Chapter 13

Day 7: EVD and PCA

13.1 Schedule

- 0900-0930: Debrief
- 0930-1015: Eigenvalue Decomposition (EVD)
- 1015-1030: Coffee
- 1030-1115: PCA and Maximum Variance
- 1115-1210: Conceptual PCA and PCA blog post
- 1210-1225: Review and Preview
- 1225-1230: Survey

13.2 Debrief [15 mins]

In the last class and in the take-home exercise, you worked on a number of different exercises involving eigenvalues and eigenvectors.

Exercise 13.1

1. With your table, identify a list of key concepts/take home messages/things you learned in the last class and take-home assignment.
2. Try to resolve your confusions with the folks at your table and by talking to an instructor.

13.3 Eigenvalue Decomposition (EVD) [45 mins]

The eigenvalue decomposition, also known as the eigendecomposition, is an operation on matrices in which a square matrix is expressed as a product of matrices made up of its eigenvalues and eigenvectors. It can be used to find inverses and powers of matrices, as well as to derive some important results in data analysis. For instance, in a prior exercise, you saw that the eigenvector corresponding to the largest eigenvalue of a covariance matrix was in the direction of greatest variance in your data set. This property can be proved using the eigendecomposition.

The eigenvalue decomposition is also helpful in dimensionality reduction, which is a process where we can represent higher-dimensional vectors as a linear combination of a smaller number of vectors than dimensions – an example of which you saw in a previous exercise where you represented pictures of peoples faces using a linear combination of vectors. The eigendecomposition is also often used to change coordinate systems.

The Big Idea

Assume that a square $n \times n$ matrix $\mathbf{A}$ has n linearly independent eigenvectors $\mathbf{v}_i$ with corresponding eigenvalues λ_i, i.e.

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i \; i = 1, 2, \ldots, n$$

Instead of thinking of these eigenvalues and eigenvector separately, let's package them into matrices as follows:

$$[\mathbf{A}\mathbf{v}_1 \; \mathbf{A}\mathbf{v}_2 \; \ldots \; \mathbf{A}\mathbf{v}_n] = [\lambda_1\mathbf{v}_1 \; \lambda_2\mathbf{v}_2 \; \ldots \; \lambda_n\mathbf{v}_n]$$

Properties of matrix multiplication suggests that we can re-write this matrix equation in the form

$$\mathbf{A}[\mathbf{v}_1 \; \mathbf{v}_2 \; \ldots \; \mathbf{v}_n] = [\mathbf{v}_1 \; \mathbf{v}_2 \; \ldots \; \mathbf{v}_n] \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix}$$

where the last matrix has each eigenvalue on the diagonal. If we now define

$$\begin{aligned} \mathbf{V} &= [\mathbf{v}_1 \; \mathbf{v}_2 \; \ldots \; \mathbf{v}_n] \\ \mathbf{D} &= \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix} \end{aligned}$$

then the previous equation becomes

$$\mathbf{AV} = \mathbf{VD}$$

Since we assumed that the eigenvectors are linearly independent this implies that the columns of $\mathbf{V}$ are linearly independent which in turn implies that the inverse of $\mathbf{V}$ exists. We can there fore write

$$\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1} \tag{13.1}$$

where the matrix $\mathbf{V}$ has the i-th eigenvector of $\mathbf{A}$ as its i-th column, and $\mathbf{D}$ is a diagonal matrix with the i-th eigenvalue of $\mathbf{A}$ as its ii-th entry. This expression is known as the *eigendecomposition* of $\mathbf{A}$. In the special case where $\mathbf{A}$ is symmetric, the eigenvalues are real, and the eigenvectors are mutually orthogonal so that

$$\mathbf{V}^{-1} = \mathbf{V}^T,$$

which is a property of $n \times n$ matrices whose column vectors are mutually orthogonal and have a length of 1 (i.e., the column vectors are orthonormal).

Exercise 13.2

1. Consider the following 2×2 matrix $\mathbf{A}$.

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

   By hand, compute its eigenvectors and eigenvalues, determine the matrices $\mathbf{V}$, $\mathbf{D}$, and $\mathbf{V}^{-1}$,

and confirm that (13.1) is correct. Use MATLAB to confirm your results by computing » [V,D]=eig(A). **Note**: you should normalize each of your eigenvectors to be unit length.

Exercise 13.3

1. The eigendecomposition can be used to change basis as follows. Consider the matrix $\mathbf{A}$ from the previous exercise as a transformation matrix.
   a) How does the matrix $\mathbf{A}$ transform the vector $\mathbf{w} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$? Draw both $\mathbf{w}$ and $\mathbf{Aw}$ on an xy-coordinate plane.
   b) Draw both eigenvectors of $\mathbf{A}$ on this coordinate plane.
   c) Decompose the vector $\mathbf{w}$ as a linear combination of both eigenvectors. You should be able to do this with a matrix-vector multiply. You are expressing the vector in a new basis.
   d) Scale each component by the relevant eigenvalue.
   e) Undo the decomposition to return to the original basis.
   f) What just happened?

Exercise 13.4

One thing that the eigendecomposition helps us compute is how to raise $\mathbf{A}$ to an integer power, without going through the process of repeated multiplication.

1. Using eigendecomposition, show the following is true
$$\mathbf{A}^2 = \mathbf{V}\mathbf{D}^2\mathbf{V}^{-1} \tag{13.2}$$
and confirm this result using the matrix from earlier the earlier exercise. Note that for any diagonal matrix $\mathbf{D}$, $\mathbf{D}^k$ is another diagonal matrix whose ii-th entry equals the ii-th entry of $\mathbf{D}$ raised to the k-th power. Hence computing $\mathbf{D}^n$ is not computationally difficult - you just raise each diagonal entry to the n-th power.
2. Show that the following is also true
$$\mathbf{A}^n = \mathbf{V}\mathbf{D}^n\mathbf{V}^{-1}$$

13.4 Principal Components Analysis (PCA)

In the night assignment you explored, in a graphical manner, the relationship between the eigenvectors of the covariance matrix and the distribution of the data. For instance, you looked at the daily temperature values in Boston versus Sao Paolo and the daily temperatures in Boston versus Washington D.C.

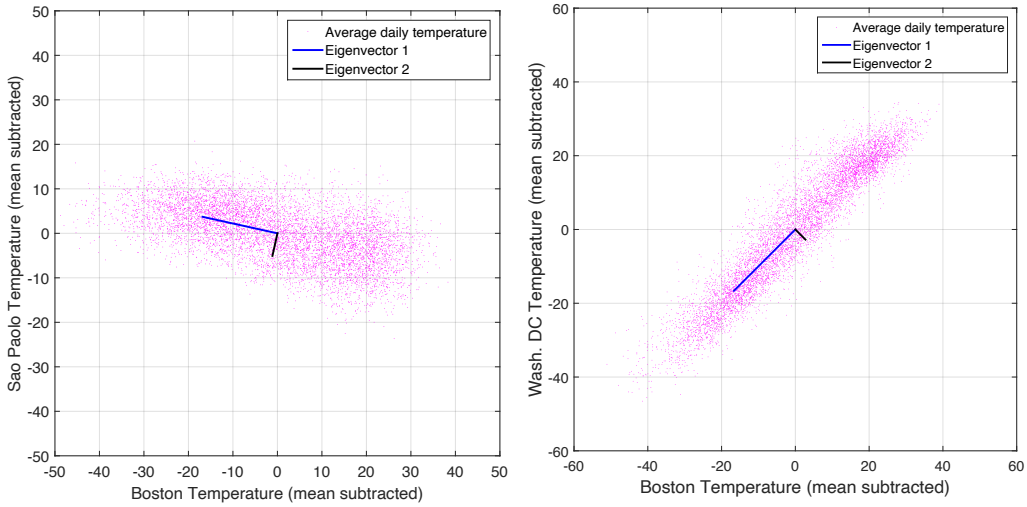

Figure 13.1: Centered average daily temperatures of Boston vs Sao Paolo (left) and Boston vs Washington DC, with the eigenvectors of the covariance matrix.

From visually inspecting these figures we saw that eigenvector 1, which corresponded to the larger of the two eigenvalues, seemed to be pointing in the direction where the data exhibited the most variability (i.e., the data was most spread out along this direction). You also looked at this for a 3D dataset consisting of the temperatures from Boston, Sao Paolo, and Washington DC.

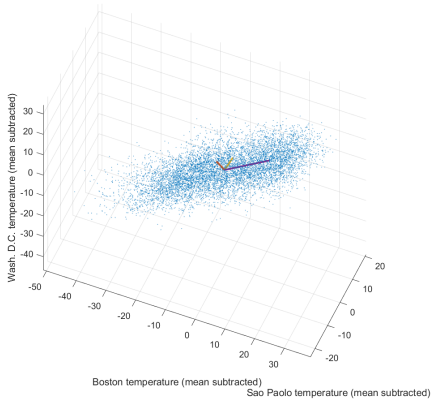

Figure 13.2: Temperatures and eigenvectors for Boston, Sao Paolo, and Washington DC

In this 3D dataset, we see the same phenomenon: that the principal eigenvector points along the direction of maximum variation in the data. It turns out that this phenomenon will hold no matter the dimensionality of the data (it works for 4D datasets, 10D datasets, and even datasets with 1,000s of dimensions)! This fact provides the basis for the principal components algorithm. In PCA, instead of working with the data in its original form, we express it in a basis given by eigenvectors of the covariance matrix that have the largest eigenvalues. We can understand the properties of using this basis through two key properties.

- *Property 1*: the principal eigenvectors of the covariance matrix will maximize the variance of the data when the data is projected onto these vectors (we can think of vectors that capture large variation in the data as representing important properties of the data).
- *Property 2*: the principal eigenvectors of the covariance matrix will allow us, in a particular sense, to optimally compress our data. That is, we will be able to recover the original data with the highest possible accuracy from the projections of the data onto the principal eigenvectors.

The power of PCA lies in its ability to achieve both of the properties described above simultaneously. For this reason, the principal components of a dataset will act as keys to unlocking the secrets lurking in the data! **Today we will be exploring property 1, and in the night assignment you will also be exploring property 2**.

The Principal Eigenvector as the Direction of Maximum Variance

The graphs of the daily temperature data show, graphically, that the principal eigenvector of the covariance matrix corresponds to the direction of maximum variation in the data. In this section we'll be formalizing this result. We've decided to structure this part of the day assignment as an extended exercise where you will be working through the proof of this fact step-by-step. While there are many ways to do this proof, we'll be walking you through one way that will connect well with the ideas we've been exploring in the last week or so of the course. We recommend that you do a part of the proof, check it against the solutions and then move onto the next piece.

Before getting started, let's look at some material from night 6 that shows that the covariance matrix can be computed using matrix multiplication.

Suppose that we have two different data variables x and y (e.g. corresponding to temperatures in Boston and Sao Paolo), with x_i and y_i being different values in the data set we can define a a matrix $\mathbf{A}$ as follows:

$$\mathbf{A} = \frac{1}{\sqrt{N-1}} \begin{pmatrix} x_1 - \mu_x & y_1 - \mu_y \\ x_2 - \mu_x & y_2 - \mu_y \\ x_3 - \mu_x & y_3 - \mu_y \\ \vdots & \vdots \\ x_N - \mu_x & y_N - \mu_y \end{pmatrix} \tag{13.3}$$

where μ_x is the mean of the first column, and N is the number of samples (rows). The covariance matrix of x and y is $\mathbf{R} = \mathbf{A}^T\mathbf{A}$. You can think of the entries of this matrix as storing the un-normalized correlations between the temperatures. Because $\mathbf{R}^T = \mathbf{R}$, this matrix is symmetric, and hence has orthogonal eigenvectors.

Let's assume that we are given a dataset with n samples and d dimensions (instead of just 2 dimensions as shown above). We can transform it into the form given in Equation 41.12 by subtracting the mean from each column and dividing the entire matrix by $\sqrt{N-1}$. We now have a mean-centered data matrix $\mathbf{A}$ with n rows and d columns and the covariance matrix of our data is given by $\mathbf{A}^\top\mathbf{A}$.

Exercise 13.5

Our overall goal is to show that if we take a unit vector $\mathbf{u}$, project our mean-centered data onto it (as $\mathbf{Au}$), and examine the variance of the projected data, that this variance is largest when $\mathbf{u}$ is the principal eigenvector of the covariance matrix $\mathbf{A}^\top\mathbf{A}$.

1. First we'll write down an expression for the variance of $\mathbf{Au}$ (we'll write this as $Var[\mathbf{Au}]$) as a matrix multiplication. We'll do this step together (i.e., we'll show you how to do it). For this part of the exercise you should make sure you understand the steps we performed.

   If $\mathbf{A}$ is in the form given in Equation 41.12, then $\mathbf{Au}$ will have 0 mean (since $\mathbf{Au}$ is a linear combination of columns with 0 mean). Using the same logic that led us to conclude that $\mathbf{A}^\top\mathbf{A}$ is the covariance matrix of the data, $(\mathbf{Au})^\top(\mathbf{Au})$ will give us the variance of the data projected onto $\mathbf{u}$ (remember that variance is just a special case of covariance where we are comparing a quantity to itself). It's worth noting that since $\mathbf{Au}$ is a vector, the expression $(\mathbf{Au})^\top(\mathbf{Au})$ is known as the inner product, which is really the same as the dot product (that is, $(\mathbf{Au})^\top(\mathbf{Au}) = \mathbf{Au} \cdot \mathbf{Au}$). Thus, the variance is given by the following equation.

   $$\begin{aligned} Var[\mathbf{Au}] &= (\mathbf{Au})^\top(\mathbf{Au}) \\ &= \mathbf{u}^\top\mathbf{A}^\top\mathbf{Au} \quad \text{note: we are applying the rule that } (\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top \end{aligned}$$

2. Substitute the eigenvalue decomposition, $\mathbf{VDV}^\top$, for the covariance matrix $\mathbf{A}^\top\mathbf{A}$ (since $\mathbf{A}^\top\mathbf{A}$ is symmetric and real, we can substitute $\mathbf{V}^\top$ for the inverse of $\mathbf{V}$ in the eigenvalue decomposition).

3. Define the vector $\mathbf{y} = \mathbf{V}^\top\mathbf{u}$ and substitute it into the expression from part 2.

4. Expand out the expression in part 3 so that it is in terms of the squares of the elements of $\mathbf{y}$ and the diagonal entries of $\mathbf{D}$ in order of largest to smallest.
5. Show that $\mathbf{y}$ is a unit vector by taking the inner product with itself and showing that it is equal to 1 (recall that the inner product is the same as the dot product). *Hint:* $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$ *since* $\mathbf{V}$ *is orthonormal and has* d *linearly independent columns.*
6. Argue that since $\mathbf{y}$ is a unit vector (which implies $\sum_{i=1}^{d} y_i^2 = 1$), that the expression in part 4 is maximized when $y_i = 1$ when i is the index of the principal eigenvector and $y_i = 0$ when i is any other index. To get a feel for why this is true, try writing out a specific case where, perhaps, $\mathbf{y}$ has two or three dimensions.
7. Show that we achieve the value of $\mathbf{y}$ in part 5 (that is where $y_i = 1$ when i is the index of the principal eigenvector and $y_i = 0$ when i is any other index) when $\mathbf{u}$ is the principal eigenvector of $\mathbf{A}^\top\mathbf{A}$.
8. What have you just shown?!? Make sure you have a sense of what you just did (don't get lost in the mathematical symbols).

Beyond the first principal component

We've now gone into depth in understanding the first principal component and its amazing property of maximizing variance. The second principal component is simply going to be the direction that maximizes variance subject to the requirement that it is orthogonal to the first principal component. With a slight modification to your proof you can show that the second principal component will be in the direction of the eigenvector with the second largest eigenvalue. The trend continues for other principal components (i.e., the ith principal component is the eigenvector with the ith largest eigenvalue).

Applications of PCA (thinking it through conceptually)

In this section you're going to be thinking about what the PCA algorithm might do when applied in different domains. The focus of this section will be on trying to understand at a conceptual level what might happen when we apply PCA. In the next section, you'll be reading through an example of applying PCA to some actual data.

Exercise 13.6

For each application, hypothesize what the first principal component might be. That is, for each particular scenario what would the direction be that maximizes the variance of the data projected onto that direction? What might the second principal component be (that is a vector orthogonal to the first that maximizes the variance of the data)?

1. Consider a dataset consisting of ratings from n users of m movies. Let's assume that the ratings are numerical and are on a scale of 1 to 5 (5 being the best). Consider some collection of movies (they could be some specific movies or you could just think of movie genres) and a particular population of users (could be college students, QEA professors, or just the general population). Draw the data matrix $\mathbf{A}$ and label the rows and columns (e.g., with movies or users). In a qualitative sense, make a prediction as to what the first principal component would look like for this dataset. What might the second principal component look like? No numbers... just guess at which dimensions would be positive, negative, or close to 0 for your principal components.

2. Consider a dataset consisting of the prevalence of the flu in various parts of the US. The CDC maintains an animated map of the flu activity over time, which you can (and should) access at https://www.cdc.gov/flu/weekly/usmap.htm. To simplify this data, let's think about the number of flu cases in each of the six major major geographical regions of the US.

If we think about our data matrix as consisting of a row for each week of measured flu activity and each column as a region of the US, in a qualitative sense, make a prediction as to what the first principal component would look like for this dataset. What might the second principal component look like? No numbers... just guess at which dimensions would be positive, negative, or close to 0 for your principal components.

Exercise 13.7

With your table-mates, read through this post that shows the application of PCA to understanding the US political leanings (if you are viewing this in DropBox preview and can't click the link, go to http://bit.ly/37n9qwe). Before, starting here are some process suggestions.

- Checkin with folks at your table as to how they'd like to go through this document (e.g., read

the entire thing individually and come together and ask questions, read it individually but stop after each major section to ask questions, read it aloud as a table).

- If you don't understand something, you can either call over an instructor or note your confusion on the whiteboard and keep going (e.g., if its something that doesn't impede your understanding of the main points in the article).

Solution 13.2

1. The eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 3$ with corresponding eigenvectors $\mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix}$ and $\mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$. This gives

$$\mathbf{V} = \frac{1}{\sqrt{2}}\begin{bmatrix}-1 & 1\\1 & 1\end{bmatrix},\ \mathbf{D} = \begin{bmatrix}1 & 0\\0 & 3\end{bmatrix},\ \mathbf{V}^{-1} = \sqrt{2}\begin{bmatrix}-1/2 & 1/2\\1/2 & 1/2\end{bmatrix}.$$

   and if you multiply them all together you will get the original matrix $\mathbf{A}$. Running "eig" in MATLAB gives the same eigenvalues and eigenvectors, although every eigenvector could be multiplied by -1. MATLAB may also place your eigenvalues and eigenvectors in a different order.

Solution 13.3

1. The vector becomes $\begin{bmatrix}5\\4\end{bmatrix}$.
2. The eigenvectors were $\mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix}$ and $\mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$.
3. Decomposing the vector $\mathbf{w}$ as linear combination of the eigenvectors is equivalent to solving

$$\mathbf{V}\mathbf{c} = \mathbf{w}$$

   for the vector $\mathbf{c}$. This is the coordinates of the vector $\mathbf{w}$ in the new basis. You should find that $\mathbf{c} = \sqrt{2}\begin{bmatrix}-0.5\\1.5\end{bmatrix}$.
4. We multiply the first component by 1 and the second component by 3 to give $\sqrt{2}\begin{bmatrix}-0.5\\4.5\end{bmatrix}$.
5. In order to undo the change of basis we hit this vector with $\mathbf{V}$ which gives $\begin{bmatrix}5\\4\end{bmatrix}$ as expected.
6. The eigendecomposition can be be thought of as a change of basis followed by a scaling matrix followed by the change back to the original basis.

Solution 13.4

1. Since $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$, we know that

$$\mathbf{A}^2 = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}\mathbf{V}\mathbf{D}\mathbf{V}^{-1} = \mathbf{V}\mathbf{D}^2\mathbf{V}^{-1}.$$

2. Similar reasoning to the previous problem shows that

$$\mathbf{A}^n = \mathbf{V}\mathbf{D}^n\mathbf{V}^{-1}$$

Solution 13.5

1. Solution is already given in the problem

2.

$$Var[\mathbf{Au}] = \mathbf{u}^\top \mathbf{VDV}^\top \mathbf{u}$$

3.

$$\begin{aligned} Var[\mathbf{Au}] &= (\mathbf{V}^\top \mathbf{u})^\top \mathbf{D}(\mathbf{V}^\top \mathbf{u}) \\ &= \mathbf{y}^\top \mathbf{Dy} \end{aligned}$$

4.

$$\begin{aligned} Var[\mathbf{Au}] &= \mathbf{y}^\top \mathbf{Dy} \\ &= \mathbf{y}^\top \begin{bmatrix} y_1 D_{1,1} \\ y_2 D_{2,2} \\ \vdots \\ y_d D_{d,d} \end{bmatrix} \\ &= \sum_{i=1}^{d} y_i^2 D_{i,i} \end{aligned}$$

5.

$$\begin{aligned} \mathbf{y}^\top \mathbf{y} &= (\mathbf{V}^\top \mathbf{u})^\top (\mathbf{V}^\top \mathbf{u}) \\ &= \mathbf{u}^\top \mathbf{VV}^\top \mathbf{u} \\ &= \mathbf{u}^\top \mathbf{u} \\ &= 1 \end{aligned}$$

6. If we choose $y_i = 1$ where i is the index of the principal eigenvector, then the expression in part 4 will give us $D_{i,i}$. Any other choice of $\mathbf{y}$ will result in some weighted combination of the eigenvalues (the diagonal elements of $\mathbf{D}$) where the weights are all positive and add up to 1. It is easy to see that putting any weight on a non-maximal eigenvalue will result in a lower variance as computed by the expression in part 4.

7. Since $\mathbf{y} = \mathbf{V}^\top \mathbf{u}$, y_i is the dot product of $\mathbf{u}$ and the ith eigenvector, $\mathbf{v_i}$, with $\mathbf{u}$. Since we assume all of the eigenvectors are unit vectors and mutually orthogonal, if we set $\mathbf{u}$ to be the principal eigenvector of $\mathbf{A}^\top \mathbf{A}$, then the dot product of $\mathbf{u}$ and $\mathbf{v_i}$ will be 1 for i corresponding to the principal eigenvector and 0 for all other indices.

8. You just showed that the direction along the principal eigenvector of the covariance matrix maximizes the variance of the projected data. That's pretty cool!