
Probabilistic Conformal Prediction Using Conditional Random Samples

Zhendong Wang*

UT Austin

Department of Statistics and Data Sciences

zhendong.wang@utexas.edu

Ruijiang Gao*

UT Austin

McCombs School of Business

ruijiang@utexas.edu

Mingzhang Yin*

University of Florida

Warrington College of Business

m.yin@ufl.edu

Mingyuan Zhou

UT Austin

McCombs School of Business

mingyuan.zhou@mccombs.utexas.edu

David M. Blei

Columbia University

Department of Statistics and Computer Science

david.blei@columbia.edu

Abstract

This paper proposes probabilistic conformal prediction (PCP), a predictive inference algorithm that estimates a target variable by a discontinuous predictive set. Given inputs, PCP constructs the predictive set based on random samples from an estimated generative model. It is efficient and compatible with conditional generative models with either explicit or implicit density functions. We show that PCP guarantees correct marginal coverage with finite samples and give empirical evidence of conditional coverage. We study PCP on a variety of simulated and real datasets. Compared to existing conformal prediction methods, PCP provides sharper predictive sets.

1 INTRODUCTION

A core problem in supervised machine learning is to predict a target variable $Y \in \mathcal{Y}$ given a vector of inputs $X \in \mathbb{R}^p$. In this problem, a predictive function $q(Y | X)$ is fitted on an observed dataset $\mathbf{D} = \{(X_i, Y_i)\}_{i=1}^N$ and then used to predict the target Y_{N+1} of a new data point with inputs X_{N+1} . While much of machine learning focuses on point predictions of Y , the problem of predictive inference aims at more robust prediction. In predictive inference, our goal is to create a *predictive set* that is likely to contain the unobserved target (Geisser, 1993).

In particular, the field of *conformal prediction* develops pre-

dictive inference algorithms that aim for calibrated coverage probabilities (Papadopoulos et al., 2002; Vovk et al., 2005). Assume the data pairs (X_i, Y_i) are sampled independent and identically distributed (i.i.d) from a population distribution $\mathbb{P}(X, Y)$. Given an input X , a conformal prediction algorithm provides a set $C_\alpha(X)$ such that

$$\mathbb{P}_{X,Y}(Y \in \hat{C}_\alpha(X)) \geq 1 - \alpha. \quad (1)$$

The scalar $\alpha \in [0, 1]$ is a predefined miscoverage rate and $\hat{C}_\alpha(X) \subset \mathcal{Y}$ is the predictive set. A set that satisfies Eq. (1) is called a *valid* predictive set. Since the trivial set $\hat{C}_\alpha(X) = \mathcal{Y}$ is valid, a major desideratum is to keep the size of the predictive set small and thus informative. This property is known as *sharpness* (Lei et al., 2015). In this paper, we develop a new method for conformal prediction that produces valid and sharp predictive sets.

Existing conformal prediction methods often produce a continuous interval as the predictive set (Lei and Wasserman, 2014; Barber et al., 2019b; Romano et al., 2019; Sesia and Romano, 2021; Messoudi et al., 2021). Such intervals are appropriate in some predictive situations. However, consider a target distribution with separated high-density regions. In this setting, to ensure validity the set must include all of the high-density regions; but since it is continuous it must also include the low-density regions between them. In a nutshell, the continuity of the predictive sets posits a tension between the validity and sharpness.

For example, consider a prediction problem that estimates the drop-off location of a taxi passenger based on the passenger's information. The target distribution is likely to be multimodal, centered around locations such as tourist attractions and transit centers. A continuous predictive set have to encompass these regions, regardless of how far apart they are. A more informative set would contain the regions themselves, but not the areas between them. The multimodal target distribution is prevalent in practice such as the effects

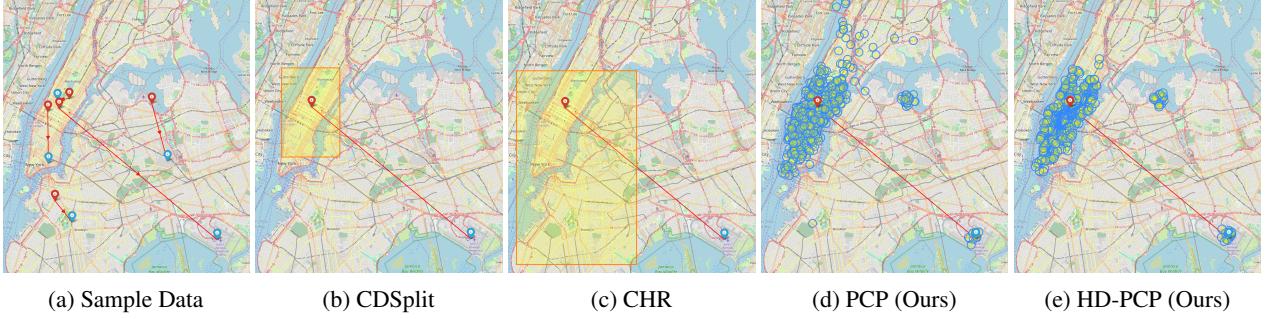


Figure 1: NYC Taxi data. The covariates are pickup location (red pin) and other passenger information; The target is the dropoff location (blue pin). Left to right: Five random samples from Travel Data; Predictive sets output by CDSplit (Izbicki et al., 2020), CHR (Sesia and Romano, 2021), PCP, and HD-PCP for one travel record.

of a stroke on brain regions (Gillmann et al., 2021) and the action rewards of a robot (Myers et al., 2022).

Fig. 1 provides an example of the taxi application. Panel (a) illustrates the data, the destinations of NYC taxi passengers. Given a new set of inputs, panels (b) and (c) show existing conformal prediction methods, which predict large regions for the possible destinations. Panels (d) and (e) show the results of our algorithms (*probabilistic conformal prediction* (PCP)), which form sharper predictive sets from distinct subregions of the map. Fig. 2 illustrates the algorithm.

In more detail, PCP builds on the split conformal prediction framework (Lei and Wasserman, 2014; Papadopoulos et al., 2002). It begins by randomly splitting the observed data \mathbf{D} into a preliminary set \mathbf{D}_{pre} and a calibration set \mathbf{D}_{cal} . It then has three stages. (1) It fits a conditional generative model $q(Y | X)$ to the preliminary data \mathbf{D}_{pre} . (2) For each point (X_i, Y_i) in the calibration set \mathbf{D}_{cal} , it generates K independent samples of predictions $\hat{\mathbf{Y}}_{X_i} = \{\hat{Y}_{i1}, \dots, \hat{Y}_{iK}\}$ from the fitted model $q(Y | X_i)$. It then calculates the distance between each sampled prediction and the true label Y_i . These quantities are called the *nonconformity scores* and measure the goodness-of-fit of the generative model. (3) Finally, it calculates and records the $(1 - \alpha)$ empirical quantile of the nonconformity scores. The quantile will be used to construct the predictive sets.

To form the predictive set of a new datapoint, first PCP generates prediction samples from the fitted target distribution. Then each sample is expanded to a ball that centers at the sample point and has a radius equal to the quantile computed from the calibration set. Finally, the predictive set is defined as the union of the balls over the samples. Because it is centered at high-density regions, this predictive set is sharp. Further, as we prove below, it is valid.

There are several advantages to PCP (and a related extension, high-density PCP). First, it adapts automatically to the landscape of the target distribution, providing sharp and valid predictive sets regardless of the underlying distribution. Second, the generative model for PCP may have an explicit

or implicit density function as long as it can generate random samples. Without requiring an explicit density, PCP is compatible with the likelihood-free prediction (Alsing et al., 2019; Chan et al., 2018) and is less prone to model misspecification (Mirza and Osindero, 2014). Last, (HD-)PCP can be applied to multi-target regression where the target variable $Y \in \mathbb{R}^T$, $T \geq 1$ (Breiman and Friedman, 1997; Messoudi et al., 2021). As we shall see, (HD-)PCP scales efficiently with the target dimension and creates a sharp predictive set by capturing the targets’ dependencies.

2 RELATED WORK

PCP provides a contribution to the growing field of conformal prediction. Some conformal prediction methods are based on predicting summary statistics of the target distribution, for example, by fitting a mean response function (Lei et al., 2018), conditional quantile functions (Romano et al., 2019) and approximate histograms (Sesia and Romano, 2021). However, these methods produce a single continuous interval as the predictive set, which might be too loose for multimodal targets.

Other conformal prediction methods estimate the full target distribution. Distributional conformal prediction (DCP) is based on the estimated cumulative density function (Chernozhukov et al., 2021) but its prediction is often sensitive to the tail estimation (Sesia and Romano, 2021). CDSplit uses a level set of the estimated probability density function as the predictive set (Izbicki et al., 2020). Similar to PCP, CDSplit can produce discontinuous predictive sets. However, the level set might be loose when the distribution has high dispersion and it has to be computed approximately. PCP is more computationally efficient than CDSplit, and further it is compatible with likelihood-free predictions due to the sampling-based design. Empirically, across multiple datasets, PCP creates sharper predictive sets than these existing conformal methods.

Finally, there are a few conformal methods for multi-target regression (Messoudi et al., 2020, 2021; Neeven and

Smirnov, 2018). Compared to these methods, PCP models the target variables jointly and can produce discontinuous predictive sets. As we show in the empirical studies, PCP provides sharper and more interpretable predictions.

3 PROBABILISTIC CONFORMAL PREDICTION

Problem setup Consider i.i.d. pairs of covariates X_i and a target variable Y_i , *i.e.*, $\mathbf{D} = \{(X_i, Y_i)\}_{i=1}^N$, from a distribution $P(X, Y)$. We observe data \mathbf{D} and the covariates X_{N+1} of a new data point. To form a predictive set $\hat{C}(X_{N+1})$ for the unobserved target Y_{N+1} with valid uncertainty estimation, we create a predictive set $\hat{C}_\alpha(\cdot) : \mathcal{X} \mapsto \mathcal{Y}$ that satisfies Eq. (1) for $\alpha \in [0, 1]$. Since an arbitrary wide predictive set has valid coverage, a predictive set should be as sharp as possible. Classic conformal prediction is based on leave-one-out estimation (Vovk et al., 2005), which has high computational cost due to multiple model fitting. In this paper, we adopt the split conformal prediction framework, which improves computational efficiency by data-splitting (Lei et al., 2018; Papadopoulos et al., 2002). It randomly splits the observed data to a preliminary set and a calibration set. The model is fitted on the preliminary set and kept fixed in computing the nonconformity scores on the calibration set and the test set.

3.1 Generative model fitting

The proposed PCP depends on random samples from a conditional generative model (CGM) $q(Y|X)$ that approximates the target variable distribution $p(Y|X)$. This differs from standard conformal prediction methods that are based on fitting the summary statistics such as the conditional mean and quantiles of the target (Lei and Wasserman, 2014; Romano et al., 2019) and that depend on evaluating probability densities (Izbicki et al., 2020; Chernozhukov et al., 2021; Hoff, 2021). Since the only requirement is to sample from $q(Y|X)$, we consider both typical CGMs with explicit density functions and popular CGMs with implicit density. In particular, we consider PCP with explicit models such as Kernel Mixture Network (KMN) (Ambrogioni et al., 2017), Mixture Density Network (MixD) (Bishop, 1994), and implicit models such as GAN (Mirza and Osindero, 2014). See Section D for more details about CGMs used. We regard CGMs as backbone models for PCP.

3.2 Uncertainty calibration with random samples

Suppose a conditional density model is fitted on a preliminary dataset \mathbf{D}_{pre} . We use the fitted model $q(Y|X)$ and the calibration data to construct a predictive set for a new test data point. For a data point (X_i, Y_i) in the calibration set, the algorithm first generates K random samples \hat{Y}_{ik} , $k = 1, \dots, K$ independently from $q(Y|X_i)$, denoted as

$\hat{\mathbf{Y}}_i = \{\hat{Y}_{i1}, \dots, \hat{Y}_{iK}\}$. Then, it computes the distance from the observed outcome to this set of samples as

$$E_i = \min_{1 \leq k \leq K} \|Y_i - \hat{Y}_{ik}\|. \quad (2)$$

The scalar E_i is set as the nonconformity score. The norm $\|\cdot\|$ in Eq. (2) is user-specified. For the regression problems considered in this paper, we choose the Euclidean norm.

The score in Eq. (2) adopts the standard notation of distance between a point and a set. Intuitively, a small score indicates that the speculated outcomes \hat{Y}_{ik} are close to the observed outcome Y_i , where \hat{Y}_{ik} are from the approximate density $q(Y|X_i)$ and Y_i is from the true underlying density $p(Y|X_i)$. We use the empirical quantile of the nonconformity scores from the calibration data to construct the predictive set. The α -th empirical quantile is defined as $Q_\alpha(E_{1:n}) = \inf_E \{(\sum_{i=1}^n \mathbb{1}[E_i \leq E]) / n \geq \alpha\}$ where $\alpha \in [0, 1]$ and $\mathbb{1}[\cdot]$ is the indicator function.

For a new data point with covariates X , we generate $\hat{\mathbf{Y}} = \{\hat{Y}_1, \dots, \hat{Y}_K\}$ with $\hat{Y}_k \sim q(Y|X)$. Suppose that the desired nominal coverage is $1 - \alpha$. Then, each sample \hat{Y}_k is expanded to a region $R_k = \{y : \|y - \hat{Y}_k\| \leq r\}$ with $r = Q_{1-\alpha}(\{E_1, \dots, E_n\} \cup \{\infty\})$. We call R_k an element region of the data point X . The proposed predictive set is the union of the element regions,

$$\begin{aligned} \hat{C}(X, \hat{\mathbf{Y}}) &= \bigcup_{k=1}^K R_k \\ &= \bigcup_{k=1}^K \{y : \|y - \hat{Y}_k\| \leq Q_{1-\alpha}(E_{1:n} \cup \{\infty\})\}. \end{aligned} \quad (3)$$

As a special case, when the outcome is a scalar, the predictive set can be written explicitly as

$$\begin{aligned} \hat{C}(X, \hat{\mathbf{Y}}) &= \bigcup_{k=1}^K \left[\hat{Y}_k - Q_{1-\alpha}(E_{1:n} \cup \{\infty\}), \right. \\ &\quad \left. \hat{Y}_k + Q_{1-\alpha}(E_{1:n} \cup \{\infty\}) \right]. \end{aligned} \quad (4)$$

The proposed PCP algorithm is summarized in Algorithm 1.

3.3 Properties of the nonconformity score

The nonconformity score in Eq. (2) measures the local conformity with several desired properties. First, the minimization operator in Eq. (2) leads to high conformity when one of the generated samples is close to the observed target, regardless of the positions of other samples. Considering a multimodal target distribution, some samples can capture the modes that the observed Y is not at but do not increase the nonconformity score. The points at the low density regions between the modes would have high nonconformity scores and hence are not included in the predictive set. This is in contrast to other ways of score design such as the mean distance between $\{\hat{Y}_{ik}\}$ and Y_i . Second, the score in Eq. (2) reflects the goodness-of-fit. Consider a CGM with high density and a CGM with low density at the region where

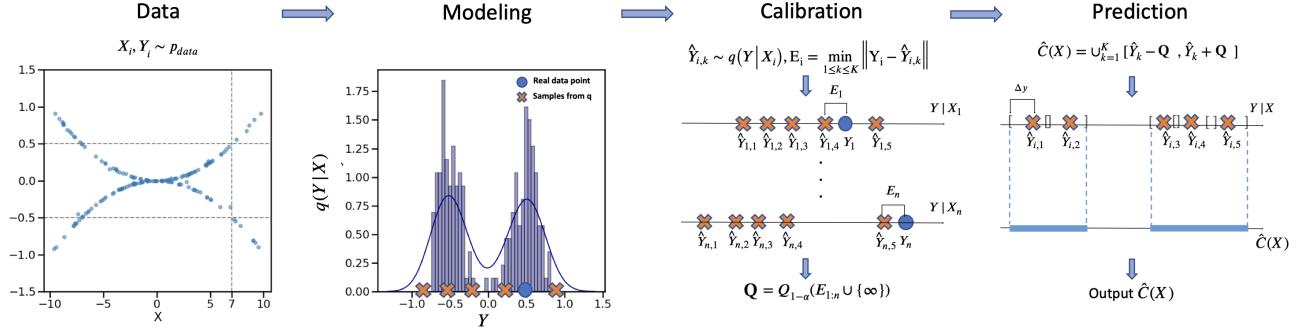


Figure 2: Illustration of PCP. Data: i.i.d data $\mathbf{D} = \{(X_i, Y_i)\}_{i=1}^N$; Modeling: K random samples generated from a fitted $q(Y|X)$; Calibration: compute scores E_i and the quantile Δy ; Prediction: construct the predictive set $\hat{C}(X)$ for a test data.

$p(Y|X)$ is high. Though at a first glance, for a *specific* data point, both CGMs might produce a set of samples leading to similar scores. However, the quantile $Q_{1-\alpha}(E_{1:n} \cup \{\infty\})$ is computed over multiple points. The CGM with a better fit has more samples around the observed target, thus producing a smaller score in expectation due to the minimization operator, a lower $Q_{1-\alpha}$, and finally a sharper predictive set.

As shown in Eq. (3), the predictive set can be either continuous or discontinuous. Therefore, it can produce a sharp estimate by automatically adapting to the target distribution. When the generative model is not fitted well, PCP maintains a valid marginal coverage, properly quantifying the predictive uncertainty. When the generative model fits well, the predictive set allocates its volume according to the random samples. For example, if $p(Y|X)$ is multimodal and the multimodality is captured by the estimated $q(Y|X)$, the predictive set would consist of discontinuous sets around the modes where each set is relatively small.

Though in some situations, a continuous interval prediction is preferred in terms of interpretability (Sesia and Romano, 2021), when the target is multimodal, a discontinuous set might be more interpretable. For example, when predicting a watch price based on its appearance without knowing the brand, a price range $(\$100, \$200) \cup (\$1000, \$1200)$ might be more informative than $(\$100, \$1200)$. Nevertheless, one can take the convex hull of a discontinuous set to form a continuous interval but not vice versa.

By the construction of the predictive set in Eq. (3), the estimated density $q(Y|X)$ can be explicit or implicit, allowing for a wide range of CGMs. Moreover, the predictive set in Eq. (3) can be computed without approximation, making PCP scalable to a high dimensional target Y .

3.4 Statistical guarantees

PCP has a guaranteed marginal coverage as in Theorem 1.

Theorem 1. Suppose (X_i, Y_i) , $i \in \{1, 2, \dots, n\}$ and (X, Y) are exchangeable, then

Algorithm 1 Probabilistic Conformal Prediction

Input: Data $\mathbf{D} = \{(X_i, Y_i)\}_{i=1}^N$, model $q(Y|X)$, nominal level α , test point X , sample size K .

Step I: Conditional generative model

- 1: Split the data into three folds \mathcal{Z}_{tr} , \mathcal{Z}_{val} , \mathcal{Z}_{cal} with set of index as \mathbf{I}_{tr} , \mathbf{I}_{val} , \mathbf{I}_{cal} respectively
- 2: Fit $q(Y|X)$ on \mathcal{Z}_{tr} with hyper-parameter chosen by cross validation on \mathcal{Z}_{val}

Step II: Predictive set for a test point

- 1: For $i \in \mathbf{I}_{\text{cal}}$, sample $\hat{Y}_{i1}, \dots, \hat{Y}_{iK} \sim q(Y|X_i)$
- 2: For test point X , sample $\hat{Y}_1, \dots, \hat{Y}_n \sim q(Y|X)$
- 3: Compute nonconformity score $\{E_i\}_{i \in \mathbf{I}_{\text{cal}}}$ by Eq. (2), $E_{N+1} = \infty$, $\tilde{\mathbf{E}}_{\text{cal}} = \mathbf{I}_{\text{cal}} \cup \{N+1\}$
- 4: Set r as the $(1 - \alpha)$ empirical quantile of $\{E_i\}_{i \in \tilde{\mathbf{E}}_{\text{cal}}}$
- 5: Compute the predictive set $\hat{C}(X, \hat{Y})$ by Eq. (3)

Output: Predictive set $\hat{C}(X, \hat{Y})$

(1) the predictive set in Eq. (3) satisfies

$$\mathbb{P}_{X, Y, \hat{Y}}(Y \in \hat{C}(X, \hat{Y})) \geq 1 - \alpha; \quad (5)$$

(2) when the scores E_1, \dots, E_n are distinct almost surely,

$$\mathbb{P}_{X, Y, \hat{Y}}(Y \in \hat{C}(X, \hat{Y})) \leq 1 - \alpha + \frac{1}{n+1}. \quad (6)$$

Theorem 1 demonstrates that the marginal coverage of PCP is valid and tight. In particular, the condition of the upper bound is satisfied when $p(Y|X)$ is continuous in the regression setting. Note that the coverage probability in Theorem 1 is with respect to both (X, Y) and \hat{Y} , which requires new adaptations of the standard conformal inference proof. The details are in Section A.

In practice, we take the quantile of $E_{1:n}$ instead of the inflated scores $E_{1:n} \cup \{\infty\}$ in Eq. (3). The following corollary offers the coverage guarantee under such modification.

Corollary 1. Under the conditions of Theorem 1 and suppose $\alpha \geq 1/(n+1)$, if the quantile in Eq. (3) is $Q_{1-\alpha}(E_{1:n})$,

then $\mathbb{P}(Y \in \hat{C}(X, \hat{Y})) \in [1 - \alpha - 1/(n+1), 1 - \alpha + 1/(n+1)]$; if the quantile in Eq. (3) is $Q_{(1-\alpha)(1+\frac{1}{n})}(E_{1:n})$, then $\mathbb{P}(Y \in \hat{C}(X, \hat{Y})) \in [1 - \alpha, 1 - \alpha + 1/(n+1)]$

Conditional coverage. The guarantees in Theorem 1 is not for a fixed X_i . The miss of conditional coverage has been shown as a consequence of the distribution-free assumption (Barber et al., 2019a). Some existing work establish conditional coverage under strong assumptions such as unimodal and bounded target distribution (Sesia and Romano, 2021), asymptotic assumption with $n \rightarrow \infty$, and the consistency of the predictive model (Chernozhukov et al., 2021; Sesia and Romano, 2021). In this paper, instead of invoking strong assumptions, we empirically evaluate the conditional coverage of PCP with metrics developed and adopted by a variety of conformal prediction studies (Romano et al., 2020; Gibbs and Candès, 2022; Taufiq et al., 2022; Xu and Xie, 2022).

3.5 High Density Probabilistic Conformal Prediction

Ideally, we may want the predictive sets to contain only high density regions to offer interpretable predictions. As shown in Section B, for different sets with the same coverage probability under a multimodal distribution, the high density region has the smallest size.

In PCP, the generated random samples include low density samples. This may lead to many isolated sets and make interpretation difficult. To mitigate this problem, we propose High Density Probabilistic Conformal Prediction (HD-PCP) to filter out β fraction low-density samples to identify the high density regions when $q(Y|X)$ is explicit. Instead of sampling K samples from $q(Y|X)$ like in PCP, we keep $(1 - \beta)$ fraction of K samples with the highest estimated density. The HD-PCP algorithm is summarized in Section B. The marginal coverage guarantee still holds for HD-PCP.

Corollary 2. Under the conditions of Theorem 1, HD-PCP has the same marginal coverage as PCP.

4 EXPERIMENTS

In this section, we conduct a comprehensive analysis demonstrating the advantages of PCP compared to previously proposed conformal prediction methods. We aim to answer the following *questions*: **(a)** how does PCP perform in terms of coverage and predictive set size when compared with baseline models on synthetic datasets? **(b)** Does the filtering improve the predictive set of HD-PCP? **(c)** How well do PCP and HD-PCP perform on real datasets with a single target? **(d)** How do the backbone models impact the performance of PCP? **(e)** Does PCP provide better predictive sets in tasks with multi-dimensional targets?

We first conduct experiment on classic 2D synthetic data to answer question **(a)** and **(b)**. Then, we compare PCP

and HD-PCP with a full set of baseline methods on several selected real datasets to address question **(b)**, **(c)** and **(d)**. Finally, we conduct experiments on multi-dimensional regression tasks to address question **(e)**.

Baselines. We consider CHR (Sesia and Romano, 2021), DistSplit (Izbicki et al., 2020), CDSplit (Izbicki et al., 2020), DCP (Chernozhukov et al., 2021), and CQR (Romano et al., 2019) as our comparison baselines. For CHR, we use two different conditional density estimation models based on neural networks and random forest, and we denote them as CHR-NN and CHR-QRF. We evaluate all baselines with their public implementations except for CDSplit. We implement a python-based CDSplit based on the official R code to use the same backbone generative model for a fair comparison, denoted as CDSplit-KMN and CDSplit-MixD.

Choosing the hyperparameter K. We conduct an ablation study on the effect of the sample size K of PCP. As shown in Fig. 5, empirically we find when K increases, the average size of the predictive sets first reduces fast and then reduces slow. In practice, we set K moderately large to balance the sharpness and the computational cost, *i.e.*, $K = 40$ or $K = 1000$ (two-dimensional targets). $K = 1000$ is chosen since it is at the same magnitude as 40^2 .

4.1 Synthetic data experiments

To evaluate the effectiveness of the proposed methods, we compare the predictive set of PCP and HD-PCP with other baseline methods on synthetic data. We show the evaluation results of the s-curve and the 25-Gaussians in Fig. 3 and place detailed results in Appendix E.

Fig. 3 illustrates that when the dataset has multimodal $p(Y|X)$ distribution, models considering multimodality, such as CDSplit and (HD-)PCP, work apparently better than the models that can only provide unimodal predictions (question **(a)**). Quantitatively, all models achieve the target marginal coverage $(1 - \alpha)$, while the average set sizes of CDSplit and (HD-)PCP are several times smaller than that from CHR. CDSplit and PCP both can provide sharp and informative predictive sets for these multimodal datasets and PCP is slightly better with respect to the set size. The right two panels show the effect of filtering high density samples. The predictive sets by HD-PCP become cleaner and concentrated on the correct modes. Correspondingly, the average set size of HD-PCP is smaller than PCP. The histogram moving from blue bins to orange bins also shows the effectiveness of the filtering (question **(b)**).

4.2 Real data experiments

We study regression tasks on several real datasets to evaluate PCP and HD-PCP. We consider multiple types of generative models $q(Y | X)$, including implicit models (GAN Goodfellow et al. (2014)) and semi-implicit model (Yin and Zhou,

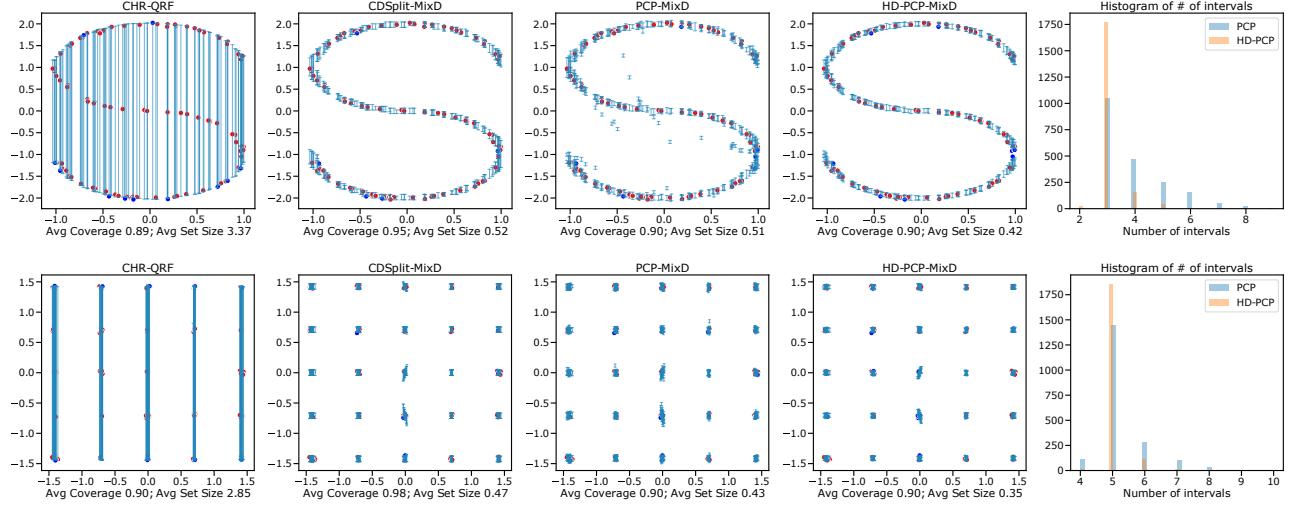


Figure 3: Visualization of predictive sets ($\alpha = 0.1$) on 2D toy datasets: s-curve and 25-Gaussians. We show the predictive sets on 100 test data samples. Blues lines: the predictive sets from each method; Blue dots: test points that are not covered by the predictive sets; Reds dots: test points covered. We report the marginal coverage and the average set size across test datapoints in the x-axis label. The fifth column shows the histogram of the number of predicted intervals of PCP and HD-PCP. We set $K = 40$ for (HD-)PCP, $\beta = 0.2$. Detailed experiments are in Section E.

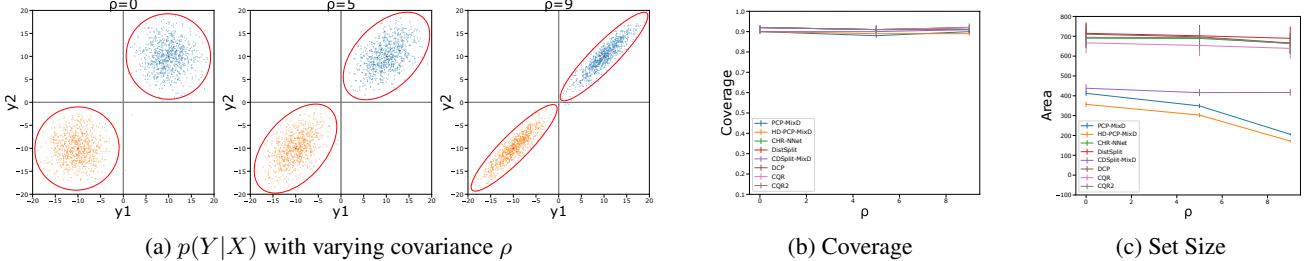


Figure 4: (a): Conditional data distribution $p(Y|X)$ for multi-target synthetic dataset. (b), (c): Marginal coverage and set size for baselines. Though all methods achieve similar coverage in (b), PCP produces the smallest set size when the covariance ρ increases as shown in (c).

2018a,b), explicit models (KMN (Ambrogioni et al., 2017), MixD (Bishop, 1994), and QRF (Meinshausen, 2006)). We denote them as PCP-GAN, PCP-SIVI, PCP-KMN, PCP-MixD and PCP-QRF respectively.

Datasets. We conduct real data experiments on 9 public-domain datasets: bike sharing data (bike), physicochemical properties of protein (bio), blog feedback (blog), and Facebook comment volume, variants one (fb1) and two (fb2), medical expenditure panel survey number 19 (meps19), number 20 (meps20), and number 21 (meps21) (Romano et al., 2019) and temperature forecast data (Cho et al., 2020). See Section F for detailed data statistics.

Evaluation Protocol. We compute the marginal coverage, conditional coverage (approximated by the worst-slab coverage (Cauchois et al., 2021; Romano et al., 2020)), and the predictive set size for all datasets. We report results based on 50 random splits for all datasets.

Table 1 shows numerical results. For our methods, we report PCP-MixD and HD-PCP-MixD; for baselines, we report the backbone model that works generally the best across the 9 datasets with respect to the set size in the main paper. See detailed results in Section F: Table 1 reports the best results among the variants of each method in terms of predictive sets ; Table 5, Table 6 and Table 7 report full experiment results.

We observe that all conformal methods achieve $(1 - \alpha)$ marginal coverage. Thus, our comparison focuses on the size of predictive sets. As shown in Table 1, HD-PCP-MixD outperforms all the other baselines on 7 out of 9 datasets in terms of the predictive set size. If choosing an optimal backbone model for each dataset, our methods outperform baselines on all datasets. Comparing HD-PCP with PCP, we find that the filtering technique brings consistent performance improvement. Table 1 shows that PCP outperforms the baselines by a large margin, especially on blog, face-

Table 1: Summary results of real data experiments, where Marg. C and Cond. C denotes the marginal coverage and approximated conditional coverage. The results are averaged over 50 random cross-validation splits. We report the set size mean and standard error (inside the parentheses, the default is 0.00) based on the same 50 splits. The nominal coverage rate ($1 - \alpha$) is 90%, the K for (HD-)PCP is set as 40. To save space and keep consistency, here we report PCP-MixD, HD-PCP-MixD, CHR-QRF, CDSplit-MixD and CQR, which include the variant that works generally the best across the 9 datasets. The detailed results are placed in Section F.

Data	Metric	PCP (ours)	HD-PCP (ours)	CHR	DistSplit	CDSplit	DCP	CQR
bike	Marg. C	0.90	0.90	0.90	0.90	0.92	0.90	0.90
	Cond. C	0.86	0.88	0.88	0.87	0.91	0.88	0.89
	Set Size	128.13(0.53)	102.92(0.48)	204.10(1.03)	423.13(1.51)	115.74(0.50)	443.76(1.36)	403.88(0.86)
bio	Marg. C	0.90	0.90	0.90	0.90	0.90	0.90	0.90
	Cond. C	0.89	0.90	0.89	0.89	0.90	0.89	0.89
	Set Size	11.47(0.04)	10.06(0.05)	10.21(0.04)	13.19(0.04)	9.58(0.04)	12.95(0.04)	13.00(0.02)
blog	Marg. C	0.89	0.90	0.90	0.90	0.96	0.90	0.90
	Cond. C	0.85	0.87	0.87	0.87	0.95	0.88	0.87
	Set Size	10.78(0.17)	9.44(0.19)	10.81(0.17)	16.27(0.23)	39.00(0.40)	1422.36(0.03)	15.15(0.26)
facebook1	Marg. C	0.90	0.90	0.90	0.90	0.95	0.90	0.90
	Cond. C	0.82	0.85	0.86	0.89	0.95	0.89	0.88
	Set Size	9.99(0.14)	8.93(0.12)	11.21(0.12)	14.03(0.16)	33.69(0.16)	1303.01(0.04)	13.79(0.15)
facebook2	Marg. C	0.90	0.90	0.90	0.90	0.97	0.90	0.90
	Cond. C	0.82	0.84	0.87	0.89	0.96	0.89	0.89
	Set Size	9.93(0.11)	8.84(0.10)	10.81(0.14)	13.48(0.19)	45.75(0.16)	1963.68(0.03)	13.00(0.17)
meps19	Marg. C	0.90	0.90	0.90	0.90	0.93	0.90	0.90
	Cond. C	0.87	0.88	0.89	0.89	0.92	0.88	0.89
	Set Size	19.28(0.16)	17.78(0.18)	18.26(0.15)	29.96(0.28)	23.86(0.17)	559.23(0.01)	28.71(0.18)
meps20	Marg. C	0.90	0.90	0.90	0.90	0.92	0.90	0.90
	Cond. C	0.87	0.88	0.90	0.90	0.92	0.88	0.90
	Set Size	19.52(0.16)	18.19(0.17)	17.94(0.18)	29.35(0.23)	22.93(0.16)	520.25(0.01)	27.57(0.15)
meps21	Marg. C	0.90	0.90	0.90	0.90	0.93	0.90	0.90
	Cond. C	0.87	0.88	0.90	0.89	0.92	0.88	0.89
	Set Size	19.18(0.12)	17.91(0.15)	18.65(0.16)	30.32(0.31)	23.63(0.17)	531.25(0.01)	29.89(0.20)
temperature	Marg. C	0.90	0.90	0.90	0.90	0.92	0.90	0.90
	Cond. C	0.90	0.89	0.89	0.89	0.91	0.88	0.87
	Set Size	2.10(0.01)	1.85(0.01)	3.24(0.01)	3.07(0.01)	2.23(0.01)	3.10(0.02)	3.55(0.03)

Table 2: Summary results of Multi-Target Regression experiments, where Marg. C and Cond. C denote the marginal coverage and approximated conditional coverage. Results are averaged over 10 random cross-validation splits. We report the standard error inside the parentheses. The nominal coverage rate ($1 - \alpha$) is 90%, the K for (HD-)PCP is set as 1000.

Data	Metric	PCP (ours)	HD-PCP (ours)	CHR	DistSplit	CDSplit	DCP	CQR
Taxi	Marg. C	0.90(0.01)	0.89(0.01)	0.87(0.01)	0.91(0.01)	0.91(0.01)	0.89(0.01)	0.90(0.01)
	Cond. C	0.89(0.03)	0.87(0.02)	0.84(0.04)	0.92(0.03)	0.89(0.03)	0.84(0.03)	0.86(0.04)
	Set Size	0.0089(0.0001)	0.0064(0.0002)	0.0245(0.0009)	0.0202(0.0009)	0.0097(0.0023)	0.0354(0.0013)	1.9302(0.5202)
Energy	Marg. C	0.89(0.01)	0.90(0.01)	0.93(0.01)	0.92(0.01)	0.93(0.01)	0.93(0.01)	0.91(0.01)
	Cond. C	0.87(0.04)	0.94(0.03)	0.93(0.05)	0.83(0.10)	0.89(0.06)	0.97(0.02)	0.94(0.03)
	Set Size	19.22(2.55)	14.13(3.55)	27.41(2.65)	36.31(3.04)	36.88(3.94)	34.18(3.61)	45.20(3.19)

book1 and facebook2 datasets (question **(c)**).

Moreover, note that PCP can work with any CGMs. The flexibility of PCP makes it achieve good performance by choosing a proper generative model according to datasets. For example, PCP-SIVI works well in bike and facebook data with implicit generative models, as shown in Table 5 (question **(d)**).

The limitation of CDSplit may be because the method needs to make partitions of data based on K-means algorithm, which is known to be unstable due to local minima. It also needs to approximate the level set on a grid of the target space to form the predictive set. It may be sensitive to the range and coarseness of the grid. Thus, we notice that CDSplit produces large predictive sets on facebook data and

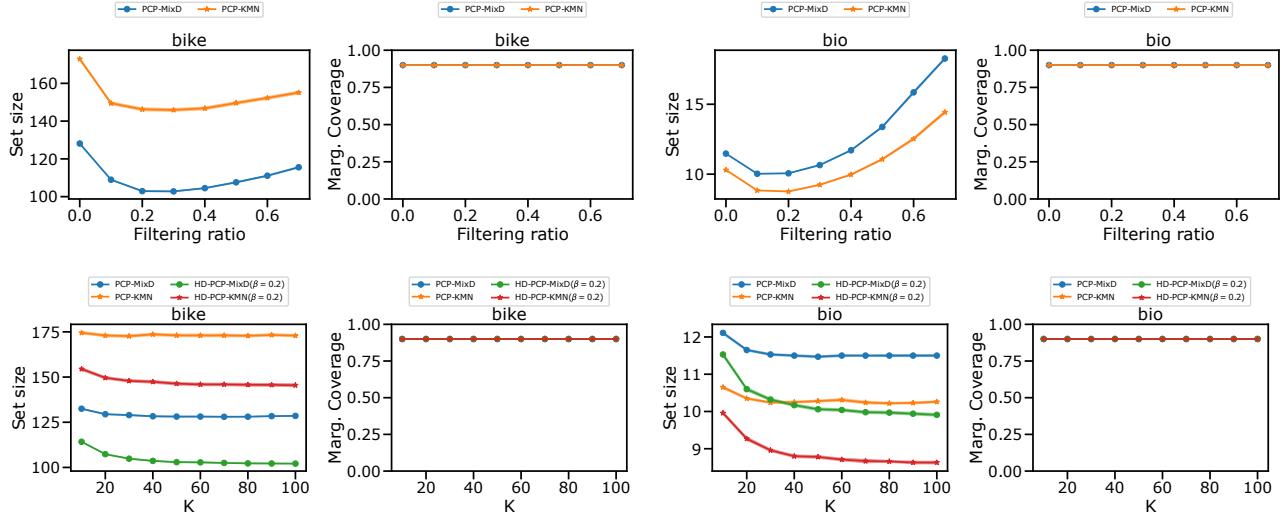


Figure 5: Ablation study on choosing hyperparameter K and filtering ratio β . We run experiments of PCP and HD-PCP with MixD and KMN as backbone models on two datasets, bike and bio. The K grid is $[10, 20, 30, \dots, 100]$ and the β grid is $[0.1, 0.2, \dots, 0.7]$. In the first row, we show predictive set sizes with $K = 50$ and varying β . Second row shows how predictive set size varies with K .

meps data. Similar to Sesia and Romano (2021), we observe that DCP is sensitive to the estimation of the distribution tails, which makes it unstable for some datasets. CHR is more robust but it could only provide a single continuous interval and produces a loose predictive set when the data exhibits multimodality. CQR predicts intervals based on the learned lower and upper quantiles, which leads to large intervals when the data distribution is multimodal.

4.3 Multi-dimensional targets

We further study PCP and HD-PCP on multi-target datasets. We adapt previous baselines to multi-target conformal algorithms by fitting each dimension separately with coverage level $(1 - \alpha)/d$ (Neeven and Smirnov, 2018), where d is the dimension of target Y (this ensures the coverage of the target vector is $1 - \alpha$ (Lei and Candès, 2021)).

We construct a synthetic dataset to illustrate the benefit of PCP that models the targets dependently. Covariates $X \sim \mathcal{N}(0, I_5)$ and the target $Y \sim 0.5\mathcal{N}(\mu_1, \Sigma) + 0.5\mathcal{N}(\mu_2, \Sigma)$. $\mu_i = (x^\top, 1)\beta_i$ and $\beta_i \sim \mathcal{N}(0, I_6)$, $\Sigma = \begin{pmatrix} 10 & \rho \\ \rho & 10 \end{pmatrix}$. The synthetic data distribution in Figure 4a shows that the distribution concentrates as ρ increases. As shown in Fig. 4b and Fig. 4c, PCP achieves the best performance in terms of average predictive set size. We observe when ρ gets higher, only PCP shrinks the predictive set accordingly while the predictions from other methods have little change and become loose. The detailed results are in Table 8, Section E.

We further study two multi-target real datasets. Taxi Data are the taxi trip records of NYC. For energy dataset,

we predict the heating load and cooling load for energy efficiency analysis (Tsanas and Xifara, 2012). To calculate the predictive set size of PCP, due to the overlapped regions, the set size cannot be calculated exactly. We estimate it by Monte Carlo simulation with a grid size of 100 on each dimension. For (HD-)PCP and CDSplit, we use MixD as the backbone model.

As shown in Table 2, algorithms considering multimodality have significantly better performance compared to other baselines. We visualize the conformal region predicted by (HD-)PCP in Fig. 1 (See Section G for more results). PCP can capture the most popular regions of New York city for drop-off such as downtown of Manhattan, LaGuardia airport and JFK airport, while methods with continuous predictive sets would learn a wide bounding box. Furthermore, PCP has smaller predictive set than CDSplit, potentially because the joint estimation of the high-dimensional targets can capture dependencies between the target elements. Applying the filtering, HD-PCP provides a cleaner and more interpretable predictive set and further reduces the set size (question (e)).

We include more experiments with two other multi-target real datasets in Section H (8 and 3 targets). PCP and HD-PCP consistently offer significantly sharper predictive sets compared to baselines, and HD-PCP can further improve the performance of PCP.

4.4 MNIST experiments

We further evaluate PCP on high dimensional data, MNIST dataset (Deng, 2012). We first fit a VAE model on handwrit-

ten digit images of (3, 5) and (1, 7) respectively. The upper half of the digits is masked as the input for the VAE. Then, we build our PCP onto the $14 \times 28 = 392$ dimensional target. As shown in Fig. 6, PCP could produce reasonable predictive sets, which not only cover the true data samples but also covers both modes and the high density regions around them, *e.g.* PCP covers both 3 and 5 in its predictive set. See Section I for details and additional experiment.

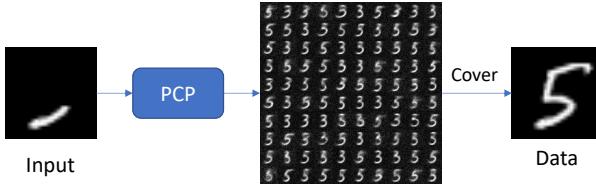


Figure 6: MNIST data. We show that PCP could produce proper predictive sets for high dimensional image data.

5 DISCUSSION AND FUTURE WORK

We proposed PCP and HD-PCP as new conformal prediction algorithms that find valid and sharp predictive sets using random samples from a conditional generative model. PCP and HD-PCP outperform existing methods for the sharpness of the predictive set, particularly with multimodal data and multi-dimensional targets.

There are several limitations that we consider as future work. First, we focus on the regression tasks here. Future research might adapt (HD-)PCP to causal effect estimation and classification problems (Gao et al., 2021; Biggs et al., 2021; Romano et al., 2020; Yin et al., 2021). For classification problems, since the target is categorical without ordering, one may need to create a class embedding by the methods such as multidimensional scaling (Looveren and Klaise, 2021). Second, we focus on the Euclidean distance in computing the (HD-)PCP nonconformity score. It would be interesting to study the geometry of the predictive sets under different distance measures. Finally, it is intriguing to combine (HD-)PCP with the most recent deep generative models to provide statistical guarantees (Song et al., 2020; Ho et al., 2020).

Acknowledgements

D. M. Blei acknowledges the support of NSF IIS 2127869, ONR N00014-17-1-2131, ONR N00014-15-1-2209, Simons Foundation, Open Philanthropy. M. Yin acknowledges the computing support by HiPerGator.

References

- Alsing, J., Charnock, T., Feeney, S., and Wandelt, B. (2019). Fast likelihood-free cosmology with neural density esti-
- mators and active learning. *Monthly Notices of the Royal Astronomical Society*, 488(3):4440–4458.
- Ambrogioni, L., Güçlü, U., van Gerven, M. A., and Maris, E. (2017). The kernel mixture network: A nonparametric method for conditional density estimation of continuous random variables. *arXiv preprint arXiv:1705.07111*.
- Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2019a). The limits of distribution-free conditional predictive inference.
- Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2019b). Predictive inference with the jackknife+. *arXiv*.
- Biggs, M., Gao, R., and Sun, W. (2021). Loss functions for discrete contextual pricing with observational data. *arXiv preprint arXiv:2111.09933*.
- Bishop, C. M. (1994). Mixture density networks. Technical report, Aston University.
- Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54.
- Cauchois, M., Gupta, S., and Duchi, J. C. (2021). Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.*, 22:81:1–81:42.
- Chan, J., Perrone, V., Spence, J., Jenkins, P., Mathieson, S., and Song, Y. (2018). A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in neural information processing systems*, 31.
- Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118.
- Cho, D., Yoo, C., Im, J., and Cha, D. (2020). Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. *Earth and Space Science*, 7.
- Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.
- Gao, R., Biggs, M., Sun, W., and Han, L. (2021). Enhancing counterfactual classification via self-training. *arXiv preprint arXiv:2112.04461*.
- Geisser, S. (1993). *Predictive Inference*, volume 55. CRC Press.
- Gibbs, I. and Candès, E. (2022). Conformal inference for online prediction with arbitrary distribution shifts. *arXiv preprint arXiv:2208.08401*.
- Gillmann, C., Peter, L., Schmidt, C., Saur, D., and Scheuermann, G. (2021). Visualizing multimodal deep learning

- for lesion prediction. *IEEE Computer Graphics and Applications*, 41(5):90–98.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In *NIPS*.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.
- Hoff, P. (2021). Bayes-optimal prediction with frequentist coverage control. *arXiv*.
- Izbicki, R., Shimizu, G., and Stern, R. (2020). Flexible distribution-free conditional predictive bands using density estimators. In *International Conference on Artificial Intelligence and Statistics*, pages 3068–3077. PMLR.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kuchibhotla, A. K. (2020). Exchangeability, conformal prediction, and rank tests. *arXiv*.
- 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.
- Lei, J., Rinaldo, A., and Wasserman, L. (2015). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1):29–43.
- Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96.
- Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B*.
- Looveren, A. V. and Klaise, J. (2021). Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer.
- Meinshausen, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999.
- Messoudi, S., Destercke, S., and Rousseau, S. (2020). Conformal multi-target regression using neural networks. In *Conformal and Probabilistic Prediction and Applications*, pages 65–83. PMLR.
- Messoudi, S., Destercke, S., and Rousseau, S. (2021). Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101.
- Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Myers, V., Biyik, E., Anari, N., and Sadigh, D. (2022). Learning multimodal rewards from rankings. In *Conference on Robot Learning*, pages 342–352. PMLR.
- Neeven, J. and Smirnov, E. (2018). Conformal stacked weather forecasting. In *Conformal and Probabilistic Prediction and Applications*, pages 220–233. PMLR.
- 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.
- Romano, Y., Patterson, E., and Candès, E. J. (2019). Conformalized quantile regression. In *NeurIPS*.
- Romano, Y., Sesia, M., and Candès, E. J. (2020). Classification with valid and adaptive coverage. *arXiv: Methodology*.
- Sesia, M. and Romano, Y. (2021). Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34.
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., and Vlahavas, I. (2016). Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104(1):55–98.
- Taufiq, M. F., Ton, J.-F., Cornish, R., Teh, Y. W., and Doucet, A. (2022). Conformal off-policy prediction in contextual bandits. *arXiv preprint arXiv:2206.04405*.
- Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Tsanas, A. and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and buildings*, 49:560–567.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.
- Xu, C. and Xie, Y. (2022). Conformal prediction set for time-series. *arXiv preprint arXiv:2206.07851*.
- Yin, M., Shi, C., Wang, Y., and Blei, D. M. (2021). Conformal sensitivity analysis for individual treatment effects. *arXiv*.
- Yin, M. and Zhou, M. (2018a). Semi-implicit generative model. *Proceedings of the NeurIPS Workshop on Bayesian Deep Learning*.
- Yin, M. and Zhou, M. (2018b). Semi-implicit variational inference. In *International Conference on Machine Learning*, pages 5660–5669. PMLR.

A Proof

For completeness, we first present a lemma adapted from Tibshirani et al. (2019); Romano et al. (2019).

Lemma 1. Suppose Z_1, \dots, Z_n, Z_{n+1} are scalar random variables that are exchangeable and almost surely distinct, then for $\beta \in [0, 1]$

$$\beta \leq \mathbb{P}(Z_{n+1} \leq Q_\beta(Z_{1:n} \cup \{\infty\})) \leq \beta + \frac{1}{n+1}.$$

Proof of Lemma 1. Denote the inflated $Z_{1:n}$ as

$$\tilde{Z}_{1:n} = Z_{1:n} \cup \{\infty\}. \quad (7)$$

The β -th empirical quantile is defined as

$$Q_\beta(Z_{1:n}) := \inf\{x : (\sum_{i=1}^n \mathbb{1}[Z_i \leq x])/n \geq \beta\}. \quad (8)$$

By the definition in (8), $Q_\beta(Z_{1:n}) = Z_{(\lceil n\beta \rceil)}$ where $Z_{(k)}$ is the k-th order statistics. By Lemma 1 of Tibshirani et al. (2019), the events

$$Z_{n+1} \leq Q_\beta(\tilde{Z}_{1:n}) \Leftrightarrow Z_{n+1} \leq Q_\beta(Z_{1:n+1}). \quad (9)$$

Furthermore, by exchangeability and the definition of empirical quantile,

$$\begin{aligned} \mathbb{P}(Z_{n+1} \leq Q_\beta(Z_{1:n+1})) &= \mathbb{P}(Z_{n+1} \leq Z_{(\lceil (n+1)\beta \rceil)}) \\ &\geq \frac{\lceil (n+1)\beta \rceil}{n+1} \geq \beta. \end{aligned} \quad (10)$$

When $Z_{1:n+1}$ are almost surely distinct,

$$\begin{aligned} \mathbb{P}(Z_{n+1} \leq Q_\beta(Z_{1:n+1})) &= \mathbb{P}(Z_{n+1} \leq Z_{(\lceil (n+1)\beta \rceil)}) \\ &= \frac{\lceil (n+1)\beta \rceil}{n+1} \leq \frac{\lfloor (n+1)\beta \rfloor + 1}{n+1} \leq \beta + \frac{1}{n+1}. \end{aligned} \quad (11)$$

By Eq. (9), the proof is completed. \square

Proof of Theorem 1. Given the estimated conditional density $q(Y|X)$ on an independently sampled training set D_{tr} . By assumption, the calibration set $\{(X_i, Y_i)\}_{i=1}^n$ and the test point (X_{n+1}, Y_{n+1}) are exchangeable. Denote $D_i = (X_i, Y_i, \hat{Y}_i)$ for $i = 1, \dots, n, n+1$. Then $D_i \sim p(X, Y)q^K(Y|X_i)$ and $\{D_{1:n+1}\}$ are exchangeable because $\hat{Y}_i \perp\!\!\!\perp \hat{Y}_j$.

The nonconformity score E_i in Eq. (2) is defined as a deterministic function of D_i . Therefore $\{E_i\}_{i=1}^{n+1}$ are exchangeable (Kuchibhotla, 2020) and are almost surely distinct. By Lemma 1,

$$1 - \alpha \leq \mathbb{P}(E_{n+1} \leq Q_\alpha(E_{1:n} \cup \{\infty\}) | D_{\text{tr}}) \leq 1 - \alpha + \frac{1}{n+1}. \quad (12)$$

Next we demonstrate that for

$$\hat{C}(X_{n+1}, \hat{Y}_{n+1}) = \cup_{k=1}^K \{y : \|y - \hat{Y}_{n+1,k}\| \leq \hat{Q}\}, \quad \hat{Q} = Q_{1-\alpha}(E_{1:n} \cup \{\infty\}),$$

the following statement holds

$$Y_{n+1} \in \hat{C}(X_{n+1}, \hat{Y}_{n+1}) \Leftrightarrow E_{n+1} \leq \hat{Q}.$$

Suppose the LHS is true, then $\exists m, 1 \leq m \leq K$, s.t. $Y_{n+1} \in \{y : \|y - \hat{Y}_{n+1,m}\| \leq \hat{Q}\}$. This means $\|Y_{n+1} - \hat{Y}_{n+1,m}\| \leq \hat{Q}$. Hence $E_{n+1} = \min_k \|Y_{n+1} - \hat{Y}_{n+1,k}\| \leq \|Y_{n+1} - \hat{Y}_{n+1,m}\| \leq \hat{Q}$.

Algorithm 2 High Density Probabilistic Conformal Prediction

Input: Data $\mathbf{D} = \{(X_i, Y_i)\}_{i=1}^N$, nominal level α , test point X , generative model class \mathcal{Q} , sample size K , β grid B (For HD-PCP).

Step I: Conditional generative model

- 1: Split the data into three folds \mathcal{Z}_{tr} , \mathcal{Z}_{val} , \mathcal{Z}_{cal} with set of index as \mathbf{I}_{tr} , \mathbf{I}_{val} , \mathbf{I}_{cal} respectively
- 2: Estimate $q(Y|X)$ on \mathcal{Z}_{tr} with hyper-parameter chosen by cross validation on \mathcal{Z}_{val}

Step II: Predictive set for a test point

- 1: For $i \in \mathbf{I}_{\text{cal}}$, sample $\hat{Y}_{i1}, \dots, \hat{Y}_{iK} \sim q(Y|X_i)$
- 2: For $\beta \in B$, Filtering out β fraction of $\{\hat{Y}_{ik}\}_{k=1}^K$ with the lowest density. Repeat Line 3-7 for $x \in \mathbf{I}_{\text{cal}}$. $\beta_0 = \arg \min_{\beta} \lambda(\sum_{x \in \mathbf{I}_{\text{cal}}} \hat{C}_{\beta}(x, \hat{\mathbf{Y}}))$
- 3: For a test point, sample $\hat{Y}_1, \dots, \hat{Y}_K \sim q(Y|X)$
- 4: Filtering out β fraction of $\{\hat{Y}_k\}_{k=1}^K$ with the lowest density
- 5: Compute nonconformity score $\{E_i\}_{i \in \mathbf{I}_{\text{cal}}}$ by Eq. (2), $E_{N+1} = \infty$, $\tilde{\mathbf{I}}_{\text{cal}} = \mathbf{I}_{\text{cal}} \cup \{N+1\}$
- 6: Set r as the $(1 - \alpha)$ empirical quantile of $\{E_i\}_{i \in \tilde{\mathbf{I}}_{\text{cal}}}$
- 7: Compute the predictive set $\hat{C}_{\beta}(X, \hat{\mathbf{Y}})$ by Eq. (3)

Output: Predictive set $\hat{C}_{\beta_0}(X, \hat{\mathbf{Y}})$

On the other hand, suppose the RHS is true, letting $t = \arg \min_k \|Y_{n+1} - \hat{Y}_{n+1,k}\|$, we have $\|Y_{n+1} - \hat{Y}_{n+1,t}\| \leq \hat{Q}$, i.e., $Y_{n+1} \in \{y : \|y - \hat{Y}_{n+1,t}\| \leq \hat{Q}\}$. Therefore, $Y_{n+1} \in \hat{C}(X_{n+1}, \hat{\mathbf{Y}}_{n+1})$.

Then by Eq. (12), we have

$$1 - \alpha \leq \mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}, \hat{\mathbf{Y}}_{n+1}) | D_{\text{tr}}) \leq 1 - \alpha + \frac{1}{n+1}. \quad (13)$$

Marginalizing out D_{tr} the statement is proved. \square

Proof of Corollary 1. Suppose $0 \leq \beta \leq n/(n+1)$, then $\lceil (n+1)\beta \rceil \neq n+1$. We have

$$Q_{\beta}(\tilde{Z}_{1:n}) = \tilde{Z}_{(\lceil (n+1)\beta \rceil, n+1)} = Z_{(\lceil n \frac{n+1}{n} \beta, n \rceil)} = Q_{(1+\frac{1}{n})\beta}(Z_{1:n}) \quad (14)$$

where $\tilde{Z}_{1:n}$ is defined in Eq. (7). By Eqs. (9) to (11),

$$\beta \leq \mathbb{P}(Z_{n+1} \leq Q_{(1+\frac{1}{n})\beta}(Z_{1:n})) \leq \beta + \frac{1}{n+1}. \quad (15)$$

By Eq. (15), we have

$$\beta - \frac{1}{n+1} \leq \mathbb{P}(Z_{n+1} \leq Q_{\beta}(Z_{1:n})) \leq \beta + \frac{1}{n+1}. \quad (16)$$

\square

Proof of Corollary 2. With the notations in the proof of Theorem 1, $D_i = (X_i, Y_i, \hat{\mathbf{Y}}_i)$ are i.i.d. variables. For a fixed conditional density function $q(y|x)$, the nonconformity score E_i is fully determined by $X_i, Y_i, \hat{\mathbf{Y}}_i$. So $E_i = g(D_i)$ where $g(\cdot)$ is a deterministic function including the filtering step of HD-PCP. By Kuchibhotla (2020), since $\{D_i\}_{i=1}^{n+1}$ are i.i.d., $\{E_i\}_{i=1}^{n+1}$ are exchangeable. The other parts of proof follow the same as the proof of Theorem 1. \square

B High Density Probabilistic Conformal Prediction

Figure 7 shows the different predictive sets with 95% coverage when the underlying distribution is bi-mode normal.

We summarize our HD-PCP algorithm in Algorithm 2.

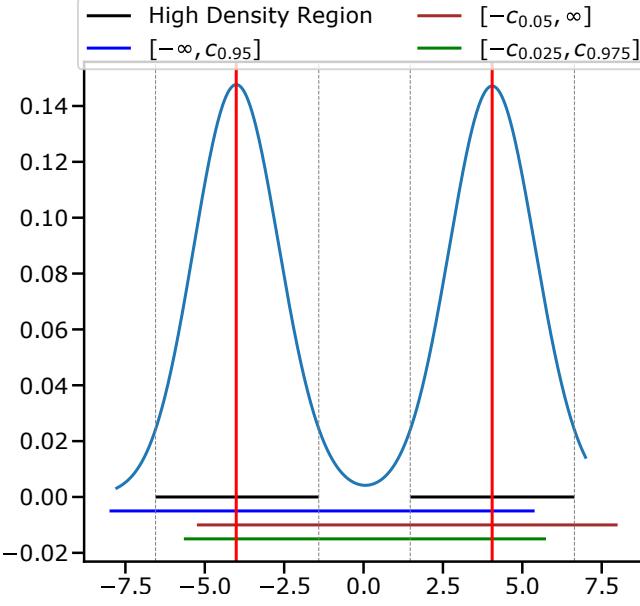


Figure 7: High Density Region can capture multi-mode easier comparing to other predictive set.

C Hyperparameters

PCP introduces one additional hyperparameter, the sample size K . We conduct ablation study on the effect of K in Fig. 5 and find that as long as K is set moderately large, i.e., $K = 40$, the predictive set size is near optimal. K is not a very sensitive hyperparameter that needs much effort for tuning.

D Summary of Conditional generative models

SIVI Model. Following (Yin and Zhou, 2018b), we build a conditional distribution estimator by using semi-implicit variational inference and we call it SIVI model. Specifically, we approximate $\mathbb{P}(Y | X)$ by an inference distribution $q_\phi(Y | X)$ with respect to parameter ϕ . We construct the inference distribution as a hierarchical model,

$$y \sim q_{\phi_1}(y|x, z), z \sim q_{\phi_2}(z|x, \psi), \psi \sim p(\psi)$$

Here $z \in \mathbb{R}^d$ is an auxiliary latent variable and $p(\psi)$ is a known noise distribution, i.e., $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The inference distribution is the marginal distribution of the hierarchical model, $q_\phi(y | x) = \int q_{\phi_1}(y|z)q_{\phi_2}(z|x)dz$ with $\phi = (\phi_1, \phi_2)$. We model $q_{\phi_1}(y|z)$ and $q_{\phi_2}(z|x)$ as Gaussian distributions, whose mean and standard deviation are the outputs of neural networks feed with corresponding (x, z) and (x, ψ) . The marginal distribution $q(y | x)$ can thus be constructed with flexibility in modeling multimodality, skewness and kurtosis. We learn the ϕ by maximizing the ELBO,

$$\mathcal{L}_K = \mathbb{E}_{\psi^{(0)}, \dots, \psi^{(K)} \stackrel{iid}{\sim} p(\psi)} \mathbb{E}_{z \sim q(\cdot | \psi^{(0)}, x)} \left[\ln q_{\phi_1}(y | x, z) + \ln \frac{p(z)}{\frac{1}{K+1} \sum_{k=0}^K q_{\phi_2}(z | \psi^{(k)}, x)} \right],$$

where $p(z)$ is a prior distribution for latent variable z , i.e., $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and K is set as 20.

GAN model. To fit a conditional distribution, we follows (Mirza and Osindero, 2014) to build a Conditional GAN model with a generator $G(x, z)$ and a discriminator $D(x, y)$. For simplicity, we call it GAN model. Here, the z is a latent variable, which is usually set as $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $G(x, z)$ is modeled by a neural network, whose outputs are the samples of y , and $D(x, y)$ is another neural network, which outputs the probability that the given y is from the true data distribution. We train G and D with the following adversarial loss,

$$\min_G \max_D V(G, D) = \mathbb{E}_{x, y \sim p_{data}(x, y)} [\log D(x, y)] + \mathbb{E}_{z \sim p(z), x \sim p_{data}(x)} [\log(1 - D(x, G(x, z)))]$$

Kernel Mixture Network. Ambrogioni et al. (2017) model arbitrarily complex conditional densities as linear combinations of a family of kernel functions centered at a subset of training points. The weights are determined by the outer layer of a

deep neural network, trained by minimizing the negative log likelihood. The conditional density function is modeled as follows,

$$q(y|x) = \frac{1}{\sum_{p,j} w_{p,j}(x; W)} \sum_{p,j} w_{p,j}(x; W) \mathcal{K}_j(y, y^{(p)}),$$

where p denotes the index of the observed data points, \mathcal{K}_j is the pre-set kernel function, j is the index of selected bandwidth for \mathcal{K}_j , and $w_{p,j}(x; W)$ represents the weight of each kernel. A common choice for \mathcal{K}_j is the Gaussian kernel,

$$\mathcal{K}_j(y, y'; \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp^{-\frac{(y-y')^2}{2\sigma_j^2}}.$$

The weights $w_{p,j}(x; W)$ are determined by a deep neural network (DNN), with covariates x as the inputs and W as the parameters. All weights are non negative by applying non-negative activation functions on the output layer of DNN. We train the KMN model by minimizing the loss function,

$$\mathcal{L}(W) = - \sum_q \left[\log \sum_{p,j} w_{p,j}(x; W) \mathcal{K}_j(y, y^{(p)}) - \log \sum_{p,j} w_{p,j}(x; W) \right]$$

Mixture Density Network. Bishop (1994) proposes the mixture density network as follows,

$$q(y|x) = \sum_{k=1}^K \pi_k(x) \mathcal{N}(y|\mu_k(x), \sigma_k^2(x))$$

where $\pi_k(\cdot)$, $\mu_k(\cdot)$ and $\sigma_k(\cdot)$ are all modeled by neural networks. $\sum_{k=1}^K \pi_k(x) = 1$ is guaranteed by using softmax activation function. The model is trained by minimizing the loss function,

$$\mathcal{L} = - \sum_{i=1}^N \log \left[\sum_{k=1}^K \pi_k(x_i) \mathcal{N}(y_i|\mu_k(x_i), \sigma_k^2(x_i)) \right],$$

where $\{(x_i, y_i)\}_{i=1}^N$ are the observed data points.

Quantile Regression Forest Meinshausen (2006) shows that random forests provide information about the full conditional distribution of the response variable, not only about the conditional mean. Conditional quantiles can be inferred with quantile regression forests, a generalisation of random forests. Quantile regression forests give a non-parametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables. We refer to (Meinshausen, 2006) for more details about QRF model. PCP needs to get samples from QRF. We first sample a percentile $\tau \sim U[0, 1]$, the uniform distribution on the unit interval, and then use QRF to get the estimated conditional quantile value y_τ as a y sample.

E Full synthetic experiment results

We include the Full synthetic experiment results for 2D toy datasets, s-curve, half-moons, 25-Gaussians, 8-Gaussians, circle and swiss-roll, in Fig. 8 and Fig. 9. We compare conformal prediction with mean estimation (CP-MeanPred), CHR-QRF and CDSplit-MixD with our method (HD-)PCP.

The data used to plot Fig. 4b and Fig. 4c is included in Table 8. When ρ increases, the set size decreases for PCP and HD-PCP while keeping nearly constant for other baselines, which overlooks the joint relationship between targets.

F Full real data experiment results

We report all experiment results for our single target real data regression tasks in Table 5, Table 6 and Table 7. Table 4 illustrates the best performance of each method with best backbone model picked for each dataset respectively.

G Additional Plots for NYC Taxi Data

Fig. 10 shows the additional plots for NYC Taxi data for PCP, HD-PCP, CHR and CDSplit. Each row represents one individual record in the test set and we show the predictive set generated by each algorithm. Clearly, PCP and HD-PCP generate the

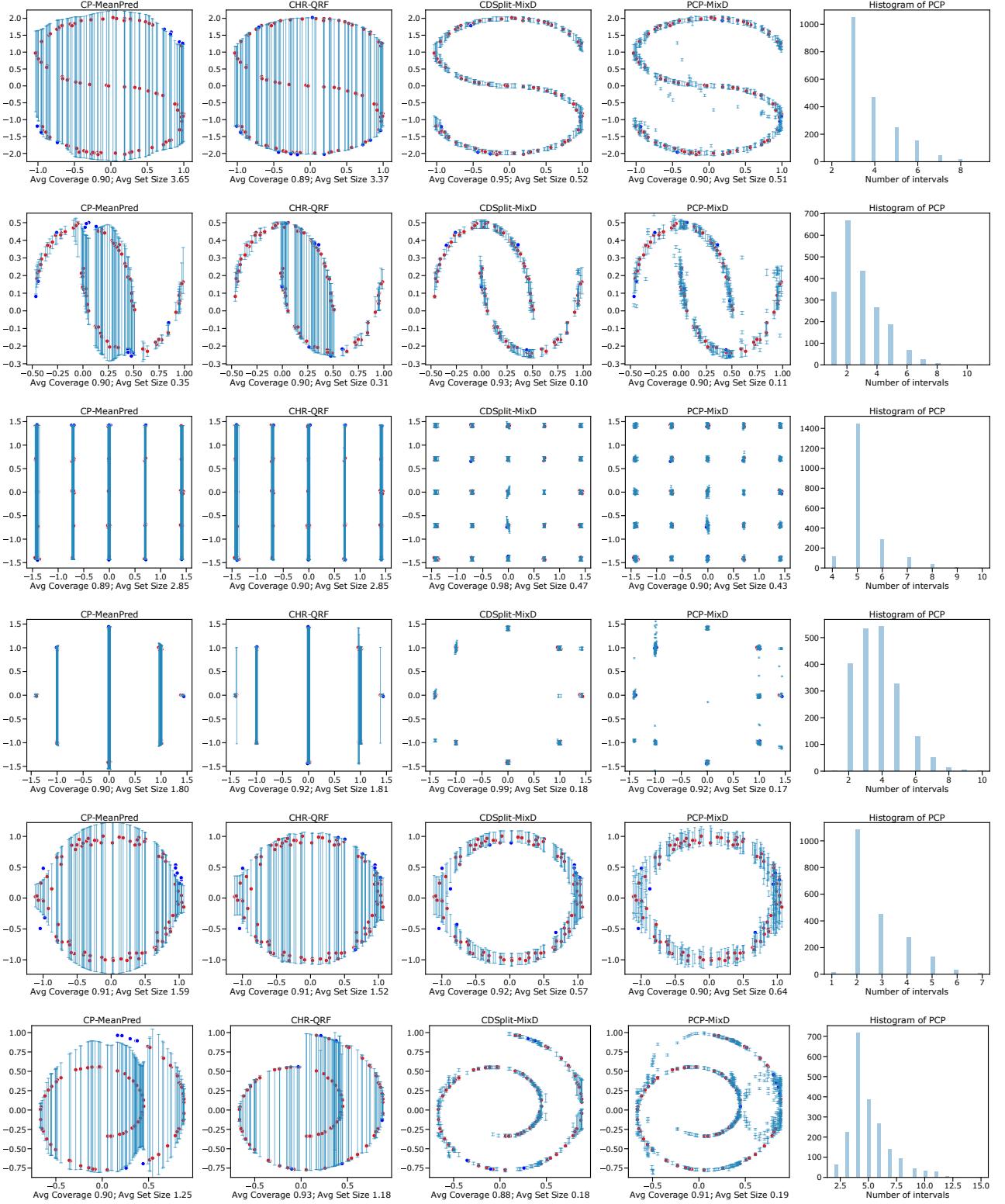


Figure 8: Visualization of predictive sets ($\alpha = 0.1$) on 2D toy datasets: s-curve, half-moons, 25-Gaussians, 8-Gaussians, circle and swiss-roll. For ours, we show the PCP in the last two columns. We show the predictive sets on 100 test data samples, where blues lines represent the predictive sets, blue dots are test points that are not covered by the predictive sets and reds dots are the test points covered. We show the marginal coverage and the average interval length across test datapoints in the x-axis label. The fifth column shows the histogram of the number predicted intervals of PCP.

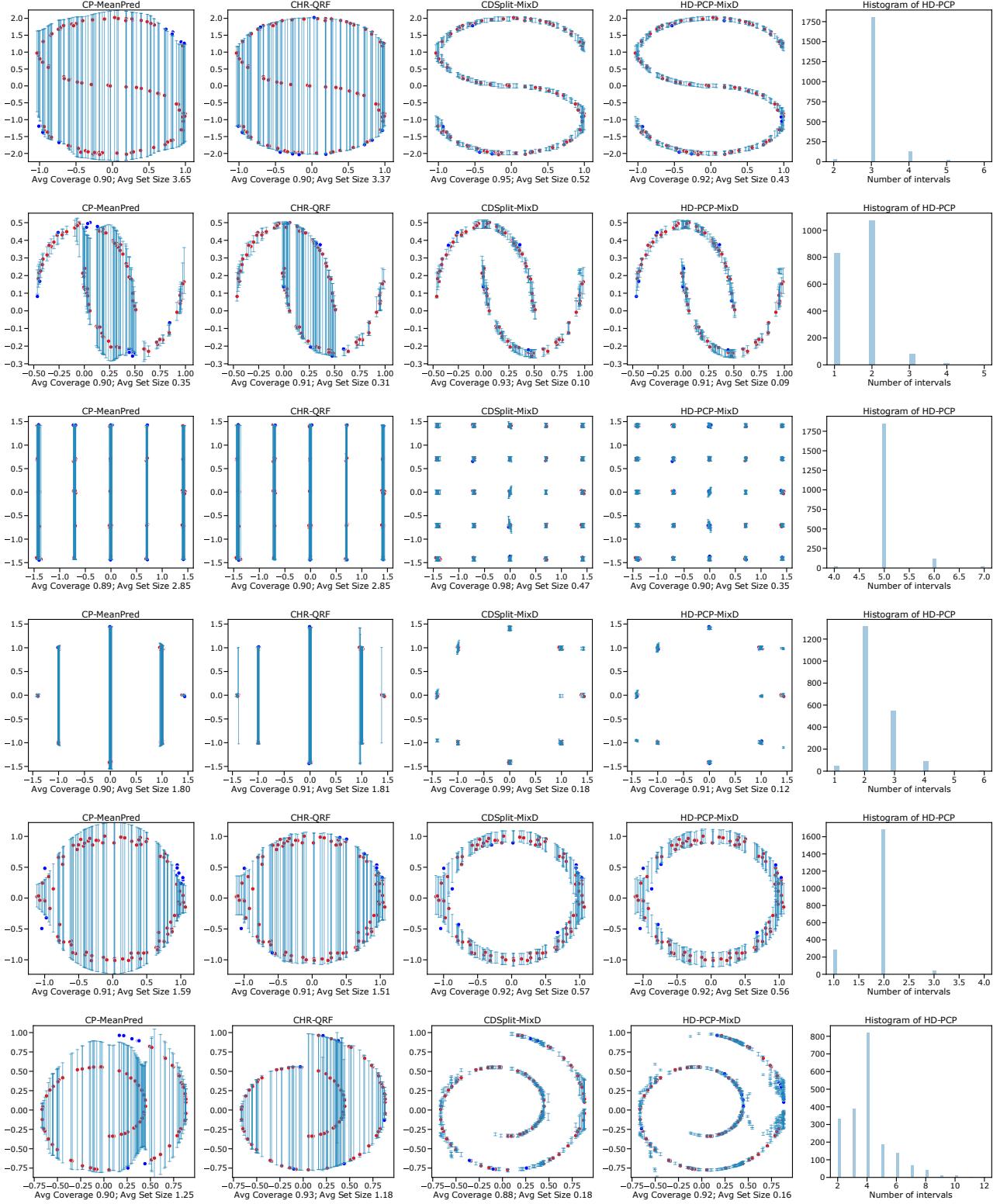


Figure 9: Visualization of predictive sets ($\alpha = 0.1, \beta = 0.2$) on 2D toy datasets: s-curve, half-moons, 25-Gaussians, 8-Gaussians, circle and swiss-roll. For ours, we show the HD-PCP in the last two columns. We show the predictive sets on 100 test data samples, where blues lines represent the predictive sets, blue dots are test points that are not covered by the predictive sets and reds dots are the test points covered. We show the marginal coverage and the average interval length across test datapoints in the x-axis label. The fifth column shows the histogram of the number predicted intervals of HD-PCP.

Data	n_train	n_calib	n_test
bike	6886	2000	2000
bio	41730	2000	2000
blog	48397	2000	2000
facebook1	36948	2000	2000
facebook2	77311	2000	2000
meps_19	11785	2000	2000
meps_20	13541	2000	2000
meps_21	11656	2000	2000
temperature	5314	1138	1138

Table 3: Dataset splits for training, calibration and testing.

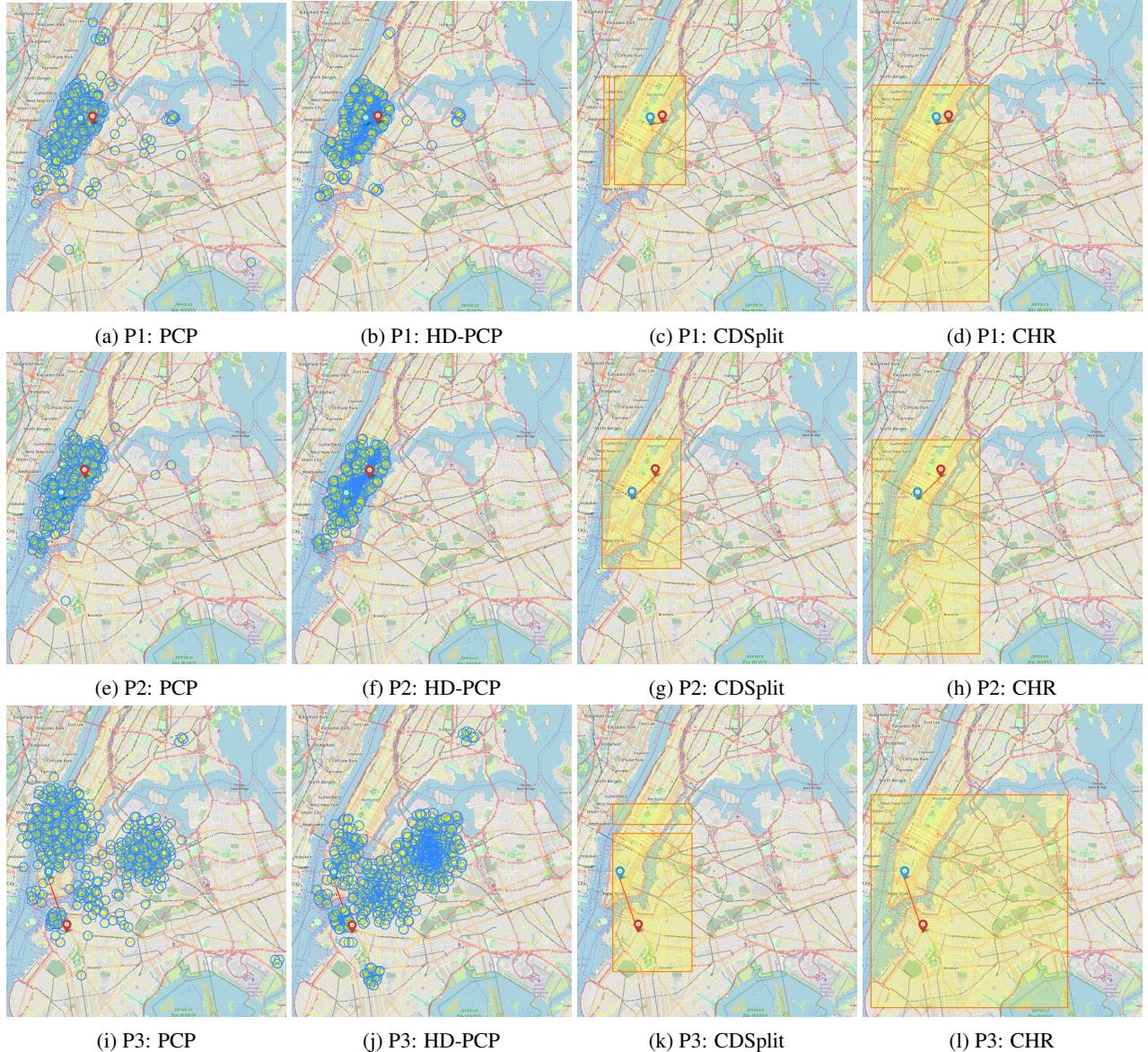


Figure 10: NYC Taxi data. Red Pin: pickup location; Blue Pin: dropoff location. Left to right: Predictive set output by PCP, HD-PCP, CDSplit and CHR. Each row represents a random selected individual record sampled from the dataset.

Probabilistic Conformal Prediction Using Conditional Random Samples

Data	Method	Marg. Coverage	Cond. Coverage	Set Size Mean	Set Size SE
bike	PCP-SIVI	0.90	0.86	134.55	1.13
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.88	102.92	0.48
	CHR-QRF	0.90	0.88	204.10	1.03
	DistSplit	0.90	0.87	423.13	1.51
	CDSplit-MixD	0.92	0.92	115.74	0.50
	DCP	0.90	0.88	443.76	1.36
	CQR	0.90	0.89	403.88	0.86
bio	PCP-KMN	0.90	0.89	10.30	0.05
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.89	8.76	0.05
	CHR-QRF	0.90	0.89	10.21	0.04
	DistSplit	0.90	0.89	13.19	0.04
	CDSplit-KMN	0.9	0.9	9.13	0.04
	DCP	0.90	0.89	12.95	0.04
	CQR2	0.90	0.89	12.88	0.05
blogdata	PCP-QRF	0.89	0.85	3.68	0.65
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.87	9.44	0.19
	CHR-QRF	0.90	0.87	10.81	0.17
	DistSplit	0.90	0.87	16.27	0.23
	CDSplit-MixD	0.96	0.95	39.00	0.40
	DCP	0.90	0.88	1422.36	0.03
	CQR2	0.90	0.87	13.91	0.27
facebook1	PCP-QRF	0.90	0.82	4.52	0.76
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.83	8.62	0.06
	CHR-NNNet	0.90	0.87	9.96	0.11
	DistSplit	0.90	0.89	14.03	0.16
	CDSplit-MixD	0.95	0.95	33.69	0.16
	DCP	0.90	0.89	1303.01	0.04
	CQR2	0.90	0.88	12.17	0.15
facebook2	PCP-QRF	0.90	0.82	3.62	0.72
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.82	8.34	0.07
	CHR-NNNet	0.90	0.87	10.15	0.13
	DistSplit	0.90	0.89	13.48	0.19
	CDSplit-KMN	0.95	0.94	44.53	0.26
	DCP	0.90	0.89	1963.68	0.03
	CQR2	0.90	0.89	11.41	0.17
meps19	PCP-GAN	0.90	0.86	18.41	0.17
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.88	17.78	0.18
	CHR-QRF	0.90	0.89	18.26	0.15
	DistSplit	0.90	0.89	29.96	0.28
	CDSplit-MixD	0.93	0.92	23.86	0.17
	DCP	0.90	0.88	559.23	0.01
	CQR	0.90	0.89	28.71	0.18
meps20	PCP-MixD	0.90	0.87	19.52	0.16
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.88	18.19	0.17
	CHR-QRF	0.90	0.90	17.94	0.18
	DistSplit	0.90	0.90	29.35	0.23
	CDSplit-MixD	0.92	0.92	22.93	0.16
	DCP	0.90	0.88	520.25	0.01
	CQR	0.90	0.90	27.57	0.15
meps21	PCP-MixD	0.90	0.87	19.18	0.12
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.88	17.91	0.15
	CHR-QRF	0.90	0.90	18.65	0.16
	DistSplit	0.90	0.89	30.32	0.31
	CDSplit-MixD	0.93	0.92	23.63	0.17
	DCP	0.90	0.88	531.25	0.01
	CQR	0.90	0.89	29.89	0.2
temperature	PCP-MixD	0.90	0.90	2.10	0.01
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.89	1.85	0.01
	CHR-NNNet	0.90	0.89	3.17	0.01
	DistSplit	0.90	0.89	3.07	0.01
	CDSplit-MixD	0.92	0.91	2.23	0.01
	DCP	0.90	0.88	3.1	0.02
	CQR2	0.90	0.88	3.14	0.02

Table 4: Best results of real data experiments (the best variant of each method for each dataset is selected).

Data	Method	Marg. Coverage	Cond. Coverage	Set Size Mean	Set Size SE
bike	PCP-SIVI	0.90	0.86	134.55	1.13
	PCP-GAN	0.90	0.88	399.32	2.48
	PCP-QRF	0.90	0.87	241.11	3.35
	PCP-MixD	0.90	0.87	128.13	0.53
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.88	102.92	0.48
	PCP-KMN	0.90	0.88	172.92	1.04
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.88	146.24	1.06
	CHR-NNNet	0.90	0.89	353.51	1.59
	CHR-QRF	0.90	0.88	204.10	1.03
	DistSplit	0.90	0.87	423.13	1.51
	CDSplit-KMN	0.92	0.91	161.16	0.72
	CDSplit-MixD	0.92	0.92	115.74	0.50
	DCP	0.90	0.88	443.76	1.36
	CQR	0.90	0.89	403.88	0.86
	CQR2	0.90	0.88	416.75	1.57
bio	PCP-SIVI	0.90	0.89	14.08	0.06
	PCP-GAN	0.90	0.89	13.11	0.05
	PCP-QRF	0.90	0.89	11.08	0.16
	PCP-MixD	0.90	0.89	11.47	0.04
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.90	10.06	0.05
	PCP-KMN	0.90	0.89	10.30	0.05
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.89	8.76	0.05
	CHR-NNNet	0.90	0.89	11.74	0.04
	CHR-QRF	0.90	0.89	10.21	0.04
	DistSplit	0.90	0.89	13.19	0.04
	CDSplit-KMN	0.90	0.90	9.13	0.04
	CDSplit-MixD	0.90	0.90	9.58	0.04
	DCP	0.90	0.89	12.95	0.04
	CQR	0.90	0.89	13.00	0.02
	CQR2	0.90	0.89	12.88	0.05
blog	PCP-SIVI	0.90	0.85	11.21	0.32
	PCP-GAN	0.90	0.86	11.67	0.16
	PCP-QRF	0.89	0.85	3.68	0.65
	PCP-MixD	0.90	0.85	10.78	0.17
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.87	9.44	0.19
	PCP-KMN	0.90	0.85	10.67	0.13
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.86	10.51	0.14
	CHR-NNNet	0.90	0.88	11.1	0.19
	CHR-QRF	0.90	0.87	10.81	0.17
	DistSplit	0.90	0.87	16.27	0.23
	CDSplit-KMN	0.96	0.95	45.90	0.62
	CDSplit-MixD	0.96	0.95	39.00	0.40
	DCP	0.90	0.88	1422.36	0.03
	CQR	0.90	0.87	15.15	0.26
	CQR2	0.90	0.87	13.91	0.27
facebook1	PCP-SIVI	0.90	0.83	8.8	0.06
	PCP-GAN	0.90	0.85	9.22	0.05
	PCP-QRF	0.90	0.82	4.52	0.76
	PCP-MixD	0.90	0.82	9.99	0.14
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.85	8.93	0.12
	PCP-KMN	0.90	0.82	10.60	0.06
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.83	8.62	0.06
	CHR-NNNet	0.90	0.87	9.96	0.11
	CHR-QRF	0.90	0.86	11.21	0.12
	DistSplit	0.90	0.89	14.03	0.16
	CDSplit-KMN	0.95	0.94	33.88	0.19
	CDSplit-MixD	0.95	0.95	33.69	0.16
	DCP	0.90	0.89	1303.01	0.04
	CQR	0.90	0.89	13.79	0.15
	CQR2	0.90	0.88	12.17	0.15

Table 5: Detailed results of experiments on data: bike, bio, blog and facebook1.

Data	Method	Marg. Coverage	Cond. Coverage	Set Size Mean	Set Size SE
facebook2	PCP-SIVI	0.90	0.83	8.69	0.17
	PCP-GAN	0.90	0.84	9.47	0.1
	PCP-QRF	0.90	0.82	3.62	0.72
	PCP-MixD	0.90	0.82	9.93	0.11
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.84	8.84	0.10
	PCP-KMN	0.90	0.81	10.42	0.07
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.82	8.34	0.07
	CHR-NNet	0.90	0.87	10.15	0.13
	CHR-QRF	0.90	0.87	10.81	0.14
	DistSplit	0.90	0.89	13.48	0.19
	CDSplit-KMN	0.95	0.94	44.53	0.26
	CDSplit-MixD	0.97	0.96	45.75	0.16
	DCP	0.90	0.89	1963.68	0.03
	CQR	0.90	0.89	13	0.17
	CQR2	0.90	0.89	11.41	0.17
meps19	PCP-SIVI	0.90	0.85	26.93	0.3
	PCP-GAN	0.90	0.86	18.41	0.17
	PCP-QRF	0.90	0.86	20.16	0.48
	PCP-MixD	0.90	0.87	19.28	0.16
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.88	17.78	0.18
	PCP-KMN	0.90	0.85	23.24	0.21
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.84	23.48	0.20
	CHR-NNet	0.90	0.90	20.17	0.2
	CHR-QRF	0.90	0.89	18.26	0.15
	DistSplit	0.90	0.89	29.96	0.28
	CDSplit-KMN	0.93	0.91	31.10	0.33
	CDSplit-MixD	0.93	0.92	23.86	0.17
	DCP	0.90	0.88	559.23	0.01
	CQR	0.90	0.89	28.71	0.18
	CQR2	0.90	0.89	30.78	0.36
meps20	PCP-SIVI	0.90	0.86	23.87	0.16
	PCP-GAN	0.90	0.86	19.92	0.18
	PCP-QRF	0.90	0.86	20.47	0.52
	PCP-MixD	0.90	0.87	19.52	0.16
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.88	18.19	0.17
	PCP-KMN	0.90	0.85	22.96	0.17
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.85	23.35	0.18
	CHR-NNet	0.90	0.9	19.43	0.18
	CHR-QRF	0.90	0.90	17.94	0.18
	DistSplit	0.90	0.90	29.35	0.23
	CDSplit-KMN	0.93	0.90	29.05	0.30
	CDSplit-MixD	0.92	0.92	22.93	0.16
	DCP	0.90	0.88	520.25	0.01
	CQR	0.90	0.90	27.57	0.15
	CQR2	0.90	0.90	29.94	0.31
meps21	PCP-SIVI	0.90	0.85	23.74	0.27
	PCP-GAN	0.90	0.86	19.73	0.16
	PCP-QRF	0.89	0.86	18.52	0.45
	PCP-MixD	0.90	0.87	19.18	0.12
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.88	17.91	0.15
	PCP-KMN	0.90	0.85	23.13	0.17
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.85	23.70	0.19
	CHR-NNet	0.90	0.90	20.07	0.22
	CHR-QRF	0.90	0.90	18.65	0.16
	DistSplit	0.90	0.89	30.32	0.31
	CDSplit-KMN	0.92	0.91	30.42	0.41
	CDSplit-MixD	0.93	0.92	23.63	0.17
	DCP	0.90	0.88	531.25	0.01
	CQR	0.90	0.89	29.89	0.2
	CQR2	0.90	0.89	31.78	0.36

Table 6: Detailed results of experiments on data: facebook2, meps19, meps20 and meps21.

Data	Method	Marg. Coverage	Cond. Coverage	Set Size Mean	Set Size SE
temperature	PCP-SIVI	0.90	0.90	3.27	0.06
	PCP-GAN	0.90	0.89	3.51	0.04
	PCP-QRF	0.88	0.86	3.78	0.09
	PCP-MixD	0.90	0.90	2.10	0.01
	HD-PCP-MixD($\beta = 0.2$)	0.90	0.89	1.85	0.01
	PCP-KMN	0.90	0.89	2.68	0.01
	HD-PCP-KMN($\beta = 0.2$)	0.90	0.89	2.43	0.01
	CHR-NNNet	0.90	0.89	3.17	0.01
	CHR-QRF	0.90	0.89	3.24	0.01
	DistSplit	0.90	0.89	3.07	0.01
	CDSplit-KMN	0.91	0.90	2.84	0.02
	CDSplit-MixD	0.92	0.91	2.23	0.01
	DCP	0.90	0.88	3.1	0.02
	CQR	0.90	0.87	3.55	0.03
	CQR2	0.90	0.88	3.14	0.02

Table 7: Detailed results of experiments on data: temperature.

Synthetic Data	Method	Cond. Coverage	Marg. Coverage	Set Size Mean	Set Size SE
$\rho = 0$	HD-PCP-MixD	0.92 (0.03)	0.90 (0.01)	356.82	10.78
	PCP-MixD	0.94 (0.02)	0.90 (0.01)	412.40	12.70
	CHR-NNNet	0.88 (0.05)	0.92 (0.01)	690.49	51.55
	DistSplit	0.93 (0.01)	0.92 (0.01)	714.92	54.81
	CDSplit-MixD	0.87 (0.03)	0.90 (0.01)	437.98	20.20
	DCP	0.90 (0.03)	0.92 (0.01)	710.64	53.99
	CQR	0.95 (0.02)	0.92 (0.01)	667.01	51.47
	CQR2	0.89 (0.03)	0.92 (0.01)	694.64	52.40
$\rho = 5$	HD-PCP-MixD	0.88 (0.03)	0.89 (0.01)	302.64	9.88
	PCP-MixD	0.87 (0.03)	0.88 (0.01)	348.93	10.70
	CHR-NNNet	0.92 (0.03)	0.91 (0.01)	689.22	57.47
	DistSplit	0.93 (0.02)	0.91 (0.01)	702.56	54.02
	CDSplit-MixD	0.87 (0.03)	0.90 (0.01)	415.86	17.49
	DCP	0.89 (0.05)	0.91 (0.02)	697.18	54.01
	CQR	0.87 (0.03)	0.91 (0.02)	653.64	53.59
	CQR2	0.88 (0.05)	0.91 (0.02)	693.10	56.18
$\rho = 9$	HD-PCP-MixD	0.83 (0.07)	0.89 (0.02)	171.61	8.36
	PCP-MixD	0.89 (0.03)	0.90 (0.01)	205.63	5.76
	CHR-NNNet	0.92 (0.02)	0.92 (0.02)	664.31	51.59
	DistSplit	0.91 (0.05)	0.91 (0.01)	689.72	59.96
	CDSplit-MixD	0.92 (0.02)	0.91 (0.01)	416.86	16.96
	DCP	0.90 (0.03)	0.91 (0.02)	666.13	50.61
	CQR	0.93 (0.04)	0.92 (0.01)	639.00	50.19
	CQR2	0.92 (0.02)	0.91 (0.01)	663.54	54.59

Table 8: Detailed results for multidimensional target synthetic dataset. The set size for PCP and HD-PCP decreases when ρ increases while the set sizes for other baselines are similar for different ρ since the marginal distribution remains the same.

most informative predictive set where popular neighborhoods and airports are tagged, while HD-PCP offers a more sparse and clean set. As expected, CHR offers a wide predictive set. Both CDSplit and CHR fail to provide predictive set with clear interpretation. Fig. 12 shows the predictive set and heatmap for pickup from SOHO and Chinatown. Most popular spots in NYC have higher density, which means passengers are more likely to be dropped off there.

H Additional Results for Multi-Target Regression Task

We include more experiment results for multi-target regression task in this section. We use two datasets for river flow prediction (Spyromitros-Xioufis et al., 2016) and stock prediction from StatLib repository. River flow dataset predicts the river network flows for future 48 hours for 8 sites (8 targets) and the stock dataset has stock price for 10 aerospace companies and we try to predict 3 companies' price using remaining companies'. There are 64 features including past river flow information for river flow predictions. For train, calibration and test size, we use 6925, 2000 and 200 for river flow prediction and 750, 100 and 100 for stock prediction respectively.

Since the Monte-Carlo estimation of overlapping hypersphere suffers from curse of dimensionality. We convert each dataset into two-dimensional pairwise comparisons to evaluate the robustness of each method (8 targets result in 28 pairs). We plot the pairwise comparison of PCP and HD-PCP against CHR and CDSplit, the two baselines that performs generally the best among other datasets. We use Mixture density Network for CDSplit, PCP and HD-PCP and Neural Netork based CHR, the results are averaged over 5 runs.

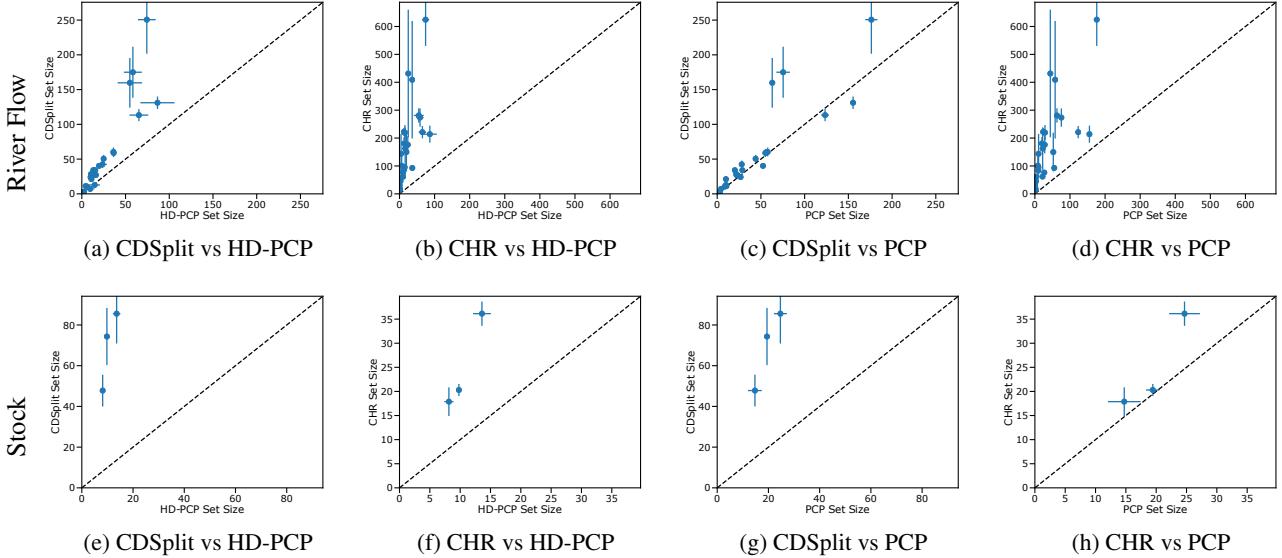


Figure 11: Additional Results for Multi-Target Regression Task. Each point corresponds to the size of predictive set for two elements of the target vector.

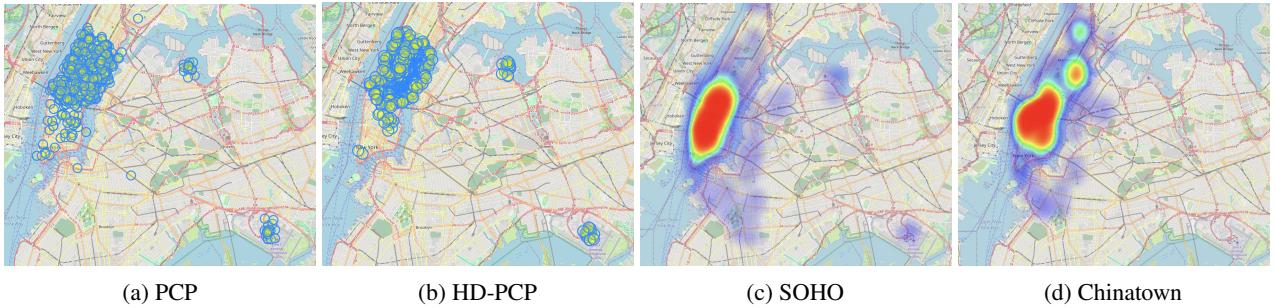


Figure 12: NYC Taxi data. (a), (b): predictive set for an individual using PCP and HD-PCP; (c): predictive set for riders from SoHo; (d): predictive set for riders from Chinatown.

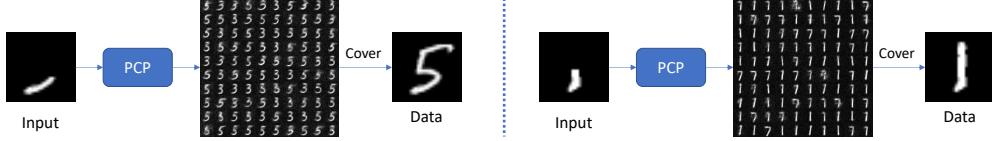


Figure 13: MNIST data. We show that PCP could produce proper predictive sets for high dimensional MNIST datasets.

For X-axis, we plot the set size of PCP / HD-PCP and for Y-axis, we plot the set size for CDSplit and CHR, and we also show the $Y = X$ line. If all points fall into the left region, it means PCP / HD-PCP outputs a sharper predictive set. For PCP, almost all points fall into the left region, which indicates PCP has a better or comparable performance with CDSplit and CHR in all pairwise comparisions. HD-PCP has a much better performance and the points all fall into the far left part in the figure, which shows HD-PCP offers a much sharper predictive set.

I MNIST Experiments

In this experiment, we aim to evaluate whether PCP could produce reasonable predictive sets for high dimensional data. We use the MNIST dataset (Deng, 2012), which contains 70000 (28×28) images (60000 for training and 10000 for testing) of handwritten digits from 0 to 9. We choose digits (3, 5) and (1, 7) from the dataset. The upper half of the digits is masked as the input X and the original unmasked images are the outputs Y .

We first fit a Conditional VAE model onto the selected data (X, Y) . We follow the original VAE (Kingma and Welling, 2013) architecture, where the encoder and decoder are multilayer perceptron (MLP) based with 500 hidden units for each layer. The latent variable dimension is set as 50. We train the VAE model with 10 epochs on the selected training data.

We build our PCP onto the pretrained VAE model, with additional 500 calibration datapoints. Since now Y has $14 \times 28 = 392$ dimensions, we compute the \mathcal{L}_2 distance between the \hat{Y} samples and the true Y , and select the minimum \mathcal{L}_2 as our nonconformity score. Then, we compute the $(1 - \alpha)$ quantile of the nonconformity scores over the calibration data to obtain the predictive set radius r . Finally, we evaluate PCP on test data points. For each test datapoint, we sample $K = 10$ random samples from the VAE model, build the predictive set via radius r and then sample from the predictive set to get PCP samples, as shown in Fig. 13.