UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Small and time-efficient distribution-free predictive regions

Victor Cândido Reis

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Victor Cândido Reis

Small and time-efficient distribution-free predictive regions

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Rafael Izbicki

USP – São Carlos
May 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

C375s

Candido Reis, Victor
Small and time-efficient distribution-free predictive regions / Victor Candido Reis; orientador Rafael Izbicki. -- São Carlos, 2023.
44 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Conformal prediction. I. Izbicki, Rafael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Victor Cândido Reis

Regiões preditivas flexíveis, eficientes e livres-de-suposição

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Rafael Izbicki

USP – São Carlos
Maio de 2023

This work is dedicated to my mother Marisa and my father Osny.

ACKNOWLEDGEMENTS

The acknowledgments are directed to my family, my advisor and my friends who helped me to create favorable conditions to carry out this work and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001).

"If I have seen further,
it is by standing on the shoulders of giants."
(Sir Isaac Newton)

RESUMO

REIS, V. C. **Regiões preditivas flexíveis, eficientes e livres-de-suposição**. 2023. 44 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Frequentemente, prever uma variável alvo (resposta) é objeto de interesse de investigações e estudos. Nesse cenário, é comum existirem variáveis mais acessíveis (covariáveis) que podem ajudar no processo de previsão. Métodos de regressão e classificação surgem então com o objetivo de usar as associações estatísticas entre todas as informações disponíveis para modelar a variável de interesse. Há um grande foco, durante tal modelagem, em estimar regiões que descrevam a flutuação da resposta, possibilitando, por exemplo, quantificar a incerteza de estimativas pontuais.

Conformal prediction é uma classe de métodos derivada de Vovk, Gammerman and Shafer (2005) que busca fornecer regiões com formas gerais e garantia de alta probabilidade, assumindo, basicamente, apenas permutabilidade das observações, suposição mais fraca do que dados independentes e identicamente distribuídos, o que permite seu uso extensivo. Novas metodologias têm sido desenvolvidas para aprimorar as propriedades teóricas dessa classe, bem como a aplicabilidade das ideias originais do ponto de vista prático de execução e custo computacional.

Este trabalho objetivou enriquecer a classe de *Conformal prediction* com foco em problemas de regressão, propondo uma nova abordagem que reúne um melhor aproveitamento dos dados com uma maior generalidade no formato das regiões, em uma perspectiva de custo computacional mais eficiente.

Resultados competitivos foram encontrados ao comparar o método proposto com trabalhos anteriores via estudos de simulação.

Palavras-chave: Regressão, *conformal prediction methods*, regiões, eficiência, custo de execução.

ABSTRACT

REIS, V. C. **Small and time-efficient distribution-free predictive regions**. 2023. 44 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Predicting a target variable (response) is often the main objective of many studies and investigations. In such scenarios, there are usually other variables, known as covariates, that are more readily available and can assist in the prediction process. Regression and classification methods aim to utilize the statistical associations between all available information to model the variable of interest. During such modeling, there is a significant emphasis on estimating regions that describe the fluctuations of the response, allowing for the quantification of the uncertainty of point estimates.

Conformal prediction methods (VOVK; GAMMERMAN; SHAFER, 2005) are a class of methods that aim to provide regions with general shapes and high probability guarantees, assuming only exchangeability, which is a weaker assumption than independent and identically distributed data. This allows for extensive use in various applications. New methodologies have been developed to improve the theoretical properties and applicability of the original ideas, with a practical perspective on execution and computational cost.

Motivated by this context, this work aims to enrich the class of conformal prediction methods, with a particular focus on regression problems and proposes a new method that better utilizes available information, provides greater generality in the format of the regions, and is more efficient in terms of computational cost.

The proposed method was compared with previous works using simulation studies, and it achieved competitive results.

Keywords: Regression, conformal prediction methods, regions, efficiency, cost of execution.

LIST OF ALGORITHMS

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

The goal of many statistics and machine learning problems is to predict a response variable Y. In this scenario, it is common to have more accessible variables (covariates $\boldsymbol{x}$) that can help in the prediction process.

Regression and classification methods arise with an interest in using statistical associations between all the information available to model the variable of interest. There is a strong focus, in the modeling, on estimating regions $R(\boldsymbol{x})$ that describe the fluctuation of the response, enabling to quantify the uncertainty of point estimates. Many methods were created with this goal, for example, confidence intervals in generalized linear models (NETER *et al.*, 1996), but the probabilistic guaranties of the methods depend of strong assumptions, such as parametric distributions, shape of regression function, and even in favorable situations intervals may not be the best type of region to get information.

Conformal prediction methods is a class of methods derived from Vovk, Gammerman and Shafer (2005) that seeks to provide regions with general shapes and guarantee of high probability of coverage, assuming, basically, only exchangeability, which is a weaker assumption than independent and identically distributed data, allowing its extensive use. New methodologies have been developed to improve the theoretical properties as well as the applicability of the original ideas from the practical point of view of execution and computational cost.

Two particular strategies presented in Lei *et al.* (2018) and Barber *et al.* (2021) have guided other conformal methods, but while the first in general provides larger regions as consequence of splitting the data into two sets, the case of Split Conformal, the second, Jackknife+, is still computationally expensive depending the method used to estimate the regression, for example.

The work developed by Gupta, Kuchibhotla and Ramdas (2022) establishes an intermediate view balancing the efficiency of using information with the cost of execution,

using bagging. But, unfortunately, this method only provides regions of interval type, unfavorable, for example, in multimodality situations.

The work developed by Izbicki, Shimizu and Stern (2022) makes use of the Split Conformal, presenting deficiencies in the use of available data, but on the other hand, it displays advances in terms of the generality and quality of the regions with solid findings that elucidated propitious conditions and strategies to reach optimal results in terms of size and probabilistic properties.

This work, motivated by the previous context, aims to enrich the class of *conformal prediction methods* exploring aspects of the proposals of Gupta, Kuchibhotla and Ramdas (2022) and Izbicki, Shimizu and Stern (2022), proposing a new one that, in short, brings together a better use of available information with greater generality in the format of the regions in a more efficient perspective of computational cost. The remaining of the work is divided as follows: In Chapter 2 details of the class of conformal prediction methods are presented. Chapter 3 describes the proposed method. In Chapter 4 simulation studies are made to verify the performance of the proposal. Finally, Chapter 5 shows the reached goals and future directions to improve the approach.

CHAPTER

2

CONFORMAL PREDICTION

2.1 Overview

Conformal prediction methods compose a class of methods derived from Vovk, Gammerman and Shafer (2005). Methodologies in this class provide, in a regression problem with response $Y \in \mathbb{R}$ and a random vector of covariates $\boldsymbol{X} \in \mathbb{R}^d$, regions $R(\boldsymbol{x})$ satisfying the marginal coverage property: The probability $\mathbb{P}(Y \in R(\boldsymbol{X})) \geq (1-\alpha)$, where coverage level α is a prefixed small number. The class has, essentially, only one assumption: *exchangeability* of the observations. This assumption has independent and identically distributed data as a particular case. Given a sample $D = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ composed by n observations of random vectors named as labeled data, Algorithm 1 shows one way to group the ideas of the class through a practical example: Split Conformal (LEI *et al.*, 2018).

Algorithm 1 – Split Conformal

procedure SPLIT-CONFORMAL($\boldsymbol{x}, \alpha, D, \mathscr{A}$) ▷ Input: new vector of covariates, coverage level, labeled data and regression algorithmic
  $D_1 = sample(D, \frac{n}{2})$ ▷ Random sample of size $\frac{n}{2}$
  $D_2 = D \setminus D_1$ ▷ Observations in D not present in D_1
  for $(y_i, \boldsymbol{x}_i) \in D_2$ **do**
    $s(y_i, \boldsymbol{x}_i, D_1, \mathscr{A}) = |y_i - \hat{\mu}(\boldsymbol{x}_i)|$ ▷ Estimate the regression function $\hat{\mu}(\cdot)$ through $\mathscr{A}$ with D_1 and evaluate the score in $(y_i, \boldsymbol{x}_i)$
  end for
  $d = quantile(\{s(y_i, \boldsymbol{x}_i, D_1, \mathscr{A}) : (y_i, \boldsymbol{x}_i) \in D_2\}, (\frac{n}{2}+1)(1-\alpha))$ ▷ The kth smallest value in $\{s(y_i, \boldsymbol{x}_i, D_1, \mathscr{A}) : (y_i, \boldsymbol{x}_i) \in D_2\}$, where k is the smallest integer bigger than $(\frac{n}{2}+1)(1-\alpha)$
  $R_{split}(\boldsymbol{x}) \leftarrow [\hat{\mu}(\boldsymbol{x}) - d, \hat{\mu}(\boldsymbol{x}) + d]$ ▷ Estimate the regression function $\hat{\mu}(\cdot)$ through $\mathscr{A}$ with D_1 and evaluate that in $\boldsymbol{x}$
  return $R_{split}(\boldsymbol{x})$ ▷ Output: region
end procedure

Algorithm 1 basically uses one part of labeled data to estimate a centrality measure,

more precisely $\hat{\mu}(\cdot)$ (estimated regression function), and the other to quantify the fluctuation around that with the *nonconformity score* $s(y_i, \boldsymbol{x}_i, D_1, \mathscr{A})$.

2.2 Marginal Coverage

Based on Algorithm 1, assuming independent and identically distributed data and $\hat{\mu}(\cdot)$ as a symmetric function for D_1, it is easy to see through the definition of D_1, the probabilistic guarantee, that is, $\mathbb{P}(Y \in R(\boldsymbol{X})) \geq (1-\alpha)$. Although only exchangeability is assumed, the guarantee still holds; this assumption does not require independence. The term "marginal" refers to the joint distribution of $(Y, \boldsymbol{X})$ which includes the labeled data $D = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^{n}$, splitted in two parts in this example, and the new observation, and is integrated across all of them.

2.3 Other score functions

In Algorithm 1 another important component of conformal methods is exemplified: the score $s(y_i, \boldsymbol{x}_i, D_1, \mathscr{A})$. Scores are the tool chosen to define a relationship of order between responses y given the vector of covariates $\boldsymbol{x}$ with labeled data and some estimation algorithm $\mathscr{A}$; the algorithm $\mathscr{A}$ typically produces a point estimate. This component (score) can be exchanged without problems with the marginal coverage of methods, allowing great flexibility for the class. For instance, the score $s(y_i, \boldsymbol{x}_i, D_1, \mathscr{A}) = |y_i - \hat{\mu}(\boldsymbol{x}_i)|$ (LEI *et al.*, 2018) and the resulting interval region defined in Algorithm 1, may not be suitable for a problem with bimodality, then a method in this class can be used with other score to get better results and interpretations in different contexts.

Another intuitive nonconformity score developed by Lei *et al.* (2018) is:

$$\frac{|\hat{\mu}(\boldsymbol{x}) - y|}{\hat{\sigma}(\boldsymbol{x})},$$

by utilizing $\hat{\sigma}(\boldsymbol{x})$, the estimated standard deviation of the conditional distribution of Y given $\boldsymbol{x}$, the score takes the conditional dispersion as a reference to quantify the distance between the conditional mean and observed value, minimizing a possible effect of the heteroscedasticity.

Scores enable the combination of information about each collected observation $(Y, \boldsymbol{X})$, specifically the features of the conditional distributions given the vector of covariates. Mathematically, the score

$$s : \mathscr{Y} \times \mathscr{X} \times \mathscr{D} \to \mathbb{R}$$

$$(y, \boldsymbol{x}, D, \mathscr{A}) \mapsto s(y, \boldsymbol{x}, D, \mathscr{A}),$$

is a function to express how plausible a response value is compared to another, establishing a logical relationship of order (with domain equals to the cartesian product between the support of Y,$\boldsymbol{X}$ and labeled data). In some nonconformity score for example, responses with more plausibility compared with a fixed response need to have a score less than the score of the fixed response.

This function needs to be symmetric for the elements in $\boldsymbol{D}$ to not interfere in the exchangeability.

The last two arguments from labeled data and estimation algorithm will be omitted to simplify the notation with focus in the new response and new vector of covariates:

$$(y,\boldsymbol{x}) \mapsto s(y,\boldsymbol{x}).$$

The estimation algorithm $\mathscr{A}$ can also be considered a fixed component within the score. The previous separation was used to emphasize the potential of combination with other methods, for example, many regression estimators can be used in Algorithm 1. This component is usually evaluated on the labeled data, thus the estimation algorithm needs to be symmetric on it to allow the same for the score.

The QOOB method (Quantile Out Of Bag), developed by Gupta, Kuchibhotla and Ramdas (2022), utilizes a non-conformity score, developed by Romano, Patterson and Candes (2019), which utilizes more robust statistics to describe the shape of conditional distributions. The expression for this score is as follows:

$$s_Q(y,\boldsymbol{x}) = \max(\hat{q}_\beta(\boldsymbol{x}) - y, y - \hat{q}_{1-\beta}(\boldsymbol{x})) =$$

$$(\mathbb{I}\{y \notin [\hat{q}_\beta(\boldsymbol{x}), \hat{q}_{1-\beta}(\boldsymbol{x})]\} - \mathbb{I}\{y \in [\hat{q}_\beta(\boldsymbol{x}), \hat{q}_{1-\beta}(\boldsymbol{x})]\}) \min(|\hat{q}_\beta(\boldsymbol{x}) - y|, |\hat{q}_{1-\beta}(\boldsymbol{x}) - y|),$$

where $\hat{q}_\beta(\boldsymbol{x})$ and $\hat{q}_{1-\beta}(\boldsymbol{x})$ denote the estimated quantiles with cumulated probability β and $1-\beta$ respectively. This score, in absolute value, represents the minimum distance between the response and the points in the interval defined adaptively for each $\boldsymbol{x}$ by the quantiles; observations outside the interval receive a positive score, while those inside receive a negative score.

The last example is the score of HPD-Split (Highest Predictive Density) created by Izbicki, Shimizu and Stern (2022):

$$s_{HPD}(y,\boldsymbol{x}) = \int \mathbb{I}[\hat{f}_{Y|\boldsymbol{x}}(y') \leq \hat{f}_{Y|\boldsymbol{x}}(y)] \hat{f}_{Y|\boldsymbol{x}}(y') dy',$$

with $\hat{f}_{Y|\boldsymbol{x}}(\cdot)$ as the estimated conditional density of y given $\boldsymbol{x}$.

One critical aspect of Conformal Prediction Methods lies in the evaluation and formulation of scores. It is essential to address the challenge, for example, posed by observations with different covariates, which can exhibit distinct conditional densities.

Consequently, it becomes important to ensure a favorable ordering of these observations, even when confronted with such complex scenarios.

Some scores consider even more aspects, for example, size, format and conditional coverage. The optimal size for a fixed probability is achieved only if the format of the region is flexible and can be adapted to multimodal distributions, which is another desirable feature. The HPD score incorporates all these perspectives seeking better regions. This represents an estimate of probability of getting a new response with density less or equal than the observed given the same vector of covariates. Using this score, the set of y's that satisfies the indicator function can have many shapes, not just interval format.

The flexibility of the scores has encouraged the search for the best region, as previously mentioned. As a result, properties were discovered that aid in obtaining better regions.

2.4 Desirable properties

2.4.1 Conditional and local coverage

Achieving good regions requires more than just marginal coverage. Conditional coverage is a desirable property that can improve the region and make it more informative, but in general it is not attainable (VOVK, 2012; LEI; WASSERMAN, 2014). That consists in:

$$\mathbb{P}(Y \in R(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}) \geq (1-\alpha) \text{ for all } \boldsymbol{x}.$$

A weaker and more tangible property, in the sense of not requiring strong assumptions related with conditional coverage is the local coverage. Informally, instead of giving just a point, that is, a vector of covariates, it is given an arbitrary region of possible vectors of covariates in the probability, for example, a small neighborhood around some point. Asymptotic conditional coverage can be reached using local coverage.

It is interesting to note that conditional coverage implies marginal coverage but the opposite is not true being possible to introduce some examples. Indeed, conditional coverage implies marginal coverage because:

$$\mathbb{P}(Y \in R(\boldsymbol{X})) = \int \mathbb{P}(Y \in R(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}.$$

Assuming conditional coverage with probability prefixed equals to $(1-\alpha)$, that is, $\mathbb{P}(Y \in R(\boldsymbol{X})|\boldsymbol{X}=\boldsymbol{x}) \geq (1-\alpha)$ for all $\boldsymbol{x}$:

$$\begin{aligned}\int \mathbb{P}(Y \in R(\boldsymbol{X})|\boldsymbol{X}=\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} &\geq \int (1-\alpha) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \Rightarrow \\ \mathbb{P}(Y \in R(\boldsymbol{X})) &\geq (1-\alpha) \int f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \Rightarrow \\ \mathbb{P}(Y \in R(\boldsymbol{X})) &\geq (1-\alpha).\end{aligned}$$

2.4.2 Oracle as a possible consequence

Oracle is, by definition, the best region $R(\boldsymbol{x})$ to reach. Oracle provides the smallest region for a fixed conditional coverage (IZBICKI; SHIMIZU; STERN, 2022). Thus, trying to get closer to the conditional coverage, for example, asymptotically with local coverage, can guide the way to reach informative regions. Izbicki, Shimizu and Stern (2022) proved with some additional assumptions that the HPD-Split using $s_{HPD}(\cdot,\cdot)$ reaches the oracle asymptotically, showing the generality of this score.

2.5 Examples of conformal methods

2.5.1 HPD-Split

HPD-Split is an approach to create conformal regions developed by Izbicki, Shimizu and Stern (2022). In this method the dataset is divide in two parts, one to estimate conditional densities $f_{Y|\boldsymbol{x}}(\cdot)$ and another part to evaluate the score preventing bias.

The conformity score of the method presented in section 2.3 is calculated for observations in the second part, then, other observations are used to estimate the conditional densities similar to Algorithm 1.

Let S denote the set of scores that fall below a threshold, determined by previous calculations. The region for a new vector of covariates $\boldsymbol{x}$ is then given by:

$$R_{HPD}(\boldsymbol{x}) = \{y \in \mathscr{Y} \mid s_{HPD}(y,\boldsymbol{x}) \in S\}.$$

2.5.2 Jackknife+

Algorithm 2 presents an alternative method for defining a region with marginal coverage $(1-\alpha)$. This method involves dividing the data, leaving out one observation for future evaluation, and using it for estimation. Despite the efficient use of data, the cost of estimating n regression functions is usually high.

In general, the leave-one-out approach is computationally expensive, requiring n complex estimations using $n-1$ observations to evaluate n scores; in the context of

Algorithm 2 – Jackknife+

procedure JACKKNIFE+($\boldsymbol{x}, \alpha, D$) ▷ Input: new covariates, coverage level, labeled data.
  for $i = (1, \cdots, n)$ **do**
    $s_{-i}(y_i, \boldsymbol{x}_i) = |y_i - \hat{\mu}_{-i}(\boldsymbol{x}_i)|$ ▷ Estimate the regression function with D_{-i}, the set of all observations except i and evaluate the score in $(y_i, \boldsymbol{x}_i)$, saving $\hat{\mu}_{-i}(\boldsymbol{x})$ as well
  end for
  $d_1 = quantile(\{\hat{\mu}_{-i}(\boldsymbol{x}) - s_{-i}(y_i, \boldsymbol{x}_i) : i = (1, \cdots, n)\}, (n+1)(1-\alpha))$ ▷ The jth smallest value in $\{\hat{\mu}_{-i}(\boldsymbol{x}) - s(y_i, \boldsymbol{x}_i) : i = (1, \cdots, n)\}$, where j is the biggest integer lower than $(n+1)(1-\alpha)$
  $d_2 = quantile(\{\hat{\mu}_{-i}(\boldsymbol{x}) + s_{-i}(y_i, \boldsymbol{x}_i) : i = (1, \cdots, n)\}, (n+1)(1-\alpha))$ ▷ The kth smallest value in $\{\hat{\mu}_{-i}(\boldsymbol{x}) + s_{-i}(y_i, \boldsymbol{x}_i) : i = (1, \cdots, n)\}$, where k is the smallest integer bigger than $(n+1)(1-\alpha)$
  $R_{jack+}(\boldsymbol{x}) \leftarrow [d_1, d_2]$
  return $R_{jack+}(\boldsymbol{x})$ ▷ Output: region
end procedure

conditional density estimation with a high-dimensional covariate space, for example, the speed of the method is crucial.

Two types of strategies can be explored to save time: The use of fast estimation methods; or the use of methods where it is not necessary recalculate the components completely when one observation is removed. The QOOB follows the second perspective, estimating trees to build quantile regressions. Each tree is built using a subset of observations and possibly covariates. Next, for each observation i, a quantile regression is estimated using the trees where observation i does not belong to the respective subsets. Additionally, Gupta, Kuchibhotla and Ramdas (2022) conducted a theoretical study that defines a class of scores that share a strategy to ensure marginal coverage.

2.5.3 QOOB

The QOOB method (GUPTA; KUCHIBHOTLA; RAMDAS, 2022) involves the following procedure: First, a random forest is trained using all the data. Next, the trees that do not include each observation i are grouped, and each set of trees is used to estimate two quantiles $q_{\beta_{-i}}(\cdot)$ and $q_{1-\beta_{-i}}(\cdot)$ of the associated conditional distribution. Formally, the number of trees in the forest needs to be randomized for each region construction according to Gupta, Kuchibhotla and Ramdas (2022). However, in practice, Kim, Xu and Barber (2020) found similar results for a fixed number of trees B in many cases, providing insight into this aspect.

The score of this method was presented in section 2.3. To achieve marginal coverage, it is important in the strategy applied in this work not to use the whole random forest in

the estimation of quantiles, instead only trees trained without observation $(Y_i, \boldsymbol{X}_i)$.

$$s_{Q_{-i}}(y, \boldsymbol{x}) = \max(\hat{q}_{\beta_{-i}}(\boldsymbol{x}) - y, y - \hat{q}_{1-\beta_{-i}}(\boldsymbol{x})) =$$

$$(\mathbb{I}\{y \notin [\hat{q}_{\beta_{-i}}(\boldsymbol{x}), \hat{q}_{1-\beta_{-i}}(\boldsymbol{x})]\} - \mathbb{I}\{y \in [\hat{q}_{\beta_{-i}}(\boldsymbol{x}), \hat{q}_{1-\beta_{-i}}(\boldsymbol{x})]\}) \min(|\hat{q}_{\beta_{-i}}(\boldsymbol{x}) - y|, |\hat{q}_{1-\beta_{-i}}(\boldsymbol{x}) - y|).$$

To define the region with coverage level α and marginal coverage of $(1-2\alpha)$, the scores related with possible new responses must satisfy:

$$R_{QOOB}(\boldsymbol{x}) = \left\{ y \in \mathscr{Y} \;\middle|\; \sum_{i=1}^{n} \mathbb{I}[s_{Q_{-i}}(y_i, \boldsymbol{x}_i) < s_{Q_{-i}}(y, \boldsymbol{x})] < (1-\alpha)(n+1) \right\}.$$

CHAPTER

3

PROPOSED APPROACH

Previously, the main concepts and specific examples of Conformal Prediction Methods were discussed, highlighting their particular focus, advantages, and limitations. Many of the negative underlined points of them allow improvements, creating opportunities for future works.

Indeed, the class of Conformal Prediction Methods incorporates desirable characteristics such as wide applicability, modularity, and strong probabilistic properties. However, the use of data in HPD-Split, the interval-only regions provided by QOOB, and the computational expense associated with leave-one-out approaches can impact the final results. Our proposed method tries to overcome these sensitive topics by at least minimizing their effect, ensuring better regions with respect to size and shape.

3.1 The method

In details, our procedure uses a random forest composed by B trees $\mathscr{T} = \{T_j\}_{j=1}^{B}$ trained with the labeled data D to, for each observation $(Y_i, \boldsymbol{X}_i)$, estimate the conditional density for the vector of covariates and the new vector of covariates, $f_{Y|\boldsymbol{x}_i}(\cdot)$ and $f_{Y|\boldsymbol{x}}(\cdot)$, using $\mathscr{T}_{-i}$, the set of trees T trained that did not use the observation i. Using these densities, we compute $s_{HPD_{-i}}(\cdot,\cdot)$. Finally, the region with marginal coverage of $(1-2\alpha)$ is given by:

$$R(\boldsymbol{x}) = \left\{ y \in \mathscr{Y} \;\middle|\; \sum_{i=1}^{n} \mathbb{I}[(1 - s_{HPD_{-i}}(y_i, \boldsymbol{x}_i)) < (1 - s_{HPD_{-i}}(y, \boldsymbol{x}))] < (1-\alpha)(n+1) \right\}.$$

The execution of the method can be thought of in two steps. In the first step, the random forest is trained and scores for labeled data are evaluated. Algorithm 3 describes this stage.

Algorithm 3 – Random forest and scores for labeled data

procedure TRAIN(D, B) ▷ Input: labeled data and number of trees
  $\mathscr{T} \leftarrow \{T\}_{j=1}^{B}$ ▷ Train the random forest
  for $i = (1, \cdots, n)$ **do**
    $\hat{f}_{Y|\boldsymbol{x}_i}(\cdot) \leftarrow density(\boldsymbol{x}_i, \mathscr{T}_{-i})$ ▷ Estimate the density of Y given $\boldsymbol{X} = \boldsymbol{x}_i$ with $\mathscr{T}_{-i}$
    $s_{HPD_{-i}}(y_i, \boldsymbol{x}_i) \leftarrow \int \mathbb{I}[\hat{f}_{Y|\boldsymbol{x}_i}(y) \leq \hat{f}_{Y|\boldsymbol{x}_i}(y_i)] \hat{f}_{Y|\boldsymbol{x}_i}(y) dy$
  end for
  return $\mathscr{T}, \{s_{HPD_{-i}}(y_i, \boldsymbol{x}_i)\}_{i=1}^{n}$ ▷ Output: random forest and scores for each observation
end procedure

In the second step, the region $R(\boldsymbol{x})$ is built using the results of the previous algorithm. Algorithm 4 shows the whole process.

Algorithm 4 – Predictive region for a new vector of covariates

procedure PREDICTION($\boldsymbol{x}, \alpha, \mathscr{T}, \{s_{HPD_{-i}}(y_i, \boldsymbol{x}_i)\}_{i=1}^{n}$) ▷ Input: new vector of covariates, random forest and scores for each observation
  $R(\boldsymbol{x}) \leftarrow \emptyset$
  for $y \in \mathscr{Y}$ **do**
    for $i = (1, \cdots, n)$ **do**
      $\hat{f}_{Y|\boldsymbol{x}}(\cdot) \leftarrow density(\boldsymbol{x}, \mathscr{T}_{-i})$ ▷ Estimate the density of Y given $\boldsymbol{X} = \boldsymbol{x}$ with $\mathscr{T}_{-i}$
      $s_{HPD_{-i}}(y, \boldsymbol{x}) \leftarrow \int \mathbb{I}[\hat{f}_{Y|\boldsymbol{x}}(y') \leq \hat{f}_{Y|\boldsymbol{x}}(y)] \hat{f}_{Y|\boldsymbol{x}}(y') dy'$
    end for
    if $\sum_{i=1}^{n} \mathbb{I}[(1 - s_{HPD_{-i}}(y_i, \boldsymbol{x}_i)) < (1 - s_{HPD_{-i}}(y, \boldsymbol{x}))] < (1 - \alpha)(n + 1)$ **then**
      $R(\boldsymbol{x}) \leftarrow R(\boldsymbol{x}) \cup \{y\}$
    end if
  end for
  return $R(\boldsymbol{x})$ ▷ Output: region
end procedure

Essentially, our method combines the HPD score, allowing better regions in terms of size and shape, with the QOOB's framework, improving the way of using data. A natural follow-up question is how we use the trees for density estimation. Pospisil and Lee (2019) presented RFCDE (Radom Forest for Conditional Density Estimation and Functional Data), the method used in this work, through the package of the same authors in R (<https://www.r-project.org>) with some adaptations. Some codes of the proposed method in this work are available in <https://github.com/victorcandidoreis/conformal202305>. The essence of the method is to model the density as a mixture of distributions (using for example the gaussian kernel) centered at the observed responses, using trees to estimate the weights. The mixture weights (more details later) and a bandwidth summarize the estimation:

$$f_(y|\boldsymbol{x}) = \frac{1}{\sum_j w_j(\boldsymbol{x})} \sum_j w_j(\boldsymbol{x}) K_h(Y_j - y).$$

Pospisil and Lee (2019) use a specific loss for this task (L), the integral of squared distance between the actual and predicted density instead of the traditional mean squared error (MSE). The authors of this work provide a implementation of the method, but the leave-one-out context motivated some adaptations.

$$L = \int (f(y) - \hat{f}(y))^2 dy.$$

Considering this context, the relationship between the weights estimated by Pospisil and Lee (2019) using L and MSE as the loss function, to optimize the forest, was investigated. Simulation studies indicate the normalized square root of MSE weights as a not bad substitute for the density estimation. With the evaluated weights, the implementation proposed by Pospisil and Lee (2019) was used to estimate the bandwidth with the plug-in option.

In the sequence, details are shown to formalize the procedure:

To calculate the score, it is necessary to get the density estimation without using the observation i:

$$f_{-i}(y|\boldsymbol{x}) = \frac{1}{\sum_{j \neq i} w_j(\boldsymbol{x})} \sum_{j \neq i} w_j(\boldsymbol{x}) K_h(Y_j - y),$$

where $K_h(Y_j - y)$ is a normal density (a gaussian kernel) centered in Y_j and bandwidth h. The weights and bandwidth translate the estimated density using the approach in Pospisil and Lee (2019):

$$w_j^*(\boldsymbol{x}) = \frac{1}{|\mathscr{T}_{-i}|} \sum_{T \in \mathscr{T}_{-i}} \frac{\mathbb{I}[\boldsymbol{x}_j \in N_T(\boldsymbol{x})]}{\sum_{m \neq i} \mathbb{I}[\boldsymbol{x}_m \in N_T(\boldsymbol{x})]},$$

where $N_T(\boldsymbol{x})$ the region of the terminal node that includes $\boldsymbol{x}$ for the tree T. In the sequence, the transformation proposed in this work is applied:

$$w_j(\boldsymbol{x}) = \frac{\sqrt{w_j^*(\boldsymbol{x})}}{\sum_{m \neq i} \sqrt{w_m^*(\boldsymbol{x})}}.$$

Lastly, applying the implementation provided by Pospisil and Lee (2019), the bandwidth and the numeric evaluation of the mixture of normal distributions are obtained.

The reason why our approach gives the correct coverage is the following: $s_{HPD}(y, \boldsymbol{x})$ is a member of the class of nested sets (GUPTA; KUCHIBHOTLA; RAMDAS, 2022). Essentially, given $\mathscr{T}_t(\boldsymbol{x})$, a sequence of regions for y indexed by t, such that for $t < t' \Rightarrow \mathscr{T}_t(x) \subset \mathscr{T}_{t'}(x)$; a score defined as $r(x,y) := \inf\{t \in \boldsymbol{T} : y \in \mathscr{T}_t(x)\}$ is a member of this class.

Let's verify the used function of the HPD score as a member. Starting with the HPD score:

$$s_{HPD}(y, \boldsymbol{x}) = \int \mathbb{I}[\hat{f}_{Y|\boldsymbol{x}}(y') \leq \hat{f}_{Y|\boldsymbol{x}}(y)] \hat{f}_{Y|\boldsymbol{x}}(y') dy',$$

define:

$$\mathscr{T}_t^{HPD}(\boldsymbol{x}) = \left\{ y' : s_{HPD}(y', \boldsymbol{x}) \geq (1-t) \right\}.$$

If $t < t'$ then $(1-t) > (1-t')$; for any y that satisfies $s_{HPD}(y,\boldsymbol{x}) \geq (1-t)$ it will satisfy $s_{HPD}(y,\boldsymbol{x}) \geq (1-t')$ too, because $s_{HPD}(y,\boldsymbol{x}) \geq (1-t) > (1-t')$, thus:

$$t < t' \Rightarrow \mathscr{F}_t^{HPD}(\boldsymbol{x}) \subset \mathscr{F}_{t'}^{HPD}(\boldsymbol{x}),$$

and:

$$r(x,y) := \inf\{t \in \boldsymbol{T} : y \in \mathscr{F}_t^{HPD}(x)\} = \inf\left\{t \in \boldsymbol{T} : y \in \left\{y' : s_{HPD}(y',\boldsymbol{x}) \geq (1-t)\right\}\right\} =$$

$$(1 - s_{HPD}(y,\boldsymbol{x})).$$

Informally, $y \in \mathscr{F}_t^{HPD}(\boldsymbol{x})$ if at least its estimated conditional density $\hat{f}_{Y|\boldsymbol{x}}(y)$ is threshold to integrate greater estimated conditional densities of Y, following the definition of $s_{HPD}(y,\boldsymbol{x})$.

CHAPTER

4

EXPERIMENTS

Simulations are made to compare the previous approaches with the proposed in this work. Considering the scenarios and methods present in Izbicki, Shimizu and Stern (2022), two different sample sizes, $n = 1,000$ and $n = 2,500$, were explored. In all of them, $\boldsymbol{X} = (X_1, \cdots, X_d)$, with iid $X_i \sim Unif(-1.5, 1.5)$ (iid - independent identically distributed), $d = 20$ and $\alpha = 0.05$, to get the marginal coverage of at least 0.95. All scenarios have as a challenge irrelevant features (19 covariates).

- (Homoscedastic) $Y|\boldsymbol{x} \sim N(0.3x_1, 1)$.

- (Bimodal) $Y|\boldsymbol{x} \sim 0.5N(f(\boldsymbol{x}) - g(\boldsymbol{x}), \sigma^2(\boldsymbol{x})) + 0.5N(f(\boldsymbol{x}) + g(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$, with $f(\boldsymbol{x}) = (x_1 - 1)^2(x_1 + 1)$, $g(\boldsymbol{x}) = 2\mathbb{I}(x_1 \geq -0.5)\sqrt{x_1 + 0.5}$, and $\sigma^2(\boldsymbol{x}) = 0.25 + |x_1|$.

- (Heteroscedastic) $Y|\boldsymbol{x} \sim N(0.3x_1, 1 + 0.3|x_1|)$.

- (Asymmetric) $Y|\boldsymbol{x} = 1.5x_1 + \varepsilon$, where $\varepsilon \sim Gamma(1 + 0.6|x_1|, 1 + 0.6|x_1|)$.

A brief explanation of some methods in the comparison was adapted from Izbicki, Shimizu and Stern (2022):

- (Reg-split) The regression-split method (LEI *et al.*, 2018), based on the conformal score $|Y_i - \hat{\mu}(\boldsymbol{x}_i)|$.

- (Local Reg-split) The local regression-split method (LEI *et al.*, 2018), based on the conformal score $\dfrac{|Yi - \hat{\mu}(\boldsymbol{x}_i)|}{\hat{p}(\boldsymbol{x}_i)}$, where $\hat{p}(\boldsymbol{x}_i)$ is an estimate of the conditional $(\boldsymbol{X}_i)$ mean absolute deviation of $(Y_i - \mu(\boldsymbol{x}_i))$.

Method	Estimated marginal coverage			
	Bimodal	Heterocedastic	Homocedastic	Asymmetric
HPD-split-FlexCode	0.984	0.974	0.976	0.976
Dist-split+	0.952	0.954	0.934	0.932
Quantile-split	0.94	0.954	0.952	0.936
Reg-split	0.932	0.952	0.938	0.944
Local Reg-split	0.946	0.952	0.938	0.938
HPD-split-Forest	0.948	0.95	0.94	0.936
Proposed Approach	0.948	0.956	0.944	0.95

Table 1 – Estimated marginal coverage for $n = 1,000$. All values are close to 0.95.

• (Quantile-split) The conformal quantile regression method (ROMANO; PATTERSON; CANDES, 2019; SESIA; CANDèS, 2020), based on conformalized quantile regression.

• (Dist-split+) The conformal method from Izbicki, Shimizu and Stern (2020) that uses the cumulative distribution function, $F(y|\boldsymbol{x})$, to create prediction intervals.

For each scenario, 100 vectors of covariates were chosen randomly and fixed. In the sequence, $n+1$ random vectors were taken. Finally, all methods are applied using n vectors to construct the region for the remaining one and for all the 100 fixed 500 times, getting information to check the marginal and conditional coverage, respectively.

The same strategy used to estimate the density in the proposed approach was combined with HPD-split, adding another method named as HPD-Split-Forest to the comparison besides the original way with FlexCode.

Previous simulations indicated overcoverage for the proposed method, allowing the comparison of all methods with $\alpha = 0.05$. Defining $\alpha = 0.05$, the proposed method has as guarantee of marginal coverage $1 - 2\alpha = 0.9$ at least, differently of other methods with a fixed coverage equals to $1 - \alpha = 0.95$. But the actual coverages of proposed approach for all scenarios was around 0.95, indicating the existence of conditions with the threshold greater than $1 - 2\alpha$.

4.1 Marginal coverage

Table 1 shows the estimated marginal coverage for $n = 1,000$. No values indicate deviation of the required coverage. Similar results were obtained for $n = 2,500$ (Table 2). The next results confirm this conclusion since the conditional coverage implies the marginal coverage.

Method	Estimated marginal coverage			
	Bimodal	Heterocedastic	Homocedastic	Asymmetric
HPD-split-FlexCode	0.954	0.958	0.984	0.97
Dist-split+	0.942	0.940	0.952	0.944
Quantile-split	0.948	0.95	0.940	0.942
Reg-split	0.922	0.946	0.932	0.924
Local Reg-split	0.944	0.942	0.946	0.916
HPD-split-Forest	0.944	0.942	0.948	0.938
Proposed Approach	0.952	0.946	0.948	0.962

Table 2 – Estimated marginal coverage for $n = 2,500$. All values are close to 0.95.

4.2 Conditional coverage

In the sequence, the conditional coverage for $n = 1,000$ and $n = 2,500$ was investigated alongside the size of the provided region. Table 3 and Table 4 summarize the simulations, respectively. The average region's size - Avg. size, the average of absolute deviations of estimated conditional coverage to 0.95 Avg. abs. dev. and the average only for negative deviations Avg. of abs. neg. dev. (under coverage compared with 0.95) were obtained. The measures highlight the proposed approach as the best method regarding robustness, with the best value or the second one in each measure and all scenarios.

Scenario	Bimodal			
Measure	Est. coverage	Avg. size	Avg. abs. dev.	Avg. of abs.neg dev.
HPD-split-FlexCode	0.9528	6.5518	0.0426	0.0618
Dist-split+	0.9534	6.5367	0.0394	0.0550
Quantile-split	0.9588	6.8242	0.0410	0.0536
Reg-split	0.9640	7.3593	0.0566	0.0986
Local Reg-split	0.9447	6.2224	0.0434	0.0629
HPD-split-Forest	0.9588	6.7553	0.0460	0.0651
Proposed Approach	0.9613	6.3810	0.0392	0.0492
Scenario	Heterocedastic			
Measure	Est. coverage	Avg. size	Avg. abs. dev.	Avg. of abs.neg dev.
HPD-split-FlexCode	0.9658	8.7756	0.0359	0.0412
Dist-split+	0.9501	7.8919	0.0384	0.0487
Quantile-split	0.9536	7.3356	0.0342	0.0396
Reg-split	0.9493	7.1837	0.0412	0.0513
Local Reg-split	0.9455	7.3421	0.0436	0.0591
HPD-split-Forest	0.9510	7.1135	0.0358	0.0418
Proposed Approach	0.9538	7.0092	0.0305	0.0329
Scenario	Homocedastic			
Measure	Est. coverage	Avg. size	Avg. abs. dev.	Avg. of abs.neg dev.
HPD-split-FlexCode	0.9758	5.2884	0.0320	0.0270
Dist-split+	0.9466	4.4126	0.0272	0.0324
Quantile-split	0.9489	4.2385	0.0216	0.0263
Reg-split	0.9364	3.8783	0.0169	0.0212
Local Reg-split	0.9369	4.1561	0.0368	0.0502
HPD-split-Forest	0.9446	4.1857	0.0214	0.0281
Proposed Approach	0.9481	4.1121	0.0174	0.0221
Scenario	Asymmetric			
Measure	Est. coverage	Avg. size	Avg. abs. dev.	Avg. of abs.neg dev.
HPD-split-FlexCode	0.9802	4.3130	0.0354	0.0341
Dist-split+	0.9444	3.3801	0.0443	0.0747
Quantile-split	0.9590	8.7848	0.0746	0.2992
Reg-split	0.9403	4.1559	0.0628	0.1634
Local Reg-split	0.9285	3.8072	0.0589	0.0963
HPD-split-Forest	0.9465	4.1095	0.0537	0.0990
Proposed Approach	0.9512	3.5430	0.0412	0.0647

Table 3 – Average region's size - Avg. size, the average of absolute deviations of estimated conditional coverage to 0.95 Avg. abs. dev. and the average only for negative deviations Avg. of abs. neg. dev. for $n = 1,000$ and scenarios in Izbicki, Shimizu and Stern (2022).

Scenario	Bimodal			
Measure	Est. coverage	Avg. size	Avg. abs. dev.	Avg. of abs.neg dev.
HPD-split-FlexCode	0.9535	6.2650	0.0361	0.0467
Dist-split+	0.9530	6.2211	0.0335	0.0424
Quantile-split	0.9591	6.4501	0.0338	0.0398
Reg-split	0.9656	7.3048	0.0567	0.0961
Local Reg-split	0.9425	5.9330	0.0397	0.0535
HPD-split-Forest	0.9590	6.3142	0.0390	0.0500
Proposed Approach	0.9600	5.9965	0.0310	0.0353
Scenario	Heterocedastic			
Measure	Est. coverage	Avg. size	Avg. abs. dev.	Avg. of abs.neg dev.
HPD-split-FlexCode	0.9718	8.9534	0.0335	0.0325
Dist-split+	0.9511	7.6975	0.0338	0.0417
Quantile-split	0.9539	7.1028	0.0290	0.0312
Reg-split	0.9509	7.1584	0.0399	0.0488
Local Reg-split	0.9449	7.0455	0.0392	0.0503
HPD-split-Forest	0.9514	6.8967	0.0308	0.0334
Proposed Approach	0.9532	6.8266	0.0253	0.0256
Scenario	Homocedastic			
Measure	Est. coverage	Avg. size	Avg. abs. dev.	Avg. of abs.neg dev.
HPD-split-FlexCode	0.9788	5.5318	0.0328	0.0209
Dist-split+	0.9488	4.5060	0.0281	0.0339
Quantile-split	0.9500	4.1402	0.0173	0.0196
Reg-split	0.9373	3.8255	0.0136	0.0157
Local Reg-split	0.9372	4.0304	0.0319	0.0423
HPD-split-Forest	0.9445	4.0377	0.0166	0.0215
Proposed Approach	0.9478	4.0278	0.0146	0.0183
Scenario	Asymmetric			
Measure	Est. coverage	Avg. size	Avg. abs. dev.	Avg. of abs.neg dev.
HPD-split-FlexCode	0.9686	3.6771	0.0287	0.0245
Dist-split+	0.9534	5.6518	0.0417	0.0553
Quantile-split	0.9574	6.6890	0.0695	0.2531
Reg-split	0.9343	3.2614	0.0501	0.0953
Local Reg-split	0.9276	3.2126	0.0527	0.0774
HPD-split-Forest	0.9453	3.3726	0.0422	0.0663
Proposed Approach	0.9486	3.1386	0.0342	0.0487

Table 4 – Average region's size - Avg. size, the average of absolute deviations of estimated conditional coverage to 0.95 Avg. abs. dev. and the average only for negative deviations Avg. of abs. neg. dev. for $n = 2,500$ and scenarios in Izbicki, Shimizu and Stern (2022).

CHAPTER

5

CONCLUSION AND FUTURE WORKS

5.1 Conclusions

An overview about conformal prediction was presented, showing the main ideas, components and some examples of methods. The definition, the concept of score, conditional coverage and the HPD Split are examples of topics covered.

In addition, a new approach was proposed, allying good properties of two previous methods: Izbicki, Shimizu and Stern (2022) and Gupta, Kuchibhotla and Ramdas (2022), with respect to the flexibility of the score, region and the efficient use of data and evaluation. The flexibility of the score and region is related with the aim of obtaining the optimal region in a convenient shape. The efficiency arises from estimation of densities with more data, compared to approaches that split part of the data only to evaluate the score, using some already estimated trees to save computing time.

The goal of estimating densities in a leave-one-out context and using trees with MSE loss and to get a member of conformal prediction class with competitive performance was reached, associating the HPD score with nested sets and theoretical results in Gupta, Kuchibhotla and Ramdas (2022).

Finally, simulation studies confirm the advantages of the new method. Based on previous scenarios in the literature, comparisons showed very competitive measures, with the proposed method being one of the best approaches in all of them.

5.2 Future works

There is potential in investigating the theoretical proof of marginal coverage with a possible better inequality in some conditions and in improving and seeking other methods to estimate the conditional densities with trees.

BIBLIOGRAPHY

BARBER, R. F.; CANDES, E. J.; RAMDAS, A.; TIBSHIRANI, R. J. Predictive inference with the jackknife+. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 49, n. 1, p. 486–507, 2021. Citation on page 21.

GUPTA, C.; KUCHIBHOTLA, A. K.; RAMDAS, A. Nested conformal prediction and quantile out-of-bag ensemble methods. **Pattern Recognition**, v. 127, p. 108496, 2022. ISSN 0031-3203. Available: <https://www.sciencedirect.com/science/article/pii/S0031320321006725>. Citations on pages 21, 22, 25, 28, 33, and 41.

IZBICKI, R.; SHIMIZU, G.; STERN, R. Flexible distribution-free conditional predictive bands using density estimators. In: PMLR. **International Conference on Artificial Intelligence and Statistics**. [S.l.], 2020. p. 3068–3077. Citation on page 36.

IZBICKI, R.; SHIMIZU, G.; STERN, R. B. Cd-split and hpd-split: efficient conformal regions in high dimensions. **Journal of Machine Learning Research**, v. 23, p. 1–32, 2022. Citations on pages 17, 22, 25, 27, 35, 38, 39, and 41.

KIM, B.; XU, C.; BARBER, R. F. Predictive inference is free with the jackknife+-after-bootstrap. **arXiv preprint arXiv:2002.09025**, 2020. Citation on page 28.

LEI, J.; G'SELL, M.; RINALDO, A.; TIBSHIRANI, R. J.; WASSERMAN, L. Distribution-free predictive inference for regression. **Journal of the American Statistical Association**, Taylor & Francis, v. 113, n. 523, p. 1094–1111, 2018. Citations on pages 21, 23, 24, and 35.

LEI, J.; WASSERMAN, L. Distribution-free prediction bands for non-parametric regression. **Journal of the Royal Statistical Society: Series B: Statistical Methodology**, JSTOR, p. 71–96, 2014. Citation on page 26.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN, W. *et al.* Applied linear statistical models. Irwin Chicago, 1996. Citation on page 21.

POSPISIL, T.; LEE, A. B. (f) rfcde: Random forests for conditional density estimation and functional data. **arXiv preprint arXiv:1906.07177**, 2019. Citations on pages 32 and 33.

ROMANO, Y.; PATTERSON, E.; CANDES, E. Conformalized quantile regression. **Advances in Neural Information Processing Systems**, v. 32, p. 3543–3553, 2019. Citations on pages 25 and 36.

SESIA, M.; CANDèS, E. J. A comparison of some conformal quantile regression methods. **Stat**, v. 9, n. 1, p. e261, 2020. E261 sta4.261. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.261>. Citation on page 36.

VOVK, V. Conditional validity of inductive conformal predictors. In: PMLR. **Asian conference on machine learning**. [S.l.], 2012. p. 475–490. Citation on page 26.

VOVK, V.; GAMMERMAN, A.; SHAFER, G. **Algorithmic learning in a random world**. [S.l.]: Springer Science & Business Media, 2005. Citations on pages 11, 13, 21, and 23.


ICMC
USP
SÃO CARLOS