Analytics for Observational Data (IT142IU)

*Lab 3-4: PCA*

* 1. Objectives
* Calculate covariances and correlations given datasets.
* Apply PCA to select features.
* Dataset sources:
  + Sample.dataset.csv (provided on the Blackboard)
  + silicon-wafer-thickness.csv (provided on the Blackboard)
* Programming languages: Python/Java
  1. Analyzing the data

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | Sample.dataset.csv |
| Correlation before PCA | * Corrlation: 0.831 |
| Covariance before PCA |  |
| Eigenvalues and  Eigenvectors |  |
| Data after standardizing |  |
| New data after PCA |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| Remarks | Data before PCA:  Strong relationship (corr = 0.831) and positive relationship X1 and X2  Var X1: 0.199  Var X2 : 0.017  X1 is more indicating than X2 |

***Part 2.***

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | silicon-wafer-thickness |
| Correlation matrix before PCA |  |
| Covariance matrix before PCA |  |
| Eigenvalues and  Eigenvectors |  |
| Data after standardizing |  |
| New data after PCA using **4** components |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| New data after PCA using **8** 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? | Using boxplot to show outliers:      Regular Points:   * Most of the data points (blue) are concentrated near the origin, indicating that the principal   components PC1 and PC2 do not have large dispersion among the points in the main cluster.  Outliers:   * There are several key data analysis clusters, such as points with   indices 154, 39, 38, 145, 110, and 25. |
| Exclude the unusual observations and refit the PCA model. |  |
| Remarks | Strong co-variation among the variables was shown by the correlations and covariances before to PCA. These variables were converted into uncorrelated principal components (PCs) via PCA with four components, which maximized variance. Information content per PC was captured by the diagonal components (variances) in the covariance matrix.   The uncorrelated structure (near-zero off-diagonals) and consistent variances were preserved when the PCA components were increased to eight. This indicates that the first four PCs captured the majority of the important data, with other components accounting for the remaining variation. |

Part 3

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset name | silicon-wafer-thickness |
| Use sklearn.decomposition  to build a PCA model on all the data |  |
| Correlation after PCA |  |
| Covariance after PCA |  |
| Compare with the above results and evaluate | The relationship between the 9 components decreases as the core decreases 0 and the covariance increases after PCA, creating gradually stronger fluctuations for the raw data. |
| 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? | Using boxplot to identity the outliers:    Regular Points:   * Most of the data points (blue) are concentrated near the center, indicating that the principal   components PC1 and PC2 do not have large dispersion among the points in the main cluster.  Outliers:   * There are several key data analysis clusters, such as points with   indices 154, 65,25,22, 134, 39, 38, 110. |
| Exclude the unusual observations and refit the PCA model. |  |
| Remarks | Strong co-variation among the variables was shown by the correlations and covariances  before to PCA. These variables were converted into uncorrelated principal components (PCs)  via PCA with four components, which maximized variance. Information content per PC was  captured by the diagonal components (variances) in the  covariance matrix.   The uncorrelated structure (near-zero off-diagonals) and consistent variances were preserved  when the PCA components were increased to eight. This indicates that the first four PCs  captured the majority of the important data, with other components accounting for the remaining  variation. |