would expect most of our points to lie in an ellipse, whose axes are the eigenvectors of Σ, scaled by the corresponding eigenvectors. We've drawn that ellipse below, with one area shaded so we can see its orientation.



For each of the following matrices:

- Verify that the resulting matrix is still $n \times d$, and therefore can be interpreted as modifying the datapoints of X
- Draw what the resulting data set would look like (i.e. how would the ellipse representing the covariance move?).

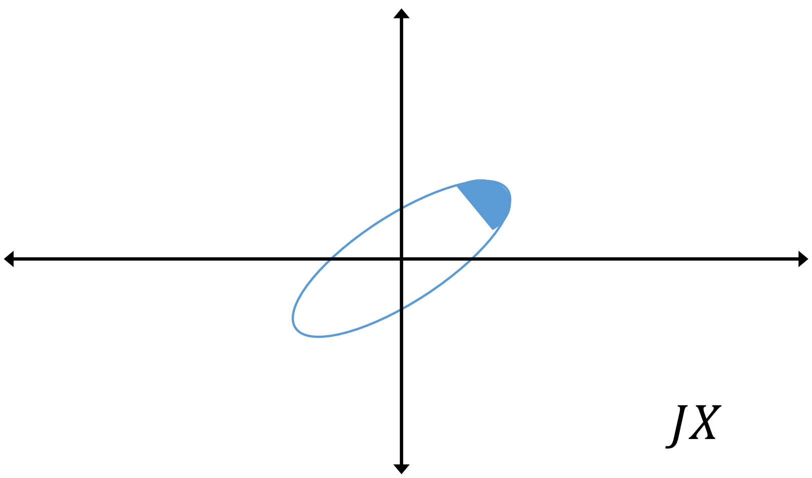
(a) JX **Solution:**

Since J is $n \times n$, the resulting matrix is, indeed $n \times d$.

Notice that $\mathbf{1}\mathbf{1}^T/n$ is an $n \times n$ matrix where every entry is $1/n$.

$$\begin{aligned} JX &= (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)X \\ &= X - \mathbf{1}\mathbf{1}^T X/n \\ &= X - \mathbf{1}\left(\sum_{i=1}^{n} x_i^T/n\right) \\ &= X - \mathbf{1}\overline{x} \end{aligned}$$

where $\overline{x}$ is the average of the rows of X. Thus JX is just X "demeaned"



(b) JXV **Solution:**

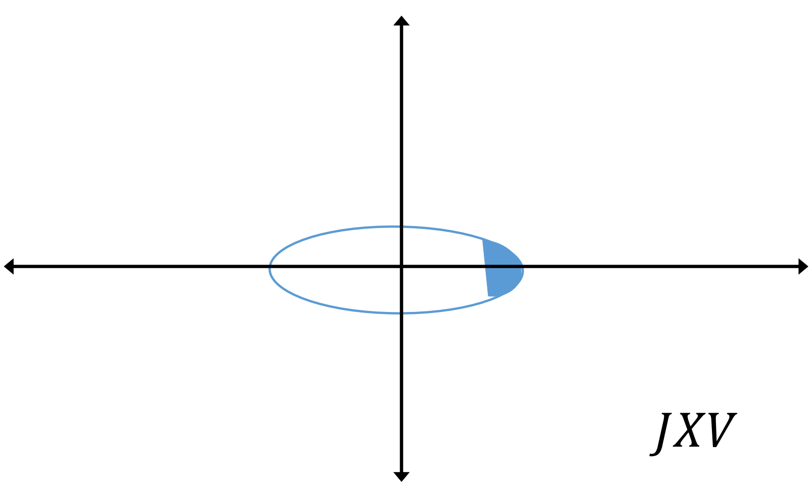
Since V is $d \times d$, JXV is still $n \times d$.

Below we will give a few different views of the calculation (they all say the same thing, but alternative wordings may give a different perspective)

We want to think about row i of JXV. Define y_i to be the i^{th} row of JX written as a column vector. When we multiply JXV, we will find $y_i^T v_j$ and put that number in entry i, j of the matrix. Since the columns of V form an orthnormal basis of $\mathbb{R}^d$, the i^{th} row of JXV is just y rewritten in the basis of V.

Said differently, if row i of JXV is the vector z then $y_i = \sum z[j]v_j$.

But then what does JXV look like? Well the j^{th} entry of row i is its dot product with v_j. I.e. in each direction of the standard basis we are going to go as far as we went in the principal component directions in y. So we have rotated the ellipse to now be on the standard basis.



(c) $JXVS^{-1}$, where S^{-1} is the $d \times d$ diagonal matrix, where $S^{-1}_{i,i} = 1/S_{i,i}$ (note that since S is not square, calling a matrix S^{-1} is an abuse of notation) **Solution:**

Since S^{-1} is $d \times d$, the matrix remains $n \times d$.

Note that S^{-1} just renormalizes each column j by $1/\sigma_j$. Thus row i of $JXVS^{-1}$ is the i^{th} demeaned data point, written in the singular vector basis, now normalized, so the most extreme data points have length at most 1 in the new basis.

How do we know the lengths become at most 1? The easiest way is to look at a single row, let e_i be the vector with a 1 in entry i and all 0's everywhere else. Note that $e_i^T A$ is the i^{th} row of the matrix A. To understand the length of the i^{th} row we want:

$$\begin{aligned}
||(e_i^T JXVS^{-1})||_2^2 &= (e_i^T JXVS^{-1})^T (e_i^T JXVS^{-1}) \\
&= (e_i^T USV^T VS^{-1})^T (e_i^T USV^T VS^{-1}) \\
&= (e_i^T USS^{-1})^T (e_i^T USS^{-1}) \\
&= (e_i^T UI')^T (e_i^T UI')
\end{aligned}$$

Where I' is the $n \times d$ matrix, which has 1s in every entry on the diagonal and 0's everywhere else. (Recall that S^{-1} isn't really an inverse – S isn't square so it can't have a real inverse)

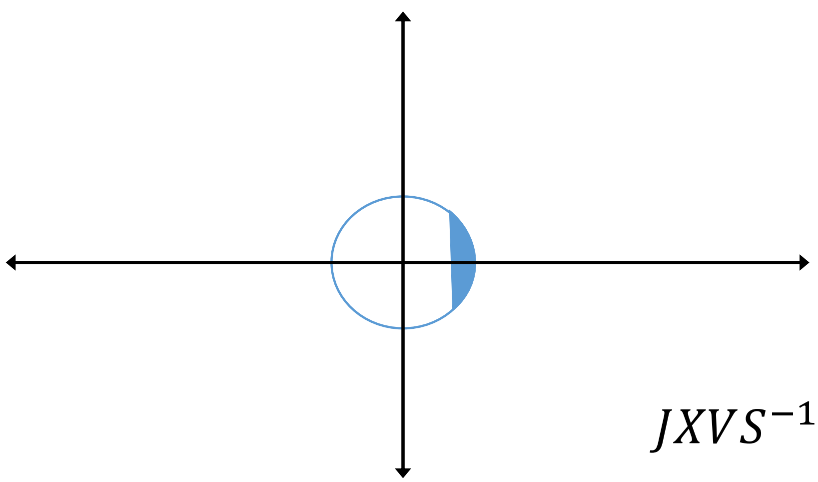
Let U' be the $n \times d$ matrix formed by deleting columns $d+1, d+2, \ldots, n$ of U.

$$\begin{aligned}
(e_i^T UI')^T (e_i^T UI') &= (e_i^T U')^T (e_i^T U') \\
&= U_i'^T U_i' \leq 1
\end{aligned}$$

Where the last inequality follows from the fact that (the full row) U_i is an orthonormal vector. Since we've just deleted entries from it, the length of the vector only decreased, and so the length is still at most 1.

Vectors of length at most 1 lie inside a circle, so we've "squashed" our vectors. Notice that since we're shrinking each direction according to its singular value, we are shrinking each vector by a different amount, such that we end up with a circle.

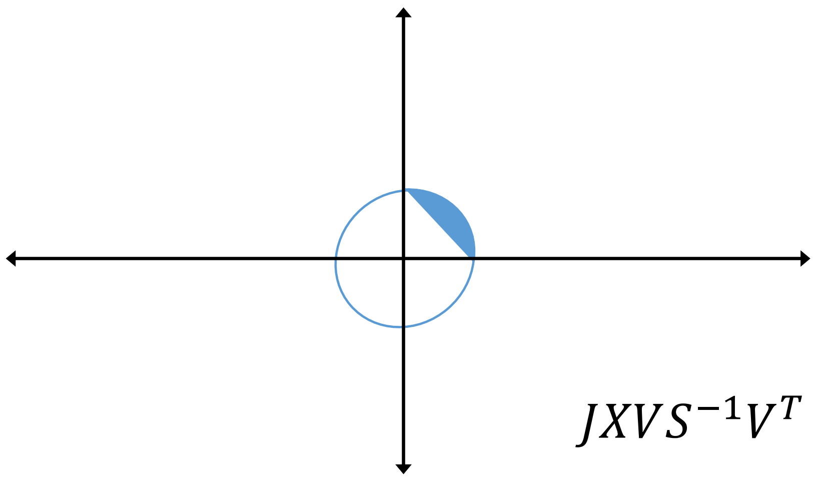
What's the radius of our circle? It turns out it's about $\frac{1}{\sqrt{n}}$ – try drawing some real data for various n and see what happens!



(d) $JXVS^{-1}V^T$ **Solution:**

Row i of $JXVS^{-1}V^T$: is harder to understand than the previous ones, let's do a calculation. Recall that entry i, j of $JXVS^{-1}$ is $\frac{1}{\sigma_j} y_i^T v_j$, where y_i is the demeaned version of x_i.

Then entry i, j of $JXVS^{-1}V^T$ is: $\sum_{k=1}^{d} \frac{1}{\sigma_k} y_i^T v_k v_k[j]$
so we can write row i as: $\sum_{k=1}^{d} \frac{1}{\sigma_k} y_i^T v_k v_k^T = \sum_{k=1}^{d} \frac{1}{\sigma_k} y_i I$ So we have just y_i but normalized by the σs. Thus we have "rotated" the points back to the original space, but kept them the same length as before.



3. Principal Component

Consider the following dataset, which is represented as three points in $\mathbb{R}^2$. Note that in this problem we will **not** demean the dataset.

$$\begin{bmatrix} 1 & 2 \\ 1.5 & 3 \\ 6 & 12 \end{bmatrix}$$

(a) What is the first principal component vector, v_1?

Solution:

Each point has second coordinate twice the first, so every point is on the line $y = 2x$, or equivalently is a multiple of the vector $[1, 2]$.

That direction, normalized, is the first principal component, so $v_1 = [1/\sqrt{5}, 2/\sqrt{5}]$.

(b) What is the second principal component, v_2?

Solution:

Since every data is in the span of the first principal component, any unit norm vector perpendicular to v_1 is an acceptable choice. One such vector is $[-2\sqrt{5}, 1/\sqrt{5}]$.

(c) If we use only the first principal component to compress the dataset, what will the representation of each point be?

Solution:

The first point is $\sqrt{5}v_1$, the second one is $1.5 \cdot \sqrt{5}v_1$, and the third one is $6 \cdot \sqrt{5}v_1$.

(d) Will this representation be lossy, or perfectly preserve the dataset?

Solution:

In this particular dataset, we perfectly preserve this dataset (the points are all multiples of v_1).

Answer the same questions for the following, slightly larger dataset:

$$\begin{bmatrix} 1 & 1 \\ 1.5 & 1.5 \\ -2 & 2 \\ 4 & -4 \\ 6 & -6 \\ 2 & 2 \end{bmatrix}$$

(a) What is the first principal component vector, v_1?

Solution:

Notice that every point is either a multiple of $[1, 1]$ or $[1, -1]$, so some of those must be our principal component. The norms of the multiples of $[1, -1]$ are much larger, so $[1/\sqrt{2}, -1/\sqrt{2}]$ is v_1.

(b) What is the second principal component, v_2?

Solution:

We need a vector perpendicular to v_1, which can best describe our remaining data. Since we're in two dimensions, we don't have choices after we chose the first principal component. Therefore, the second principal component is $[1/\sqrt{2}, 1/\sqrt{2}]$.

(c) If we use only the first principal component to compress the dataset, what will the representation of each point be?

Solution:

Data points 1, 2, and 6 are all perpendicular to v_1, so are represented as $[0, 0]$ (i.e. $0 \cdot v_1$). The other points are multiples of v_1, which are $-2\sqrt{2}v_1$, $4\sqrt{2}v_1$, and $6\sqrt{2}v_1$, respectively.

(d) Will this representation be lossy, or perfectly preserve the data?

Solution:

The data representation is lossy. Points 1, 2, and 6 have lost information.

4. Using the Eigenbasis

It's a very useful fact that for any symmetric n×n matrix A you can find a set of eigenvectors u_1, ..., u_n for A such that:

- $\|u_i\|_2 = 1$
- $u_i^T u_j = 0, \forall i \neq j$
- $u_1, \ldots, u_n$ form a basis of $\mathbb{R}^n$

One of the reasons this fact is useful is that facts about these matrices are easier to prove if you think about the vectors in terms of their "eigenbasis" components, instead of their components in the standard basis. As a trivial example, we'll show that you can calculate Ax for a vector x without having to do the matrix multiplication.

(a) Consider the matrix $A = \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix}$. Verify that $u_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ and $u_2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ are eigenvectors and meet the definitions. Find the eigenvalues associated with u_1 and u_2

Solution:

They are eigenvectors: $Au_1 = \begin{bmatrix} 4/\sqrt{2} - 1/\sqrt{2} \\ -1/\sqrt{2} + 4/\sqrt{2} \end{bmatrix} = 3u_1 Au_2 = 5u_2$ by a similar calculation.
They are unit norm: $u_1^T u_1 = (1/\sqrt{2})^2 + (1/\sqrt{2})^2 = 1/2 + 1/2 = 1$. The calculation for u_2 is similar.
They are orthogonal: $u_1^T u_2 = (1/\sqrt{2})(-1/\sqrt{2}) + (1/\sqrt{2})(1/\sqrt{2}) = -1/2 + 1/2 = 0$
They form a basis (since they're 2 linearly independent vectors in $\mathbb{R}^2$)

(b) since $\{u_1, u_2\}$ are a basis, we can write any vector as a linear combination of them. Write $x = \begin{bmatrix} -1/\sqrt{2} \\ 3/\sqrt{2} \end{bmatrix}$ in this basis.

Solution:

$x^T u_1 = -1/2 + 3/2 = 1 . x^T u_2 = 1/2 + 3/2 = 2$. So $x = u_1 + 2u_2$

(c) Based on the eigenvectors and eigenvalues your found in part a, diagonalize A, i.e, find matrix U and D such that $UDU^T = A$ where all entries of D are 0 except for the ones on the diagonal.

Solution:

$U = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} D = \begin{bmatrix} 3 & 0 \\ 0 & 5 \end{bmatrix}$

(d) Use the decomposition and the eigenvalues you calculated to calculate Ax wihtout doing matrix-vector multiplication.

Solution:

$Ax = A\,(u_1 + 2u_2) = Au_1 + 2Au_2 = 3u_1 + 10u_2 = \begin{bmatrix} -7/\sqrt{2} \\ 13/\sqrt{2} \end{bmatrix}$

This method of calculating a matrix vector product won't actually be more computationally efficient – but it's what's "really" happening when you do the multiplication, so this will be useful intuition under certain circumstances. Expressing vectors in an eigenbasis is also a useful proof technique, as we'll see in some later problems.

5. Sets of Eigenvectors

(a) Prove that if A is a symmetric matrix with n distinct eigenvalues, then its eigenvectors are orthogonal. Hint: if u and v are eigenvectors, calculate u^T Av two different ways.

Solution:

Here's one possible proof. Let u, v be eigenvectors with eigenvalues λ_u and λ_v respectively (with $\lambda_u \neq \lambda_v$) Consider the quantity $u^T Av$ On one hand,

$$u^T Av = u^T (\lambda_v v) = \lambda_v \cdot u^T v$$

On the other hand, since A is symmetric:

$$u^T Av = u^T A^T v = (Au)^T v = \lambda_u \left(u^T v\right)$$

Combinbing we have $\lambda_v u^T v = \lambda_u u^T v$, since $\lambda_u \neq \lambda_v$, we must have $u^T v = 0$ for all eigenvectors u, v

(b) Suppose that A is a symmetric matrix. Prove, without appealing to calculus, that the solution to $\arg\max_x x^T Ax$ s.t. $\|x\|_2 = 1$ is the eigenvector x_1 corresponding to the largest eigenvalue λ_1 of A. (Hint: the eigenvectors of a symmetric matrix can be chosen to be an orthonormal basis, i.e. unit vectors spanning all of $\mathbb{R}^n$.)

Solution:

Let $u_1, \ldots, u_n$ be an orthonormal set of unit vectors (which are guaranteed to exist by symmetry of A. Let x be a unit vector. We can write x as $\sum_{i=1}^n \alpha_i u_i$. We claim that $\sum \alpha_i^2 = 1$. Indeed:

$$\|x\|^2 = \left\|\sum \alpha_i u_i\right\|^2 \stackrel{*}{=} \sum \|\alpha_i u_i\|^2 = \sum \alpha_i^2 \|u_i\|^2 = \sum \alpha_i^2$$

Where the starred equality is a result of observing that any cross-terms are 0 by orthogonality of u_i (see a more detailed explanation in Section 4 of the solution).
Now let's examine $x^T Ax$

$$\begin{aligned} x^T Ax &= x^T A \left(\sum \alpha_i u_i\right) \\ &= x^T \left(\sum \alpha_i \lambda_i u_i\right) \\ &= \left(\sum \alpha_i u_i^T\right)\left(\sum \alpha_i \lambda_i u_i\right) \\ &\stackrel{*}{=} \sum \alpha_i^2 \lambda_i \|u_i\|^2 \\ &= \sum \alpha_i^2 \lambda_i \end{aligned}$$

Where again the starred equality uses that cross terms are 0 by orthoganality. since $\sum \alpha_i^2 = 1$, we are just taking a convex combination of the λ_i. This is clearly maximized by making $\alpha_1 = 1$ where λ_1 is the maximum eigenvalue. Observe that this is indeed possible by setting $x = u_1$ as claimed.

(c) Let A and B be two $\mathbb{R}^{n \times n}$ symmetric matrices. Suppose A and B have the exact same set of eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n$ with the corresponding eigenvalues $\alpha_1, \alpha_2, \cdots, \alpha_n$ for $\boldsymbol{A}$, and $\beta_1, \beta_2, \cdots, \beta_n$ for $\boldsymbol{B}$. Please write down the eigenvectors and their corresponding eigenvalues for the following matrices:

(i) $D = A - B$

Solution:

Eigenvectors u_i with eigenvalues $\alpha_i - \beta_i$ since $(A-B)x = Ax - Bx = (\alpha_i - \beta_i)\,x$

(ii) $E = AB$

Solution:

Eigenvectors u_i with eigenvalues $\alpha_i\beta_i$ since $ABu_i = A\beta_i u_i = \beta_i A u_i = \beta_i \alpha_i u_i$

(iii) $\boldsymbol{F} = \boldsymbol{A}^{-1}\boldsymbol{B}$ (assume $\boldsymbol{A}$ is invertible)

Solution:

Observe that $A^{-1}u_i = \frac{1}{a_i}u_i$
To show this we examine $A^{-1}Au_i$ On the one hand: $A^{-1}Au_i = A^{-1}\alpha_i u_i = \alpha_i A^{-1}u_i$
On the other hand: $A^{-1}Au_i = Iu_i = u_i$ Setting both equal to each other, we have $\alpha_i A^{-1}u_i = u_i$, so $A^{-1}u_i = \frac{1}{\alpha_i}u_i$
Then we have $A^{-1}Bu_i = A^{-1}\beta_i u_i = \beta_i A^{-1}$
$u_i = \frac{\beta_i}{\alpha_i}u_i$
Thus $A^{-1}B$ has eigenvectors u_i with eigenvalues $\frac{\partial_i}{a_i}$