POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Conformal prediction and copula based methods for profile monitoring

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Niccolò Donadini

Advisor: Prof. Simone Vantini

Co-advisor: Teresa Bortolotti

Academic year: 2023-2024

1. Introduction

In recent years, industrial processes of every type have become increasingly complex, bringing to light the necessity to monitor them continuously, in order to find anomalous patterns in an easy and fast way. For this reason, profile monitoring, a sub-field of Statistical Process Control that deals with anomaly detection, is attracting more and more researchers. The aim of this thesis is to provide a new method, based on conformal prediction and copulae, for recognizing unusual behaviour in functional data. In particular, our methodology is able to associate to a new profile the so called p-value function, which asses how much a new function is strange with respect to in-control profiles, on a scale that ranges from 0 to 1. By testing our proposal with application to well known data and with a simulation study, a problem arises: the current version of the p-value function detects easily an anomalous behaviour if it occurs in the function values, i.e. if we are in presence of very high/low values with respect to the in-control ones. However, the p-value function is useless in spotting unusual patterns that occur at higher order of derivatives; for example, it is not able to distinguish whether a new function has an anomalous increasing/decreasing behaviour. Consequently, in order to solve this problem, we decide to extend the p-value function to include higher order of derivatives. At this point, a copula adjustment is required to obtain a joint coverage level, given the marginal coverage used to build the conformal regions respectively for the functions and their derivatives. After having tested the latter version of the p-value function with applications and simulations, the methodology is applied to a real case study to underline its potential in real scenarios. In particular, we are referring to the Vertical Density Profile data, where each profile measures the density over the vertical axes of a given particleboards.

2. Methods

The basis of our methodology is represented by the Conformal prediction applied in the framework of functional analysis. Specifically, we are going to use the conformal prediction bands, inspired by the work of Diquigiovanni, Fontana and Vantini (2021), to build our p-value function. The idea is the following: a new function y_{n+1} is compared with respect to the conformal predictions bands of level $1-\alpha$, built from a funcional dataset $y_1, \dots, y_n$. Imagine to set $1-\alpha = 0.6$ and, consequently, to compute the conformal region of level 0.6. If we find some

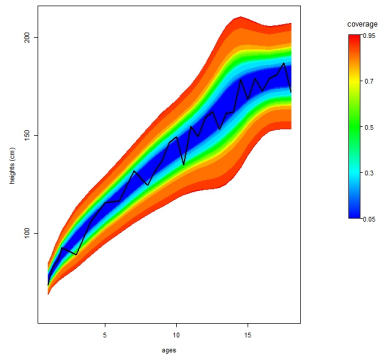
points over the domain such that y_{n+1} intersects the upper or the lower bound of the 0.6 level conformal region, then we will assign to these points the p-value of 0.4. By applying this technique, the more the p-value function is close to 1, the more its corresponding function conforms to the others. On the other hand, a p-value function with low values, near 0, put evidence on the fact that the corresponding function presents an unexpected behaviour. A strength of our method is that the p-value function not only recognizes if a function is an anomaly, but also indicates which points, over the entire domain, are responsible of an unusual pattern. Formally, the p-value function is defined as follows:

$$\forall t \in \mathcal{T} \quad p(t) = \min_{\alpha \in [0,1]} \alpha \quad s.t.$$
$$(u_{1-\alpha} - y_{n+1})^+(t) = 0 \quad \vee$$
$$(y_{n+1} - l_{1-\alpha})^+(t) = 0 \qquad (1)$$

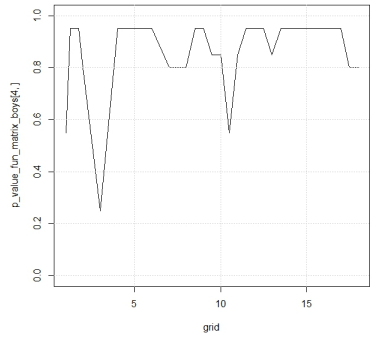
where $u_{1-\alpha}$ and $l_{1-\alpha}$ are respectively the upper and the lower bound of the conformal region of level $1-\alpha$, i.e. $C_{n,1-\alpha}$. $\mathcal{T}$ is the domain of our considered function y_{n+1} and p is the desired output: the p-value function.

Our proposed methodology is now applied to a specific case, built from the Berkeley Growth Study data, which collects the heights of 39 boys and 54 girls and the corresponding ages, included in a range from 1 to 18 years. In particular, we are considering a function where we manually added some noise. This is a perfect example to show the problem of the method and to motivate the reason why a further correction is necessary. Fig. 1 displays the results: it is obvious that the considered function has an anomalous behaviour, since it increase/decrease rapidly. On the other hand, by looking at the colored regions, one can understand that the other functions have a strict increasing trend. Nevertheless, the p-value function has high values and does not detect the presence of any unusual behaviour.

The only way to solve this problem is to consider not only the function itself, but also higher order of derivatives. Following this path, we take into consideration multivariate functional dataset, where the i^{th} multivariate function can be indicated as $\mathbf{y}_i = (y_i^0, y_i^1, \ldots, y_i^m)$ (m indicates the last order of derivative we want to consider, thus the multivariate dimension of the



(a) anomalous function over the conformal regions



(b) p_value function

Figure 1: Berkeley Growth Study data, but only the boys functions. We build an anomaly by adding noise to the boy04 function. This anomaly is plotted over the conformal regions in panel (a), while The corresponding p_value functions is shown in panel (b)

data is m+1). We are now left to the problem of computing the marginal coverage levels, $1-\alpha_0, 1-\alpha_1, \ldots, 1-\alpha_m$, when the joint level, $1-\alpha_g$ is fixed. Inspired by the work of Messoudi, Destercke and Rousseau (2021), we decide to estimate the marginal coverage levels from the following matrix:

$$R = \begin{bmatrix} R_1^0 & R_1^1 & \ldots & R_1^m \\ \vdots & \vdots & & \vdots \\ R_l^0 & R_l^1 & \ldots & R_l^m \end{bmatrix} \qquad (2)$$

where R_i^j is the non conformity scores associated to the i^{th} function in the calibration set, whose dimension will be indicated with l, taken at the j^{th} order of derivative, $j = 0, \ldots, m$. (We are referring to the Split Conformal prediction method, based on the random split of the data

into a training and a calibration set).
From (2), with the help of non parametric copula estimation methods, namely the empirical and/or the kernel copula, we manage to estimate the desired quantities. Without loss of generality, we put every marginal level $\alpha_0, \alpha_1, \ldots, \alpha_m$ equal to the value of α_t, meaning that the conformal prediction bands will have the same coverage $1-\alpha_t$ for every order of derivative. Following the above mentioned procedure, we manage to map every $1-\alpha_t \in [0,1]$ into its related joint value, $1-\alpha_g$. Finally, we are ready to define in a formal way the ultimate version of our work, i.e the p-value function copula adjusted of a function $\mathbf{y}_{n+1}$ with respect to the functional dataset $\mathcal{D}$:

$$\begin{aligned} &\forall t \in \mathcal{T}, \quad \forall j = 0, \ldots, m, \\ &\qquad p^j(t) = \min_{\alpha \in \mathcal{A}_g} \alpha \quad s.t. \\ &(u^j_{1-\alpha} - y^j_{n+1})^+(t) = 0 \quad \vee \\ &\qquad (y^j_{n+1} - l^j_{1-\alpha})^+(t) = 0 \end{aligned} \tag{3}$$

where $\mathcal{A}_g$ is the set with all the corrected values of α, obtained by mapping every α_t in [0,1] into its corresponding joint value. $u^j_{1-\alpha}$ and $l^j_{1-\alpha}$ are, respectively, the upper and the lower bound of the conformal region with coverage $1-\alpha$ associated to the j^{th} order of derivative. $p^j(t) \in [0,1]$ is the value in t of the j^{th} order *p_ value function.*

3. Simulation study

The development of our methodology is supported by a simulation study. For this purpose, we decide to analyze two different simulation scenarios: sinusoidal functions with amplitude and phase variation and splines with different variability over the domain. In both scenarios, we consider a fixed test made of 200 functions and a training set, i.e. the set used to compute the conformal regions, with an increasing dimension on the following logarithmic scale 2,4,16,32,64,128. Our aim is to estimate the empirical coverage, namely the probability that a new function will be all inside a conformal region of level $1-\alpha$, using the p-value function. It is the simplest way to check if our method preserves the theoretical properties behind the conformal prediction approach. In the first part of the simulation, the p-value function is tested against a similar literature version, which we call not adjusted p-value function (because it is computed point by point focusing only on the local behaviour, without referring to conformal prediction). The results shown in Fig. 2 confirm that our p-value function outscores the literature one in estimating the coverage. Thus, as the training set dimension gets bigger, our method reaches the expected target, i.e the theoretical coverage (the red line), while the not adjusted p-value function underestimates it.(for simplicity, we present only the scenario of amplitude and phase variation sinusoidal functions, focusing on the case in which the training set dimension is equal to 128).

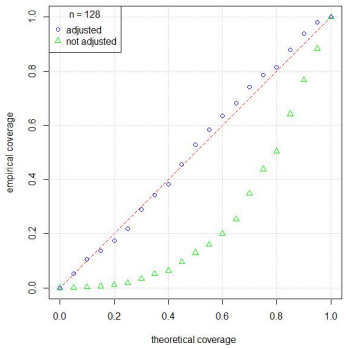
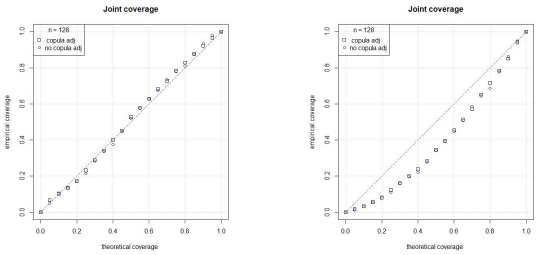


Figure 2: Simulation study considering amplitude and phase variation sinusoidal functions, the test is fixed at M=200 functions, while, for simplicity, we present only the case in which the training set dimension is equal to 128. In particular, the blue circles refer to our version of the p-value function, while the green triangles indicate the estimate in the not adjusted version.

In the second part of the simulation study the codomain extension comes into play, leading us to study functions and first derivatives. So, the focus is now on the differences that come out when the copula adjustment, estimated with the empirical/kernel copula, is taken into consideration. Consequently, by using the same two simulation scenarios presented before, we check the pro and cons of the *copula adjusted* p-value function with respect to the *not copula adjusted* p-value function. To avoid any confusion, we underline that both the two methods are based on conformal prediction bands. Nevertheless, the first one relies also on the copula adjustment while the second does not. In particular,

we compute both the marginal coverage, where functions and first derivatives are investigated separately, and the joint coverage, where we calculate the probability that both a given function and its first derivative are contained in the conformal region of level $1-\alpha$. The goal is to show that the copula adjustment is necessary, in order to obtain p-value functions with the desired joint coverage. The output is shown in Fig.3. The first simulation scenario, sinusoidal functions, is in line with our expectations because both the two versions reach the expected target. However, the splines scenario exhibits a problematic coverage *underestimation* also when the copula correction is considered. We will reserve the right to go deep into this issue, which with high probability is at coding level, later in the future.



(a) (b)

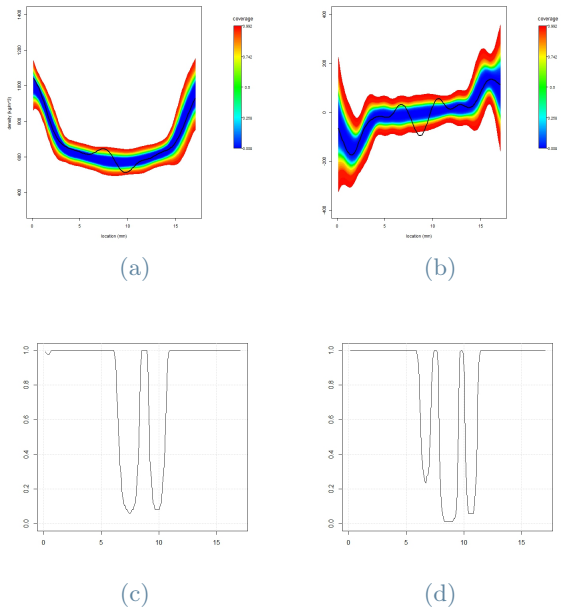
Figure 3: Simulation study considering amplitude and phase variation sinusoidal functions (a) and splines with different variability over the domain (b). The test is fixed at M=200 functions, while, for simplicity, we present only the case in which the training set dimension is equal to 128. In both cases, the empirical copula estimation method is used. The black squares refer to the copula adjusted p-value function, while the blue circles refer to the not copula adjusted p-value function. Only the joint coverage estimates are shown.

4. Real case study

The p-value function, with both the conformal and the copula adjustment, is applied to a real case study. Specifically, we consider the Vertical Density Profiles (VDP), that play a fundamental role in measuring the quality of particleboards. The latter are build through a complex manufacturing process, which needs a frequent control. Taking advantage of the strict relation between the process condition and the VDP curve, it is usual to apply profile monitoring to the VDP data, in order to find anomalous patterns that correspond to failure of the industrial process. The scope of this section is to carry on the monitoring of the above mentioned data, using our copula adjusted p-value function. The dataset under observation is composed of $N = 263$ VDP profiles, each of them measuring the density (kg/m^3) over the vertical axes of a given particleboards, considering a grid of $p = 189$ equally spaced locations. As in the rest of the thesis, we compute the derivatives up to the first order. By looking at Fig. 4, one can observe that profile's copula adjusted p-value function is close to the highest value, except for 2 points in the middle region. Concerning the first derivative, a strange pattern is still occurring in the central part of the domain. One can recognize it by the low values taken by the corresponding p-value function. Putting all together, our analysis underlines the fact that, in the middle of the domain, an anomaly is present not only with respect to the values assumed by the profiles, but also with respect to the increasing/decreasing trend, that is far from the one observed in the other in control functions. We recall that, in this particular example, without the codomain extension, we wouldn't have been able to spot the unusual pattern in the first derivative.

5. Conclusions

Our methodology is born as a technique, based on conformal prediction, not only to detect an anomalous pattern, but also to indicate which are the points, over the domain, responsible of such an unusual pattern. On one hand, simulation study confirm the improvements that our method brings with respect to the not adjusted p-value function. On the other hand, the application on testing data shows all the problems of our proposed methodology in detecting unexpected behaviours at higher order of derivatives. Consequently, we decide to extend our framework, in order to include higher order of derivative in our analysis. This additional step requires a further adjustment, performed by the copula estimation, to ensure an acceptable joint coverage levels. The simulation study on the ultimate version of the p-value function high-



(a) (b)

(c) (d)

Figure 4: VDP dataset, the function considered is an out of control one, obtained manually by modifying the amplitude of the VDP254 in the central part of the domain. In the first row the function and its first derivative are plotted over the copula adjusted conformal prediction bands, respectively in panel (a) and (b). The second row displays the two copula adjusted p-value function, for the function, (c), and for its first derivative, (d).

lights that our proposal is affected by some problems with respect to the achievement of the expected theoretical joint coverage. Nevertheless, the application to the real case study, VDP data, shows all the potential of our copula adjusted p-value function in detecting strange behaviour at any order of derivative. All considered, our proposed methodology can be an additional and useful tool in the field of profile monitoring or, more in general, in every other framework, where the detection of anomaly, unexpected pattern is strongly required.

References

[1] J. Diquigiovanni, M. Fontana and S. Vantini, (2021), The Importance of Being a Band: Finite-Sample Exact Distribution-Free Prediction Sets for Functional Data, arXiv:2102.06746v2

[2] S. Messoudi, S. Destercke and S. Rousseau, (2021), Copula-based conformal prediction for multi-target regression, Pattern Recognition, Université de Technologie de Compiègne

[3] B.M. Colosimo, M. Meneses and Q. Semeraro, (2013), Vertical density profile monitoring using mixed-effects model, ScienceDirect, p. 2.