NTNU

Norwegian University of Science and Technology

Principal Component Analysis (PCA) for Outlier Detection

Week 02 – Advanced Topic 3

Agenda

- Purpose
- Method
- Results
  - Continuous variable
  - Categorical variable
- Conclusion

Why outlier detection?

- **Data Quality Improvement:** Outlier detection helps identify data errors and ensures data integrity.

- **Better Model Performance:** In machine learning, outliers can have a significant impact on model training and prediction. If outliers are causing the model to perform poorly, removing them might be necessary to improve model accuracy and generalization.

- **Anomaly Discovery:** Outliers often represent unique events or behaviours, providing valuable insights.

Methods: Python PCA package

Hotelling’s T-Square:

Hotelling's T2 is a multivariate extension of the T-test, used to measure the statistical distance of each data point from the centroid of the PCA model, which can identify outliers in the principal components space. A lower p-value T2 indicates that the observation is an outlier in terms of the principal components.

SPE/DmodX (Squared Prediction Error or Distance to Model):

SPE/DmodX measures the squared difference between the original data points and their reconstructed values from the PCA model, focusing on the residuals not captured by the principal components. A high SPE value indicates that the data point lies far from the PCA model, suggesting it may be an outlier that doesn't fit well within the identified components.

Dataset: Continuous variable

- **Wine dataset** from sklearn by Forina, M. et al, as part of the PARVUS project.
- Contains the results of a **chemical analysis** of wines grown in **three different regions** in Italy.

Number of Instances:	178
Number of Attributes:	13 numeric, predictive attributes and the class
Attribute Information:	• Alcohol • Malic acid • Ash • Alcalinity of ash • Magnesium • Total phenols • Flavanoids • Nonflavanoid phenols • Proanthocyanins • Color intensity • Hue • OD280/OD315 of diluted wines • Proline

```
     alcohol  malic_acid   ash  alcalinity_of_ash  magnesium  total_phenols  \
0      14.23        1.71  2.43               15.6      127.0           2.80
0      13.20        1.78  2.14               11.2      100.0           2.65
0      13.16        2.36  2.67               18.6      101.0           2.80
0      14.37        1.95  2.50               16.8      113.0           3.85
0      13.24        2.59  2.87               21.0      118.0           2.80
..       ...         ...   ...                ...        ...            ...
2      13.71        5.65  2.45               20.5       95.0           1.68
2      13.40        3.91  2.48               23.0      102.0           1.80
2      13.27        4.28  2.26               20.0      120.0           1.59
2      13.17        2.59  2.37               20.0      120.0           1.65
2      14.13        4.10  2.74               24.5       96.0           2.05

     flavanoids  nonflavanoid_phenols  proanthocyanins  color_intensity   hue  \
0          3.06                  0.28             2.29             5.64  1.04
0          2.76                  0.26             1.28             4.38  1.05
0          3.24                  0.30             2.81             5.68  1.03
0          3.49                  0.24             2.18             7.80  0.86
0          2.69                  0.39             1.82             4.32  1.04
..          ...                   ...              ...              ...   ...
2          0.61                  0.52             1.06             7.70  0.64
2          0.75                  0.43             1.41             7.30  0.70
2          0.69                  0.43             1.35            10.20  0.59
2          0.68                  0.53             1.46             9.30  0.60
2          0.76                  0.56             1.35             9.20  0.61

     od280/od315_of_diluted_wines  proline
0                            3.92   1065.0
0                            3.40   1050.0
0                            3.17   1185.0
0                            3.45   1480.0
0                            2.93    735.0
..                            ...      ...
2                            1.74    740.0
2                            1.56    750.0
2                            1.56    835.0
2                            1.62    840.0
2                            1.60    560.0

[178 rows x 13 columns]
```

Result: Continuous variable

Hotelling T2 | SPE/DmodX

	y_proba	p_raw	y_score	y_bool	y_bool_spe	y_score_spe
0	0.982875	0.376726	21.351215	False	False	3.617239
0	0.982875	0.624371	17.438087	False	False	2.234477
0	0.982875	0.589438	17.969195	False	False	2.719789
0	0.982875	0.134454	27.028857	False	True	4.659735
0	0.982875	0.883264	12.861094	False	False	1.332104
..	...	...	...	...	...	...
2	0.982875	0.147396	26.583414	False	True	4.033903
2	0.982875	0.771408	15.087004	False	False	3.139750
2	0.982875	0.244157	23.959708	False	True	3.846217
2	0.982875	0.333600	22.128104	False	False	3.312952
2	0.982875	0.138437	26.888278	False	True	4.238283

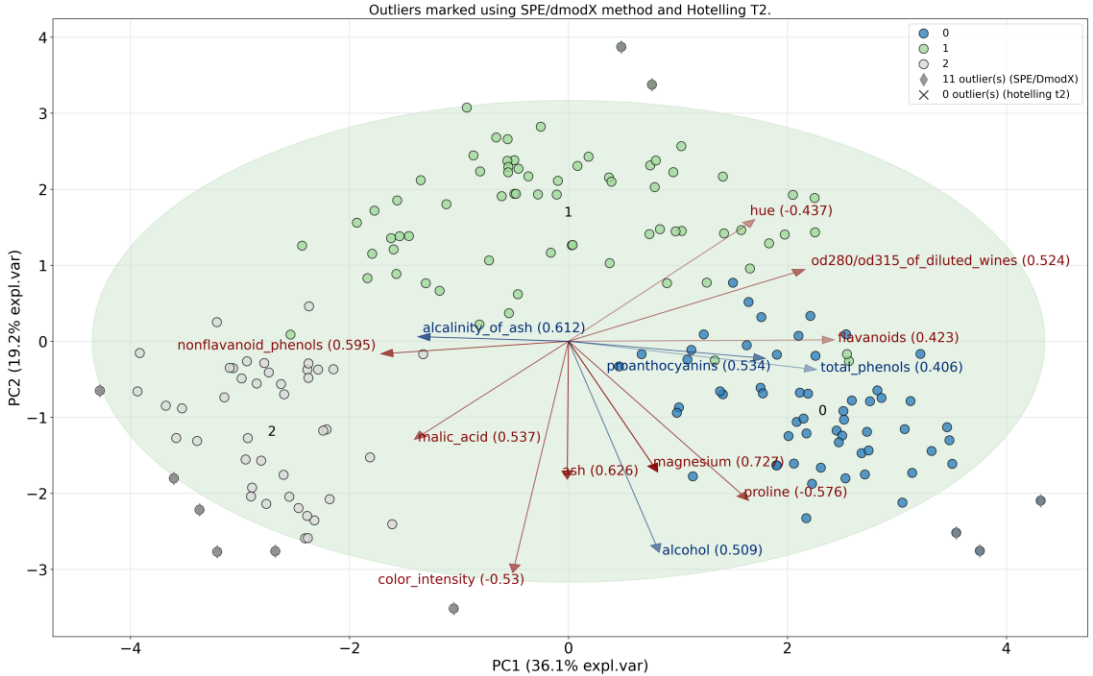
SPE Results:

- **Y_score_spe** = SPE score. Higher value → larger error → outlier
- **Y_bool_spe** = SPE outlier flag. True = Outlier

HT2 Results:

- **Y_proba** = Probability of each observation being an inlier. Close to 1 → low likelihood to be outlier
- **P_raw** = P-value associated with outlier detection. P-value < 0.05 more likely to be an outlier
- **Y_score** = Distance to PCA centroid. Higher values → further from center →outlier
- **Y_bool** = HT2 outlier flag. True = Outlier

Result: PCA plot



alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue
14.37	1.95	2.5	16.8	113.0	3.85	3.49	0.24	2.18	7.8	0.86
14.38	1.87	2.38	12.0	102.0	3.3	3.64	0.29	2.96	7.5	1.2
14.19	1.59	2.48	16.5	108.0	3.3	3.93	0.32	1.86	8.7	1.23
12.0	0.92	2.0	19.0	86.0	2.42	2.26	0.3	1.43	2.5	1.38
11.03	1.51	2.2	21.5	85.0	2.46	2.17	0.52	2.01	1.9	1.71
13.88	5.04	2.23	20.0	80.0	0.98	0.34	0.4	0.68	4.9	0.58
13.17	5.19	2.32	22.0	93.0	1.74	0.63	0.61	1.55	7.9	0.6
14.34	1.68	2.7	25.0	98.0	2.8	1.31	0.53	2.7	13.0	0.57
13.71	5.65	2.45	20.5	95.0	1.68	0.61	0.52	1.06	7.7	0.64
13.27	4.28	2.26	20.0	120.0	1.59	0.69	0.43	1.35	10.2	0.59
14.13	4.1	2.74	24.5	96.0	2.05	0.76	0.56	1.35	9.2	0.61

11 outliers

Dataset: Categorical variable

- Student performance data set, contains 649 samples and 33 variables
- Source:
  Using Data Mining to Predict Secondary School
  ISBN: 978–9077381–39–7

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	\
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	
1	GP	F	17	U	GT3	T	1	1	at_home	other	
2	GP	F	15	U	LE3	T	1	1	at_home	other	
3	GP	F	15	U	GT3	T	4	2	health	services	
4	GP	F	16	U	GT3	T	3	3	other	other	
..	...	..	...	...	...	...	...	...	...	...	
644	MS	F	19	R	GT3	T	2	3	services	other	
645	MS	F	18	U	LE3	T	3	1	teacher	services	
646	MS	F	18	U	GT3	T	1	1	other	other	
647	MS	M	17	U	LE3	T	3	1	services	services	
648	MS	M	18	R	LE3	T	3	2	services	other	

	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	...	4	3	4	1	1	3	4	0	11	11
1	...	5	3	3	1	1	3	2	9	11	11
2	...	4	3	2	2	3	3	6	12	13	12
3	...	3	2	2	1	1	5	0	14	14	14
4	...	4	3	2	1	2	5	0	11	13	13
..	...	...	...	...	...	...	...	...	..	..	..
644	...	5	4	2	1	2	5	4	10	11	10
645	...	4	3	4	1	1	1	4	15	15	16
646	...	1	1	1	1	1	5	6	11	12	9
647	...	2	4	5	3	4	2	6	10	10	10
648	...	4	4	1	3	4	5	4	10	11	11

[649 rows x 33 columns]

NTNU

Transforming dataset: One-hot package

```
     school_GP  school_MS  sex_F  sex_M  age_15.0  age_16.0  age_17.0  \
0         True      False   True  False     False     False     False
1         True      False   True  False     False     False      True
2         True      False   True  False      True     False     False
3         True      False   True  False      True     False     False
4         True      False   True  False     False      True     False
..         ...        ...    ...    ...       ...       ...       ...
644      False       True   True  False     False     False     False
645      False       True   True  False     False     False     False
646      False       True   True  False     False     False     False
647      False       True  False   True     False     False      True
648      False       True  False   True     False     False     False

     age_18.0  age_19.0  age_20.0  ...  G3_14.0  G3_15.0  G3_16.0  G3_17.0  \
0        True     False     False  ...    False    False    False    False
1       False     False     False  ...    False    False    False    False
2       False     False     False  ...    False    False    False    False
3       False     False     False  ...     True    False    False    False
4       False     False     False  ...    False    False    False    False
..        ...       ...       ...  ...      ...      ...      ...      ...
644     False      True     False  ...    False    False    False    False
645      True     False     False  ...    False    False     True    False
646      True     False     False  ...    False    False    False    False
647     False     False     False  ...    False    False    False    False
648      True     False     False  ...    False    False    False    False

     G3_18.0  G3_19.0  G3_6.0  G3_7.0  G3_8.0  G3_9.0
0      False    False   False   False   False   False
1      False    False   False   False   False   False
2      False    False   False   False   False   False
3      False    False   False   False   False   False
4      False    False   False   False   False   False
..       ...      ...     ...     ...     ...     ...
644    False    False   False   False   False   False
645    False    False   False   False   False   False
646    False    False   False   False   False    True
647    False    False   False   False   False   False
648    False    False   False   False   False   False

[649 rows x 166 columns]
```

```python
from df2onehot import df2onehot

# One hot encoding
df_hot = df2onehot(df)['onehot']

print(df_hot)
```

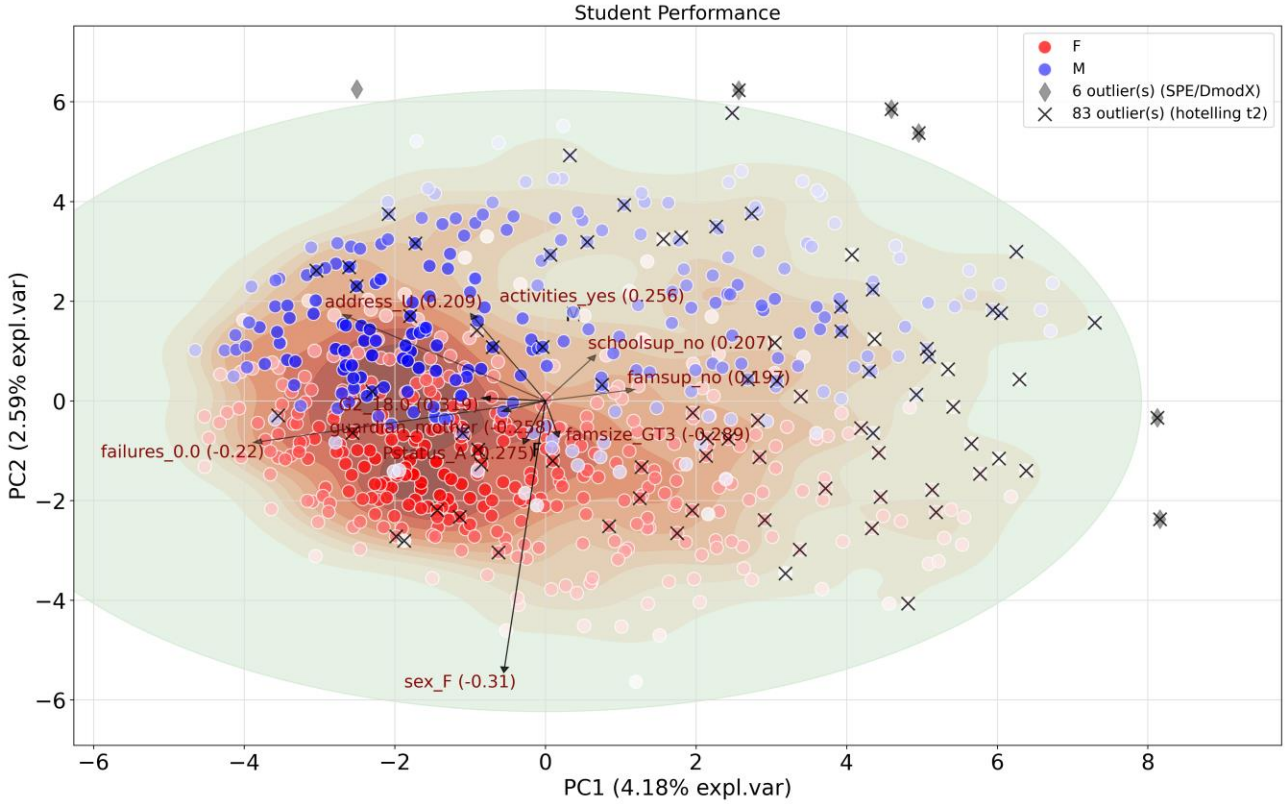
Hotelling T2

SPE/DmodX

	y_proba	p_raw	y_score	y_bool	y_bool_spe	y_score_spe
0	1.000000	0.977574	176.351770	False	False	2.474179
1	1.000000	0.999953	144.148118	False	False	1.626835
2	1.000000	0.958262	181.391445	False	False	1.441919
3	1.000000	0.995238	165.927401	False	False	3.799497
4	1.000000	0.999984	140.221797	False	False	2.975651
5	1.000000	0.999723	151.411072	False	False	4.056198
6	1.000000	0.999999	129.975245	False	False	2.236291
7	1.000000	0.707196	204.219328	False	False	3.594414
8	1.000000	0.841183	195.260585	False	False	3.609387
9	1.000000	0.999829	149.314240	False	False	2.970198
10	1.000000	0.946816	183.547297	False	False	4.538295
11	1.000000	0.996676	163.825436	False	False	2.690649
12	1.000000	0.999243	156.052035	False	False	4.046950
13	1.000000	0.999903	146.981384	False	False	2.750240

Result: Categorical variable

Result: PCA plot



school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet
GP	M	22	U	GT3	T	3	1	services	services	other	mother	1	1	3	no	no	no	no	no	no	yes
GP	M	18	U	GT3	T	2	1	services	services	other	mother	1	1	2	no	no	no	no	no	no	yes
MS	M	18	U	LE3	T	4	4	at_home	health	home	mother	1	4	0	no	yes	no	yes	yes	no	yes
MS	F	19	U	GT3	T	1	1	at_home	services	other	father	2	1	1	no	no	no	no	yes	no	no
MS	F	19	R	GT3	A	1	1	at_home	at_home	course	other	2	2	3	no	yes	no	yes	yes	no	no

5 outliers

Conclusion

- The PCA package, utilizing both Hotelling's T2 (HT2) and Squared Prediction Error (SPE) methods, is an **effective tool for detecting outliers** in datasets containing continuous and categorical variables.
- For **categorical variables**, it is essential to first **transform the data using one-hot encoding** to make them compatible with PCA.
- Although HT2 and SPE are distinct methods, they can be used in **tandem to enhance** the robustness of outlier detection.
- The PCA package yields two significant outcomes: a separated dataset with **identified outliers** and **visual plots** that illustrate the extent to which these outliers deviate within the PCA model.

Pros and Cons:

- Hotelling's T2:
  + Detecting outliers globally and consistent
  - Specific for normal distributed data, sensitive and bias for high-correlated variables

- SPE/DmodX:
  + Strongly detect outlier than modelled conventional PCA, able to detect local outlier
  - Complex, sensitive at selected components

References

1. Source: https://medium.com/dataman-in-ai/handbook-of-anomaly-detection-with-python-outlier-detection-5-pca-d1acbdba1b7e
2. Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010): 433-459.
3. Shlens, Jonathon. "A tutorial on principal component analysis." *arXiv preprint arXiv:1404.1100* (2014).
4. For our revised version of jupyter notebook see: https://github.com/tsaqifwismadi/PCA_outliers/blob/main/PCA%20Outlier.ipynb