Statistical Models

Course 2024-2025

SEMINAR 4. Principal Components and Multidimensional Scaling

1 Basic Problems

PROBLEM 1.- For the following covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

a) Determine the population principal components.

b) Calculate the proportion of the total variance explained by the first principal component.

PROBLEM 2.- For the following covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

a) Determine the population principal components.

b) Calculate the proportion of the total variance explained by each of the components.

PROBLEM 3.- Convert the covariance matrix from Problem 1 to a correlation matrix. Using the resulting matrix:

a) Determine the (correlation) principal components.

b) Calculate the proportion of the total standardized variance explained by the first principal component.

PROBLEM 4.- Given a random sample $\mathbf{X}$ below:

$$\mathbf{X} = \begin{bmatrix} 2 & 9 & 6 & 8 \\ 12 & 9 & 9 & 10 \end{bmatrix}$$

a) Compute the sample covariance matrix $\mathbf{S}$

b) Determine the sample principal components.

c) Determine the percentage of variance explained by each component.

PROBLEM 5.- Using the data from Problem 4:

a) Calculate a 95% confidence ellipse for the mean of the population.

b) Use the confidence ellipse to decide whether $H_0 : \mu = [4, 7.5]^T$ can be rejected at $\alpha = 0.05$ significance level.

c) Calculate individual 95% confidence intervals of the mean for each dimension of $\mathbf{X}$ (Hint: use Bonferroni correction). Do your reach the same conclusion as (b) using these intervals?

d) Calculate individual 95% confidence intervals of the mean for each **principal component**. Do your reach the same conclusion as (b) using these intervals?

PROBLEM 6.- When decomposing the sample covariance matrix from a certain dataset $\mathbf{X}$ consisting of a total of 6 samples, we obtain the following eigenvalues and eigenvector matrices ($\mathbf{\Lambda}$ and $\mathbf{P}$, respectively):

$$\mathbf{\Lambda} = \begin{bmatrix} 9.8892 & 0 & 0 \\ 0 & 4.8579 & 0 \\ 0 & 0 & 0.4530 \end{bmatrix} \qquad \mathbf{P} = \begin{bmatrix} -0.3874 & -0.5496 & 0.7401 \\ 0.8941 & -0.0283 & 0.4470 \\ -0.2247 & 0.8349 & 0.5024 \end{bmatrix}$$

We are also given the coordinates of our 6 samples in PCA space (i.e. the projections of the original samples in the direction of the eigenvectors indicated above):

$$\mathbf{Y} = \begin{bmatrix} 2.0129 & -4.1265 & 0.9561 & 1.6069 & 3.2944 & -3.7438 \\ -0.8916 & -1.8209 & -2.2478 & 3.2904 & -0.3703 & 2.0401 \\ 0.3915 & -0.8100 & 0.1823 & -0.7686 & 0.0983 & 0.9065 \end{bmatrix}$$

a) Can you reconstruct the coordinates of the samples $\mathbf{X}$ (in the *original* space)? (Yes / No - Justify).

b) If you answer YES to (a), reconstruct the samples. If you answer NO, then indicate what else you need to assume to be able to reconstruct the original samples. Make a reasonable assumption and proceed to the calculation.

c) What percentage of variance from the original data can you explain if you retain only the first 2 principal components?

d) Reconstruct the original sample coordinates as done in (b), but using only the first 2 principal components.

e) Quantify how much is the difference between your results in (d) with respect to those in (b).

f) How do your answers to (e) and (c) relate to each other?

PROBLEM 7.- Table 1 shows the approximate straight-line distances between a few Spanish cities.

a) Re-arrange the information from the table in the form of a 6×6 matrix of pair-wise distances.

b) Convert the matrix obtained in (a) to the matrix of inner products $\mathbf{B}$, i.e. one with entries $\{b_{ij} = \mathbf{x}_i^T \mathbf{x}_j\}$.

c) Perform eigendecomposition of matrix $\mathbf{B}$ and determine the number of dimensions to retain.

d) Obtain the configuration of the cities listed in the table in the *embedding* space.

e) Compute the distances between all pairs of cities in the embedding space. Compare the results to those indicated in the table (hint: make a scatter plot between them).

Table 1: Straight-line city distances [km]

City	Barcelona	Lleida	Madrid	Sevilla	Zaragoza
Badajoz	825	713	328	187	601
Barcelona		130	505	829	257
Lleida			384	737	126
Madrid				390	273
Sevilla					644

PROBLEM 8.- Table 2 shows the travel distance (by road) between the same cities as the previous problem. Repeat the calculations using these new distances and compare to the results obtained before.

Table 2: Road distances between cities [km]

City	Barcelona	Lleida	Madrid	Sevilla	Zaragoza
Badajoz	1015	862	398	244	711
Barcelona		156	619	992	308
Lleida			458	992	153
Madrid				533	312
Sevilla					843

PROBLEM 9.- Table 3 shows the same data as Problem 7 but adding the Earth Centre.

a) Obtain the configuration of the 7 resulting points in an *embedding* space of 2 dimensions.

b) Compute the proportion of variance explained by the first 2 dimensions.

c) Compute the distances between all pairs of points in the 2D embedding space. Compare the results to those indicated in the table (hint: make a scatter plot between them).

d) Obtain the configuration of the 7 resulting points in an *embedding* space of 3 dimensions.

e) Compute the proportion of variance explained by the first 2 dimensions.

f) Compute the distances between all pairs of points in the 3D embedding space. Compare the results to those indicated in the table (hint: make a scatter plot between them).

Table 3: Straight-line distances [km]

Point	Barcelona	Lleida	Madrid	Sevilla	Zaragoza	Earth Centre
Badajoz	825	713	328	187	601	6371
Barcelona		130	505	829	257	6371
Lleida			384	737	126	6371
Madrid				390	273	6371
Sevilla					644	6371
Zaragoza						6371

PROBLEM 10.-

(a) Provide the definition of the first principal component (PC).

(b) Demonstrate that, based on your definition in (a), the first PC can be obtained by projection into a unit-norm eigenvector of the covariance matrix of the data.

(c) Demonstrate that, from all eigenvectors of the covariance matrix, the one needed in (b) is the eigenvector associated to the largest eigenvalue.

PROBLEM 11.- Demonstrate that after transforming a given dataset onto its principal components, if we retain all the components then the total variance of the data is preserved.

PROBLEM 12.- We would like to perform PCA to the following sample of $n = 7$ points in $p = 2$ dimensions:

$$\mathbf{X} = \begin{bmatrix} -3 & -2 & -1 & 0 & 1 & 2 & 3 \\ 10 & 5 & 2 & 1 & 2 & 5 & 10 \end{bmatrix}$$

a) Compute the principal components of the data.

b) If we retain only the first component we can no longer represent the input data perfectly; make a plot showing the original input data as well as the approximation obtained by retaining only the first component. Assess visually the fitting of your approximation to the actual data and provide a conclusion.

2 Lab Practice Problems

PROBLEM 13.- File SM22_Seminar_4_ExpressionShapes.xlsx contains 260-dimensional vectors of facial shapes for 70 subjects, each of them displaying 4 different facial expressions: 1. Neutral, 2. Smile, 3. Anger, 4. Scream. For further clarity, the first row of the file indicates the expression being displayed (i.e. from 1 to 4) and the data of each subject corresponds to consecutive columns. Thus, the first 4 columns correspond to shapes from subject 1, columns 5 to 8 correspond to shapes from subject 2, etc; and the shape information is contained from rows 2 to 261.

The total number of samples is therefore $(70$ individuals$) \times (4$ expressions$) = 280$. For each sample, the facial shape is described by a set of 130 landmark points, each of them consisting of two coordinates; therefore, each facial shape is a 260-dimensional vector. The function PlotFaceXY_130.m (provided within the seminar data folder) can be used to plot any of those 260-dimensional vectors and display the facial shape that it represents[1].

We wish to answer the following research question: *on average, do the facial shapes of the 4 expressions indicated above differ significantly?*

a) State the null and alternative hypotheses for the research question indicated above.

b) Reduce the dimensionality of your input data by projecting it linearly into the direction of highest variance.

c) Indicate what method would you use to test the hypotheses stated in (a) using the 1-dimensional representation obtained in (b). State the assumptions required by the method to make sure that the result of the analysis will be valid.

d) Perform the test selected in (c) at significance level $\alpha = 0.01$. Clearly indicate your conclusion and your answer to the research question.

e) Use bootstrap or permutation tests to assess whether the direction identified in (b) can be considered significant, at $\alpha = 0.01$. Clearly explain the procedure that you follow and report the estimated p-value. Hint: to determine significance you should not look at the direction vector, but at the magnitude of its associated eigenvalue.

[1] In case you wish to plot the points with your own function, simply note that the first 130 dimensions correspond to the horizontal coordinates and the remaining 130 dimensions correspond to the vertical coordinates, maintaining the same ordering.

PROBLEM 14.- When we reduce the dimensionality, it is informative to inspect the new low dimensional representation to gain some insight of what we have obtained. We will do so using again the data from Problem 10, and we will see how such inspection reveals a shortage of our analysis in Problem 10. Then, we will propose some solution and will repeat the analysis.

To accomplish that, follow these steps:

a) Use the direction of maximum variance from Problem 10(b) together with the sample mean, to synthesize new facial shapes showing the variation along the selected direction. In other words: generate sample points in the low-dimensional space (in this case, 1-dimensional) and convert them back to the 260-dimensional space that you used as input in Problem 10, so that for each generated sample point you get one facial shape. The generated sample points should go from -3 to $+3$ standard deviations along the direction of maximum variance. This is popularly referred to as the *first mode of variation* of your data.

b) Describe qualitatively what is the variation that the direction of maximum variance has captured. Analyze how does this relate to the research question of Problem 10 (i.e. do you think it is convenient or not?).

Most likely, you have just found that the direction of maximum variance identified in Problem 10(b) was strongly affected by *global variation*, such as rotation, translation or scaling. Such variation is not really considered *shape* variation.[2]

One possible solution to the above is that, instead of using the original data, we try to estimate some measure of deformation between shapes. One such estimate can be obtained by means of the Procrustes distance, which might be understood as a measure of deformation between shapes after removing any effect of rotation, translation and scaling. File `SM22_Seminar_4_ExpressionDistances.xlsx` provides a distance matrix computed in this manner for the same 280 facial shapes from Problem 10 (and also in the same order). Reducing the dimensionality using this matrix seems more promising than what we have done in Problem 10:

c) Apply multidimensional scaling to the distance matrix just described and make a scatter plot of the estimated configuration of the data points using only the first 2 dimensions of highest variance. Now we will not see facial shapes but points in the embedding space; therefore, to make sense of it, we need to at least differentiate between the points that correspond to each of the 4 expressions that we are analyzing (e.g. using colors). Make sure the plot is clear and well presented.

d) Repeat the analysis from Problem 10(d). In order to have results that are comparable to those in 10(d), use only one of the dimensions obtained in 11(c), which should obviously be the one with highest variance.

[2] E.g. a tennis ball and a bowling ball are both approximately spherical, regardless of their size. The same reasoning applies to rotation and translation.

PROBLEM 15.- File SM22_Seminar_4_14dims.xlsx contains a random sample from a $p = 14$ dimensional distribution. We are interested in assessing how many factors shall be used to model the data. For this purpose we will perform permutation tests on the data matrix to determine what would be the magnitude of the covariance eigenvalues given a random data structure:

a) Compute the sample covariance matrix $\mathbf{S}$; compute the sample principal components $\mathbf{P}$ and their associated eigenvalues, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$

b) Perform R random permutations of the data and, for each permutation, compute again the covariance matrix and its eigenvalues, $\lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_p^*$. When doing the permutations, to be compatible with the null hypothesis of random data structure, you must permute samples within each dimension. You cannot permute samples between different dimensions. Display the resulting eigenvalues (for all R permutations) in a scree plot.

c) Compare the eigenvalues obtained in (a) and (b) to determine the number of components that should be retained (and used to derive factor loadings). To do so, you should test if each eigenvalue λ_j is significantly different from the distribution of the corresponding random eigenvalues λ_j^*

PROBLEM 16.- Many psychological studies have found evidence that emotions can be represented in an abstract 2-dimensional (2D) space. This 2D space can be obtained by means of the following experiment:

- Select a group of volunteers for the experiment.
- Select a set of facial pictures[3] displaying the different emotions that you want to investigate (several pictures per emotion are recommended).
- Show each participant some random pairs of images and ask them to rate each pair according to how similar are the emotions in these two pictures. For example, $10 =$ the two pictures show exactly the same emotion; $0 =$ the pictures show completely opposite emotions.
- Apply Multi-Dimensional Scaling (MDS) to determine the relative configuration of these emotions (pictures) in a hypothetical embedding space (for example, of dimension 2).

File SM22_Seminar_4_PictureSimilarities.xlsx contains similarity data as the one described above, corresponding to 42 facial pictures showing 7 emotions (6 pictures for each emotion). The emotions used in the experiment were: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise.

In order to apply MDS, we need to convert this matrix of similarities, $\mathbf{C}$, to a distance matrix, $\mathbf{D}$. The standard way to do so is as follows:

$$d_{ij} = \sqrt{c_{ii} - 2\, c_{ij} + c_{jj}}$$

where c_{ij} are the similarity scores between pictures i and j (i.e. those provided in the excel file), and d_{ij} are the resulting distances between pictures i and j.

a) Using the similarity scores from file SM22_Seminar_4_PictureSimilarities.xlsx, convert them into a distance matrix $\mathbf{D}$ and apply MDS.

b) From the coordinates obtained above, $\mathbf{Y}$ in the *embedding* space, display the resulting configuration using the first 2 dimensions (i.e. those associated to the highest eigenvalues). According to the dimensional theory of emotions, the resulting configuration should display an approximately circular arrangement. Make sure to indicate the emotion label for each of the displayed points (e.g. with different colors).

[3] Actually, any other stimuli can be used as long as they clearly relate to an emotion.

PROBLEM 17.- In the previous problem you have displayed a 2D configuration, in which each point corresponds to the embedding representation of the input pictures used in the experiment. Actually, from the full embedding $\mathbf{Y}$ we were interested only in the first two dimensions, which we may call $\mathbf{Y_2}$, i.e. a 2D approximation of $\mathbf{Y}$.

- Test whether there are significant differences between these 7 emotions, in terms of the 2 main dimensions represented by $\mathbf{Y_2}$.
- If you find a significant difference, determine what pairs of expressions differ.
- Perform all tests at significance level $\alpha = 0.05$.
- If the method(s) that you use require the data to fulfill certain assumptions, you are allowed to assume that the data complies with them, as long as you indicate which are these assumptions.

PROBLEM 18.- Files `SM22_Seminar_4_MaleShapes.xlsx` and `_FemaleShapes.xlsx` contain samples that describe the facial shape of 58 men and 54 women, respectively. For each individual, the facial shape is described by a set of 130 landmark points, each of them consisting of two coordinates; therefore, each facial shape is a 260-dimensional vector. The function `PlotFaceXY_130.m` (provided with the lab material) can be used to plot any of those 260-dimensional vectors and display the facial shape that it represents[4].

We wish to answer the following research question: *on average, are there significant differences between the facial shapes of men and women?*

a) State the null and alternative hypotheses for the research question indicated above.

b) Reduce the dimensionality of your input data by projecting it linearly into only 2 dimensions, such that those 2 dimensions retain as much variance as possible. Estimate what percentage of variance is retained by those 2 dimensions.

c) Estimate a 95% confidence interval for the percentage of variance retained by the 2 dimensions that you have computed in (b).

d) Indicate what method would you use to test the hypotheses stated in (a) using the 2-dimensional representation obtained in (b). State and verify the assumptions required by the method to make sure that the result of the analysis will be valid.

e) Perform the test selected in (d) at significance level $\alpha = 0.05$. Clearly indicate your conclusion and your answer to the research question.

[4] In case you wish to plot the points with your own function, simply note that the first 130 dimensions correspond to the horizontal coordinates and the remaining 130 dimensions correspond to the vertical coordinates, maintaining the same ordering.

3 Answers to Selected Problems

PROBLEM 1.-

a) The principal components are:

$$\mathbf{e}_1 = \begin{bmatrix} -0.8944 \\ -0.4472 \end{bmatrix} \qquad \mathbf{e}_2 = \begin{bmatrix} 0.4472 \\ -0.8944 \end{bmatrix}$$

b) The first component explains 85.7% of the total variance.

PROBLEM 4.-

a) $\mathbf{S} = \frac{1}{12} \begin{bmatrix} 155 & -44 \\ -44 & 24 \end{bmatrix}$

b) $\mathbf{e}_1 = [-0.9271, 0.3749]^T \qquad \mathbf{e}_2 = [-0.3749, -0.9271]^T$

c) 95.54% and 4.46% (first and second component, respectively).

PROBLEM 5.-

a) See figure 1.

b) Reject.

c) C.I. dim 1: $(-0.2146; \quad 12.7146$
C.I. dim 2: $7.0467; \quad 12.9533)$. Cannot reject H_0.

d) C.I. first principal component: $-8.9916; \quad 4.9020)$
C.I. second principal component: $(-13.1155; \quad -10.1122$
The hypothesized mean in PCA space is: $(-0.8962, -8.4526)^T$ $\Rightarrow$ reject H_0 because it falls out of the C.I. on the second component.

PROBLEM 6.-

a) With the information provided, we formally cannot reconstruct the original samples (we mess the mean), but we can reconstruct them up to an unknown translation.

b) Assuming centered data (i.e. the sample mean coincides with the origin):

$$\mathbf{X} = \begin{bmatrix} 0 & 2 & 1 & -3 & -1 & 1 \\ 2 & -4 & 1 & 1 & 3 & -3 \\ -1 & -1 & -2 & 2 & -1 & 3 \end{bmatrix}$$

c) Approximately 97.02%

d)

$$\hat{\mathbf{X}} = \begin{bmatrix} -0.2898 & 2.5994 & 0.8650 & -2.4309 & -1.0727 & 0.3291 \\ 1.8250 & -3.6380 & 0.9185 & 1.3436 & 2.9560 & -3.4051 \\ -1.1967 & -0.5930 & -2.0915 & 2.3861 & -1.0494 & 2.5445 \end{bmatrix}$$

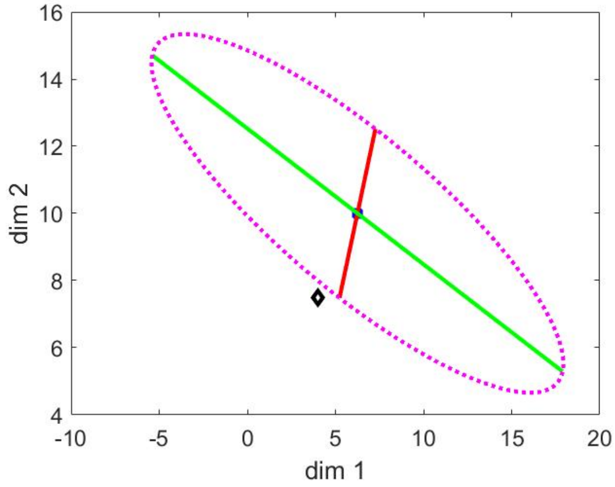

Figure 1: Resulting 95% confidence region for Problem 5-a. The position of the hypothesized mean in 5-b is also displayed.

e)

$$\mathbf{X}-\hat{\mathbf{X}}=\begin{bmatrix}0.2898 & -0.5994 & 0.1350 & -0.5691 & 0.0727 & 0.6709\\ 0.1750 & -0.3620 & 0.0815 & -0.3436 & 0.0440 & 0.4051\\ 0.1967 & -0.4070 & 0.0915 & -0.3861 & 0.0494 & 0.4555\end{bmatrix}$$

f)

$$\frac{\sum_i\sum_j (X_{ij}-\hat{X}_{ij})^2}{\sum_i\sum_j (X_{ij})^2}\simeq\frac{2.2648}{76}\simeq 2.98\% = 100\% - 97.02\%$$

PROBLEM 7.-

b)

$$\mathbf{B}=1000\times\begin{bmatrix}169.1850 & -170.0365 & -123.4144 & 35.7218 & 163.3227 & -74.7786\\ -170.0365 & 171.3670 & 123.4111 & -36.9077 & -161.7223 & 73.8884\\ -123.4144 & 123.4111 & 92.3553 & -22.6290 & -129.1921 & 59.4691\\ 35.7218 & -36.9077 & -22.6290 & 9.8426 & 25.0861 & -11.1138\\ 163.3227 & -161.7223 & -129.1921 & 25.0861 & 192.4295 & -89.9239\\ -74.7786 & 73.8884 & 59.4691 & -11.1138 & -89.9239 & 42.4588\end{bmatrix}$$

c) Two components are enough to explain 99.9% of the variance.

d) See Figure 2.

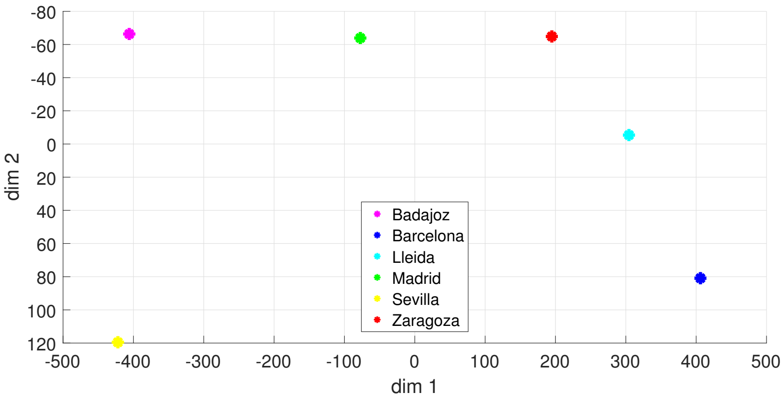

Figure 2: Resulting configuration for Problem 7.