

---

# Diagnostics for Conditional Density Models and Bayesian Inference Algorithms

---

David Zhao<sup>1</sup>

Niccolò Dalmaso<sup>1</sup>

Rafael Izbicki<sup>2</sup>

Ann B. Lee<sup>1</sup>

<sup>1</sup>Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

<sup>2</sup>Department of Statistics, Federal University of São Carlos (UFSCar), São Carlos, Brazil

## Abstract

There has been growing interest in the AI community for precise uncertainty quantification. Conditional density models  $f(y|\mathbf{x})$ , where  $\mathbf{x}$  represents potentially high-dimensional features, are an integral part of uncertainty quantification in prediction and Bayesian inference. However, it is challenging to assess conditional density estimates and gain insight into modes of failure. While existing diagnostic tools can determine whether an approximated conditional density is compatible overall with a data sample, they lack a principled framework for identifying, locating, and interpreting the nature of statistically significant discrepancies over the entire feature space. In this paper, we present rigorous and easy-to-interpret diagnostics such as (i) the “Local Coverage Test” (LCT), which distinguishes an arbitrarily misspecified model from the true conditional density of the sample, and (ii) “Amortized Local P-P plots” (ALP) which can quickly provide interpretable graphical summaries of distributional differences at any location  $\mathbf{x}$  in the feature space. Our validation procedures scale to high dimensions and can potentially adapt to any type of data at hand. We demonstrate the effectiveness of LCT and ALP through a simulated experiment and applications to prediction and parameter inference for image data.

## 1 INTRODUCTION

There has been growing interest in the AI community for precise uncertainty quantification (UQ), with conditional density models playing a key role in UQ in prediction and Bayesian inference. For instance, 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 infor-

mative than point predictions. Indeed, in prediction settings,  $f$  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  $\theta$  of interest after observing data  $\mathbf{x}$ .

Recently, a large body of work in machine learning has been developed for estimating conditional densities  $f$  for all possible values of  $\mathbf{x}$ , or to generate predictions that follow the unknown conditional density (see Uria et al. 2014, Sohn et al. 2015, Papamakarios et al. 2017, Dutordoir et al. 2018, Papamakarios et al. 2021 and references therein). With the advent of high-precision data and simulations, simulation-based inference (SBI; Cranmer et al. [2020]) has also played a growing role in disciplines ranging from physics, chemistry and engineering to the biological and social sciences. The SBI category includes 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].

Inevitably, 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, especially for high-dimensional or mixed-type data  $\mathbf{x}$ . There does not currently 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.

**Related work.** Large AI models, such as deep generative autoregressive models or Bayesian networks, are typically fit using global loss functions like the Kullback-Leibler divergence or the  $L^2$  loss [Izbicki et al., 2017, Rothfuss et al., 2019]. 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 (using larger training samples, training times, etc.), or if the current estimate is “close enough”. Another line of work assesses goodness-of-

fit of a conditional density model via a two-sample test that compare samples from  $\hat{f}$  and  $f$ . Earlier tests involve a conditional version of the standard Kolmogorov test [Andrews, 1997, Zheng, 2000] in one dimension, or are 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. While these tests are consistent, they do not provide insight on how the distributions of  $\hat{f}$  and  $f$  differ locally. Kernel approaches also require the user to specify an appropriate kernel and tuning parameters, which can be challenging in practice. Finally, existing diagnostics that do describe the nature of inconsistencies between  $\hat{f}$  and  $f$  only test for a form of overall coherence between a data-averaged conditional (posterior) distribution and its marginal (prior) distribution. Typically, they compute 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 some clearly misspecified conditional models including models that ignore the dependence on the covariates altogether [Schmidt et al., 2020]. (Our Theorem 1 details different failure modes of existing diagnostics.)

**Contribution and novelty.** Our work provides diagnostic tools for UQ and calibration of predictive models that provide insight in simple, explainable terms like coverage, bias, dispersion, and multimodality in  $y$  (output of interest) as a function of  $\mathbf{x}$  (observed inputs). Having interpretable diagnostics is crucial for scientific collaborators and end users to build trust in large AI models.

Existing diagnostics for conditional density models cannot detect every kind of misspecified model and give insight into local quality of fit at any given  $\mathbf{x}$ . Our method quantifies deviations between actual and nominal coverage in  $y$ . It (i) detects arbitrarily misspecified models and (ii) assesses and visualizes quality of fit anywhere in feature space, even at points without observed data, in terms of easy-to-explain diagnostics. To the best of our knowledge, no other method in the literature provides both consistency and diagnostics for complex high-dimensional data.

To enrich our vocabulary for desired properties of CDEs, we begin our paper by defining global and local consistency (see Definitions 1 and 3, respectively). We then describe our diagnostic framework, which 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.)
- **[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.)

- **[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 1 for examples).

Our diagnostics are easy and fast to compute, and can identify, locate, and interpret the nature of (statistically significant) discrepancies over the entire feature space. At the heart of our approach is the realization that the local coverage of a CDE model is itself a conditional probability (see Equation 5) 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 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 Figures 4-6 for examples).

## 2 EXISTING DIAGNOSTICS ARE INSENSITIVE TO COVARIATE TRANSFORMATIONS

**Notation.** Let  $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$  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 3.3,  $Y$  is multivariate), and a random vector  $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ . In a prediction setting,  $\mathcal{D}$  represents a hold-out set not used to train  $\hat{f}$ . In a Bayesian setting,  $Y$  represents the parameter of interest (sometimes also denoted with  $\theta$ ), 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  $\mathbf{X}|Y_i$ .

Ideally, a test should be able to distinguish *any* given alternative conditional density model  $\hat{f}(y|\mathbf{x})$  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  $\hat{f}(y|\mathbf{x})$  is *globally consistent with the density  $f(y|\mathbf{x})$*  if the following null hypothesis holds:

$$H_0 : \hat{f}(y|\mathbf{x}) = f(y|\mathbf{x}) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } y \in \mathcal{Y}. \quad (1)$$

Note that  $\hat{f}$  is a particular fixed conditional density estimate, and we test whether samples from  $\hat{f}$  are consistent with samples from  $f$ . Existing diagnostics typically validate density models by computing PIT values on independent data, which were not used to estimate  $\hat{f}(y|\mathbf{x})$ :

**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  $\hat{f}(y|\mathbf{x})$ , is

$$PIT(y; \mathbf{x}) = \int_{-\infty}^y \hat{f}(y'|\mathbf{x}) dy'. \quad (2)$$

See Figure 2, top panel for an illustration of this calculation.

**Remark 1.** For implicit models of  $\hat{f}(y|\mathbf{x})$  (that is, generative models that via e.g. MCMC can sample from, but not directly evaluate  $\hat{f}$ ), we can approximate the PIT values by forward-simulating data: For fixed  $\mathbf{x} \in \mathcal{X}$  and  $y \in \mathcal{Y}$ , draw  $Y_1, \dots, Y_L \sim f(\cdot|\mathbf{x})$ . 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  $\hat{f}(y|\mathbf{x})$  is globally consistent, then the PIT values are uniformly distributed. More precisely, if  $H_0$  (Equation 1) is true, then the random variables  $PIT(Y_1; \mathbf{X}_1), \dots, PIT(Y_n; \mathbf{X}_n) \stackrel{i.i.d.}{\sim} \text{Unif}(0, 1)$ . This result is 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].

Our first point is that unfortunately, such random variables can be uniformly distributed even if global consistency does not hold. This is shown in the following theorem.

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

$$\hat{f}(y|\mathbf{x}) = f(y|g(\mathbf{x})). \quad (3)$$

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

Many models naturally lead to estimates that could satisfy the condition in Equation 3, even without being globally consistent. In fact, clearly misspecified models  $\hat{f}$  can yield uniform PIT values and “pass” an associated goodness-of-fit test regardless of the sample size. For example: if  $\hat{f}(y|\mathbf{x})$  is based on a linear model, then  $\hat{f}(y|\mathbf{x})$  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  $\hat{f}(y|\mathbf{x}) = f(y|g(\mathbf{x}))$  even when  $\hat{f}(y|\mathbf{x})$  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 and Lee, 2017, Dalmaso et al., 2020] could satisfy  $\hat{f}(y|\mathbf{x}) = f(y|g(\mathbf{x}))$  for  $g(\mathbf{x}) := (\mathbf{x})_S$ , where  $S \subset \{1, \dots, d\}$  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 for an illustration.

Our second point is that current diagnostics also do not pinpoint the locations in feature space  $\mathcal{X}$  where the estimates of  $f$  should be improved. Hence, in addition to global consistency, we need diagnostics that test the following property:

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

$$H_0(\mathbf{x}) : \hat{f}(y|\mathbf{x}) = f(y|\mathbf{x}) \text{ for every } y \in \mathcal{Y}. \quad (4)$$

In the next section, we introduce new diagnostics that are able to test whether a conditional density model  $\hat{f}$  is both globally and locally consistent with the underlying conditional distribution  $f$  of the data. Our diagnostics are still based on PIT, and hence 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 NEW DIAGNOSTICS TEST LOCAL AND GLOBAL CONSISTENCY

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  $H_0(\mathbf{x}) : \hat{f}(\cdot|\mathbf{x}) = f(\cdot|\mathbf{x})$  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  $\hat{f}$  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  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$  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  $\alpha \in (0, 1)$ , we consider the cumulative distribution function (CDF) of PIT at  $\mathbf{x}$ ,

$$r_\alpha(\mathbf{x}) := \mathbb{P}(PIT(Y; \mathbf{x}) < \alpha | \mathbf{x}), \quad (5)$$

which is the regression of the random variable  $W^\alpha := \mathbb{I}(PIT(Y; \mathbf{X}) < \alpha)$  on  $\mathbf{X}$ .

From Theorem 2, it follows that the estimated density is locally consistent at  $\mathbf{x}$  if and only if  $r_\alpha(\mathbf{x}) = \alpha$  for every  $\alpha$ :

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

Our new diagnostics are able to test for both local and global consistency. They rely on the simple idea of estimating  $r_\alpha(\mathbf{x})$  and then evaluating how much it deviates from  $\alpha$  (see Section 3.1). Note that

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

where  $\hat{q}_\alpha(\mathbf{x})$  is the  $\alpha$ -quantile of  $\hat{f}$ . That is,  $r_\alpha(\mathbf{x})$  evaluates the local level- $\alpha$  coverage of  $\hat{f}$  at  $\mathbf{x}$ . In Section 3.2, we explore the connection between test statistics and coverage, for interpretable descriptions of how conditional density models  $\hat{f}$  may fail to approximate the true conditional density  $f$ .

### 3.1 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  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$  from  $F_{\mathbf{X}, Y}$  (which was not used to construct  $\hat{f}$ ), we compute  $W_i^\alpha := \mathbb{I}(\text{PIT}(Y_i; \mathbf{X}_i) < \alpha)$ . To estimate the coverage  $r_\alpha(\mathbf{x})$  (Equation 5) for any  $\mathbf{x} \in \mathcal{X}$ , we then simply regress  $W$  on  $\mathbf{X}$  using the transformed data  $(\mathbf{X}_1, W_1), \dots, (\mathbf{X}_n, W_n)$ . Numerous classes of regression estimators can be used, from kernel smoothers to random forests to neural networks.

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

$$T(\mathbf{x}) := \frac{1}{|G|} \sum_{\alpha \in G} (\hat{r}_\alpha(\mathbf{x}) - \alpha)^2,$$

where  $\hat{r}_\alpha$  denotes the regression estimator and  $G$  is a grid of  $\alpha$  values. Large values of  $T(\mathbf{x})$  indicate a large discrepancy between  $\hat{f}$  and  $f$  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 1 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 advocate using a method like Benjamini-Hochberg to control the false discovery rate.

Similarly, we can also test global consistency (Definition 1) with a Monte Carlo strategy. Algorithm 3 in Supp. Mat. B. details our procedure. We introduce the *Global Coverage Test* (GCT) based on the following test statistic:

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

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 maximize power by choosing the regression model with the smallest MSE on validation data.

### 3.2 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

---

#### Algorithm 1 P-values for Local Coverage Test

---

**Require:** conditional density model  $\hat{f}$ ; test data  $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ ; test point  $\mathbf{x} \in \mathcal{X}$ ; regression estimator  $\hat{r}$ ; number of null training samples  $B$

**Ensure:** estimated p-value  $\hat{p}(\mathbf{x})$  for any  $\mathbf{x} \in \mathcal{X}$

```

1: // Compute test statistic at  $\mathbf{x}$ :
2: Compute values  $\text{PIT}(Y_1; \mathbf{X}_1), \dots, \text{PIT}(Y_n; \mathbf{X}_n)$ 
3:  $G \leftarrow$  grid of  $\alpha$  values in  $(0, 1)$ .
4: for  $\alpha$  in  $G$  do
5:   Compute indicators  $W_1^\alpha, \dots, W_n^\alpha$ 
6:   Train regression method  $\hat{r}_\alpha$  on  $\{\mathbf{X}_i, W_i^\alpha\}_{i=1}^n$ 
7: end for
8: Compute test statistic  $T(\mathbf{x})$ 
9: // Recompute test statistic under null distribution:
10: for  $b$  in  $1, \dots, B$  do
11:   Draw  $U_1^{(b)}, \dots, U_n^{(b)} \sim \text{Unif}[0, 1]$ .
12:   for  $\alpha$  in  $G$  do
13:     Compute indicators  $\{W_{\alpha,i}^{(b)} = \mathbb{I}(U_i^{(b)} < \alpha)\}_{i=1}^n$ 
14:     Train regression method  $\hat{r}_\alpha^{(b)}$  on  $\{\mathbf{X}_i, W_{\alpha,i}^{(b)}\}_{i=1}^n$ 
15:   end for
16:   Compute  $T^{(b)}(\mathbf{x}) := \frac{1}{|G|} \sum_{\alpha \in G} (\hat{r}_\alpha^{(b)}(\mathbf{x}) - \alpha)^2$ 
17: end for
18: return  $\hat{p}(\mathbf{x}) := \frac{1}{B} \sum_{b=1}^B \mathbb{I}(T(\mathbf{x}) < T^{(b)}(\mathbf{x}))$ 

```

---

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 1 illustrates how a local P-P plot of  $\hat{r}_\alpha(\mathbf{x})$  against  $\alpha$  (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 the estimated density  $\hat{f}$  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  $\hat{f}$  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 3.1 and the local P-P plots only require initially training  $\hat{r}_\alpha$  on the observed data; the regression estimator can then be used to compute  $\hat{r}_\alpha(\mathbf{x}_{val})$  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 2 details the construction of ALPs, including how to compute confidence bands by a Monte Carlo algorithm.

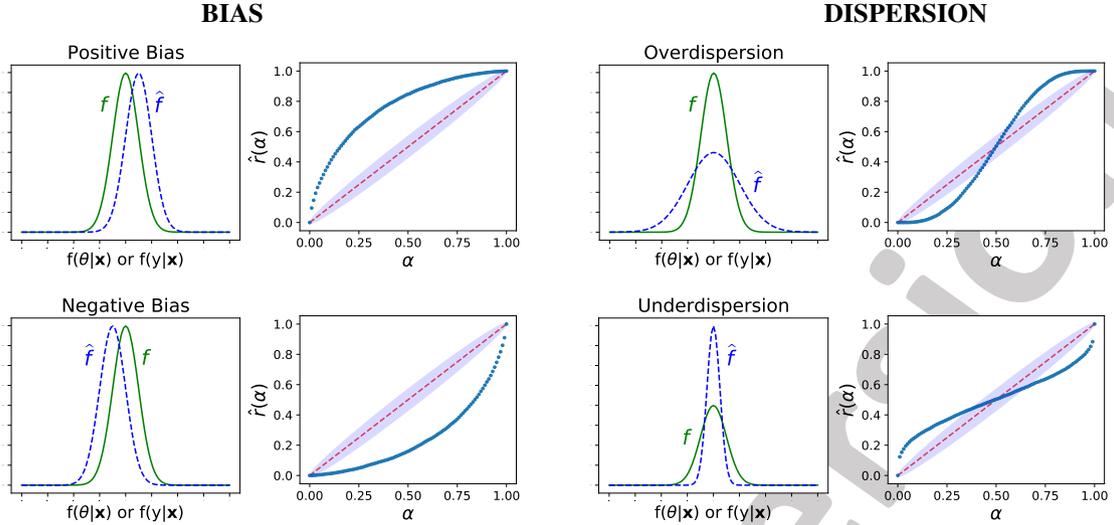


Figure 1: P-P plots are commonly used to assess how well a density model fits actual data. Such plots display, in a clear and interpretable way, effects like bias (left panel) and dispersion (right panel) in an estimated distribution  $\hat{f}$  vis-a-vis the true data-generating distribution  $f$ . Our framework yields a computationally efficient way to construct “amortized local P-P plots” for comparing conditional densities  $\hat{f}(\theta|\mathbf{x})$  and  $\hat{f}(y|\mathbf{x})$  at any location  $\mathbf{x}$  of the feature space  $\mathcal{X}$ . See text for details and Sections 4-6 for examples.

---

#### Algorithm 2 Confidence bands for local P-P plot

---

**Require:** test data  $\{\mathbf{X}_i\}_{i=1}^n$ ; test point  $\mathbf{x} \in \mathcal{X}$ ; regression estimator  $\hat{r}$ ; number of null training samples  $B$ ; confidence level  $\eta$

**Ensure:** estimated upper and lower confidence bands  $U(\mathbf{x}), L(\mathbf{x})$  at level  $1 - \eta$  for any  $\mathbf{x} \in \mathcal{X}$

```

1: // Recompute regression under null distribution:
2:  $G \leftarrow$  grid of  $\alpha$  values in  $(0, 1)$ .
3: for  $b$  in  $1, \dots, B$  do
4:   Draw  $U_1^{(b)}, \dots, U_n^{(b)} \sim \text{Unif}[0, 1]$ .
5:   for  $\alpha$  in  $G$  do
6:     Compute indicators  $\{W_{\alpha,i}^{(b)} = \mathbb{I}(U_i^{(b)} < \alpha)\}_{i=1}^n$ 
7:     Train regression method  $\hat{r}_\alpha^{(b)}$  on  $\{\mathbf{X}_i, W_{\alpha,i}^{(b)}\}_{i=1}^n$ 
8:   end for
9:   Compute  $\hat{r}_\alpha^{(b)}(\mathbf{x})$ 
10: end for
11: // Compute  $(1 - \eta)$  confidence bands for  $\hat{r}_\alpha(\mathbf{x})$ :
12:  $U(\mathbf{x}), L(\mathbf{x}) \leftarrow \emptyset$ 
13: for  $\alpha$  in  $G$  do
14:    $U(\mathbf{x}) \leftarrow U(\mathbf{x}) \cup (1 - \frac{\eta}{2})$ -quantile of  $\{\hat{r}_\alpha^{(b)}(\mathbf{x})\}_{b=1}^B$ 
15:    $L(\mathbf{x}) \leftarrow L(\mathbf{x}) \cup \frac{\eta}{2}$ -quantile of  $\{\hat{r}_\alpha^{(b)}(\mathbf{x})\}_{b=1}^B$ 
16: end for
17: return  $U(\mathbf{x}), L(\mathbf{x})$ 

```

---

### 3.3 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. [2021] for the prediction setting. That is, the PIT values can be computed using the estimate  $\hat{f}(h(\mathbf{Y})|\mathbf{x})$  induced by  $\hat{f}(\mathbf{Y}|\mathbf{x})$  for some chosen  $h: \mathbb{R}^p \rightarrow \mathbb{R}$ . 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  $\theta$ . Thus, it is natural to evaluate PIT values of  $h(\theta) = \theta_i$  for each parameter of interest. Another useful projection is copPIT [Ziegel and Gneiting, 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  $\hat{f}$  is through highest predictive density values (HPD values; Harrison et al. 2015, Dalmaso et al. 2020), which are defined by

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

(see Figure 2, bottom, for an illustration).  $\text{HPD}(\mathbf{y}; \mathbf{x})$  is a measure of how plausible  $\mathbf{y}$  is according to  $\hat{f}(\mathbf{y}|\mathbf{x})$  (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 [Dalmaso 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 4, Supp. Mat. A). Fortunately, HPD values are uniform under the local consistency hypothesis:

**Theorem 3.** For any  $\mathbf{x} \in \mathcal{X}$ , if the local null hypothesis  $H_0(\mathbf{x}) : \hat{f}(\cdot|\mathbf{x}) = f(\cdot|\mathbf{x})$  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 3.1 and 3.2 can be used with HPD values to check global and local consistency for multivariate responses, as well as to construct local P-P plots. (Supp. Mat. F showcases multivariate extensions via HPD.) The HPD statistic is especially appealing if one wishes to construct predictive regions with  $\hat{f}$  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  $\hat{f}(\mathbf{y}|\mathbf{x})$  is larger than a certain threshold (in the Bayesian case, these are the highest posterior credible regions). More precisely, if  $HPD_\alpha(\mathbf{x})$  is the  $\alpha$ -level HPD set for  $\mathbf{y}$ , then

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

Thus, by testing local consistency of  $\hat{f}$  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  $\hat{f} \neq f$ .

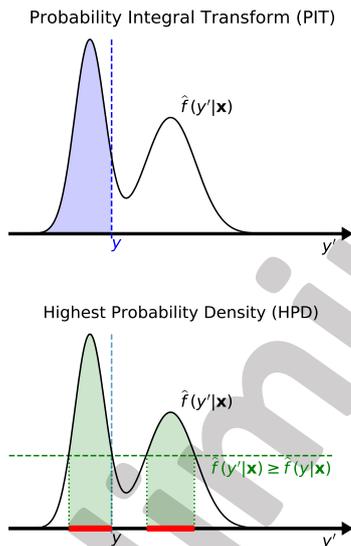


Figure 2: Schematic diagram of the construction of PIT (top panel, shaded blue area) and HPD value (bottom panel, shaded green area) for an estimated density  $\hat{f}$  evaluated at  $(y, \mathbf{x})$ . The highlighted red intervals in the bottom panel correspond to the highest density region (HDR) of  $y|\mathbf{x}$ .

## 4 EXAMPLE 1: OMITTED VARIABLE BIAS IN CDE MODELS

Our first example involves omitted but clearly relevant variables in a prediction setting. Inspired by Section 2.2.2 of

Shalizi [2021], we generate  $\mathbf{X} = (X_1, X_2) \sim N(0, \Sigma) \in \mathbb{R}^2$ , with  $\Sigma_{1,1} = \Sigma_{2,2} = 1$  and  $\Sigma_{1,2} = 0.8$ , 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:  $\hat{f}_1$ , trained only on  $X_1$ , and  $\hat{f}_2$ , trained on  $\mathbf{X}$ . Both models are fitted using a nearest-neighbor kernel CDE [Dalmaso et al., 2020] with hyperparameters chosen by data splitting: we use 10000 training, 5000 validation, and 200 test points.

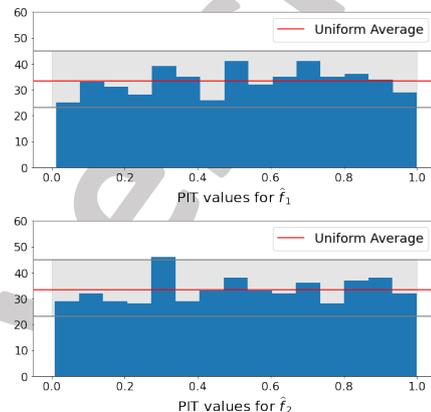


Figure 3: Standard diagnostics for Example 1 showing histograms of PIT values computed on 200 test points (with 95% confidence bands for a  $Unif[0,1]$  distribution). *Top*: Results for  $\hat{f}_1$ , which has only been fit to the first of two covariates. *Bottom*: Results for  $\hat{f}_2$ , which has been fit to both covariates. The top panel shows that standard PIT diagnostics cannot tell that  $\hat{f}_1$  is a poor approximation to  $f$ . GCT, on the other hand, detects that  $\hat{f}_1$  is misspecified ( $p=0.004$ ), while not rejecting the global null for  $\hat{f}_2$  ( $p=0.894$ ).

This is a toy 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 4) 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  $\hat{f}_1$  versus the model  $\hat{f}_2$  at any given  $\mathbf{x} \in \mathcal{X}$ ; see Figure 4 panels (c) and (d)). Despite the clear relationship between  $Y$  and  $X_2$ , both  $\hat{f}_1$  (which omits  $X_2$ ) and  $\hat{f}_2$  pass existing goodness-of-fit tests based on PIT (Figure 3). This result can be explained by Theorem 1: because PIT is insensitive to covariate transformations and  $\hat{f}_1(y|\mathbf{x}) \approx f(y|x_1)$ , PIT values are uniformly distributed, even though  $\hat{f}_1$  omits a key variable. The GCT, however, detects that  $\hat{f}_1$  is misspecified ( $p = 0.004$ ), while the global null (Equation 1) is not rejected for  $\hat{f}_2$  ( $p = 0.894$ ).

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 4 shows p-values from local cov-

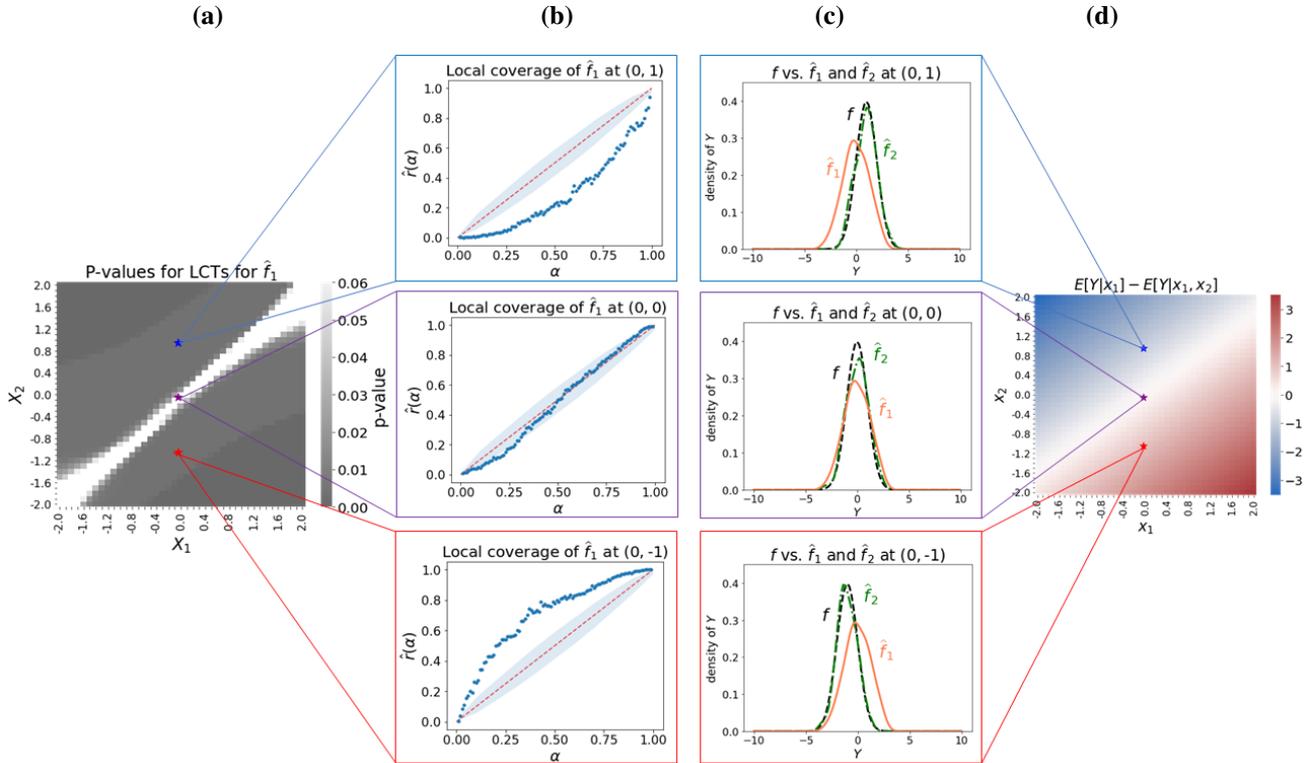


Figure 4: New diagnostics for Example 1. (a) P-values for LCTs for  $\hat{f}_1$  indicate a poor fit across most of the feature space. (b) Amortized local P-P plots at selected points show the density  $\hat{f}_1$  as negatively biased (blue), well estimated at significance level  $\alpha = 0.05$  with barely perceived overdispersion (purple), and positively biased (red). (Gray regions represent 95% confidence bands under the null.) (c)  $\hat{f}_1$  and  $\hat{f}_2$  vs. the true (unknown) conditional density  $f$  at the selected points.  $\hat{f}_1$  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.  $\hat{f}_2$  fits well at all three points. The difference on average in the predictions of  $Y$  from  $\hat{f}_1(\cdot|x)$  vs. the true distribution  $f(\cdot|x)$  for fixed  $x$  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.)

erage tests for  $\hat{f}_1$  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  $\hat{f}_1$  and the true conditional density  $f$  at a point  $x$  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 4 zooms in on a few different locations  $x$  with local P-P plots that depict and interpret distributional deviations. At the blue point,  $\hat{f}_1$  underestimates the true  $Y$ : we reject the local null (Equation 4), and the P-P plot indicates negative bias. Conversely, at the red point,  $\hat{f}_1$  overestimates the true  $Y$ ; we reject the local null, and the P-P plot indicates positive bias. At the purple point,  $\hat{f}_1$  is close to  $f$ , so the local null hypothesis is not rejected.

This toy 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.

## 5 EXAMPLE 2: CONDITIONAL NEURAL DENSITIES FOR GALAXY IMAGES

In this example of CDE in a prediction setting, 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 “good” and “bad” 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 [Schmidt et al., 2020], and our method is the first to properly address them.

Here,  $x$  represents a  $20 \times 20$ -pixel image of an elliptical galaxy generated by `GalSim`, an open-source toolkit for simulating realistic images of astronomical objects [Rowe

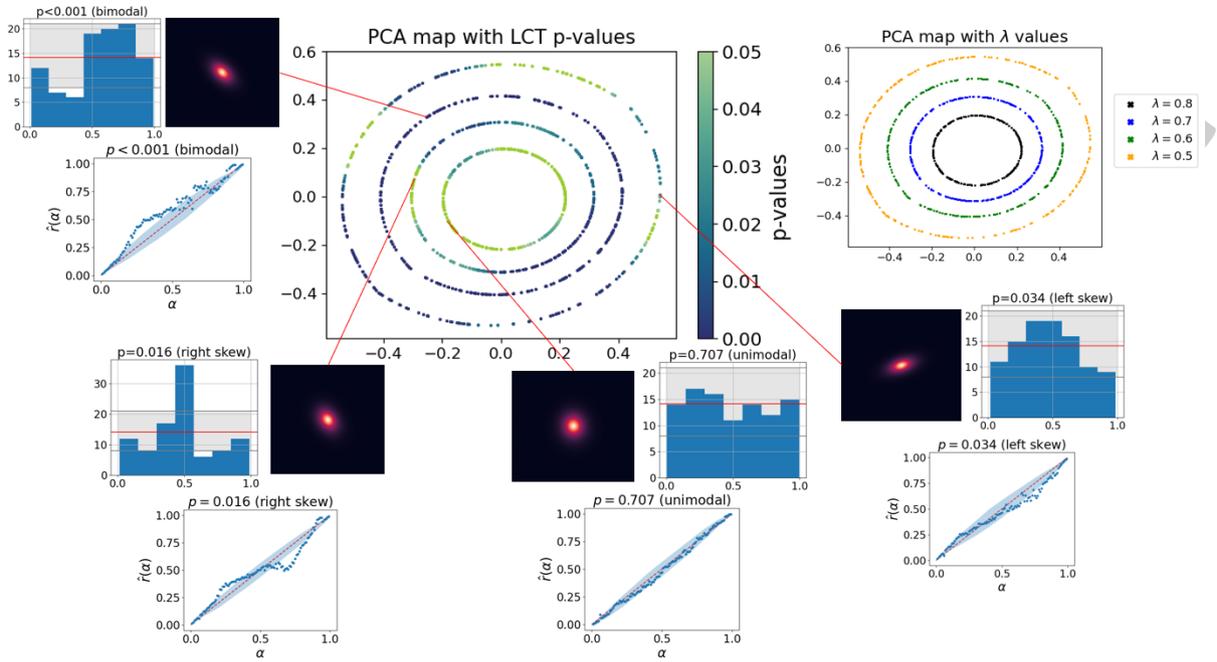


Figure 5: New diagnostics for Example 2. For visualization, we show the location of the test galaxy points in  $\mathbb{R}^{400}$  along the first two principal components (see center panel “PCA map with LCT p-values”). Test statistics from the LCTs indicate that the unimodal density model generally 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 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.

et al., 2015]. In *GalSim*, we can vary the axis ratio  $\lambda$ , 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:

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

See Figure 8 in Supp. Mat. D for a plot of these distributions.

For illustration, we fit a unimodal Gaussian neural density model to estimate the conditional density  $Z|\mathbf{X}$ . Our diagnostics 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 flaws of this CDE model. First, we perform the GCT which rejects the global null ( $p < 0.001$ ). Next, we turn to LCTs and P-P plots to explore where and how the fit is inadequate. Figure 5 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.

## 6 EXAMPLE 3: NEURAL POSTERIOR INFERENCE FOR GALAXY IMAGES

Our final example tests for image data  $\mathbf{x} \in \mathbb{R}^{400}$  whether a Bayesian posterior model  $\hat{f}(\theta|\mathbf{x})$  fits the true posterior. As in Example 2,  $\mathbf{x}$  represents an image of an elliptical galaxy generated by *GalSim*. As before,  $\lambda$  is the galaxy’s axis ratio, but now the quantity of interest  $\theta$  is the galaxy’s rotation angle with respect to the x-axis; that is, an unknown *internal* parameter. For illustration, we create a mixture of a larger population with  $\lambda = 0.7$  (spheroidal galaxies), and a smaller population with  $\lambda = 0.1$  (elongated galaxies). We then simulate a sample of images as follows: first, we draw

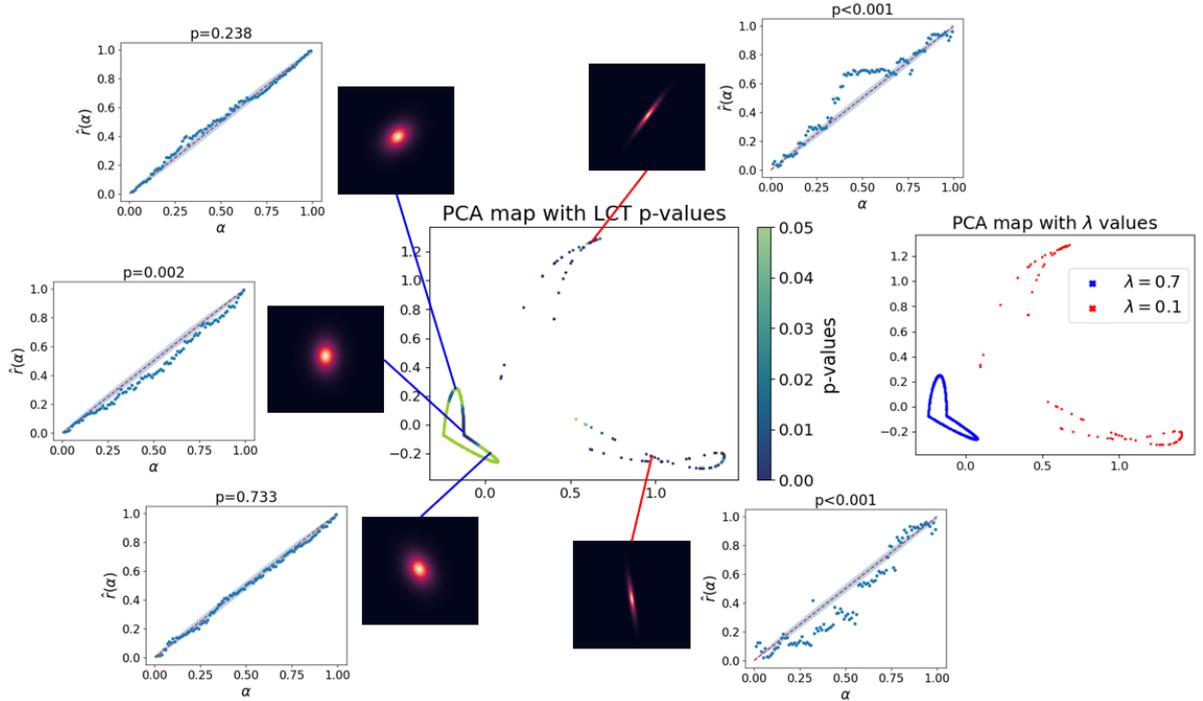


Figure 6: New diagnostics for simulation-based inference algorithm in Example 3. For visualization, we show the location of the test galaxy points in  $\mathbb{R}^{400}$  along its first two components (see center panel “PCA map with LCT p-values”). P-values for LCTs indicate that the ConvMDN generally fits well for the dominant 90% population of spheroidal galaxies ( $\lambda = 0.7$ ), while fitting poorly for the smaller 10% subpopulation of elongated galaxies ( $\lambda = 0.1$ ). Local P-P plots show statistically significant deviations in the CDEs (gray regions are 95% confidence bands under the null) for the latter population, suggesting we need better approximations of the posterior for this group.

$\lambda$  and  $\theta$  from a prior distribution given by

$$\mathbb{P}(\lambda = 0.7) = 1 - \mathbb{P}(\lambda = 0.1) = 0.9$$

$$\theta \sim \text{Unif}(-\pi, \pi)$$

Then we sample  $20 \times 20$  galaxy images  $\mathbf{X}$  according to the data model  $\mathbf{X}|\lambda, \theta \sim \text{GalSim}(a, \lambda)$ , where

$$a|\lambda = 0.7 \sim N(\theta, 0.05)$$

$$a|\lambda = 0.1 \sim 0.5\text{Laplace}(\theta, 0.05) + 0.5\text{Laplace}(\theta, 0.0005).$$

As in Example 2, we fit a convolutional mixture density network (ConvMDN); in this case, it gives us an estimate of the posterior distribution  $f(\theta|\mathbf{x})$ . This time, we allow  $K$ , the number of mixture components, to vary. According to the KL divergence loss computed on a separate test sample with 1000 images, the best fit of  $f(\theta|\mathbf{x})$  is achieved by a ConvMDN model with  $K = 7$  (see Table 1 in Supp. Mat. E). Here, the ConvMDN model with the smallest KL loss fails the GCT ( $p < 0.001$ ), so we turn to LCTs and P-P plots to understand why. Figure 6 plots the test galaxy images along their first two principal components. The LCTs show that the ConvMDN model generally fits the density well for the main population of spheroidal galaxies ( $\lambda = 0.7$ ), but fails to properly model the smaller population of elongated galaxies ( $\lambda = 0.1$ ). P-P plots at selected test points indicate severe

bias in the posterior estimates for the  $\lambda = 0.1$  population. These plots suggest that an effective way of obtaining a better approximation of the posterior is by improving the fit for the  $\lambda = 0.1$  population (by obtaining more data in that region of the feature space, using a different model class, etc). For instance, CDE models not based on mixtures [Papamakarios et al., 2019] could be more effective.

**Conclusion.** Conditional density models are widely used for uncertainty quantification in prediction and Bayesian inference. In this work, we offer practical procedures (GCT, LCT, ALP) for identifying, locating, and interpreting modes of failure for an approximation of the true conditional density. Our tools can be used in conjunction with loss functions, which are useful for performing model selection, but not good at evaluating whether a practitioner should keep looking for better models, or at providing information as to how a model could be improved. Finally, because LCT pinpoints hard-to-train regions of the feature space, our framework can provide guidance for active learning schemes.

**Acknowledgments.** This work is supported by NSF DMS-2053804 and NSF PHY-2020295. RI is grateful for the financial support of CNPq (309607/2020-5) and FAPESP (2019/11321-9).

## References

- D. W. K. Andrews. A conditional Kolmogorov test. *Econometrica*, 65(5):1097–1128, 1997.
- Rongmon Bordoloi, Simon J. Lilly, and Adam Amara. Photo-z performance for precision cosmology. *Monthly Notices of the Royal Astronomical Society*, 406(2):881–895, 08 2010. doi: 10.1111/j.1365-2966.2010.16765.x.
- Yanzhi Chen and Michael U. Gutmann. Adaptive gaussian copula ABC. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1584–1592. PMLR, 16–18 Apr 2019.
- Samantha R. Cook, Andrew Gelman, and Donald B. Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006.
- Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.
- Niccolò Dalmaso, Taylor Pospisil, Ann B. Lee, Rafael Izbicki, Peter E. Freeman, and Alex I. Malz. Conditional density estimation tools in Python and R with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, 30:100362, Jan 2020. ISSN 2213-1337. doi: 10.1016/j.ascom.2019.100362.
- Carlos Alberto de Bragança Pereira and Julio Michael Stern. Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy*, 1(4):99–110, 1999.
- Antonio D’Isanto and Kai Lars Polsterer. Photometric redshift estimation via deep learning. generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*, 609:A111, 2018.
- Vincent Dutordoir, Hugh Salimbeni, Marc Peter Deisenroth, and James Hensman. Gaussian process conditional density estimation. In *Advances in Neural Information Processing Systems 31*, Neural Information Processing Systems. Curran Associates, Inc., 2018.
- Peter E. Freeman, Rafael Izbicki, and Ann B. Lee. 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, 2017. doi: 10.1093/mnras/stx764.
- Fah F. Gan and Kenneth J. Koehler. Goodness-of-fit tests based on p-p probability plots. *Technometrics*, 32(3):289–303, 1990. doi: 10.1080/00401706.1990.10484682.
- Christian Genest and Louis-Paul Rivest. On the multivariate probability integral transformation. *Statistics & probability letters*, 53(4):391–399, 2001.
- Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. JMLR Workshop and Conference Proceedings.
- David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2404–2414, Long Beach, California, USA, 09–15 Jun 2019. PMLR.
- Diana Harrison, David Sutton, Pedro Carvalho, and Michael Hobson. Validation of Bayesian posterior distributions using a multidimensional Kolmogorov–Smirnov test. *Monthly Notices of the Royal Astronomical Society*, 451(3):2610–2624, 06 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv1110.
- Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.
- Rafael Izbicki and Ann B. Lee. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831, 2017.
- Rafael Izbicki, Ann B. Lee, and Peter E. Freeman. Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. *Annals of Applied Statistics*, 11(2):698–724, 2017.
- Rafael Izbicki, Ann B. Lee, and Taylor Pospisil. ABC–CDE: Toward Approximate Bayesian Computation With Complex High-Dimensional Data and Limited Simulations. *Journal of Computational and Graphical Statistics*, pages 1–20, 2019. doi: 10.1080/10618600.2018.1546594.
- Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 221–230. PMLR, 03–06 Aug 2020.
- Ilmun Kim, Ann B. Lee, and Jing Lei. Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305, 2019. doi: 10.1214/19-EJS1648. URL <https://doi.org/10.1214/19-EJS1648>.

- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Jan-Matthis Lueckmann, Pedro J. Gonçalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H. Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 1289–1299, Red Hook, NY, USA, 2017. Curran Associates Inc.
- Jean-Michel Marin, Louis Raynal, Pierre Pudlo, Mathieu Ribatet, and Christian Robert. ABC random forests for Bayesian parameter inference. *Bioinformatics (Oxford, England)*, 35, 05 2016. doi: 10.1093/bioinformatics/bty867.
- M. J. Moreira. A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027 – 1048, 2003.
- S. Mucesh, W. G. Hartley, A. Palmese, O. Lahav, L. White-way, A. F. L. Bluck, A. Alarcon, A. Amon, K. Bechtol, G. M. Bernstein, A. Carnero Rosell, M. Carrasco Kind, and DES Collaboration. A machine learning approach to galaxy properties: joint redshift–stellar mass probability distributions with random forest. *Monthly Notices of the Royal Astronomical Society*, 502(2):2770–2786, 01 2021. doi: 10.1093/mnras/stab164.
- George Papamakarios and Iain Murray. Fast  $\epsilon$ -free Inference of Simulation Models with Bayesian Conditional Density Estimation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
- George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, Red Hook, NY, USA, 2017. Curran Associates Inc.
- George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *22nd International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 837–848. PMLR, 2019.
- George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.
- Jonas Rothfuss, Fabio Ferreira, Simon Walther, and Maxim Ulrich. Conditional density estimation with neural networks: Best practices and benchmarks. *arXiv preprint arXiv:1903.00954*, 2019.
- Barnaby Rowe, Mike Jarvis, Rachel Mandelbaum, Gary M. Bernstein, James Bosch, Melanie Simet, Joshua E. Meyers, Tomasz Kacprzak, Reiko Nakajima, Joe Zuntz, et al. GALSIM: The modular galaxy image simulation toolkit. *Astronomy and Computing*, 10:121–150, 2015.
- S. J. Schmidt, A. I. Malz, J. Y. H. Soo, I. A. Almosallam, M. Brescia, S. Cavuoti, J. Cohen-Tanugi, et al. 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, 2020.
- Cosma Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press, 2021.
- Motoki Shiga, Voot Tangkaratt, and Masashi Sugiyama. Direct conditional probability density estimation with sparse feature selection. *Machine Learning*, 100(2):161–182, 2015. doi: 10.1007/s10994-014-5472-x.
- Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
- W. Stute and L. X. Zhu. Model checks for generalized linear models. *Scandinavian Journal of Statistics*, 29(3):535 – 545, 2002. ISSN 0303-6896.
- Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.
- Masayuki Tanaka, Jean Coupon, Bau-Ching Hsieh, Sogo Mineo, Atsushi J Nishizawa, Joshua Speagle, Hisanori Furusawa, Satoshi Miyazaki, and Hitoshi Murayama. Photometric redshifts for Hyper Suprime-Cam Subaru Strategic Program Data Release 1. *Publications of the Astronomical Society of Japan*, 70(SP1), 01 2018. doi: 10.1093/pasj/psx077.
- Bengio Uria, Iain Murray, and Hugo Larochelle. 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, 09–15 Jun 2014. JMLR.
- J. X. Zheng. A consistent test of conditional parametric distributions. *Econometric Theory*, 16(5):667 – 691, 2000.
- Johanna F. Ziegel and Tilmann Gneiting. Copula calibration. *Electronic Journal of Statistics*, 8(2):2619–2638, 2014. doi: 10.1214/14-EJS964.