Università degli Studi di Pavia

Università della Svizzera italiana

Joint PhD program in Computational Mathematics and Decision Sciences

XXXIV Cycle

Statistical and Machine Learning Models for Neurosciences

PhD Dissertation of:

Elena Ballante

PhD Coordinator:

Prof. Luca Pavarino

Supervisor:

Prof. Silvia Figini

Academic year 2020-2021

Una realtà non ci fu data e non c'è,
ma dobbiamo farcela noi, se vogliamo essere:
e non sarà mai una per tutti, una per sempre,
ma di continuo e infinitamente mutabile.
(Luigi Pirandello)

Acknowledgements

First of all, I would like to thank my advisor Silvia Figini that gave me the opportunity to do this experience and who gave me all the support I needed.
I would like also to thank the IRCCS Mondino Foundation for making my doctorate possible, its scientific director Fabio Blandini for the trust he has placed in me and all the colleagues I have collaborated with for giving me continuous stimuli and transmit their passion for their work.
I would like also to thank the other people whose collaboration has been fundamental to my academic growth: Dott. Aymeric Stamm, Prof. Lise Bellanger, Prof. Giuseppe Toscani, Prof. Pietro Muliere, and Prof. Pierpaolo Uberti. A big thank you goes to my colleagues and friends Marta and Chiara Bardelli, whose presence and support was essential throughout my entire PhD. Without them I am not sure that I would reach this milestone.
I thank my parents for the constant trust and support, especially my mother that has always been my biggest fan.
I have to thank my lifelong friends because with their presence they have given me the serenity I needed: Sharon, Maurizio, Chiara, Mariachiara, Marina, Francesca, Irene and Teresa.
A big thank you also to all the PhD students and Post Docs with whom I shared moments of work and relax and with their presence have contributed to make more enjoyable this time.

Abstract

This thesis addresses several problems encountered in the field of statistical and machine learning methods for data analysis in neurosciences. The thesis is divided into three parts. The first part of the thesis is related to classification tree models. In the research field of polarization measures, a new polarization measure is defined. The function is incorporated in the decision tree algorithm as a splitting function in order to tackle some weaknesses of classical impurity measures. The new algorithm is called Polarized Classification Tree model. The model is tested on simulated and real data sets and compared with decision tree models where the classical impurity measures are deployed.

In the second part of the thesis a new index for assessing and selecting the best model in a classification task when the target variable is ordinal is developed. The index proposed is compared to the traditional measures on simulated data sets and it is applied in a real case study related to Attenuated Psychosis Syndrome.

The third part covers the topic of smoothing methods for quaternion time series data in the context of motion data classification.

Different proper methods to smoothing time series in quaternion algebra are reviewed and a new method is proposed.

The new method is compared with a method proposed in the literature in terms of classification performances on a real data set and five data sets obtained introducing different degrees of noise. The results confirmed the hypothesis made on the basis of the theoretical information available from the two methods, i.e. the logarithm is smoother and generally provides better results than the existing method in terms of classification performances.

List of Papers

Paper I

Ballante, E., Galvani, M., Uberti, P., Figini, S. "Polarized Classification Tree Models: theory and computational aspects". In: *Journal of Classification* (2021) DOI: 10.1007/s00357-021-09383-8.

Paper II

Ballante, E., Figini, S., Uberti, P. "A new approach in model selection for ordinal target variables". In: *Computational Statistics* (2021) DOI: 10.1007/s00180-021-01112-4.

Paper III

Ballante, E., Molteni, S., Mensi, M.M., Figini, S. "At risk mental status analysis: a comparison of model selection methods for ordinal target variable". In proceedings of *SIS 2020*.

Paper IV

Ballante, E., Bellanger, L., Drouin, P., Figini, S., Stamm, A. "Smoothing Method for Unit Quaternion Time Series: An application to motion data". *Submitted.*

Contents

List of Figures

List of Tables

Introduction

This thesis addresses several problems encountered in the field of statistical and machine learning methods for data analysis in neurosciences.

The thesis is divided into three parts. Part 1 is related to the study and improvement of a classical supervised machine learning model, the decision tree model. Part 2 is about the definition of a model assessment and selection method in a classification task when the target variable is ordinal. Part 3 describes a new method to smooth quaternion time series data in order to improve classification performances.

The topics covered in this thesis fall within the framework of supervised machine learning. Consider a learning set $\mathcal{L} = \{(\mathbf{x}_i, y_i)_{i=1,...N}\}$, where $\mathbf{x}_i$ are the p-dimensional vectors of covariates (input variables) and y_i is the value of the target variable Y for each subject $i = 1, ..., N$. The objective of the supervised machine learning models is to make inference about the function f that links the covariates with the target variable, i.e. it tries to reconstruct the relation $Y = f(X) + \epsilon$ between the predictors and the target variable, where ϵ is the random error with $E(\epsilon) = 0$ and independent from X. In the framework of classification the target variable is qualitative.

Part 1 treats classification tree models. These models are defined in Breiman, Friedman, and Olsen 1984 as a recursive procedure through which a set of N statistical units are progressively divided into groups, according to a splitting rule that aims to maximize homogeneity or purity of the response variable in each of the obtained groups. Classification tree models are able to handle both numerical and categorical predictors without requiring any assumption on the target variable.

The main distinctive element of a classification tree model is the choice of the

splitting rule. A splitting rule selects a predictor from those available and chooses the best partition of its levels. The choice is generally made using a goodness measure which, in standard classification trees, is evaluated observing the pureness in terms of target variable of the descendant nodes.

In the literature on classification trees (see Mingers 1989), it is recognized that splitting rules based on the impurity measures (i.e. Gini impurity index and Information Gain) suffer from some weaknesses. One of them is that most popular splitting criteria are mainly focus on nodes impurity and do not take into account the predictors' distribution.

We propose a new splitting function starting from the definition of polarization measures. The concept of polarization measures was introduced in Esteban and Ray 1991, Esteban and Ray 1994, Foster and Wolfson 1992 and in Wolfson 1994, and it is typically adopted in the socio-economic context to measure inequality in income distribution.

In our proposal, we define a multidimensional polarization measure, which considers one continuous variable and exogenously defined groups represented by a categorical variable.

Since we would like to introduce a measure which treats variables coming from different contexts (not only from the economic one), a measure of variability which is not of inequality is introduced. Furthermore, a generalization of the axioms defined in Duclos, Esteban, and Ray 2004 is required to derive our multidimensional index of polarization which works on continuous explanatory variables. The new measure is incorporated in the decision tree algorithm as a splitting function in order to take into account the distribution of the covariates instead of the only nodes' impurity. The new algorithm is called Polarized Classification Tree model. A toy example is described in order to show the peculiarity of the new measure as splitting method with respect to the classical splitting functions (Gini impurity index and Information Gain). The model was tested on simulated and real data sets and compared with decision tree models, where the cited classical impurity measures are deployed. Results confirm that the new model proposed is competitive with respect to the classical measures

and in some cases it shows significantly better performances.

In the Part 2 of the thesis, we developed a new index for assessing and selecting the best model in a classification task when the target variable is ordinal.

Evaluation measures are widely used in predictive models to compare different algorithms, thus providing the selection of the best model for the data at hand. Different indicators can be used to assess the performance of a model in terms of accuracy, discriminatory power and stability of the results. The choice of indicators to perform model selection is a fundamental point and several methods have been proposed depending on the nature of the data and the problem domain (see e.g. Adams and Hand 2000; Bradley 1997; Hand 2009).

While in the model definition stage for ordinal target variable there are different approaches in the literature (see Agresti 2010; Ahmad and Brown 2015; Kotłowski et al. 2008; Torra et al. 2006), for the model selection there is a lack of adequate tools (Cardoso and Sousa 2011). Moreover, performance indicators should take into account the nature of the target variable, especially when the dependent variable is ordinal. This motivates the proposal of a new class of measures to select the best model in predictive contexts characterized by a multi-class ordinal target variable, using the misclassification errors coupled with a measure of uncertainty on the prediction.

Two toy examples show how the index works and its advantages with respect to the classical evaluation measures (accuracy, AUC, MSE). The index is tested on a simulated data set and compared with the cited classical measures in the evaluation of the performances of different models. Results confirm that the new index can capture peculiar aspects compared to the traditional measures.

The index proposed is also applied in a real case study. A data set related to the study of Attenuated Psychosis Syndrome is analysed in terms of classification task.

The Part 3 covers the topic of smoothing methods for quaternion time series data in the context of motion data classification.

Smoothing is a fundamental step in time series analysis and even more when sensors are used for the acquisition and measurements. In fact, sensors tend

to collect a certain amount of noise that is difficult to isolate and manage. The smoothing process attempts to capture important patterns in the data, while leaving out noise or other fine-scale information that can make the main structures hardly analysable.

While in the literature many different techniques to perform smoothing are present for real signals, in a non-standard algebra such as quaternionic algebra the topic is still developing. Some interesting works related to smoothing time series in quaternion algebra are Ginzberg and Walden 2012, Janiak, Szczęsna, and Słupik 2014, Hsieh 2002. Despite this, the lack of availability of the code makes these methods leaving open the problem of applications to the real world cases.

Starting from the method proposed in Hsieh 2002, a new method that deploys the logarithm transformation instead of angular velocity to transform the quaternion time series into a real three dimensional time series. These two methods are compared in terms of classification performances on a real data set and five derived data set where different degrees of noise are introduced. In order to confirm the validity of the proposed method, logistic regression models on accuracy and AUC are performed, where the influence of the transformation function and the smoothing method is evaluated on simulated data sets. The results confirm the hypothesis made on the basis of the theoretical information available on the two methods, i.e. the logarithm is smoother and generally provides better results than the existing method in terms of classification performances.

Summary of Papers

Paper I proposes a new splitting method to deploy in the decision tree models, taking into account the distribution of the explanatory variables in the space partitioning. The new algorithm proposed is called Polarized Classification Tree model.

Paper II proposes a new index for the assessment and selection of classification models when the target variable is ordinal.

Paper III shows an application to real world problem of the index proposed in Paper II.

Paper IV proposes a new method to smooth unit quaternion time series in the classification framework.

Part 1

Polarized Classification Tree Models

Introduction

Classification trees are classical supervised machine learning algorithms, introduced in Breiman, Friedman, and Olsen 1984.

They are defined as a recursive procedure through which a set of N statistical units are progressively divided into groups, according to a splitting rule that aims to maximize homogeneity or purity of the response variable in each of the obtained groups. Classification tree models are able to handle both numerical and categorical predictors without requiring any assumption on the target variable.

In order to build a classification tree model, let Y be a target variable referring to a set of N samples falling into M classes, e.g. Y can take value in $\{1, 2, ..., M\}$. Let $\mathbb{X}$ be the $N \times K$ data matrix with N independent observations and K explanatory variables. The features $X_1, ..., X_K$ are random variables that take values in K different sets $A_1, ..., A_K$. In order to build a binary classification tree a process of partition of the input space is required.

The main distinctive element of a classification tree model is the choice of the splitting rule. A splitting rule selects a predictor from those available and chooses the best partition of its levels. The choice is generally made using a goodness measure which, in standard classification trees, is evaluated observing the pureness of the descendant nodes in terms of target variable.

Despite the great popularity of tree models and the large amount of modification proposed also in recent years (see for example Nerini and Ghattas 2007, D'Ambrosio et al. 2017, Aria et al. 2018, and Iorio et al. 2019), we underline that most popular splitting criteria are mainly focus on nodes impurity and do not take into account the variables distribution.

Polarization measures can be a good alternative to evaluate the homogeneity within classes and heterogeneity between classes to identify the best split to be performed at a specific node. The concept of polarization measures was introduced in Esteban and Ray 1991, Foster and Wolfson 1992, Esteban and Ray 1994 and in Wolfson 1994, and it is typically adopted in the socio-economic

context to measure inequality in income distribution.

In our proposal we define a multidimensional polarization measure, which considers one continuous variable and exogenously defined groups represented by a categorical variable. Since we would like to introduce a measure which treats variables coming from different contexts (not only from the economic one), a measure of variability which is not of inequality is introduced. Furthermore, a generalization of the axioms defined in Duclos, Esteban, and Ray 2004 is required to derive our multidimensional index of polarization which works on continuous explanatory variables. The new measure is incorporated in the decision tree algorithm as a splitting function in order to take into account the distribution of the covariates instead of the only impurity of the nodes. In fact, the classical impurity measures applied as splitting criteria are Gini impurity index, defined as $i(t) = \sum_{j \neq i} p(j|t)p(i|t)$, and Information Gain that is calculated by subtracting the weighted entropies of each branch from the original entropy, where the entropy of the node is $H(t) = -\sum_{i=1}^{n} p(i|t)\log(p(i|t))$.

The new algorithm proposed is called Polarized Classification Tree model. A toy example is described in order to show the peculiarity of the new measure as splitting method with respect to the classical splitting functions (Gini impurity index and Information Gain). The model was tested on simulated and real data sets and compared with decision tree models where the cited classical impurity measures are deployed. Results confirm that the new model proposed is competitive with respect to the classical measures and in some cases it shows significantly better performances.

The main contributions of the work are multiple. From a theoretical point of view we define a new measure that can generalize polarization measures in order to be applied in an efficient way in classification trees. From a computational point of view we develop a new classification algorithm based on decision trees improving decision tree splitting criteria observing not only the pureness of a new node but also including variables distribution overpassing some of the decision trees weaknesses.

Paper I

Polarized Classification Tree Models: theory and computational aspects

Elena Ballante[1], Marta Galvani[1], Pierpaolo Uberti[2], Silvia Figini[3]

Published in *Journal of Classification*, February 2021, DOI: 10.1007/s00357-021-09383-8

Abstract

In this paper a new approach in classification models, called Polarized Classification Tree model, is introduced.

From a methodological perspective a new index of polarization to measure the goodness of splits in the growth of a classification tree is proposed. The new introduced measure tackles weaknesses of the classical ones used in classification trees (Gini and Information Gain), because it does not only measure the impurity but it reflects the distribution of each covariate in the node, i.e. employing more discriminating covariates to split the data at each node.

From a computational prospective, a new algorithm is proposed and implemented employing the new proposed measure in the growth of a tree. In order to show how our proposal works, a simulation exercise has been

[1] Department of Mathematics, University of Pavia, Pavia, Italy
[2] Department of Economics, University of Genova, Genova, Italy
[3] Department of Political and Social Sciences, University of Pavia, Pavia, Italy

carried out. The results obtained in the simulation framework suggest that our proposal significantly outperforms impurity measures commonly adopted in classification tree modeling. Moreover, the empirical evidence on real data shows that Polarized Classification Tree models are competitive and sometimes better with respect to classical classification tree models.

Keywords: Classification trees, Polarization Measures, Splitting rules

I.1 Introduction

Classification trees are non parametric predictive methods obtained by recursively partitioning the data space and fitting a simple prediction model within each partition Breiman, Friedman, and Olsen 1984.
The idea is to divide the entire X-space into rectangles such that each rectangle is as homogeneous or pure as possible in terms of the dependent variable (binary or categorical), thus containing points that belong to just one class Shneiderman 1992.
As decision tree models are simple and easy interpretable models able to obtain good predictive performance, they are of interest in many recent works in literature (see for example Aria et al. 2018, Iorio et al. 2019, Nerini and Ghattas 2007 and D'Ambrosio, Aria, et al. 2017).
One of the main distinctive element of a classification tree model is how the splitting rule is chosen for the units belonging to a group, which corresponds to a node of the tree, and how an index of impurity is selected to measure the variability of the response values in a node of the tree.
The main used splitting rules are the Gini index, introduced in the CART algorithm proposed in Breiman, Friedman, and Olsen 1984, and the Information Gain, employed in the C4.5 algorithm Quinlan 2014. Other different splitting criteria have been proposed in literature as alternatives to these two ones. A faster alternative to the Gini index is proposed in Mola and Siciliano 1997 employing the predictability index τ of Goodman and Kruskal 1979 as a splitting rule. In Ciampi et al. 1987, Clark and Pregibon 2017 and Quinlan 2014 the likelihood is used as splitting criterion, while the mean posterior improvement (MPI) is used as an alternative to the Gini rule in Taylor and Silverman 1993. Statistical tests are also introduced as splitting criteria in Loh and Shin 1997 and Loh and Vanichsetakul 1988. Different splitting criteria are combined with a weighted sum in Shih 1999. A more recent work (see D'Ambrosio and Tutore 2011) proposes a new splitting criterion based on a weighted Gini impurity measure. Mola and Siciliano 1992 introduces a two-stage approach to find the

best split as to optimize a predictability function. On this approach is based the splitting rule proposed by Tutore, Siciliano, and Aria 2007, which introduce an instrumental variable called Partial Predictability Trees. In Cieslak et al. 2012 the Hellinger distance is used as splitting rules, this method is shown to be very efficient for imbalanced datasets but works only for binary target variables. See Fayyad and Irani 1992, Buntine and Niblett 1992 and Loh and Shin 1997 for a comparison of different splitting rules. Despite many different splitting rules have been proposed in literature, the most used in application problems are still the Information Gain and the Gini index and they are also used in literature as benchmark to compare the performance of new proposed splitting rules, see for example Chandra, Kothari, and Paul 2010 and Zhang and Jiang 2012.

In this paper a new measure of goodness of a split, based on an extension of polarization indices introduced by Esteban and Ray 1994, is proposed for classification tree modelling.

The contribution of the paper is twofold: from a methodological perspective a new multidimensional polarization measure is proposed; in terms of computation, a new algorithm for classification tree models is derived which the authors call *Polarized Classification Tree*. The new measure, based on polarization, tackles weaknesses of the classical measures used in classification trees (e.g. Gini index and Information Gain) by reflecting the distribution of each covariate in the node.

The rest of the paper is structured as follows: Section I.2 describes impurity and polarization measures; Section I.3 shows our methodological proposal; Section I.4 integrates the new measure inside decision tree algorithm. Section I.5 and Section I.6 reports the empirical evidences obtained on simulated and real datasets respectively. Conclusions and further ideas for research are summarized in Section I.7.

I.2 Impurity and polarization measures

In the literature on classification trees (see Mingers 1989), it is recognized that splitting rules based on the impurity measures (i.e. the Gini impurity index, the Information Gain) suffer from some weaknesses. Firstly, impurity measures are equivalent one to one another and they are also equivalent to random splits, in terms of the accuracy of the resulting model, see Mingers 1989. Secondly, impurity measures do not take into account the distribution of the features, but only the pureness of the descendant nodes in terms of the target variable and this fact could lead to an high dependence on the data at hand, see Aluja-Banet and Nafria 2003. The algorithms proposed in classification tree analytics tend to select the same variables for the splitting in different nodes, especially when these variables could be splitted in a variety of ways, making it difficult to draw conclusions about the tree structure.

As explained in previous section, the problem of finding an efficient splitting rules has been considered in different research papers. The aim of our contribution is to propose a new class of measures to evaluate the goodness of a split which tackles the previous mentioned weaknesses. In order to consider both the impurity and distribution of the features in the growth of the tree, our idea is to replace the impurity measure with a polarization index.

Polarization measures, introduced in Esteban and Ray 1994, Foster and Wolfson 1992 and Wolfson 1994, are typically adopted in the socio-economic context to measure inequality in income distribution. In Esteban and Ray 1994 and Duclos, Esteban, and Ray 2004 the authors provide an axiomatic definition for the class of polarization measures and a characterization theorem. In Esteban and Ray 1994, polarization is viewed as a clustering of an observed variable (typically ordinal) around an arbitrary number of local means, while in Duclos, Esteban, and Ray 2004, a definition of income polarization is proposed considering a continuous variable. In Esteban and Ray 1994 and Duclos, Esteban, and Ray 2004 the results of polarization measures are related to one variable, thus they can be considered univariate approaches.

In Zhang and Kanbur 2001 a multidimensional measure of polarization is proposed which considers within-groups inequality to capture internal heterogeneity and between-groups inequality to measure external heterogeneity. The index is composed by the ratio of the between groups and the within groups inequality. In Gigliarano and Mosler 2008 a general class of indices of multivariate polarization is derived starting from a matrix $\mathbb{X}$ of size $N \times K$, where N is the total number of individuals with their endowments classified in K attributes. The class of indices can be written as: $P(\mathbb{X}) = \zeta(B(\mathbb{X}), W(\mathbb{X}), S(\mathbb{X}))$ where B and W reflect the measure of between and within groups inequality respectively and S takes into account the size of each group. In details B and W can be chosen among different multivariate inequality indices present in literature, e.g. Tsui 1995 and Maasoumi 1986, and they can be applied only to variables that are transferable among individuals. ζ is a function $\mathbb{R}^3 \to \mathbb{R}$ increasing on B and S and decreasing on W. Gigliarano and Mosler 2008 discuss the possibility of extending the discrete version of the axioms proposed in Esteban and Ray 1994 to their proposed measure, stating also some properties of the measure.
Our idea is to define a multidimensional polarization measure, which considers one continuous variable when groups are exogenously defined coupled with a generalization of the continuous version of the axioms defined in Duclos, Esteban, and Ray 2004, opportunely adapted for our measure, as described in Section I.3 and proved in the Appendix.

I.3 A new impurity measure of polarization for classification analytic

Our measure of polarization is evaluated measuring the homogeneity/heterogeneity of the population with the use of variability between and within groups.
The new proposed index is a function of four inputs:

$$P(\mathbb{X}) = \zeta(B, W, \mathbf{p}, M) \tag{I.1}$$

where B and W are the variability between and within groups respectively, $\mathbf{p} = (p_1, ..., p_M)$ is the vector describing the proportion of elements in each group and M is the number of groups. Since we would like to introduce a measure which treats variables coming from different contexts (not only transferable variable), a measure of variability instead of inequality is introduced, thus making our proposal different from the one in Gigliarano and Mosler 2008.
Following the intuition on polarization, $P(\mathbb{X})$ is high for large values of B (i.e. the groups strongly differ from each other), for small values of W (i.e. the elements of the groups are homogeneous), for large values of $\max\{p_j\}$ and for small values of M (i.e. the population is divided into few groups with an unbalanced proportion of elements in one single group).
On the other hand, we expect $P(\mathbb{X})$ to take small values when the population is divided into numerous balanced groups with small variability between groups B and high variability within groups W.
Suppose that there are M groups exogenously defined, and that each observation is classified into one group through the use of a categorical variable with M levels. Let n_j be the number of individuals in the j^{th} group, N the total number of observations in the population, $p_j = \frac{n_j}{N}$ the proportion of population in the j^{th} group. Let f_j be the probability density function of the interesting feature x in the j^{th} group with expected value μ_j; the expected value of the global distribution f of the population is μ.
We set the following assumptions:

Hyp. 1 $M > 1$

Hyp. 2 $p_j > 0 \quad \forall j \in 1, ..., M$

Hyp. 3 $\{supp(f_j)\}_{j=1,...,M}$ are connected and
$\{supp(f_i)\} \cap \{supp(f_j)\} = \emptyset$ for $i, j = 1, \ldots, M$ with $i \neq j$.

Hyp. 4 $\int_{supp(f_j)} f_j \, dx = 1$.

Assumptions 1 and 2 exclude trivial cases, respectively a unique group for the entire population and the existence of empty groups.

Assumption 3 directly refers to the basic definition proposed in Duclos, Esteban, and Ray 2004 for the axiomatic theory of polarization measures. From an empirical point of view, assumption 3 translates the idea that the M groups of the original population are separated such that there is no uncertainty about the belonging of a single element to a certain group. As for the original definition of polarization measures, also in the case of multidimensional polarization measures, this assumption is not always verify in real application problems.
Assumption 4 requires that the functions f_j are probability densities; this assumption is necessary to provide an axiomatic definition of the polarization measure as pointed out in the appendix.
Our polarization measure is defined as follows.

Definition I.3.1. Given a population $\mathbb{X}$ and M groups, the polarization is:

$$P(B, W, \mathbf{p}, M) = \eta(B, W) \cdot \psi(\mathbf{p}, M) \tag{I.2}$$

where

$$\eta(B, W) = \frac{B}{B + W} = 1 - \frac{W}{B + W} \tag{I.3}$$

with

$$B = \sum_{j=1}^{M} (\mu_j - \mu)^2 \tag{I.4}$$

and

$$W = \sum_{j=1}^{M} \int_{supp\{f_j\}} (x - \mu_j)^2 f_j(x)\, dx \tag{I.5}$$

and

$$\psi(\mathbf{p}, M) = \frac{\max_{j=1,\dots,M} (p_j) - \frac{1}{N}}{\frac{N-2}{N}} \tag{I.6}$$

The measure proposed in Definition I.3.1 is the product of two components: $\eta(B, W)$ accounts for the variability between and within groups, while $\psi(\mathbf{p}, M)$ considers the number of the groups and their cardinality.
The measure P is normalized and takes values in the interval $[0, 1]$ as proved in the following proposition.

Proposition I.3.2. *Given a population* $\mathbb{X}$ *and* M *groups,* $P(B, W, \boldsymbol{p}, M) \in [0, 1]$.

Proof. Considering Definition I.3.1, the measure
$P(B, W, \mathbf{p}, M)$ is the product of the two components $\eta(B, W)$ and $\psi(\mathbf{p}, M)$.
The quantity $\eta(B, W)$ is defined as a ratio of the non-negative variability measures B and W, see equation I.3; by construction $\eta(B, W) \leq 1$. Moreover, the variability B is strictly positive, and using assumption 1 and 3 at least one of the elements in the sum defining B is strictly positive. So we have $\eta(B, W) \in (0, 1]$.
The quantity $\psi(\mathbf{p}, M)$ is a non-negative ratio; the minimum value is achieved when $\max_{j=1,\dots,M}(p_j) = \frac{1}{N}$ so that $\psi(\mathbf{p}, M) = 0$. The maximum value is obtained when $M = 2$ and $\max_{j=1,\dots,M}(p_j) = \frac{N-1}{N}$; in this case $\psi(\mathbf{p}, M) = 1$. In general, $\psi(\mathbf{p}, M) \in [0, 1]$.
As a consequence $P(B, W, \mathbf{p}, M) = \eta(B, W) \cdot \psi(\mathbf{p}, M) \in [0, 1]$ and the proposition is proved. ■

The following Corollary holds.

Corollary I.3.3. *The maximum and minimum values for* $P(B, W, \boldsymbol{p}, M)$ *are respectively equal to* 1 *and* 0 .

Proof. Trivial from Property I.3.2. ■

We note that $P(B, W, \mathbf{p}, M) = 0$ if and only if $\psi(\mathbf{p}, M) = 0$, or equivalently $\max_{j=1,\dots,M}(p_j) = \frac{1}{N}$. The condition is verified exclusively when $M = N$; considering assumption 2, this is the case where each group contains one single element of the original population supporting the intuition of absence of polarization.
On the other hand, note that $P(B, W, \mathbf{p}, M) = 1$ if and only if $\eta(B, W) = 1$ and $\psi(\mathbf{p}, M) = 1$. The condition on $\eta(B, W)$ requires $W = 0$ while $\psi(\mathbf{p}, M) = 1$ is equivalent to the case of $M = 2$ and one of the groups containing $N-1$ elements. In other words, the maximum polarization is achieved when the number of

groups is minimum, the original population except for one element belongs to one single group and the variance within groups is null such that the groups show maximum internal homogeneity.
Moreover, we should underline that the proposed measure is invariant for any permutation of the vector $\mathbf{p}$; intuitively the polarization of a population does not depend on the order in which we take the groups into account. We provide the axiomatic base for multidimensional polarization measures as a generalization of the axioms proposed by Duclos, Esteban, and Ray 2004.

Axiom I.3.4. *For any number of groups and any distribution of observations into the groups, a global squeeze (as defined in Duclos, Esteban, and Ray 2004) can not modify the polarization.*

Axiom I.3.4 requires the polarization measure to be invariant with respect to a global reduction of the variance of the population.

Axiom I.3.5. *If the population is divided symmetrically into three groups, each one composed of a basic density with the same root and mutually disjoint supports, then a symmetric squeeze of the side densities can not reduce polarization.*

Axiom I.3.5 requires the polarization measure to increase when the variability within groups W decreases. Note that the values of B, $\mathbf{p}$ and M are invariant with respect to the transformation described.

Axiom I.3.6. *Consider a symmetric distributed population divided into four groups, each one composed of a basic density with the same root and mutually disjoint supports. Slide the two middle densities to the side (keeping all supports disjointed). Then polarization must increase.*

Axiom I.3.6 requires the polarization measure to increase when the variability between groups B increases, when W, $\mathbf{p}$ and M are given.

Axiom I.3.7. *If $P_F \geq P_G$ and q is a non negative integer value, then $P_{qF} \geq P_{qG}$, where qF and qG represent population scaling of F and G respectively.*

Axiom I.3.7 describes a transformation that changes the sample size of the population with no consequences on the proportion of individuals in each group. In the appendix we prove that our proposal respects all four axioms, thus can be classified as a multidimensional polarization measure.

I.4 Polarized Classification tree

In this section we show how the multidimensional polarization measure introduced in Section I.3 can be used in classification tree models as a new measure of goodness of a split in the growth of a classification tree.
The new approach, which the authors call *Polarized Classification Tree* (PCT) has been implemented in R software. In Breiman, Friedman, and Olsen 1984 a split is defined as "good" if it generates "purer" descendant nodes then the goodness of a split criterion can be summarized from an impurity measure.
In our proposal a split is good if descendant nodes are more polarized, i.e. the polarization inside two sub-nodes is maximum. In order to evaluate the polarization in one sub-node as in I.1 we consider:

- The function $\psi(\mathbf{p}, M)$ which takes into account, the "pureness" of the sub-node. A sub-node is "purer" if one class of the target variable is more represented with respect to the others and the polarization is higher.

- The function $\eta(B, W)$ which measures homogeneity and heterogeneity among groups. $\eta(B, W)$, and consequently the polarization, is higher if the groups are "well characterized" by the variable X, selecting a split that obtains sub-nodes where the variable clearly discriminates well between different groups.

To clarify how our measure works with respect to the indices used in the literature a toy example is described.
As shown in Figure I.1, two explanatory variables X_1 and X_2 are considered. The target variable Y assumes three possible values *a, b* and *c*, corresponding to

three different groups. Figure I.1 shows the distribution of the two explanatory variables in the three groups determined by Y.
In this example the three groups are well distinguishable in both the distributions of X_1 and X_2, but it is evident that X_2 has an higher discriminatory power compared to X_1. The four best splits, in terms of pureness of the descendant

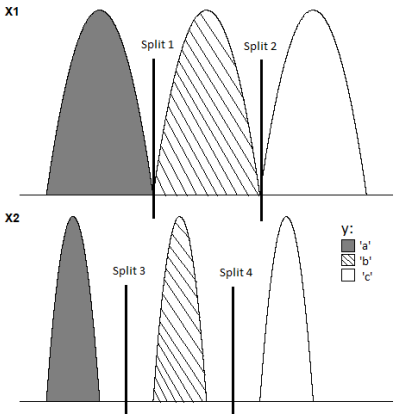


Figure I.1: Distributions of two explanatory variables for a three-class target variable.

nodes, are: Split 1 and 3, dividing group a from b and c respectively; Split 2 and 4, dividing a and b from c, as shown in Figure I.1. When evaluating the goodness of these possible splits, Gini and Information Gain criteria can not discriminate; indeed, when the tree is estimated on the training set, all the considered splits generate the same situation of impurity in the descendant nodes, thus making impossible to discriminate between the different splits.
When evaluating the goodness of the splits using our polarization measure, the distribution of the explanatory variables in the groups is taken into account. The goodness is higher for Split 3 and 4 with respect to Split 1 and 2, because the groups are more 'characterized' by variable X_2, thus leading to selecting a split on X_2 rather than on X_1.

Since classification trees can treat both numerical and categorical variables, we will extend the measure introduced in Section I.3 to deal with categorical

variables.

Consider a categorical variable X which assumes I different values, e.g. $X \in \{1, ..., I\}$ and suppose that there are M groups exogenously defined and each observation is assigned to a group.

Let n_{ij} be the number of observations taking value in the i^{th} category and assigned to the j^{th} group, $n_{i\cdot}$ be the number of observations taking value in the i^{th} category and $n_{\cdot j}$ be the number of observations assigned to the j^{th} group.

The polarization index can be written as in equation (I.2): $P(B, W, \mathbf{p}, M) = \eta(B, W) \cdot \psi(\mathbf{p}, M)$ where $W = \frac{N}{2} - \frac{1}{2} \sum_{j=1}^{M} \frac{1}{n_{\cdot j}} \sum_{i=1}^{I} n_{ij}^2$ and $B = M$.

Assumptions on the polarization index are described in Section I.3. We note that the theoretical definition of the measure requires that $M > 1$. Obviously this assumption can not always be satisfied in the computational stage when a pure node is obtained at some step. To handle this case we set $P(B, W, \mathbf{p}, M) = 1$ when $M = 1$. In addition some clarification has to be done on Assumption 3; from an empirical point of view this assumption reflects the idea that observing the distribution of a covariate we are able to clearly discriminate among the groups defined in the target variable. Of course, in real application problems, this assumption is not always satisfied. We show, in the empirical evaluation on both simulated and real datasets, that the relaxation of this hypothesis does not invalidate the performance of the proposed measure as splitting criteria.

Algorithm 1 shows the procedure used to build the PCT model. Let S be the set of all possible splits defined on the training set T. For each possible split $s \in S$, all samples can be divided into sub-node t_L^s the condition s is satisfied, otherwise t_R^s. The best split s^* is identified maximizing the polarization in the two sub-nodes. The growing procedure is stopped in one node if the node is pure in terms of target variable or if other stopping conditions are met (i.e. the number of samples in the node is less than a fixed threshold). Following the same procedure adopted in CART model, when the tree is built, the most representative class in each final node is assigned to that final node.

In the next sections we show how the proposed method works on different simulated and real datasets. Results obtained using the PCT model are

PCT: Polarized Classification Tree

Input: Training set T
if *T is "pure" OR other stopping conditions met* **then**
  return
end
forall *possible splits $s \in S$ referring to all attribute $x \in T$* **do**
  $s^* = \arg\max_{s \in S}(P(f_{x|t_L^s}, \mathbf{p}) + P(f_{x|t_R^s}, \mathbf{p}))$ *Select best split*;
end
PCT = Create a classification node from T based on s^* generating the two sub-nodes $t_L^{s^*}$ and $t_R^{s^*}$
for *each $t' \in \{t_L^{s^*}, t_R^{s^*}\}$* **do**
  PolTree$_{s^*}$ = PCT(T_{s^*});
  Attach Tree$_{s^*}$ to the corresponding branch of the tree
end

Algorithm 1: PCT: Polarized Classification Tree

compared to the ones obtained using the Gini index and the Information Gain measure as splitting rule, which are procedures most used as benchmark to compare new proposed splitting rules, as already underlined in Section I.1.

I.5 Empirical evaluation on simulated data

In order to show how our new impurity measure works inside PCT, this section reports the empirical results achieved on different simulated datasets. The performance of the PCT algorithm is compared with respect to the classification tree based on different splitting criteria. In particular the Polarization splitting criteria is compared to the Gini impurity index and the Information Gain in terms of the Area Under the ROC Curve (AUC) value. The results reported in the rest of the paper are based on a cross validation exercise and expressed in terms of out of sample performance.

The simulation framework considered in this paper is inspired by the paper of Loh and Shin 1997 where different impurity measures are compared for classification tree modeling. The data are sampled from four pairs of distributions that

are represented by the solid density curves represented in Figure I.2, where each distribution represents the covariate of a group G_i defined by the associated target variable. $N(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2, $T2(\mu)$ is a t-distribution with 2 degrees of freedom centered at μ and $Chisq(\nu)$ is a chi-square distribution with ν degrees of freedom. The 100 observations of the two groups represented by the target variable Y are sampled respectively from the first and from the second distribution as shown in Figure I.2. Results

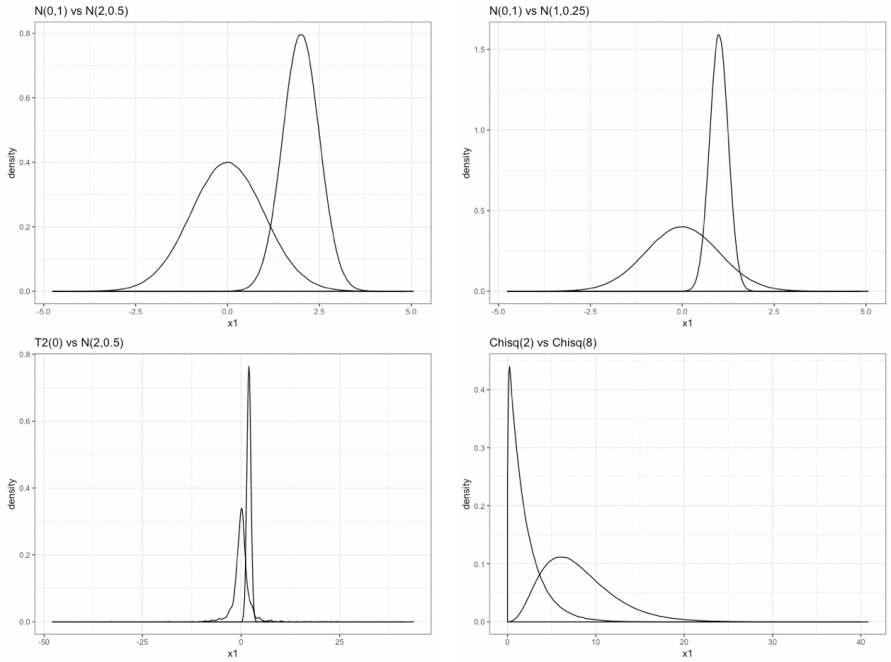


Figure I.2: Simulation and representation of the different class populations used for the classifiers comparison.

obtained by the three classification models under comparison are expressed in terms of the AUC value. Averaged AUC values (i.e. mean (AUC)) and the corresponding confidence intervals at 95% (i.e. CI (AUC)) for each simulated dataset obtained using Monte Carlo simulation with 100 iterations are reported in Table I.1.

In the reported examples AUC values obtained for PCT are better with respect to the classical splitting methods based on the Gini index and

Split Criteria	Distribution G1	Distribution G2	# Simulations	mean (AUC)	CI (AUC)
Polarization	N(0,1)	N(1,0.25)	100	0.909	(0.908;0.909)
Gini	N(0,1)	N(1,0.25)	100	0.872	(0.872;0.873)
Information Gain	N(0,1)	N(1,0.25)	100	0.882	(0.881;0.883)
Polarization	N(0,1)	N(2,0.5)	100	0.964	(0.964;0.965)
Gini	N(0,1)	N(2,0.5)	100	0.919	(0.918;0.920)
Information Gain	N(0,1)	N(2,0.5)	100	0.926	(0.925;0.927)
Polarization	T2(0)	N(2,0.5)	100	0.946	(0.945;0.946)
Gini	T2(0)	N(2,0.5)	100	0.910	(0.909;0.911)
Information Gain	T2(0)	N(2,0.5)	100	0.918	(0.917;0.919)
Polarization	Chisq(2)	Chisq(8)	100	0.955	(0.954;0.957)
Gini	Chisq(2)	Chisq(8)	100	0.922	(0.921;0.923)
Information Gain	Chisq(2)	Chisq(8)	100	0.928	(0.927;0.930)

Table I.1: Confidence intervals for AUC values obtained through a 100 iteration Monte Carlo procedure to compare the performance of classifiers on different simulated datasets

Information gain, as shown in Table I.1. In all cases the confidence intervals for the AUC derived using the polarization splitting criteria do not intersect those obtained using the Gini index and Information gain. For each simulated dataset a De Long test (E. R. DeLong, D. M. DeLong, and Clarke-Pearson 1988) is performed to compare obtained results, in terms of AUC, among PCT and trees employing respectively the Gini index and the Information Gain. In Table I.2 the average pvalue of the De Long test obtained along the 100 simulations for each dataset are shown. We also applied a one side Wilcoxon test to compare the AUC values obtained with PCT and decision trees employing Gini and Information Gain, in both cases obtained pvalues for all the datasets are lower than 0.05, showing that AUC values obtained with PCT are significantly higher. On the basis of the results at hand, the polarization measure introduced in

Distributions of G1 and G2	Average pvalue PCT vs Gini	Average pvalue PCT vs IG
N(0,1) ; N(1,0.25)	0.0357	0.0006
N(0,1) ; N(2,0.5)	0.0080	0.0033
T2(0) ; N(2,0.5)	0.0294	0.00002
Chisq(2) ; Chisq(8)	0.0070	0.0305

Table I.2: Average obtained pvalue of the De Long test to compare the AUC values of PCT against trees employing Gini index and Information Gain

this paper, shows a statistical significant superiority with respect to the other considered splitting criteria in terms of predictive performance observing the

obtained AUC values.

I.6 Empirical evaluation on real data

The performance of the splitting criteria under comparison are evaluated on 18 different real datasets. The considered datasets come from the UCI repository Dua and Graff 2017.
In order to have a complete comparison among classifiers, different datasets characterized by binary or multiple classes target variable are considered. The datasets are made up of categorical and/or numerical explanatory variables. In Table I.3 different information on the datasets are reported: sample size (Samples), total number of variables (Var), number of categorical (Cat) and numerical (Num) variables, number of classes in the target variable (Num Class) and the normalized Shannon entropy (Balance). The normalized Shannon entropy is evaluated on the target variable to measure the level of imbalance of each dataset (i.e. the value is equal to 0 if the dataset is totally unbalanced and equal to 1 if the samples are equi-distributed among the classes). See Appendix I.B for more details on the datasets.

A 10-fold cross-validation procedure for the datasets reported in Table I.3 is performed to evaluate the different approaches under comparison. All the classifiers are trained and evaluated on the same 10-folds. In addition, the same stopping condition is used for all the models, i.e. the minimum number of observations inside a node is set at 10% of the number of observations in the training set.
As suggested in Demsar 2006, since datasets are different, the evaluated performance metrics can not be compared directly, but for each dataset the metrics are used to rank the classifiers. On the basis of the AUC, each classifiers is ranked assigning value 1 to the best one, considering the mean value between two ranks if the classifiers perform equally. A Dunn test with Bonferroni correction is then applied to compare the obtained rankings with

Dataset	Samples	Var	Cat	Num	Num Class	Balance
banknotes	1372	4	0	4	2	0.99
breast	699	9	0	9	2	0.93
breast cancer	286	9	8	1	2	0.79
breast coimbra	116	9	0	9	2	0.99
car	1728	6	6	0	4	0.60
crx	690	15	10	5	2	0.99
fertility	100	9	6	3	2	0.52
glass	214	9	0	9	6	0.84
haberman	306	3	0	3	2	0.83
hepatitis	155	19	13	6	2	0.73
horse colic	300	27	17	10	2	0.91
krkp	3196	36	36	0	2	1
lymph	148	18	18	0	4	0.61
post operative	87	8	8	0	3	0.85
scale	625	4	4	0	3	0.83
sonar	208	60	0	60	2	1
spectheart	80	22	22	0	2	1
wine	178	13	0	13	3	0.99

Table I.3: Dataset descriptions

confidence at 95%. Table I.4 shows the ranking of each model registered on the datasets. The polarized classification tree works better with respect to Gini and Information Gain assuming different kind of target variables (i.e. *banknotes authentication* and *glass*). We note that classification trees based on the Gini index and Information Gain are superior in terms of performance for only two datasets each.

A Dunn test with Bonferroni correction show a significant difference between obtained results for PCT and Gini index (the adjust pvalue is equal to 0.03), while no differences are present between Information Gain and the other two splitting methods. Hence, we can affirm that PCT is competitive and sometimes better with respect to the most two used splitting rules (i.e. Gini index and Information Gain) and can be considered as a valid alternative to be employed and compared when looking for the model that better suits the data at hand. It can be noticed that PCT model obtains good performance when dataset covariates are mainly numerical, as they perform better or equal to the other methods (see for example *banknotes, glass or breast coimbra*). Obtained results

suggest instead that the balancing of the target variable and the presence of multiclass target variable do not influence the performance of the introduced method.

Dataset	Rank AUC		
	Gini	**InfoGain**	**Pol**
bank note authentication	3	2	1
breast	3	1.5	1.5
breast cancer	1.5	1.5	3
breast coimbra	2	3	1
car	1.5	1.5	3
crx	2	3	1
fertility	1	2	3
glass	3	2	1
haberman	3	1.5	1.5
hepatitis	3	2	1
horse colic	3	2	1
krkp	1.5	1.5	3
lymph	2.5	2.5	1
postoperative	1	2	3
scale	3	1.5	1.5
sonar	3	1	2
spectheart	3	1.5	1.5
wine	2.5	1	2.5
Mean Rank	2.36	1.83	1.80

Table I.4: Mean rank values for AUC for each classifier

I.7 Conclusions

This paper introduces a new index of polarization to measure the goodness of a split in the growth of a classification tree. Definition and properties of the new multidimensional polarization index are described in detail in the paper and proved in the appendix.

The new measure tackles weaknesses of the classical measures used in classification tree modeling, taking into account the distribution of each covariate in the node. From a computational point of view, the new measure proposed is evaluated inside a classification tree model and implemented in R software and is available from the authors upon request.

The results obtained in the simulation framework suggest that our proposal significantly outperforms classical impurity measure commonly adopted in classification tree modeling (i.e. Gini and Information Gain).
The performance registered running polarized classification tree models on real data extracted from the UCI repository, confirms the competitiveness of our methodological approach. More precisely, the empirical evidence at hand, shows that Polarized Classification Tree models are competitive and sometimes better with respect to classification tree models based on Gini or Information Gain.
A further analysis on this topic should compare the introduced Polarized Classification Trees with other splitting measures present in literature and to include this new splitting measure in ensemble three modelling (e.g. Random Forest).

Appendix I.A Appendix section

Let f be a basic density, as defined in Duclos, Esteban, and Ray 2004 , i.e. an unnormalized, symmetric and unimodal function, with compact support.
Some transformations can be performed on these functions s:

- λ-squeeze, with $\lambda \in (0,1)$ $f^{\lambda} = \frac{1}{\lambda} f\left(\frac{x-(1-\lambda)\mu}{\lambda}\right)$ where μ is the mean of f.

- δ-slide, $\delta > 0$ $g(x) = f(x \pm \delta)$

- population rescaling of a non negative integer q $g(x) = qf(x)$

- income rescaling to a new mean μ' $g(x) = \frac{\mu}{\mu'} f(\frac{x\mu}{\mu'})$

These transformations preserve symmetry and unimodality and the resulting transformed function is still a basic density.

On the basis of the Hypothesis 1-4, stated in Section I.3, we prove that the index introduced in this paper verifies the axiomatic definition of polarization given in Section I.3.
Some preliminary observations are needed for the proof. Let f be a density function of a continuous. Suppose that $supp\, f = [a, b]$ and μ is the expected

value of the population.

Let f^λ be the squeeze of f with $\lambda \in (0,1)$, then:

Observation I.A.1. *The support of* f^λ *is:* $supp\, f^\lambda = [\lambda a + (1-\lambda)\mu, \lambda b + (1-\lambda)\mu] \subset [a,b]$

Observation I.A.2. $\int_{\lambda a+(1-\lambda)\mu}^{\lambda b+(1-\lambda)\mu} \frac{1}{\lambda} f(\frac{x-(1-\lambda)\mu}{\lambda})\, dx = \int_a^b f(x)\, dx = 1$

Observation I.A.3. $\mu' = \int_{\lambda a+(1-\lambda)\mu}^{\lambda b+(1-\lambda)\mu} x \frac{1}{\lambda} f(\frac{x-(1-\lambda)\mu}{\lambda})\, dx = \int_a^b (\lambda y + (1-\lambda)\mu) f(y)\, dy =$

$= \lambda\mu + (1-\lambda)\mu = \mu$

Observation I.A.4. $V(f^\lambda) = \int_{\lambda a+(1-\lambda)\mu}^{\lambda b+(1-\lambda)\mu} (x-\mu)^2 \frac{1}{\lambda} f(\frac{x-(1-\lambda)\mu}{\lambda})\, dx =$

$= \lambda^2 \int_a^b (y-\mu)^2 f(y)\, dy = \lambda^2 V(f)$

Axiom 1

Let f_j be the density function of each group $j = 1, ..., M$. Since by assumption the f_j have disjoint supports, we can define the global distribution as $f = \frac{1}{M}\Big(f_1 + ... + f_M\Big)$. A global squeeze on the entire population is defined as:

$$f^\lambda = \frac{1}{M\lambda} f\big(\frac{x-(1-\lambda)\mu}{\lambda}\big) = \frac{1}{M\lambda} f_1\big(\frac{x-(1-\lambda)\mu}{\lambda}\big) + ... + \frac{1}{M\lambda} f_M\big(\frac{x-(1-\lambda)\mu}{\lambda}\big) =$$
$$= \frac{1}{M}\Big(f_1^\lambda + ... + f_M^\lambda\Big)$$

If $supp(f_j) = [a_j, b_j]$, then $supp(f_j^\lambda) = [\lambda a_j + (1-\lambda)\mu, \lambda b_j + (1-\lambda)\mu]$ (for Obs. I.A.1). The mean of each group is defined as: $\mu_j = \int_{a_j}^{b_j} x f_j(x)\, dx$ The mean of each group after the squeeze becomes: $\mu_j' = \frac{1}{\lambda} \int_{\lambda a_j+(1-\lambda)\mu}^{\lambda b_j+(1-\lambda)\mu} x f_j(\frac{x+(1-\lambda)\mu}{\lambda})\, dx =$

$\int_{a_j}^{b_j} (\lambda y + (1-\lambda)\mu) f_j(y)\, dy$

$= \lambda \int_{a_j}^{b_j} y f_j(y)\, dy + (1-\lambda)\mu \int_{a_j}^{b_j} f_j(y)\, dy = \lambda\mu_j + (1-\lambda)\mu$ So we can evaluate the variability between groups after the squeeze as follow: $B' = \sum_j (\mu_j' - \mu)^2 =$

$\sum_j (\lambda\mu_j + (1-\lambda)\mu - \mu)^2 = \sum_j (\lambda\mu_j - \lambda\mu)^2 = \lambda^2 B$ The variability within groups after the squeeze becomes: $W' = \sum_j \frac{1}{\lambda} \int_{\lambda a_j+(1-\lambda)\mu}^{\lambda b_j+(1-\lambda)\mu} (x-\mu_j')^2 f_j\big(\frac{x+(1-\lambda)\mu}{\lambda}\big)\, dx =$

$\sum_j \int_{a_j}^{b_j} (\lambda y + (1-\lambda)\mu - \lambda\mu_j - (1-\lambda)\mu)^2 f_j(y)\, dy =$

$\sum_j \int_{a_j}^{b_j} \lambda^2 (y - \mu_j)^2 f_j(y)\, dy = \lambda^2 W$ So the polarization becomes: $P(B', W', \mathbf{p}, M) = \frac{B'}{B'+W'} \cdot \psi(\mathbf{p}, M) = \frac{\lambda^2 B}{\lambda^2 B + \lambda^2 W} \cdot \psi(\mathbf{p}, M) = P(B, W, \mathbf{p}, M)$

Axiom 1 is proved .

Axiom 2

Let f_1, f_2, f_3 be three basic densities of the population corresponding to three different groups and P the total polarization value. The global distribution is completely symmetric, so groups 1 and 3 have the same population and group 2 is exactly midway between them. If we operate the same squeeze to f_1 and f_3, we can prove that the polarization value is not decreasing.

First, it is possible to observe that as the squeeze is performed on f_1 and f_3 separately, the expected values μ_1 and μ_3 do not change (for Obs. I.A.4). $P(B, W, \mathbf{p}, M) = \frac{B}{B+W} \cdot \psi(\mathbf{p}, M)$ where $W = \sum_{j=1}^{3} \int_{suppf_j} (x-\mu_j)^2 f_j(x)\, dx$ and $P'(B, W', \mathbf{p}, M) = \frac{B}{B+W'} \cdot \psi(\mathbf{p}, M)$ where $W' = K_2 \int_{suppf_2} (x-\mu_2)^2 f_2(x)\, dx +$
$\lambda^2 \Big(K_1 \int_{suppf_1} (x-\mu_1)^2 f_1(x)\, dx$
$+ K_3 \int_{suppf_3} (x-\mu_3)^2 f_3(x)\, dx \Big) < W$
So we can conclude that $P'(B, W', \mathbf{p}, M) \geq P(B, W, \mathbf{p}, M)$.

Axiom 3

Let f_1, f_2, f_3, f_4 be four basic densities referred to four different groups, with mutually disjoint supports, and let the distribution of the entire population be completely symmetric. A symmetric slide of f_2 and f_3 to the side must increase the polarization.

Before the slide: $P(B, W, \mathbf{p}, M) = \Big(1 - \frac{W}{B+W}\Big) \cdot \psi(\mathbf{p}, M)$ where $B = \sum_{j=1}^{4} (\mu_j - \mu)^2 = (\mu_1 - \mu)^2 + (\mu_2 - \mu)^2 + (\mu_3 - \mu)^2 + (\mu_4 - \mu)^2$ After the slide, because of the symmetry of the transformation, the global mean μ does not change while the means of f_2 and f_3 become respectively $\mu_2 - \delta$ and $\mu_3 + \delta$.
So we obtain: $B' = (\mu_1 - \mu)^2 + (\mu_2 - \delta - \mu)^2 + (\mu_3 + \delta - \mu)^2 + (\mu_4 - \mu)^2$
$= (\mu_1 - \mu)^2 + (\mu_2 - \mu)^2 + \delta(\delta - 2\mu_2 + 2\mu)$
$+ (\mu_3 - \mu)^2 + \delta(\delta + 2\mu_3 - 2\mu) + (\mu_4 - \mu)^2$ where $\delta(\delta - 2\mu_2 + 2\mu) > 0$ and $\delta(\delta + 2\mu_3 - 2\mu) > 0$ under the hypothesis that $\mu_2 < \mu$ and $\mu_3 > \mu$ and $B' > B$.
Thus we obtain that $P'(B', W, \mathbf{p}, M) > P(B, W, \mathbf{p}, M)$.

Axiom 4

Considering two different distributions referred to the same population, the function $\eta(B, W)$ is not affected by the scaling transformation of the two distributions.

So if $P_F(B_F, W_F, \mathbf{p}_F, M) > P_G(B_G, W_G, \mathbf{p}_G, M)$,

then $\frac{\max_j p_j^F - \frac{1}{N}}{\frac{N-2}{N}} > \frac{\max_j p_j^G - \frac{1}{N}}{\frac{N-2}{N}}$.

Then we can trivially show that

$P_{qF}(B_{qF}, W_{qF}, \mathbf{p}_{qF}, M) > P_{qG}(B_{qG}, W_{qG}, \mathbf{p}_{qG}, M)$ with q a non negative integer value. Indeed: $\frac{\max_j p_j^{qF} - \frac{1}{qN}}{\frac{qN-2}{qN}} > \frac{\max_j p_j^{qG} - \frac{1}{qN}}{\frac{qN-2}{qN}}$. We conclude that the measure proposed is a multidimensional polarization measure.

Appendix I.B Appendix section

The performance of the splitting criteria under comparison are evaluated on 18 different real datasets, coming from the UCI repository (Dua and Graff 2017). In this section detailed information on each dataset are reported.

Banknotes authentication Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

Breast The dataset contains information about samples that arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data and contains 8 groups of patients.

Breast cancer This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature (see also lymphography and primary-tumor). It contains clinical informations about patient with breast cancer. This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Breast coimbra The dataset contains clinical features that were observed or measured for 64 patients with breast cancer and 52 healthy controls. There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis. Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

Car Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX, as described in Bohanec and Rajkovic 1990. The model evaluates cars according to the following concept structure: car acceptability is estimate by overall price (that is divided in buying price and price of the maintenance) and technical characteristics detailed as confort (number of doors, capacity in terms of persons to carry and size of luggage boot) and safety.

CRX This dataset contains informations that concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. This dataset is interesting because there is a good mix of attributes: continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values.

Fertility 100 volunteers provide a semen sample analyzed according to the WHO 2010 criteria. Sperm concentration are related to socio-demographic data,

environmental factors, health status, and life habits

Glass This data are from USA Forensic Science Service; it contains 6 types of glass defined in terms of their oxide content (i.e. Na, Fe, K, etc). The study of classification of types of glass was motivated by criminological investigation.

Haberman The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Hepatitis This dataset contains medical information about a group of 155 people with acute and chronic hepatitis, initially studied by Peter B. Gregory of the Stanford University School of Medicine. Among this 155 patients 33 died and 122 survived, and for each of them 19 variables, such as age, sex and the results of standard biochemical measurements, are collected. The aim of the dataset is to discover whether the data could be combined in a model that could predict a patient's chance of survival. See Diaconis and Efron 1983.

Horse Colic This dataset contains health information about horses in order to predict whether or not a horse can survive, based upon past medical conditions.

Krkp The Chess Endgame Database for White King and Rook against Black King (KRK) contains information on chess end game, where a pawn on a7 is one square away from queening. The main aim is to predict the outcome of the chess endgames, thus the target variable contains two possible values: White-can-win ("won") and White-cannot-win ("nowin").

Lymph This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. The aim of this dataset is to make a lymphatic diseases diagnosis observing different information extracted through medical imaging techniques; four different diagnosis are possible: normal, arched, deformed, displaced.

Post Operative Because hypothermia is a significant concern after surgery, in this dataset different attributes which correspond roughly to body temperature measurements are collected from 87 different patients. The aim of this dataset is to determine where patients in postoperative recovery area should be sent to next. In particular three different decisions can be taken: I (patient sent to Intensive Care Unit), S (patient prepared to go home) and A (patient sent to general hospital floor).

Scale In this dataset results of a psychological experiment are collected observing tips of 625 patients. Four attributes are collected for each sample: the left weight, the left distance, the right weight, and the right distance. Each example is then classified as having the balance scale tip to the right, tip to the left, or be balanced.

Sonar This dataset is composed by 208 sonar signals bounced off a metal cylinder or a roughly cylindrical rock. For each signal we have a set of 60 numbers in the range 0.0 to 1.0, representing the energy within a particular frequency band, integrated over a certain period of time. The integration aperture for higher frequencies occur later in time, since these frequencies are transmitted later during the chirp. The target variable associated to each record contains the letter "R" if the signal is bounced off a rock and "M" if it is bounced off a mine (metal cylinder).

Spectheart Diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images are describer in the dataset. The database of 80 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature pattern was created for each patient and then each pattern was further processed to obtain 22 binary feature patterns. Each of the patients is classified into two categories: normal and abnormal, contained in the target variable.

Wine The wine dataset contains the results of a chemical analysis performed on three different types of wines grown in a specific area of Italy. 178 samples are analysed and 13 different attributes are recorded for each sample. The target variable is a three classes categorical variable representing the analysed type of wine.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*. Vol. 656. John Wiley & Sons.

Ahmad, A. and Brown, G. (2015). "Random Ordinality Ensembles: Ensambles methods for multi-valued categorical data". In: *Information Sciences* vol. 296, pp. 75–94.

Aluja-Banet, T. and Nafria, E. (2003). "Stability and scalability in decision trees". In: *Computational Statistics* vol. 18, no. 3, pp. 505–520.

Aria, M. et al. (2018). "Dynamic recursive tree-based partitioning for malignant melanoma identification in skin lesion dermoscopic images". In: *Statistical papers*, pp. 1–17.

Bohanec, M. and Rajkovic, V. (1990). "DEX: An Expert System Shell for Decision Support". In: *Sistemica* vol. 1, pp. 145–157.

Breiman, L., Friedman, J., and Olsen, R. (1984). *Classification and Regression Trees*.

Buntine, W. and Niblett, T. (1992). "A Further Comparison of Splitting Rules for Decision-Tree Induction". In: *Machine Learning* vol. 8, pp. 75–85.

Cardoso, J. and Sousa, R. (2011). "Measuring the Performance of Ordinal Classification". In: *International Journal of Pattern Recognition and Artificial Intelligence* vol. 25, no. 08, pp. 1173–1195.

Chandra, B., Kothari, R., and Paul, P. (2010). "A new node splitting measure for decision tree construction". In: *Pattern Recognition* vol. 43, no. 8, pp. 2725–2731.

Ciampi, A et al. (1987). “Recursive partitioning: a versatile method for exploratory data analysis in biostatistics”. In: *Biostatistics. The University of Western Ontario Series in Philosophy of Science*, pp. 23–50.

Cieslak, D. A. et al. (2012). “Hellinger distance decision trees are robust and skew-insensitive”. In: *Data Mining and Knowledge Discovery* vol. 24, no. 1, pp. 136–158.

Clark, L. A. and Pregibon, D. (2017). “Tree-based models”. In: *Statistical Models in S*, pp. 377–419.

D’Ambrosio, A., Aria, M., et al. (2017). “Regression trees for multivalued numerical response variables”. In: *Expert Systems with Applications* vol. 69, pp. 21–28.

D’Ambrosio, A. and Tutore, V. A. (2011). “Conditional Classification Trees by Weighting the Gini Impurity Measure”. In: *New Perspectives in Statistical Modeling and Data Analysis. Studies in Classification, Data Analysis and Knowledge Organization*, pp. 377–419.

DeLong, E. R., DeLong, D. M, and Clarke-Pearson, D. L. (1988). “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach”. In: *Biometrics*, pp. 837–845.

Demsar, J. (2006). “Statistical comparisons of classifiers over multiple data sets”. In: *Journal of Machine Learning* vol. 7, pp. 1–30.

Diaconis, P. and Efron, B. (1983). “Computer-Intensive Methods in Statistics”. In: *Scientific American* vol. 248.

Dua, D. and Graff, C. (2017). *UCI Machine Learning Repository*.

Duclos, J. Y., Esteban, J. M., and Ray, D. (2004). “Polarization: Concepts, Measurement, Estimation”. In: *Econometrica* vol. 72, no. 6, pp. 1737–1772.

Esteban, J. M. and Ray, D. (1994). “On the Measurement of Polarization”. In: *Econometrica* vol. 62, no. 4, pp. 819–851.

Fayyad, U. M. and Irani, K. B. (1992). “The attribute selection problem in decision tree generation”. In: *AAAI*, pp. 104–110.

Foster, J. and Wolfson, M. C. (1992). “Polarization and the Decline of the Middle Class: Canada and the US”. In: *OPHI Working Paper, University of Oxford* vol. 31.

Galimberti, G., Soffritti, G., and Di Maso, M. (2012). “Classification trees for ordinal responses in R: The rpartScore package”. In: *Journal of Statistical Software* vol. 47.

Gigliarano, C. and Mosler, K. (2008). “Constructing Indices of Multivariate Polarization”. In: *The Journal of Economic Inequality* vol. 7, pp. 435–460.

Goodman, L. A. and Kruskal, W. H. (1979). “Measures of association for cross classifications”. In: *Measures of association for cross classifications*. Ed. by Springer, pp. 2–34.

Hornung, R. (2020). “Ordinal forests”. In: *Journal of Classification* vol. 37, pp. 4–17.

Iorio, C. et al. (2019). “Informative trees by visual pruning”. In: *Expert Systems with Applications* vol. 127, pp. 228–240.

Kotłowski, W. et al. (2008). “Stochastic dominance-based rough set model for ordinal classification”. In: *Information Sciences* vol. 178, no. 21, pp. 4019–4037.

Loh, W.-Y. and Shin, Y.-S (1997). “Split selection methods for classification trees”. In: *Statistica Sinica* vol. 7, pp. 815–840.

Loh, W.-Y. and Vanichsetakul, N. (1988). “Tree-structured classification via generalized discriminant analysis”. In: *Journal of the American Statistical Association* vol. 83, no. 403, pp. 715–725.

Mingers, J. (1989). “An Empirical Comparison of Selection Measures for Decision-Tree Induction”. In: *Machine Learning* vol. 3, no. 4, pp. 319–342.

Mola, F. and Siciliano, R. (1992). “A Two-Stage Predictive Splitting Algorithm in Binary Segmentation”. In: *Computational Statistics*. Ed. by Dodge, Yadolah and Whittaker, Joe. Heidelberg: Physica-Verlag HD, pp. 179–184.

— (1997). “A fast splitting procedure for classification trees”. In: *Statistics and Computing* vol. 7, pp. 209–216.

Morrone, A., Piscitelli, A., and D'Ambrosio, A. (2019). "How disadvantages shape life satisfaction: an alternative methodological approach". In: *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement* vol. 141, no. 1, pp. 477–502.

Maasoumi, E. (1986). "The measurement and decomposition of multi-dimensional inequality". In: *Econometrica: Journal of the Econometric Society*, pp. 991–997.

Nerini, D. and Ghattas, B. (2007). "Classifying densities using functional regression trees: Applications in oceanology". In: *Computational Statistics & Data Analysis* vol. 51, no. 10, pp. 4984–4993.

Piccarreta, R. (2004). "Ordinal Classification Trees Based on Impurity Measures". In.

Quinlan, J. R. (2014). *C4.5: programs for machine learning*. Elsevier.

Shih, Y.S. (1999). "Families of splitting criteria for classification trees". In: *Statistics and Computing* vol. 9, no. 4, pp. 309–315.

Shneiderman, B. (1992). "Tree visualization with tree-maps: 2-d space-filling approach". In: *ACM Transactions on graphics (TOG)* vol. 11, no. 1, pp. 92–99.

Taylor, P. C. and Silverman, B. W. (1993). "Block diagrams and splitting criteria for classification trees". In: *Statistics and Computing* vol. 3, no. 4, pp. 147–161.

Torra, V. et al. (2006). "Regression for ordinal variables without underlying continuous variables". In: *Information Sciences* vol. 176, no. 4, pp. 465–474.

Tsui, K.-Y. (1995). "Multidimensional generalizations of the relative and absolute inequality indices: the Atkinson-Kolm-Sen approach". In: *Journal of Economic Theory* vol. 67, no. 1, pp. 251–265.

Tutore, V. A., Siciliano, R., and Aria, M. (2007). "Conditional classification trees using instrumental variables". In: *International Symposium on Intelligent Data Analysis*. Ed. by Springer, pp. 163–173.

Wolfson, M. C. (1994). "When Inequalities Diverge". In: *The American Economic Review* vol. 84, no. 2, pp. 353–358.

Zhang, X. and Jiang, S. (2012). “A Splitting Criteria Based on Similarity in Decision Tree Learning”. In: *Journal of Software* vol. 7, pp. 1775–1782.

Zhang, X. and Kanbur, R. (2001). “What difference do polarisation measures make? An application to China”. In: *Journal of development studies* vol. 37, no. 3, pp. 85–98.

Part 2

Model selection methods for ordinal target variables

Introduction

Classification and regression are among the founding tasks of the machine learning field. A lot of tools are developed in literature and applied in many different domains. In this context, less attention was paid to tasks where the target variable is ordinal.

Ordinal data are those categorical data where a natural order exists between levels. This kind of data was often considered as pure nominal or converted to numeric to reflect the natural order of categories. Increasing the number of real problems that involve this type of data, the development of suitable tools is attracting more and more interest.

Concerning ordinal response variables modelling, different approaches are described in literature, both parametric (see Agresti 2010; Kotłowski et al. 2008; Torra et al. 2006) and non-parametric (see Ahmad and Brown 2015; Galimberti, Soffritti, and Di Maso 2012; Hornung 2020; Morrone, Piscitelli, and D'Ambrosio 2019; Piccarreta 2004), for the model selection stage there is a lack of adequate tools (Cardoso and Sousa 2011).

Moreover, performance indicators should take into account the nature of the target variable, especially when the dependent variable is ordinal. This motivates the proposal of a new class of measures to select the best model in predictive contexts characterized by a multi-class ordinal target variable, using the misclassification errors coupled with a measure of uncertainty on the prediction.

Two toy examples show how the index works and its advantages with respect to the classical evaluation measures deployed in literature in this context (accuracy, AUC, MSE). The index is applied on a simulated data set and compared with the cited classical measures in the evaluation of the performances of different models. Results confirm that the new index can capture peculiar aspects compared to the traditional measures.

The index proposed is also deployed in a real case study. A data set related to the study of Attenuated Psychosis Syndrome is analysed in terms of classifica-

tion task.

The main contribution of this work is the definition of a new class of measures to select the best model in predictive contexts characterized by a multi-class ordinal target variable, using the misclassification errors coupled with a measure of uncertainty on the prediction. This new approach takes into account the ordered nature of the variable and, in addition, takes into account the uncertainty that the model assigns to the prediction in order to obtain the maximum of interpretability. This second aspect is particularly important in medical applications.

In the Paper II we propose the new method for model selection when the target variable is ordinal. Paper III presents an application of the index to a real case study in the neuropsychiatric field: a classification task on three different groups of subjects (subject with psychosis, subjects at risk and subjects not at risk) is performed.

Paper II

A new approach in model selection for ordinal target variables

Elena Ballante[1], Silvia Figini[2], Pierpaolo Uberti[3]

Published in *Computational Statistics*, May 2021, DOI: 10.1007/s00180-021-01112-4.

Abstract

Multi-class predictive models are generally evaluated averaging binary classification indicators without a distinction between nominal and ordinal dependent variables. This paper introduces a novel approach to assess performances of predictive models characterized by an ordinal target variable and a new index for model evaluation is proposed. The new index satisfies mathematical properties and it can be applied to the evaluation of parametric and non parametric models. In order to show how our performance indicator works, empirical evidences obtained on toy examples and simulated data are provided. On the basis of the results achieved, we underline that our approach can be a more suitable criterion for model selection than the performance indexes currently suggested in the literature

Keywords: Classification, Ordinal Data, Performance Index, Model Assessment

[1] Department of Mathematics, University of Pavia, Pavia, Italy
[2] Department of Political and Social Sciences, University of Pavia, Pavia, Italy
[3] Department of Economics, University of Genova, Genova, Italy

II.1 Introduction

Evaluation measures are widely used in predictive models to compare different algorithms, thus providing the selection of the best model for the data at hand. Performance indicators can be used to assess the performance of a model in terms of accuracy, discriminatory power and stability of the results. The choice of indicators to perform model selection is a fundamental point and many approaches have been proposed over the years (see e.g. Bradley 1997, Adams and Hand 2000, Hand 2009).

Concerning binary target variables, distinct criteria to compare the performance of classification models are available (see Hand 1997, Hand 2001, Sokolova, Japkowicz, and Szpakowicz 2006, Hossin and Sulaiman 2015).

Multi-class classification models are generally evaluated averaging binary classification indicators (see Hand and Till 2001, Sokolova and Lapalme 2009, Hossin and Sulaiman 2015) and in literature there is not a clear distinction among them with respect to multi-class nominal and ordinal targets (e.g. Frank and Hall 2001, Pang and Lee 2005, Gaudette and Japkowicz 2009).

In the model definition stage for ordinal response variables there are different approaches described in literature, both parametric (see Agresti 2010; Kotłowski et al. 2008; Torra et al. 2006) and non-parametric (see Ahmad and Brown 2015; Hornung 2020; Morrone, Piscitelli, and D'Ambrosio 2019; Piccarreta 2004), while for the model selection stage there is a lack of adequate tools.

In our opinion, performance indicators should take into account the nature of the target variable, especially when the dependent variable is ordinal. This leads us to propose a new class of measures to select the best model in predictive contexts characterized by a multi-class ordinal target variable, using the misclassification errors coupled with a measure of uncertainty on the prediction.

The paper is structured as follows: Section II.2 reviews the metrics most used in literature; Section II.3 shows our methodological proposal and proves mathematical properties; Section II.4 explains how our proposed index works in two toy examples; Section II.5 reports the empirical evidence obtained on

simulated data. Conclusions and further research ideas are summarized in Section II.6.

II.2 Review of the literature for ordinal dependent variables

The most popular measures of performance in ordinal predictive classification models are based on AUC (Area Under the Receiver Operating Characteristic (ROC) Curve), accuracy (expressed in terms of correct classification) and MSE (Mean Square Error), see Gaudette and Japkowicz 2009 and Huang and Ling 2007 among others. The accuracy, measured as percentage of correct predictions over total instances, is the most used evaluation metric for binary and multi-class classification problems (Sokolova, Japkowicz, and Szpakowicz 2006), assuming that the costs of the different misclassifications are equal.

The AUC for multi-class classification is defined in Hand and Till 2001 as a generalization of the AUC (based on the probabilistic definition of AUC); it suffers of different weaknesses also in the binary classification problem (Gigliarano, Figini, and Muliere 2014) and it is cost-independent, assumption that can be viewed as a weakness when the target is ordinal.

The mean square error (MSE) measures the difference between prediction values and observed values in regression problems using an Euclidean distance. MSE can be used in ordinal predictive models, converting the classes of the ordinal target variable y in integers and computing the difference between them and it does not take into account the ordering in a predictive model characterized by ordinal classes in the response variable.

Furthermore, it is well known that in imbalanced data characterized by underfitting or over-fitting the mean square error could provide trivial results (see Hossin and Sulaiman 2015).

II.3 A new index for model performances evaluation and comparison for ordinal target

Let $\mathbf{y} = \{y_1, .., y_N\}$ be a test set for the ordinal target variable Y, where $y_i \in \{1, ..., M\}$ (with M number of classes ordered of the target variable) and let $\mathbb{X}$ be the $N \times p$ data matrix, where N is the number of observations and p the number of covariates.
The output of a predictive model is a matrix $P = \{p_{ij}\}$, where $0 \leq p_{ij} \leq 1$, which contains the probability that observation i belong to the class j estimated by the model under evaluation.
Standard multi-class classification rules assign the observation i to the class $j = \text{argmax}_l\{p_{i,l}\}$.
In order to introduce our proposal, the definitions of classification function and error interval are required.

Definition II.3.1 (Classification function)**.** Let observations $\{1, ..., N\}$ be grouped by the estimated classes $\hat{y_i} = j$. For each class, sort the observations in a non-increasing order with respect to $p_{i,j}$. The vector of indexes i of the observations is a permutation of the original vector, according to the ordering defined above. For a given model, the classification function is a piecewise constant function $f_{mod} : [0, 1] \to \{1, .., M\}$ such that $f_{mod}([\frac{i-1}{N}, \frac{i}{N})) = y_i$ for $i \in \{1, ..., N\}$.

As a special case, the *perfect classification function*, is a piecewise constant function $f_{exact} : [0, 1] \to \{1, .., M\}$ such that each estimated class corresponds to the real class identified by $\mathbf{y}$.
Note that the function f_{exact} is unique except for permutation of the observations in the same estimated class.

The error interval in each class can be derived as the interval between the first misclassified observation and the end of the observations in that estimated

class.

Definition II.3.2 (Error Interval)**.** Considered the vector of observations ordered as described in Definition II.3.1. Suppose that the range corresponding to the estimated class j in that vector has indexes in $[n_{j-1}, n_j)$. Let $\tilde{i_j} \in \{n_{j-1}, ..., n_j\}$ the index of the first misclassified observation. So the error interval is defined as $[\frac{\tilde{i_j}}{N}, \frac{n_j}{N})$, i.e. the interval between the first misclassified observation and the last observation of the estimated class j, and its length is defined as $e_j = \frac{n_j - \tilde{i_j}}{N}$.
If no misclassification occurs in $[n_{j-1}, n_j)$, the error interval is defined as an empty set and the length is $e_j = 0$.

Consider an artificial example. Let $N = 10$ be the number of observations and each of these belongs to a class defined by a three levels target variable ($M = 3$). Suppose that a (hypothetical) predictive model returns the predictions as in Table II.1.

Observation	Probabilities			Estimated Class	Real Class
	Class 1	Class 2	Class 3		
1	0.288	0.174	**0.538**	3	1
2	0.325	**0.478**	0.197	2	2
3	**0.828**	0.013	0.159	1	1
4	0.310	0.106	**0.584**	3	3
5	0.120	0.262	**0.618**	3	3
6	**0.426**	0.167	0.407	1	3
7	**0.849**	0.126	0.025	1	2
8	**0.520**	0.401	0.079	1	1
9	0.147	**0.670**	0.183	2	2
10	0.142	**0.593**	0.265	2	3

Table II.1: Example. The probabilities are randomly generated, the estimated class is the class with the maximum of probability assigned, the real class are generated starting from the estimated class with some classification errors artificially introduced.

The classification function is derived grouping the observations in the estimated class as: {3,6,7,8} in Class 1, {2,9,10} in Class 2 and {1,4,5} in Class 3. In each group the observations are sorted with respect to the probability of

the estimated class. For the group 1 the probabilities are 0.828, 0.426, 0.849, 0.520 respectively, then the ordered group is: {7,3,8,6}. Following the same rule the group 2 becomes {9,10,2} and group 3 is {5,4,1}.
The final sequence of observations can be written as in Table II.2.

i	7	3	8	6	9	10	2	5	4	1
i	1	2	3	4	5	6	7	8	9	10
y	2	1	1	3	2	3	2	3	3	1
$\hat{y}$	1	1	1	1	2	2	2	3	3	3

Table II.2: Index construction

The classification function and the corresponding perfect classification function are depicted in Figure II.1 and Figure II.2 respectively.

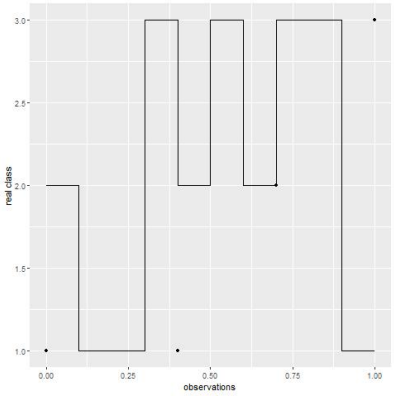


Figure II.1: Classification function

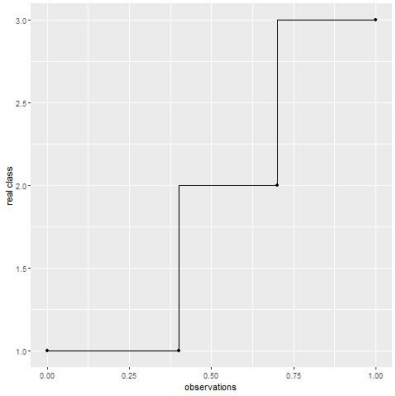


Figure II.2: Perfect classification function

In order to define the three error intervals, as a preliminary step we identify the intervals of observations related to each estimated class: $[0, 0.4)$ for Class 1, $[0.4, 0.7)$ for Class 2, $[0.7, 1)$ for Class 3. From Table II.2, in the estimated Class 1 the first error corresponds to the first observation, so the error interval is $[0, 0.4)$, in the estimated Class 2 the first error corresponds to the observation 6, then the error interval is $[0.5, 0.7)$ and in the estimated Class 3

the first error corresponds to the observation 10 and the error interval is $[0.9, 1)$.

Starting from Definition II.3.1 and Definition II.3.2, Definition II.3.3 introduces a new index for model performance evaluation in predictive models characterized by an ordinal target variable.

Definition II.3.3 (Index)**.** Consider for each class $\{1, ...M\}$ the corresponding weight $w_j = \frac{e_j}{l_j}$, where e_j is the j^{th} error interval length and $l_j = n_j - n_{j-1}$ is the length of the j^{th} estimated class in the domain, such that $0 \leq w_j \leq 1$. We define the new index as:

$$I = \sum_{j=1}^{M} w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod}(x) - f_{exact}(x))|dx$$

i.e. the new index is defined as the weighted sum of the distance between classification function and perfect classification function.

On the basis of the previous example, we can compute the value for the index introduced in Definition II.3.3: the three integral results are (0.3, 0.1, 0.2) and the corresponding weights are (1, 0.67, 0.33), thus $I = 0.433$.
The index satisfies the following properties.

Propr II.3.4. $I \in [0, +\infty)$.
$I = 0$ *if and only if* $f_{mod} = f_{exact}$.

Proof.

$$I = \sum_{j=0}^{M-1} w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx \geq \sum_{j=0}^{M-1} \frac{n_j - \tilde{i}_j}{N} |f_{mod} - f_{exact}| \frac{n_j - n_{j-1}}{N}$$

and

- $n_j \geq \tilde{i}_j$,
- $n_j > n_{j-1}$

by definition, than we can conclude that $I \geq 0$.
We prove also that $I = 0$ if and only if $f_{mod} = f_{exact}$.

$$I = 0 \implies w_j = 0 \text{ or } \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx = 0 \quad \forall\, j \text{ in } \{1, ..., M-1\}.$$

- $w_j = 0 \iff \tilde{i}_j = n_j$, i.e there are not classification errors, so $f_{mod} = f_{exact}$ in class j.

- $\int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx = 0 \iff f_{mod} = f_{exact}$ in the class j.

So we can conclude that $I = 0 \implies f_{mod} = f_{exact}$.

The other implication is trivial. ■

Propr II.3.5. *I has a sharp upper bound $M - 1$*

The upper bound $M - 1$ is reached if and only if $M = 2$ (binary classification).

Proof.

$$I = \sum_{j=0}^{M-1} w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx \leq \sum_{j=0}^{M-1} 1 \cdot \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx \leq$$

$$\leq \max_x |(f_{mod} - f_{exact})(x)| \sum_{j=0}^{M-1} \frac{n_j - n_{j-1}}{N} \leq M - 1$$

If $M = 2$ we obtain $|(f_{mod} - f_{exact})(x)| = 1\ \forall x \in [0, 1]$ so that $I = M - 1$. If $M > 2$, $|(f_{mod} - f_{exact})(x)| > 1$ for at least one class (by construction) the inequality is strict. ■

Proposition II.3.6. *$I \leq K$,*

where K is defined as

$$K = \sum_{i=1}^{M} l_i \max\{M - i, i - 1\}$$

Proof. The maximum value is reached when the worst classification is obtained, i.e. when all observations are associated to the farthest class. If this happens, the error interval is as long as the class domain, so $w_j = 1 \forall j = 1, ..., M$ and each integral is the area of a rectangle with basis the class domain l_j and height the maximum height reachable.

■

Definition II.3.7 (Normalized index)**.**

$$I_n = \frac{1}{K} \sum_{j=0}^{M-1} w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx$$

where K is the maximum defined in the Proposition II.3.6.
So $0 \leq I_n \leq 1$.

In the previous example, $K = 1.7$ and the corresponding value of the defined normalized index is 0.255.

Proposition II.3.8. *The accuracy is a special case of the index introduced in Definition II.3.3.*

Proof. The accuracy is $acc = p_{err} = \frac{\#\{\text{misclassified observations}\}}{N}$ i.e. the proportion of misclassified observations.
Setting $M = 2$, from the Proposition II.3.6, $K = 1$.
$max_x|f_{mod}(x) - f_{exact}(x)| = 1$, each weight is $w_j = \frac{1}{N}$ if $w_1 = w_2 = 1$ and $I_n = p_{err}$. ■

Propr II.3.9 (Monotonicity)**.** *Consider a classification C with ϵ misclassifications and N observations. Operating a transformation of the classification C in C' where an observation right classified is changed in a misclassification, the index I_n becomes higher.*

Proof. In the classification C', $\epsilon'=\epsilon + 1$ are misclassified observations: the ϵ observations misclassified in C plus a new misclassification. Suppose that the new misclassification is the observation i that is classified in the class j' instead of the real class j.
All the components in the sum of the index I_n remain unchanged except for the j^{th}, thus obtaining I_n^j. So

$$I_n^j = w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |f_{mod}(x) - f_{exact}(x)|dx$$

Looking at each of the two elements in the product:

- $w'_j \geq w_j$
  Two different cases are possible: if the probability associated to the i^{th} observations is less or equal than the probability of the first error, the error interval $w'_j = w_j$; on the other hand, the error interval become larger, thus $w'_j > w_j$.

- $|f'_{mod} - f_{exact}| > |f_{mod} - f_{exact}|$
  In C' there is one misclassification more than in C, so the distance between f_{mod} and f_{exact} increases.

We can conclude that $I'^{j}_{n} \geq I^{j}_{n}$. ■

We remark that in the Property II.3.9 the vice versa does not hold, i.e. if $I_{mod1} \geq I_{mod2}$ we can not make conclusions on the number of misclassified observations in the two classifications.

II.4 Toy examples

In order to show how our index works with respect to the indexes proposed in the literature, toy examples are reported in this section with the main aim of discussing the behaviour in terms of model selection of our index with respect to AUC, accuracy and MSE.
Y is a target variable characterized by $M = 3$ levels $y_i \in \{1, 2, 3\}$ and Model 1 and Model 2 are two competitive models under comparison. The numerical setting of both examples is stated in Appendix II.A.

II.4.1 First toy example

In the first toy example we take into account the ordinal structure of the target variable Y. Table II.3 and Table II.4 are the corresponding confusion matrices for Model 1 and Model 2. It is clear that the Model 2 makes a better classification than Model 1.

		Actual		
		1	2	3
Predict	1	5	0	1
	2	0	7	0
	3	0	0	7

Table II.3: Confusion matrix Model 1

		Actual		
		1	2	3
Predict	1	5	1	0
	2	0	6	0
	3	0	0	8

Table II.4: Confusion matrix Model 2

Model	Proposed Index	Normalized Index	AUC	accuracy	MSE
1	0.08	0.05	0.95	0.95	0.20
2	0.04	0.03	0.95	0.95	0.05

Table II.5: Results

For the sake of comparison, for each model the AUC, the accuracy, the MSE and our index are computed as summarized in Table II.5.
We remark that looking at Table II.5 the values obtained for the AUC and the accuracy indexes for Model 1 and Model 2 are exactly equal, thus, in terms of model choice, Model 1 and Model 2 are not different. Our index highlights a difference in terms of performance between the two models under comparison and it selects Model 2 as the best one. Further details about the setting are given in Table II.11 in Appendix II.A.

II.4.2 Second toy example

The second toy example considers the probability assigned to each observation. In practical applications where we need also to evaluate how much uncertainty is associated to a prediction, the starting point considers the probability that the new observation belongs to the estimated class.

From Table II.6, both Model 1 and Model 2 assign an observation of the third class to the first one. The first classification assigns a higher probability to the misclassified observation than the second (p=0.866 vs p=0.400), see Table II.12 in Appendix II.A. Table II.12 reports set probabilities and consequent assigned classes. Then we can conclude that Model 2 is better than Model 1 for data at hands.

		Actual		
		1	2	3
Predict	1	5	0	1
	2	0	7	0
	3	0	0	7

Table II.6: Confusion matrix

From Table II.7 both models are equivalent in terms of MSE and accuracy, thus on the basis of classical measures Model 1 and Model 2 are not different. Our index reports different values for the models under comparison and select Model 2 as the best one.

Model	Proposed Index	Normalized Index	AUC	accuracy	MSE
1	0.083	0.051	0.956	0.950	0.200
2	0.017	0.010	0.983	0.950	0.200

Table II.7: Results

II.5 Empirical evaluation on simulated data

In order to show how our proposal works in model selection, this section reports the empirical results achieved on a simulated dataset.

The simulated dataset is composed of three covariates obtained by a Monte Carlo simulation and an ordinal target variable with $M = 5$, as reported in Table II.8. The sample size is $N = 7500$. The dataset is exactly balanced in terms of response variable: 1500 observations are generated for each level of y.

y	1	2	3	4	5
x1	N(2,1.5)	N(3,1)	N(4,1.5)	N(5,1)	N(6,1)
x2	N(1,2.5)	N(5,2)	N(7,2.5)	N(8.5,2)	N(9.5,2)
x3	U(0,3)				

Table II.8: Simulated data structure.

Five different models are under comparison:

- Ordinal logistic regression (Ord Log),
- Conditional inference tree (Tree),
- Support vector machine (SVM),
- Ordinal Random forest (RFor),
- k- Nearest Neighbour with k=20 (kNN-20),
- k- Nearest Neighbour with k=50 (kNN-5),
- Naive Bayes (NaiveB).

For each model AUC, accuracy, MSE and our index are computed using a 10-fold cross validation. More specifically, the dataset is randomly partitioned into 10 equal sized subsamples (of 750 observations), each of which is retained as validation data and the remaining 9 subsamples are used as training data. The process is then repeated 10 times, with each of the subsamples used exactly once for validation. The resulting metrics are averaged and than reported in Table II.9.

For the sake of clarity, Table II.10 shows the resulting ranks for the models, using the results obtained for the four metrics under comparison.

We can see that the k-nearest neighbour with $k = 5$ is classified as the best model according to all the indexes employed for model choice except for the AUC metric, but the values of AUC are extremely similar to the best model (the difference is less than 0.001). Furthermore, from Table II.9 k-nearest neighbour

Model	Proposed Index	Normalized index	AUC	Accuracy	MSE
Ord Log	0.450	0.141	0.864	0.581	0.580
Tree	1.569	0.491	0.875	0.586	0.643
SVM	0.446	0.137	0.869	0.592	0.581
RFor	0.469	0.143	0.875	0.589	0.643
kNN-20	0.003	0.0009	0.999	0.976	0.025
kNN-5	0.002	0.0006	0.999	0.993	0.008
NaiveB	0.434	0.132	0.877	0.604	0.594

Table II.9: Model comparison

Model	Proposed Index/Normalized	AUC	Accuracy	MSE
Ord Log	5	7	7	3
Tree	7	4	6	6
SVM	4	6	4	4
RFor	6	5	5	7
kNN-20	2	1	2	2
kNN-5	1	2	1	1
NaiveB	3	3	3	5

Table II.10: Results in terms of ranking.

outperforms the other models (with both choices of k). The Naive Bayes is ranked as the second-best model after kNN with respect to all performance indicators except for MSE (with minimum differences from Ord Log and SVM). When the performance differences between models are macroscopic all the indexes agree in model selection. The interest of a new metric come out when other indexes can not individuate differences between performances, then the natural structure of data and prediction probabilities become fundamental for the selection of the best model.

II.6 Conclusions

A new performance indicator is proposed to compare predictive classification models characterized by ordinal target variable.

Our index is based on the definition of a classification function and an error interval. A normalized version of the index is derived. The empirical evidence at hands underlined that our index discriminates better among different models

with respect to classical measures available in the literature.

Our index can be used coupled with other metrics for assessing model performances for model selection.

From a computational point of view a further idea of research will consider the implementation of our index in a new R package. In terms of application we think that our index could be directly incorporate in the process of assessment for predictive analytics.

Appendix II.A Toy example settings

In order to make the toy examples reproducible, numerical settings are reported. Table II.11 and in Table II.12 contain the hypothetical output of the two models described in Section II.4: a progressive ID of observations, probabilities assigned for each class (p_1, p_2, p_3) by Model 1 and Model 2, the resulting Estimated class for each model and the Real class assigned arbitrary by the author.

Observation	Model 1			Model 2			Estimated class Model 1	Estimated class Model 2	Real class
	p_1	p_2	p_3	p_1	p_2	p_3			
1	0.114	0.473	0.413	0.114	0.473	0.413	2	2	2
2	0.068	0.184	0.747	0.068	0.184	0.747	3	3	3
3	**0.750**	**0.125**	**0.125**	**0.125**	**0.750**	**0.125**	**1**	**2**	**3**
4	0.587	0.212	0.201	0.587	0.212	0.201	1	1	1
5	0.0583	0.623	0.319	0.0583	0.623	0.319	2	2	2
6	0.371	0.063	0.565	0.371	0.063	0.565	3	3	3
7	0.329	0.179	0.491	0.329	0.179	0.491	3	3	3
8	0.114	0.444	0.442	0.114	0.444	0.442	2	2	2
9	0.936	0.014	0.050	0.936	0.014	0.050	1	1	1
10	0.116	0.229	0.655	0.116	0.229	0.655	3	3	3
11	0.376	0.398	0.226	0.376	0.398	0.226	2	2	2
12	0.435	0.438	0.128	0.435	0.438	0.128	2	2	2
13	0.452	0.226	0.321	0.452	0.226	0.321	1	1	1
14	0.740	0.173	0.087	0.740	0.173	0.087	1	1	1
15	0.180	0.796	0.0243	0.180	0.796	0.0243	2	2	2
16	0.343	0.392	0.265	0.343	0.392	0.265	2	2	2
17	0.049	0.073	0.878	0.049	0.073	0.878	3	3	3
18	0.522	0.076	0.403	0.522	0.076	0.403	1	1	1
19	0.012	0.194	0.794	0.012	0.194	0.794	3	3	3
20	0.128	0.380	0.491	0.128	0.380	0.491	3	3	3

Table II.11: First toy example

Observation	Model 1			Model 2			Estimated class	Real class
	p_1	p_2	p_3	p_1	p_2	p_3		
1	0.114	0.473	0.413	0.114	0.473	0.413	2	2
2	0.068	0.184	0.747	0.068	0.184	0.747	3	3
3	**0.866**	**0.012**	**0.121**	**0.400**	**0.300**	**0.300**	**1**	**3**
4	0.587	0.212	0.201	0.587	0.212	0.201	1	1
5	0.0583	0.623	0.319	0.0583	0.623	0.319	2	2
6	0.371	0.063	0.565	0.371	0.063	0.565	3	3
7	0.329	0.179	0.491	0.329	0.179	0.491	3	3
8	0.114	0.444	0.442	0.114	0.444	0.442	2	2
9	0.936	0.014	0.050	0.936	0.014	0.050	1	1
10	0.116	0.229	0.655	0.116	0.229	0.655	3	3
11	0.376	0.398	0.226	0.376	0.398	0.226	2	2
12	0.435	0.438	0.128	0.435	0.438	0.128	2	2
13	0.452	0.226	0.321	0.452	0.226	0.321	1	1
14	0.740	0.173	0.087	0.740	0.173	0.087	1	1
15	0.180	0.796	0.0243	0.180	0.796	0.0243	2	2
16	0.343	0.392	0.265	0.343	0.392	0.265	2	2
17	0.049	0.073	0.878	0.049	0.073	0.878	3	3
18	0.522	0.076	0.403	0.522	0.076	0.403	1	1
19	0.012	0.194	0.794	0.012	0.194	0.794	3	3
20	0.128	0.380	0.491	0.128	0.380	0.491	3	3

Table II.12: Second toy example

References

Adams, N.M. and Hand, D.J. (2000). "Improving the Practice of Classifier Performance Assessment". In: *Neural Computation* vol. 12, pp. 305–311.

Agresti, A. (2010). *Analysis of ordinal categorical data*. Vol. 656. John Wiley & Sons.

Ahmad, A. and Brown, G. (2015). "Random Ordinality Ensembles: Ensambles methods for multi-valued categorical data". In: *Information Sciences* vol. 296, pp. 75–94.

Bradley, A.P. (1997). "The use of the area under the ROC curve in evaluation of machine learning algorithms". In: *Pattern Recognition* vol. 30, pp. 1145–1159.

Frank, E. and Hall, M. (2001). "A Simple Approach to Ordinal Classification". In: vol. 2167, pp. 145–156.

Gaudette, L. and Japkowicz, N. (2009). “Evaluation Methods for Ordinal Classification”. In: Gao, Y. and Japkowicz, N. *Advances in Artificial Intelligence*, pp. 207–210.

Gigliarano, C., Figini, S., and Muliere, P. (2014). “Making classifier performance comparisons when ROC curves intersect”. In: *Computational Statistics and Data Analysis* vol. 77, pp. 300–312.

Hand, D.J. (1997). *Construction and Assessment of Classification Rules*. Ed. by Wiley. Wiley Series in Probability and Statistics.

— (2001). “Measuring diagnostic accuracy of statistical prediction rules”. In: *Statistica Neerlandica* vol. 55, pp. 3–16.

— (2009). “Measuring classifier performance: a coherent alternative to the area under the ROC curve”. In: *Machine Learning* vol. 77, no. 1, pp. 103–123.

Hand, D.J. and Till, R.J. (2001). “A simple generalisation of the area under the ROC curve for multiple class classification problems”. In: *Machine Learning* vol. 45, no. 2, pp. 171–186.

Hornung, R. (2020). “Ordinal forests”. In: *Journal of Classification* vol. 37, pp. 4–17.

Hossin, M. and Sulaiman, M.N. (2015). “A review on evaluation metrics for data classification evaluations”. In: *International Journal of Data Mining & Knowledge Management Process* vol. 5, no. 2, pp. 171–186.

Huang, J. and Ling, C.X. (2007). “Constructing New and Better Evaluation Measures for Machine Learning”. In: *IJCAI International Joint Conference on Artificial Intelligence (IJCAI’07)*, pp. 859–864.

Kotłowski, W. et al. (2008). “Stochastic dominance-based rough set model for ordinal classification”. In: *Information Sciences* vol. 178, no. 21, pp. 4019–4037.

Morrone, A., Piscitelli, A., and D’Ambrosio, A. (2019). “How disadvantages shape life satisfaction: an alternative methodological approach”. In: *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement* vol. 141, no. 1, pp. 477–502.

Pang, B. and Lee, L. (2005). “Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales”. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 115–124.

Piccarreta, R. (2004). “Ordinal Classification Trees Based on Impurity Measures”. In.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). “Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation”. In: *Advances in Artificial Intelligence. Lecture Notes in Computer Science (AI 2006)*. Vol. 4304, pp. 1015–1021.

Sokolova, M. and Lapalme, G. (2009). “A systematic analysis of performance measures for classification tasks”. In: *Information Processing & Management* vol. 45, no. 4, pp. 427–437.

Torra, V. et al. (2006). “Regression for ordinal variables without underlying continuous variables”. In: *Information Sciences* vol. 176, no. 4, pp. 465–474.

Paper III

At risk mental status analysis: a comparison of model selection methods for ordinal target variable

Elena Ballante[1], Silvia Molteni[2], Martina Mensi[3], Silvia Figini[4]

Conference short paper published in *Book of Short Papers SIS 2020*

Abstract

The presence of an ordinal variable poses several problems that are not deepen in the literature. In this work we aim to analyse a dataset on mental state at risk, a delicate and extremely debatable definition in the psychiatric field, in order to produce a model that is of double value. On the one hand it allows to understand which are the most influential characteristics on the mental state, on the other hand it allows to predict the correct diagnosis and possibly also to assess the risk of a possible transition to the psychotic state. With this aim we focus on model selection methods in the framework of ordinal target variable.

Keywords: Ordinal Classification; Psychosis; Model Selection

[1] Department of Mathematics, University of Pavia, Pavia, Italy
[b] Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy
[c] IRCCS Mondino Foundation, Pavia, Italy
[d] Department of Political and Social Sciences, University of Pavia, Pavia, Italy

III.1 Introduction

Despite advances in pharmacological and psychotherapeutic interventions over the last decades, psychotic disorders continue to be among the most severe disorders in medicine. In children and adolescents, schizophrenia is one of the ten main causes of disability-adjusted life years (DALYs) in 10 to 14-year-old boys and 15 to 19-year-old girls Gore et al. 2011). The identification of people at high-risk of developing psychosis is one of the most promising strategies to improve outcomes. Indeed, retrospective studies indicate that the onset of full psychosis is commonly preceded by a prodromal phase lasting up to several years Hafner et al. 1999; Schultze-Lutter et al. 2010. Recently, the importance of research in persons at high risk has been increasingly recognized to such an extent that Attenuated Psychosis Syndrome has been introduced in section III ("Emerging Measures and Models") of the Diagnostic and Statistical Manual of Mental Disorder, fifth Edition. Some concerns raised regarding the introduction of this new syndrome need to be addressed with special attention in children and adolescents, where research on the high risk state is still in its infancy Schimmelmann, Walger, and Schultze-Lutter 2013. In particular, criticism about pathologization of non-ill behaviours and experiences has been voiced. In fact, during adolescence, the assessment of psychiatric symptoms and disorders is challenging. Several authors have underlined the difficulty in discriminating between normal behaviours and psychiatric symptoms Welsh and Tiffin 2013. Normative adolescent experiences can make the clinical picture blurred and lead to false positive psychotic diagnoses Carol and Mittal 2015. Overall, in children and adolescents research on the high risk state and attenuated psychotic symptoms is still in its infancy and the clinical validity of at risk criteria appears understudied. Furthermore, only few studies have evaluated the psychopathological and neuropsychological characteristics of adolescents with attenuated psychotic symptoms (APS).

In this context an accurate data analysis is fundamental for knowledge extraction and prediction: it is extremely important to deploy adequate models

and to perform a suitable selection procedure.
If there are several models that are suitable for the analyses, it can be noted that the evaluation measures for model selection are few and inadequate to the problem under analysis.
Performance indicators can be used to assess the performance of a model in terms of accuracy, discriminatory power and stability of the results. The choice of indicators to made model selection is a fundamental point and many approaches have been proposed over the years (see e.g. Bradley 1997; Hand 2009).
Multi-class classification models are generally evaluated averaging binary classification indicators (see Hossin and Sulaiman 2015) and in the literature there is not a clear distinction among them with respect to multi-class nominal and ordinal targets (e.g. Gaudette and Japkowicz 2009).
While in the model definition stage for ordinal target variable there are different approaches in the literature (see Agresti 2010; Kotłowski et al. 2008), for the model selection there is a lack of adequate tools (Cardoso and Sousa 2011).
In our opinion, performance indicators should consider the nature of the target variable, especially when the dependent variable is ordinal. In medical application is also fundamental to take into account the uncertainty that the model assign to the prediction in order to obtain the maximum of interpretability. This leads us to apply and compare different measures to select the best model to predict the mental status, contexts characterized by a multi-class ordinal target variable.

The rest of the paper is organized as follow: Section III.2 describes data at hand and the analysis performed, Section III.3 describes preliminary results obtained.

III.2 Data description and analysis

The dataset is composed by 240 observation (corresponding to 240 patients under examination). The target variable is the membership to one of three cat-

egories: subject not at risk, at risk, psychotic. This variable is clearly ordinal and in this context is extremely important to consider the order of levels.
The covariates considered are some personal information like age at onset, sex, ethnicity, some medical history of the patients and familiarity to mental illnesses, some information about symptoms, duration of the psychotherapy and of the drugs assumption, the IQ index (Intelligence Quotient) and the SOFAS (Social and Occupational Functioning Assessment Scale).
On this dataset, five different models are implemented and compared: Ordinal logistic regression McCullagh 1980, Classification tree Breiman, Friedman, and Olsen 1984, Support vector machine Drucker et al. 1997, Random forest Breiman 2001, k-Nearest Neighbour Cover and Hart 1967.
In the model selection step we compare the information given by an index for ordinal target proposed in Ballante, Uberti, and Figini 2020 with standard indexes used in literature that are AUC (Area Under the ROC curve), accuracy (expressed in terms of correct classification) and MSE (Mean Square Error) (see Gaudette and Japkowicz 2009 and Huang and Ling 2007 among others), another index for ordinal target variable proposed in Cardoso and Sousa 2011 and the total misclassification cost used in Piccarreta 2008. The index proposed in Ballante, Uberti, and Figini 2020 is defined basing on a classification function, i.e. a function which represents the actual classification made by the model under evaluation, compared with an exact classification function that is the goal of each model. This index takes into account the ordinal structure of the target variable and the probability assigned from the model at each observation. This first aspect has obvious advantages in this context, whereas the second aspect is extremely useful in medical application, when we need to consider the uncertainty of the prediction as well as the prediction itself. In the rest of the paper we refer to this index as OPI (Ordinal Probability Index). The AUC for multi-class classification is defined in Hand and Till 2001 as a generalization of the AUC (based on the probabilistic definition of AUC); it suffers of different weaknesses also in the binary classification problem (Gigliarano, Figini, and Muliere 2014) and it is cost-independent, assumption that can be viewed as a

weakness when the target is ordinal.

Accuracy (percentage of correct predictions over total instances) is the most used evaluation metric for binary and multi-class classification problems (Sokolova, Japkowicz, and Szpakowicz 2006), assuming that the costs of the different misclassifications are equal.

Mean square error (MSE) measures the difference between prediction values and observed values in regression problems using an Euclidean distance. MSE can be used in ordinal predictive models, converting the classes of the ordinal target variable y in integers and computing the difference between them and it does not takes into account the ordering in a predictive model characterized by ordinal classes in the response variable. Furthermore, it is well known that in imbalanced data characterized by under-fitting or over-fitting the mean square error could provide trivial results (see Hossin and Sulaiman 2015).

Total misclassification cost is simply defined as the sum of absolute values of the differences between the real class and the predicted class, transformed integer values as in MSE. It was used in Piccarreta 2008 to prune a new classification tree algorithm in ordinal framework.

Ordinal Index proposed in Cardoso and Sousa 2011 is based on confusion matrix and on the concept of non-discordant pair of points, i.e. when the relative order of the predicted classes of two observations is the same of the relative order of the real classes. The advantage of this method with respect to AUC, accuracy and MSE is that consider the ordinal structure of the target variable, but it does not take into account the probabilities assigned to the prediction like the first index proposed.

III.3 Preliminary results

Data at hands are composed by 240 observations and fifteen covariates both qualitative and quantitative as possible explanatory variables and predictors. The target variable has three ordered level (not at risk, at risk, psychotic). On

this dataset six different predictive models are implemented:

- Ordinal logistic regression (Ord Log),
- Classification tree (Tree),
- Support vector machine (SVM),
- Random forest (RFor),
- k- Nearest Neighbour (kNN),
- Naive Bayes (NBayes).

In order to select the best model a 5-fold cross validation is implemented. The models are compared in terms of out of sample performance on the basis of OPI, AUC, accuracy, MSE, the ordinal index (Ord Ind) and misclassification cost (Misc Cost). Table III.1 reports the mean values of the metrics under comparison derived from the cross validation exercise. For sake of clarity, Table III.2 shows the resulting ranks for the models, using the results obtained for the four metrics under comparison.

Model	OPI	AUC	Accuracy	MSE	Ord Ind	Misc Cost
Ord Log	0.176	0.847	0.713	0.316	0.372	12.6
Tree	0.215	0.850	0.685	0.399	0.410	14.6
SVM	0.150	0.987	0.760	0.310	0.335	11.2
RFor	0.138	0.902	0.774	0.282	0.315	10.4
kNN	0.351	0.735	0.624	0.404	0.451	16.4
NBayes	0.165	0.912	0.756	0.301	0.341	11.2

Table III.1: Model selection.

Model	OPI	AUC	Accuracy	MSE	Ord Ind	Misc Cost
Ord Log	4	5	4	4	4	4
Tree	5	4	5	5	5	5
SVM	2	2	2	3	2	2.5
RFor	1	3	1	1	1	1
kNN	6	6	6	6	6	6
NBayes	3	1	3	2	3	2.5

Table III.2: Results in terms of ranking.

The performances achieved on the model under comparison underline that there is a group of models with worst performances (Ord Log, Tree, kNN) and a group of model with better performances (SVM, RFor, NBayes). On the basis of OPI, Accuracy, MSE, Ord Ind and Misc Cost the Random Forest model is the best one. Looking at AUC the best model is Naive Bayes but the performance exactly the same of Random Forest in terms of De Long test ($p > 0.2$).

On the basis of the results obtained in Table III.1 a further analysis considers the results obtained using Random Forest. A necessary aspect to consider is the analysis of features involved in the classification process. The next steps will be divide the data in 60% and 40% to deploy and train the single model selected on the dataset, on this model the variable importance can be analyse in order to understand which aspects have more weight in the classification. Random Forest model produces as output an index of variable importance; on the basis of the data at hand the variable with higher degree of importance are also interesting in the clinical domain, as for example SOFAS score, CGIS score, depressive disorders. Also SES score is one of the variables with greater importance, underlining the influence of the socio-economic level on mental status.

A further analysis will consider longitudinal behaviour of the patient thus providing the opportunity to model also time dependent covariates. Moreover in a different dataset we have information about the transition to psychosis of some patient included in this study and further analysis will focus on this groups of patients that make the transition to psychosis in order to evaluate the performances of the model selected only on this category.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*. Vol. 656. John Wiley & Sons.

Ballante, E., Uberti, P., and Figini, S. (2020). *A new approach in model selection for ordinal target variables*. arXiv: 2003.02761.

Bradley, A.P. (1997). "The use of the area under the ROC curve in evaluation of machine learning algorithms". In: *Pattern Recognition* vol. 30, pp. 1145–1159.

Breiman, L. (2001). "Random Forests". In: *Machine Learning* vol. 45, no. 1, pp. 5–32.

Breiman, L., Friedman, J., and Olsen, R. (1984). *Classification and Regression Trees*.

Cardoso, J. and Sousa, R. (2011). "Measuring the Performance of Ordinal Classification". In: *International Journal of Pattern Recognition and Artificial Intelligence* vol. 25, no. 08, pp. 1173–1195.

Carol, E. and Mittal, V. (2015). "Normative adolescent experiences may confound assessment of positive symptoms in youth at ultra-high risk for psychosis". In: *Schizophrenia Research* vol. 166, pp. 358–359.

Cover, T.M. and Hart, P.E. (1967). "Nearest Neighbor Pattern Classification". In: *IEEE Transactions on Information Theory* vol. 13, pp. 21–27.

Drucker, H. et al. (1997). "Support vector regression machines". In: *Advances in neural information processing systems* vol. 28, pp. 779–784.

Gaudette, L. and Japkowicz, N. (2009). "Evaluation Methods for Ordinal Classification". In: Gao, Y. and Japkowicz, N. *Advances in Artificial Intelligence*, pp. 207–210.

Gigliarano, C., Figini, S., and Muliere, P. (2014). "Making classifier performance comparisons when ROC curves intersect". In: *Computational Statistics and Data Analysis* vol. 77, pp. 300–312.

Ginzberg, P. and Walden, A.T. (2012). "Matrix-valued and quaternion wavelets". In: *IEEE transactions on signal processing* vol. 61, no. 6, pp. 1357–1367.

Gore, F.M. et al. (2011). "Global burden of disease in young people aged 10-24 years: a systematic analysis". In: *The Lancet* vol. 377, pp. 486–486.

Hafner, H. et al. (1999). “Depression, negative symptoms, social stagnation and social decline in the early course of schizophrenia”. In: *Acta Psychiatr. Scand.* vol. 100, pp. 105–118.

Hand, D.J. (2009). “Measuring classifier performance: a coherent alternative to the area under the ROC curve”. In: *Machine Learning* vol. 77, no. 1, pp. 103–123.

Hand, D.J. and Till, R.J. (2001). “A simple generalisation of the area under the ROC curve for multiple class classification problems”. In: *Machine Learning* vol. 45, no. 2, pp. 171–186.

Hossin, M. and Sulaiman, M.N. (2015). “A review on evaluation metrics for data classification evaluations”. In: *International Journal of Data Mining & Knowledge Management Process* vol. 5, no. 2, pp. 171–186.

Hsieh, C. C. (2002). “Motion Smoothing Using Wavelets”. In: *Journal of Intelligent and Robotic Systems* vol. 35, pp. 157–169.

Huang, J. and Ling, C. (Jan. 2007). “Constructing New and Better Evaluation Measures for Machine Learning.” In: pp. 859–864.

Janiak, M., Szczęsna, A., and Słupik, J. (2014). “Implementation of Quaternion Based Lifting Scheme for Motion Data Editor Software”. In: *Intelligent Information and Database Systems (ACIIDS)*, pp. 515–524.

Kotłowski, W. et al. (2008). “Stochastic dominance-based rough set model for ordinal classification”. In: *Information Sciences* vol. 178, no. 21, pp. 4019–4037.

McCullagh, P. (1980). “Regression Models for Ordinal Data”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* vol. 42, no. 42, pp. 109–142.

Piccarreta, R. (Feb. 2008). “Classification trees for ordinal variables”. In: *Computational Statistics* vol. 23, pp. 407–427.

Schimmelmann, B.G., Walger, P., and Schultze-Lutter, F. (2013). “The Significance of At-Risk Symptoms for Psychosis in Children and Adolescents”. In: *The Canadian Journal of Psychiatry* vol. 58, no. 1, pp. 32–40.

Schultze-Lutter, F. et al. (2010). “Basic symptoms and ultrahigh risk criteria: symptom development in the initial prodromal state”. In: *Schizophrenia Bullettin* vol. 36, pp. 182–191.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). “Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation”. In: *Advances in Artificial Intelligence. Lecture Notes in Computer Science (AI 2006)*. Vol. 4304, pp. 1015–1021.

Welsh, P. and Tiffin, P.A. (2013). “Attitudes of patients and clinicians in relation to the at-risk state for psychosis”. In: *Early Interv Psychiatry* vol. 7, pp. 361–367.

Remarks

The Paper III is a short proceeding and describes only preliminary results in order to show the potential of the application of the ordinal index proposed in Paper II in a real world problem. Further results obtained analysing the same cohort of patients of the short paper can be found in Mensi et al. 2021[1], where the data set was extended with longitudinal information and additional analysis.

The aims of this study were to characterize the profile of DSM-5 APS (Attenuated Psychosis Syndrome as introduced in Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) adolescents, at presentation, compared with adolescents suffering from early-onset psychosis (EOP) and with other psychiatric disorders and to estimate their long-term risk of transition to psychosis and prognostic accuracy of DSM-5 APS.

[1]Mensi, M.M., Molteni, S., Iorio, M., Filosi, E., Ballante, E., Balottin, U., Fusar-Poli, P., Borgatti, R. (2021). "Prognostic Accuracy of DSM-5 Attenuated Psychosis Syndrome in Adolescents: Prospective Real-World 5-Year Cohort Study", *Schizophrenia Bulletin*, vol. 47, no. 6, pp. 1663–1673

Part 3

Quaternion time series analysis

Introduction

Quaternion algebra and quaternion time series analysis are fields of interest for researchers as the need to study motion data is emerging in several kinds of applications: computer animation, virtual reality, robotics, and biomedical sensors are only some examples.

The quaternion algebra was introduced by W.R. Hamilton in 1843 as a generalization of complex numbers to describe rotations in a three dimensional space. The popularity of quaternions is due to the possibility to use them to describe orientations in a convenient way that tackle different limits of the classical representation of orientations as Euler angles through rotation matrices.

In general, in time series analysis, the smoothing process is often a crucial step and even more in this context when sensors are used for the acquisition and measurements. In fact, sensors tend to collect a certain amount of noise that is difficult to isolate and manage.

Different methods to smoothing time series in quaternion algebra were developed in literature and some of these were applied to real or simulate data sets, see for example Ginzberg and Walden 2012, Janiak, Szczęsna, and Słupik 2014, Hsieh 2002. In general, these methods are based on the generalizations of classical smoothing methods as Fourier transform and wavelet analysis. Despite this, the lack of availability of the code makes these methods leaving open the problem of applications to the real world cases.

In the following work, we reviewed the existing methods in literature for smoothing unit quaternion time series. We considered the method proposed in Hsieh 2002, that consists in transforming the quaternion time series in the corresponding angular velocity time series. The angular velocity time series was smoothed and then transformed back to quaternion space. This method seems promising, easy to apply and particularly useful because it allows to implement all the theories developed in Euclidean spaces. With reference to this idea, we proposed a new method that deploys the logarithm transformation instead of angular velocity calculation to transform the quaternion time series in a real

three dimensional time series. The advantage of the proposed method is that logarithm is in general a smoother transformation than angular velocity, so it can introduce a lower degree of transformation errors.

These two methods are compared in terms of classification performances on a real data set and five derived data sets where different degrees of noise are introduced. The results confirm the hypothesis made on the basis of the theoretical information available from the two methods, i.e. the logarithm is smoother and generally provides better results than the existing method. These results are strengthened by a regression model that confirms this conclusion from a statistical point of view.

The real data set considered for the analysis contains the measurements of motion of the hip joint of 27 healthy subjects registered under two different conditions: natural walking and a walking made difficult by an impediment to simulate a walking impairment. This work is part of a bigger project that aims to detect first signs of walking impairments in patients with ALS, MS and other neurodegenerative diseases to personalize the therapeutic approach.

The main contribution of this new method is to manage unit quaternion in a proper way, transforming the time series in an Euclidean space in order to take advantage of all the literature of smoothing techniques, with better results than the angular velocity transformation method already defined in Hsieh 2002.

Paper IV

Smoothing Method for Unit Quaternion Time Series: An application to motion data

Elena Ballante[1,2], Lise Bellanger[3], Pierre Drouin[3,4], Silvia Figini [5], Aymeric Stamm[3]

Submitted

Abstract

Smoothing orientation data is a fundamental task in different fields of research. Different methods of smoothing time series in quaternion algebras have been described in the literature, but their application to real world problems is still an open point. This paper develops an effective method, which is easy to apply, for smoothing quaternion time series in order to obtain good performance in classification tasks.

Following the idea described in C. C. Hsieh 2002, which involves an angular velocity transformation of unit quaternion time series, we propose a new method based on the idea of employing the logarithm function to transform the quaternion time series to a real three-dimensional time series that can be smoothed with classical methods.

The results on classification tasks involving both a real data set and 10

[1] Department of Mathematics, University of Pavia, Pavia, Italy
[2] BioData Science Unit, IRCCS Mondino Foundation, Pavia, Italy
[3] Department of Mathematics Jean Leray, UMR CNRS 6629, Nantes University, France
[4] UmaniIT, Department of Research and Development, France
[5] Department of Political and Social Science, University of Pavia, Pavia, Italy

artificially noisy data sets confirm the effectiveness of the proposed method compared with the angular velocity one. These results are strengthened by a regression model that confirms this conclusion from a statistical point of view.

Keywords: Quaternion time series, Smoothing method, Classification task

IV.1 Introduction

The representation and analysis of the motion of human body is a research subject which has been constantly expanding with the increasing use of sensors. In time series analysis, smoothing is a fundamental step in real world applications, especially when sensors are involved, because a certain amount of noise is always captured. The presence of noise can lead to inconclusive results or even wrong conclusions when data are analyzed and classification or clustering algorithms are applied. Instead, preprocessing the data can extract meaningful features and patterns.

In this paper, we analyse motion data registered by a motion sensor called MetaMotionR (MMR), from Mbientlab, that measures the spatial orientation of the hips and stores them as a quaternion time series. The motion of the hip joint is registered under two different conditions: natural walking and a walking made difficult by an impediment to simulate a walking impairment due to ALS, MS and other neurodegenerative diseases.

In this context, different smoothing techniques for quaternion time series are reviewed. The smoothing technique proposed in C. C. Hsieh 2002 has been selected for its simplicity of implementation and its power in making available all the techniques developed in Euclidean spaces. On the basis of this method, a new technique is proposed and compared with the previous one. They were applied to real and artificially noisy data to understand the influence of the level of the noise on the performance of a smoothing methods.

The rest of this paper is organized as follows: in Section IV.2 different approaches to quaternion smoothing present in the literature and suitable for the specific problem of smoothing a 1D quaternion time series are described and the new method is described. In Section IV.3 a theoretical comparison of some of the methods is presented. Section IV.4 shows the experimental settings and results of the quaternion wavelet smoothing of real and noisy data in terms of classification performance. Conclusions and further ideas for research are summarized in Section IV.5.

IV.2 Quaternion time series smoothing methods

We are interested in smoothing methods suitable for one-dimensional quaternion-valued signals. Consider a signal $f \in L^2(\mathbb{R}, \mathbb{H})$. Most of the existing smoothing methods for this type of signal are generalizations of classical smoothing techniques originally meant for Euclidean spaces: the Fourier transform, spline functions, and wavelets. These methods have been adapted to quaternion time series in different ways.

Spline functions are often used in quaternion algebras to interpolate signals (see Ramamoorthi and Barr 1997 and Nielson 2004, for example), while there are no examples of applications where they are directly applied to smoothing signals.

The Fourier transform was extended in Hitzer 2007 and Li, Leng, and Fei 2018 as a proper transform in quaternion space. Instead, the application of the Fourier transform to quaternionic signals in real world examples involves a transformation from $\mathbb{H}$ to a real vector space where the Fourier transform is applied. The real spaces involved were an angular velocity space in Fang et al. 1998 and C. Hsieh et al. 1998 and a frequency space in Kenwright 2015.

Regarding the Quaternion Wavelet Transform (QWT), extensive reviews of the techniques related to it can be found in Xu et al. 2010 and Fletcher and Sangwine 2017. With the exception of the naïve approach of smoothing each component of the quaternions independently (see Traversoni 1995, for example), the other methods exploit different isomorphisms between $\mathbb{H}$ and other spaces with well known properties.

One of the earliest attempts to apply wavelet methods to quaternion time series in a proper way is Mitrea 1994, where Clifford wavelets and a Clifford multiresolution analysis were introduced. The application to quaternion time series is possible because $\mathbb{H}$ is isomorphic to the Clifford Algebra *Cl(0,2)*. The author limited his considerations to the theoretical statement of the method and the definition of the Haar wavelet.

In Traversoni 2001, the idea of Mitrea was explored in the context of image analysis, and the Haar wavelet was applied to biomedical data (tomography

images) written in terms of quaternions and compressed via wavelets, but the idea was not further explored.

The idea of matrix-valued wavelets (MVWs) was explored in Ginzberg 2013 and Ginzberg and Walden 2012, where they demonstrate that the wavelets defined in the previous literature, such as He and Yu 2005 and Peng and Zhao 2004, were trivial, and so they proposed new matrix-valued wavelets, using the isomorphism between $\mathbb{H}$ and the space of matrices $\mathbb{R}^{4x4}$ with quaternion-structure conditions on the coefficients. The isomorphism is defined as in Equation (IV.1).

$$q = w + x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \mapsto \begin{bmatrix} w & -x & -y & -z \\ x & w & -z & y \\ y & z & w & -x \\ z & -y & x & w \end{bmatrix} \tag{IV.1}$$

In this framework, they designed quaternion-structured MVWs and hence quaternion wavelets. The MATLAB code to compute the wavelet filter coefficients was presented, but no code was provided to perform a wavelet analysis on quaternion signals.

Ginzberg and Walden 2012 applied MVWs to a simulated quaternion time series. Fletcher 2017 extended this work, adding new wavelet filters of different lengths and explained how to arrange the filters in a matrix for analysing images and applying it to the analysis of a colour vector image.

Szczęsna, Słupik, and Janiak 2012 presented a different approach to analysing a quaternion signal, with multi-resolution techniques, based on second generation wavelet transform. The quaternion lifting scheme is defined as follows. The input data set is split into two disjoint sets of even and odd indexed samples. Samples with odd indices are predicted based on the sample with even indices (using the SLERP or SQUAD methods for quaternion time series, as reviewed in IV.B). Next, the input value with the odd index is replaced by the offset (difference) between its value and its prediction. The outputs are updated, so that coarse-scale coefficients have the same average value as the input samples. This step is necessary for the stability of the wavelet transform.

In this procedure, the wavelet function used can be reconstructed, but it is not necessary for the computation.

Another approach to quaternion signal smoothing through wavelets is described in C. C. Hsieh 2002, recalling the methods explored in Fang et al. 1998 and C. Hsieh et al. 1998 for the application of the Fourier transform to quaternion signals. The analysis is now focused on unit quaternion time series (see section IV.B for definition an properties) in $\mathbb{H}_1 \subset \mathbb{H}$.

The underlying idea is that if a unit quaternion time series is smooth, the changes of the angular velocities should be small. With this rationale, the smoothing process can be applied in the angular velocity space. As the angular velocities are in three-dimensional Euclidean space, all the real wavelet techniques, and even more, in general, all smoothing techniques for Euclidean spaces can be deployed. After the smoothing process, the unit quaternion time series are reconstructed.

In order to obtain angular velocities without employing derivatives, the following approximation formulas are used. Given a unit quaternion time series $q_1, ..., q_N$ and the time step h at which they were measured, the angular velocity is approximated as follows:

$$\mathbf{v}_i = \frac{log(q_i^{-1} q_{i+1})}{h},\ i = 1, ..., N-1. \tag{IV.2}$$

Further details about the derivation of this expression from the definition of angular velocity are provided in IV.B.

With these approximate angular velocities $\tilde{\mathbf{v}}_1, ..., \tilde{\mathbf{v}}_N$, the quaternion time series is reconstructed as

$$\tilde{q}_i = q_1 \prod_{j=2}^{i} \exp(\tilde{\mathbf{v}}_j h),\ i = 1, ..., N-1. \tag{IV.3}$$

Following this idea, our proposed approach considers the logarithm transformation in order to go from $\mathbb{H}_1$ to its tangent space $\mathbb{R}^3$. The logarithm of a unit quaternion time series is a time series in 3-dimensional Euclidean space defined as follows (see IV.B for further details):

$$log(q) = (\frac{x}{|\mathbf{v}|} arccos(\frac{w}{|q|}), \frac{y}{|\mathbf{v}|} arccos(\frac{w}{|q|}), \frac{z}{|\mathbf{v}|} arccos(\frac{w}{|q|})) \in \mathbb{R}^3 \tag{IV.4}$$

The idea is to employ suitable smoothing process to each component of the logarithm of the quaternions and then compute the unit quaternions by taking the quaternionic exponential, as follows:

$$\exp(q) = \exp(w)(\cos(|\mathbf{v}|), \frac{x}{|\mathbf{v}|}\sin(|\mathbf{v}|), \frac{y}{|\mathbf{v}|}\sin(|\mathbf{v}|), \frac{z}{|\mathbf{v}|}\sin(|\mathbf{v}|)) \qquad \text{(IV.5)}$$

where the w component of the logarithm is always 0.

This transformation is smoother with respect to the angular velocity and it has some intrinsic differences that will be explored in Section IV.3. These differences in the definition of the transformed space, where the smoothing is performed, affect the performance in classification tasks in ways that will be described in Section IV.4.

IV.3 Comparison of the methods

Firstly, we are interested in the comparison of the images of the two transformations involved.

The quaternionic logarithm function, for a unit quaternion, is $f : \mathbb{H}_1 \mapsto \mathbb{R}^3$ such that $q = (w, x, y, z) \mapsto \log(q) = \frac{(x,y,z)}{|(x,y,z)|} arccos(\frac{w}{|q|}) \in \mathbb{R}^3$.

For $\mathbf{v} = (x, y, z)$, write $\frac{\mathbf{v}}{|\mathbf{v}|}$ for the vector of unit norm in $\mathbb{R}^3$ in the same direction as $\mathbf{v}$ and $arccos(w/|q|) \in [0, \pi]$.

As a consequence, $Image(f) = \{\mathbf{v} \in \mathbb{R}^3 : |\mathbf{v}| \leq \pi\}$ is the ball of radius π in $\mathbb{R}^3$.

Given a unit quaternion time series $q_1, ..., q_N$, the angular velocities are approximated as in Equation (IV.2): $\mathbf{v}_i = \frac{log(q_i^{-1} q_{i+1})}{h}$, where $q_i^{-1} q_{i+1} \in \mathbb{H}_1$. Therefore, the same considerations can be applied to the numerator and $Image(f) = \{\mathbf{v} \in \mathbb{R}^3 : |\mathbf{v}| \leq \frac{\pi}{h}\}$ is the ball of radius $\frac{\pi}{h}$ in $\mathbb{R}^3$. In the theoretical framework, the angular velocity is calculated as a derivative and in the limit for h that goes to 0, the image of the transformation is all of $\mathbb{R}^3$.

Since in our application the angular velocity is approximated, h is a small positive constant and the image of the transformation is a ball in $\mathbb{R}^3$ with a radius larger than that with the logarithm transformation.

To further develop this comparison, we consider the geometric interpretation of

the transformations involved: angular velocity and logarithm.

The logarithm function applied to a unit quaternion q, gives the point corresponding to q in the tangent space at the identity quaternion. So when we take the logarithm of a quaternion time series, we obtain a series lying entirely in that one specific tangent space.

The angular velocity transformation $log(q_i^{-1}q_{i+1})$ gives the point in the tangent space at q_i corresponding to q_{i+1}. As a consequence, the corresponding time series in $\mathbb{R}^3$ is a collection of points lying in tangent spaces at different points.

Another critical issue that must be taken into account is that in the space $\mathbb{H}_1$ of unit quaternions, the product is not commutative.

As is well known, the formula $\exp(p)\exp(q) = \exp(p+q)$ does not hold in general when p and q do not commute. In this case, the Cambell-Baker-Hausdorff formula (see Baker 1905) for the product of two non-commuting exponentials is applied and in the general case it provides an infinite correction term within the right-hand side of the exponential.

The problem was exactly solved for rotational data in $SO(3)$ (see Condurache and Ciureanu 2020) and in $SU(N)$ (note that $SU(2)$ is isomorph to $\mathbb{H}_1$), see Weigert 1997.

An exact formula to determine the value of the quaternion α such that $\exp(p)\exp(q) = \exp(\alpha)$ is stated in Froelich and Salingaros 1984.

IV.4 Experimental results

In this section we will describe how the different smoothing methods and transformations affect the classification, in order to suggest a rationale with which to proceed in the smoothing of unit quaternion time series.

IV.4.1 Data description

The original data set consists of 54 unit quaternion time series of 101 observations each. The time goes from 0 to 100 (%) in steps of 1%. The data

were recorded by a wearable motion sensor called MetaMotionR (MMR) an Inertial Measurement Unit (IMU) from Mbientlab. It is a device that combines a three-axis accelerometer, a gyroscope, and magnetometer, to determine its orientation in the form of a unit quaternion. It is worn at the level of the hip to measure the angle of rotation of the hip during walking movements, at a frequency of 100 Hz. The signal captured by a motion sensor is periodic and composed of actual walking steps referred to as gait cycles. A gait cycle is defined as the sequence of movements performed by the body during the phase delimited by two successive contacts of a given foot with the ground. We therefore compute an average gait cycle, referred to as the individual gait pattern, by jointly aligning in time and pointwise averaging the segmented gait cycles.

Data related to 27 healthy subjects were collected under two different conditions. The first evaluation was made letting the subject perform a natural walking movement. Another record was made using a knee immobilizer orthosis to simulate a walking impairment.

To represent 3D rotations we choose a unit quaternion representation for convenience, as suggested in the literature on 3D rotation analysis (see, e.g. Dam, Koch, and Lillholm 2000).

A unit quaternion represents a 3D rotation between a given object's frame, or coordinate system (the IMU's coordinate system), and a fixed coordinate system defined as the reference. We choose the first orientation observed of the Individual Gait Pattern (IGP) as the reference, and each unit quaternion of the IGP represents the rotation between this first orientation and the one observed at a given time. For this reason, in the original dataset, the first element of each time series is the quaternion (1 0 0 0), representing the identity rotation. We also processed the data in order to 'straighten' the IGP, so that the first and the last element of the IGP are the identity rotation. In order to apply wavelet methods, the original time series is re-sampled to have 128 time points.

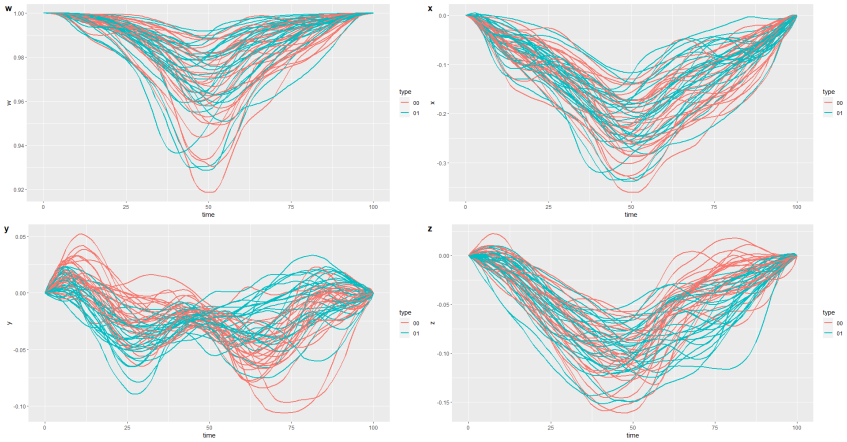


Figure IV.1: Component-wise representation of the individual gait pattern data. Color represents the two conditions.

In order to understand the influence of noise on the performance of the different smoothing methods, we applied the methods described in Section IV.4.2 to data to which different levels of noise had been added. We generated these noisy data sets by adding Gaussian noise to the logarithm of the curves in the original data set, following Ieva et al. 2019. The quaternion time series were transformed to $\mathbb{R}^3$ through the logarithm transformation and then Gaussian noise was added independently to each component. Consider the observation $X_{i,k}$, where i corresponds to the i-th subject and k correspond to the k-th component of the multidimensional time series, $m(t)$ identifies the median line, and ϵ is a Gaussian error term:

$$X_{i,k} = m_k(t) + \epsilon_k(t), \quad Cov(\epsilon_k(s), \epsilon_k(t)) = C(s,t), \quad \forall i = 1, ..., N, \forall i = k, ..., L$$

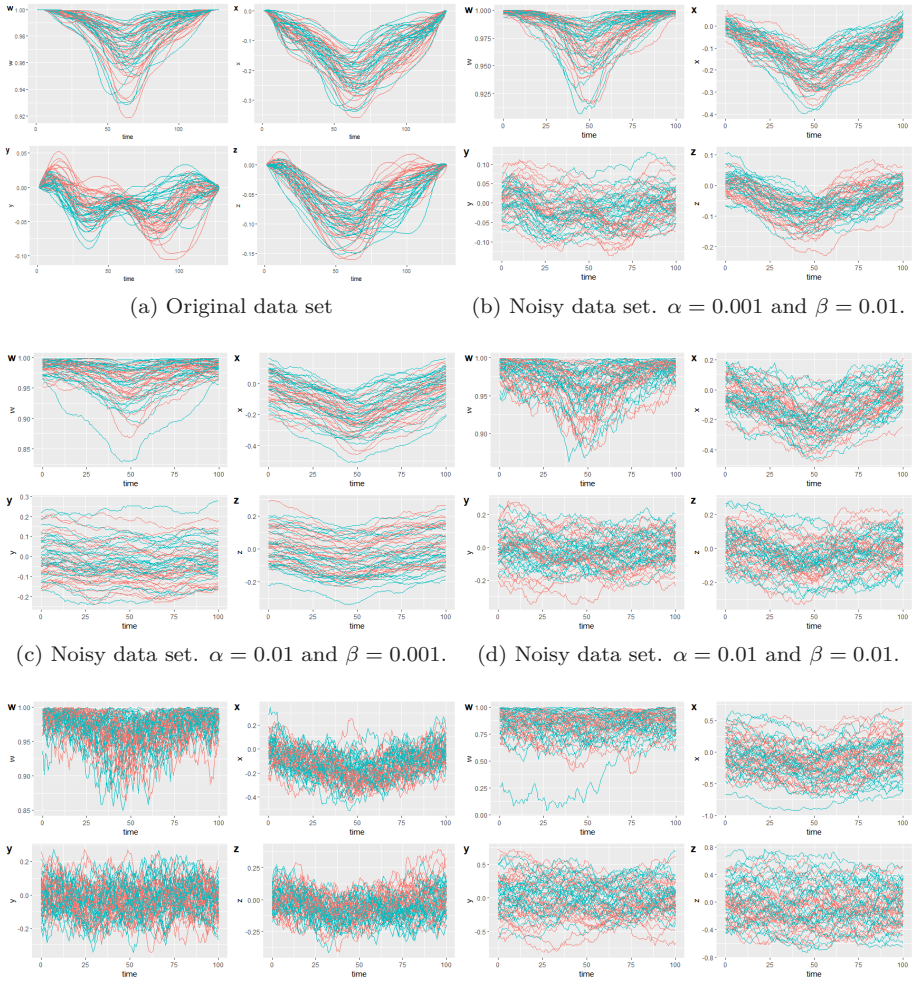
where $Cov(\epsilon_k(s), \epsilon_k(t)) = C(s,t)$ is generated as an exponential-like covariance function with two parameters:

$$C(s,t) = \alpha e^{-\beta|s-t|}. \tag{IV.6}$$

Different degrees and types of noise were simulated by varying the parameters α and β as described in Table IV.1 and visually represented in Figure IV.2.

α	β	Noise
0.001	0.01	Low noise, moderately correlated
0.01	0.001	Moderate noise, highly correlated
0.01	0.01	Moderate noise, moderately correlated
0.01	0.1	Moderate noise, weakly correlated
0.1	0.01	High noise, moderately correlated

Table IV.1: Combination of parameters for the generation of the noisy data.



(a) Original data set

(b) Noisy data set. $\alpha = 0.001$ and $\beta = 0.01$.

(c) Noisy data set. $\alpha = 0.01$ and $\beta = 0.001$.

(d) Noisy data set. $\alpha = 0.01$ and $\beta = 0.01$.

(e) Noisy data set. $\alpha = 0.01$ and $\beta = 0.1$.

(f) Noisy data set. $\alpha = 0.1$ and $\beta = 0.01$.

Figure IV.2: Component-wise representation of the individual gait pattern data with the different levels of noise. The colour indicates which of the two conditions.

IV.4.2 Methods and experimental settings

We compare the wavelet smoothing method with Fourier and spline smoothing, each one embedded in one of the two transformations from $\mathbb{H}_1$ to $\mathbb{R}^3$.

In order to smooth signals using wavelets, the discrete wavelet transform was applied, with soft thresholding in its generalized sense for multidimensional signals, as defined in Pigoli and Sangalli 2012:

$$\bar{\mathbf{w}} = \begin{cases} \mathbf{0}, & \text{if } ||\mathbf{w}|| \leq t_p; \\ (1 - \frac{t_p}{||\mathbf{w}||})\mathbf{w}, & \text{if } ||\mathbf{w}|| > t_p; \end{cases}$$

where $\mathbf{w}$ are the p-dimensional vectors of the detail coefficients of the DWT. The chosen threshold was the universal threshold as generalized in Pigoli and Sangalli 2012: $t_p = \sigma\sqrt{3\log(N)}$ where σ is the standard deviation of the noise. As σ is generally unknown in practical situations, it must be estimated following the idea described in Donoho, Johnstone, and Picard 1995, where the Median Absolute Deviation (MAD) of the details coefficients was proposed ($MAD(\mathbf{x}) = median(|x - median(x)|)$). The estimated standard deviation in the multidimensional case is as follows:

$$\hat{\sigma} = \frac{MAD(\mathbf{d_1})}{0.6745}$$

where $\mathbf{d_1} = \{d^i_{1,k}\}_{k,i}$ is the vector of detail coefficients obtained from the first level of decomposition of each component function (all pooled together).
The following mother wavelets and decomposition levels are considered:

- Mother wavelets: Haar, Daubechies 4 (d4), Daubechies 6 (d6), Daubechies 8 (d8), Daubechies 16 (d16), Least Asymmetric 8 (la8), Least Asymmetric 16 (la16), Least Asymmetric 20 (la20), Best Localized 14 (bl14), Best Localized 20 (bl20).
- Decomposition levels (DLs): from 1 to 6.

The Fourier smoothing was performed through a non-parametric regression smoothing using 20, 40 and 60 basis elements. No covariates and no roughness penalty was used.

Linear, cubic and quintic splines were employed with cross validated parameters for each curve. The number of knots considered is 71. The parameters selected are not optimal, because their optimization it is outside the scope of this paper. For each combination of parameters, the smoothing process is evaluated in terms of classification performance. A k-nearest neighbours (k-NN) model is used to perform classification on the original and on the smoothed quaternion time series in order to select the smoothing method that removes noise while preserving the most important features that can distinguish between two groups.

The k-NN algorithm is a non-parametric classification method first developed in Fix and Hodges 1951. An observation is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours. Here, k is a positive integer, typically small, that we will set to the standard value of 5. The value of k is generally optimized based on data at hand, but that is outside the scope of this paper.

Being a distance-based algorithm, it is easily generalized to quaternion time series using the Dynamic Time Warping distance as defined in Equation IV.8 (IV.B), as suggested in Świtoński, Josiński, and Wojciechowski 2019.

The results presented in the present paper are based on a cross validation exercise, where 5 folds are defined to obtain stable results working with a small sample size.

For each series in the test fold, the distances from the series in the training fold are computed. The 5 nearest time series in the training set are considered and the majority label is assigned to the tested series.

The results are evaluated in terms of accuracy and AUC (area under the ROC curve) and presented as the average taken over the 5 folds.

To increase the robustness of the conclusions, linear regression models were studied to model the influence of the transformations and of the choice of smoothing methods on the performance indices. Each level of noise described in subsection IV.4.1 was simulated three times and the original data were simulated adding a minimal noise setting $\alpha = 0.0001$ and $\beta = 0.0001$. The accuracy and the AUC were evaluated and considered as target variables, and smoothing

method and type of transformation were considered as covariates.
All the computations were performed using the R software (R Core Team (2017)), the figures are generated with the ggplot2 package (v3.3.3; Wickham, 2016) and the plotly package (Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015. https://plot.ly.). The quaternion related functions are provided in the squat package (https://github.com/astamm/squat).

IV.4.3 Classification results

The performances reached by k-NN on the original individual gait pattern data set have an accuracy of 0.8200 and an AUC of 0.9149. When we applied a smoothing process to the data after the logarithm transformation, for all the methods and all the choices of parameters, the accuracy is 0.8100 and almost all the AUCs are 0.8531, with small differences for some combinations of the parameters, as can be seen in Table IV.8, Table IV.9 and Table IV.11. All the values reached using smoothed data are below the performances with the original data set without any type of smoothing.
Considering the angular velocity transformation, the performances are lower than both the original data and he logarithm smoothing process (see Table IV.12, Table IV.13, Table IV.14 and Table IV.15).
This shows that the smoothing process does not improve the classification performance when the curves considered are already nearly smooth. Instead, in some cases, the performances are lower, which seems to suggest that the smoothing process removes some important features in the data that are already exploitable.
The comparison between angular velocity methods and the logarithm shows that when smooth functions are involved, the logarithm better preserves the characteristics of the curves, as can be seen in Figure IV.3. An explanation for this could be that the logarithm transformation is smoother than the angular velocity, which presents a higher variability also for regular curves.

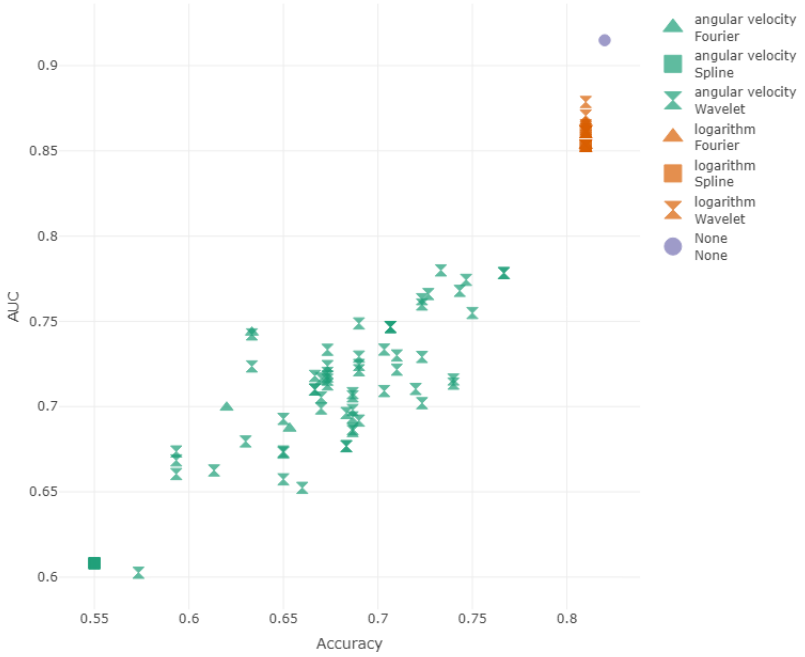


Figure IV.3: Results on original data. Performances of the different methods are evaluated in terms of accuracy and AUC. The shape distinguishes between Fourier, spline or wavelet smoothing methods and colours distinguish between logarithm and angular velocity transformations.

IV.4.3.1 Classification results on noisy data: Analysis of variance at fixed autocorrelation

Now consider the performances reached with noisy data sets as defined in Table IV.1. We start with a fixed value of the autocorrelation ($\beta = 0.01$), increasing the value of the variance parameter ($\alpha = 0.001$, $\alpha = 0.01$, $\alpha = 0.1$). Consider at first the data set generated with $\alpha = 0.001$ and $\beta = 0.01$. We are introducing low levels of noise (the noise has low variance) and a moderate correlation between the nearest points.

The performances reached without any smoothing have an accuracy of 0.550 and an AUC of 0.601. The results regarding the logarithm method are shown in Table IV.16, Table IV.17, Table IV.18 and Table IV.19.

The best methods can be identified as the wavelet smoothing method with

different combinations of parameters: wavelet d6 with 4 in terms of accuracy (accuracy=0.5433 and AUC=0.6527) and wavelet d4 with 5 and 6 decomposition levels in terms of AUC (accuracy=0.49 and AUC=0.6958).

The results regarding the angular velocity method are presented in Table IV.20, Table IV.21, Table IV.22 and Table IV.23.

The angular velocity method achieves poorer results in terms of classification accuracy for all the smoothing functions and choices of parameters (accuracy$\leq$0.47) as can be seen in Figure IV.4. The highest values of AUC are reached with wavelet d16 with 1 decomposition level (accuracy=0.3933 and AUC=0.7434).

It should be noted that almost all the smoothing methods and transformation yield higher AUC but lower accuracy. The only method competitive with the non-smoothed data set is the best of the logarithm transformation.

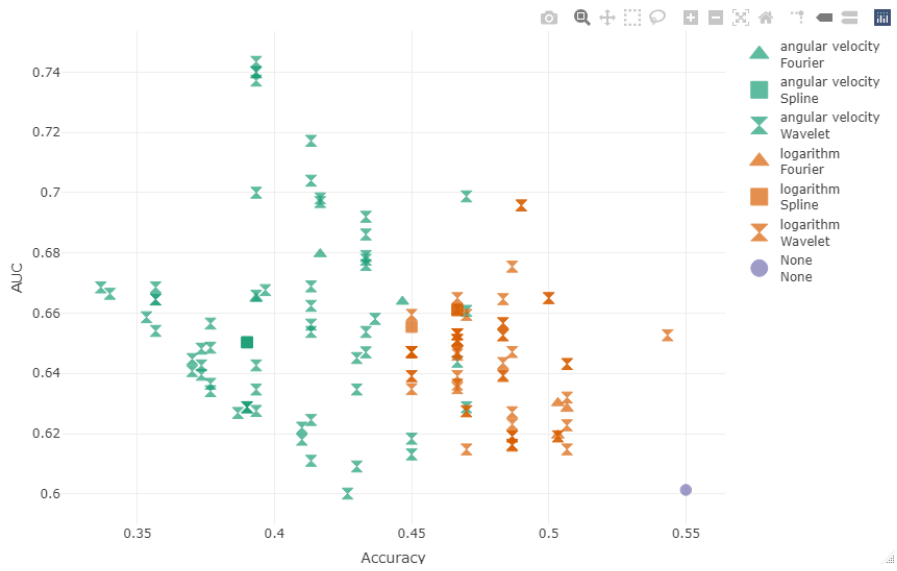


Figure IV.4: Results on noisy data with low levels of noise and moderate correlation between close points. Performances of the different methods are evaluated in terms of accuracy and AUC. Shape distinguishes between Fourier, spline or wavelet smoothing methods and colours distinguish between logarithm and angular velocity transformations.

Now consider the data set with moderate noise variance and moderate correlation ($\alpha = 0.01$ and $\beta = 0.01$). Data classification without any smoothing obtains an accuracy of 0.407 and an AUC of 0.608.

The results regarding the logarithm method are presented in Table IV.32, Table IV.33, Table IV.34 and Table IV.35 and the results regarding the angular velocity method are presented in Table IV.36, Table IV.37, Table IV.38 and Table IV.39.

We can observe that in this case the logarithm transformation performs similarly to the angular velocity and it is difficult to identify the best method as can be seen in Figure IV.5. Almost all the smoothing methods obtain better results than the non-smoothed data set.

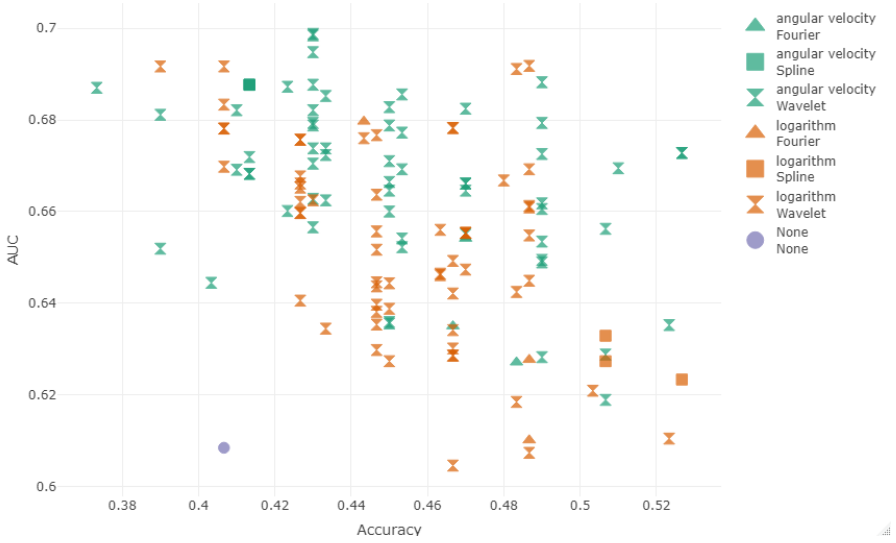


Figure IV.5: Results on noisy data with moderate levels of noise and moderate correlation between close points. Performances of the different methods are evaluated in terms of accuracy and AUC. Shape distinguishes between Fourier, spline or wavelet smoothing methods and colours distinguish between logarithm and angular velocity transformations.

Now consider the noisy data set generated with high noise variance and moderate correlation between close points (α=0.1 and β=0.01). Data

classification without any smoothing reaches an accuracy of 0.5 and an AUC of 0.609.

The results regarding the logarithm method are presented in Table IV.48, Table IV.49, Table IV.50 and Table IV.51 and the results regarding the angular velocity method are presented in Table IV.52, Table IV.53, Table IV.54 and Table IV.55.

We can see that one method reaches better results than the original data classification in terms of the AUC, but with the same accuracy (wavelet la20 with 3 decomposition levels, accuracy=0.5 and AUC=0.6322). Higher values of AUC are reached with lower levels of accuracy: for this reason it is difficult to identify the best method. The logarithm transformation seems to obtain better results in terms of AUC than does the use of the angular velocity, with similar values of accuracy.

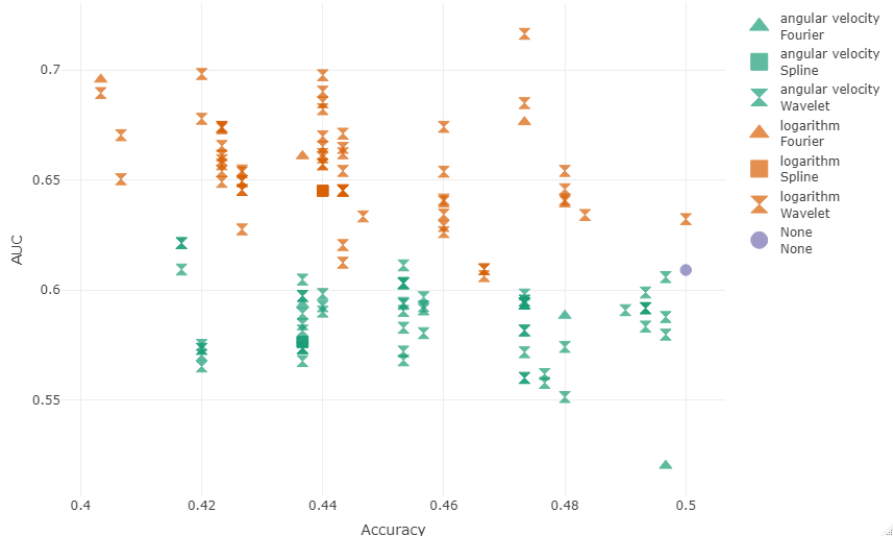


Figure IV.6: Results on noisy data with high levels of noise and moderate correlation between close points. Performances of the different methods are evaluated in terms of accuracy and AUC. Shape distinguishes between Fourier, spline or wavelet smoothing methods and colours distinguish between logarithm and angular velocity transformations.

IV.4.3.2 Classification results on noisy data: Analysis of autocorrelation at fixed variance

Now consider the performances reached with noisy data sets generated with a fixed value of variance ($\alpha = 0.01$), increasing the value of the autocorrelation parameter ($\beta = 0.001$, $\beta = 0.01$, $\beta = 0.1$).

Consider the data set with moderate noise variance ($\alpha = 0.01$) and an high correlation between close points ($\beta = 0.001$). Data classification without any smoothing reaches an accuracy of 0.630 and an AUC of 0.591.

The results regarding the logarithm method are presented in Table IV.24, Table IV.25, Table IV.26 and Table IV.27 and the results regarding the angular velocity method are presented in Table IV.28, Table IV.29, Table IV.30 and Table IV.31.

In this data set the angular velocity performs better than logarithm in terms of AUC, but worse in terms of accuracy. Both the transformations with all the methods obtain worse performances than the raw data classification and no smoothing is suggested, as can be seen in Figure IV.7.

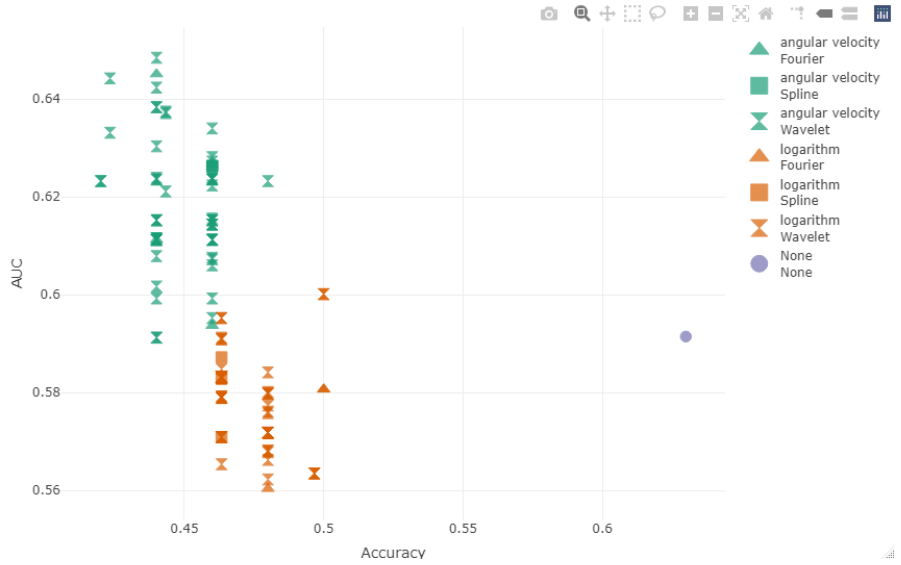


Figure IV.7: Results on noisy data with moderate levels of noise and high correlation between close points. Performances of the different methods are evaluated in terms of accuracy and AUC. Shape distinguishes between Fourier, spline or wavelet smoothing methods and colours distinguish between logarithm and angular velocity transformations.

In general we can observe that when noise levels are low, the smoothing process is not necessary and it increases the risk of removing important features of the data set.

Now consider the data set with moderate noise variance and moderate correlation ($\alpha = 0.01$ and $\beta = 0.01$). As seen before, data classification without any smoothing obtains an accuracy of 0.407 and an AUC of 0.608.

The results regarding the logarithm method are presented in Table IV.32, Table IV.33, Table IV.34 and Table IV.35 and the results regarding the angular velocity method are presented in Table IV.36, Table IV.37, Table IV.38 and Table IV.39.

We can observe that in this case the logarithm transformation performs similarly to the angular velocity and it is difficult to identify the best method as can be seen in Figure IV.5. Almost all the smoothing methods obtain better results than the non-smoothed data set.

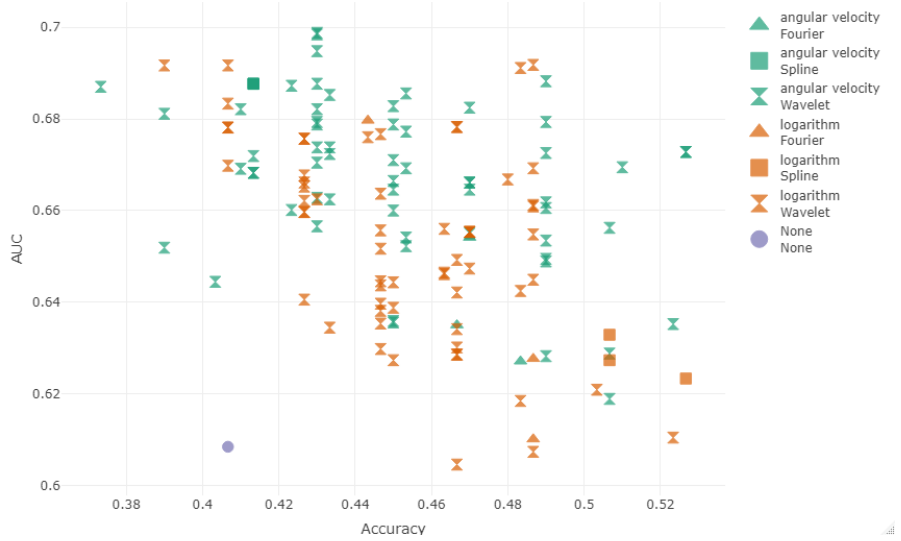


Figure IV.8: Results on noisy data with moderate levels of noise and moderate correlation between close points. Performances of the different methods are evaluated in terms of accuracy and AUC. Shape distinguishes between Fourier, spline or wavelet smoothing methods and colours distinguish between logarithm and angular velocity transformations.

Now consider the noisy data generated with a moderate noise variance and low correlation between close points ($\alpha = 0.01$ and $\beta = 0.1$). Data classification without any smoothing reaches an accuracy of 0.640 and an AUC of 0.678.
The results regarding the logarithm method are presented in Table IV.40, Table IV.41, Table IV.42 and Table IV.43 and the results regarding the angular velocity method are presented in Table IV.44, Table IV.45, Table IV.46 and Table IV.47.
We can see that a lot of smoothing methods reach better results than the original data classification, but only if we consider a smoothing transformation. The angular velocity transformation seems to have lower results. The best result in terms of accuracy is reached by wavelet d4 with 1 decomposition level (accuracy=0.7200, AUC=0.7136). In terms of AUC the best method is wavelet la8 with 4 decomposition levels, accuracy=0.6467 and AUC=0.7416

(see Figure IV.9).

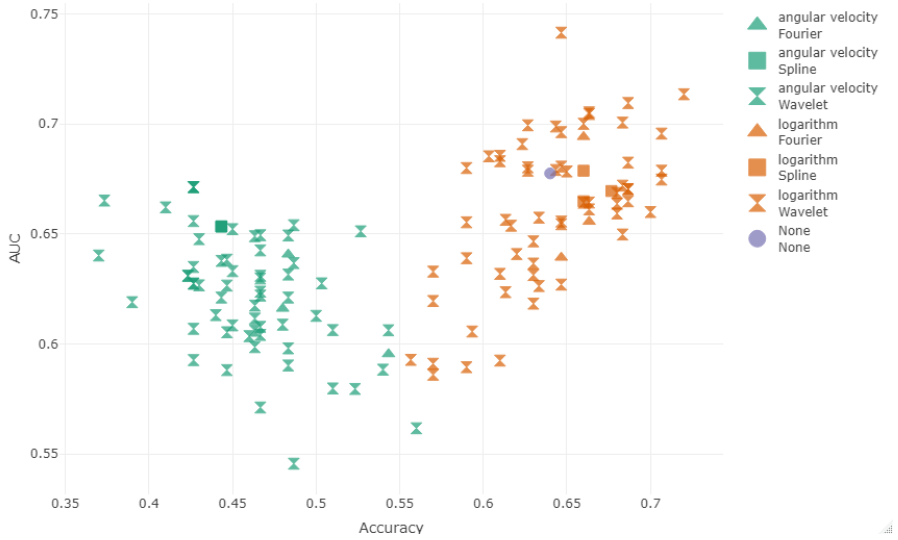
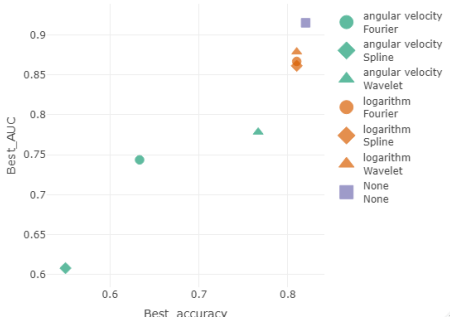


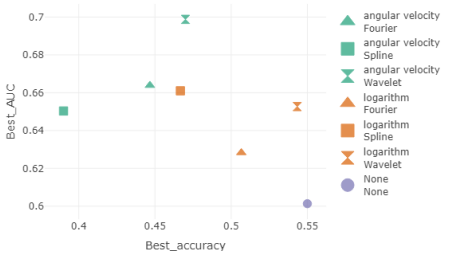
Figure IV.9: Results on noisy data with moderate levels of noise and low correlation between close points. Performances of the different methods are evaluated in terms of accuracy and AUC. Shape distinguishes between Fourier, spline or wavelet smoothing methods and colours distinguish between logarithm and angular velocity transformations.

IV.4.4 Final results

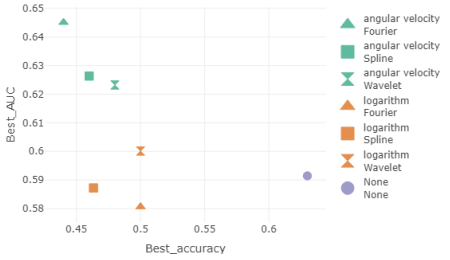
The influence of noise is clear: when the curves are nearly smooth, the smoothing methods can not improve in the classification, whereas when we introduce noise (both in terms of high variance and low autocorrelation), the need for applying smoothing methods becomes clear and the performance can be improved by the process, as we can see in Figure IV.10.



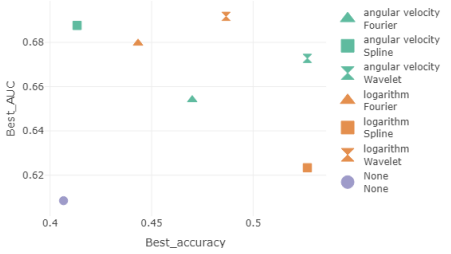
(a) Original data set



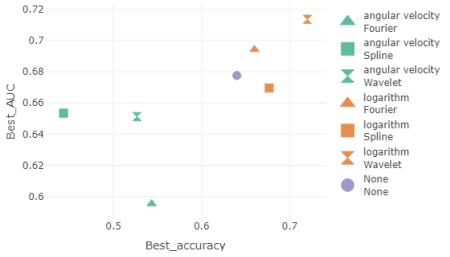
(b) Noisy data set. $\alpha = 0.001$ and $\beta = 0.01$.



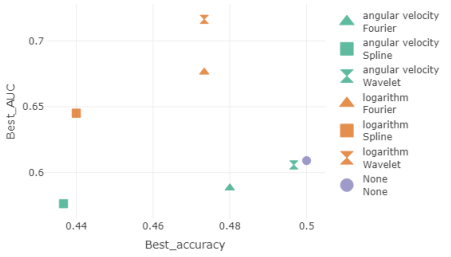
(c) Noisy data set. $\alpha = 0.01$ and $\beta = 0.001$.



(d) Noisy data set. $\alpha = 0.01$ and $\beta = 0.01$.



(e) Noisy data set. $\alpha = 0.01$ and $\beta = 0.1$.



(f) Noisy data set. $\alpha = 0.1$ and $\beta = 0.01$.

Figure IV.10: For each method (Fourier, spline and wavelet) and transformation (logarithm and angular velocity) the best result is presented, where the best result is identified by using the sum of the accuracy and the AUC. Shape distinguishes between Fourier, spline or wavelet methods and colours distinguish between logarithm and angular velocity transformations.

In order to confirm the validity of the proposed method, a linear regression analysis of the accuracy and the AUC has been performed, where the influence of the transformation function and the smoothing method is evaluated. Ten data sets have been generated for each combination of parameters α and β, as defined in Equation IV.6, and the original data set is simulated with the

parameters $\alpha = \beta = 0.0001$. The covariates of the three models are:

- The variable 'transformation' indicates if the transformation function is angular velocity or logarithm. It is a factor variable with reference value 'angular velocity' (two levels).

- The variable 'smoothing_method' indicates if the smoothing method is Fourier, spline or wavelet. It is a factor variable with reference value 'Fourier' (three levels).

- Par alpha and par beta correspond to the noise parameters as defined in Equation (IV.6) and are numerical variables.

The target variables of the three models are the accuracy and the AUC. The logit transformation is applied to each of the target variables to transform the range from [0,1] to $(-\infty, +\infty)$. This produces a larger range of values than the other common transformations. Because the target variables still do not satisfy the normality assumption, a bootstrap procedure is applied to obtain the coefficients and confidence intervals.
The results are summarized in Table IV.2 and Table IV.5. For each model, the ANOVA tables for the linear model are presented in Table IV.3, Table IV.4, Table IV.6, Table IV.7.

Variables	Coefficients	stdev	CI	
Intercept	0.088	0.020	(0.051, 0.125)	*
Transformation logarithm	0.121	0.008	(0.106,0.138)	*
Smoothing method spline	0.011	0.028	(-0.043,0.068)	
Smoothing method wavelet	0.022	0.020	(-0.043,0.068)	
Noise variance (α)	-2.965	0.114	(-3.196, -2.756)	*
Noise autocorrelation (β)	-0.752	0.097	(-0.945,-0.562)	*

Table IV.2: Linear regression model for accuracy target variable with a bootstrap procedure.

Variables	Df	Sum of Sq	Mean Sq	F value	p value	
Transformation	1	28.92	28.921	201.3827	$< 2\,10^{-16}$	***
Smoothing method	2	0.20	0.101	0.7023	0.4955	
Noise variance (α)	1	81.94	81.941	570.5719	$< 2\,10^{-16}$	***
Noise autocorrelation (β)	1	5.50	5.501	8.3024	$6.36\,10^{-10}$	***
Residuals	7914	1136.55	0.144			

Table IV.3: ANOVA table for linear model with accuracy target variable.

	Res Df	RSS	Df	Sum of Sq	F value	p value	
Null model	7919	1253.1					
Final model	7914	1136.5	5	116.56	162.33	$< 2\,10^{-16}$	***

Table IV.4: ANOVA table for linear model with accuracy target variable.

Variables	Coefficients	stddev	CI	
Intercept	0.676	0.015	(0.645,0.707)	*
Transformation logarithm	0.052	0.007	(0.038,0.064)	*
Smoothing method spline	0.002	0.021	(-0.035,0.043)	
Smoothing method wavelet	0.004	0.015	(-0.027,0.033)	
Noise variance (α)	-1.564	0.061	(-1.679,-1.433)	*
Noise autocorrelation (β)	-1.838	0.065	(-1.969,-1.714)	*

Table IV.5: Linear regression model for AUC target variable with a bootstrap procedure.

Variables	Df	Sum of Sq	Mean Sq	F value	p value	
Transformation	1	5.31	5.312	70.2698	$< 2\,10^{-16}$	***
Smoothing method	2	0.01	0.003	0.0406	0.9602	
Noise variance (α)	1	18.69	18.689	247.2439	$< 2\,10^{-16}$	***
Noise autocorrelation (β)	1	32.83	32.829	434.3070	$< 2\,10^{-16}$	***
Residuals	7914	501.55	0.063			

Table IV.6: ANOVA table for linear model with AUC target variable.

	Res Df	RSS	Df	Sum of Sq	F value	p value	
Null model	7919	655.06					
Final model	7914	598.22	5	56.837	150.38	$< 2\,10^{-16}$	***

Table IV.7: ANOVA table for linear model with AUC target variable.

There are certain commonalities of these two target variables. The smoothing methods (wavelet, spline and Fourier) do not seem to have a global impact on the quality of the smoothing process in terms of the classification

performance: the coefficients of the wavelet and spline methods compared to the reference level (Fourier) are not significant. Instead, the coefficient related to the logarithm transformation with respect to the angular velocity transformation is significantly different from zero and positive. The results confirm the positive effects of the logarithm transformation on that target variable. We can also observe that the variance and autocorrelation parameters in the noise generation are significant, with negative coefficients. Higher levels of noise have a negative impact on the classification performances, as can be expected.

IV.5 Conclusions

In this paper, we presented a new method to smooth unit quaternion time series and compared it with the method proposed in C. C. Hsieh 2002.

The main contribution of this new method is to manage unit quaternions in a proper way, transforming the time series to an Euclidean space in order to take advantage of all the existing smoothing techniques. More specifically, we considered wavelet methods and compared this with Fourier and spline smoothing methods.

The results were evaluated in terms of their classification performance on a data set of unit quaternion time series describing walking cycles with a binary outcome variable. Another 5 versions of this data set were created by adding noise to the original data, in order to evaluate the influence of different degrees of noise on the smoothing process.

The results on the original data set and on the noisy ones confirm the need for applying smoothing techniques when the data are noisy and the opportuneness of deploying the proposed method (namely, using the logarithm transformation of unit quaternion time series) to obtain in general better results. Which one of the different smoothing techniques in $\mathbb{R}^3$ should be used depends on the particular data set to be analyzed and should be evaluated on a case by case basis.

Further avenues of research include the application of different noise models to

evaluate the influence of the particular nature of the data set, the application of other classification models, and a deeper analysis of the classical smoothing methods applied in this context. The approach described in this paper can be exploited in terms of the functional representation of quaternion time series, but this aspect needs further study.

The R functions developed for this work will be provided in the squat package.

Appendix IV.A Detailed results

IV.A.1 Original dataset

	Linear	Cubic	Quintic
Accuracy	0.8100	0.8100	0.8100
AUC	0.8611	0.8531	0.8531

Table IV.8: Original data, spline smoothing method, logarithm transformation

	20 basis	40 basis	60 basis
Accuracy	0.8100	0.8100	0.8100
AUC	0.8667	0.8531	0.8531

Table IV.9: Original data, Fourier smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100
d4	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100
d6	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100
d8	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100
d16	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100
la8	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100
la16	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100
la20	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100
bl14	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100
bl20	0.8100	0.8100	0.8100	0.8100	0.8100	0.8100

Table IV.10: Original data, accuracy, wavelet smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.8639	0.8651	0.8707	0.8787	0.8627	0.8627
d4	0.8531	0.8611	0.8611	0.8611	0.8611	0.8611
d6	0.8531	0.8531	0.8531	0.8611	0.8611	0.8611
d8	0.8531	0.8611	0.8611	0.8611	0.8531	0.8611
d16	0.8531	0.8531	0.8531	0.8531	0.8531	0.8531
la8	0.8531	0.8531	0.8611	0.8611	0.8611	0.8611
la16	0.8531	0.8531	0.8531	0.8531	0.8531	0.8531
la20	0.8531	0.8531	0.8531	0.8531	0.8531	0.8531
bl14	0.8531	0.8531	0.8611	0.8611	0.8611	0.8611
bl20	0.8531	0.8531	0.8531	0.8531	0.8531	0.8531

Table IV.11: Original data, AUC, wavelet smoothing method, logarithm transformation

	Linear	Cubic	Quintic
Accuracy	0.5500	0.5500	0.5500
AUC	0.6081	0.6081	0.6081

Table IV.12: Original data, spline smoothing method, angular velocity transformation

	20 basis	40 basis	60 basis
Accuracy	0.6200	0.6333	0.6533
AUC	0.6996	0.7436	0.6872

Table IV.13: Original data, Fourier smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.5733	0.5933	0.5933	0.6133	0.6300	0.5933
d4	0.6833	0.6333	0.6733	0.6333	0.6900	0.6900
d6	0.6500	0.7200	0.7400	0.7033	0.7400	0.7233
d8	0.6867	0.6600	0.7233	0.6867	0.6500	0.6867
d16	0.7100	0.7500	0.7333	0.7467	0.7667	0.7667
la8	0.7033	0.7100	0.6733	0.6733	0.6733	0.6733
la16	0.7067	0.6733	0.6500	0.6667	0.6667	0.6667
la20	0.7067	0.6700	0.6867	0.6500	0.6833	0.6833
bl14	0.6900	0.7233	0.6900	0.7267	0.7233	0.7433
bl20	0.7067	0.6700	0.6900	0.6700	0.6867	0.6867

Table IV.14: Original data, accuracy, wavelet smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.6026	0.6737	0.6684	0.6626	0.6796	0.6604
d4	0.6962	0.7423	0.7130	0.7236	0.7293	0.7213
d6	0.6928	0.7103	0.7134	0.7091	0.7159	0.7020
d8	0.7079	0.6523	0.7291	0.6869	0.6737	0.7060
d16	0.7300	0.7549	0.7799	0.7743	0.7783	0.7783
la8	0.7334	0.7217	0.7177	0.7238	0.7198	0.7158
la16	0.7467	0.7333	0.6730	0.7180	0.7100	0.7100
la20	0.7467	0.7053	0.6980	0.6573	0.6768	0.6768
bl14	0.7244	0.7630	0.7488	0.7660	0.7599	0.7679
bl20	0.7467	0.7164	0.6918	0.6989	0.6937	0.6857

Table IV.15: Original data, AUC, wavelet smoothing method, angular velocity transformation

IV.A.2 First noisy dataset, alpha=0.001, beta=0.01

	Linear	Cubic	Quintic
Accuracy	0.4500	0.4667	0.4667
AUC	0.6554	0.6610	0.6610

Table IV.16: First noisy data set, spline smoothing method, logarithm transformation

	20 basis	40 basis	60 basis
Accuracy	0.5033	0.5033	0.5067
AUC	0.6194	0.6302	0.6283

Table IV.17: First noisy data set, Fourier smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.4833	0.4833	0.4833	0.4833	0.4667	0.4667
d4	0.4500	0.4700	0.4667	0.4867	0.4900	0.4900
d6	0.4500	0.4667	0.4667	0.5433	0.5067	0.5067
d8	0.4667	0.4833	0.4667	0.4867	0.4867	0.4667
d16	0.4500	0.5000	0.5000	0.5033	0.4867	0.5033
la8	0.4500	0.4833	0.4833	0.5067	0.4700	0.4500
la16	0.4500	0.4667	0.4867	0.4867	0.4867	0.4700
la20	0.4500	0.4667	0.4667	0.4867	0.4867	0.4700
bl14	0.4500	0.4667	0.4500	0.5067	0.5067	0.4867
bl20	0.4500	0.4833	0.4667	0.4667	0.4667	0.4667

Table IV.18: First noisy data set, accuracy, wavelet smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.6527	0.6391	0.6394	0.6434	0.6351	0.6462
d4	0.6391	0.6594	0.6613	0.6754	0.6958	0.6958
d6	0.6391	0.6530	0.6530	0.6527	0.6148	0.6228
d8	0.6530	0.6647	0.6511	0.6471	0.6163	0.6363
d16	0.6471	0.6650	0.6650	0.6191	0.6163	0.6191
la8	0.6471	0.6527	0.6567	0.6431	0.6148	0.6348
la16	0.6471	0.6471	0.6271	0.6160	0.6191	0.6274
la20	0.6471	0.6530	0.6391	0.6191	0.6191	0.6274
bl14	0.6471	0.6650	0.6594	0.6431	0.6320	0.6231
bl20	0.6471	0.6567	0.6471	0.6511	0.6511	0.6511

Table IV.19: First noisy data set, AUC, wavelet smoothing method, logarithm transformation

	Linear	Cubic	Quintic
Accuracy	0.3900	0.3900	0.3900
AUC	0.6503	0.6503	0.6503

Table IV.20: First noisy data set, spline smoothing method, angular velocity transformation

	20 basis	40 basis	60 basis
Accuracy	0.3933	0.4167	0.4467
AUC	0.6649	0.6797	0.6639

Table IV.21: First noisy data set, Fourier smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.3933	0.4133	0.3967	0.3900	0.4267	0.3767
d4	0.4333	0.4667	0.4333	0.3933	0.3867	0.3700
d6	0.4133	0.4333	0.3767	0.4700	0.3567	0.3933
d8	0.4700	0.4133	0.4500	0.4367	0.3567	0.3767
d16	0.3933	0.4333	0.4133	0.4700	0.3400	0.3767
la8	0.4133	0.4333	0.4300	0.3733	0.3367	0.3533
la16	0.3933	0.4133	0.4300	0.3933	0.3733	0.3700
la20	0.3933	0.4167	0.4133	0.3733	0.3900	0.3900
bl14	0.4333	0.4333	0.4500	0.4133	0.3567	0.3567
bl20	0.3933	0.4167	0.4300	0.3933	0.4100	0.4100

Table IV.22: First noisy data set, accuracy, wavelet smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.7000	0.6689	0.6677	0.6288	0.6001	0.6566
d4	0.6920	0.6439	0.6470	0.6658	0.6269	0.6448
d6	0.7040	0.6861	0.6341	0.6288	0.6541	0.6347
d8	0.6988	0.6624	0.6183	0.6581	0.6686	0.6486
d16	0.7434	0.6790	0.6246	0.6608	0.6664	0.6366
la8	0.7172	0.6538	0.6451	0.6482	0.6686	0.6587
la16	0.7372	0.6538	0.6347	0.6427	0.6427	0.6408
la20	0.7400	0.6969	0.6110	0.6396	0.6288	0.6288
bl14	0.6760	0.6778	0.6131	0.6562	0.6646	0.6646
bl20	0.7400	0.6981	0.6091	0.6276	0.6220	0.6180

Table IV.23: First noisy data set, AUC, wavelet smoothing method, angular velocity transformation

IV.A.3 Second noisy dataset, alpha=0.01, beta=0.001

	Linear	Cubic	Quintic
Accuracy	0.4633	0.4633	0.4633
AUC	0.5872	0.5709	0.5829

Table IV.24: Second noisy data set, spline smoothing method, logarithm transformation

	20 basis	40 basis	60 basis
Accuracy	0.4800	0.5000	0.5000
AUC	0.5603	0.5807	0.5807

Table IV.25: Second noisy data set, Fourier smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.4633	0.4633	0.4800	0.4800	0.5000	0.5000
d4	0.4633	0.4633	0.4800	0.4967	0.4967	0.4967
d6	0.4633	0.4800	0.4633	0.4800	0.4800	0.4800
d8	0.4633	0.4633	0.4633	0.4800	0.4800	0.4800
d16	0.4633	0.4633	0.4633	0.4800	0.4800	0.4800
la8	0.4633	0.4633	0.4633	0.4633	0.4800	0.4800
la16	0.4633	0.4633	0.4633	0.4633	0.4633	0.4633
la20	0.4633	0.4633	0.4800	0.4800	0.4800	0.4800
bl14	0.4633	0.4633	0.4633	0.4800	0.4800	0.4800
bl20	0.4633	0.4633	0.4633	0.4800	0.4800	0.4800

Table IV.26: Second noisy data set, accuracy, wavelet smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.5952	0.5952	0.5761	0.5841	0.6001	0.6001
d4	0.5832	0.5848	0.5681	0.5634	0.5634	0.5634
d6	0.5832	0.5681	0.5832	0.5758	0.5773	0.5718
d8	0.5912	0.5789	0.5653	0.5662	0.5678	0.5622
d16	0.5832	0.5709	0.5709	0.5718	0.5718	0.5718
la8	0.5832	0.5829	0.5789	0.5909	0.5798	0.5798
la16	0.5832	0.5789	0.5789	0.5829	0.5829	0.5829
la20	0.5832	0.5792	0.5801	0.5718	0.5718	0.5718
bl14	0.5832	0.5792	0.5829	0.5718	0.5718	0.5718
bl20	0.5832	0.5709	0.5909	0.5798	0.5718	0.5718

Table IV.27: Second noisy data set, AUC, wavelet smoothing method, logarithm transformation

	Linear	Cubic	Quintic
Accuracy	0.4400	0.4600	0.4400
AUC	0.6112	0.5937	0.6451

Table IV.28: Second noisy data set, spline smoothing method, angular velocity transformation

	20 basis	40 basis	60 basis
Accuracy	0.5800	0.5600	0.5833
AUC	0.6514	0.6598	0.6798

Table IV.29: Second noisy data set, Fourier smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.4600	0.4400	0.4400	0.4233	0.4600	0.4600
d4	0.4800	0.4400	0.4600	0.4200	0.4400	0.4400
d6	0.4400	0.4233	0.4400	0.4400	0.4400	0.4400
d8	0.4600	0.4600	0.4400	0.4200	0.4400	0.4600
d16	0.4600	0.4400	0.4400	0.4600	0.4400	0.4600
la8	0.4600	0.4600	0.4600	0.4600	0.4400	0.4600
la16	0.4600	0.4600	0.4400	0.4433	0.4600	0.4600
la20	0.4600	0.4400	0.4400	0.4433	0.4600	0.4600
bl14	0.4600	0.4400	0.4400	0.4400	0.4400	0.4400
bl20	0.4600	0.4400	0.4400	0.4433	0.4600	0.4600

Table IV.30: Second noisy data set, accuracy, wavelet smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.6112	0.6112	0.6112	0.6331	0.6239	0.6239
d4	0.6232	0.6152	0.5992	0.6232	0.6116	0.6239
d6	0.6484	0.6442	0.6152	0.6236	0.5912	0.6017
d8	0.6340	0.6223	0.6112	0.6232	0.6079	0.6112
d16	0.6272	0.6303	0.6112	0.6112	0.5912	0.6072
la8	0.6152	0.6143	0.5952	0.6076	0.6116	0.6156
la16	0.6236	0.6143	0.6112	0.6374	0.6112	0.6152
la20	0.6236	0.6383	0.6112	0.6371	0.6112	0.6152
bl14	0.6281	0.6423	0.6112	0.6152	0.5992	0.6236
bl20	0.6236	0.6383	0.6112	0.6211	0.6143	0.6060

Table IV.31: Second noisy data set, AUC, wavelet smoothing method, angular velocity transformation

IV.A.4 Third noisy dataset, alpha=0.01, beta=0.01

	Linear	Cubic	Quintic
Accuracy	0.5267	0.5067	0.5067
AUC	0.6233	0.6273	0.6329

Table IV.32: Third noisy data set, spline smoothing method, logarithm transformation

	20 basis	40 basis	60 basis
Accuracy	0.4433	0.4867	0.4867
AUC	0.6797	0.6101	0.6277

Table IV.33: Third noisy data set, Fourier smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.4867	0.4867	0.4833	0.4867	0.4667	0.4667
d4	0.4867	0.4633	0.4667	0.4800	0.5033	0.5233
d6	0.4333	0.4067	0.4667	0.4833	0.4667	0.4667
d8	0.4500	0.4067	0.3900	0.4467	0.4500	0.4467
d16	0.4867	0.4267	0.4467	0.4867	0.4667	0.4267
la8	0.4700	0.4867	0.4267	0.4833	0.4667	0.4633
la16	0.4700	0.4467	0.4467	0.4067	0.4267	0.4267
la20	0.4700	0.4467	0.4467	0.4067	0.4267	0.4267
bl14	0.4500	0.4067	0.4467	0.4433	0.4300	0.4633
bl20	0.4700	0.4467	0.4267	0.4267	0.4667	0.4467

Table IV.34: Third noisy data set, accuracy, wavelet smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.6609	0.6449	0.6911	0.6918	0.6782	0.6782
d4	0.6609	0.6449	0.6911	0.6918	0.6782	0.6782
d6	0.6344	0.6833	0.6286	0.6424	0.6286	0.6341
d8	0.6443	0.6698	0.6917	0.6767	0.6388	0.6298
d16	0.6692	0.6757	0.6446	0.6548	0.6421	0.6757
la8	0.6553	0.6073	0.6406	0.6184	0.6046	0.6464
la16	0.6553	0.6353	0.6437	0.6917	0.6677	0.6597
la20	0.6553	0.6517	0.6397	0.6781	0.6652	0.6597
bl14	0.6273	0.6781	0.6557	0.6760	0.6624	0.6560
bl20	0.6473	0.6637	0.6621	0.6661	0.6301	0.6381

Table IV.35: Third noisy data set, AUC, wavelet smoothing method, logarithm transformation

	Linear	Cubic	Quintic
Accuracy	0.4133	0.4133	0.4133
AUC	0.6877	0.6877	0.6877

Table IV.36: Third noisy data set, spline smoothing method, angular velocity transformation

	20 basis	40 basis	60 basis
Accuracy	0.4700	0.4667	0.4833
AUC	0.6541	0.6350	0.6271

Table IV.37: Third noisy data set, Fourier smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.4500	0.5067	0.4900	0.4500	0.4500	0.4300
d4	0.3900	0.4900	0.4900	0.4500	0.4533	0.4300
d6	0.4233	0.4900	0.4700	0.4500	0.4333	0.4500
d8	0.4033	0.5233	0.4300	0.4500	0.4900	0.4700
d16	0.4100	0.5067	0.4500	0.4300	0.4900	0.4533
la8	0.4133	0.5267	0.5100	0.4300	0.4533	0.4300
la16	0.3733	0.5267	0.4333	0.4300	0.4700	0.4133
la20	0.3900	0.4900	0.4333	0.4300	0.4700	0.4133
bl14	0.4233	0.5067	0.4900	0.4300	0.4533	0.4333
bl20	0.4100	0.4900	0.4533	0.4300	0.4700	0.4300

Table IV.38: Third noisy data set, accuracy, wavelet smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.6359	0.6562	0.6489	0.6600	0.6356	0.6821
d4	0.6519	0.6606	0.6282	0.6646	0.6772	0.6738
d6	0.6601	0.6793	0.6646	0.6664	0.6852	0.6710
d8	0.6444	0.6352	0.6628	0.6828	0.6882	0.6550
d16	0.6821	0.6288	0.6788	0.6988	0.6726	0.6522
la8	0.6719	0.6728	0.6694	0.6566	0.6856	0.6793
la16	0.6870	0.6728	0.6723	0.6704	0.6661	0.6682
la20	0.6811	0.6494	0.6624	0.6788	0.6661	0.6682
bl14	0.6872	0.6189	0.6534	0.6984	0.6692	0.6738
bl20	0.6691	0.6618	0.6541	0.6948	0.6824	0.6877

Table IV.39: Third noisy data set, AUC, wavelet smoothing method, angular velocity transformation

IV.A.5 Fourth noisy dataset, alpha=0.01, beta=0.1

	Linear	Cubic	Quintic
Accuracy	0.6600	0.6600	0.6767
AUC	0.6649	0.6788	0.6696

Table IV.40: Fourth noisy data set, spline smoothing method, logarithm transformation

	20 basis	40 basis	60 basis
Accuracy	0.6633	0.6600	0.6467
AUC	0.6557	0.6942	0.6393

Table IV.41: Fourth noisy data set, Fourier smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.6200	0.6500	0.6467	0.5700	0.6300	0.5700
d4	0.7200	0.6600	0.5933	0.6133	0.6100	0.6633
d6	0.7067	0.6467	0.5567	0.6133	0.6333	0.6333
d8	0.6633	0.5900	0.6100	0.6467	0.6433	0.6467
d16	0.6833	0.7000	0.5900	0.6100	0.5900	0.6300
la8	0.7067	0.6233	0.6433	0.6467	0.6867	0.6833
la16	0.6867	0.6600	0.6267	0.6633	0.6267	0.6467
la20	0.6867	0.6800	0.5700	0.6633	0.6300	0.6800
bl14	0.7067	0.5900	0.6033	0.6867	0.6267	0.6100
bl20	0.6867	0.6800	0.5700	0.6300	0.6167	0.6833

Table IV.42: Fourth noisy data set, accuracy, wavelet smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.6409	0.6784	0.6270	0.5860	0.6183	0.6196
d4	0.7136	0.6642	0.6057	0.6236	0.6319	0.7046
d6	0.6748	0.6544	0.5928	0.6564	0.6574	0.6263
d8	0.6612	0.6800	0.6831	0.6963	0.6791	0.6557
d16	0.6720	0.6600	0.6390	0.5924	0.5894	0.6467
la8	0.6957	0.6909	0.6989	0.7416	0.7096	0.7007
la16	0.6704	0.7001	0.6994	0.7053	0.6788	0.6807
la20	0.6649	0.6593	0.6329	0.6643	0.6313	0.6638
bl14	0.6788	0.6553	0.6853	0.6824	0.6803	0.6856
bl20	0.6704	0.6687	0.5910	0.6366	0.6538	0.6498

Table IV.43: Fourth noisy data set, AUC, wavelet smoothing method, logarithm transformation

	Linear	Cubic	Quintic
Accuracy	0.4433	0.4433	0.4433
AUC	0.6534	0.6534	0.6534

Table IV.44: Fourth noisy data set, spline smoothing method, angular velocity transformation

	20 basis	40 basis	60 basis
Accuracy	0.4833	0.5433	0.4800
AUC	0.6408	0.5954	0.6161

Table IV.45: Fourth noisy data set, Fourier smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.4633	0.5433	0.5400	0.4267	0.4633	0.4400
d4	0.5267	0.4833	0.4600	0.3900	0.4800	0.4233
d6	0.4867	0.4500	0.4667	0.4300	0.4667	0.4267
d8	0.3733	0.4667	0.4467	0.4300	0.4267	0.4267
d16	0.4833	0.5100	0.4633	0.4500	0.4467	0.4433
la8	0.4667	0.4867	0.4267	0.3700	0.4633	0.4233
la16	0.4467	0.4833	0.4633	0.4833	0.4833	0.4267
la20	0.4467	0.5600	0.4667	0.5000	0.5033	0.4267
bl14	0.4867	0.5100	0.4267	0.4100	0.4667	0.4433
bl20	0.4667	0.5233	0.4667	0.4500	0.4667	0.4267

Table IV.46: Fourth noisy data set, accuracy, wavelet smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.6490	0.6062	0.5883	0.6276	0.6174	0.6131
d4	0.6512	0.5980	0.6036	0.6190	0.6088	0.6310
d6	0.6540	0.6084	0.6220	0.6477	0.6494	0.6273
d8	0.6652	0.5711	0.6053	0.6267	0.6559	0.6350
d16	0.6493	0.5798	0.6060	0.6331	0.6383	0.6211
la8	0.6426	0.5456	0.5927	0.6402	0.6116	0.6310
la16	0.6266	0.5902	0.5986	0.6211	0.6316	0.6713
la20	0.5881	0.5617	0.6078	0.6128	0.6276	0.6713
bl14	0.6368	0.6063	0.6069	0.6621	0.6310	0.6378
bl20	0.6041	0.5796	0.6238	0.6522	0.6298	0.6713

Table IV.47: Fourth noisy data set, AUC, wavelet smoothing method, angular velocity transformation

IV.A.6 Fifth noisy dataset, alpha=0.1, beta=0.01

	Linear	Cubic	Quintic
Accuracy	0.4400	0.4400	0.4400
AUC	0.6451	0.6451	0.6451

Table IV.48: Fifth noisy data set, spline smoothing method, logarithm transformation

	20 basis	40 basis	60 basis
Accuracy	0.4033	0.4367	0.4733
AUC	0.6957	0.6609	0.6763

Table IV.49: Fifth noisy data set, Fourier smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.4600	0.4200	0.4400	0.4400	0.4600	0.4600
d4	0.4233	0.4267	0.4267	0.4233	0.4433	0.4433
d6	0.4267	0.4267	0.4800	0.4433	0.4267	0.4067
d8	0.4600	0.4833	0.4800	0.4400	0.4200	0.4600
d16	0.4400	0.4667	0.4400	0.4400	0.4733	0.4733
la8	0.4233	0.4233	0.4233	0.4433	0.4233	0.4400
la16	0.4433	0.4800	0.4800	0.4433	0.4233	0.4033
la20	0.4433	0.4667	0.5000	0.4433	0.4400	0.4400
bl14	0.4267	0.4467	0.4600	0.4400	0.4233	0.4067
bl20	0.4267	0.4667	0.4600	0.4433	0.4233	0.4433

Table IV.50: Fifth noisy data set, accuracy, wavelet smoothing method, logarithm transformation

DLs	1	2	3	4	5	6
haar	0.6291	0.6778	0.6852	0.6602	0.6263	0.6343
d4	0.6741	0.6276	0.6534	0.6611	0.6451	0.6451
d6	0.6494	0.6454	0.6411	0.6710	0.6543	0.6703
d8	0.6411	0.6341	0.6541	0.6821	0.6981	0.6741
d16	0.6698	0.6063	0.6901	0.6976	0.6849	0.7163
la8	0.6741	0.6491	0.6571	0.6451	0.6540	0.6651
la16	0.6541	0.6402	0.6458	0.6646	0.6734	0.6894
la20	0.6621	0.6094	0.6322	0.6204	0.6571	0.6571
bl14	0.6494	0.6334	0.6538	0.6618	0.6590	0.6503
bl20	0.6454	0.6094	0.6402	0.6454	0.6654	0.6124

Table IV.51: Fifth noisy data set, AUC, wavelet smoothing method, logarithm transformation

	Linear	Cubic	Quintic
Accuracy	0.4367	0.4367	0.4367
AUC	0.5764	0.5764	0.5764

Table IV.52: Fifth noisy data set, spline smoothing method, angular velocity transformation

	20 basis	40 basis	60 basis
Accuracy	0.4367	0.4800	0.4967
AUC	0.5933	0.5884	0.5203

Table IV.53: Fifth noisy data set, Fourier smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.4567	0.4567	0.4533	0.4200	0.4533	0.4367
d4	0.4733	0.4900	0.4733	0.4733	0.4533	0.4200
d6	0.4400	0.4933	0.4733	0.4733	0.4367	0.4200
d8	0.4967	0.4400	0.4733	0.4533	0.4367	0.4367
d16	0.4533	0.4800	0.4733	0.4733	0.4567	0.4200
la8	0.4933	0.4800	0.4167	0.4567	0.4533	0.4367
la16	0.4933	0.4967	0.4533	0.4167	0.4733	0.4367
la20	0.4933	0.4767	0.4367	0.4367	0.4733	0.4367
bl14	0.4400	0.4967	0.4167	0.4733	0.4533	0.4200
bl20	0.4933	0.4767	0.4367	0.4733	0.4533	0.4533

Table IV.54: Fifth noisy data set, accuracy, wavelet smoothing method, angular velocity transformation

DLs	1	2	3	4	5	6
haar	0.5927	0.5912	0.5942	0.5751	0.5829	0.5678
d4	0.5718	0.5909	0.5940	0.5940	0.6032	0.5653
d6	0.5927	0.5989	0.5952	0.5940	0.6048	0.5733
d8	0.6059	0.5982	0.5980	0.5721	0.5817	0.5844
d16	0.5906	0.5514	0.5952	0.5601	0.5804	0.5706
la8	0.5834	0.5742	0.6213	0.5968	0.6032	0.5737
la16	0.5918	0.5878	0.6112	0.6093	0.5817	0.5737
la20	0.5918	0.5579	0.5973	0.5893	0.5817	0.5737
bl14	0.5902	0.5798	0.6213	0.5940	0.5937	0.5733
bl20	0.5918	0.5619	0.5973	0.5601	0.6029	0.5681

Table IV.55: Fifth noisy data set, AUC, wavelet smoothing method, angular velocity transformation

Appendix IV.B The theory of quaternions

IV.B.1 Introduction to quaternions and quaternion algebra

Quaternion algebra has been fully studied and applied since the development of computer graphics, specifically to approach the problem of the motion of a

rigid body.

Quaternions were first defined by Hamilton in 1843, as a generalization of the complex numbers, with three imaginary units: **i, j, k**.

Each quaternion can be represented as $q = w + x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ where w, x, y, and z are real numbers: w is referred to as the real part of q and x, y, and z are the imaginary parts.

Imaginary units satisfy the conditions

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$$

This implies that quaternion multiplication is not commutative. Quaternions are often represented as the 4-dimensional vector of their real components: $q = (w, x, y, z)$.

The norm of a quaternion is defined to be $|q| = \sqrt{w^2 + x^2 + y^2 + z^2}$.

These different types of quaternions and operations on quaternions include:

- The pure quaternion $q = (0, x, y, z)$.
- The identity quaternion $q = (1, 0, 0, 0)$.
- The unit quaternion $q = (w, x, y, z)$, where $|q| = 1$.
- The conjugate quaternion $\bar{q} = (w, -x, -y, -z)$.
- The quaternion inverse $q^{-1} = \bar{q}/|q|$
  If we define $\mathbf{v} = (x, y, z) \in \mathbb{R}^3$:
- The logarithm $\log(q) = (\log(|q|), \frac{x}{|\mathbf{v}|} arccos(\frac{w}{|q|}), \frac{y}{|\mathbf{v}|} arccos(\frac{w}{|q|}), \frac{z}{|\mathbf{v}|} arccos(\frac{w}{|q|}))$,
- The exponential $\exp(q) = \exp(w)(\cos(|\mathbf{v}|), \frac{x}{|\mathbf{v}|}\sin(|\mathbf{v}|), \frac{y}{|\mathbf{v}|}\sin(|\mathbf{v}|), \frac{z}{|\mathbf{v}|}\sin(|\mathbf{v}|))$.

Unit quaternions are specifically used for describing rotations in three dimensions. Their intrinsic properties confer on them a number of advantages over the other classical representations, such as Euler angles and rotation matrices (see Dam, Koch, and Lillholm 2000).

A unit quaternion (i.e. a finite rotation) can also be represented as a single

rotation about an appropriately chosen axis. So the quaternion can also be defined as an angle $\theta \in \mathbb{R}$ and a three element vector $\mathbf{v} = (v_x, v_y, v_z) \in \mathbb{S}^2$, where $\mathbb{S}^2 := \{\mathbf{v} \in \mathbb{R}^3 : \|\mathbf{v}\| = 1\}$:

$$q = q(\theta, \mathbf{v}) = (\cos\frac{\theta}{2}, v_x \sin\frac{\theta}{2}, v_y \sin\frac{\theta}{2}, v_z \sin\frac{\theta}{2})$$

The inverse mapping is defined by the equations:

$$\begin{cases} \theta(q) := 2arccos(w) \\ \mathbf{v}(q) = \frac{(x,y,z)}{|(x,y,z)|} = \frac{(x,y,z)}{\sqrt{1-w^2}} \end{cases}$$

We remark that the map from the unit quaternions to the rotations is not injective: for every rotation, two quaternions, +q and –q, lying at antipodal points of a hypersphere, correspond to it.

For unit quaternions, the exponential and logarithm maps assume specific formulations and meanings. In fact, Euler's identity for complex numbers generalizes to quaternions, i.e. $\exp(\mathbf{v}\theta) = \cos\theta + \mathbf{v}\sin\theta$, results derived from the power series representation for $\exp(x)$. From this formulation is also possible to define the logarithm of a unit quaternion, $\log(q) = \mathbf{v}\theta \in \mathbb{R}^3$. It is important to note that the noncommutativity of quaternion multiplication invalidates the standard identities for the exponential and logarithm functions, as described in Baker 1905.

Let us suppose that $\mathbf{r} = (r_x, r_y, r_z)$ is a point in 3D space and q_r represents the same vector in quaternionic form, $q_r = (0, r_x, r_y, r_z)$.

The vector q'_r, resulting from the rotation by an angle θ around the axis $\mathbf{v}$, can be calculated by quaternion multiplication as $q'_r = qq_rq^{-1}$

Let $q_1 = (w_1, x_1, y_1, z_1)$ and $q_2 = (w_2, x_2, y_2, z_2)$ be unit quaternions. The distance between them is the geodesic distance defined as follows:

$$d(q_1, q_2) = 2arccos(|q_1 \cdot q_2|) = 2arccos(w_1w_2 + x_1x_2 + y_1y_2 + z_1z_2). \quad \text{(IV.7)}$$

This definition is functionally equivalent to the geodesic distance on the unit sphere, defined by $d(q_1, q_2) = \|log(q_1q_2^T)\|$.

The average of n quaternions $q_1, ..., q_n$ is defined as the sample Fréchet mean:

$$\bar{q} = avg(q_i) = argmin_q(\sum_{i=1}^{n} d^2(q, q_i))$$

IV.B.2 Quaternion time series

A quaternion time series is a sequence of (unit) quaternions $q^{(i)}$, $i \in \{1, ..., n\}$. One of the distances between two quaternion time series $\mathbf{q_1}$ and $\mathbf{q_2}$ is defined as the l^2 geodesic distance:

$$d(\mathbf{q_1}, \mathbf{q_2}) = \sqrt{\sum_{i=1}^{n} d^2(q_1^{(i)}, q_2^{(i)})} =$$

$$= \sqrt{\sum_{i=1}^{n} (arccos(w_1^{(i)} w_2^{(i)} + x_1^{(i)} x_2^{(i)} + y_1^{(i)} y_2^{(i)} + z_1^{(i)} z_2^{(i)}))^2}$$

A more elastic distance measure defined in Jablonski 2011 is called Dynamic Time Warping (DTW).

$$DTW(\mathbf{q}_1, \mathbf{q}_2) = \min_{\omega} d_\omega(\mathbf{q}_1, \mathbf{q}_2) = \min_{\omega} \sum_{i=1}^{T} d(q_1^{(\omega(i))}, q_2^{(\omega(i))}) \tag{IV.8}$$

where $\omega(i)$ is the warping path, i.e. a function which defines a mapping between index i to index j. The distances implemented in Equation IV.8 can be chosen according to the purpose, but the geodesic distance as defined in Equation IV.7 is the most common choice.

In the context of quaternion time series, quaternion spherical linear interpolation (SLERP) is an extension of linear interpolation along a plane to spherical interpolation in three dimensions, as first proposed in Shoemake 1985. Given two quaternions, q_1 and q_2, SLERP interpolates a new quaternion, q_0, along the great circle that connects q_1 and q_2:

$$q_0 = \frac{\sin((1-T)\theta)}{\sin(\theta)} q_1 + \frac{\sin(T\theta)}{\sin(\theta)} q_2$$

where T is the interpolation coefficient that determines how close the new quaternion is to either q_1 and q_2, and θ is one-half the angular distance between q_1 and q_2.

The spherical quadrangle interpolation (SQUAD) is a spline-based interpolation of rotations (unit quaternion), also known as spherical cubic interpolation.
If $\{q_i\}_{i=1,...N}$ is a sequence of N quaternions, then define an 'helper' quaternion $s_i = \exp(-\frac{log(q_{i+1}q_i^{-1}) + log(q_{i-1}q_i^{-1})}{4})q_i$.
Then the interpolation is given by

$$squad(q_i, q_{i+1}, s_i, s_{i+1}, T) = slerp(slerp(q_i, q_{i+1}, T), slerp(s_i, s_{i+1}, T), 2T(1-T))$$

where q_i, q_{i+1} represent the start and destination rotations and T is the interpolation parameter, which lies in the interval $[0, 1]$.
Another crucial point in our work is the definition of angular velocity for a quaternion time series. A unit quaternion time series is the discrete representation of a curve $q : T \to \mathbb{H}_1$. The space of unit quaternions can be thought of as the sphere $\mathbb{S}^3 \subseteq \mathbb{R}^4$ so the linear properties can be exploited to define the derivative of a quaternion curve as follows:

$$\dot{q}(t) = \frac{d}{dt}q(t) = \lim_{h \to 0} \frac{q(t+h) - q(t)}{h}$$

As $\mathbb{S}^3$ is a Lie group, the angular velocities of $q(t)$ can be represented as a vector $\Omega(t) \subseteq \mathbb{R}^3$ such that:

$$\dot{q}(t) = \frac{1}{2}q(t) * \begin{bmatrix} 0 \\ \Omega(t) \end{bmatrix}$$

Inverting this equation we obtain that

$$\Omega(t) = \lim_{h \to 0} 2Im(\frac{\bar{q}(t) * q(t+h)}{h})$$

where the imaginary part excludes only the first component, which is always zero in this expression.
As a consequence, $\Omega(t) = 2Im(\frac{\bar{q}(t)*q(t+\delta)}{\delta})$ is a good approximation of the angular velocity for small values of δ.
In order to go back from angular velocity to the unit quaternion time series we can exploit a forward Lie-group Euler method to calculate the next orientation $q(t+\delta)$ from $q(t)$ and $\Omega(t)$, defined as follows:

$$q(t+\delta) = q(t) * e\tilde{x}p(\delta\Omega(t))$$

where $e\tilde{x}p : \mathbb{R}^3 \to \mathbb{S}^3$ is the exponential function in Lie groups, defined as power series. For unit quaternions it has the closed form $e\tilde{x}p(v) = \cos(\frac{1}{2}||v||) + \frac{v}{||v||}\sin(\frac{1}{2}||v||)$ where $v \in \mathbb{R}^3$ and $||\bullet||$ is the standard Euclidean norm (see Rico-Martinez and Gallardo-Alvarado 2000 and Boyle 2017 for further details).

Appendix IV.C Wavelet theory

IV.C.1 Introduction to real wavelets

Wavelet analysis is a mathematical theory developed in the late 1900s to analyse signals from a time–frequency point of view.
Compared to the traditional Fourier method for frequency analysis, wavelets have several advantages. Some of them are that wavelets provide different possibilities for the choice of the basis, so as to better fit the signal, and they analyse a resolution matched to scale. For these reasons, they are a more flexible tool when the signal contains discontinuities and sharp spikes. Moreover, in most of the applications, if the best wavelet adapted to the data is chosen, or if the coefficients are truncated below a threshold, the data are sparsely represented. The procedure for wavelet analysis is to adopt a wavelet to represent the noise, called the analysing wavelet or the mother wavelet, and another wavelet to represent the rest of the signal, called the scaling function or the father wavelet.

For the Fourier transform, this new domain contains basis functions that are sines and cosines. For the wavelet transform, this new domain contains more complicated basis functions called wavelets, mother wavelets, or analyzing wavelets.
While both types of basis function are localized in frequency, only wavelet functions are localized in space. This localization makes many functions and operators using wavelets sparse when transformed into the wavelet domain. This sparseness, in turn, results in a number of useful applications such as data compression, detecting features in images, and removing noise from time series. An advantage of wavelet transforms is that the windows can vary. This is crucial

in order to isolate signal discontinuities and to obtain a detailed frequency analysis. In fact, with wavelet analysis, we have short high-frequency basis functions and long low-frequency ones. The different families of wavelet make different trade-offs between how compactly the basis functions are localized in space and how smooth they are.

IV.C.2 Multiresolution analysis

The time–frequency resolution problem is caused by the Heisenberg uncertainty principle and exists regardless of the technique of analysis used. By using an approach called multiresolution analysis (MRA), it is possible to analyse a signal at different frequencies with different resolutions.
It is assumed that low frequencies last for the entire duration of the signal and give the basic information, whereas high frequencies represent the noise component of the signal. This is often the case in practical applications. Wavelet analysis calculates the correlation between the signal under consideration and a wavelet function $\psi(t)$. The similarity between the signal and the analysing wavelet function is computed separately for different time intervals, resulting in a two-dimensional representation. The analysing wavelet function $\psi(t)$ is also referred to as the mother wavelet.

Definition IV.C.1. Let $\{V_j\}_{j\in\mathbb{Z}}$ be a sequence of closed subspaces $V_j \subseteq L^2(\mathbb{R})$ and let $\phi \in V_0$. An orthogonal multiresolution for $L^2(\mathbb{R})$ is a couple $(\{V_j\}_j, \phi)$ such that:

IV.C.1.1. $V_j \subset V_{j+1}$

IV.C.1.2. $\overline{\cup_j V_j} = L^2(\mathbb{R})$ and $\cap_{j=-\infty}^{+\infty} V_j = \{0\}$

IV.C.1.3. $\{l \mapsto f(l)\} \in V_j \iff \{l \mapsto f(2l)\} \in V_{j+1}$

IV.C.1.4. $\{\phi(l-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_0 and $\int_{\mathbb{R}} \phi \neq 0$.

The projection of $f \in L^2(\mathbb{R})$ on the sequence $\{V_j\}_j$ gives a progressively better approximation of f as j increases. The function ϕ is called the *scaling*

function or the *father wavelet*. Due to properties 3 and 4, $\{2^{\frac{j}{2}}\phi(2^j l-k)\}_k$ is an orthonormal basis for V_j.

It is more useful for exploring the detailed information needed to go from the space V_j to the space V_{j+1}, starting from a coarse space V_0.
For this reason the sequence of complement spaces $W_j = V_{j+1} \setminus V_j$ is introduced.
A mother wavelet is a function $\psi \in W_0$ so that $\{\psi(l-k)\}_k$ is a basis for W_0.
Moreover, the mother wavelet ψ must be a measurable function in $L^2(\mathbb{R}) \cap L^1(\mathbb{R})$, i.e.

$$\int_{-\infty}^{+\infty} |\psi(x)|\, dx < \infty \quad \text{and} \quad \int_{-\infty}^{+\infty} |\psi(x)|^2\, dx < \infty.$$

IV.C.3 The discrete wavelet transform

Dilations and translations of the mother function, or the analyzing wavelet $\psi(x)$, define an orthogonal basis, our wavelet basis:

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j t - k)$$

The parameters j and k are integers that scale and dilate the mother function ψ to generate wavelets. The scale index j indicates the wavelet's width, and the location index l gives its position. Note that the mother functions are rescaled, or dilated by powers of two, and translated by integers. What makes wavelet bases especially interesting is the self-similarity caused by the scales and dilations. Once we know about the mother functions, we know everything about the basis. To span our data domain at different resolutions, the analysing wavelet is used in a scaling equation:

$$\psi(x) = \sum_n h_\psi[n]\sqrt{2}\phi(2t-n)$$

where $\phi(x)$ is the scaling function for the mother function and $h_\psi[n]$ are the wavelet coefficients.
As a consequence of the properties stated in the last paragraph, $L^2(\mathbb{R}) = \oplus_{j\in\mathbb{Z}} W_j$ and $\{\psi_{j,k}(l)\}_k = \{2^{\frac{j}{2}}\psi(2^j l-k)\}_k$ is an orthonormal basis for $L^2(\mathbb{R})$.

Moreover $L^2(\mathbb{R}) = V_0 \oplus W_0 \oplus W_1 \oplus W_2 \oplus ....$

Therefore, for each $f \in L^2(\mathbb{R})$, we have

$$f = \sum_j \sum_k \langle f, \psi_{j,k} \rangle \psi_{j,k} = \sum_k \langle f, \phi_{0,k} \rangle \phi_{0,k} + \sum_{j=0}^{+\infty} \sum_k \langle f, \psi_{j,k} \rangle \psi_{j,k} =$$
$$= \sum_k s_{0,k} \phi_{0,k} + \sum_{j=0}^{+\infty} \sum_k d_{j,k} \psi_{j,k}$$

where $\langle \cdot, \cdot \rangle$ is the scalar product in $L^2(\mathbb{R})$, $s_{0,k} := \langle f, \phi_{0,k} \rangle$ are called the *approximation (or scaling) coefficients* and $d_{j,k} := \langle f, \psi_{j,k} \rangle$ are the *details (or wavelet) coefficients.*

The parameter j represents all the possible *decomposition levels.*

The coefficients $\{s_{0,k}\}_{j \in \mathbb{Z}}$ and $\{d_{j,k}\}_{j \in \mathbb{Z} \cap \{j \geq 0\}, k \in \mathbb{Z}}$ are called the *discrete wavelet transform* (DWT) of f. It can be shown that ϕ and ψ satisfy the dilation equations:

$$\phi(l) = \sum_k \sqrt{2} h_k \phi(2l - k) \quad \text{and} \quad \psi(l) = \sum_k \sqrt{2} g_k \phi(2l - k)$$

for some sequences $\{h_k\}_k$ and $\{g_k\}_k$, called, respectively, the *scaling filter* and the *wavelet filter.*

The iterative process that determines the scaling function is sometimes called the cascade algorithm. The calculation requires an initial $\phi^{(0)}(l)$. In practical examples the initial iteration function is a constant.

The two sets of coefficients $H = \{h_k\}_k$ and $G = \{g_k\}_k$ are known in the signal processing literature as *quadrature mirror filters.*

The scaling function and its defining coefficients H detect localized low frequency information and are called low-pass filters (LPF). The wavelet function and its defining coefficients G detect localized high frequency information and are called high-pass filters (HPF).

For Daubechies wavelets the number of coefficients in H and G, or the length of the filters H and G, denoted by L, is related to the number of vanishing moments M by $2M = L$.

The number of vanishing moments is a regularity property and is expressed as:

$$\int_{-\infty}^{+\infty} \psi(x) x^m \, dx = 0,$$

for $m = 0, ..., M - 1$.

IV.C.4 The Mallat pyramidal algorithm

The theory of wavelets as already presented gives a representation of a real continuous signal $f \in L^2(\mathbb{R})$ as a projection onto an orthonormal basis.
If we assume given a discrete sample of the original signal f, this analysis give us an approximate wavelet representation of that signal (see Resnikoff and Wells 1998) and is called the discrete wavelet transform (DWT).
An efficient way to implement the DWT was proposed in Mallat 1989 and is called the Mallat pyramidal algorithm.
The signal f is convolved with two filters: a high pass filter and a low pass filter. The high pass filter retains the high frequency components (details) and the low pass filter retain low frequency components (approximation). The resulting coefficients are down-sampled to maintain the original size of the data set. The decomposition process can be iterated and successive approximations can be further decomposed. This is called the wavelet decomposition tree. The number of iterations performed is called the *maximum decomposition level* and, as a consequence, is the number of detail levels considered.
While in theory the process can continue indefinitely, in reality the maximum value of decomposition level is $\log_2(N)$. This is due to the Mallat Algorithm and specifically to the down-sampling process: at each step the length of the signal is one-half of the previous length.

IV.C.5 Smoothing using the wavelet transform

The smoothing process in the framework of DWT is essentially carried out by shrinking the wavelet coefficients.

In fact we suppose that the wavelet coefficients can be described by:

$$d_{j,k} = d^0_{j,k} + \rho_{j,k}$$

where $d_{j,k}$ are the empirical wavelet coefficients extracted from the data, $d^0_{j,k}$ are the true wavelet coefficients of the signal without noise, and $\rho_{j,k}$ are the wavelet transforms of the noise.

The general idea of shrinkage is to subtract from the empirical coefficients the values related to noise. This can be done using different approaches.

Hard thresholding :

$$\overline{d}_{j,k} = \begin{cases} 0, & \text{if } |d_{j,k}| \leq T; \\ d_{j,k}, & \text{otherwise} \end{cases}$$

Soft thresholding:

$$\overline{d}_{j,k} = \begin{cases} 0, & \text{if } |d_{j,k}| \leq T; \\ sign(d_{j,k})||d_{j,k}| - T|, & \text{otherwise} \end{cases}$$

Mid thresholding:

$$\overline{d}_{j,k} = \begin{cases} 0, & \text{if } |d_{j,k}| \leq T_1; \\ sign(d_{j,k})||d_{j,k}| - T_2|, & \text{if } T_1 < |d_{j,k}| \leq T_2 \\ d_{j,k} & \text{otherwise} \end{cases}$$

The next step, following the choice of the thresholding method, is to determine the best value for the threshold T.

In general, T can be manually chosen to give what appears to be the right amount of smoothing.

A grid-search approach can be implemented to compare the smoothed signals obtained with different threshold values.

There are also data driven methods to calculate the threshold value and the most well known is the universal threshold as defined in Donoho, Johnstone, and Picard 1995.

The universal threshold is defined so that if the original time series is nothing

but Gaussian noise, then all the wavelet coefficients are (correctly) set to zero using a hard thresholding scheme.
The universal threshold is:

$$\lambda = \sigma\sqrt{2\log(N)}$$

where N is the length of the series and σ is the –unknown– noise variance.

$$\hat{\sigma} = \frac{MAD(\mathbf{d_1})}{0.6745}$$

where $\mathbf{d_1} = \{d_{1,k}\}_k$ is the vector of detail coefficients obtained from the first level of decomposition and MAD is the Median Absolute Deviation: $MAD(\mathbf{x}) = median(|\mathbf{x} - median(\mathbf{x})|)$.

In order to perform a smoothing procedure using DWT, the following steps are necessary:

- Choosing the wavelet function ψ (specifying the family and length L).
- Choosing the decomposition level (i.e. the maximum value of j in the decomposition of f, in practice the number of detail coefficients used in the decomposition).
- Applying the wavelet transform to the data and extracting the wavelet and detail coefficients.
- Choosing the shrinkage method and threshold value.
- Applying the inverse wavelet transform to the smoothed coefficients.

IV.C.6 Evaluation of the smoothing process

There are two main approaches to evaluating the smoothing process, depending on the aim of the research.
The first approach is the estimation of the similarity (or dissimilarity) between the original signal f and the smoothed one f_{smooth}. Here are some examples of common metrics:

- Residual autocorrelation (Horgan 1999)
- P, Q, MS criteria (Sharie, Mosavi, and Rahemi 2020)
- Reconstruction Square Error (Pasti et al. 1999)
- Minimum Descriptor Length (Pasti et al. 1999)
- Energy based methods (Y. Sang 2012)
- Entropy based methods (Y. F. Sang et al. 2009)

The second approach consists in the evaluation of the smoothing process on the basis of the performances of a classification model, as proposed in Zhang et al. 2016.

References

Baker, H. F. (1905). "Alternants and Continuous Groups". In: *Proceedings of the London Mathematical Society* vol. s2-3, no. 1, pp. 24–47.

Boyle, M. (2017). "The Integration of Angular Velocity". In: *Advances in Applied Clifford Algebras* vol. 27, pp. 2345–2374.

Condurache, D. and Ciureanu, I.-A. (2020). "Baker–Campbell–Hausdorff–Dynkin Formula for the Lie Algebra of Rigid Body Displacements". In: *Mathematics* vol. 8, no. 7.

Dam, E., Koch, M., and Lillholm, M. (2000). "Quaternions, Interpolation and Animation". In.

Donoho, D.L., Johnstone I.M.and Kerkyacharian, G., and Picard, D. (1995). "Wavelet Shrinkage: Asymptopia ?" In: *Journal of the Royal Statistical Society. Series B (Methodological)* vol. 57, no. 2, pp. 301–369.

Fang, Y. et al. (1998). "Real time motion fairing with unit quaternions". In: *Computer-Aided Design* vol. 30, no. 3, pp. 191–198.

Fix, E. and Hodges, J.L. (1951). *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, Randolph Field, Texas.

Fletcher, P. (2017). “Quaternion Wavelet transforms of Colour Vector Images”. In: *9th Computer Science and Electronic Engineering (CEEC)*, pp. 168–171.

Fletcher, P. and Sangwine, S.J. (2017). “The development of the quaternion wavelet transform”. In: *Signal Processing* vol. 136, pp. 2–15.

Froelich, J. and Salingaros, N. (1984). “The exponential mapping in Clifford algebras”. In: *Journal of Mathematical Physics* vol. 25 (8), pp. 2347–2350.

Ginzberg, P. (2013). “Quaternion matrices: statistical properties and applications to signal processing and wavelets”. PhD thesis. Imperial College London.

Ginzberg, P. and Walden, A.T. (2012). “Matrix-valued and quaternion wavelets”. In: *IEEE transactions on signal processing* vol. 61, no. 6, pp. 1357–1367.

He, J. X. and Yu, B. (2005). “Wavelet analysis of quaternion-valued time-series”. In: *International Journal of Wavelets, Multiresolution and Information Processing* vol. 03, no. 02, pp. 233–246.

Hitzer, E. (2007). “Quaternion Fourier Transform on Quaternion Fields and Generalizations”. In: *Advances in Applied Clifford Algebras* vol. 17, pp. 497–517.

Horgan, G.W. (1999). “Using wavelets for data smoothing: A simulation study”. In: *Journal of Applied Statistics* vol. 26, no. 8, pp. 923–932.

Hsieh, C. et al. (July 1998). “Noise smoothing for VR equipment in quaternions”. In: *IIE Transactions* vol. 30, pp. 581–587.

Hsieh, C. C. (2002). “Motion Smoothing Using Wavelets”. In: *Journal of Intelligent and Robotic Systems* vol. 35, pp. 157–169.

Ieva, F. et al. (2019). “roahd Package: Robust Analysis of High Dimensional Data”. In: *The R Journal* vol. 11, no. 2, pp. 291–307.

Jablonski, B. (2011). “Quaternion dynamic time warping”. In: *IEEE transactions on signal processing* vol. 60, no. 3, pp. 1174–1183.

Kenwright, B. (2015). “Quaternion Fourier Transform for Character Motions”. In: *Workshop on Virtual Reality Interaction and Physical Simulation*. Ed. by

Jaillet, Fabrice, Zara, Florence, and Zachmann, Gabriel. The Eurographics Association.

Li, S., Leng, J., and Fei, M. (2018). "The quaternion-Fourier transform and applications". In: *International Conference on Communications and Networking in China*. Springer, pp. 157–165.

Mallat, S. G. (1989). "A theory for multiresolution signal decomposition: the wavelet representation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 11, no. 7, pp. 674–693.

Mitrea, M. (1994). *Clifford wavelets, singular integrals, and Hardy spaces*. Springer.

Nielson, Gregory M (2004). "ν-Quaternion splines for the smooth interpolation of orientations". In: *IEEE transactions on visualization and computer graphics* vol. 10, no. 2, pp. 224–229.

Pasti, L. et al. (1999). "Optimization of signal denoising in discrete wavelet transform". In: *Chemometrics and Intelligent Laboratory Systems* vol. 48, no. 1, pp. 21–34.

Peng, L. and Zhao, J. (2004). "Quaternion-valued smooth orthogonal wavelets with short support and symmetry". In: *Advances in Analysis and Geometry*. Springer, pp. 365–376.

Pigoli, D. and Sangalli, L.M. (2012). "Wavelets in functional data analysis: Estimation of multidimensional curves and their derivatives". In: *Computational Statistics & Data Analysis* vol. 56, no. 6, pp. 1482–1498.

Ramamoorthi, R. and Barr, A.H. (1997). "Fast Construction of Accurate Quaternion Splines". In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '97. ACM Press/Addison-Wesley Publishing Co., pp. 287–292.

Resnikoff, H.L. and Wells, R.O.Jr. (1998). *Wavelet Analysis*. Springer-Verlag New York.

Rico-Martinez, J. and Gallardo-Alvarado, J. (2000). "A Simple Method for the Determination of Angular Velocity and Acceleration of a Spherical Motion Through Quaternions". In: *Meccanica* vol. 35, pp. 111–118.

Sang, Y. (2012). “A Practical Guide to Discrete Wavelet Decomposition of Hydrologic Time Series”. In: *Water Resources Management* vol. 26, pp. 3345–3365.

Sang, Y. F. et al. (2009). “Entropy Based Wavelet Denoising Method for Time Series Analysis”. In: *Entropy* vol. 11, pp. 1123–1148.

Sharie, M., Mosavi, M.R., and Rahemi, N. (2020). “Determination of an appropriate mother wavelet for de noising of weak GPS correlation signals based on similarity measurements”. In: *Engineering Science and Technology, an International Journal* vol. 23, no. 2, pp. 281–288.

Shoemake, K. (1985). “Animating Rotation with Quaternion Curves”. In: *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*. Vol. 19. 3, pp. 245–254.

Świtoński, Ad., Josiński, H., and Wojciechowski, K. (July 2019). “Dynamic time warping in classification and selection of motion capture data”. In: *Multidimensional Systems and Signal Processing* vol. 30.

Szczęsna, A., Słupik, J., and Janiak, M. (2012). “The Smooth Quaternion Lifting Scheme Transform for Multi-resolution Motion Analysis”. In: *Proceedings of the 2012 international conference on Computer Vision and Graphics (ICCVG)*, pp. 657–668.

Traversoni, L. (1995). “Quaternions on wavelets problems”. In: *Series in Approximations and Decompositions* vol. 6, pp. 391–398.

— (2001). “Image analysis using quaternion wavelets”. In: *Geometric Algebra with Applications in Science and Engineering*. Springer, pp. 326–345.

Weigert, S. (Dec. 1997). “Baker-Campbell-Hausdorff relation for special unitary groups SU(N)”. In: *Journal of Physics A General Physics* vol. 30, p. 8739.

Xu, Y. et al. (2010). “QWT: Retrospective and New Applications”. In: *Geometric Algebra Computing: in Engineering and Computer Science*, pp. 249–273.

Zhang, Z. et al. (2016). “Choosing Wavelet Methods, Filters, and Lengths for Functional Brain Network Construction”. In: *PLOS ONE* vol. 11, no. 6, pp. 1–24.

Conclusion

This thesis addresses several problems encountered in the field of statistical and machine learning methods for data analysis in neurosciences. The thesis is divided into three parts.

Part 1 is related to the study and improvement of a classical supervised machine learning model, the decision tree model. A new algorithm called Polarized Classification Tree model is defined in order to tackle some weaknesses of classical tree models. In the research field of polarization measures, a new measure is defined and incorporated in the decision tree algorithm as a splitting function. The polarization measure proposed allows the model to take into account the distribution of the predictors instead of the only impurity of the nodes. Results confirm that the new model proposed is competitive with respect to the classical measures and in some cases it shows significantly better performances.

The main contributions of this work are two folds: from a theoretical point of view a generalization of polarization axioms in the multidimensional case is provided and a new measure is defined, from a computational point of view a new classification model is provided. Further ideas of research include the use of Polarized Classification Tree model into ensemble models starting from Random Forest model. An additional planned work is the application of the new model proposed and the comparison with the existing models in a real world problem in order to show the real potential of the proposal for the applications.

Part 2 is about the definition of a model assessment and selection method in a classification task when the target variable is ordinal. The new index can compensate for the lack of appropriate performance evaluation measures for classification models with ordinal target variables. Two toy examples show

how the index works and its advantages with respect to the classical evaluation measures (accuracy, AUC, MSE). Results on simulated data confirm that the new index can capture peculiar aspects compared to the traditional measures. The index proposed is also applied in a real case study. A data set related to the study of Attenuated Psychosis Syndrome is analysed in terms of classification task.

The main contribution of this work is the proposal of an index to perform model selection in the case of ordinal target variables that, coupled with other metrics, can be incorporated in the routine process of model selection. From a computational point of view, the next step will be the implementation of the index in a R package to make it easily available. From a theoretical point of view, a possible extension is the definition of a family of indices where different distances defined in literature are employed and the different consequences of this choice should be discussed.

Part 3 describes a new method to smooth motion data represented as quaternion time series. Different methods to smoothing time series in quaternion algebra were developed in literature and some of these were applied to real or simulate data sets, see for example Ginzberg and Walden 2012, Janiak, Szczęsna, and Słupik 2014, Hsieh 2002. Despite this, the lack of availability of the code makes these methods leaving open the problem of applications to the real world cases. For this reason, starting from the method proposed in Hsieh 2002, a new method that deploys the logarithm function instead of angular velocity to transform the quaternion time series in a real three dimensional time series.

These two methods are compared in terms of classification performances on simulated data sets where different degrees of noise are introduced. The results confirm the hypothesis made on the basis of the theoretical information available from the two methods, i.e. the proposed method generally provides better results than the existing one in terms of classification performances evaluated through accuracy and AUC measures.

From a computational point of view, the R functions developed for this work will be made available.

From a theoretical point of view, further ideas of research include the application of different noise models to evaluate the influence on data sets of different nature, the application of other classification models and a deeper analysis on classical smoothing methods applied in this context. The approach described in this paper can be exploited in terms of the functional representation of quaternion time series, but this aspect needs further study.

Moreover, this work is part of a bigger project and data collected from patients with neurodegenerative diseases will be analysed with the suitable statistical and machine learning methods starting from the ones developed in this thesis.