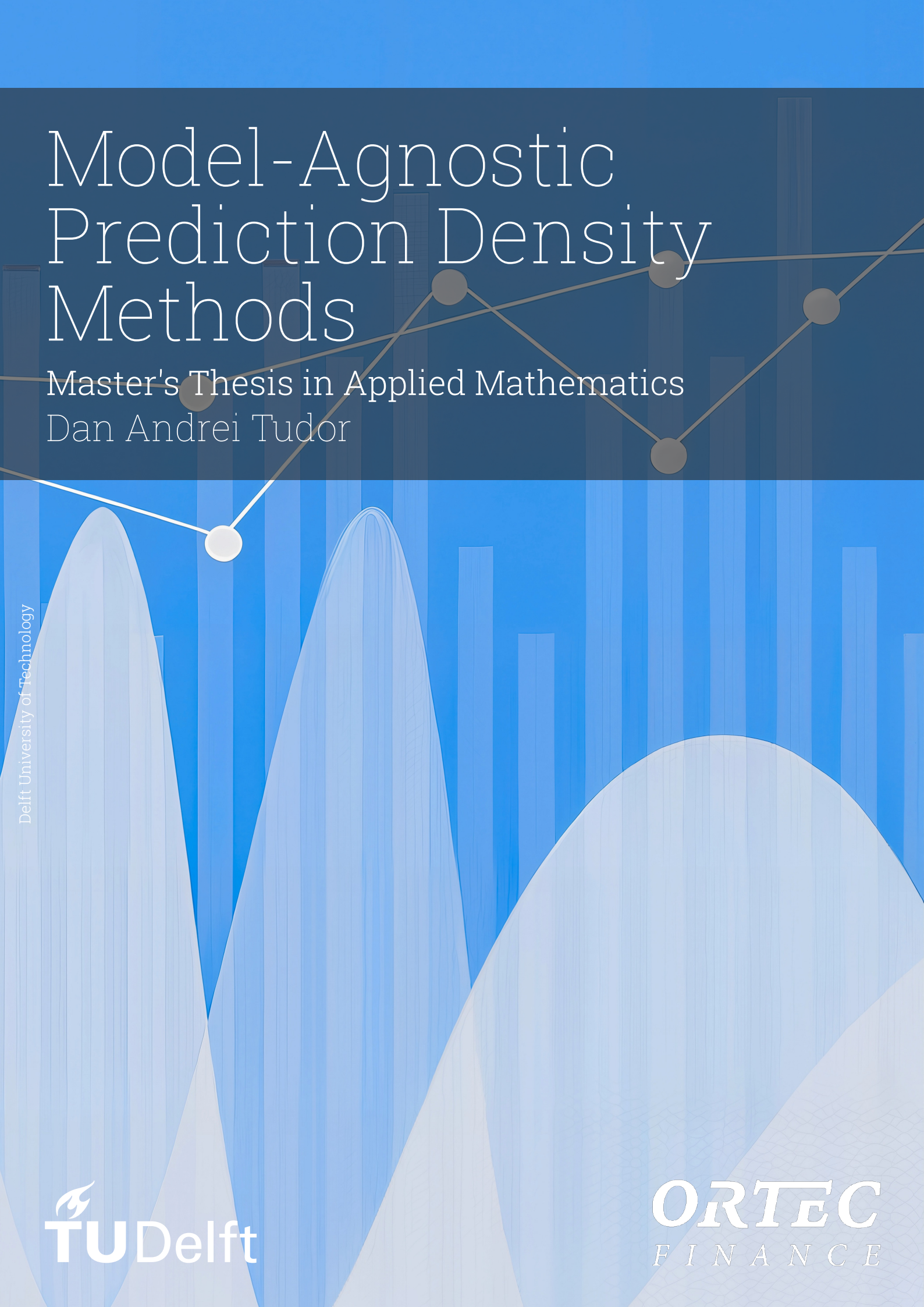
Model-Agnostic Prediction Density Methods

Master's Thesis in Applied Mathematics

Dan Andrei Tudor

Delft University of Technology



Model-Agnostic Prediction Density Methods

by

Dan Andrei Tudor

to obtain the degree of Master of Science in Applied Mathematics
at Delft University of Technology,
to be defended publicly on Thursday June 4, 2026 at 1:00 PM.

Student number:	6172040	
Project duration:	September 1, 2025 – May 30, 2026	
Thesis committee:	dr. Dorota Kurowicka,	TU Delft, supervisor
	dr. Balint Negyesi,	Ortec Finance, supervisor
	Valerii Zoller MSc,	Ortec Finance, supervisor
	dr. Jakob Soehl,	TU Delft

Cover: Adobe Stock photos under CC BY-NC 2.0 (Modified)
Style: TU Delft Report Style, with modifications by Daan Zwaneveld

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

ORTEC FINANCE

Preface

It has been a satisfying journey writing my Master's thesis at Ortec Finance over the past nine months. Over weekly meetings with my two caring company supervisors, dr. Balint Negyesi and Valerii Zoller, I have explored countless directions for my thesis project, and our discussions have turned out to be very productive. Thank you for your guidance, availability and for the freedom you have offered me in giving the project its current form.

Thank you to my great university supervisor, dr. Dorota Kurowicka, for the structured guidance throughout the project, continuous involvement, and valuable feedback, which significantly improved the flow and structure of this thesis.

I am also thankful to dr. Jakob Soehl for graciously agreeing to serve on my thesis committee and for taking the time to review this work.

To my family and friends, wherever you may be, thank you for being understanding and supporting throughout my studies abroad.

Over the past five years in the Netherlands, obtaining my Bachelor's degree in Mathematics at Vrije Universiteit Amsterdam and pursuing a Master's in Applied Mathematics at Delft University of Technology, I kept coming back to the realization that exploring and discovering new ideas gives me great fulfillment. Starting next fall, I will be pursuing a PhD in Statistics at the University of Wisconsin-Madison, where I look forward to joining a world-class research environment.

Dan Andrei Tudor
Delft, May 2026

Abstract

The present work focuses on constructing predictive densities, conditional on a set of features, for one-dimensional real-valued random variables. We approach the problem in a model-agnostic manner, aiming for methods that can be applied to arbitrary models without imposing parametric assumptions on the underlying distribution.

We first present the standard kernel density estimation approach and discuss its limitations. In this context, the conformal framework, originally developed for prediction intervals, is particularly appealing due to its finite-sample marginal validity guarantees. We then introduce conformal predictive distributions, a recent development in the literature that ensures the associated predictive system is marginally $\mathrm{Unif}[0, 1]$ under the Probability Integral Transform (PIT).

However, these distributions tend to be highly fuzzy. As a result, directly applying finite differencing to conformal predictive distributions produces noisy predictive densities that may obscure important underlying features.

To address this issue, we propose two solutions. First, Gaussian filtering yields the smoothest densities and empirically maintains PIT perturbations within a satisfactory range, although no theoretical bounds on the perturbation are derived. Second, we introduce a new method, termed *quantile-matching*, which produces less fuzzy densities while providing a sharp theoretical upper bound on PIT perturbation.

Furthermore, we show that when the number of quantiles is allowed to equal the size of the calibration set, the distribution induced by quantile-matching coincides with the crisp modification of conformal predictive distributions, thereby yielding an upper bound on their PIT perturbation as well.

Finally, we evaluate the proposed methods on a large simulated real estate transactions dataset based on the Hierarchical Trend Model. Our results indicate that the quantile-matching approach outperforms competing methods across several metrics, including the Mean Absolute Error of the associated tail means and running time per transaction price.

Contents

1

Introduction

Evaluating risk is a fundamental topic in domains as diverse as finance, healthcare, engineering, and environmental policy. In all of these fields, decisions are often made with incomplete information, and one must assess and manage risk to prevent losses or operational disruptions and make better decisions in a competitive environment. In this sense, for practitioners, *predictions* should serve as a fundamental tool for assessing risk and describing uncertainty. Point predictions, which provide a single estimate of a future value, although common in the industry, do not convey information about uncertainty; recent research has shifted towards prediction intervals [38] and predictive distributions [23] to address this.

What is a prediction?

To assess future risk and uncertainty, one must use available information in the present to make informed "guesses" about the future. Unlike estimation, which focuses on inferring population parameters, *prediction* aims to forecast future observations and their characteristics. More formally, let (X, Y) be a dataset, where each X represents a vector of *features* and Y represents the random variable of interest, i.e., the one for which the prediction should be applied, which we shall call the *target variable*. In the present thesis, we focus on a real-valued *one-dimensional* target variable, whereas the features may be multi-dimensional. Suppose we also have a *statistical model* relating X and Y, as defined below [25, Definition 1.2].

Definition 1.1. A statistical model is a family of probability measures $\{\mathbb{P}_\theta : \theta \in \Theta\}$ on some measurable space. The set Θ is called the parameter space.

The model can be written as $Y \sim \mathbb{P}_\theta,\ \theta \in \Theta$, or alternatively $f_Y(y \mid \theta),\ \theta \in \Theta$, where f_Y is the probability density function of $\mathbb{P}_\theta$. A statistical model can be parametric (θ is finite-dimensional) or nonparametric (θ is infinite-dimensional).

In prediction problems, the target of study is not a parameter, but the unobserved value of a random variable. This can include, however, estimating the value of a random parameter attached to a particular portion of the data. A *point prediction* refers to a single estimate of a future outcome derived from observed data and a specified *statistical model*, which we formalize in the following definition.

Definition 1.2. Let Y be the target variable, $X = (X_1, X_2, \ldots, X_p)$ a vector of features, and $f(Y \mid X, \theta)$ the conditional distribution of Y given X and model parameters θ. A *point prediction* of Y given $X = x$ is an estimator $\hat{Y}(x)$ chosen to minimize the conditional expected value of some loss function $L(Y, \hat{Y}(x))$ in the following sense

$$\hat{Y}(x) = \arg\min_{\hat{y}} \mathbb{E}[L(Y, \hat{y}) \mid X = x].$$

Example 1.3. The most common loss function is the squared loss $L_2(Y, \hat{Y}(x)) := (Y - \hat{Y}(x))^2$, which gives rise to the optimal point prediction $\hat{Y}(x) = \mathbb{E}(Y \mid X = x)$. Alternatively, if the loss function is taken to be the absolute loss $L_1(Y, \hat{Y}(x)) := |Y - \hat{Y}(x)|$, then the optimal point prediction is given by $\hat{Y}(x) = \text{median}(Y \mid X = x)$.

Example 1.4. In the context of quantile regression, which we shall introduce in Chapter 3, the pinball loss function is defined as [34]

$$\rho_\alpha(Y, \hat{Y}(x)) := \begin{cases} \alpha(Y - \hat{Y}(x)) & \text{if } Y - \hat{Y}(x) > 0 \\ (1-\alpha)(Y - \hat{Y}(x)) & \text{otherwise.} \end{cases}$$

Under $\rho_\alpha(Y, \hat{Y}(x))$, the optimal point prediction is the α-quantile of the target variable's distribution.

Let us now give a simple example of a point prediction estimate where the optimal point prediction is found exactly.

Example 1.5. Suppose we have the simple model

$$Y \mid X = x \sim N(2x + 3, \sigma^2),$$

where σ^2 is known. Then, the optimal point prediction under the squared loss is given by

$$\hat{Y}(x) = \mathbb{E}(Y|X = x) = 2x + 3.$$

The above example, however, is artificial. In practice, the true conditional distribution is unknown, and we must rely on approximations based on estimating parameters from the data. The following example illustrates this idea in the linear regression setting.

Example 1.6. Suppose we have n observations and p features expressed as

$$(X_{i1}, \ldots, X_{ip}, Y_i), \quad \text{for } i = 1, \ldots, n.$$

Based on this, we formulate the multiple linear regression model as follows, for $i, j = 1, \ldots, n$ with $i \neq j$,

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \\ \mathbb{E}(\epsilon_i) &= 0, \\ \mathbb{E}(\epsilon_i \epsilon_j) &= 0, \\ \mathbb{E}(\epsilon_i^2) &= \sigma^2, \end{aligned}$$

where $\beta_0, \ldots, \beta_p$ and σ^2 are unknown parameters to be estimated and ϵ_i are error terms. Alternatively, we can use the matrix-vector notation for simplicity

$$\begin{aligned} Y &= X\beta + \epsilon, \\ \mathbb{E}(\epsilon) &= 0 \\ \mathsf{Cov}(\epsilon) &= \sigma^2 I_n, \end{aligned}$$

where $Y = (Y_1, \ldots, Y_n)^T$, X is the $n \times (p+1)$ matrix with i-th row $X_i = (1, X_{i1}, \ldots, X_{ip})$, $\beta = (\beta_0, \ldots, \beta_p)^T$ is the vector of unknown coefficients, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is the stochastic vector of errors. Least squares estimation of parameters gives [7] the *unbiased* estimators

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - p - 1}, \end{aligned}$$

where $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T$ is the vector of *predicted* responses computed as $\hat{Y} = X\hat{\beta}$. Then, upon observing a new features vector $X_{n+1} = x$, we obtain the optimal point prediction estimate in terms of squared loss

$$\hat{Y}(x) = \mathbb{E}(Y_{n+1}|X_{n+1} = x) = x^T \mathbb{E}(\hat{\beta}) = x^T \beta.$$

Therefore, the optimal point prediction estimate depends on the unknown β. This is a problem in practice, since β is unknown. One then has to settle on the estimate $\tilde{Y}(x) = x^T \hat{\beta}$ which retains good asymptotic properties, namely that $\tilde{Y}(x) \to \hat{Y}(x)$ as $n \to \infty$ [7].

From point predictions to prediction intervals and beyond

As we have seen above, optimal point prediction estimates (in the squared loss sense) might not be computed explicitly when the parameters of the model are not known, and we instead have to rely on estimates that retain "good" asymptotic properties. Additionally, point predictions fail to retain information about prediction uncertainty. A point prediction gives a single estimate, but it does not capture how confident we are in the prediction we have made. In practice, noise, estimation error, and even model misspecification can cause the true outcome to vary widely from the single-point prediction, with important negative consequences. It becomes apparent, then, that uncertainty needs to be quantified, particularly for high-risk applications such as medicine and finance. For example, a point prediction for an expected loss of, say, one million euros, gives no insight into the probability that the loss is significantly higher, say, ten million. These problems, ubiquitous in applied fields, need a better metric for *predicting* that captures uncertainty.

Additionally, in the above examples, we have seen predictions made under specific model assumptions. For instance, in the linear regression setting, we assume *unbiasedness* of error terms, which does not necessarily happen in practice. If the underlying model is misspecified, predictions might be misleading. It is desirable, then, that we develop methods we shall call *model-agnostic*. These methods do not assume a fixed form of the underlying model, allowing for great flexibility in the relationship between features and target variables. Model-agnostic prediction methods focus instead on the prediction accuracy itself and are especially useful in high-risk applications when the true underlying models are unknown.

It is then natural to extend point prediction estimates to prediction intervals, as real-world prediction always involves a degree of uncertainty. Whereas a single value hides that uncertainty, intervals express it explicitly, depending on how wide the interval is. Where the risk of loss matters, a range of outcomes at a given confidence level is preferable. However, even if prediction intervals are superior in this regard to point prediction estimates, uncertainty can still be hidden. For instance, a prediction interval at a given confidence level does not indicate whether certain values within the interval are more likely than others. In high-risk settings such as finance, decisions do depend on the shape of the uncertainty, and with a full predictive distribution, extreme-event probabilities and tail risks can be assessed.

Recent research has increasingly seen interest in predictive distributions [23, 41]. In certain contexts, however, extending predictive distributions to predictive densities is desirable, and is a current goal of the present thesis. Firstly, predictive densities provide a qualitative description of the underlying model that is not readily obtainable from the distribution, such as multimodality, skewness, and tail heaviness. Such hidden features are easily retrievable from a density, and many practitioners are interested in them. Secondly, computing non-linear (conditional) expectations of functionals of the target variable is most easily done using a density. Perhaps of most interest to industry experts, at least within Ortec Finance, is a model feature update based on the predictive density. Let us explain this idea in more detail in the context of real estate applications, the primary area of implementation within the company.

Suppose we wish to obtain the predictive distribution of the price of a house with characteristics $x := (x_1, \ldots, x_p)$. Assume that one characteristic j is uncertain or completely missing. This situation arises when features correspond to outdated attributes, such as energy label or condition. Denote then the uncertain feature by $\theta := x_j$ and let the remaining fully observed characteristics be

$$x_{-j} := (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p),$$

so that, up to permutation, $x \cong (\theta, x_{-j})$. Given a predictive model that produces a house price prediction $\hat{y}(\theta, x_{-j})$, define the function $h(\theta) := \hat{y}(\theta, x_{-j})$, which represents the prediction as a function of the uncertain feature alone.

Let $p(\theta \mid x_{-j})$ denote some prior distribution encoding our beliefs about the missing characteristic given the observed features. With access to a predictive density $q(y \mid \theta, x_{-j})$, describing the uncertainty of the predicted outcome, conditional on the features, one can then proceed to do posterior inference on the missing feature. Upon observing a realized house price y, Bayes' rule yields the posterior distribution of the uncertain feature:

$$p(\theta \mid x_{-j}, y) \propto q(y \mid \theta, x_{-j})\, p(\theta \mid x_{-j}).$$

Hence, the predictive density acts as a likelihood function that allows the prediction model to be inverted, transforming outcome information into information about missing inputs. The use of a predictive *density* is essential, since it assigns a likelihood to each possible observed outcome y. This formulation shows that predictive densities provide more than uncertainty quantification: they enable probabilistic inference over latent or missing covariates by serving as likelihood functions within a Bayesian updating framework. Although this is not a focus of our current thesis, it highlights the relevance of predictive densities and will most definitely be a topic of further research within Ortec Finance.

Conformal prediction

At the turn of the millennium, a new framework for prediction intervals emerged that has become a ”gold standard” in many applications, including machine learning: *conformal prediction* [38]. The main advantage of the framework, compared with standard predictions, is that it provides a *finite sample marginal coverage guarantee*. By introducing an additional recalibration step of the model, in which one adjusts their prediction based on how non-conforming the observations are compared to the prediction made, this framework can, on average, provide well-calibrated prediction intervals, something that other standard prediction methods are unable to do [38, 27]. The conformal outline also extends to predictive distributions, where it remains marginally well-calibrated in finite samples. Additionally, conformal predictions are entirely model-agnostic and achieve desirable results even under model misspecification.

Outline of present thesis

We aim to present a unified and coherent framework for properly quantifying uncertainty: first through conformal prediction intervals, then through conformal predictive distributions. Our overarching goal in the present work is to extend conformal predictive distributions [39, 40, 41, 23] to be able to retrieve *predictive densities*, that is, conditional probability density functions of the target variable of interest, based on the newly observed feature, which are sufficiently accurate and broadly general. We will stay within the conformal framework to retain as many of the marginal guarantees inherent to these methods as possible.

We shall first provide a brief overview of *Kernel Density Estimation* in Chapter 2, a classical nonparametric method of constructing an unconditional probability density function. This method is based on the assumption that the observations $Y_1, \ldots, Y_n$ are independent and identically distributed. Hence, it does not account for any of the model's features.

We shall then turn our attention towards prediction intervals with features in the model. Prediction intervals give a range within which the future outcome is expected to fall at a given level. They communicate uncertainty in simple terms and allow decision makers to interpret results easily. Model-agnostic prediction intervals have uses, for example, in medicine, where they have been applied to estimate changes in lipid and lipoprotein levels [20]. Chapter 3 gives an overview of current conformal prediction interval methods and expands upon the present literature by presenting a proof of the fact that the Split Conformal Direct Prediction Method [12] is well-calibrated. The conformal prediction framework has been mainly used to construct prediction intervals [38, 27, 12], but it has also been extended to construct *predictive distributions* [41, 40, 39]. This framework is notable in two principal ways: it is model-agnostic, that is, it does not rely on any parametric assumptions on the underlying model, and it retains a finite-sample marginal coverage guarantee.

The extension from conformal prediction intervals to conformal predictive distributions [41, 39] is given in Chapter 4. We adapt the finite-sample marginal coverage guarantee of conformal prediction intervals to conformal predictive distributions in this chapter, and extend this concept to allow for *asymptotic* conformal predictive distributions. This extension is needed to produce true cumulative distribution functions and to act as a bridge to retrieving *predictive densities*. We contribute to the existing literature by providing several results that quantify the distance from the true uniform distribution of the probability integral transform of asymptotic conformal predictive distributions.

The extension to predictive densities is given in Chapter 5, where we discuss four approaches, including a novel method we call *quantile-matching*. Our method essentially functions as a compromise between a direct finite-differencing approach, which can be fuzzy and can obscure underlying characteristics of

the predictive density, and a Gaussian filtering approach, which creates the smoothest densities but damages the marginal calibration inherent in conformal predictive distributions. Our method successfully manages to decrease the fuzziness of the raw finite-differenced conformal predictive densities, while remaining within a reasonable bound from the theoretical $\mathrm{Unif}[0, 1]$ distribution in Probability Integral Transform. Furthermore, we draw a valuable connection between the quantile-matching induced distribution and the crisp modification of conformal predictive distributions. The quantile-matching method we propose allows users to choose their quantile resolution, which can be beneficial when practitioners are more interested in, say, more extreme quantiles. We believe there is room for further research on our methodology, as it is a promising addition to the literature due to its flexibility and ease of implementation, especially with standard conformity scores.

Applications to a large simulated dataset of real estate prices are presented in Chapter 6, where the performance of the different approaches is evaluated across multiple metrics. We find that running the quantile-matching algorithm is more efficient than running the full conformal prediction distribution algorithm, particularly in large calibration sizes. It is particularly well-suited for computing tail means, while maintaining similar performance across various metrics, including the mean integrated squared error and the continuous ranked probability score.

2

Kernel Density Estimation

The most common choice for estimating density functions is *kernel density estimators* (KDEs). They were introduced independently by Parzen [26] and Rosenblatt [28]. The present chapter is dedicated to presenting the idea behind Kernel Density Estimators and understanding their uses and limitations.

Let us now focus on a set of independent and identically distributed random variables $Y_1, \ldots, Y_n$ with *continuous* density function $f(y)$ (except for finitely many points). In this very simple setting, building a prediction density for a new observation Y_{n+1} is the same as estimating the density function $f(y)$ at all points y in a sufficiently large interval for our application. However, even in this simplified setting, the problem of constructing predictive densities is nontrivial. For starters, Rosenblatt [28] observes that there is no unbiased estimator for all continuous density functions $f(y)$ for all $y \in [a, b]$, irrespective of the interval $[a, b]$. Therefore, one needs to consider alternative metrics for the performance of our prediction densities. Here, we provide a proof of this statement, filling in some gaps in the original paper [28].

Theorem 2.1. *There is no unbiased estimator $T(\mathbf{Y}, \mathbf{y})$ of $f(y)$ for all continuous density functions f on $\mathbb{R}$.*

Proof. The proof hinges on an argument by contradiction and the fact that the unordered set $\{Y_1, \ldots, Y_n\}$ is a sufficient and complete statistic for the set of all continuous density functions (except for finitely many points).

To see that this is the case, observe that the likelihood function can be rewritten as

$$L(y_1, \ldots, y_n \mid f) = \prod_{i=1}^{n} f(y_i) = \prod_{y \in \{y_1, \ldots, y_n\}} f(y).$$

Define

$$g(\{y_1, \ldots, y_n\}, f) := \prod_{y \in \{y_1, \ldots, y_n\}} f(y), \quad h(y_1, \ldots, y_n) := 1.$$

Then, $L \cong g \cdot h$, so by the Fisher-Neyman factorization theorem (see, for instance, [1]), indeed $\{Y_1, \ldots, Y_n\}$ is a sufficient statistic for the problem.

Furthermore, $\{Y_1, \ldots, Y_n\}$ is also a complete statistic for our problem. This step fills a gap in Rosenblatt's [28] proof, which relies on citing lecture notes from 1950. These lecture notes appear to be lost, so we reason here independently to complete this step. To proceed, let $g(Y_1, \ldots, Y_n)$ be any symmetric function of $Y_1, \ldots, Y_n$ mapping to the real numbers. We show directly that the definition of completeness holds: if $\mathbb{E}_f g(\{Y_1, \ldots, Y_n\}) = 0$ for all continuous densities f, then $g(\{Y_1, \ldots, Y_n\}) = 0$ f-a.s. Suppose that indeed $\mathbb{E}_f g(Y_1, \ldots, Y_n) = 0$ for all continuous densities f. To proceed, let us fix a point $(y_1, \ldots, y_n) \in \mathbb{R}^n$. Consider, for each y_i, the sequence of density functions $(f_{m,y_i})_{m=1}^{\infty}$ defined by

$$f_{m,y_i}(y) = \frac{m}{2} \mathbb{1}_{[y_i - \frac{1}{m}, y_i + \frac{1}{m}]}(y),$$

each corresponding to a $\mathrm{Unif}[y_i - \frac{1}{m}, y_i + \frac{1}{m}]$ random variable. Next, mix these densities to form the sequence

$$f_{m,(y_1,\ldots,y_n)}(y) := \sum_{i=1}^{n} \frac{1}{n} f_{m,y_i}(y) = \frac{m}{2n} \mathbb{1}_{[y_1-\frac{1}{m},y_1+\frac{1}{m}]\cup\cdots\cup[y_n-\frac{1}{m},y_n+\frac{1}{m}]}(y).$$

For this particular sequence, we have, by assumption,

$$\begin{aligned}
\mathbb{E}_{f_{m,(y_1,\ldots,y_n)}} g(\{Y_1,\ldots,Y_n\}) &= \int_{\mathbb{R}^n} g(\{x_1,\ldots,x_n\}) \prod_{i=1}^{n} f_{m,(y_1,\ldots,y_n)}(x_i)\, dx_1 \cdots dx_n \\
&= \left(\frac{m}{2n}\right)^n \int_{\mathbb{R}^n} g(\{x_1,\ldots,x_n\}) \mathbb{1}_{\left(\bigcup_{i=1}^{n}[y_i-\frac{1}{m},y_i+\frac{1}{m}]\right)^n}(x_1,\ldots,x_n) d(\{x_1,\ldots,x_n\}) \\
&= 0.
\end{aligned}$$

Using the Lebesgue differentiation theorem [22], taking the limit in the above, we obtain that, for almost all $(y_1,\ldots,y_n) \in \mathbb{R}^n$, $g(\{y_1,\ldots,y_n\}) = 0$. This is independent of the underlying density f, as the integration is done with respect to the Lebesgue measure, and, consequently, $\mathbb{P}_f(g(Y_1,\ldots,Y_n) = 0) = 1$ for all f. Thus, $\{Y_1,\ldots,Y_n\}$ is a complete statistic for the problem.

Now, suppose, for the sake of contradiction, that there exists an estimator $T(\mathbf{Y}, y)$ such that $\mathbb{E}\, T(\mathbf{Y}, y) = f(y)$ for all continuous f and all y. We may assume, moreover, that the estimator is symmetric in each component $y_1,\ldots,y_n$, since the unordered set $\{Y_1,\ldots,Y_n\}$ is a sufficient statistic for f. Now, observe that if $T(\mathbf{Y}, y)$ is a symmetric estimator of $f(y)$, then also $\int_a^b T(\mathbf{Y}, y)dy$ is a symmetric estimator of $F(b) - F(a)$. Moreover, the estimator is unbiased, since, by Fubini's theorem,

$$\mathbb{E} \int_a^b T(\mathbf{Y}, y)dy = \int_a^b \mathbb{E}\, T(\mathbf{Y}, y)dy = \int_a^b f(y)dy = F(b) - F(a).$$

However, $\hat{F}_n(b) - \hat{F}_n(a)$, where $\hat{F}_n$ is the empirical cumulative distribution function, defined as in Definition 3.3, is also an unbiased symmetric estimator of $F(b) - F(a)$. This is easy to see, since in the case of i.i.d random variables $\mathbb{E}\, \hat{F}_n(b) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(Z_i \le b) = F(b)$.Therefore, we obtain that

$$\mathbb{E}_f \left(\int_a^b T(\mathbf{Y}, y)dy - (\hat{F}_n(b) - \hat{F}_n(a)) \right) = 0 \quad \text{for all continuous } f,$$

implying that $\int_a^b T(\mathbf{Y}, y)dy = \hat{F}_n(b) - \hat{F}_n(a)$ for all a and b and almost all $Y_1,\ldots,Y_n$, by completeness of $\{Y_1,\ldots,Y_n\}$. This, in turn, implies that $\hat{F}_n(y)$ is continuous in y for almost all $Y_1,\ldots,Y_n$, which leads to a contradiction.

We conclude that there exists no unbiased estimator for the class of continuous density functions. □

The above theorem highlights the inherent difficulty in tracking the performance of predictive densities, even in simple settings such as independent and identically distributed observations of the target variables. In pursuit of quantifying this performance, Rosenblatt [28] and Parzen [26] settle for the good asymptotic properties of Kernel Density Estimators.

We now proceed to describe the technique of Kernel Density Estimation. Firstly, let us note that to estimate the probability density function $f(y)$ given n i.i.d. observations $Y_1,\ldots,Y_n$, a natural choice starts from the unbiased estimator of the cumulative distribution function, namely, the empirical cumulative distribution function, and applying a central difference scheme of the type

$$\hat{f}_n(y) = \frac{\hat{F}_n(y+h) - \hat{F}_n(y-h)}{2h}, \tag{2.1}$$

where h is chosen suitably (more details on how to choose h are discussed later). Then, what lies at the basis of the technique is a simple observation Parzen made in [26], namely that (2.1) can be rewritten

as the weighted average

$$\hat{f}_n(y) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{y-x}{h}\right) d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{y-Y_i}{h}\right), \tag{2.2}$$

where

$$K(y) = \begin{cases} \frac{1}{2}, & |y| \leq 1 \\ 0, & |y| > 1. \end{cases} \tag{2.3}$$

Rewriting (2.1) in the form of (2.2) opens up the study of Kernel Density Estimation, by replacing the representation of $K(y)$ in (2.3) with other functions. In the original works, Rosenblatt [28] and Parzen [26] investigate suitable choices for h and $K(y)$ so that the representation in (2.2) is asymptotically unbiased and consistent at every point of continuity y. Furthermore, by varying the size of h with the sample size n, one would also desire that $\lim_{n\to\infty} h_n = 0$, as to be consistent with the original central difference scheme framework. Under this condition, we look for choices of $K(y)$ such that

$$\lim_{n\to\infty} \mathbb{E}\hat{f}_n(y) = f(y). \tag{2.4}$$

The sufficient conditions on K are given below. [32]

$$K(y) \geq 0 \quad \text{for all } y \in \mathbb{R} \qquad (\text{non-negativity}), \tag{2.5}$$

$$K(y) = K(-y) \quad \text{for all } y \in \mathbb{R} \qquad (\text{symmetry}), \tag{2.6}$$

$$\sup_{y\in\mathbb{R}} K(y) < \infty \qquad (\text{uniform boundedness}), \tag{2.7}$$

$$\lim_{y\to\infty} yK(y) = 0 \qquad (\text{regular tail behavior}), \tag{2.8}$$

$$\int_{-\infty}^{\infty} K(y)dy = 1 \qquad (\text{normalization}), \tag{2.9}$$

$$\int_{-\infty}^{\infty} y^2 K(y)dy < \infty \qquad (\text{finite second moment}). \tag{2.10}$$

From now on, we shall need these conditions for K. It is convenient to give such functions a suitable name.

Definition 2.2 (Kernel)**.** A function K satisfying (2.5)-(2.10) is called a kernel.

Observe that

$$\mathbb{E}\hat{f}_n(y) = \mathbb{E}\left(\frac{1}{h_n} K\left(\frac{y-Y_1}{h_n}\right)\right) = \int_{-\infty}^{\infty} \frac{1}{h_n} K\left(\frac{y-x}{h_n}\right) f(x)dx. \tag{2.11}$$

The following theorem, which follows from a broader framework in [5], shows that a *kernel* implies asymptotic unbiasedness of estimators of the type described in (2.2).

Theorem 2.3. *Let K be a Borel function satisfying* (2.5)*,* (2.7) *and* (2.8) *and*

$$\int_{-\infty}^{\infty} K(y)dy < \infty,$$

and h_n a sequence of positive constants satisfying (2.12)*. Let f satisfy*

$$\int_{-\infty}^{\infty} |f(y)|dy < \infty.$$

Define the sequence

$$f_n(y) = \frac{1}{h_n} \int_{-\infty}^{+\infty} K\left(\frac{y-x}{h_n}\right) f(x)dx.$$

Then, at every point y of continuity of f, we have

$$\lim_{n\to\infty} f_n(y) = f(y) \int_{-\infty}^{\infty} K(x)dx$$

Proof. First, observe that, after a change of variables in f_n,

$$f_n(y) - f(y)\int_{-\infty}^{\infty} K(x)dx = \int_{-\infty}^{\infty}(f(y-x)-f(y))\frac{1}{h_n}K\left(\frac{x}{h_n}\right)dx.$$

Now fix $\delta > 0$ and split the region of integration into two parts: $|x| \leq \delta$ and $|x| > \delta$. Then, using the triangle inequality,

$$\begin{aligned}
&\left|f_n(y) - f(y)\int_{-\infty}^{\infty} K(x)dx\right| \leq \max_{|x|\leq\delta}|f(y-x)-f(y)|\int_{|z|\leq\frac{\delta}{h_n}} K(z)dz \\
&+ \int_{|x|>\delta}(|f(y-x)|+|f(y)|)\frac{1}{h_n}K\left(\frac{x}{h_n}\right)dx \\
&\leq \max_{|x|\leq\delta}|f(y-x)-f(y)|\int_{|z|\leq\frac{\delta}{h_n}} K(z)dz + \int_{|x|>\delta}\frac{|f(y-x)|}{x}\frac{x}{h_n}K\left(\frac{x}{h_n}\right)dx \\
&+ |f(y)|\int_{|x|>\delta}\frac{1}{h_n}K\left(\frac{x}{h_n}\right)dx \\
&\leq \max_{|x|\leq\delta}|f(y-x)-f(y)|\int_{-\infty}^{\infty} K(z)dz + \frac{1}{\delta}\sup_{|z|>\frac{\delta}{h_n}}|zK(z)|\int_{-\infty}^{\infty}|f(x)|dx + |f(y)|\int_{|z|>\frac{\delta}{h_n}} K(z)dz.
\end{aligned}$$

When one lets n tend to ∞, the quantity $\frac{\delta}{h_n}$ goes to ∞, therefore, by the regular tail behavior property of (2.8), the second term vanishes, and by the hypothesis that $\int_{-\infty}^{\infty} K(y)dy < \infty$, we also get that $\int_{|z|>\frac{\delta}{h_n}} K(z)dz$ goes to 0, and so the third term vanishes. Upon letting $\delta \to 0$, by continuity of f at y, the first term vanishes as well. Therefore, we conclude that

$$\lim_{n\to\infty} f_n(y) = f(y)\int_{-\infty}^{\infty} K(x)dx.$$

□

Corollary 2.3.1. *Let $K(y)$ be a kernel and h_n a sequence of positive constants satisfying* (2.12)*. Define $\hat{f}_n(y)$ as in* (2.2)*. Then*

$$\lim_{n\to\infty}\mathbb{E}\hat{f}_n(y) = f(y).$$

Proof. In Theorem 2.3, take $f_n(y) = \mathbb{E}\hat{f}_n(y)$. Using (2.11) and the fact that the kernel is normalized by (2.9), the conclusion follows. □

Another desirable property of the class of kernel density estimators is consistency. In this way, one can ensure convergence in probability to the true density function. To obtain this, we need one more condition on the sequence h_n, namely that $\lim_{n\to\infty} nh_n = +\infty$.

It is convenient to name such a sequence, and the common name in the literature is a sequence of *bandwidths* [32].

Definition 2.4 (Bandwidth)**.** A sequence h_n satisfying

$$\lim_{n\to\infty} h_n = 0 \tag{2.12}$$

and

$$\lim_{n\to\infty} nh_n = +\infty. \tag{2.13}$$

is called a sequence of bandwidths.

We now show that indeed kernels, together with a sequence of bandwidths, make the estimators in (2.2) consistent [26].

Theorem 2.5. *Let K be a kernel and h_n a sequence of bandwidths. Then the estimators in* (2.2) *are consistent.*

Proof. First, note that

$$\mathsf{Var}(\hat{f}_n(y)) = \frac{1}{n}\mathsf{Var}\left(\frac{1}{h_n}K\left(\frac{y-Y_1}{h_n}\right)\right).$$

Additionally, we have, provided that K^2 satisfy the properties in Theorem 2.3, that, as $n \to \infty$,

$$h_n\mathbb{E}\left(\frac{1}{h_n}K\left(\frac{y-Y_1}{h_n}\right)\right)^2 = \frac{1}{h_n}\int_{-\infty}^{\infty} K^2\left(\frac{y-x}{h_n}\right) f(x)dx \to f(y)\int_{-\infty}^{\infty} K^2(x)dx,$$

in view of Theorem 2.3. Indeed, K^2 satisfies non-negativity (2.5) and uniform boundedness (2.7) trivially, and (2.8) is inherited from the regular tail behavior of K combined with its uniform boundedness. Note also that $\int_{-\infty}^{+\infty} K^2(y)dy \leq \sup_{y\in\mathbb{R}} K(y)\int_{-\infty}^{\infty} K(y)dy < \infty$, so that all conditions of Theorem 2.3 are satisfied. Additionally, we get from Corollary 2.3.1 and (2.11) that $\left[\mathbb{E}\left(\frac{1}{h_n}K\left(\frac{y-Y_1}{h_n}\right)\right)\right]^2 \to f^2(y)$ and consequently, as $n \to \infty$,

$$h_n\left[\mathbb{E}\left(\frac{1}{h_n}K\left(\frac{y-Y_1}{h_n}\right)\right)\right]^2 \to 0 \quad \text{as } n \to \infty.$$

Therefore,

$$\lim_{n\to\infty} nh_n\mathsf{Var}(\hat{f}_n(y)) = f(y)\int_{-\infty}^{\infty} K^2(x)dx,$$

and consequently, under condition (2.13), we get $\lim_{n\to\infty}\mathsf{Var}(\hat{f}_n(y)) = 0$. Therefore, we have, in terms of the mean squared error

$$\mathsf{MSE}(\hat{f}_n(y), f(y)) := \mathbb{E}(\hat{f}_n(y) - f(y))^2 = \mathsf{Var}(\hat{f}_n(y)) + (\mathbb{E}(\hat{f}_n(y)) - f(y))^2 \to 0 \quad \text{as } n \to \infty,$$

in view of Corollary 2.3.1. Thus, $\hat{f}_n(y) \to f(y)$ in L^2, and so also in probability, making the estimator consistent. □

We have seen above that, under certain assumptions, Kernel Density Estimators have *locally* good asymptotic properties. Namely, they are asymptotically unbiased and consistent at every point of continuity of the underlying density function. However, the scope of density estimation is to have a good prediction density for *every* point y in the domain. It becomes evident that we need better criteria for quantifying the error made by density estimators that take into account their global behavior. Such a criterion, which is widely used in the literature, is the Mean Integrated Squared Error (MISE) [28, 32].

Definition 2.6 (MISE)**.** The Mean Integrated Squared Error of a density estimator $\hat{f}$ of f is defined as

$$\mathrm{MISE}(\hat{f}, f) := \int \mathbb{E}(\hat{f}(x) - f(x))^2 dx = \int \mathsf{Var}(\hat{f}(x))dx + \int (\mathbb{E}\hat{f}(x) - f(x))^2 dx.$$

The study of Kernel Density Estimators now translates into appropriate choices of the kernel and the bandwidths that minimize the mean integrated squared error.

The optimal choice of bandwidths

For this section, we follow the exposition in [32]. For convenience, we can also drop the subscripts h_n and $\hat{f}_n(x)$ when it is clear that we are dealing with a sample size of n and that the bandwidth depends on the sample size. Let us first focus on the bias term in the MISE formula. We can write

$$\begin{aligned}\mathbb{E}\hat{f}(x) - f(x) &= \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f(y)dy - f(x)\\ &= \int K(t)\left[f(x-ht) - f(x)\right]dt,\end{aligned}$$

after a change of variables $t = \frac{x-y}{h}$ and using the normalization property (2.9) of kernels. Assume that the true density function is twice continuously differentiable, so that we can use a Taylor series expansion around x,

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + O(h^3). \tag{2.14}$$

Then, in the bias formula, we get

$$\mathbb{E}\hat{f}(x) - f(x) = -hf'(x)\int tK(t)dt + \frac{1}{2}h^2f''(x)\int t^2K(t)dt + O(h^3) = \frac{1}{2}h^2f''(x)\int t^2K(t)dt + O(h^3), \tag{2.15}$$

since $\int tK(t)dt = 0$ by the symmetry property (2.6) of kernels. Let

$$k_2(K) := \int t^2K(t)dt.$$

Then the integrated bias-squared term can be approximated as

$$\int (\mathbb{E}\hat{f}(x) - f(x))^2dx \approx \frac{1}{4}h^4k_2(K)^2\int f''(x)^2dx \tag{2.16}$$

We now turn to the integrated variance term. First, note that

$$\mathbb{E}\hat{f}(x) = \int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy \tag{2.17}$$

and

$$\mathsf{Var}(\hat{f}(x)) = \frac{1}{n}\int \frac{1}{h^2}K^2\left(\frac{x-y}{h}\right)f(y)dy - \frac{1}{n}\left[\int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy\right]^2, \tag{2.18}$$

which, under the same t change of variable and the Taylor expansion (2.14) approximates as follows, when n is large and h is small,

$$\begin{aligned}
\mathsf{Var}(\hat{f}(x)) &= \frac{1}{n}\int \frac{1}{h}K^2(t)\,f(x-ht)dt - \frac{1}{n}[f(x) + \mathbb{E}\hat{f}(x) - f(x)]^2 \\
&= \frac{1}{n}\int \frac{1}{h}K^2(t)\,f(x-ht)dt - \frac{1}{n}[f(x) + O(h^2)]^2 \quad \text{(using (2.15))} \\
&= \frac{1}{nh}\int (f(x) - htf'(x) + O(h^2))K^2(t)dt + O(1/n) \\
&= \frac{1}{nh}f(x)\int K^2(t)dt + O(1/n) \\
&\approx \frac{1}{nh}f(x)\int K^2(t)dt.
\end{aligned}$$

Integrating over x and using the fact that f is a density function, we obtain the integrated variance approximation

$$\int \mathsf{Var}(\hat{f}(x))dx \approx \frac{1}{nh}\int K^2(t)dt. \tag{2.19}$$

Combining (2.16) and (2.19) yields the following approximation for MISE.

$$\mathrm{MISE}(\hat{f}, f) \approx \frac{1}{nh}\int K^2(t)dt + \frac{1}{4}h^4k_2(K)^2\int f''(x)^2dx. \tag{2.20}$$

Observe the bias-variance tradeoff in the above formula. The first term grows like $\frac{1}{h}$, whereas the second term grows like h^4. In other words, the bias can be minimized by choosing a small bandwidth, but this would lead to a large variance, and vice versa. One has to instead balance both terms simultaneously. We may now optimize for the bandwidth that minimizes the approximate value of the MISE. This lemma was originally stated without proof by Parzen in [26].

Lemma 2.7. *The optimal value that minimises the* MISE *approximation in* (2.20) *is given by*

$$h_{opt} = n^{-1/5} k_2(K)^{-2/5} \left(\int K^2(t)dt \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5}. \tag{2.21}$$

Proof. Let $A = \frac{1}{n} \int K^2(t)dt$ and $B = \frac{1}{4} k_2(K)^2 \int f''(x)^2 dx$ Then the optimization problem can be cast as

$$\min_{h>0} Ah^{-1} + Bh^4.$$

Taking the first derivative with respect to h and setting it to zero yields $-Ah^{-2} + 4Bh^3 = 0$, which gives the optimal value for h as $h_{opt} = A^{1/5}(4B)^{-1/5}$. As the second derivative is given by $\frac{d^2}{dh^2}(Ah^{-1} + Bh^4) = 2Ah^{-3} + 12Bh^2 > 0$ for all $h > 0$, $h_{opt} = A^{1/5}(4B)^{-1/5}$ is a global minimum. Retrieving back A and B gives exactly the optimum in (2.21). □

Note that the formula for the optimal bandwidth comes with an intrinsic issue: it depends on the unknown density function itself. However, important conclusions can be drawn from this formula [32]. First, the optimal bandwidth indeed converges to zero as the sample size increases and also satisfies condition (2.13) that $\lim_{n\to\infty} nh_n = \infty$. However, the rate of convergence to zero is rather slow at a rate of $n^{-1/5}$. Additionally, since h_{opt} depends on the term $\int f''(x)dx$, which measures the speed of fluctuations in the underlying density, this indicates that smaller values of h are more appropriate for highly fluctuating densities.

With the optimal bandwidth value now known, we now move to the problem of choosing an appropriate kernel.

The optimal choice of kernels

We can now substitute the value of h_{opt} found in (2.21) back into (2.20) to obtain the approximated MISE

$$\widehat{\mathrm{MISE}}(\hat{f}, f) = \frac{5}{4} n^{-4/5} k_2(K)^{2/5} \left(\int K^2(t)dt \right)^{4/5} \left(\int f''(x)^2 dx \right)^{1/5}. \tag{2.22}$$

All other things being equal, one should choose the kernel that minimizes the quantity

$$C(K) := k_2^{2/5}(K) \left(\int K^2(t)dt \right)^{4/5}. \tag{2.23}$$

Additionally, observe that if $k_2(K) \neq 1$, one can substitute the kernel K by $\tilde{K} = k_2^{1/2}(K) K(k_2(K)^{1/2} t)$. Then $k_2(\tilde{K}) = 1$ and $C(\tilde{K}) = C(K)$. In other words, we can assume without loss of generality that $k_2(K) = 1$. Then the problem can be reduced to the following optimization problem

$$\begin{aligned} \min & \int K^2(t)dt \\ \text{subject to} & \int K(t)dt = 1 \\ & \int t^2 K(t)dt = 1 \\ & K(t) = K(-t) \\ & K(t) \geq 0. \end{aligned} \tag{2.24}$$

The cast optimization problem is now independent of f, the bandwidth h, and the sample size n. Epanechnikov [11] and, in a different context, Hodges and Lehmann [18] give the solution to this optimization problem directly. Here, we provide a detailed proof of this fact, as the original papers give the result as a known fact, without going into details.

Theorem 2.8. *The kernel optimizing* (2.24) *is given by the Epanechnikov kernel [11]*

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right), & -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 \quad \textit{otherwise}. \end{cases} \tag{2.25}$$

Proof. We proceed by using the Lagrange multipliers method with integral (isoperimetric) constraints [24]. Since the problem we are dealing with does not involve a first-derivative condition on K, the Lagrange augmented integral becomes, in the unconstrained sense,

$$\int (K^2(t) + \lambda_1 K(t) + \lambda_2 t^2 K(t))dt$$

and the Euler condition becomes

$$0 = \frac{d}{dK}(K^2(t) + \lambda_1 K(t) + \lambda_2 t^2 K(t)) = 2K(t) + \lambda_1 + \lambda_2 t^2,$$

pointwise in t. We can then solve for λ_1 and λ_2 by plugging in the formula found for $K(t)$ from the Euler equation

$$K(t) = \frac{-\lambda_1 - \lambda_2 t^2}{2} \tag{2.26}$$

back into the constraints $\int K(t)dt = 1$ and $\int t^2 K(t)dt = 1$. To ensure that $\int K(t)dt = 1$ and K is symmetric, we have to truncate the domain of integration on a symmetric interval $[-a, a]$. Therefore, the conditions become

$$\begin{aligned} -\lambda_1 a - \lambda_2 \frac{a^3}{3} &= 1, \\ -\lambda_1 \frac{a^3}{3} - \lambda_2 \frac{a^5}{5} &= 1. \end{aligned}$$

Solving this linear system for λ_1 and λ_2 yields

$$\begin{aligned} \lambda_1 &= \frac{3(5 - 3a^2)}{4a^3}, \\ \lambda_2 &= \frac{15(a^2 - 3)}{4a^5}. \end{aligned}$$

Plugging these values back into (2.26), we get, after simplifying,

$$K(t) = \frac{3}{4a}\left(1 - \frac{t^2}{a^2}\right) \quad \text{on } [-a, a], \tag{2.27}$$

and integrating $K^2(t)$ to retrieve the objective function yields

$$\int_{-a}^{a} K^2(t)dt = \frac{3}{8}\left(\frac{3}{a} - \frac{10}{a^3} + \frac{15}{a^5}\right).$$

The last step is to minimize over a to achieve the global minimum. Taking the first derivative and setting it to zero yields the equation $-3a^4 + 30a^2 - 75 = 0$, or, alternatively, $-3(a^2 - 5)^2 = 0$. Therefore, the unique nonnegative solution is $a = \sqrt{5}$. Plugging this back into (2.27) yields exactly the Epanechnikov kernel in (2.25). □

Comparison of different kernels

It is now illustrative to compare different kernels to the optimal Epanechnikov kernel in terms of their efficiency [32], which we define below.

Definition 2.9 (Kernel efficiency)**.** The efficiency of a kernel K is defined by

$$\text{eff}(K) = \left[\frac{C(K)}{C(K_e)}\right]^{5/4}, \tag{2.28}$$

where $C(K)$ is defined as in (2.23).

We scale by the power $5/4$ above since, as Kendall explains in [32, Section 3.3.2], for large n, the MISE will be the same whether we use n observations and the kernel K or whether we use $n\,\text{eff}(K)$ observations and the kernel K_e. Table 2.1 gives the efficiency of different kernels.

Table 2.1: Comparison of Common Kernel Functions and Their Efficiencies

Kernel	$K(t)$	**Efficiency**
Epanechnikov	$K(t) = \frac{3}{4\sqrt{5}}\left(1 - \frac{t^2}{5}\right) \quad \lvert t\rvert \le \sqrt{5},$ $0 \quad \text{otherwise}$	1
Biweight	$K(t) = \frac{15}{16}(1-t^2)^2 \quad \lvert t\rvert \le 1,$ $0 \quad \text{otherwise}$	0.9939
Triangular	$K(t) = 1 - \lvert t\rvert \quad \lvert t\rvert \le 1$ $0 \quad \text{otherwise}$	0.9859
Gaussian	$K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$	0.9512
Rectangular	$K(t) = \frac{1}{2} \quad \lvert t\rvert \le 1$ $0 \quad \text{otherwise}$	0.9295

Table 2.1 shows that even though the Epanechnikov kernel is indeed theoretically optimal, the differences are quite small compared to other kernels. It is desirable then to choose kernels based on smoothness, and in many practical applications, the normal kernel is preferred.

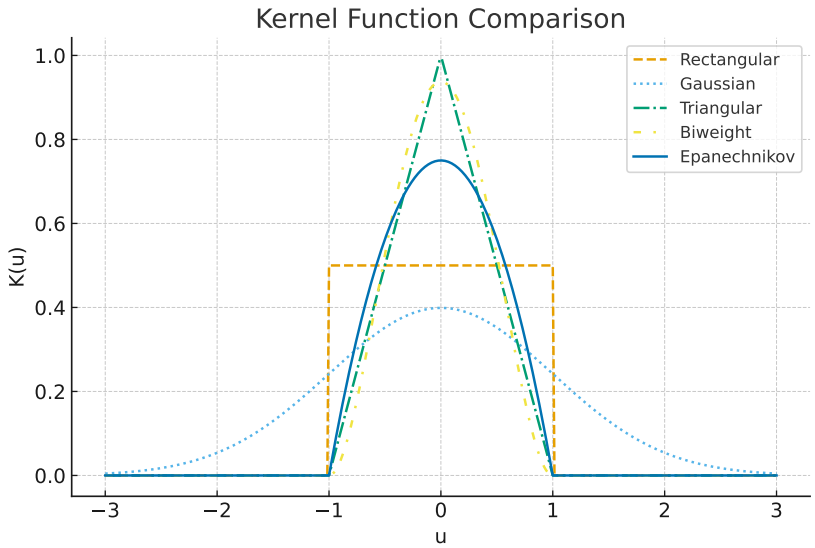


Figure 2.1: Comparison of the five kernel functions

Estimating the optimal bandwidth

Since, as we've seen before, the optimal bandwidth (2.21) depends on the unknown density, one must rely on estimating it instead. Put into context, the problem of choosing an appropriate bandwidth is crucial in Kernel Density Estimation, as it influences the end result significantly more than the choice of the kernel. Here, we discuss two different approaches for estimating the bandwidth: a *subjective* choice and an *automated* choice. For simple purposes of data exploration, a subjective choice is often sufficient. However, in high-risk applications, the automated choice is preferable, as it provides a standardized method for selecting the bandwidth. We once again follow the exposition in [32].

The subjective choice

As a subjective choice, we can start with the parametric assumption that f is normally distributed with mean μ and variance σ^2. Then

$$\int f''(x)^2 dx = \frac{3}{8}\pi^{-1/2}\sigma^{-5} \approx 0.212\sigma^{-5}.$$

Using a Gaussian kernel and estimating the variance σ from the data gives the estimated optimal bandwidth in (2.21) as

$$\hat{h}_{opt} = (4\pi)^{-1/10}\left(\frac{3}{8}\pi^{-1/2}\right)^{-1/5}\hat{\sigma}n^{-1/5} \approx 1.06\hat{\sigma}n^{-1/5}. \tag{2.29}$$

A fast approach to estimate the bandwidth under this paradigm is to approximate the standard deviation by the sample standard deviation and plug it into (2.29). This will work fine in the case of unimodal distributions that are "somewhat" normal, but it will oversmooth the distribution if the distribution is multimodal, as $\int f''(x)^2 dx$ will be much larger relative to the normal distribution. To adjust for this issue, the following heuristic change for $\hat{\sigma}$ in (2.29) is suggested

$$A = \min(\text{sample standard deviation}, \text{interquartile range}/1.34),$$

combined with reducing the factor of 1.06 in (2.29) to 0.9 leading to

$$\hat{h}_{Silverman} = 0.9An^{-1/5}. \tag{2.30}$$

This will do well with unimodal densities and will not do too badly with bimodal densities. In most simulations, Silverman [32] notes that this estimated bandwidth performs well for a wide variety of parametric underlying distributions. From now on, we shall call the choice of bandwidth as in (2.30) *Silverman's rule of thumb*. Unfortunately, there is no sound mathematical reason to picking the constant 1.34 to divide by in the standard deviation estimation, but rather it is seen through trial-and-error done by Silverman in [32] that for a wide range of underlying distributions, the MISE of the estimation using (2.30) as a bandwidth is within 10% of the optimum.

The automated choice: likelihood cross-validation

The regime of cross-validation requires us to introduce another goodness-of-fit criterion, namely the Integrated Squared Error (ISE).

Definition 2.10 (ISE)**.** For a given density estimator $\hat{f}$, the Integrated Squared Error is defined by

$$\mathrm{ISE}(\hat{f}, f) := \int (\hat{f}(t) - f(t))^2 dt = \int \hat{f}(t)^2 dt - 2\int \hat{f}(t)f(t)dt + \int f^2(t)dt. \tag{2.31}$$

Remark 1. Note that, by Fubini's theorem, $\mathbb{E}\left[\mathrm{ISE}(\hat{f}, f)\right] = \mathrm{MISE}(\hat{f}, f)$.

Observe that the density estimator that minimizes the integrated squared error also minimizes

$$R(\hat{f}) := \int \hat{f}(t)^2 dt - 2\int \hat{f}(t)f(t)dt, \tag{2.32}$$

as the last term in the ISE formula depends only on the underlying density f. To be able to solve the minimization problem without knowing f, we need to replace $R(\hat{f})$ by an estimator $\hat{R}(\hat{f})$. Herein lies the *cross-validation* principle.

If a random variable X has the same density f as $Y_1, \ldots, Y_n$ and is also independent of them, then

$$\mathbb{E}(\hat{f}(X) \,|\, Y_1, \ldots, Y_n) = \int \hat{f}(t) f(t) dt,$$

which is, up to scaling, exactly the second term in $R(\hat{f})$. Therefore, if we had additional observations $X_1, \ldots, X_m$, we could estimate $\int \hat{f}(t) f(t) dt$ by $\frac{1}{m} \sum_{i=1}^{m} \hat{f}(X_i)$, as an estimator of the associated conditional expected value above. Unfortunately, such additional observations are not available, and we rely on the cross-validation principle. Let us define $\hat{f}_{-i}$ to be the density estimator obtained from all observations $Y_1, \ldots, Y_n$ except for the i-th. We can then replace X_i with Y_i and $\hat{f}$ with $\hat{f}_{-i}$ and apply the conditional expectation principle above with these estimated densities instead. We now work with $n-1$ observations instead of n, leading to the estimated value for $R(f)$ as

$$\hat{R}(\hat{f}) = \int \hat{f}(t)^2 dt - \frac{2}{n} \sum_{i=1}^{n-1} \hat{f}_{-i}(Y_i). \tag{2.33}$$

Minimizing $\hat{R}(\hat{f})$ is now possible as it depends entirely on the estimated densities. Choosing an optimal bandwidth that minimizes $\hat{R}(\hat{f})$ is possible once we choose a specific kernel. Additionally, the use of n instead of $n-1$ in the denominator of (2.33) stems from a heuristic argument in the literature [32] that the difference is small for large n.

Finally, we point out that cross-validation is a computationally expensive method, as it requires recomputing the density estimator n times and then solving an optimization problem. While for small sample sizes this might be feasible ($n < 10000$), in the context of large samples, an appropriate approach is to select a small subsample to compute the cross-validated bandwidth on, and then apply it to the original dataset.

Examples

We conclude this chapter by showcasing some examples of kernel density estimation in various contexts. We first showcase the performance of KDEs for the purposes for which they were built: independent and identically distributed random variables. We then visualize what happens when we deviate from the independence assumption.

Let us first compare the performance of four different kernels for a random sample of 5000 i.i.d $N(0,1)$ random variables. Since the underlying distribution is exactly normal, Silverman's rule of thumb is an appropriate choice of bandwidth, thereby avoiding the costly cross-validated bandwidth computations. Either way, the comparison of the four kernels is shown in Figures 2.2 and 2.3 for both bandwidth choices. The associated MISE values are given in Table 2.2.

Kernel	Silverman's Rule	Cross-Validation (CV)
Gaussian	0.0020	0.0017
Epanechnikov	0.0055	0.0017
Biweight	0.0071	0.0017
Uniform	0.0044	0.0017

Table 2.2: MISE values of KDE for the $N(0,1)$ sample of size 5000 with four different kernels and different bandwidths.

We observe that the Gaussian Kernel performs best, perhaps unsurprisingly, irrespective of the chosen bandwidth, since the underlying true distribution is Gaussian. A more interesting comparison arises when the distribution is bimodal, which violates the underlying assumptions of Silverman's rule of thumb. In this case, we once more compare the performances of the four kernels, first with a fixed bandwidth chosen using Silverman's rule of thumb, and second with a cross-validated bandwidth for each kernel.

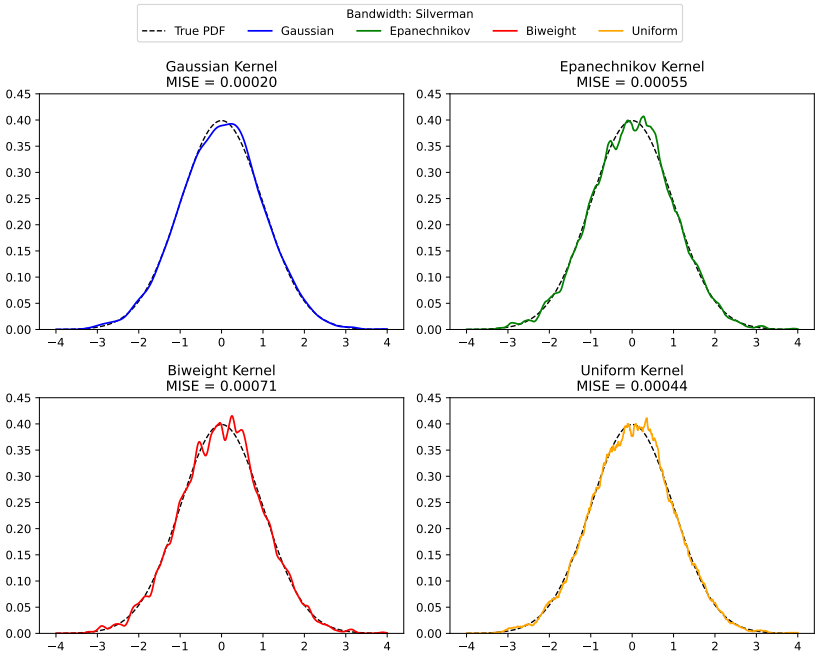


Figure 2.2: A comparison of KDEs for a $N(0,1)$ sample of size 5000 with four different kernels with a fixed bandwidth chosen through Silverman's rule of thumb.

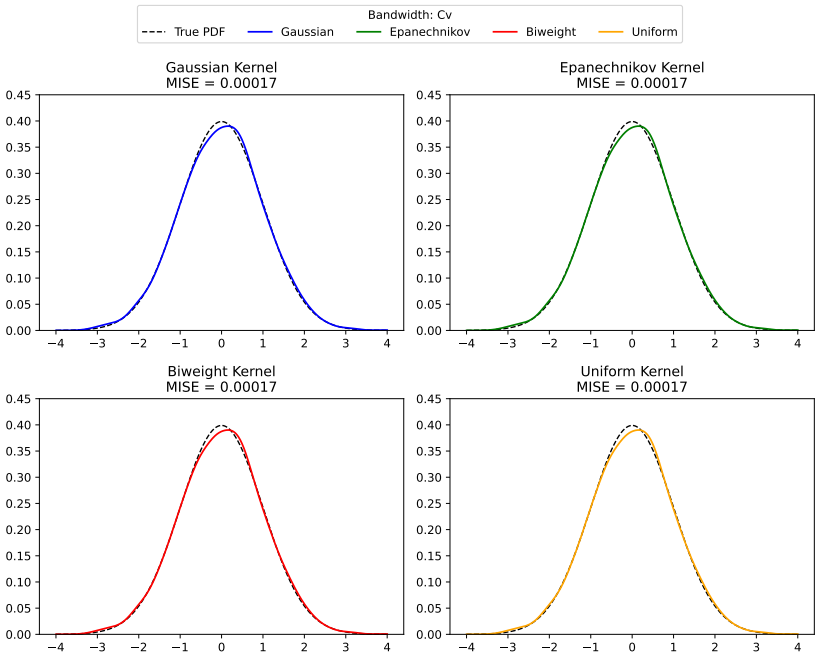


Figure 2.3: A comparison of KDEs for a $N(0,1)$ sample of size 5000 with four different kernels with cross-validated bandwidths.

The underlying bimodal distribution is an equally weighted mixture of $N(-2,0.8)$ and $N(2,0.8)$ random variables, with true probability density function given by

$$f(x) = 0.5\frac{1}{\sqrt{2\pi 0.8^2}}e^{-(x+2)^2/(2\cdot 0.8^2))} + 0.5\frac{1}{\sqrt{2\pi 0.8^2}}e^{-(x-2)^2/(2\cdot 0.8^2))}.$$

Figures 2.4 and 2.5 showcase the performance of KDE in this context.

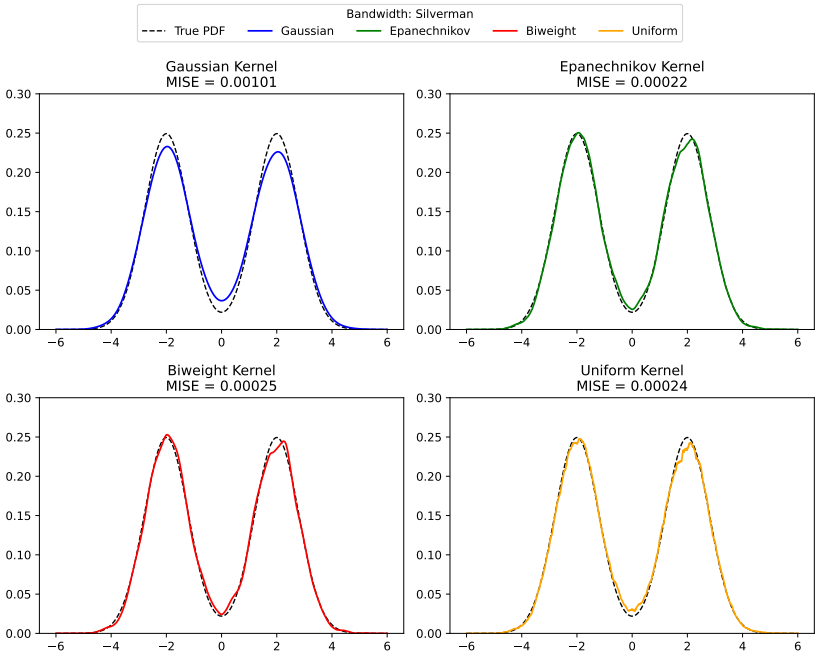


Figure 2.4: A comparison of KDEs for a bimodal sample of size 5000 with four different kernels with a fixed bandwidth chosen through Silverman's rule of thumb.

In line with the theoretical framework, one observes that an appropriate choice of bandwidths makes the question of choosing the "best" kernel largely irrelevant - the four kernels have an identical MISE when the bandwidth is cross-validated. However, when the bandwidth is chosen using Silverman's rule of thumb, Epanechnikov indeed yields the most accurate kernel density approximation to the true underlying density, but the improvement over the biweight or uniform kernels is incremental. The Gaussian oversmooths in combination with Silverman's bandwidth and, in doing so, deviates significantly from the peaks of the bimodal distribution, leading to a MISE that is fourfold larger than that of the optimal Epanechnikov kernel.

Kernel	Silverman's Rule	Cross-Validation (CV)
Gaussian	0.00101	0.00022
Epanechnikov	0.00022	0.00022
Biweight	0.00025	0.00022
Uniform	0.00024	0.00022

Table 2.3: MISE values of KDEs for a bimodal sample of size 5000 with four different kernels and different bandwidths.

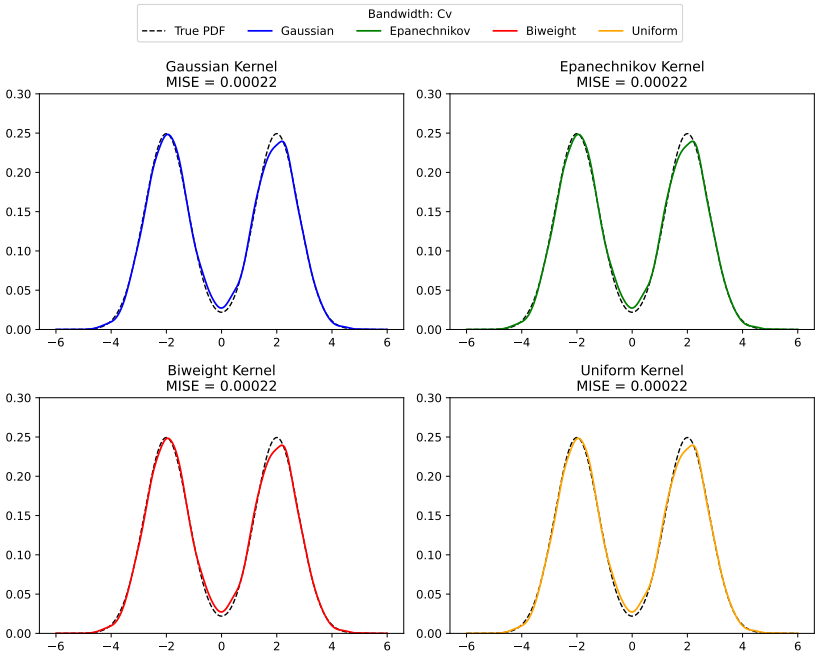


Figure 2.5: A comparison of KDEs for a bimodal sample of size 5000 with four different kernels with cross-validated bandwidths.

Introducing dependencies: the AR(1) model

Let us now observe how Kernel Density Estimation performs when the independence assumption is violated. To showcase this, we will use KDE as a prediction density estimator for an AR(1) model.

The AR(1) model [31] with mean μ is formed by a sequence $\{Y_t\}_{t=1}^n$ defined by the following recursion

$$Y_{t+1} = \mu + \phi(Y_t - \mu) + \epsilon_t, \tag{2.34}$$

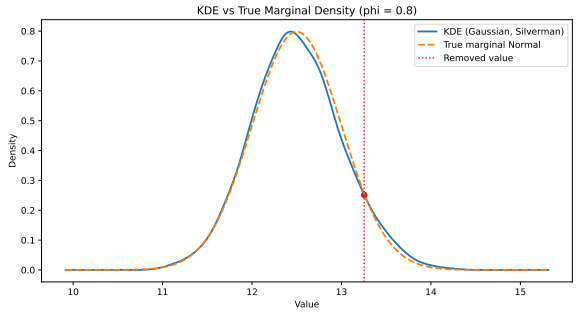
where ϕ is the parameter of the model and $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. Observe that if $Y_1 \sim N(\mu, \sigma^2)$ with $\sigma^2 := \frac{\sigma_\epsilon^2}{1-\phi^2}$, then $Y_t \sim N(\mu, \sigma^2)$ for all t. The proof is by induction. First, the base case is satisfied by assumption, so we can move directly to the induction step and assume that $Y_t \sim N(\mu, \sigma^2)$. The normality follows from the fact that a linear combination of normal random variables is also normal. The mean is given by $\mathbb{E}(Y_{t+1}) = \mu + \phi(\mathbb{E}(Y_t) - \mu) = \mu$, and the variance is given by

$$\mathsf{Var}(Y_{t+1}) = \phi^2 \mathsf{Var}(Y_t) + \sigma_\epsilon^2 = \phi^2 \frac{\sigma_\epsilon^2}{1-\phi^2} + \sigma_\epsilon^2 = \frac{\sigma_\epsilon^2}{1-\phi^2} = \sigma^2.$$

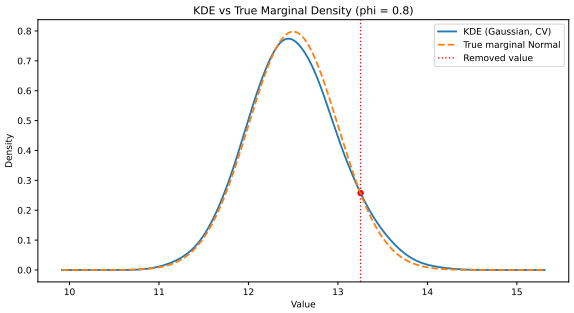
Observe then that the process is weak-sense stationary if $|\phi| < 1$ in the sense that the AR(1) model with $Y_1 \sim N(\mu, \sigma^2)$ retains the same marginal distribution over all Y_t, while introducing sequential dependencies between the random variables. This makes it a perfect candidate for our kernel density estimation showcase and for seeing what happens when dependencies are introduced between random variables.

We simulate three AR(1) models where each random variable has a marginal $N(12.5, 0.5^2)$ distribution with different ϕ's: $\phi = 0.8$, $\phi = 0.99$ and $\phi = 0.999$. For each of these ϕ's, we use a time horizon of 5000, and we aim to naively estimate the marginal distribution at step 5000 using Kernel Density Estimation based on the first 4999 samples, which are treated as independent samples. Due to the way the AR(1) process is obtained, we expect that the larger the ϕ, the worse KDE will perform, as $\phi = 1$ is the critical point of weak-sense stationarity, on the one hand, and the larger the ϕ, the more 'dependent' on the previous value the process is, on the other.

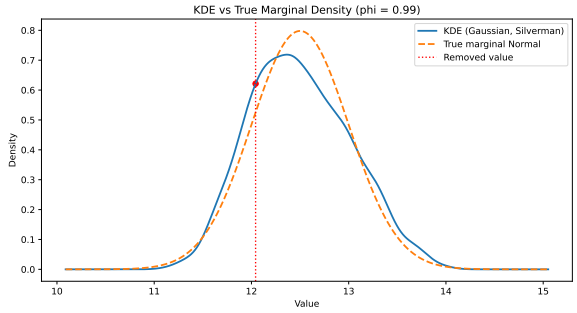
KDE is performed using a Gaussian kernel for smoothness, with both Silverman's bandwidth choice and cross-validated bandwidths (on a subsample of size 900 for computational efficiency).



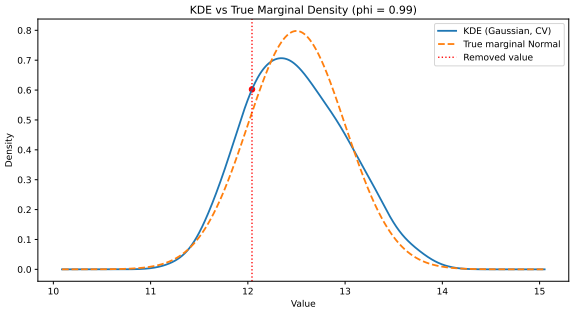
(a) $\phi = 0.8$, Silverman



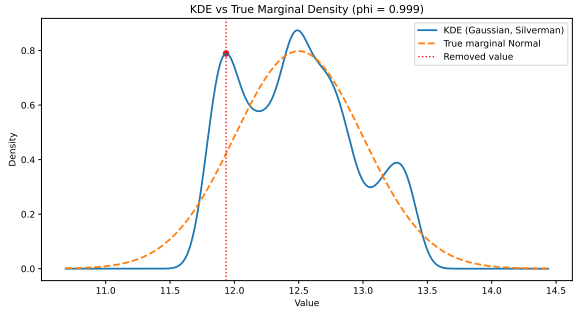
(b) $\phi = 0.8$, Cross-validated



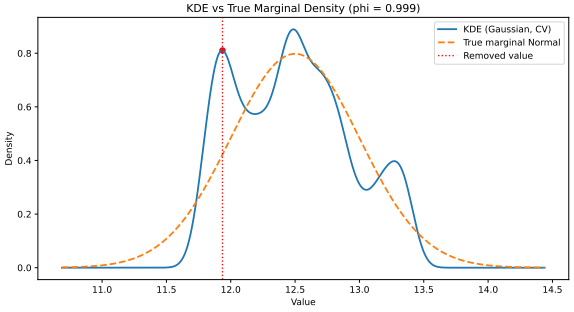
(c) $\phi = 0.99$, Silverman



(d) $\phi = 0.99$, Cross-validated



(e) $\phi = 0.999$, Silverman



(f) $\phi = 0.999$, Cross-validated

Figure 2.6: KDE prediction density estimates based on samples of size 5000 using Silverman's rule of thumb and cross-validated bandwidths for different values of ϕ.

ϕ	**Cross-Validation (CV)**	**Silverman's Rule**
0.8	0.001607	0.001305
0.99	0.011846	0.011545
0.999	0.044694	0.040616

Table 2.4: MISE values for different ϕ values and bandwidth selection methods

From these figures, our intuition is confirmed. We observe that KDE performs relatively well on the AR(1) model when ϕ is relatively small, but starts to break down when ϕ approaches 1. Observe that KDE can no longer retain even the basic shape of the underlying marginal distribution, and, irrespective of the bandwidth chosen, it produces a trimodal distribution at $\phi = 0.999$ and a slightly skewed distribution at $\phi = 0.99$. We note in Table 2.4 a significant almost four-fold increase in the MISE values from $\phi = 0.99$ to $\phi = 0.999$.

We have seen that Kernel Density Estimation is suitable only in cases when we are interested in the marginal distribution of our target variable and the sequence of observations is independent and identically distributed. In addition, the main advantageous properties of this method, such as consistency, hold asymptotically, and no conclusions can be drawn in finite samples. A different framework is required in this case, which retains a finite sample marginal validity guarantee. From now on, we shall focus on the *conformal* framework, first presenting the main ideas for retrieving prediction intervals, then extending them to predictive distributions, and finally to predictive densities.

3

Conformal prediction intervals

The present chapter provides a broad overview of model-agnostic conformal prediction interval methods as a gentle transition to the more involved conformal distribution methods. Additionally, the chapter includes a proof of the theoretical coverage guarantee of the split conformal direct prediction method, a new addition to the literature on prediction intervals.

In this chapter, we work in the following setting. Let us have a sample set $\{(X_i, Y_i)_{i=1}^n\}$, where each X_i is a random vector describing features of the target *real-valued* random variable Y_i. Assume we have p features and define the corresponding observation space as $\mathbb{R}^{p+1} = \mathbb{R}^p \times \mathbb{R}$. An element $z = (x, y)$ with $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$ of the observation space is called an *observation* consisting of features $x \in \mathbb{R}^p$ and target $y \in \mathbb{R}$. We aim, given a sequence of observations $(z_i)_{i=1}^n$ and a new test feature x_{n+1}, to predict an interval in which the target y_{n+1} might fall. Furthermore, our prediction has to be marginally accurate, that is, we aim to construct a prediction interval $PI(X_{n+1})$ at *miscoverage rate* α. THis means that the probability of the random Y_{n+1} being in the prediction interval is at least $1 - \alpha$, taken *marginally* over the randomness of the test feature X_{n+1} and *all* train and calibration points,

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1})) \geq 1 - \alpha. \tag{3.1}$$

If the method satisfies the above *marginal* coverage guarantee, it is also sometimes said to satisfy the $(1 - \alpha)$-*MC* guarantee [2].

Unless mentioned otherwise, we will ultimately focus on retaining a *marginal* coverage guarantee as in (3.1). All throughout, probabilities are taken with respect to all samples $\{(X_i, Y_i)\}_{i=1}^n$ and test point (X_{n+1}, Y_{n+1}).

We ask that the method is model-agnostic, in the sense that it works irrespective of the joint distribution P_{XY} of each observation and the sample size n. The main framework for prediction intervals that satisfies the $(1 - \alpha)$-*MC* guarantee is the *conformal* one, introduced in [38].

The four main methodologies we present in the current chapter for constructing prediction intervals are: *(Split) Conformal Prediction* [38], *Quantile Regression* [21], *Conformalized Quantile Regression* [27], and *(Split) Conformal Direct Prediction Method* [12]. Of the four, only the conformal ones satisfy the $(1 - \alpha)$-MC guarantee.

Since all methods involve the use of *quantiles*, the following subsection is dedicated to summarizing the key definitions and lemmas about quantiles, which will be of use in this chapter, particularly proving marginal coverage guarantees.

A short review of (empirical) quantiles

To start, let us introduce some notation concerning quantiles. We then present two useful lemmas for a subsequent proof in this chapter, variants of which can be consulted in [38]. First, we will define quantiles and empirical quantiles.

Definition 3.1 (True quantile)**.** Denote the cumulative distribution function of a real-valued random variable Z as $F(z) := \mathbb{P}(Z \leq z)$, and the associated *true quantile function* at level $\alpha \in (0,1)$ as $Q_\alpha := \inf\{z \in \mathbb{R} : F(z) \geq \alpha\}$.

Similarly, we can define the right-quantile function.

Definition 3.2 (Right-quantile)**.** The *right-quantile function* is defined as $RQ_\alpha := \sup\{z \in \mathbb{R} : F^-(z) \leq \alpha\}$, where $F^-(z) = \mathbb{P}(Z < z)$.

When n observations are being made, the empirical cumulative distribution function and its associated empirical quantiles can be defined as follows.

Definition 3.3 (Empirical quantiles)**.** In the case of identically distributed random variables $Z_1, \ldots, Z_n$, we additionally define the *empirical cumulative distribution function* as $\hat{F}_n(z) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Z_i \leq z\}}$. We then define the *empirical quantile function* $\tilde{Q}_\alpha^n$ as the true quantile function with respect to the empirical CDF.

Definition 3.4 (Right empirical quantiles)**.** The right empirical quantile $\widetilde{RQ}_\alpha^n$ is the true right quantile function with respect to $\hat{F}_n^-(z) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Z_i < z}$.

The empirical and right-empirical quantile functions can be written down explicitly [27] with respect to the order statistics $Z_{(1)}, \ldots, Z_{(n)}$ as

$$\tilde{Q}_\alpha^n = Z_{(\lceil n\alpha \rceil)}, \quad \widetilde{RQ}_\alpha^n = Z_{(\lfloor n\alpha \rfloor + 1)}. \tag{3.2}$$

We now present two lemmas concerning exchangeable random variables and their respective quantiles [38]. These two lemmas provide a bridge between the properties of quantiles and the theoretical marginal coverage guarantee (3.1) we seek in the methods we present. Most notably, they lie at the foundation of the proof at the end of this chapter, our personal contribution to this chapter.

Lemma 3.5. *Let $Z_1, \ldots, Z_n$ be exchangeable random variables. Then*

$$\mathbb{P}(Z_n \leq \tilde{Q}_\alpha^n) \geq \alpha.$$

If, in addition, $Z_1, \ldots, Z_n$ are almost surely distinct, then

$$\mathbb{P}(Z_n \leq \tilde{Q}_\alpha^n) \leq \alpha + \frac{1}{n}.$$

Proof. By exchangeability of $Z_1, \ldots, Z_n$ and the fact that the empirical quantile is a function of the order statistics (so the initial order of Z_i's does not matter), we have $\mathbb{P}(Z_n \leq \tilde{Q}_\alpha^n) = \mathbb{P}(Z_i \leq \tilde{Q}_\alpha^n)$. Additionally, observe that

$$\mathbb{E}\hat{F}_n(\tilde{Q}_\alpha^n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Z_i \leq \tilde{Q}_\alpha^n\}}\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{P}(Z_i \leq \tilde{Q}_\alpha^n) = \mathbb{P}(Z_n \leq \tilde{Q}_\alpha^n). \tag{3.3}$$

Moreover, by definition of quantiles, it holds that $\hat{F}_n(\tilde{Q}_\alpha^n) \geq a$. Taking expectation on both sides, we obtain that $\mathbb{P}(Z_n \leq \tilde{Q}_\alpha^n) \geq \alpha$, which is the first result of the lemma.

In addition, if $Z_1, \ldots, Z_n$ are almost surely distinct, we observe that, for every $z \in \mathbb{R}$, we have

$$\hat{F}_n(z) - \hat{F}_n^-(z) = \frac{1}{n}\left(\sum_{i=1}^n \mathbb{1}_{Z_i \leq z} - \sum_{i=1}^n \mathbb{1}_{Z_i < z}\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Z_i = z} \leq \frac{1}{n},$$

since at most one of the indicators can take the value one. Observe additionally that $Z_{(\lceil n\alpha \rceil)} = \tilde{Q}_\alpha^n \leq \widetilde{RQ}_\alpha^n = Z_{(\lfloor n\alpha \rfloor + 1)}$. Therefore, we have that $\hat{F}_n(\tilde{Q}_\alpha^n) \leq \hat{F}_n(\widetilde{RQ}_\alpha^n) \leq \hat{F}_n^-(\widetilde{RQ}_\alpha^n) + \frac{1}{n} \leq \alpha + \frac{1}{n}$, where the last inequality is due to the construction of $\widetilde{RQ}_\alpha^n$ so that it satisfies $\hat{F}_n^-(\widetilde{RQ}_\alpha^n) \leq \alpha$. Taking once again expectations on both sides and using (3.3) yields $\mathbb{P}(Z_n \leq \tilde{Q}_\alpha^n) \leq \alpha + \frac{1}{n}$. □

The next lemma concerns a similar result when another random variable is added to the set. Note that the statement still refers to the empirical quantile of the first n observations. This is crucial for proving the theoretical coverage guarantee of the split conformal direct prediction method.

Lemma 3.6. *Let $Z_1, \ldots, Z_{n+1}$ be exchangeable random variables. Then*

$$\mathbb{P}(Z_{n+1} \leq \tilde{Q}^n_{\alpha(1+\frac{1}{n})}) \geq \alpha.$$

If, in addition, $Z_1, \ldots, Z_{n+1}$ are almost surely distinct, then

$$\mathbb{P}(Z_{n+1} \leq \tilde{Q}^n_{\alpha(1+\frac{1}{n})}) \leq \alpha + \frac{1}{n+1}.$$

Proof. We first have to distinguish between the order statistics $Z_{(1)}, \ldots, Z_{(n)}$ of the first n observations and the order statistics $Z^*_{(1)}, \ldots, Z^*_{(n+1)}$ of all $n+1$ observations. Observe that $Z_{n+1} \leq Z_{(k)}$ if and only if $Z_{n+1} \leq Z^*_{(k)}$. Using the explicit formulas for the empirical quantiles, we have $\tilde{Q}^n_{\alpha(1+\frac{1}{n})} = Z_{(\lceil (n+1)\alpha \rceil)}$ and $\tilde{Q}^{n+1}_{\alpha} = Z^*_{(\lceil (n+1)\alpha \rceil)}$. Therefore, $Z_{n+1} \leq \tilde{Q}^n_{\alpha(1+\frac{1}{n})}$ if and only if $Z_{n+1} \leq \tilde{Q}^{n+1}_{\alpha}$. Consequently, $\mathbb{P}(Z_{n+1} \leq \tilde{Q}^n_{\alpha(1+\frac{1}{n})}) = \mathbb{P}(Z_{n+1} \leq \tilde{Q}^{n+1}_{\alpha})$. The conclusion follows by applying Lemma 3.5 with $n+1$ instead of n. □

(Split) Conformal Prediction

The scope of the Split Conformal Prediction Method is to construct a prediction interval satisfying the theoretical guarantee (3.1), irrespective of the sample size and the joint distribution of the samples. This method relies on loose *exchangeability* assumption of the observations. The term "conformal" was coined by Vovk et al. in [38] to describe a method that *conforms* to the underlying distribution without any parametric assumptions, highlighting its *model-agnosticity*. We present the more streamlined split version of the method here for brevity, which separates the data into a training set, used for fitting the predictive model, and a calibration set, used for calibrating the prediction interval. In the context of predictive distributions in the upcoming chapter, both the theoretical foundations of general conformal methods and their split variants will be formalized. An acute reader will observe similarities between the cross-validation principle presented in Chapter 2 and the conformalization principle for prediction intervals presented here. Both, roughly speaking, improve upon existing methods by eliminating some observations and then accounting for them later on, but cross-validation does so sequentially. Both methods draw on the ideas of Stone [35], but the more general conformal principle for prediction intervals was first introduced by Vovk in the early 2000s [38].

Under the *exchangeability* assumptions, the method begins by splitting the observations into two disjoint subsets: a training set $\{(x_i, y_i) : i \in \mathcal{I}_1\}$ and a calibration set $\{(x_i, y_i) : i \in \mathcal{I}_2\}$, where $\mathcal{I}_1$ and $\mathcal{I}_2$ form a partition of $\{1, \ldots, n\}$. For example, we can assume without loss of generality that $\mathcal{I}_1 = \{1, \ldots, m\}$ and $\mathcal{I}_2 = \{m+1, \ldots, n\}$. Given any regression algorithm $\mathcal{A}$, the training set is used for regression

$$\hat{y}(x) \leftarrow \mathcal{A}(\{(x_i, y_i) : i \in \mathcal{I}_1\}).$$

Then, on $\mathcal{I}_2$, *calibration scores* are computed between the target variable and the fitted regression model, in the form of absolute residuals

$$E_i = |y_i - \hat{y}(x_i)|, \quad i \in \mathcal{I}_2. \tag{3.4}$$

The empirical quantile of the absolute residuals is then computed, adjusted to the size of the calibration set, i.e. we calculate the $(1-\alpha)(1+\frac{1}{|\mathcal{I}_2|})$ empirical quantile $\tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)$ of $\{E_i : i \in \mathcal{I}_2\}$. At the end, the prediction interval for a new observation x_{n+1} is computed as

$$PI(x_{n+1}) = [\hat{y}(x_{n+1}) - \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2),\ \hat{y}(x_{n+1}) + \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)]. \tag{3.5}$$

The procedure described here is also summarized in Algorithm 1 below.

Algorithm 1 Split Conformal Prediction Method

Input: Dataset $\{(x_i, y_i)\}_{i=1}^n$, new observation x_{n+1}, miscoverage rate $\alpha \in (0,1)$, regression algorithm $\mathcal{A}$.
Algorithm: Partition $\{1,\dots,n\}$ into a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$.
Fit $\hat{y}(x) \leftarrow \mathcal{A}(\{x_i, y_i\} : i \in \mathcal{I}_1)$.
Compute conformity scores $E_i = |y_i - \hat{y}(x_i)|$ for $i \in \mathcal{I}_2$.
Compute the $(1-\alpha)(1+\frac{1}{|\mathcal{I}_2|})$ empirical quantile $\tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)$ of $\{E_i : i \in I_2\}$
Output:

$$\mathrm{PI}(x_{n+1}) = [\hat{y}(x_{n+1}) - \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2),\ \hat{y}(x_{n+1}) + \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)].$$

It is guaranteed that the marginal coverage (3.1) is achieved for exchangeable variables, irrespective of the sample size, with this method, as proven in [23]. Furthermore, an explicit upper bound can be given. This result is summarized in the following theorem.

Theorem 3.7. *Under the assumption that $(X_i, Y_i)_{i=1}^n$ are exchangeable, the prediction interval $PI(X_{n+1})$ constructed by the Split Conformal Prediction Method as described in Algorithm 1 is marginally well-calibrated, i.e., it satisfies the $(1-\alpha) - MC$ guarantee,*

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1})) \geq 1 - \alpha.$$

If, in addition, the conformity scores are almost surely distinct, then the prediction interval is nearly perfectly calibrated

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1})) \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2| + 1}.$$

The fact that the length of the prediction interval is fixed to $2\tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)$ is, however, a major limitation of this algorithm, as it does not contain any information on how uncertain the model is at the new observation. The following three methods we present overcome this limitation by constructing variable-width prediction intervals.

Quantile regression

Conditional quantile regression [21] estimates a given quantile of Y given X. This is cast as an optimization problem over the"pinball" loss [34]

$$\rho_\alpha(y, \hat{y}) := \begin{cases} \alpha(y - \hat{y}) & \text{if } y - \hat{y} > 0 \\ (1-\alpha)(y - \hat{y}) & \text{otherwise.} \end{cases}$$

The estimated quantile function of Y_{n+1} given $X_{n+1} = x$, denoted by $\hat{q}_a(x)$, is found by solving the following optimization problem

$$\hat{q}_\alpha(x) = f(x, \hat{\theta})$$
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(Y_i, f(X_i, \theta)),$$

where $f(x, \theta)$ is the quantile regression fitting function (this can take many forms and is up to the user to specify). The strategy to construct a prediction interval at miscoverage rate α for a new observation x_{n+1} is to then estimate the conditional quantiles $\hat{q}_{\alpha/2}(x_{n+1})$ and $\hat{q}_{1-\alpha/2}(x_{n+1})$ and output the prediction interval

$$PI(x_{n+1}) = [\hat{q}_{\alpha/2}(x_{n+1}),\ \hat{q}_{1-\alpha/2}(x_{n+1})].$$

The procedure is also summarized in Algorithm 2 below.

The main drawback of this algorithm is that the prediction interval constructed is not guaranteed to satisfy the finite sample marginal coverage guarantee described in (3.1). In [27] and [12], it is shown

Algorithm 2 Quantile Regression Method

Input: Dataset $\{(x_i, y_i)\}_{i=1}^n$, new observation x_{n+1}, miscoverage rate $\alpha \in (0,1)$, quantile regression algorithm $\mathcal{B}$.
Algorithm: Fit $\{\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)\} \leftarrow \mathcal{B}(\{x_i, y_i\} : i \in \{1, \ldots, n\})$.
Output:

$$\mathrm{PI}(x_{n+1}) = [\hat{q}_{\alpha/2}(x_{n+1}),\ \hat{q}_{1-\alpha/2}(x_{n+1})].$$

through simulated examples that the constructed intervals through quantile regression can significantly undercover. Only under regularity conditions and certain specific function forms is the theoretical coverage satisfied asymptotically. The following two methods recover the finite-sample theoretical coverage guarantee by drawing on ideas from conformal prediction.

Conformalized Quantile Regression

The Conformalized Quantile Regression method successfully addresses each of the main drawbacks of the above methods by combining their virtues: one provides a theoretical finite-sample guarantee (3.1), and the other yields variable-dependent widths for the prediction intervals. Taken together, it is shown in [27] that the theoretical finite-sample guarantee is satisfied while maintaining the variable-width.

Conformalized Quantile Regression, as in Split Conformal Prediction, starts by splitting the data into a training set indexed by $\mathcal{I}_1$ and a calibration set indexed by $\mathcal{I}_2$. The two conditional quantile functions $\{\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}\}$ are fitted on the training set, given a quantile regression algorithm $\mathcal{B}$.

$$\{\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}\} \leftarrow \mathcal{B}(\{(x_i, y_i) :\ i \in \mathcal{I}_1\}).$$

The next step computes *conformity scores*. They quantify the error made by the "dummy" prediction intervals $C(x_i) = [\hat{q}_{\alpha/2}(x_i), \hat{q}_{1-\alpha/2}(x_i)]$ for $i \in \mathcal{I}_2$, that is, on the calibration set. The *conformity scores* are computed as

$$E_i := \max\{\hat{q}_{\alpha/2}(x_i) - y_i,\ y_i - \hat{q}_{1-\alpha/2}(x_i)\}. \tag{3.6}$$

The interpretation of the conformity score is as follows: if Y_i falls outside the predicted interval, then the conformity score quantifies the magnitude of the error made by this mistake, and E_i is non-negative. If Y_i does belong to the predicted interval, the E_i is the larger of two non-positive numbers, and so is itself non-positive. In this sense, E_i penalizes undercoverage of the interval and also accounts for overcoverage of the interval.

On the set of conformity scores, the empirical quantile is computed, adjusted to the size of the calibration set, i.e. we compute the $(1-\alpha)(1+\frac{1}{|\mathcal{I}_2|})$ empirical quantile $\tilde{Q}_{1-a}(E, \mathcal{I}_2)$ of $\{E_i :\ i \in \mathcal{I}_2\}$. Finally, the prediction interval for a new observation X_{n+1} is computed as

$$\mathrm{PI}(x_{n+1}) = [\hat{q}_{\alpha/2}(x_{n+1}) - \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2), \hat{q}_{1-\alpha/2}(x_{n+1}) + \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)].$$

The procedure described above is also summarized in Algorithm 3.

Algorithm 3 Conformalized Quantile Regression

Input: Dataset $\{(x_i, y_i)\}_{i=1}^n$, observation x_{n+1}, miscoverage rate $\alpha \in (0,1)$, quantile regression algorithm $\mathcal{B}$.
Algorithm: Partition $\{1, \ldots, n\}$ into a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$.
Fit $\{\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)\} \leftarrow \mathcal{B}(\{x_i, y_i\} : i \in \mathcal{I}_1)$.
Compute conformity scores $E_i = \max\{\hat{q}_{\alpha/2}(x_i) - y_i,\ y_i - \hat{q}_{1-\alpha/2}(x_i)\}$ for $i \in I_2$.
Compute the $(1-\alpha)(1+\frac{1}{|\mathcal{I}_2|})$ empirical quantile $\tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)$ of $\{E_i : i \in I_2\}$
Output:

$$\mathrm{PI}(x_{n+1}) = [\hat{q}_{\alpha/2}(x_{n+1}) - \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2),\ \hat{q}_{1-\alpha/2}(x_{n+1}) + \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)].$$

The following theorem [27, Theorem 1] provides a finite-sample marginal guarantee of the produced conformal prediction interval. We present it here without proof. A similar statement will be given anyhow for the Conformal Direct Prediction Method in the following section, which we will prove.

Theorem 3.8. *Under the assumption that $(X_i, Y_i)_{i=1}^n$ are exchangeable, the prediction interval $PI(X_{n+1})$ constructed by the Split Conformal Quantile Regression Method as described in Algorithm 3 is marginally well-calibrated, i.e., it satisfies the $(1-\alpha)-MC$ guarantee,*

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1})) \geq 1 - \alpha.$$

If, in addition, the conformity scores are almost surely distinct, then the prediction interval is nearly perfectly calibrated

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1})) \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2| + 1}.$$

Split Conformal Direct Prediction method

In this section, we describe the conformal direct prediction algorithm in detail, following the approach introduced in [12]. In practice, the conformal direct prediction method has been observed to produce the smallest intervals on average among the four methods, while also yielding variable-dependent widths for the prediction intervals. Until now, no theoretical finite-sample guarantee has been proven, and hereby we expand on the original work by proving that the split conformal direct prediction method also gives a well-calibrated prediction interval.

Let us now describe the procedure for the split conformal direct prediction method. As with all other conformal methods, we shall split the data into a training set indexed by $\mathcal{I}_1$ and a calibration set indexed by $\mathcal{I}_2$. However, unlike the methods described above, we need two regression algorithms. With the first regression algorithm $\mathcal{A}$, we regress the target variable on the features in the training set, and then compute the predicted value at the new observation.

$$\hat{y}(x) \leftarrow \mathcal{A}(\{x_i, y_i\} : i \in \mathcal{I}_1).$$

Next, we compute the absolute-valued residuals $r_i = |y_i - \hat{y}(x_i)|$ for $i \in \mathcal{I}_1$. We then fit, given a quantile regression model $\mathcal{B}$, the quantiles of the residuals r_i on the set of features x_i and compute the $(1-\alpha)$-th quantile

$$\hat{q}^r_{1-\alpha}(x_i) \leftarrow \mathcal{B}(\{x_i, r_i\} : i \in \mathcal{I}_1).$$

As with conformalized quantile regression, we compute conformity scores E_i, where the formula is given as

$$E_i = \max\{\hat{y}(x_i) - \hat{q}^r_{1-a}(x_i) - y_i, y_i - \hat{y}(x_i) - \hat{q}^r_{1-a}(x_i)\} \quad \text{for } i \in I_2. \tag{3.7}$$

Finally, we compute the empirical quantile of conformity scores, adjusted to the size of the calibration set, i.e. we compute the $(1-\alpha)(1+\frac{1}{|\mathcal{I}_2|})$ empirical quantile $\tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)$ of $\{E_i : i \in \mathcal{I}_2\}$. The prediction interval for a new observation x_{n+1} is computed as

$$\mathrm{PI}(x_{n+1}) = [\hat{y}(x_{n+1}) - \hat{q}^r_{1-\alpha}(x_{n+1}) - \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2), \hat{y}(x_{n+1}) + \hat{q}^r_{1-\alpha}(x_{n+1}) + \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)].$$

The procedure described above is also described in Algorithm 4 below.

The Split Conformal Direct Prediction Method distinguishes itself from other conformal prediction methods by its inherent multi-layered nature. Additionally, the method assumes a symmetric distribution of residuals, which can usually be achieved via a variable transformation and is not considered a severe limitation of the algorithm.

We will now show that, under the assumption of data exchangeability, the prediction interval is nearly perfectly calibrated.

Algorithm 4 Split Conformal Direct Prediction Method

Input: Dataset $\{(x_i, y_i)\}_{i=1}^n$, new observation x_{n+1}, miscoverage rate $\alpha \in (0,1)$, regression algorithm for point prediction $\mathcal{A}$, quantile regression algorithm for residuals $\mathcal{B}$.
Algorithm: Partition $\{1,\dots,n\}$ into a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$.
Fit $\hat{y}(x) \leftarrow \mathcal{A}(\{x_i, y_i\} : i \in \mathcal{I}_1)$.
Compute (absolute-valued) residuals $r_i = |y_i - \hat{y}(x_i)|$ for $i \in \mathcal{I}_1$.
Fit the quantile regression model on the training set $\hat{q}^r_{1-\alpha}(x_i) \leftarrow \mathcal{B}(\{x_i, r_i\} : i \in \mathcal{I}_1)$.
Compute conformity scores $E_i = \max\{\hat{y}(x_i) - \hat{q}^r_{1-\alpha}(x_i) - y_i, y_i - \hat{y}(x_i) - \hat{q}^r_{1-\alpha}(x_i)\}$ for $i \in I_2$.
Compute the $(1-\alpha)(1+\frac{1}{|\mathcal{I}_2|})$ empirical quantile $\tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)$ of $\{E_i : i \in I_2\}$.
Output:

$$\mathrm{PI}(X_{n+1}) = [\hat{y}(x_{n+1}) - \hat{q}_{1-\alpha}(x_{n+1}) - \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2),\ \hat{y}(x_{n+1}) + \hat{q}_{1-\alpha}(x_{n+1}) + \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)].$$

Theoretical coverage of split conformal direct prediction method

Theorem 3.9. *Under the assumption that $(X_i, Y_i)_{i=1}^n$ are exchangeable, the prediction interval $PI(X_{n+1})$ constructed by the Split Conformal Direct Prediction Method as described in Algorithm 4 is marginally well-calibrated, i.e., it satisfies the $(1-\alpha) - MC$ guarantee,*

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1})) \geq 1 - \alpha.$$

If, in addition, the conformity scores are almost surely distinct, then the prediction interval is nearly perfectly calibrated

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1})) \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2| + 1}.$$

Proof. Observe that

$$Y_{n+1} \geq \hat{y}(X_{n+1}) - \hat{q}_{1-\alpha}(X_{n+1}) - \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2) \Leftrightarrow \hat{y}(X_{n+1}) - \hat{q}_{1-\alpha}(X_{n+1}) - Y_{n+1} \leq \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)$$

and

$$Y_{n+1} \leq \hat{y}(X_{n+1}) + \hat{q}_{1-\alpha}(X_{n+1}) + \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2) \Leftrightarrow Y_{n+1} - \hat{y}(X_{n+1}) - \hat{q}_{1-\alpha}(X_{n+1}) \leq \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2).$$

Consequently, using the definition of E_{n+1}, we obtain

$$Y_{n+1} \in PI(X_{n+1}) \Leftrightarrow E_{n+1} \leq \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2). \tag{3.8}$$

Since the original pairs $(X_i, Y_i)_{i=1}^n$ are exchangeable, so are the calibration variables E_i, as measurable functions of exchangeable random variables. Therefore, we can apply Lemma 3.6 to get that

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1})) = \mathbb{P}(E_{n+1} \leq \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)) \geq 1 - \alpha.$$

Under the additional assumptions that the conformity scores are almost surely distinct, we also get the upper bound.

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1})) = \mathbb{P}(E_{n+1} \leq \tilde{Q}_{1-\alpha}(E, \mathcal{I}_2)) \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2| + 1}.$$

□

Remark 2. The proofs of Theorems 3.7 and 3.8 follow by a similar argument that leads to the equivalence (3.8).

Discussion on coverage guarantees

In the present chapter, we have focused on a marginal type of coverage, taken over the joint distribution of (X_{n+1}, Y_{n+1}). The $(1-\alpha)$-MC guarantee essentially suggests that on average over all values of X_{n+1}, the constructed interval $PI(X_{n+1})$ is accurate with probability at least $(1-\alpha)$. However, this means that we cannot guarantee that for a specific new value $X_{n+1} = x$ the prediction interval constructed by a specific method is correct with a certain probability, but only that on average over many draws of X_{n+1} it will be.

We can, however, wonder whether the coverage also holds conditionally, in the sense of

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1}) \,|X_{n+1} = x) \geq 1 - \alpha.$$

If the above holds for almost all x, it is said that the method satisfies the $(1-\alpha)$-CC (conditional coverage) guarantee.

Unfortunately, [2] shows that if such a method exists, then the prediction intervals produced must always be of infinite expected length, irrespective of the underlying distribution. Such a method is evidently meaningless, leading up to the realisation that no algorithm can satisfy the $(1-\alpha)$-CC guarantee. The theorem is stated below without proof [2, Proposition 1].

Theorem 3.10. *Suppose that a method satisfies the $(1-\alpha)$-CC guarantee. Then for all underlying distributions, the output prediction interval $PI(X_{n+1})$ must be of infinite expected length at almost all points x aside from the atoms of its distribution P_X, i.e.*

$$\mathbb{E}[\lambda(PI(x))] = \infty,$$

where $\lambda(\cdot)$ denotes the one-dimensional Lebesgue measure.

To better understand the differences between marginal and conditional coverage guarantees, we illustrate these concepts with two example applications, one given in [2], and one in [12].

The first example application is in healthcare [2]. Suppose that at each observation point i we have a patient with x_i relevant features (age, family history, etc.). At the same time, the target variable Y_i represents a measurable outcome, for instance, the reduction in blood pressure after administering a specific drug. Upon the arrival of a new patient at the doctor's office with features x_{n+1}, the doctor would like to predict the reduction in blood pressure Y_{n+1} within a specific range with a certain confidence. A doctor's statement for a patient may look along the lines of "Based on your age and family history, you can expect your blood pressure to go down by 10 to 15 mmHg." The difference between marginal and conditional coverage guarantees in this context is as follows. With $\alpha = 0.05$, the marginal coverage guarantee suggests that the doctor's statement should hold with probability 95% *on average* over all patients arriving at the clinic. This means, for example, that the statement might be underperforming (or, in fact, not performing at all) in some age subgroups, as long as it's offset by overcoverage in others. The conditional coverage guarantee, on the other hand, suggests that the doctor's statement holds with probability 95% for every individual patient.

Similarly, for the real estate application in [12], at each data point i, the sale price of a house is given as the target variable Y_i, whereas X_i enshrines various features that may be socio-economic (neighborhood population density, crime rates, etc), physical environmental (noise intensity levels, air pollution levels, etc) and functional environmental (proximities to services and amenities). The total number of features is 73. One is then interested in providing a financially accurate prediction interval estimate for the house price Y_{n+1} upon collecting the features x_{n+1}. Similarly, at level $\alpha = 0.05$, the prediction interval will be accurate 95% of the time, averaging over all possible features under the marginal coverage guarantee. In contrast, the interval is accurate with 95% probability for each specific house, under the conditional coverage guarantee. In practice, conformal prediction intervals for real estate prices can be very accurate in a specific region of the Netherlands but underperform in another, as long as the accuracy balances out across the entire country.

The inherent drawback of being unable to make accurate predictions conditional on a specific feature is particularly disadvantageous in certain situations, such as healthcare, but marginal guarantees suffice in many practical settings. In more sensitive situations, an asymptotic conditional coverage guarantee

can be obtained at a higher computational cost. This extended method, called Distributional Conformal Prediction [8], requires first training a conditional distribution function on the available data. However, this method falls outside of the scope of our current work and we refer interested readers to [8] for a more detailed discussion.

As a middle ground between *marginal* and *conditional* coverage, in [36] the concept of *training conditional coverage* is considered, namely

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1}) \,|\, \{(X_i, Y_i)\}_{i=1}^n) \geq 1 - \alpha. \tag{3.9}$$

By mimicking the proof of Theorem 3.9 directly, but conditional on the training data, we can see that training conditional validity holds [36] for split CP, CQR, and split CDP, further enhancing the theoretical coverage guarantees of these three methods.

Proposition 3.11. *Under the assumption that* $(X_i, Y_i)_{i=1}^n$ *are exchangeable, the prediction interval* $PI(X_{n+1})$ *constructed by either the Conformalized Quantile Regression, as described in Algorithm 3, or the Split Conformal Direct Prediction Method, as described in Algorithm 4, is training conditionally well-calibrated, i.e.,*

$$\mathbb{P}(Y_{n+1} \in PI(X_{n+1}) \,|\, \{(X_i, Y_i)\}_{i=1}^n) \geq 1 - \alpha.$$

The discussion here highlights the main advantage of conformal prediction: even if the model is misspecified, the calibration procedure specific to conformal prediction ensures a marginal coverage guarantee, and also a training conditional coverage guarantee. This lies in contrast with, for instance, quantile regression, which does not have such a marginal coverage guarantee, and empirically, it is observed that for real estate applications, the quantile regression has, on aggregate, higher miscoverage rates with wider intervals [12]. In the next chapter, we extend the conformal framework to predictive distributions and show that such conformal procedures, if embedded with appropriate modifications to the conformity scores, achieve an almost equivalent marginal validity guarantee to that of conformal prediction intervals.

4

Conformal predictive distributions

We continue with the framework of Chapter 3, where we have a sample set $\{(X_i, Y_i)_{i=1}^n\}$, where each X_i is a random vector describing features of the real-valued target random variable Y_i. We aim to develop a framework to compute conformal predictive distributions for a new Y_{n+1} given the arrival of X_{n+1}. In this chapter, we formalize the idea of *conformal predictive distributions* and their split variant [41, 39]. We furthermore extend this concept to allow for *asymptotic* predictive distributions, and we prove new results that quantify the perturbation induced by the probability integral transform, most prominently Theorem 4.25 and Theorem 4.26.

Foundational work in the theory of predictive distributions [30, 40, 41] seeks a conditional (pseudo)-distribution function that satisfies a similar type of marginal coverage guarantee as for prediction intervals. In fact, this guarantee should be such that one can retrieve a marginally well-calibrated prediction interval from the constructed predictive distribution. To this end, we shall formalize an equivalent guarantee for conformal distributions. First, let us recall some defining properties of distribution functions [41, Section 2].

Defining properties of distribution functions

Let us first consider the case when the distribution function is everywhere continuous, with no points of discontinuity. The following lemma identifies uniquely the distribution function of a random variable.

Lemma 4.1. *Let F be a continuous distribution function on $\mathbb{R}$, Y a random variable with distribution F, and $Q : \mathbb{R} \to \mathbb{R}$ a non-decreasing function. If $Q(Y) \sim \mathrm{Unif}[0, 1]$, then $Q = F$.*

Proof. Suppose, for the sake of contradiction, that there exists a $y \in \mathbb{R}$ such that $Q(y) \neq F(y)$. Let $y^* = \sup\{z : Q(z) = Q(y)\}$. Then, by the definition of F and its continuity, $\mathbb{P}(Q(Y) \leq Q(y)) = F(y^*)$. On the other hand, since $Q(Y)$ is uniform, we also have $\mathbb{P}(Q(Y) \leq Q(y)) = Q(y)$. Then, $Q(y) = F(y^*)$. Since, by assumption $Q(y) \neq F(y)$, it must be that $y^* > y$ by monotonicity of Q, and consequently $F(y*) > F(y)$. Since $Q(y) = Q(y^*)$ and the interval $[y, y^*)$ of positive probability $F(y^*) - F(y)$ is mapped to a single point through Q, we reach a contradiction with the uniform distribution assumption. □

If the distribution function is not everywhere continuous, we can state a similar statement, but, in this case, we have to account for randomization. Below we recall the definition of lexicographic order on $\mathbb{R} \times [0, 1]$.

Definition 4.2 (Lexicographic order)**.** The lexicographic order $(y, \tau) \leq (y', \tau')$ on $\mathbb{R} \times [0, 1]$ is defined to mean that $y < y'$ or both $y = y'$ and $\tau \leq \tau'$.

With this definition in mind, we can now adjust the above lemma for general distribution functions [41, Lemma 2]. We approach the final step of the proof (proving the statement for general $\tau \in [0, 1]$) somewhat differently than the original paper, and explain this step more minutely.

Lemma 4.3. *Let $\mathbb{P}_F$ be a probability measure on $\mathbb{R}$ with distribution function F, and Y is a random variable distributed as $\mathbb{P}_F$. Let U be the induced probability measure of the uniform distribution on $[0,1]$ and $\tau_U \sim \mathrm{Unif}[0,1]$ independent of Y.Denote by $\mathbb{P}$ the product measure $\mathbb{P}_F \times U$. Define $Q : \mathbb{R}\times[0,1] \to \mathbb{R}$ as a non-decreasing function with respect to the lexicographic order on $\mathbb{R} \times [0,1]$. Suppose the image $(\mathbb{P}_F \times U)Q^{-1}$ of the product $\mathbb{P}_F \times U$ under the mapping Q is uniform on $[0,1]$, that is, $Q(Y, \tau_U) \sim \mathrm{Unif}[0,1]$. Then, for all y and τ,*

$$Q(y,\tau) = (1-\tau)F(y-) + \tau F(y). \tag{4.1}$$

Here, $F(y-) = \lim_{z\to y^-} F(z)$.

Proof. We first prove that $Q(y,1) = F(y)$ for all $y \in \mathbb{R}$. Assume for the sake of contradiction that there exists $y \in \mathbb{R}$ such that $Q(y,1) \neq F(y)$ and set, similarly to Lemma 4.1,

$$y^* = \sup\{z : Q(z,1) = Q(y,1)\}. \tag{4.2}$$

By a similar argument to Lemma 4.1, we have for $(Y, \tau_U) \sim \mathbb{P}_F \times U$,

$$\begin{aligned} Q(y,1) &= \mathbb{P}(Q(Y,\tau_U) \le Q(y,1)) \\ &\ge \mathbb{P}(Q(Y,1) \le Q(y,1)) \quad (\text{as } (Y,\tau_U) \le (Y,1) \text{ a.s. and } Q \text{ is non-decreasing}) \\ &\ge \mathbb{P}((Y,1) \le (y,1)) \\ &= \mathbb{P}_F(Y \le y) = F(y). \end{aligned}$$

Since by assumption, we have $Q(y,1) \neq F(y)$, we must have $Q(y,1) > F(y)$. If the supremum in (4.2) is attained, then

$$F(y) < Q(y,1) = \mathbb{P}(Q(Y,1) \le Q(y,1)) = \mathbb{P}((Y,1) \le (y^*,1)) = F(y^*),$$

and the lexicographic interval $((y,1),(y^*,1)]$ of positive probability $F(y^*) - F(y)$ gets mapped through Q into one point. Similarly, if the supremum in (4.2) is not attained, then

$$F(y) < Q(y,1) = \mathbb{P}(Q(Y,1) \le Q(y,1)) = \mathbb{P}((Y,1) \le (y^*,1)) = F(y^*-),$$

and once again the lexicographic interval $((y,1),(y^*,0))$ of positive probability $F(y^*-) - F(y)$ gets mapped through Q into one point. Both cases contradict the fact that the distribution of Q is uniform, which shows that $Q(y,1) = F(y)$ for all $y \in \mathbb{R}$. Analogously, we prove that $Q(y,0) = F(y-)$ for all $y \in \mathbb{R}$.

Observe now that (4.1) holds trivially for all τ at any point of continuity of F, by lexicographic monotonicity of Q. Now fix y such that $F(y-) < F(y)$ and $\alpha \in (F(y-), F(y))$. The lexicographic monotonicity of Q implies that the preimage of the interval $(F(y-), \alpha]$ is

$$Q^{-1}\left((F(y-),\alpha]\right) = \{y\} \times (0, \tau_\alpha],$$

where τ_α satisfies $Q(y,\tau_\alpha) = \alpha$. Since $Q(Y,\tau_U) \sim \mathrm{Unif}[0,1]$, the measure of the preimage must equal the length of the interval, that is,

$$(F(y) - F(y-)) \cdot \tau_\alpha = \alpha - F(y-).$$

Solving for α yields $\alpha = Q(y,\tau_\alpha) = (1-\tau_\alpha)F(y-) + \tau_\alpha F(y)$. Since this holds for all $\alpha \in (F(y-), F(y))$, we conclude that $Q(y,\tau) = (1-\tau)F(y-) + \tau F(y)$ for all $\tau \in [0,1]$. □

The two lemmas above show that the function being non-decreasing and that the mapping of Y through its distribution F being uniformly distributed are defining properties of distribution functions of probability measures. The mapping $F(Y)$ is called the *Probability Integral Transform*. We are now equipped with the necessary tools to define and understand conformal predictive distributions properly.

Conformal predictive distributions

We once again proceed as in [30, Definition 1] and [41]. Assume we have p features and define the corresponding observation space as $\mathbb{R}^{p+1} = \mathbb{R}^p \times \mathbb{R}$. An element $z = (x, y)$ with $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$ of the observation space is called an *observation* consisting of features $x \in \mathbb{R}^p$ and target $y \in \mathbb{R}$. We aim, given a sequence of observations $(z_i)_{i=1}^n$ and a new test feature x_{n+1}, to predict the target y_{n+1}. We furthermore assume, as in Chapter 3, that the random observations are exchangeable. We shall continue with the notation above in this chapter.

At a high level, Vovk et al. [41] aim to construct an object based on conformity scores that acts similarly to a cumulative distribution function, while retaining a similar marginal validity guarantee as the ones we obtained in Chapter 3. However, as we shall see later on, this object, called a conformal transducer, includes randomization to break ties between conformity scores, on the one hand, and to account for its own discontinuities on the other. With randomization in mind, the overarching aim of this chapter is to construct a conformal transducer that is also a randomized predictive system, which we define below, as in [41, Definition 1].

Definition 4.4 (Randomized predictive system)**.** A right-continuous function $Q : (\mathbb{R}^{p+1})^{n+1} \times [0,1] \to [0,1]$ is called a *randomized predictive system* (RPS) if it satisfies:

1. *Lexicographic Monotonicity*: For any fixed training sequence $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$ and test feature $x_{n+1} \in \mathbb{R}^p$, the function $Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau)$ is non-decreasing in (y, τ), which is understood in the sense of the lexicographic order on $\mathbb{R} \times [0,1]$ as defined in Definition 4.2.
2. *Boundary Stability:* For all $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$ and $x_{n+1} \in \mathbb{R}^p$,

$$\lim_{y \to -\infty} Q(z_1, \ldots, z_n, (x_{n+1}, y), 0) = 0 \quad \text{and} \quad \lim_{y \to \infty} Q(z_1, \ldots, z_n, (x_{n+1}, y), 1) = 1.$$

3. *Validity:* For any exchangeable sequence of random variables $Z_1, \ldots, Z_{n+1}$ taking values in $\mathbb{R}^{p+1}$ and any $\tau \sim \mathrm{Unif}[0,1]$ independent of the sequence, the random variable $Q(Z_1, \ldots, Z_n, Z_{n+1}, \tau)$ follows the uniform distribution on $[0,1]$, i.e.,

$$\mathbb{P}(Q(Z_1, \ldots, Z_n, Z_{n+1}, \tau) \leq \alpha) = \alpha \quad \text{for all } \alpha \in [0,1]. \tag{4.3}$$

Definition 4.5 (Randomized predictive distribution)**.** A randomized predictive distribution (RPD) is defined as the function

$$Q_n : (y, \tau) \in \mathbb{R} \times [0,1] \mapsto Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau), \tag{4.4}$$

which is the output of the randomized predictive system Q on a training sequence $z_1, \ldots, z_n$ and a test feature x_{n+1}, coupled with a random number $\tau \sim \mathrm{Unif}[0,1]$.

In Chapter 3, we presented three different (split) conformal prediction methods. A main point of divergence between the three is the (split) conformity score used, which was loosely defined at that point. Here, we can now formally define how those scores arise using conformity measures.

Definition 4.6 (Conformity measure)**.** A *conformity measure* is a function $E : (\mathbb{R}^{p+1})^{n+1} \to \mathbb{R}$ which is measurable, and invariant under permutations of the first n arguments, that is, for any n-tuple $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$, any $z_{n+1} \in \mathbb{R}^{p+1}$, and any permutation π of $\{1, \ldots, n\}$,

$$E(z_1, \ldots, z_n, z_{n+1}) = E(z_{\pi_1}, \ldots, z_{\pi_n}, z_{n+1}).$$

Definition 4.7 (Conformity score)**.** For a fixed $y \in \mathbb{R}$ and conformity measure $E : (\mathbb{R}^{p+1})^{n+1} \to \mathbb{R}$, the corresponding *conformity scores* are defined by

$$\begin{aligned} E_i &:= E(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, (x_{n+1}, y), z_i), \quad i = 1, \ldots, n, \\ E^y &:= E(z_1, \ldots, z_n, (x_{n+1}, y)). \end{aligned} \tag{4.5}$$

As we have briefly described in the previous chapter, conformity scores quantify how unusual a specific observation is. We need one more ingredient to formally define conformal predictive distributions,

namely the concept of a *conformal transducer*. Intuitively, conformal transducers compute a "randomized" p-value of a test of exchangeable observations. The conformal transducer breaks ties in the conformity scores by introducing a uniform random variable as extra input. Since conformal transducers depend only on the conformity scores, one can find the necessary and sufficient conditions on these conformity scores to ensure that the conformal transducer is a randomized predictive system. We shall see what these conditions are in the following section for the split variant. As mentioned above, we construct a conformal predictive distribution to be aligned with the conformal prediction intervals framework: that is, we aim to be able to retrieve a marginally accurate prediction interval from the predictive distribution.

Definition 4.8 (Conformal transducer)**.** The *conformal transducer* determined by a conformity measure E and its corresponding conformity scores E_i is defined as

$$Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau) := \frac{|\{i = 1, \ldots, n : \ E_i < E^y\}| + \tau |\{i = 1, \ldots, n : E_i = E^y\}| + \tau}{n+1}.$$

Conversely, a function is called a *conformal transducer* if it is the conformal transducer of some conformity measure.

Definition 4.9 (Conformal predictive system)**.** A *conformal predictive system* is a function that is both a conformal transducer and a randomized predictive system.

Definition 4.10 (Conformal predictive distribution)**.** For a conformal predictive system Q, its randomized predictive distribution Q_n is called a *conformal predictive distribution*.

Upon constructing a conformal predictive distribution, one can now retrieve the conformal predictor that produces a prediction region of the type described in Chapter 3, which is marginally valid.

Definition 4.11 (Conformal predictor)**.** For a conformal predictive system Q and Borel set $A \subset [0, 1]$, the conformal predictor is given by

$$\Gamma^A(z_1, \ldots, z_n, x_{n+1}, \tau) := \{y \in \mathbb{R} : Q(z_1, \ldots, z_n, (x_{n+1}, y)), \tau) \in A\}.$$

Remark 3. The standard property of validity for a conformal predictive system, as outlined in (4.3), is that its associated p-values $Q(z_1, \ldots, z_{n+1}, \tau)$ are uniformly distributed on $[0, 1]$, which implies that the coverage probability of $Y_{n+1} \in \Gamma^A(Z_1, \ldots, Z_n, X_{n+1})$ is $\lambda(A)$, where λ is the usual Lebesgue measure. Consequently, one can retrieve the marginal guarantee of conformal prediction intervals automatically from the construction of conformal predictive distributions. Indeed, if Q is a conformal predictive system, then the prediction interval

$$\Gamma^{1-\alpha}(z_1, \ldots, z_n, x_{n+1}, \tau) = \left\{y \in \mathbb{R} : Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau) \in \left[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right]\right\},$$

is exactly $(1 - \alpha)$ marginally well calibrated, in view of (4.3).

The nuances of conformity measures

In the present context, the usual interpretation of a conformal transducer is that it represents a randomized p-value for testing the null hypothesis of the observations being exchangeable. For conformal predictive distributions, the informal alternative hypothesis is that $y_{n+1} = y$ is smaller than expected under the exchangeable model. In this case, the selected conformity measure quantifies how well the observation (x_{n+1}, y_{n+1}) conforms to the remaining observations. Observe the one-sided nature of this notion of conformity; an observation can only be *non-conforming* if it is too small relative to the model's prediction. This restricts the conformity measures one can choose to confer this information on conformal transducers.

For example, the conformity measures used in Chapter 3 do not satisfy the necessary requirements to produce conformal predictive distributions. That is, the core requirement of lexicographic monotonicity is lost when unsigned measures are being used. In the present context, an unsigned measure is one in which positive and negative deviations are treated as equally non-conforming. Specifically, the output of a conformal transducer for a candidate label $y \in \mathbb{R}$ is typically unimodal rather than monotonic in the

case of unsigned measures. As $y \to \pm\infty$, the conformity score E^y increases without bound, leading to $\lim_{y\to+\infty} Q(z_1,\ldots,z_n,(x_{n+1},y),1) = 0$, which goes against boundary stability.

Consequently, the conformal transducer associated with an unsigned conformity measure usually does not satisfy the necessary requirements of a conformal predictive system. To ensure that the output $Q(z_1,\ldots,z_{n+1},\tau)$ satisfies the axiomatic requirements of a conformal predictive system, the conformity measure must be signed and monotonic in y, and a common choice is

$$E(z_1,\ldots,z_{n+1}) = y_{n+1} - \hat{y}_{n+1}.$$

We discuss more conformity measure choices in the following section, once we account for the usual practical notion of splitting the data into a training set and a calibration set. For split conformity measures, one can actually give exactly the necessary and sufficient conditions to be imposed on the split conformity measures as to ensure that the associated conformal transducer is a randomized predictive system.

The split variant

We now modify the general setting above to specifically permit split variants of the algorithms, as presented in Chapter 3. As full conformal systems require retraining the underlying algorithm for each test feature with associated postulated target, they are rather computationally inefficient. The split version offers better computational efficiency, whereas its predictive efficiency is empirically lower than that of the full version [39]. The main reference point for the terminology introduced here is [39].

The framework here will not be unfamiliar to the attentive reader, and it should feel rather natural in light of the high-level overview of conformal prediction interval methods presented in Chapter 3. We begin by defining a split conformity measure, as a slight modification of the conformity measure defined in Definition 4.6.

Definition 4.12 (Split conformity measure)**.** A family of measurable functions $E_m : (\mathbb{R}^{p+1})^{m+1} \to \mathbb{R}$ with $m = 1, 2, \ldots$, is called a *split conformity measure*.

As before, we shall split the sequence $z_1,\ldots,z_n$ into two: a training set $\{z_i : i \in \mathcal{I}_1\}$ and a calibration set $\{z_i : i \in \mathcal{I}_2\}$ where $\mathcal{I}_1$ and $\mathcal{I}_2$ form a partition of the index set $\{1,\ldots,n\}$. For ease of presentation, we shall assume without loss of generality, $\mathcal{I}_1 = \{1,\ldots,m\}$ and $\mathcal{I}_2 = \{m+1,\ldots,n\}$. Suppose we fix m; from now on, we omit the subscript m when referring to a split conformity measure corresponding to a split with m training objects. The corresponding split conformity scores are now slightly adjusted and defined below. Notice that one does not need to permute between observations anymore, and the formula is thus simpler than that of the full conformity scores.

Definition 4.13 (Split conformity score)**.** For a split conformity measure $E : (\mathbb{R}^{p+1})^{m+1} \to \mathbb{R}$ and a fixed $y \in \mathbb{R}$, the corresponding *split conformity scores* are defined by

$$\begin{aligned} E_i &= E(z_1,\ldots,z_m,(x_{m+i},y_{m+i})), \quad i = 1,\ldots,n-m. \\ E^y &= E(z_1,\ldots,z_m,(x_{n+1},y)). \end{aligned} \tag{4.6}$$

Remark 4. We can now formally define the split conformity measures used in the conformal methods presented in Chapter 3. Under the notation used in this chapter, the standard Split Conformal Prediction method uses the split conformity measure $E_{CP}(z_1,\ldots,z_{n+1}) = |y_{n+1} - \hat{y}(x_{n+1})|$, whereas Conformalized Quantile Regression uses $E_{CQR}(z_1,\ldots,z_{n+1}) = \max\{\hat{q}_{\alpha/2}(x_{n+1}) - y_{n+1}, y_{n+1} - \hat{q}_{1-\alpha/2}(x_{n+1})\}$ and Conformal Direct Prediction uses $E_{CDP}(z_1,\ldots,z_{n+1}) = \max\{\hat{y}(x_{n+1}) - \hat{q}^r_{1-\alpha}(x_{n+1}) - y_{n+1}, y_{n+1} - \hat{y}(x_{n+1}) - \hat{q}^r_{1-\alpha}(x_{n+1})\}$. The corresponding split conformity scores are then as described in (3.4), (3.6) and (3.7).

Definition 4.14 (Split conformal transducer)**.** The *split conformal transducer* determined by a conformity measure E and its corresponding conformity scores is defined as

$$Q(z_1,\ldots,z_n,(x_{n+1},y),\tau) := \frac{|\{i = 1,\ldots,n-m : E_i < E^y\}| + \tau|\{i = 1,\ldots,n-m : E_i = E^y\}| + \tau}{n-m+1}. \tag{4.7}$$

Conversely, a function is called a *split conformal transducer* if it is the split conformal transducer of some split conformity measure.

From now on in this chapter, we introduce the following notation. Let

$$\begin{aligned}\mathcal{E} &:= \{E_1, \ldots, E_{n-m}, E^y\},\\ \mathcal{E}_< &:= \{i \in \{1, \ldots, n-m\} : E_i < E^y\},\\ \mathcal{E}_= &:= \{i \in \{1, \ldots, n-m\} : E_i = E^y\},\\ \mathcal{E}_> &:= \{i \in \{1, \ldots, n-m\} : E_i > E^y\},\\ N_< &:= |\mathcal{E}_<|,\ N_= := |\mathcal{E}_=|,\ N_> := |\mathcal{E}_>|.\end{aligned}$$

Then, the split conformal transducer formula (4.7) rewrites as

$$Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau) := \frac{N_< + \tau N_= + \tau}{n-m+1} \tag{4.8}$$

It is noteworthy to observe that the standard property of validity (4.3) adapted to split conformity measures is satisfied automatically. A statement similar to this one is found in [38, Theorem 11.1], where it is given without proof, and it is also mentioned as a known fact in [39, Section 3]. In our current work, we relax the assumption of IID to exchangeability for what appears to be the first time in the literature and give a complete proof of the statement.

Theorem 4.15. *Let $Q : (\mathbb{R}^{p+1})^{m+1} \times [0,1] \to \mathbb{R}$ be a split conformal transducer associated with some conformity measure E. If $Z_1, \ldots . Z_n, Z_{n+1}$ are exchangeable, where $Z_{n+1} = (X_{n+1}, Y_{n+1})$, and $\tau \sim \mathrm{Unif}[0,1]$ is independent of $Z_1, \ldots, Z_n, Z_{n+1}$, then $Q(Z_1, \ldots, Z_n, Z_{n+1}, \tau)$ follows the uniform distribution on $[0,1]$.*

Proof. By the exchangeability of $Z_1, \ldots, Z_n, Z_{n+1}$ and the independence of the model construction from the calibration indices, the sequence $(E_{m+1}, \ldots, E_n, E^y)$ is also exchangeable.

We first compute the distribution of Q conditional on $\mathcal{E}$. Denote by $v_1 < v_2 < \cdots < v_k$ the k distinct values in $\mathcal{E}$, arranged in ascending order, with multiplicities $c_1, c_2, \ldots, c_k$ respectively. Note that $\sum_{i=1}^{k} c_i = n-m+1$. Under the assumption of exchangeability, which implies that the variables in $\mathcal{E}$ are identically distributed, we have, for any $j \in \{1, \ldots, k\}$,

$$\mathbb{P}(E^y = v_j \mid \mathcal{E}) = \frac{c_j}{n-m+1}.$$

We observe that if $E^y = v_j$, then $N_< = \sum_{l=1}^{j-1} c_l$ and $N_= = c_j - 1$. Substituting these into the definition of Q, we get, conditional on $E^y = v_j$ and $\mathcal{E}$, that

$$Q(Z_1, \ldots, Z_n, Z_{n+1}, \tau) = \frac{\sum_{i=1}^{j-1} c_i + \tau c_j}{n-m+1}.$$

To simplify, let us call $Q_{Z_{n+1}} := Q(Z_1, \ldots, Z_n, Z_{n+1}, \tau)$. Since $\tau \sim \mathrm{Unif}[0,1]$, the conditional distribution of $Q_{Z_{n+1}}$ given $E^y = v_j$ and $\mathcal{E}$ is uniform on the interval $I_j = \left(\frac{\sum_{i=1}^{j-1} c_i}{n-m+1}, \frac{\sum_{i=1}^{j} c_i}{n-m+1}\right]$. The length of this interval is exactly $\frac{c_j}{n-m+1}$. Summing over all j and using the law of total probability, we get

$$\begin{aligned}\mathbb{P}(Q_{Z_{n+1}} \le u \mid \mathcal{E}) &= \sum_{j=1}^{k} \mathbb{P}(Q_{Z_{n+1}} \le u \mid E^y = v_j, \mathcal{E}) \mathbb{P}(E^y = v_j \mid \mathcal{E})\\ &= \sum_{j=1}^{k} \mathbb{P}(Q_{Z_{n+1}} \le u \mid E^y = v_j, \mathcal{E}) \cdot \frac{c_j}{n-m+1}\\ &= \sum_{j=1}^{k} \left[\mathbb{1}_{\{u \in I_j\}} \cdot \frac{\left(u - \frac{\sum_{i=1}^{j-1} c_i}{n-m+1}\right)}{c_j/(n-m+1)} + \mathbb{1}_{\{u > \frac{\sum_{i=1}^{j} c_i}{n-m+1}\}} \right] \cdot \frac{c_j}{n-m+1}.\end{aligned}$$

Note that the intervals I_j partition $(0,1]$. For any $u \in [0,1]$, let j^* be the index such that $u \in I_{j^*}$. Then, the probability rewrites as

$$\mathbb{P}(Q_{Z_{n+1}} \leq u \mid \mathcal{E}) = \sum_{j=1}^{j^*-1} \frac{c_j}{n-m+1} + \frac{\left(u - \frac{\sum_{i=1}^{j^*-1} c_i}{n-m+1}\right)}{c_j^*/(n-m+1)} \cdot \frac{c_j^*}{n-m+1} = u.$$

Finally, to conclude the statement unconditionally, we use the law of total expectation, as follows

$$\mathbb{P}(Q_{Z_{n+1}} \leq u) = \mathbb{E}(\mathbb{1}_{\{Q_{Z_{n+1}} \leq u\}}) = \mathbb{E}(\mathbb{E}(\mathbb{1}_{\{Q_{Z_{n+1}} \leq u\}} \mid \mathcal{E})) = \mathbb{E}(\mathbb{P}(Q_{Z_{n+1}} \leq u \mid \mathcal{E})) = \mathbb{E}(u) = u,$$

which is what we wanted to prove. □

Now that we know the standard property of validity is automatically satisfied, the question of whether a split conformal transducer is a Randomized Predictive System boils down to checking whether the lexicographic monotonicity and boundary stability conditions are satisfied. Vovk et al. give in [39] the necessary and sufficient conditions for this to happen, which are presented below. The following definition is as in [41, Section 2.2].

Definition 4.16 (Monotonic split conformity measure)**.** A split conformity measure E is called *monotonic* if $E(z_1, \ldots, z_m, (x,y))$ is monotonically increasing in y, i.e.,

$$y \leq y' \Rightarrow E(z_1, \ldots, z_m, (x,y)) \leq E(z_1, \ldots, z_m, (x,y')),$$

for all x.

Although this definition is rather straightforward, an additional condition is needed to ensure that a split conformal transducer is an RPS. To this end, we must define the concept of a convex set and convex hull. These two definitions are based on [3].

Definition 4.17 (Convex set)**.** Let S be a vector space and C a subset of S. We say that C is *convex* if, for all $x, y \in C$, we have $(1-t)x + ty \in C$, for all $t \in [0,1]$.

Definition 4.18 (Convex hull)**.** The *convex hull* of a set C, denoted by $\mathrm{conv}(C)$, is the minimal convex set containing C.

Example 4.19. Let $C = \{2,3\} \subset \mathbb{R}$. Then the convex hull $\mathrm{conv}(C) = [2,3]$.

In our present context, we are primarily interested in the convex closures of subsets of $\mathbb{R}$. These can take one of the following four forms: (a,b), $[a,b)$, $(a,b]$, or $[a,b]$, with $a, b \in \mathbb{R} \cup \{-\infty, +\infty\}$. We are now equipped to define *balanced* split conformity measures, following [39].

Definition 4.20 (Balanced split conformity measure)**.** A monotonic split conformity measure E is *balanced* if, for all x and, for any m and $z_1, \ldots, z_m$, the set

$$\mathrm{conv}\, E(z_1, \ldots, z_m, (x, \mathbb{R})) := \mathrm{conv}\{E(z_1, \ldots, z_m, (x,y)) : y \in \mathbb{R}\}$$

does not depend on x and is an open interval in $\mathbb{R}$.

Remark 5. In the context of conformal prediction, an underlying assumption is that the (split) conformity measure is *coercive*, in the sense of $\mathrm{conv} E(z_1, \ldots, z_m, \mathbb{R}^{p+1}) = (-\infty, +\infty)$. Intuitively, this means that as y drifts further away from our proposed model prediction, the (split) conformity measure notices this aspect and reflects it in its associated conformity score. We carry this assumption from now on.

We are now prepared to show that balanced and monotonic split conformity measures give split conformal transducers that are RPS, as defined in Definition 4.4. The following statement is originally given in [39, Proposition 1 and 2].

Theorem 4.21. *The split conformal transducer based on a split conformity measure E is an RPS if and only if E is balanced and monotonic.*

Proof. We first show that if E is balanced and monotonic, then the associated split conformal transducer is an RPS. Since Theorem 4.15 gives that the validity requirement is automatically satisfied, we only need to check for lexicographic monotonicity and boundary stability. It is clear that the split conformal transducer is non-decreasing in τ by the construction given in (4.7).

We now show that it is also non-decreasing in y. Since the split conformity measure is monotonic, it follows that as y increases, so will E^y. Then, as E^y increases, the cardinality $N_>$ can only decrease, while $N_<$ can not decrease. The associated coefficients in the formula (4.8) are 1 for $N_<$, $\tau \in [0,1]$ for $N_=$, and 0 for $N_>$, and so the overall split conformal transducer value has to increase in y. This demonstrates lexicographic monotonicity.

Furthermore, since E is balanced, there exists y^* and y' such that $E^y < \min_{i\in\{1,\dots,n-m\}} E_i$ for all $y < y^*$ and $E^y > \max_{i\in\{1,\dots,n-m\}} E_i$ for all $y > y'$ respectively. Consequently, $N_< = 0$ and $N_= = 0$ for all $y < y^*$, and plugging this back into (4.8) yields

$$\lim_{y\to-\infty} Q(z_1,\dots,z_n,(x_{n+1},y),0) = 0.$$

Analogously, we get that $N_< = n-m+1$ and $N_= = 0$ for all $y > y'$, and so

$$\lim_{y\to\infty} Q(z_1,\dots,z_n,(x_{n+1},y),1) = 1,$$

which proves the Boundary Stability condition. Therefore, the associated split conformity transducer Q is an RPS.

Conversely, suppose the split conformal transducer based on E is an RPS. Fix m, $z_1,\dots,z_m$ and x, and take $n = m+1$, that is, we have m training points and a single calibration point. Suppose, for the sake of contradiction, we have $y < y'$ such that $E(z_1,\dots,z_m,(x,y)) > E(z_1,\dots,z_m,(x,y'))$. Then, for $z_{m+1} = (x,y)$, we have $E_1 = E^y$ and $E_1 > E^{y'}$. Then,

$$Q(z_1,\dots,z_n,(x,y),1) = Q(z_1,\dots,z_m,z_{m+1},(x,y),1) = \frac{|N_<| + |N_=| + 1}{2} = 1$$
$$Q(z_1,\dots,z_n,(x,y'),1) = Q(z_1,\dots,z_m,z_{m+1},(x,y'),1) = \frac{|N_<| + |N_=| + 1}{2} = \frac{1}{2},$$

which contradicts the lexicographic monotonicity condition of an RPS. Therefore, E is monotonic.

Suppose now that E is not balanced. That is, for some fixed $m, z_1,\dots,z_m$ and $x, x' \in \mathbb{R}^{p+1}$, we have

$$\operatorname{conv} E(z_1,\dots,z_m,(x,\mathbb{R})) \neq \operatorname{conv} E(z_1,\dots,z_m,(x',\mathbb{R})).$$

We may, without loss of generality, let $y \in \operatorname{conv} E(z_1,\dots,z_m,(x,\mathbb{R}))$ such that $y < y'$ for all $y' \in \operatorname{conv} E(z_1,\dots,z_m,(x',\mathbb{R}))$ (recall that all convex hulls on $\mathbb{R}$ are intervals). Then, for $z_{m+1} = (x,y)$, $\tau = 0$ and any test pair (x',y'), where x' is fixed and y' is varying, we have $|N_<| = 1$, and consequently, plugging this into (4.8),

$$Q(z_1,\dots,z_m,z_{m+1},(x',y'),0) = \frac{1}{2},$$

which, by letting y' go to $-\infty$, contradicts the boundary stability condition of

$$\lim_{y'\to-\infty} Q(z_1,\dots,z_m,z_{m+1},(x',y'),0) = 0.$$

Therefore, the split conformity measure E must be balanced, which is what we wanted to prove. □

Corollary 4.21.1. *The split conformal transducer based on either of the following modified split conformity measures, stemming from the conformal prediction method, conformalized quantile regression, and the conformal direct prediction method, respectively, is an RPS.*

$$E_{CP}(z_1,\dots,z_m,(x_{n+1},y)) = y - \hat{y}(x_{n+1}) \tag{4.9}$$

$$E_{CQR}(z_1,\dots,z_m,(x_{n+1},y)) = y - \hat{q}_{1-\alpha}(x_{n+1}) \tag{4.10}$$

$$E_{CDP}(z_1,\dots,z_m,(x_{n+1},y)) = y - \hat{y}(x_{n+1}) - \hat{q}^r_{1-\alpha}(x_{n+1}) \tag{4.11}$$

Proof. All three conformity measures are linear in y with positive coefficients, and therefore monotonic. Keeping everything else fixed, we see that as $y \to -\infty$, the conformity measures go to $-\infty$ as well, and as $y \to +\infty$, the conformity measures go to $+\infty$. Therefore, we have that the convex hull of $E(z_1, \ldots, z_m, (x_{n+1}, y))$ over $y \in \mathbb{R}$ is, in all three cases, $(-\infty, +\infty)$ and therefore does not depend on x_{n+1}. Therefore, the three split conformity measures are also balanced. By Theorem 4.21, each of the split conformal transducers based on these three split conformity measures is an RPS. □

Now that we have described the types of split conformity measures that can be used to produce a conformal predictive system, we present below the general algorithm for producing a conformal predictive distribution. Note that, by construction, the conformal predictive distribution still depends on $\tau \sim \text{Unif}[0, 1]$. To account for this, we observe that, as we let τ travel from 0 to 1, we obtain an interval for each fixed y. This is a "fuzzy" distribution and not a true distribution in the sense of producing a single output for a fixed y. We first give the pseudocode algorithm to create fuzzy predictions [41, Algorithm 1], and we then adjust the construction to allow for randomized predictive distributions [39, Algorithm 1].

Algorithm 5 "Fuzzy" Split Conformal Predictive System

Input: Dataset $\{z_i = (x_i, y_i)\}_{i=1}^n$, observation (test object) x_{n+1}.
Algorithm: Partition $\{1, \ldots, n\}$ into a training set $\mathcal{I}_1 = \{1, \ldots, m\}$ and a calibration set $\mathcal{I}_2 = \{m + 1, \ldots, n\}$.
for $i \in \{1, \ldots, n - m\}$ **do**
Solve for C_i in the equation $E(z_1, \ldots, z_m, (x_{m+i}, y_{m+i})) = E(z_1, \ldots, z_m, (x_{n+1}, C_i))$.
end for
Sort $C_1, \ldots, C_{n-m}$ in ascending order to obtain $C_{(1)}, \leq \cdots \leq C_{(n-m)}$, and set $C_{(0)} = -\infty$ and $C_{(n-m+1)} = +\infty$.
Output: Return a "fuzzy" predictive distribution for the label y of x_{n+1}

$$Q_n(y) := \begin{cases} \left[\frac{i}{n-m+1}, \frac{i+1}{n-m+1}\right] & \text{if } y \in (C_{(i)}, C_{(i+1)}) \text{ for } i \in \{0, 1, \ldots, n-m\}. \\ \\ \left[\frac{i'-1}{n-m+1}, \frac{i''+1}{n-m+1}\right] & \text{if } y = C_{(i)} \text{ for } i \in \{0, 1, \ldots, n-m\}, \end{cases} \tag{4.12}$$

where $i' := \min\{j : C_{(j)} = C_{(i)}\}$ and $i'' := \max\{j : C_{(j)} = C_{(i)}\}$.

The "fuzzy" split conformal predictive system can be easily modified to obtain a pseudo-distribution function by simply drawing a τ value and using it as an input. The procedure is described in Algorithm 6.

Algorithm 6 Split Conformal Predictive System

Input: Dataset $\{z_i = (x_i, y_i)\}_{i=1}^n$, observation (test object) x_{n+1},
Algorithm: Draw $\tau \sim \text{Unif}[0, 1]$.
Partition $\{1, \ldots, n\}$ into a training set $\mathcal{I}_1 = \{1, \ldots, m\}$ and a calibration set $\mathcal{I}_2 = \{m + 1, \ldots, n\}$.
for $i \in \{1, \ldots, n - m\}$ **do**
Solve for C_i in the equation $E(z_1, \ldots, z_m, (x_{m+i}, y_{m+i})) = E(z_1, \ldots, z_m, (x_{n+1}, C_i))$.
end for
Sort $C_1, \ldots, C_{n-m}$ in ascending order to obtain $C_{(1)} \leq \cdots \leq C_{(n-m)}$, and set $C_{(0)} = -\infty$ and $C_{(n-m+1)} = +\infty$.
Output: Return a predictive distribution for the label y of x_{n+1}

$$Q_n(y) := \begin{cases} \frac{i+\tau}{n-m+1} & \text{if } y \in (C_{(i)}, C_{(i+1)}) \text{ for } i \in \{0, 1, \ldots, n-m\}. \\ \\ \frac{i'-1+(i''-i'+2)\tau}{n-m+1} & \text{if } y = C_{(i)} \text{ for } i \in \{0, 1, \ldots, n-m\}. \end{cases} \tag{4.13}$$

where $i' := \min\{j : C_{(j)} = C_{(i)}\}$ and $i'' := \max\{j : C_{(j)} = C_{(i)}\}$.

In the two algorithms above, one can see that there are $n - m$ equations to be solved for C_i (which we

call the *conformal atoms* from now on). This is not a computationally expensive step, as it often just amounts to rearranging the output of the scores. For instance, using the standard conformal prediction score amounts to solving

$$C_i - \hat{y}(x_{n+1}) = y_{m+i} - \hat{y}(x_{m+i}) \Leftrightarrow C_i = \hat{y}(x_{n+1}) + y_{m+i} - \hat{y}(x_{m+i}).$$

For our three classic split conformity scores (4.9),(4.10) and (4.11), the conformal atoms are then merely a shifted version of the conformity scores E_i. In other words, we have the relation

$$C_i = \hat{y}(x_{n+1}) + E_i. \tag{4.14}$$

Furthermore, in view of (4.14), under the common assumptions in Chapter 3 that the conformity scores are almost surely distinct, we get that the conformal atoms are also almost surely distinct. The output of the conformal predictive system in Algorithm 6 simplifies to

$$Q_n(y) := \begin{cases} \frac{i+\tau}{n-m+1} & \text{if } y \in (C_{(i)}, C_{(i+1)}) \text{ for } i \in \{0, 1, \ldots, n-m\}. \\ \\ \frac{i-1+2\tau}{n-m+1} & \text{if } y = C_{(i)} \text{ for } i \in \{0, 1, \ldots, n-m\}. \end{cases} \tag{4.15}$$

The split conformal predictive system with output Q_n thus induces an underlying probability measure μ_n on $\bar{\mathbb{R}}$ satisfying

$$\mu_n(\cdot \mid x_{n+1}) := \frac{\tau}{n-m+1}\delta_{-\infty} + \sum_{i=1}^{n-m} \frac{1}{n-m+1}\delta_{C_{(i)}(x_{n+1})} + \frac{1-\tau}{n-m+1}\delta_{+\infty}, \tag{4.16}$$

where δ is the usual Dirac delta measure. We denote in the formula $C_{(i)}(x_{n+1})$ instead of the usual $C_{(i)}$ to highlight the dependence on the observed x_{n+1}. Then, the randomized predictive system $Q(\cdot, (x_{n+1}, y), \tau)$, in view of (4.7), satisfies

$$Q(\cdot, (x_{n+1}, y), \tau) = \mu_n((-\infty, y)|x_{n+1}) + \tau\mu_n(\{y\}), \tag{4.17}$$

where we use the short-handed notation $Q(\cdot, (x_{n+1}, y), \tau) = Q(z_1, \ldots, z_n, (x_{n+1}, y), \tau)$.

We call the output $Q_n(y)$ a pseudo-distribution in Algorithm 6 since on the interval $(-\infty, C_{(1)})$ the output is $\frac{\tau}{n-m+1}$, whereas on the interval $(C_{(n-m)}, +\infty)$ the output is $\frac{n-m+\tau}{n-m+1}$. In particular, for every $\tau \in [0, 1]$, the conformal predictive system output does not satisfy $\lim_{y\to-\infty} Q_n(y) = 0$ and $\lim_{y\to\infty} Q_n(y) = 1$ simultaneously.

This acts as an obstacle for us, since we aim to eventually obtain a predictive density function. However, cumulative distribution functions have underlying measures defined on $\mathbb{R}$ and they do not give weight at $\pm\infty$. It is unclear how one can retrieve a probability density function from an underlying measure defined on the extended real line $\bar{\mathbb{R}}$. It appears, then, that some of the conditions for randomized predictive systems must be relaxed. We aim for a small measure-correcting solution by reassigning the tail mass at $\pm\infty$ to $C_{(1)}$ and $C_{(n-m)}$ respectively. We then aim to have an asymptotic randomized predictive system, as defined below.

Definition 4.22 (Asymptotic RPS)**.** A right-continuous function $Q : (\mathbb{R}^{p+1})^{n+1} \times [0, 1] \to [0, 1]$ is called an *asymptotic* randomized predictive system if it satisfies the lexicographic monotonicity and boundary stability conditions as in Definition 4.4 and the validity condition (4.3) holds asymptotically as $n \to \infty$, that is, for any exchangeable sequence of random variables $Z_1, \ldots, Z_{n+1}$ taking values in $\mathbb{R}^{p+1}$ and any $\tau \sim \mathrm{Unif}[0, 1]$ independent of the sequence, the random variable $Q(Z_1, \ldots, Z_n, Z_{n+1}, \tau)$ converges in distribution to $\mathrm{Unif}[0, 1]$.

A measure correction solution

Denote by $\bar{\mu}_n(\cdot \mid x_{n+1})$ the tail-corrected measure with formula

$$\bar{\mu}_n(\cdot \mid x_{n+1}) := \frac{1+\tau}{n-m+1}\,\delta_{C_{(1)}(x_{n+1})} + \sum_{i=2}^{n-m-1} \frac{1}{n-m+1}\,\delta_{C_{(i)}(x_{n+1})} + \frac{2-\tau}{n-m+1}\,\delta_{C_{(n-m)}(x_{n+1})}. \tag{4.18}$$

This measure now lives on the real line $\mathbb{R}$ and admits a proper distribution function. The randomized predictive distribution associated with the tail-corrected measure is then given by

$$\bar{Q}(\cdot,(x_{n+1},y),\tau)=\bar{\mu}_n((-\infty,y)|x_{n+1})+\tau\bar{\mu}_n(\{y\}). \tag{4.19}$$

Observe that this is in the same format as (4.17). We now show that applying this measure correction yields a predictive distribution that is approximately uniform and precisely quantify the deviation from the uniform distribution. To do so, we introduce the concept of total variation between measures [4].

Definition 4.23. The *total variation distance* between two probability measures μ,ν on the measurable space $(\Omega,\mathcal{F})$ is defined as

$$\mathrm{TV}(\mu,\nu)=\sup_{A\in\mathcal{F}}|\mu(A)-\nu(A)|. \tag{4.20}$$

Alternatively, the total variation can be rewritten in the following dual form over $f:\Omega\to\mathbb{R}$

$$\mathrm{TV}(\mu,\nu)=\frac{1}{2}\sup_{\|f\|_\infty\le 1}\left|\int f\,d\mu-\int f\,d\nu\right|. \tag{4.21}$$

In our present context, however, the measures are purely atomic, so they are supported on a finite space. In this particular case, the total variation formula has a closed form, which we show in the following lemma. This lemma is a simple corollary to the Hahn-Jordan decomposition theorem [6].

Lemma 4.24. *Let $(\Omega,\mathcal{F})$ be a measurable space and let μ and ν be probability measures on Ω. Assume that there exists a finite set $S=\{z_1,\dots,z_k\}\subset\Omega$ such that both measures are supported on S, i.e. $\mu(\Omega\setminus S)=0$ and $\nu(\Omega\setminus S)=0$.*

Then

$$\mathrm{TV}(\mu,\nu)=\frac{1}{2}\sum_{z\in S}\left|\mu(\{z\})-\nu(\{z\})\right|. \tag{4.22}$$

Proof. Since both measures are supported on S, for any measurable set A, we have

$$\mu(A)-\nu(A)=\sum_{z\in A\cap S}\Delta(z),$$

where $\Delta(z):=\mu(\{z\})-\nu(\{z\})$. Observe that

$$\sum_{z\in S}\Delta(z)=\mu(\Omega)-\nu(\Omega)=1-1=0,$$

since μ and ν are probability measures. Define the subsets of positive and negative differences:

$$P:=\{z\in S:\Delta(z)>0\},\qquad N:=\{z\in S:\Delta(z)<0\}.$$

Then

$$\mu(A)-\nu(A)=\sum_{z\in A\cap P}\Delta(z)+\sum_{z\in A\cap N}\Delta(z).$$

Including any element of N decreases the sum, and excluding any element of P also decreases the sum. Hence, the supremum is attained at $A=P$, yielding

$$\sup_A(\mu(A)-\nu(A))=\sum_{z\in P}\Delta(z).$$

Similarly,

$$\nu(A)-\mu(A)=\sum_{z\in A\cap P}(-\Delta(z))+\sum_{z\in A\cap N}(-\Delta(z)).$$

is maximized by taking $A=N$, yielding

$$\sup_A(\nu(A)-\mu(A))=\sum_{z\in N}(-\Delta(z)).$$

However, since $\sum_{z\in S}\Delta(z)=0$, we have $\sum_{z\in P}\Delta(z)=\sum_{z\in N}(-\Delta(z))$. Hence both suprema are equal, and we can write

$$\mathrm{TV}(\mu,\nu)=\sum_{z\in P}\Delta(z)=\frac{1}{2}\sum_{z\in S}|\Delta(z)|,$$

since $\sum_{z\in P}\Delta(z)+\sum_{z\in N}(-\Delta(z))=\sum_{z\in S}|\Delta(z)|$.

Substituting the definition of $\Delta(z)$ yields (4.22), which is what we wanted to prove. □

Theorem 4.25. *Let $\mu_n(\cdot \mid x_{n+1})$ be defined as in* (4.16) *and its associated tail-corrected measure $\bar{\mu}_n(\cdot \mid x_{n+1})$ as in* (4.18)*. The associated randomized predictive systems are constructed as in* (4.17) *and* (4.19)*, respectively.*

If $n-m\to\infty$ as $n\to\infty$, then $\bar{Q}(\cdot,(X_{n+1},Y_{n+1}))$ is an asymptotic randomized predictive system. Furthermore, we have the bound

$$\sup_{u\in[0,1]}\left|\mathbb{P}(\bar{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau)\le u)-u\right|\le\frac{2}{n-m+1}.$$

Proof. We first embed $\bar{\mu}_n(\cdot \mid x_{n+1})$ into $\overline{\mathbb{R}}$ by assigning zero mass to $\pm\infty$, so both measures live on the same measurable space. Since both measures are purely atomic, their total variation distance is given by

$$\mathrm{TV}(\mu_n(\cdot \mid x_{n+1}),\bar{\mu}_n(\cdot \mid x_{n+1}))=\frac{1}{2}\sum_{z\in\overline{\mathbb{R}}}\left|\mu_n(\{z\} \mid x_{n+1})-\bar{\mu}_n(\{z\} \mid x_{n+1})\right|,$$

in view of Lemma 4.24. The two measures differ only at four atoms: $-\infty,+\infty,C_{(1)},C_{(n-m)}$. A direct computation shows

$$\begin{aligned}
\left|\mu_n(\{-\infty\} \mid x_{n+1})-\bar{\mu}_n(\{-\infty\} \mid x_{n+1})\right| &= \left|\frac{\tau}{n-m+1}-0\right|=\frac{\tau}{n-m+1},\\
\left|\mu_n(\{C_{(1)}\} \mid x_{n+1})-\bar{\mu}_n(\{C_{(1)}\} \mid x_{n+1})\right| &= \left|\frac{1}{n-m+1}-\frac{1+\tau}{n-m+1}\right|=\frac{\tau}{n-m+1},\\
\left|\mu_n(\{C_{(n-m)}\} \mid x_{n+1})-\bar{\mu}_n(\{C_{(n-m)}\} \mid x_{n+1})\right| &= \left|\frac{1}{n-m+1}-\frac{2-\tau}{n-m+1}\right|=\frac{1-\tau}{n-m+1},\\
\left|\mu_n(\{+\infty\} \mid x_{n+1})-\bar{\mu}_n(\{+\infty\} \mid x_{n+1})\right| &= \left|\frac{1-\tau}{n-m+1}-0\right|=\frac{1-\tau}{n-m+1}.
\end{aligned}$$

Therefore,

$$\mathrm{TV}(\mu_n(\cdot \mid x_{n+1}),\bar{\mu}_n(\cdot \mid x_{n+1})=\frac{1}{n-m+1}\quad\text{for all } x_{n+1}\in\mathbb{R}.$$

As a function of the random variable X_{n+1}, we read the above equality as holding for every realization x_{n+1} of X_{n+1}. Therefore, the equality also holds in the almost sure sense, that is, marginally on X_{n+1}, we get

$$\mathrm{TV}(\mu_n(\cdot \mid X_{n+1}),\bar{\mu}_n(\cdot \mid X_{n+1}))=\frac{1}{n-m+1}\quad\mathbb{P}-\text{a.s.}\tag{4.23}$$

Now consider the randomized predictive systems of (4.17) and (4.19). For each y, define the function

$$f_{y,\tau}(z)=\mathbb{1}_{\{z<y\}}+\tau\,\mathbb{1}_{\{z=y\}}.\tag{4.24}$$

Then, we can write

$$Q(\cdot,(X_{n+1},y))=\int f_{y,\tau}\,d\mu_n,\quad \bar{Q}(\cdot,(X_{n+1},y))=\int f_{y,\tau}d\bar{\mu}_n.\tag{4.25}$$

Furthermore, by construction, we have $|f_{y,\tau}| \leq 1$ a.e., since it can only take the values 0, τ, and 1. Then, using the dual version of the total variation (4.21), we can write, for every $y \in \mathbb{R}$,

$$\begin{aligned} |Q(\cdot,(X_{n+1},y),\tau) - \bar{Q}(\cdot,(X_{n+1},y),\tau)| &= \left| \int f_{y,\tau}\, d\mu_n - \int f_{y,\tau}\, d\bar{\mu}_n \right| \\ &\leq \sup_{\|f\|_\infty \leq 1} \left| \int f\, d\mu_n - \int f\, d\bar{\mu}_n \right| \\ &= 2\,\mathrm{TV}(\mu_n(\cdot \mid X_{n+1}), \bar{\mu}_n(\cdot \mid X_{n+1})) = \frac{2}{n-m+1}. \end{aligned} \tag{4.26}$$

Let $u \in [0,1]$. Then, we have, in view of the above inequality,

$$\begin{aligned} \{\bar{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u\} &\subseteq \left\{ Q(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u + \frac{2}{n-m+1} \right\}, \\ \left\{ Q(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u - \frac{2}{n-m+1} \right\} &\subseteq \left\{ \bar{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u \right\}. \end{aligned}$$

Upon taking probabilities and using the fact that $Q(\cdot,(X_{n+1},Y_{n+1})) \sim \mathrm{Unif}[0,1]$, we obtain

$$\max\left\{0, u - \frac{2}{n-m+1}\right\} \leq \mathbb{P}(\bar{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u) \leq \min\left\{1, u + \frac{2}{n-m+1}\right\}.$$

This inequality holds for all $u \in [0,1]$. Therefore, we have the bound

$$\sup_{u\in[0,1]} \left| \mathbb{P}(\bar{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u) - u \right| \leq \frac{2}{n-m+1},$$

which immediately gives that $\bar{Q}(\cdot,(X_{n+1},Y_{n+1}))$ converges in distribution to $\mathrm{Unif}[0,1]$ as $n \to \infty$. □

A step towards predictive densities

So far, we have presented the standard conformal approach for producing marginally accurate predictive distributions [41] using randomized predictive systems. A downside of this approach is that, while the marginal validity is satisfied exactly, the output of the predictive system is not a true cumulative distribution function. In the previous section, we relaxed the definition of a randomized predictive system to allow the validity condition (4.3) to be satisfied asymptotically. We then explicitly give a tail-corrected version of the underlying measure (4.16) that creates a proper predictive distribution function. However, the output CDF is still piecewise constant with jumps. Our end goal remains to retrieve a predictive density function, and one straightforward approach is to have a continuous distribution function and directly differentiate to obtain the probability density function, while retaining an approximate marginal validity statement. The following statement shows that, if we can control the absolute-value difference between a continuous version $\widehat{Q}$ of the tail-corrected predictive distribution $\bar{Q}$, then we can control the distance from the uniform distribution of the marginal distribution of $\widehat{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau)$.

Theorem 4.26. *Let $\widehat{Q}(\cdot,(X_{n+1},Y_{n+1}))$ be a continuous version of $\bar{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau)$, as defined in Theorem 4.25. Suppose that, for every $y \in \mathbb{R}$, we have*

$$|\widehat{Q}(\cdot,(X_{n+1},y),\tau) - \bar{Q}(\cdot,(X_{n+1},y),\tau)| \leq \varepsilon, \tag{4.27}$$

where $\varepsilon > 0$. Then, the following bound holds

$$\sup_{u\in[0,1]} \left| \mathbb{P}(\widehat{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u) - u \right| \leq \varepsilon + \frac{2}{n-m+1}. \tag{4.28}$$

Proof. We use an almost identical argument to the previous theorem. First, we combine (4.26) and (4.27) with the triangle inequality to obtain

$$\begin{aligned}|\widehat{Q}(\cdot,(X_{n+1},y),\tau) - Q(\cdot,(X_{n+1},y),\tau)| &\leq |\widehat{Q}(\cdot,(X_{n+1},y),\tau) - \bar{Q}(\cdot,(X_{n+1},y),\tau)| \\ &\quad + |\bar{Q}(\cdot,(X_{n+1},y),\tau) - Q(\cdot,(X_{n+1},y),\tau)| \\ &\leq \varepsilon + \frac{2}{n-m+1}.\end{aligned}$$

Letting $u \in [0,1]$, we have, in view of the above inequality, the following inclusions

$$\begin{aligned}\{\widehat{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u\} &\subseteq \left\{Q(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u + \varepsilon + \frac{2}{n-m+1}\right\}, \\ \left\{Q(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u - \varepsilon - \frac{2}{n-m+1}\right\} &\subseteq \left\{\widehat{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u\right\}.\end{aligned}$$

Upon taking probabilities and using, once again, the fact that $Q(\cdot,(X_{n+1},Y_{n+1})) \sim \text{Unif}[0,1]$, we obtain

$$\max\left\{0, u - \varepsilon - \frac{2}{n-m+1}\right\} \leq \mathbb{P}(\widehat{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u) \leq \min\left\{1, u + \varepsilon + \frac{2}{n-m+1}\right\}.$$

This inequality holds for all $u \in [0,1]$. Therefore, we obtain the desired bound

$$\sup_{u\in[0,1]} \left|\mathbb{P}(\widehat{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u) - u\right| \leq \varepsilon + \frac{2}{n-m+1}.$$

□

The above theorem remains of paramount importance when we extend our concepts to predictive densities, and we will come back to it in Chapter 5. A natural suggestion would be to obtain a predictive density by directly differentiating (in closed form) a continuous version of the conformal predictive density. In the following chapter, we present such an approach and we explore its advantages and limitations. We furthermore introduce two alternative derivations of predictive densities via filtering and via quantile-matching.

5

Retrieving predictive densities from conformal distributions

In the present chapter, we extend conformal predictive distributions to predictive densities. The simplest and most direct approach is to apply a straightforward finite differencing step directly at the conformal atoms. Another straightforward approach is to smooth the piecewise-constant predictive distributions constructed in Chapter 4 so that we can control the difference in (4.27). Then, Theorem 4.26 can be applied, and we have a measure of the marginal distance of the probability transform from the true uniform distribution. By directly differentiating, we then obtain a predictive density. This method has the advantage of providing a closed-form expression for the predictive density at every point, without relying on a finite-difference step that can introduce additional errors. Both of these approaches suffer at the density level by being noisy and sensitive to small changes between conformal atoms.

It will then appear evident that we must approach the issue of highly fluctuating densities from a different angle. To reduce noise, we propose two solutions: the first uses Gaussian filtering. However, after filtering, the associated predictive distribution may no longer satisfy theoretical guarantees regarding the distance of the Probability Integral Transform from the true uniform distribution. The second approach is original in the conformal context, and we shall call it the *quantile-matching* method. It merges the two worlds of conformal prediction intervals with conformal predictive distribution in a principled way. To reduce noise, we preselect a set of quantile levels of interest and compute the associated conformal quantiles. As we shall see later on, this essentially amounts to selecting a subset of the conformal atoms $C_{(i)}$. Then, finite differencing is applied on this subset. With this approach, we can reduce the noise in the predictive density while maintaining an upper bound on the perturbation in the Probability Integral Transform of the associated predictive distribution.

Direct finite differencing

The most direct approach is to implement a forward finite-difference scheme on the tail-corrected conformal predictive distribution.

$$\hat{f}(\cdot,(x,C_{(i)})) = \frac{\bar{Q}(\cdot,(x,C_{(i+1)}),\tau) - \bar{Q}(\cdot,(x,C_{(i)}),\tau)}{C_{(i+1)} - C_{(i)}} = \frac{1}{(n-m+1)(C_{(i+1)} - C_{(i)})}, \tag{5.1}$$

Then, simple linear interpolation is applied to fill in the values between atoms.

$$\hat{f}(\cdot,(x,y)) = \hat{f}(\cdot,(x,C_{(i)})) + (y - C_{(i)})\frac{\hat{f}(\cdot,(x,C_{(i+1)})) - \hat{f}(\cdot,(x,C_{(i)}))}{C_{(i+1)} - C_{(i)}}.$$

This idea, however, leads to very noisy densities, even when a large number of conformal atoms are used. For our use case, the number of conformal atoms is upwards of hundreds of thousands.

Monotonic cubic interpolation

The approach of monotonic interpolation ensures that the output is increasing, while maintaining the same values as the original function at the jump points. Furthermore, by construction, since between jumps the output function is monotonic, the distance from the original piecewise-constant function cannot be larger than the size of the jumps, meaning we can obtain a closed-form upper bound of the perturbation from the uniform distribution of the probability integral transform. As our goal is to obtain a continuous density function upon differentiating our distribution, cubic splines appear, at first glance, to be a good candidate. Monotonic cubic interpolation is a standard approach introduced in 1980 [15], and it appears promising in our context, so we briefly summarize it here. We follow [15] and [10]. For more streamlined notation, we also use [42] as a reference. We present the method here for a general set of points, but in our case, we will apply it to the jump points of the predictive distributions, which occur exactly at the ordered conformal atoms.

Let $I = [a, b]$ be an interval and

$$a = x_1 < x_2 < \cdots < x_n = b$$

be a partition of I. Furthermore, assume we have an increasing sequence $(f_i)_{i=1}^n$, that is $f_i \leq f_{i+1}$ for $i = 1, \ldots, n-1$. We aim to construct a continuously differentiable piecewise cubic function $p(x)$ on I such that

$$p(x_i) = f_i \quad \text{for } i = 1, \ldots, n.$$

Furthermore, we impose that p is monotonically increasing.

On each subinterval $I_i = [x_i, x_{i+1}]$, $p(x)$ is represented as a cubic polynomial taking the following form

$$p(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i, \tag{5.2}$$

where

$$\begin{aligned}
a_i &= \frac{1}{\Delta x_i^2}\left(-2\frac{\Delta f_i}{\Delta x_i} + f_i' + f_{i+1}'\right), \\
b_i &= \frac{1}{\Delta x_i}\left(3\frac{\Delta f_i}{\Delta x_i} - 2f_i' - f_{i+1}'\right), \\
c_i &= f_i' \\
d_i &= f_i,
\end{aligned}$$

with

$$\begin{aligned}
f_i' &= p'(x_i), \quad i = 1, \ldots, n, \\
\Delta f_i &= f_{i+1} - f_i, \quad i = 1, \ldots, n-1, \\
\Delta x_i &= x_{i+1} - x_i, \quad i = 1, \ldots, n-1.
\end{aligned}$$

To preserve scale invariance of the derivatives, we further introduce the following relationship between the derivative and the slope $m_i = \frac{\Delta f_i}{\Delta x_i}$:

$$\begin{aligned}
f_i' &= \alpha_i m_i, \\
f_{i+1}' &= \beta_i m_i,
\end{aligned}$$

for $\alpha_i \geq 0$ and $\beta_i \geq 0$. It becomes apparent that, despite the points f_i being fixed, one can change the derivatives f_i' to accommodate a large family of interpolating cubic splines. That is, the procedure of interpolating with cubic splines becomes essentially a procedure for calculating the values $(f_i')_{i=1}^n$. We then are interested in finding those values $(f_i')_{i=1}^n$ such that the resulting interpolant is monotonic.

On each subinterval $[x_i, x_{i+1}]$, the curve is monotonic if and only if there is no sign change in the derivative value along any path in the interval. Then, a necessary condition to ensure monotonicity is that

$$\operatorname{sgn}(f_i') = \operatorname{sgn}(f_{i+1}') = \operatorname{sgn}(m_i), \tag{5.3}$$

where sgn is the usual sign function. Furthermore, if $m_i = 0$, then p is monotone on the interval if and only if $f'_i = f'_{i+1} = 0$. Additionally, if $m_i = 0$, then p is constant if and only if $y'_i = y'_{i+1} = 0$. For the remainder of this section, assume $m_i \neq 0$ and (5.3) is satisfied. Then, $p(x)$ is monotonic on $[x_i, x_{i+1}]$ if $p'(x) \neq 0$ for all $x \in [x_i, x_{i+1}]$, which implies that there are no local extrema on this interval. Direct calculations, together with the above restrictions, yield the necessary and sufficient conditions in the following lemma. A detailed derivation can be consulted in [15].

Lemma 5.1. *1. If $\alpha_i + \beta_i - 2 \leq 0$, $p(x)$ is monotone on $[x_i, x_{i+1}]$ if and only if* (5.3) *is satisfied.*

2. If $\alpha_i + \beta_i - 2 > 0$, then $p(x)$ is monotone on $[x_i, x_{i+1}]$ if and only if (5.3) *is satisfied and one of the following conditions is satisfied:*

(a) $2\alpha_i + \beta_i - 3 \leq 0$,

(b) $\alpha_i + 2\beta_i - 3 \leq 0$,

(c) $\alpha_i^2 + \alpha_i(\beta_i - 6) + (\beta_i - 3)^2 \leq 0$.

As a practical way to ensure the conditions in the above lemma are satisfied, one can cast the following optimization problem [42].

$$
\begin{aligned}
\min \quad & \sum_{k=0}^{n-2} \left(12a_k^2\,\Delta x_k^3 + 12a_k b_k\,\Delta x_k^2 + 4b_k^2\,\Delta x_k\right) \\
\text{subject to} \quad & a_k\Delta x_k^3 + b_k\Delta x_k^2 + c_k\Delta x_k + y_k = y_{k+1}, \\
& 3a_k\Delta x_k^2 + 2b_k\Delta x_k + c_k = c_{k+1}, \\
& -c_k \leq 0, \\
& -3a_k\Delta x_k^2 - 2b_k\Delta x_k - c_k \leq 0, \\
& -3a_k\Delta x_k^2 - 2b_k\Delta x_k - 3m_k \leq 0, \\
& 3a_k\Delta x_k^2 + 2b_k\Delta x_k + 3c_k - 9m_k \leq 0, \\
& 6a_k\Delta x_k^2 + 4b_k\Delta x_k + 3c_k - 9m_k \leq 0, \\
& 3a_k\Delta x_k^2 + 2b_k\Delta x_k - 3m_k \leq 0.
\end{aligned}
$$

In the above optimization problem, the first condition denotes interpolation, the second enforces first derivative continuity, and the rest enforce monotonicity. The above optimization problem can be solved numerically, and modern numerical software packages readily provide the monotonic cubic interpolation algorithm.

Quantifying the difference

We now return to the framework of conformal predictive distributions, where monotonic cubic interpolation can be used at the discontinuities $C_{(i)}$ of the constructed conformal predictive distributions in Algorithm 7. Theorem 4.26 gives us an explicit bound of how far away from the true conformal predictive distribution we can be after creating a continuous version of the tail-corrected distribution. Before applying the monotonic cubic interpolation algorithm to construct a predictive distribution $\widehat{Q}_{MCI}$, we fix, for all $x \in \mathbb{R}$,

$$
\begin{aligned}
\widehat{Q}_{MCI}(\cdot, (x, C_{(1)}), \tau) &= \bar{Q}(\cdot, (x, C_{(1)}+), \tau) = \frac{1+\tau}{n-m+1}, \\
\widehat{Q}_{MCI}(\cdot, (x, C_{(i)}), \tau) &= \bar{Q}(\cdot, (x, C_{(i)}+), \tau) = \frac{i+\tau}{n-m+1}, \quad i = 2, \ldots, n-m-1, \\
\widehat{Q}_{MCI}(\cdot, (x, C_{(n-m)}), \tau) &= \bar{Q}(\cdot, (x, C_{(n-m)}+), \tau) = 1, \\
\widehat{Q}_{MCI}(\cdot, (x, C_{(1)} - \gamma), \tau) &= 0,
\end{aligned}
$$

where $\gamma > 0$ is a small value to ensure a smooth continuous transition on $(-\infty, C_{(1)})$. Then the monotonic cubic interpolation algorithm can be applied on the interval $[C_{(1)} - \gamma,\, C_{(n-m)}]$ to obtain a smooth

distribution on this interval (the algorithm requires bounded support to be applied). The extension to the entire real line is then done by setting

$$\widehat{Q}_{MCI}(\cdot,(x,y),\tau) = 0 \quad \text{for } y < C_{(1)} - \gamma,$$
$$\widehat{Q}_{MCI}(\cdot,(x,y),\tau) = 1 \quad \text{for } y > C_{(n-m)}.$$

With this construction, we obtain the following bounds, as a direct consequence of Theorem 4.26.

Proposition 5.2. *Let $\widehat{Q}_{MCI}(\cdot,(X_{n+1},Y_{n+1},\tau))$ be the monotonic cubic interpolated version of the tail-corrected distribution $\bar{Q}(\cdot,(X_{n+1},Y_{n+1}),\tau)$, as defined in Theorem 4.25. Then, the following bound holds*

$$\sup_{u\in[0,1]} \left|\mathbb{P}(\widehat{Q}_{MCI}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u) - u\right| \leq \frac{4}{n-m+1}.$$

Proof. By construction, since $\widehat{Q}_{MCI}$ is monotonic, and $\bar{Q}$ is piecewise constant on $(C_{(i)},C_{(i+1)})$, the distance between $\widehat{Q}_{MCI}$ and $\bar{Q}$ in y cannot be greater than the largest of jumps between the discontinuities of $\bar{Q}$. The largest such jump is of size $\frac{1+\tau}{n-m+1}$ at $C_{(1)}$ and $C_{(n-m)}$ respectively. Furthermore, $\frac{1+\tau}{n-m+1} \leq \frac{2}{n-m+1}$ a.s., as $\tau \sim \mathrm{Unif}[0,1]$. Therefore, we have the bound

$$|\widehat{Q}_{MCI}(\cdot,(X_{n+1},y),\tau) - \bar{Q}(\cdot,(X_{n+1},y),\tau)| \leq \frac{2}{n-m+1} \quad \text{a.s.}, \tag{5.4}$$

for all $y \in \mathbb{R}$. Consequently, Theorem 4.26 can be applied to obtain the desired bound

$$\sup_{u\in[0,1]} \left|\mathbb{P}(\widehat{Q}_{MCI}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u) - u\right| \leq \frac{4}{n-m+1}.$$

□

Retrieving predictive densities via MCI

Once a continuous predictive distribution function is constructed via MCI, the associated predictive probability density function can be retrieved by differentiating the distribution function. In our case, the monotonic cubic interpolated distribution admits a closed-form derivative. Recall that $\widehat{Q}_{MCI}(\cdot,(x,y).\tau)$ is of the form

$$\widehat{Q}_{MCI}(\cdot,(x,y),\tau) = a_i(y-C_{(i)})^3 + b_i(y-C_{(i)})^2 + c_i(y-C_{(i)}) + d_i \tag{5.5}$$

on any interval $(C_{(i)},C_{(i+1)})$, where we now use the convention $C_{(0)} = C_{(1)} - \gamma$. Then, the associated predictive density $\hat{f}_{MCI}(\cdot,(x,y))$ on $[C_{(i)},C_{(i+1)}]$ is computed as

$$\hat{f}_{MCI}(\cdot,(x,y)) = 3a_i(y-C_{(i)})^2 + 2b_i(y-C_{(i)}) + c_i. \tag{5.6}$$

However, as the number of samples n increases, so does the number of calibration points $n-m$ (assuming a split that scales with the number of calibration points, which is standard practice). Consequently, the coefficients a_i, b_i, c_i change on $n-m+1$ subintervals, leading to fluctuating derivatives of the distribution function. This phenomenon can be observed in the following section, in Figure 5.2. The density function will then appear highly fluctuating as well, even though, in its closed form, it satisfies

$$\left|\int_{-\infty}^{y'} \hat{f}_{MCI}(\cdot,(X_{n+1},y))\,\mathrm{d}y - \bar{Q}(\cdot,(X_{n+1},y'),\tau)\right| \leq \frac{2}{n-m+1},$$

for all $y' \in \mathbb{R}$, directly by (5.4). Consequently, from Proposition 5.2, we obtain

$$\sup_{u\in[0,1]} \left|\mathbb{P}(\widehat{Q}_{MCI}(\cdot,(X_{n+1},Y_{n+1}),\tau) \leq u) - u\right| \leq \frac{4}{n-m+1}.$$

This remains true if, instead, a central finite-difference scheme is used and numerical integration is applied to reverse the operation.

It is up to the practitioner whether using MCI is worth the extra step. Especially in large samples, the difference between evaluating closed-form via MCI and directly finite differencing becomes virtually negligible.

For clearer visualization purposes and to better understand the underlying shape and modality of the density function, one can apply a Gaussian filtering step pre-differentiation. The predictive density is then obtained via a central-difference scheme on an evenly spaced grid. This, however, breaks the theoretical guarantees, but our numerical experiments show good empirical performances of this technique, which we will elaborate on shortly after describing the Gaussian filtering technique.

Gaussian filtering: the return of the kernel

Let $g(y)$ denote a continuous function that we wish to smooth. Gaussian filtering provides a principled method to obtain a smooth approximation of g by performing a convolution with a Gaussian kernel. This section is brief and is inspired mainly by the Python documentation on one-dimensional Gaussian filtering [29]. For a more detailed explanation, we refer readers to classical books on filtering, such as [33].

Recall the continuous Gaussian kernel with standard deviation σ defined as

$$K_\sigma(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{y^2}{2\sigma^2}\Big), \quad y \in \mathbb{R}. \tag{5.7}$$

The Gaussian-filtered version of a continuous function $g(y)$ is given by the convolution

$$g_{\mathrm{GF}}(y) = (g * K_\sigma)(y) = \int_{-\infty}^{\infty} g(z)\, K_\sigma(y-z)\, dz, \tag{5.8}$$

which can be interpreted as a locally weighted average of g, with weights decaying exponentially with distance from y.

However, in practice, we only have a finite set of samples of g on a uniform grid. Let y_i, $i = 0, \dots, N-1$, denote the grid points, defined by

$$y_i = y_{\min} + i\,\Delta y, \qquad \Delta y = \frac{y_{\max} - y_{\min}}{N-1}.$$

In our implementation, we use $N = 2000$ equally spaced points.

The continuous convolution integral is then approximated by the discrete sum

$$g_{\mathrm{GF}}(y_i) \approx \sum_{j=-R}^{R} g(y_{i+j})\, k_j, \qquad k_j = \frac{\exp\!\Big(-\frac{(j\Delta y)^2}{2\sigma^2}\Big)}{\sum_{l=-R}^{R} \exp\!\Big(-\frac{(l\Delta y)^2}{2\sigma^2}\Big)}, \tag{5.9}$$

where k_j is the normalized discrete Gaussian kernel with standard deviation σ, and R is the truncation radius chosen so that the kernel is negligible beyond $|j| > R$. The sum therefore spans $2R+1$ grid points centered at y_i. Since the convolution requires samples $g(y_{i+j})$ that may fall outside the grid when $i + j \notin \{0, \dots, N-1\}$, boundary conditions must be imposed. In our implementation, the smoothed CDF is clamped to 0 for indices below 0 and to 1 for indices above $N-1$, consistent with the boundary treatment described in the MCI section.

The problem of Gaussian filtering a noisy CDF approximation, such as filtering $\bar{Q}(\cdot, (x, y), \tau)$, reduces to the appropriate choice of σ that best preserves the underlying shape, while properly filtering noise. We do not have a principled choice for σ, but our empirical experiments show that a choice of σ around 8-12 represents a good balance between filtering noise and preserving underlying shape, especially in

small samples. It is also important to note that the post-filtered distribution approximation no longer has theoretical guarantees with respect to Theorem 4.26, but we can still empirically quantify the distance of the probability transform from the uniform distribution. Our experiments show that, even in large samples of around 1 million data points, the distance (4.28) is below $\frac{3}{\sqrt{n-m+1}}$. It remains an open question for further research whether theoretical guarantees can still be recovered after filtering at the CDF level.

Quantile matching

In the previous chapter, we have seen the notion of a randomized predictive system as the ”gold standard” for constructing conformal predictive distributions. The notion of conformal transducers that are RPS is rich and powerful, as it allows for constructing distributions that can retain a finite-sample marginal validity guarantee, as explained through Theorem 4.15 and Theorem 4.21. However, such constructions do not create a true conditional CDF, as their underlying measure lives on the extended real line $\bar{\mathbb{R}}$ and gives weight at $\pm\infty$. Furthermore, they rely on the inclusion of an additional uniform random variable τ, which yields fuzzy distributions.

Let us, for the sake of the quantile-matching method, remove the randomization and instead approach the problem in a hybrid fashion. We match a set of quantile levels to the output of a one-sided prediction interval, as in Chapter 3, and recompute the predictive distribution for this subset of quantile levels, as in Chapter 4. Under this change, we propose a new framework for constructing predictive densities directly by computing conformal prediction quantiles, which we shall call the *quantile-matching* method. The main upside of this method is that, to obtain prediction densities, the user has an additional layer of freedom in choosing the number of quantile levels needed to extract sufficient information about the underlying shape, while controlling violations of the validity condition in finite samples and preserving asymptotic validity. We can also disregard the Gaussian filtering step for densities if a small enough number of quantile levels is chosen. This allows us to retain, at all times, a principled worst-case-scenario upper bound on the distance from the true uniform distribution.

First, let us adapt the definitions of a randomized predictive system to a predictive system (without randomization). We shall also allow for the validity condition to hold asymptotically.

Predictive systems

Here, we define predictive systems, following the framework presented in [30, Section 2]. Predictive systems exclude randomization, removing one layer of complexity of randomized predictive systems while imposing similar regularity conditions as an RPS. Crucially, the validity guarantee remains virtually unchanged. The advantage of predictive systems is that they construct predictive distributions on the real line, but excluding randomization makes them unsuitable for handling ties in the underlying data, on the one hand, and giving an exact validity coverage in the conformal sense, on the other. The first point does not act as an obstacle in our case, under the standard assumption that the conformity scores are almost surely distinct, as we shall shortly explain in the upcoming section.

Definition 5.3 (Predictive system)**.** A right-continuous function $Q : (\mathbb{R}^{p+1})^{n+1} \to [0,1]$ is called a predictive system if it satisfies:

1. Monotonicity: For any fixed training sequence $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$ and test feature $x_{n+1} \in \mathbb{R}^p$, the function $Q(z_1, \ldots, z_n, (x_{n+1}, y))$ is non-decreasing in y.
2. Boundary stability: For all $(z_1, \ldots, z_n) \in (\mathbb{R}^{p+1})^n$ and $x_{n+1} \in \mathbb{R}^p$,
$$\lim_{y\to-\infty} Q(z_1, \ldots, z_n, (x_{n+1}, y)) = 0 \quad \text{and} \quad \lim_{y\to\infty} Q(z_1, \ldots, z_n, (x_{n+1}, y)) = 1.$$
3. Validity: For any exchangeable sequence of random variables $Z_1, \ldots, Z_{n+1}$ taking values in $\mathbb{R}^{p+1}$, the random variable $Q(Z_1, \ldots, Z_{n+1})$ follows the uniform distribution on $[0,1]$, i.e.
$$\mathbb{P}(Q(Z_1, \ldots, Z_{n+1}) \leq \alpha) = \alpha \quad \text{for all } \alpha \in [0,1]. \tag{5.10}$$

It is called an *asymptotic* predictive system if (5.10) holds asymptotically as $n \to \infty$, i.e., the random variable $Q(Z_1, \ldots, Z_{n+1})$ converges in distribution to the uniform distribution on $[0,1]$ as $n \to \infty$.

Definition 5.4 (Predictive distribution)**.** The output of an (asymptotic) predictive system Q is called an (asymptotic) predictive distribution and is defined as the function

$$Q_n : y \in \mathbb{R} \mapsto Q(z_1, \ldots, z_n, (x_{n+1}, y)).$$

Remark 6. Observe that now, by construction, Q_n is a true cumulative distribution function.

Constructing an asymptotic predictive system via quantile-matching

We now describe an original construction of an asymptotic predictive system, starting from conformal prediction intervals. The procedure is suitable with the three main conformal procedures of Chapter 2, for which Theorem 3.7, Theorem 3.8, and Theorem 3.9 hold, establishing marginal coverage guarantees. To align with the notation of Chapter 4, we assume again that $\mathcal{I}_2 = \{m+1, \ldots, n\}$, so that $|\mathcal{I}_2| = n - m$.

The main idea of our procedure is to slightly modify the previously used conformity scores to obtain one-sided prediction intervals. These essentially act as approximate prediction quantiles in this context, which can be recomputed over a fine grid of quantile levels. The following proposition formalizes this notion. Its proof is essentially identical to that of Theorem 3.9.

Proposition 5.5. *Define the following modified split conformity measures at α-level for split CP, split CQR, and split CDP:*

- $E_{CP}(z_1, \ldots, z_m, (x_{n+1}, y)) = y - \hat{y}(x_{n+1})$.
- $E_{CQR}(z_1, \ldots, z_m, (x_{n+1}, y) = y - \hat{q}_\alpha(x_{n+1})$
- $E_{CDP}(z_1, \ldots, z_m, (x_{n+1}, y)) = y - \hat{y}(x_{n+1}) - \hat{q}^r_\alpha(x_{n+1})$.

Then the prediction interval created by each of the conformity scores computed from either of the conformity measures is one-sided of the type $(-\infty, y^\alpha_{X_{n+1}}]$ and

$$\mathbb{P}\left(Y_{n+1} \leq y^\alpha_{X_{n+1}}\right) \in \left[\alpha, \alpha + \frac{1}{n-m+1}\right].$$

Remark 7. Recalling the algorithms from Chapter 2, we can explicitly give the formula for $y^\alpha_{X_{n+1}}$ for each of the three conformity score, as follows

$$y^{\alpha, CP}_{X_{n+1}} = \hat{y}(X_{n+1}) + \tilde{Q}_\alpha(E, \mathcal{I}_2), \tag{5.11}$$

$$y^{\alpha, CQR}_{X_{n+1}} = \hat{q}_\alpha(X_{n+1}) + \tilde{Q}_\alpha(E, \mathcal{I}_2), \tag{5.12}$$

$$y^{\alpha, CDP}_{X_{n+1}} = \hat{y}(X_{n+1}) + \hat{q}^r_\alpha(X_{n+1}) + \tilde{Q}_\alpha(E, \mathcal{I}_2), \tag{5.13}$$

where the notation is aligned with Algorithms 1, 3 and 4, respectively.

Now suppose that we introduce an evenly-spaced grid $(\alpha_i)_{i=1}^K$ on $[0, 1]$, that is

$$0 < \alpha_1 < \cdots < \alpha_K = 1 \quad \text{with } \alpha_i = \frac{i}{K} \text{ for } i = 1, \ldots, K,$$

with the convention that $\alpha_0 = 0$ and $y^{\alpha_0}_{X_{n+1}} = -\infty$, as well as $y^{\alpha_K}_{X_{n+1}} = +\infty$. Furthermore, to avoid overlap between predicted quantiles, we impose the strict condition $K < n - m + 1$, in view of Proposition 5.5.

We can recompute the predicted quantiles $(y^{\alpha_i}_{X_{n+1}})_{i=1}^K$ using any of the three conformal procedures with the modified conformity scores described above. Then, using Proposition 5.5, we can approximate

$$\mathbb{P}\left(Y_{n+1} \leq y^{\alpha_i}_{X_{n+1}}\right) \approx \alpha_i.$$

Plugged in for an observed test point $X_{n+1} = x_{n+1}$ would translate in the candidate predictive system

$$Q(z_1, \ldots, z_n, (x_{n+1}, y^{\alpha_i}_{x_{n+1}})) = \alpha_i,$$

and we keep it constant at points in between the computed quantile levels. That is, for a general y, we give the candidate asymptotic predictive system

$$Q(z_1, \ldots, z_n, (x_{n+1}, y)) = \inf\{\alpha_i : y \leq y_{x_{n+1}}^{\alpha_i}\}. \tag{5.14}$$

Note, however, that in its current form, the predictive system suffers from a similar problem as in the previous chapter and still does not output a true predictive distribution, as for $y < y_{x_{n+1}}^{\alpha_1}$, we have $Q_n(y) = \alpha_K = \frac{1}{K} > 0$ and so this is not a true CDF. To account for this problem, we apply the following correction.

$$\bar{Q}(z_1, \ldots, z_n, (x_{n+1}, y)) = \begin{cases} \inf\{\alpha_i : y \leq y_{x_{n+1}}^{\alpha_i}\} & \text{if } y > C_{(1)}, \\ 0 & \text{if } y \leq C_{(1)}. \end{cases} \tag{5.15}$$

The correction being done at $C_{(1)}$ makes sense as, by construction, $y_{x_{n+1}}^{\alpha_1} = \tilde{Q}_{\alpha_1}(C, \mathcal{I}_2) \geq C_{(1)}$. This will become clearer in the subsequent subsection. The procedure described above, which leads to (5.15), will be called **quantile-matching** from now on. The quantile-matching procedure is also summarized in Algorithm 7.

The following theorem gives an upper bound on the perturbation from the uniform distribution of the approximated system constructed above. The result is stated below for the standard conformal prediction algorithm. For CQR and CDP, an additional condition on the quantile regression algorithm must be imposed to preserve monotonicity, which will be explained in the remark following the proof of this theorem.

Theorem 5.6. *Suppose standard conformal prediction is used to construct the sequence of quantiles* $(y_{X_{n+1}}^{\alpha_i})_{i=1}^K$. *If* $n - m \to \infty$ *as* $n \to \infty$ *and* $K \to \infty$ *as* $n \to \infty$, *then the construction in* (5.15) *is an asymptotic predictive system. Furthermore, we have the bound*

$$\sup_{u \in (0,1)} \left| \mathbb{P}(\bar{Q}(Z_1, \ldots, Z_n, (X_{n+1}, Y_{n+1})) \leq u) - u \right| \leq \frac{1}{K} + \frac{1}{n - m + 1}. \tag{5.16}$$

Proof. Note that, by construction, the sequence $(\alpha_i)_{i=1}^K$ is increasing, and so is $(y_{x_{n+1}}^{\alpha_i})_{i=1}^K$. This follows directly from the formula $y_{x_{n+1}}^{\alpha_i} = \hat{y}(x_{n+1}) + \tilde{Q}_{\alpha_i}(E, \mathcal{I}_2)$ and the fact that the empirical quantile $\tilde{Q}_\alpha(E, \mathcal{I}_2)$ is increasing in α. Then, $\bar{Q}$ is monotonic in y, as an increase in y can only overcome a superior threshold level $y_{x_{n+1}}^{\alpha_i}$. Throughout this proof, we will denote $\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) := \bar{Q}(Z_1, \ldots, Z_n, (X_{n+1}, Y_{n+1}))$.

The boundary stability condition is satisfied by construction, as for $y < C_{(1)}$, we have

$$\bar{Q}(z_1, \ldots, z_n, (x_{n+1}, y)) = 0,$$

and for $y > y_{x_{n+1}}^{\alpha_{K-1}}$, we have

$$\bar{Q}(z_1, \ldots, z_n, (x_{n+1}, y)) = 1.$$

It remains to check the asymptotic validity property. First, note that, for each $i = 1, \ldots, K$, we have

$$\bar{Q}(z_1, \ldots, z_n, (x_{n+1}, y)) \leq \alpha_i \text{ if and only if } y \leq y_{x_{n+1}}^{\alpha_i}. \tag{5.17}$$

Indeed, if $y \leq y_{x_{n+1}}^{\alpha_i}$, then $\inf\{\alpha_j : y \leq y_{x_{n+1}}^{\alpha_j}\} \leq \alpha_i$, and the left-hand side is exactly the value that $\bar{Q}$ takes on this branch. Vice versa, if $\bar{Q}(\cdot, (x_{n+1}, y)) \leq \alpha_i$, then $\inf\{\alpha_j : y \leq y_{x_{n+1}}^{\alpha_j}\} \leq \alpha_i$, so $y \leq y_{x_{n+1}}^{\alpha_i}$ -otherwise, $y > y_{x_{n+1}}^{\alpha_i}$ would imply that $\inf\{\alpha_j : y \leq y_{x_{n+1}}^{\alpha_j}\} > \alpha_i$. Then, we have, using Proposition 5.5,

$$\mathbb{P}(\bar{Q}(Z_1, \ldots, Z_n, (X_{n+1}, Y_{n+1})) \leq \alpha_i) = \mathbb{P}(Y_{n+1} \leq y_{X_{n+1}}^{\alpha_i}) \in \left[\alpha_i, \alpha_i + \frac{1}{n - m + 1} \right]. \tag{5.18}$$

Therefore, the validity requirement holds at $u \in \{\alpha_1, \ldots, \alpha_K\}$. In fact, the statement is even stronger, that is, we have

$$\left|\mathbb{P}(\bar{Q}(Z_1, \ldots, Z_n, (X_{n+1}, Y_{n+1})) \leq \alpha_i) - \alpha_i\right| \leq \frac{1}{n-m+1} \quad \text{for all } i = 1, \ldots, K. \tag{5.19}$$

Now fix u such that $\alpha_i < u < \alpha_{i+1}$ for some $i = 1, \ldots, K-1$. Then, we have

$$\mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq \alpha_i) \leq \mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq u) \leq \mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq \alpha_{i+1}).$$

Then, using (5.18), we obtain the upper and lower bounds

$$\alpha_i \leq \mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq u) \leq \alpha_{i+1} + \frac{1}{n-m+1}.$$

Subtracting u from both sides yields

$$\alpha_i - u \leq \mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq u) - u \leq \alpha_{i+1} + \frac{1}{n-m+1} - u,$$

and using that $\alpha_i < u < \alpha_{i+1}$, we get

$$\alpha_i - \alpha_{i+1} \leq \mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq u) - u \leq \alpha_{i+1} + \frac{1}{n-m+1} - \alpha_i.$$

Once we take absolute values, we obtain

$$|\mathbb{P}(\bar{Q}(Z_1, \ldots, Z_n, (X_{n+1}, Y_{n+1})) \leq u) - u| \leq \alpha_{i+1} - \alpha_i + \frac{1}{n-m+1} = \frac{1}{K} + \frac{1}{n-m+1}. \tag{5.20}$$

Finally, fix u such that $0 < u < \alpha_1$. In this case, we have, similarly to the steps above

$$\mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq 0) \leq \mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq u) \leq \mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq \alpha_1),$$

which translates, using (5.17) and Proposition 5.5, to

$$0 \leq \mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq u) \leq \alpha_1.$$

Subtracting u from both sides gives

$$-u \leq \mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq u) - u \leq \alpha_1 - u.$$

Upon taking absolute values and using $0 < u < \alpha_1$, we obtain the upper bound

$$|\mathbb{P}(\bar{Q}(\cdot, (X_{n+1}, Y_{n+1})) \leq u) - u| \leq \max\{\alpha_1 - u, u\} \leq \alpha_1 = \frac{1}{K}. \tag{5.21}$$

Combining (5.18), (5.20) and (5.21), we obtain the bound

$$\sup_u \left|\mathbb{P}(\bar{Q}(Z_1, \ldots, Z_n, (X_{n+1}, Y_{n+1})) \leq u) - u\right| \leq \frac{1}{K} + \frac{1}{n-m+1}.$$

If we let $n \to \infty$ and then $K \to \infty$, we get that

$$\sup_u \left|\mathbb{P}(\bar{Q}(Z_1, \ldots, Z_n, (X_{n+1}, Y_{n+1})) \leq u) - u\right| \to 0$$

as $n \to \infty$, which implies that $\bar{Q}(Z_1, \ldots, Z_n, Z_{n+1})$ converges in distribution to Unif$[0, 1]$, and so the validity condition (5.10) holds asymptotically.

Therefore, since all three conditions of Definition 5.3 are satisfied, $\bar{Q}$ is an asymptotic predictive system. □

Remark 8. For the split CQR and split CPD method, the additional condition to ensure that the construction is an asymptotic predictive system is to ensure that the quantile regression algorithm used to produce $\hat{q}_\alpha(x_{n+1})$ is increasing in α for all x_{n+1}. Then, we can ensure that the sequence $(y^{\alpha_i}_{x_{n+1}})_{i=1}^K$ is increasing as well, again using the explicit formula. It then follows that the construction (5.15) is increasing in y, which is the only condition we might violate in the absence of a monotonic in α quantile regression algorithm.

Remark 9. The upper bound (5.16) holds also for Q in (5.14), i.e., without applying tail-correction.

Algorithm 7 Split Quantile-Matched Predictive Distribution

Input: Dataset $\{z_i = (x_i, y_i)\}_{i=1}^n$, observation (test object) x_{n+1},
Algorithm: Partition $\{1, \ldots, n\}$ into a training set $\mathcal{I}_1 = \{1, \ldots, m\}$ and a calibration set $\mathcal{I}_2 = \{m + 1, \ldots, n\}$.
Introduce evenly-spaced grid $(\alpha_i)_{i=1}^K$ with $\alpha_i = \frac{i}{K}$ for all $i = 1, \ldots, K$
for $i \in \{1, \ldots, K\}$ **do**
Compute conformal prediction quantiles $y_{x_{n+1}}^{\alpha_i}$.
end for
Output: Return a predictive distribution for the label y of x_{n+1}

$$Q_n(y) := \begin{cases} \inf\{\alpha_i : y \leq y_{x_{n+1}}^{\alpha_i}\} & \text{if } y > C_{(1)}, \\ 0 & \text{if } y \leq C_{(1)}. \end{cases} \tag{5.22}$$

Relation with conformal predictive distributions

First, let us introduce the *crisp* modifications of randomized predictive distributions [39, Section 5].

$$Q_{crisp}(z_1, \ldots, z_n, (x_{n+1}, y)) := \frac{i}{n-m} \quad \text{if } y \in (C_{(i)}, C_{(i+1)}]. \tag{5.23}$$

The crisp modification no longer depends on τ and essentially acts as the *empirical CDF* of the conformal atoms C_i. Vovk et al. [39] introduce crisp modifications of randomized predictive distributions in a heuristic fashion to remove the fuzziness of their randomized counterparts and compute the continuous ranked probability score.

Let us now recall the formulation of the empirical quantiles and the connection between conformal atoms and conformity scores highlighted in (4.14), $C_i = E_i + \hat{y}(x_{n+1})$. Then, for instance using the explicit formula for standard conformal prediction quantiles (5.11) and using that the $\alpha(1 + \frac{1}{n-m})$ empirical quantile $\tilde{Q}_\alpha(E, \mathcal{I}_2)$ of conformity scores is shifted by the fixed $\hat{y}(x_{n+1})$, the quantile rewrites to

$$y_{x_{n+1}}^{\alpha_i} = \tilde{Q}_{\alpha_i}(C, \mathcal{I}_2). \tag{5.24}$$

That is, with quantile matching, we preselect a group of quantile levels we are most interested in, and subsequently compute the empirical quantiles of the conformal atoms at these levels. But the empirical quantiles, by (3.2), have the closed form $\tilde{Q}_{\alpha_i}(C, \mathcal{I}_2) = C_{(\lceil (n-m)\alpha_i + \alpha_i \rceil)}$. Therefore, the procedure essentially amounts to preselecting a subset of the conformal atoms to be matched with each quantile level.

In contrast, the conformal predictive distribution of Chapter 4 also essentially acts as an empirical CDF of the conformal atoms up to including randomization. Furthermore, when $K = n - m$ (which is also the largest admissible K), the grid induced is simply $\alpha_i = \frac{i}{n-m}$, and the construction through quantile-matching (5.15) yields the formula (5.23) for the *crisp* predictive distribution. Consequently, we can state the following corollary.

Corollary 5.6.1. *Let Q_{crisp} be the crisp modification of randomized predictive distributions as defined in* (5.23)*. Then, the following upper bound holds*

$$\sup_{u \in (0,1)} |\mathbb{P}(Q_{crisp}(Z_1, \ldots, Z_n, (X_{n+1}, Y_{n+1})) \leq u) - u| \leq \frac{1}{n-m} + \frac{1}{n-m+1}. \tag{5.25}$$

Proof. In Theorem 5.6, take $K = n - m$. Then, we have the closed-form $y_{X_{n+1}}^{\alpha_i} = C_{(i+1)}$ and a quick check ensures that the constructed predictive system in (5.15) matches with the crisp modificification of (5.23). □

We consider this corollary one of the most important contributions of our thesis. The result brings the connection between randomized predictive distributions, conformal prediction intervals, and quantile-matching on predictive systems full circle. In the original paper, the authors concede that crisp modifications do not satisfy the usual uniform property of a conformal predictive distribution ("they, however, do not satisfy any validity properties" [39, Page 10]). Here, we provide a sharp estimate of how much uniformity has been lost with the crisp modifications via this corollary. It is also noteworthy that the crisp version arises simply by setting $K = n - m$ in the quantile-matching construction.

A different coverage notion

Let us briefly highlight and compare the coverage notions of predictive systems and randomized predictive systems. Namely, note that the validity guarantee (5.10) excludes randomization, that is, it is taken marginally over the randomness of $Z_1, \ldots, Z_{n+1}$, similarly to the notion of marginal coverage guarantee of prediction intervals (3.1). The validity guarantee of RPS is taken marginally over the randomness of $Z_1, \ldots, Z_{n+1}$ *and* an auxiliary uniform random variable τ. In this sense, the validity guarantee for predictive systems is stronger, as it marginalizes over one less random variable. However, note that neither simple nor randomized predictive systems guarantee a *training conditional* coverage guarantee (see the discussion at the end of Chapter 2) by design, and there is little research to see whether certain procedures give such a training conditional coverage guarantee. Similarly to conformal prediction intervals, it is unachievable to obtain a conditional validity guarantee for a certain procedure, as this is not possible in the simplified case of prediction intervals (of any given fixed length), let alone in the distributional sense.

Finite differencing via quantile-matching

Given that the user has an additional layer of freedom in choosing the number of quantiles used, as well as their levels, the trade-off between a possibly larger deviation from the uniform distribution and greater flexibility appears to become more manageable. Indeed, the distribution constructed in the *quantile-matching step* benefits from allowing the users to select both the number of quantiles to be estimated through conformal prediction *and* the levels at which they are evaluated. For instance, certain practitioners might be interested in more extreme scenarios, and one could, in principle, use a non-uniform grid, concentrating more mass where precision is needed.

For the quantile-matching method, since the user sets the quantiles, one can use a forward finite difference scheme to approximate the density at a given quantile level. The following density approximation arises

$$\hat{f}_{QM}(\cdot, (x, y_x^{\alpha_i})) = \frac{\bar{Q}(\cdot, (x, y_x^{\alpha_{i+1}})) - \bar{Q}(\cdot, (x, y_x^{\alpha_i}))}{y_x^{\alpha_{i+1}} - y_x^{\alpha_i}} = \frac{\alpha_{i+1} - \alpha_i}{y_x^{\alpha_{i+1}} - y_x^{\alpha_i}}, \tag{5.26}$$

which, under an evenly spaced grid, translates to

$$\hat{f}_{QM}(\cdot, (x, y_x^{\alpha_i})) = \frac{\bar{Q}(\cdot, (x, y_x^{\alpha_{i+1}})) - \bar{Q}(\cdot, (x, y_x^{\alpha_i}))}{y_x^{\alpha_{i+1}} - y_x^{\alpha_i}} = \frac{1}{K(y_x^{\alpha_{i+1}} - y_x^{\alpha_i})}.$$

Then, once again, simple linear interpolation is applied to fill in the values between atoms.

$$\hat{f}_{QM}(\cdot, (x, y)) = \hat{f}_{QM}(\cdot, (x, y_x^{\alpha_i})) + (y - y_x^{\alpha_i}) \frac{\hat{f}_{QM}(\cdot, (x, y_x^{\alpha_{i+1}})) - \hat{f}_{QM}(\cdot, (x, y_x^{\alpha_i}))}{y_x^{\alpha_{i+1}} - y_x^{\alpha_i}}.$$

Using fewer quantiles removes the very noisy patterns at the density level, as we shall shortly see in the following chapter. The user is responsible for selecting the appropriate number of quantiles, depending on their tolerance for perturbation of the Probability Integral Transform, in view of Theorem 5.6.

Examples

In this section, we explore a simple model on which we apply monotonic cubic interpolation, Gaussian filtering, and the quantile-matching method to illustrate the following three postulates. Firstly, monotonic cubic interpolation is largely unjustified for large sample sizes, as it does not yield better predictive

densities than finite differencing. Secondly, for visualization purposes, Gaussian filtering is quite precise and manages to recover key features of the underlying density even from very small samples. Lastly, we show that indeed, the quantile-matched distribution approaches the crisp modification of conformal predictive distributions as the number of quantile levels approaches the number of calibration points.

Monotonic cubic interpolation and filtering

To illustrate the use of monotonic cubic interpolation and Gaussian filtering, we use the following simple example. Suppose we have only one feature represented by the random variable X and we want to predict the target variable Y. The model is governed by the following law.

$$\begin{aligned} X &\sim N(2,1) \\ Y &= 2X + 3 + \varepsilon. \end{aligned} \tag{5.27}$$

Then, conditional on $X = x$, we have

$$Y \mid X = x \overset{d}{=} 2x + 3 + \varepsilon.$$

The conditional distribution is then entirely determined by the underlying distribution of ε, and it is, in fact, just a shifted version of ε. For our experiments, we will use the following three distributions for ε:

$$\begin{aligned} &\varepsilon \sim N(0,1), \\ &\varepsilon \sim \frac{1}{2}N(-2, 0.6^2) + \frac{1}{2}N(2, 0.6^2), \\ &\varepsilon \sim \mathrm{Rayleigh}(2). \end{aligned}$$

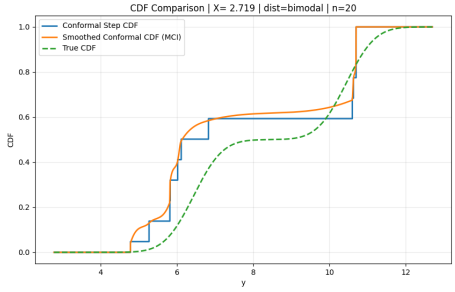
These three distributions are used to illustrate different shapes of the density function. The normal distribution is symmetric, unimodal, and has the real line as its support. The mixture distribution is bimodal, and the Rayleigh distribution is right-tailed skewed with right tails decaying faster than the normal distribution. Additionally, the Rayleigh distribution is supported only on the positive real line.

For our experiments, we shall use n independent and identically distributed samples $(x_i, y_i)_{i=1}^n$ from the model (5.27), which we split $50 : 50$ between training and calibration. Evidently, the underlying model is treated as unknown, and we use a linear regression algorithm as the prediction algorithm $\mathcal{A}$, as well as a LightGBM model [19] for the quantile regression algorithm $\mathcal{B}$. Then, one more x_{n+1} is sampled independently from the $N(2,1)$ distribution and used to compute the conformal predictive distribution and subsequently obtain the predictive density. As a conformity score for calibration, we use the modified CDP score defined in (4.11). We have experimented with the other monotonic and balanced conformity scores suggested in Chapter 4 (stemming from the standard conformal prediction method and conformalized quantile regression), and the differences in the predictive distributions are minuscule for our simple example.

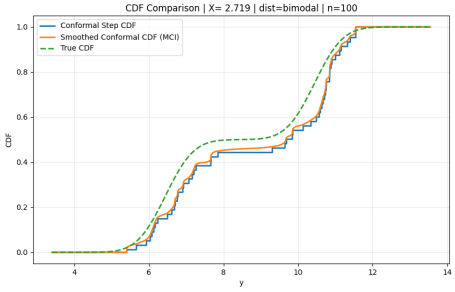
In our experiments, we use $20, 100, 1000, 10000, 100000$, and 1 million samples for n. In the following figures, the true conditional CDF at a random test point is compared with the tail-corrected conformal distribution and its MCI-smoothed counterpart for different values of n and the three proposed distributions.

The difference between the true and conformal CDF, with or without smoothing, seems to vanish with increasing sample size. One can already observe that a relatively small sample size $(n = 1000)$ is sufficient to capture the general shape of the underlying distribution. In Figure 5.1, one can see this phenomenon for the bimodal distribution. The CDF comparison for the normal and Rayleigh distribution can be consulted in Figures A.1 and A.2 in Appendix A. Even with relatively few samples ($n = 1000$), the practical use of monotonic cubic interpolation becomes questionable, as the step function becomes very jagged, and this symptom is not treated by interpolating between conformal atoms. However, Gaussian filtering appears to be highly advantageous for visualization.

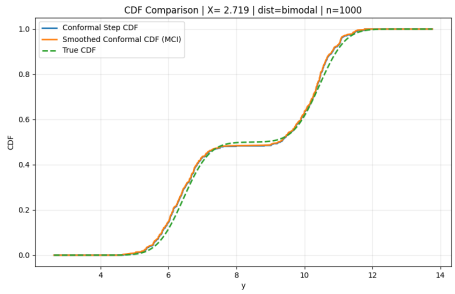
Upon retrieving the predictive density, we indeed observe the expected noisy behavior of the unfiltered version. However, the Gaussian-filtered predictive density is surprisingly accurate for the bimodal distribution, even with small sample sizes. For instance, we can observe that even with 20 sample points



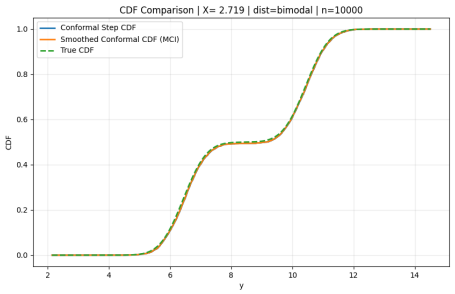
(a) $n = 20$



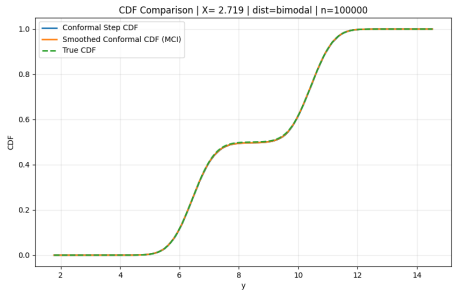
(b) $n = 100$



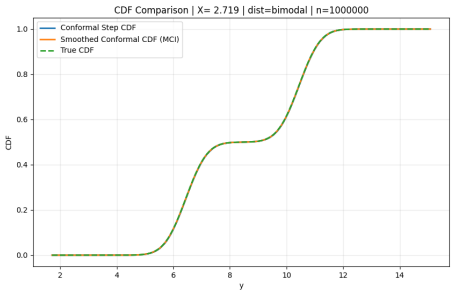
(c) $n = 1,000$



(d) $n = 10,000$



(e) $n = 100,000$



(f) $n = 1,000,000$

Figure 5.1: CDF comparison for the bimodal distribution across different sample sizes

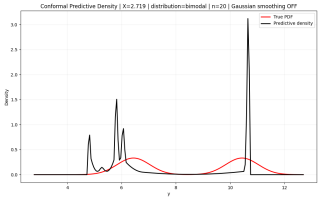
(and a 50:50 split), the predictive density retrieves key features of the underlying shape, such as bimodality. The filtered predictive density becomes more accurate as the sample size increases, to the point of being virtually indistinguishable at one million samples. Note also that, even though the unfiltered PDF remains noisy throughout, there is a decrease in the height of the jumps with an increase in sample size. The results for the bimodal distribution are summarized in Figure 5.2, whereas the results for the normal and Rayleigh can be consulted in Figures A.3-A.4 in Appendix A.

Quantile-matching and the crisp modification

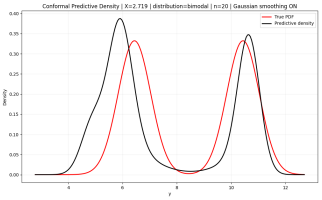
With a sample size of $n = 200$ (so 100 calibration points) using the same simple model as above, we showcase how the quantile-matching procedure approaches the crisp modification of conformal predictive distributions (5.23) as K goes to $n - m$. In Figure 5.3, we observe this precise behavior, where we use $K = 25$, $K = 50$, and $K = 100$ respectively. Indeed, the error between the quantile-matching-induced distributions and the crisp modifications decreases as K approaches 100. When $K = 100$, there is a perfect overlap between the two distributions, as expected. As before, here we give the results for the bimodal distribution, whereas those for the normal and Rayleigh distributions can be consulted in Appendix A.

In the next chapter, we apply the proposed methodologies on a large simulated real estate transaction dataset, where we compare them based on different performance metrics.

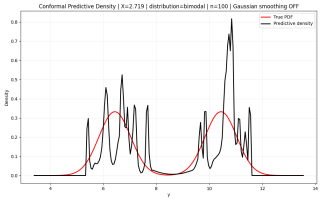
Unfiltered **Filtered**



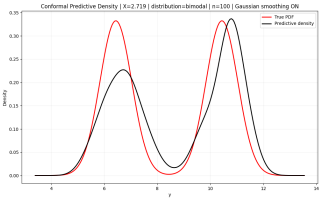
(a) $n = 20$



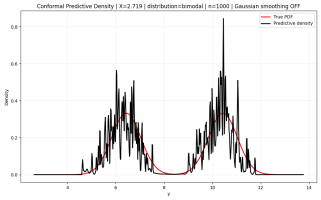
(b) $n = 20$



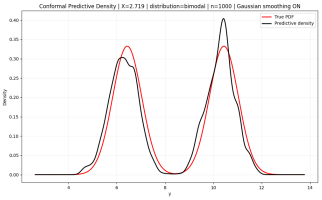
(c) $n = 100$



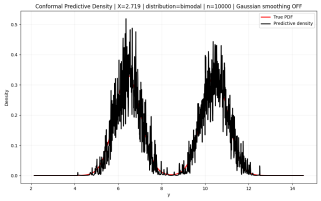
(d) $n = 100$



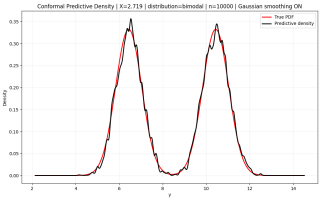
(e) $n = 1000$



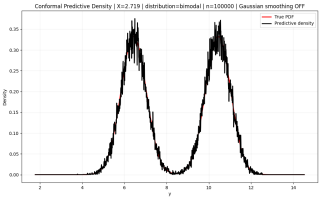
(f) $n = 1000$



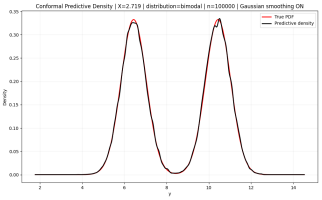
(g) $n = 10{,}000$



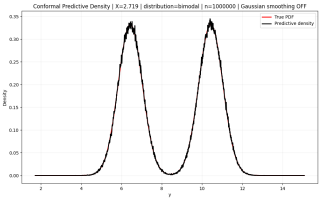
(h) $n = 10{,}000$



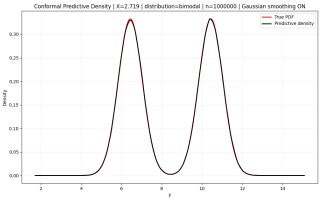
(i) $n = 100{,}000$



(j) $n = 100{,}000$

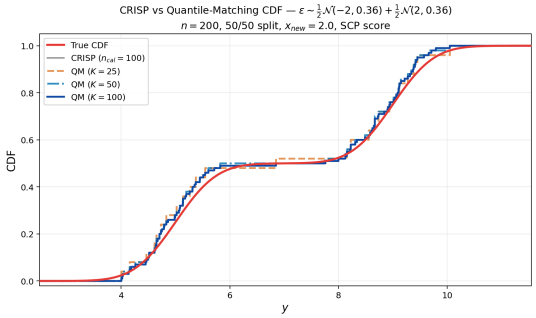


(k) $n = 1{,}000{,}000$

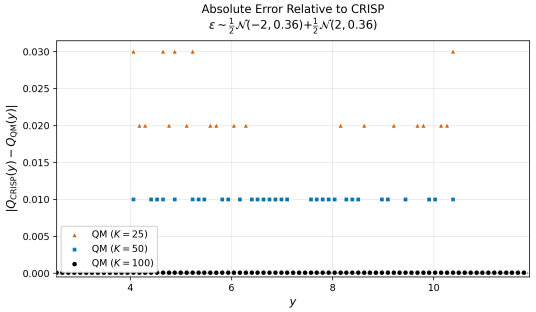


(l) $n = 1{,}000{,}000$

Figure 5.2: Unfiltered and Gaussian-filtered PDFs for the bimodal distribution across sample sizes $n \in \{20,\ 100,\ 1000,\ 10^4,\ 10^5,\ 10^6\}$.



(a) Comparison of CDFs between quantile-matching, crisp, and true CDF



(b) Absolute error at conformal atoms between quantile-matching and crisp distributions

Figure 5.3: Comparison of quantile-matching and crisp distributions with $n = 200$.

6

Applications

The use of predictive distributions and, in certain contexts, predictive densities is of importance to industry practitioners. Having a principled, model-agnostic way to return a predictive distribution is relevant in many financial applications, where minimal parametric assumptions can be made about either the features or the target variable.

The extension to predictive densities, which has been an important research direction in the present work, is interesting in three main ways. Firstly, a qualitative analysis of the features of the underlying conditional distribution, such as shape, modality, skewness or kurtosis, can really be only visualized through a density function. Secondly, computing conditional expectations of functions of the target variable requires knowledge of the probability density function. Thirdly, and perhaps most relevant to practitioners, the likelihood function also requires knowledge of a density function. At financial institutions such as Ortec Finance, parameter recalibration is often used to update a model, and knowledge of the underlying density is relevant in this context.

In the present chapter, we present an important comparison between the three main approaches to retrieve predictive densities: directly finite differencing (or computing the derivative in closed-form via MCI), applying Gaussian filtering on the noisy CDF, or the quantile-matching procedure, where we preselect a number of quantile levels (e.g. 100) and build a distribution based on these values. We empirically showcase various peculiarities and compare the methods' performance on a large dataset of simulated housing prices with relevant features, which resembles the true data used in the Department of Real Estate Valuation at Ortec Finance. The model used to simulate prices is detailed in the following section and is the same as the one used in [12]. We note that throughout this chapter, our target random variable Y represents transaction prices of houses *after* a $\log$ transformation.

The Hierarchical Trend Model

The transaction prices are simulated using the Hierarchical Trend Model (HTM) [14, 13],

$$\mathbf{y}_t \sim \mathcal{N}\big(\mathbb{1}_{n_t}\mu_t + D_{\boldsymbol{\lambda},t}\boldsymbol{\lambda}_t + D_{\boldsymbol{\theta},t}\boldsymbol{\theta}_t + X_t\beta + D_{\eta,t}\eta + \varepsilon_t,\ \sigma_\varepsilon^2 I_{n_t}\big)\,,$$

$$\Delta\mu_t \sim \mathcal{N}\left(\rho\Delta\mu_{t-1} + \alpha(1-\rho),\ \sigma_\mu^2\right), \quad \Delta\mu_1 \sim \mathcal{N}\left(\alpha,\ \frac{\sigma_\mu^2}{1-\rho^2}\right), \quad \mu_1 = 0,$$

$$\boldsymbol{\lambda}_{t+1} \sim \mathcal{N}\left(\boldsymbol{\lambda}_t,\ \sigma_{\boldsymbol{\lambda}}^2 I_{n_t}\right),$$

$$\boldsymbol{\theta}_{t+1} \sim \mathcal{N}\left(\boldsymbol{\theta}_t,\ \sigma_{\boldsymbol{\theta}}^2 I_{n_t}\right),$$

$$\boldsymbol{\eta} \sim \mathcal{N}\left(0, \sigma_{\boldsymbol{\eta}}^2\right).$$

Here, $\mathbf{y}_t$ is a $n_t \times 1$ vector of $\log$ prices, and t indicates time measured in quarters of a year. The one-dimensional variable μ_t represents the log value of some common index. The index return $\Delta\mu_t$ follows

the AR(1) model (as detailed at the end of Chapter 2), with lag coefficient ρ and unconditional mean α. The vectors $\boldsymbol{\lambda}_t$ and $\boldsymbol{\theta}_t$ represent a collection of log indexes, specified as random walks, for different regions and house types, in deviation from the common log index μ_t. The property characteristics are stored in X, such as floor area, lot area, and year of construction, with corresponding coefficients β. The matrices $D_{\boldsymbol{\lambda},t}$, $D_{\boldsymbol{\theta},t}$, and $D_{\eta,t}$ are selection matrices containing zeros and ones which are used to select the appropriate region, house type, and neighborhood for each transaction, respectively. Finally, the vector $\boldsymbol{\eta}$ contains neighborhood random effects.

As in [12], the posterior predictive distribution for each transaction price in the train and test set is constructed, and for each transaction, 4000 samples are simulated, and a random one is used as the simulated transaction price to be tested y_{n_t+1}. Note that the conditional distribution of transactions is still normal, conditional on all features. Then, a true underlying distribution is still known and can serve as a benchmark for comparing different approaches.In our experiments, we found that monotonic cubic interpolation does not yield a significant improvement; therefore, we decided to exclude it from the results presented.

To train our model, unless mentioned otherwise, the dataset is split 65/25/10, that is 65% of the dataset is used for training, on which we use a LightGBM model [19] for the point prediction and the quantile prediction (if necessary), 25% is used for calibration and 10% is made available for testing, out of which we draw transaction prices at random. This translates to training on 715111 data points, calibrating on 275043 data points, and having 110018 data points available for testing.

Mean Integrated Squared Error Comparison

We compare the performance of the three predictive densities using the Mean Integrated Squared Error formula (2.6). For each of the three approaches, we used the CDP score (4.11), the CQR score (4.10), and the standard CP score (4.9), which we shorten *SCP* here. Our results show small MISE variations among the three methods, with significantly worse performance when using the CQR score. The differences between the SCP score and the CDP score are very small (they are the same up to the third decimal point), which suggests there is no significant advantage in choosing CDP over SCP, whereas choosing the standard conformity score can reduce computation time significantly, as we do not require the retraining of a quantile regression algorithm at various levels anymore. Therefore, we recommend using the standard conformity score. Among the three approaches, Gaussian filtering has a slightly smaller MISE than the other two. Compared to direct finite-differencing, the average MISE is similar to that of quantile-matching, making this method preferable because it reduces fuzziness. Compared to Gaussian filtering, which has no theoretical guarantee, quantile-matching can also be preferable when we are interested in controlling the perturbation in Probability Integral Transform.

Table 6.1: Average MISE across twenty random transaction prices

Method	Score Type	Avg. MISE
QM (K=100)	CDP	0.159300
	CQR	0.218269
	SCP	0.159538
Direct FD	CDP	0.159533
	CQR	0.218332
	SCP	0.159629
Gaussian Filtering	CDP	0.157770
	CQR	0.217019
	SCP	0.157954

Figures

Figures 6.1-6.3 showcase the conformal predictive distribution compared to the true distribution for a random transaction price with given characteristics, and its induced predictive density, which is retrieved via either direct finite differencing or applying Gaussian filtering. We use either the standard 65/25/10 split, 10000, or 1000 calibration points, and adjust the size of the training set accordingly, to showcase

the differences. The red dotted line represents the point prediction $\hat{y}$ made, whereas the grey dotted line represents the true sampled value from the HTM model [14].

We observe that the height of jumps increases when we use fewer calibration points, and a Gaussian kernel with a fixed variance cannot keep up with the fuzziness with fewer calibration points, retaining some of the roughness with $\sigma = 8$. As a consequence of the marginal validity guarantee of conformal predictive systems, which can be associated with a vision of being accurate ”on average” across all possible observations, the predictive distribution here, although it retains some characteristics of the underlying distributions, does not actually approach the true distribution. Depending on the target transaction price and features, there might be a significant shift away from the true distribution, such as the one visible in Figures 6.1-6.3. For other test points, however, the distribution can be significantly more accurate, and such figures can be consulted in Appendix B.

Alternatively, Figures 6.4-6.6 show the quantile-matched predictive distribution with $K = 100$ and the same calibration sizes as for conformal predictive distributions. It successfully reduces the fuzziness of the conformal counterparts across all calibration sizes, particularly for the large 65-25-10 split. For comparison, Figures 6.7-6.9 apply the quantile-matching with $K = 500$. In these figures, we observe the perturbation-fuzziness tradeoff. Increasing the number of selected quantiles produces distributions that are marginally closer to the true uniform distribution in PIT, but this also increases the fluctuations of the associated predictive density.

Additionally, a Gaussian filtering step can be applied on top of the quantile-matched distribution, whose induced density is the smoothest out of all variants, and which produces the densities with the lowest MISE. They, however, have no theoretical finite-sample validity properties.

Diagnostics

A common and relevant diagnostic is the so-called *calibration* check, introduced in [39]. We compare the conformal predictive distribution values $Q(\cdot, (x, C_{(i)})$ at the conformal atoms $C_{(i)}$, against a uniform grid on $[0, 1]$. If the distribution is almost perfectly calibrated, we should expect these points to lie close to the diagonal.

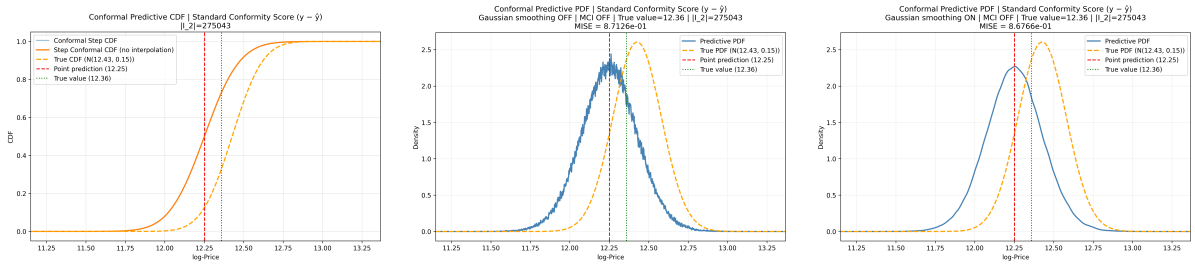
Additionally, one can quantify precisely the distance from the diagonal at these points. Up to scaling to a certain threshold, we observe empirically that, over 1000 simulated house prices, the filtered CDF lies at a distance below $\frac{3}{\sqrt{n-m+1}}$ (but above $3/(n-m+1)$). This indicates that, indeed, filtering damages the uniform validity at a faster rate than simple monotonic cubic interpolation (which is within $\frac{3}{n-m+1}$ as per Proposition 5.2), but remains within an acceptable distance from the true uniform, even in large calibration sizes (our dataset is over one million samples). The calibration and distance check for a single random transaction price using 1000 calibration points are shown in Figures 6.10 and 6.11.

For the quantile-matched distribution, we additionally evaluate the distance at the midpoints between estimated quantile levels, as in light of Theorem 5.6, the method is almost perfectly calibrated at estimated quantile levels, and the error increases in between these levels up to at most $\frac{1}{K} + \frac{1}{n-m+1}$. This is exactly the behavior we observe in Figures 6.12 and 6.13.

Scoring rules for forecasting evaluation

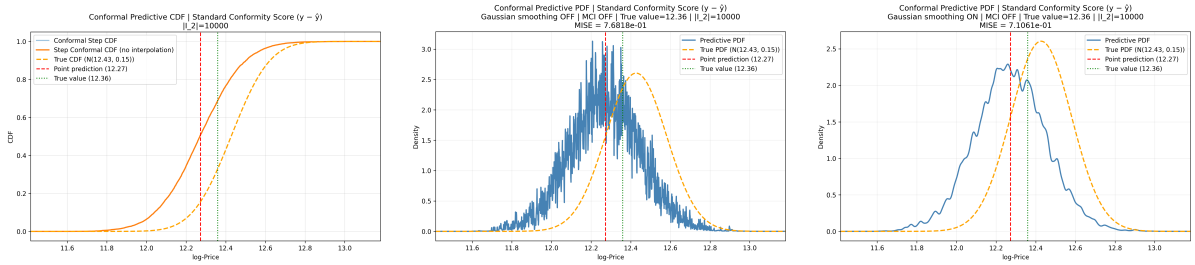
Apart from calibrating our predictive distribution, which we evaluate using the Probability Integral Transform, we are also interested in the sharpness of our predictions, which we assess by the concentration of the predictive densities around the true realized value.

To assess the calibration and sharpness of our predictive distributions and, subsequently, predictive densities, one introduces proper scoring rules [16] to assess these qualities. The propriety property is essential in ensuring that forecasters provide honest and careful quotes [17]. Here, we briefly discuss proper scoring rules and highlight those most relevant in our context. The main reference point for the following subsection is [16].



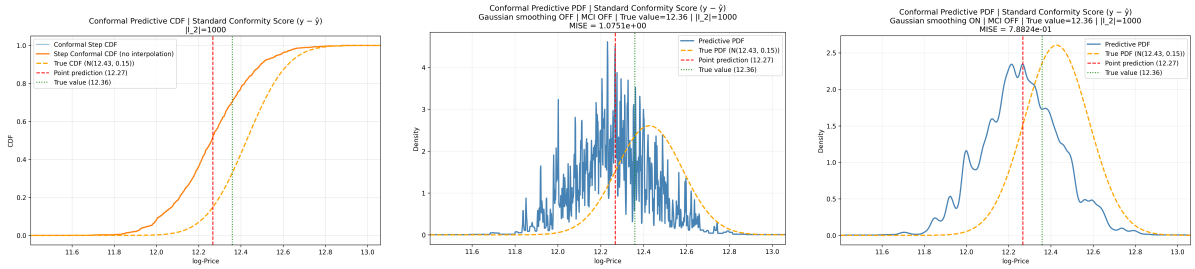
(a) Predictive distribution **(b)** PDF via finite differencing **(c)** PDF via Gaussian filtering

Figure 6.1: Conformal predictive distributions using a 65/25/10 train–calibration–test split and the SCP score.



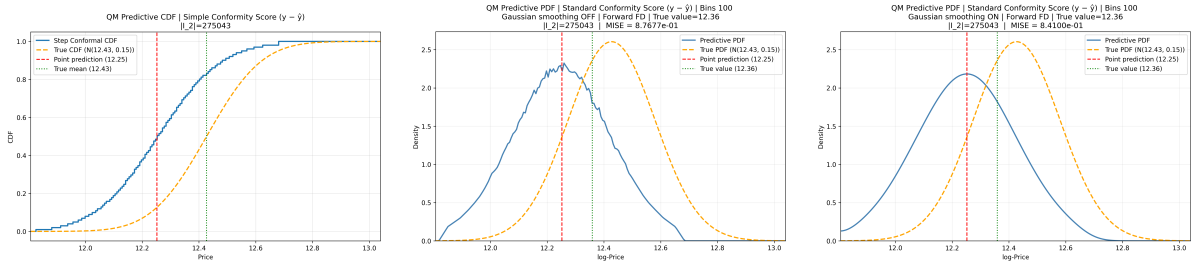
(a) Predictive distribution **(b)** PDF via finite differencing **(c)** PDF via Gaussian filtering

Figure 6.2: Conformal predictive distributions using 10000 calibration points and the SCP score.



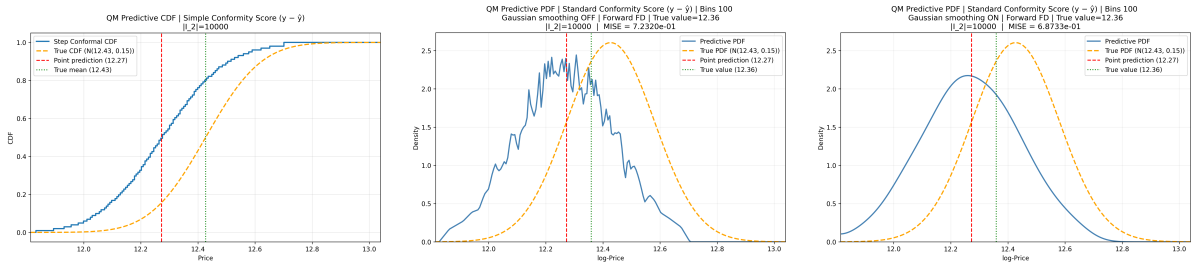
(a) Predictive distribution **(b)** PDF via finite differencing **(c)** PDF via Gaussian filtering

Figure 6.3: Conformal predictive distributions using 1000 calibration points and the SCP score.



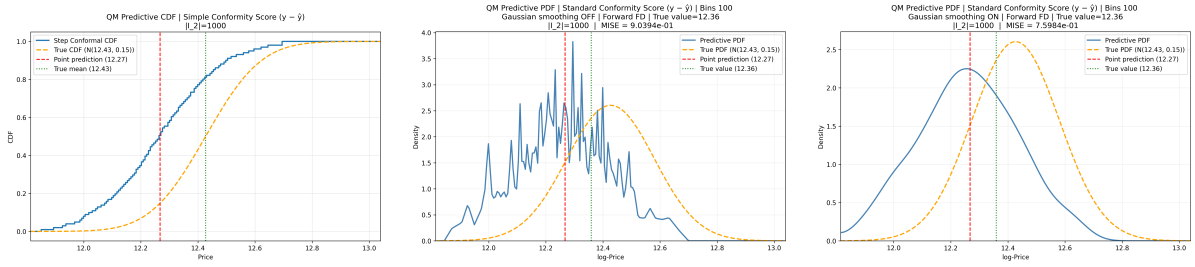
(a) Predictive distribution **(b)** PDF via finite differencing **(c)** PDF via Gaussian filtering

Figure 6.4: Quantile-matched predictive distributions using 100 quantile levels, a 65/25/10 train–calibration–test split, and the SCP score.



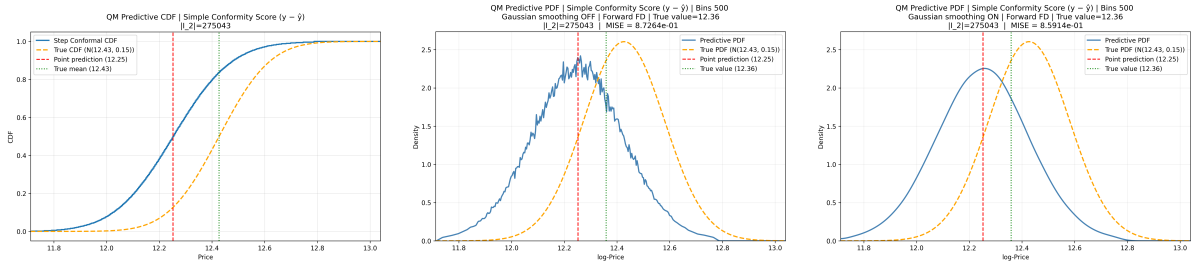
(a) Predictive distribution **(b)** PDF via finite differencing **(c)** PDF via Gaussian filtering

Figure 6.5: Quantile-matched predictive distributions using 100 quantile levels, 10000 calibration points, and the SCP score.



(a) Predictive distribution **(b)** PDF via finite differencing **(c)** PDF via Gaussian filtering

Figure 6.6: Quantile-matched predictive distributions using 100 quantile levels, 1000 calibration points, and the SCP score.

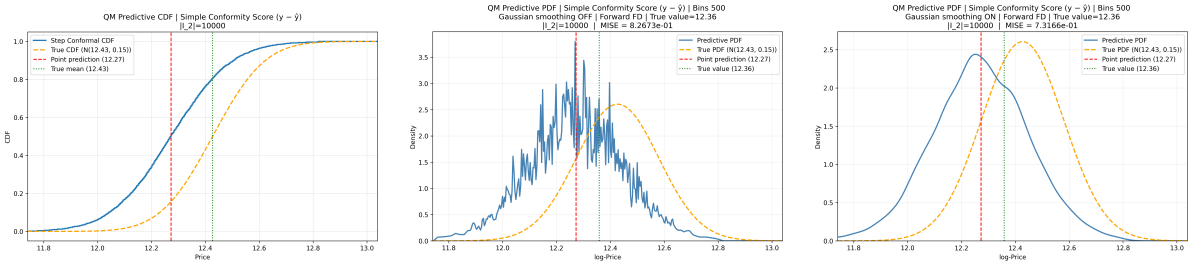


(a) Predictive distribution **(b)** PDF via finite differencing **(c)** PDF via Gaussian filtering

Figure 6.7: Quantile-matched predictive distributions using 500 quantile levels, a 65/25/10 train–calibration–test split, and the SCP score.

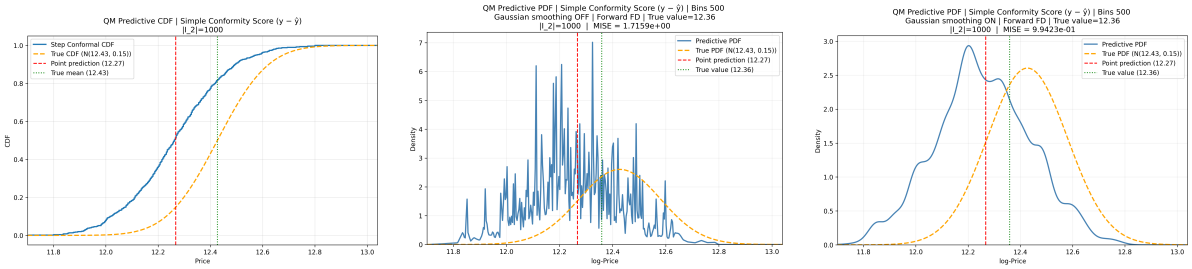
Proper scoring rules

Let $\mathcal{F}$ be a generic convex class of probability distributions on $\mathbb{R}$, identified with their respective CDF and PDF. A scoring rule assigns a numerical score $S(F, y)$ to every (F, y), where $F \in \mathcal{F}$ is a probabilistic forecast, that is, a predictive distribution, and $y \in \mathbb{R}$ is the actual realized value.



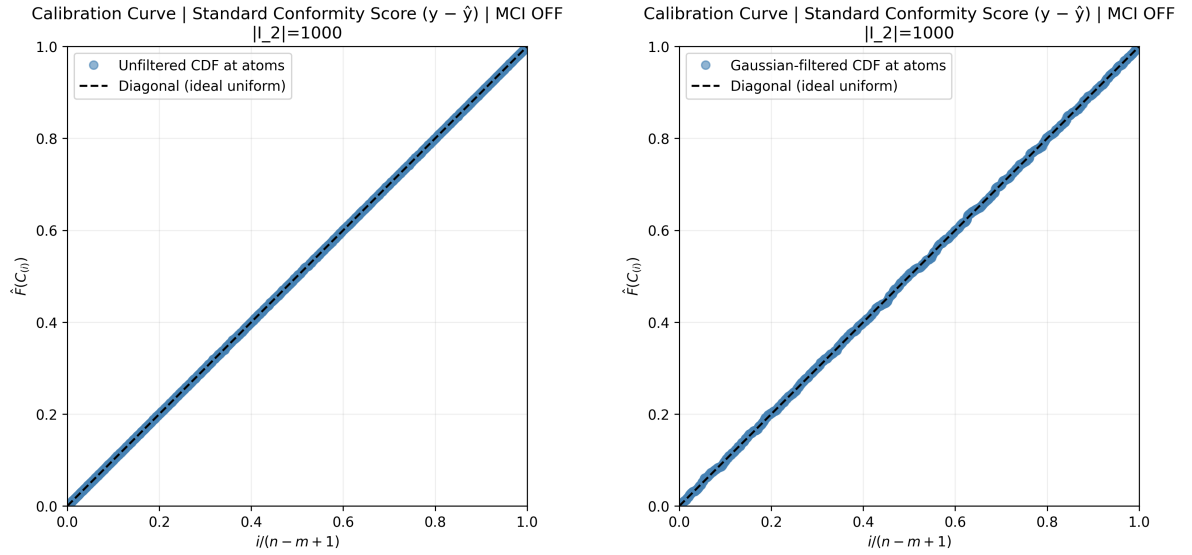
(a) Predictive distribution **(b)** PDF via finite differencing **(c)** PDF via Gaussian filtering

Figure 6.8: Quantile-matched predictive distributions using 500 quantile levels, 10000 calibration points, and the SCP score.



(a) Predictive distribution **(b)** PDF via finite differencing **(c)** PDF via Gaussian filtering

Figure 6.9: Quantile-matched predictive distributions using 500 quantile levels, 1000 calibration points, and the SCP score.



(a) Calibration curve for the conformal predictive distribution **(b)** Calibration curve for Gaussian filtered distribution

Figure 6.10: Calibration curves for the conformal predictive distribution with 1000 calibration points of a random transaction price

Definition 6.1 (Proper scoring rule)**.** The scoring rule $S : \mathcal{F} \times \mathbb{R} \to \bar{\mathbb{R}}$ is *proper* relative to $\mathcal{F}$ if

$$\mathbb{E}_F(S(F,Y)) \leq \mathbb{E}_G(S(F,Y)) \quad \text{for all } F, G \in \mathcal{F}. \tag{6.1}$$

It is called *strictly proper* if the above holds with equality if and only if $F = G$.

From the above definition, we understand that a proper scoring rule is one such that using the true distribution as the predictive distribution is optimal in expectation. One can state the exact necessary and sufficient conditions for a scoring function to be proper [16, Theorem 3], but this is not of interest to us in this case. Alternatively, the above definition can be relaxed to *locally proper scoring rules*.

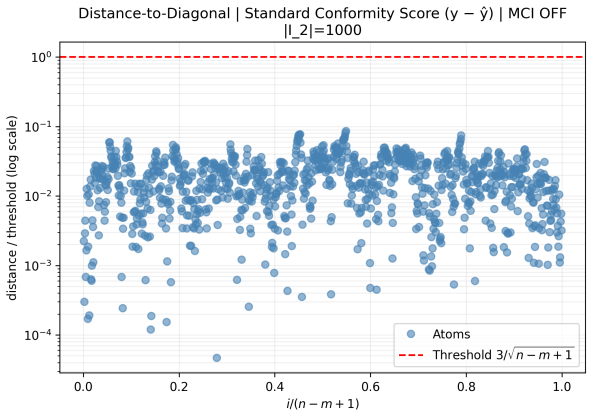


Figure 6.11: Absolute error of Gaussian-filtered predictive distribution versus $\mathrm{Unif}[0, 1]$

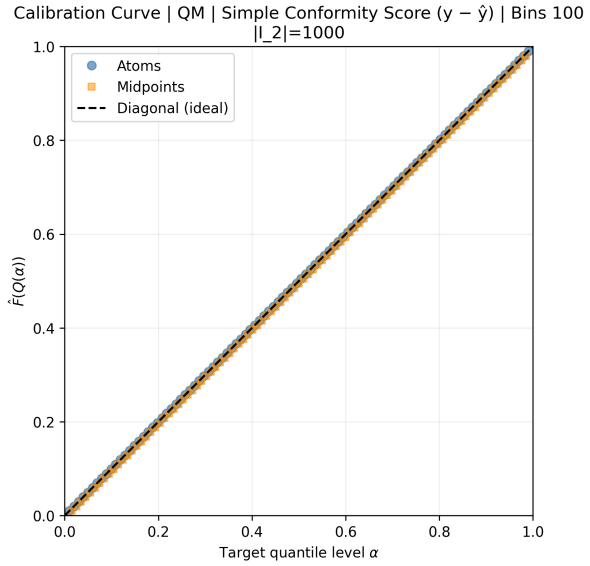


Figure 6.12: Calibration curve for quantile-matched predictive distribution with 1000 calibration points

Definition 6.2 (Locally proper scoring rule)**.** Suppose the convex class $\mathcal{F}$ contains only probability measures on $\mathbb{R}$ that admit a probability density f with continuous derivatives up to order k. Then, S is *locally proper of order* k if there exists a function $s : \mathbb{R}^{2+k} \to \bar{\mathbb{R}}$ such that

$$S(f, y) = s(y, f(y), f'(y), \ldots, f^{(k)}(y)) \quad \text{for all densities } f \in \mathcal{F} \text{ and } y \in \mathbb{R}. \tag{6.2}$$

Three common (locally) proper scoring rules arise in the literature to evaluate prediction forecasts.

Definition 6.3. (Logarithmic score) The *logarithmic score* (LS) is a locally proper scoring rule of order 0, defined by

$$\mathrm{LS}(f, y) = -\log f(y). \tag{6.3}$$

The logarithmic score measures how much density has been assigned to the actual realized value, and a smaller score is considered more favorable. However, this score does not account for the overall shape of the density and will always reward densities that are very concentrated at the actual realization. As an alternative that penalizes overconfidence in a single value, the quadratic score is introduced.

Definition 6.4 (Quadratic score)**.** The *quadratic score* (QS) is a proper scoring rule defined by

$$\mathrm{QS}(f, y) = -2f(y) + \int f^2(z)dz. \tag{6.4}$$

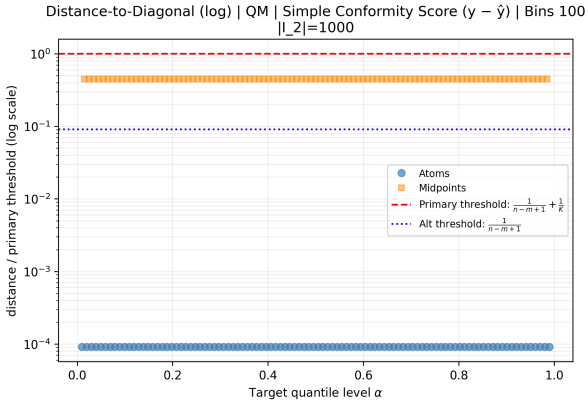


Figure 6.13: Absolute error of quantile-matched distribution versus $\mathrm{Unif}[0,1]$ with 100 calibration points

The Quadratic score again rewards those densities that put higher mass at the actual realized value, while punishing densities that are too concentrated around this value via the integral term.

One can also compare the quality of the associated predictive distributions via the continuous ranked probability score (CRPS).

Definition 6.5 (Continuous ranked probability score)**.** The *continuous ranked probability score* (CRPS) is a proper scoring rule assigned to predictive distributions defined by

$$\mathrm{CRPS}(F, y) = \int (F(z) - \mathbb{1}_{\{z \geq y\}})^2 dz. \tag{6.5}$$

CRPS rewards those distributions that are concentrated at the actual realized value. It attains the minimum of 0 when F is concentrated at y, and it is always non-negative. Therefore, the closer to zero, the better.

Since conformal predictive distributions are fuzzy, the *crisp* modifications as defined in (5.23) must be used instead to compute CRPS for conformal predictive distributions [39, Section 5].

Finally, an assessment based solely on the first two moments of the predictive distributions μ_F, σ_F^2 can be done using the proper Dawid-Sebastiani score [9].

Definition 6.6 (Dawid-Sebastiani score)**.** The *Dawid-Sebastiani score* (DSS) is a proper scoring rule defined by

$$\mathrm{DSS}(F, y) = \frac{(y - \mu_F)^2}{\sigma_F^2} + 2 \log \sigma_F. \tag{6.6}$$

The first term can be associated with the calibration performance of the prediction by computing how many standard deviations the true value is from the predictive mean. It equals one on average for a perfectly calibrated prediction, since $\mathbb{E}\left[\frac{(Y-\mu)^2}{\sigma^2}\right] = 1$ when Y has mean μ and standard deviation σ. The log term can be interpreted as a sharpness term, where a distribution with a smaller standard deviation is rewarded more strongly. Like with the other proper scoring rules, a smaller DSS value is desirable.

Tail mean absolute error

In addition to the scoring rules summarized in the above section, we compare the performances of the different approaches by computing the 0.05-left-tail conditional mean

$$\mathbb{E}(Y_{n_t+1} \mid Z_1, \ldots, Z_{n_t}, Y_{n_t+1} \leq \hat{q}_{0.05}).$$

Since the true underlying conditional density is normal for our simulated house prices dataset, its associated tail mean can be computed analytically using the formula

$$\mathbb{E}(X \mid X \leq q_\alpha) = \mu - \sigma \frac{\varphi(z_\alpha)}{\alpha}, \quad \text{for } X \sim N(\mu, \sigma^2).$$

In the above formula, φ is the standard normal PDF with Φ the standard normal CDF, $z_\alpha = \Phi^{-1}(\alpha)$, and $q_\alpha = \mu + \sigma z_\alpha$.

We then take the mean absolute error (MAE) between the true and approximated tail means of each method.

Performance comparison

Table 6.2 summarizes an overview of the scoring results compared between direct finite-differencing and quantile-matching, with or without Gaussian filtering applied. The methods are compared across calibration sizes of 1000, 10000, and 275043 (corresponding to the 65/25/10 split), respectively. We find that the quantile-matching method successfully decreases the MISE values in a small calibration size compared to direct finite-differencing without Gaussian filtering, and applying an additional Gaussian filtering step on top of it produces the smallest MISE values on average across all calibration sizes, albeit the improvement is incremental with a 65/25/10 split. The CRPS performance is very similar across all approaches and calibration sizes. The quadratic score is somewhat smaller for quantile-matching than for direct finite-differencing, and Gaussian filtering further reduces the score.

The Dawid-Sebastiani score is significantly smaller by directly finite-differencing the conformal predictive distribution, and Gaussian filtering on top offers no significant improvement to either the conformal or quantile-matched distribution in this case. On the other hand, the tail mean MAE is smaller across all calibration sizes for the quantile-matching method than for the conformal predictive distribution. Applying a Gaussian filtering step actually severely damages these errors due to spreading out tail mass across a larger range, which gets inflated in expectation, and we find that a naked quantile-matched distribution is best suited for computing tail means.

Since the quantile-matched density with 100 quantile levels gives no weight below or above the 0.01 and 0.99 quantile levels respectively, the log score is significantly inflated by rare tail points when we approximate $\hat{f}_{QM}(y) \approx 0$. We expect the score to be reduced by introducing a finer resolution for the tails, for instance by using an uneven grid $[0.001, \ldots, 0.009, 0.01, 0.02, \ldots, 0.99, 0.991, \ldots, 0.999]$. Exploring the quantile-matching method with uneven grids is open to further research, however, and we do not focus on it here.

The quantile-matching method is significantly faster to run for large calibration sets, since the resolution of the grid is far smaller (100) compared to evaluating the distribution for, e.g., 275000 conformal atoms.

Overall, we find that the new quantile-matching method is robust across various metrics and is especially well-suited for evaluating tail statistics. It achieves similar performance levels to conformal distributions under the quadratic score, CRPS and MISE, while underperforming when using the log-score or the Dawid-Sebastiani score. The main advantage of the newly developed method lies in its ability to reduce fuzziness at the density level and in its faster runtime for large calibration sizes. We are looking forward to further developments in the quantile-matching method, including, but not limited to, the exploration of adaptive quantile grid sizes or optimal choices of the level K.

Table 6.2: Average scoring metrics over 1000 test observations for three calibration set sizes: $|\mathcal{I}_2| = 1{,}000$ $|\mathcal{I}_2| = 10{,}000$ and $|\mathcal{I}_2| = 275{,}043$ (full 25% split), using the SCP score. Here, quantile matching is applied using 100 evenly spaced quantile levels. The tail means are evaluated at 0.05-quantile level. †LS for QM without filtering is inflated by rare events where $\hat{f}_{QM}(y) \approx 0$.

		Calibration size $\lvert\mathcal{I}_2\rvert = 1{,}000$						
Method	**Filtering**	**MISE**	**CRPS**	**LS**	**QS**	**DSS**	**Tail mean MAE**	**Time (s)**
Direct FD	No	0.4525	0.1013	-0.171	-1.229	-2.422	0.0716	1.480
Direct FD	Yes	0.1684	0.1013	-0.281	-1.572	-2.421	0.1684	1.449
QM	No	0.2993	0.1013	$13.451^{\dagger}$	-1.405	-1.045	0.0599	1.731
QM	Yes	0.1573	0.1016	$1.069^{\dagger}$	-1.582	-1.045	0.5543	1.720
		Calibration size $\lvert\mathcal{I}_2\rvert = 10{,}000$						
Method	**Filtering**	**MISE**	**CRPS**	**LS**	**QS**	**DSS**	**Tail mean MAE**	**Time (s)**
Direct FD	No	0.2112	0.1014	-0.269	-1.547	-2.418	0.0708	1.317
Direct FD	Yes	0.1546	0.1015	-0.289	-1.584	-2.418	0.0817	1.294
QM	No	0.1629	0.1015	$13.456^{\dagger}$	-1.583	-1.550	0.0597	1.324
QM	Yes	0.1524	0.1015	$1.109^{\dagger}$	-1.584	-1.554	0.5120	1.299
		Calibration size $\lvert\mathcal{I}_2\rvert = 275{,}043$ (25% split)						
Method	**Filtering**	**MISE**	**CRPS**	**LS**	**QS**	**DSS**	**Tail mean MAE**	**Time (s)**
Direct FD	No	0.1684	0.1023	-0.280	-1.572	-2.398	0.0718	6.661
Direct FD	Yes	0.1667	0.1023	-0.283	-1.578	-2.398	0.1423	6.525
QM	No	0.1687	0.1023	$14.843^{\dagger}$	-1.574	-1.262	0.0613	1.628
QM	Yes	0.1665	0.1023	$1.125^{\dagger}$	-1.577	-1.264	0.4874	1.624

7

Conclusion

Summary of main contributions

Throughout this thesis, we have built on the theory of conformal predictions, most notably the very rich contributions of Vladimir Vovk [38, 41, 40, 39]. Along the way, we have approached certain classical results differently, relaxed restrictive assumptions in previous work, or, as far as we know, produced original results not presented before in the literature. In this section, we provide an extensive overview of the most important contributions of our thesis to the existing literature.

- The proof of Theorem 2.1 relies on a completeness argument of the unordered statistic $\{Y_1, \ldots, Y_n\}$. This argument is cited as being proven in lecture notes from 1950, which have been lost to time. We prove this independently from scratch here.
- We give a detailed proof of Theorem 2.8 that the Epanechnikov kernel is optimal, which is not given in the original papers [11, 18].
- We state and prove in Theorem 3.9 that the Split Conformal Direct Prediction Method also retains a finite sample marginal coverage guarantee.
- We relax the IID assumption in Theorem 4.15 to exchangeability, and we give an original proof of the theorem.
- We adjust the tails of the classical conformal predictive distributions [41] to obtain true CDFs and quantify the error in PIT using Theorem 4.25.
- We provide a principled framework for quantifying the distance from the true uniform distribution in probability integral transform with Theorem 4.26.
- We extend the concept of conformal predictive distributions to be able to retrieve predictive densities, using monotonic cubic interpolation and/or Gaussian filtering.
- We develop a novel method with *quantile-matching*, which unites both approaches of conformal prediction intervals and conformal predictive distributions. We give an upper bound on the distance between its associated Probability Integral Transform and the true uniform distribution via Theorem 5.6. We connect the quantile-matching method to the crisp modifications [39] of randomized predictive distributions, and we quantify the error in the PIT of the crisp versions via Corollary 5.6.1. We show empirically that the quantile-matching approach is better suited for computing tail means, having a smaller mean absolute error than the conformal counterparts, while maintaining similar performance under MISE, CRPS, and QS.
- We apply the methods on a real estate simulated dataset and show empirically the error in probability integral transform of filtered conformal distributions to be of order $O\left(\frac{1}{\sqrt{n-m+1}}\right)$ for our real estate dataset.

Key findings

The present work has explored model-agnostic prediction density methods, with a particular focus on the *conformal* framework. Chapter 2 introduced Kernel Density Estimation for independent and identically distributed random variables without features. This method performs well asymptotically, but no meaningful conclusions can be drawn from finite samples.

In contrast, several conformal prediction interval methods are presented in Chapter 3, including the standard version, conformalized quantile regression, and conformal direct prediction method. All of these methods satisfy a finite-sample marginal validity guarantee that remains true even under model misspecification. We discuss different coverage notions and highlight that one cannot construct meaningful prediction intervals for which the validity guarantee holds *conditionally*.

The extension of conformal prediction intervals to conformal predictive distributions is discussed in Chapter 4, where we adjust the marginal validity guarantee appropriately. The validity guarantee essentially says that conformal predictive systems are $\mathrm{Unif}[0, 1]$ marginally in PIT over all observations. The distributions are randomized to ensure that this condition holds exactly in finite samples, resulting in fuzzy distributions. We relax the concept of a randomized predictive system to allow for the marginal validity guarantee to hold asymptotically. Within this framework, we obtain upper bounds on the perturbation from the uniform distribution in the probability integral transform for tail-corrected distributions and their continuous versions.

To retrieve predictive densities, several approaches are discussed in Chapter 5, including directly finite-differencing the conformal predictive distributions at the conformal atoms, which produces very rough densities that can hide underlying features; Gaussian filtering the predictive distribution, which produces the smoothest densities but gives no theoretical upper bound on the perturbation in probability integral transform; and quantile-matching, a novel approach that lies at the intersection of prediction intervals and predictive distributions and that retains a theoretical upper bound of the deviation in probability integral transform from uniformity. We draw a connection between quantile-matched distributions and crisp modifications of conformal predictive distributions, providing a sharp upper bound in the absolute error in PIT of the crisp version from the $\mathrm{Unif}[0, 1]$.

The methods are illustrated on a large dataset of simulated house prices based on the Hierarchical Trend Model in Chapter 6. We observe empirically that the Gaussian filtered distribution, even in large calibration sizes, remains within $\frac{3}{\sqrt{n-m+1}}$ distance from the true uniform distribution in the probability integral transform. Across different calibration scores, we find no advantage to using the CQR or CDP scores over the standard conformity score, and we advise practitioners to use the standard version due to faster computation time, which stems from the absence of a quantile regression algorithm. Across different performance metrics, we find that quantile-matching performs similarly to conformal predictive densities obtained via direct finite-differencing or Gaussian filtering, including MISE, CRPS, and QS. It overperforms in terms of tail-expectation mean absolute error compared to its counterparts, while underperforming on the log-score and Dawid-Sebastiani scores. We expect that the log-score can be further reduced by using an adaptive grid size around the extreme tails, which should reduce the number of outliers where $\hat{f}_{QM}(y) \approx 0$. Such an extension remains an open question for further research regarding the quantile-matching method.

Other directions for further research include examining the asymptotic performance of these methods, particularly the Gaussian-filtered version with fixed variance. We suspect that the empirical bound observed might be violated. Alternatively, exploring a principled (i.e., optimal) approach in choosing the kernel smoothing parameter also remains an open question. In terms of consistency, empirically, we see that using the standard conformity score leads to predictive systems that are not universally consistent, given the shift in predictive distributions from the true underlying one. We believe further research can be done on constructing predictive distributions and associated predictive densities that are also universally consistent, and we refer the reader to [37] as a good starting point, where a universally consistent conformal predictive system is constructed. We also invite further exploration of other conformity scores associated with the quantile-matching method to observe whether performance can be improved. Finally, for practitioners, model reparametrization based on predictive densities is of particular interest, and we believe this to be an important next step to the model-agnostic predictive densities framework explored here.

References

[1] F. Abramovich and Y. Ritov. *Statistical Inference: A Concise Introduction*. Taylor and Francis, 2022.

[2] R.F. Barber et al. *The limits of distribution-free conditional predictive inference*. 2020. arXiv: 1903.04684 [math.ST]. URL: https://arxiv.org/abs/1903.04684.

[3] M. de Berg et al. "Computational Geometry". In: *Computational Geometry: Algorithms and Applications*. Springer Berlin Heidelberg, 2008, pp. 1–17. ISBN: 978-3-540-77974-2. DOI: 10.1007/978-3-540-77974-2_1. URL: https://doi.org/10.1007/978-3-540-77974-2_1.

[4] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1995, pp. 242–243.

[5] S. Bochner. *Harmonic Analysis and the Theory of Probability*. Dover Books on Mathematics. Dover Publications, 2013. ISBN: 9780486154800.

[6] V.I. Bogachev. "Operations on measures and functions". In: *Measure Theory*. Springer Berlin Heidelberg, 2007, pp. 175–248. ISBN: 978-3-540-34514-5. DOI: 10.1007/978-3-540-34514-5_3. URL: https://doi.org/10.1007/978-3-540-34514-5_3.

[7] G. Casella and R. Berger. *Statistical Inference*. 2nd. Taylor and Francis, 2024.

[8] V. Chernozhukov et al. "Distributional conformal prediction". In: *Proceedings of the National Academy of Sciences of the United States of America* 118.48 (2021), e2107794118. DOI: 10.1073/pnas.2107794118.

[9] A. Dawid. "Coherent disperson criteria for optimal experimental design". In: *Annals of Statistics* 27 (Mar. 1999). DOI: 10.1214/aos/1018031101.

[10] R. L. Dougherty, A. Edelman, and J. M. Hyman. "Nonnegativity-, Monotonicity-, or Convexity-Preserving Cubic and Quintic Hermite Interpolation". In: *Mathematics of Computation* 52.186 (1989), pp. 471–494. ISSN: 00255718, 10886842. URL: http://www.jstor.org/stable/2008477 (visited on 03/03/2026).

[11] V. A. Epanechnikov. "Non-Parametric Estimation of a Multivariate Probability Density". In: *Theory of Probability & Its Applications* 14.1 (1969), pp. 153–158. DOI: 10.1137/1114019.

[12] M. Francke, A. Kadhum, and D. Kroon. "Creating Model-Agnostic Prediction Intervals". In: *SSRN Electronic Journal* (Jan. 2024). DOI: 10.2139/ssrn.4872625.

[13] M. Francke and A. F. de Vos. "Efficient Computation of Hierarchical Trends". In: *Journal of Business & Economic Statistics* 18.1 (2000), pp. 51–57. DOI: 10.1080/07350015.2000.10524847.

[14] M. Francke and G.A. Vos. "The Hierarchical Trend Model for Property Valuation and Local Price Indices". In: *The Journal of Real Estate Finance and Economics* 28 (Mar. 2004). DOI: 10.1023/B:REAL.0000011153.04496.42.

[15] F. N. Fritsch and R. E. Carlson. "Monotone Piecewise Cubic Interpolation". In: *SIAM Journal on Numerical Analysis* 17.2 (1980), pp. 238–246. DOI: 10.1137/0717021. URL: https://doi.org/10.1137/0717021.

[16] T. Gneiting and M. Katzfuss. "Probabilistic Forecasting". In: *Annual Review of Statistics and Its Application* 1 (Jan. 2014), pp. 125–151. DOI: 10.1146/annurev-statistics-062713-085831.

[17] T. Gneiting and A.E. Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378. ISSN: 01621459. URL: http://www.jstor.org/stable/27639845 (visited on 05/08/2026).

[18] J. L. Hodges and E. L. Lehmann. "The Efficiency of Some Nonparametric Competitors of the t-Test". In: *The Annals of Mathematical Statistics* 27.2 (1956), pp. 324–335. ISSN: 00034851, 21688990.

[19] G. Ke et al. “LightGBM: a highly efficient gradient boosting decision tree”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. 2017, pp. 3149–3157. ISBN: 9781510860964.

[20] G. A Kelley and K. S Kelley. “Impact of progressive resistance training on lipids and lipoproteins in adults: another look at a meta-analysis using prediction intervals”. In: *Preventive medicine* 49.6 (Dec. 2009), pp. 473–475. ISSN: 0091-7435. DOI: 10.1016/j.ypmed.2009.09.018. URL: https://doi.org/10.1016/j.ypmed.2009.09.018.

[21] R. Koenker and G. Bassett. “Regression Quantiles”. In: *Econometrica* 46.1 (1978), pp. 33–50. URL: http://www.jstor.org/stable/1913643.

[22] H. Lebesgue. “Sur l’intégration des fonctions discontinues”. In: *Annales scientifiques de l’École Normale Supérieure* (1910).

[23] J. Lei et al. *Distribution-Free Predictive Inference For Regression*. 2017. arXiv: 1604.04173 [stat.ME]. URL: https://arxiv.org/abs/1604.04173.

[24] D. Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton: Princeton University Press, 2012. ISBN: 9781400842643. DOI: doi:10.1515/9781400842643.

[25] F. van der Meulen. *Statistical Inference - Lecture Notes for the course WI4455*. 2022. URL: https://github.com/fmeulen/fmeulen.github.io/blob/main/ln-wi4455.pdf.

[26] E. Parzen. “On Estimation of a Probability Density Function and Mode”. In: *The Annals of Mathematical Statistics* 33.3 (1962), pp. 1065–1076. URL: http://www.jstor.org/stable/2237880 (visited on 10/07/2025).

[27] Y. Romano, E. Patterson, and E. J. Candès. “Conformalized Quantile Regression”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (2019). URL: https://arxiv.org/abs/1905.03222.

[28] M. Rosenblatt. “Remarks on Some Nonparametric Estimates of a Density Function”. In: *The Annals of Mathematical Statistics* 27.3 (1956), pp. 832–837. ISSN: 00034851, 21688990. URL: http://www.jstor.org/stable/2237390 (visited on 10/07/2025).

[29] SciPy Developers. *scipy.ndimage.gaussian_filter1d — Multi-dimensional Gaussian filter*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.gaussian_filter1d.html. Accessed: 2026-03-25. 2026.

[30] J. Shen, Y.R. Liu, and M. Xie. “Prediction with confidence—A general framework for predictive inference”. In: *Journal of Statistical Planning and Inference* 195 (2018), pp. 126–140. ISSN: 0378-3758. DOI: https://doi.org/10.1016/j.jspi.2017.09.012. URL: https://www.sciencedirect.com/science/article/pii/S0378375817301696.

[31] R.H. Shumway and D.S. Stoffer. “ARIMA Models”. In: *Time Series Analysis and Its Applications: With R Examples*. Cham: Springer Nature Switzerland, 2025, pp. 85–175. ISBN: 978-3-031-70584-7. DOI: 10.1007/978-3-031-70584-7_3. URL: https://doi.org/10.1007/978-3-031-70584-7_3.

[32] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Feb. 2018, pp. 1–175. ISBN: 9781315140919. DOI: 10.1201/9781315140919.

[33] S.W. Smith. *The Scientist and Engineer’s Guide to Digital Signal Processing*. California Technical Pub., 1997. ISBN: 9780966017632. URL: https://books.google.nl/books?id=rp2VQgAACAAJ.

[34] I. Steinwart and A. Christmann. “Estimating conditional quantiles with the help of the pinball loss”. In: *Bernoulli* 17.1 (2011), pp. 211–225.

[35] M. Stone. “Cross-Validatory Choice and Assessment of Statistical Predictions”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (Dec. 2018), pp. 111–133. DOI: 10.1111/j.2517-6161.1974.tb00994.x.

[36] V. Vovk. *Conditional validity of inductive conformal predictors*. 2012. arXiv: 1209.2673 [cs.LG]. URL: https://arxiv.org/abs/1209.2673.

[37] V. Vovk. “Universal predictive systems”. In: *Pattern Recognition* 126 (2022), p. 108536. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2022.108536. URL: https://www.sciencedirect.com/science/article/pii/S0031320322000176.

[38] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Jan. 2005. DOI: 10.1007/b106715.

[39] V. Vovk et al. “Computationally efficient versions of conformal predictive distributions”. In: *Neurocomputing* 397 (2020), pp. 292–308. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.10.110. URL: https://www.sciencedirect.com/science/article/pii/S0925231219316042.

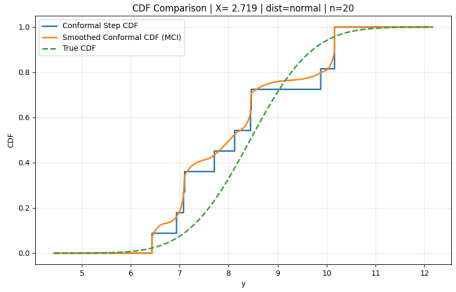
[40] V. Vovk et al. *Conformal predictive distributions with kernels*. 2017. arXiv: 1710.08894 [cs.LG]. URL: https://arxiv.org/abs/1710.08894.

[41] V. Vovk et al. “Nonparametric predictive distributions based on conformal prediction”. In: *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*. Ed. by Alex Gammerman et al. Vol. 60. Proceedings of Machine Learning Research. PMLR, 2017, pp. 82–102. URL: https://proceedings.mlr.press/v60/vovk17a.html.

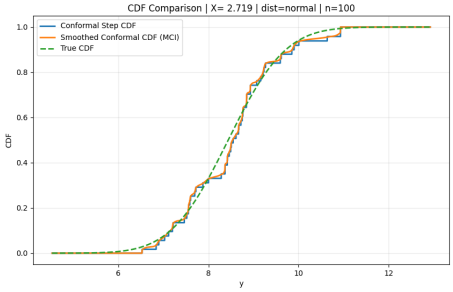
[42] G. Wolberg and I. Alfy. “Monotonic Cubic Spline Interpolation”. In: *Proceedings of the International Conference on Computer Graphics*. CGI ’99. USA: IEEE Computer Society, 1999, p. 188. ISBN: 0769501850.

A

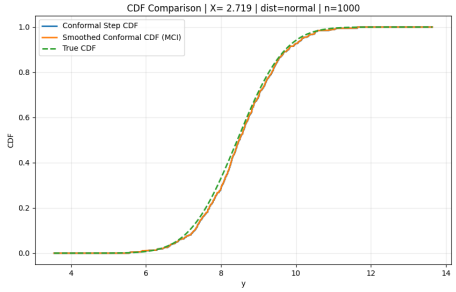
Chapter 5: Additional figures



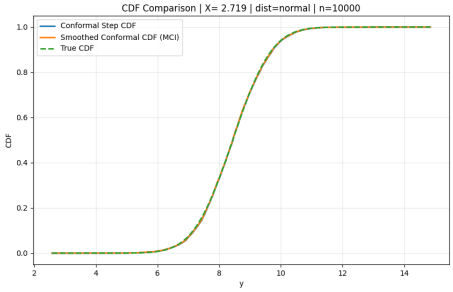
(a) $n = 20$



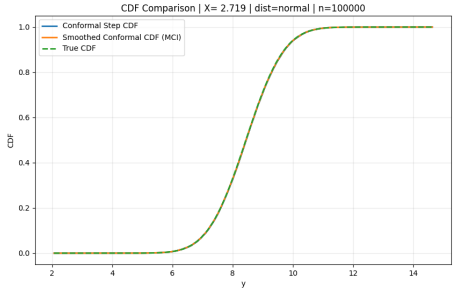
(b) $n = 100$



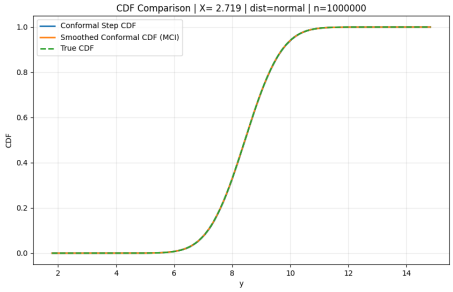
(c) $n = 1000$



(d) $n = 10000$

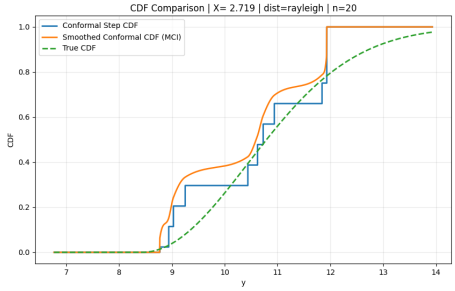


(e) $n = 100000$

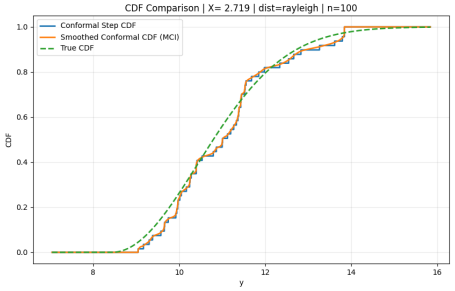


(f) $n = 1000000$

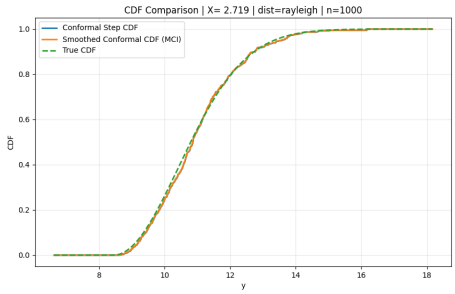
Figure A.1: CDF comparison for the $N(0, 1)$ distribution across different sample sizes



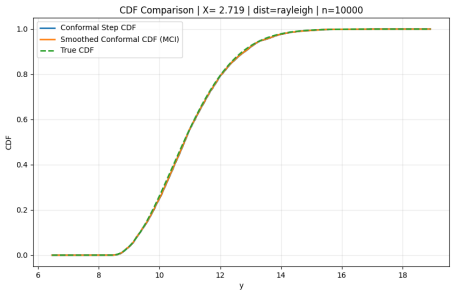
(a) $n = 20$



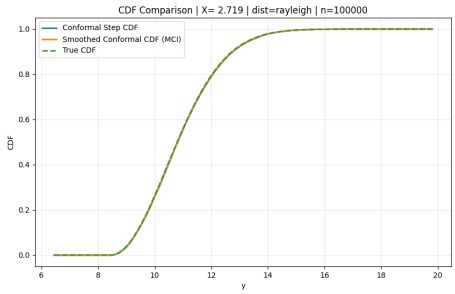
(b) $n = 100$



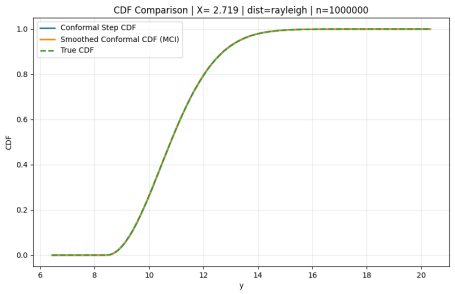
(c) $n = 1000$



(d) $n = 10000$



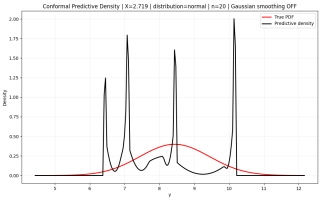
(e) $n = 100000$



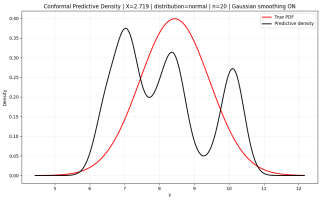
(f) $n = 1000000$

Figure A.2: CDF comparison for the $\mathrm{Rayleigh}(2)$ distribution across different sample sizes

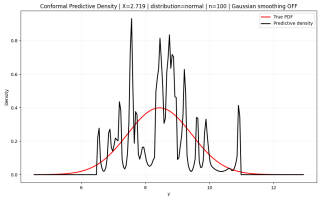
Unfiltered **Filtered**



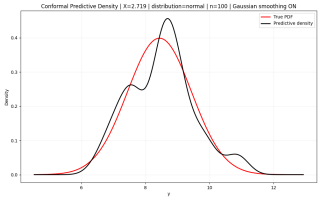
(a) $n = 20$



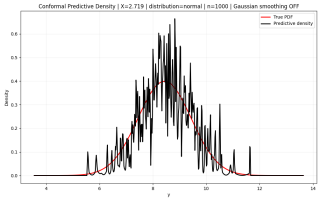
(b) $n = 20$



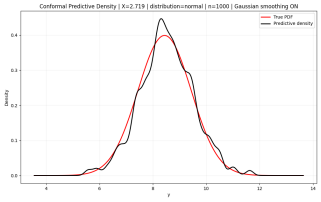
(c) $n = 100$



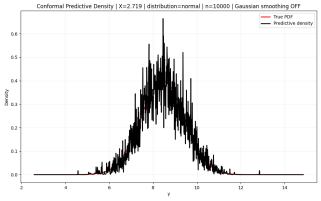
(d) $n = 100$



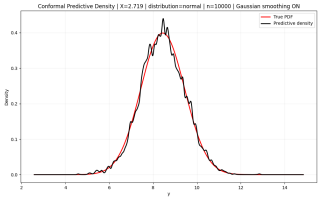
(e) $n = 1000$



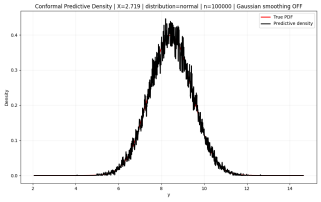
(f) $n = 1000$



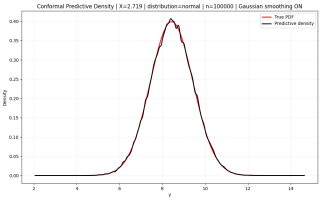
(g) $n = 10{,}000$



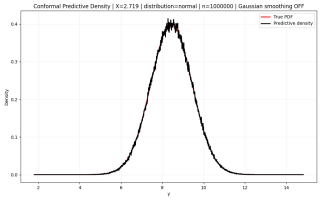
(h) $n = 10{,}000$



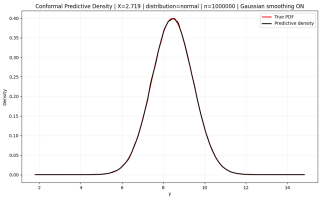
(i) $n = 100{,}000$



(j) $n = 100{,}000$



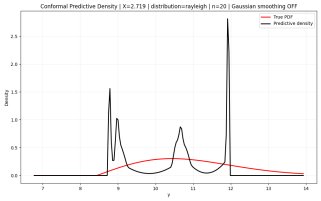
(k) $n = 1{,}000{,}000$



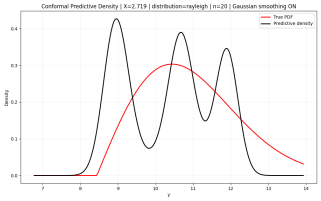
(l) $n = 1{,}000{,}000$

Figure A.3: Unfiltered and Gaussian-filtered PDFs for the $N(0, 1)$ distribution across sample sizes $n \in \{20,\ 100,\ 1000,\ 10^4,\ 10^5,\ 10^6\}$. Each row corresponds to one sample size; the left column shows the raw MCI estimate and the right column the Gaussian-filtered result.

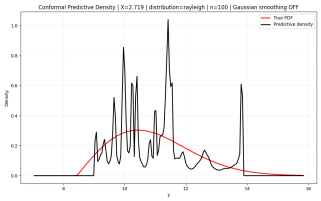
Unfiltered | **Filtered**



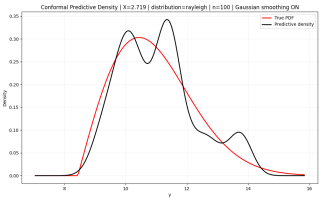
(a) $n = 20$



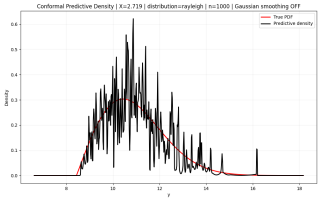
(b) $n = 20$



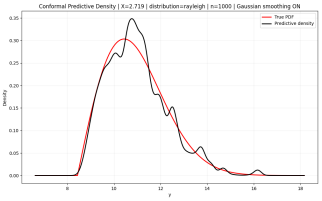
(c) $n = 100$



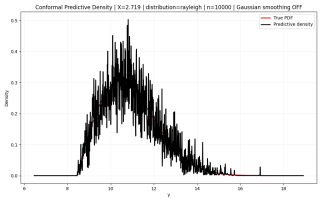
(d) $n = 100$



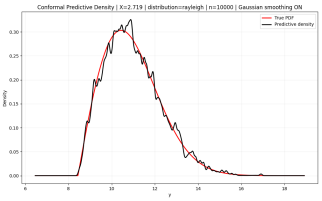
(e) $n = 1000$



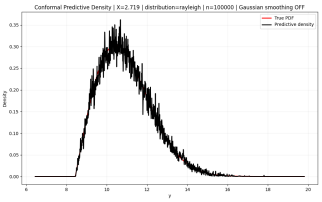
(f) $n = 1000$



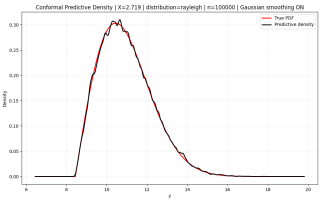
(g) $n = 10{,}000$



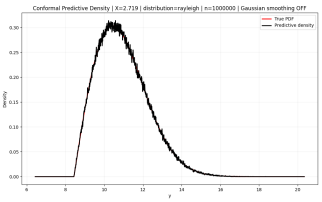
(h) $n = 10{,}000$



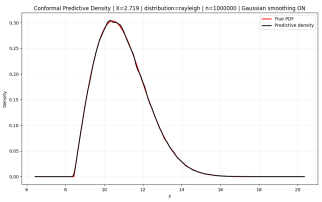
(i) $n = 100{,}000$



(j) $n = 100{,}000$



(k) $n = 1{,}000{,}000$



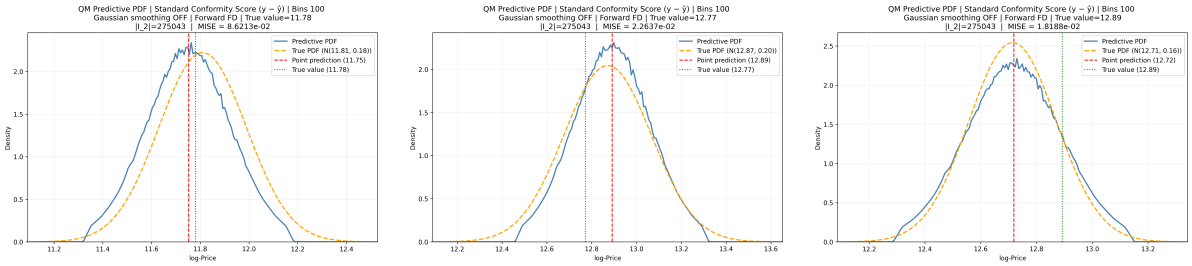
(l) $n = 1{,}000{,}000$

Figure A.4: Unfiltered and Gaussian-filtered PDFs for the $\mathrm{Rayleigh}(2)$ distribution across sample sizes $n \in \{20,\ 100,\ 1000,\ 10^4,\ 10^5,\ 10^6\}$. Each row corresponds to one sample size; the left column shows the raw MCI estimate and the right column the Gaussian-filtered result.

B

Chapter 6: Additional figures

A few predictive densities for some other transaction prices

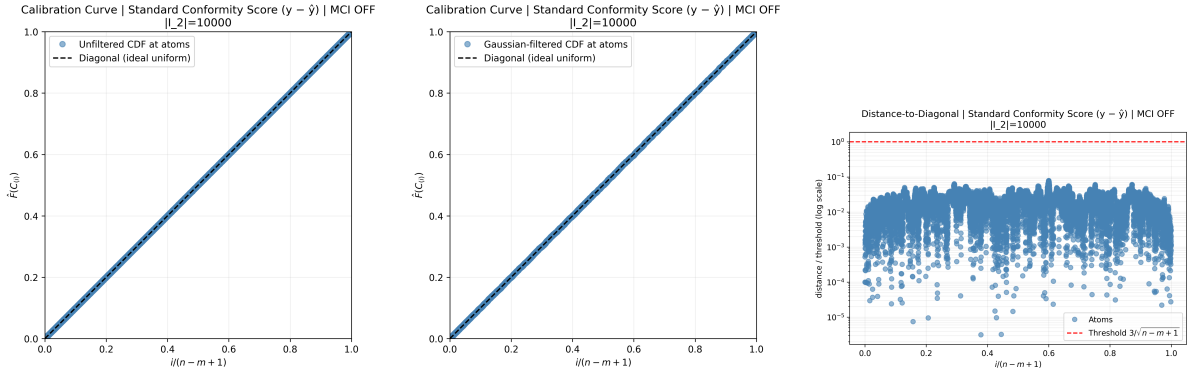


(a) True value=11.78 **(b)** True value=12.77 **(c)** True value=12.89

Figure B.1: Quantile-matched predictive densities with a 65/25/10 split

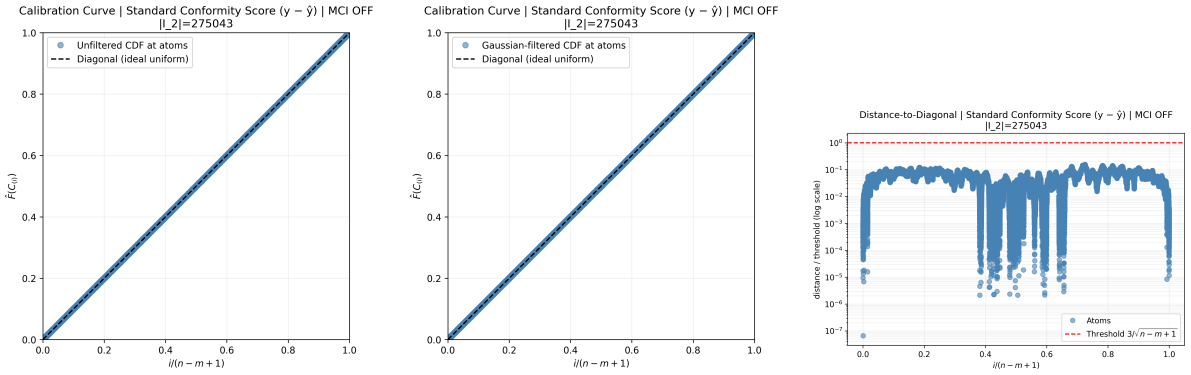
The above figures show that it is very difficult to minimize multiple criteria simultaneously in a model-agnostic sense. For instance, the point prediction in the right-most figure is almost perfectly aligned with the true mean (12.72) of the underlying true distribution, leading to a very small MISE. However, the actual observed value is 12.89, and on forecasting scores such as CRPS and LS, the distribution is penalized for assigning less mass at the realization.

Calibration curves



(a) Calibration curve for the conformal predictive distribution **(b)** Calibration curve for Gaussian filtered distribution **(c)** Absolute error of Gaussian-filtered distribution versus $\mathrm{Unif}[0, 1]$

Figure B.2: Calibration curves and absolute error in PIT for the conformal predictive distribution with 10000 calibration points of a random transaction price

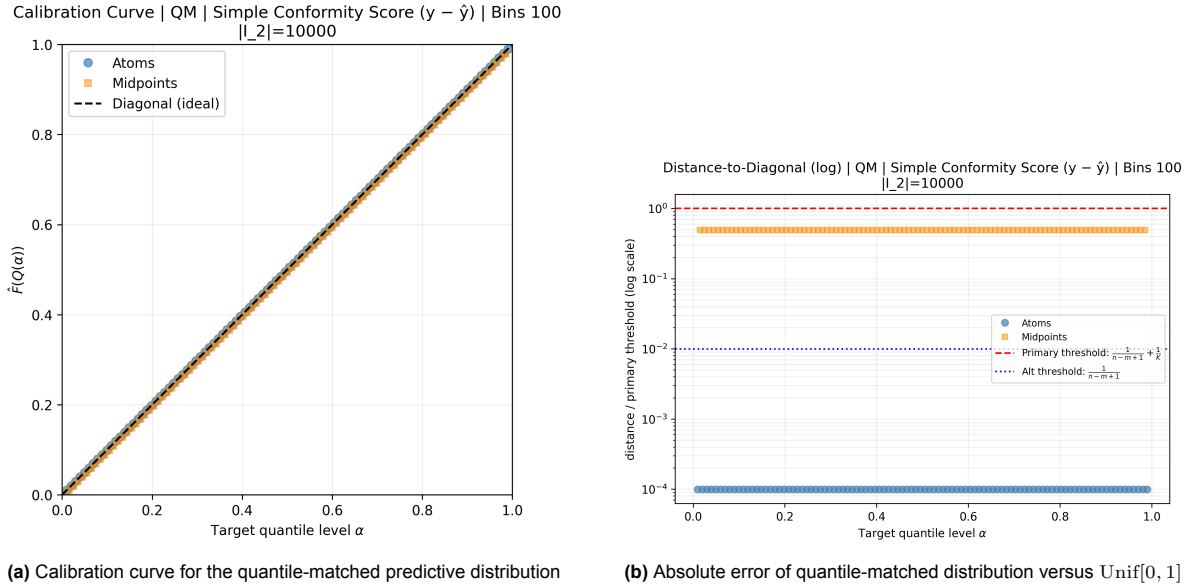


(a) Calibration curve for the conformal predictive distribution

(b) Calibration curve for Gaussian filtered distribution

(c) Absolute error of Gaussian-filtered distribution versus $\mathrm{Unif}[0, 1]$

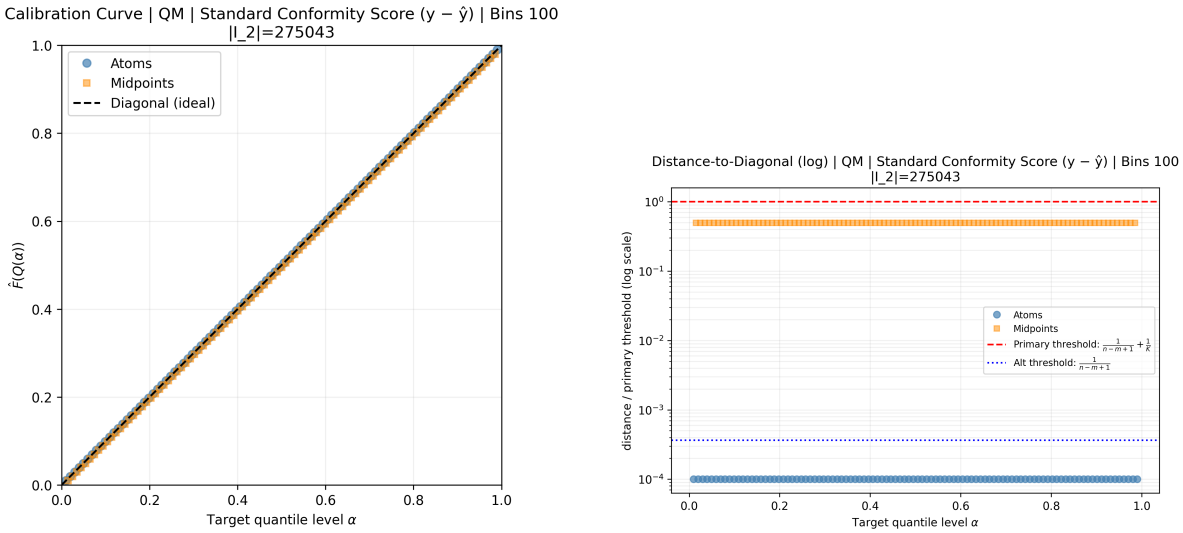
Figure B.3: Calibration curves and absolute error in PIT for the conformal predictive distribution with a 65/25/10 split of a random transaction price



(a) Calibration curve for the quantile-matched predictive distribution

(b) Absolute error of quantile-matched distribution versus $\mathrm{Unif}[0, 1]$

Figure B.4: Calibration curves and absolute error in PIT for the quantile-matched distribution with 10000 calibration points of a random transaction price



(a) Calibration curve for the quantile-matched predictive distribution

(b) Absolute error of quantile-matched distribution versus $\mathrm{Unif}[0, 1]$

Figure B.5: Calibration curves and absolute error in PIT for the quantile-matched distribution with $K = 100$ and a 65/25/10 split for a random transaction price