CARNEGIE MELLON UNIVERSITY

Calibrated Conditional Density Models and Predictive Inference via Local Diagnostics

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

DOCTOR OF PHILOSOPHY
IN
STATISTICS AND MACHINE LEARNING
BY

DAVID ZHAO
DEPARTMENT OF STATISTICS
DEPARTMENT OF MACHINE LEARNING
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PA 15213

Carnegie Mellon University

MAY 2023

Dedicated to my grandparents.

Acknowledgements

First of all, I thank my brilliant advisor Ann Lee and collaborator Rafael Izbicki. I have learned and grown substantially from working with them on challenging, impactful real-world problems in the physical sciences, and I am very grateful that our time together has been so productive. Ann has been tirelessly committed and supportive of my work from the beginning, and I could not possibly have written this thesis without her extraordinary vision and thoughtful guidance.

I thank my rockstar thesis committee of Aaditya Ramdas, Jing Lei, and Rafael Stern. They have provided insightful and constructive feedback that helped me improve my work, see it from fresh perspectives, and better communicate it to others. I am honored to have them on my committee.

I thank all professors and staff from the Statistics & Data Science and Machine Learning Departments, for everything they have taught me over the past five years. Thanks to Siva Balakrishnan, Larry Wasserman, Nina Balcan, and Stephanie Rosenthal for especially rewarding classes. Special thanks to Alessandro Rinaldo for being a fantastic mentor and teacher, and for his wisdom, positivity, and great sense of humor.

I thank my fellow students I had the privilege of collaborating with: Biprateep Dey, Niccolo Dalmasso, Tria McNeely, Benjamin LeRoy, Luca Masserano, Wanshan Li, and Riccardo Fogliato. I learned many valuable skills from working together on projects, and I am proud of everything we accomplished together.

I thank my PhD cohort, a diverse group of extremely impressive people I am proud to know. I look fondly back on our many hours working together on campus and our delicious potluck dinners. Shoutout to Tim Barry, my weightlifting and blue-sky conversation buddy—I look forward to his first album release and his first genomics startup.

Finally, I thank my family and my partner. I cannot express how much their love and encouragement have meant to me over the years. My partner has enriched my life immeasurably and made the past five years here in Pittsburgh so special; it feels like home now, and I know we will miss it. The love from my family has always been the greatest gift I've received, giving me the strength to pursue all my goals.

It is better to solve the right problem approximately
than to solve the wrong problem exactly. $\sim$*John Tukey*

Abstract

Conditional densities, $f(y|\mathbf{x})$, are integral to uncertainty quantification when predicting a target y from covariates $\mathbf{x}$, but they are challenging to estimate well. It is therefore difficult to ensure that $(1-\alpha)$-level prediction sets for y constructed from a conditional density model have the correct conditional coverage; that is, they contain the observed y with probability $(1-\alpha)$ at all locations $\mathbf{x}$ in feature space. We investigate how we can, with access to an observed "ground truth" sample of $(\mathbf{x}, y)$, develop diagnostics that specify how an estimated conditional density errs from the true density, at any location $\mathbf{x}$. With this detailed insight into quality of fit at any $\mathbf{x}$, we are then naturally able to develop a calibration method, to correct an initial estimated conditional density using a ground truth data sample. This method produces calibrated densities for $f(y|\mathbf{x})$ that are approximately accurate across all locations $\mathbf{x}$, yielding calibrated prediction sets with accurate conditional coverage.

In the first part of this thesis, we present practical procedures for identifying, localizing, and interpreting the nature of (statistically significant) discrepancies between an approximated and true conditional density, over the entire feature space. Our flexible framework is more discerning than previous diagnostics, in that we can distinguish an arbitrarily misspecified model from the true conditional density of an observed sample. We also provide "Amortized Local P-P plots" (ALP), which are interpretable graphical summaries of distributional differences at any location in the feature space. In the second part of this thesis, we leverage this diagnostic framework to correct misspecified conditional density models, from which we can then construct calibrated prediction sets that have desired conditional coverage. Because our diagnostics directly specify where in feature space and how an estimated and true conditional CDF may differ, we can use this information to directly correct the model towards the target conditional coverage. We explore an application to the real-world astrophysical problem of photometric redshift ("photo-z") prediction, where conditional density models are difficult to estimate and conditional coverage is of practical significance. In the third part of this thesis, we explore an extension of this calibration method that hybridizes it with local conformal inference, allowing it to achieve finite-sample marginal and local validity at the expense of some precision.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the increasing complexity and popularity of black box algorithms, there has been growing interest in the machine learning community for precise uncertainty quantification for their predictions, beyond just point estimates of the conditional mean. The conditional density $f(y|\mathbf{x})$ of the response variable y given features $\mathbf{x}$ can be used to build predictive regions for y, which are more informative than point predictions. Indeed, in prediction settings, $f(y|\mathbf{x})$ provides a full account of the uncertainty in the outcome y given new observations $\mathbf{x}$. Conditional densities are also central to Bayesian parameter inference, where the posterior distribution $f(\theta|\mathbf{x})$ is key to quantifying uncertainty about the parameters θ of interest after observing data $\mathbf{x}$. Any downstream analysis in predictive modeling or Bayesian inference depends on the trustworthiness of the assumed conditional density model.

Validating such models can be challenging. Prior to our work, there did not exist a comprehensive and rigorous set of diagnostics that describe, for all values of $\mathbf{x}$, the quality of fit of a conditional density model. It is especially challenging to gain insight into different modes of failure of conditional density models, particularly for high-dimensional or mixed-type data $\mathbf{x}$. While previous diagnostic tools could determine whether an approximated conditional density is compatible overall with a data sample, they lacked a principled framework for identifying, locating, and interpreting the nature of statistically significant discrepancies over the entire feature space. While there is rapidly growing interest in both methods and applications for describing the entire predictive distribution of Y given $\mathbf{X}$ (see Section 2.1 for a survey), most current research on predictive inference concerns constructing calibrated prediction sets only. It is often believed that the problem of obtaining and assessing entire conditionally calibrated predictive distributions is too challenging. Here, conditional calibration refers to the property that $(1-\alpha)$-level prediction sets for y have the correct conditional coverage; that is, they contain the observed y with probability $(1-\alpha)$ at all locations $\mathbf{x}$ in feature space.

In this thesis, we show that comprehensive and interpretable diagnostics of entire predictive distributions, and moreover recalibration of those distributions towards conditional coverage, are indeed attainable goals in practice. Our proposed method relies on the idea of regressing probability integral transform (PIT) scores, computed given an

initial conditional density model $\widehat{f}(y|\mathbf{x})$, against $\mathbf{X}$. This regression gives full and granular diagnostics of conditional coverage across the entire feature space, which can then be used to recalibrate the entire model.

In the first part of this thesis, we present a rigorous framework for assessing the quality of conditional density models, and diagnosing where locally and how exactly the models may be fitting poorly. Our "Global Coverage Test" (GCT) can distinguish an arbitrarily misspecified model from the true conditional density of the sample, while our "Local Coverage Test" (LCT) can pinpoint where specifically in the feature space the model is fitting poorly. Furthermore, we develop "Amortized Local P-P plots" (ALP), which efficiently provide simple, interpretable graphical summaries of distributional differences at any location $\mathbf{x}$ in the feature space. These diagnostics provide rich insight into model quality in simple, explainable terms like coverage, bias, dispersion, and multimodality in y as a function of $\mathbf{x}$. Our validation procedures scale to high dimensions and can potentially adapt to any type of data at hand. We demonstrate the effectiveness of our tests and local plots through simulated experiments including a prediction problem for high-dimensional image data.

In the second part of this thesis, we describe how we can leverage the practical goodness-of-fit diagnostics described above to *calibrate* conditional densities to be accurate across all locations $\mathbf{x}$. In many scientific contexts, the calibrated conditional PDFs themselves are of vital interest. Additionally, we can use calibrated conditional PDFs to construct calibrated prediction sets with accurate conditional coverage. Note that our diagnostic framework, which quantify deviations between actual and nominal coverage in y at any given $\mathbf{x}$ for an initial CDE $\widehat{f}(y|\mathbf{x})$, is directly measuring the *achieved conditional coverage* of the initial model. Therefore, these diagnostics can guide us in directly targeting (and fixing) what is wrong with the conditional coverage. Compared to other methods that do not directly try to optimize for conditional validity directly, we can achieve better empirical results. We also observe that if regression estimates for coverage are consistent, then our method should achieve asymptotic conditional validity and asymptotic conditional efficiency, even if the initial CDE $\widehat{f}(y|\mathbf{x})$ is not consistent. We benchmark our corrected prediction bands against oracle bands and state-of-the-art predictive inference algorithms for synthetic data, including settings with a distributional shift. Furthermore, we produce calibrated predictive distributions for two scientific applications: probabilistic nowcasting based on sequences of satellite images, and estimation of galaxy distances based on imaging data (photometric redshifts).

In the third part of this thesis, we explore an extension of the calibration method that hybridizes it with local conformal inference. Instead of directly using a different regression of PIT scores against $\mathbf{X}$ at every different location in feature space, we partition the feature space and estimate a single conformal correction within each partition element. This alternative framework, with the appropriate data splitting scheme, is able to achieve certain finite-sample coverage guarantees that are characteristic of conformal prediction methods (see Section 2.4 for an overview of conformal prediction), at the expense of requiring an additional split of the data. We also observe that if regression estimates for coverage are consistent, then our method should achieve asymptotic conditional validity. We demonstrate the practical advantages and disadvantages of this approach on a simulated experiment as well as an application to photometric redshift estimation.

This thesis is structured as follows. Chapter 2 provides an in-depth review of relevant literature, beginning broadly with the general problem of uncertainty quantification, then focusing on related work on assessing quality of conditional density models, quantile regression, and conformal inference. Chapter 3 presents our method for diagnostics of conditional density models, explains how it improves upon the limitations of previous diagnostics, and highlights its properties using two stylized examples. Chapter 4 presents our method for calibrating conditional density models using this diagnostic framework, describes its theoretical properties, and highlights its practical properties on two stylized examples as well as two applications to real-world scientific problems. Chapter 5 presents our hybrid conformalized calibration approach, describes its theoretical properties, and compares its practical properties with our original calibration framework using a stylized example and a real-world application. Finally, Chapter 6 provides a summary and discussion of the thesis and potential future work.

Chapter 2

Literature Review

This chapter provides a review of existing work relevant to this thesis. We start with an overview of the literature on uncertainty quantification, highlighting the recent interest in the machine learning community in estimating full conditional densities or predictive distributions $f(y|\mathbf{x})$. We then examine the challenges of assessing the quality of these estimated conditional densities and the limitations of existing methods. Next, we describe quantile regression methods, a popular related approach to uncertainty quantification that estimates conditional quantile functions instead of conditional densities. Finally, we give an overview of conformal prediction methods, which share some of the goals of our work while starting from a different framework.

2.1 Uncertainty Quantification for Predictive Inference

The term "uncertainty quantification" is often used as an umbrella term to describe all approaches that go beyond using point estimation of a variable of interest to assess the predictive accuracy of models (Berger and Smith, 2019; Abdar et al., 2021). In scientific applications, uncertainty quantification is sometimes more important than point predictions, as often the goal is to construct a probabilistic forecast, which takes the form of a predictive probability distribution over future quantities or events of interest (Gneiting and Katzfuss, 2014; Chen et al., 2022). In engineering and finance, uncertainty quantification can also be essential for decision-making tasks like optimizing supply chains for actual demand (Farmer, 2017; Göttlich and Knapp, 2020). In this work, we consider the problem of assessing the uncertainty about a continuous response or "target" variable $Y \in \mathbb{R}$ given input features or covariates $\mathbf{X} \in \mathcal{X}$.

Figure 2.1 shows a simple illustration of why point estimation of a continuous response Y given a feature X may be insufficient. For some values of x, the conditional distribution of Y is a unimodal Gaussian, while in other parts of the feature space the conditional distribution of Y is a bimodal Gaussian. Point estimation of the conditional mean or median would simply predict $Y = 0$ for every input value of x, and merely including an estimate of the conditional variance (or a confidence interval) is not enough to fully describe the behavior of $f(y|x)$.

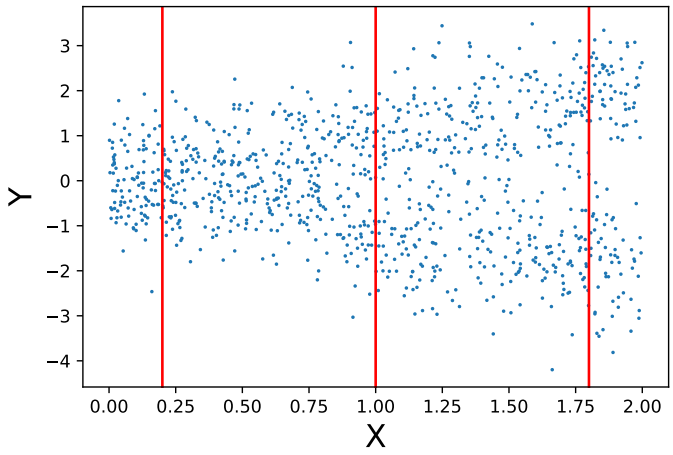


Figure 2.1: Scatterplot of the joint distribution of $f(x, y)$ reveals that the conditional density $f(y|x)$ is unimodal for some values of x and bimodal for other values of x, as highlighted by vertical red lines at $x = 0.2$, $x = 1.0$, and $x = 1.8$. Point estimation of the conditional mean of $y|x$, even accompanied by conditional variance estimates, is insufficient for fully describing the behavior of $f(y|x)$.

The literature on uncertainty quantification for predictive inference is vast, but much existing research, such as the burgeoning field of conformal prediction, is concerned with constructing calibrated prediction intervals and prediction sets only, rather than obtaining and assessing entire predictive distributions that are well-calibrated. It is often believed that the latter problem is too challenging. While prediction sets are certainly useful in quantifying uncertainties in predictive models, we have been witnessing a transformation across scientific disciplines in terms of interest in the entire predictive distribution of Y given $\mathbf{x}$. For instance, see Gneiting (2008) for a discussion of probabilistic forecasting in weather predictions, Timmermann (2000) for a survey of conditional density estimation in financial risk management, Alkema et al. (2007) for applications of computing Bayesian posteriors in epidemiological projections, and Mandelbaum et al. (2008); Malz and Hogg (2022) for the importance of predictive distributions for astrophysical studies.

To meet this demand for more precise uncertainty quantification across scientific fields, recently a large body of work in machine learning has been developed for estimating conditional densities $f(y|\mathbf{x})$ for all possible values of $\mathbf{x}$, or to generate predictions that follow the unknown conditional density. Standard approaches for directly estimating the conditional density functions $f(y|\mathbf{x})$ using neural networks include mixture density networks (Bishop, 1994) and kernel mixture networks (Ambrogioni et al., 2017). Alternatively, one can quantify uncertainty by training multiple models on the same data; for instance, by ensembling neural networks (Lakshminarayanan et al., 2017; Pearce et al., 2020) or using dropout methods (Srivastava et al., 2014; Gal and Ghahramani, 2016) that randomly drop units from the network during training. Another approach is Bayesian neural networks (MacKay, 1992; Neal, 2012; Goan and Fookes, 2020), which incorporate priors on the neural network weights and output posterior predictive distributions

instead of point predictions. Much recent work has improved the quality, flexibility, and efficiency of Bayesian neural networks (Graves, 2011; Blundell et al., 2015; Louizos and Welling, 2017; Pawlowski et al., 2017).

Additional approaches include normalizing flows, a model class that includes neural autoregressive models (Uria et al., 2014; Sohn et al., 2015; Papamakarios et al., 2019; Kobyzev et al., 2021) and Gaussian process CDEs (Dutordoir et al., 2018). Furthermore, one can use simpler nonparametric CDE methods (Izbicki and Lee, 2016; Izbicki et al., 2017; Dalmasso et al., 2020), as well as implicit CDE methods that encode the predictive distribution implicitly, such as conditional generative adversial networks (Mirza and Osindero, 2014)) and quantile regression methods that estimate all quantiles simultaneously (Chung et al., 2021; Fasiolo et al., 2021; Tagasovska and Lopez-Paz, 2019; Amerise, 2018; Liu and Wu, 2011). In addition, with the advent of high-precision data and simulations, simulation-based inference (Cranmer et al., 2020) has also played a growing role in disciplines ranging from physics, chemistry and engineering to the biological and social sciences. Advances in simulation-based-inference have led to the development of machine-learning based methods to learn an explicit surrogate model of the posterior (Marin et al., 2016; Papamakarios and Murray, 2016; Lueckmann et al., 2017; Chen and Gutmann, 2019; Izbicki et al., 2019; Greenberg et al., 2019).

Though there are numerous ways one can obtain predictive distributions, the models are only useful in practice if they are approximately *individually or conditionally calibrated*, meaning that the estimated cumulative distribution function (CDF)

$$\widehat{F}(y|\mathbf{x}) \approx F(y|\mathbf{x}) \text{ for all } y \in \mathbb{R} \text{ at every } \mathbf{x} \in \mathcal{X}.$$

In words, the predicted conditional probability of an event happening given input $\mathbf{x}$ should match its observed probability. Instance-wise uncertainties are crucial in practical applications. For example, weather forecasts may predict the probability of rainfall given the current state of environmental predictors. Similarly, medical research may estimate the efficacy of a drug for individuals of specific demographics after taking a given dose. Achieving instance-wise uncertainties can be important for algorithmic fairness so as not to over- or under-predict risks for certain groups of individuals (Kleinberg et al., 2016; Zhao et al., 2020).

When one obtains an entire predictive distribution that is conditionally calibrated, one can also derive various quantities of interest, such as conditional moments, prediction intervals, or even more general prediction sets. Figure 2.2 illustrates Highest Probability Density (HPD) sets, which include all values of y for which the estimated conditional density $\widehat{f}(y|\mathbf{x})$ exceeds a certain threshold. HPD sets are more general than intervals, in that they need not be contiguous. By construction, individually well-calibrated predictive distributions $\widehat{F}(Y|\mathbf{X})$ lead to conditionally valid prediction sets. Indeed, if $C_\alpha(\mathbf{X})$ is a prediction set derived from $\widehat{F}(Y|\mathbf{X})$ with nominal coverage $1-\alpha$, then having individually calibrated $\widehat{F}(Y|\mathbf{X})$ implies that

$$\mathbb{P}(Y \in C_\alpha(\mathbf{X})|\mathbf{X} = \mathbf{x}) = 1-\alpha\,,\ \forall \mathbf{x} \in \mathcal{X}. \tag{2.1}$$

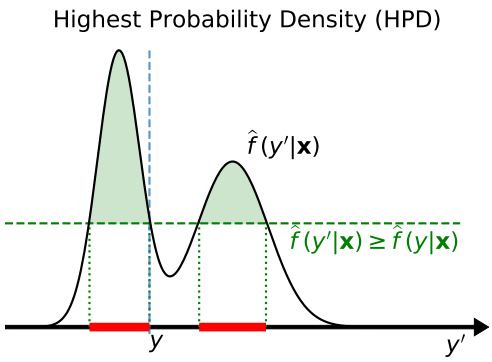


Figure 2.2: The Highest Probability Density (HPD) regions defined by a specific value y for input features $\mathbf{x}$ include all values $y' \in \mathbb{R}$ for which the estimated conditional density $\widehat{f}(y'|\mathbf{x})$ exceeds the threshold defined by $\widehat{f}(y|\mathbf{x})$. Note that these regions, highlighted in red in this example, need not form a single contiguous interval.

Unfortunately, off-the-shelf methods for obtaining predictive distributions, even the numerous complex ones cited above, are usually far from being calibrated. The fundamental issue is that conditional density estimators are typically fitted by minimizing a loss function, such as the Kullback-Leibler divergence or integral probability metrics, that do not directly depend upon calibration (Papamakarios et al., 2019; Dalmasso et al., 2020; Rothfuss et al., 2019). This is especially true of large machine learning models, such as deep generative autoregressive models or Bayesian networks. Loss functions are useful for training models but only provide relative comparisons of overall model fit. Hence, a practitioner may not know whether he or she should keep looking for better models (by e.g., using larger training samples, training times, etc.), or if the current estimate is "close enough". For example, none of the above Bayesian neural network methods provide frequentist coverage guarantees, and empirically they can often generate prediction sets with poor coverage (Yao et al., 2019).

2.2 Assessing Quality of Conditional Density Models

To ensure that an estimated predictive distribution is well-calibrated, we would naturally seek statistical tests or diagnostics that indicate how well the model fits on a held-out data sample. However, existing approaches to goodness-of-fit testing are usually not tailored for assessing conditional density models, and do not provide the kind of granular insight we desire.

One line of work assesses goodness-of-fit of a conditional density model via a two-sample test that compares samples from $\widehat{f}(y|\mathbf{x})$ and $f(y|\mathbf{x})$. Earlier tests involved a conditional version of the standard Kolmogorov test (Andrews, 1997; Zheng, 2000) in one dimension, or were tailored to specific families of conditional densities (Stute and Zhu, 2002; Moreira, 2003). Recently, Jitkrittum et al. (2020) developed a fast kernel-based approach that can also identify local regions of poor fit. Kernel approaches also require the user to specify an appropriate kernel and

tuning parameters, which can be challenging in practice. In general, we note that while these two-sample tests are useful for deciding whether or not an estimated predictive distribution needs to be improved, they do not provide any means to actually correct discrepancies. That is, while these tests are consistent, they do not provide insight on how the distributions of $\widehat{f}(y|\mathbf{x})$ and $f(y|\mathbf{x})$ differ locally.

Another approach which does allow for recalibration of predictive distributions is to assess how the marginal distribution of probability integral transform (PIT) values differs from a uniform distribution (Cook et al., 2006; Freeman et al., 2017; Talts et al., 2018; D'Isanto and Polsterer, 2018) and apply corrections to bring them into agreement (Bordoloi et al., 2010). Figure 2.3 shows that the probability integral transform at location $\mathbf{x}$ for value y measures the probability mass of the estimated conditional density $\widehat{f}(\cdot|\mathbf{x})$ that lies below y.

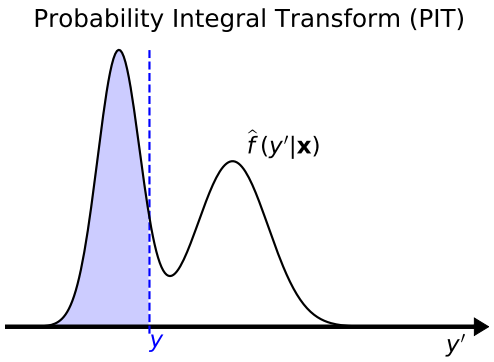


Figure 2.3: The probability integral transform (PIT) measures the amount of probability mass that lies below a certain threshold. If a conditional density model $\widehat{f}(y|\mathbf{x})$ is well-calibrated, then PIT values computed using this model on an i.i.d. sample from $F(\mathbf{x}, y)$ should be uniformly distributed on average.

This approach does provide us with diagnostics that describe the nature of inconsistencies between $\widehat{f}(y|\mathbf{x})$ and $f(y|\mathbf{x})$. If, for example, the marginal distribution of PIT values contains too many large values relative to the uniform distribution, this indicates that the estimated $\widehat{f}(y|\mathbf{x})$ models are on average positively biased, and all of them should be adjusted downwards. While informative, these diagnostics were originally developed for assessing *unconditional* density models (Gan and Koehler, 1990). As such, they are known to fail to detect some clearly misspecified conditional models including models that ignore the dependence on the covariates altogether (Schmidt et al., 2020). This is because they are only testing for a form of overall coherence between a data-averaged conditional (posterior) distribution and its marginal (prior) distribution. In other words, they only assess average or marginal calibration over the entire distribution of $\mathbf{X} \in \mathcal{X}$. Average calibration is often simply referred to as just "calibration" (Naeini et al., 2015; Guo et al., 2017; Kuleshov et al., 2018), although it is a well-known problem that one can achieve marginally calibrated distributions, $\mathbb{E}_{\mathbf{X}\sim F_X}\left[\widehat{F}\left(y|\mathbf{x}\right)\right] = \mathbb{E}_{\mathbf{X}\sim F_X}\left[F\left(y|\mathbf{x}\right)\right]$, which completely ignore the input $\mathbf{x}$.

In particular, the marginal distribution of PIT values obtained from a model $\widehat{f}(y|\mathbf{x})$ will be uniformly distributed if $\widehat{f}(y|\mathbf{x}) = f(y)$, even though $\widehat{f}(y|\mathbf{x})$ is a terrible estimator that has not learned anything about the conditional

distribution of y given $\mathbf{x}$, and is simply asserting the marginal distribution $f(y)$ everywhere while ignoring $\mathbf{x}$. Figure 2.4, taken from Figure 2 of Schmidt et al. (2020), shows a real-world application where existing diagnostics used by scientists are susceptible to precisely this sort of problem. The figure compares various methods that astronomers have developed for estimating the predictive distribution of galaxy distances based on imaging data (photometric redshifts), a problem known as photo-z estimation. The method in the bottom right, "TrainZ", is a pathological model developed by the authors that simply asserts the empirical marginal $\widehat{f}(y)$ everywhere instead of trying to learn the true function $f(y|\mathbf{x})$. The distribution of marginal PIT values, as visualized with Q-Q plots and histograms, looks perfectly uniform for TrainZ. Thus, existing PIT diagnostics would indicate that the model fits well, even though we know the model is completely misspecified.

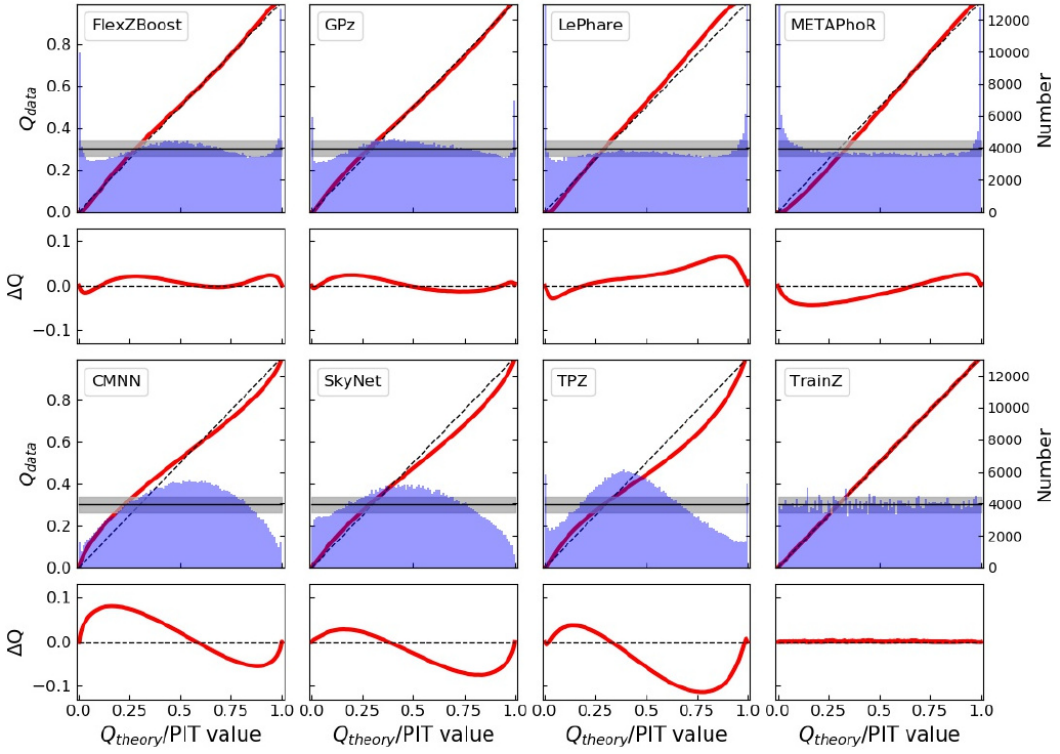


Figure 2.4: Taken from Figure 2 of Schmidt et al. (2020). Each of the eight panels shows, for a different CDE method, the QQ plot (red) and histogram (blue) of the marginal PIT distributions based on the estimated photo-z PDFs on a held-out data sample, along with the ideal QQ (black dashed diagonal) and ideal uniform histogram (gray horizontal) curves. The lower inset in each panel shows a difference plot for the QQ deviation from the ideal diagonal. "TrainZ", the bottom right panel, is a misspecified model that simply predicts the empirical marginal $\widehat{f}(y)$. A precise diagnostic method should be able to easily tell that TrainZ fits the data poorly. However, existing diagnostics based on the marginal distribution of PIT values fail to detect any problems with TrainZ, as the distribution is perfectly uniform.

More generally, inconsistencies in various regions of the feature space can cancel out when looked at as an ensemble (Jitkrittum et al., 2020; Luo et al., 2021). In other words, the local distribution of PIT values can be far from

uniform in various parts of the feature space, while the overall marginal distribution of PIT values is still uniform. For a stylized illustration of this, see Figure 2.5.

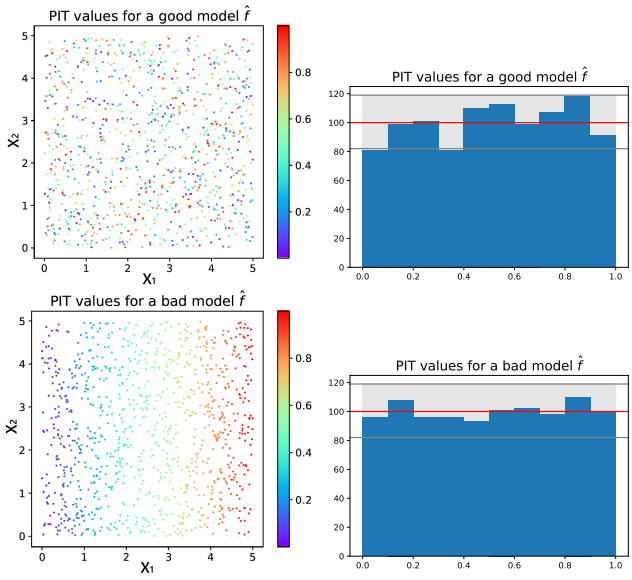


Figure 2.5: *Top row:* For a well-calibrated "good" model, PIT values computed anywhere in the feature space (consisting of X_1, X_2 in this example) have distribution $Unif[0, 1]$, and when aggregated across the feature space are also uniform. Existing diagnostics based on the marginal uniformity of PIT values will correctly indicate that this model fits well. *Bottom row:* This "bad" model fits the data poorly everywhere in the feature space; PIT values are too low in some places, too high in others, and never well-calibrated locally. However, when aggregated across the feature space, these errors cancel out, yielding a uniform marginal PIT distribution. Thus, existing diagnostics cannot determine that there is anything wrong with this clearly misspecified model.

By construction, any recalibration scheme based on the marginal distribution of PIT values would only fit the data on average, and thus only improve marginal calibration. The existing literature does not provide a method for recalibrating CDEs towards individual or conditional calibration of the predictive distributions.

2.3 Quantile Regression

A popular alternative to directly estimating the conditional density is instead to estimate conditional quantiles, which is a way of implicitly estimating the conditional CDF. This approach is known as quantile regression (Koenker and Bassett Jr., 1978; Koenker and Hallock, 2001). In a regression setting, being able to specify the full quantile function is an appealing goal; a model that predicts the true conditional quantiles $F^{-1}(\alpha|\mathbf{x})$ of Y, for each input $\mathbf{X}$ at all quantile levels $\alpha \in [0, 1]$, presents a full and efficient account of the uncertainty in predicting Y given $\mathbf{X}$. Quantiles are easily interpretable, and can directly be used to construct prediction intervals. Indeed, the conditional quantiles of $Y|\mathbf{X} = \mathbf{x}$ are perhaps the most natural way to produce intervals satisfying conditional coverage (Equation 2.1). Given the true conditional quantiles, $F^{-1}(0.5\alpha|\mathbf{x})$ and $F^{-1}(1 - 0.5\alpha|\mathbf{x})$, we can simply build the $(1 - 0.5\alpha)$-level oracle prediction

interval

$$C^*_\alpha(x) = \left[F^{-1}(0.5\alpha|\mathbf{x}), F^{-1}(1-0.5\alpha|\mathbf{x})\right]. \tag{2.2}$$

The true conditional quantiles are unknown but can be estimated with quantile regression, yielding the prediction interval

$$C_\alpha(x) = \left[\widehat{F}^{-1}(0.5\alpha|\mathbf{x}), \widehat{F}^{-1}(1-0.5\alpha|\mathbf{x})\right]. \tag{2.3}$$

To achieve this, standard quantile regression methods focus on optimizing the so-called pinball loss, a tilted transformation of the absolute value function. Given a target y, an estimate $\widehat{y}$, and quantile level $\tau \in [0,1]$, the pinball loss is defined as

$$L_\tau(y, \widehat{y}) = (\widehat{y} - y)\left(\mathbb{I}(\widehat{y} > y) - \tau\right) \tag{2.4}$$

Note that when setting $\tau = 1/2$, Equation 2.4 becomes the L^1 norm that is used to estimate the conditional median of Y given $\mathbf{X}$. By setting $\tau \neq 1/2$, we get a tilted version that instead estimates the specified conditional quantile τ of Y given $\mathbf{X}$.

Because of the flexibility of the pinball loss, quantile regression can be used to model complex distributions in a non-parametric way, yielding prediction intervals that are adaptive to heteroscedasticity without requiring parametric assumptions (Takeuchi et al., 2006; Zhou and Portnoy, 1996). Moreover, as mentioned, quantiles provide an attractive representation for uncertainty because they are directly interpretable with units in the target output space, allowing for easy construction of prediction intervals. Quantiles can also be used to efficiently sample from the predictive distribution via inverse transform sampling (Hao and Naiman, 2007).

Despite the numerous appealing properties of quantile regression, these methods face the same fundamental limitations as models that entire predictive distribution of Y given $\mathbf{X}$, when it comes to the question of individual or conditional calibration. Namely, the prediction intervals $C_\alpha(\mathbf{X})$ in Equation 2.3 for various $\mathbf{x}$ might not satisfy the conditional coverage statement in Equation 2.1. While these intervals do converge in the large-sample limit to the oracle intervals $C^*_\alpha(\mathbf{X})$ (Koenker and Bassett Jr., 1978; Taylor and Bunn, 1999), frequentist coverage of these intervals is only guaranteed for specific model classes, under certain regularity and asymptotic conditions (Takeuchi et al., 2006; Meinshausen, 2006). Even though $C^*_\alpha(\mathbf{X})$ satisfies Equation 2.1 in theory, quantile regression with the standard pinball loss has been empirically shown to sometimes yield highly miscalibrated quantile estimates with incorrect conditional coverage in practice (Chung et al., 2021; Feldman et al., 2021).

Recent alternatives to standard quantile regression have been proposed, with the goal of achieving better calibration. Conformalized quantile regression (Romano et al., 2019) combines the standard method with conformal prediction (to be discussed in Section 2.4), in order to guarantee marginal calibration of prediction intervals. They achieve this using a custom conformity score function based on the learning the 0.5α and $1 - 0.5\alpha$ quantiles. As previously discussed, marginal calibration is no guarantee of individual calibration, especially when using the standard pinball loss. Another method proposed by Chung et al. (2021) empirically estimates conditional quantiles

from a kernel density conditional density estimator, then fits a regression model to interpolate between the estimated quantiles. Conditional calibration of this method relies on the consistency of this CDE, and hence reintroduces the familiar problem of how to assess and ensure the quality of conditional density models. Orthogonal quantile regression (Feldman et al., 2021) amends the pinball loss with an orthogonality loss function to promote independence between the size of prediction intervals and their achieved conditional coverage, the idea being that these two quantities are independent for the true conditional quantiles. However, this orthogonality loss is an indirect metric that is not the same as actually optimizing for conditional coverage. We show empirically in Section 4.4 why using the orthogonality loss does not necessarily lead to individual calibration. The bottom line here is that quantifying predictive uncertainty in ways that achieve conditional coverage in practice remains a challenging open problem, whether one tackles it by directly estimating CDEs or by estimating conditional quantile functions.

We also note that while estimating full predictive distributions yields all quantiles by construction, quantile regression does not directly yield the full predictive distributions. On the one hand, recent work in quantile regression has improved the ability to train for all quantiles simultaneously (Rodrigues and Pereira, 2020; Tagasovska and Lopez-Paz, 2019), and in principle one can invert and differentiate the quantile functions in order to obtain the full predictive distributions. Indeed, under certain regularity assumptions, estimates of conditional quantile functions using the standard pinball loss are proven to be asymptotically consistent (Steinwart and Christmann, 2011). However, in practice for finite samples, quantile crossing is major concern. Quantile crossing, commonly observed when modeling multiple quantiles simultaneously, is the failure of the estimated conditional quantile function to obey the required monotonicity constraint. In other words, for some values $\alpha_1 > \alpha_2$, there exists $\mathbf{x}$ such that

$$\widehat{F}^{-1}(\alpha_1|\mathbf{x}) < \widehat{F}^{-1}(\alpha_2|\mathbf{x})$$

Mitigating quantile crossing is a difficult problem and is still an active area of research, with recent proposed approaches including sorting-based post-processing at the end (Chernozhukov et al., 2010; Kim et al., 2021), expensive constrained optimization (Liu and Wu, 2009), or some linear or non-linear monotonic interpolation between quantile estimates (Gasthaus et al., 2019; Park et al., 2022). Despite theoretical consistency, because quantile regression is not guaranteed to be conditionally calibrated (or even monotonic) in practice, the implicitly estimated predictive distributions are unlikely to be of high quality.

2.4 Conformal Inference

In this section, we give a brief overview of conformal inference (Vovk et al., 1999), a framework for uncertainty quantification that is primarily appealing for two reasons: (i) it provides a mathematical guarantee of marginal calibration in finite samples, and (ii) it requires minimal distributional assumptions on the data-generating process, and therefore can be applied to a range of black box models. Consider a setting with covariates $\mathbf{X} \in \mathcal{X}$ and a

continuous response $Y \in \mathbb{R}$ drawn from a joint distribution $F_{\mathbf{X},Y}$. Conformal prediction methods have the special property of yielding prediction sets $C_\alpha(\mathbf{X})$ with finite-sample marginal validity; that is,

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}(Y \in C_\alpha(\mathbf{X})) = 1 - \alpha, \tag{2.5}$$

as long as the data are exchangeable (Vovk et al., 2005).

One creates conformal prediction sets by evaluating potential candidate values y, and including them in the set if they are not too extreme relative to previous observations. The measure of extremeness comes from a conformal or non-conformal score. A conformal score has smaller values indicating more extremeness; a common example is the conditional density estimate $\widehat{f}(y|\mathbf{x})$. A non-conformal score has larger values indicating more extremeness; a common example is the quantity $|y - \mu(\mathbf{x})|$ defined using a regression function $\mu(\mathbf{x})$. (Multiplying by -1 would convert a conformal score to a non-conformal score, and vice-versa.) For brevity, we refer to either conformal or non-conformal scores as "conformity scores" for the rest of this thesis.

The original, or full, conformal procedure often requires conformity scores to be calculated an impractically large number of times (Lei et al., 2018), so we focus on inductive or split conformal prediction (Papadopoulos et al., 2002), which is more efficient at the cost of splitting the data. After defining a conformity score function $cs(\cdot)$, split conformal prediction decides whether a candidate value y should be included in a prediction region for $\mathbf{x}$ in three steps. First, split the observed data into disjoint training and calibration sets, which we denote $\mathcal{D}_{train}$ and $\mathcal{D}_{cal}$. Second, build a prediction model using the training set that is used to calculate conformity scores, $cs\left(y_i|\mathbf{x}_i\right)$, for observations in the calibration set $\mathcal{D}_{cal}$. Third, for a test point $\mathbf{x}_{n+1}$, evaluate candidate values $\widehat{y}_{n+1}$, and include the value in the output set if the conformity score $cs\left(\widehat{y}_{n+1}|\mathbf{x}_{n+1}\right)$ is less extreme than $100\alpha\%$ of the conformal scores for the calibration set. Algorithm 1 formally presents the split conformal approach.

Why is this procedure able to mathematically guarantee finite-sample marginal validity? It results from the assumption of exchangeability of points in the calibration set and the new test point. Given this exchangeability, the rank of the conformal score for the true observation y_{n+1} with respect to the conformal scores from the calibration set is distributed uniformly between 1 and $1 + |\mathcal{D}_{cal}|$. Therefore, including all candidate values $\widehat{y}_{n+1}$ that are less extreme than $100\alpha\%$ of the conformal scores for the calibration set will by construction contain the true point y_{n+1} with probability $1 - \alpha$ as desired.

Obtaining marginal calibration in finite samples, under minimal parametric assumptions on the data generating process, is the hallmark of conformal prediction. None of the various methods we have previously considered in this chapter (except those that incorporate the conformal framework) provide such a mathematical guarantee. However, there is no guarantee that conformal prediction sets will satisfy conditional calibration (Equation 2.1), even approximately. Indeed, it is provably impossible to obtain exact conditional validity in finite samples, unless one imposes strong assumptions on the underlying distribution, which goes against the "distribution-free" spirit of the method (Vovk, 2012; Lei and Wasserman, 2014; Barber et al., 2020).

Algorithm 1 Split Conformal Prediction

Input: Conformity score function $cs(\cdot)$; significance level α; data points $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$ split into disjoint training and calibration sets $\mathcal{D}_{train}$ and $\mathcal{D}_{cal}$; new test point x_{n+1}; candidate value $\widehat{y}_{n+1}$.
Task: Decide if $\widehat{y}_{n+1}$ should be included in a $(1-\alpha)$-level prediction set for $\mathbf{x}_{n+1}$.
Algorithm:

1. Train conformity score function $cs(\cdot)$ on training set $\mathcal{D}_{train}$.
2. Compute conformity scores on the calibration dataset:
   For all $(\mathbf{x}_j, y_j) \in \mathcal{D}_{cal}$, set
   $$t_j = cs((\mathbf{x}_j, y_j))$$
3. Compute conformity score for the new observation:
   $$t_{n+1} = cs((\mathbf{x}_{n+1}, \widehat{y}_{n+1}))$$
4. Compute
   $$\tau = \frac{\#\left\{(\mathbf{x}_j, y_j) \in \mathcal{D}_{cal} : t_j \geq t_{n+1}\right\}}{1 + |\mathcal{D}_{cal}|}$$
5. Include $\widehat{y}_{n+1}$ in the new prediction region if $\tau > \alpha$.

In light of this, recent efforts in conformal prediction have attempted to achieve approximate conditional validity by designing conformity scores with an approximately homogeneous distribution across the feature space $\mathcal{X}$. For example, `reg-split` (Lei et al., 2018) constructs conformity scores using the residuals (evaluated on $\mathcal{D}_{cal}$) from a regression model (trained on $\mathcal{D}_{train}$), and may use a second regression to account for heteroskedasticity. Another method `dist-split` (Izbicki et al., 2020) constructs conformity scores using empirical PIT scores (evaluated on $\mathcal{D}_{cal}$) based on a conditional density model (trained on $\mathcal{D}_{train}$). In the large-sample limit, if the models used to build conformity scores are statistically consistent, then the conformity scores will be homogeneously distributed across (that is, independent of) $\mathcal{X}$, and the prediction sets will achieve conditional coverage. This property is known as asymptotic conditional validity.

Unfortunately, it is difficult to verify whether these methods provide good conditional coverage in practice for finite samples. We also note that conformal prediction sets are not conditionally valid, even asymptotically, if the initial model for assigning conformity scores is misspecified. Furthermore, conformal prediction methods cannot obtain, even implicitly, full predictive distributions. Therefore, in the context of this thesis, which seeks to assess and recalibrate entire predictive distributions, conformal prediction is best viewed as an adjacent field with some shared goals and insights but also key differences.

Chapter 3

Local Diagnostics for Conditional Density Models

This chapter presents a diagnostic framework for identifying, locating, and interpreting the nature of statistically significant discrepancies between a fitted conditional density model $\widehat{f}(y|\mathbf{x})$ and the true conditional density $f(y|\mathbf{x})$ over the entire feature space (refer also to our published work, Zhao et al. (2021)). These diagnostic tools provide insight into fundamental properties such as coverage, bias, dispersion, and multimodality in y (output of interest) as a function of $\mathbf{x}$ (observed inputs). Existing diagnostics for conditional density models, described in Section 2.2, are known to be unable to detect every kind of misspecified model, and furthermore they do not give insight into local quality of fit of the model at any given $\mathbf{x}$. Our proposed method is sensitive enough to detect arbitrarily misspecified models. Moreover, because our method quantifies deviations between actual and nominal coverage in y as a function of $\mathbf{x}$, it is able to assess and visualize quality of fit anywhere in the feature space, even at points without observed data, in terms of easy-to-explain diagnostics. Having interpretable diagnostics is crucial for scientific collaborators and end users to build trust in large machine learning models.

Section 3.1 provides a theoretical framework for our problem, and details different failure modes of previous diagnostic methods. Section 3.2 explains how our new diagnostics address these limitations to achieve desirable properties. Section 3.3 describes the procedure for testing local and global consistency of CDEs, while Section 3.4 describes how to construct local plots that visualize the nature of deviations from local consistency in different parts of the feature space $\mathcal{X}$. Section 3.5 explores a stylized example that demonstrates the favorable properties of our method, and Section 3.6 applies our method to a high-dimensional example with image data. Finally, Section 3.7 discusses an extension of our framework to handle multivariate responses. All code used to produce our experiments is available at `https://github.com/zhao-david/CDE-diagnostics`. We have also included an installable Python

package `cde-diagnostics` with a detailed tutorial.

Notation. Let $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ denote an i.i.d. sample from $F_{\mathbf{X},Y}$, the joint distribution of $(\mathbf{X}, Y)$ for a random variable $Y \in \mathcal{Y} \subseteq \mathbb{R}$ (in Section 3.7, Y is multivariate), and a random vector $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$. Let $\widehat{f}(y|\mathbf{x})$ denote a conditional density estimate (CDE) for which we seek goodness-of-fit diagnostics. In a prediction setting, $\mathcal{D}$ represents a hold-out "calibration" set not used to train $\widehat{f}(y|\mathbf{x})$. In a Bayesian setting, Y represents the parameter of interest (sometimes also denoted as θ), and each element of $\mathcal{D}$ is obtained by first drawing Y_i from the prior distribution, and then drawing $\mathbf{X}_i$ from the statistical model of $\mathbf{X}|Y_i$.

3.1 Previous Diagnostics were Insensitive to Covariate Transformations

Prior to our work, there did exist diagnostic methods that could describe the nature of inconsistencies between $\widehat{f}(y|\mathbf{x})$ and $f(y|\mathbf{x})$ (see Section 2.2 for a more in-depth discussion). But these previous diagnostics only tested for a form of overall coherence between a data-averaged conditional (or posterior) distribution and its marginal (or prior) distribution. Typically, they relied on computing probability integral transform (PIT) values (Cook et al., 2006; Freeman et al., 2017; Talts et al., 2018; D'Isanto and Polsterer, 2018). While informative, these diagnostics were originally developed for assessing *unconditional* density models (Gan and Koehler, 1990). As such, they are known to fail to detect clearly misspecified conditional models, including models that ignore the dependence of the response on the covariates altogether (Schmidt et al., 2020).

Ideally, a test should be able to distinguish *any* given alternative conditional density model $\widehat{f}(y|\mathbf{x})$ from the true density $f(y|\mathbf{x})$, as well as locate discrepancies in the feature space $\mathcal{X}$. More precisely, a test should be able to identify what we in this section define as global and local consistency.

Definition 1 (Global Consistency). *An estimate $\widehat{f}(y|\mathbf{x})$ is globally consistent with the density $f(y|\mathbf{x})$ if the following null hypothesis holds:*

$$H_0 : \widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x}) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } y \in \mathcal{Y}. \tag{3.1}$$

Note that $\widehat{f}(y|\mathbf{x})$ is a particular fixed conditional density estimate, and we test whether samples from $\widehat{f}(y|\mathbf{x})$ are consistent with samples from $f(y|\mathbf{x})$.

Previous diagnostics typically validated density models by computing PIT values on an independent calibration dataset, which was not used to estimate $\widehat{f}(y|\mathbf{x})$:

Definition 2 (PIT). *Fix $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$. The probability integral transform of y at $\mathbf{x}$, as modeled by the conditional density estimate $\widehat{f}(y|\mathbf{x})$, is*

$$\textit{PIT}(y; \mathbf{x}) = \int_{-\infty}^{y} \widehat{f}(y'|\mathbf{x}) dy'. \tag{3.2}$$

See Figure 2.3 for an illustration of this calculation.

Remark 1. *For implicit models of $\widehat{f}(y|\mathbf{x})$ (that is, generative models that via e.g. MCMC can sample from, but not directly evaluate $\widehat{f}$), we can approximate the PIT values by forward-simulating data: For fixed $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, draw $Y_1, \ldots, Y_L \sim \widehat{f}(\cdot|\mathbf{x})$. Then, approximate PIT$(y; \mathbf{x})$ via the cumulative sum $L^{-1} \sum_{i=1}^{L} \mathbb{I}(y_i \leq y)$.*

If the conditional density model $\widehat{f}(y|\mathbf{x})$ is globally consistent, then the PIT values are uniformly distributed. More precisely, if H_0 (Equation 3.1) is true, then the random variables

$$\text{PIT}(Y_1; \mathbf{X}_1), \ldots, \text{PIT}(Y_n; \mathbf{X}_n) \overset{i.i.d.}{\sim} \text{Unif}(0, 1).$$

This result was often used to test goodness-of-fit of conditional density models in practice (Cook et al., 2006; Bordoloi et al., 2010; Tanaka et al., 2018).

Unfortunately, such random variables can be uniformly distributed even if global consistency does not hold. This is shown in the following theorem, which essentially details different failure modes of previous diagnostics.

Theorem 1 (Insensitivity to Covariate Transformations). *Suppose there exists a function $g : \mathcal{X} \longrightarrow \mathcal{Z}$, where $\mathcal{Z} \subseteq \mathbb{R}^k$ for some k, that satisfies*

$$\widehat{f}(y|\mathbf{x}) = f(y|g(\mathbf{x})). \tag{3.3}$$

Let $(\mathbf{X}, Y) \sim F_{\mathbf{X},Y}$. Then PIT$(Y; \mathbf{X}) \sim$ Unif$(0, 1)$.

Many models naturally lead to estimates that could satisfy the condition in Equation 3.3, even without being globally consistent. In fact, clearly misspecified models $\widehat{f}(y|\mathbf{x})$ can yield uniform PIT values and "pass" an associated goodness-of-fit test regardless of the sample size. For example, if $\widehat{f}(y|\mathbf{x})$ is based on a linear model, then $\widehat{f}(y|\mathbf{x})$ will by construction depend on $\mathbf{x} \in \mathbb{R}^d$ only through $g(\mathbf{x}) := \beta^T \mathbf{x}$ for some $\beta \in \mathbb{R}^d$. As a result, we could have $\widehat{f}(y|\mathbf{x}) = f(y|g(\mathbf{x}))$ even when $\widehat{f}(y|\mathbf{x})$ is potentially very different from $f(y|\mathbf{x})$. As another example, a conditional density estimator that performs variable selection (Shiga et al., 2015; Izbicki et al., 2017; Dalmasso et al., 2020) could satisfy $\widehat{f}(y|\mathbf{x}) = f(y|g(\mathbf{x}))$ for $g(\mathbf{x}) := (\mathbf{x})_S$, where $S \subset \{1, \ldots, d\}$ is a subset of the covariates. A test of the overall uniformity of PIT values is no guarantee that we are correctly modeling the relationship between y and the predictors $\mathbf{x}$; see Figure 3.2 for an illustration.

Previous diagnostics were also unable to pinpoint the locations in the feature space $\mathcal{X}$ where the estimates of $f(y|\mathbf{x})$ should be improved. Hence, in addition to global consistency, we need diagnostics that test for the following property:

Definition 3 (**Local Consistency**). *Fix $\mathbf{x} \in \mathcal{X}$. An estimate $\widehat{f}(y|\mathbf{x})$ is locally consistent with the density $f(y|\mathbf{x})$ at fixed $\mathbf{x}$ if the following null hypothesis holds:*

$$H_0(\mathbf{x}) : \widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x}) \text{ for every } y \in \mathcal{Y}. \tag{3.4}$$

In the next three sections, we introduce new diagnostics that are able to test whether a conditional density model $\widehat{f}(y|\mathbf{x})$ is both globally and locally consistent with the underlying conditional distribution $f(y|\mathbf{x})$ of the data. Our diagnostics are still based on the probability integral transform, and hence they retain the properties (e.g., interpretability, ability to provide graphical summaries, and so on) that have made PIT a popular choice in model validation.

3.2 New Diagnostics Test Local and Global Consistency

We now present a new diagnostic framework that addresses the shortcomings of previous methods. The framework has three main components:

- **[GCT - Global Coverage Test]** A statistical hypothesis test that can distinguish *any* misspecified density model from the true conditional density. (This is a test of global consistency; see Definition 1.)
- **[LCT - Local Coverage Test]** A statistical hypothesis test that identifies *where* in the feature space the model fits poorly. (This is a test of local consistency; see Definition 3.)
- **[ALP - Amortized Local P-P plots]** Interpretable graphical summaries of the fitted model that show *how* it deviates from the true density at any location in feature space (see Figure 3.1 for examples). We also provide amortized PIT histograms that contain the same information as ALPs but in a different format (see Section 3.4 for details).

This section provides the conceptual and theoretical basis for our approach. Section 3.3 discusses the GCT and LCT, and Section 3.4 discusses ALPs.

At the heart of our approach is the realization that the local coverage of a CDE model is itself a conditional probability (see Equation 4.2) that often varies smoothly with $\mathbf{x}$. Hence, we can estimate the local coverage at any given $\mathbf{x}$ by leveraging a suitable regression method using sample points in a neighborhood of $\mathbf{x}$. Thanks to the impressive arsenal of existing regression methods, we can adapt our approach to different types of potentially high-dimensional

data to obtain computationally and statistically efficient validation. Finally, because we specifically evaluate local coverage (rather than other types of discrepancies), the practitioner can "zoom in" on statistically significant local discrepancies flagged by the LCT, and identify common modes of failure in the fitted conditional density (see Figure 3.3 for an example).

Our new diagnostics rely on the following key result:

Theorem 2 (Local Consistency and Pointwise Uniformity). *For any $\mathbf{x} \in \mathcal{X}$, the local null hypothesis $H_0(\mathbf{x})$: $\widehat{f}(\cdot|\mathbf{x}) = f(\cdot|\mathbf{x})$ holds if, and only if, the distribution of PIT$(Y;\mathbf{x})$ given $\mathbf{x}$ is uniform over $(0,1)$.*

Theorem 2 implies that if we had a sample of Y's at the fixed location $\mathbf{x}$, we could test the local consistency (Definition 3) of $\widehat{f}$ by determining whether the sample's PIT values come from a uniform distribution. In addition, for global consistency we need local consistency at every $\mathbf{x} \in \mathcal{X}$. Clearly, such a testing procedure would not be practical: typically, we have data of the form $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ with at most one observation at any given $\mathbf{x} \in \mathcal{X}$.

Our solution is to instead address this problem as a regression. For fixed $\gamma \in (0,1)$, we consider the cumulative distribution function (CDF) of PIT at $\mathbf{x}$,

$$r(\gamma;\mathbf{x}) := \mathbb{P}\left(\text{PIT}(Y;\mathbf{x}) < \gamma | \mathbf{x}\right), \tag{3.5}$$

which is the regression of the random variable

$$W^{\gamma} := \mathbb{I}(\text{PIT}(Y;\mathbf{X}) < \gamma) \tag{3.6}$$

on $(\mathbf{X}, \gamma)$.

From Theorem 2, it follows that the estimated density is locally consistent at $\mathbf{x}$ if and only if $r(\gamma;\mathbf{x}) = \gamma$ for every $\gamma \in (0,1)$:

Corollary 1. *Fix $\mathbf{x} \in \mathcal{X}$. Then $r(\gamma;\mathbf{x}) = \gamma$ for every $\gamma \in (0,1)$ if, and only if, $\widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x})$ for every $y \in \mathcal{Y}$.*

Of course, in practice we do not have access to the true $r(\gamma;\mathbf{x})$ function, which describes the true coverage of $\widehat{f}(y|\mathbf{x})$. But the key idea is that we can estimate this function. In this way, our new diagnostics are able to test for both local and global consistency, by relying on the simple idea of estimating $r(\gamma;\mathbf{x})$ via regression, and then evaluating how much it deviates from γ (see Section 3.3 for details).

Note that

$$\text{PIT}(Y;\mathbf{x}) < \alpha \iff Y \in (-\infty, \widehat{q}_{\gamma}(\mathbf{x}))$$

where $\widehat{q}_\gamma(\mathbf{x})$ is the γ-quantile of $\widehat{f}(y|\mathbf{x})$. That is, $r(\gamma;\mathbf{x})$ assesses the local level-γ **coverage** of $\widehat{f}(y|\mathbf{x})$ at $\mathbf{x}$. In Section 3.4, we explore the connection between test statistics and coverage, for interpretable descriptions of how conditional density models $\widehat{f}(y|\mathbf{x})$ may fail to approximate the true conditional density $f(y|\mathbf{x})$.

3.3 Local and Global Coverage Tests

Our procedure for testing local and global consistency is very simple and can be adapted to different types of data. For an i.i.d. test sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ from $F_{\mathbf{X},Y}$ (which was not used to construct $\widehat{f}(y|\mathbf{x})$), we compute

$$W_i^\gamma := \mathbb{I}(\text{PIT}(Y_i; \mathbf{X}_i) < \gamma). \tag{3.7}$$

That is, for each test point i, we compute W_i^γ over multiple sampled values of $\gamma \sim G$ from some distribution G over $(0,1)$. By default, we take G to be the $Unif[0,1]$ distribution. To estimate the coverage $r(\gamma;\mathbf{x})$ (Equation 3.5) for any $\mathbf{x} \in \mathcal{X}$, we then simply regress W on $(\mathbf{X}, \gamma)$. Numerous classes of regression estimators can be used, from kernel smoothers to random forests to neural networks.

Algorithm 2 P-values for Local Coverage Test

Input: conditional density model $\widehat{f}$; test data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$; test point $\mathbf{x} \in \mathcal{X}$; regression estimator $\widehat{r}$; distribution G of nominal coverage values γ over $(0,1)$; number of null training samples B

Output: estimated p-value $\widehat{p}(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$

Compute test statistic at $\mathbf{x}$:
Compute values $\text{PIT}(Y_1; \mathbf{X}_1), \ldots, \text{PIT}(Y_n; \mathbf{X}_n)$
for $\gamma \sim G$ **do**
    Compute indicators $W_1^\gamma, \ldots, W_n^\gamma$
    Train regression estimator $\widehat{r}(\gamma;\mathbf{x})$ on $\{\mathbf{X}_i, W_i^\gamma\}_{i=1}^n$
end for
Compute test statistic $T(\mathbf{x})$ (Eq. 3.8)
Recompute test statistic under null distribution:
for b in $1, \ldots, B$ **do**
    Draw $U_1^{(b)}, \ldots, U_n^{(b)} \sim \text{Unif}[0,1]$.
    for $\gamma \sim G$ **do**
        Compute indicators $\left\{W_i^{\gamma,(b)} := \mathbb{I}(U_i^{(b)} < \gamma)\right\}_{i=1}^n$
        Train new regression estimator $\widehat{r}^{(b)}(\gamma;\mathbf{x})$ on $\left\{\mathbf{X}_i, W_i^{\gamma,(b)}\right\}_{i=1}^n$
    end for
    Compute $T^{(b)}(\mathbf{x}) := \sum_\gamma \left(\widehat{r}^{(b)}(\gamma;\mathbf{x}) - \gamma\right)^2$
end for
return $\widehat{p}(\mathbf{x}) := \dfrac{1}{B}\sum_{b=1}^B \mathbb{I}\left(T(\mathbf{x}) < T^{(b)}(\mathbf{x})\right)$

To test local consistency (Definition 3), we introduce the *Local Coverage Test* (LCT) with the test statistic

$$T(\mathbf{x}) := \sum_{\gamma} \left(\widehat{r}(\gamma; \mathbf{x}) - \gamma\right)^2, \tag{3.8}$$

where $\widehat{r}(\gamma; \mathbf{x})$ denotes the regression estimator and we evaluate over a fixed set of γ values (which may or may not be evenly spaced). Large values of $T(\mathbf{x})$ indicate a large discrepancy between $\widehat{f}(y|\mathbf{x})$ and $f(y|\mathbf{x})$ at $\mathbf{x}$ in terms of coverage, and Corollary 1 links coverage to consistency. To decide on the correct cutoff for rejecting $H_0(\mathbf{x})$, we use a Monte Carlo technique that simulates $T(\mathbf{x})$ under H_0. Algorithm 2 details our procedure.

For the LCT, note that we are performing multiple hypothesis tests at different locations $\mathbf{x}$. After obtaining LCT p-values, we recommend using a multiple testing correction procedure like Benjamini-Hochberg to control the false discovery rate.

Algorithm 3 P-values for Global Coverage Test

Input: conditional density model $\widehat{f}$; test data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$; regression estimator $\widehat{r}$; distribution G of nominal coverage values γ over $(0, 1)$; number of null training samples B

Output: estimated p-value $\widehat{p}(\mathbf{x})$ across all $\mathbf{x} \in \mathcal{X}$

Compute test statistic over $\mathbf{X}_1, \ldots, \mathbf{X}_n$:
Compute values $\text{PIT}(Y_1; \mathbf{X}_1), \ldots, \text{PIT}(Y_n; \mathbf{X}_n)$
for $\gamma \sim G$ **do**
    Compute indicators $W_1^\gamma, \ldots, W_n^\gamma$
    Train regression estimator $\widehat{r}(\gamma; \mathbf{x})$ on $\{\mathbf{X}_i, W_i^\gamma\}_{i=1}^n$
end for
Compute test statistic $T(\mathbf{X}_i)$ (Eq. 3.8) for $i = 1, \ldots .n$
Compute test statistic $S = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} T(\mathbf{X}_i)$
Recompute test statistic under null distribution:
for b in $1, \ldots, B$ **do**
    Draw $U_1^{(b)}, \ldots, U_n^{(b)} \sim \text{Unif}[0, 1]$.
    for $\gamma \sim G$ **do**
        Compute indicators $\left\{W_i^{\gamma,(b)} := \mathbb{I}(U_i^{(b)} < \gamma)\right\}_{i=1}^n$
        Train new regression estimator $\widehat{r}^{(b)}(\gamma; \mathbf{x})$ on $\left\{\mathbf{X}_i, W_i^{\gamma,(b)}\right\}_{i=1}^n$
    end for
    Compute $T^{(b)}(\mathbf{X}_i) := \displaystyle\sum_{\gamma} \left(\widehat{r}^{(b)}(\gamma; \mathbf{X}_i) - \gamma\right)^2$ for $i = 1, \ldots, n$
    Compute $S^{(b)} := \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} T^{(b)}(\mathbf{X}_i)$
end for
return $\widehat{p}(\mathbf{x}) := \dfrac{1}{B} \displaystyle\sum_{b=1}^{B} \mathbb{I}\left(S < S^{(b)}\right)$

Similarly, we can also test global consistency (Definition 1) with a Monte Carlo strategy. We introduce the *Global Coverage Test* (GCT) based on the following test statistic:

$$S := \frac{1}{n} \sum_{i=1}^{n} T(\mathbf{X}_i). \tag{3.9}$$

Algorithm 3 details our procedure.

We recommend performing the global test first and, if the global null is rejected, investigating further with local tests. Empirically, we have found that the power of our tests is related to the MSE (a measurable quantity) of the regression method we use. This observation is in line with similar results in Kim et al. (2019, Theorems 3.3 and 4.1). Hence, as a practical strategy, we can maximize power by choosing the regression model that achieves the smallest MSE on validation data.

We now give sufficient conditions under which the p-values from the local coverage test are approximately valid. We rely on the following assumption.

Assumption 1 (**Local regression estimator**). *There exists $\epsilon > 0$ such that $\widehat{r}$ only uses the sample points in $\mathcal{D}'$ with $\mathbf{X}_i \in B(\mathbf{x}; \epsilon)$, where $B(\mathbf{x}; \epsilon)$ is a ball of radius ϵ centered at $\mathbf{x}$.*

Assumption 1 holds for regression estimators that are based on partitions, such as tree-based estimators (e.g., random forests, boosting methods) or smoothing kernel estimators with kernels with bounded support.

Theorem 3. *Under the null hypothesis*

$$H_0^{\epsilon}(\mathbf{x}) : \widehat{F}(y|\mathbf{x}) = F(y|\mathbf{x}) \text{ for every } y \in \mathcal{Y} \text{ for all } \mathbf{x}' \in B(\mathbf{x}; \epsilon)$$

and under Assumption 1, for any $0 < \alpha < 1$,

$$\lim_{B \longrightarrow \infty} \mathbb{P}\left(p(\mathbf{x}) \leq \alpha\right) = \alpha.$$

3.4 Amortized Local P-P Plots

Our diagnostic framework does not just give us the ability to identify deviations from local consistency in different parts of the feature space $\mathcal{X}$. It also provides us with insight into the nature of such deviations at any given location $\mathbf{x}$. For unconditional density models, data scientists have long favored using P-P plots (which plot two cumulative distribution functions against each other) to assess how closely a density model agrees with actual observed data. What makes our work unique is that we are able to construct "amortized local P-P plots" (ALPs) with similar interpretations to assess *conditional* density models over the entire feature space.

Figure 3.1 illustrates how a local P-P plot of $\widehat{r}(\gamma; \mathbf{x})$ against γ (that is, the estimated CDF against the true CDF at $\mathbf{x}$) can identify different types of deviations in a conditional density model. For example, positive or negative bias in

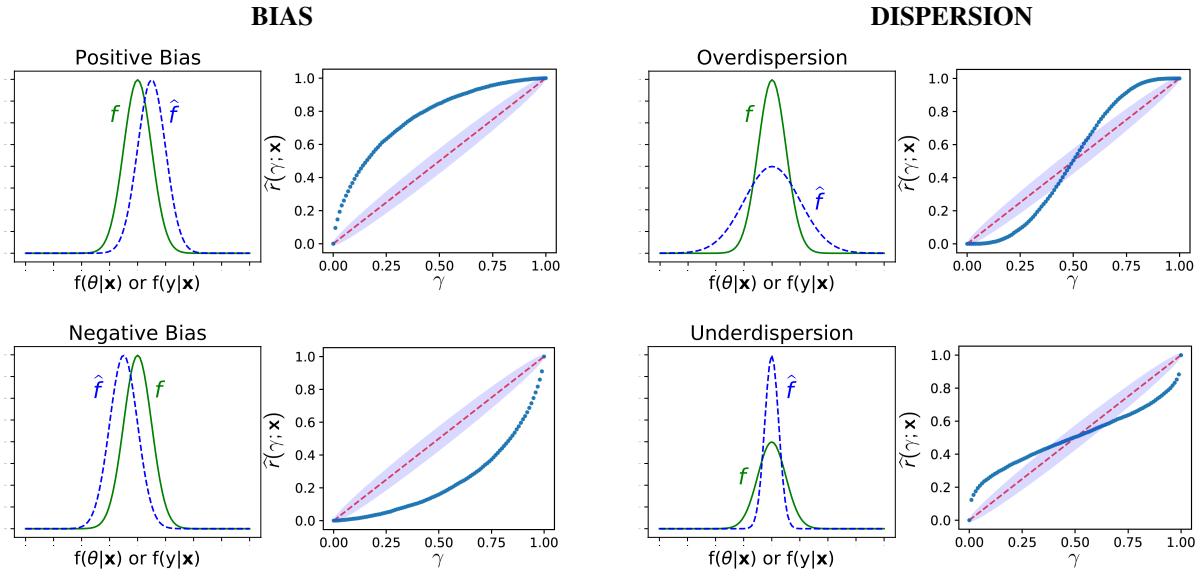


Figure 3.1: P-P plots are commonly used to assess how well a density model fits actual data. Such plots display, in a clear and interpretable way, effects like bias (left panel) and dispersion (right panel) in an estimated distribution $\widehat{f}$ vis-a-vis the true data-generating distribution f. Our framework yields a computationally efficient way to construct "amortized local P-P plots" for comparing conditional densities $\widehat{f}(\theta|\mathbf{x})$ and $\widehat{f}(y|\mathbf{x})$ at any location $\mathbf{x}$ of the feature space $\mathcal{X}$. See text for details and Sections 3.5 and 3.6 for examples.

the estimated density $\widehat{f}$ relative to f leads to P-P plot values that are too high or too low, respectively. We can also easily identify overdispersion or underdispersion of $\widehat{f}$ from an "S"-shaped P-P plot.

Of particular note is that our local P-P plots are "amortized", in the sense that computationally expensive steps do not have to be repeated with e.g Monte Carlo sampling at each $\mathbf{x}$ of interest. Both the consistency tests in Section 3.3 and the local P-P plots or ALPs only require initially training $\widehat{r}(\gamma; \mathbf{x})$ on the observed data; the regression estimator can then be used to compute $\widehat{r}(\gamma; \mathbf{x}_{val})$ at any new evaluation point $\mathbf{x}_{val}$. Because of the flexibility in the choice of regression method, our construction also potentially scales to high-dimensional or different types of data $\mathbf{x}$. Algorithm 4 details the construction of confidence bands for ALPs (under the null) using a Monte Carlo algorithm.

Algorithm 4 Confidence band for ALP under H_0

Input: test data $\{\mathbf{X}_i\}_{i=1}^n$; test point $\mathbf{x} \in \mathcal{X}$; regression estimator $\widehat{r}$; distribution G of nominal coverage values γ over $(0,1)$; number of null training samples B; confidence level $1-\eta$

Output: estimated $(1-\eta)$ confidence band $\{L(\mathbf{x}), U(\mathbf{x})\}$ for $\widehat{r}(\gamma;\mathbf{x})$ under the null, for any $\mathbf{x} \in \mathcal{X}$, and $\gamma \in (0,1)$

Recompute regression under null distribution:
for b in $1,\ldots,B$ **do**
  Draw $U_1^{(b)},\ldots,U_n^{(b)} \sim$ Unif$[0,1]$.
  for $\gamma \sim G$ **do**
    Compute indicators $\left\{W_i^{\gamma,(b)} := \mathbb{I}\left(U_i^{(b)} < \gamma\right)\right\}_{i=1}^n$
    Train regression method $\widehat{r}^{(b)}(\gamma;\mathbf{x})$ on $\left\{\mathbf{X}_i, W_i^{\gamma,(b)}\right\}_{i=1}^n$
  end for
  Compute $\widehat{r}^{(b)}(\gamma;\mathbf{x})$ for test point $\mathbf{x}$.
end for
Compute $(1-\eta)$ **confidence band for** $\widehat{r}^{(b)}(\gamma;\mathbf{x})$**:**
$L(\mathbf{x}), U(\mathbf{x}) \leftarrow \emptyset$
for $\gamma \sim G$ **do**
  $L(\mathbf{x}) \leftarrow L(\mathbf{x}) \cup \frac{\eta}{2}$-quantile of $\left\{\widehat{r}^{(b)}(\gamma;\mathbf{x})\right\}_{b=1}^B$
  $U(\mathbf{x}) \leftarrow U(\mathbf{x}) \cup \left(1-\frac{\eta}{2}\right)$-quantile of $\left\{\widehat{r}^{(b)}(\gamma;\mathbf{x})\right\}_{b=1}^B$
end for
return $L(\mathbf{x}), U(\mathbf{x})$

As an alternative to ALPs, one can also visualize the same information in local PIT histograms. Algorithm 5 describes how to construct local PIT histograms. Note that if one has already obtained an estimator $\widehat{r}(\gamma;\mathbf{x})$ for the local PIT distribution (CDF) via regression (by, for example, running Algorithm 2), then one can generate a local histogram at any $\mathbf{x} \in \mathcal{X}$ by simply evaluating those $\widehat{r}$ functions at $\mathbf{x}$, without needing to rerun any regressions. Similarly, there is no need to repeat the MC sampling under the null in Algorithm 6 to create confidence bands for the local PIT histograms.

Algorithm 5 Local PIT histograms

Input: conditional density model $\widehat{f}$; test data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$; test point $\mathbf{x} \in \mathcal{X}$; regression estimator $\widehat{r}$; distribution G of nominal coverage values γ over $(0,1)$; number of bins n_{bin}

Output: local PIT histogram $H(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$

Compute estimated local coverage at $\mathbf{x}$**:**
Compute values PIT$(Y_1;\mathbf{X}_1),\ldots,PIT(Y_n;\mathbf{X}_n)$
for $\gamma \sim G$ **do**
  Compute indicators $\{W_i^\gamma := \mathbb{I}(U_i < \gamma)\}_{i=1}^n$
  Train regression method $\widehat{r}(\gamma;\mathbf{x})$ on $\{\mathbf{X}_i, W_i^\gamma\}_{i=1}^n$
  Compute value $\widehat{r}(\gamma;\mathbf{x})$ for test point $\mathbf{x}$
end for
Compute local PIT histogram:
Create histogram $H(\mathbf{x})$ of $\widehat{r}(\gamma;\mathbf{x})$ values by dividing $[0,1]$ into n_{bin} equal-sized bins
return histogram $H(\mathbf{x})$

Algorithm 6 Confidence band for local PIT histogram under H_0

Input: test data $\{\mathbf{X}_i\}_{i=1}^n$; test point $\mathbf{x} \in \mathcal{X}$; regression estimator $\widehat{r}$; number of bins n_{bin}; distribution G of nominal coverage values γ over $(0,1)$; number of null training samples B; confidence level 1-η; number of bins n_{bin}

Output: estimated $(1-\eta)$ confidence band $\{L(\mathbf{x}), U(\mathbf{x})\}$ for local PIT histogram $H(\mathbf{x})$ under the null, for any $\mathbf{x} \in \mathcal{X}$

Recompute regression under null distribution:
for b in $1, \ldots, B$ **do**
  Draw $U_1^{(b)}, \ldots, U_n^{(b)} \sim \text{Unif}[0,1]$.
  for $\gamma \sim G$ **do**
    Compute indicators $\left\{W_i^{\gamma,(b)} := \mathbb{I}(U_i^{(b)} < \gamma)\right\}_{i=1}^n$
    Train regression method $\widehat{r}^{(b)}(\gamma; \mathbf{x})$ on $\left\{\mathbf{X}_i, W_i^{\gamma,(b)}\right\}_{i=1}^n$
  end for
end for
Compute confidence band:
for b in $1, \ldots, B$ **do**
  Create histogram $H^{(b)}(\mathbf{x})$ of $\left\{\widehat{r}^{(b)}(\gamma; \mathbf{x})\right\}$ by dividing $[0,1]$ into n_{bin} equal-sized bins
end for
$L(\mathbf{x}) \leftarrow \dfrac{\eta}{2}$-quantile of $\left\{H^{(b)}(\mathbf{x})\right\}_{b=1}^B$
$U(\mathbf{x}) \leftarrow \left(1 - \dfrac{\eta}{2}\right)$-quantile of $\left\{H^{(b)}(\mathbf{x})\right\}_{b=1}^B$
return $L(\mathbf{x}), U(\mathbf{x})$

3.5 Example: Omitted Variable Bias in CDE Models

This stylized example involves omitted but clearly relevant variables in a prediction setting. Inspired by Section 2.2.2 of Shalizi (2013), we generate

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right)$$

and take the response to be

$$Y|\mathbf{X} \sim N(X_1 + X_2, 1).$$

To mimic the variable selection procedure common in high-dimensional inference methods, we fit two conditional density models: $\widehat{f}_1$, trained only on X_1, and $\widehat{f}_2$, trained on $\mathbf{X}$. Both models are fitted using a nearest-neighbor kernel CDE (Dalmasso et al., 2020) with hyperparameters chosen by data splitting.

This is a stylized example where omitting one of the variables might lead to unwanted bias when predicting the outcome Y for new inputs $\mathbf{X}$. As an indication of this bias, we have included a heat map (see panel (d) of Figure 3.3) of the difference in the true (unknown) conditional means, $\mathbb{E}[Y|x_1] - \mathbb{E}[Y|x_1, x_2]$ as a function of x_1 and x_2. (In this example, the omitted variable bias is approximately the same as the difference in the averages of the predictions of Y when using the model $\widehat{f}_1$ versus the model $\widehat{f}_2$ at any given $\mathbf{x} \in \mathcal{X}$; see Figure 3.3 panels (c) and (d)). Despite the clear relationship between Y and X_2, both $\widehat{f}_1$ (which omits X_2) and $\widehat{f}_2$ pass previous goodness-of-fit tests based on PIT (Figure 3.2). This result can be explained by Theorem 1: because PIT is insensitive to covariate transformations and

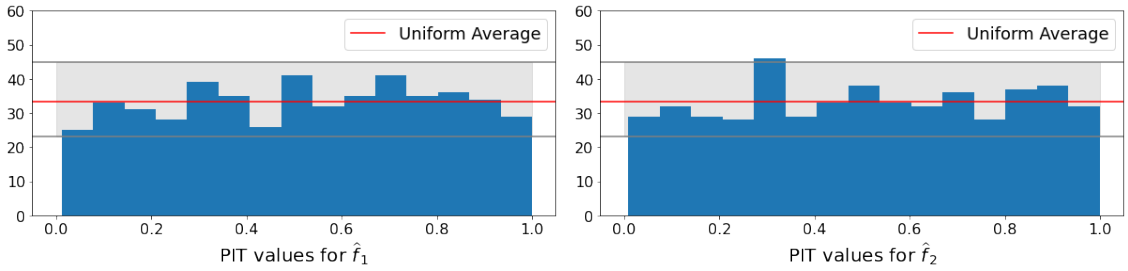


Figure 3.2: Standard diagnostics for showing histograms of PIT values computed on 200 test points (with 95% confidence bands for a Unif[0,1] distribution). *Left:* Results for $\widehat{f}_1$, which has only been fit to the first of two covariates. *Right:* Results for $\widehat{f}_2$, which has been fit to both covariates. The top panel shows that standard PIT diagnostics cannot tell that $\widehat{f}_1$ is a poor approximation to f. GCT, on the other hand, detects that $\widehat{f}_1$ is misspecified (p=0.004), while not rejecting the global null for $\widehat{f}_2$ (p=0.894).

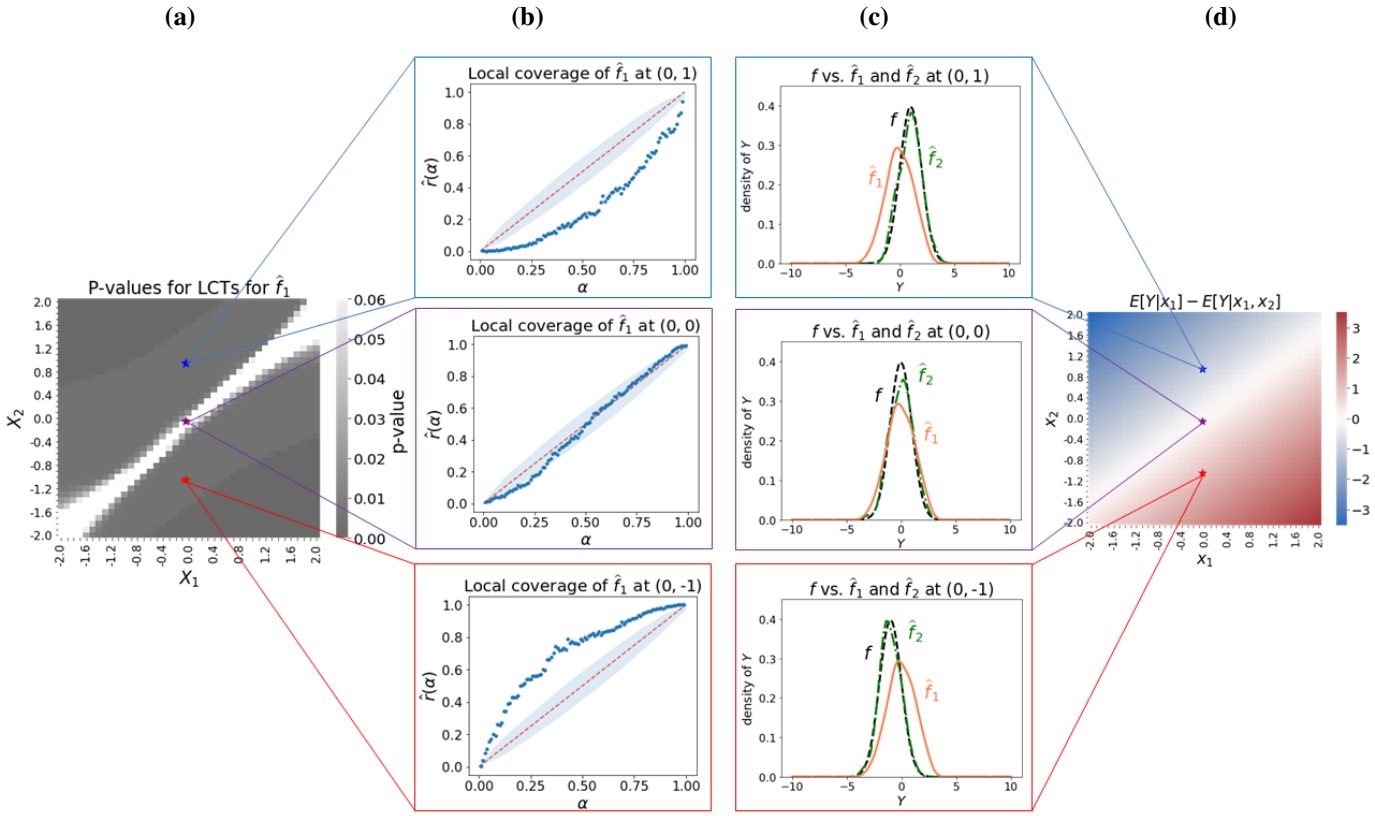


Figure 3.3: Diagnostics for omitted variable bias example. **(a)** P-values for LCTs for $\widehat{f}_1$ indicate a poor fit across most of the feature space. **(b)** Amortized local P-P plots at selected points show the density $\widehat{f}_1$ as negatively biased (blue), well estimated at significance level $\alpha = 0.05$ with barely perceived overdispersion (purple), and positively biased (red). (Gray regions represent 95% confidence bands under the null.) **(c)** $\widehat{f}_1$ and $\widehat{f}_2$ vs. the true (unknown) conditional density f at the selected points. $\widehat{f}_1$ is clearly negatively and positively biased at the blue and red points, respectively, while the model does not reject the local null at the purple point. $\widehat{f}_2$ fits well at all three points. The difference on average in the predictions of Y from $\widehat{f}_1(\cdot|\mathbf{x})$ vs. the true distribution $f(\cdot|\mathbf{x})$ for fixed $\mathbf{x}$ indeed corresponds to the "omitted variable bias" $\mathbb{E}[Y|x_1] - \mathbb{E}[Y|x_1, x_2]$. (*Note:* Panels (c) and (d) require knowledge of the true f, which would not be available to the practitioner.)

$\widehat{f}_1(y|\mathbf{x}) \approx f(y|x_1)$, PIT values are uniformly distributed, even though $\widehat{f}_1$ omits a key variable. The GCT, however, detects that $\widehat{f}_1$ is misspecified ($p = 0.004$), while the global null (Equation 3.1) is not rejected for $\widehat{f}_2$ ($p = 0.894$).

The next question a practitioner might ask is: "What exactly is wrong with the fit?". LCTs and local P-P plots can pinpoint the locations of discrepancies and describe the failure modes. Panel (a) of Figure 3.3 shows p-values from local coverage tests for $\widehat{f}_1$ across the entire feature space of $\mathbf{X}$. The patterns in these p-values are largely explained by panel (d), which shows the difference between the conditional means of Y given x_1 and given x_1, x_2. The detected level of discrepancy between the estimate $\widehat{f}_1$ and the true conditional density f at a point $\mathbf{x}$ directly relates to the omitted variable bias $\mathbb{E}[Y|x_1] - \mathbb{E}[Y|x_1, x_2] = 0.8x_1 - x_2$: the LCT p-values close to the line $x_2 = 0.8x_1$ are large (indicating no statistically significant deviations from the true model), and p-values decrease as we move away from this line.

Panel (b) of Figure 3.3 zooms in on a few different locations $\mathbf{x}$ with local P-P plots that depict and interpret distributional deviations. At the blue point, $\widehat{f}_1$ underestimates the true Y: we reject the local null (Equation 3.4), and the P-P plot indicates negative bias. Conversely, at the red point, $\widehat{f}_1$ overestimates the true Y; we reject the local null, and the P-P plot indicates positive bias. At the purple point, $\widehat{f}_1$ is close to f, so the local null hypothesis is not rejected.

For reference, we show the results of the local test for the well-specified model $\widehat{f}_2$, which does pass the global test. Figure 3.4, right panel, shows p-values from LCTs across the feature space for the model $\widehat{f}_2$. Unlike model $\widehat{f}_1$, which was fit on X_1 alone, $\widehat{f}_2$ was fit on both X_1 and X_2. Hence, $\widehat{f}_2$ is able to pass all tests, with local P-P plots indicating a good fit (with two examples shown in the Figure 3.4, left panel).

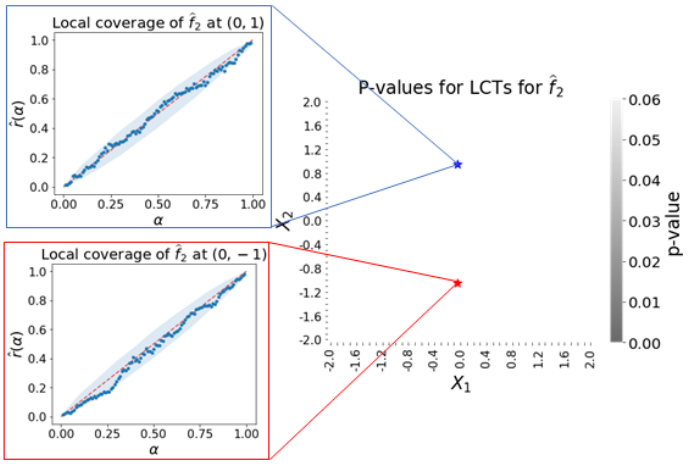


Figure 3.4: P-values for LCTs for $\widehat{f}_2$ suggest an adequate fit everywhere in the feature space; local coverage plots at selected points also suggest a good fit.

This stylized example is a simple illustration of the general phenomenon of potentially unwanted omitted variable bias, which can be difficult to detect without testing for local and global consistency of models. Our proposed diagnostics identify this issue and provide insight into how the omitted variable distorts the fitted model relative to the true conditional density, across the entire feature space.

3.6 Example: Conditional Neural Densities for Galaxy Images

In this high-dimensional example of conditional density estimation in a prediction setting, we assess the quality of CDEs fit to image data $\mathbf{x} \in \mathbb{R}^{400}$. In particular, we apply neural density models to estimate the distribution of synthetic "redshift" Z (a proxy for distance; the response) assigned to photometric or "photo-z" galaxy images $\mathbf{X}$ (the predictors). We then illustrate how our methods distinguish between well-fitting and poorly fitting CDEs. This toy example is motivated by the urgent need for metrics to assess photo-z probability density function accuracy. Diagnostics currently used by astronomers have known shortcomings as we have discussed (Schmidt et al., 2020), and our method is the first to properly address them.

Here, $\mathbf{x}$ represents a 20×20-pixel image of an elliptical galaxy generated by `GalSim`, an open-source toolkit for simulating realistic images of astronomical objects (Rowe et al., 2015). In `GalSim`, we can vary the axis ratio λ, defined as the ratio between the minor and major axes of the projection of the elliptical galaxy. We create four equally sized populations of galaxies, with $\lambda \in \{0.8, 0.7, 0.6, 0.5\}$. We then assign a response variable Z according to different distributions (unimodal, skewed and bimodal) as follows:

$$
\begin{aligned}
Z|\lambda = 0.8 &\sim N(0.1, 0.02) \\
Z|\lambda = 0.7 &\sim \text{Beta}(3, 7) \\
Z|\lambda = 0.6 &\sim 0.6N(0.3, 0.05) + 0.4N(0.7, 0.05) \\
Z|\lambda = 0.5 &\sim \text{Beta}(7, 3).
\end{aligned}
$$

Figure 3.5 shows the true conditional densities of the simulated "redshift" Z, indexed by the axis ratio λ of the corresponding galaxy image.

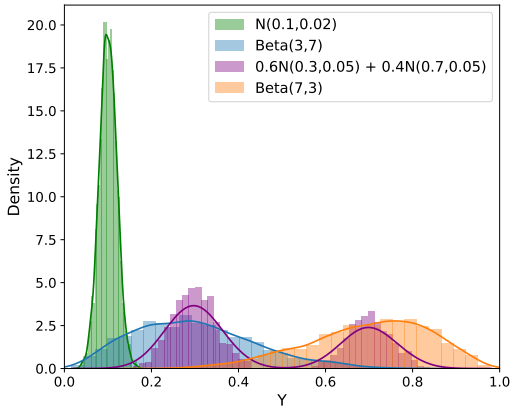


Figure 3.5: We assign a unimodal distribution of "redshift" Z for to the galaxy population with $\lambda = 0.8$, and higher, more skewed and bimodal distributions of Z to the populations with $\lambda = 0.7, 0.6, 0.5$.

For illustration, we fit a unimodal Gaussian neural density model to estimate the conditional density $Z|\mathbf{X}$. Our diagnostics then pinpoint where in the feature space the density is bimodal or skewed, and thus a fit with one Gaussian is inadequate. We know of no other diagnostics that can provide such insight when fitting neural density models. Specifically, we fit a convolutional mixture density network (ConvMDN, D'Isanto and Polsterer (2018)) with a single Gaussian component, two convolutional and two fully connected layers with ReLU activations (Glorot et al., 2011). (We train on 10000 images using the Adam optimizer (Kingma and Ba, 2014) with learning rate 10^{-3}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.) This gives an estimate of $f(z|\mathbf{x})$. We expect this CDE model to fit well for the $\lambda = 0.8$ unimodal population, and fit poorly for the other bimodal or skewed populations.

Our diagnostic framework effectively detects the shortcomings of this conditional density model. First, we perform the global coverage test, which rejects the global null ($p < 0.001$). Next, we turn to local coverage tests and local P-P plots to explore where and how the fit is inadequate. Figure 3.6 shows a principal component map of the test data. The LCTs are able to identify a unimodal Gaussian model fits well for the $\lambda = 0.8$ population, but that the same model fails to adequately estimate the PDFs of the remaining populations. P-P plots at selected test points indicate significant distributional deviations and suggest the need to consider more flexible model classes that incorporate bimodal and skewed distributions. For instance, CDE models not based on mixtures (Papamakarios et al., 2019) could be more effective for this data.

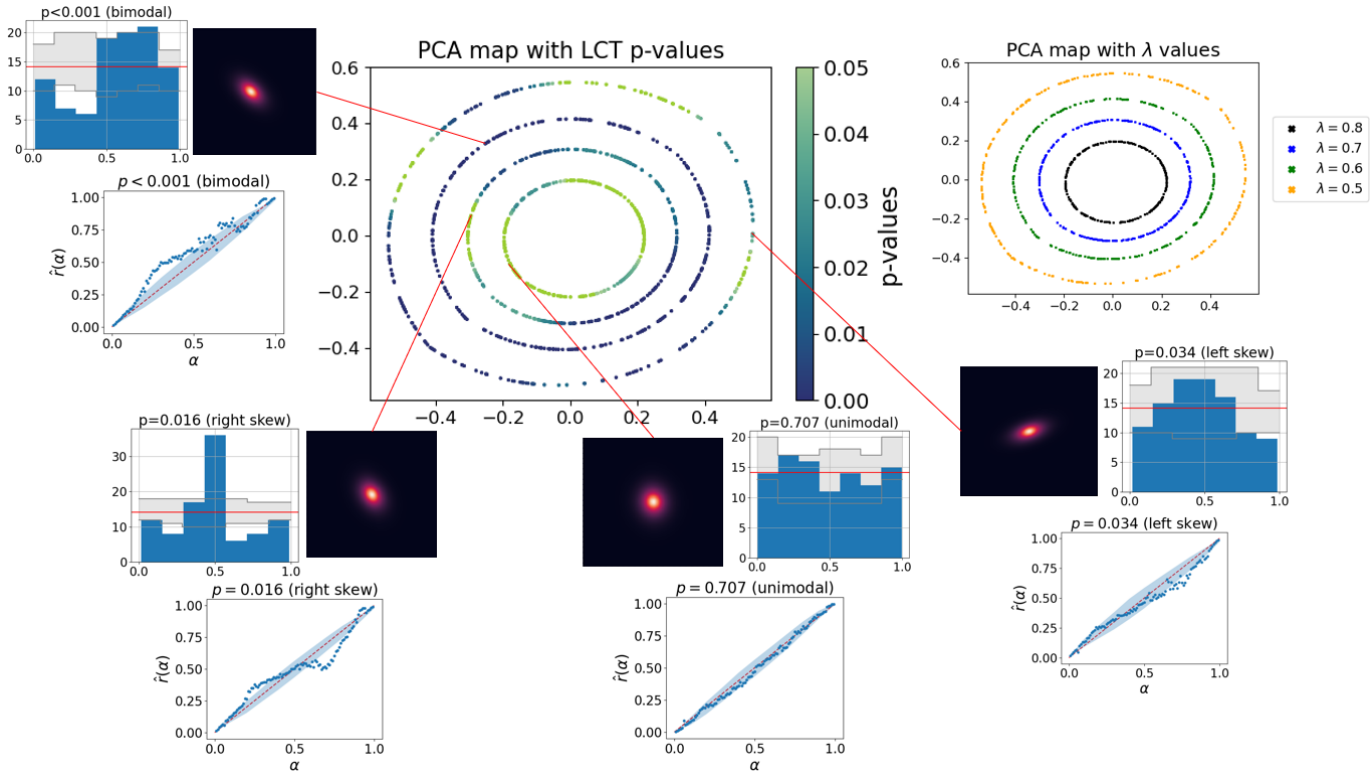


Figure 3.6: Diagnostics for galaxy images example. For visualization, we show the location of the test galaxy points in $\mathbb{R}^{400}$ along the first two principal components (see center panel "PCA map with LCT p-values"). Test statistics from the LCTs indicate that the unimodal density model generally fits well for the $\lambda = 0.8$ population, while fitting poorly for the other three populations with skewed and bimodal true redshift distributions. Local P-P plots or ALPs show statistically significant deviations in the CDEs (gray regions are 95% confidence bands under the null) for the latter population, suggesting the need for more flexible model classes. We also display local PIT histograms with confidence bands under the null, as a different way to present the same information as in the ALPs. (The histograms are computed from the $\widehat{r}_\alpha$ values according to Algorithm 4; no additional regression is needed.)

3.7 Handling Multivariate Responses

If the response $\mathbf{Y}$ is multivariate, then the random variable $F_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X})$ is not uniformly distributed (Genest and Rivest, 2001), so PIT values cannot be trivially generalized to higher dimensions.

One way to overcome this is to evaluate the PIT statistic of univariate projections of $\mathbf{Y}$, as done by Talts et al. (2018) for Bayesian consistency checks and Mucesh et al. (2020) for the prediction setting. That is, the PIT values can be computed using the estimate $\widehat{f}(h(\mathbf{Y})|\mathbf{x})$ induced by $\widehat{f}(\mathbf{Y}|\mathbf{x})$ for some chosen $h : \mathbb{R}^p \longrightarrow \mathbb{R}$. Different projections can be used depending on the context. For instance, in Bayesian applications, posterior distributions are often used to compute credible regions for univariate projections of the parameters θ. Thus, it is natural to evaluate PIT values of $h(\theta) = \theta_i$ for each parameter of interest. Another useful projection is copPIT (Ziegel et al., 2014), which creates a unidimensional projection that has information about the joint distribution of $\mathbf{Y}$.

Our diagnostic techniques are not enough to consistently assess the fit to $f(\mathbf{Y}|\mathbf{x})$ if applied to these projections, but they do consistently evaluate the fit to $f(h(\mathbf{Y})|\mathbf{x})$, which is often good enough in practice.

An alternative approach to assessing $\widehat{f}$ is through highest predictive density values (HPD values; Harrison et al. 2015; Dalmasso et al. 2020), which are defined by

$$\text{HPD}(\mathbf{y};\mathbf{x}) = \int_{\mathbf{y}':\widehat{f}(\mathbf{y}'|\mathbf{x})\geq\widehat{f}(\mathbf{y}|\mathbf{x})} \widehat{f}(\mathbf{y}'|\mathbf{x})d\mathbf{y}'$$

(see Figure 2.2 for an illustration). $\text{HPD}(\mathbf{y};\mathbf{x})$ is a measure of how plausible $\mathbf{y}$ is according to $\widehat{f}(\mathbf{y}|\mathbf{x})$ (in the Bayesian context, this is the complement of the e-value (de Bragança Pereira and Stern, 1999); small values indicate high plausibility). As with PIT values, HPD values are uniform under the global null hypothesis (Dalmasso et al., 2020). However, standard goodness-of-fit tests based on HPD values share the same problem as those based on PIT: they are insensitive to covariate transformations (see Theorem 5 in Appendix A). Fortunately, HPD values are uniform under the local consistency hypothesis:

Theorem 4. *For any $\mathbf{x} \in \mathcal{X}$, if the local null hypothesis $H_0(\mathbf{x}) : \widehat{f}(\cdot|\mathbf{x}) = f(\cdot|\mathbf{x})$ holds, then the distribution of HPD$(Y;\mathbf{x})$ given $\mathbf{x}$ is uniform over $(0,1)$. (The reverse is however not true.)*

It follows that the same techniques developed in Sections 3.3 and 3.4 can be used with HPD values to check global and local consistency for multivariate responses, as well as to construct local P-P plots.

The HPD statistic is especially appealing if one wishes to construct predictive regions with $\widehat{f}$ as HPD values are intrinsically related to highest predictive density sets (Hyndman, 1996). HPD sets are region estimates of $\mathbf{y}$ that contain all $\mathbf{y}$'s for which $\widehat{f}(\mathbf{y}|\mathbf{x})$ is larger than a certain threshold (in the Bayesian case, these are the highest posterior credible regions). More precisely, if $\text{HPD}_\alpha(\mathbf{x})$ is the α-level HPD set for $\mathbf{y}$, then

$$\text{HPD}(\mathbf{y};\mathbf{x}) < \alpha \iff Y \in \text{HPD}_\alpha(\mathbf{x}).$$

Thus, by testing local consistency of $\widehat{f}$ via HPD values, we assess the coverage of HPD sets. It should be noted, however, that even if the HPD values are uniform (conditional on $\mathbf{x}$), it may be the case that $\widehat{f} \neq f$.

Chapter 4

Calibration of Conditional Density Models

This chapter presents a calibration framework for conditional density models that leverages the diagnostic framework developed in Chapter 3 for evaluating the quality of such models. Because the diagnostics introduced in Section 3.2 assess the empirical conditional coverage of a given model $\widehat{f}(y|\mathbf{x})$, we can directly use the detailed information they provide in order to recalibrate $\widehat{f}(y|\mathbf{x})$ towards achieving individual calibration across all locations $\mathbf{x} \in \mathcal{X}$. Our non-parametric and easily interpretable approach is unique both in how it directly targets conditional coverage and, moreover, in how it allows for the construction of entire calibrated predictive distributions rather than just prediction sets; of course, well-calibrated prediction sets are a natural byproduct of well-calibrated predictive distributions. Although estimating entire distributions non-parametrically is difficult, the performance of our method is on par with state-of-the-art predictive inference algorithms for constructing prediction sets, in terms of achieving conditional coverage in practice.

Section 4.1 introduces our approach for obtaining full calibrated conditional PDFs, by using the insights from our new diagnostic framework. Section 4.2 explains how our algorithm for calibrated predictive distributions can easily be adapted for constructing calibrated prediction intervals and HPD sets. Section 4.3 discusses the theoretical properties of our regression approach and the resulting calibrated prediction sets. Section 4.4 presents a stylized example in which our method successfully recovers the full predictive distribution, and achieves state-of-the-art performance for conditional coverage. Section 4.5 shows how our method can handle model misspecifications, in a stylized example of diagnostics and recalibration in a setting with distributional shift. Section 4.6 features an application to a high-impact physics problem that requires good estimates of multimodal distributions. Section 4.7 explores an application to a high-impact nowcasting problem in climate science. Finally, Section 4.8 describes an alternative approach to recalibration that relies on estimates of local HPD coverage rather than local PIT coverage. All code used to produce our experiments is available at `https://github.com/zhao-david/Cal-PIT`.

4.1 Calibrated Full Conditional PDFs

This section introduces the `Cal-PIT` method for a calibrating conditional density model $\widehat{f}(y|\mathbf{x})$, using the probability integral transform (PIT) and a calibration set $\mathcal{D}$ that was not used to train the initial $\widehat{f}(y|\mathbf{x})$. Our approach leverages the key observation that an estimated conditional CDF $\widehat{F}(Y|\mathbf{X})$ is conditionally calibrated if and only if its PIT value is uniformly distributed *conditionally on* $\mathbf{x}$.

For fixed $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, recall that the local probability integral transform (PIT) of y at $\mathbf{x}$ (see Equation 3.2) is given by

$$\text{PIT}(y; \mathbf{x}) := \int_{-\infty}^{y} \widehat{f}(y'|\mathbf{x}) dy' = \widehat{F}(y|\mathbf{x}). \tag{4.1}$$

where $\widehat{F}$ is CDF associated with $\widehat{f}$.

The diagnostics developed in Chapter 3 require the estimation of the true CDF of the PIT values, which we refer to as the PIT-CDF. Note that we defined this earlier in Equation 3.5, but now we need to amend the notation to indicate which CDE $\widehat{f}$ we are computing the PIT coverage for.

Definition 4 (PIT-CDF)**.** *For every $\mathbf{x} \in \mathcal{X}$ and $\gamma \in (0, 1)$, the CDF of the local PIT distribution, for conditional density model $\widehat{f}(y|\mathbf{x})$, is given by*

$$r^{\widehat{f}}(\gamma; \mathbf{x}) := \mathbb{P}\left(PIT(Y; \mathbf{x}) \leq \gamma \mid \mathbf{x}\right) = \mathbb{P}\left(\widehat{F}(Y|\mathbf{x}) \leq \gamma \;\middle|\; \mathbf{x}\right). \tag{4.2}$$

Recall that the PIT-CDF values $r^{\widehat{f}}(\gamma; \mathbf{x})$ characterize the local consistency (Definition 3) of $\widehat{f}$. For a fixed location $\mathbf{x}$, $\widehat{f}(\cdot|\mathbf{x})$ is locally consistent if and only if $r^{\widehat{f}}(\gamma; \mathbf{x}) = \gamma$ for every $\gamma \in (0, 1)$. Hence, by plotting an estimate of $r^{\widehat{f}}(\gamma; \mathbf{x})$ versus γ via amortized local P-P plots (ALPs) (see Section 3.4), we can assess how close $\widehat{f}$ is to f across the entire feature space. We can also describe the type of deviations that might occur; refer back to Figure 3.1 for some representative examples.

As described in Section 3.3, our approach involves learning $r^{\widehat{f}}(\gamma; \mathbf{x})$ using regression. We first augment the calibration data by drawing multiple values of $\gamma \sim U(0, 1)$ for each data point in $\mathcal{D}$ to form the indicator random variable W_i^{γ} (Equation 3.7), then regress W_i^{γ} on both $\mathbf{X}_i$ and γ. As $r^{\widehat{f}}(\gamma; \mathbf{x})$ is a non-decreasing function of γ, we use monotonic neural networks (Wehenkel and Louppe, 2019) as our regression algorithm, though our framework is flexible and any other suitable regression method may be used. Again, thanks to the impressive arsenal of existing regression methods, we can adapt to different types of potentially high-dimensional data. In practice, estimating empirical coverage via regression may be "easier" (or more statistically efficient) than training a consistent CDE. We then obtain the regression function

$$\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x}) := \widehat{\mathbb{P}}(\text{PIT}(Y; \mathbf{x}) \leq \gamma). \tag{4.3}$$

We note that if the initial estimate is perfectly accurate ($\widehat{f} = f$), then we have the simple function $r^{\widehat{f}}(\gamma; \mathbf{x}) = \gamma$, which is very smooth and easy to estimate (indeed, it is constant in $\mathbf{x}$). Thus, the extra step of using our recalibration

procedure should not cost much when the initial CDE estimates are decent. It is when the initial CDE is not well estimated that $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$ becomes nontrivial to learn, but in this case the extra effort is well worth it, because the original CDE estimates are not properly calibrated.

The `Cal-PIT` method uses the estimated regression function $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$ (Equation 4.3) to correct the initial CDE $\widehat{f}(y|\mathbf{x})$, so that the recalibrated CDE, which we denote $\widetilde{f}$, is approximately locally consistent across the feature space. Our strategy for calibration is to "adjust" or "relabel" the quantiles of $\widehat{F}(y|\mathbf{x})$, which indicate nominal coverage according to the initial CDE, so that they match their true achieved coverage as estimated by $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$. In particular, we will perform this relabeling via the direct composition of the $\widehat{r}^{\widehat{f}}$ and $\widehat{F}(y|\mathbf{x})$ functions. Given these two functions, we define the new calibrated CDF $\widetilde{F}$ as follows:

$$\widetilde{F}\left(y|\mathbf{x}\right) := \widehat{r}^{\widehat{f}}\left(\widehat{F}(y|\mathbf{x}); \mathbf{x}\right). \tag{4.4}$$

The intuition behind our recalibration procedure is the following. Suppose that we know the true CDF of $\widehat{F}$; that is, we know $r^{\widehat{f}}(\gamma; \mathbf{x}) := \mathbb{P}\left(\widehat{F}(Y|\mathbf{x}) \leq \gamma \;\middle|\; \mathbf{x}\right)$ for all $\gamma \in (0, 1)$ and every $\mathbf{x} \in \mathcal{X}$. Then, by construction, we have:

$$\begin{aligned} r^{\widehat{f}}\left(\widehat{F}(y|\mathbf{x}); \mathbf{x}\right) &= \mathbb{P}\left(\widehat{F}(Y|\mathbf{x}) \leq \widehat{F}(y|\mathbf{x}) \;\middle|\; \mathbf{x}\right) \\ &= \mathbb{P}\left(Y \leq y|\mathbf{x}\right) \\ &= F(y|\mathbf{x}). \end{aligned}$$

Hence, if the regression is perfectly estimated (that is, $\widehat{r}^{\widehat{f}} = r^{\widehat{f}}$), then, as long as both F and $\widehat{F}$ are continuous and $\widehat{F}$ dominates F (see Assumptions 2 and 3 in Section 4.3 for details), the recalibrated CDF is locally consistent across the feature space:

$$\begin{aligned} \widetilde{F}\left(y|\mathbf{x}\right) &= \widehat{r}^{\widehat{f}}\left(\widehat{F}(y|\mathbf{x}); \mathbf{x}\right) \\ &= r^{\widehat{f}}\left(\widehat{F}(y|\mathbf{x}); \mathbf{x}\right) \\ &= F\left(y|\mathbf{x}\right). \end{aligned}$$

Algorithm 7 details the `Cal-PIT` procedure for constructing re-calibrated CDEs. In practice, for each $\mathbf{x}$ of interest, we first evaluate $\widetilde{F}\left(y|\mathbf{x}\right)$ across a grid G of values y, then use smoothing splines to interpolate between these values. Afterwards, we can differentiate these spline functions to obtain $\widetilde{f}\left(y|\mathbf{x}\right)$, which is our estimate for the calibrated conditional PDF at $\mathbf{x}$.

After calibration, the nominal quantiles of $\widetilde{F}(y|\mathbf{x})$ should reflect the true coverage level; that is, we expect that $\widehat{r}^{\widetilde{f}}(\gamma; \mathbf{x}) \approx \gamma$ for any test point $\mathbf{x}$. Note that $\widehat{r}^{\widetilde{f}}$ refers to the estimated PIT coverage of the new calibrated conditional

PDF, $\widetilde{f}(y|\mathbf{x})$. Indeed, we verify that if $\widehat{r}$ is perfectly estimated (that is, $\widehat{r}^{\widehat{f}} = r^{\widehat{f}}$ and $\widehat{r}^{\widetilde{f}} = r^{\widetilde{f}}$), then:

$$\begin{aligned}
\widehat{r}^{\widetilde{f}}(\gamma; \mathbf{x}) &= r^{\widetilde{f}}(\gamma; \mathbf{x}) \\
&= \mathbb{P}\left(\widetilde{F}(Y|\mathbf{x}) \leq \gamma \;\middle|\; \mathbf{x}\right) \\
&= \mathbb{P}\left(F(Y|\mathbf{x}) \leq \gamma|\mathbf{x}\right) \\
&= \mathbb{P}\left(Y \leq F^{-1}(\gamma|\mathbf{x})|\mathbf{x}\right) \\
&= F\left(F^{-1}(\gamma|\mathbf{x})|\mathbf{x}\right) \\
&= \gamma.
\end{aligned}$$

Algorithm 7 `Cal-PIT` for Calibrated PDFs

Input: initial CDE $\widehat{f}(y|\mathbf{x})$ evaluated at $y \in G$; calibration set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$; oversampling factor K; test points $\mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$; nominal miscoverage level α, flag `HPD` (true if computing HPD sets)
Output: calibrated CDF $\widetilde{F}(y|\mathbf{x})$, calibrated CDE $\widetilde{f}(y|\mathbf{x})$, for all $\mathbf{x} \in \mathcal{V}$

1: **Learn PIT-CDF from augmented and upsampled calibration data $\mathcal{D}'$**
2: Set $\mathcal{D}' \leftarrow \emptyset$
3: **for** i in $\{1, ..., n\}$ **do**
4: **for** j in $\{1, ..., K\}$ **do**
5: Draw $\gamma_{i,j} \sim U(0, 1)$
6: Compute $W_{i,j} \leftarrow \mathbb{I}\left(\text{PIT}(Y_i; \mathbf{X}_i) \leq \gamma_{i,j}\right)$
7: Let $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(\mathbf{X}_i, \gamma_{i,j}, W_{i,j})\}$
8: **end for**
9: **end for**
10: Use $\mathcal{D}'$ to learn $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x}) := \widehat{\mathbb{P}}\left(\text{PIT}(Y; \mathbf{x}) \leq \gamma \mid \mathbf{x}\right)$ via a regression of W on $\mathbf{X}$ and γ, which is monotonic w.r.t. γ.
11:
12: **Calibration using PIT-CDF**
13: **for** $\mathbf{x} \in \mathcal{V}$ **do**
14: **Construct recalibrated CDF and CDE**
15: Compute $\widehat{F}(y|\mathbf{x}) \leftarrow \text{cumsum}(\widehat{f}(y|\mathbf{x}))$ for $y \in G$
16: Let $\widetilde{F}(y|\mathbf{x}) \leftarrow \widehat{r}^{\widehat{f}}(\widehat{F}(y|\mathbf{x}); \mathbf{x})$ for $y \in G$
17: Apply interpolating (or smoothing) splines to obtain $\widetilde{F}(\cdot|\mathbf{x})$
18: Differentiate $\widetilde{F}(y|\mathbf{x})$ to obtain recalibrated PDF $\widetilde{f}(y|\mathbf{x})$ for $y \in G$
19: Renormalize $\widetilde{f}(y|\mathbf{x})$ according to Izbicki and Lee (2016, Section 2.2)
20: **end for**
21: **return** $\widetilde{F}(y|\mathbf{x})$, $\widetilde{f}(y|\mathbf{x})$, for all $\mathbf{x} \in \mathcal{V}$

Note that Line 10 of Algorithm 7 computes exactly the local PIT regression $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$ used to construct the local P-P plots described in Section 3.4. Our procedure applies a correction that, if ideally estimated, would result in perfect calibration of the conditional density at every location $\mathbf{X}$. In other words, the new corrected densities $\widetilde{f}$ would have $\widehat{r}^{\widetilde{f}}(\gamma; \mathbf{x}) = \gamma$ for every γ, across all X, and all the local P-P plots would be perfectly calibrated straight 45° lines.

4.2 Calibrated Predictive Inference

After computing calibrated conditional densities $\widetilde{f}(y|\mathbf{x})$, it is straightforward to construct calibrated prediction sets $C_\alpha(\mathbf{x})$ for $Y|\mathbf{X}=\mathbf{x}$ that are individually calibrated (Equation 2.1). There are various ways to derive valid prediction sets from some full predictive distribution $\widehat{f}(y|\mathbf{x})$. One simple and common formulation is the prediction interval:

$$\text{INT}(\mathbf{x}, l_{\mathbf{x}}, u_{\mathbf{x}}) = \left\{ y : l_{\mathbf{x}} < \widehat{F}(y|\mathbf{x}) < u_{\mathbf{x}} \right\}.$$

To build prediction intervals using our method, we do not even need to explicitly compute the full calibrated conditional PDF $\widetilde{f}(y|\mathbf{x})$; the calibrated conditional quantile function is sufficient. After defining $\widetilde{F}^{-1}(\cdot|\mathbf{x})$ using interpolating or smoothing splines (refer to Algorithm 8), we can, for each $\mathbf{x}$ of interest, obtain the `Cal-PIT` prediction interval at $\mathbf{x}$, defined as

$$C_\alpha(\mathbf{x}) := \left[\widetilde{F}^{-1}(0.5\alpha|\mathbf{x}),\ \widetilde{F}^{-1}(1-0.5\alpha|\mathbf{x}) \right], \tag{4.5}$$

which approximately achieves $1-\alpha$ conditional coverage.

Alternatively, we can use `Cal-PIT` to form Highest Predictive Density (HPD) prediction sets instead of prediction intervals:

$$\text{HPD}(\mathbf{x}, h_{\mathbf{x}}) = \left\{ y : \widehat{f}(y|\mathbf{x}) > h_{\mathbf{x}} \right\}.$$

The oracle (1-α)-level HPD set (based on the true density $f(y|\mathbf{x})$) is defined as

$$\text{HPD}_\alpha(\mathbf{x}) = \{ y : f(y|\mathbf{x}) \geq t_{\mathbf{x},\alpha} \},$$

where $t_{\mathbf{x},\alpha}$ is such that

$$\int_{y \in \text{HPD}_\alpha(\mathbf{x})} f(y|\mathbf{x}) dy = 1-\alpha.$$

HPDs are the smallest prediction sets that have coverage $1-\alpha$, and thus they may be more informative and considerably smaller than quantile-based intervals, while maintaining the conditional coverage at the nominal level; see Section 4.4 for an example with a bimodal predictive distribution.

The `Cal-PIT` estimate of $\text{HPD}_\alpha(\mathbf{x})$ is given by

$$C_\alpha(\mathbf{x}) = \left\{ y : \widetilde{f}(y|\mathbf{x}) \geq \widetilde{t}_{\mathbf{x},\alpha} \right\}, \tag{4.6}$$

where $\widetilde{f}$ is the `Cal-PIT` calibrated CDE derived in Algorithm 7) and $\widetilde{t}_{\mathbf{x},\alpha}$ is such that

$$\int_{y \in C_\alpha(\mathbf{x})} \widetilde{f}(y|\mathbf{x}) dy = 1 - \alpha.$$

Algorithm 8 details the `Cal-PIT` procedure for constructing either prediction intervals or HPD prediction sets.

Algorithm 8 `Cal-PIT` for Calibrated Prediction Sets

Input: initial CDE $\widehat{f}(y|\mathbf{x})$ evaluated at $y \in G$; calibration set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$; oversampling factor K; test points $\mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$; nominal miscoverage level α, flag `HPD` (true if computing HPD sets)
Output: `Cal-PIT` prediction set $C(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{V}$

Learn PIT-CDF from augmented and upsampled calibration data $\mathcal{D}'$
Set $\mathcal{D}' \leftarrow \emptyset$
for i in $\{1, ..., n\}$ **do**
  for j in $\{1, ..., K\}$ **do**
    Draw $\gamma_{i,j} \sim U(0, 1)$
    Compute $W_{i,j} \leftarrow \mathbb{I}\left(\text{PIT}(Y_i; \mathbf{X}_i) \leq \gamma_{i,j}\right)$
    Let $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(\mathbf{X}_i, \gamma_{i,j}, W_{i,j})\}$
  end for
end for
Use $\mathcal{D}'$ to learn $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x}) := \widehat{\mathbb{P}}\left(\text{PIT}(Y; \mathbf{x}) \leq \gamma \mid \mathbf{x}\right)$ via a regression of W on $\mathbf{X}$ and γ, which is monotonic w.r.t. γ.

Calibration using PIT-CDF
for $\mathbf{x} \in \mathcal{V}$ **do**
  Construct recalibrated CDF and CDE
  Compute $\widehat{F}(y|\mathbf{x}) \leftarrow \text{cumsum}(\widehat{f}(y|\mathbf{x}))$ for $y \in G$
  Let $\widetilde{F}(y|\mathbf{x}) \leftarrow \widehat{r}^{\widehat{f}}(\widehat{F}(y|\mathbf{x}); \mathbf{x})$ for $y \in G$
  Apply interpolating (or smoothing) splines to obtain $\widetilde{F}(\cdot|\mathbf{x})$ and $\widetilde{F}^{-1}(\cdot|\mathbf{x})$
  Differentiate $\widetilde{F}(y|\mathbf{x})$ to obtain recalibrated PDF $\widetilde{f}(y|\mathbf{x})$ for $y \in G$
  Renormalize $\widetilde{f}(y|\mathbf{x})$ according to Izbicki and Lee (2016, Section 2.2)

  Construct Cal-`PIT` prediction set with conditional coverage $1 - \alpha$
  Compute interval $C(\mathbf{x}) \leftarrow [\widetilde{F}^{-1}(0.5\alpha|\mathbf{x}); \widetilde{F}^{-1}(1 - 0.5\alpha|\mathbf{x})]$
  if `HPD` **then**
    Obtain HPD sets $C(\mathbf{x}) \leftarrow \{y : \widetilde{f}(y|\mathbf{x}) \geq \widetilde{t}_{\mathbf{x},\alpha}\}$, where $\widetilde{t}_{\mathbf{x},\alpha}$ is such that $\int_{y \in C_\alpha(\mathbf{x})} \widetilde{f}(y|\mathbf{x}) dy = 1 - \alpha$
  end if
end for
return $C(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{V}$

4.3 Theoretical Properties

In this section, we provide convergence rates for the recalibrated CDF estimator $\widetilde{F}$. We also show that `Cal-PIT` intervals achieve asymptotic conditional validity even if the initial CDE $\widehat{f}$ is not consistent, and prove the Fisher

consistency of `Cal-PIT` HPD sets.

We note that the following results are conditional on $\widehat{f}$; that is, all uncertainty comes from the calibration sample, and we take the initial CDE $\widehat{f}$ as given.

We assume that the true distribution of $Y|\mathbf{x}$ and its initial estimate are continuous, and that $\widehat{F}$ places its mass on a region that is at least as large as that of F:

Assumption 2 (Continuity of the cumulative distribution functions)**.** *For every $\mathbf{x} \in \mathcal{X}$, $\widehat{F}(\cdot|\mathbf{x})$ and $F(\cdot|\mathbf{x})$ are strictly continuous functions.*

Assumption 3 ($\widehat{F}$ dominates F)**.** *For every $\mathbf{x} \in \mathcal{X}$, $\widehat{F}(\cdot|\mathbf{x})$ dominates $F(\cdot|\mathbf{x})$.*

To provide convergence rates for the recalibrated CDF, we assume that $F(\cdot|\mathbf{x})$ cannot place too much mass in regions where the initial estimate $\widehat{F}(\cdot|\mathbf{x})$ places little mass:

Assumption 4 (Bounded density)**.** *There exists $K > 0$ such that, for every $\mathbf{x} \in \mathcal{X}$, the Radon-Nikodym derivative of $F(\cdot|\mathbf{x})$ with respect to $\widehat{F}(\cdot|\mathbf{x})$ is bounded above by K.*

Finally, we assume that the regression method converges at a rate $O(n^{-\kappa})$:

Assumption 5 (Convergence rate of the regression method)**.** *The regression method used to estimate $r^{\widehat{f}}$ is such that its convergence rate is given by*

$$\mathbb{E}\left[\int\int \left(\widehat{r}^{\widehat{f}}(\gamma;\mathbf{x}) - r^{\widehat{f}}(\gamma;\mathbf{x})\right)^2 d\gamma dP(\mathbf{x})\right] = O\left(\frac{1}{n^{\kappa}}\right)$$

for some $\kappa > 0$.

Many methods satisfy Assumption 5 for some value κ, which is typically rated to the dimension of $\mathcal{X}$ and the smoothness of the true regression r (see, for instance, Györfi et al. 2002).

Under these assumptions, we can derive the rate of convergence for $\widetilde{F}$:

Theorem 6. *Under Assumptions 2, 3, 4 and 5,*

$$\mathbb{E}\left[\int\int \left(\widetilde{F}(y|\mathbf{x}) - F(y|\mathbf{x})\right)^2 dP(y,\mathbf{x})\right] = O\left(\frac{1}{n^{\kappa}}\right).$$

Next, we show that with a uniformly consistent regression estimator $\widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})$ (see Bierens 1983; Hardle et al. 1984; Liero 1989; Girard et al. 2014 for some examples), `Cal-PIT` intervals achieve asymptotic conditional validity, even if the initial CDE $\widehat{f}(y|\mathbf{x})$ is not consistent.

Assumption 6 (Uniform consistency of the regression estimator)**.** *The regression estimator is such that*

$$\sup_{\mathbf{x}\in\mathcal{X},\gamma\in[0,1]}\left|\widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})-r^{\widehat{f}}(\gamma;\mathbf{x})\right|\xrightarrow[n\longrightarrow\infty]{a.s.}0,$$

where the convergence is with respect to the calibration set $\mathcal{D}$ only; $\widehat{f}$ is fixed.

Theorem 7 (Consistency and conditional coverage of `Cal-PIT` intervals)**.** *Let $C^*_\alpha(\mathbf{x})=\left[F^{-1}(0.5\alpha|\mathbf{x});F^{-1}(1-0.5\alpha|\mathbf{x})\right]$ be the oracle prediction band, and let $C^n_\alpha(\mathbf{x})$ denote the `Cal-PIT` interval. Under Assumptions 2, 3 and 6,*

$$\lambda\left(C^n_\alpha(\mathbf{X})\Delta C^*_\alpha(\mathbf{X})\right)\xrightarrow[n\longrightarrow\infty]{a.s.}0,$$

where λ is the Lebesgue measure in $\mathbb{R}$ and Δ is the symmetric difference between two sets. It follows that $C^n_\alpha(\mathbf{X})$ has asymptotic conditional coverage of $1-\alpha$ (Lei et al., 2018).

Finally, we show the Fisher consistency of HPD prediction sets estimated with the `Cal-PIT` method. For every $\mathbf{x}\in\mathcal{X}$, let

$$C_\alpha(\mathbf{x})=\left\{y:\widetilde{f}(y|\mathbf{x})\geq\widetilde{t}_{\mathbf{x},\alpha}\right\},$$

where $\widetilde{t}_{\mathbf{x},\alpha}$ is such that

$$\int_{y\in C_\alpha(\mathbf{x})}\widetilde{f}(y|\mathbf{x})dy=1-\alpha$$

be the `Cal-PIT` HPD set. Similarly, let

$$\text{HPD}_\alpha(\mathbf{x})=\{y:f(y|\mathbf{x})\geq t_{\mathbf{x},\alpha}\},\tag{4.7}$$

where $t_{\mathbf{x},\alpha}$ is such that

$$\int_{y\in\text{HPD}_\alpha(\mathbf{x})}f(y|\mathbf{x})dy=1-\alpha$$

be the true oracle HPD set. The next theorem shows that if the probabilistic classifier is well estimated, then `Cal-PIT` HPD sets are exactly equivalent to oracle HPD sets.

Theorem 8 (Fisher consistency of `Cal-PIT` HPD sets)**.** *Fix $\mathbf{x}\in\mathcal{X}$. If $\widehat{r}(\gamma;\mathbf{x})=r(\gamma;\mathbf{x})$ for every $\gamma\in[0,1]$, $C_\alpha(\mathbf{x})=HPD_\alpha(\mathbf{x})$ and $\mathbb{P}(Y\in C_\alpha(\mathbf{X})|\mathbf{x})=1-\alpha$.*

4.4 Example: IID Data, No Model Misspecification

Our first example is a low-dimensional stylized version of the galaxy photometric redshift (photo-z) application in Section 4.6, for which we expect the true predictive distributions to be multimodal in some parts of the feature space, as multiple widely different distances (redshifts) can be consistent with the observed features (colors) of a galaxy.

Motivated by the photo-z application, we modify the two-group example of Feldman et al. (2021) to have a bimodal structure. The data for this example consist of two groups with different spreads:

$$
\begin{aligned}
\epsilon_1 &\sim N(0,1), \\
\epsilon_2 &\sim N(0,0.1^2), \\
X_0 &\sim \text{Bern}(0.2), \\
X_{1,2} &\overset{\text{i.i.d.}}{\sim} \text{Unif}[-5,5]^2, \\
Y &= \begin{cases}
3\epsilon_2 + 0.2(X_1+5)\epsilon_1, & X_0 = 0, X_1 < 0 \\
3\epsilon_2 - 0.2(X_1-5)\epsilon_1, & X_0 = 1, X_1 < 0 \\
3\epsilon_2 + 0.2(X_1+5)\epsilon_1 + X_1, & X_0 = 0, X_1 > 0 \\
3\epsilon_2 - 0.2(X_1-5)\epsilon_1 - X_1, & X_0 = 1, X_1 > 0
\end{cases}
\end{aligned}
$$

The target variable Y depends on three variables (X_0, X_1, X_2), with one of the variables (X_0) indicating group membership. However, the practitioner only has access to the predictors X_1 and X_2, resulting in the CDE being bimodal in the regime $X_1 > 0$ with one branch for each class (see "Majority" versus "Minority" in Figure 4.1).

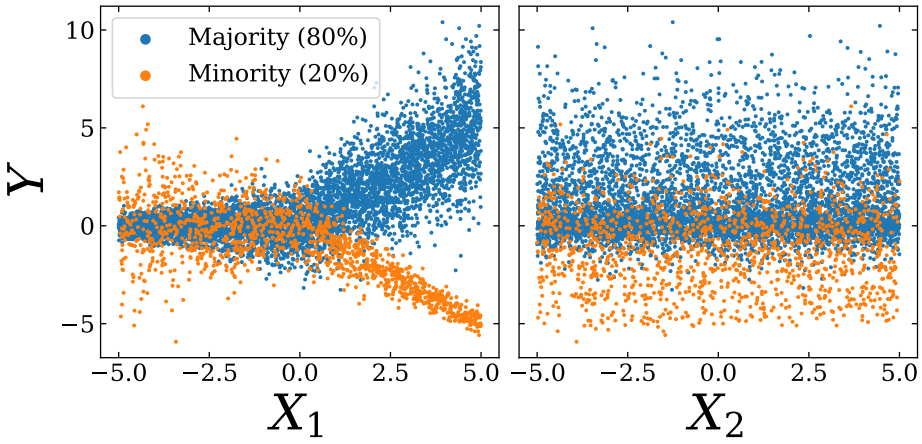


Figure 4.1: Visualization of one random instance of the data used for this example. There are two covariates (X_1, X_2), and a target variable Y. The analytic form of the true data distribution is described in the text. The data set consists of two groups with different spreads. Y splits into two branches for $X_1 > 0$; that is, the true CDE is bimodal in this region.

Our primary goal is to calibrate entire predictive distributions, but because of the rich literature on constructing calibrated prediction sets, we also benchmark `Cal-PIT` prediction sets against results from state-of-the-art quantile regression and conformal inference methods, in addition to comparing our results to oracle bands from Monte Carlo simulations from the true data-generating process.

For benchmarking purposes, we compute 90% prediction sets for Y with `Cal-PIT (INT)` and `Cal-PIT (HPD)`, and then compare the results to prediction sets from five state-of-the-art methods, namely: (i) quantile

regression (QR; Koenker and Bassett Jr. 1978) with a pinball loss; (ii) conformalized quantile regression (`CQR`; Romano et al. 2019); (iii) orthogonal quantile regression (`OQR`; Feldman et al. 2021) which introduces a penalty on the pinball loss to improve conditional coverage; (iv) `Reg-split` (Lei et al., 2018); and (v) distributional conformal prediction (`DCP`; Chernozhukov et al. 2018). Methods (i)-(iv) are all trained with XGBoost (Chen and Guestrin, 2016). Our `Cal-PIT` methods use an initial CDE trained using FlexCode with an XGBoost regressor (Izbicki et al., 2017; Dalmasso et al., 2020) and a monotonic neural network (Wehenkel and Louppe, 2019) for learning $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$. Finally, `DCP` computes a conformity score based on PIT values derived from the same initial CDE as `Cal-PIT`. We split data of total size n equally into train and calibration sets (except for `QR` and `OQR` which use all data for training).

It is mathematically impossible in practice to exactly verify whether conditional validity has been achieved in finite datasets (Zhao et al., 2020). But here in this stylized example, we have knowledge of and are able to simulate from the true data generating process for the synthetic data (which would not in reality be available to a practitioner). We can therefore use Monte Carlo simulations to compute the true conditional coverage at a fixed set of 1000 uniformly sampled test points in $\mathbf{X}$. Similarly, we can compute "oracle" prediction sets.

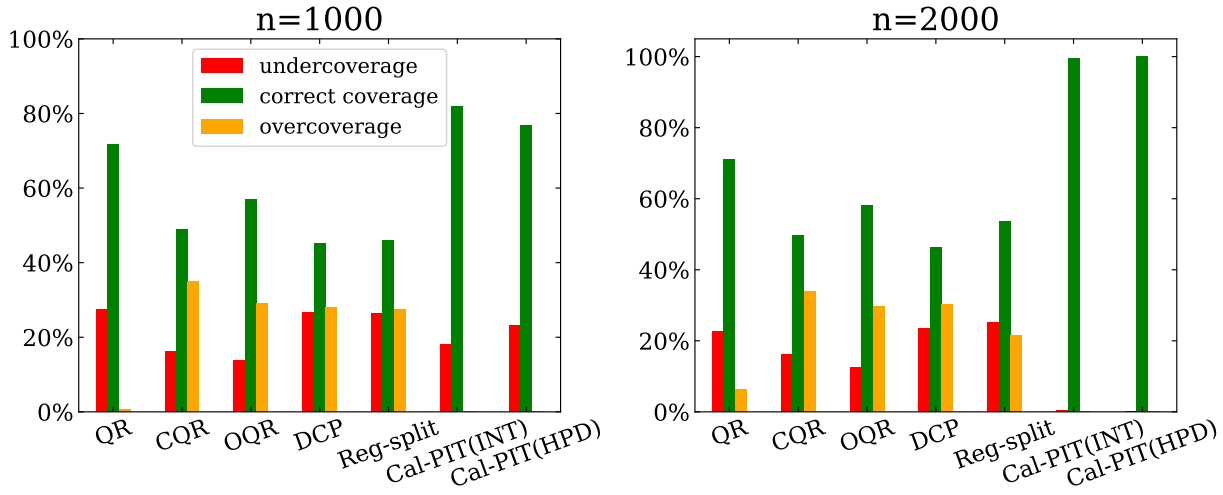


Figure 4.2: The proportion of test points with correct conditional coverage for different methods. Data of total size n are split equally into train and calibration sets (except for QR which uses all data for training). While conformal methods improve upon QR, `Cal-PIT` leads to better conditional coverage, even for smaller sample sizes.

Figure 4.2 compares the conditional coverage of each method, evaluated using Monte Carlo simulations as described above. Test points with coverage within two standard deviations (SD) of $1 - \alpha = 0.9$ based on 100 random realizations are labeled as having "correct" coverage. All methods improve in terms of conditional coverage with increasing sample size, but only `Cal-PIT` consistently attains the nominal 90% coverage across the feature space for $n \geq 2000$.

Figure 4.3 shows the calibrated CDEs from `Cal-PIT`. These estimates reveal that the true conditional density is bimodal for $X_1 > 0$; thus, the most efficient prediction sets in this feature subspace would not be single intervals,

but rather pairs of intervals. Indeed, Figure 4.4 shows that `Cal-PIT (HPD)` yields smaller prediction sets than `Cal-PIT (INT)`. Because HPD sets can capture the bimodality in the data while intervals cannot, this is a case where `Cal-PIT (HPD)` has better efficiency. This qualitative insight is only possible because `Cal-PIT` estimates the entire predictive distributions.

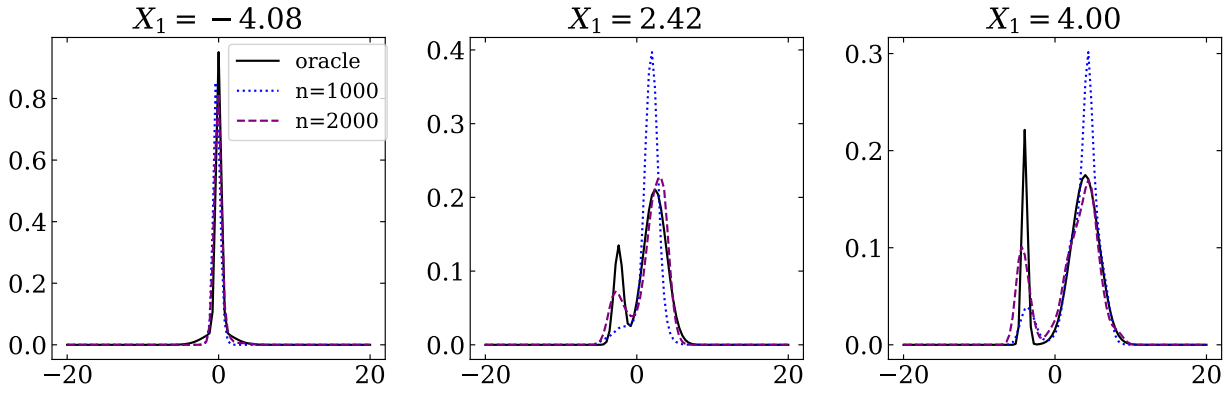


Figure 4.3: Conditional PDFs for sample points at different locations of X_1. The true "oracle" PDF is bimodal for $X_1 > 0$; thus, the most efficient prediction sets in this feature subspace are not single intervals, but pairs of intervals. `Cal-PIT` estimates entire predictive distributions, which converge to oracle predictive distributions as the sample size increases.

Figure 4.4 shows that both `Cal-PIT (INT)` and `Cal-PIT (HPD)` have set sizes that are as small as their optimal counterparts ("Oracle Band" and "Oracle HPD", respectively), and that `Cal-PIT (HPD)` sets are indeed more informative (that is, the regions are smaller) than `Cal-PIT (INT)`.

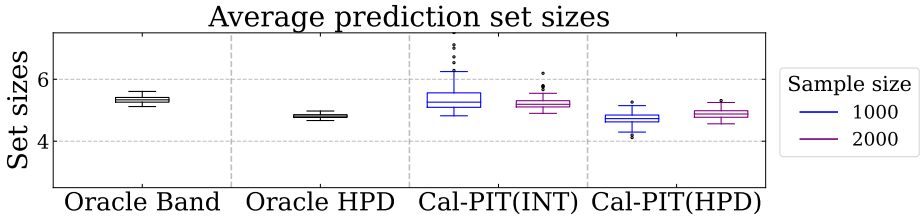


Figure 4.4: Average prediction set sizes for test points for different methods along with the ideal "Oracle Band" and "Oracle HPD". Box plots show the size distribution for multiple trials of the experiment. `Cal-PIT` achieves prediction sets that are at least as tight as those by other methods, while simultaneously providing more accurate coverage.

We saw that this example is difficult for both quantile regression (`QR`) and orthogonal quantile regression (`OQR`) to learn. As described in Section 2.3, the `OQR` method augments the standard pinball loss of `QR` with a penalty on the correlation between prediction set size and coverage, which can improve conditional coverage in certain settings (Feldman et al., 2021), but is not very helpful in this example. Figure 4.5 shows that in this example, the initial prediction sets learned by `QR` have bad conditional coverage, but also do not have much correlation between size and coverage. Thus, the penalty applied by `OQR` is unable to substantially improve upon the `QR` results.

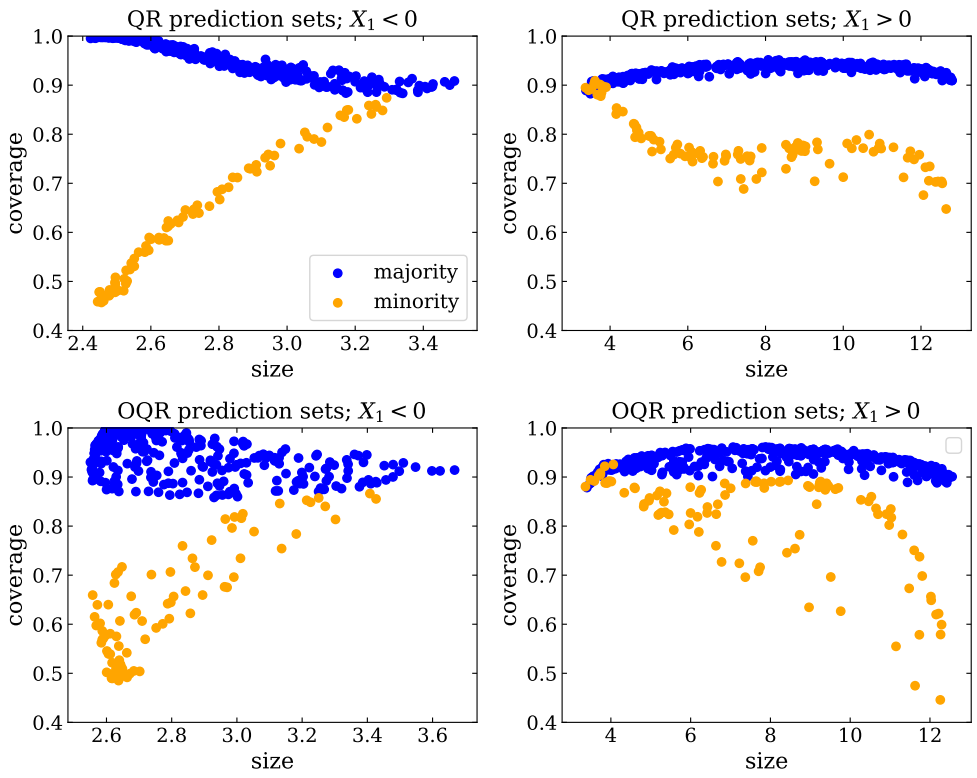


Figure 4.5: *Top:* Prediction sets from quantile regression (`QR`). We see clear correlations between size and coverage, but note that X_0 is not actually available as a predictor, i.e. we cannot "see" the blue and orange colors. The overall correlations, without the colors, are weak. *Bottom:* Prediction sets from orthogonalized quantile regression (`OQR`). Because the overall correlation between size and coverage is weak, penalizing it does not change the results very much. In particular, we still see high correlations (and bad conditional coverage) in the minority group.

We emphasize that methods like `OQR` target *proxies* for conditional coverage, while our `Cal-PIT` method *directly* targets conditional coverage. Therefore, our method succeeds in more general settings. This example is a case where penalizing the correlation between prediction set size and coverage is not a good proxy for achieving conditional coverage, so `OQR` is not as successful as `Cal-PIT` at achieving conditional coverage.

4.5 Example: Misspecified Models

This example demonstrates that our method can effectively diagnose and correct model misspecifications, yielding prediction sets that still achieve conditional coverage when the training and calibration datasets are not the same. We explore a problem with a single predictor X in two different settings: One in which the true target distribution $f(y|x)$ is skewed and a second for which $f(y|x)$ is kurtotic. In both cases, the initial estimate of the distribution, $\widehat{f}(y|x)$ (used for the inputs to `Cal-PIT`) is a Gaussian.

In particular, the training distribution is a Gaussian centered on the true conditional mean,

$$Y_0|X \sim \mathcal{N}(\mu = X, \sigma = 2),$$

while the two target distributions have skew and excess kurtosis relative to the Gaussian training distribution:

$$Y_1|X \sim \text{sinh-arcsinh}(\mu = X, \sigma = 2 - |X|, \gamma = X, \tau = 1),$$
$$Y_2|X \sim \text{sinh-arcsinh}(\mu = X, \sigma = 2, \gamma = 0, \tau = 1 - X/4).$$

These different initial (training) and target (calibration) distributions are illustrated in the left panel of Figure 4.6. The family of sinh-arcsinh normal distributions (Jones and Pewsey, 2009) has been suggested before by Barnes et al. (2021) as a flexible parametric model that supports estimation of the type of heteroscedastic, asymmetric uncertainties often observed in climate data.

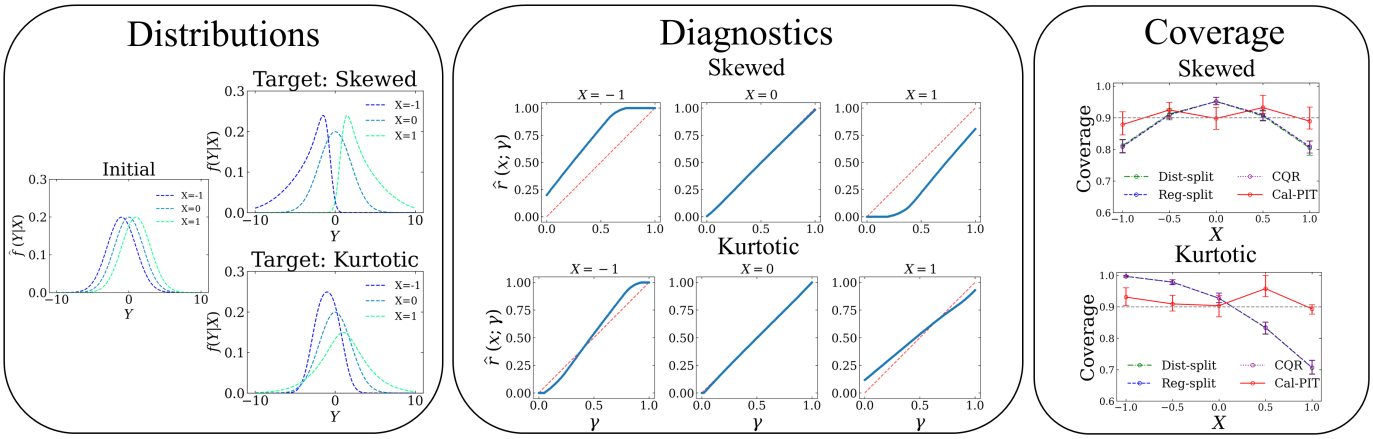


Figure 4.6: *Left*: Initial and target distributions model misspecifications example. The initial fit is Gaussian, but the target distributions are skewed and kurtotic, so the model is mis-specified. Conditional densities for each distribution are shown at slices of X. *Center*: Diagnostic local P-P plots. `Cal-PIT` identifies that, relative to the training density, the skewed observed data are biased at $X = -1/X = 1$ but well estimated at $X = 0$, and that the observed data for the kurtotic target are well estimated at $X = 0$ but under-dispersed at $X = -1$ and over-dispersed at $X = 1$. These insights allow `Cal-PIT` to correct the initial model. *Right*: Conditional coverage obtained via different calibration methods on target data; nominal coverage level $1 - \alpha = 0.9$. `Cal-PIT` is the only method to achieve conditional validity for all inputs X.

We split the data evenly between training and calibration sets with each having 10,000 data points. Using a monotonic neural network regression function for local PIT coverage trained on the calibration set, $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$, we construct "amortized local P-P plots" (ALPs) to show *how* the estimated conditional density $\widehat{f}(y|x)$ deviates from the true density in each setting. These insights are visualized in the center panel of Figure 4.6. We then use `Cal-PIT` to recalibrate the initial predictions on a test set, which comes from the same distribution as the calibration set (but not the training set).

As with the previous example, we are able to draw Monte Carlo samples from the true data generating processes to assess the conditional coverage for various values of X on an independent test set, which is shown in the right panel

of Figure 4.6. The predictive distributions achieve nominal conditional coverage after recalibration using `Cal-PIT` whereas `reg-split`, `CQR` and `DCP` fail to achieve conditional coverage, even though they are calibrated using data from the true data-generating process. Our method is the only one that pinpoints the nature of the discrepancy from the estimated distribution and then directly corrects for deviations in conditional coverage.

Predictive inference has proven challenging in the presence of dataset shift, where the training and test distributions are not the same. While recent work has adapted conformal inference to achieve *marginal* coverage in specific settings, such as covariate shift (Tibshirani et al., 2019) and label shift (Podkopaev and Ramdas, 2021), the assumptions on the specific type of distributional shift can be hard to verify in practice. The method of (Cauchois et al., 2020) estimates the amount of future dataset shift based on the variability in the training data, and achieves approximate marginal validity in the worst-case scenario given a specified amount of shift. A similar method developed by Gibbs and Candès (2021) adaptively forms prediction sets that are robust to dataset shift in an online setting, and achieves approximate marginal validity at most time steps. The related work described above concerns settings where the test set differs from *both* training and calibration sets; dataset shift is such that new test data comes from a distribution that has not been encountered before, and therefore one must either estimate the amount of dataset shift or make distributional assumptions about it, in order to obtain (approximate or exact) marginal validity.

In contrast, this example highlights a situation where the test set is drawn from the same distribution as the calibration set, but different from the training set. In other words, the exchangeability assumption necessary for finite-sample *marginal* validity in conformal prediction still holds, but conformal methods do not obtain *conditional* validity in practice because the conformity scores are constructed using the training distribution, which is misspecified for the calibration and test sets. Our method presents a novel way of diagnosing and correcting this model misspecification using the calibration set, allowing it to achieve much better conditional coverage in practice.

4.6 Application: Photo-z PDF Recalibration

This section applies our method to the real-world problem of photo-z estimation. Redshift, which we denote as z, is a measure of distance to a galaxy and is essential for estimating intrinsic luminosity and 3D location in space, which is crucial information for many astrophysical studies. However, obtaining direct redshift measurements of a large number of objects is prohibitively resource-intensive. Therefore, redshift estimates often must be derived from easier-to-obtain imaging data, resulting in measurements called photometric redshifts or photo-z's. We have chosen this application because of its high impact in a scientific field, and because the photo-z data challenge (DC1) of Schmidt et al. (2020) provides us with clear benchmarks against state-of-the-art CDE methods.

Because images contain limited information about redshifts, multiple redshifts can be consistent with a given photo-z measurement. That is, galaxies at very different redshifts can have similar image properties Conditional density models are commonly used to represent photo-z estimates and their associated uncertainties; the predictive distributions are often multimodal (because of degeneracies), and do not conform to any of the standard models

(Benítez, 2000; Mandelbaum et al., 2008; Malz and Hogg, 2022). Machine learning-based methods are widely used to predict photo-z-distributions when adequate training data are available (e.g., (Beck et al., 2016; Zhou et al., 2021; Dalmasso et al., 2020; Almosallam et al., 2016; Dey et al., 2021)). However, none of these methods can ensure that PDFs are well-calibrated at every point in feature space. Bordoloi et al. (2010) described a method to recalibrate PDFs using a single correction factor based on the overall distribution of PIT values; this ensures a uniform global distribution of PIT values, but does not ensure PDFs are well-calibrated locally in feature space.

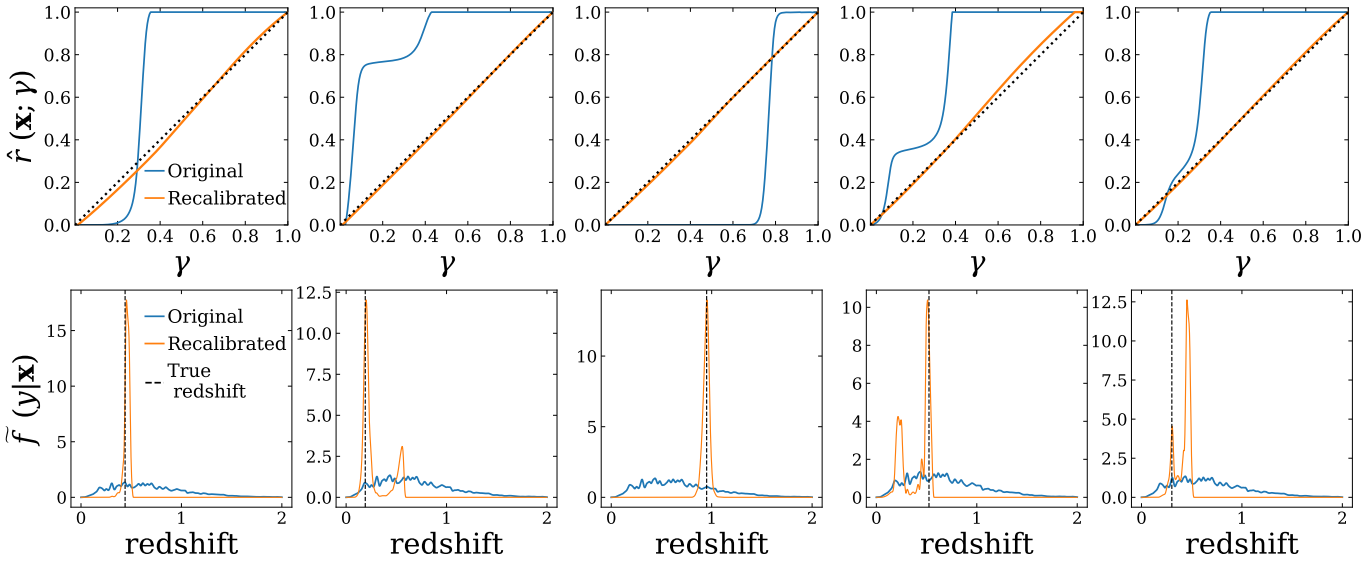


Figure 4.7: *Top:* Diagnostic local P-P plot for five galaxies before and after Cal-PIT is applied. *Bottom:* CDEs for the corresponding galaxies before and after calibration along with their true redshifts. Recalibration using `Cal-PIT` can recover multimodalities while ensuring good conditional coverage.

For this application, we use the simulated data from Schmidt et al. (2020), which has been used to benchmark photo-z CDE prediction methods in the past. The features used to train the models are called apparent magnitudes and colors which are various measures of total light in an image. We use the "training set" from Schmidt et al. (2020) with about 44,000 instances as our calibration set; then split the remaining data into two sets: a validation set (twice as large as the calibration set) and a larger test set comprised of roughly 250,000 instances. We start with the marginal distribution of redshifts as our initial CDE estimate. As described in Section 2.2, Schmidt et al. (2020) demonstrated that such a pathologically misspecified CDE estimate performs well on many commonly used metrics that only check for marginal coverage, while it does not provide information about conditional coverage.

We learn the local distribution of PIT values by training $r^{\widehat{f}}$ on the calibration set and use it to recalibrate the CDEs in our validation and test sets via our `Cal-PIT` method. To assess the quality of our recalibrated CDEs, we train another regression model using the validation set and its recalibrated CDEs. We also use the CDE loss (Izbicki et al., 2017) as another independent metric of conditional coverage. We infer the local CDF of PIT for every instance in the test set before and after recalibration using the two trained models. The top panel of Figure 4.7 shows the diagnostic

local P-P plot for five sample galaxies in the test set. The local CDF of PIT for these instances follows the identity line closely (i.e., the CDF of a uniform distribution), indicating good conditional coverage. The bottom panel of Figure 4.7 also shows that multimodal CDEs can be recovered, even when the input CDE before calibration is unimodal.

Table 4.1: Comparison with methods benchmarked in the LSST-DESC Photo-z Data Challenge (Schmidt et al., 2020). In terms of CDE loss, `Cal-PIT` performs better than all the other methods compared including one approach which was specifically optimized for minimum CDE loss (`FlexZBoost`).

Photo-z Algorithm	CDE Loss
`ANNz2` (Sadeh et al., 2016)	-6.88
`BPZ` (Benítez, 2000)	-7.82
`Delight` (Leistedt and Hogg, 2017)	-8.33
`EAZY` (Brammer et al., 2008)	-7.07
`FlexZBoost` (Izbicki et al., 2017)	**-10.60**
`GPz` (Almosallam et al., 2016)	-9.93
`LePhare` (Arnouts et al., 1999)	-1.66
`METAPhoR` (Cavuoti et al., 2017)	-6.28
`CMNN` (Graham et al., 2018)	-10.43
`SkyNet` (Graff et al., 2014)	-7.89
`TPZ` (Carrasco Kind and Brunner, 2013)	-9.55
`trainZ` (Schmidt et al., 2020)	-0.83
`Cal-PIT`	**-10.71**

We also see a large improvement in the value of the CDE Loss, with a decrease from -0.84 to -10.71 after recalibration. Table 4.1 shows that `Cal-PIT` yields lower CDE loss than any of the cutting-edge methods benchmarked by Schmidt et al. (2020) as part of the LSST-DESC data challenge. The Cramér-von Mises statistic between the local PIT CDF and the uniform distribution is another measure of the quality of conditional coverage (Schmidt et al., 2020). Figure 4.8 shows that the Cramér-von Mises statistic decreases significantly on the entire test set when comparing both fits, with a mean decrease of about 4.5 times.

As discussed, due to the noisy and limited information about redshift contained in galaxy images, galaxies with similar imaging data may have different redshifts and vice versa. We want this property to be captured in photo-z predictive distributions, requiring them to be multimodal. Since we do not know the "ground truth" CDEs, we generally have to rely on indirect methods to assess coverage.

Here, we provide a rudimentary but direct demonstration that the CDEs we predict are indeed meaningful. We compare the CDEs of the five galaxies shown in Figure 4.7 with the distribution of true redshifts of other galaxies with similar imaging data. What does similar imaging data mean? We identify similar galaxies by searching for other galaxies in the training set whose colors and magnitudes (rescaled by subtracting the mean and dividing by the standard deviation for each feature) lie within a Euclidean distance of 0.5 units of our selected galaxies. Figure 4.9 shows their redshift distribution as an inverse-distance weighted histogram along with their CDEs. We observe that the histograms show bimodal distributions when our inferred CDEs are bimodal, and unimodal distributions when our inferred CDEs are unimodal, matching expectations.

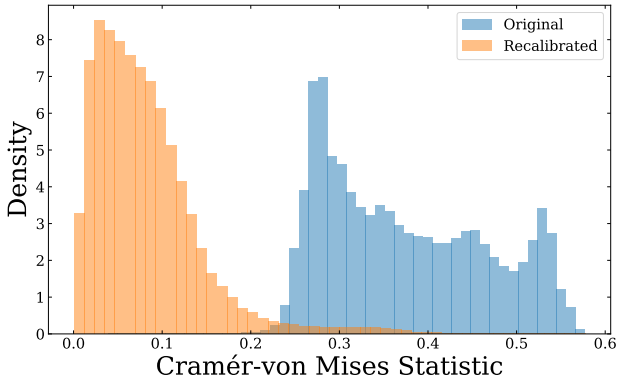


Figure 4.8: Distribution of the Cramér-von Mises (CvM) Statistic (i.e., mean squared difference) between the local PIT CDF of each galaxy in the test set and the CDF of a Uniform distribution. As the "ground truth" CDEs are unknown, we assess conditional coverage by training regression models to predict the local PIT CDFs on the calibration and validation sets. We observe a significant decrease in the value of CvM statistic for the entire test set, with the average value decreasing by $\sim 4.5\times$. The value of CDE loss (Izbicki et al., 2017) which is another independent measure of conditional coverage decreases from -0.84 to -10.71 after recalibration.

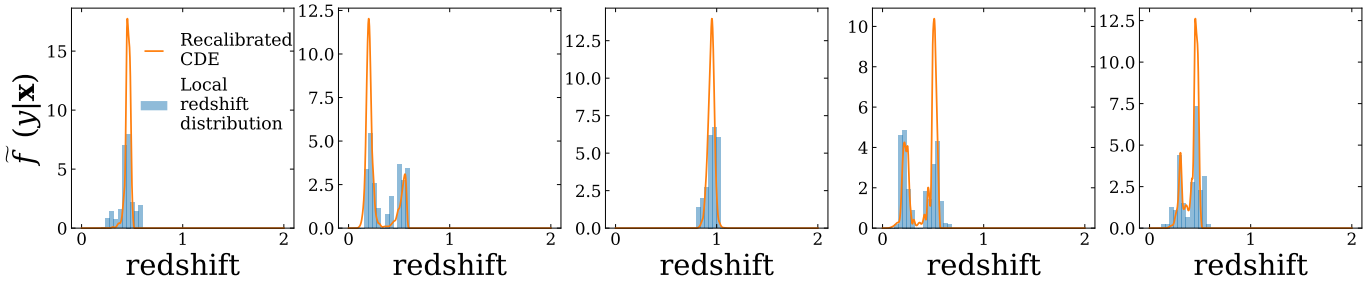


Figure 4.9: Comparison of photo-z CDEs for the galaxies shown in Figure 4.7 with the distribution of true redshifts of other galaxies having similar imaging properties. We observe that the histograms show bimodal distributions only when our inferred CDEs are bimodal.

4.7 Application: Tropical Cyclone Intensity Nowcasting

This section presents an application to probabilistic nowcasting of tropical cyclone intensities. We illustrate `Cal-PIT` calibration of entire predictive distributions of $Y_t|\mathbf{S}_{<t}$ for dependent high-dimensional sequence data $\{(\mathbf{S}_{<t}, Y_t)\}$, which are based on satellite images of tropical cyclones (TCs). The target variable Y_t represents TC intensity at time t, and the predictor $\mathbf{S}_{<t}$ is an entire 24-hour sequence of one-dimensional functions summarizing the spatio-temporal evolution of TC convective structure leading up to time t. The sequence data are strongly correlated as the image sequences from time t to $t+1$ are only shifted by 30 minutes. Here, we simulate from a model fit to observed data so that we can directly assess conditional coverage, via Monte Carlo simulations from a known data generating process.

The original data capture TC convective structure, as observed every 30 minutes by Geostationary Operational Environmental Satellite (GOES) infrared imagery (Janowiak et al., 2020) of storms from the North Atlantic (NAL) and Eastern North Pacific (ENP) basins between 2000-2020; in addition, we have TC intensities at a 6-hour time

resolution from NHC's HURDAT2 best track database (Landsea and Franklin, 2013). Every thirty minutes during the lifetime of a storm, we record a ∼800 km×800 km "stamp" of IR imagery surrounding the TC location, showing cloud-top temperatures for the storm. The left panel of Figure 4.10 shows two such stamps.

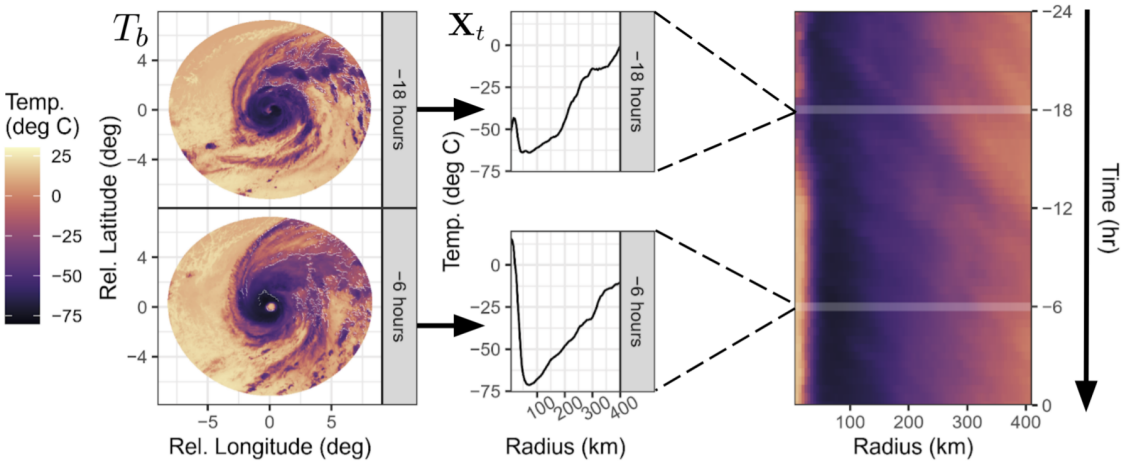
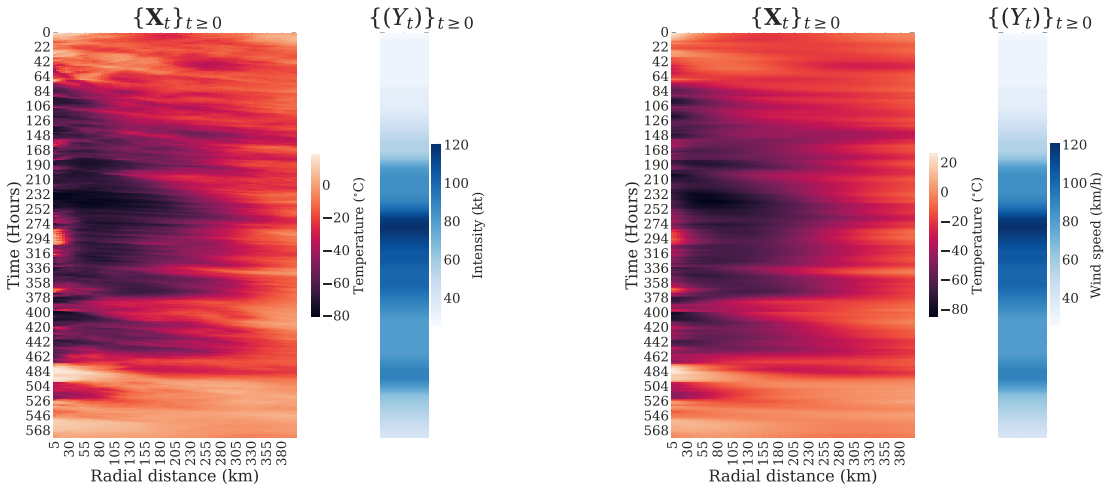


Figure 4.10: *Left:* The raw data is a sequence of TC-centered cloud-top temperature images from GOES. *Center:* We convert each GOES image into a radial profile. *Right:* The 24-hour sequence of consecutive radial profiles, sampled every 30 minutes, defines a structural trajectory or Hovmöller diagram. These trajectories serve as high-dimensional inputs for predicting TC intensity. Figure from (McNeely et al., 2022).

The radial profile, defined as $T(r) = \frac{1}{2\pi}\int_0^{2\pi} T_b(r,\theta)d\theta$, captures the structure of cloud-top temperatures T_b as a function of radius r from the TC center and serves as an easily interpretable description of the depth and location of convection near the TC core (McNeely et al., 2020; Sanabia et al., 2014). The radial profiles are computed at 5-km resolution from 0-400km (d = 80); see Figure 4.10, center. Finally, at each time t we stack the preceding 24 hours (48 profiles) into a structural trajectory, $\mathbf{S}_{<t}$, consisting of an image of the most recent 48 rows of the data. We visualize these summaries over time with Hovmöller diagrams (Hovmöller (1949); see Figure 4.10, right).

Our goal is to create a synthetic data generating process that has a similar dependency structure as actual TCs. The left panel of Figure 4.11 shows an example sequence of observed radial profiles every 30 minutes for a real TC, along with observed wind speed Y. We interpolate Y, which is available every 6 hours, to a 30-minute resolution. Using the radial profiles from all TC data, we perform a principal component analysis (PCA). Figure 4.12 shows the first three principal components, or empirical orthogonal functions (EOFs). The right panel of Figure 4.11 shows the reconstruction of the TC using just these three EOFs.

First, we create a synthetic model for the high-dimensional sequences of TC image data, using PCA reconstruction. Let $\Delta PC_t := PC_t - PC_{t-30m}$ be the 30-minute change in a principal component (PC) coefficient at time t for observed data. We fit a vector autoregression (VAR) model to $(\Delta PC1_t, \Delta PC2_t, \Delta PC3_t)$ to capture the dependence of each component on its own lags as well as the lags of the other components. The model chosen by the BIC criterion has order 3, for a lag of 90 minutes. With the fitted VAR model, we can jointly simulate synthetic time series data for



(a) Observed **(b)** Reconstructed

Figure 4.11: Observed and reconstructed radial profiles $\mathbf{X}_t$ over time for Hurricane Teddy 2020 (*left*). These are recorded every 30 mins. We obtain a decent reconstruction by using the first 3 PCs. Observed wind speed values Y_t, recorded every 6 hours but interpolated on the same 30 min grid (*right*).

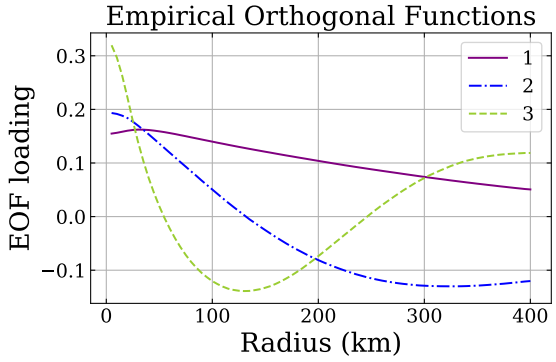


Figure 4.12: Top 3 PCA components, or empirical orthogonal functions (EOFs), for TC radial profiles.

$PC1, PC2, PC3$. A TC structural trajectory is constructed by multiplying simulated time series of PCA coefficients with their corresponding eigenvectors, which are illustrated in Figure 4.12.

Next, we model the time evolution of intensities Y, by fitting a time series regression of intensity change on its past values together with PC coefficients for present and past TC structure. Let $Z := \text{logit}(Y/200)$ so that simulated values of intensities Y are reasonable, i.e. fall between 0 and 200 knots. We then define $\Delta Z_t = Z_t - Z_{t-6h}$. Finally, we fit the following linear regression model for ΔZ:

$$\begin{aligned}\Delta Z_t =&\beta_0 + \beta_1 Z_{t-6h} + \beta_2 \Delta Z_{t-6h} + \beta_3 PC1_t + \beta_4 PC2_t \beta_5 PC3_t + \beta_6 PC1_{t-6h} + \beta_7 PC2_{t-6h} \\ &+ \beta_8 PC3_{t-6h} + \beta_9 PC1_{t-12h} + \beta_{10} PC2_{t-12h} + \beta_{11} PC3_{t-18h} + \beta_{12} PC2_{t-24h} + \epsilon_t\end{aligned} \tag{4.8}$$

where ϵ_t is Gaussian noise with mean 0 and standard deviation set to the root mean squared error between the real and predicted radial profiles in the training set. Note that ΔZ_t has dependencies on its own lagged values as well as lagged values of PC_t. As a sanity check, Figure 4.13 shows that the marginal distributions of the simulated and real wind speed values (Y) look similar.

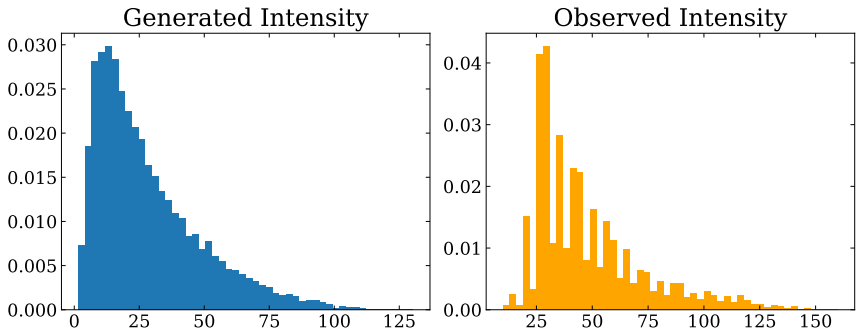


Figure 4.13: *Left:* Marginal distribution of generated wind speed values Y, based on the model in Equation 4.8. *Right:* Marginal distribution of observed wind speed values.

Figure 4.14 shows an example of data from a simulated storm. On the left, we have a Hovmöller diagram of the evolution of TC convective structure $\{(\mathbf{X}_t)\}_{t\geq 0}$, with each row representing the radial profile $\mathbf{X}_t \in \mathbb{R}^{120}$ of cloud-top temperatures as a function of radial distance from the TC center; time evolution is top-down in hours. On the right, we have $\{Y_t\}_{t\geq 0}$, the simulated TC "intensities" at corresponding times t. Let a sequence $\mathbf{S}_{<t} := (\mathbf{X}_{t-48}, \mathbf{X}_{t-47}, \ldots, \mathbf{X}_t)$ include the 24-hour history of convective structure (49 radial profiles). We simulate 800 "storms" from a fitted TC length distribution. Sequence data $\{(\mathbf{S}_{<t}, Y_t)\}$ from the same storm are shifted by 30 minutes; hence, they are *strongly correlated*. Sequence data from different storms, on the other hand, are independent.

Our goal is to construct prediction sets for $Y_t|\mathbf{S}_{<t}$, and illustrate how `Cal-PIT` improves upon an initial MDN fit. With our trained VAR model, we generate a very long time series for $PC1, PC2, PC3$ with a value of the PC's randomly selected from the training set of storms as the initial point. The time series is then divided into 24-hour-long chunks and the structural trajectory and intensities are reconstructed. We create 8000 such instances for our training set, 8000 more for our calibration set, and 4000 instances for our test sets. We rejected a 24-hour long window between each chunk of the time series to ensure that each instance has no memory of the previous ones. Train, calibration, and testing were performed on *different* simulated "storms".

We then fit a unimodal Gaussian neural density model to estimate the conditional density $f(y|\mathbf{s})$ of TC intensities given past radial profiles. Specifically, we fit a convolutional mixture density network (ConvMDN, D'Isanto and Polsterer (2018)) with a single Gaussian component, two convolutional and two fully connected layers which gives an initial estimate of $f(y|\mathbf{s})$. We then use a convolutional neural network with two convolutional layers followed by 5 fully connected layers which take the structural trajectory images and the coverage level (α) as inputs training. The network output is restricted to be monotonic w.r.t. α Wehenkel and Louppe (2019). For both the models we use ReLU

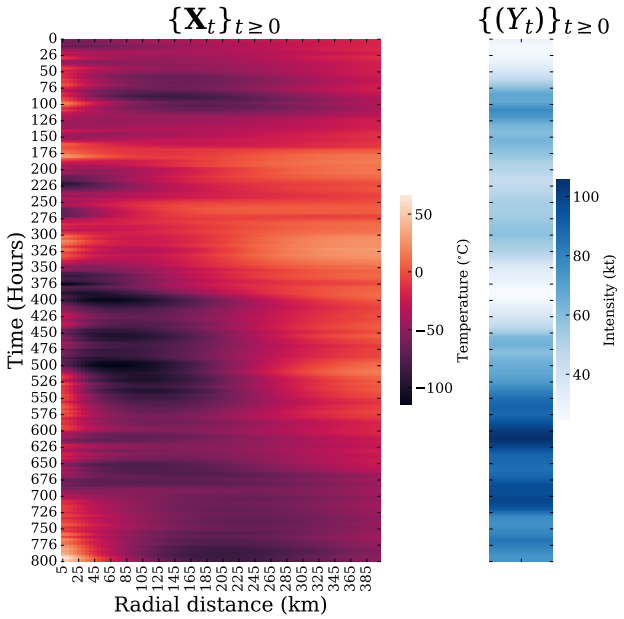


Figure 4.14: Simulated radial profiles and intensities for an example TC. *Left:* Hovmöller diagram of the evolution of TC convective structure $\{\mathbf{X}_t\}_{t\geq 0}$; each row represents the radial profile $\mathbf{X}_t$ of cloud-top temperatures as a function of radial distance from the TC center at time t. Our predictors are 48-hour overlapping sequences $\{\mathbf{S}_t\}_{t\geq 0}$ with data from the same "storm" being highly dependent. *Right:* The target response, here shown as a time series $\{(Y_t)\}_{t\geq 0}$ of simulated TC intensities.

activations (Glorot et al., 2011) for intermediate layers and train using the Adam optimizer (Kingma and Ba, 2014) with learning rate 10^{-3}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

Next, we apply `Cal-PIT` to learn $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{s})$ using 8000 calibration points. Finally, we evaluate the conditional coverage of the initial CDE and `Cal-PIT` on 4000 test points, using Monte Carlo samples from the known data generating process. Figure 4.15 shows that `Cal-PIT` recalibration improves upon the initial ConvMDN fit: The left panel shows prediction sets for $Y_t|\mathbf{S}_{<t}$ for a sample simulated TC, before and after calibration. The calibrated prediction sets track the behavior of the observed trajectory more closely. Moreover, the right panel shows `Cal-PIT` achieves better conditional coverage, even though the effective sample size is small because of dependencies between intensities in the same storm.

The ConvMDN struggles in this example because of the conditional distribution of $Y|\mathbf{S}$ sometimes being skewed towards larger intensities; this phenomenon can partly be observed in Figure 4.16, where we show the distribution of Y_t at fixed values of t for some example simulated TCs. `Cal-PIT` is able to adjust for the model misspecification (similar to the example in Section 4.5), resulting in narrower prediction bands that are still conditionally valid.

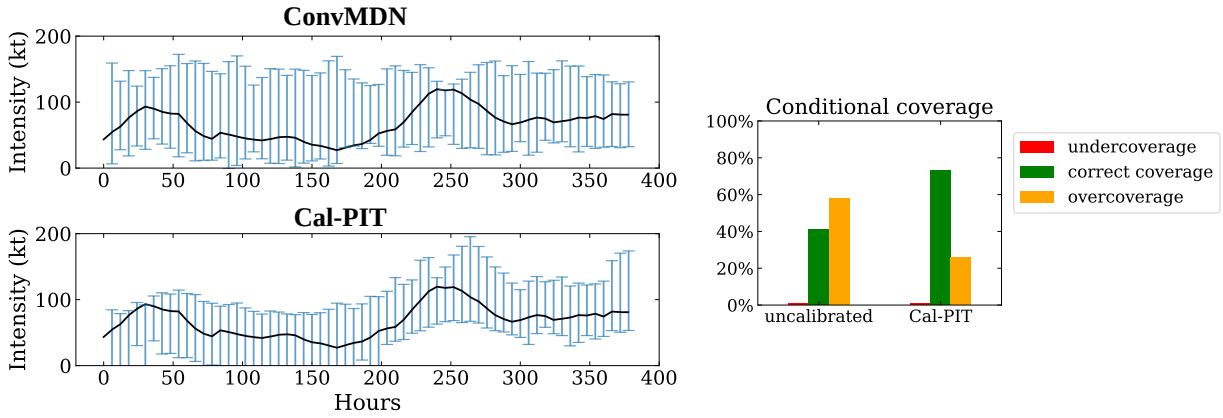


Figure 4.15: *Left:* Simulated TC example with dependent high-dimensional sequence data. Prediction sets for TC intensities, before and after calibration (blue bars), together with the actual trajectory of intensities $\{Y_t\}_t$ (solid black lines). `Cal-PIT` tracks the behavior of the trajectories more closely. *Right:* Conditional coverage of both methods across sequences $\mathbf{s}$. The initial ConvMDN fit with a single Gaussian component over-covers in certain regions of the feature space. `Cal-PIT` partly corrects for the over-coverage and returns more precise prediction sets.

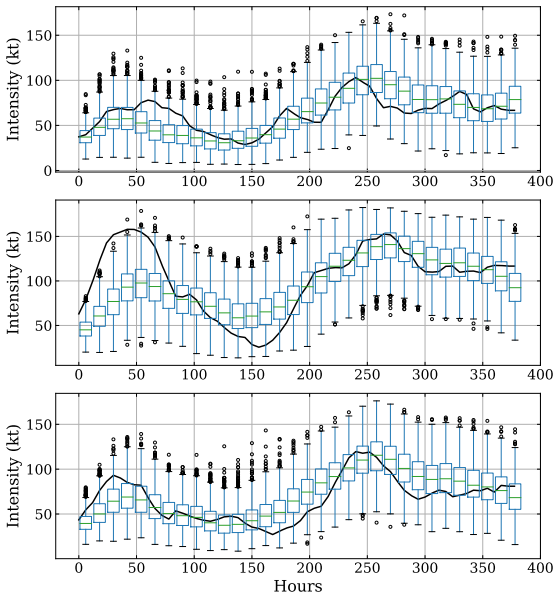


Figure 4.16: Boxplots of the distribution of Y_t at fixed values of t, for simulated TCs. The distributions show skewness, which may explain why the uncalibrated ConvMDN does not fit perfectly. Moreover, the calibrated prediction sets appear to track the observed trajectories (black curves) more closely than the ConvMDN.

4.8 Calibration with Local HPD Coverage

This section proposes an alternative method called `Cal-HPD` that performs diagnostics and recalibration by leveraging local HPD coverage instead of PIT coverage.

We define HPD values as follows:

$$\widehat{H}(y;\mathbf{x}) := \int_{\left\{y:\widehat{f}(y'|\mathbf{x})\leq\widehat{f}(y|\mathbf{x})\right\}} \widehat{f}(y'|\mathbf{x})dy'.$$

Analogously to `Cal-PIT`, the idea of `Cal-HPD` is to estimate the local HPD coverage at each $\mathbf{x}$,

$$h^{\widehat{f}}(\gamma;\mathbf{x}) := \mathbb{P}(\widehat{H}(Y;\mathbf{x}) \leq \gamma|\mathbf{x}),$$

by regression, just as we would estimate the PIT-CDF in `Cal-PIT`. Let $\widehat{h}^{\widehat{f}}(\gamma;\mathbf{x})$ denote such an estimate. The recalibrated $(1-\alpha)$-level HPD set at a location $\mathbf{x}$ is given by the $(1-\alpha^*(\mathbf{x}))$-level HPD set of the original density $\widehat{f}(y|\mathbf{x})$, where $\alpha^*(\mathbf{x})$ is such that

$$\widehat{h}^{\widehat{f}}(\alpha^*(\mathbf{x});\mathbf{x}) = \alpha.$$

One drawback of this framework is that it does not yield full predictive distributions. Moreover, although the approach corrects HPD sets, aiming for conditional coverage, the constructed sets will not be optimal if the initial model $\widehat{f}$ is misspecified. That is, because `Cal-HPD` does not learn and correct the entire predictive distribution, it relies on the consistency of the initial model $\widehat{f}$ and provides no mechanism for, e.g., correcting an initial unimodal CDE into a recalibrated bimodal CDE. In this work, we only report results for `Cal-PIT (INT)` and `Cal-PIT (HPD)`, and we do not report results for Cal-HPD.

Chapter 5

Local Conformalized Calibration

This chapter presents a "hybrid" framework that combines the structure of local conformal prediction with the `Cal-PIT` calibration methodology developed in Chapter 4. At the expense of requiring an additional split of the data, the hybrid approach yields prediction sets with the same finite-sample marginal and local coverage guarantees as local split-conformal methods. The idea is, instead of directly using the calibrated conditional PDF $\widetilde{f}(y|\mathbf{x})$ for building prediction sets, we partition the feature space and perform a conformal adjustment to these prediction sets within each partition element. When defining local regions with such a partition, the hope is to cluster together parts of the feature space where the data points $\mathbf{x}$ are similar, so that estimating a single conformal adjustment to the $\widetilde{f}(y|\mathbf{x})$ function in a partition element leads to good conditional coverage in practice. In finite-sample settings where we have no guarantee of estimating $\widehat{r}^{\widehat{f}}$ (and therefore $\widetilde{f}(y|\mathbf{x})$) perfectly, this hybrid method may lead to better robustness, but the tradeoff is that we need an additional data split if we seek finite-sample coverage guarantees.

Section 5.1 provides background on the structure and goals of local conformal prediction, and describes the `CD-split+` methodology of Izbicki et al. (2022) that in some ways is a foundational framework for our hybrid method. Section 5.2 introduces our hybrid method that essentially puts a local conformal "wrapper" around the calibrated predictive inference method detailed in Section 4.2. Section 5.3 discusses the theoretical properties of our approach, namely that it achieves finite-sample marginal and local validity as well as asymptotic conditional validity. Section 5.4 presents a stylized example that illustrates how finite-sample coverage guarantees can improve conditional coverage when comparing the hybrid method to the original method in Section 4.2, despite the tradeoff of additional sample splitting. Finally, Section 5.5 features an application to a high-impact physics problem where we form prediction sets for multimodal distributions.

5.1 Local Conformal Prediction

Recall that conformal prediction methods have the special property of yielding prediction sets $C_\alpha(\mathbf{X})$ with finite-sample marginal validity (see Equation 2.5); that is,

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}(Y \in C_\alpha(\mathbf{X})) = 1 - \alpha,$$

as long as the data are exchangeable (Vovk et al., 2005). In practice, we usually care more about obtaining conditional validity (see Equation 2.1),

$$\mathbb{P}(Y \in C_\alpha(\mathbf{X}) | \mathbf{X} = \mathbf{x}) = 1 - \alpha\,,\ \forall \mathbf{x} \in \mathcal{X},$$

but exact finite-sample conditional validity is provably impossible to obtain unless one imposes unrealistically strong assumptions on the underlying distribution (Vovk, 2012; Barber et al., 2020).

Local conformal prediction (Lei and Wasserman, 2014; Guan, 2019) offers a practical "middle ground" between practically insufficient marginal validity and practically unattainable conditional validity. This method provides *locally valid* prediction regions. Local validity is a weaker guarantee than conditional validity, but a stronger guarantee than marginal validity. Given a partition $\mathcal{A}$ of the feature space $\mathcal{X}$, it ensures $1 - \alpha$ level coverage not only on average across $\mathbf{X}$, but also within each predefined local region $A \in \mathcal{A}$ of the partition. That is,

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}(Y \in C_\alpha(\mathbf{x}) | \mathbf{x} \in A) > 1 - \alpha\,,\ \forall A \in \mathcal{A}, \tag{5.1}$$

where $\mathcal{A}$ is a partition of $\mathcal{X}$.

After defining a partition $\mathcal{A}$ over the feature space $\mathcal{X}$, local conformal prediction then computes a conformal adjustment for a new instance $\mathbf{x}$ using only the subset of a conformal calibration dataset $\mathcal{D}' = \{(\mathbf{X}'_1, Y'_1), \ldots, (\mathbf{X}'_n, Y'_n)\}$ that falls in the same partition element as $\mathbf{x}$, as defined below:

Definition 5. *Let $\mathcal{A}$ be a partition of $\mathcal{X}$. For each partition element $A \in \mathcal{A}$, let*

$$A(\mathbf{x}) := \{(\mathbf{X}'_i, Y'_i) \in \mathcal{D}' :\ \exists A \in \mathcal{A} \textit{ s.t. } \mathbf{x} \in A \textit{ and } \mathbf{X}'_i \in A\}. \tag{5.2}$$

In the large-sample limit of an infinite number of infinitesimal partition elements, local validity effectively becomes equivalent to conditional validity ($A(\mathbf{x})$ in the limit will only contain $\mathbf{x}$ itself). Hence, local conformal prediction methods typically satisfy asymptotic conditional validity (Lei and Wasserman, 2014; Guan, 2019; Izbicki et al., 2022). In practice, with finite data samples, local conformal prediction is only useful if there are enough data points in each partition element A to perform a meaningful conformal adjustment. Therefore, the key practical challenge is how to define the partition $\mathcal{A}$ such that data points within each partition element are relatively homogeneous (and therefore benefit similarly from the conformal adjustment), while ensuring there are not too many partition elements so that

there is still a sufficient sample size within each element. For instance, if the partition were defined simply using the Euclidean distance over the feature space $\mathcal{X}$, this scheme would not scale to high-dimensional feature spaces (Lei and Wasserman, 2014; Barber et al., 2020; Tibshirani et al., 2019). The curse of dimensionality means that in high-dimensional feature spaces, Euclidean neighborhoods need to be large in order to contain sufficiently many data points. Therefore, partition elements are likely to be heterogeneous, which means there may still be strong deviations from conditional validity even after ensuring local validity across each partition element.

The `CD-split+` method of Izbicki et al. (2022) addresses the issue of scaling to high dimensions by partitioning the feature space based on a fitted conditional density model $\widehat{f}(y|\mathbf{x})$. Essentially, it group two points $\mathbf{x}_i$ and $\mathbf{x}_j$ into the same partition element if their fitted conditional PDFs $\widehat{f}(Y|\mathbf{x}_i)$ and $\widehat{f}(Y|\mathbf{x}_j)$ are sufficiently similar overall (comparing across various quantiles with the notion of a profile distance; see Defn.16 in Izbicki et al. (2022)). After defining this partition, `CD-split+` then takes the fitted CDE scores $\widehat{f}(y_i|\mathbf{x}_i)$ themselves as the conformity score, and computes a conformal adjustment within each partition element. Since the fitted CDE itself is used for the conformity score, and points with similar fitted CDEs are grouped together, this method hopes to produce relatively homogeneous partition elements and achieve approximate conditional coverage.

The `CD-split+` method depends crucially upon having a well-fitted initial CDE $\widehat{f}(y|\mathbf{x})$. Partition elements will only be relatively homogeneous if $\widehat{f}(y|\mathbf{x})$ is a decent approximation of the true predictive distribution $f(y|\mathbf{x})$. In the next section, we introduce a new local conformal prediction method that similarly relies on CDEs for conformity scores and for defining a partition, but leverages the calibration framework we developed in Chapter 4 so that instead of using an initial CDE $\widehat{f}(y|\mathbf{x})$, we compute everything using a calibrated CDE $\widetilde{f}(y|\mathbf{x})$. This allows us to achieve asymptotic conditional validity in theory as well as better conditional coverage in practice, even when the initial CDE $\widehat{f}(y|\mathbf{x})$ is misspecified or poorly fit.

5.2 Local Conformalized Calibrated Predictive Inference

This section presents our local conformalized approach to calibrated predictive inference. As discussed in Section 5.1, one way of viewing this approach is that it is essentially the `CD-split+` method of Izbicki et al. (2022) but using the calibrated CDE $\widetilde{f}(y|\mathbf{x})$ instead of the initial CDE $\widehat{f}(y|\mathbf{x})$. Another perspective is that it is essentially the same as the calibration method detailed in Section 4.2, except that we use an extra data split in order to put a conformal "wrapper" around the prediction sets to ensure finite-sample marginal and local validity.

Recall that after we obtain a calibrated conditional PDF $\widetilde{f}(y|\mathbf{x})$, it is straightforward to compute either calibrated prediction intervals (see Equation 4.5)

$$C_\alpha(\mathbf{x}) := \left[\widetilde{F}^{-1}(0.5\alpha|\mathbf{x}),\ \widetilde{F}^{-1}(1-0.5\alpha|\mathbf{x})\right],$$

or calibrated HPD sets (see Equation 4.6),

$$C_\alpha(\mathbf{x}) = \left\{ y : \widetilde{f}(y|\mathbf{x}) \geq \widetilde{t}_{\mathbf{x},\alpha} \right\},$$

where $\widetilde{t}_{\mathbf{x},\alpha}$ is such that

$$\int_{y \in \text{HPD}_\alpha(\mathbf{x})} \widetilde{f}(y|\mathbf{x}) dy = 1 - \alpha.$$

If the regression function $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$ is well estimated, the PDFs $\widetilde{f}(y|\mathbf{x})$ should be approximately correct, which means the calibration prediction intervals or HPD sets should be approximately conditionally valid. While we derive asymptotic conditional validity and efficiency results in Section 4.3, we cannot guarantee that $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$ is always well estimated in practice. By incorporating the framework of local conformal prediction, we can obtain a sensible partition $\mathcal{A}$ of the feature space and ensure exact finite-sample validity within each partition element $A \in \mathcal{A}$, which may also improve conditional coverage in cases where $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$ is not well estimated.

To form the partition $\mathcal{A}$, we first use the calibration framework presented in Chapter 4 to obtain calibrated PDFs $\widetilde{f}(y|\mathbf{x})$, computed over the calibration set. Then, we define a *model diagnostic distance*, similar in spirit to LeRoy and Zhao (2021), between two points $\mathbf{x}_i$ and $\mathbf{x}_j$, computed over a grid of y values covering the support of $\widetilde{f}(y|\mathbf{x})$:

$$d^2_{\text{md}}(\mathbf{x}_i, \mathbf{x}_j) := \sum_{\substack{y \in \\ [y_0, y_0+\delta, y_0+2\delta, \ldots, y_1]}} \left(\widetilde{f}(y|\mathbf{x}_i) - \widetilde{f}(y|\mathbf{x}_j) \right)^2 . \tag{5.3}$$

With this distance, we can use a clustering algorithm to obtain a partition $\mathcal{A}$ over the calibration set, based on a specified number of clusters K. In our method, we use the k-means++ algorithm, which identifies K cluster centroids $C_1, \ldots, C_K$. Each calibration point $\mathbf{X}_i$ is then assigned to the cluster that has the closest centroid C_k based on the distance d^2_{md}. This idea is similar to how Izbicki et al. (2022) partitions the feature space for local conformal inference, by defining a *profile distance* between the initial CDEs $\widehat{f}(y|\mathbf{x}_i)$ and $\widehat{f}(y|\mathbf{x}_j)$. Our method can yield a sensible partition $\mathcal{A}$ even when the initial CDE is poorly fit.

So far, we have implicitly required a training set $\mathcal{D}_{train}$ to learn the initial CDE $\widehat{f}(y|\mathbf{x})$ as well as a calibration set $\mathcal{D}$ to learn the regression function $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$, from which we can derive calibrated PDFs $\widetilde{f}(y|\mathbf{x})$ anywhere in the feature space. Section 4.2 describes how we can then obtain calibrated prediction sets C_α. To implement our local conformalized approach, we now need an additional data split to provide a wrapper around the calibrated prediction sets; call this the "conformal data set", $\mathcal{D}' = (\mathbf{x}'_1, y'_1), \ldots, (\mathbf{x}'_n, y'_n)$. Across this third data set, we now compute a set of conformity scores, using $\widetilde{f}(y'_1|\mathbf{x}'_1), \ldots, \widetilde{f}(y'_n|\mathbf{x}'_n)$.

Finally, consider a new test point $\mathbf{x}$, and let $A_{\mathbf{x}}$ be the partition element that $\mathbf{x}$ belongs to (see Definition 5). To obtain $(1-\alpha)$-level prediction intervals, we first compute conformity scores $\widetilde{F}(y'_1|\mathbf{x}'_1), \ldots, \widetilde{F}(y'_n|\mathbf{x}'_n)$ across $\mathcal{D}'$.

Algorithm 9 `Local Conformalized Calibrated Prediction Sets`

Input: regression function $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$ trained on calibration set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ according to Algorithm 7; conformal data set $\mathcal{D}' = \{(\mathbf{x}'_1, y'_1), \ldots, (\mathbf{x}'_n, y'_n)\}$; test points $\mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$; initial CDE $\widehat{f}(y|\mathbf{x})$ evaluated at $y \in G$; nominal miscoverage level α; flag `HPD` (true if computing HPD sets)

Output: local conformalized `Cal-PIT` prediction set $C(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{V}$

for $(\mathbf{x}', y') \in \mathcal{D}'$ **do**
  Compute conformity scores on conformal data set $\mathcal{D}'$
  Compute $\widehat{F}(y|\mathbf{x}') \leftarrow \text{cumsum}(\widehat{f}(y|\mathbf{x}'))$ for $y \in G$
  Let $\widetilde{F}(y|\mathbf{x}') \leftarrow \widehat{r}^{\widehat{f}}(\widehat{F}(y|\mathbf{x}'); \mathbf{x}')$ for $y \in G$
  Apply interpolating (or smoothing) splines to obtain $\widetilde{F}(\cdot|\mathbf{x}')$ and $\widetilde{F}^{-1}(\cdot|\mathbf{x}')$
  Differentiate $\widetilde{F}(y|\mathbf{x}')$ to obtain recalibrated PDF $\widetilde{f}(y|\mathbf{x}')$ for $y \in G$
  Renormalize $\widetilde{f}(y|\mathbf{x}')$ according to Izbicki and Lee (2016, Section 2.2)
  if `HPD` **then**
    Compute conformity score $\widetilde{H}(y'; \mathbf{x}') = \displaystyle\int_{\left\{y: \widetilde{f}(y|\mathbf{x}') \geq \widetilde{f}(y'|\mathbf{x}')\right\}} \widetilde{f}(y|\mathbf{x}')dy$
  else
    Compute conformity score $\widetilde{F}(y'|\mathbf{x}')$
  end if
end for

Calibration using PIT-CDF and local conformity scores
Define partition $\mathcal{A}$ via k-means++ clustering algorithm, using $d^2_{md}(\mathbf{x}_i, \mathbf{x}_j)$ as the distance metric (see Def.5).
for $\mathbf{x} \in \mathcal{V}$ **do**
  Compute $\widehat{F}(y|\mathbf{x}) \leftarrow \text{cumsum}(\widehat{f}(y|\mathbf{x}))$ for $y \in G$
  Let $\widetilde{F}(y|\mathbf{x}) \leftarrow \widehat{r}^{\widehat{f}}(\widehat{F}(y|\mathbf{x}); \mathbf{x})$ for $y \in G$
  Apply interpolating (or smoothing) splines to obtain $\widetilde{F}(\cdot|\mathbf{x})$ and $\widetilde{F}^{-1}(\cdot|\mathbf{x})$
  Differentiate $\widetilde{F}(y|\mathbf{x})$ to obtain recalibrated PDF $\widetilde{f}(y|\mathbf{x})$ for $y \in G$
  Renormalize $\widetilde{f}(y|\mathbf{x})$ according to Izbicki and Lee (2016, Section 2.2)
  Determine partition element $A(\mathbf{x})$.
  if `HPD` **then**
    Define α' as the α-quantile of $\left\{\widetilde{H}(y'_j; \mathbf{x}'_j) : \mathbf{x}'_j \in A_{\mathbf{x}}\right\}$
    Obtain HPD sets $C(\mathbf{x}) = \left\{y : \widetilde{f}(y|\mathbf{x}) \geq T_{\mathbf{x},\alpha}\right\}$, where $T_{\mathbf{x},\alpha}$ is s.t. $\displaystyle\int_{\left\{y: \widetilde{f}(y|\mathbf{x}) \geq T_{\mathbf{x},\alpha}\right\}} \widetilde{f}(y|\mathbf{x})dy = 1 - \alpha'$
  else
    Obtain intervals $C(\mathbf{x}) = \left\{y : L_{\mathbf{x},\alpha} \leq \widetilde{F}(y|\mathbf{x}) \leq U_{\mathbf{x},\alpha}\right\}$, where $L_{\mathbf{x},\alpha}$ is the 0.5α-quantile and $U_{\mathbf{x},\alpha}$ is the $(1 - 0.5\alpha)$-quantile of $\left\{\widetilde{F}(y'_j|\mathbf{x}'_j) : \mathbf{x}'_j \in A_{\mathbf{x}}\right\}$
  end if
end for
return $C(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{V}$

Then, the local conformalized calibrated prediction interval at $\mathbf{x}$ is:

$$C_\alpha(\mathbf{x}) = \left\{ y : L_{\mathbf{x},\alpha} \leq \widetilde{F}(y|\mathbf{x}) \leq U_{\mathbf{x},\alpha} \right\} \tag{5.4}$$

where $L_{\mathbf{x},\alpha}$ is the 0.5α-quantile of the conformity scores from $\mathcal{D}'$ that are *in the same partition as* $\mathbf{x}$, namely $\left\{ \widetilde{F}(y'_j|\mathbf{x}'_j) : \mathbf{x}'_j \in A_{\mathbf{x}} \right\}$, and $U_{\mathbf{x},\alpha}$ is the $(1 - 0.5\alpha)$-quantile of those scores. Similarly, to obtain $(1-\alpha)$-level HPD sets, we first compute conformity scores $\widetilde{H}(y'_1; \mathbf{x}'_1), \ldots, \widetilde{H}(y'_n; \mathbf{x}'_n)$ across $\mathcal{D}'$, where

$$\widetilde{H}(y'; \mathbf{x}') = \int_{\left\{ y : \widetilde{f}(y|\mathbf{x}') \geq \widetilde{f}(y'|\mathbf{x}') \right\}} \widetilde{f}(y|\mathbf{x}') dy.$$

Then, the local conformalized calibrated prediction HPD set at $\mathbf{x}$ is

$$C_\alpha(\mathbf{x}) = \left\{ y : \widetilde{f}(y|\mathbf{x}) \geq T_{\mathbf{x},\alpha} \right\} \tag{5.5}$$

where $T_{\mathbf{x},\alpha}$ is such that

$$\int_{\left\{ y : \widetilde{f}(y|\mathbf{x}) \geq T_{\mathbf{x},\alpha} \right\}} \widetilde{f}(y|\mathbf{x}) dy = 1 - \alpha'$$

and α' is the α-quantile of the conformity scores from $\mathcal{D}'$ that are *in the same partition as* $\mathbf{x}$, namely the set $\left\{ \widetilde{H}(y'_j; \mathbf{x}'_j) : \mathbf{x}'_j \in A_{\mathbf{x}} \right\}$. In other words, the thresholds for determining prediction intervals and HPD sets are not directly taken from $\widetilde{f}(y|\mathbf{x})$, but first undergo a conformal adjustment based on appropriate conformity scores computed on the subset of $\mathcal{D}'$ that falls in the same partition element $A_{\mathbf{x}}$ as the test point $\mathbf{x}$.

To ensure local conformal validity, the data used to obtain the local region partition $\mathcal{A}$ must be independent of the data used to compute conformity scores. That is, we reiterate that compared to our original calibration framework, we now need an additional "conformal data set", $\mathcal{D}' = \{(\mathbf{X}'_1, Y'_1), \ldots, (\mathbf{X}'_n, Y'_n)\}$, on which to compute conformity scores. In total, the procedure requires a three-part split of the data: (1) $\mathcal{D}_{train}$ for training the initial CDE $\widehat{f}$; (2) $\mathcal{D}$ for training $\widehat{r}^{\widehat{f}}$, which is used to construct calibrated PDFs $\widetilde{f}(y|\mathbf{x})$ as well as to define the partition $\mathcal{A}$; (3) $\mathcal{D}'$ over which to compute conformity scores $\widetilde{f}(y|\mathbf{x})$, then used to calibrate thresholds in each partition element. In practice, the comparison between our original calibration framework and this local conformalized version is the tradeoff between guaranteeing finite-sample local validity vs. requiring an extra data split.

Algorithm 9 details the local conformalized `Cal-PIT` procedure for constructing either prediction intervals or HPD prediction sets.

5.3 Theoretical Properties

In this section, we show that the proposed method achieves marginal and local validity. We also show that the method obtains asymptotic conditional validity, even if the initial CDE $\widehat{f}$ is not consistent.

Consider a new test point $\mathbf{X}_{n+1}$. Let $cs : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ be a conformity score function (in our case, either $\widetilde{F}(y|\mathbf{x})$ or $\widetilde{H}(y'\mathbf{x})$), which was trained on the calibration set $\mathcal{D}$. What we call the conformal set $\mathcal{D}'$ is then used to calculate *split residuals*, $U_i = cs(\mathbf{X}'_i, Y'_i),\ i = 1, \ldots, n$. Under the assumption that the split residuals are exchangeable (with each other and with the new test point), the rank of $U_{n+1} = cs(\mathbf{X}_{n+1}, Y_{n+1})$ is uniform among $1, \ldots, n+1$. Let $U_{\lfloor \alpha \rfloor}$ denote the $\lfloor n\alpha \rfloor$-th order statistic of $U_1, \ldots, U_n$. We then have

$$\begin{aligned} \mathbb{P}\left(U_{n+1} \geq U_{\lfloor \alpha \rfloor}\right) &\geq 1 - \alpha \\ \mathbb{P}\left(cs(\mathbf{X}_{n+1}, Y_{n+1}) \geq U_{\lfloor \alpha \rfloor}\right) &\geq 1 - \alpha \\ \mathbb{P}\left(Y_{n+1} \in \left\{y : cs(\mathbf{X}_{n+1}, y) \geq U_{\lfloor \alpha \rfloor}\right\}\right) &\geq 1 - \alpha. \end{aligned}$$

In other words, $\left\{y : cs(\mathbf{X}_{n+1}, y) \geq U_{\lfloor \alpha \rfloor}\right\}$ is a marginally valid prediction region for Y given $\mathbf{X}_{n+1}$.

Now, let $\mathcal{A}$ be a partition of $\mathcal{X}$, and for each $A \in \mathcal{A}$, let $A(\mathbf{X}_{n+1})$ be as defined in Definition 5. Let $U^A_{\lfloor \alpha \rfloor}$ denote the $\lfloor |A(\mathbf{X}_{n+1})| \cdot \alpha \rfloor$-th order statistic of the U_i for only those points $(\mathbf{X}_i, Y_i)$ that fall within $A(\mathbf{X}_{n+1})$.

Theorem 9. *If $\mathcal{D}' = \{(\mathbf{X}'_1, Y'_1, \ldots, (\mathbf{X}'_n, Y'_n))\}$ are exchangeable with each other and with new test point $(\mathbf{X}_{n+1}, Y_{n+1})$, then*

$$C(\mathbf{X}_{n+1}) = \left\{y : cs(\mathbf{X}_{n+1}, y) \geq U^A_{\lfloor \alpha \rfloor}\right\}$$

satisfies marginal validity as well as local validity with respect to $\mathcal{A}$.

Since our local conformalized calibrated prediction method is directly based on the split conformal method with adaptive cutoffs, it follows from Theorem 9 that, so long as the points in the conformal set $\mathcal{D}'$ are exchangeable, our method satisfies marginal and local validity:

Theorem 10. *If the points in conformal set $\mathcal{D}' = \{(\mathbf{X}'_1, Y'_1, \ldots, (\mathbf{X}'_n, Y'_n))\}$ are exchangeable with each other and with new test point $(\mathbf{X}_{n+1}, Y_{n+1})$, then the local conformalized calibration prediction method, as detailed in Algorithm 9, satisfies marginal validity as well as local validity with respect to its defined partition $\mathcal{A}$.*

Proof. This follows directly from Theorem 9 and the definition of Algorithm 9. □

Next, we show that with a uniformly consistent regression estimator $\widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$, our method produces prediction intervals and HPD sets that achieve asymptotic conditional validity, even if the initial CDE $\widehat{f}(y|\mathbf{x})$ is not consistent.

Theorem 11. *Let $C^*_\alpha(\mathbf{x}) = \left[F^{-1}(0.5\alpha|\mathbf{x}); F^{-1}(1 - 0.5\alpha|\mathbf{x})\right]$ be the oracle prediction band, and let $C^n_\alpha(\mathbf{x})$ denote the prediction interval output by Algorithm 9. Under Assumptions 2, 3, and 6 (refer to Section 4.3),*

$$\lambda\left(C^n_\alpha(\mathbf{X}) \Delta C^*_\alpha(\mathbf{X})\right) \xrightarrow[n \longrightarrow \infty]{a.s.} 0,$$

where λ is the Lebesgue measure in $\mathbb{R}$ and Δ is the symmetric difference between two sets. It follows that $C^n_\alpha(\mathbf{X})$ has asymptotic conditional coverage of $1 - \alpha$ (Lei et al., 2018).

Similarly, the next theorem shows that if the probabilistic classifier is well estimated, then our local conformalized HPD sets converge asymptotically to oracle HPD sets, $\text{HPD}_\alpha(\mathbf{x}) = \{y : f(y|\mathbf{x}) \geq t_{\mathbf{x},\alpha}\}$ (see Equation 4.7).

Theorem 12. *Fix $\mathbf{x} \in \mathcal{X}$, and let $C_\alpha^n(\mathbf{x})$ denote the prediction HPD set output by Algorithm 9. If $\widehat{r}(\gamma; \mathbf{x}) = r(\gamma; \mathbf{x})$ for every $\gamma \in [0, 1]$, then $\lambda\left(C_\alpha^n(\mathbf{X}) \Delta \mathit{HPD}_\alpha(\mathbf{x})\right) \xrightarrow[n \longrightarrow \infty]{a.s.} 0$. It follows that $C_\alpha^n(\mathbf{X})$ has asymptotic conditional coverage of $1 - \alpha$ (Lei et al., 2018).*

5.4 Example: Low-dimensional Feature Space Partitioning

Our first example is the same low-dimensional stylized version of the galaxy photometric redshift (photo-z) problem presented in Section 4.4; see the text for equations and Figure 4.1 for a visualization of the data. The task is to predict the target variable Y based on predictors X_1 and X_2. We compute 90% prediction sets for Y using this hybrid local conformal calibration method detailed in Algorithm 9, and we obtain both prediction intervals and HPD sets, which we refer to as `Hybrid INT` and `Hybrid HPD` respectively. For a total sample size n, we split the data equally into three sets: training (to learn $\widehat{f}$), calibration (to learn $\widehat{r}^{\widehat{f}}$, then used to construct $\widetilde{f}$), and conformal (to compute conformity scores). We compare the results to our original `Cal-PIT (INT)` and `Cal-PIT (INT)` prediction sets, as well as to `CD-split+` prediction intervals and HPD sets; recall that for these algorithms, we only split the n data points into two equal training and calibration sets.

As before, it is impossible to exactly verify conditional validity in finite datasets in practice, but this stylized example gives us control over the true data generating process. We use Monte Carlo simulations of synthetic data to compute the true conditional coverage at a fixed set of 1000 uniformly sampled points in $\mathbf{X}$, as well as to compute "oracle" prediction sets.

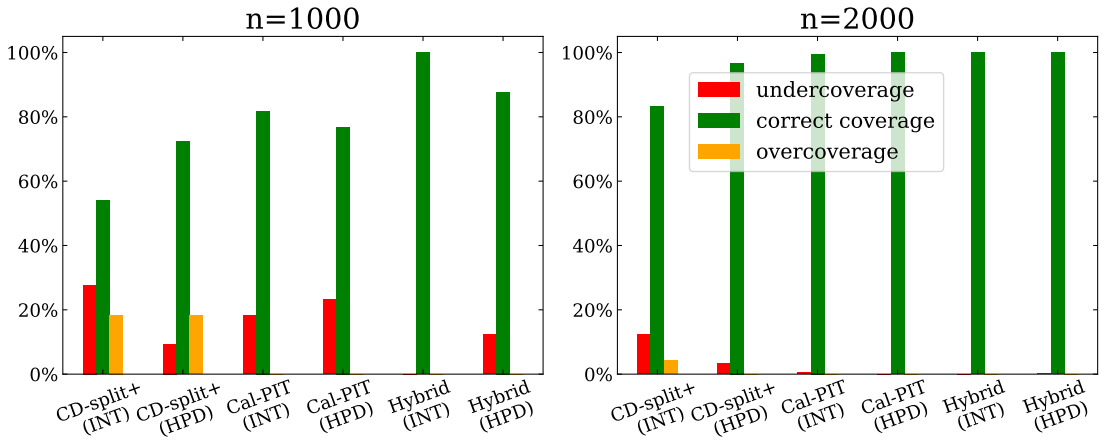


Figure 5.1: The proportion of test points with correct conditional coverage for different methods. Data of total size n are split equally three ways into train, calibration, conformal sets for `Hybrid` methods, and equally into train and calibration sets for `Cal-PIT` and `CD-split+`. For $n = 1000$, `Hybrid` methods obtain improved conditional coverage.

Figure 5.1 compares the conditional coverage of each method, evaluated using Monte Carlo simulations as described above. Test points with coverage within two standard deviations (SD) of $1 - \alpha = 0.9$ based on 100 random realizations are labeled as having "correct" coverage. For $n = 2000$, all methods approximately attain the nominal 90% coverage across the feature space. For $n = 1000$, the data size is too small for $\widetilde{f}(y|\mathbf{x})$ to always be well estimated, and therefore `Cal-PIT` methods exhibit some undercoverage. By performing local conformal corrections, `Hybrid` methods are able to attain improved conditional coverage, even though the extra data split means they have fewer points with which to estimate $\widehat{f}$ and $\widehat{r}^{\widehat{f}}$.

Why exactly are the local conformal corrections of the `Hybrid` methods useful in this example? Figure 5.2 shows which of the test points achieve undercoverage vs. correct coverage for the `Cal-PIT` methods for data size $n = 1000$. We see that the difficulty of estimating $\widetilde{f}(y|\mathbf{x})$ does not appear to be uniform across X_1; rather, the parts of the feature space with more complex bimodality tend to be more challenging to calibrate.

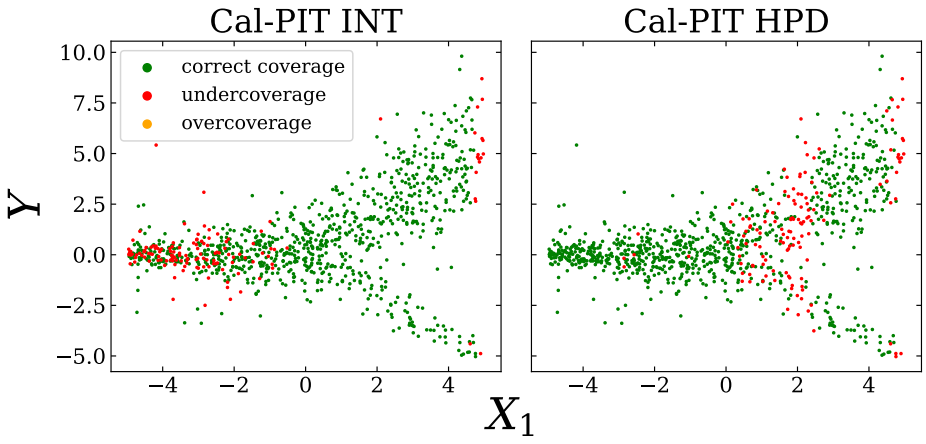


Figure 5.2: Performance of `Cal-PIT` methods on test set, for total data size of $n = 1000$. Test points are color-coded by whether they are on average correctly covered, undercovered, or overcovered. Points with incorrect coverage are not uniformly dispersed across the feature space, but tend to be clustered in the regions of X_1 with more challenging bimodality.

Now, consider the partitions learned by the `Hybrid` methods; three random instances of these clusterings are shown in Figure 5.3. While the group labels shift at random and are not meaningful, the learned clusters are mostly consistent across random iterations. This is expected since the partition reflects how the structure of the learned $\widetilde{f}(y|\mathbf{x})$ changes depending on X_1: in the leftmost X_1 region, the true $f(y|\mathbf{x})$ is two Gaussians with the same mean but different variances; in the center X_1 region, $f(y|\mathbf{x})$ is essentially a single Gaussian; in the rightmost X_1 region, $f(y|\mathbf{x})$ is a bimodal mixture of two separated Gaussians. Within each partition element, or cluster, we perform a single conformal correction so that marginal coverage is achieved across each partition element.

Comparing Figures 5.2 and 5.3, we see that the test points for which the `Cal-PIT` methods (for $n = 1000$) achieve incorrect conditional coverage tend to fall into the same cluster. Therefore, performing a single conformal correction within each cluster is able to improve conditional coverage. Intuitively, if a given partition element A

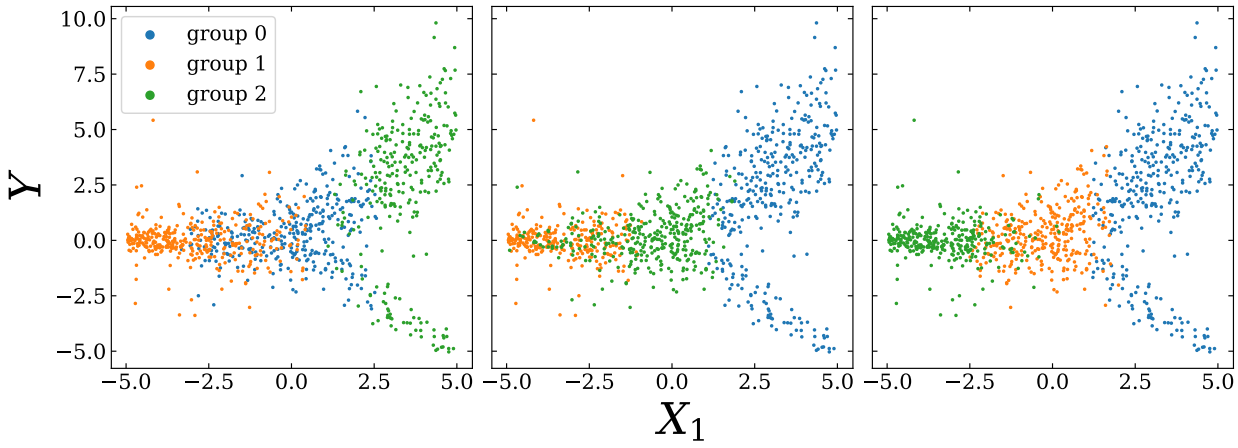


Figure 5.3: Partitions of the test set learned by the `Hybrid` methods, for $n = 1000$. Three random instances are shown; group labels are not meaningful, but the learned clusters themselves are stable across iterations. Essentially the same three clusters, reflecting the changing structure of $f(y|\mathbf{x})$ across X_1, are learned each time, and they are structured such that the points that on average have incorrect conditional coverage in the `Cal-PIT` methods, for $n = 1000$, tend to fall into the same cluster.

happens to contain most of the points with undercoverage, then the local conformal correction to A would adjust the thresholds ($L_{\mathbf{x}}$ and $U_{\mathbf{x}}$, or α') to be "wider", thus raising the average coverage of all the points in A.

In contrast, the local conformal corrections of `CD-split+` are less useful in this example, because the initial CDE $\widehat{f}(y|\mathbf{x})$ is not very good. Figure 5.4 shows that the partitions learned by `CD-split+` are less meaningful, and do not really capture either the structure of X_1. Furthermore, local conformal corrections only ensure correct coverage *on average* within each partition element. The quality of the CDE is still what primarily matters when it comes to conditional coverage, so the main advantage of `Hybrid` over `CD-split+` is that it works with the calibrated CDEs $\widetilde{f}$ which are more accurate than the original CDEs $\widehat{f}$.

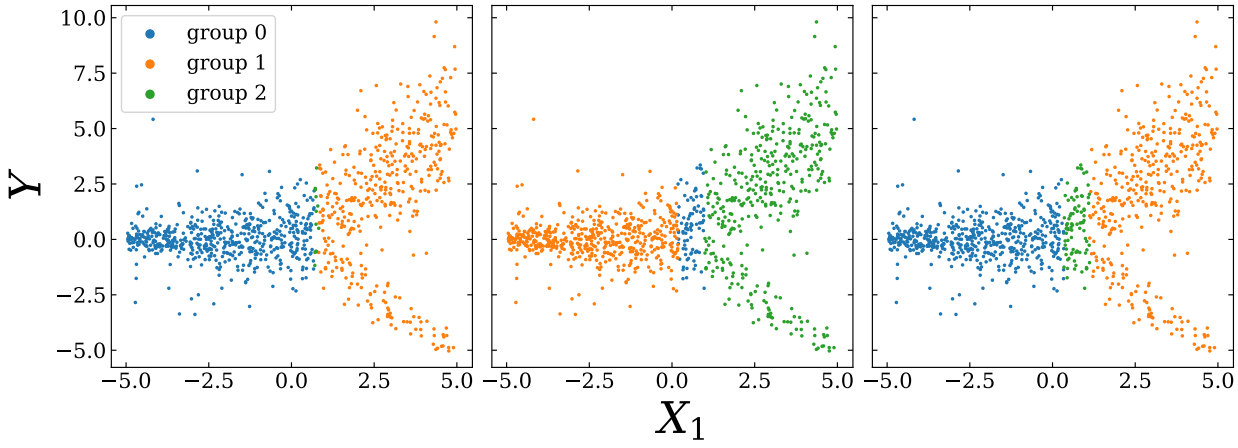


Figure 5.4: Partitions of the test set learned by the `CD-split+` methods, for $n = 1000$. Three random instances are shown. The learned clusters are not as meaningful as those learned by the `Hybrid` method, and are less reflective of the changing structure of $f(y|\mathbf{x})$ across X_1.

5.5 Application: Photometric Redshift Prediction

This section considers the photo-z estimation problem described previously in Section 4.6. Here, we use the Happy A dataset, which contains 74,944 galaxies based on the Sloan Digital Sky Survey DR12 (Beck et al., 2017). Our goal is to construct $1 - \alpha = 90\%$ prediction sets for redshifts based on corresponding photo-z estimates. We hold out 5000 galaxies for the test set, and divide the remaining $n = 69,944$ galaxies equally as required by sample splitting for each method. That is, for the `Hybrid` method developed in this chapter, we split n into three sets: training (to learn $\widehat{f}$), calibration (to learn $\widehat{r}^{\widehat{f}}$, then used to construct $\widehat{f}$), and conformal (to compute conformity scores). For `Cal-PIT`, we need only split n into two sets, because conformity scores are not needed. We also compare our methods with `CD-split+` (Izbicki et al., 2022), which is a local conformal method based on $\widehat{f}$ (rather than $\widetilde{f}$, as the `Hybrid` method is), and with `Reg-split` (Lei et al., 2018), a regression-based conformal method that incorporates local adaptivity by estimating the conditional mean absolute deviation; for these methods, we split n into two sets, for training the conformity score and calibration. For illustrative purposes, we use a Gaussian neural density model for the initial CDE $\widehat{f}$ for `CD-split+`, `Cal-PIT`, and `Hybrid`. We will see that `Cal-PIT` and `Hybrid` can recalibrate the unimodal Gaussian densities to be bimodal if necessary, leading to possibly discontiguous prediction sets, while `CD-split+` can only provide prediction intervals. We use XGBoost to fit `Reg-split`. (A similar experiment was performed on the same dataset by Izbicki et al. (2022), with a key difference being that here we are deliberately starting with a misspecified initial CDE $\widehat{f}$ in order to demonstrate the advantages of our method in this type of setting.)

Because we do not have the true data generating process for this dataset, we cannot assess conditional coverage via Monte Carlo sampling as we did for toy examples. Instead, we consider local coverage of the methods in two regions of the feature space: bright galaxies and faint galaxies. Galaxies are classified based on the r-magnitude values of their photometric estimates as either bright (low r-magnitude, in the units of this dataset) or faint (high r-magnitude). Table 5.1 compares the achieved coverage and average size of the prediction sets of different methods on bright vs. faint galaxies. `Reg-split` achieves the smallest prediction sets among the methods, but has undercoverage for faint galaxies. In this example, `CD-split+` can only provide prediction intervals when starting from an initial unimodal CDE, so it has the largest prediction sets. It also undercovers for faint galaxies. Only the `Hybrid` method achieves approximate local coverage; `Cal-PIT` has slightly larger prediction sets than `Hybrid`, which results in overcoverage.

Table 5.1: Coverage and average size of the prediction sets for various methods, along with their standard errors. Only the `Hybrid` method achieves nominal local coverage among bright and faint galaxies.

	Galaxy	Reg-split	CD-split+	Cal-PIT	Hybrid
Coverage	Bright	0.924 (0.009)	0.945 (0.007)	0.983 (0.004)	0.908 (0.009)
	Faint	0.818 (0.022)	0.799 (0.023)	0.914 (0.016)	0.871 (0.019)
Size	Bright	0.131 (0.005)	0.408 (0.008)	0.267 (0.007)	0.217 (0.007)
	Faint	0.181 (0.011)	0.456 (0.014)	0.297 (0.013)	0.255 (0.012)

Figure 5.5 shows examples of prediction sets produced by each method for sample bright and faint galaxies, along with an horizontal line indicating the true redshift in each instance. Faint galaxies have larger prediction sets across all

methods, since they often have multimodal densities (Wittman, 2009; Kügler et al., 2016; Dalmasso et al., 2020). We see that while bright galaxies have consistently low redshifts, the redshifts of faint galaxies often appear to be bimodal. However, `Reg-split` and `CD-split+` produce prediction intervals that do not reflect this bimodal structure. Only the `Cal-PIT` and `Hybrid` methods are able to recalibrate an initial unimodal CDE to yield bimodal densities $\widetilde{f}$ and therefore discontiguous prediction sets.

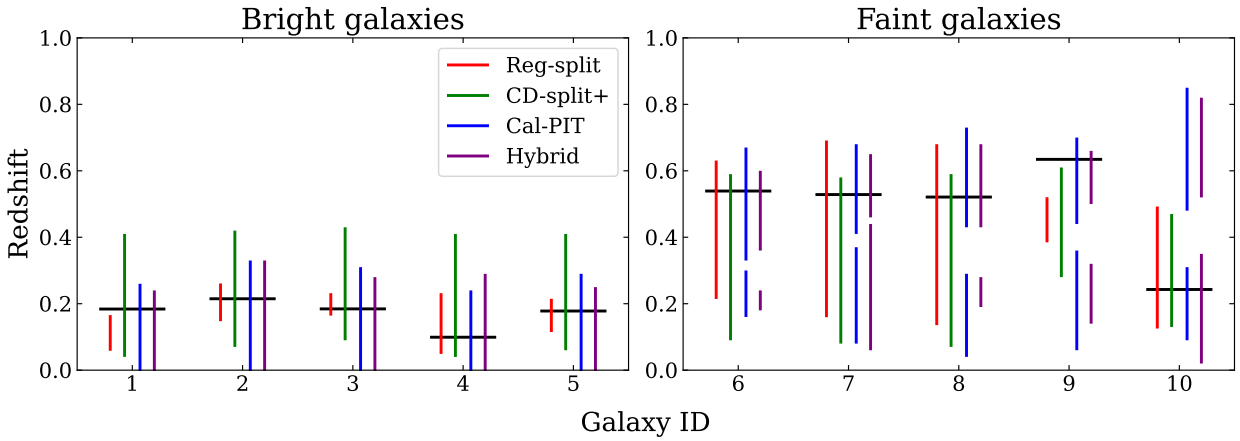


Figure 5.5: Based on Figure 11 in Izbicki et al. (2022). Prediction sets from various methods on sample bright and faint galaxies from the test set. Horizontal lines indicate the true redshift of each galaxy. For faint galaxies, the true density is often bimodal, but only `Cal-PIT` and `Hybrid` can provide prediction sets that are not single intervals.

Why does `CD-split+` in this example not achieve local coverage, while the `Hybrid` method does? Figure 5.6 compares the partitions learned by the two methods. The clusters learned by the `Hybrid` method, based on the structure of $\widetilde{f}(y|\mathbf{x})$, roughly correspond to r-magnitude, which is the basis for defining bright vs. faint galaxies. In contrast, the clusters learned by `CD-split+`, based on a poorly fitting initial CDE $\widehat{f}(y|\mathbf{x})$, do not cleanly separate bright and faint galaxies; therefore the conformal adjustment within each learned partition element does not ensure nominal coverage for the faint galaxies.

Of course, here we deliberately chose a unimodal Gaussian CDE for $\widehat{f}$ to highlight the unique ability of our `Hybrid` method to overcome a misspecified initial CDE. `CD-split+` would likely perform better on this dataset if given a better choice of $\widehat{f}$, such as a non-parametric CDE method. In real world applications, it is often difficult to be confident that an initial CDE is well-specified and decently fit. When the initial CDE fits poorly, the `Hybrid` method is able to recalibrate the CDE via the `Cal-PIT` procedure we have developed, while also guaranteeing finite-sample marginal and local coverage by virtue of its conformal inference framework.

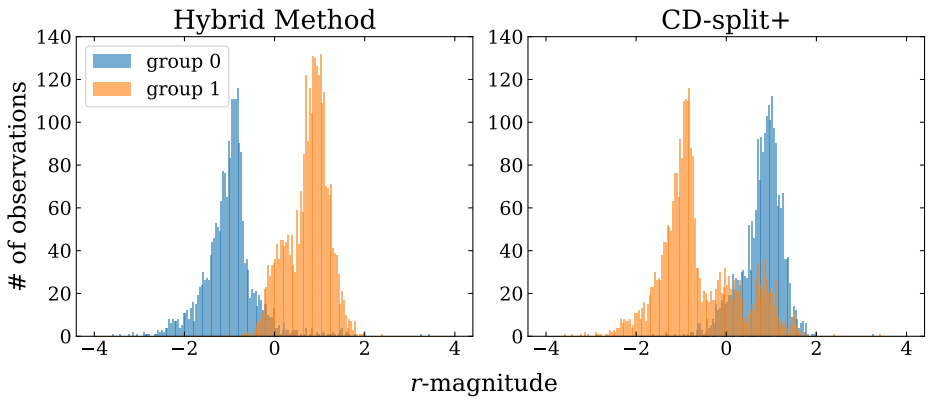


Figure 5.6: Identified clusters of test set galaxies based on the `Hybrid` and `CD-split+` methods, visualized by r-magnitude on the x-axis. Because the initial CDE $\widehat{f}(y|\mathbf{x})$ is a misspecified unimodal Gaussian, the clusters identified by `CD-split+` are not optimal and do not correspond to bright vs. faint galaxies (a classification determined by r-magnitude). However, the `Hybrid` clusters are based on the recalibrated CDE $\widetilde{f}(y|\mathbf{x})$, so they reflect a cleaner separation based on r-magnitude.

Chapter 6

Conclusion and Future Work

This chapter has three subsections. Section 6.1 summarizes the conclusions of the previous chapters and the general contributions of the work in this thesis. Section 6.2 discusses limitations to our methods, as well as potential extensions and future work that could address those limitations. Finally, Section 6.3 describes potential future applications that could incorporate the work in this thesis into future research directions.

6.1 Summary

This thesis introduced a novel framework for (i) diagnosing the quality of conditional density estimators in terms of conditional calibration (as described in Chapter 3), and (ii) leveraging these diagnostics via the `Cal-PIT` method to correct those CDEs towards better conditional coverage (as described in Chapter 4). Our work represents a small contribution to the vast field of uncertainty quantification, and we hope that others will find our methods practically useful, as well as complementary to the many other promising lines of related research that seek to accurately describe full predictive distributions $f(y|\mathbf{x})$.

As discussed in Chapter 2, there is much scientific interest in estimation beyond point predictions to capturing full predictive distributions, and the machine learning community has developed many new, complex models to meet this demand. Less attention has been devoted to assessing the conditional calibration of these models, and our work is an original examination of this problem that also addresses some known limitations of previous diagnostic methods. Recent work in the complementary field of quantile regression has also sought to promote conditional calibration of estimated conditional quantiles, but our method is unique in how it directly targets conditional coverage as a goal. Conformal inference is also a related field that provides mathematical finite-sample coverage guarantees for prediction sets; in Chapter 5, we were able to synthesize ideas from conformal inference to develop an extension of `Cal-PIT` that combines desirable properties of both methods.

To our knowledge, ours is the first and only diagnostic framework that allows one, at any location $\mathbf{x} \in \mathcal{X}$ in the feature space, to identify and describe the distributional differences between an estimated conditional density and the true distribution of y at $\mathbf{x}$. Our ALP plots visualize this information in simple, intuitive graphical summaries; from these plots, the practitioner directly sees an overall picture of the quality of conditional calibration at $\mathbf{x}$, and can easily identify patterns such as bias, dispersion, and multimodality in y as a function of $\mathbf{x}$. This rich diagnostic information powers our recalibration framework of `Cal-PIT`, which is unique in how it directly optimizes for conditional calibration and is able to produce full recalibrated predictive distributions $\widetilde{f}(y|\mathbf{x})$.

We demonstrated the effectiveness of our diagnostics and recalibration through experiments on synthetic data, where our methods compare favorably on the metric of achieved conditional coverage with state-of-the-art predictive inference algorithms. Furthermore, we explored an application to the real-world astrophysical problem of estimating galaxy distances based on imaging data (photometric redshifts), a problem for which conditional density models are difficult to estimate and conditional calibration is of practical significance. In particular, we have developed a solution to the problem examined in Schmidt et al. (2020) (see Section 2.2 for details), which seeks improved diagnostics for evaluating probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST), a large-scale synoptic astronomical survey that is expected to launch in 2024. Indeed, astronomers affiliated with the Dark Energy Science Collaboration (DESC), which is the international science collaboration that will make high accuracy measurements of fundamental cosmological parameters using data collected from the LSST, have expressed interest in incorporating our diagnostics into their evaluation pipeline, which is very gratifying and exciting for us.

6.2 Limitations and Extensions

One limitation of our method is that it requires reasonably large sample sizes in order to learn the regression function $\widehat{r}(\gamma; \mathbf{x})$, which can be a complex function of both γ and $\mathbf{x}$. This means that sometimes it may not be practically useful to apply our method, if one cannot expect to obtain a decent estimate of $\widehat{r}(\gamma; \mathbf{x})$ given the available data sample size and complexity of the task at hand. One possible avenue for mitigating this issue is to apply active learning (Settles, 2009) in order to choose the calibration sample $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ in a smarter way. The framework of active learning allows the practitioner to iteratively choose the data points on which to train a machine learning algorithm. By iteratively identifying points that are expected to bring greatest improvement to the model, active learning helps the model focus on more critical regions of the feature space $\mathcal{X}$ (e.g., regions whose distribution is more difficult to learn), thereby leading to greater overall accuracy with fewer required training instances. Currently, we assume that the points in $\mathcal{D}$ are drawn i.i.d. from $F_{\mathbf{X},Y}$. Future work could try to incorporate active learning to sample more intentionally, at locations $\mathbf{x}$ that would be most useful for improving the fit of $\widehat{r}(\gamma; \mathbf{x})$. One idea for determining such locations would be to try to evaluate where there is greatest uncertainty in the currently trained $\widehat{r}(\gamma; \mathbf{x})$ function; this would be similar in spirit to recent work that uses active learning to more efficiently train deep Gaussian processes (Sauer et al., 2022).

Another limitation of our method is that it operates under the assumption that Y is continuous, so it only applies to regression tasks and not to classification tasks. There has been substantial recent interest on how to improve the conditional calibration of classification algorithms. Recognized approaches include Platt scaling (Platt, 1999), histogram binning (Zadrozny and Elkan, 2001), temperature scaling (Guo et al., 2017), robust conditionals Wald and Globerson (2017), and Dirichlet calibration (Kull et al., 2019), among others. It would be interesting if future work could extend the unique perspective developed in this thesis to the classification setting, to obtain analogous diagnostics that directly estimate the empirical conditional coverage of classification algorithms. Unlike the other methods mentioned above, our method would implicitly assume and leverage the smoothness of conditional coverage across the feature space. It is not obvious how we would develop such a method. We might consider as a starting point whether, given a binary probabilistic classifier $\widehat{f}(y|\mathbf{x})$ for $Y \in \{0, 1\}$, we could train a regression function analogous to $\widehat{r}(\gamma; \mathbf{x})$ that assesses the empirical quality of the predicted class probabilities. One suggestion is to use a monotonic probabilistic classifier to regress the (now binary) output Y_i on the two variables $(\mathbf{X}_i, \widehat{P}(Y_i = 1|\mathbf{X}_i))$; whereas previously we had sampled multiple γ quantiles for each $\mathbf{X}_i$, here we would pair each $\mathbf{X}_i$ with a single predicted class probability $\widehat{P}(Y_i = 1|\mathbf{X}_i)$. It would be exciting if future work could augment and refine this process, and ultimately be able to successfully diagnose and recalibrate binary classification algorithms (which could then naturally extend to general classification settings).

6.3 Future Applications

We hope that the work in this thesis will be viewed as a practical "toolkit" that is complementary to and can be incorporated into other lines of research. There are natural directions for future research that investigate how our method may assist in the development or improvement of other uncertainty quantification methods. If the practitioner already has a parametric or pre-specified model class in mind, they could potentially still use our diagnostics (which are inherently non-parametric) to obtain insight into quality of fit and guidance on where optimally to sample next. More generally, we can also apply our full `Cal-PIT` recalibration scheme to complex, non-parametric modeling scenarios with intricate dependency structures beyond what we have examined in this thesis. We elaborate on an example of each of these future directions below.

An example of a modeling scenario where our non-parametric diagnostic toolkit could give guidance on improving the fit, without seeking to replace the original model, is Gaussian process regression, which is commonly used in the physical sciences. The practitioner may have good reason to stay within the Gaussian process model class, which is flexible enough to capture model nonlinearities while being simple enough to yield closed-form variance expressions. Our diagnostic ALP plots may be able to guide model fitting by identifying, say, regions or potential underfitting or overfitting, both of which are common issues with Gaussian process modeling. This could potentially form the basis of an active learning scheme where the practitioner would sample future points in a more informed way; that is, by choosing the points where the current fit is most in need of improvement, based on the metric of empirical conditional

coverage. In contrast, standard active learning for Gaussian processes is based on choosing the new sample point that would decrease the overall variance or uncertainty the most, whereas our approach would be trying to optimally improve the conditional calibration of the model.

A key consideration here is that our diagnostic method applies to stochastic simulators with aleatoric uncertainty, but typically Gaussian processes that perform emulation only have epistemic uncertainty, as the target of inference is a non-stochastic function f from which one obtains non-stochastic, noise-free evaluations $y_i = f(x_i)$. If the Gaussian process is an epistemic model and the uncertainties are epistemic uncertainties, it is unclear whether it is conceptually meaningful to talk about "validation" of those uncertainties. However, there are indeed situations where f truly is a random function, so that observations $y_i = f(x_i)$ for fixed points x_i are now random. Because the target of inference for new points x^* is now the random quantity $y^* = f(x^*)$, this is now a prediction problem for which the Gaussian process formalism yields prediction intervals for y^*.

Recent work in climate science (Kuusela and Stein, 2018) leverages just this Gaussian process formalism to model seawater temperature from Argo floats (Roemmich et al., 2015) in the upper 2000 meters of the global ocean. The authors use locally stationary Gaussian process regression to obtain moving-window estimates of covariance parameters that adapt to the large, non-stationary spatiotemporal dataset at hand. Because temperature and salinity variation represent true stochastic variation, it is a well-defined problem to construct prediction sets for these quantities and assess their frequentist coverage. That is, if the underlying stochastic process truly follows a locally Gaussian process with known parameters, then the obtained prediction intervals will have correct unconditional (marginal) coverage, where "coverage" is typically understood in the sense of getting new realizations of the (entire) random function. Figure 6.1, taken from Figure 5 of Kuusela and Stein (2018), assesses the marginal coverage for temperature prediction intervals achieved by different methods. The "student nugget" uncertainty quantification methods developed by the authors have less deviation between sample and theoretical quantiles of the predictive distribution, and thus obtain substantially better empirical coverage than prior (reference) procedure.

This prediction setting is very complicated due to the high-noise ocean environment, complex spatio-temporal dynamics, and in situ nature of the Argo floats (which drift randomly in the ocean, rather than being on a fixed grid). Recent work by Park et al. (2021) strives to improve the quality of prediction intervals via debiasing the Gaussian process. It would be interesting to explore whether our diagnostic framework could say something about conditional coverage in this type of setting, and potentially improve the prediction intervals so that they are not only better marginally calibrated but also better conditionally calibrated with respect to the state vector $\mathbf{x}$ (or certain key variables within that vector). Again, we see our diagnostic framework not as a non-parametric competitor or replacement for pre-specified models like Gaussian processes, which may have well-known inherent advantages, but rather as a complementary tool for identifying potential issues and assisting in the modeling process.

The other direction for future work we have identified is to apply `Cal-PIT` to more complex non-parametric modeling situations, because we have the flexibility to choose the regression function for $\widehat{r}(\gamma; \mathbf{x})$ to adapt to various complex forms that may be needed for the task at hand. Here, we discuss one specific example of a promising extension

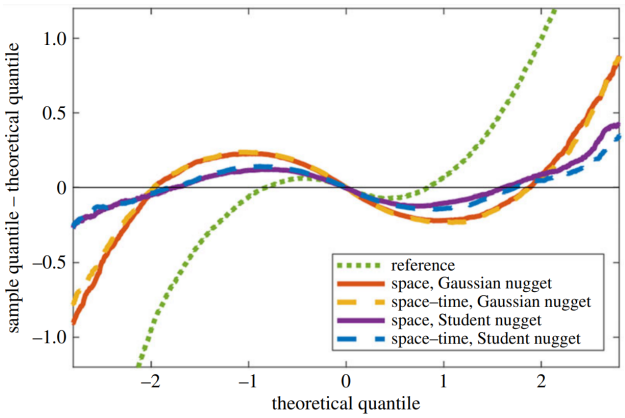


Figure 6.1: Taken from Figure 5 of Kuusela and Stein (2018). The y-axis plots the difference between the cross-validated sample quantile and the corresponding standard Gaussian theoretical quantile for temperature prediction intervals at 300 dbar. The closer the curves are to a horizontal straight line at 0, the better the marginal calibration of the predictive distributions. The models with a Student nugget achieve better marginal coverage than the Roemmich–Gilson-like reference model.

that comes from the following observation: `Cal-PIT` can potentially be extended to multivariate output vectors $\mathbf{Y}$ by the decomposition $f(\mathbf{y}|\mathbf{x}) = \prod_i f(y_i|\mathbf{x}, \mathbf{y}_{<i})$; thus, we could perform `Cal-PIT` corrections on autoregressive components of the conditional distribution, allowing us to validate the conditional calibration of $f(\mathbf{y}|\mathbf{x})$ even when it has this highly complex dependence structure. This is a particularly promising direction for deep Pixel-CNN and Pixel-RNN models (Van den Oord et al., 2016; Van Den Oord et al., 2016).

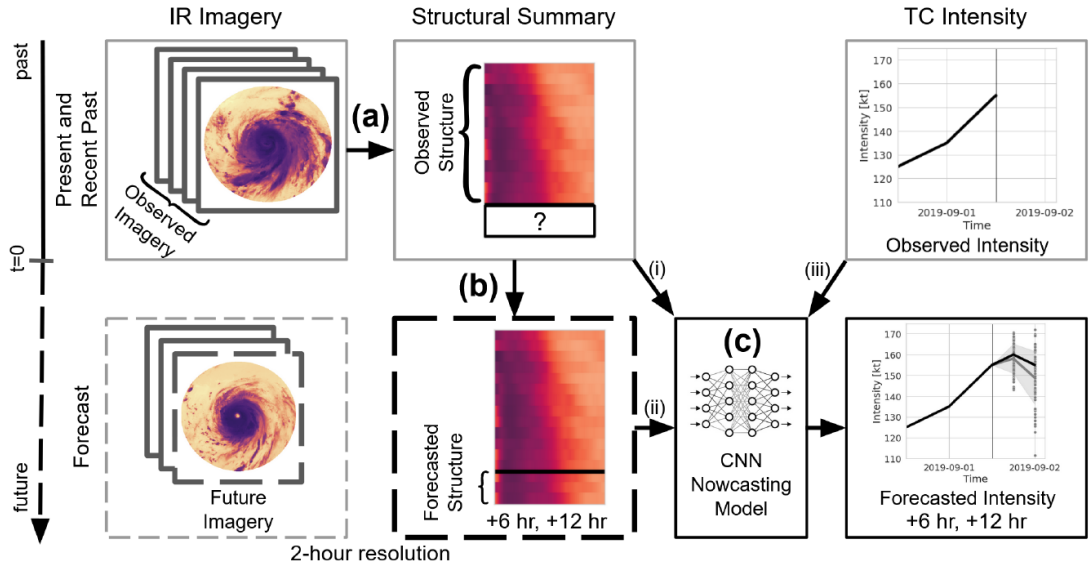


Figure 6.2: Taken from Figure 2 of McNeely et al. (2023). (a) As described in Figure 4.10, radial profiles quantify the evolution of spatio-temporal convective structure. (b) The authors generate structural forecasts by projecting the radial profiles into the future via a PixelSNAIL model. (c) A CNN nowcasting model generates forecasted intensities at +6 to +12 hours from three sources of inputs: (i) observed structure, (ii) forecasted structure, and (iii) observed storm intensity.

Such models have recently been leveraged for a much more realistic and sophisticated version of the synthetic time series trajectories for tropical cyclones that we simulated in Section 4.7. Using a deep autoregressive Generative Model (PixelSNAIL), McNeely et al. (2023) develop a novel structural forecasting method, where not only TC intensity but also TC convective structure is propagated 6 to 12 hours forward in time based on past observed infrared imagery ($\{\mathbf{S}_{<t}\}$) and operational intensities ($\{Y_{<t}\}$). See Figure 6.2 for an illustration of this. The PixelSNAIL model achieves this by stochastically simulating possible trajectories of radial profiles, as captured by Hovmöller diagrams (refer to the right panel of Fig. 4.10), forward in time by forecasting one pixel at a time. The authors then use a nowcasting model to predict TC intensity as a function of observed intensities and the observed and forecasted structural trajectories. This approach is successful and performs well alongside state-of-the-art TC forecasting techniques, but it relies on a large spatiotemporal model with hundreds of thousands of parameters, trained on a limited dataset of real hurricane trajectories. Therefore, it may be desirable to validate how well calibrated the model is, i.e. to validate the PixelSNAIL model used for structural forecasts, which is really the crux of the method.

Since Pixel-RNNs have complex autoregressive dependency structure (including long short-term memory connections to past observations), it is not immediately obvious how one would diagnose and correct potential issues with the structural forecasts. It would be interesting to explore whether we could apply our `Cal-PIT` framework to this complex non-parametric task. Figure 6.3, taken from Figure 3 of McNeely et al. (2023), shows how the PixelSNAIL model essentially models a new CDE for each pixel, one pixel at a time while respecting the autoregressive structure. By using the decomposition stated above, $f(\mathbf{y}|\mathbf{x}) = \prod_i f(y_i|\mathbf{x}, \mathbf{y}_{<i})$, we could also potentially fit `Cal-PIT` by learning the $\widehat{r}(\gamma; \mathbf{x})$ regression function one pixel at a time. Specifically, we could fit the regression function using a deep Pixel-CNN architecture that respects the same autoregressive structure. This would yield diagnostic information about the quality of conditional CDEs located at each pixel that is forecasted forward in time.

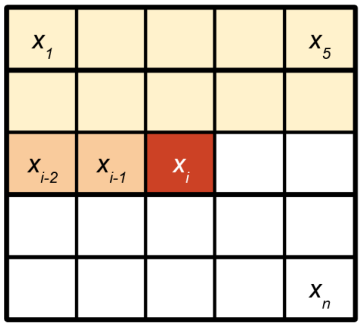


Figure 6.3: Taken from Figure 3 of McNeely et al. (2023). Masking in Pixel Autoregression. Illustration of raster-scan ordering and the causal masking. Convolutions at index i only have access to pixel values in previous rows (earlier time points, color coded by yellow), and pixel values in the same row but to the left of pixel x_i (same time point, color coded by orange).

We have described just two potential applications to frequentist uncertainty quantification in the physical sciences where the work developed in this thesis may yield synergies, by helping to assess and possibly correct issues in

complex models of full predictive distributions. Encouraged by the success we saw in applying our method to photo-z estimation, as well as its success on illustrative synthetic examples, we are eager to publicize our ideas and optimistic that in some small way they may prove useful for future lines of research in uncertainty quantification.

Bibliography

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76(C):243–297. 5

Alkema, L., Raftery, A. E., and Clark, S. J. (2007). Probabilistic projections of HIV prevalence using Bayesian melding. *The Annals of Applied Statistics*, 1(1):229–248. 6

Almosallam, I. A., Jarvis, M. J., and Roberts, S. J. (2016). GPZ: non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 462(1):726–739. 49, 50

Ambrogioni, L., Güçlü, U., van Gerven, M. A. J., and Maris, E. (2017). The Kernel Mixture Network: A Nonparametric Method for Conditional Density Estimation of Continuous Random Variables. *arXiv e-prints*, page arXiv:1705.07111. 6

Amerise, I. L. (2018). Quantile regression estimation using non-crossing constraints. *Journal of Mathematics and Statistics*, 14(1):107–118. 7

Andrews, D. W. K. (1997). A conditional Kolmogorov test. *Econometrica*, 65(5):1097–1128. 8

Arnouts, S., Cristiani, S., Moscardini, L., Matarrese, S., Lucchin, F., Fontana, A., and Giallongo, E. (1999). Measuring and modelling the redshift evolution of clustering: the Hubble Deep Field North. *Monthly Notices of the Royal Astronomical Society*, 310(2):540–556. 50

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2020). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*. 14, 60, 61

Barnes, E. A., Barnes, R. J., and Gordillo, N. (2021). Adding uncertainty to neural network regression tasks in the geosciences. *arXiv e-prints*, page arXiv:2109.07250. 47

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., and Csabai, I. (2016). Photometric redshifts for the SDSS Data Release 12. *Monthly Notices of the Royal Astronomical Society*, 460(2):1371–1381. 49

Beck, R., Lin, C. A., Ishida, E. E. O., Gieseke, F., de Souza, R. S., Costa-Duarte, M. V., Hattab, M. W., and Krone-Martins, A. (2017). On the realistic validation of photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 468(4):4323–4339. 69

Benítez, N. (2000). Bayesian Photometric Redshift Estimation. *The Astrophysical Journal*, 536(2):571–583. 49, 50

Berger, J. O. and Smith, L. A. (2019). On the statistical formalism of uncertainty quantification. *Annual Review of Statistics and Its Application*, 6(1):433–460. 5

Bierens, H. J. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association*, 78(383):699–707. 41

Bishop, C. M. (1994). Mixture density networks. 6

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1613–1622. PMLR. 7

Bordoloi, R., Lilly, S. J., and Amara, A. (2010). Photo-z performance for precision cosmology. *Monthly Notices of the Royal Astronomical Society*, 406(2):881–895. 9, 19, 49

Brammer, G. B., van Dokkum, P. G., and Coppi, P. (2008). EAZY: A Fast, Public Photometric Redshift Code. *The Astrophysical Journal*, 686(2):1503–1513. 50

Carrasco Kind, M. and Brunner, R. J. (2013). TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Monthly Notices of the Royal Astronomical Society*, 432(2):1483–1501. 50

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust validation: Confident predictions even when distributions shift. *arXiv*, pages 1–35. 48

Cavuoti, S., Amaro, V., Brescia, M., Vellucci, C., Tortora, C., and Longo, G. (2017). METAPHOR: a machine-learning-based method for the probability density estimation of photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 465(2):1959–1973. 50

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM. 44

Chen, T. Y., Dey, B., Ghosh, A., Kagan, M., Nord, B., and Ramachandra, N. (2022). Interpretable Uncertainty Quantification in AI for HEP. *Proceedings of the US Community Study on the Future of Particle Physics (Snowmass 2021)*, page arXiv:2208.03284. 5

Chen, Y. and Gutmann, M. U. (2019). Adaptive Gaussian copula ABC. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1584–1592. PMLR. 7

Chernozhukov, V., Fernandez-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125. 13

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference on Learning Theory*, Proceedings of Machine Learning Research, pages 732–749. PMLR. 44

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. (2021). Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 10971–10984. 7, 12

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692. 9, 18, 19

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062. 7

Dalmasso, N., Pospisil, T., Lee, A., Izbicki, R., Freeman, P., and Malz, A. (2020). Conditional density estimation tools in python and r with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, 30:100362. 33

Dalmasso, N., Pospisil, T., Lee, A. B., Izbicki, R., Freeman, P. E., and Malz, A. I. (2020). Conditional density estimation tools in python and R with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, 30:100362. 7, 8, 19, 27, 44, 49, 70

de Bragança Pereira, C. A. and Stern, J. M. (1999). Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy*, 1(4):99–110. 33

Dey, B., Andrews, B. H., Newman, J. A., Mao, Y.-Y., Rau, M. M., and Zhou, R. (2021). Photometric Redshifts from SDSS Images with an Interpretable Deep Capsule Network. *arXiv e-prints*, page arXiv:2112.03939. 49

D'Isanto, A. and Polsterer, K. L. (2018). Photometric redshift estimation via deep learning. generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*, 609:A111. 9, 18, 31, 54

Dutordoir, V., Salimbeni, H., Hensman, J., and Deisenroth, M. P. (2018). Gaussian process conditional density estimation. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2391–2401. 7

Farmer, C. (2017). Uncertainty quantification and optimal decisions. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2200):20170115. 5

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., and Goude, Y. (2021). Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 116(535):1402–1412. 7

Feldman, S., Bates, S., and Romano, Y. (2021). Improving Conditional Coverage via Orthogonal Quantile Regression. In *Advances in Neural Information Processing Systems*, volume 34, pages 2060–2071. Curran Associates, Inc. 12, 13, 43, 44, 45

Freeman, P., Izbicki, R., and Lee, A. B. (2017). A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. *Monthly Notices of the Royal Astronomical Society*, 468(4):4556–4565. 9, 18

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1050–1059. PMLR. 6

Gan, F. F. and Koehler, K. J. (1990). Goodness-of-Fit Tests Based on P-P Probability Plots. *Technometrics*, 32(3):289–303. 9, 18

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. (2019). Probabilistic forecasting with spline quantile function RNNs. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the 22nd Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1901–1910. PMLR. 13

Genest, C. and Rivest, L.-P. (2001). On the multivariate probability integral transformation. *Statistics & Probability Letters*, 53(4):391–399. 32

Gibbs, I. and Candès, E. (2021). Adaptive Conformal Inference Under Distribution Shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc. 48

Girard, S., Guillou, A., and Stupfler, G. (2014). Uniform strong consistency of a frontier estimator using kernel regression on high order moments. *ESAIM: Probability and Statistics*, 18:642–666. 41

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In Gordon, G., Dunson, D., and Dudik, M., editors, *Proceedings of the 14th Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323. PMLR. 31, 55

Gneiting, T. (2008). Probabilistic forecasting. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 319–321. 6

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151. 5

Goan, E. and Fookes, C. (2020). Bayesian neural networks: An introduction and survey. *CoRR*, abs/2006.12024. 6

Göttlich, S. and Knapp, S. (2020). Uncertainty quantification with risk measures in production planning. *Journal of Mathematics in Industry*, 10(1):1–21. 5

Graff, P., Feroz, F., Hobson, M. P., and Lasenby, A. (2014). SKYNET: an efficient and robust neural network training tool for machine learning in astronomy. *Monthly Notices of the Royal Astronomical Society*, 441(2):1741–1759. 50

Graham, M. L., Connolly, A. J., Ivezić, Ž., Schmidt, S. J., Jones, R. L., Jurić, M., Daniel, S. F., and Yoachim, P. (2018). Photometric Redshifts with the LSST: Evaluating Survey Observing Strategies. *The Astronomical Journal*, 155(1):1. 50

Graves, A. (2011). Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems 24*, Neural Information Processing Systems. Curran Associates, Inc. 7

Greenberg, D., Nonnenmacher, M., and Macke, J. (2019). Automatic posterior transformation for likelihood-free inference. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2404–2414, Long Beach, California, USA. PMLR. 7

Guan, L. (2019). Conformal prediction with localization. *arXiv e-prints*, page arXiv:1908.08558. 60

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR. 9, 75

Györfi, L., Kohler, M., Krzyzak, A., Walk, H., et al. (2002). *A distribution-free theory of nonparametric regression*, volume 1. Springer. 41

Hao, L. and Naiman, D. Q. (2007). *Quantile regression*. Number 149. Sage. 12

Hardle, W., Luckhaus, S., et al. (1984). Uniform consistency of a class of regression function estimators. *The Annals of Statistics*, 12(2):612–623. 41

Harrison, D., Sutton, D., Carvalho, P., and Hobson, M. (2015). Validation of Bayesian posterior distributions using a multidimensional Kolmogorov-Smirnov test. *Monthly Notices of the Royal Astronomical Society*, 451(3):2610–2624. 33, 98

Hovmöller, E. (1949). The trough-and-ridge diagram. *Tellus*, 1(2):62–66. 52

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126. 33

Izbicki, R. and Lee, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4):1297–1316. 7, 38, 40, 63

Izbicki, R., Lee, A. B., et al. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831. xviii, 7, 19, 44, 49, 50, 51

Izbicki, R., Lee, A. B., and Pospisil, T. (2019). ABC–CDE: Toward approximate Bayesian computation with complex high-dimensional data and limited simulations. *Journal of Computational and Graphical Statistics*, pages 1–20. 7

Izbicki, R., Shimizu, G., and Stern, R. B. (2020). Distribution-free conditional predictive bands using density estimators. In *International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3068–3077. PMLR. 15

Izbicki, R., Shimizu, G., and Stern, R. B. (2022). CD-split and HPD-split: Efficient Conformal Regions in High Dimensions. *Journal of Machine Learning Research*, 23(87):1–32. xix, 59, 60, 61, 62, 69, 70, 101

Janowiak, J., Joyce, B., and Xie, P. (2020). NCEP/CPC L3 half hourly 4km global (60S - 60N) merged IR v1. 51

Jitkrittum, W., Kanagawa, H., and Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In Adams, R. P. and Gogate, V., editors, *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 221–230. AUAI Press. 8, 10

Jones, M. C. and Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, 96(4):761–780. 47

Kim, I., Lee, A. B., and Lei, J. (2019). Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305. 24

Kim, T., Fakoor, R., Mueller, J., Smola, A. J., and Tibshirani, R. J. (2021). Deep quantile aggregation. *arXiv e-prints*, page arXiv:2103.00083. 13

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv e-prints*, page arXiv:1412.6980. 31, 55

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv e-prints*, page arXiv:1609.05807. 7

Kobyzev, I., Prince, S. J. D., and Brubaker, M. A. (2021). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979. 7

Koenker, R. and Bassett Jr., G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50. 11, 12, 44

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156. 11

Kügler, S., Gianniotis, N., , and Polsterer, K. L. (2016). A spectral model for multimodal redshift estimation. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE. 70

Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2801–2809. PMLR. 9

Kull, M., Perelló-Nieto, M., Kängsepp, M., de Menezes e Silva Filho, T., Song, H., and Flach, P. A. (2019). Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *CoRR*, abs/1910.12656. 75

Kuusela, M. and Stein, M. L. (2018). Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proceedings of the Royal Society A*, 474(2220). xix, 76, 77

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 31, pages 6402–6413. Curran Associates, Inc. 6

Landsea, C. W. and Franklin, J. L. (2013). Atlantic hurricane database uncertainty and presentation of a new database format. *Monthly Weather Review*, 141(10):3576–3592. 52

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111. 14, 15, 42, 44, 65, 66, 69

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(1):71–96. 14, 60, 61

Leistedt, B. and Hogg, D. W. (2017). Data-driven, Interpretable Photometric Redshifts Trained on Heterogeneous and Unrepresentative Data. *The Astrophysical Journal*, 838(1):5. 50

LeRoy, B. and Zhao, D. (2021). MD-split+: Practical Local Conformal Inference in High Dimensions. *arXiv e-prints*, page arXiv:2107.03280. 62

Liero, H. (1989). Strong uniform consistency of nonparametric regression function estimates. *Probability theory and related fields*, 82(4):587–614. 41

Liu, Y. and Wu, Y. (2009). Stepwise multiple quantile regression estimation using non-crossing constraints. *Statistics and its Interface*, 2(3):299–310. 13

Liu, Y. and Wu, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of Nonparametric Statistics*, 23(2):415–437. PMID: 22190842. 7

Louizos, C. and Welling, M. (2017). Multiplicative normalizing flows for variational Bayesian neural networks. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 2218–2227. PMLR. 7

Lueckmann, J.-M., Gonçalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. (2017). Flexible Statistical Inference for Mechanistic Models of Neural Dynamics. In *Advances in Neural Information Processing Systems*, volume 29, pages 1289–1299. Curran Associates, Inc. 7

Luo, R., Bhatnagar, A., Wang, H., Xiong, C., Savarese, S., Bai, Y., Zhao, S., and Ermon, S. (2021). Localized calibration: Metrics and recalibration. *CoRR*, abs/2102.10809. 10

MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472. 6

Malz, A. I. and Hogg, D. W. (2022). How to Obtain the Redshift Distribution from Probabilistic Redshift Estimates. *The Astrophysical Journal*, 928(2):127. 6, 49

Mandelbaum, R., Seljak, U., Hirata, C. M., Bardelli, S., Bolzonella, M., Bongiorno, A., Carollo, M., Contini, T., Cunha, C. E., Garilli, B., Iovino, A., Kampczyk, P., Kneib, J. P., Knobel, C., Koo, D. C., Lamareille, F., Le Fèvre, O., Le Borgne, J. F., Lilly, S. J., Maier, C., Mainieri, V., Mignoli, M., Newman, J. A., Oesch, P. A., Perez-Montero, E., Ricciardelli, E., Scodeggio, M., Silverman, J., and Tasca, L. (2008). Precision photometric redshift calibration for galaxy-galaxy weak lensing. *Monthly Notices of the Royal Astronomical Society*, 386(2):781–806. 6, 49

Marin, J.-M., Raynal, L., Pudlo, P., Ribatet, M., and Robert, C. (2016). ABC random forests for Bayesian parameter inference. *Bioinformatics (Oxford, England)*, 35. 7

McNeely, T., Khokhlov, P., Dalmasso, N., Wood, K. M., and Lee, A. B. (2023). Structural Forecasting for Short-term Tropical Cyclone Intensity Guidance. *Weather and Forecasting*. xix, xx, 77, 78

McNeely, T., Lee, A. B., Wood, K. M., and Hammerling, D. (2020). Unlocking GOES: A statistical framework for quantifying the evolution of convective structure in tropical cyclones. *Journal of Applied Meteorology and Climatology*, 59(10):1671–1689. 52

McNeely, T., Vincent, G., Lee, A. B., Izbicki, R., and Wood, K. M. (2022). Detecting Distributional Differences in Labeled Sequence Data with Application to Tropical Cyclone Satellite Imagery. *arXiv e-prints*, page arXiv:2202.02253. xviii, 52

Meinshausen, N. (2006). Quantile regression forests. *The Journal of Machine Learning Research*, 7:983–999. 12

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *CoRR*, abs/1411.1784. 7

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027 – 1048. 8

Mucesh, S., Hartley, W., Palmese, A., Lahav, O., Whiteway, L., Amon, A., Bechtol, K., Bernstein, G., Rosell, A. C., Kind, M. C., et al. (2020). A machine learning approach to galaxy properties: Joint redshift-stellar mass probability distributions with random forest. *arXiv e-prints*, page arXiv:2012.05928. 32

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 9

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media. 6

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer. 14

Papamakarios, G. and Murray, I. (2016). Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc. 7

Papamakarios, G., Nalisnick, E., Jimenez Rezende, D., Mohamed, S., and Lakshminarayanan, B. (2019). Normalizing Flows for Probabilistic Modeling and Inference. *arXiv e-prints*, page arXiv:1912.02762. 7, 8, 31

Park, B., Kuusela, M., Giglio, D., and Gray, A. (2021). Spatio-Temporal Local Interpolation of Global Ocean Heat Transport Using Argo Floats: A Debiased Latent Gaussian Process Approach. *arXiv e-prints*, page arXiv:2105.09707. 76

Park, Y., Maddix, D., Aubet, F. X., Kan, K., Gasthaus, J., and Wang, Y. (2022). Learning quantile functions without quantile crossing for distribution-free time series forecasting. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the 25th Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8127–8150. PMLR. 13

Pawlowski, N., Brock, A., Lee, M. C., Rajchl, M., and Glocker, B. (2017). Implicit weight uncertainty in neural networks. *arXiv e-prints*, page arXiv:1711.01297. 7

Pearce, T., Leibfried, F., and Brintrup, A. (2020). Uncertainty in neural networks: Approximately Bayesian ensembling. In *International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 234–244. PMLR. 6

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74. 75

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 125 of *Proceedings of Machine Learning Research*. PMLR. 48

Rodrigues, F. and Pereira, F. C. (2020). Beyond expectation: deep joint mean and quantile regression for spatiotemporal problems. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5377–5389. 13

Roemmich, D., Church, J., Gilson, J., Monselesan, D., Sutton, P., and Wijffels, S. (2015). Unabated planetary warming and its ocean structure since 2006. *Nature Climate Change*, 5(3):240–245. 76

Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 3543–3553. Curran Associates, Inc. 12, 44

Rothfuss, J., Ferreira, F., Walther, S., and Ulrich, M. (2019). Conditional density estimation with neural networks: Best practices and benchmarks. 8

Rowe, B., Jarvis, M., Mandelbaum, R., Bernstein, G. M., Bosch, J., Simet, M., Meyers, J. E., Kacprzak, T., Nakajima, R., Zuntz, J., et al. (2015). GALSIM: The modular galaxy image simulation toolkit. *Astronomy and Computing*, 10:121–150. 30

Sadeh, I., Abdalla, F. B., and Lahav, O. (2016). ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning. *Publications of the Astronomical Society of the Pacific*, 128(968):104502. 50

Sanabia, E. R., Barrett, B. S., and Fine, C. M. (2014). Relationships between tropical cyclone intensity and eyewall structure as determined by radial profiles of inner-core infrared brightness temperature. *Monthly Weather Review*, 142(12):4581–4599. 52

Sauer, A., Gramacy, R. B., and Higdon, D. (2022). Active learning for deep Gaussian process surrogates. *Technometrics*, 64:1–15. 74

Schmidt, S. J., Malz, A. I., Soo, J. Y. H., Almosallam, I. A., Brescia, M., Cavuoti, S., Cohen-Tanugi, J., Connolly, A. J., DeRose, J., Freeman, P. E., Graham, M. L., Iyer, K. G., Jarvis, M. J., Kalmbach, J. B., Kovacs, E., Lee,

A. B., Longo, G., Morrison, C. B., Newman, J. A., Nourbakhsh, E., Nuss, E., Pospisil, T., Tranin, H., Wechsler, R. H., Zhou, R., Izbicki, R., and LSST Dark Energy Science Collaboration (2020). Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). *Monthly Notices of the Royal Astronomical Society*, 499(2):1587–1606. xiii, xv, 9, 10, 18, 30, 48, 49, 50

Schmidt, S. J., Malz, A. I., Soo, J. Y. H., Almosallam, I. A., Brescia, M., Cavuoti, S., Cohen-Tanugi, J., et al. (2020). Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). *Monthly Notices of the Royal Astronomical Society*, 499(2):1587–1606. 49, 74

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 74

Shalizi, C. (2013). *Advanced data analysis from an elementary point of view*. 27

Shiga, M., Tangkaratt, V., and Sugiyama, M. (2015). Direct conditional probability density estimation with sparse feature selection. *Machine Learning*, 100(2):161–182. 19

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc. 7

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958. 6

Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1). 13

Stute, W. and Zhu, L. X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics*, 29(3):535–545. 8

Tagasovska, N. and Lopez-Paz, D. (2019). Single-model uncertainties for deep learning. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 6414–6425. 7, 13

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. *The Journal of Machine Learning Research*, 7:1231–1264. 12

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv e-prints*, page arXiv:1804.06788. 9, 18, 32

Tanaka, M., Coupon, J., Hsieh, B.-C., Mineo, S., Nishizawa, A. J., Speagle, J., Furusawa, H., Miyazaki, S., and Murayama, H. (2018). Photometric redshifts for hyper suprime-cam subaru strategic program data release 1. *Publications of the Astronomical Society of Japan*, 70(SP1). 19

Taylor, J. W. and Bunn, D. W. (1999). A quantile regression approach to generating prediction intervals. *Management Science*, 45(2):225–237. 12

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. 48, 61

Timmermann, A. (2000). Density forecasting in economics and finance. *Journal of Forecasting*, 19(4):231. 6

Uria, B., Murray, I., and Larochelle, H. (2014). A deep and tractable density estimator. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, Beijing, China. JMLR. 7

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. (2016). Conditional image generation with PixelCNN decoders. *Advances in Neural Information Processing Systems*, 29. 77

Van Den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1747–1756. PMLR. 77

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490. PMLR. 14, 60

Vovk, V., Gammerman, A., and Saunders, C. (1999). Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*, pages 444–453. 13

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer Science & Business Media. 14, 60

Wald, Y. and Globerson, A. (2017). Robust Conditional Probabilities. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6359–6368. 75

Wehenkel, A. and Louppe, G. (2019). Unconstrained monotonic neural networks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1543–1553. 36, 44, 54

Wittman, D. (2009). What lies beneath: Using p (z) to reduce systematic photometric redshift errors. *The Astrophysical Journal Letters*, 700(2):L174. 70

Yao, J., Pan, W., Ghosh, S., and Doshi-Velez, F. (2019). Quality of uncertainty quantification for Bayesian neural network inference. *arXiv e-prints*, page arXiv:1906.09686. 8

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning, ICML 2001*, Proceedings of Machine Learning Research, pages 609–616. PMLR. 75

Zhao, D., Dalmasso, N., Izbicki, R., and Lee, A. B. (2021). Diagnostics for conditional density models and Bayesian inference algorithms. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161 of *Proceedings of Machine Learning Research*, pages 1830–1840. PMLR. 17

Zhao, S., Ma, T., and Ermon, S. (2020). Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pages 11387–11397. PMLR. 7, 44

Zheng, J. X. (2000). A consistent test of conditional parametric distributions. *Econometric Theory*, 16(5):667 – 691. 8

Zhou, K. Q. and Portnoy, S. L. (1996). Direct use of regression quantiles to construct confidence sets in linear models. *The Annals of Statistics*, 24(1):287–306. 12

Zhou, R., Newman, J. A., Mao, Y.-Y., Meisner, A., Moustakas, J., Myers, A. D., Prakash, A., Zentner, A. R., Brooks, D., Duan, Y., Landriau, M., Levi, M. E., Prada, F., and Tarle, G. (2021). The clustering of DESI-like luminous red galaxies using photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 501(3):3309–3331. 49

Ziegel, J. F., Gneiting, T., et al. (2014). Copula calibration. *Electronic Journal of Statistics*, 8(2):2619–2638. 32

Appendix

Appendix A

Proofs

In this section, we show proofs of the results stated in this thesis.

A.1 Proof of Theorem 1

Proof. Let $z = g(\mathbf{x})$ and $Z = g(\mathbf{X})$. Notice that Equation 3.3 implies $\widehat{F}(Y|\mathbf{x}) = F(Y|g(\mathbf{x})) = F(Y|z)$. Thus,

$$\widehat{F}(Y|\mathbf{X}) = F(Y|g(\mathbf{X})) = F(Y|Z) \tag{A.1}$$

If $(\mathbf{X}, Y) \sim F_{\mathbf{X},Y}$, then, for every $0 \leq a \leq 1$,

$$\begin{aligned}
\mathbb{P}(\text{PIT}(Y, \mathbf{X}) \leq a) &= \mathbb{P}\left(\widehat{F}(Y|\mathbf{X}) \leq a\right) \\
&= \int_{\mathcal{Z}} \mathbb{P}\left(\widehat{F}(Y|\mathbf{X}) \leq a | Z = z\right) f(z)dz \\
&= \int_{\mathcal{Z}} \mathbb{P}(F(Y|Z) \leq a | Z = z) f(z)dz \qquad \text{(Eq. A.1)} \\
&= \int_{\mathcal{Z}} \mathbb{P}(F(Y|z) \leq a | Z = z) f(z)dz \\
&= \int_{\mathcal{Z}} \mathbb{P}\left(Y \leq F^{-1}(a|z) | Z = z\right) f(z)dz \\
&= \int_{\mathcal{Z}} F\left(F^{-1}(a|z) | Z = z\right) f(z)dz \\
&= \int_{\mathcal{Z}} a f(z)dz \\
&= a.
\end{aligned}$$

□

A.2 Proof of Theorem 2

Proof. Assume that $\widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x})$. It follows that, for any $0 < \alpha < 1$,

$$\begin{aligned}
\mathbb{P}\left(\text{PIT}(Y;\mathbf{X}) < \alpha|\mathbf{x}\right) &= \mathbb{P}\left(F_{Y|\mathbf{x}}(Y) \leq \alpha|\mathbf{x}\right) \\
&= \mathbb{P}\left(Y \leq F^{-1}_{Y|\mathbf{x}}(\alpha)|\mathbf{x}\right) \\
&= F_{Y|\mathbf{x}}\left(F^{-1}_{Y|\mathbf{x}}(\alpha)\right) \\
&= \alpha,
\end{aligned}$$

which shows that the distribution of $\text{PIT}(Y;\mathbf{X})$, conditional on $\mathbf{x}$, is uniform.

Now, assume that $\mathbb{P}\left(\text{PIT}(Y;\mathbf{X}) < \alpha|\mathbf{x}\right) = \alpha$ for every $0 < \alpha < 1$ and let

$$\widehat{F}_{y|\mathbf{x}}(y) = \int_{-\infty}^{y} \widehat{f}(y'|\mathbf{x})dy'$$

Then, we have

$$\begin{aligned}
\alpha &= \mathbb{P}\left(\text{PIT}(Y;\mathbf{X}) < \alpha|\mathbf{x}\right) \\
&= \mathbb{P}\left(\widehat{F}_{Y|\mathbf{x}}(Y) \leq \alpha|\mathbf{x}\right) \\
&= \mathbb{P}\left(Y \leq \widehat{F}^{-1}_{Y|\mathbf{x}}(\alpha)|\mathbf{x}\right) \\
&= F_{Y|\mathbf{x}}\left(\widehat{F}^{-1}_{Y|\mathbf{x}}(\alpha)\right).
\end{aligned}$$

It follows that

$$F_{Y|\mathbf{x}}\left(\widehat{F}^{-1}_{Y|\mathbf{x}}(\alpha)\right) = \alpha,$$

and thus

$$\widehat{F}^{-1}_{Y|\mathbf{x}}(\alpha) = F^{-1}_{Y|\mathbf{x}}\left(\alpha\right),\ \forall \alpha \in (0,1).$$

The conclusion follows from the fact that the CDF characterizes the distribution of a random variable. □

A.2.1 Proof of Corollary 1

Proof. Notice that

$$r_\alpha(\mathbf{x}) = \mathbb{E}\left[Z^\alpha|\mathbf{x}\right] = \mathbb{P}\left(\text{PIT}(Y;\mathbf{X}) < \alpha|\mathbf{x}\right)$$

It follows that $r_\alpha(\mathbf{x}) = \alpha$ for every $\alpha \in (0,1)$ if, and only if, the distribution of $\text{PIT}(\mathbf{Y};\mathbf{X})$, conditional on $\mathbf{X}$, is uniform over $(0,1)$. The conclusion follows from Theorem 2. □

A.3 Proof of Theorem 3

Proof. Notice that, under $H_0^\epsilon(\mathbf{x})$, for every $\mathbf{x}' \in B(\mathbf{x};\epsilon)$,

$$\text{PIT}(Y_i;\mathbf{X}_i)|\mathbf{X}_i = \mathbf{x}' \sim \text{Unif}(0,1),$$

and therefore

$$(\mathbf{X}_i, \gamma_{i,j}, W_{i,j}^{(b)})|\mathbf{X}_i = \mathbf{x}', \gamma_{i,j} \overset{\text{i.i.d.}}{\sim} (\mathbf{X}_i, \gamma_{i,j}, W_{i,j})|\mathbf{X}_i = \mathbf{x}', \gamma_{i,j}.$$

It follows that

$$\mathcal{D}'_{\mathbf{x}}|\mathcal{C} \overset{\text{i.i.d.}}{\sim} \mathcal{D}^{(b)}_{\mathbf{x}}|\mathcal{C},$$

where $\mathcal{C} = \{(\mathbf{X}_i, \gamma_{i,j})\}_{i,j}$, $\mathcal{D}'_{\mathbf{x}} = \{(\mathbf{X}_i, \gamma_{i,j}, W_{i,j}) \in \mathcal{D}' : \mathbf{X}_i \in B(\mathbf{x};\epsilon)\}$ and
$\mathcal{D}^{(b)}_{\mathbf{x}} := \left\{(\mathbf{X}_i, \gamma_{i,j}, W_{i,j}^{(b)}) \in \mathcal{D}^{(b)} : \mathbf{X}_i \in B(\mathbf{x};\epsilon)\right\}$.

Now, by Assumption 1, $T(\mathbf{x})$ (similary, $T^{(b)}(\mathbf{x})$) is a function of $\mathcal{D}'_{\mathbf{x}}$ (similarly, $\mathcal{D}^{(b)}_{\mathbf{x}}$). It follows that

$$T(\mathbf{x})|\mathcal{C} \overset{\text{i.i.d.}}{\sim} T^{(b)}(\mathbf{x})|\mathcal{C}.$$

Thus, by the law of large numbers, for every fixed augmented dataset $\mathcal{D}' = d_{\text{obs}}$,

$$\begin{aligned} p(\mathbf{x})|\mathcal{C}, \mathcal{D}' = d_{\text{obs}} &\xrightarrow[B\longrightarrow\infty]{\text{a.s.}} \mathbb{P}\left(T_{\mathcal{D}'=d_{\text{obs}}}(\mathbf{x}) < T_{\mathcal{D}'}(\mathbf{x})|\mathcal{C}\right) \\ &= 1 - F_{T_{\mathcal{D}'}(\mathbf{x})|\mathcal{C}}(T_{\mathcal{D}'=d_{\text{obs}}}(\mathbf{x})), \end{aligned}$$

where $T_{\mathcal{D}'=d_{\text{obs}}}$ is the test statistic computed at $\mathcal{D}' = d_{\text{obs}}$, and hence

$$p(\mathbf{x})|\mathcal{C} \xrightarrow[B\longrightarrow\infty]{\mathcal{L}} 1 - F_{T_{\mathcal{D}'}(\mathbf{x})|\mathcal{C}}(T_{\mathcal{D}'}(\mathbf{x})),$$

The conclusion follows from the fact that $F_{T_{\mathcal{D}'}(\mathbf{x})|\mathcal{C}}(T_{\mathcal{D}'}(\mathbf{x}))$ is a uniform random variable, and therefore so is $1 - F_{T_{\mathcal{D}'}(\mathbf{x})|\mathcal{C}}(T_{\mathcal{D}'}(\mathbf{x}))$. □

A.4 Proof of Theorem 4

Proof. Under the null hypothesis $H_0(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$, we have that:

$$\text{HPD}(\mathbf{y};\mathbf{x}) = \int_{\mathbf{y}':\widehat{f}(\mathbf{y}'|\mathbf{x}) \geq \widehat{f}(\mathbf{y}|\mathbf{x})} \widehat{f}(\mathbf{y}'|\mathbf{x}) d\mathbf{y} \tag{A.2}$$

$$= \int_{\mathbf{y}':f(\mathbf{y}'|\mathbf{x}) \geq f(\mathbf{y}|\mathbf{x})} f(\mathbf{y}'|\mathbf{x}) d\mathbf{y}. \tag{A.3}$$

Applying the results about uniformity of HPD for $f(\cdot|\mathbf{x})$ from Harrison et al. (2015, Section A.2) (also reproduced in the proof of Theorem 5) proves the theorem.

□

A.5 Theorem 5

We show that using HPD values are also insensitive to covariate transformations, and therefore replacing PIT values with HPD values in previous diagnostic methods does not address their fundamental limitations.

Theorem 5 (HPD values are insensitive to covariate transformations). *Let* $(\mathbf{X}, \mathbf{Y}) \sim F_{\mathbf{X},\mathbf{Y}}$. *If there exists a function* $g : \mathcal{X} \to \mathcal{Z}$ *such that* $\widehat{f}(\mathbf{y}|\mathbf{x}) = f(\mathbf{y}|g(\mathbf{x}))$, *then HPD*$(\mathbf{Y}; \mathbf{X}) \sim Unif(0, 1)$.

Proof of Theorem 5. Under the assumption, we can rewrite the HPD value as:

$$\begin{aligned}
\text{HPD}(\mathbf{y}, \mathbf{x}) &= \int_{\mathbf{y}' : f(\mathbf{y}'|g(\mathbf{x})) > f(\mathbf{y}|g(\mathbf{x}))} f(\mathbf{y}'|g(\mathbf{x}))dy' \\
&= \int_{y' : f(\mathbf{y}'|\mathbf{z}) > f(\mathbf{y}|\mathbf{z})} f(\mathbf{y}'|\mathbf{z})dy' \\
&= \text{HPD}(\mathbf{y}, \mathbf{z}),
\end{aligned}$$

with $g(\mathbf{x}) = \mathbf{z}$.

Following the proof structure by Harrison et al. (2015) closely, we define the random variable

$$\xi_{\mathbf{z},\mathbf{y}} = \text{HPD}(\mathbf{z}, \mathbf{y}),$$

equipped with the probability density function $h : (\mathcal{Z} \times \mathcal{Y}) \to \mathbb{R}$. Dropping the subscripts for simplicity, let $\xi^* = \text{HPD}(\mathbf{z}^*, \mathbf{y}^*)$ the HPD value of a specific pair $(\mathbf{z}^*, \mathbf{y}^*)$; ξ^* is the probability mass of f above the level set $f(\mathbf{y}^*|\mathbf{z}^* = g(\mathbf{x}^*))$. Without loss of generality, if we show that $h(\xi^*) = 1$, then we can conclude that $\xi(y, z)$ is uniformly distributed $Unif[0, 1]$.

Using the fundamental theorem of calculus we can write:

$$\begin{aligned}
h(\xi^*) &= \frac{\partial}{\partial \xi^*} \int_{-\infty}^{\xi^*} g(\epsilon) d\epsilon \\
&= \frac{\partial}{\partial \xi^*} \int_{-\infty}^{\xi^*} \int_{\mathcal{Z}\times\mathcal{Y}} \delta(\xi(y,z) - \epsilon) dF(z,y) d\epsilon \\
&= \frac{\partial}{\partial \xi^*} \int_{\mathcal{Z}\times\mathcal{Y}} \Phi(\xi(y,z) - \xi^*) dF(z,y) \\
&= \frac{\partial}{\partial \xi^*} \int_{\mathcal{Z}} \left[\int_{\mathcal{Y}} \Phi(\xi(y,z) - \xi^*) f(y|z) dy \right] f(z) dz \\
&= \frac{\partial}{\partial \xi^*} \int_{\mathcal{Z}} \xi^* f(z) dz \\
&= \frac{\partial}{\partial \xi^*} \xi^* \\
&= 1
\end{aligned}$$

where Φ is the Heavyside function, which is 1 when the argument is positive and 0 otherwise. □

A.6 Proof of Theorem 6

Lemma 1. *Let G and H be two cumulative distribution functions such that G dominates H, and let μ_G and μ_H be their associated measures over $\mathbb{R}$. Then, for every fixed $y \in \mathbb{R}$,*

$$\mu_H\left(\{y' \in \mathbb{R} : y' \leq y\}\right) = \mu_H\left(\{y' \in \mathbb{R} : G(y') \leq G(y)\}\right).$$

Proof. Fix $y \in \mathbb{R}$ and let $A = \{y' \in \mathbb{R} : y' \leq y\}$ and $B = \{y' \in \mathbb{R} : G(y') \leq G(y)\}$. Because $A \subseteq B$,

$$\mu_H(A) \leq \mu_H(B). \tag{A.4}$$

We note that $\mu_G(B \cap A^c) = 0$. From this and the assumption that G dominates H, we conclude that $\mu_H(B \cap A^c) = 0$. It follows that

$$\begin{aligned}
\mu_H(B) &= \mu_H(B \cap A) + \mu_H(B \cap A^c) && \text{(A.5)} \\
&\leq \mu_H(A) + 0 \\
&= \mu_H(A). && \text{(A.6)}
\end{aligned}$$

From Equations A.4 and A.5, we conclude that $\mu_H(A) = \mu_H(B)$.

□

Lemma 2. *Fix $y \in \mathbb{R}$ and let $\gamma := \widehat{F}(y|\mathbf{x})$. Then, under Assumptions 2 and 3, $\widetilde{F}(y|\mathbf{x}) = \widehat{r}^{\widehat{f}}(\gamma; \mathbf{x})$ and $F(y|\mathbf{x}) = r^{\widehat{f}}(\gamma; \mathbf{x})$.*

Proof. We note that $\gamma = \widehat{F}(y|\mathbf{x})$ implies that $y = \widehat{F}^{-1}(\gamma|\mathbf{x})$. It follows then by construction,

$$\widetilde{F}(y|\mathbf{x}) = \widetilde{F}\left(\widehat{F}^{-1}(\gamma|\mathbf{x})|\mathbf{x}\right) = \widehat{r}^{\widehat{f}}(\gamma;\mathbf{x}).$$

Moreover,

$$\begin{aligned} F(y|\mathbf{x}) &= \mathbb{P}(Y \leq y|\mathbf{x}) && \text{(A.7)} \\ &= \mathbb{P}\left(\widehat{F}(Y|\mathbf{x}) \leq \widehat{F}(y|\mathbf{x})|\mathbf{x}\right) \\ &\quad \text{(Assumption 3 and Lemma 1)} \\ &= \mathbb{P}\left(\text{PIT}(Y;\mathbf{x}) \leq \widehat{F}(y|\mathbf{x})|\mathbf{x}\right) && \text{(A.8)} \\ &= \mathbb{P}\left(\text{PIT}(Y;\mathbf{x}) \leq \gamma|\mathbf{x}\right) \\ &= r^{\widehat{f}}(\gamma;\mathbf{x}), && \text{(A.9)} \end{aligned}$$

which concludes the proof. □

Proof of Theorem 6. Consider the change of variables $\gamma = \widehat{F}(y|\mathbf{x})$, so that $d\gamma = \widehat{f}(y|\mathbf{x})dy$. Lemma 2 implies that $\widetilde{F}(y|\mathbf{x}) = \widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})$ and $F(y|\mathbf{x}) = r^{\widehat{f}}(\gamma;\mathbf{x})$. It follows from that and Assumption 4 that

$$\begin{aligned} &\int\int \left(\widetilde{F}(y|\mathbf{x}) - F(y|\mathbf{x})\right)^2 dP(y,\mathbf{x}) \\ &\quad \leq K \int\int \left(\widetilde{F}(y|\mathbf{x}) - F(y|\mathbf{x})\right)^2 \widehat{f}(y|\mathbf{x})dyP(\mathbf{x}) \\ &\quad = K \int\int \left(\widehat{r}^{\widehat{f}}(\gamma;\mathbf{x}) - r^{\widehat{f}}(\gamma;\mathbf{x})\right)^2 d\gamma dP(\mathbf{x}). \end{aligned}$$

The conclusion follows from Assumption 5. □

A.7 Proof of Theorem 7

Proof. From Lemma 2,

$$\begin{aligned} &\sup_{\mathbf{x}\in\mathcal{X}, y\in\mathbb{R}} \left|\widetilde{F}(y|\mathbf{x}) - F(y|\mathbf{x})\right| \\ &= \sup_{\mathbf{x}\in\mathcal{X}, \gamma\in[0,1]} \left|\widehat{r}^{\widehat{f}}(\gamma;\mathbf{x}) - r^{\widehat{f}}(\gamma;\mathbf{x})\right| \xrightarrow[n\longrightarrow\infty]{\text{a.s.}} 0, \end{aligned}$$

where the last step follows from Assumption 6. It then follows from Assumption 2 that

$$\sup_{\mathbf{x}\in\mathcal{X}, \gamma\in[0,1]} \left|\widetilde{F}^{-1}(\gamma|\mathbf{x}) - F^{-1}(\gamma|\mathbf{x})\right| \xrightarrow[n\longrightarrow\infty]{\text{a.s.}} 0,$$

and, in particular,

$$\sup_{\mathbf{x}\in\mathcal{X}, \alpha\in\{.5\alpha, 1-.5\alpha\}} \left|\widetilde{F}^{-1}(\alpha|\mathbf{x}) - F^{-1}(\alpha|\mathbf{x})\right| \xrightarrow[n\longrightarrow\infty]{\text{a.s.}} 0,$$

from which the conclusion of the theorem follows. □

A.8 Proof of Theorem 8

Proof. Fix $y \in \mathbb{R}$ and let $\gamma = \widehat{F}(y|\mathbf{x})$, so that $y = \widehat{F}^{-1}(\gamma|\mathbf{x})$. It follows that

$$\begin{aligned}
\widetilde{F}(y|\mathbf{x}) &= \widetilde{F}\left(\widehat{F}^{-1}(\gamma|\mathbf{x})|\mathbf{x}\right) \\
&= \widehat{r}(\gamma;\mathbf{x}) \\
&= r(\gamma;\mathbf{x}) \\
&= \mathbb{P}\left(\widehat{F}(Y|\mathbf{x}) \leq \widehat{F}(y|\mathbf{x})|\mathbf{x},\gamma\right) \\
&= \mathbb{P}\left(Y \leq y|\mathbf{x},\gamma\right) \\
&= F(y|\mathbf{x}),
\end{aligned}$$

and therefore $\widetilde{f}(y|\mathbf{x}) = f(y|\mathbf{x})$ for almost every $y \in \mathbb{R}$. It follows that $C_\alpha(\mathbf{x}) = \text{HPD}_\alpha(\mathbf{x})$. The claim about conditional coverage follows from the definition of the HPD. □

A.9 Proof of Theorem 9

Proof. Following the proof structure by Izbicki et al. (2022) closely, we note that cs is a function of $\mathcal{D}'$. Therefore, $cs(\mathbf{X}_i, Y_i)$ can be written as $u(\mathbf{X}_i, Y_i, \mathcal{D}')$. Since the points in $\mathcal{D}'$ are exchangeable, it follows that $U_i := u(\mathbf{X}_i, Y_i, \mathcal{D}')$ are exchangeable. Therefore, for every $A \in \mathcal{A}$, $\{U_i : \mathbf{X}_i \in A\}$ are exchangeable. In particular, given that $\mathbf{X}_{n+1} \in A$, we have that $\{U_i : \mathbf{X}_i \in A(\mathbf{X}_{n+1})\}$ are exchangeable. Now, we have that for every $A \in \mathcal{A}$,

$$\begin{aligned}
\mathbb{P}\left(U_{n+1} \geq U^A_{\lfloor\alpha\rfloor}(\mathbf{X}_{n+1})|\mathbf{X}_{n+1} \in A)\right) &\geq 1-\alpha \\
\mathbb{P}\left(cs(\mathbf{X}_{n+1}, Y_{n+1}) \geq U^A_{\lfloor\alpha\rfloor}(\mathbf{X}_{n+1})|\mathbf{X}_{n+1} \in A)\right) &\geq 1-\alpha \\
\mathbb{P}\left(Y_{n+1} \in \left\{y : cs(\mathbf{X}_{n+1}, y) \geq U^A_{\lfloor\alpha\rfloor}(\mathbf{X}_{n+1})\right\}|\mathbf{X}_{n+1} \in A)\right) &\geq 1-\alpha \\
\mathbb{P}\left(Y_{n+1} \in C(\mathbf{X}_{n+1})|\mathbf{X}_{n+1} \in A)\right) &\geq 1-\alpha
\end{aligned}$$

It follows that $C(X_{n+1})$ satisfies local validity with respect to $\mathcal{A}$, and therefore by construction also satisfies marginal validity. □

A.10 Proof of Theorem 11

Proof. From Lemma 2,

$$\sup_{\mathbf{x}\in\mathcal{X},y\in\mathbb{R}}\left|\widetilde{F}(y|\mathbf{x})-F(y|\mathbf{x})\right|$$
$$=\sup_{\mathbf{x}\in\mathcal{X},\gamma\in[0,1]}\left|\widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})-r^{\widehat{f}}(\gamma;\mathbf{x})\right|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0,$$

where the last step follows from Assumption 6.

Let $U_{\lfloor\gamma\rfloor}$ denote the empirical γ-quantile of the set $\left\{\widetilde{F}(y'|\mathbf{x}'):\mathbf{x}'\in A_{\mathbf{x}}\right\}$ for the calibrated CDF. Let $V_{\lfloor\gamma\rfloor}$ denote the empirical γ-quantile of the set $\{F(y'|\mathbf{x}'):\mathbf{x}'\in A_{\mathbf{x}}\}$ for the true CDF. We have that

$$\sup_{\mathbf{x}\in\mathcal{X},\gamma\in[0,1]}\left|U_{\lfloor\gamma\rfloor}-V_{\lfloor\gamma\rfloor}\right|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0.$$

But the conformal set $(\mathbf{X}'_1,Y'_1),\ldots,(\mathbf{X}'_n,Y'_n)$ is drawn i.i.d. from $F_{X,Y}$, so

$$\sup_{\mathbf{x}\in\mathcal{X},\gamma\in[0,1]}\left|V_{\lfloor\gamma\rfloor}-\gamma\right|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0,$$

and thus

$$\sup_{\mathbf{x}\in\mathcal{X},\gamma\in[0,1]}\left|U_{\lfloor\gamma\rfloor}-\gamma\right|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0.$$

That is, the empirical distribution of $\left\{\widetilde{F}(y'|\mathbf{x}'):\mathbf{x}'\in A_{\mathbf{x}}\right\}$ converges to the $Unif[0,1]$ distribution.

Since $L_{\mathbf{x},\alpha}$ is defined to be the empirical 0.5α-quantile of $\left\{\widetilde{F}(y'|\mathbf{x}'):x'\in A_{\mathbf{x}}\right\}$,

$$\sup_{\mathbf{x}\in\mathcal{X}}|L_{\mathbf{x},\alpha}-0.5\alpha|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0.$$

Similarly, since $U_{\mathbf{x},\alpha}$ is defined to be the empirical $1-0.5\alpha$-quantile of $\left\{\widetilde{F}(y'|\mathbf{x}'):x'\in A_{\mathbf{x}}\right\}$,

$$\sup_{\mathbf{x}\in\mathcal{X}}|U_{\mathbf{x},\alpha}-(1-0.5\alpha)|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0.$$

Recall that we showed in the proof of Theorem 7, under the same assumptions, that

$$\sup_{\mathbf{x}\in\mathcal{X},\alpha\in\{.5\alpha,1-.5\alpha\}}\left|\widetilde{F}^{-1}(\alpha|\mathbf{x})-F^{-1}(\alpha|\mathbf{x})\right|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0.$$

Therefore, we have

$$\sup_{\mathbf{x}\in\mathcal{X}}\left|\widetilde{F}^{-1}(L_{x,\alpha}|\mathbf{x})-F^{-1}(0.5\alpha|\mathbf{x})\right|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0,$$
$$\sup_{\mathbf{x}\in\mathcal{X}}\left|\widetilde{F}^{-1}(U_{x,\alpha}|\mathbf{x})-F^{-1}(1-0.5\alpha|\mathbf{x})\right|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0,$$

from which the conclusion of the theorem follows. □

A.11 Proof of Theorem 12

Proof. Recall that we showed in the proof of Theorem 8 under the same assumptions, that $\widetilde{F}(y|\mathbf{x}) = F(y|\mathbf{x})$ and therefore $\widetilde{f}(y|\mathbf{x}) = f(y|\mathbf{x})$ for almost every $y \in \mathbb{R}$.

Let $U_{\lfloor\gamma\rfloor}$ denote the empirical γ-quantile of the set of HPD values $\left\{\widetilde{H}(y';\mathbf{x}') : \mathbf{x}' \in A_{\mathbf{x}}\right\}$ for the calibrated density, and let $V_{\lfloor\gamma\rfloor}$ denote the empirical γ-quantile of the set of HPD values $\{H(y'|\mathbf{x}') : \mathbf{x}' \in A_{\mathbf{x}}\}$ for the true density. We have that $U_{\lfloor\gamma\rfloor} = V_{\lfloor\gamma\rfloor}$ for almost every $y \in \mathcal{R}$.

But the conformal set $(\mathbf{X}'_1, Y'_1), \ldots, (\mathbf{X}'_n, Y'_n)$ is drawn i.i.d. from $F_{X,Y}$, so

$$\sup_{\mathbf{x}\in\mathcal{X},\gamma\in[0,1]}\left|V_{\lfloor\gamma\rfloor}-\gamma\right|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0,$$

and thus

$$\sup_{\mathbf{x}\in\mathcal{X},\gamma\in[0,1]}\left|U_{\lfloor\gamma\rfloor}-\gamma\right|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0.$$

That is, the empirical distribution of $\left\{\widetilde{H}(y'|\mathbf{x}') : \mathbf{x}' \in A_{\mathbf{x}}\right\}$ converges to the $Unif[0,1]$ distribution.

Since α' is defined to be the empirical α-quantile of $\left\{\widetilde{H}(y'|\mathbf{x}') : x' \in A_{\mathbf{x}}\right\}$,

$$\sup_{\mathbf{x}\in\mathcal{X}}|\alpha'-\alpha|\xrightarrow[n\longrightarrow\infty]{\text{a.s.}}0,$$

from which the conclusion of the theorem follows. □