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

Course 2023-2024

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

In the intricate realm of breast cancer, the quest for precise diagnosis demands a sophisticated analytical approach that transcends conventional methods. This study embarks on a meticulous exploration, leveraging advanced statistical techniques—discriminant analysis, dimensionality reduction through factorial analysis, and principal component analysis. Our focus centers on unraveling the intricate tapestry of morphological indicators within breast cancer datasets, specifically those linked to diverse sample forms across various cancers.

The amalgamation of genetic, clinical, and morphological indicators necessitates a nuanced statistical lens. Through discriminant analysis, we aim to unveil distinctive morphological patterns characterizing benign and malignant tumors, pushing the boundaries of traditional diagnostic methods. Simultaneously, the application of factorial and principal component analyses enables us to distill essential information from the complex dataset, shedding light on the mathematical factors that truly drive breast cancer classification.

This statistical exploration transcends immediate diagnostic applications. By identifying morphological indicators tied to diverse cancer samples, we not only enhance diagnostic precision but also contribute to a deeper mathematical understanding of the underlying relationships governing the manifestation of different forms of breast cancer. This statistical lens empowers us to navigate the complexities of the disease, unveiling hidden mathematical patterns and connections that may inform novel avenues of research.

In this journey, our objective is twofold: to refine breast cancer diagnosis by identifying morphological indicators associated with various cancer forms through advanced statistical analyses and to contribute to a broader mathematical understanding of the intricate relationships within the disease. As we dissect the statistical nuances associated with morphological variations, we envision a future where mathematical 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 mathematical paradigm of breast cancer diagnosis, providing a foundation for more effective interventions and advancing the collective mathematical understanding of this complex and multifaceted disease.

This study aims to refine breast cancer diagnosis by identifying morphological indicators associated with various cancer forms through advanced statistical analyses, simultaneously contributing to a broader mathematical understanding of the intricate relationships within the disease.

3. Materials and Methods

3.1. Materials

For this study, we have used a public Data Base published by the University of California, at its Machine Learning Repository. This database contains information about 30 'different' features measured through 569 instances of breast cancers. In fact, we do not have 30 features, we have 10, but because most of them refers to spatial measures, we have 3 for each one of this type (x, y, z). The ten real-valued feautres are computed for each cell nuclues:

- 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 ($perimeter^2$ / $area - 1{,}0$).
- Concavity (severity of concave portions of the contour).
- Symmetry.
- Fractal dimension (çoastline approximation 1).

Due to the size of the features, we recommend to consult the `R` report of this work to see a summary of the principal position, dispersion and form measures of the features of the database. The table is at section 2.3, where we can find all this interest data.

3.2. Statiscal methods

3.2.1. Univariate exploratory analysis

First of all we have conducted an **univariate exploratory analysis**. Regarding the types of values we have on the dataset, we are not interested on perform any type of data-grouping.
We computed the porcentge of missing values, and we saw that there are not missing values.

Secondly, we went trough a classical numeric descriptive analysis where the basic numerical descriptive statistics such as principal position, dispersion and form measures were given. We also computed their histograms, boxplots and density graphs.

Thirdly, we tried to identify extreme values or outliers. Firstly, we standarized the data to avoid the different scales of our feautres. The decision took for the outliers were to substitute them with the median, due it works better than substituting with the mean.

Lastly, to avoid the assumption of normality, we checked the distribution of each feature of the dataset. We did it in a graphical way and using the Shapiro-Wilk test.

3.2.2. Multivariate Exploratory Analysis

First of all, we checked some conditions that are necessary to apply the different multivariate analysis techniques:

- It is necessary to check if the variables are or no independent. At popultation level, we checked **correlation** using *Bartlett's Test*.
- We also checked the multivariant normality, using the *Henze-Zirkler* test and *Royston* test. It is important to remark that these tests could be perturbated because the presence of outliers, so we had to eliminate them.

Once checked all this conditions, we had applied different techniques of Multivariant Analysis. We conducted a **Principal Component Analysis (PCA)** to reduce the dimension of the problem using observable variables. After it, we made a **Factorial Analysis** to identify potential *latent variables* (no observable) that have a high correlation with a group of observable variables and no correlation with the others. Lastly, we performed a **Discriminant Analysis** (**Linear** and **Quadratic**) to establish a classification method of new observations of a cualitative variable according to its characteristics (predictors).