Analytics for Observational Data (IT142IU)

Lab 3-4: PCA

## Objectives

* Calculate covariances and correlations given datasets.
* Apply PCA to select features.
* Dataset sources:
  + Mid-term.dataset.csv (provided on the Blackboard)
  + <https://archive.ics.uci.edu/ml/datasets/climate+model+simulation+crashes>
* Programming languages: Python/Java

## Analyzing the data

***Part 1.***

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | Mid-term.dataset.csv |
| Correlation before PCA |  |
| Covariance before PCA |  |
| Eigenvalues and  Eigenvectors |  |
| Data after standardizing |  |
| New data after PCA |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| Remarks | The correlation and covariance decrease after PCA |

***Part 2.***

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | climate+model+simulation+crashes  using columns 3-20 |
| Correlation matrix before PCA |  |
| Covariance matrix before PCA |  |
| Eigenvalues and  Eigenvectors |  |
| Data after standardizing |  |
| New data after PCA using **5** components |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| New data after PCA using **10** components |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| Plot the scores for the first two components. What do you notice? Investigate the outliers, and the raw data for each of these unusual observations. What do you conclude about those observations? | * It is low |
| Exclude the unusual observations and refit the PCA model. |  |
| Remarks | Increasing in the score of two first component after excluded the outliers |

***Part 3***

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | climate+model+simulation+crashes  using columns 3-20 |
| Use sklearn.decomposition  to build a PCA model on all the data |  |
| Correlation after PCA | 5 Components    10 Components |
| Covariance after PCA | 5 Components    10 Components |
| Compare with the above results and evaluate | It differs in some way from the previous manual process. |
| Plot the scores for the first two components. What do you notice? Investigate the outliers, and the raw data for each of these unusual observations. What do you conclude about those observations? |  |
| Exclude the unusual observations and refit the PCA model. | Ie: 5 component |
| Remarks | * It’s somehow different from the previous manual process. |

## Some references

* PCA: <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
* Standardization: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
* PCA in Python: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

<https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60>

* Learn more about PCA: <https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/pca-exercises>