Problem #1 Using spectral decomposition (SVD) to perform principal component analysis (PCA)

Pertinent reading for Problem #1:

R. Briandet, E.K. Kemsley, and R.H. Wilson, "Discrimination of *Arabica* and *Robusta* in Instant Coffee by Fourier Transform Infrared Spectroscopy and Chemometrics," *J. Agric. Food Chem*. 44 (1) 170 – 174 (1996)

As you progress into this problem, here's a reminder of what your giant data matrix looks like !

$$\underline{Variables}$$

$$X = \begin{bmatrix} \overrightarrow{X_1} & \overrightarrow{X_2} & \cdots & \overrightarrow{X_{10}} \\ 20 & 10 & \square & 1 \\ 3 & 43 & \square & 5 \\ 0 & 25 & \cdots & 97 \\ 4 & 90 & \square & 65 \\ \vdots & \vdots & \square & \vdots \end{bmatrix} = \left[\begin{array}{c|c|c|c} \square & \square & \square & \square \\ \overrightarrow{y_1} & \overrightarrow{y_2} & \cdots & \overrightarrow{y_{10}} \\ \square & \square & \square & \square \end{array}\right]$$

$p = 10$
Experimental factors
(10 coffee bean samples)

$n = 286$
# of observations
(wavenumber values)

Part #1: Combining multiple IR spectroscopy data into 1 giant data matrix X

In this problem, you are given 10 bags of coffee beans: 5 arabica beans (*C. arabica*)
5 robusta beans (*C. Canephora v. Robusta*)

and you want to know whether you can use infrared (IR) spectroscopy to differentiate between the 2 kinds of beans. To satisfy your curiosity, you used the FTIR machine in the basement of 44 Cummington and measured the IR spectrum of the 10 coffee grinds near the "fingerprint" (1/wavelength) region between 800 – 1900 cm^{-1}. Figure 1a is what your IR ***transmission*** results look like, where Figure 1b reports the same data plotted in IR ***absorbance***

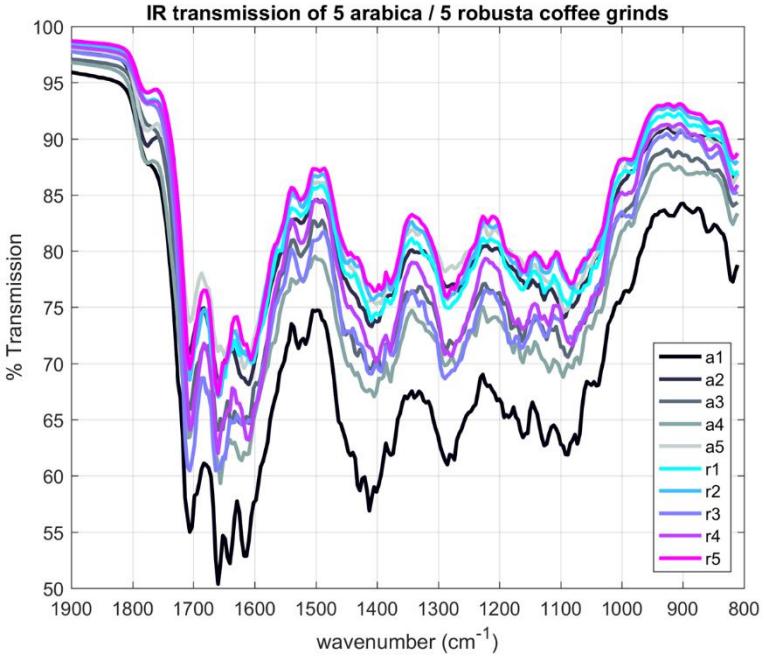


Figure 1a: The IR transmission spectra for:

5 Arabica grinds (a1 – a5)
5 Robusta grinds (r1 – r5)

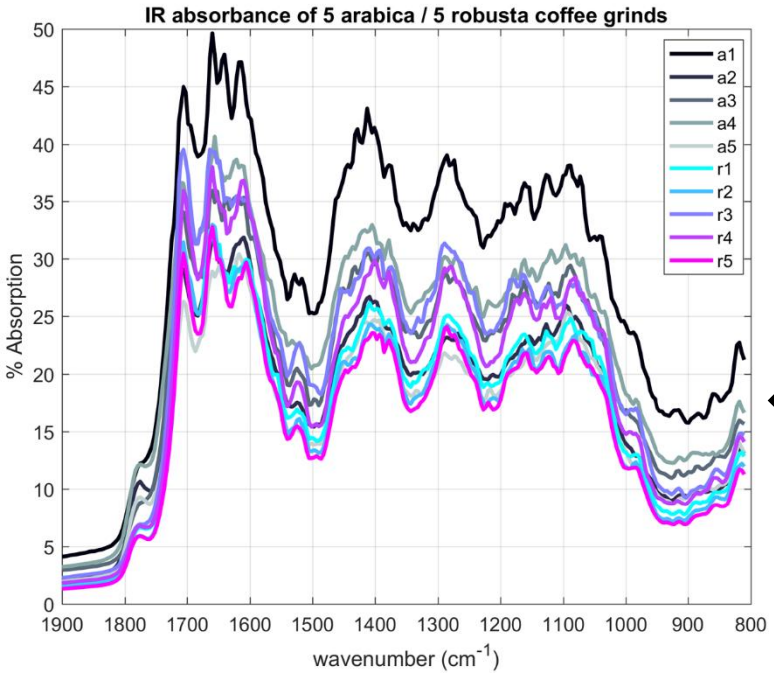


Figure 1b: The same data as in Figure 1a,but plotted in term of IR ***absorbance***

The file:

"Problem2_coffee_FTIR_data.txt"

will be the ***absorbance*** data set shown here !!

Your Tasks for this section: As usual, please turn in all **.m matlab scripts*, plots, or requested echoed values (in diary files) shaded in yellow.

1. Using matlab's *textread* function, load in 10 columns of IR data stored in "Problem2_coffee_FTIR_data.txt." Since reading in 10 columns worth of stuff will make the command really long, you can use the 3 - period ellipsis notation " . . ." to join several lines of code into 1 single line of command. For instance, I would spread out the *textread* command into 4 lines of code:

```
[wavenumber, arabica1, arabica2, arabica3, arabica4, arabica5, ...
        robusta1, robusta2, robusta3, robusta4, robusta5 ] ...
        = textread('Problem1_coffee_FTIR_data.txt', ...
                '%f%f%f%f%f%f %f%f%f%f%f', 'headerlines', 3 )
```

** Note: There are 3 lines of headers in this file, and also:

a) $The\ 1st\ data\ column \ = \ wavenumber \ \ (units \ = \ cm^{-1})$

b) $The\ next\ 5\ columns \ = \ IR\ absorption\ data\ from\ 5\ different\ sources\ of\ arabica\ beans$

c) $The\ last\ 5\ columns \ = \ IR\ absorption\ data\ from\ 5\ different\ sources\ of\ robusta\ beans$

2. Build a giant data matrix X by joining the 10 columns of IR data side-by-side. The matrix X should be in the form:

$$X = \left[\begin{array}{c|c|c|c|c|c|c|c} arabica1\ data & arabica2\ data & \cdots & arabica5\ data & robusta1\ data & robusta2\ data & \cdots & robusta5\ data \end{array}\right]$$

3. Using matlab's *svd* function, calculate the orthogonal matrices U, V, and the singular value matrix Σ. Please echo the following:

a) The smaller eigenvector matrix V.

b) The 10 non-zero singular values $\sigma_1\ ,\ \sigma_2\ ,\ \cdots\ \ ,\ \sigma_{10}$ that's inside your Σ matrix. Notice that matlab has automatically ranked them in descending order !

4. Using matlab's *semilogy* function, plot all the singular values σ_p on the y-axis versus a dummy index on the x-axis that represents the p[th] - singular value of your data matrix X. Also, add grid lines to your plot !

5. Recall the equation for SVD spectral decompositions for your data matrix X:

$$X = \underbrace{\sigma_1 (u_1 v_1{}^T)}_{1^{st}\text{ spectrum (matrix)}} + \underbrace{\sigma_2 (u_2 v_2{}^T)}_{2^{nd}\text{ spectrum (matrix)}} + \cdots + \underbrace{\sigma_{10} (u_{10} v_{10}{}^T)}_{10^{th}\text{ spectrum (matrix)}}$$

Using *pcolor* , plot the *values of the 4 following matrices (X and the 3 SVD spectra)* and save them as *.jpg files. Check out Figure 2 on the next page to see an example of what these 4 plots should look like !

SVD spectrum	**Matrix**	$[\boldsymbol{cmin} \quad \boldsymbol{cmax}]$ **for the caxis command (% absorbance colors)**
Original data	X	$[0 \quad 50]$
#1 (Dominant)	$\sigma_1 (u_1 v_1{}^T)$	$[0 \quad 50]$
#2 (Secondary)	$\sigma_2 (u_2 v_2{}^T)$	$[0 \quad 10]$ (increases the data contrast)
#3 (Tertiary)	$\sigma_3 (u_3 v_3{}^T)$	$[0 \quad 10]$

For each of these plots :

a) Using the *set* command, set the 'EdgeColor' to 'none' to remove the black borders for each pixels.

b) Please color-label the plot using *colorbar*, and also, limit the color values using *caxis([cmin cmax])* command. Then, add an "% absorbance" label to your colorbar (check out the example matlab script to see how to do this).

c) Reverse the y-axis values by invoking the command:

set(gca, 'YDir', 'reverse')

d) The title of the original data matrix X should say, "Original data X." The title for each SVD spectrum plot should say, "SVD submatrix p with σ_p = $xxxx.xxxx$" , where:

i) $p \quad = \quad 1,\ 2,\ or\ 3$

ii) $\sigma_p \quad = \quad Greek\ letter\ with\ a\ dyanmically - changing\ 'p'\ value$

iii) $xxxx.xxxx \ = \ The\ actual\ singular\ value\ \sigma_p\ \ in\ you\ S - matrix$

** *Hint*: You should use the *strcat* and *num2str* functions to do this !!
Again, check out the example matlab script to see how to do this efficiently !

e) Next, the x-axis should have labels "a1 a2 a3 a4 a5 r1 r2 r3 r4 r5 " aligned with each pixel columns. This corresponds to the 10 different coffee bean types. Consult the example matlab script on how to do this.

f) Moreover, the y-axis should be labeled as "wavenumber (cm^{-1})." This corresponds to the wavenumber used in our IR scans.

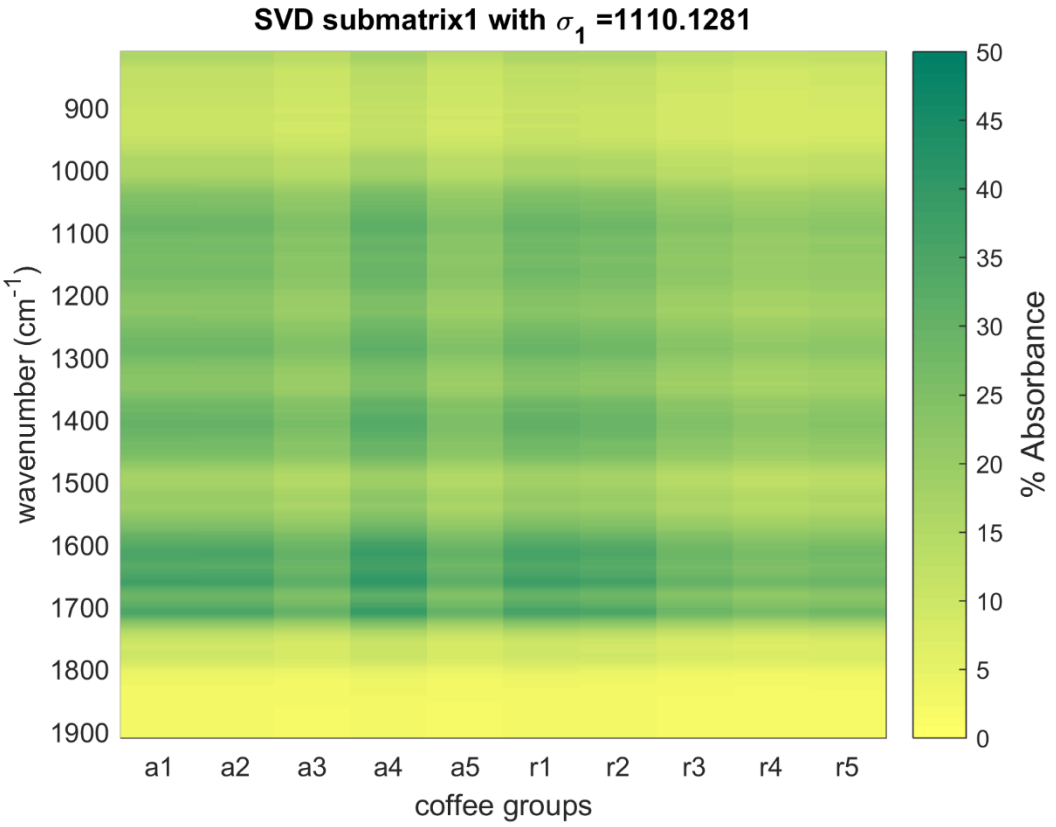


Figure 2: An example of what your SVD submatrix pcolor plots should look like !

Task #2: Analyze the SVD component that can "separate" arabica vs. robusta beans

Yay ! You've actually done the hardest part of this problem ! =)

Now, stare at each of the 3 SVD submatrix plots for a minute. You'll notice that 1 of those 3 plots will exhibit the following characteristics:

a) The vertical SVD patterns of the arabica group (a1 – a5) will look similar to each other
b) The vertical SVD patterns of the robusta group (r1 – r5) will also look similar to each other
c) Group (a1-a5) looks completely different than Group (r1 – r5).

An example of what you may see is depicted in Figure 3.

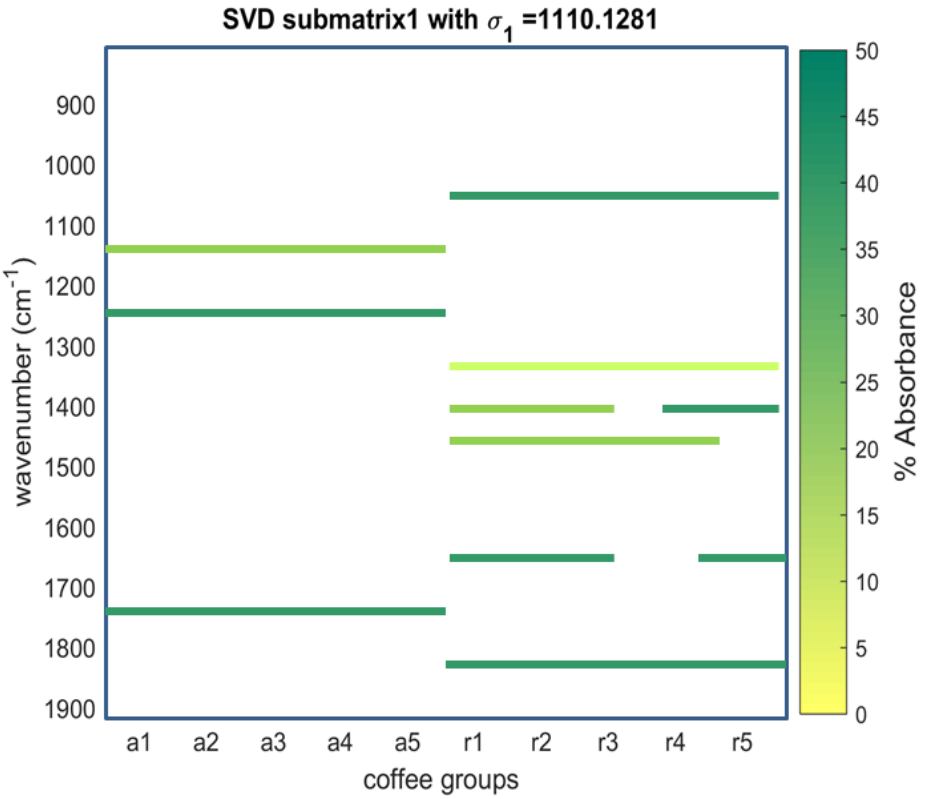


Figure 3: An example in which the vertical SVD patterns of the arabica beans (left, a1-a5) can be differentiated from the robusta beans (right, r1-r5).

Yes !!! This is the SVD component that we want ! Now, we are ready to extract the robusta-specific IR signature from the SVD submatrix in Figure 3. As depicted in Figure 4, we will take a vertical slice of data from either coffee beans r1 or r2 that best represents the robusta group's IR signature. Then, we can plot the resulting IR absorbance versus the wavenumber.

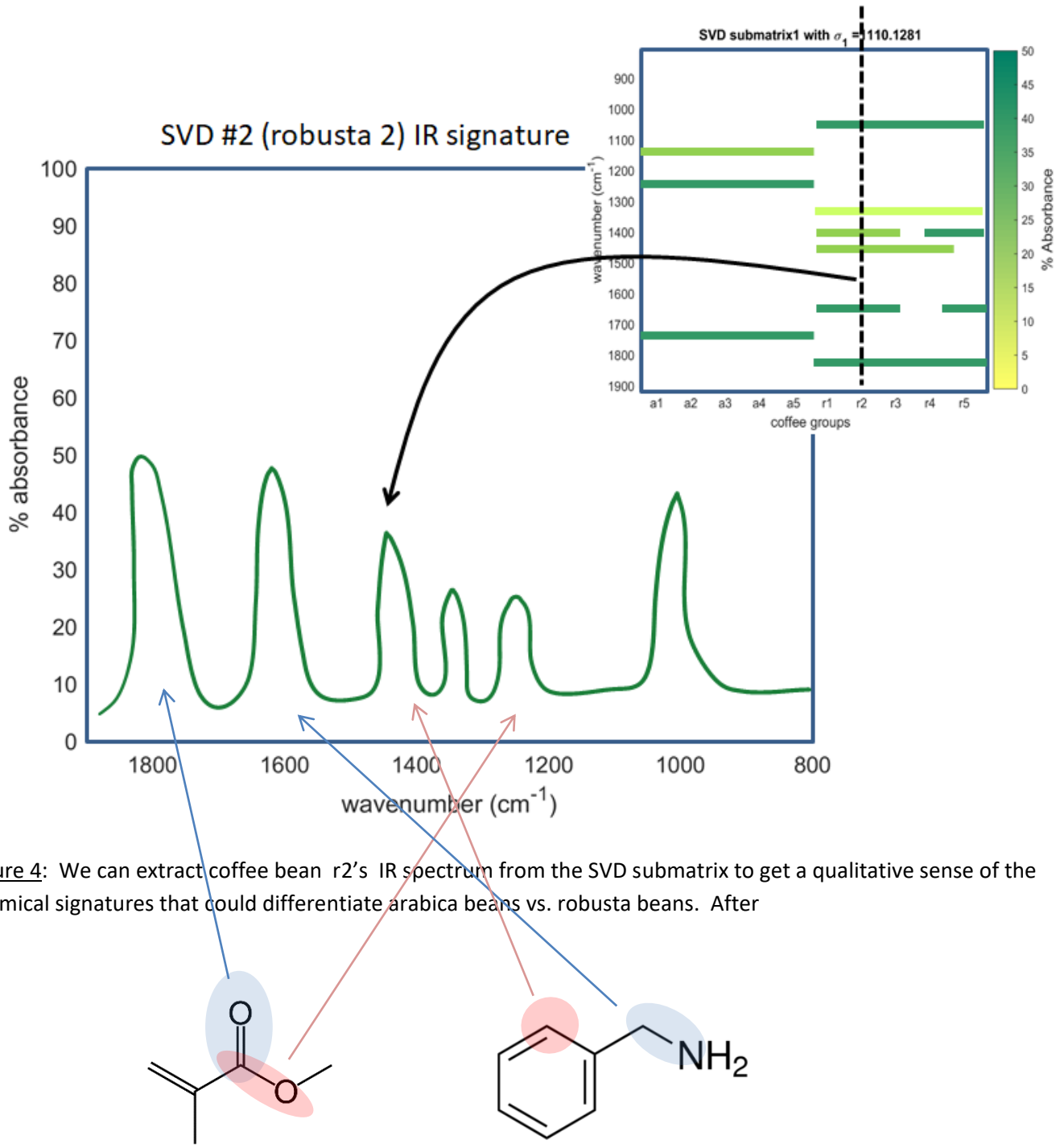


Figure 4: We can extract coffee bean r2's IR spectrum from the SVD submatrix to get a qualitative sense of the chemical signatures that could differentiate arabica beans vs. robusta beans. After

If you were a coffee biochemist, your job is now is to stare at the IR signature in Figure 4 and try to identify:

1) Any possible molecules that could be in your robusta beans
2) Any "dominant" IR signatures from various functional groups within these molecules

Your Tasks for this section: As usual, please turn in all **.m matlab scripts*, plots, or requested echoed values (in diary files) shaded in yellow.

1. Stare at your SVD submatrix plots $\sigma_1 (u_1 {v_1}^T)$, $\sigma_2 (u_2 {v_2}^T)$, and $\sigma_3 (u_3 {v_3}^T)$ for a little bit. Identify the one that exhibits patterns that could best differentiate between the 2 coffee bean groups.

2. From your chosen SVD submatrix, choose the best robusta bean type that best represents the entire family of robusta IR spectra. Cookie-cut out that data column from your SVD matrix.

3. Using matlab's *plot* function, plot the absorbance vs. wavenumber spectrum from your chosen data column in step 2 and save it as a *.jpg file. Again, the format of your plot should look very similar to Figure 4. On that plot, you should have:

a) Title = "SVD # p, robusta m IR signature" , where:

$$\begin{cases} p = The \quad \sigma_p \left(u_p {v_p}^T\right) \quad SVD\ submatrix\ you\ chose \\ m = The \quad m^{th} \quad robusta\ bean\ you\ chose\ that\ best\ represents\ the\ robusta\ group\ ! \end{cases}$$

b) X-axis label = "wavenumber" , where it is shown in *reverse axis order* by using:

set(gca, 'XDir', 'reverse')

c) Y-axis label = "% Absorbance"

d) Add *grids* to your plot. This will be important for visualizing your data in the next step.

4. Now, read the accompanied research paper by *R. Briandet, et. al.* carefully. In the text, the authors suggests that robusta beans may be differentiated from arabica beans because of an increase in IR signatures from 2 chemical ingredients. Please identify these 2 chemicals (type the answer in your diary file).

Hint: Your plotted IR signature should look really similar to one of the figures in that paper, and also, your plot should contain a linear combination of the peaks that are associated with these 2 chemicals !! =)

Task #3: Confirm your "best IR signature" findings using the old-school covariance S

Recall that if we had a giant data matrix X, we can immediately calculate the covariance matrix S :

$$X = \text{Data matrix} \xrightarrow{\text{subtract the means of each column of data}} D = \text{Deviation matrix} \xrightarrow{\text{Find covariance matrix } S} S = \frac{1}{n-1} D^T D$$

And once you know the eigenvectors and eigenvalues of S (and ranked them in order of descending values of λ), you can identify the "dominant," secondary, tertiary… etc. principal components of your data:

$$S\ V_{cov} = V_{cov}\ \Lambda, \quad where \quad \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_{10} \end{bmatrix}$$

$Big\ \lambda$ → $Small\ \lambda$

$$and \quad V_{cov} = \begin{bmatrix} \vec{v_1} & \vec{v_2} & \cdots & \vec{v_{10}} \end{bmatrix}$$

$Dominant\ principal\ axis$ → $Un-imporant\ axes$

Now, if you recall from Problem 1's overview file:

The V-matrix derived from the SVD decomposition:

$$X = U\ \Sigma\ V^T$$

⟷ They will be close to each other !

The V_{cov} matrix derived from the covariance matrix S

(The main difference is that V_{cov} contains the *zero-meaned* data matrix D instead of the raw data matrix X)

Your Tasks for this section: As usual, please turn in all **.m matlab scripts*, plots, or requested echoed values (in diary files) shaded in yellow.

1. From your original data matrix X, calculate (no need to echo these):

   a) The deviation matrix D:

$$D = \left[\overrightarrow{d_1} \;\middle|\; \overrightarrow{d_2} \;\middle|\; \cdots \middle|\; \overrightarrow{d_8} \;\middle|\; \overrightarrow{d_9} \;\middle|\; \overrightarrow{d_{10}} \right], \quad where \quad \overrightarrow{d_p} = \overrightarrow{x_p} - mean\left(\overrightarrow{x_p}\right)$$

   b) The covariance matrix S:

$$S = \frac{1}{n-1} D^T D$$

2. Then, using matlab's *eig* function, find the eigenvalue matrix Λ and the eigenvector matrix V_{cov} from the covariance matrix S and echo them in your diary. Remember, Λ and V_{cov} should obey the eigenvalue equation:

$$S \quad V_{cov} \quad = \quad V_{cov} \quad \Lambda$$

*** Look at this matrix, and note that the first 3 eigenvalues λ_1 , λ_2 and , λ_3 are much larger than the other ones !*

3. Normalize all entries in V_{cov} such that *the smallest entry in each column is 1.000*. Echo $V_{Normalized}$ in your diary. Here's a quick example: If V_{cov} was a 3×3 matrix:

$$V_{cov} = \begin{bmatrix} 2.828 & 2.828 & 1.500 \\ 1.414 & -14.14 & -1.500 \\ 0.707 & 0.707 & -0.500 \end{bmatrix} \xrightarrow{Make\ the\ smallest\ entry=1} V_{normalized} = \begin{bmatrix} 4 & 4 & 3 \\ 2 & -10 & -3 \\ 1 & 1 & -1 \end{bmatrix}$$

Part #4: Interpretation of your PCA results from $V_{normalized}$

As you stare at your $V_{Normalized}$ matrix, notice the following characteristics:

- Every entry within the dominant-eigenvalue's corresponding eigenvector (otherwise known as the ***1st Principal Axis***) have the same signs right now
- Moreover, these entries are all small in magnitude
- Then, notice that for other eigenvectors, in particular the one that corresponds to the *2nd-most-dominant* eigenvalue (ie. the ***2nd Principal Axis***), the entries within that vector are similar in magnitude….. but some *have different signs*.

Again, let's suppose V_{cov} was a 3×3 matrix (yours will be a 10×10 matrix !!), where the eigenvectors $\overrightarrow{v_p}$ are listed in descending order with respect to the eigenvalue of S. Once we've calculated $V_{normalized}$, you can actually visualize the principal axes characteristics *by eye !!!!!!!*

$$V_{cov} = \begin{matrix} \lambda_1 \\ (largest) \end{matrix} \quad \begin{matrix} \lambda_2 \end{matrix} \quad \begin{matrix} \lambda_3 \end{matrix}$$

$$\begin{bmatrix} 2.828 & 2.828 & 1.500 \\ 1.414 & -14.14 & -1.500 \\ 0.707 & 0.707 & -0.500 \end{bmatrix} \xrightarrow{smallest\ entry=1} V_{normalized} = \begin{matrix} \lambda_1 \\ (largest) \end{matrix} \quad \begin{matrix} \lambda_2 \end{matrix} \quad \begin{matrix} \lambda_3 \end{matrix}$$

$$\begin{bmatrix} 4 & 4 & 3 \\ 2 & -10 & -3 \\ 1 & 1 & -1 \end{bmatrix}$$

The relative weight factors in the 1st PC will usually give you a qualitative measure of the "similarities" between different coffee beans.

Usually, the 1st PC is not so useful in differentiating between different "groups" of data columns =(

The 2nd and 3rd PCs are usually the more interesting ones. When there are a lot of sign changes, they will usually tell you:

a) Is one or more *GROUPS* of photos (or variables) differ from other GROUPS ?

b) Are there *super-oddballs, or super-"opposite-signed" oddballs* within your entire collections of photos ? You will see instances of this when you have a huge +/- magnitude change between 2 or 3 entries, and the same time, all other entries remain small in magnitude.

The reason why you want to do this will only be apparent when you check out *Johnson & Wichern, Example 8.5, pp. 451-452*. Here's that example copied from p. 452 of the book. My commentaries are on the side…

and

$$\mathbf{R} = \begin{bmatrix} 1.000 & .632 & .511 & .115 & .155 \\ .632 & 1.000 & .574 & .322 & .213 \\ .511 & .574 & 1.000 & .183 & .146 \\ .115 & .322 & .183 & 1.000 & .683 \\ .155 & .213 & .146 & .683 & 1.000 \end{bmatrix}$$

We note that $\mathbf{R}$ is the covariance matrix of the standardized observations

$$z_1 = \frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}}, z_2 = \frac{x_2 - \bar{x}_2}{\sqrt{s_{22}}}, \ldots, z_5 = \frac{x_5 - \bar{x}_5}{\sqrt{s_{55}}}$$

The eigenvalues and corresponding normalized eigenvectors of $\mathbf{R}$, determined by a computer, are

$\hat{\lambda}_1 = 2.437$, $\hat{\mathbf{e}}_1' = [\ .469, \ .532, \ .465, \ .387, \ .361]$

$\hat{\lambda}_2 = 1.407$, $\hat{\mathbf{e}}_2' = [-.368, -.236, -.315, \ .585, \ .606]$

$\hat{\lambda}_3 = .501$, $\hat{\mathbf{e}}_3' = [-.604, -.136, \ .772, \ .093, -.109]$

$\hat{\lambda}_4 = .400$, $\hat{\mathbf{e}}_4' = [\ .363, -.629, \ .289, -.381, \ .493]$

$\hat{\lambda}_5 = .255$, $\hat{\mathbf{e}}_5' = [\ .384, -.496, \ .071, \ .595, -.498]$

Using the standardized variables, we obtain the first two sample principal components:

$$\hat{y}_1 = \hat{\mathbf{e}}_1'\mathbf{z} = .469z_1 + .532z_2 + .465z_3 + .387z_4 + .361z_5$$

$$\hat{y}_2 = \hat{\mathbf{e}}_2'\mathbf{z} = -.368z_1 - .236z_2 - .315z_3 + .585z_4 + .606z_5$$

These components, which account for

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p}\right)100\% = \left(\frac{2.437 + 1.407}{5}\right)100\% = 77\%$$

of the total (standardized) sample variance, have interesting interpretations. The first component is a roughly equally weighted sum, or "index," of the five stocks. This component might be called a *general stock-market component*, or, simply, a *market component*.

The second component represents a contrast between the banking stocks (JP Morgan, Citibank, Wells Fargo) and the oil stocks (Royal Dutch Shell, Exxon-Mobil). It might be called an *industry component*. Thus, we see that most of the variation in these stock returns is due to market activity and uncorrelated industry activity. This interpretation of stock price behavior also has been suggested by King [12].

The remaining components are not easy to interpret and, collectively, represent variation that is probably specific to each stock. In any event, they do not explain much of the total sample variance. ■

Our eigenvector matrix (derived from the covariance matrix S, or equivalently, from matrix R)

$$V_{cov} = \left[\ \vec{e_1}\ \middle|\ \vec{e_2}\ \middle|\ \vec{e_3}\ \middle|\ \vec{e_4}\ \middle|\ \vec{e_5}\ \right]$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \lambda_3 & & \\ & & & \lambda_4 & \\ & & & & \lambda_5 \end{bmatrix}$$

PC #1 (not useful…)

PC #2 (useful !!)

The sum of these Eigenvalues = 5

The 1st PC explains 48.7% of the data Variations

Adding a 2nd PC will gain another 28.1% increase in explaining our data variations.

The 3 sign changes in the original coordinates ($z_i \sim x_i$) signify that the experimental variables z_1, z_2, and z_3 are maybe fundamentally different than z_4 and z_5.

$Z_1 \sim X_1$ = JP Morgan
$Z_2 \sim X_2$ = Citybank
$Z_3 \sim X_3$ = Wells Fargo

$Z_4 \sim X_4$ = Shell
$Z_5 \sim X_5$ = Exxon Mobil

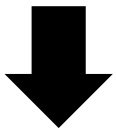
Your Tasks for this section: As usual, please turn in all **.m matlab scripts*, plots, or requested echoed values (in diary files) shaded in yellow.

I would like to examine the first 3 PCs (the largest 3 eigenvalues and their corresponding eigenvectors). To be specific, I want to know the following:

1. What % of the total variance can be explained by taking the 1st three PC's ?

2. In the 2nd PC eigenvector, please *use your intuition* and give a plausible explanation to:

   - Why 5 out of the 10 entries in appear to have a "-" sign and
   - Why do the other 5 entries appear to have a "+" sign ?



By answering this question, you have just proved (statistically) that:

The p^{th} SVD component that you chose in Task 2.1

(that could differentiate between arabica vs. robusta coffee beans)

$$\sigma_p \left(u_p {v_p}^T\right)$$

Is the same p^{th} principal component from your analysis of $V_{normalized}$ …

YES !!!! =)