Received March 11, 2021, accepted April 9, 2021, date of publication April 13, 2021, date of current version April 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3072941

Text Analytics and Mixed Feature Extraction in Ovarian Cancer Clinical and Genetic Data

LUIS BOTE-CURIEL[1], SERGIO RUIZ-LLORENTE[2], SERGIO MUÑOZ-ROMERO[1,3], MÓNICA YAGÜE-FERNÁNDEZ[2], ARANTZAZU BARQUÍN[2], JESÚS GARCÍA-DONAS[2], AND JOSÉ LUIS ROJO-ÁLVAREZ[1,3], (Senior Member, IEEE)

[1]Department of Signal Theory and Communications, Universidad Rey Juan Carlos, 28942 Fuenlabrada, Spain
[2]Unit of Gynecological, Genitourinary and Skin Tumors, Hospital HM Sanchinarro, Fundación Investigación HM Hospitales, 28050 Madrid, Spain
[3]Persei Vivarium, 28013 Madrid, Spain

Corresponding author: José Luis Rojo-Álvarez (joseluis.rojo@urjc.es)

This work was supported in part by the Ministry of Science and Innovation Grants through meHeart-RisBi and Beyond under Grant PID2019-104356RB-C42 and Grant PID2019-106623RB-C41, the FINALE Project under Grant TEC2016-75161-C2-1-R and Grant TEC2016-75161-C2-2-R, and the KERMES Project under Grant TEC2016-81900-REDT, and in part by the European Regional Development Fund (FEDER) under Project PID2019-106623RB-C41.

ABSTRACT Developments of richer integrative analysis methods for oncological studies are needed for efficiently leveraging the amount of clinical and genetic data available to provide the clinicians with better information. However, analyses of this nature often require mixing data of different types, which are not immediate to address jointly with classical methods. In this work, our aim is to find relationships between clinical and genetic features of different types (metric, categorical, and text) and the ovarian cancer (OC) disease progression. To this end, we first propose a univariate statistical method for text type applying bootstrap resampling to Bag of Words and Latent Dirichlet Allocation in order to include as features the free-text fields of the health recordings. Secondly, we extend bootstrap resampling for metric and categorical feature extraction with Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA), respectively. We subsequently formulate a novel and integrative method for jointly considering metric, categorical, and text features. Results obtained in text analysis indicate individual differences in some words between two OC patients groups categorised according to their sensitivity to platinum drugs. These results indicate separability between both groups for text features. Also, regarding the multivariate analysis, clinical data results showed separability patterns for the three methods analysed according to the platinum-sensitivity degree. The use of these analytical tools in our OC cohort has allowed us to demonstrate their strengths by confirming the predictive and prognostic role of widely-known clinical and genetic variables (BRCA status, value of adjuvant therapy and optimal resection, or family history) and demonstrating significant associations in other variables whose role in OC development has been studied to a lesser extent (such as PMS1, GPC3, and SLX4 genes). These results highlight the value of implementing these approaches for the identification of novel biomarkers in the context of OC.

INDEX TERMS Bag of words, bootstrap resampling, clinical data, feature extraction, genetic data, latent Dirichlet allocation, multiple component analysis, ovarian cancer, principal component analysis.

I. INTRODUCTION

High throughput sequencing strategies have been widely applied for gaining insights into the genomic profile of tumors. The comprehensive characterization of genetic alterations, combined with improvements in sequencing techniques, functional interpretation of genetic results, and *in silico* analysis tools helped us to define predictive/prognostic biomarkers and resulted in the clinical application of these molecular findings (e.g. FDA approved panels evaluating the status of cancer predisposing genes) [1].

However, several challenges could be mentioned as limiting factors for the translation of molecular data into clinically actionable markers. First, the need for establishing standardized laboratory and analytical practices to reduce the bias of the experimental workflow followed by each scientific group and to facilitate the integration, sharing and comparison of data of interest. Second, there is a lack of information

The associate editor coordinating the review of this manuscript and approving it for publication was Gina Tourassi.

regarding the biological role of unknown significance genetic variants (USVs) in the development of cancer, which could behave either as pathogenic alterations (causative), as variants slightly increasing the risk for cancer development or modifying the clinical presentation of the disease (age of onset, developed symptoms, or aggressiveness of the tumor) or as passenger alterations (not causative). In addition, despite that tumor mutational burden (TMB) has become a promising biomarker for both prognosis and immunotherapy, several challenges still compromise the adoption of TMB for clinical decision making. Lastly, the development of better integrative analysis models for studies including both extensive clinical and genomics or other -omics data (transcriptomics, proteomics, or epigenomics, among others) are needed for efficiently leveraging the massive amount of molecular data in the benefit of providing treatment recommendations to the clinicians [2].

Given that Data Science analyses often requires mixing sources with different nature, which are not immediate to analyse jointly, the purpose of our study was to use existing or novel analysis methods to identify significant relationships between different types of clinical and genetic features (metric, categorical, and text) and consequently define more reliable predictive and prognostic biomarkers in the context of ovarian cancer (OC) disease. To achieve this goal, we propose, in the first part of this work, a univariate statistical method for text type applying bootstrap resampling to Bag of Words (BoW) in order to inspect the free-text fields of the health recordings. This method is an extension, for text variables, of a previous work that enables to scrutinize, with a unique framework, differences in metric and categorical variables [3]. To complete the text analysis, we introduce the use of Latent Dirichlet Allocation (LDA) method for topic discovery in the same text fields of the health recordings, and, as a novelty, for using it as a method to represent text in which each observation of the dataset is encoded as a real-valued vector. But, although univariate analysis methods are extremely useful and provide us with very relevant information, they have the limitation of not considering interactions between variables. To overcome this limitation, more advanced or sophisticated multivariate methods need to be used.

For this reason, in the second part of this work, we propose a framework which consists of the application of bootstrap resampling to the classical Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) methods to use them as a feature extraction framework of metric and categorical variables, respectively. In detail, this creates a new set of features which capture most of the useful information contained in the initial set of variables. By interpreting this new set of features, we can have an idea of which original variables are more relevant. In addition, the representation of these new features in 2 or 3 dimensions (the most relevant eigendirections) often provides useful information about patterns present in the data. However, neither PCA nor MCA can be used straightforwardly with mixed variables. Therefore, in order to be able to conduct a joint exploration of variables of different nature (metric, categorical, and text), we introduce a new method based on the principles of the two previous ones, which we call MCAPCA method. Overall, these methods are used to extract information and intrinsic patterns in an OC dataset consisting of clinical, text, and genetic features.

The scheme of the paper is as follows. In Section II, we describe the OC dataset used in this work. Then, in Section III, we extend the use of bootstrap resampling principles for textual variables in terms of BoW, and we also present LDA method in this setting. Furthermore, we expand the bootstrap resampling principles to multivariate analysis, using PCA for metric variables, MCA for categorical variables, and a novel variation of these methods for combinations of metric, categorical, and text variables. After that, results with OC data are provided in Section IV. Discussion and conclusions are established in Section V. Additionally, we present simple examples of the use of the described multivariate methods using synthetic data in Appendix.

II. DATASET DESCRIPTION

The dataset used in this work was created as part of the *BRCAness* initiative from the Innovation Oncology Laboratory of the Gynecologycal, Genitourinary, and Skin Cancer Department, at Clara Campal Comprehensive Cancer Center Madrid (Hospital HM Sanchinarro, Spain). Based on the degree of sensitivity to platinum-based drugs, 54 patients were selected from an OC cohort including approximately 300 cases. Genomic DNA was extracted from formalin-fixed embedded paraffin tumors of these 54 patients to perform Next Generation Sequencing (NGS) profiling, either by means of whole-exome sequencing (WES) or predesigned targeted gene panels (Onco80).[1]

Clinical variables cover different aspects of the disease such as: (1) Information provided by the patient prior to the diagnosis, including the symptoms developed or the oncological personal or familial antecedents, which are key variables considering the genetic factors involved in the heritability of ovarian carcinoma; (2) The molecular profiling of potential familial OC patients and the results obtained in the genetic diagnosis screening; (3) Main features of the OC developed, namely the anatomical location of the studied samples, histological features of the primary tumor, tumor stage, grade and perineural invasion; (4) Type of surgical procedure (primary debulking vs. interval surgery) and information regarding the administration of neoadyuvant and/or adyuvant chemotherapies (number of cycles, toxicities, degree of response, or progression, among others); (5) Data regarding the subsequent chemotherapeutic lines of treatment (drug scheme, number of cycles, toxicities, degree of response, or progression); And (6) current status of the patient or information related to disease progression variables (platinum free interval [PFI],

[1] An informed consent was obtained from all the study participants, and the study was approved by the Institutional Review Board of HM Hospitals Ethics committee.

TABLE 1. List of the most relevant variables of the clinical part in the OC dataset and their type and description.

Name	Type	Description
Oncological_History	Categorical	Presence of cancer in personal medical history regardless of the type.
Oncological_History_Description	Text	Description of the type of cancer in personal medical history other than ovarian.
Gynecological_Family_History	Categorical	Presence of gynecological cancer in family medical history.
Gynecological_Family_History_Description	Text	Description of the presence of gynecological cancer in family medical history.
Status_BRCA	Categorical	Mutational status of the BRCA1 and BRCA2 genes.
Age_at_Diagnosis	Metric	Age at diagnosis.
Anatomical_Location	Categorical	Anatomical location of the tumor.
Histology_1st_Component	Categorical	Type of ovarian tumor developed by the patient.
Grade	Categorical	Constitutes a metric feature reflecting the microscopic cell appearance abnormality of the tumoral cells, being the highest score associated to more dedifferentiated or advanced tumors.
Perineural_Vascular_Invasion	Categorical	Existence of vascular or perineural invasion. It is an indicative that the tumor has begun to invade the surrounding tissues.
Stage	Categorical	Stage of cancer in relation to its spread throughout the body.
Surgery	Categorical	Type of surgery representing if it is primary or interval.
HIPEC_in_Surgery	Categorical	Hyperthermic intraperitoneal chemotherapy treatment.
Type_of_Primary_Surgery	Categorical	Type of primary surgery representing an optimal (R0) or suboptimal resection (R1) of the tumor.
Neoadjuvance	Categorical	Chemotherapy treatment prior to primary surgery.
Response_of_Neoadjuvance	Categorical	Observed response to neoadjuvant chemotherapy.
Attitude_of_Interval_Surgery	Categorical	Decision after neoadjuvancy.
Type_of_Interval_Surgery	Categorical	Type of interval surgery representing an optimal (R0) or suboptimal resection (R1) of the tumor.
Adjuvance	Categorical	Chemotherapy treatment after the primary surgery.
Cycles_of_Adjuvance	Metric	Number of cycles of chemotherapy received in adjuvant therapy.
Response_of_Adjuvance	Categorical	Observed response to adjuvant chemotherapy.
PFS	Metric	Progression free survival. It is the time from the first date of pharmacological treatment until radiological or biochemical progression.
PFI	Response variable	Platinum free interval. It is the time between the last cycle of platinum and evidence of disease progression; depending on the length of platinum drugs sensitivity, patients could be categorized as platinum resistant (<6 months) or sensitive (>6 months).
OS	Metric	Overall survival. It estimates the duration of patient survival from the date of diagnosis or treatment initiation.
Bevacizumab_Maintenance	Categorical	Antiangiogenic treatment.
Attitude_after_1er_Line	Text	Descriptions of the patient's progress after first chemotherapy cycle.
Attitude_after_2nd_Line	Text	Descriptions of the patient's progress after second chemotherapy cycle.
Attitude_after_3rd_Line	Text	Descriptions of the patient's progress after third chemotherapy cycle.
Attitude_after_4th_Line	Text	Descriptions of the patient's progress after fourth chemotherapy cycle.
Attitude_after_5th_Line	Text	Descriptions of the patient's progress after fifth chemotherapy cycle.
Attitude_after_6th_Line	Text	Descriptions of the patient's progress after sixth chemotherapy cycle.

TABLE 2. List of the most relevant variables of the genetic part of the OC dataset and their types and description.

Name	Type	Description
HGNC_Symbol	Categorical	Gene nomenclature (Human Genome Nomenclature Committee).
Gene_Description	Text	Gene description.
Chr	Categorical	Chromosome.
Genetic_Change	Categorical	Genetic change from the reference allele to the variant allele.
Genotype	Categorical	Genotype.
VarDepth	Metric	Number of times that reading a specific region, the variant allele has been read.
Conservation_Score	Metric	Conservation of the region under study at an evolutionary level.
Grantham_Distance	Metric	Variable that reflects how different are the amino acids that are changed in missense mutations.
Condel_Prediction	Categorical	Prediction of pathogenicity of the variant according to the Condel tool.
Condel_Prediction_Score	Metric	Score of the degree of pathogenicity of the variant according to the Condel tool.
Sift_Prediction	Categorical	Prediction of pathogenicity of the variant according to the Sift tool.
Sift_Prediction_Score	Metric	Score of the degree of pathogenicity of the variant according to the Sift tool.
PolyPhen_Prediction	Categorical	Prediction of pathogenicity of the variant according to the PolyPhen tool.
PolyPhen_Prediction_Score	Metric	Score of the degree of pathogenicity of the variant according to the PolyPhen tool.
IMPACT	Categorical	Pathogenicity prediction.
Amino_Acids	Categorical	Reference amino acid to amino acid variant translated by the variant.
PFI	Response variable	Platinum free interval. It is the time between the last cycle of platinum and evidence of disease progression; depending on the length of platinum drugs sensitivity, patients could be categorized as platinum resistant (<6 months) or sensitive (>6 months).

progression free survival [PFS], overall survival [OS]) and exitus of the studied cases. The names of the clinical variables used in this work, together with their types and descriptions, are shown in Table 1.

Genetic variables provide information for each of the somatic or germline detected variants. Relevant variables included the gene and chromosome harbouring the mutation, the coding strand, the chromosomic position of the variant, and the detected alleles (genome reference vs. mutant allele). NGS platforms also provided information regarding the sequencing depth of the position of considered nucleotides and the detection frequency for each allele (reference vs. mutation), which is an indirect estimate of the variant status (heterozygous or homozygous), the clonality of the mutation, and the potential infiltration of non-tumoral cells in the tumoral sample. The functional implication of the mutations was covered by different genetic variables such as the location of the variant (coding regions vs. non translated regions, 5' or 3' UTR), the effect on the codified protein (punctual [missense], early truncation, frameshift, or alternative splicing), the phylogenetic conservation score of the DNA region including the mutated nucleotide, and the prediction of pathological defects based on several *in silico* tools (Condel, Polyphen, or Shift). Names, types, and descriptions of the most relevant genetic variables used in this work are presented in Table 2.

III. DATA ANALYSIS

In a preceding work [3], we established a bootstrap resampling analysis framework using simple and univariate statistical descriptions of categorical, metric, and date features in datasets, making results easily interpretable by the users (managers, clinicians, and others). However, text data type was left out of that work. In this current work, we complete the proposed framework by providing an extension of the univariate resample analysis on free-text features using BoW. Besides, this text analysis is supplemented with LDA, a method that discovers latent topics in text and the most relevant words of each of them. Also, and given the need to consider interactions among features, we expand the bootstrap resampling principles to linear multivariate analysis to use them as a feature extraction framework: First, for PCA for subsets of metric variables; Second, for MCA for subsets of categorical variables; And third, for analysis of a mix of metric, categorical, and text variables using a new variant of PCA and MCA adapted to this type of problems.

A. TEXT FEATURES

To address text-analytic problems, BoW method is often used [4]. This method basically consists of histograms for the amount of appearances of each word in a text. We denote by T_j a text feature F_j, that is, $F_j.type = \mathfrak{T}$. With this, $\{(w_j^k, T_j^k), k = 1, \cdots, K_j\}$ is the set of the K_j different words, w_j^k, that can be found in the text of that feature, and their corresponding relative frequency, T_j^k. This relative frequency, T_j^k, which represents the proportion of presence of the word w_j^k in the text, can be written as a probability mass function (*pmf*), denoted here by $P(w_j^k)$. Considering two groups, G_1 and G_2, the conditional *pmf*s for that feature are as follows:

$$P(w_j^k|G_1), \quad P(w_j^k|G_2).$$

With this, we can define a statistic as the difference in conditional *pmf*s,

$$\Delta P(w_j^k) = P(w_j^k|G_1) - P(w_j^k|G_2),$$

which could be used to perform hypothesis tests.

B. LATENT DIRICHLET ALLOCATION

LDA is a generative probabilistic model describing a collection of documents called corpus. Its basic idea is that documents in the corpus are represented as a distribution over latent topics, where each topic is characterized by a distribution over fixed words called vocabulary [5].

We represent the corpus of M documents as $\mathbf{D} = \{d_1, \cdots, d_i, \cdots, d_M\}$, where d_i is a document. Each document d_i has a sequence of N_i words denoted by $\mathbf{w}_i = \left\{w_{i1}, \cdots, w_{ij}, \cdots, w_{iN_i}\right\}$, where w_{ij} is the jth word in the document d_i. The number of latent topics in this corpus is written as K. The probability distribution of the kth topic over the vocabulary is represented as $\boldsymbol{\phi}^{(k)}$, and $\boldsymbol{\theta}_i$ is a probability distribution over the topics in the document d_i. We also define z_{ij} as the topic index for word w_{ij}. With this notation, the generative process for a corpus $\mathbf{D}$ under a LDA model is as follows:

1) For $k = 1, \cdots, K$:
   a) we choose $\boldsymbol{\phi}^{(k)}$ from a symmetric Dirichlet distribution with parameter β.
2) For each document d_i of the corpus $\mathbf{D}$:
   a) We choose $\boldsymbol{\theta}_i$ from a symmetric Dirichlet distribution with parameter α.
   b) For each word w_{ij}:
      i) We choose z_{ij} from a multinomial distribution with probabilities $\boldsymbol{\theta}_i$ and number of trials $n = 1$.
      ii) We choose w_{ij} from a multinomial distribution with probabilities $\boldsymbol{\phi}^{(z_{ij})}$ and number of trials $n = 1$.

The value z_{ij} represents the topic for the jth word in document d_i and is grouped in a set $\mathbf{z}_i = \left\{z_{i1}, \cdots, z_{ij}, \cdots, z_{iN}\right\}$.

For a document d_i, if we know the parameters α and β, the probability distributions of the topics over the vocabulary, $\boldsymbol{\phi} = \left\{\boldsymbol{\phi}^{(1)}, \cdots, \boldsymbol{\phi}^{(k)}, \cdots, \boldsymbol{\phi}^{(K)}\right\}$, the probability distribution over the topics, $\boldsymbol{\theta}_i$, the set of N words, $\mathbf{w}_i$, and the set of N topic per word, $\mathbf{z}_i$, we can define the joint distribution as

$$p(\mathbf{w}_i, \mathbf{z}_i, \boldsymbol{\theta}_i, \boldsymbol{\phi}|\alpha, \beta) = p(\boldsymbol{\phi}|\beta)p(\boldsymbol{\theta}_i|\alpha)p(\mathbf{z}_i|\boldsymbol{\theta})p(\mathbf{w}_i|\boldsymbol{\phi}).$$

However, variables $\mathbf{z}_i$, $\boldsymbol{\theta}_i$, and $\boldsymbol{\phi}$ are unobserved or latent variables. The key problem is to reversing the defined generative process and learning the distributions of these latent variables in the model using the observed data $\mathbf{w}_i$ and the given parameters α and β. In LDA, this amounts to solving the posterior distribution

$$p(\boldsymbol{\theta}_i, \boldsymbol{\phi}, \mathbf{z}_i|\mathbf{w}_i, \alpha, \beta) = \frac{p(\mathbf{w}_i, \mathbf{z}_i, \boldsymbol{\theta}_i, \boldsymbol{\phi}|\alpha, \beta)}{p(\mathbf{w}_i|\alpha, \beta)}.$$

This distribution is intractable to solve analytically. However, there are a number of approximate inference techniques available that they can be applied to the problem, including collapsed variational Bayes [6] and collapsed Gibbs sampling [7].

All in all, LDA is primarily used as a method for topic discovery in text. These latent topics can be well displayed in a chart with the most relevant words of each topic using the probability distribution of each topic over the vocabulary, $\boldsymbol{\phi}$. In addition, and as a novelty, we propose to use LDA as a method to represent text where each document is encoded as a real-valued vector using the probability distribution over the topics in this document, $\boldsymbol{\theta}_i$.

C. PCA WITH BOOTSTRAP RESAMPLING

PCA is one of the best known linear multivariate statistical technique. It analyses a dataset representing observations described by several metric variables, which are, in general, mutually correlated. Its goal is to represent this set of observed metric variables in terms of a smaller set of new orthogonal variables [8]. Therefore, PCA is often used as a dimension reduction technique for data compression or visualisation.

Suppose a dataset with N observations and Q metric variables available. In PCA, the interest is to find the projections of the observations in a lower dimensional space, P, known as principal subspace, with $P < Q$, so that variance is maximised to retain as much variability as possible [9].

Given a dataset in form of table of $N \times Q$, we can represent it by a data matrix $\mathbf{X} \in \mathbb{R}^{N \times Q}$, which we assume that is centered (the mean of each of the Q metric variables is 0). If $\mathbf{B} \in \mathbb{R}^{N \times P}$ is a orthogonal matrix whose columns form an orthonormal basis of the principal subspace, PCA problem is reduced to

$$\underset{\mathbf{B}}{\text{argmin}} \left\| \mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}^\top \right\|_F^2 \quad \text{subject to } \mathbf{B}^\top\mathbf{B} = \mathbf{I},$$

where $\mathbf{X}\mathbf{B}\mathbf{B}^\top$ is the reconstruction of projections of the $\mathbf{X}$ over the subspace spanned by the columns of $\mathbf{B}$ and $\|\cdot\|_F$ is the Frobenius norm [10].

Knowing that $\|\mathbf{A}\|_F^2 = \text{Tr}\left\{\mathbf{A}\mathbf{A}^\top\right\}$, we can write

$$\begin{aligned}\left\| \mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}^\top \right\|_F^2 &= \text{Tr}\left\{(\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}^\top)(\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}^\top)^\top\right\} \\ &= \text{Tr}\left\{\mathbf{X}^\top\mathbf{X}\right\} - \text{Tr}\left\{\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathbf{B}\right\}.\end{aligned}$$

Therefore, the minimization problem can be seen as a maximization problem,

$$\underset{\mathbf{B}}{\text{argmax}}\ \text{Tr}\left\{\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathbf{B}\right\} \quad \text{subject to } \mathbf{B}^\top\mathbf{B} = \mathbf{I}.$$

Since $\mathbf{X}$ is a centered matrix, its covariance matrix (regardless of the scale factor $\frac{1}{N}$) is $\mathbf{S} = \mathbf{X}^\top\mathbf{X}$. Then,

$$\underset{\mathbf{B}}{\text{argmax}}\ \text{Tr}\left\{\mathbf{B}^\top\mathbf{S}\mathbf{B}\right\}, \quad \text{subject to } \mathbf{B}^\top\mathbf{B} = \mathbf{I}.$$

Using Lagrange multipliers, we can rewrite the above equation as

$$\underset{\mathbf{B}}{\text{argmax}}\ \text{Tr}\left\{\mathbf{B}^\top\mathbf{S}\mathbf{B}\right\} - \text{Tr}\left\{(\mathbf{B}^\top\mathbf{B} - \mathbf{I})\Lambda\right\},$$

where Λ is a diagonal matrix containing the Lagrange multipliers. If we derive with respect to $\mathbf{B}$ and set equal to zero, we can solve this problem as a standard eigenvalues problem

$$\mathbf{S}\mathbf{B} = \mathbf{B}\Lambda,$$

where $\mathbf{B} \in \mathbb{R}^{N\times P}$ is the eigenvectors matrix and $\Lambda \in \mathbb{R}^{P\times P}$ is the diagonal eigenvalues matrix.

Therefore, eigenvectors give the directions of maximum variation of the data and eigenvalues quantifying the amount of variation of the data projected on its corresponding eigenvector. So, if we want to find a lower dimensional space to represent the data, we must retain only the eigenvectors where data has more variation, that is, those with larger eigenvalues. For visualisation, the lower dimensional space, P, would be equal to 2 (for 2 dimensions) or equal to 3 (for 3 dimensions).

To find the projections of the observations $\mathbf{X}$ in the lower dimensional space, P, spanned by $\mathbf{B}$, we just perform

$$\mathbf{F} = \mathbf{X}\mathbf{B},$$

where $\mathbf{F} \in \mathbb{R}^{N\times P}$ is the projected data matrix whose columns are the new uncorrelated variables with standard deviation equal to zero.

Projections of the observations can be seen, according to the above equation, as a liner combination of each component of eigenvectors with the original variables, resulting the new (projected) variables. Thus, a large component in an eigenvector (in absolute value) means that the corresponding original variable has more influence in the creation of the new variable. With this, we can have an idea of which original variables are most relevant in each new variable. Also, visualizing the projections of the observations in 2 or 3 dimensions, we can see if the dataset has intrinsic or natural separability.

It is sometimes important to visualise not only the projections of the observations in the principal subspace, but also the original variables. These variables can be plotted as points using the $\mathbf{B}$ matrix as coordinates, with P equal 2 or 3 [8]. More specifically, each of the N rows of the matrix is an original variable that can be represented by its P coordinates. In this way, we can detect the relationships between them.

In some cases, centered data matrix $\mathbf{X}$ is normalised by the standard deviation of each of the Q metric variables. In this situation, we can write the new centered and normalised data matrix as $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is the diagonal matrix of the variances of the Q metric variables. The covariance matrix without the scale factor $\frac{1}{N}$ of $\widetilde{\mathbf{X}}$ is $\widetilde{\mathbf{S}} = \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}$, where $\widetilde{\mathbf{S}}$ is also the correlation matrix of $\mathbf{X}$. Just like before, we can solve this problem as a standard eigenvalues problem $\widetilde{\mathbf{S}}\widetilde{\mathbf{B}} = \widetilde{\mathbf{B}}\widetilde{\Lambda}$, with the projections of the observations as $\widetilde{\mathbf{F}} = \widetilde{\mathbf{X}}\widetilde{\mathbf{B}}$.

On another note, bootstrap resampling method is a statistical technique which is based on the idea that if we want to make an inference from a population in terms of some statistic whose calculation is known, but its actual distribution is not easy to obtain analytically, we can resample with replacement the sample data and make inferences on resamples [3]. This allows us to check how reliable is the statistic, since we could estimate the confidence interval (CI) of it [11]. In our case, we are interested in the CI of each eigenvalue and of each component of each eigenvector resulting from PCA. To do this, we first calculate PCA on each bootstrap resample of the dataset and then we calculate the CI for each eigenvalue and each component of the eigenvectors.

However, several problems can happen with eigenvectors in each bootstrap resample [12]. The first one is that the sign of eigenvectors in PCA is arbitrary, so they could be multiplied by -1 (reflection). The second one is that it could be an inversion in the order of the eigenvectors when two or more eigenvalues are similar (re-order). To reverse these problems, we calculate two distance matrices: One is the distance matrix between the resample eigenvectors and the empirical eigenvectors; The other is the distance matrix between the inverted resample eigenvectors and the empirical eigenvectors. The order and the sign of the resample eigenvectors is decided based on the maximum absolute distance position in each of distance matrix.

In order to illustrate the functioning of this method, we present a simple example with a synthetic dataset in Appendix.

D. MCA WITH BOOTSTRAP RESAMPLING

Correspondence Analysis (CA) is a technique for exploration of two categorical variables (with several categories for each of the two variables) used with a two-way contingency table [13]. This technique is also known as Simple CA [14]. As in PCA with metric variables, the key idea in CA is to reduce the dimensionality of the two-way contingency table and simplify it in a subspace of low-dimensionality. Commonly, two or three dimensions are used for visualization [15], in a way that the categories of the two variables are depicted as coordinates, revealing associations between them [14].

MCA is often seen as an extension of CA which allows to analyse the relationships of more than two categorical variables [16]. However, it can also be seen as a form of PCA applicable to categorical rather than metric variables [17].

In the presence of several categorical variables, there are two possible ways to organize the data for MCA analysis, namely, in an indicator (binary) table or in a Burt table [14]. For our case, we use the former representation.

Let us consider an original data table of $N \times Q$ dimensions, where N is the number of observations and Q is the number of categorical variables. Each categorical variable q_n has J_{q_n} categories, so that total number of categories of all the variables is $J = \sum_{n=1}^{Q} J_{q_n}$. With this, we are able to transform the original data table into an indicator matrix $\mathbf{X} \in \{0, 1\}^{N \times J}$.

The goal of MCA is to reduce the dimensionality of the observations (rows) when variables are categorical. To get this, firstly, we center and normalise the indicator matrix $\mathbf{X}$. If $\mathbf{x} = \frac{1}{N}X^{\top}\ 1$ is a $J \times 1$ vector with the means of each variable, being $\mathbf{1}$ a $N \times 1$ vector of ones, and $\mathbf{D}$ is the diagonal matrix of these means, the centered and normalised matrix $\mathbf{X}$ is $\widetilde{\mathbf{X}} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\top})\mathbf{D}^{-\frac{1}{2}}$. Matrix $\widetilde{\mathbf{X}}$ is the data matrix on which we want to find the orthogonal matrix $\mathbf{B}$ where this data will be projected. As it is demonstrated in PCA and generalizing for this case, we can solve this issue as a standard eigenvalues problem

$$\widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}}\mathbf{B} = \mathbf{B}\Lambda.$$

To find the projections of $\widetilde{\mathbf{X}}$ in the lower dimensional space spanned by $\mathbf{B}$, we perform

$$\mathbf{F} = \widetilde{\mathbf{X}}\mathbf{B}.$$

Thus, MCA can be seen as basically an adaptation of PCA for categorical data. The only difference is that the data matrix to which we apply the standard eigenvalues problem is a binary matrix that is centered and normalised by the means (rather than by the standard deviation as in PCA). Therefore, as in the case of PCA, a large component in an eigenvector (in absolute value) means that the corresponding original variable has more influence in the creation of the new variable, given information of which original variables are most relevant in each new variable. Likewise, a visualization of the projections gives us an idea of the intrinsic separability of the dataset.

Also in this case, we are interested in the CI of each eigenvalue and of each component of each eigenvector as a result of applying MCA. The process is similar to that already explained, i.e., we initially apply MCA to each bootstrap resample of the dataset and then we calculate the CI for each eigenvalue and each component of the eigenvectors. Problems with reflections and with the order of the eigenvectors can also appear when MCA is applied to resamples, solving them using the same method as in PCA. A simple example of this method on synthetic data is presented in Appendix.

E. MCAPCA WITH BOOTSTRAP RESAMPLING

After presenting PCA for metric variables and MCA for categorical variables, we are interested in a method for feature extraction that is able to analyse variables of both types at the same time. Therefore, we propose a new method that we have called MCAPCA. This method pursues the principal idea of reducing the dimensionality of a mixed data matrix with metric and categorical variables, obtaining a subspace of low-dimensionality. Our proposal here is basically a variant of MCA (which in turn is a variant of PCA) where the metric variables have to be properly transformed. Besides, using the LDA text encoding method based on the probability distribution over the topics of observations (documents), $\boldsymbol{\theta}_i$, we can encode each observation in a real-valued vector so that it is possible to analysis text variables together with categorical and metric variables using MCAPCA. The proposed procedure is explained next.

Let us consider an original data table of $N \times Q$ dimensions, where N is the number of observations and Q is the sum of metric (Q_M) and categorical (Q_C) variables. Each categorical variable q_n has J_{q_n} categories, so that the total number of categories of all the categorical variables is $J = \sum_{n=1}^{Q_C} J_{q_n}$. With this, we transform the categorical variables into an indicator matrix $\mathbf{X}_C \in \{0, 1\}^{N \times J}$. We also have to transform the metric variables. In this case, we use the Escofier transform [18], that allows the metric variables to be analysed by MCA producing the exact same results as PCA. After this transformation, we obtain a metric matrix $\mathbf{X}_M \in \mathbb{R}^{N \times 2Q_M}$. Finally, we can join this matrix $\mathbf{X}_M$ with the indicator matrix $\mathbf{X}_C$ obtaining a data matrix $\mathbf{X} \in \mathbb{R}^{N \times (J+2Q_M)}$.

To get the goal of reducing the dimensionality of the observations of this matrix $\mathbf{X}$, we firstly have to center and normalise it. To center, we simply subtract from each column its mean. To normalise, we calculate the first singular value of each variable of $\mathbf{X}$, this is, each metric variable is a sub-matrix of 2 columns (given by the Escofier transform) and each categorical variable will be a sub-matrix of J_{q_n} columns. So, if the number of columns of the sub-matrix of each variable is k_i, and its first eigenvalue is d_i, we can calculate a factor $\alpha_i = \frac{k_i}{d_i}$ [19]. If we define $\mathbf{D}_\alpha$ as the diagonal matrix of these factors, $\mathbf{x} = \frac{1}{N}X^{\top}\ 1$ as a $J \times 1$ vector with the means of each variable, and $\mathbf{D}$ as the diagonal matrix of these means, then the centered and normalised matrix is $\widetilde{\mathbf{X}} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\top})(\mathbf{D}\mathbf{D}_\alpha)^{-\frac{1}{2}}$.

Matrix $\widetilde{\mathbf{X}}$ is the data matrix on which we have to find the orthogonal matrix $\mathbf{B}$ to project the data, solving this, like in PCA and MCA, as a standard eigenvalues problem given by

$$\widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}}\mathbf{B} = \mathbf{B}\Lambda.$$

Finally, we are able to find the projections of $\widetilde{\mathbf{X}}$ in the lower dimensional space extended by $\mathbf{B}$ doing

$$\mathbf{F} = \widetilde{\mathbf{X}}\mathbf{B}$$

As in the two previous cases, we are interested in determining how is the CI of each eigenvalue and each component of eigenvectors obtained from using MCAPCA. To do this, we use the bootstrap resampling method as we have done in PCA and MCA. Same problems of reflection and in the order of the eigenvectors can also arise when MCAPCA is applied to resamples, solving them using the same method as in PCA and MCA.

IV. EXPERIMENTS AND RESULTS

In this section, we apply the methods exposed in the previous section to analyse the OC dataset. We first explore the text features of medical comments present in the OC dataset using BoW method and LDA. Then, we analyse the metric features of the OC dataset applying PCA method, the categorical features of the OC dataset with MCA method, and the new MCAPCA method to metric, categorical, and text features of the dataset together.

A. TEXT ANALYSIS IN OC DATASET

In this part, we analyse text features of medical comments present in the OC dataset, using BoW method and LDA for this aim.

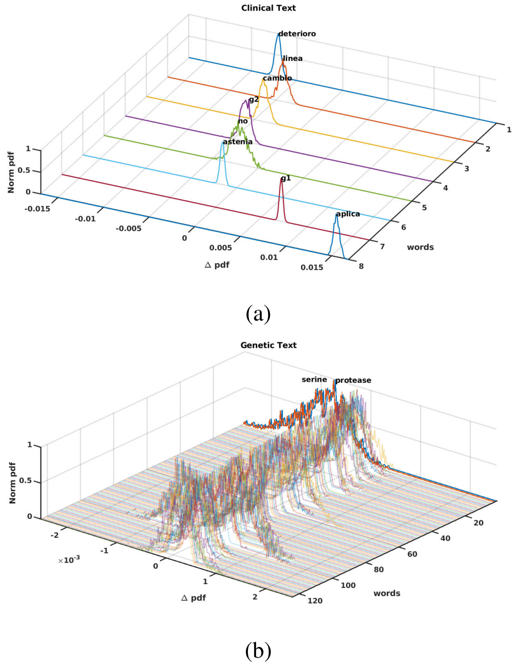
Each observation of the genetic part of the OC dataset has a text feature providing the description of the sequenced gene (*Gene_Description*) (Table 2). In the clinical part, there are more than one text feature per observation which we analyse jointly, namely, description of the type of cancer developed by the patient other than ovarian (*Oncological_History_Description*), presence of a family history of gynaecological cancer (*Ginecological_Family_History_Description*), and descriptions of the patient's progress after each chemotherapy cycles (*Attitude_after_1st_Line, Attitude_after_2nd_Line, others*) (Table 1).

Before using BoW or LDA with the above text variables, a data preprocessing is necessary. In detail, we convert words to lowercase, we tokenize the text, that is, we represent the text as a collection of words (also known as tokens), we remove the stop words that can add noise to data like ''a'',''and'', ''to'', or ''the'', we erase punctuation symbols, and we remove words that have do not appear more than 5 times in total. With all this, the clinical part of the OC dataset has a vocabulary of 64 words and the genetic part has a vocabulary of 130 words.

For BoW analysis, we calculate, for the text features of the clinical and genetic part of the OC dataset, the *pmf* of each word, w_j^k, for G_1 and G_2 groups, namely, $P(w_j^k|G_1)$ and $P(w_j^k|G_2)$. As we are interested in the analysis of the disease progression, we separate the OC dataset into two interest groups based on the PFI feature, an indicator of disease progression. Concretely, PFI is defined as the time (in months) between the last cycle of platinum and evidence of disease progression [20]. In this setting, depending on the length of platinum drugs sensitivity, patients could be categorized as platinum resistant (<6 months) or platinum sensitive (>6 months). In the following, G_1 corresponds to the platinum sensitive group and G_2 to the platinum resistant group. After that, we could calculate an statistic of difference in conditional *pmf*s, $\Delta P(w_j^k) = P(w_j^k|G_1) - P(w_j^k|G_2)$. With this, we propose the following hypothesis test:

- Null hypothesis, $H_0 : \Delta P(w_j^k) = 0$, so that there is no difference between both groups for the word w_j^k.
- Alternative hypothesis, $H_A : \Delta P(w_j^k) \neq 0$, so that there is difference between both groups for the word w_j^k. If $\Delta P(w_j^k) > 0$ (if $\Delta P(w_j^k) < 0$), then the relative frequency of this word is larger in G_1 (G_2).

However, when there is some difference between both groups, we need to establish whether this difference is large enough to support statistical significance. To deal with this, we calculate an estimation of the *pdf* of $\Delta P(w_j^k)$, by employing a bootstrap resampling method. If the CI over the estimation of the *pdf* of $\Delta P(w_j^k)$ overlaps 0, we do not reject the null hypothesis, H_0. Nevertheless, if the CI over the estimation of the *pdf* of $\Delta P(w_j^k)$ does not overlap 0, we reject the null hypothesis, H_0, and accept the alternative hypothesis, H_A. With this, if the CI over the estimation is located at positive values, this word is a relevant property of G_1 (platinum sensitive group). Conversely, if the CI over the estimation is located at negative values, this word is relevant for G_2 (platinum resistant group).



(a)

(b)

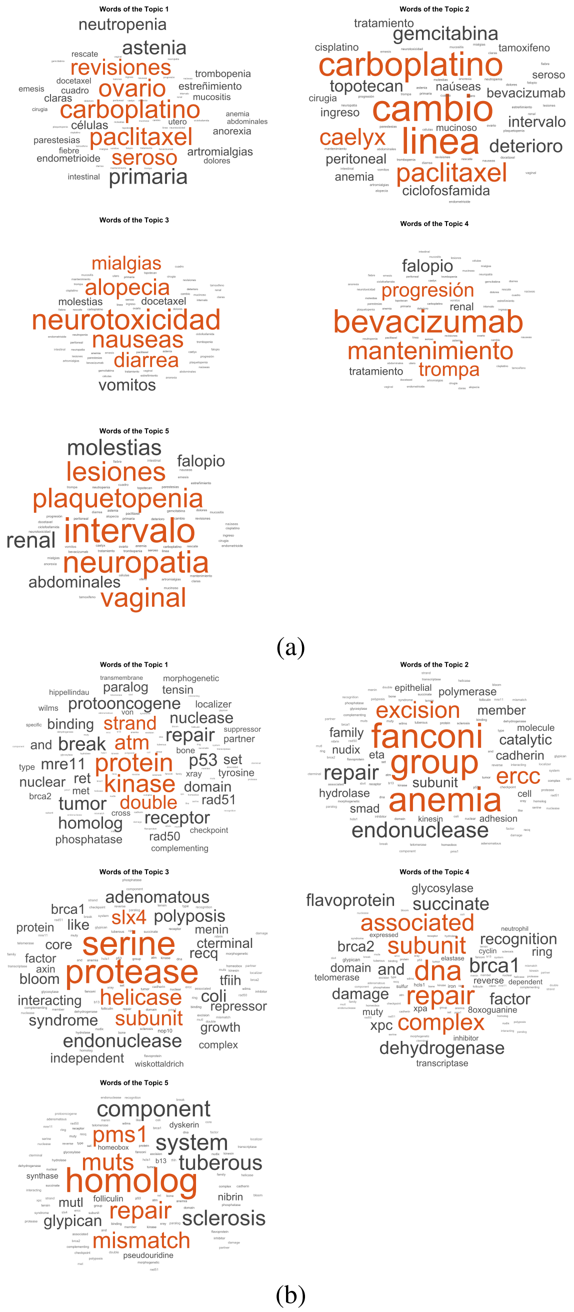
FIGURE 1. Bootstrap estimation of the *pdf* of each $\Delta P(w_j^k)$ in (a) clinical part and (b) genetic part of the OC dataset.

Results of applying the hypothesis test with a 99% confidence level both in the clinical and genetic part are shown in Fig. 1. The clinical part, displayed in Panel (a), reveals several significant words. Among them, the most relevant ones are *astenia* (asthenia in English) and *deterioro* (deterioration). These words are placed in the negative part of Δpdf, being their frequencies larger in the platinum resistant group, which is consistent with a clinical deterioration of the OC patients. For the genetic part, results showed in Panel (b) demonstrated an association with *serine proteases*, key enzymes in the extracellular matrix remodeling and consequently in the biological changes adopted by the tumoral cells to promote their metastatic properties.

Also, we applied the LDA method to the text features of both the clinical and genetic part of the OC dataset with the aim of discovering latent topics in the text fields and as a method to encode each document (each observation of the OC dataset) as a real-valued vector. This codification enables text features to be analysed together with categorical and metric features using multivariate methods like MCAPCA.

Therefore, in our case we consider the text fields of each observation both in the clinical and in the genetic part as a document d_i, following the LDA explanation in Subsection III-B. In order to discover latent topics in these text fields of the OC dataset, we fix the number of topics to discover in 5 ($K = 5$). To choose this number of topics, we have calculated the corpus topic probabilities of LDA models with different values of model order K, that is, the probabilities of observing each topic in the entire data set used to fit the LDA model for several K. With this, using bootstrap resampling, we were able to estimate the distribution for each topic of the corpus topic probabilities for different K. Using this distribution, we observed that, from $K = 5$ and for higher values, the bootstrap distributions of the last topics exhibit a bimodality with one mode corresponding to several 0 values (not shown), meaning that the last topics can be considered as empty.

The latent topics discovered can be shown using the word distributions per topics, $\boldsymbol{\phi}^{(k)}$, which can be used to create word cloud charts, where the more probability of occurrence of a specific word, the bigger and bolder it appears in the word cloud. Results can be seen in Fig. 2. For the clinical part, where the results are shown in the Panel (a), some words that stand out for each topic are: For Topic 1, *carboplatino* (carboplatin), *paclitaxel*, *ovario* (ovary), or *seroso* (serous); for Topic 2, *caelyx*; For Topic 3, *neurotoxicidad* (neurotoxicity), *naúseas* (nausea), *alopecia*, or *diarrea* (diarrhea); For Topic 4, *bevacizumab*, *mantenimiento* (maintenance), or *progresión* (progression); And for Topic 5, *lesiones* (injuries), *neuropatía* (neuropathy), or *intervalo* (interval). Otherwise, Panel (b) shows results for the genetic part, where some important words are: For Topic 1, *protein*, *kinase*, or *strand*; For Topic 2, *group*, *anemia*, or *Fanconi*; For Topic 3, *protease* or *serine*; For Topic 4, *DNA*, *repair*, or *subunit*; And for Topic 5, *homolog*, *repair*, or *mismatch*. Regarding the previously mentioned variables, both clinical and genetic factors represent text features previously described to be associated with the development of OC. Clinical variables included key terms related with either the treatment (carboplatin, paclitaxel, caelyx, bevacizumab) or the adverse effects of such drugs (alopecia, nauseas, or neurotoxicity, among others). Among the genetic factors significantly associated to the disease, we observed concepts linked to genes (*BRCA1, BRCA2, or Fanconi anemia* related genes), or DNA-repair pathways (HDR, homology directed repair or MR, mismatch repair) commonly altered in OC. In addition to these widely OC-related terms, a significant association was observed for other genes whose role in the development has been studied to a lesser extent (*SLX4* and *PMS1*). Therefore, the strategies



(a)

(b)

FIGURE 2. Word cloud charts of each of the 5 topics corresponding to (a) clinical part and (b) genetic part of the OC dataset.

considered are able to extract key terms involved in the pathogenesis of the tumor under study.

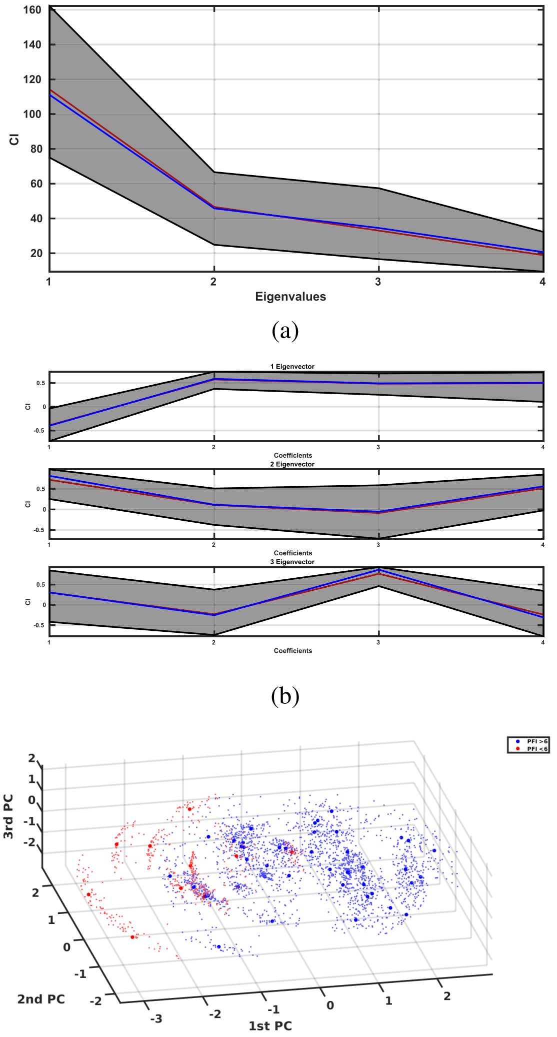
Besides, latent topics discovered can be used as a method to encode text where each observation of the OC dataset, called a document, is coded as a real-valued vector using the probability distribution over the topics in this document, $\boldsymbol{\theta}_i$. Thus, with $K = 5$, each observation in the clinical and genetic part of the OC dataset is encoded with a vector of real numbers of size 5. We use this codification to analyse together text with categorical and metric features using MCAPCA method. Results are presented in Subsection IV-D.

B. PCA ANALYSIS IN OC DATASET

In this experiment, we focus on the analysis of metric features of the OC dataset described in Section II. For this purpose, we use PCA method, which creates new uncorrelated features that are linear combinations of the original ones, providing an idea of: (1) Which new variables retain more information (through eigenvalues); And (2) which original variables are most relevant or influential in the creation of each new variable (through eigenvectors). We check how reliable is this information by calculating those CI with bootstrap resampling. Also, we can have a representation of the OC dataset in 3 dimensions, given by the most relevant eigendirections, thus providing us with information on the presence of natural and non-supervised patterns in the data.

In Fig. 3, we can see results of applying PCA to metric features of the clinical part of the OC dataset (Table 1). In Panel (a), we can see that the first eigenvalue is considerably greater than the others, which means that the direction of the first eigenvector retains much more variation of the data. In Panel (b), the first three eigenvectors are shown. We can see that, in the first eigenvector, all the components are relevant (none of the them overlap zero), meaning that original features that correspond to these components are influentially combined to form this new variable. These are the age at diagnosis (*Age_at_Diagnosis*), the progression free survival (*PFS*), which represents the time from the first date of pharmacological treatment until radiological or biochemical progression, the number of cycles of chemotherapy received in adjuvant therapy (*Cycles_of_Adjuvance*), and the overall survival (*OS*), which estimates the duration of patient survival from the date of diagnosis or treatment initiation. In the second eigenvector, the most important component is the first one, corresponding to the age at diagnosis, and, in the third eigenvector, the third, which is the cycles of chemotherapy in adjuvant therapy. This information is summarised in Table 3. In Panel (c), we can observe the projections, in 3 dimensions, of the clinical observations of the OC dataset (thick points) and their bootstrap resamples (thin points). In this case, as we are interested in the PFI as indicator of disease progression, we split OC dataset into platinum resistant patients (red points) and platinum sensitive patients (blue points) to try to find clusters which tell us if there is separability over the new features. It can be seen that both groups of OC patients are practically separated. In this regard, the ability of these variables to differentiate between responders vs. non responders is consistent with the expected, since the platinum-free interval positively correlates with features related to disease progression (overall survival and progression free survival) and with the number of platinum-based cycles that OC patients would receive.

In Fig. 4, we can observe, in this case, results of applying PCA to the metric variables of the genetic part of the OC dataset (Table 2). In Panel (a), it shows that the first two



(a)

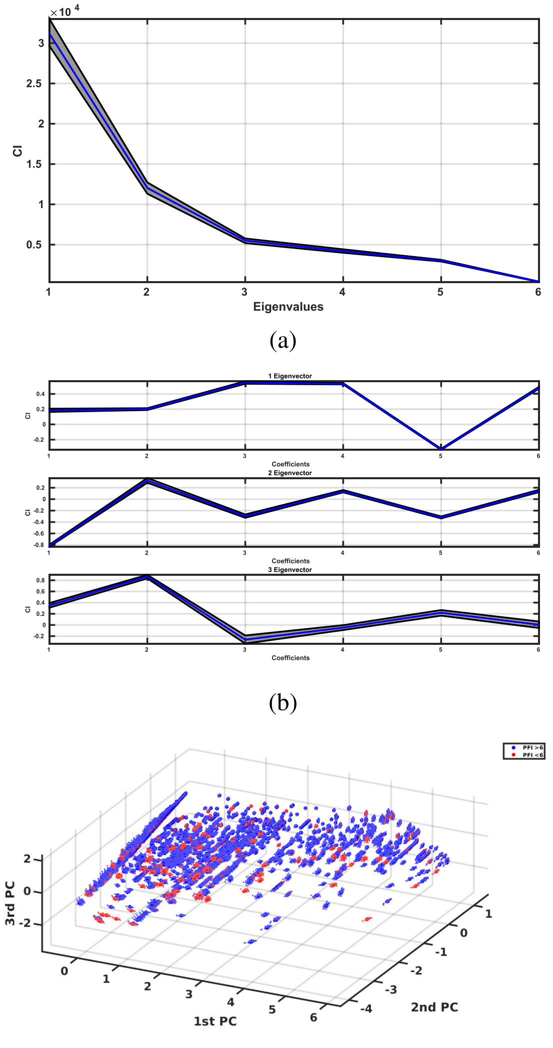
(b)

(c)

FIGURE 3. **Bootstrap estimation of the CI of (a) eigenvalues and (b) eigenvectors resulting of applying PCA to the clinical metric variables of the OC dataset. In (c), we can observe the 3-D projections of observations along with their bootstrap resamples after applying PCA to the clinical metric variables of the OC dataset.**

TABLE 3. **Relevant variables in each of the first three eigenvectors resulting from applying PCA to the clinical part of the OC dataset.**

1st Eigenvalue	2nd Eigenvalue	3rd Eigenvalue
Age_at_Diagnosis PFS Cycles_of_Adjuvance OS	Age_at_Diagnosis	Cycles_of_Adjuvance



(a)

(b)

(c)

FIGURE 4. Bootstrap estimation of the CI of (a) eigenvalues and (b) eigenvectors resulting of applying PCA to the genetic metric variables of the OC dataset. In (c), we can observe the 3-D projections of observations along with their bootstrap resamples after applying PCA to the genetic metric variables of the OC dataset.

TABLE 4. Relevant variables in each of the first three eigenvectors resulting from applying PCA to the genetic part of the OC dataset.

1st Eigenvalue	2nd Eigenvalue	3rd Eigenvalue
VarDepth	VarDepth	VarDepth
Conservation_Score	Conservation_Score	Conservation_Score
Grantham_Distance	Grantham_Distance	Grantham_Distance
Condel_Prediction_Score	Condel_Prediction_Score	Condel_Prediction_Score
Sift_Prediction_Score	Sift_Prediction_Score	Sift_Prediction_Score
PolyPhen_Prediction_Score	PolyPhen_Prediction_Score	PolyPhen_Prediction_Score

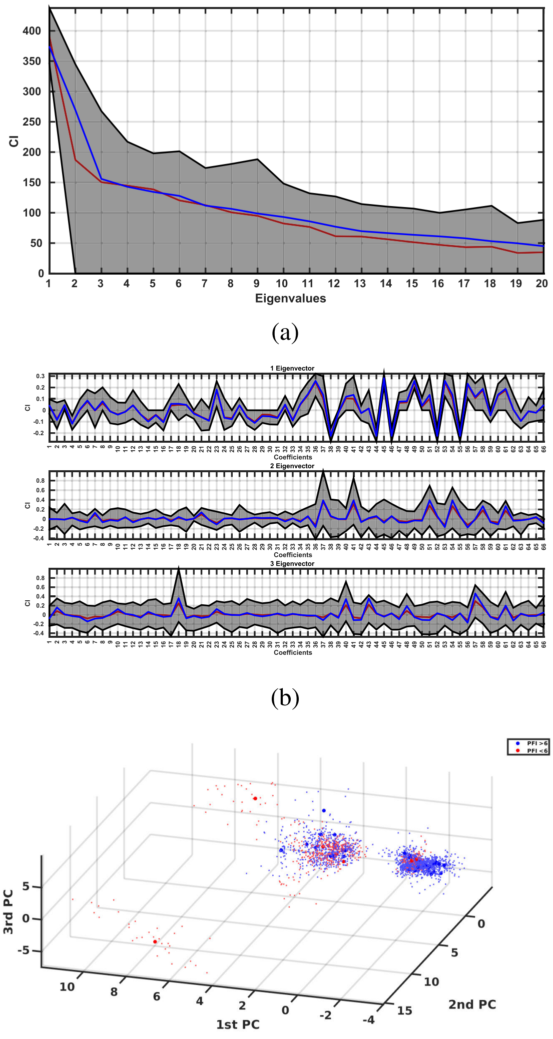
eigenvalues can be considered important, and that CI are much narrower than in the clinical part. In Panel (b), we can observe that, for the first three eigenvectors, all components are important, so that original features that correspond to these components are very influential in these three new variables. These original variables are: the number of sequencing reads which contains an specific variant allele (*VarDepth*); the conservation of the region under study at an evolutionary level (*Conservation_Score*) [21]; feature that reflects how different are the amino acids that are changed in missense mutations (*Grantham_Distance*) [22]; score of the degree of pathogenicity of the variant according to different *in silico* tools (*Condel_Prediction_Score*, *Sift_Prediction_Score* and *PolyPhen_Prediction_Score*) [23]–[25]. Table 4 exposes these information. In Panel (c), we can see the genetic observations of the OC dataset (thick points) and their bootstrap resamples (thin points) projected in 3 dimensions. It can be seen that platinum sensitive observations (blue points) and platinum resistant observations (red points) are very mixed.

In this setting, the terms included in these eigenvalues represent features directly associated with the relevance of nucleotides, and consequently of their corresponding amino acids, altered by the mutations under study. The incapacity of these genetic factors to discern between the two groups established on the basis of the response to platinum agents may be due, however, to the fact that to a great extent these metric values constitute *in silico* estimates. Therefore, the biological effect of these alterations in the endogenous activity of the protein and the role of such variants in OC pathogenesis have not been functionally validated. It is also necessary a greater concretion in the genetic datasets and deeper bioinformatic studies to obtain more solid results.

C. MCA ANALYSIS IN THE OC DATASET

In a similar way that we use PCA for metric variables, we use MCA for categorical variables. Likewise, new features obtained from applying MCA to the OC dataset are linear combinations of the original ones ranked in order of the amount of data variation retained. This allows us to know which original features are most important in each new variable and their reliability calculating the CI with bootstrap resampling. In addition to this, we can plot the projections of the OC dataset observations in 3 dimensions to get information on the presence of intrinsic patterns.

Results of applying MCA to the categorical features of the clinical part of the OC dataset (Table 1) are presented in Fig. 5. In this case, we represent the first 20 eigenvalues and the first three eigenvectors. In the first eigenvector, there are several components that are important (none of the them overlap zero), meaning that original feature categories that correspond to these components are combined in an important way to form this new variable. These are: In the information about the presence of gynecological cancer in family medical history (*Gynecological_Family_History*), *Yes* category; In the type of surgery (*Surgery*), both *Interval* and *Primary* categories; In the chemotherapy treatment prior to primary surgery (*Neoadjuvance*), *Yes* and *No* categories; In the observed response to neoadjuvant chemotherapy (*Response_of_Neoadjuvance*), *RC* category, meaning absence of all detectable cancer after treatment administration; In the decision after neoadjuvant treatment (*Attitude_of_Interval_Surgery*), *Yes* category, which implies the logical continuation of the disease treatment in

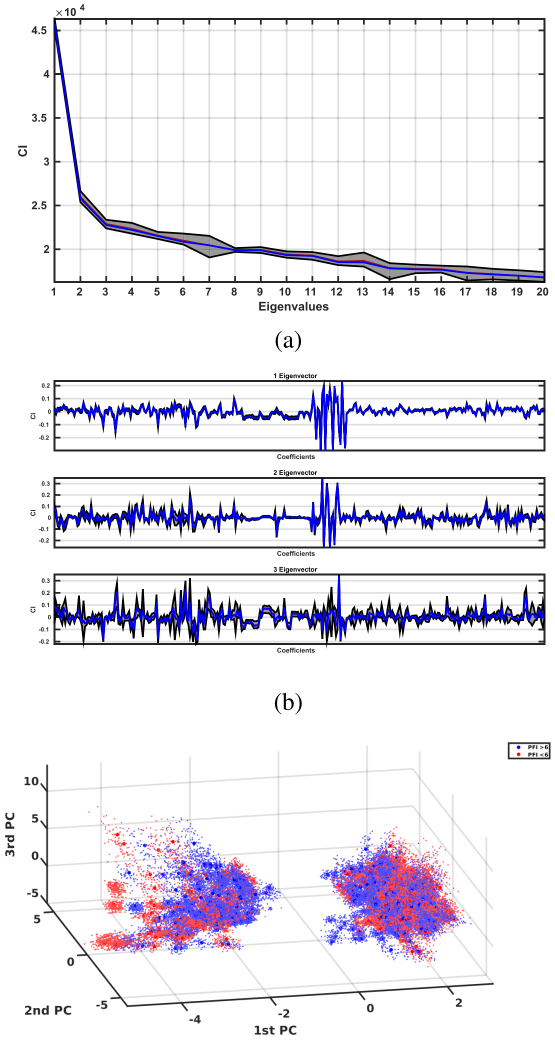


(a)

(b)

(c)

FIGURE 5. Bootstrap estimation of the CI of eigenvalues (a) and eigenvectors (b) resulting of applying MCA to the clinical categorical variables of the OC dataset. In (c), we can observe the 3-D projections of observations and their bootstrap resamples after applying MCA to the clinical categorical variables of the OC dataset.



(a)

(b)

(c)

FIGURE 6. Bootstrap estimation of the CI of (a) eigenvalues and (b) eigenvectors resulting of applying MCA to the genetic categorical variables of the OC dataset. In (c), we can observe the 3-D projections of observations and their bootstrap resamples after applying MCA to the genetic categorical variables of the OC dataset.

a two-step procedure (neoadyuvancy plus interval surgery); In the type of interval surgery (*Type_of_Interval_Surgery*), *R0* category, which means a complete resection of the tumor during surgical procedures; In the chemotherapy treatment after the primary surgery *Adjuvance*), *Yes* category; And in the observed response to adjuvant chemotherapy *Response_of_Adjuvance*), *RC category*, as in the neoadjuvant chemotherapy. The second and third eigenvector do not show relevant features categories. In this sense, all the terms considered relevant according to the first eigenvector are related to the surgical procedures and the type of treatments received by OC patients, which have been largely correlated with the duration and degree of response of OC patients to platinum-based drugs.

Likewise, MCA results of the genetic part of the OC dataset (Table 2) are displayed in Fig. 6. For the first eigenvector, relevant feature categories are: *GPC3*, *MSH2*, and *TSC1* genes (*HGNC_Symbol*); In genotype (*Genotype*), *P_Homo_ref*, and *UNC_Hetero* categories; Neutral or tolerated categories for *in silico* tools predicting the pathogenicity of the considered variants (*Condel_Prediction*, *Sift_Prediction*, and *PolyPhen_Prediction*) [23]–[25]. For the second eigenvector, we find these relevant feature categories: *chr5* chromosome (*Chr*); *Deleterious* and *Neutral* in *Condel_Prediction*; *Deleterious* and *Tolerated* in *Sift_Prediction*; And *Benign* and *Probably damaging* in *PolyPhen_Prediction*. For the third eigenvector, the following features categories stand out: *MSH6* gene; *Chr2* chromosome; Genetic changes from the reference allele to the variant allele (*Genetic_Change*) represented by $AC > A$, $GTAAAAAAA > GAAAA$, and $T > TTCTC$; And *Low* category in *IMPACT*. Both Table 5 and Table 6 expose each one of these features categories for clinical and genetic, respectively.

TABLE 5. Relevant categories of variables in each of the first three eigenvectors resulting from applying MCA to the clinical part of the OC dataset.

1st Eigenvalue	2nd Eigenvalue	3rd Eigenvalue
Gynecological_Family_History: Yes Surgery: Interval Surgery: Primary Neoadjuvance: No Neoadjuvance: Yes Response_of_Neoadjuvance: RC Attitude_of_Interval_Surgery: Yes Type_of_Interval_Surgery: R0 Adjuvance: Yes Response_of_Adjuvance: RC		

TABLE 6. Relevant categories of variables in each of the first three eigenvectors resulting from applying MCA to the genetic part of the OC dataset.

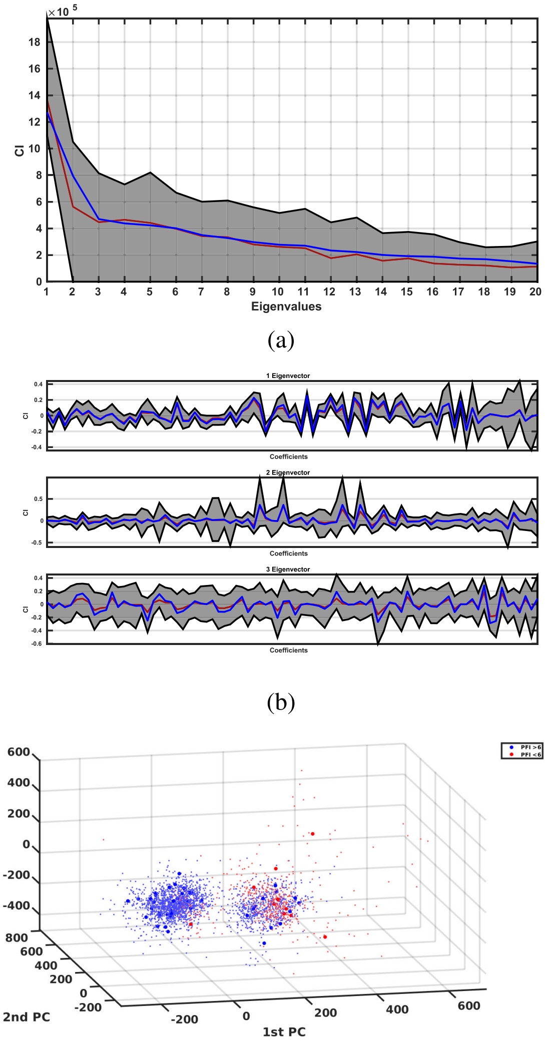
1st Eigenvalue	2nd Eigenvalue	3rd Eigenvalue
HGNC_Symbol: GPC3 HGNC_Symbol: MSH2 HGNC_Symbol: TSC1 Genotype: P_Homo_ref Genotype: UNC_Hetero Condel_Prediction: Neutral Sift_Prediction: Tolerated PolyPhen_Prediction: Benign IMPACT: Low IMPACT: Moderate	Chr: chr5 Condel_Prediction: Deleterious Condel_Prediction: Neutral Sift_Prediction: Deleterious Sift_Prediction: Tolerated PolyPhen_Prediction: Benign PolyPhen_Prediction: Probably damaging	HGNC_Symbol: MSH6 Chr: chr2 Genetic_Change: AC>A Genetic_Change: GTAAAAAAA>GAAAA Genetic_Change: T>TTCTC IMPACT: Low

In Panel (c) of both Fig. 5 and Fig. 6, we can observe the clinical and genetic projections of the OC dataset observations in 3 dimensions (thick points) and their bootstrap resamples (thin points) for platinum sensitive (blue points) and platinum resistant (red points) groups. For the clinical part, a certain separability between groups can be observed. However, for the genetic part, there is not a apparent pattern of segregation. Interestingly, some of the categorical genetic features included genes involved in tumor, either in general, or in ovarian carcinoma progression (for example, *MSH2 and MSH6*), as well as certain categories of several *in silico* tools (Sift, PolyPhen, or Condel) able to predict the pathogenicity of single nucleotide variants. In addition, our strategy allowed us to detect significant association with less-studied genetic variables, as it is the case of *GPC3* gene, which has been suggested to modulate the clinical response of ovarian clear cell carcinomas to standard therapies.

However, it seems recommendable to establish new analytical tools and to expand to better consolidated genetic datasets in order to continue in this path and to identify new variables behaving as efficient predictive or prognosis biomarkers.

D. MCAPCA ANALYSIS IN OC DATASET

Likewise we use PCA in metric variables and MCA in categorical variables, we use the new method MCAPCA in both types of variables. Furthermore, we can analyse text variables together with categorical and metric using the LDA text encoding method in which each observation is encoded, in our case, in a real-valued vector of dimension 5 (number of latent topics), using the probability distribution over the topics in this document, $\boldsymbol{\theta}_i$, calculated in Subsection IV-A.



(a)

(b)

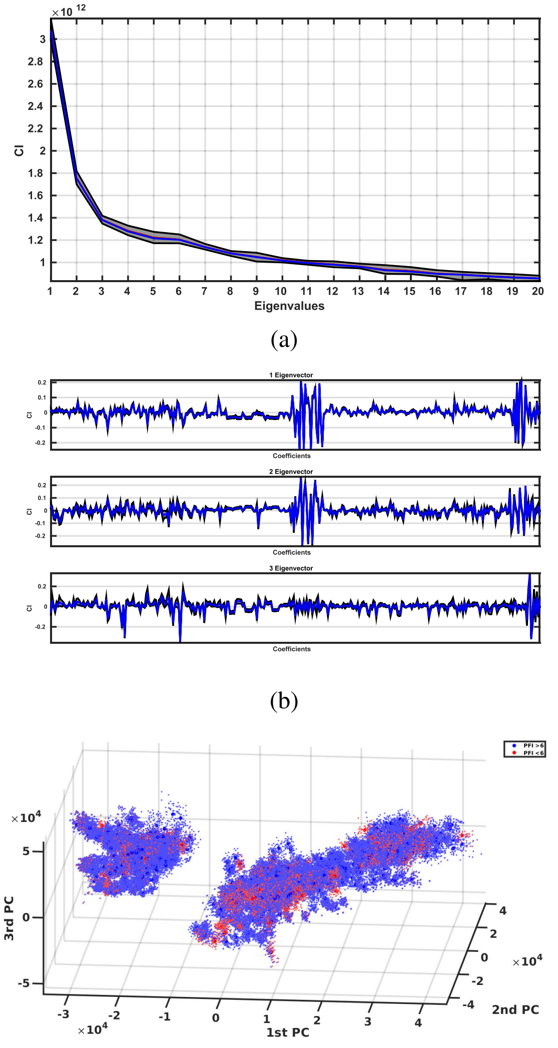
(c)

FIGURE 7. Bootstrap estimation of the CI of (a) eigenvalues and (b) eigenvectors resulting of applying MCAPCA to the clinical metric and categorical variables of the OC dataset. In (c), we can observe the 3-D projections of observations and their bootstrap resamples after applying MCAPCA to the clinical metric and categorical variables of the OC dataset.

Results obtained after applying MCAPCA to the clinical part of the OC dataset (Table 1) are shown in Fig. 7. We show the first 20 eigenvalues and the first three eigenvectors. In the clinical part, results are the same as those obtained with MCA. The only difference is that, for the first eigenvector, the numerical variables selected with PCA also appear. In this case, the MCAPCA method behaves as the union of PCA and MCA, which is just what we want to achieve, since this could facilitate an analysis of relationships between numerical and categorical variables.

In the same way, MCAPCA results of the genetic part of the OC dataset (Table 2) are displayed in Fig. 8. As in the clinical part, the method behaves like the union of MCA and PCA, since the results are the same as each of the PCA and MCA methods separately. These results are listed in Table 7 and Table 8.

In Panel (c) of Fig. 7 and Fig. 8, we can observe the clinical and genetic projections of the OC dataset observations in



(a)

(b)

(c)

FIGURE 8. Bootstrap estimation of the CI of (a) eigenvalues and (b) eigenvectors resulting of applying MCAPCA to the genetic metric and categorical variables of the OC dataset. In (c), we can observe the 3-D projections of observations and their bootstrap resamples after applying MCAPCA to the genetic metric and categorical variables of the OC dataset.

TABLE 7. Relevant categories of variables in each of the first three eigenvectors resulting from applying MCAPCA to the clinical part of the OC dataset.

1st Eigenvalue	2nd Eigenvalue	3rd Eigenvalue
Gynecological_Family_History: Yes		
Surgery: Interval		
Surgery: Primary		
Neoadjuvance: No		
Neoadjuvance: Yes		
Response_of_Neoadjuvance: RC		
Attitude_of_Interval_Surgery: Yes		
Type_of_Interval_Surgery: R0		
Adjuvance: Yes		
Response_of_Adjuvance: RC		
Age_at_Diagnosis		
PFS<		
Cycles_of_Adjuvance		
OS		

TABLE 8. Relevant categories of variables in each of the first three eigenvectors resulting from applying MCAPCA to the genetic part of the OC dataset.

1st Eigenvalue	2nd Eigenvalue	3rd Eigenvalue
HGNC_Symbol: GPC3	Chr: chr5	HGNC_Symbol: MSH6
HGNC_Symbol: MSH2	Condel_Prediction: Deleterious	Chr: chr2
HGNC_Symbol: TSC1	Condel_Prediction: Neutral	Genetic_Change: AC>A
Genotype: P_Homo_ref	Sift_Prediction: Deleterious	Genetic_Change: GTAAAAAAA>GAAAA
Genotype: UNC_Hetero	Sift_Prediction: Tolerated	Genetic_Change: T>TTCTC
Condel_Prediction: Neutral	PolyPhen_Prediction: Benign	IMPACT: Low
Sift_Prediction: Tolerated	PolyPhen_Prediction: Probably damaging	Conservation_Score
PolyPhen_Prediction: Benign	Conservation_Score	Grantham_Distance
IMPACT: Low	Grantham_Distance	Condel_Prediction_Score
IMPACT: Moderate	Condel_Prediction_Score	Sift_Prediction_Score
Conservation_Score	Sift_Prediction_Score	PolyPhen_Prediction_Score
Grantham_Distance	PolyPhen_Prediction_Score	
Condel_Prediction_Score		
Sift_Prediction_Score		
PolyPhen_Prediction_Score		

3 dimensions (thick points) and their bootstrap resamples (thin points) for platinum sensitive (blue points) and platinum resistant (red points) groups. As was the case for PCA and MCA, a certain pattern of separability can be seen for the clinical part, something that does not happen in the genetic part.

V. DISCUSSION AND CONCLUSION

OC represents a serious health problem since is the second most common gynecological neoplasm, with an estimated annual incidence of 225 000 women worldwide [26]. It is also the fifth cause of cancer associated mortality and the gynecological tumor with the worst prognosis (140 000 deaths per year), with a 5-year overall survival close to 15%. This problem is marked by the lack of robust predictive and prognostic molecular biomarkers underpining *a priori* knowledge of the evolution of the disease [2].

The most well-known genetic factors associated to the development of OC is the presence of pathogenic mutations in genes involved in the DNA damage repair by homologous recombination (HR; >60% of high grade serous OC patients) and more specifically, *BRCA1* and *BRCA2 loci*. It is worth highlighting that loss of function alterations in both genes significantly correlates with a better therapeutic response to conventional chemotherapies (platinum-based agents) and personalized treatments which cause increased cell DNA damage (poly ADP ribose polymerase (PARP) inhibitors). These findings led to the definition of a *BRCAness* phenotype, which include patients presenting uncommon pathogenic variants in other HR-related genes different than *BRCA1* and *BRCA2*, but developing a clinical course similar to cases carrying alterations in these *BRCA* genes. As a consequence, multiple experimental approaches (loss of heterozigosity, genomic scars of MyChoice platform (Myriad), mutational signature 3, among others) [27]–[29] have been implemented in recent years to indirectly estimate the cellular activity of the HR pathway and therefore define subsets of OC patients who could respond more efficiently to the current therapies. Despite its applicability, limited genetics-clinical correlations have been described on basis to HR pathway activity since clinical benefit is observed regardless of the *BRCA* status. Therefore, it is highly desirable the adoption of experimental approaches that include the integration of -omics data massively obtained from tumor samples under study with clinical information of interest.

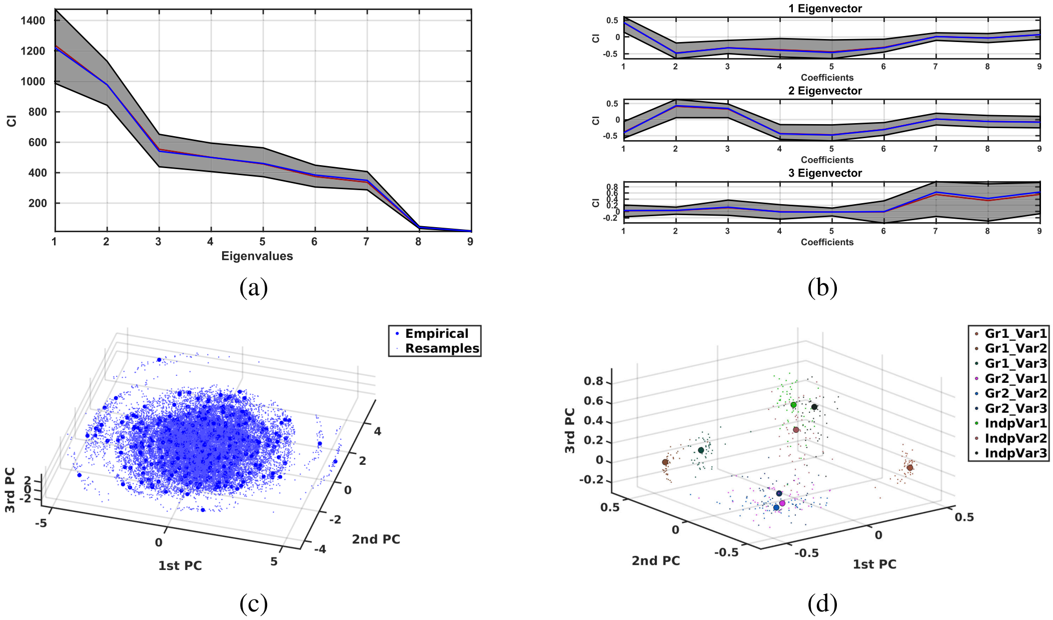


FIGURE 9. Bootstrap estimation of the CI of (a) eigenvalues and (b) the first three eigenvectors resulting of applying PCA to a simple example with synthetic data. In (c) and (d), we can observe the 3-D projections of observations and variables along with their resamples after applying PCA to simple example with synthetic data.

In the text analysis, a method for text type applying bootstrap resampling to BoW has been used in order to inspect the free-text fields present in the clinical and genetic datasets available for our OC cohort. In parallel, LDA method has been applied to discover latent topics in the text fields of such dataset. Given that both strategies were able to detect both clinical and genetic differences according to the degree of therapeutic response to platinum drugs (PFI resistant, $<$6 months vs. PFI sensitive, $>$6 months), we strongly consider that our OC datasets are discriminatory for text features. As expected, these methods showed significant association with topics widely associated to the development of OC, such as key terms related with the treatment, the adverse effects of such therapies and the most frequently mutated genes and altered molecular pathways (*BRCA loci*, DNA-repair pathways, or Mismatch Repair, among others). In addition, significant associations were detected with other variables whose role in the etiology of ovarian cancer has been studied to a much lesser extent, such as the presence of alterations in the *SLX4* and *PMS1* genes. Regarding *SLX4*, rare loss-of-fuction variants have been described to contribute to the development of non BRCA mutated gynecologycal carcinomas, both ovarian and breast tumors [30], [31]. Significant association detected for *PMS1 locus* could be explained by the role of this factor in the DNA mismatch repair pathway and its involvement in the pathogenesis of hereditary tumors. Consequently, these methods were able to extract both well known OC-related terms as well as low-risk OC concepts involved in the ethiology of the tumor under study, reinforcing the fortress of our approach.

In the linear multidimensional analyses, we have presented a framework consisting of the application of bootstrap resampling to PCA and MCA, and we have also contributed with MCAPCA, a new method that we propose to analyse mixed features. This framework has helped us to explore both the clinical part and the genetic part of the available OC dataset. In detail, we have used this framework as a feature extraction method to create a new set of features that captures most of the useful information contained in the initial set of variables of the OC dataset. Interpreting this new set of features, we have an idea of which original variables are most relevant. In addition, representation of these new features in 2 or 3 dimensions provides us with information about the presence of patterns in the data. Therefore, in this work we do not use feature selection or supervised methods, since our idea is to perform a linear multivariate exploration of the OC dataset just to try to understand the features and their interactions in relation to the OC disease progression in the feature extraction stage. This type of data interpretability analysis is often ignored, and researchers often move directly to a supervised analysis, which could result in missing relevant information provided by existing multivariate and nonsupervised feature extraction methods. The results obtained showed us that the clinical data has a certain pattern of separability for platinum resistant and platinum sensitive groups in the three methods. There results are in line with the predictive and prognostic value of certain widely studied clinical variables (family history of gynecological cancer, presence of residual disease after surgery, or degree of response to adjuvant therapy).

Additionally, this may be a indicator of success in using the new set of features obtained for each method to predict disease progression.

Despite that such pattern of separability does not appear so evident for the genetic part, different terms were suggested to behave as potential predictors of the response to platinum-drugs. This is the case of different mismatch repair genes (*MSH2* and *MSH6*) and the scores/categories of *in silico* pathogenicity predictors. Worth of mention, *GPC3* gene, which has been described to modulate the therapeutic response of clear cell ovarian carcinomas [32], [33]), showed significant association despite of the reduced frequency of such histology in our cohort. The lack of power to detect additional correlations could be explained by the fact of having an over-stratified genetic dataset, which prevents the extraction by the analytical methods used in this study of terms relevant for the development of OC.

Clearly, these analytical tools should be able to recognize genetic factors widely described to participate in OC oncogenesis, such as *BRCA1* and *BRCA2* genes. It should also be noted that ovarian cancer is an aetiologically complex disease, whose development is also determined by other molecular alterations studied by other -omics approaches (methylation or proteomic profilings, among others). Therefore, the integration of results obtained by multiple -omics molecular profiling could be the ideal scenario for an efficient discrimination of OC cases with the most favorable prognosis. Finally, it is also necessary to subsequently explore the prediction information provided by existing non-linear and supervised methods to fully evaluate the information coveyed by the analyzed dataset.

APPENDIX

SYNTHETIC EXAMPLES

A. PCA SYNTHETIC EXAMPLE

To show how PCA works with bootstrap resampling, we implement a simple example. We first generate a synthetic dataset composed of 500 individuals and 9 metric variables, in which there are 2 groups of 3 variables that are dependent on each other, and 3 variables that are completely independent. We apply PCA on this dataset, and also on each bootstrap resample. In this way, we can estimate the CI of eigenvalues and each component of eigenvectors, and make an analysis of them.

Results of the simple example with synthetic data presented above are shown in Fig. 9. In Panel (a) are represented the CI of eigenvalues in gray, the mean value depicted with a red line, and the empirical value with a blue line. We can see that the first two eigenvalues are much greater than the others, meaning that the direction of the corresponding eigenvectors retains much more variation of the data. In Panel (b) are shown the CI of each component of the first three eigenvectors in gray, the mean value with a red line, and the empirical value with a blue line. We can check that, for the first eigenvector (which represents the first new variable), components that have most importance are all except the seventh, eighth, and ninth, meaning that these original variables are the ones that provide the most information for this new variable. In the second eigenvector (which represents the second new variable), the most important components are also the first six, which means that these are the original variables which provide the most information to this new variable. In the same way, in the third eigenvector (which represents the third new variable), there are no important variables. In Panel (c) are depicted, in 3 dimensions, the projections of each observation of the dataset (big blue points) and its resamples (small blue points). Also, in Panel (d) are plotted the original variables in 3 dimensions.

B. MCA SYNTHETIC EXAMPLE

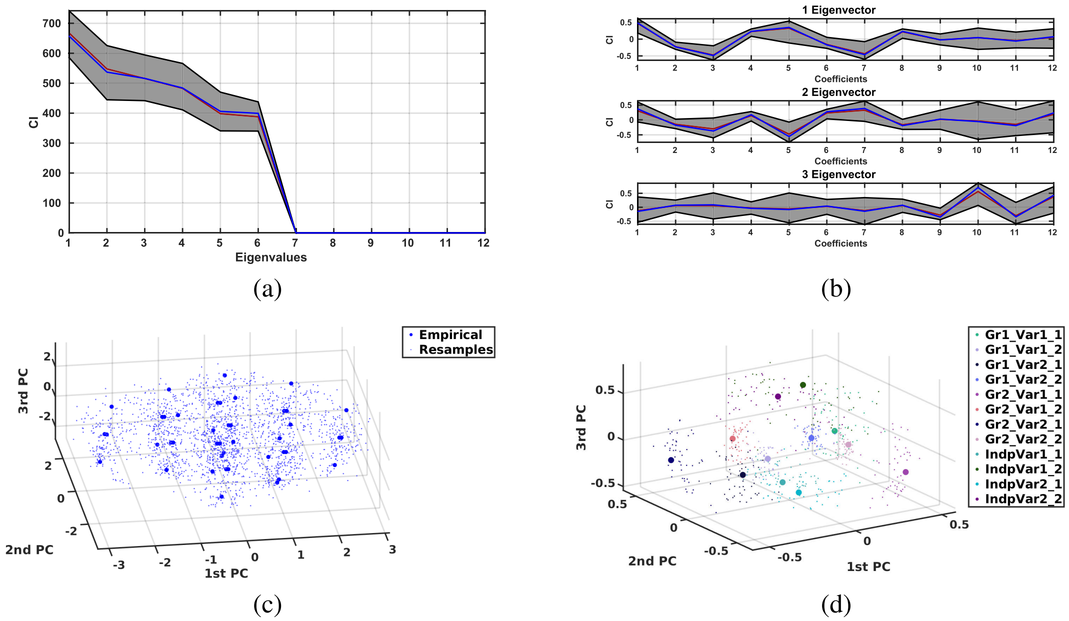
We present a simple example with the aim of enlightening the functioning of MCA method with bootstrap resampling. We create a synthetic dataset of 500 individuals and 6 categorical variables, each with 2 categories. This generates an indicator matrix of 500 × 12 dimensions. These variables are 2 groups of 2 dependent variables, and 2 completely independent variables. With this synthetic dataset, we calculate MCA on it and also on each resample of it. Thus, we can estimate the CI of eigenvalues and eigenvectors.

The simple example described above has the results shown in Fig. 10. We can see in Panel (a) how eigenvalues from the sixth position have values of zero, meaning that the direction of the corresponding eigenvectors do not retain any variation of the data, and therefore any information, and they could be dispensed with. In Panel (b) are presented the first three eigenvectors with their CI. In the first eigenvector, which represents the first new variable, components that have most importance are the first four and the seventh and eighth, so these original variables are the ones that provide the most information for this new variable. In the second eigenvector, the most important component is the sixth, and in the third eigenvector, the ninth. In Panel (c) are plotted the projections of each observation of the dataset (big blue points) and its resamples (small blue points). Additionally, in Panel (d) are displayed the representations of the original variables in 3 dimensions.

C. MCAPCA SYNTHETIC EXAMPLE

We develop a simple example in order to show how MCAPCA works with bootstrap resampling. To this effect, we generate a synthetic dataset of 500 individuals and 6 mixed variables. There are one group of 2 dependent metric variables, another group of 2 dependent categorical variables with 2 categories each one, and 2 completely independent variables, one metric and one categorical (with 2 categories, too). With this synthetic dataset, we apply MCAPCA on it and also on each resample of it, being able to estimate the CI of eigenvalues and eigenvectors.

Results obtained in the simple example described are displayed in Fig. 11. In Panel (a), we can observed how eigenvalues from the seventh position have values of zero, so we can get rid of them due to the direction of the corresponding



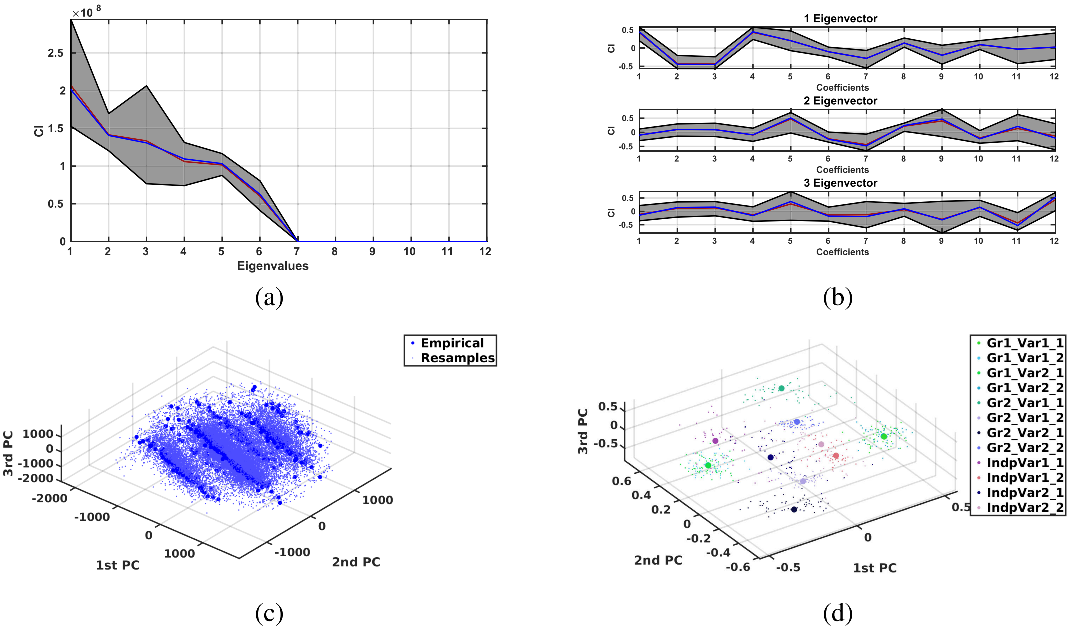
(a)

(b)

(c)

(d)

FIGURE 10. **Bootstrap estimation of the CI of (a) eigenvalues and the (b) first three eigenvectors resulting of applying MCA to a simple example with synthetic data. In (c) and (d), we can observe the 3-D projections of observations and variables along with their resamples after applying MCA to simple example with synthetic data.**



(a)

(b)

(c)

(d)

FIGURE 11. **Bootstrap estimation of the CI of (a) eigenvalues and (b) the first three eigenvectors resulting of applying MCAPCA to a simple example with synthetic data. In (c) and (d), we can observe the 3-D projections of observations and variables along with their resamples after applying MCAPCA to simple example with synthetic data.**

eigenvectors do not retain any variation of the data. In Panel (b) are presented the CI of each component of the first three eigenvectors. In the first eigenvector, the first four components and the eighth have importance, so these original variables are the ones that provide the most information for this new variable. In the second eigenvector, the most influential components are sixth, seventh, and eighth, and in the third eigenvector, the last two components. In Panel (c) are represented the projections of each observation of the dataset (big blue points) and its resamples (small blue points) using the first three new variables generated with MCAPCA as coordinates. Moreover, representations of the original variables in 3 dimensions are displayed in Panel (d).

AUTHOR CONTRIBUTIONS

L. B.-C., S. M.-R., and J. L. R.-Á. designed and organized the article. L. B.-C. performed the data analysis and wrote the methods and part of the introduction, results, and conclusion. S. R.-L. wrote the dataset description and part of the introduction, results, and conclusion. S. M.-R., J. G.-D, and J. L. R.-Á. contributed writing some parts and reviewing the manuscript. M. Y.-F. and A. B. provided the data processing. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

Authors declare no conflict of interest.

REFERENCES

[1] J. Millstein, T. Budden, E. L. Goode, and M. S. Anglesio, "Prognostic gene expression signature for high-grade serous ovarian cancer," *Ann. Oncol.*, vol. 31, no. 9, pp. 1240–1250, 2020.

[2] C. Stewart, C. Ralyea, and S. Lockwood, "Ovarian cancer: An integrated review," *Seminars Oncol. Nursing*, vol. 35, no. 2, pp. 151–156, Apr. 2019.

[3] L. Bote-Curiel, S. Ruiz-Llorente, S. Munoz-Romero, M. Yague-Fernandez, A. Barquin, J. Garcia-Donas, and J. L. Rojo-Alvarez, "A resampling univariate analysis approach to ovarian cancer from clinical and genetic data," *IEEE Access*, vol. 9, pp. 25959–25972, 2021.

[4] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, nos. 1–4, pp. 43–52, Dec. 2010.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[6] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," in *Proc. Adv. Neural Inf. Process. Syst.*, Cambridge, MA, USA: MIT Press, 2006, pp. 1353–1360.

[7] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, Apr. 2004.

[8] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[9] A. Fisher, B. Caffo, B. Schwartz, and V. Zipunnikov, "Fast, exact bootstrap principal component analysis for p > 1 million," *J. Amer. Stat. Assoc.*, vol. 111, no. 514, pp. 846–860, 2016.

[10] S. Muñoz-Romero, "Análisis multivariante: Soluciones eficientes e interpretables," Ph.D. dissertation, Dept. Signal Theory Commun., Universidad Carlos III de Madrid, Getafe, Spain, May 2015.

[11] H. Babamoradi, V. Frans, and R. Åsmund, "Bootstrap based confidence limits in principal component analysis—A case study," *Chemometrics Intell. Lab. Syst.*, vol. 120, pp. 97–105, Jan. 2013.

[12] A. Zabala and U. Pascual, "Bootstrapping Q methodology to improve the understanding of human perspectives," *PLoS ONE*, vol. 11, no. 2, 2016, Art. no. e0148087.

[13] D. Corral-De-Witt, E. V. Carrera, S. Muñoz-Romero, K. Tepe, and J. L. Rojo-Álvarez, "Multiple correspondence analysis of emergencies attended by integrated security services," *Appl. Sci.*, vol. 9, no. 7, p. 1396, 2019.

[14] G. Michailidis, *Correspondence Analysis*. Thousand Oaks, CA, USA: SAGE Publications, 2007, pp. 191–193.

[15] O. Nenadic and M. Greenacre, "Correspondence analysis in R, with two- and three-dimensional graphics: The ca package," *J. Stat. Softw.*, vol. 20, no. 3, pp. 1–13, 2007.

[16] H. Abdi and L. Williams, *Correspondence Analysis*. Thousand Oaks, CA, USA: SAGE Publications, 2010, pp. 267–278.

[17] J. C. Gower and D. J. Hand, *Biplots*. Boca Raton, FL, USA: Chapman & Hall, 1st ed., 1995.

[18] B. Escofier, "Traitement simultané de variables qualitatives et quantitatives en analyse factorielle," *Cahiers de l'analyse des données*, vol. 4, no. 2, pp. 137–146, 1979.

[19] H. Abdi, L. J. Williams, and D. Valentin, "Multiple factor analysis: Principal component analysis for multitable and multiblock data sets," *WIREs Comput. Statist.*, vol. 5, no. 2, pp. 149–179, 2013.

[20] E. Pujade-Lauraine and P. Combe, "Recurrent ovarian cancer," *Ann. Oncol.*, vol. 27, pp. i63–i65, Sep. 2016.

[21] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using GERP++," *PLOS Comput. Biol.*, vol. 6, no. 12, pp. 1–13, 2010.

[22] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, vol. 185, no. 4154, pp. 862–864, 1974.

[23] A. González-Pérez and N. López-Bigas, "Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel," *Amer. J. human Genet.*, vol. 88, no. 4, pp. 440–449, 2011.

[24] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nature Protocols*, vol. 4, no. 7, pp. 1073–1081, Jul. 2009.

[25] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010.

[26] M. Moschetta, A. George, S. B. Kaye, and S. Banerjee, "BRCA somatic mutations and epigenetic BRCA modifications in serous ovarian cancer," *Ann. Oncol.*, vol. 27, no. 8, pp. 1449–1455, Aug. 2016.

[27] M. L. Telli *et al.*, "Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer," *Clin. Cancer Res.*, vol. 22, no. 15, pp. 3764–3773, Aug. 2016.

[28] D. C. Gulhan, J. J.-K. Lee, G. E. M. Melloni, I. Cortés-Ciriano, and P. J. Park, "Detecting the mutational signature of homologous recombination deficiency in clinical samples," *Nature Genet.*, vol. 51, no. 5, pp. 912–919, May 2019.

[29] E. M. Swisher *et al.*, "Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma (ARIEL2—Part 1): An international, multicentre, open-label, phase 2 trial," *Lancet Oncol.*, vol. 18, no. 1, pp. 75–87, Jan. 2017.

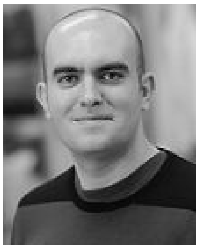
[30] H. Song *et al.*, "Population-based targeted sequencing of 54 candidate genes identifies PALB2 as a susceptibility gene for high-grade serous ovarian cancer," *J. Med. Genet.*, May 2020, doi: 10.1136/jmedgenet-2019-106739.

[31] S. Shah, Y. Kim, I. Ostrovnaya, R. Murali, K. A. Schrader, F. P. Lach, K. Sarrel, R. Rau-Murthy, N. Hansen, L. Zhang, T. Kirchhoff, Z. Stadler, M. Robson, J. Vijai, K. Offit, and A. Smogorzewska, "Assessment of slx4 mutations in hereditary breast cancers," *PLoS ONE*, vol. 8, no. 6, pp. 1–5, 2013.

[32] C. Luo, K. Shibata, S. Suzuki, H. Kajiyama, T. Senga, Y. Koya, M. Daimon, M. Yamashita, and F. Kikkawa, "GPC3 expression in mouse ovarian cancer induces GPC3-specific t cell-mediated immune response through m1 macrophages and suppresses tumor growth," *Oncol. Rep.*, vol. 32, no. 3, pp. 913–921, Sep. 2014.

[33] K. Wiedemeyer, M. Köbel, H. Koelkebeck, Z. Xiao, and K. Vashisht, "High glypican-3 expression characterizes a distinct subset of ovarian clear cell carcinomas in Canadian patients: An opportunity for targeted therapy," *Human Pathol.*, vol. 98, pp. 56–63, Apr. 2020.

LUIS BOTE-CURIEL received the degree in telecommunication engineering from the Universidad de Valladolid, Spain, in 2012. He is interested in machine learning and deep learning, focused on interpretability and their application to the healthcare field.



SERGIO RUIZ-LLORENTE received the degree in biology from the Universidad de Alcalá, in 2000, and the Ph.D. degree in human genetics at the Spanish National Cancer Center (CNIO), in 2005.

Later on, he worked with different national and international research centers, including the Instituto de Investigaciones Biomédicas-UAM, Memorial Sloan Kettering Cancer Center, and HM-CIOCC. He has probed experience in the use of genetic diagnostic tools, the applicability and management of high throughput molecular platforms, and the development of *in vivo* and *in vitro* preclinical assays. This research activity has resulted in the publication of 25 scientific articles in international journals, seven of them as the first author.

SERGIO MUÑOZ-ROMERO received the degree in engineering and the Ph.D. degree in machine learning from the Universidad Carlos III de Madrid, Spain, in 2009 and 2015, respectively. His current research interests include machine learning algorithms and statistical learning theory, mainly dimensionality reduction and feature selection methods, for real-world problems, especially, for aging and oncology.

MÓNICA YAGÜE-FERNÁNDEZ received the bachelor's degree in biotechnology and the master's degree in clinical and applied research in oncology from Universidad CEU San Pablo, in 2018 and 2019, respectively, and the master's degree in administration and management of pharmaceutical, biotechnological, and health companies from CESIF, in 2020. She is developing her career and working at the Clara Campal Comprehensive Oncology Center (CIOCC), focusing on gynecological, genitourinary, and skin tumors from a clinical and basic point of view, developing clinical trials, and research studies.

ARANTZAZU BARQUÍN received the master's degree in medical oncology sponsored by the Spanish Medical Oncologist Association (SEOM) and in oncology molecular biology sponsored by the National Centre of Oncological Investigation (CNIO). She finished her specialization in medical oncology in 2019. She successfully passed ESMO examination in 2017. She has completed a short rotation at the Princess Margaret Cancer Center, Toronto, Canada, in 2018. She is affiliation with the Centro Oncologico Clara Campal, Spain, specialized in the Gynaecological, Genitourinary, and Skin Tumour Unit, and an Active Member of the Laboratory of Translational and Innovation in Oncology. She is currently focusing her investigation in ovarian cancer.

JESÚS GARCÍA-DONAS is currently a Medical Oncologist and the Head of the Genitourinary, Gynecological, Skin and Rare Tumors Unit, Clara Campal Comprehensive Cancer Center. With extensive experience in the realization of clinical trials, he has participated in more than 200 studies and also designed and directed multiple clinical trials without commercial interest whose promoters have been cooperative groups. The group he leads has published more than 40 articles in international scientific journals of first level, and has won awards in recognition of their activity (Merck Foundation awards to the Research in Rare Diseases) and public and private fellowships.

JOSÉ LUIS ROJO-ÁLVAREZ (Senior Member, IEEE) received the B.Sc. degree in telecommunication engineering from the Universidade de Vigo, in 1996, and the Ph.D. degree in telecommunication engineering from the Universidad Politécnica de Madrid, in 2000. He is currently a Professor with the Department of Signal Theory and Communications, Universidad Rey Juan Carlos, Spain. He has coauthored more than 140 international articles and has contributed to more than 180 conference proceedings. His research interests focus on statistical learning methods for signal and image processing, arrhythmia mechanisms, robust signal processing methods, and data science for oncology.