UNIVERSIDAD DE GRANADA

Science Faculty

Multivariate Statics

Influence of diverse indicators in the diagnosis of breast cancer

Bachelor's degree in Computer Science and Mathematics

Authors:
Julián Garrido Arana
Javier Gómez López
Juan Valentín Guerrero Cano

December 2023

Índice

1 Abstract

This study responds to the urgent need for a comprehensive understanding of factors influencing breast cancer diagnosis, a disease with substantial global health implications. Employing advanced statiscal techniques such as discriminant analysis, factorial analysis for dimensional reduction, and principal component analysis, the research aims to identify key indicators for more accurate breast cancer classification.

By improving diagnostic precision, the study contributes to early and personalized interventions. Additionaly, it seeks to unravel complex interrelationships among various indicators, offering potential insights for advancements in breast cancer research. In essence, this work addresses the practival imperative to enhance breast cancer diagnostics and contributes to ongoing efforts in understanding the intricate aspects of this disease, ultimately advancing medical knowledge and breast cancer treatment.

2 Introduction

Within the intricate realm of breast cancer, achieving precise diagnosis stands as a pivotal frontier with profound implications for patient outcomes. Recognizing the limitations of conventional diagnostic paradigms, this study embarks on a meticulous exploration, employing advanced statistical techniques—discriminant analysis, dimensionality reduction through factorial analysis, principal component analysis, etc. The focal point is the unraveling of the intricate tapestry of indicators inherent in breast cancer datasets, particularly those associated with the diverse morphological forms exhibited across various cancer samples.

The amalgamation of genetic, clinical, and morphological indicators demands a nuanced analytical approach. Through discriminant analysis, we aim to uncover distinctive patterns characterizing benign and malignant tumors, pushing beyond the constraints of traditional diagnostic methods. Simultaneously, the application of factorial and principal component analyses allows us to distill essential information from the complex dataset, shedding light on the factors that truly drive breast cancer classification.

This statistical exploration extends beyond immediate diagnostic applications. By identifying morphological indicators related to diverse cancer samples, we not only refine diagnostic precision but also offer insights into the underlying biological relationships governing the manifestation of different forms of breast cancer. This statistical lens enables us to navigate the complexities of the disease, revealing hidden patterns and connections that may inform future avenues of research.

In this journey, our objective is twofold: to enhance the accuracy of breast cancer diagnosis by identifying morphological indicators specific to various cancer forms and to contribute to a broader understanding of the disease's intricacies. As we dissect the statistical nuances associated with morphological variations, we envision a future where diagnostic tools are not only more precise but also inherently adaptable to the diverse manifestations of breast cancer. Through this statistical lens, our exploration aspires to redefine the paradigm of breast cancer diagnosis, providing a foundation for more effective interventions and advancing the collective understanding of this complex and multifaceted disease

3 Materials and Methods

3.1 Database's description

This dataset contains information about breast cancer cell nuclei, with features computed from digitized images of fine needle aspirates (FNA) of breast masses. The dataset is collected from 569 patients based on 10 different indicators of breast cancer. It has three different measures for each indicator, each one corresponding to a spacial dimensions (XYZ axis).

Therefore, the whole dataset is made up of 569 samples and 30 features, plus two additional features that are the ID (it is not interesting in our case) and the Diagnosis (target of the study).

The ten real-valued features for each cell nucleusa are the followings:

- Radius (mean of distances from center to points on the perimeter)
- Texture (standard deviation of gray-scale values)
- Perimeter
- Area
- Smoothness (local variation in radius lengths)
- Compactness ($\frac{\text{perimeter}^2}{\text{area}-1.0}$)
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry
- Fractal dimension ("coastline approximation" - 1)

These features describe characteristics of cell nuclei. The dataset was obtained from images available at `http://www.cs.wisc.edu/~street/images/`.

On the following table there are some important statistics measures for our dataset that could be interesting to know.

breast_cancer_dataset

N: 569

	area1	area2	area3	compactness1	compactness2	compactness3	concave_ points1	concave_ points2	concave_ points3	concavity1	concavity2	concavity3	fractal_ dimension1	fractal_ dimension2	fractal_ dimension3	perimeter1	perimeter2	perimeter3	radius1
Mean	654.89	40.34	880.58	0.10	0.03	0.25	0.05	0.01	0.11	0.09	0.03	0.27	0.06	0.00	0.08	91.97	2.87	107.26	14.13
Std.Dev	351.91	45.49	569.36	0.05	0.02	0.16	0.04	0.01	0.07	0.08	0.03	0.21	0.01	0.00	0.02	24.30	2.02	33.60	3.52
Min	143.50	6.80	185.20	0.02	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.06	43.79	0.76	50.41	6.98
Q1	420.30	17.85	515.30	0.06	0.01	0.15	0.02	0.01	0.06	0.03	0.02	0.11	0.06	0.00	0.07	75.17	1.61	84.11	11.70
Median	551.10	24.53	686.50	0.09	0.02	0.21	0.03	0.01	0.10	0.06	0.03	0.23	0.06	0.00	0.08	86.24	2.29	97.66	13.37
Q3	782.70	45.19	1084.00	0.13	0.03	0.34	0.07	0.01	0.16	0.13	0.04	0.38	0.07	0.00	0.09	104.10	3.36	125.40	15.78
Max	2501.00	542.20	4254.00	0.35	0.14	1.06	0.20	0.05	0.29	0.43	0.40	1.25	0.10	0.03	0.21	188.50	21.98	251.20	28.11
MAD	227.28	13.63	319.65	0.05	0.01	0.13	0.03	0.01	0.07	0.06	0.02	0.20	0.01	0.00	0.01	18.84	1.14	25.01	2.82
IQR	362.40	27.34	568.70	0.07	0.02	0.19	0.05	0.01	0.10	0.10	0.03	0.27	0.01	0.00	0.02	28.93	1.75	41.29	4.08
CV	0.54	1.13	0.65	0.51	0.70	0.62	0.79	0.52	0.57	0.90	0.95	0.77	0.11	0.70	0.22	0.26	0.71	0.31	0.25
Skewness	1.64	5.42	1.85	1.18	1.89	1.47	1.17	1.44	0.49	1.39	5.08	1.14	1.30	3.90	1.65	0.99	3.43	1.12	0.94
SE.Skewness	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
Kurtosis	3.59	48.59	4.32	1.61	5.02	2.98	1.03	5.04	-0.55	1.95	48.24	1.57	2.95	25.94	5.16	0.94	21.12	1.04	0.81
N.Valid	569	569	569	569	569	569	569	569	569	569	569	569	569	569	569	569	569	569	569
Pct.Valid	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Figure 1: Principal measures for some features.

NOTE: As the number of features is too high to show all the measures, we have just attached the table with some of them.

3.2 Statistical methods

Our proyect has been divided in different sections, where we apply different statistical methods, which will help us to interpretate the results and achieve a better performance.

At the beginning, along the second section we have applied a **univariate** exploratory analysis, to be aware of what types of data we are working on. More precisely:

- **Missing values:** we have studied the presence of missing values on all the features of our dataset, although we have not found anyone. If there were any missing value we would analyze the random patterns on them. For that as we have continuous variables we would have applied a Student test, and then replacing them by the mean or median.
- **Outliers:** As some methods we apply along our proyect are sensitive to the presence of outliers we have identified the outliers in each feature. As we have found too many outliers, we have decided to replace them by the mean, as this option was the one which gave us better results.
- **Classical numeric descriptive analysis:** For each feature we have computed a "summary" of some interesting statistics measures such as Minimun value, First Quantile, Median, Mean, Third Quantile and Maximum value. Furthermore, we have plot the histograms of each feature, their density and boxplot for a better understanding of the dataset.
- **Univariate normality:** as many techniques can not avoid the assumption of normality, we have checked the distribution of each feature of our dataset using two different methods. One which is graphical, using the function qqplot, and the second which apply the Shapiro-Wilk test.

Secondly we have performed a **multivariate** exploratory analysis using different techniques such as:

- **Correlation:** we have studied the correlation between the features at two different levels; samplet level, with the corresponding correlation matrix, and at population level, with the Berlett's test.
- **Multivariate normality:** in order to check if the distribution of our data follows a multivariate normal distribution we have performed some test to check this assumption. We have applied the Royston's test and the Henze-Zirkler's test. Previuosly, we have dealt with the presence of the outliers as we know that these tests are sensitive to the presence of the outliers.

After checking all these assumptions, we have applied some other methods that will give us the principal results of the proyect, such as Principal Componente Analysis, Factorial Analysis, Linear Discriminant Analysis and Quadratic Discriminant Analysis for building a classification model. Additionally, the Cluster Analysis has been done to complete our research. We will explain all these techniques more precisely along the following sections.

4 Results

In this section, we will examine the results of the previously mentioned methods from an objective standpoint.

Firstly, it is remarkable that we have not dealt with missing values as we have not them in our dataset. Meanwhile, for the outliers, as explained before, we have replace them by the mean value of each feature. We have considered as outliers the values that are not inside the interquartile range (IQR).

The **univariate normality analysis** have shown that almost every feature of our dataset does not follow a normal distribution. Apart of graphics plot (qq-plots) the Shapiro-Wilk test gave us for each feature a p-value smaller than 0.05 (except for the feature "smoothness1").

	variable	value	NORMALITY
1	radius1	7.460958e-12	NO
2	texture1	5.195521e-05	NO
3	perimeter1	3.028810e-12	NO
4	area1	1.206725e-17	NO
5	smoothness1	5.323514e-02	YES
6	compactness1	9.871618e-13	NO
7	concavity1	3.043752e-18	NO
8	concave_points1	8.317034e-18	NO
9	symmetry1	1.315040e-02	NO
10	fractal_dimension1	1.132071e-08	NO
11	radius2	6.060747e-18	NO
12	texture2	9.830276e-09	NO
13	perimeter2	7.199905e-17	NO
14	area2	7.553385e-23	NO
15	smoothness2	2.431925e-09	NO
16	compactness2	3.783127e-15	NO
17	concavity2	1.437786e-12	NO
18	concave_points2	1.522610e-04	NO
19	symmetry2	1.842650e-12	NO
20	fractal_dimension2	5.191323e-14	NO
21	radius3	4.115530e-15	NO
22	texture3	1.471121e-04	NO
23	perimeter3	4.780865e-15	NO
24	area3	5.140758e-20	NO
25	smoothness3	1.133407e-02	NO
26	compactness3	9.694684e-14	NO
27	concavity3	6.459841e-14	NO
28	concave_points3	1.984878e-10	NO
29	symmetry3	3.773075e-03	NO
30	fractal_dimension3	9.326099e-11	NO

Figure 2: Results of Shapiro-Wilk test.

Regarding the independence of the variables, the **correlation** method at sample level has given us a correlation matrix with almost non-zero coefficients. Even, we can find high correlation coefficients between some features. Applying **the Bartlett's sphericity test** for studying the correlation at population level, we have reach that the null hypothesis is rejected (we obtained a p-value of 0). Therefore, we can conclude after these results that the features of our dataset are correlated, and therefore not independent.

Therefore, as the features are mainly correlated, it is logical to apply a dimensionaly reduction through PCA or FA:

- **Principal Component Analysis:**
- **Factorial Analysis:**

Finally, we have proceeded with the **Discriminant Analysis**. For ensuring the veracity of the results of the different discriminant analysis methods that we have applied, we have studied again (but this time with the target variable) the univariate normality of the data. The results have been the same, and for the multivariate normality too. Therefore, despite the lack of normality in our whole dataset, we built the two models (linear and quadratic) for the Discriminant Analysis. It is remarkable that in the following section we will address these issues and how can we interpretate these results.

- **Linear Discriminant Analysis:**
- **Quadratic Discriminant Analysis:**

The following attached images presents a comparison between the results obtained from these Discriminant Analysis methods.

5 Discussion

Along this section we will interpretate the different results obtained, and which decisions we have made in each step of the process.

Once the data have been preprocessed, we have studied the correlation between variables. As the results were clearly enough to say that our features were not independent, we have made a PCA and FA in order to reduce the dimension of our dataset.

Before continue with the next steps, we needed to decide if we should finally reduce the dimension of our dataset or not. We decided not to reduce the dimension of our dataset, basing our decision in two key factos: the dimension of our dataset (30 features) is not huge enough to make the computational cost too high, so regarding this factor it is not necessary to reduce dimension. The other factor was the nature of our dataset; as we were stuying medical features, we were not interesting in taking less features by making combination of them. This would lead in "not interpetable" features and in this study we are not interesting in losing the capacity of interpretate each feature. Therefore, despite we have made the study of PCA and FA, the following results are obtained regarding that we have not made dimensionaly reduction.

Finally,

6 Conclusion

After, this riguruous analysis of the data we can conclude some imporant facts. Firstly, the dataset was hard to work with, as the most part of the results obtained from the different test were the opposite of what we could expect. Secondly, we have learnt that despite applying dimensionality reduction could be interesting in some cases, in our case it was not too much interesting, as we wanted to be able to interprate the features at the final, as the variables of the dataset were precise measures of the diseases.

Another fact we have noticed along the proyect has been that not every dataset will produce the same results, and not what we "want" to get. In real life, the studies will be done in datasets that probably will not follow the "rules" we expect, and therefore, we have to be careful with the results we obtained for the models for example, as the might be not too much reliable.

To sum up, this proyect has allowed us to know more about the multivariate analysis of real problems, and to learn how to adapt our interpretation to unexpected results. It is remarkable, that the way this proyect it is focused will help us in the future too for machine learning proyect, to make the right statistical analysis for the data, and to make right interpretations of the prediction model.