Setting a generalized functional linear model (GFLM) for the classification of different types of cancer

Miguel Flores, Guido Saltos and Sergio Castillo-Páez

***Abstract* — This work aims to classify the DNA sequences of healthy and malignant cancer respectively. For this, supervised and unsupervised classification methods from a functional context are used; i.e. each strand of DNA is an observation. The observations are discretized, for that reason different ways to represent these observations with functions are evaluated. In addition, an exploratory study is done: estimating the mean and variance of each functional type of cancer. For the unsupervised classification method, hierarchical clustering with different measures of functional distance is used. On the other hand, for the supervised classification method, a functional generalized linear model is used. For this model the first and second derivatives are used which are included as discriminating variables. It has been verified that one of the advantages of working in the functional context is to obtain a model to correctly classify cancers by 100%. For the implementation of the methods it has been used the fda.usc R package that includes all the techniques of functional data analysis used in this work. In addition, some that have been developed in recent decades. For more details of these techniques can be consulted Ramsay, J. O. and Silverman (2005) and Ferraty et al. (2006).**

***Index Terms*— Depth of functional data, DNA, functional data analysis, functional distances, statistical classification**

I. Introduction

THE DNA Microarray chips and high-density oligonucleotide are widely used in modern biomedical research and can serve as a guide for the diagnosis and treatment of some diseases.

One of the most interesting and current applications is the characterization and classification of different types of cancer Singh D. et al. (2002). Microarray data show expression levels of many genes with respect to a number of observations (samples) and therefore can be considered as functional data or data with high dimension.

To this effect, it is very common to use multivariate methods to classify or create groups, for example according to Romualdi et al., (2003); Wessels et al., (2005); Tárraga et al. (2008) the best methods are: the K nearest neighbor method (KNN) and Diagonal Linear Discriminant Analysis (DLDA). Also, in the work of Dudoit et al. (2002) you can see a comparison of discrimination methods for the classification of tumors using gene expression data.

However, these methods of classical statistics do not perform well when the dimension of the data is very high relative to the size of the sample. (López-Pintado et al., 2010)

In this paper, a new approach is proposed for the classification of different types of cancer using models of Functional Data Analysis (FDA). This recent field of statistics allows processing data with high dimension and take advantage of their functional character.

Specifically, it is used a generalized functional linear model fit to classify the levels of expression of a set of genes in a type of tumor that affects a group of individuals.

To illustrate procedures of Functional Analysis of data is used, the database "prostate” belonging to the package " depthTools " R, which contains a random sample of 25 non-tumor samples (healthy) and 25 tumor samples (malignant), in which have been measured the expression levels of 100 genes. For more details on the data, you can consult Singh D. et al. (2002).

A glossary follows explained in Table I.

TABLE I
TERMS GLOSSARY

Term	Definition
AIC	Akaike Information Criteria
DNA	Deoxyribonucleic Acid
DLDA	Diagonal Linear Discriminant Analysis
FDA	Functional Data Analysis
FPLS	Functional Partial Least Squared - Principal Component
FM depth	Fraiman and Muniz depth
GCV	Generalized Cross-Validation
GFLM	Generalized Functional Linear Model
KNN	K nearest neighbors estimator
LLR	Local Linear Smoothing
NW	Nadaraya Watson Kernel Estimator
PL	Partial Least - Principal Component
RP depth	Random Projection depth

Finally, to implement the FDA procedures, the R statistical software is used, because the R package fda.usc has applicable routines for functional data. This package carries out exploratory and descriptive analysis of functional data, analyzing its most important features such as depth

Miguel Flores, is a professor at the Departamento de Matemática, Escuela Politécnica Nacional, 17012759 Ecuador (e-mail: miguel.flores@epn.edu.ec)
Guido Saltos, is a professor at the Universidad de las Américas, Quito, Pichincha, Ecuador (e-mail: guido.saltos@udla.edu.ec)

Sergio Castillo-Páez, is a professor at the Universidad de las Fuerzas Armadas del Ecuador ESPE, Sangolqui, Pichincha, Ecuador (e-mail: sacastillo@espe.edu.ec)

measurements or functional outliers detection, among others. Besides, fda.usc includes the functions implemented by Ferraty et al. (2006).

II. Functional Definition And Representation

$\mathcal{X}$ is defined as functional variable of interest, level of expression of genes taking values in a normed space (or semi - normed) $\mathcal{F}$, and the set $\{\boldsymbol{x}_1, \boldsymbol{x}_2,...,\boldsymbol{x}_n\}$ is considered the functional data to be analyzed which come from n functional variables $\mathcal{X}_1, \mathcal{X}_2,...,\mathcal{X}_n$ identically distributed as $\mathcal{X}$. Functional data are discretized in a set of points $\{t_j\}_{j=1}^{d}$ not necessarily equidistant (as here).

Therefore, it has d (genes) assessments for each of the n (observations) functional variables, that is, with a matrix of 50 rows representing discretized curves and 100 columns representing points to evaluate. The first 25 rows correspond to levels of expression of normal tumors and the following 25 rows to malignant tumors.

In Figure 1, you can see in black the different levels of the genes for normal tumors and red for malignant. At first glance this figure does not distinguish differences between tumor types.

To appreciate a greater difference on the relationship of genes and their expression level, for each tumor type a panel of six graphs is presented in Figure 2, in each row there are three graphs corresponding to normal tumors, first row, and malignant tumors, second row. The graphs in each row corresponds respectively to functional data (first), first derivative (second), and second derivative (third).

The representation made in Figure 1 for the functional data implicitly assumes a space L_2 which does not allow adequate discrimination between tumor types; you can see that by studying the behavior of the level of gene expression in other spaces (see Figure 2) can have a better discrimination. Specifically, the functional space of the second derivative of the functional data provides greater features (depth and variability) to discriminate between the two tumors.

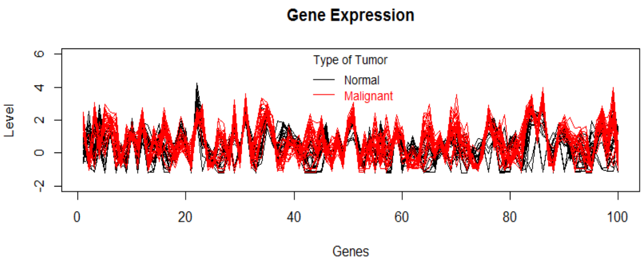

Figure 1: Graph of functional data represented by curves of black color for normal and red for malignant tumors.

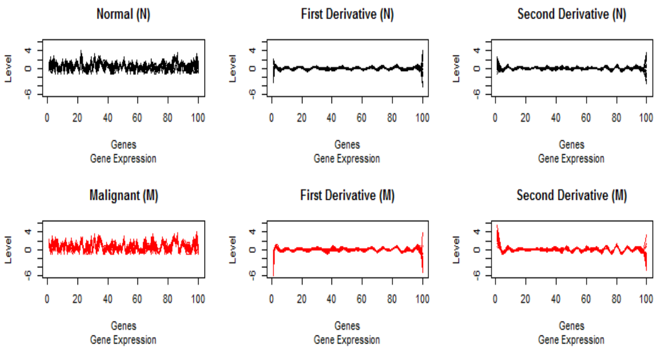

Figure 2: Panel of six graphs, in each row there are three graphs corresponding to the functional data, first derivative and second derivative; in the first row for normal tumors and in the second row for malignant tumors.

The representations in bases made for the first curve of the sample's functional data are shown in Figure 3: B-Splines (5, 20), Kernel Smoothing (KNN, LLR, NW) and Principal Components (PL and FPLS). These representations allow you to work the problem in finite dimension.

For the selection of a base, a setting parameter must be calibrated that allows a better representation; for this selection has been considered as criterion the Generalized Cross-validation (GCV) method. For more information about base types, methods and validation criteria, see Febrero-Bande, M. and Oviedo de la Fuente, M. (2012).

For the classification of tumors we work with representations in base; but for calculating distances and exploratory analysis of functional data we do not work with representation in base. The fda.usc R package is used to perform calculations using the corresponding numerical approximations.

As can be seen in Figure 3, depending on the method and the adjustment parameter representations in base, they are different. In the case of a representation by principal components we can see that there is not much difference between the PL and PLS method.

In Table II the following indicators are shown: the percentage of variance explained for each component; the correlation between the level of gene expression; and the type of tumor. These indicators are calculated for the original data, its first derivative and second derivative.

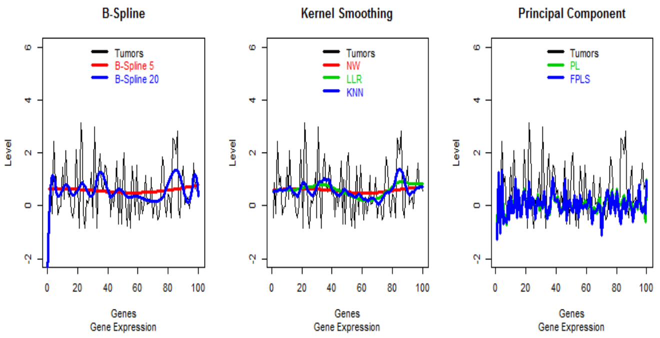

Figure 3: The base representations made to the first observation of the functional data with bases: B -Splines (5.20), Kernel Smoothing (KNN, LLR, NW) and Principal Components (PL and FPLS).

TABLE II
PERCENTAGE OF EXPLAINED VARIANCE FOR EACH COMPONENT AND THE CORRELATION BETWEEN THE EXPRESSION LEVEL OF GENES AND TUMOR TYPE FOR THE ORIGINAL DATA, ITS FIRST DERIVATIVE AND SECOND DERIVATIVE

Data	Method	Indices	PC1	PC2	PC3
Original	PL	% explained variance	75.21	17.8	6.98
		Tumor type correlation	67.8	-37.6	-11.3
	FPLS	% explained variance	79.75	14.78	5.47
		Tumor type correlation	75.1	41.5	30.9
First Derivative	PL	% explained variance	71.85	23.76	4.39
		Tumor type correlation	68.8	-35.6	41.6
	FPLS	% explained variance	75.3	18.64	6.06
		Tumor type correlation	74.1	41.4	60.4
Second Derivative	PL	% explained variance	68.8	22.22	5.96
		Tumor type correlation	66.8	-41.7	38.4
	FPLS	% explained variance	74.27	19.85	5.88
		Tumor type correlation	73.7	42.6	52.2

About 70 % of the total variability of the data is explained by the first component, regardless of the method of principal components to be used; the variability explained by the second component increases to about 20 % when working in the spaces of the functions of the first and second derivatives.

In general, we can say that the first two components explain about 90 % of the variability; the first component has a strong positive ratio of about 70 % in all methods; and the second component has a negative ratio using the method PL and a positive one using the PLS method.

III. Distance Between Functional Data

In this section it has been applied a metric for the L_2 space and 4 semi - metrics for other semi - normed spaces, in order to calculate the distance between the functional data (for more information on the definition of each measure, see Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). For calculating these measures, have been implemented the following functions developed in the fda.usc package:

1) metric.lp (for functional data represented in a L_p space, with p = 2).
2) semimetric.deriv (for functional data in the space of functions of the first and second derivative).
3) metric.pl (based on the method of principal components (PL), it calculates a PL semi- metric between functional data).
4) metric.mplsr (based on the principal component method (PLS), it calculates a FPLS semi- metric between functional data).

TABLE III
PERCENTAGE OF CORRECT CLASSIFICATION OF TUMORS FOR EACH METHOD.

Function	Space	% success
metric.deriv	First derivative	40
metric.deriv	Second derivative	78
metric.pca	Principal Component PL	84
metric.mplsr	Principal Component FLPS	98
metric.lp	L^2	40

The result of each of these functions (metric and semi- metric) is a matrix of dimension 50 x 50 containing the distances between all curves (functional data).

You can use this information as a classification rule, since it is expected that the closest curves belong to the same tumor. In Figure 4 the dendrogram for the semi - metric Principal Component FPLS is shown.

The results presented in Table III, are the percentages of correct classification and correspond to distance functions (the metric and semi- metric). The highest values are those calculated by the metric.lp and metric.deriv functions. With the semi - metric calculated by the metric.mplsr function it was achieved a 98 % of success; it should be mentioned that only came to classify erroneously one case (curve 40: a malignant cancer classified as normal). Cuevas et al., (2001) use an approach based on density estimation for doing a Cluster Analysis.

Clearly, with these results is more advisable to work in semi - normed spaces to identify differences between the expressions of genes according to cancer types for better classification.

IV. Exploratory Analysis Of Functional Data

For this section, the exploratory data analysis has been divided in two parts:

1) Variability and central tendency estimation
2) Outliers detection

Estimates of central tendency and variability for each type of cancer are done using robust methods, therefore there is not a great influence by outliers for estimates.

However, it was decided to conduct a study of outliers detection to illustrate the methodology to be used in the case of not having these robust methods.

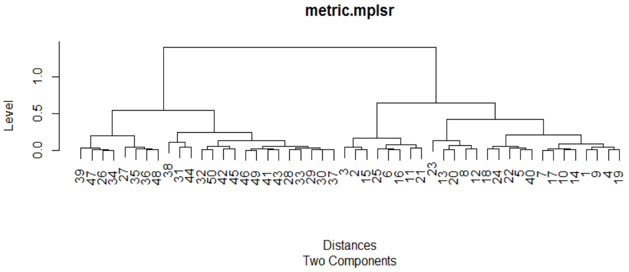

Figure 4: Dendrogram for the semi - metric Principal Component FPLS (98% correct classification)

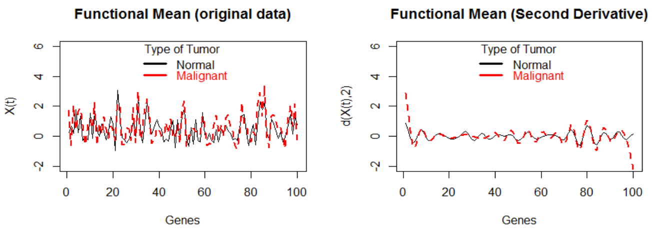

Figure 5: Functional Mean for original data and second derivative for each type of tumor.

A. *Variability and central tendency estimation*

The descriptive exploratory analysis consists in: calculating the mean, and functional variance of the expression levels of genes and its second derivative. This study is performed for each type of tumor to differentiate central tendency and variability of functional data.

In Figure 5, we can appreciate the functional mean for the original data and its second derivative. In each graph the functional mean is distinguished for each type of cancer.

As shown in the graph on the left, the difference between the curves of the functional means for the original data is not very noticeable; on the other hand, in the right picture for functional mean of the second derivative you can see a greater difference.

Overall the two graphs give us an idea that the expression levels of genes tend to values between -0.5 and 1.5, approximately; whereas, the second derivative between -0.5 and 0.5. In addition, we can see that there is greater variability in the trend of the original data than in the second derivative.

In Figure 6, it is shown a graphical representation of the confidence ball representing the estimation limits, where the functional mean oscillates for each type of functional data in each space. It has been applied the smoothing bootstrapped method using the “fda.booststrap “function included in the fda. usc package.

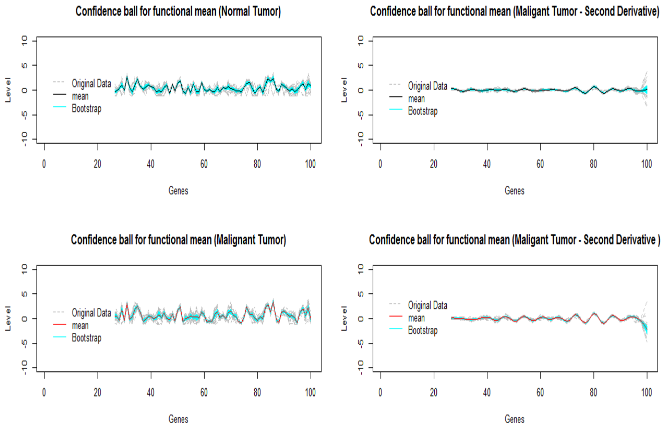

Figure 6: Confidence balls for Functional Mean for original data and second derivative for each type of tumor.

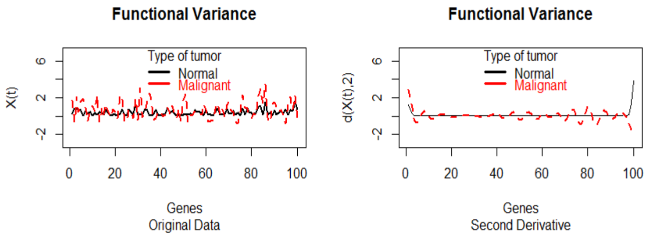

Figure 7: Functional variance for the original data and second derivative for each type of tumor

In Figure 6, light blue curves are the representation of confidence balls at a level of 95% generated by the bootstrap method for functional mean (black color curve). For graphics on the left, gray curves correspond to the original data and for graphs on the right, they are the second derivative's.

In Figure 7 the curves of variance for each type of cancer are shown in the spaces of the original data (graph on the left) and the second derivative’s (right graph). Here you can see a marked difference between the variance of normal and malignant tumors. In malignant tumors a greater range of variation is observed that in normal ones; this same behavior is similar in the two spaces of functional data.

B. *Outliers Detection*

Subsequently, a study on the presence of outliers is done because they could affect the estimation and performance (classification) of the model. The depth is a measure whose concept has emerged in the literature of robustness, measures how deep (or central) is a benchmark for a population (or sample). Therefore, those points having large depth values, will be closer to the behavior of the central data; and if they have less deep values, they will be potential candidates for outliers. For more information about the definition of a function of depth see Zuo Y. and Serfling R. (2000).

In univariate data, the median would be the deepest point of the set of points. For this study, we have applied the following depth measures which are included in the package fda.usc: Mode (mode depth); Median defined by Fraiman (Fraiman and Muniz, 2001) (FM depth); and Random Projections (RP depth).

Having studied the central tendency and variability of the data we continue with the detection of outliers in the sample. We start with an analysis with all the original data by calculating three measures of depth (shortened by 10 %) and the difference of each with respect to the median of functional data is observed (see Figure 8); subsequently, a scatter plot is made between the different depth measurements to see if there are outliers (the points with smaller depth values and that are not aligned to the general behavior of the points are considered outliers).

The analysis for the detection of outliers is accounted considering all the sample data; but this analysis is applicable for each subsample defined by the type of tumor. Table IV summarizes the curves (or outliers) considering the total sample and the subsample for each type of cancer.

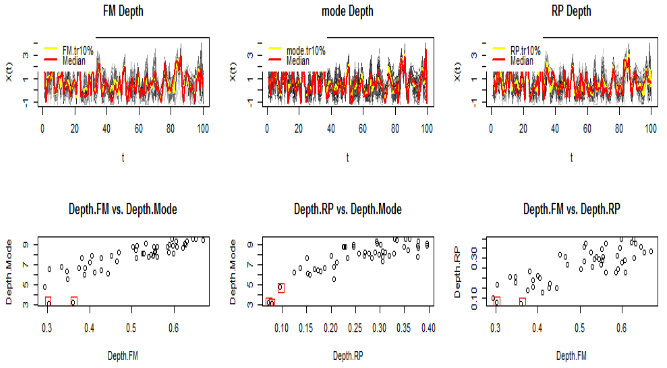

Figure 8: A panel of six graphs, the first row has the representation of depth measurements contrasted with the median of the total sample. The second row shows the scatter plots between all the depth measurements.

In the code attached to this work can be found the procedure developed for the calculation of depth measurements as their graphic representation for the entire sample and for each subsample.

In Figure 8, a panel of six charts that are distributed as follows is presented: in the first row is the representation of depth measurements contrasted with the median of the total sample; and in the second row, we have the scatter plots between all the depth measurements; it should be noted that have been marked with a red box the outliers for each graph and the order considered for reading the graphs is from left to right.

Contrasting depth measurements (second row of the graphics panel of Figure 8), clearly can be observed outliers in all three cases. In the first graph two points are identified as atypical functional data representing curves 2 and 21 belonging to the normal tumor sample; the same curves also are identified as atypical curves by observing the third graph; while in the middle graph three atypical points are observed, curves 2, 21 and 3.

To confirm this visual analysis, an analytical rule is applied which considers as atypical functional data the curves whose depth values are less than a quantile defined based on all calculated values of depth of each sample's data (curves).

In the case of the mode depth measurements to a 1% quantile, it could be identified as atypical data curves 2 and 21; this also happens with the depth measurement of random projections, i.e., the curves 2 and 21 are identified as atypical again with 1% quantile. Whereas, for the identification of atypical data in FM's median it is considered a 5% quantile and the curves are identified as 2 and 3.

Table IV summarizes functional data identified as atypical, considering each depth measurement and each sample.

TABLE IV
PERCENTAGE OF CORRECT ANSWERS IN TUMORS CLASSIFICATION

Depth	Total Sample	Normal Tumor	Malignat Tumor
Mode	2.2	2.2	40
FM	2.3	2.3	40
Rp	2.2	2.2	40.5

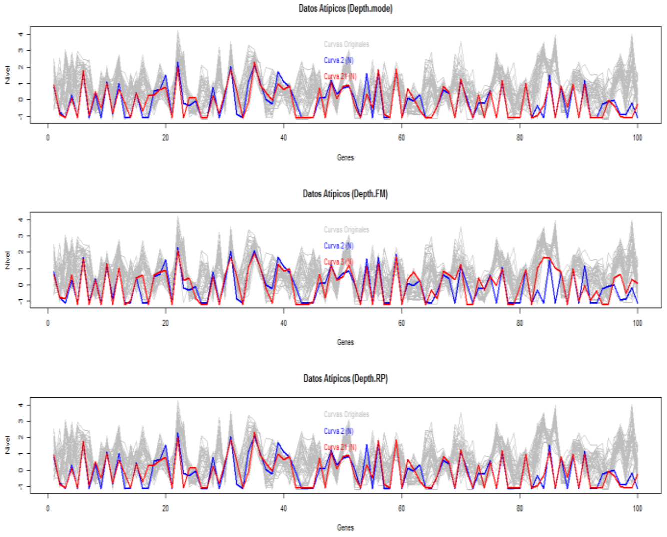

Figure 9: Curves identified as atypical functional data for each type of depth measurement.

In Figure 9, three graphs are shown. On each one, original curves are presented in gray and data identified as atypical in blue and red.

These results at first glance might indicate to us that there are only atypical data in normal tumors and that there are no atypical in malignant tumors, but performing the same analysis to identify atypical data in the sample of malignant tumors, it comes down to detect as atypical curves 40 and 48.

While in the subsample of malignant tumors curves 40 and 41 are identified as outlier. It is recalled that in the " Distance between functional data " section, in applying the distance by principal components to make a first approach to a classification rule, could not be correctly classify the curve 40.

V. Generalized Functional Linear Model (GFLM)

This section provides a Generalized Linear Functional Model (GFLM) where the functional covariates are: the level of expression of genes denoted as: **X=X (t)**, and, the first **X'(t)** and its second derivative **X''(t)** denoted as **X_1** and **X_2**, respectively; and as response scalar variable (binary) the cancer type denoted as **Y** (0 = normal tumor, 1 = malignant tumor).

In this case, as the GFLM works with a binary response variable, this model provides a classification rule for the type of cancer (Bayes' rule).

This model is also called Functional Logistic Regression (Febrero-Bande, M. and Gonzalez-Manteiga, W. 2012), i.e. the models explain the relationship between **Y** (binary response) and a functional covariate **X (t)** by base representation X (t) and β (t). The functional model of logistic regression of the probability **π_i**, the occurrence of an event, **Y_i** = 1, rather than **Y_i** = 0, conditioned on a vector of **covariates X_i (t)** is expressed as :

$$y_i = \pi_i + \epsilon_i \ , i = 1, \ldots, n$$

TABLE V
PERCENTAGE OF CORRECT CLASSIFICATION OF TUMORS

Number	Model	AIC	% classification
1	Y~X	49.7	82
2	Y~X$_1$	21.8	96
3	Y~X$_2$	24.1	96
4	Y~X+X$_1$	22.0	100
5	Y~X+X$_2$	22.0	100
6	Y~X$_1$+X$_2$	22.0	100
7	Y~X+X$_1$+X$_2$	32.0	100

Where $\boldsymbol{\pi_i}$ is the expectation of **Y** given $\mathbf{X_i}$ **(t)** modeled as follows:

$$\pi_i = P[Y=1|x_i(t) : t \in T] = \frac{exp\left\{\int_T X_i(t)\beta(t)dt\right\}}{1+exp\left\{\int_T X_i(t)\beta(t)dt\right\}} \quad ,i=1,\ldots,n$$

With $\boldsymbol{\epsilon_i}$ as independent errors with mean zero.

The functional variables used to estimate the model are:

1) Y = binary variable that identifies the type of tumor (0 = normal tumor, 1 = malignant tumor)
2) X = expression level of 100 genes of each individual
3) X1 = first derivative of the expression level of 100 genes of each individual
4) X2 = second derivative of the expression level of 100 genes of each individual

From these variables, they were estimated and compared seven models, Table III summarizes the characteristics evaluated to select the best model to use for the classification of tumor types. It is worth mentioning that the "fregre.glm" function from the R fda.usc package was used and B - Spline as representation basis for the seven models.

Additionally, it was explored with a representation based on principal component (PLS) for models 1 and 2 (it was used an R code for this) to improve results in adjustment and classification; but the results are similar to the representation in B - SPLINE therefore not proceeded to make estimates with this type of representation based .

The criteria used are: AIC (while lower is better); the percentage of tumors that are classified correctly from the total sample (% classification); and the percentage of prediction, that is, the percentage of tumors that are classified correctly from the total test sample (% Prediction). To calculate the prediction percentage, 10 test samples were used, 5 of normal tumors and 5 of malignant tumors; these were taken randomly setting a seed.

In the first model (see Table V), only are considered the original data (levels of gene expression), this is the model that explain less (AIC = 49.7) and its classification and prediction percentages are 82 % and 80 % respectively; on the other hand, with respect to the significance of the model parameters, we have that the first component (ab.bspl4.1) is significant (0.00639) to a level of significance of 5%. All parameters for the other models are not statistically significant at a level of significance lower than 1 %.

The second model (see Table V), has the lowest AIC (21.8) of all the proposed models, but there are models with better percentages of classification and prediction. From Model 4 to Model 7, the percentage of classification and prediction is 100 %, except for model 6 which has only a 70 % of prediction.

In general, when only are considered single-variable models of explanation for the type of cancer; AIC coefficient, the percentage of classification and prediction are the worst of all the proposed models.

Furthermore, it can be seen in Table V that increasing the number of variables in a model, the classification percentage improves up to 100%; however, when only considered in model 6, the functional variables: first and second derivative, the percentage of prediction is 70 %; and, when you have a more complex model with three functional variables considering the original data, its first and second derivative prediction, the model improves prediction but worsens explanation (best fit); In conclusion, one has that the complex model is good for predicting but not to explain the behavior of the cancer type variable.

In Table V, the painted yellow rows indicate the two models that have the same characteristics of explanation (best fit), classification and prediction; models 4 and 5 are the best models of seven models estimated. Therefore, the best model to classify tumors in normal and malignant is that which consider original data and one additional functional variable which can be the first or second derivative of the original data; this model comes to have a classification and prediction efficiency of 100%.

If the classification results obtained with models 4 and 5 are compared with the classification procedure by means of the distances between functional data used in section three which showed an efficiency of 98 %, we could say that for this sample, a Functional Generalized Linear Model is (GFLM) is more robust to the presence of outliers, as it allows an classification and prediction efficiency of 100%.

Finally, to complete this work Figure 10 shows the adjustments of the GLFM models for the cancer type variable when the entire sample of tumors is considered. It is worth mentioning that the graph settings for models of more than one explanatory functional variable is equal for all, because from model 4 to model 7 all have a classification percentage of 100%.

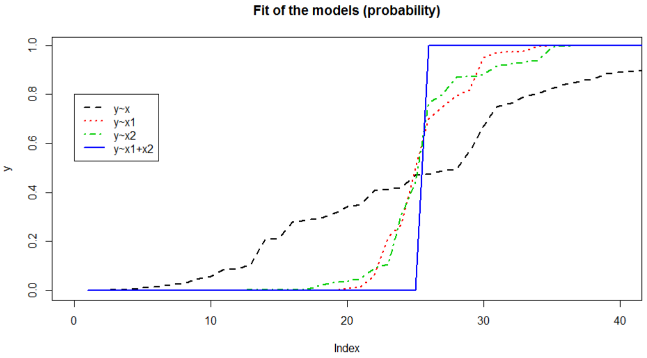

Figure 10: Functional Mean for original data and second derivative for each type of tumor.

VI. CONCLUSIONS AND FUTURE RESEARCH

1) It has greater discrimination when working with the first and second derivative of the expression of genes. This is reflected also in calculating the functional mean and variance in these spaces.

2) In the section "Outliers Detection", curves 2,3,21 and 40 are determined as outliers. These are not classified correctly using the cluster method but by using the functional generalized linear model.

3) Increasing the number of variables in the functional generalized linear model, the classification percentage improves up to 100%. The functional variables included were the first and second derivative.

4) The functional data analysis is very recent in the fields of statistics and medicine, despite this there is an increased interest in using this methodology. It is intended to continue to address problems of classification in other areas of science.

5) Specifically for the medical field, will work to make a functional generalized additive linear model that eliminates the restriction of linearity for the independent variables.

6) Besides, it is addressing functional models to describe the relationship of the expressions of genes with other variables related to cancer.

REFERENCES

[1] Cuevas A, Febrero M, Fraiman R. 2001. Cluster Analysis: a further approach based on density estimation. Computational Statisticsand Data Analysis 36: 441–456.

[2] Dudoit et al. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association, 97 (457), 77-87.

[3] Febrero-Bande, M. and Oviedo de la Fuente, M. 2012. Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software, 51(4):1-28.

[4] Febrero-Bande, M. and Gonzalez-Manteiga, W. 2012. Generalized additive models for functional data. TEST, 22(2):278-292.

[5] Ferraty, F. and Vieu, P. 2006. “Nonparametric Functional Data Analysis: Theory” and Practice. Springer-Verlag, New York., Pp. 113-146.

[6] Fraiman R. and Muniz G. 2001 Trimmed means for functional data, Test, 10(2), 419-440.

[7] Lopez-Pintado, S., Romo, J., Torrente A. 2010. “Robust depth-based tool for the analysis of gene expression data”. Biostatistics 11, 2, pp 254-264.

[8] Ramsay, J. O. and Silverman, B. W.2005. “Functional Data Analysis”, 2nd ed., Springer-Verlag, New York., pp. 147-325.

[9] Romualdi C., Campanaro S., Campagna D., Celegato B., Cannata,N, Toppo S.,Valle G. and Lanfranchi G. 2003 Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. Human Molecular Genetics **12**, 823-836.

[10] Singh D. et al. 2002. Gene expression correlates of clinincal prostate cancer behavior, Cancer cell, 1 (2), 203-209.

[11] Tárraga J., Medina I., Carbonell J., Huerta-Cepas J., Mínguez P., Alloza E., Al-Shahrour F., Vegas-Azcarate S. Gotz S. Escobar P and others 2008. GEPAS a web-based tool for microarray data analysis and interpretation. Nucleic Acids Research 36, W308-W314.

[12] Wessels L.F.A., Reinders M. J. T., Hart,A.A.M.,Veenman C.J., Dai H., He Y.D. and Van’t Veer L.J. 2005. A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics 21, 3755-3762.

[13] Zuo Y, Serfling R. 2000. General notions of statistical depth function. Annals of Statistics 28: 461–482.

Miguel Flores, is a professor at the National Polytechnic School and a researcher at the Center for Modeling Mathematics at the National Polytechnic School in Quito, Ecuador. He is a BSc. in Statistical Computing Engineer from the Polytechnic School of the Coast. In 2006 he received an in MSc. in Operations Research from the National Polytechnic School, and in 2013 received a MSc. in Technical Statistics from the University of A Coruña. He is currently a doctoral student at the University of A Coruña in the area of Statistics and Operations Research. He has over 15 years professional experience in various areas of Statistics, Computing and Optimization, multivariate data analysis, econometric, Market Research, Quality Control, definition and construction of systems indicators, development of applications and optimization modeling. ORCID ID: 0000-0002-7742-1247

Guido Saltos S. received his Engineering degree in Electronics from Escuela Politécnica del Ejército (ESPE), Quito, Ecuador in 1987. He received his M.S. degree in Applied Statistics from National Polytechnic School (EPN), Quito, Ecuador, in 2016. He worked several years in the field of industrial automation and now he is working at Universidad de las Américas (UDLA) in Quito Ecuador. His interests are related with data depth, and non-parametric statistics.

Sergio Castillo Páez, is a mathematical engineer graduated from the National Polytechnic School in 2002. He also studied finance in the Simon Bolivar Andean University, and is currently studying his PhD in Statistics at the University of Vigo, Spain. He is a professor at the ESPE Armed Forces University in Ecuador. His current lines of research are related to geostatistics and analysis of multivariate data.