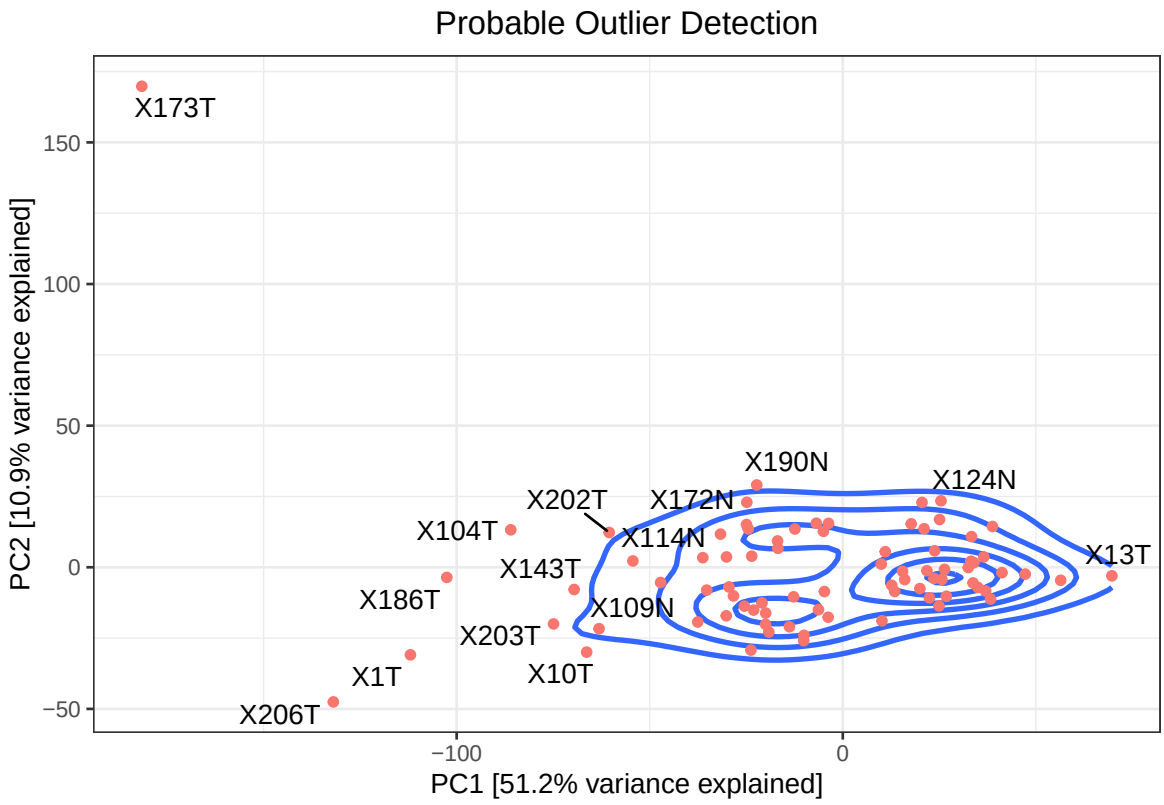
Data Cleaning with Principal Components Analysis

Identifying outliers in mRNA count data obtained from prostate cancer samples

Ifedayo Ojo | MSc Bioinformatics | Teesside University, England

June 1, 2023

Contents



Introduction

Data cleaning is a crucial preprocessing task in data analysis and machine learning. It involves identifying and correcting or removing errors, inconsistencies, and outliers from a dataset. Outliers are data points that deviate significantly from the majority of the data, potentially indicating errors, mistakes, or rare occurrences. Detecting outliers is essential for ensuring data integrity and reliability. Outliers can adversely affect statistical analyses and machine learning algorithms, leading to biased models and inaccurate predictions. Various techniques are used to identify outliers, including statistical methods (such as z-score and interquartile range), visualization techniques (scatter plots, box plots), domain knowledge, and machine learning algorithms (clustering, density-based detection, isolation forests). In this specific exercise, Principal Component Analysis (PCA), an unsupervised machine learning technique and density-based detection, will be employed to identify outliers in mRNA count data obtained from prostate cancer samples.

Principal Components Analysis

Principal Component Analysis (PCA) is a statistical technique utilized for analyzing a data table comprising observations described by multiple correlated quantitative variables. Its primary objective is to extract crucial information from the data and represent it through a new set of orthogonal variables called principal components. These

components facilitate the visualization of patterns of similarity between observations and variables on spot maps. The underlying mathematical principles of PCA rely on the eigen-decomposition of positive semi-definite matrices and the singular value decomposition (SVD) of rectangular matrices. Eigenvalues and eigenvectors, which are numeric values and vectors associated with square matrices, play a pivotal role in this decomposition process. By examining the structure of matrices such as correlation, covariance, or cross-product matrices, PCA reveals relationships within the data. PCA helps in outlier detection by transforming the data into a reduced-dimensional space where outliers become more apparent. By focusing on the principal components that capture the most variation, PCA provides a means to identify observations that deviate significantly from the norm.

Method

Data wrangling

Count data of mRNA of prostate cancer samples obtained from a study with accession number GSE229904 on the Gene expression omnibus database was used for this task. The data count was imported into the R environment. The genes were first mapped to convert the its representation from ensemble id to gene symbol. rows without corresponding gene symbols were removed from the analysis.

Data filtration

The top 10% of the highly varied gene across the samples was selected for the analysis.

Data transformation

The selected data was scaled to ensure fair representation of variables, improve PCA algorithm performance, avoid biases, and enhance interpretability. Furthermore, this data was then transformed from high dimensional to a lower-dimensional space, where the orthogonal variables (principal components) are uncorrelated and capture the most important information.

Outlier detection

A graphical representation of data density with contour lines was used to identify the outliers in the dataset. This allowed visual distinction of regions of high density from low-density regions. Samples that fall outside the high-density regions indicated by the contour lines were considered unusual or anomalous samples (Outliers).

Software

R version 4.2.3 (2023-03-15) – "Shortstop Beagle" Copyright (C) 2023 The R Foundation for Statistical Computing Platform: x86_64-pc-linux-gnu (64-bit)

Results

Data wrangling

To map the ensemble ID in the dataset to gene symbols, the AnnotationDbi package was utilized. This process resulted in 36,314 out of the 60,616 genes successfully being mapped to corresponding gene symbols.

Data filtration

The analysis included genes that exhibited a high degree of variability, specifically the top 10% of genes based on their variability across samples. A total of 3622 genes fell into this category and were considered for further analysis.

Data transformation

Data Scaling

The data was scaled and explored to ensure the density plot of the original data is comparable to scaled data. as shown in Figure 1

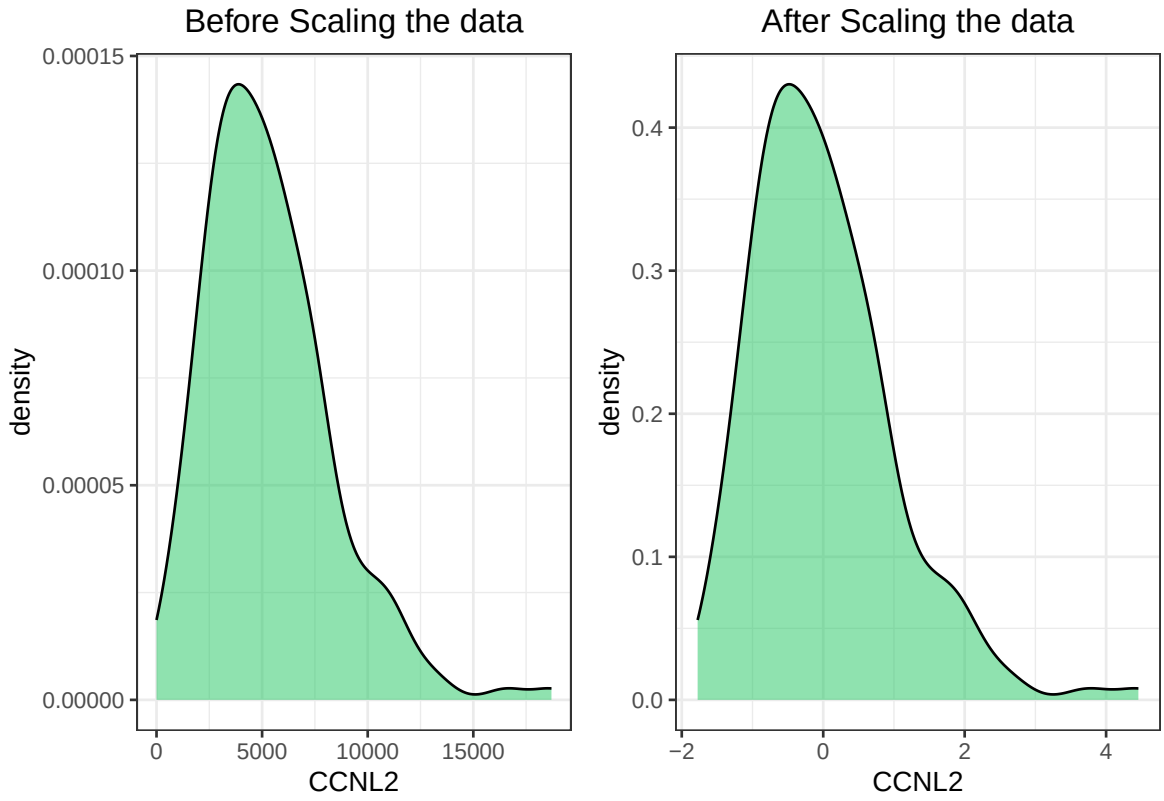


Figure 1: Density Plot of CCNL2 before and after scaling

Variance Explained

Furthermore, the high dimensional scaled data was reduced to a lower dimensional data with its variables represented by principal components. The variance explained by the first 10 principal components is depicted in Figure 2 below.

Figure 2: Variance explained first 10 PCs

Transformed data

In the same vein, Table 1 displays the weights of the genes in the first six principal components of the transformed matrix.

Table 1: Weight of genes in first 6 PCs

Principal Components By **Genes**

Weight of genes in each principal components

	PC1	PC2	PC3	PC4	PC5	
MTND2P28	0.008657295	-0.0228870423	-1.453222e-02	-0.006161125	0.0135173479	0.002470
MTCO1P12	0.007276021	0.0002586450	-5.526754e-03	0.011849770	-0.0082527706	-0.000805
MTATP6P1	0.011390973	-0.0178105829	-6.182318e-06	0.008467411	0.0159186692	0.008270
LINC02593	0.008246655	0.0059931983	7.539814e-03	-0.020360097	-0.0379265780	-0.032639
NOC2L	0.016494702	0.0271583726	1.343979e-02	-0.013571246	0.0163981150	-0.013857
AGRN	0.013121089	0.0220496956	1.537670e-02	-0.021820574	0.0063676907	-0.006414
SDF4	0.018091527	0.0207837016	2.180165e-02	0.004620258	-0.0001376603	0.010142
ACAP3	0.013648141	0.0153370009	2.882062e-02	-0.024333117	0.0107717523	-0.033731
INTS11	0.019336422	0.0155623575	1.933011e-02	-0.010769194	0.0014704701	-0.012487
DVL1	0.016220499	0.0266745452	1.573052e-02	-0.009904784	0.0236675423	-0.014058
MXRA8	0.014953761	-0.0098165943	2.927873e-02	-0.007883940	0.0094499816	-0.016170
AURKAIP1	0.014122209	0.0028304339	2.969406e-02	0.036473080	-0.0022864794	-0.005522
CCNL2	0.020602524	0.0103697541	-3.904671e-03	-0.009300445	0.0136661014	-0.009740
MRPL20	0.018693948	-0.0069578184	1.710577e-02	0.030376366	0.0032595464	0.005910
VWA1	0.011710019	0.0015221573	4.156413e-02	-0.010478281	-0.0078098974	-0.015910
SSU72	0.019872399	-0.0016815982	1.524671e-02	0.022617261	0.0037905425	0.016687
MIB2	0.010727758	0.0088327812	3.763751e-02	-0.017637431	-0.0068648869	-0.036604
SLC35E2B	0.017999468	0.0176380102	-8.052238e-03	-0.004729096	0.0280051547	0.008213
GNB1	0.022054119	-0.0007557841	7.119571e-04	-0.005263831	0.0125481757	0.019816
FAAP20	0.015516626	0.0076987684	3.015149e-02	0.022658882	-0.0040575072	-0.002313

Similarly, the contribution of the samples on the first 6 principal components is presented below in table 2

Table 2: Weight of samples in first 6 PCs

Principal Components By **Sample**

Weight of samples in each principal components

	PC1	PC2	PC3	PC4	PC5	PC6
X102N	47.181554	5.367861	15.7405393	-17.6922378	-4.0381701	4.1992483
X102T	6.290465	14.973709	-2.9628015	-0.6930583	1.2410375	8.7182839
X104N	12.399032	-13.519313	8.3551955	-24.4243784	-0.5143752	4.9722568
X104T	85.986898	-13.228654	-6.4843044	-40.2994522	-12.8083918	-11.4408132
X108N	-20.523899	-22.894232	-6.7204937	2.9014902	0.3414840	9.6293171
X108T	31.669562	-11.708959	-16.6165691	-4.5818746	-12.2079738	8.0889397
X109N	63.127173	21.689387	24.9909978	-5.9747819	8.8599119	1.2642583
X109T	23.792976	29.243804	-6.0593911	11.6369199	5.1111696	7.7462040
X10T	66.334204	29.956644	-17.1908492	-13.7561007	25.4992661	-10.1766542
X110T	-25.680572	4.309481	-11.4881458	-1.9530864	-5.8187293	3.7447877
X111T	19.146182	22.871421	3.3922158	-8.3081036	13.5165764	-0.2359581
X114N	54.360461	-2.231135	5.5106386	-19.1262481	-12.1776087	4.5043128
X114T	3.690930	-15.518810	-22.7664518	64.9304630	-9.0813114	-31.9280639
X114T2	-13.462474	8.554053	-7.2028276	-4.0037394	7.9290185	4.9504763
X115T	-22.449397	10.787623	-16.2649168	2.3993105	-0.4903777	6.8564667
X117T	-16.042391	4.351833	0.8264321	3.8929079	-7.9885948	8.2390876
X119N	25.471852	13.711952	-13.6375656	-12.3284451	16.3527516	-2.3828361
X119T	29.475746	6.948373	-4.9838606	-4.3407110	14.0130189	-6.8582677
X122T	-25.065882	-16.864629	-37.1442394	16.7283739	-13.5670346	-16.5884222
X124N	-25.438388	-23.479679	-1.1570987	12.4827424	0.6773227	-7.8613641

PCA Plot

Figure 3 displays a scatter plot illustrating the lack of correlation between the data projected onto PC1 and PC2, demonstrating the geometric orthogonality of the principal components.

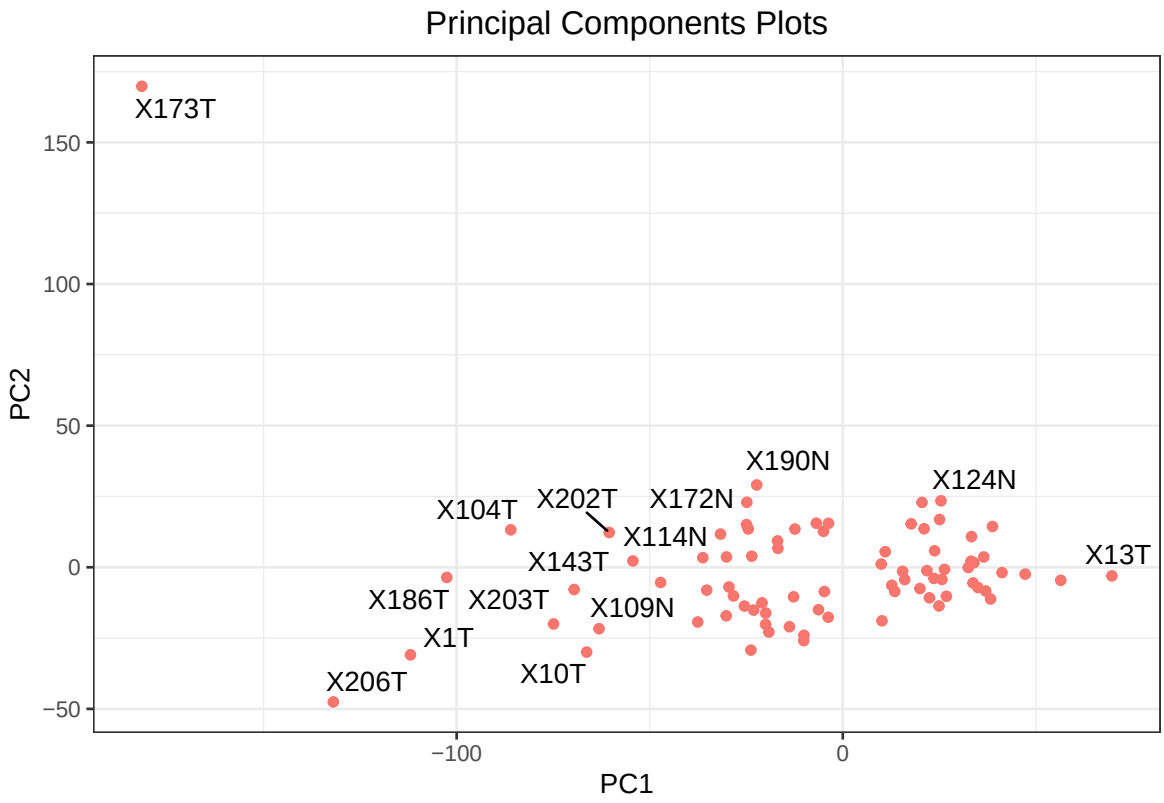


Figure 3: scatter plot of the first two PCs

Outlier detection

Analyzing the joint behavior of PC1 and PC2 involves examining the distribution of the bivariate vector (PC1, PC2). This analysis reveals regions of high density as well as regions with low or sparse density. The contour line displayed below represents the joint probability density of the first two principal components. Data points that fall outside these contour lines are considered potential outliers.

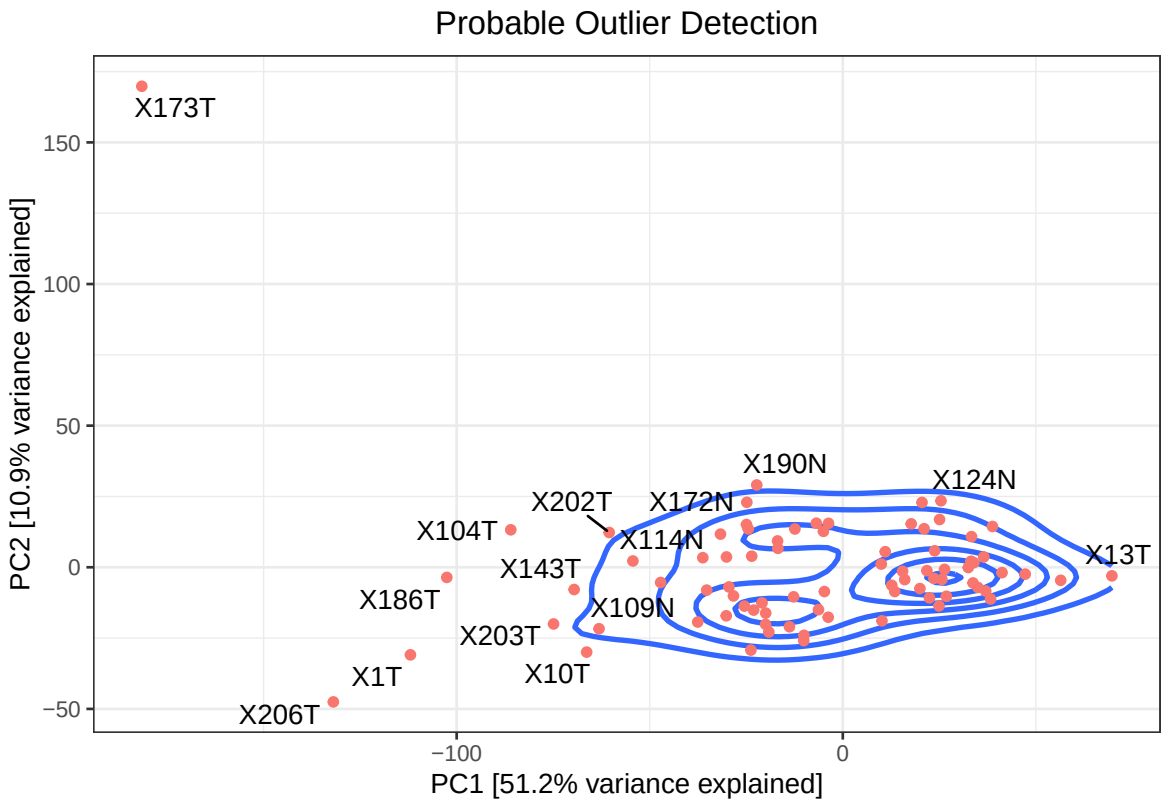


Figure 4:

conclusion

Principal component analysis (PCA) is a technique that reduces the complexity of high-dimensional data while preserving underlying trends and patterns. By transforming the data into a lower-dimensional space, PCA provides a concise summary of the original features. Performing a joint distribution density function on the first two PCs helped to reveal sample X175T,X104T,X1T, X186T and X206T as probable outlier that should be removed to imporove the result of data analysis

References

Devore, J.L., Berk, K.N. and Carlton, M.A. (2021) 'Joint Probability Distributions and Their Applications', in J.L. Devore, K.N. Berk, and M.A. Carlton (eds) Modern Mathematical Statistics with Applications. Cham: Springer International Publishing (Springer Texts in Statistics), pp. 277–356. Available at: https://doi.org/10.1007/978-3-030-55156-8_5.

Huyan, N. et al. (2022) 'Unsupervised Outlier Detection Using Memory and Contrastive Learning', IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 31, pp. 6440–6454. Available at: https://doi.org/10.1109/TIP.2022.3211476.

Lever, J., Krzywinski, M. and Altman, N. (2017) 'Principal component analysis', Nature Methods, 14(7), pp. 641–642. Available at: https://doi.org/10.1038/nmeth.4346.

Mishra, S. et al. (2017) 'Principal Component Analysis', International Journal of Livestock Research, p. 1. Available at: https://doi.org/10.5455/ijlr.20170415115235.