cemmap

centre for microdata methods and practice

Prediction sets and conformal inference with censored outcomes

Weiguang Liu
Áureo de Paula
Elie Tamer

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP04/25

Economic and Social Research Council

Prediction Sets and Conformal Inference with Censored Outcomes*

Weiguang Liu
UCL

Áureo de Paula
UCL, IFS and CeMMAP

Elie Tamer
Harvard University

January 19, 2025

Abstract

Given data on a scalar random variable Y, a prediction set for Y with miscoverage level α is a set of values for Y that contains a randomly drawn Y with probability $1-\alpha$, where $\alpha \in (0,1)$. Among all prediction sets that satisfy this coverage property, the oracle prediction set is the one with the smallest volume. This paper provides estimation methods of such prediction sets given observed conditioning covariates when Y is *censored* or *measured in intervals*. We first characterise the oracle prediction set under interval censoring and develop a consistent estimator for the shortest prediction *interval* that satisfies this coverage property. We then extend these consistency results to accommodate cases where the prediction set consists of multiple disjoint intervals. Second, we use conformal inference to construct a prediction set that achieves a particular notion of finite-sample validity under censoring and maintains consistency as sample size increases. This notion exploits exchangeability to obtain finite sample guarantees on coverage using a specially constructed conformity score function. The procedure accomodates the prediction uncertainty that is irreducible (due to the stochastic nature of outcomes), the modelling uncertainty due to partial identification and also sampling uncertainty that gets reduced as samples get larger. We conduct a set of Monte Carlo simulations and an application to data from the Current Population Survey. The results highlight the robustness and efficiency of the proposed methods.

*Liu: weiguang.liu@ucl.ac.uk; de Paula: a.paula@ucl.ac.uk; Tamer: elietamer@fas.harvard.edu. We thank participants at the 7th International Conference on Econometrics and Statistics, the CeMMAP Econometrics Research Day at UCL, the 2024 INFORMS Annual Meeting, as well as the Econometrics Seminars at University College London, the University of Cambridge, and the University of Manchester for their valuable feedback. We thank the generous funding from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant Ref: EP/X02931X/1) and the Economic and Social Research Council (ESRC) funding through the ESRC Institute for the Microeconomic Analysis of Public Policy (Grant Ref: ES/T014334/1).

1 Introduction

Interval data is pervasive. Surveys usually employ "bracketing" strategies to address item non-response. For example, respondents are often asked whether their money in savings accounts, in investments or their income fall in a sequence of brackets such as $[\$10K, \$20K]$, $[\$20K, \$50K]$ etc., while also given the option of providing a point answer to such questions (see Heeringa and Suzman (1995) and Moore and Loomis (2001)). Interval censored data also accommodate often found patterns such as right censoring, where an outcome variable is greater than a specified value (i.e., $Y \in [Y^L, \infty))$); left censored data, where an outcome variable is less than a certain value (i.e., $Y \in (-\infty, Y^U])$); and data from competing risks, where one typically observes the max or min of variables of interest. Such data typically involve outcomes of interest such as income, wealth, unemployment duration and cause-specific mortality. This paper provides inference procedures for the prediction of a variable of interest Y when we observe $[Y^L, Y^U] \ni Y$ instead.

More specifically, suppose we have a random variable Y defined on $\mathcal{Y} \subset \mathbb{R}$ with distribution P. One parameter of interest is a *prediction set* C defined as the set of $y \in \mathcal{Y}$ such that $P(y \in C) \geq 1 - \alpha$ for $\alpha \in (0, 1)$ where we say that such a set C has miscoverage level $\alpha \in (0, 1)$. The oracle prediction set is the set with the smallest volume that has this (mis)coverage property and is related to the level set of the density of Y (when this density exists). With covariates, we have access to a sample $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$, with $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$. We use these data to construct a prediction set $C_n(x) \equiv C(x; X_1, Y_1, \ldots, X_n, Y_n) \subset \mathcal{Y}$ for a given miscoverage level α at $x \in \mathcal{X}$. It is desirable that suc prediction set C_n is both consistent for the oracle prediction set and has a *finite sample coverage property:*

$$\mathrm{P}(Y_{n+1} \in C_n(X_{n+1})) \geq 1 - \alpha, \tag{1}$$

where P is the distribution over (X_i, Y_i) for $i = 1, \ldots, n+1$. This paper constructs such a set when Y is interval censored. The exact coverage property that our construction provides is given in more details in Section 2 and Section 4 below.

Such an "oracle prediction" exercise is familiar in econometrics especially in the context of forecasting in linear regressions (see Diebold 2015). For example, assuming that the outcome Y is such that $Y_i|X_i \sim \mathcal{N}(X_i'\beta, \sigma^2)$ with known σ^2, the population oracle prediction set is $X_{n+1}'\beta \pm 1.96\sigma$. Given data $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$, a consistent estimator for such a set would be $X_{n+1}'\hat{\beta} \pm 1.96\sigma$ where $\hat{\beta}$ is the least squares estimator for β. On the other hand, "an operational density forecast that accounts for parameter uncertainty is" (Diebold 2015 p. 250)

$$N\left(X_{n+1}'\hat{\beta},\ \sigma^2\left(1 + X_{n+1}'(\mathbf{X}'\mathbf{X})^{-1}X_{n+1}\right)\right),$$

This "operational density" summarizes both the sampling uncertainty in estimating $\hat{\beta}$ and the disturbance uncertainty that comes from the intrinsic randomness in the conditional distribution $Y_{n+1} \mid X_{n+1}$. One can use it to derive a forecast density with a particular coverage property. In this paper, we use the recent theory on conformal inference to show formally that a particularly constructed predictive set C_n has finite sample coverage property when Y is censored and without parametric assumptions, while enjoying the asymptotic property of being consistent for the oracle prediction set under the partially identified true population distribution. This conformal inference procedure in turn requires the definition of particular score functions to handle interval censoring. This score is similar to ones used in the conformal literature but is here modified to handle censoring. In particular, the score function we use is negative when the constructed prediction set contains the interval outcomes, but is otherwise positive, and as a result accounts for both over- and under-coverage.

In order to obtain the proposed conformal prediction set, we first characterise the oracle prediction sets under interval censoring and provide a feasible optimisation problem that relates the oracle prediction sets to the conditional distribution of the observed upper and lower brackts, (Y^L, Y^U). We use a kernel based conditional distribution function estimator as a basis for the construction of a consistent set estimator for the oracle prediction set. Interestingly, the oracle prediction set may be the union of a collection of disjoint intervals in cases for instance where the prediction density is multimodal. We provide an approach that allows for the set estimator to be a union of disjoint intervals with the prescribed coverage. In cases with a multimodal density, a union of intervals can have a smaller volume than one interval.

We then adapt recent work on conformal inference to propose a conformal prediction set using a sample of exchangeable data. This proposed construction is shown to have guarantees for finite sample coverage and to achieve asymptotic efficiency in terms of the volume of the prediction set. As a competing alternative prediction set estimator, we show in the appendix that aconformal quantile regression method provides a valid prediction set for interval censored outcomes which can be easier to implement in practice, but is in general not efficient in terms of the volume of the prediction set. In addition, quantile regression based methods are not suitable when the prediction set is not a single interval.

Interval censoring may arise from fixed bracketing and random censoring from "unfolding brackets" or survey schemes, see Heeringa and Suzman (1995) and Moore and Loomis (2001). See also the bracketing approach to outcomes using the Heatlh and Retirement Survey in Manski and Tamer (2002). We thus illustrate our contruction using a set of Monte Carlo experiments, where we consider both fixed and random censoring. Our proposed prediction set is shown to have adequate finite sample coverage properties and to have smaller volume compared with the alternative prediction sets constructed with quantile regression methods. In addition, we construct prediction sets using the Current Population Survey (CPS) data. Among people who report their income, a significant fraction (more than 20%) of the observations are interval censored. Interval censored observations are sometimes discarded or handled by imputation methods that rely on assumptions such as missing at random. Our proposed method prediction set is robust in the sense that we do not require a specification of the censoring mechanism and we demonstrate using a hold-out subset of the CPS data that our prediction set has better finite sample coverage of the interval censored incomes compared with prediction sets constructed with hot-deck imputation in place. Naturally, without relying on imputation algorithms, our oracle prediction sets are wider than ones obtained with uncensored outcomes, but our constructed sets are more robust and hence more trustworthy.

Literature review

The paper combines insights from the partial identification literature in econometrics and conformal inference results from the recent statistics literature. The partial identification literature in econometrics (see Manski 2003) takes the view that one should try and make inferences under minimal assumptions, even if these assumptions do not allow for point identification of the parameter of interest. The motivation for such a view is that inferences are less credible and conclusions are suspect when these rely on non-testable or non credible assumptions. In our context for example, if we are interested in $\beta = E[Y]$ and we observe $[Y^L, Y^U]$ rather than Y, then a partial identification approach would just say that, given the data, the identified set for β (or all we can learn about β given the data) is the interval $\left[E[Y_0], E[Y_1]\right]$. Imputation based inference uses models to predict a Y^P within $[Y^L, Y^U]$ and takes $E[Y^P]$ as a proxy for $\beta = E[Y]$. The credibility of this exercise depends on the imputation model, i.e. the set of assumptions it relies on. In the context of interval data, Manski and Tamer (2002) study identification of linear regression coefficients when the data (on regressors or outcomes) is interval censored using a partial identification approach. See Tamer (2010), Molinari (2020), and Kline and Tamer (2023) for reviews on the developments in the partial identification literature.

In order to obtain a prediction interval that has finite-sample coverage guarantees, we use the method of conformal inference, which is a distribution-free method that builds upon the idea of exchangeability and permutation tests (Vovk, Gammerman, and Shafer 2005, Vovk, Nouretdinov, and Gammerman 2009). The idea of conformal inference has become increasingly popular in the machine learning literature since it requires no assumptions on the underlying distribution other than the samples being exchangeable and it also allows the use of any predictor that treats the data symmetrically and hence preserves exchangeability (Barber, Candès, et al. (2023) discusses the validity of conformal inference when the data is not i.i.d. and the predictor is not symmetric). Conformal inference has been applied to obtain valid distribution-free prediction sets (Lei, Robins, and Wasserman 2013) and has been extended to non-parametric regression (Lei and Wasserman 2014) and high-dimensional settings (Lei, G'Sell, et al. 2018). Romano, Patterson, and Candes (2019) considers conformal quantile regression to construct prediction intervals instead of conditional mean regression. Sadinle, Lei, and Wasserman (2019) proposes a way to predict the label of Y. Chernozhukov, Wüthrich, and Zhu (2021) considers construction of valid prediction intervals with distributional regression method. We add to this literature by allowing the outcome observations to be interval-censored or interval-valued.

The classical way to construct estimator for the oracle prediction set relies on the level set estimation, (see, for example, Wilks 1941, Polonik 1995 and Samworth and Wand 2010). The uniform consistency of the nonparametric estimator of the conditional distribution is an extension of classical results on the uniform convergence rate of the kernel-based density and regression estimators such as Einmahl and Mason (2000), Gine (2002) and Hansen (2008). This paper is also related to the literature on quantile regression (see Koenker 2017 for a review).

Outline

This paper is organised as follows. In Section 2, we provide the setup and definitions of optimal sets and approaches to obtaining estimators of such sets with censoring. Section 3 provides a study of consistency of such sets using kernel estimators of the conditional distribution. Section 4 provides results on finite sample coverage using a conformalized procedure under interval censoring. Section 5 provides some Monte Carlo evidence and Section 6 applies our methods to US CPS Data. Section 7 concludes. The proofs for the theorems in the main text as well as some further results and comments are collected in the Appendix.

2 Setup and definitions

Suppose there is a random sample $\bar{\mathcal{D}} = \{\bar{Z}_i = (X_i, Y_i, Y_i^L, Y_i^U) \in \mathbb{R}^{d+3} : i \in \mathcal{I}\}$ with sample size $|\mathcal{I}| = n$ drawn independently from a joint distribution P.[1] We only observe the subvector $Z_i = (X_i, Y_i^L, Y_i^U)$, and the latent outcome Y_i is not directly observed but assumed to satisfy the condition $Y_i^L \leq Y_i \leq Y_i^U$. For each $i \in \mathcal{I}$, the predictor X_i is a d-dimensional random vector with support $\mathcal{X} \subset \mathbb{R}^d$. The latent outcome Y_i and the observed lower and upper bounds Y_i^L and Y_i^U are one-dimensional. Let $\bar{Z} = (X, Y, Y^L, Y^U)$ denote a generic sample from the joint distribution P. We will use notation such as P_Y to denote the marginal distribution of Y, and $P_{Y|X}$ to denote the distribution of Y conditional on X. Note here that the joint distribution P_{X,Y^L,Y^U} of (X, Y^L, Y^U) is identified given the observed sample $\mathcal{D} = \left\{Z_i = (X_i, Y_i^L, Y_i^U) \in \mathbb{R}^{d+2} : i \in \mathcal{I}\right\}$, but the joint distribution $P_{X,Y}$ of predictors X and latent outcome Y is not point identified given the

[1] In order to perform conformal inference, it is sufficient that the sample satisfies the weaker condition of exchangeability. Here independence is assumed for simplicity and for establishing the asymptotics for the nonparametric estimators we will use. Barber, Candès, et al. (2023) studies conformal inference beyond exchangeability and is applicable for time series data. Our results can be readily extended to the time series settings combined with their procedure.

observed sample $\mathcal{D}$ without restrictions on the censoring mechanism.

A prediction set is defined as a set $C \subset \mathcal{X} \times \mathcal{Y}$, with the interpretation that, given a realisation of the predictor $X = x$, we predict that Y is likely to fall inside the section $C(x) = \{y : (x, y) \in C\}$. In the absence of interval censoring, estimation of a prediction set that satisfies certain optimality criteria (i.e. the minimal prediction set with coverage guarantee) has been studied in the literature; see Lei and Wasserman (2014), Barber, Candes, et al. (2021) among others. As we discuss later in this section, the oracle prediction set without interval censoring is related to the upper level set of the density function of Y given X under point identification.

The situation is significantly different when we are faced with interval censoring and $P_{X,Y}$ is partially identified. In the rest of this section, we first define an optimality criterion for a prediction set under censoring, and analyse why the classical oracle prediction set (which corresponds to an upper level set) is no longer applicable under censoring. We then propose a feasible estimation procedure for the oracle prediction set. We discuss the potential reasons for why the procedure can result in conservative prediction sets and show that the proposed procedure does not incur any unnecessary conservativeness.

There are two criteria for an oracle prediction set: validity and efficiency (see Vovk, Nouretdinov, and Gammerman 2009; Lei, Robins, and Wasserman 2013). Let $\mathcal{P}_I$ be the identified set of the joint distribution of (X, Y) under censoring. That is, $\mathcal{P}_I$ is the set of joint distributions of (X, Y) that are compatible with the joint distribution P_{X,Y^L,Y^U} and some censoring mechanism. A prediction set C is valid if it satisfies the following coverage properties for a new sample (X_{n+1}, Y_{n+1}) drawn from the same distribution P at a given miscoverage level α.

Definition 2.1 (Validity under Partial Identification). We say a prediction set C is *marginally valid under partial identification* with miscoverage level α, if,

$$\inf_{P \in \mathcal{P}_I} P\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha, \tag{2}$$

and C is *conditionally valid under partial identification* with miscoverage level α if for almost everywhere $x \in \mathcal{X}$,

$$\inf_{P \in \mathcal{P}_I} P\left(Y_{n+1} \in C(x) \mid X_{n+1} = x\right) \geq 1 - \alpha. \tag{3}$$

When the model is point identified, $\mathcal{P}_I = \{P_{X,Y}\}$ is a singleton set, the conditions in Equations (2) and (3) reduce to the standard conditions of validity, see for example, Lei and Wasserman (2014). We will simply refer to these two conditions as marginal validity and conditional validity henceforth.

It is natural to require a prediction set to be valid, so that we have confidence it will contain the true value of the outcome with a certain probability. However, the requirement of validity by itself is vacuous. For example, the prediction set C such that $C(x) = (-\infty, \infty)$, for all $x \in \mathcal{X}$ is technically valid at any miscoverage level but it is also uninformative. Therefore, our target parameter of interest is an *oracle prediction set*, denoted by $C^*_{\mathcal{P}_I}$ which is defined as the minimal-volume prediction set that satisfies the validity conditions in Definition 2.1. For simplicity, we will omit the subscript and write $C^* = C^*_{\mathcal{P}_I}$ whenever the dependence on the identified set $\mathcal{P}_I$ is clear. Since the minimal-volume prediction set is intuitively the most informative among all valid prediction sets, this property is also called *efficiency.*[2]

For the sake of measurability and interpretability, we will focus on the prediction sets such that $C(x)$ takes the form of a union of a (potentially unknown) number of disjoint intervals for each $x \in \mathcal{X}$. Specifically, we assume that $C(x) \in \mathcal{C}$, P_X-almost surely, where

$$\mathcal{C} = \left\{ \bigsqcup_{m=1}^{M} [a_m, b_m] : a_m \leq b_m < a_{m+1}, M < \infty \right\}$$

denotes the collection of all finite disjoint union of closed intervals, and the symbol $\sqcup$ indicates that the union is disjoint. The following oracle prediction set under partial identification will be our estimation target.

Definition 2.2 (Oracle prediction Set under Partial Identification). The *oracle prediction set under partial identification* for a given miscoverage level α is defined as $C^* = C^*_{\mathcal{P}_I} \subset \mathcal{X} \times \mathbb{R}$ which solves the following problem for P_X-almost surely $x \in \mathcal{X}$,

$$C^*(x) = \arg \min_{C(x) \in \mathcal{C}} \mu(C(x)) \quad \text{s.t.} \quad \inf_{P \in \mathcal{P}_I} P\left(Y \in C(x) \mid X = x\right) \geq 1 - \alpha, \tag{4}$$

[2]This notion of efficiency differs from the usual definition, where an estimator is considered efficient if it has the smallest variance within a class of estimators. However, the concepts are related, as an estimator with smaller variance generally leads to the construction of tighter and more informative confidence intervals.

where μ denotes the Lebesgue measure. Since conditional validity implies marginal validity, the oracle predition set will also be marginally valid under partial identification.

Before proposing a feasible estimation procedure for the oracle prediction set C^*, we explain why the existing estimation methods are not applicable for prediction sets under point identification in the case of censoring. This analysis highlights the challenges posed by partial identification and the difference between our approach and the classical level set method studied in the literature.

When the joint distribution of (X, Y) is point identified, and $\mathcal{P}_I = \left\{P_{X,Y}\right\}$, it is well known in the literature that the oracle prediction set is related to the upper level set of the conditional density of Y assuming this density is well defined. Let $p(y|x)$ denote the conditional density of Y at y given $X = x$. An upper level set for $p(y \mid x)$ with threshold λ at $x \in \mathcal{X}$ is defined as,

$$L(x, \lambda) = \left\{y : p(y \mid x) \geq \lambda\right\}.$$

Let λ_x^α be chosen such that

$$\int \mathbb{I}\left\{p(y \mid x) \geq \lambda_x^\alpha\right\} p(y \mid x)\, \mathrm{d}y = 1 - \alpha.$$

The oracle prediction set under point identification is $C^*_{P_{Y|X}} = \left\{(x, y) : x \in \mathcal{X}, y \in L(x, \lambda_x^\alpha)\right\}$ which is both conditionally and marginally valid. For a given x, it collects the values of y with sufficiently high density and intuitively assembles the set of y's with the highest probability for a given volume. Once this is calibrated to yield the desired miscoverage level, one obtains a minimal-volume set. Estimating the oracle prediction set $C^*_{P_{Y|X}}$ therefore reduces to the problem of estimating the upper level sets $L(x, \lambda)$ of the conditional density $p(y \mid x)$ and selecting an appropriate threshold parameter.

However, this characterisation of the oracle prediction set in terms of level sets is no longer applicable under interval censoring as $P_{Y|X}$ is not identified. In fact, the next proposition states that no meaningful bounds can be obtained for the level set of the conditional density of Y given X under interval censoring. For brevity, we temporarily drop the conditioning predictors X in the following proposition. (It can be interpreted as conditional on $X = x$ for $x \in \mathcal{X}$.)

Proposition 2.1. *Suppose Y is interval censored, for any $\lambda < \infty$ and any $y \in [a, b]$, where $a < b$ and $(a, b) \in \mathbb{R}^2$ is in the support of (Y^L, Y^U), there exists some $P' \in \mathcal{P}_I$, such that $y \in L'(\lambda)$, where $L'(\lambda)$ is the upper level set for the density $p'(y)$ of P'.*

Proposition 2.1 states that for any point $y \in [a, b]$ such that (Y^L, Y^U) has positive density at $(a, b) \in \mathbb{R}^2$, we can find a random variable Y' that is observationally equivalent to Y under censoring, and this Y' admits a level set that contains y for any threshold λ. This is possible because one can adjust the density of Y' at any point y satisfying the stated conditions to be arbitrarily large while ensuring that this density remains within the identified set P_I.[3] This lack of identification for level sets is due to the possibility of pointwise pathological behaviour of the density function.

In contrast to the inability to bound the density function, the following two lemmas characterise the identified set of the conditional distribution of Y given X under interval censoring. Lemma 2.2 relates the sharp identified set of the conditional distribution of Y given X to the conditional distribution of (Y^L, Y^U), which can be directly identified and estimated from the data, and provides a feasible lower bound for the restriction in Equation (4). Lemma 2.3 shows that the lower bound is tight, and we are not unnecessarily conservative in constructing the oracle prediction set.

Lemma 2.2 (Theorem SIR-2.3 in Molinari (2020)). *The sharp identified set for the conditional distribution of Y given X, under interval censoring is given by*

$$\left\{P_{Y|X} : P_{Y|X}(Y \in [t_0, t_1] \mid X = x) \geq P\left([Y^L, Y^U] \subset [t_0, t_1] \mid X = x\right), \forall t_0 \leq t_1\right\}. \tag{5}$$

Lemma 2.2 suggests that we could implement a feasible version of Equation (4), by replacing the optimisation constraint $\inf_{P \in \mathcal{P}_I} P\left(Y \in C(x) \mid X = x\right) \geq 1-\alpha$ with $P\left([Y^L, Y^U] \subset C(x) \mid X = x\right) \geq$

[3] For example, consider a random variable Y with support on $\mathbb{R}$ that is censored within $[0, 1]$: we do not observe the value of Y when it lies in $[0, 1]$ but only observe the indicator $1[Y \in [0, 1]]$. In this case, we cannot reject the hypothesis that *any* value $y \in [0, 1]$ belongs to the level set of Y.

$1-\alpha$, which implies the original constraint, since for $C \in \mathcal{C}$, we can write $C = \sqcup_{m=1}^{M}[a_m, b_m]$

$$\begin{aligned}
\inf_{P\in\mathcal{P}_I} P\left(Y \in C \mid X = x\right) &\geq \sum_m \inf_{P\in\mathcal{P}_I} P\left(Y \in [a_m, b_m] \mid X = x\right) \\
&\geq \sum_m P\left([Y^L, Y^U] \subset [a_m, b_m] \mid X = x\right) \\
&= P\left([Y^L, Y^U] \subset C \mid X = x\right)
\end{aligned}$$

for a disjoint union of intervals $C \in \mathcal{C}$. As a result, a prediction set that satisfies the feasible version of the constraint will have coverage guarantee since

$$\begin{aligned}
P_{Y|X}(Y \in C \mid X = x) &\geq \inf_{P\in\mathcal{P}_I} P\left(Y \in C \mid X = x\right) \\
&\geq P\left([Y^L, Y^U] \subset C \mid X = x\right) \geq 1-\alpha.
\end{aligned}$$

A prediction set C is conservative if one of the three inequalities is strict. The first inequality is due to the partial identification of the joint distribution of (X, Y) and the last inequality is determined by the distribution of (X, Y^L, Y^U). Both are intrinsic properties and cannot be improved without imposing additional restrictions. Lemma 2.3 shows that for $C \in \mathcal{C}$, the second inequality is an equality and hence implementing the feasible version of the optimisation problem induces no additional conservativeness.

Lemma 2.3. *For P_X-almost everywhere $x \in \mathcal{X}$, and any $C \in \mathcal{C}$, we have under interval censoring,*

$$\inf_{P\in\mathcal{P}_I} P\left(Y \in C \mid X = x\right) = P\left([Y^L, Y^U] \subset C \mid X = x\right).$$

The above lemma shows that even though the oracle prediction set can be conservative, i.e. overcovering Y for the distribution actually generating the data, the conservativeness is only due to partial identification and the conditional distribution of (Y^L, Y^U) and not because of the constraint we focus on differs from the original one.

In fact, given Lemmas 2.2 and 2.3, we could replace the condition in Equation (4) which depends on the (unobserved) conditional distribution of the latent outcome Y with a condition that depends on the conditional distribution of (Y^L, Y^U), which can be estimated given the observed sample $\mathcal{D}$. We then arrive at the following feasible estimation problem:

Proposition 2.4 (Feasible Estimation). *The oracle prediction set C^* is the solution to the following feasible optimisation problem*

$$C^*(x) = \arg\min_{C(x)\in\mathcal{C}} \mu(C(x)) \quad s.t. \quad P\left([Y^L, Y^U] \subset C(x) \mid X = x\right) \geq 1-\alpha. \tag{6}$$

In the special case when C^* is a single interval, we will denote it as C^*_I.

A natural estimator $\hat{C}$ for the oracle prediction set can be obtained by solving the sample version of the optimisation problem above. The next two sections study the consistency of such estimators for the oracle prediction set, and provides a modification of the estimated prediction set that has finite sample coverage guarantees using conformal inference.

3 Consistency

We first consider the estimation of the oracle prediction interval $C^*_I = [\tau_0(x), \tau_1(x)]$, for $x \in \mathcal{X}$. This analysis highlights the conditions required to ensure the consistency of the prediction set under different schemes of interval censoring. Subsequently, we will extend our analysis to the estimation of oracle prediction sets $C^* \in \mathcal{C}$ consistinig of multiple intervals. To measure the difference between two sets $A, B \subset \mathbb{R}$, we use the volume of their symmetric difference, denoted by $\mu(A \triangle B)$. The symmetric difference is defined as $A \triangle B = (A \setminus B) \cup (B \setminus A)$. This is commonly used in the literature on level set estimation and conformal inference. Another commonly used notion of set distance is the Hausdorff distance. In Appendix D, we discuss the relationship between these two metrics and why the Hausdorff distance is unsuitable in certain cases.

Let $\hat{C}_I(x) = [\hat{\tau}_0(x), \hat{\tau}_1(x)]$ denote the solution to the empirical analogue of the optimisation problem defined in Proposition 2.4 given a random sample $\mathcal{D}$

$$\min_{t_0\leq t_1} |t_1 - t_0| \quad \text{s.t.} \quad P_n(t_0, t_1; x) \geq 1-\alpha. \tag{7}$$

Here we use the shorthand notation $P(t_0, t_1; x) = P(t_0 \leq Y^L \leq Y^U \leq t_1 \mid X = x)$, and $P_n(t_0, t_1; x)$ is an estimator for $P(t_0, t_1; x)$ based on the random sample $\mathcal{D} = \{(X_i, Y_i^L, Y_i^U) : i \in \mathcal{I}\}$ with sample size n.

Assumption 3.1 (Estimation). *The estimator $P_n(t_0, t_1; x)$ is consistent for $P(t_0, t_1; x)$ uniformly over $(t_0, t_1, x) \in \mathbb{R}^2 \times \mathcal{X}_n$, for some $\mathcal{X}_n \subset \mathcal{X}$,*

$$\sup_{x \in \mathcal{X}_n} \sup_{t_0 \leq t_1} \left| P_n(t_0, t_1; x) - P(t_0, t_1; x) \right| = o_p(1).$$

This assumption does not specify the estimator P_n we use. For example, a candidate estimator $P_n(t_0, t_1; x)$ based on kernel smoothing is

$$P_n(t_0, t_1; x) := \frac{\sum_{i \in \mathcal{I}} \mathbb{I}\left\{ t_0 \leq Y_i^L \leq Y_i^U \leq t_1 \right\} K_h\left(X_i - x\right)}{\sum_{i \in \mathcal{I}} K_h\left(X_i - x\right)}, \tag{8}$$

where $K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$ for a choice of kernel smoothing function $K(\cdot)$ and bandwidth parameter h, see Tsybakov (2009) and Hansen (2008). Another candidate is the distributional regression estimator (see, e.g., Foresi and Peracchi 1995; Chernozhukov, Fernández-Val, and Melly 2013; Chernozhukov, Wüthrich, and Zhu 2021).

Given a miscoverage level $\alpha \in (0, 1)$, the next assumption we make is that $C_I^*(x) = [\tau_0(x), \tau_1(x)]$ is the unique interval that satisfies the $1 - \alpha$ coverage condition with the shortest length.

Assumption 3.2 (Identification). *There exists a unique solution $C_I^*(x) = [\tau_0(x), \tau_1(x)]$ to the optimisation problem in Proposition 2.4: that is, $P(\tau_0(x), \tau_1(x); x) \geq 1 - \alpha$, and, for any $(t_0, t_1) \neq (\tau_0(x), \tau_1(x))$ such that $t_1 - t_0 \leq \tau_1(x) - \tau_0(x)$, then $P(t_0, t_1; x) < 1 - \alpha$. Additionally, for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all x, if $t_1(x) - t_0(x) \leq \tau_1(x) - \tau_0(x)$ and $\mu([t_0(x), t_1(x)] \triangle [\tau_0(x), \tau_1(x)]) > \varepsilon$, then $P(t_0, t_1; x) < P(\tau_0, \tau_1; x) - \delta$.*

Assumption 3.2 states that we cannot achieve the same coverage probability with a different interval $(t_0(x), t_1(x))$ that is not larger than $(\tau_0(x), \tau_1(x))$. Figure 1 shows a specific joint density of Y^L and Y^U. The set $\{(a, b) : t_0 \leq a \leq b \leq t_1\}$ is a triangular region in $\mathbb{R}^2$. In Figure 1, the plotted triangular region corresponds to $t_0 = -0.75$ and $t_1 = 0.5$. $P(t_0, t_1; x)$ will be the integral of the joint density over the triangular region. The first part of Assumption 3.2 thus requires that if either the triangular region is smaller or if we move the triangular region away from the optimal position, we will strictly lose coverage probability of (Y^L, Y^U); and the second part is a regularity condition ensuring that the "coverage loss" is bounded away from 0 over the support of X. This rules out the case when $P(t_0, t_1; x)$ gets flatter around $(\tau_0(x), \tau_1(x))$ for some x.

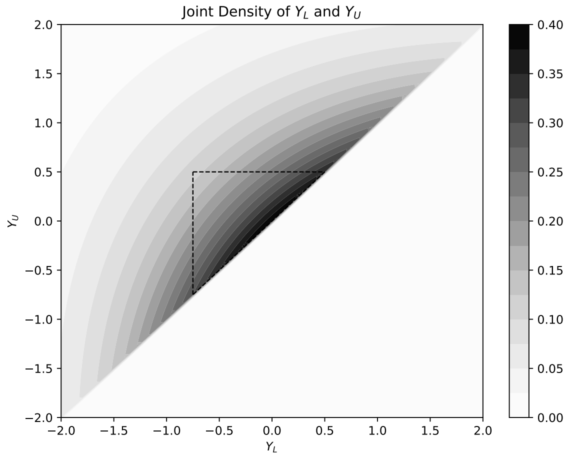


Figure 1: Joint density of Y^L and Y^U when $Y \sim N(0, 1)$, $\kappa^L, \kappa^U \sim \text{Exp}(1)$. Integral of the density over the triangular region corresponds to $P(-0.75, 0.5; x)$.

In the absence of fixed censoring, we assume the following regularity condition for the conditional distribution $P(t_0, t_1; x)$. The case with fixed censoring will be discussed later.

Assumption 3.3 (Regularity). *There exists an $\varepsilon_0 > 0$ and positive constants c_1, γ_1, such that for all $x \in \mathcal{X}$ and $\tau_0(x), \tau_1(x)$ as defined in Assumption 3.2,*

$$P\left(\tau_0(x) - \varepsilon, \tau_1(x) + \varepsilon; x\right) - P\left(\tau_0(x), \tau_1(x); x\right) \geq c_1 \varepsilon^{\gamma},$$

for any $0 < \varepsilon < \varepsilon_0$. In addition, for all $\varepsilon < \varepsilon_0$ and $t_0 \leq t_1$, $P(t_0 - \varepsilon, t_1 + \varepsilon; x) - P(t_0, t_1; x) \leq c_2 \varepsilon^{\gamma_2}$ for some constant c_2, γ_2.

Assumption 3.3 is a smoothness assumption on $P(t_0, t_1; x)$. In particular, we are assuming that by considering a slightly larger interval $[\tau_0 - \varepsilon, \tau_1 + \varepsilon]$ we can strictly increase our coverage probability. Assumption 3.3 also assumes an upper bound on the desity, which can be violated if the conditional distribution of (Y^L, Y^U) is discrete, which might happen under a fixed censoring scheme.

This assumption is related to the γ-exponent condition in the level set literature, see Polonik (1995) and Lei and Wasserman (2014).

The following theorem shows that the estimated prediction interval is consistent for the oracle prediction interval under random censoring.

Theorem 3.1. *Under Assumptions 3.1, 3.2, and 3.3, the prediction interval estimator* $\hat{C}_I(x)$ *defined in Equation* (7) *is uniformly consistent for the oracle prediction interval* $C_I^*(x)$*, that is,*

$$\sup_{x \in \mathcal{X}_n} \mu\left(\hat{C}_I(x) \triangle C_I^*(x)\right) = o_p(1)$$

as the sample size $n \longrightarrow \infty$.

Assumption 3.3 can be violated when there is fixed censoring, i.e., when there are some fixed brackets and we observe only the bracket an outcome falls in. Suppose the conditional distribution of (Y^L, Y^U) can be decomposed into a continuous part and a discrete part, where the discrete part has support over the set of points $\{(L_k, U_k) \in \mathbb{R}^2 : 1 \leq k \leq K\}$.

In the case of a mixture of random and fixed censoring, the regularity condition in Assumption 3.3 is violated. We impose the following assumption which slightly modifies the estimator to allow for non-smoothness of the conditional distribution.

Assumption 3.4 (Approximate solution). *The estimator for the oracle prediction interval* $\hat{C}_I(x) = [\hat{\tau}_0(x), \hat{\tau}_1(x)]$ *is constructed by solving the following optimization problem, for* $x \in \mathcal{X}$,

$$\min_{t_0 < t_1} t_1 - t_0 \quad s.t. \quad P_n(t_0, t_1; x) \geq 1 - \alpha - \psi_n,$$

for some $\psi_n > 0$ *and* $\psi_n \longrightarrow 0$*, such that for some* $\mathcal{X}_n$,

$$P\left(\sup_{x \in \mathcal{X}_n} \sup_{t_0 \leq t_1} \left|P_n(t_0, t_1; x) - P(t_0, t_1; x)\right| > \psi_n\right) = o(1).$$

To see the reason why fixed censoring creates a problem and we should not solve the exact constraint, consider the following example.

Example 3.1. Suppose $\alpha = 0.5$, and $(Y_i^L, Y_i^U) = (0, 1)$ or $(0, 2)$ with equal probability, which is a discrete distribution. The oracle prediction set is $[0, 1]$. With the natural estimator $P_n(a, b; x) = n^{-1} \sum \mathbb{I}\left\{[Y_i^L, Y_i^U] \subset [a, b]\right\}$, if we solve the exact empirical analog of the oracle prediction set, as in Equation 7, we will obtain

$$\check{C} = \left[0, 1 + \mathbb{I}\left\{\frac{1}{n}\sum Y_i^U < 0.5\right\}\right],$$

and the random upper bound equals 1 or 2 with equal probability, hence $\check{C}$ is not consistent.

With Assumption 3.4 in place of Assumptions 3.1 and 3.3, we obtain the following result:

Theorem 3.2. *Under Assumptions 3.2, and 3.4, we have the uniform consistency of $\hat{C}_I$ for the oracle prediction interval C_I^*, that is, as the sample size $n \longrightarrow \infty$,*

$$\sup_{x \in \mathcal{X}_n} \mu\left(\hat{C}_I(x) \triangle C_I^*(x)\right) = o_p(1).$$

We now turn to the more general case where the estimation target is the oracle prediction set $C^* \in \mathcal{C}$ defined in Proposition 2.4.

The next assumption is on the identification of the oracle prediction set, which generalises Assumption 3.2. It additionally assumes that the number of disjoint intervals in the oracle prediction set is bounded.

Assumption 3.5. *(1) For any $\varepsilon > 0$ and P_X-almost surely $x \in \mathcal{X}$, there exists $\delta = \delta(\varepsilon) > 0$, such that for any set $C'(x) \in \mathcal{C}$ with $\mu(C'(x)) \leq \mu(C^*(x))$ and $\mu(C^*(x) \triangle C'(x)) > \varepsilon$, we have $P([Y^L, Y^U] \subset C'(x) \mid X = x) < 1 - \alpha - \delta$. (2) The oracle $C^*(x) \in \mathcal{C}_{\bar{M}} = \{\sqcup_{m=1}^{M}[a_m, b_m] : M' \leq \bar{M} < \infty\}$ for some $\bar{M} < \infty$ and all $x \in \mathcal{X}$.*

With the notation $P(C(x); x) = P([Y^L, Y^U] \subset C(x) \mid X = x)$, and suppose that we have a uniform consistent estimator $P_n(C; x)$. We define the estimator $\hat{C}(x), x \in \mathcal{X}$, for the oracle prediction set C^* allowing for potential fixed censoring.

Assumption 3.6. *The estimators $\hat{C}(x), x \in \mathcal{X}$ are constructed by solving the following problem, for any $x \in \mathcal{X}$, and $M \geq \bar{M}$,*

$$\min_{C(x) \in \mathcal{C}_M} \mu(C(x)) \quad s.t. \quad P_n(C(x); x) \geq 1 - \alpha - \psi_n.$$

Here $\psi_n = o(1)$ is a sequence of positive numbers and P_n is a consistent estimator of P such that

$$P\left(\sup_{C\in\mathcal{C}_M}\sup_{x\in\mathcal{X}_n}\left|P_n(C(x);x)-P(C(x);x)\right|>\psi_n\right)=o(1),$$

as $n\longrightarrow\infty$.

Remark 1. *Appendix B states conditions under which a kernel smoothing estimator $P_n(C;x)$ is uniformly consistent and provides the corresponding convergence rate. This convergence rate can be used to inform the choice of ψ_n; in practice, it is often sufficient to simply set $\psi_n = 0$. The uniformity is restricted to the class $\mathcal{C}_M$ rather than the larger class $\mathcal{C}$ because the proof relies on arguments that are closely tied to $\mathcal{C}_M$ being a VC class of sets, whereas $\mathcal{C}$ is not. Notably, knowledge of $\bar{M}$ is not required, as it suffices to select an M that is sufficiently large. Even if M is smaller than the true $\bar{M}$, the resulting prediction set can be viewed as an approximation for the oracle.*

Theorem 3.3. *Under Assumptions 3.5 and 3.6, we have that $\hat{C}$ is a consistent estimator of C^* in the sense that, as $n\longrightarrow\infty$,*

$$\sup_{x\in\mathcal{X}_n}\mu\left(\hat{C}(x)\,\triangle\,C^*(x)\right)=o_P(1).$$

We have proposed consistent estimators for the oracle prediction set. There has been a recent interest in constructing prediction sets that are finite-sample valid via conformal inference. The next section discusses how to construct a conformal prediction set that is finite-sample valid under interval censoring.

4 Conformal inference with censored outcomes

Given a consistent estimator for the oracle prediction set $\hat{C}$, this section considers the problem of constructing a prediction set with finite-sample validity. An estimated prediction set $\tilde{C}$ with a random sample $\mathcal{D}=\left\{(X_i,Y_i^L,Y_i^U), i\in\mathcal{I}\right\}$ is said to be *finite-sample marginally valid*, if for all P,

$$\mathrm{P}\left(Y_{n+1}\in\tilde{C}(X_{n+1})\right)\geq 1-\alpha,$$

where P denotes the joint probability of $\left\{\bar{Z}_i=(X_i,Y_i,Y_i^L,Y_i^U):i\in\mathcal{I}\cup\{n+1\}\right\}$. Notice that this is the finite-sample validity condition typically considered in conformal inference literature where the probability is the joint distribution for both the observed sample and the individual to be predicted. While the oracle prediction set $C^*(x)$ depends on the true distribution P, conformal inference constructions lead to a prediction set that is finite-sample valid for all P, making it distribution-free.

In particular, we will achieve finite-sample validity using the method of *split conformal inference* (see Romano, Patterson, and Candes (2019) and Lei, G'Sell, et al. (2018)). One feasible split conformal inference procedure for the oracle prediction interval C_I^* in our setting can be described as follows. We first split the sample $\mathcal{D}$ into a training set $\mathcal{D}_1 = \{Z_i : i \in \mathcal{I}_1\}$ and a calibration set $\mathcal{D}_2 = \{Z_i : i \in \mathcal{I}_2\}$. We calculate the estimator $\hat{C}_I(x) = [\hat{\tau}_0(x), \hat{\tau}_1(x)]$ for the oracle prediction set with the training set for $x \in \mathcal{X}$ as in Equation 7. The following scores $s_j = s(Y_j^L, Y_j^U, X_j; \hat{\tau}_0, \hat{\tau}_1)$ are then computed for each $j \in \mathcal{I}_2$ in the calibration set, given $\hat{C}_I(x) = [\hat{\tau}_0(x), \hat{\tau}_1(x)]$, where

$$s\left(y^L, y^U, x; t_0(\cdot), t_1(\cdot)\right) = \max\left\{t_0(x) - y^L, y^U - t_1(x)\right\}$$

This score function is a modified version of the score function used in the case of conformalised quantile regression with no interval censoring studied in Romano, Patterson, and Candes (2019). The score function is negative when the interval $[y^L, y^U]$ is contained in $[t_0(x), t_1(x)]$ and positive otherwise. By allowing for negative scores, the conformal prediction set $\tilde{C}_I(x)$ penalises both overcoverage and undercoverage.

We can then compute the quantile

$$\vartheta_{1-\alpha} = q\left(\left\{s_j : j \in \mathcal{I}_2\right\}; (1-\alpha)(1 + 1/|\mathcal{I}_2|)\right),$$

where $q\left(A; \zeta\right)$ denotes the ζ-quantile of the set A of real numbers. Since observations in $\mathcal{I}_2$ are assumed i.i.d., if we take another point $(X_{n+1}, Y_{n+1}^L, Y_{n+1}^U)$, the hypothesis that it is drawn from the same distribution will not be rejected at $1-\alpha$ confidence level if $s_{n+1} \leq \vartheta_{1-\alpha}$. The conformal prediction set is then

$$\tilde{C}_I(x) = [\tilde{\tau}_0(x), \tilde{\tau}_1(x)] = \left[\hat{\tau}_0(x) - \vartheta_{1-\alpha}, \hat{\tau}_1(x) + \vartheta_{1-\alpha}\right]. \tag{9}$$

Intuitively speaking, when our estimated prediction set $\hat{C}_I$ is wider than necessary, the intervals $\{(Y_j^L, Y_j^U), j \in \mathcal{I}_2\}$ tend to be subsets of $\hat{C}_I$ and a higher proportion of the scores computed using the calibration set will tend to be negative. In this case, the chosen quantile of the scores $\vartheta_{1-\alpha}$ is negative, and the conformal prediction interval will correct for the over-coverage of $\hat{C}_I$.

When the prediction set $\hat{C}$ is a union of multiple intervals, we can construct a finite-sample valid conformal prediction set as follows. First we split the samples into a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$. We estimate $\hat{C} \in \mathcal{C}$ using data in the training set, and then compute the conformity scores$_j = s(Y_j^L, Y_j^U, X_j; \hat{C}_S)$, for each $j \in \mathcal{I}_2$, where for $C(x) = \sqcup_{m=1}^M [t_{0m}(x), t_{1m}(x)] \in \mathcal{C}$,

$$s(y^L, y^U, x; C) = \min_{m \leq M} \max \left\{ t_{0m}(x) - y^L, y^U - t_{1m}(x) \right\}. \tag{10}$$

Notice that $s_j \leq 0$ if and only if $[Y_j^L, Y_j^U] \subset [t_{0m}, t_{1m}]$ for some $m \leq M$. We can then compute $\vartheta_{1-\alpha} = q(\{s_j : j \in \mathcal{I}_2\}; (1-\alpha)(1 + 1/|\mathcal{I}_2|))$. The conformal prediction set is then [4]

$$\tilde{C}(x) = \bigcup_m [\hat{t}_{0m}(x) - \vartheta_{1-\alpha}, \hat{t}_{1m}(x) + \vartheta_{1-\alpha}]. \tag{11}$$

The next theorem states the finite-sample marginal validity of the conformal prediction set $\tilde{C}$ defined in Equation 11 above. It implies the validity of the conformal prediction interval $\tilde{C}_I$.

Theorem 4.1. *Under the assumption that $\{(X_i, Y_i, Y_i^L, Y_i^U) : i \in \mathcal{I} \cup \{n+1\}\}$ are i.i.d., the conformal prediction set $\tilde{C}$ in Equation 11 is finite-sample marginally valid, that is, for all P,*

$$\mathrm{P}\left(Y_{n+1} \in \tilde{C}(X_{n+1})\right) \geq 1 - \alpha.$$

where P *denotes the joint probability of* $\left\{\bar{Z}_i : i \in \mathcal{I} \cup \{n+1\}\right\}$.

The following theorem shows that the conformal prediction set $\tilde{C}(x)$ preserves the consistency of the estimator $\hat{C}(x)$ for the oracle prediction set. Recall the definition of $s(y^L, y^U, x; C)$ in Equation (10), let $s_j = s(Y_j^L, Y_j^U, X_j; \hat{C})$, and $s_j^* = s\left(Y_j^L, Y_j^U, X_j; C^*\right)$.

[4]Here, the endpoints of the estimated $\hat{C}$ are adjusted uniformly using the same $\vartheta_{1-\alpha}$. A more sophisticated adjustment method is discussed in Appendix C, where we propose a minimal Euclidean distance adjustment. We conjecture that this approach could improve the convergence rate of the conformal prediction set. The optimal choice of conformity score and conformlisation remains an interesting question.

Theorem 4.2. *Given a consistent estimator $\hat{C}$ for the oracle prediction set, if $s^* = s\left(Y^L, Y^U, X; C^*\right)$ has positive density in a neighbourhood around 0, then the conformal predictioin set $\tilde{C}$ constructed as above, for an i.i.d. sample $\mathcal{D} = \left\{\left(X_i, Y_i^L, Y_i^U\right) : i \in \mathcal{I}\right\}$ which is split into training $\mathcal{I}_1$ and calibration sets $\mathcal{I}_2$, we have*

$$\sup_x \mu\left(\tilde{C}(x) \triangle C^*(x)\right) = o_p(1)$$

as $n_1, n_2 \longrightarrow \infty$, here $n_1 = |\mathcal{I}_1|$ and $n_2 = |\mathcal{I}_2|$.

It is tempting to look for a finite-sample conditionally valid prediction set $\tilde{C}(x)$ such that for all P,

$$\mathrm{P}\left(Y_{n+1} \in \tilde{C}(x) \mid X_{n+1} = x\right) \geq 1 - \alpha.$$

However, it is impossible to construct a prediction set that is finite-sample conditionally valid and consistent for the oracle prediction set (Lei and Wasserman 2014; Barber, Candes, et al. 2021; Gibbs, Cherian, and Candès 2024), and hence we consider the weaker notion of local validity. Let $\mathcal{A} = \left\{A_k : k \leq K\right\}$ be a partition of $\mathcal{X}$, a prediction interval $\tilde{C}(x)$ is *finite-sample locally valid* with respect to $\mathcal{A}$, if for all P,

$$\mathrm{P}\left(Y_{n+1} \in \tilde{C}_{\mathcal{A}}(X_{n+1}) \mid X_{n+1} \in \mathcal{A}_k\right) \geq 1 - \alpha.$$

Let $\mathcal{I}_{2,k} = \{i \in \mathcal{I}_2 : X_i \in A_k\}$, $n_{2,k} = \left|\mathcal{I}_{2,k}\right|$, and

$$\vartheta_{1-\alpha,k} = q\left(\left\{s_j : j \in \mathcal{I}_k^{\mathcal{A}}\right\}; (1-\alpha)\left(1 + \frac{1}{n_{2,k}}\right)\right),$$

then the local conformal prediction set is

$$\tilde{C}_{\mathcal{A}}(x) = \bigcup_{m=1}^{M}\left[\hat{\tau}_{0m}(x) - \sum_{k \leq K} \mathbb{I}\left\{x \in A_k\right\} \vartheta_{1-\alpha,k}, \hat{\tau}_{1m}(x) + \sum_{k \leq K} \mathbb{I}\left\{x \in A_k\right\} \vartheta_{1-\alpha,k}\right].$$

The finite-sample local validity is a result of exchangeability, similar to Lei and Wasserman (2014).

Theorem 4.3. *Suppose $\{(X_i, Y_i^L, Y_i^U) : i \in \mathcal{I}\}$ are i.i.d. Let $\mathcal{A}$ be a partition of $\mathcal{X}$. Then the local conformal prediction set $\tilde{C}_{\mathcal{A}}(x)$ is finite-sample locally valid with respect to $\mathcal{A}$.*

In the next section, we demonstrate the performance of the conformal and local conformal prediction sets in a set of numerical experiments.[5]

5 Numerical experiments

We report here the results of a series of numerical experiments to evaluate the performance of the conformal prediction sets. In all the experiments, we randomly draw a sample of predictors $\{X_i, i = 1, 2, \ldots, n\}$, with sample size $n = 2,500$ and each $X_i \sim \text{Unif}[-1.5, 1.5]$. The miscoverage level is fixed at $\alpha = 0.1$. The unobserved outcomes Y_i and the observed intervals (Y_i^L, Y_i^U) are then generated based on one of the models described below.

- *Model A.* The unobserved outcomes Y_i are generated according to a mixture model with heteroscedasticity, similar to the simulations in Lei and Wasserman (2014).

  $$(Y_i \mid X_i = x) \sim B_i N(f(x) + g(x), \sigma^2(x)) + (1 - B_i) N(f(x) - g(x), \sigma^2(x)) \qquad (12)$$

  Here, the auxiliary random variable B_i is independently drawn from a Bernoulli distribution with probability 0.5, and the following functions are used.

  $$f(x) = 2(x-1)^2(x+1); \; g(x) = 4\sqrt{(x+0.5)\mathbb{I}\{x \geq -0.5\}}; \; \sigma^2(x) = \frac{1}{4} + |x|,$$

  The observed intervals (Y_i^L, Y_i^U) are obtained by adding independent noise to the unobserved outcomes.

  $$Y_i^L = Y_i - e_{1,i}, \quad Y_i^U = Y_i + e_{2,i}, \quad e_{1,i}, e_{2,i} \sim_{\text{i.i.d.}} |N(0,1)|. \qquad (13)$$

  In this design, the conditional distribution $P_{Y|X}$ is unimodal for smaller values of X and bi-modal for larger values of X. Regression and quantile regression based prediction sets will have difficulty capturing the bifurcation shape of the oracle prediction set.

- *Model B.* The unobserved outcomes are generated in the same way as in Equation 12 above.

[5] The Python code for implementing the proposed conformal prediction procedure, as well as replicating the numerical experiments and empirical study, is available at https://github.com/lwg342/prediction_interval_outcome.

We randomly select 20% of the unobserved outcomes and they are interval censored and in this case and (Y_i^L, Y_i^U) have a discrete distribution , while the remaining 80% are observed without censoring. Let ι_i be a random draw from a Bernoulli distribution with probability 0.2,

$$\begin{cases} Y_i^L = Y_i^U = Y_i & \text{if} \quad \iota_i = 0; \\ Y_i^L = \lfloor Y_i \rfloor, Y_i^U = Y_i^L + 1 & \text{if} \quad \iota_i = 1. \end{cases} \tag{14}$$

This model is designed to evaluate the performance of the conformalized prediction set under fixed censoring. We choose to censor 20% of the outcomes, as approximately 20% of the income data is censored in the dataset used for our empirical studies.

- *Model C.* The unobserved outcomes are generated according to the following model,

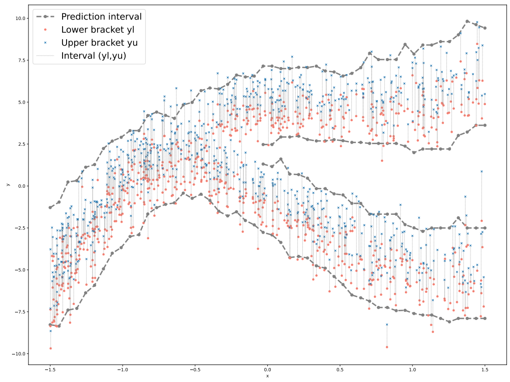
$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \sim_{\text{i.i.d.}} \chi^2_{1.5},$$

and (Y_i^L, Y_i^U) are generated accordng to Equation 14. In this case, the conditional distribution is unimodal and skewed.

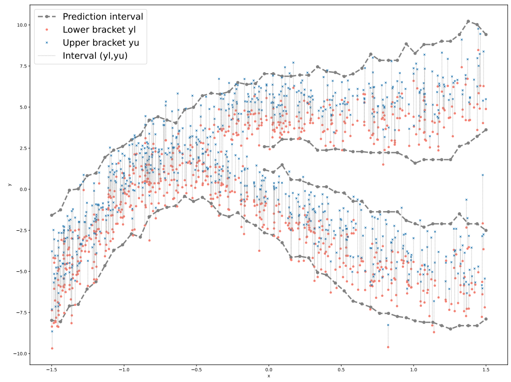
The observations $\{(X_i, Y_i^L, Y_i^U), i = 1, \ldots, n\}$, are randomly split into a training set $\mathcal{I}_1$ and a calibration set $\mathcal{I}_2$, with the training set containing 75% of the observations. We estimate the prediction set $\hat{C}_S(x)$ based on the kernel smoothing estimator $P_n(C; x)$ with the Epanechnikov kernel for the conditional distribution $P(C; x)$, as described in Equation (8), using the observations in the training set. The conformal prediction sets $\tilde{C}_S$ are then constructed according to Equation (11) using the calibration set. Local conformal prediction sets $\tilde{C}_{S,\text{loc}}$ are obtained using the local conformalization procedure, where the support $\mathcal{X}$ is divided into 5 equal-sized bins.

We then compare the average coverage and volume of the conformal prediction set $\tilde{C}$ and the local conformal prediction set $\tilde{C}_{\text{loc}}$ over 100 repetitions with two alternative conformal prediction intervals: $\tilde{C}_{q2}$, based on a quantile regression assuming a quadratic model, and $\tilde{C}_{q3}$, based on a quantile regression assuming a cubic model. See Romano, Patterson, and Candes (2019) and Appendix E for more details about construction of prediction interval based on quantile methods. The coverage is computed by randomly generating 5,000 samples of (Y^L, Y^U) and calculating the proportion of intervals covered by the prediction sets. The volume is calculated as the integrated Lebesgue measure of the prediction sets over the support $\mathcal{X}$.

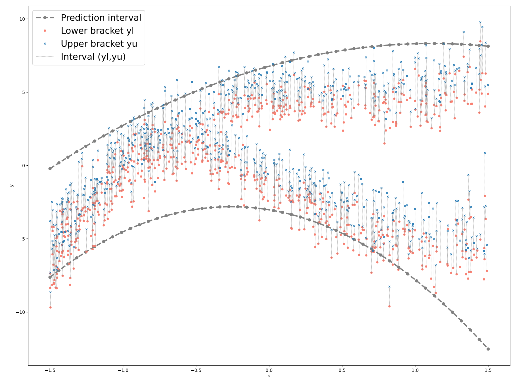
Figure 2 - 4 show the comparison of the prediction sets from the numerical experiments. The top two rows plot the observed sample (Y_i^L, Y_i^U) against X_i as well as the conformal prediction sets $C(X_i)$. The last row shows the integrated volume of the prediction sets and the coverage property. In Figure 2 and Figure 3, the prediction sets $\tilde{C}_S$ and $\tilde{C}_{S,\text{loc}}$ capture the bifurcation region of the conditional distribution and have smaller volume compared to the quantile regression-based prediction sets. Even when the conditional distribution is uni-modal, and the oracle prediction set is a single interval for each $x \in \mathcal{X}$, our proposed estimators still have smaller volume as demonstrated in Figure 4. In terms of coverage, $\tilde{C}_S, \tilde{C}_{S,\text{loc}}$ and $\tilde{C}_{q3}$ appear to have coverage close to the nominal leval $1 - \alpha = 0.9$. Notably, $\tilde{C}_{S,\text{loc}}$ appears to have better coverage than $\tilde{C}_S$ near the boundaries of the support $\mathcal{X}$.



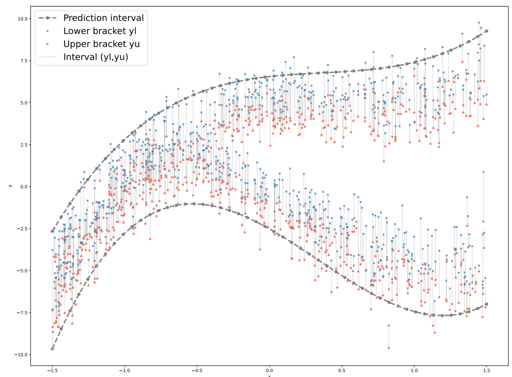
(a) Conformal prediction set $\tilde{C}_S$



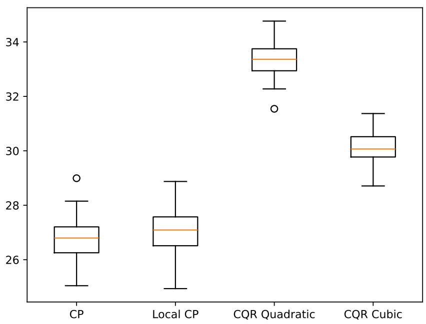
(b) Local conformal prediction set $\tilde{C}_{S,\mathrm{loc}}$



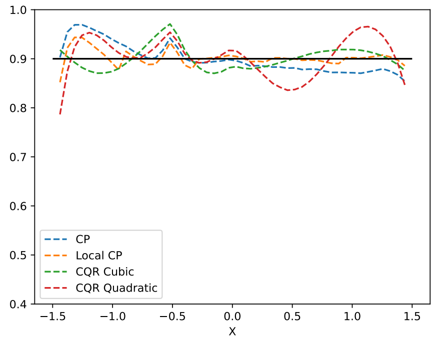
(c) Conf. pred. quadratic QR



(d) Conf. pred. cubic QR

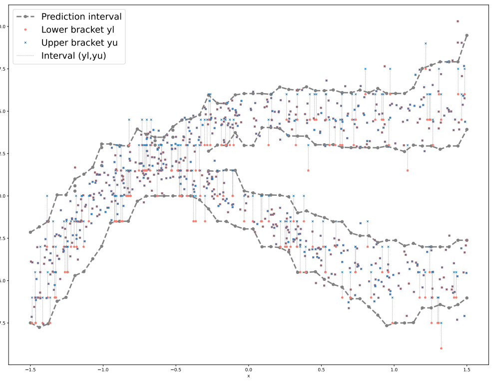


(e) Integrated volume

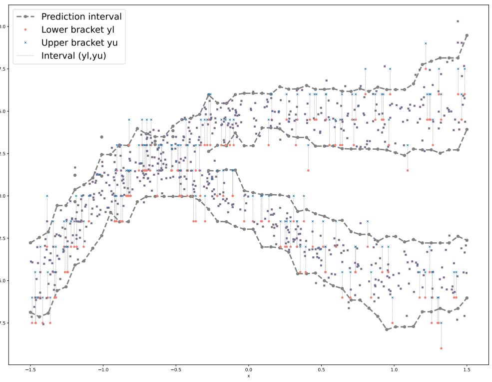


(f) Coverage

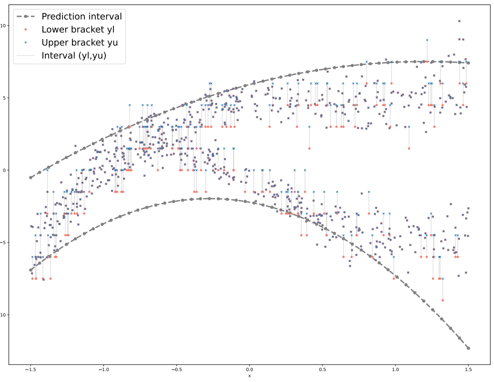
Figure 2: Comparison of prediction sets in simulation study. The first four sub-figures show the interval censored ourcomes and the conformal prediction sets. Subfigure (e) shows the integrated volume of the prediction sets, and subfigure (f) shows the coverage of the prediction sets.



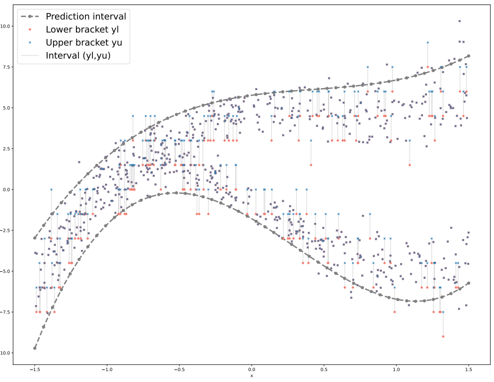
(a) Conformal prediction set



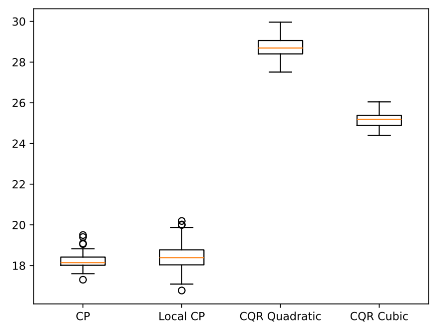
(b) Local conformal prediction set



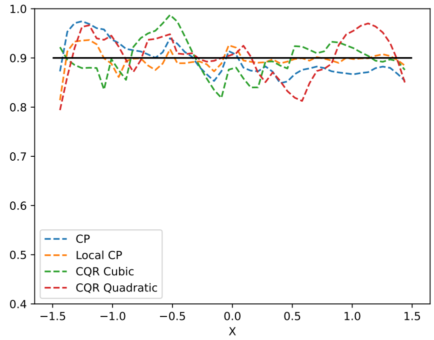
(c) Conf. pred. quadratic QR



(d) Conf. pred. cubic QR

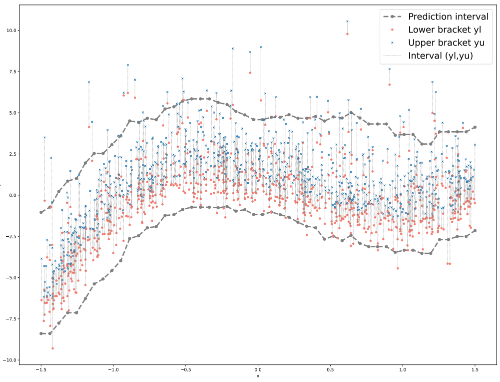


(e) Integrated volume

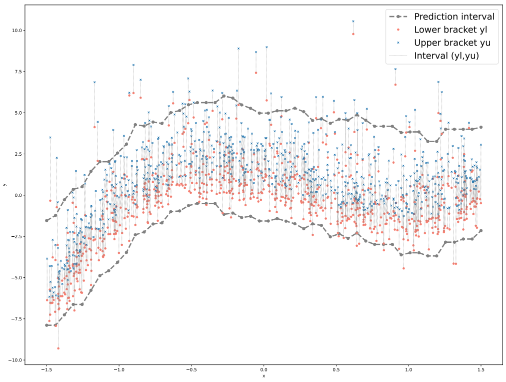


(f) Coverage

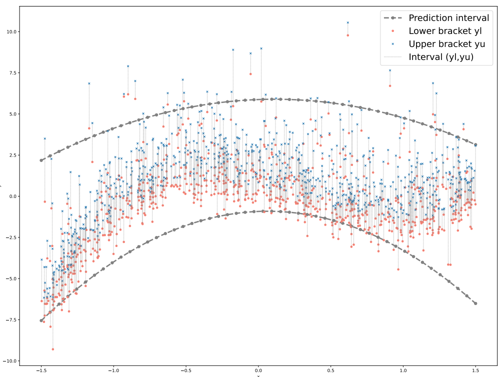
Figure 3: Comparison of prediction sets in simulation study. The first four sub-figures show the interval censored ourcomes and the conformal prediction sets. Subfigure (e) shows the integrated volume of the prediction sets, and subfigure (f) shows the coverage of the prediction sets.



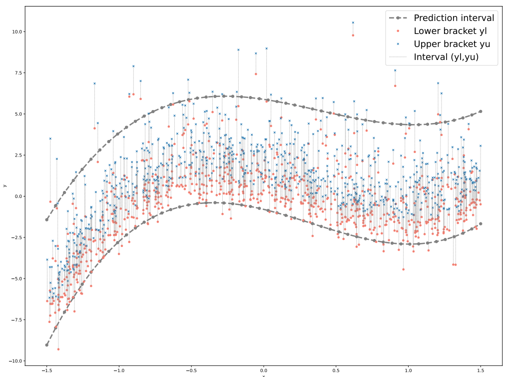
(a) Conformal prediction set



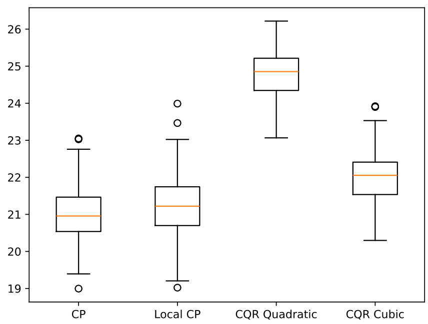
(b) Local conformal prediction set



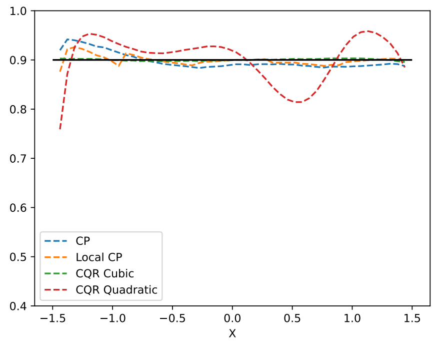
(c) Conf. pred. quadratic QR



(d) Conf. pred. cubic QR



(e) Integrated volume



(f) Coverage

Figure 4: Comparison of prediction sets in simulation study. The first four sub-figures show the interval censored ourcomes and the conformal prediction sets. Subfigure (e) shows the integrated volume of the prediction sets, and subfigure (f) shows the coverage of the prediction sets.

6 Empirical study

6.1 US CPS data

We applied our method to the 2023 US Current Population Survey (CPS) dataset and construct prediction interval for their income given years of education and experience. Following Heckman, Lochner, and Todd (2008), we construct the following variables from the dataset for individuals who report being American citizen and employed:

- *Income:* We include all individuals who reported their income or income range in the survey. Since 2014, when a person refuses to report their income, the interviewer will ask for a range, and the person's income will be imputed by matching with similar people that have income in the same range. In the 2019 ASEC Updates[6], it is stated that "if respondents did not give a value for income received from a given source, the interviewer now follows up with a question about income ranges. Respondents who did not give a specific value were asked if they received over $60,000, between $45,000 and $60,000, or less than $45,000 from this income source. Those who replied that they earned less than $45,000 from this source were further asked if they earned more than $30,000, between $15,000 and $30,000, or less than $15,000 from this income source." For people who reported their income as larger than $\$60,000$, we set the upper bound of the income as $\$400,000$, which is the topcode of the income variable in the dataset. The individuals who did not report their income nor their income range are excluded.
- *Age:* We include individuals who are between 18 and 65 years old.
- *Education:* Similar to Heckman, Lochner, and Todd (2008), we construct the years of education for each individual by converting the highest qualification they have obtained into years of education.
- *Experience:* We construct the years of experience for each individual as their age minus the years of education minus 6 (which is presumed to be the entry age into education).

The CPS dataset consists of 146,133 observations. There are 53,171 individuals who satisfy

[6] See https://cps.ipums.org/cps/asec_2019_changes.shtml

the demographic restrictions above. Among the individuals that satisfy the demographic criteria, 39,333 reported their income either exactly or as a band and the response rate is thus around 74%. The issue of nonresponse in surveys has been studied in the literature; see for example, Lillard, Smith, and Welch (1986) and Bollinger et al. (2019), where they document an increasing proportion of nonresponse of earnings in survey data and the potential bias it introduces. We will not address the nonresponse issue in this paper, but it is an important consideration in practice. Our proposed predictions could assist in the imputation of missing values in a robust manner; see Lei and Candès (2021). In fact, allowing respondents to report their income in bands could help mitigate the non-response problem. For the 39,333 individuals who reported their incomes and satisfy the demographic restrictions, 8,819 of them reported either their gross or net income as a band, which thus amounts to roughly 22.4%. of the respondents. Figure 5 shows the proportion of individuals who reported their income in a range by years of education.

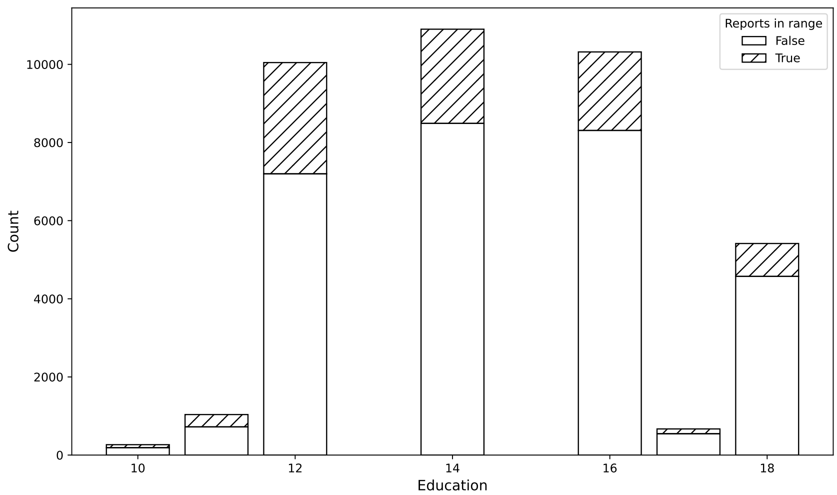


Figure 5: Proportion of individuals who reported income in a range by years of education.

In our empirical study, we first randomly reserve 20% of the data as a hold-out dataset. The remaining 80% is then further divided into training and calibration sets, with 75% of the data used for training to estimate the prediction interval $\hat{C}_I$ and 25% used for constructing the conformalized prediction interval $\tilde{C}_I$. Table 1 presents the average estimated prediction intervals across the 100 iterations. The method column indicates the way we estimate the prediction interval. "P" stands for prediction interval $\hat{C}_I$ in (7) and "CP" stands for conformal prediction interval $\tilde{C}_I$. If "M" is appended, it indicates the prediction intervals are estimated with imputed income for individuals who reported their income in an interval. It can be seen from the table that we would in general predict a higher income for individuals with more years of education.

The lower bounds of the prediction intervals increase with years of education, as well as the high-density region of the income distribution, which is shown in the columns with $\alpha = 0.9$ and a corresponding coverage of 10%. This is consistent with what has been documented in the literature. Having more years of experience tends to increase the predicted income as well. By varying the choice of α, we can gain useful insights into the predicted income distribution.

The impact of including interval-censored income data on the prediction interval is more pronounced at smaller α values, as evidenced by the results for $\alpha = 0.1$ in the table. This is because, with a 90% coverage target, the tails of the income distribution become more critical. Given that 20% of the observations are reported in ranges and many fall into the highest category, incorporating interval-censored data shifts the right tail of the income distribution. Consequently, the prediction intervals constructed with interval-censored data are significantly different from the ones constructed with imputed income data.

We also compare the coverage of the income intervals in the hold-out calibration set by prediction intervals constructed with interval-censored data and with imputed data. In Figure 6, we show the comparison of the coverages. The conformal prediction interval constructed with interval-censored data has a higher coverage at the chosen level of $1 - \alpha = 0.9$, while the prediction interval constructed with imputed data has a lower coverage. Table 2 below shows the coverage of the interval censored incomes in the hold-out dataset for the conformal prediction interval $\tilde{C}$ that are constructed using both interval-censored and exactly reported income data (shown in the "Censored" columns), as well as for the conformal prediction intervals constructed when the interval-censored incomes are imputed (shown in the "Imputed" columns). The significant difference in prediction intervals constructed with interval-censored data and with imputed data may be due to the fact that there are only five categories for individuals who report their income in a range. This seems somewhat coarse in terms of informing about the true income distribution. Notice that the conformal coverage using the censored model has coverage close to the nominal rate while the Imputed Coverage is almost always lower. Note that in Tables 1 an 2 when we display coverages of .5 and .1, this provides an empirical estimate of the 50% and

10% prediction sets, i.e., sets where respectively 50% and 90% of where the data are and should be informative about the shape of the data distribution. Notice that standard reporting of confidence intervals on parameters is meant to summarize sampling uncertainty, while here not only do these conformalized sets take into account sampling uncertainty, but also these are estimates of prediction sets of the underlying distributions of interest.

Table 1: Prediction intervals for annual income

α	Method	Edu. / Exp.	12	14	16	18
0.1	CP	10	(3,750 - 148,077)	(6,183 - 146,329)	(24,068 - 427,369)	(30,365 - 433,655)
		20	(11,321 - 457,690)	(13,725 - 426,357)	(18,722 - 425,517)	(38,276 - 440,264)
		30	(12,862 - 389,206)	(13,743 - 389,322)	(22,539 - 437,454)	(29,392 - 449,567)
		40	(12,174 - 402,666)	(11,374 - 421,875)	(13,292 - 444,424)	(11,678 - 493,910)
	CPM	10	(8,867 - 99,274)	(11,341 - 128,093)	(17,879 - 194,759)	(25,095 - 243,709)
		20	(10,518 - 145,653)	(12,617 - 159,334)	(18,680 - 270,824)	(35,708 - 301,124)
		30	(13,134 - 127,668)	(13,748 - 183,252)	(19,765 - 305,760)	(26,054 - 341,740)
		40	(10,675 - 138,292)	(9,893 - 169,031)	(11,674 - 304,466)	(11,033 - 411,907)
	P	10	(4,501 - 112,987)	(6,545 - 136,283)	(24,828 - 412,162)	(31,079 - 422,347)
		20	(12,437 - 403,886)	(14,077 - 411,421)	(19,247 - 411,791)	(38,814 - 432,214)
		30	(13,310 - 372,180)	(13,577 - 392,820)	(23,110 - 424,860)	(29,853 - 441,393)
		40	(12,495 - 389,214)	(11,455 - 417,218)	(13,512 - 435,372)	(12,725 - 446,886)
	PM	10	(8,978 - 97,593)	(11,606 - 124,455)	(18,444 - 188,033)	(26,000 - 234,411)
		20	(10,774 - 141,664)	(12,895 - 155,148)	(18,928 - 266,643)	(37,019 - 288,533)
		30	(13,284 - 125,232)	(14,218 - 176,678)	(20,387 - 294,974)	(26,773 - 329,103)
		40	(10,945 - 134,348)	(10,076 - 165,370)	(12,014 - 292,832)	(12,259 - 359,563)
0.5	CP	10	(28,403 - 69,159)	(28,451 - 61,707)	(40,785 - 93,725)	(49,968 - 124,635)
		20	(29,112 - 74,398)	(34,397 - 90,045)	(44,607 - 139,467)	(60,475 - 149,195)
		30	(29,516 - 74,826)	(37,350 - 99,430)	(44,516 - 134,520)	(48,127 - 152,613)
		40	(27,987 - 69,835)	(34,751 - 101,913)	(37,510 - 128,448)	(38,655 - 173,552)
	CPM	10	(24,957 - 55,295)	(30,970 - 64,072)	(41,826 - 90,070)	(51,771 - 114,129)
		20	(30,710 - 72,934)	(37,596 - 86,891)	(48,515 - 128,404)	(54,816 - 114,490)
		30	(32,981 - 77,018)	(37,521 - 92,111)	(55,278 - 134,171)	(61,075 - 136,151)
		40	(31,443 - 74,478)	(37,368 - 96,865)	(43,132 - 124,874)	(53,980 - 161,441)
	P	10	(27,882 - 70,433)	(27,842 - 63,044)	(40,128 - 95,239)	(49,363 - 126,132)
		20	(28,809 - 75,132)	(34,096 - 90,783)	(44,878 - 138,496)	(59,597 - 151,292)
		30	(29,089 - 75,848)	(37,067 - 100,125)	(43,679 - 136,951)	(48,118 - 152,532)

Continued on next page

Table 1: Prediction intervals for annual income

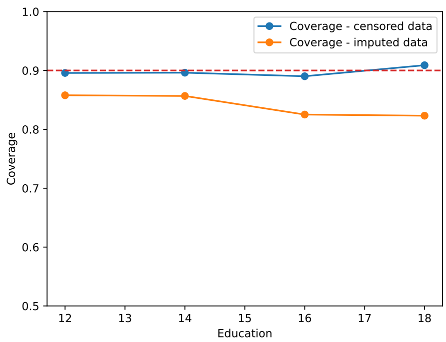
		Edu.	12	14	16	18
α	Method	Exp.				
		40	(27,731 - 70,418)	(34,913 - 101,294)	(37,209 - 129,111)	(39,201 - 170,455)
	PM	10	(24,742 - 55,759)	(30,715 - 64,589)	(41,568 - 90,614)	(51,953 - 113,663)
		20	(30,516 - 73,359)	(37,337 - 87,461)	(48,290 - 128,907)	(55,045 - 113,954)
		30	(32,793 - 77,435)	(37,620 - 91,858)	(54,999 - 134,791)	(61,070 - 136,038)
		40	(31,534 - 74,220)	(37,993 - 95,199)	(43,191 - 124,565)	(55,297 - 157,002)
0.9	CP	10	(37,280 - 47,224)	(51,494 - 64,217)	(50,508 - 59,915)	(64,290 - 77,790)
		20	(46,705 - 58,681)	(47,282 - 57,681)	(47,429 - 59,726)	(90,065 - 104,636)
		30	(44,508 - 56,170)	(48,707 - 58,304)	(60,125 - 73,829)	(86,622 - 104,886)
		40	(54,771 - 68,861)	(51,089 - 63,476)	(64,694 - 84,384)	(110,919 - 146,037)
	CPM	10	(32,945 - 39,246)	(43,036 - 49,255)	(48,741 - 55,971)	(84,551 - 99,614)
		20	(48,124 - 55,754)	(55,927 - 66,073)	(72,426 - 87,084)	(75,760 - 86,633)
		30	(43,362 - 51,148)	(50,080 - 57,792)	(74,437 - 87,542)	(87,295 - 99,780)
		40	(47,259 - 56,903)	(54,230 - 65,552)	(62,610 - 77,602)	(92,021 - 112,387)
	P	10	(36,525 - 48,183)	(50,920 - 64,919)	(49,748 - 60,842)	(64,738 - 77,259)
		20	(45,691 - 59,964)	(46,176 - 59,062)	(46,590 - 60,791)	(90,123 - 104,542)
		30	(43,507 - 57,444)	(47,394 - 59,939)	(58,687 - 75,646)	(86,500 - 105,008)
		40	(53,896 - 69,941)	(49,935 - 64,921)	(64,204 - 84,959)	(111,821 - 144,773)
	PM	10	(32,919 - 39,274)	(42,881 - 49,430)	(47,997 - 56,834)	(85,117 - 98,935)
		20	(48,083 - 55,796)	(55,770 - 66,242)	(72,782 - 86,662)	(76,003 - 86,341)
		30	(43,297 - 51,219)	(50,045 - 57,832)	(74,150 - 87,866)	(87,420 - 99,605)
		40	(47,276 - 56,869)	(54,423 - 65,315)	(62,899 - 77,205)	(93,783 - 110,213)

Note. The annual incomes are measured in US dollars. In the Method column, P stands for prediction intervals $\hat{C}_I$ at a given education level and years of experience. CP stands for the conformal prediction interval $\tilde{C}_I$. PM and CPM are prediction intervals and conformal prediction intervals, respectively, constructed with imputed income for individuals who report their income in a range.

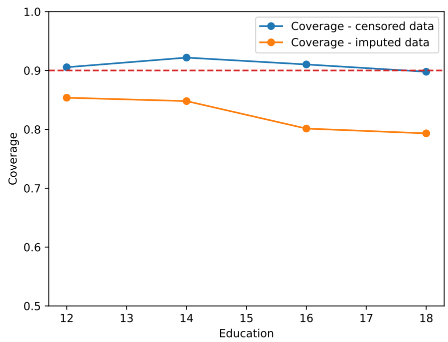
Table 2: Coverage of conformal prediction intervals

	α	0.10		0.25		0.50		0.90	
	Data Type	Censored	Imputed	Censored	Imputed	Censored	Imputed	Censored	Imputed
Education	Experience								
12	10	**0.896**	0.858	**0.742**	0.668	**0.504**	0.433	**0.104**	0.085
	20	**0.906**	0.854	**0.780**	0.693	**0.522**	0.467	**0.111**	0.074
	30	**0.911**	0.828	**0.744**	0.663	**0.509**	0.445	**0.115**	0.080
	40	**0.894**	0.838	**0.761**	0.667	**0.489**	0.458	**0.098**	0.072
14	10	**0.896**	0.857	**0.736**	0.667	**0.484**	0.444	**0.095**	0.078
	20	**0.922**	0.848	**0.759**	0.679	**0.502**	0.446	**0.107**	0.079
	30	**0.919**	0.844	**0.754**	0.698	**0.511**	0.458	**0.097**	0.080
	40	**0.910**	0.838	**0.749**	0.683	**0.505**	0.461	**0.096**	0.094
16	10	**0.890**	0.825	**0.726**	0.663	**0.488**	0.451	**0.097**	0.087
	20	**0.910**	0.801	**0.756**	0.674	**0.524**	0.451	**0.099**	0.085
	30	**0.895**	0.768	**0.765**	0.627	**0.478**	0.382	**0.105**	0.076
	40	**0.914**	0.787	**0.723**	0.632	**0.488**	0.435	0.109	**0.091**
18	10	**0.909**	0.823	**0.749**	0.681	**0.517**	0.464	0.108	**0.104**
	20	**0.898**	0.793	**0.747**	0.646	**0.496**	0.419	**0.099**	0.077
	30	**0.911**	0.800	**0.768**	0.650	**0.550**	0.421	**0.114**	0.081
	40	**0.923**	0.825	**0.822**	0.646	**0.552**	0.430	**0.092**	0.081

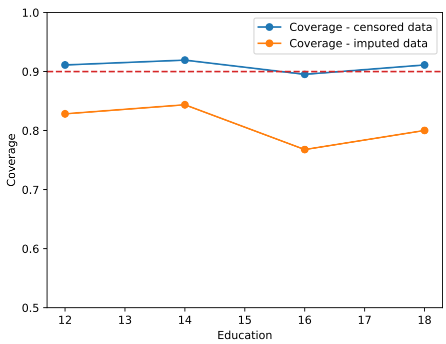
Note. This table shows the coverage of the interval censored incomes in the hold-out dataset for the conformal prediction interval $\tilde{C}$ that are constructed using both interval-censored and exactly reported income data (shown in the "Censored" columns), as well as for the conformal prediction intervals constructed when the interval-censored incomes are imputed (shown in the "Imputed" columns). The bolded values are the coverage rates that are closest to the nominal level $1-\alpha$.



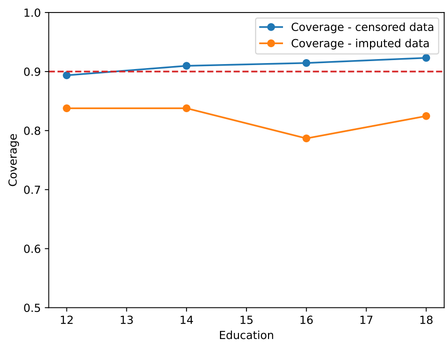
(a) Years of experience = 10



(b) Years of experience = 20



(c) Years of experience = 30



(d) Years of experience = 40

Figure 6: Comparison of coverage of prediction intervals constructed with interval-censored data and with imputed data for $\alpha = 0.1$.

7 Conclusion

The object of interest in this paper is a prediction set for an outcome given a set of covariates when this outcome is censored. Censoring is a common issue in economic and other data. We first characterise the oracle prediction set which is the smallest set that maintains a given miscoverage level under censoring. The characterisation leads to a feasible estimation strategy based on the observed sample. We then provide consistent estimators for this oracle prediction set based nonparametric estimation of the conditional distribution of the observed lower and upper brackets given the conditioning variables. We allow for the prediction set to consist of multiple intervals and consider both random and fixed censoring. Furthermore, we use recent results from conformal inference to obtain a conformal prediction set that maintains a finite sample miscoverage property using a set based score function. We find that our procedures perform well with simulated data. We also apply our procedures to data from the US Census. There is increasing interest in prediction set estimation, and several intriguing research directions are related to this paper. One promising avenue is to consider prediction sets under partial identification beyond interval censoring. While conformal inference offers finite-sample guarantees on coverage, it remains a research challenge to explore the more traditional statistical inference on the estimated prediction set. Additionally, the design of conformity scores in the context of partial identification presents an interesting and important topic for further investigation.

References

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized Random Forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/18-AOS1709.

Barber, Rina Foygel, Emmanuel J. Candes, et al. (2021). "The Limits of Distribution-Free Conditional Predictive Inference". In: *Information and Inference: A Journal of the IMA* 10.2, pp. 455–482.

Barber, Rina Foygel, Emmanuel J. Candès, et al. (2023). "Conformal Prediction beyond Exchangeability". In: *Annals of Statistics* 51.2, pp. 816–845. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/23-AOS2276.

Bollinger, Christopher R. et al. (2019). "Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch". In: *Journal of Political Economy* 127.5, pp. 2143–2185. ISSN: 0022-3808. DOI: 10.1086/701807.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly (2013). "Inference on Counterfactual Distributions". In: *Econometrica* 81.6, pp. 2205–2268. ISSN: 1468-0262. DOI: 10.3982/ECTA10582.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu (2021). "Distributional Conformal Prediction". In: *Proceedings of the National Academy of Sciences* 118.48, e2107794118. DOI: 10.1073/pnas.2107794118.

Diebold, Francis X. (2015). *Forecasting in Economics, Business, Finance and Beyond.* [iBooker itebooks].

Einmahl, Uwe and David M. Mason (2000). "An Empirical Process Approach to the Uniform Consistency of Kernel-Type Function Estimators". In: *Journal of Theoretical Probability* 13.1, pp. 1–37. ISSN: 1572-9230. DOI: 10.1023/A:1007769924157.

Folland, Gerald B. (1999). *Real Analysis: Modern Techniques and Their Applications.* Vol. 40. John Wiley & Sons.

Foresi, Silverio and Franco Peracchi (1995). "The Conditional Distribution of Excess Returns: An Empirical Analysis". In: *Journal of the American Statistical Association* 90.430, pp. 451–466. ISSN: 0162-1459. DOI: 10.2307/2291056.

Gibbs, Isaac, John J. Cherian, and Emmanuel J. Candès (2024). *Conformal Prediction With Conditional Guarantees.* DOI: 10.48550/arXiv.2305.12616. arXiv: 2305.12616 [stat]. Pre-published.

Gine, E (2002). "Rates of Strong Uniform Consistency for Multivariate Kernel Density Estimators". In: *Annales de l'Institut Henri Poincare (B) Probability and Statistics* 38.6, pp. 907–921. ISSN: 02460203. DOI: 10.1016/S0246-0203(02)01128-7.

Hansen, Bruce E. (2008). "Uniform Convergence Rates for Kernel Estimation with Dependent Data". In: *Econometric Theory* 24.3, pp. 726–748. ISSN: 0266-4666, 1469-4360. DOI: 10.1017/S0266466608080304.

Heckman, James J., Lance J. Lochner, and Petra E. Todd (2008). "Earnings Functions and Rates of Return". In: *Journal of Human Capital* 2.1, pp. 1–31. ISSN: 1932-8575. DOI: 10.1086/587037.

Heeringa, Steven G and Richard Suzman (1995). *Unfolding brackets for reducing item nonresponse in economic surveys.* National Institute on Aging, National Institutes of Health.

Kline, Brendan and Elie Tamer (2023). "Recent Developments in Partial Identification". In: *Annual Review of Economics* 15 (Volume 15, 2023), pp. 125–150. ISSN: 1941-1383, 1941-1391. DOI: 10.1146/annurev-economics-051520-021124.

Koenker, Roger (2017). "Quantile Regression: 40 Years On". In: *Annual Review of Economics* 9.1, pp. 155–176. DOI: 10.1146/annurev-economics-063016-103651.

Lei, Jing, Max G'Sell, et al. (2018). "Distribution-Free Predictive Inference for Regression". In: *Journal of the American Statistical Association* 113.523, pp. 1094–1111. ISSN: 0162-1459. DOI: 10.1080/01621459.2017.1307116.

Lei, Jing, James Robins, and Larry Wasserman (2013). "Distribution-Free Prediction Sets". In: *Journal of the American Statistical Association* 108.501, pp. 278–287. ISSN: 0162-1459. DOI: 10.1080/01621459.2012.751873. pmid: 25237208.

Lei, Jing and Larry Wasserman (2014). "Distribution-Free Prediction Bands for Non-Parametric Regression". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76.1, pp. 71–96. ISSN: 1369-7412.

Lei, Lihua and Emmanuel J. Candès (2021). "Conformal Inference of Counterfactuals and Individual Treatment Effects". In: *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 83.5, pp. 911–938. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/rssb.12445.

Li, Degui, Qi Li, and Zheng Li (2021). "Nonparametric Quantile Regression Estimation With Mixed Discrete and Continuous Data". In: *Journal of Business & Economic Statistics* 39.3, pp. 741–756. ISSN: 0735-0015. DOI: 10.1080/07350015.2020.1730856.

Lillard, Lee, James P. Smith, and Finis Welch (1986). "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation". In: *Journal of Political Economy* 94 (3, Part 1), pp. 489–506. ISSN: 0022-3808, 1537-534X. DOI: 10.1086/261386.

Manski, Charles F. (2003). *Partial Identification of Probability Distributions.* Springer Science & Business Media.

Manski, Charles F. and Elie Tamer (2002). "Inference on Regressions with Interval Data on a Regressor or Outcome". In: *Econometrica* 70.2, pp. 519–546. ISSN: 1468-0262. DOI: 10.1111/1468-0262.00294.

Molinari, Francesca (2020). "Chapter 5 - Microeconometrics with Partial Identification". In: *Handbook of Econometrics*. Ed. by Steven N. Durlauf et al. Vol. 7. Handbook of Econometrics, Volume 7A. Elsevier, pp. 355–486. DOI: 10.1016/bs.hoe.2020.05.002.

Moore, Jeffrey C and Laura S Loomis (2001). "Using alternative question strategies to reduce income nonresponse". In: *Survey Methodology* 3.

Polonik, Wolfgang (1995). "Measuring Mass Concentrations and Estimating Density Contour Clusters-an Excess Mass Approach". In: *Annals of Statistics* 23.3, pp. 855–881. ISSN: 0090-5364.

Romano, Yaniv, Evan Patterson, and Emmanuel Candes (2019). "Conformalized Quantile Regression". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Sadinle, Mauricio, Jing Lei, and Larry Wasserman (2019). "Least Ambiguous Set-Valued Classifiers with Bounded Error Levels". In: *Journal of the American Statistical Association* 114.525, pp. 223–234. ISSN: 0162-1459. DOI: 10.1080/01621459.2017.1395341.

Samworth, R. J. and M. P. Wand (2010). "Asymptotics and Optimal Bandwidth Selection for Highest Density Region Estimation". In: *Annals of Statistics* 38.3, pp. 1767–1792. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/09-AOS766.

Tamer, Elie (2010). "Partial Identification in Econometrics". In: *Annual Review of Economics* 2 (Volume 2, 2010), pp. 167–195. ISSN: 1941-1383, 1941-1391. DOI: 10.1146/annurev.economics.050708.143401.

Tsybakov, Alexandre B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. New York, NY: Springer. DOI: 10.1007/b13794.

Vaart, Aad van der and Jon Wellner (2013). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media. 523 pp. ISBN: 978-1-4757-2545-2. Google Books: zdDkBwAAQBAJ.

Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer (2005). *Algorithmic Learning in a Random World*. Cham: Springer International Publishing. DOI: 10.1007/978-3-031-06649-8.

Vovk, Vladimir, Ilia Nouretdinov, and Alex Gammerman (2009). "On-Line Predictive Linear Regression". In: *Annals of Statistics* 37.3, pp. 1566–1590. ISSN: 0090-5364.

Wilks, S. S. (1941). "Determination of Sample Sizes for Setting Tolerance Limits". In: *Annals of Mathematical Statistics* 12.1, pp. 91–96. ISSN: 0003-4851.

A Proofs

A.1 Proofs for Section 2

In this subsection, we prove the results in Section 2 and clarify the difference between our approach and the classical level set estimation.

First we state a simple and useful lemma that allows us to construct a density function that is in the identified set under interval censoring given the joint density of (Y^L, Y^U).

Lemma A.1. *Let f_{Y^L,Y^U} be the joint density of (Y^L, Y^U), for any interval $A = [a_0, a_1] \subset \mathbb{R}$, and any $\theta \in [\frac{\pi}{2}, \pi]$, define the following*

$$A_\theta = \left\{(x, y) \in \mathbb{R}^2 : x \le y, \frac{y-a}{x-a} = \tan\theta, a \in A\right\},$$

where it follows that $A_{\frac{\pi}{2}} = \left\{(x, y) : x \le y, x \in A\right\}$. Then $P_\theta(A) = P_{Y^L,Y^U}(A_\theta)$ defines a probability measure P_θ on $\mathbb{R}$, and P_θ is in the identified set of the distribution of the latent outcome Y.

With Lemma A.1 at hand, we can now state the following proposition that shows the level set of Y is not in general identified given the joint density of (Y^L, Y^U). Given the density f_Y of a random variable Y, the (upper) level set of Y at level λ is defined as $L(Y;\lambda) = \left\{y : f_Y(y) \ge \lambda\right\}$. If f_Y is identified, then the optimal $1-\alpha$ prediction set for Y can be written as $C(1-\alpha) = L(Y;\lambda_{1-\alpha})$ where $\lambda_{1-\alpha} = \sup\left\{\lambda : P_Y(Y \in L(Y;\lambda)) \ge 1-\alpha\right\}$. When the random variable Y is interval censored, we instead observe the random vector (Y^L, Y^U) and we could define the level set of (Y^L, Y^U) as,

$$L((Y^L, Y^U);\lambda) = \left\{(y^L, y^U) : f_{Y^L,Y^U}(y^L, y^U) \ge \lambda\right\}. \tag{15}$$

We note here two points. First, the level set of Y is in general not identified given the joint density of (Y^L, Y^U) and second, in terms of prediction for Y, our proposed prediction set is more efficient and has clearer interpretation than a level set of (Y^L, Y^U).

The following example shows that the level set of Y is not in general identified without additional assumptions on the mechanism of censoring.

Example A.1. Suppose for some $a_0 < b_0 \leq a_1 < b_1 \in \mathbb{R}$, and and $\lambda > 0$, the joint density of (Y^L, Y^U) satisfies the restriction that for all $(a, b) \in \mathbb{R}^2$ such that either $a_0 \leq a \leq a_1$ or $b_0 \leq b \leq b_1$,

$$f_{Y^L,Y^U}(a, b) = \lambda \mathbb{I}\left\{a \in [a_0, a_1], b \in [b_0, b_1]\right\},$$

and $f_{Y^L,Y^U}(a, b) = 0$ for all $a < b$ such that $b \in [a_0, b_0]$ and $a \in [a_1, b_1]$.

Consider the following two random variables Y_1, Y_2 that are constructed in the following ways for $y \in \mathcal{Y}$,

$$f_{Y_1}(y) = \int f_{Y^L,Y^U}(y, b)\,\mathrm{d}b; \quad f_{Y_2}(y) = \int f_{Y^L,Y^U}(a, y)\,\mathrm{d}a.$$

Then $[a_0, b_0] \subset L(Y_1, \lambda) \cup L(Y_2, \lambda)^c$, and $[a_1, b_1] \subset L(Y_2, \lambda) \cup L(Y_1, \lambda)^c$. And both Y_1, Y_2 are observationally equivalent under interval censoring, corresponding to the selection mechanism that $Y_1 = Y^L$ and $Y^2 = Y^U$.

We now apply Lemma A.1 to prove Proposition 2.1 which generalises the idea in the previous example and shows the failure of identification of the level set of Y.

Proof of Proposition 2.1. Let $(a, b) \in \mathbb{R}^2$ such that $f_{Y^L,Y^U}(a, b) > 0$ and consider sets

$$A_k = \left\{(y_l, y_u) : a - 1/k \leq y_l \leq a, b \leq y_u \leq b + 1/k\right\},$$

which shrinks nicely to (a, b) (Folland 1999), and as a reult for some k^*, $\delta = \int_{A_{k^*}} f_{Y^L,Y^U} > 0$. Let $P_{\pi/2}$ be the one defined in Lemma A.1 and $P'_{\pi/2}$ be the positive measure constructed with the same method for the density function $f_{Y^L,Y^U}\mathbb{I}\left\{A^c_{k^*}\right\}$. For any $y \in [a, b]$, and any $\lambda > 0$, let $0 < \varepsilon < \delta/\lambda$, and define the following probability distribution, for any interval $B \subset \mathbb{R}$,

$$P_{y,\varepsilon}(B) = P'_{\pi/2}(B) + \mu(B \cap B_{\frac{\varepsilon}{2}}(y)\frac{\delta}{\varepsilon}, \tag{16}$$

where $B_{\frac{\varepsilon}{2}}(y) = \left\{y' : \left|y' - y\right| \leq \frac{\varepsilon}{2}\right\}$. Then $P_{y,\varepsilon}$ is in the identified set and has density larger than λ at y. ■

Since $P(Y \in A) \geq P([Y^L, Y^U] \subset A)$, it is clear that the outer region of the level set with $1-\alpha$ covergage of (Y^L, Y^U) also has $1 - \alpha$ coverage for Y. But the outer region is in general not the smallest valid prediction set.

The following lemma shows that even though our approach is conservative in terms of coverage, its conservativeness is only due to partial identification, which is necessary without assumeing the mechanism of censoring.

Lemma A.2. *For almost surely $x \in \mathcal{X}$, and any $C = \sqcup_{m=1}^{M} [l_m, u_m] \in \mathcal{C}$, we have*

$$\inf_{P \in \mathcal{P}} P(Y \in C \mid X = x) = P\left([Y^L, Y^U] \subset C \mid X = x\right).$$

Proof. It is clear that $P(Y \in C \mid X = x) \geq P\left([Y^L, Y^U] \subset C \mid X = x\right)$ for all $P \in \mathcal{P}$, and we only need to show the reverse inequality also holds. Let $g_x = \mathrm{d}P_{Y^L,Y^U|X=x} / \mathrm{d}\mu$ be the Radon-Nikodym derivative of $P_{Y^L,Y^U|X=x}$ with respect to a dominating measure μ on $\mathbb{R}^2$. Assume without loss of generality, $-\infty = u_0 < l_1 < u_1 < \cdots < l_M < u_M < \infty = l_{M+1}$, the half space $\mathbb{R}^{2,+} = \left\{(y_l, y_u) : y_u \geq y_l\right\}$ is partitioned into the following regions,

$$\begin{aligned}
&\mathbb{R}^{2,+} = (\cup_{m=1}^{M} (A_m \cup A'_m \cup A''_m)) \cup (\cup_{m=0}^{M} B_m), \quad \text{where} \\
&A_m = \left\{(y_l, y_u) : l_m \leq y_l \leq y_u \leq u_m\right\}, \\
&A'_m = \left\{(y_l, y_u) : l_m \leq y_l \leq u_m < y_u \leq l_m + 1\right\}, \\
&A''_m = \left\{(y_l, y_u) : l_m \leq y_l \leq u_m, y > l_{m+1}\right\}, \\
&B_m = \left\{(y_l, y_u) : u_m \leq y_l \leq l_{m+1}, y_l \leq y_u\right\}.
\end{aligned}$$

Recall the definition of A_θ in Lemma A.1, we define the following positive measures on the real line, for any intervals $A \subset \mathbb{R}$,

$$\begin{aligned}
P_1(A) &= \int_{A_{\frac{\pi}{2}}} g_x \mathbb{I}\left\{\cup_m (A_m \cup B_m)\right\}, \\
P_2(A) &= \int_{A_\pi} g_x \mathbb{I}\left\{A'_m\right\}, \\
P_{3,m} &= P_{Y^L,Y^U|X=x}(A''_m) \cdot U(u_m, l_{m+1}),
\end{aligned}$$

where $U(a, b)$ is the uniform distribution over (a, b).

Then $P'(A) = P_1(A) + P_2(A) + \sum_m P_{3,m}(A)$ defines a probability measure P' such that $P'(C) =$

$P([Y^L, Y^U] \subset C)$ and for any interval A, $P'(Y \in A) \geq P([Y^L, Y^U] \subset A)$, and hence $P' \in \mathcal{P}$. ■

A.2 Proofs for Section 3

In this subsection, we first prove the uniform consistency of the proposed prediction set when the oracle prediction set is an interval under random censoring. Then the result is extended to the case allowing for fixed censoring and when oracle prediciton set is a union of multiple intervals.

When the oracle prediction set is a single interval for each $x \in \mathcal{X}$, let $C_I^*(x) = [\tau_0(x), \tau_1(x)]$.

Proof of Theorem 3.1. For any $\varepsilon > 0$, let δ be as defined in Assumption 3.2.

Let ν be chosen such that $0 < \nu < \min\left(\frac{\varepsilon}{4}, \frac{1}{2}\left(\frac{\delta}{2c_2}\right)^{\frac{1}{\gamma_2}}\right)$, and fix $0 < \kappa < \min\left(\frac{\delta}{2}, c_1\nu^{\gamma_1}\right)$. Let E denote the event $\sup_{a,b,x}\left|P_n(a, b; x) - P(a, b; x)\right| < \kappa$, under Assumption 3.1, $P(E^c) \to 0$ as $n \to \infty$.

On the event E, for any $x \in \mathcal{X}$, we want to show that $\mu\left([\hat{\tau}_0(x), \hat{\tau}_1(x)] \triangle [\tau_0(x), \tau_1(x)]\right) < \varepsilon$. Let $E_{0,x} = \left\{\omega : \mu\left([\hat{\tau}_0(x), \hat{\tau}_1(x)] \triangle [\tau_0(x), \tau_1(x)]\right) > \varepsilon\right\}$. Consider the following three cases,

1. $E_{1,x} = \left\{\omega : \hat{\tau}_1(x) - \hat{\tau}_0(x) \leq \tau_1(x) - \tau_0(x)\right\}$
2. $E_{2,x} = \left\{\omega : \hat{\tau}_1(x) - \hat{\tau}_0(x) > \tau_1(x) - \tau_0(x) + 2\nu\right\}$
3. $E_{3,x} = \left\{\omega : \tau_1(x) - \tau_0(x) < \hat{\tau}_1(x) - \hat{\tau}_0(x) \leq \tau_1(x) - \tau_0(x) + 2\nu\right\}$.

then we will show $E_{0,x} \subset E^c$ by showing that $(E_{1,x} \cup E_{2,x} \cup E_{3,x}) \cap E_{0,x} \subset E^c$ for all $x \in \mathcal{X}$.

1. In the first case, we have $\hat{\tau}_1 - \hat{\tau}_0 \leq \tau_1(x) - \tau_0(x)$, if $E_{0,x}$ also holds, then $P(\hat{\tau}_0, \hat{\tau}_1; x) \leq 1 - \alpha - \delta$ by Assumption 3.2, while $P_n(\hat{\tau}_0, \hat{\tau}_1; x) \geq 1 - \alpha$, which implies that $\left|P_n(\hat{\tau}_0, \hat{\tau}_1; x) - P\left(\hat{\tau}_0, \hat{\tau}_1; x\right)\right| > \delta > \kappa$. Hence $E_{1,x}E_{0,x} \subset E^c$.
2. In the second case, $\hat{\tau}_1 - \hat{\tau}_0 > \tau_1(x) - \tau_0(x) + 2\nu$. Let $(\bar{\tau}_0, \bar{\tau}_1) = (\tau_0 - \nu, \tau_1 + \nu)$, then by Assumption 3.3, $P(\bar{\tau}_0, \bar{\tau}_1; x) \geq 1 - \alpha + c_1\nu^{\gamma}$. Since $\bar{\tau}_1 - \bar{\tau}_0 < \hat{\tau}_1 - \hat{\tau}_0$, by Assumption 3.1, $P_n\left(\bar{\tau}_0, \bar{\tau}_1; x\right) < 1 - \alpha$. Again, this would imply that $\left|P_n(\bar{\tau}_0(x), \bar{\tau}_1(x); x) - P(\bar{\tau}_0(x), \bar{\tau}_1(x); x)\right| > c_1\nu^{\gamma_1} > \kappa$, hence $E_{2,x} \subset E^c$.
3. Using the fact that $\mu([a_0, a_1] \triangle [b_0, b_1]) \leq \left|b_0 - a_0\right| + \left|b_1 - a_1\right|$, we have either $\left|\hat{\tau}_0(x) - \tau_0(x)\right| > \frac{\varepsilon}{2}$ or $\left|\hat{\tau}_1(x) - \tau_1(x)\right| > \frac{\varepsilon}{2}$. Since $\hat{\tau}_1(x) - \hat{\tau}_0(x) - (\tau_1(x) - \tau_0(x)) \leq 2\nu < \frac{\varepsilon}{2}$, $\hat{\tau}_0(x)$ and

$\hat{\tau}_1(x)$ are on the same side of $\tau_0(x), \tau_1(x)$. As a result, either $[a_0, a_1] = [\hat{\tau}_0, \hat{\tau}_1 - 2\nu]$ or $[\hat{\tau}_0 + 2\nu, \hat{\tau}_1]$ satisfies $\mu\left([a_0, a_1] \triangle [\tau_0(x), \tau_1(x)]\right) > \varepsilon$ and $\left|[a_0, a_1]\right| \leq \tau_1(x) - \tau_0(x)$, then we have $P([a_0, a_1]; x) < 1 - \alpha - \delta$. It follows that $P(\hat{\tau}_0, \hat{\tau}_1; x) < 1 - \alpha - \delta + c_2(2\nu)^{\gamma_2}$ So $\left|P_n(\hat{\tau}_0(x), \hat{\tau}_1(x); x) - P(\hat{\tau}_0(x), \hat{\tau}_1(x); x)\right| > \frac{\delta}{2} > \kappa$. Hence $E_{3,x}E_{0,x} \subset E^c$.

As a result, on the event E, we have $\sup_x \mu([\hat{\tau}_0(x), \hat{\tau}_1(x)] \triangle [\tau_0(x), \tau_1(x)]) \leq \varepsilon$, and since $P(E^c) \to 0$ as $n \to \infty$, we have that $\sup_x \mu([\hat{\tau}_0(x), \hat{\tau}_1(x)] \triangle [\tau_0(x), \tau_1(x)]) = o_p(1)$. ∎

Proof of Theorem 3.3. For any $\varepsilon > 0$, let δ be chosen as in Assumption 3.5. Let E denote the event when $\sup_{C,x} \left|P_n(C; x) - P(C; x)\right| \leq \psi_n$. For sufficiently large n, $\psi_n \leq \frac{\delta}{2}$. For any $x \in \mathcal{X}$, if $\mu(\hat{C}(x) \triangle C^*(x)) > \varepsilon$ and $\mu(\hat{C}(x)) \leq \mu(C^*(x))$, then $P(\hat{C}(x); x) \leq 1 - \alpha - \delta$, and $P_n(\hat{C}(x); x) \geq 1 - \alpha - \psi_n$. Hence $\left|P_n(\hat{C}; x) - P(\hat{C}; x)\right| > \frac{\delta}{2} \geq \psi_n$. On the other hand, if $\mu(\hat{C}(x) \triangle C^*(x)) > \varepsilon$ and $\mu(\hat{C}(x)) > \mu(C^*(x))$, then $P_n(C(x); x) < 1 - \alpha - \psi_n$ while $P(C(x); x) \geq 1 - \alpha$ and $\left|P_n(C(x); x) - P(C(x); x)\right| > \psi_n$. That is $P(\sup_x \mu(\hat{C}(x) \triangle C^*(x)) > \varepsilon) \leq P(E^c) \to 0$. ∎

A.3 Proofs for results in Section 4

Proof of Theorem 4.1. We have, by interval censoring,

$$\mathrm{P}(Y_{n+1} \in \tilde{C}(X_{n+1})) \geq \mathrm{P}([Y^L_{n+1}, Y^U_{n+1}] \subset \tilde{C}(X_{n+1})).$$

Let $s_i = s(Y^L_i, Y^U_i, X_i; \hat{C})$, then $[Y^L_{n+1}, Y^U_{n+1}] \subset \tilde{C}(X_{n+1})$ if and only if $s_{n+1} \leq \vartheta_{1-\alpha}$. Since s_i's are exchangeable for $i \in \mathcal{I}_2$, $P\left(s_{n+1} \leq \vartheta_{1-\alpha}\right) \geq 1 - \alpha$ by Lemma 2 in Romano, Patterson, and Candes (2019). ∎

Proof of Theorem 4.2. Fix an $\varepsilon > 0$, on the event $\sup_x \mu\left(\hat{C}_I(x) \triangle C_I(x)\right) < \frac{\varepsilon}{2}$, we know that

$$\left|s(y^l, y^u, x; \hat{\tau}_0, \hat{\tau}_1) - s(y^l, y^u, x; \tau_0, \tau_1)\right| < \frac{\varepsilon}{2},$$

for all y^l, y^u and x. Let $s_{[j]}, s^*_{[j]}$ be the corresponding order statistics for s_j and s^*_j, $j \in \mathcal{I}_2$. Let $s_{[l]} = s_{[l]}$ when l is an integer and $s_{[l]} = s_{[\lfloor l \rfloor + 1]}$ when l is not an integer. We have $s_{[(n_2+1)(1-\alpha)]} > s^*_{[(n_2+1)(1-\alpha)]} - \varepsilon$ provided that $s^*_{[(n_2+1)(1-\alpha)]} - s^*_{[(n_2+1)(1-\alpha)-1]} < \frac{\varepsilon}{2}$ which happens with high probability for sufficiently large n_2. Similarly we have $s_{[(n_2+1)(1-\alpha)]} < s^*_{[(n_2+1)(1-\alpha)]} + \varepsilon$. Thanks to the consistency of sample quantile, we have that $s^*_{[(n_2+1)(1-\alpha)]} \to_p Q_{1-\alpha}(s^*)$ as $n_2 \to \infty$. By definition, $[y^L, y^U] \subset [t_0(x), t_1(x)]$ if and only if $s(y^L, y^U, x; t_0, t_1) \le 0$. Since $Q_{1-\alpha}(s^*) = \inf\left\{z : P(s^* \le z) \ge 1-\alpha\right\}$, combined with

$$P(s^* \le 0) = P\left(\tau_0 \le Y^L \le Y^U \le \tau_1\right) \ge 1-\alpha,$$

and for any $z < 0$, $P(s^* \le z) < 1-\alpha$ by Assumption 3.2, we have $Q_{1-\alpha}(s^*) = 0$. Hence $\vartheta_{1-\alpha} = o_p(1)$. ■

B Uniform consistency of the kernel smoothing estimator

In this section, we show the following kernel smoothing estimator $P_n(C;x)$ is uniformly consistent for the conditional probability $P(C;x) = P([Y^L, Y^U] \subset C \mid X = x)$ under some primitive conditions. For all $C = \bigcup_{m=1}^M [a_m, b_m]$ where $a_m \le b_m < a_{m+1}$,

$$P_n(C;x) = \frac{\sum_{i=1}^n \sum_{m=1}^M \mathbb{I}\left\{a_m \le Y_i^L \le Y_i^U \le b_m\right\} K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}.$$

Let $f(y^l, y^u; (a_m, b_m)_{m=1}^M) = \sum_{m=1}^M \mathbb{I}\left\{a_m \le Y_i^L \le Y_i^U \le b_m\right\}$ which belongs to the following class of functions indexed by the endpoints $(a_m, b_m)_{m=1}^M$,

$$\mathcal{F}_M = \left\{f(\cdot,\cdot;(a_m, b_m)_{m=1}^{M'}) : a_m \le b_m < a_{m+1}, M' \le M\right\},$$

then for $C = \bigcup[a_m, b_m]$,

$$P_n(C;x) = \frac{\sum_i f(Y_i^L, Y_i^U; (a_m, b_m)_{m=1}^M) K\left(\frac{x-X_i}{h}\right)}{\sum_i K\left(\frac{x-X_i}{h}\right)} =: P_{n,f}(x).$$

Similarly define the population version $P_f(x) = \mathrm{E}\left[f(Y^L, Y^U; (a_m, b_m)_{m=1}^M)\right]$. It is straightforward to show that $\mathcal{F}_M$ is a measurable VC class of functions, with a constant enevelop function $F = 1$

and hence for any probability measure Q over the support of (Y^L, Y^U),

$$N\Big(\varepsilon, \mathcal{F}, L_r(Q)\Big) \leq C(1/\varepsilon)^v,$$

where C, v depends only on the VC characterisitcs of $\mathcal{F}_M$ and $r \geq 1$ (Vaart and Wellner 2013). Then the next lemma follows from Corollary 1 in Einmahl and Mason (2000) with only minor modifications under some conditions on the smoothing kernel and bandwidth sequence.

Assumption B.1. *Suppose that* $K(\frac{x-X}{h}) = \prod_{j=1}^d K_0\left(\frac{x^j - X^j}{h}\right)$ *where* x^j *denote the* j*-th component of* x*, and* K_0 *is conitnuous over a compact support,* $\int K(s)\,\mathrm{d}s = 1$.

Assumption B.2. *The bandwidth sequence* $h = h_n$ *satisfies the growth conditions that , as* $n \longrightarrow \infty$,

$$h \longrightarrow 0, \quad nh^d \longrightarrow \infty, \quad \frac{|\log h|}{\log\log n} \longrightarrow 0, \quad \frac{nh^d}{\log n} \longrightarrow \infty.$$

Lemma B.1. *Let* $\mathcal{X}'$ *be a compact subset of* $\mathcal{X}$ *where* p_X *is continuous and bounded away from* 0 *on* $\mathcal{X}'$*, and suppose* $p_{(X,Y^L,Y^U)}$ *is continuous on* $\mathcal{X}' \times \mathbb{R}^2$*. Then under Assumptions B.1 and B.2, with probability* 1,

$$\sup_{x\in\mathcal{X}'} \sup_{f\in\mathcal{F}_M} \left|P_{n,f}(x) - \bar{P}_{n,f}(x)\right| = O\left(\sqrt{\frac{\log|h|}{nh^d}}\right),$$

where

$$\bar{P}_{n,f}(x) = \frac{h^{-d}\,\mathrm{E}\left[f(Y^L, Y^U; (a_m, b_m)_{m=1}^M) K\left(\frac{x-X}{h}\right)\right]}{\mathrm{E}\left[(nh^d)^{-1}\sum_i K\left(\frac{x-X}{h}\right)\right]}.$$

It remains to bound the "bias" terms. It is standard to show that, if $p_X(x)$ and $P(C;x)p_X(x)$ have uniformly continuous second derivatives with respect to x, then

$$\sup_{x\in\mathcal{X}'} \left| \mathrm{E}\left[\frac{1}{nh^d}\sum_i K\left(\frac{x - X_i}{h}\right)\right] - p_X(x)\right| = O(h^2),$$

as well as

$$\sup_{x\in\mathcal{X}'} \sup_{f\in\mathcal{F}_M} \left| h^{-d}\,\mathrm{E}\left[f(Y^L, Y^U; (a_m, b_m)_{m=1}^M) K\left(\frac{x - X}{h}\right)\right] - P_f(x)\right| = O(h^2).$$

Combining these results, we have that, with probability 1,

$$\sup_{x\in\mathcal{X}'}\sup_{f\in\mathcal{F}_M}\left|P_{n,f}(x)-P_f(x)\right| = O\left(\sqrt{\frac{\log|h|}{nh^d}}+h^2\right). \tag{17}$$

As shown in Hansen (2008) Theorem 8, we can let $\mathcal{X}' = \mathcal{X}_n$ be an expanding subsets at a suitable rate, and the uniform convergence rate needs to be multiplied by a factor of δ_n^{-1} where $\delta_n = \inf_{x\in\mathcal{X}_n} p_X(x)$.

Here we have dealt with the case when the density function p_{X,Y^L,Y^U} is smooth. For fixed censoring, suppose the discrete part is supported on a known finite set $\left\{(L_i,U_i) : i = 1,\dots,K\right\}$, as in the empirical studies where the brackets are designed by the survey, then the discrete part can be estimated at parametric rate, which is negligible compared to the nonparametric rate above.

C Differential conformal procedure

Suppose we have obtained a prediction interval $\hat{C}_I = [\hat{\tau}_0(x),\hat{\tau}_1(x)]$ with the training set $\mathcal{I}_1$. In the main text, the conformal prediction set is constructed in two steps.

1. Find the conformity score $s_j = s_j(\hat{C}_I) = \max\left(\hat{\tau}_0(X_j)-Y_j^L, Y_j^U-\hat{\tau}_1(X_j)\right)$ for $j\in\mathcal{I}_2$. The important thing is $s_j\le 0 \iff [Y_j^L,Y_j^U]\subset\hat{C}_I(X_j)$.

2. Compute the relevant $(1-\alpha)(1+\frac{1}{|\mathcal{I}_2|})$ quantile ϑ of the conformity scores $\left\{s_j : j\in\mathcal{I}_2\right\}$, and the conformal prediction set is $\tilde{C}_I$ given by

$$\tilde{C}_I(x) = \left[\hat{\tau}_0(x)-\vartheta, \hat{\tau}_1(x)+\vartheta\right], \quad x\in\mathcal{X}. \tag{18}$$

It can be seen that in the second step, the endpoints $\hat{\tau}_0(x)$ and $\hat{\tau}_1(x)$ are treated symmetrically. We could instead consider a differentiated approach on the two endpoints of $\hat{C}_I$. Define the following contour set $\mathcal{W} = \left\{(w_0,w_1) : P_{\mathcal{I}_2}(s_j(\hat{\tau}_0-w_0,\hat{\tau}_1+w_1)) \gtrsim 1-\alpha\right\}$. And we can pick $(w_0^*,w_1^*) = \min_{\mathcal{W}}\left\|(w_0^*,w_1^*)\right\|$.

Given $\hat{C}_S$ estimated with the training set $\mathcal{I}_1$, we can construct the conformity score $s_j =$

$s(Y_j^L, Y_j^U; \hat{C}_S)$, where for $C \in \mathcal{C}_M$,

$$s(y^L, y^U; C) = \min_{m \leq M} \max \left\{ \hat{\tau}_{0m} - y^L, y^U - \hat{\tau}_{1m} \right\}.$$

And we can solve the following for $w = (w_{01}, w_{11}, w_{02}, w_{12}, \ldots, w_{0M}, w_{1M})$, $\tilde{C}_M = \cup_m [\hat{\tau}_{0m} - w_{0m}, \hat{\tau}_{1m} + w_{1m}]$, by

$$\min \|w\| \quad \text{s.t.} \quad P_{\mathcal{I}_2}\left(s(Y_j^L, Y_j^U; \tilde{C}_M) \leq 0 \right) \gtrsim 1 - \alpha. \tag{19}$$

We conjecture that this method may improve the convergence rate of the conformal prediction set estimator. However, as the literature on conformal inference focuses mainly on finite-sample properties, we leave this direction for future investigation.

D Metrics for sets

Two common choices of measurement for the difference between two intervals are the Hausdorff distance and the volume of symmetric difference. The *Hausdorff distance* between two sets $A, B \subset \mathbb{R}$ is defined as

$$d_H(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} \left| a - b \right|, \sup_{b \in B} \inf_{a \in A} \left| a - b \right| \right\}. \tag{20}$$

The *symmetric difference* $\triangle$ between two sets A, B is defined as $A \triangle B = (A \setminus B) \cup (B \setminus A)$, and the volume of the symmetric difference $\mu(A \triangle B)$ has also been studied in the literature.

In general, these two metrics are not equivalent, but they are equivalent for two non-disjoint intervals on the real line. In particular, for a sequence of intervals $A_n = [a_{n0}, a_{n1}]$ and $B = [b_0, b_1]$, $b_0 < b_1$, it can be shown that $d_H(A_n, B) \to 0$ if and only if $\mu(A_n \triangle B) \to 0$ as $n \to \infty$.

Therefore, in terms of consistency for prediction intervals, we can use either metric and we will proceed with the volume of symmetric difference and show that $\hat{C}_I(x)$ is consistent in the sense that $\sup_{x \in \mathcal{X}} \mu(\hat{C}_I(x) \triangle C_I(x)) = o_p(1)$.

However, when the prediction set is a union of intervals, these two metrics are not equivalent. In fact, consider $A \subset \mathbb{R}$ and $B = A \cup \left\{ b \right\}$, for some $b \in A^c$ such that $d(b, A) > 0$, then the Hausdorff distance between A and B is $d_H(A, B) = d(b, A) > 0$, while the volume of the symmetric difference is 0. Thus, the volume of symmetric difference has the advantage of being less sensitive to a difference of negligible sets and seems more approporiate for practical prediction sets estimations.

E Estimation and conformal inference based on quantile regression

As we have suggested, there is a simpler way to construct valid prediction sets with quantile regression, although the resulting prediction set is not generally optimal in terms of length.

By Bonferroni's inequality, for two events A, B, $P(AB) \geq \max\left(0, P(A) + P(B) - 1\right)$, and given $P(Y^L \geq q^L_{\frac{\alpha}{2}}(x) \mid x) \geq 1 - \alpha/2$ and $P(Y^U \leq q^U_{1-\frac{\alpha}{2}}(x) \mid X = x) \geq 1 - \alpha/2$ by the definition of quantiles, we have

$$P\left(Y^L \geq q^L_{\frac{\alpha}{2}}(x), Y^U \leq q^U_{1-\frac{\alpha}{2}}(x) \mid X = x\right) \geq 1 - \alpha \tag{21}$$

and $\hat{C}_q(x) = [\hat{q}^L_{\frac{\alpha}{2}}(x), \hat{q}^U_{1-\frac{\alpha}{2}}(x)]$ as an estimator for $C_q(x) = [q^L_{\frac{\alpha}{2}(x)}, q^U_{1-\frac{\alpha}{2}}(x)]$ is an asymptotically valid, albeit not necessarily optimal, prediction set. Here $\hat{q}^L_{\frac{\alpha}{2}}(x), \hat{q}^U_{1-\frac{\alpha}{2}}(x)$ are the consistent quantile regression estimators. There is a vast body of research on quantile regression, see Koenker (2017), Athey, Tibshirani, and Wager (2019) for random forest based quantile regression and Li, Li, and Li (2021) for nonparametric quantile regression with mixed continuous and discrete data.

Once we have $\hat{C}_q(x)$, it is straightforward to use the conformalisation procedure to obtain finite-sample valid prediction sets. Let $s^q_j = s(Y^L_j, Y^U_j, X_j; \hat{q}^L_{\frac{\alpha}{2}}, \hat{q}^U_{\frac{\alpha}{2}})$ and $\vartheta^q_{1-\alpha} = q(\{s^q_j : j \in \mathcal{I}_2\}; (1 - \alpha)(1 + 1/|\mathcal{I}_2|))$. The conformalised quantile prediction interval is

$$\tilde{C}_q(x) = \left[\hat{q}^L_{\frac{\alpha}{2}}(x) - \vartheta^q_{1-\alpha}, \hat{q}^U_{1-\frac{\alpha}{2}}(x) + \vartheta^q_{1-\alpha}\right]. \tag{22}$$