DATA.STAT.770 Dimensionality Reduction and Visualization, Spring 2021, Exercise set 3

Part B: Linear Dimensionality Reduction

Problem B1: Mathematics of Principal Component Analysis

Derive the solution for Principal Component Analysis (PCA) using maximization of variance in the projected space as an optimization criterion. That is, given data $X_{d\times N}$ with N samples in d dimensions, find projection matrix $W_{d\times k}$, $1 \leq k \leq d$, such that $\mathrm{Var}(Z)$ is maximized for projected data $Z = W^{\mathrm{T}}X$.

You may assume that X has zero mean. Derive the solution for the first principal component w_1 by maximizing $Var(w_1^{\mathrm{T}}X)$, and using the constraint that $w_1^{\mathrm{T}}w_1 = 1$.

Problem B2: Principal Component Analysis of Wines

The Wine Quality data set "winequality-red.txt" and "winequality-white.txt" provided in this archive is a data set which can be used to predict quality of white and red wine varieties based on their chemical characteristics. However, in this exercise we do not try to predict the wine quality but simply analyze the rest of the input variables. The data set comes from the UCI Machine Learning Repository and is available there at `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`. Each wine is described by 12 feature values, in order: 1-fixed acidity, 2-volatile acidity, 3-citric acid, 4-residual sugar, 5-chlorides, 6-free sulfur dioxide, 7-total sulfur dioxide, 8-density, 9-pH, 10-sulphates, 11-alcohol, 12-quality.

- For the red wines, find the two first principal components, project the data onto them, and plot the data along them as a scatter plot. Compute the amount of variance explained by the two first components.
- Then repeat the same analysis for the white wines.
- Principal component analysis finds orthogonal projections (orthogonal axes) that contain the maximal variance. The original variables are also orthogonal axes; if PCA was restricted to use only one variable per projection, it would reduce to variable ranking of the original variables based on the variance. Perform such variable ranking for the white wines, take the top-two variables, and compare the resulting picture to the "real" (unrestricted) PCA result.

Hint: in Matlab, consider the Matlab functions "cov" and "eig". R has corresponding functions. Try to plot the wines with low quality (quality value

5 or below), medium quality (quality value 6) and high quality (quality value 7 or above) with different colors.

Problem B3: Principal Component Analysis of an Image

The file staircase.png contains a grayscale picture of the grand staircase of the RMS Olympic ocean liner; the RMS Olympic was owned by the White Star Line and was a sister ship of the Titanic. The image is 240 pixels wide and 200 pixels tall. Let's try using PCA to compress this image, just like was shown on the lecture.

A template code for this exercise, containing the parts unrelated to statistical analysis, is provided in this package in two programming languages, R (file image_compression_and_components_template.R) and Matlab (file image_compression_and_components_template.m).

1. Divide the image into 10 by 10 pixel blocks, so that the first block contains pixels in rows 1-10, columns 1-10, the second block contains pixels in rows 1-10, columns 11-20, and so on. Each block contains 100 pixel values and can be thought of as a 100-dimensional input vector, in total you have $24 \times 20 = 480$ such 100-dimensional input vectors.

2. Compute the first PCA component of this data set (480 items, 100 features) and project each input vectors onto that component - the result is one scalar value per input vector. Then project the scalar values back as a reconstruction of the origal features, as described on the lecture - the result is one 100-dimensional vector per input vector. Draw the resulting picture: for each pixel block, instead of the original pixel values, draw the approximated values.

3. Do the same for 2 components, 5, 10, 20, and 30 PCA components.

4. The PCA projection directions themselves are 100-dimensional vectors, and each element of the vectors corresponds to one pixel in the 10 by 10 image blocks. Therefore, each projection direction itself can be represented as a 10 by 10 block: it is like a filter for the image, showing which pixel values contribute positively to that principal component (high projection coefficients) and which pixel values contribute negatively (low projection coefficients). The provided template code also draws these filter images. Analyze the filters: have they identified important features of the image? Where do the features occur most strongly in the image?

Problem B4: Principal Component Analysis of an Audio File

Principal Component Analysis can be used to compress many kinds of data, not just images. The file the_entertainer.wav contains a part of "The Entertainer", a well-known 1902 piano rag by Scott Joplin (original audio available at `http://`

freemusicarchive.org/music/Scott_Joplin/Frog_Legs_Ragtime_Era_Favorites/04_-_scott_joplin_-_the_entertainer). Just like an image can be divided into spatial blocks, audio waveforms can be divided into blocks over time; each block is a high-dimensional vector of numbers, and the resulting set of high-dimensional blocks can be reduced to a lower dimensionality by PCA.

The template codes musiccompression_template.R (R) and musiccompression_template.m (Matlab) contain code to load the audio file and to transform it into a feature data set of blocks. Your task is to reduce the dimensionality of the feature data set to 15 principal components, and then reconstruct the original data from the principal components, just like was done for the image in problem B3. The template code will then arrange the reconstructed data as a reconstructed audio waveform which can be plotted as a time series, and can also be played just like the original music. You do not need to play the file, but plot it, and save it to a file; provide it as part of your answer.

Note: if you choose to play any reconstructed audio, please do so at low volumes only! In case the audio waveform is not reconstructed well, it can contain clicks or scratches or other irritating or loud sounds which could harm your hearing if played overly loudly.