Distributional conformal prediction

Victor Chernozhukov[a,b,1], Kaspar Wüthrich[c,d,e,1], and Yinchu Zhu[f,g,1,2]

[a]Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02142; [b]Center for Statistics and Data Science, Massachusetts Institute of Technology, Cambridge, MA 02142; [c]Department of Economics, University of California San Diego, La Jolla, CA 92093; [d]CESifo, 81679 Munich, Germany; [e]ifo Center for Public Finance and Political Economy, ifo Institute, 81679 Munich, Germany; [f]Department of Economics, Brandeis University, Waltham, MA 02453; and [g]International Business School, Brandeis University, Waltham, MA 02453

Edited by Emmanuel J. Candès, Stanford University, Stanford, CA, and approved October 5, 2021 (received for review April 24, 2021)

We propose a robust method for constructing conditionally valid prediction intervals based on models for conditional distributions such as quantile and distribution regression. Our approach can be applied to important prediction problems, including cross-sectional prediction, *k*–step-ahead forecasts, synthetic controls and counterfactual prediction, and individual treatment effects prediction. Our method exploits the probability integral transform and relies on permuting estimated ranks. Unlike regression residuals, ranks are independent of the predictors, allowing us to construct conditionally valid prediction intervals under heteroskedasticity. We establish approximate conditional validity under consistent estimation and provide approximate unconditional validity under model misspecification, under overfitting, and with time series data. We also propose a simple "shape" adjustment of our baseline method that yields optimal prediction intervals.

prediction intervals | conditional validity | model-free validity | quantile regression | distribution regression

We develop a robust approach for constructing prediction intervals based on models for conditional distributions. The proposed method is generic and can be implemented using a great variety of flexible and powerful methods, including conventional quantile regression (QR) (1) and distribution regression (DR) (e.g., refs. 2 and 3), as well as nonparametric and high-dimensional machine learning methods such as quantile neural networks (e.g., ref. 4) and quantile trees and random forests (e.g., refs. 5 and 6).

We observe data $\{(Y_t, X_t)\}_{t=1}^T$, where Y_t is a continuous outcome of interest and X_t is a $p \times 1$ vector of predictors. Our task is to predict Y_{T+1} given knowledge of X_{T+1}. This setting encompasses many classical cross-sectional and time series prediction problems. Moreover, our approach can be applied to synthetic control settings where the goal is to predict counterfactuals in the absence of a policy intervention (e.g., refs. 7 and 8) and to the problem of predicting individual treatment effects (e.g., refs. 9 and 10).

With independent and identically distributed (iid) (or exchangeable data), standard conformal prediction methods, which are based on modeling the conditional mean, yield prediction intervals $\widehat{\mathcal{C}}_{(1-\alpha)}$ that satisfy

$$P\left(Y_{T+1} \in \widehat{\mathcal{C}}_{(1-\alpha)}\left(X_{T+1}\right)\right) \geq 1-\alpha \qquad [1]$$

for a given miscoverage level $\alpha \in (0,1)$. A prediction interval satisfying this property is said to be unconditionally valid. Unconditionally valid prediction intervals guarantee accurate coverage on average, treating (Y_{T+1}, X_{T+1}) and $\{(Y_t, X_t)\}_{t=1}^T$ as random.

However, in many applications, unconditional validity may be unsatisfactory. Let us consider three examples; refs. 11 and 12 have further examples and discussions. First, from a fairness perspective, data-driven recommendation systems should guarantee equalized coverage across protected groups, in which case the goal is to construct prediction intervals that are valid conditional on a protected attribute such as race or gender (11). Second, as in *Predicting Stock Market Returns*, consider the problem of predicting stock returns given the realized volatility. Since the distribution of returns is more dispersed when the variance is higher, a natural prediction algorithm should yield wider prediction intervals for higher values of volatility. That is, the prediction interval should be valid conditional on the known value of realized volatility rather than on average. Third, as in *Predicting Wages Using CPS Data*, suppose our goal is to predict wages based on an individual's education and experience. An unconditionally valid prediction interval exhibits coverage 90% on average across all individuals but may contain the true wage of high school dropouts with no work experience with probability zero. A more useful prediction interval should exhibit correct coverage conditional on an individual's observed education and experience and contain the true wage with 90% probability for every single individual.

Motivated by this discussion, we develop a distributional conformal prediction (DCP) method for constructing prediction intervals that are approximately valid conditional on the full vector of predictors X_{T+1} while treating Y_{T+1} and $\{(Y_t, X_t)\}_{t=1}^T$ as random:

$$P\left(Y_{T+1} \in \widehat{\mathcal{C}}_{(1-\alpha)}\left(X_{T+1}\right) \mid X_{T+1}\right) \geq 1-\alpha+o_P(1). \qquad [2]$$

A prediction interval satisfying property Eq. **2** as $T \to \infty$ is said to be approximately conditionally valid.*

While the requirement in Eq. **2** is natural in many applications, there are also other notions of conditional validity. Instead of conditioning on X_{T+1} (object conditional), one can also study

Significance

Prediction problems are important in many contexts. Examples include cross-sectional prediction, time series forecasting, counterfactual prediction and synthetic controls, and individual treatment effect prediction. We develop a prediction method that works in conjunction with many powerful classical methods (e.g., conventional quantile regression) as well as modern high-dimensional methods for estimating conditional distributions (e.g., quantile neural networks). Unlike many existing prediction approaches, our method is valid conditional on the observed predictors and efficient under some conditions. Importantly, our method is also robust; it exhibits unconditional coverage guarantees under model misspecification, under overfitting, and with time series data.

Author contributions: V.C., K.W., and Y.Z. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]V.C., K.W., and Y.Z. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: yinchuzhu@brandeis.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2107794118/-/DCSupplemental.

Published November 24, 2021.

*See, for example, refs. 12–14 for a further discussion of the difference between conditional and unconditional validity.

the conditional coverage probability given the training sample $\{(Y_t, X_t)\}_{t=1}^{T}$ (training conditional) or given Y_{T+1} (label conditional) or combinations of them; ref. 15 has a detailed discussion. By proposition 2 of ref. 15, inductive conformal predictions (also known as split-sample conformal predictions) automatically achieve training conditional validity as long as the training sample is large enough. In classification problems (the support of Y_{T+1} is a finite set), label conditional validity is often of great interest as it is important to know the error rates for different categories and provides useful information on false-positive and false-negative rates (15). In ref. 15, label conditional validity is achieved by forming the conformity score within each category. Both training and label conditional validity can be achieved in a distribution-free way (i.e., for a given procedure, the conditional validity holds for any distribution of the data).

However, object conditional validity in the sense of Eq. **2** cannot be achieved in a distribution-free way for nontrivial predictions. By refs. 12, 13, and 15, any prediction set satisfying Eq. **2** for every probability distribution of (X_t, Y_t) has infinite Lebesgue measure with nontrivial probability. Therefore, we only aim to achieve Eq. **2** for a limited class of probability distributions. The construction of the proposed prediction set $\widehat{C}_{(1-\alpha)}$ relies on learning the conditional distribution $Y_t \mid X_t$, and we only hope for conditional validity in Eq. **2** in the class of distributions that can be learned well. In particular, this class of distributions is those satisfying our regularity conditions.

Our empirical results demonstrate the importance of using DCP instead of standard conformal prediction methods based on modeling the conditional mean. When predicting daily stock returns in *Predicting Stock Market Returns*, the coverage probability of the 90% mean-based conformal prediction interval can drop to around 50% when the realized volatility is high. By contrast, DCP provides a coverage probability close to 90% for all values of realized volatility. This finding is important since volatility tends to be high during periods of crisis when accurate risk assessments are most needed. When predicting wages in *Predicting Wages Using CPS Data*, we find that the DCP prediction intervals contain the true wage with probability close to 90% for most individuals, whereas standard mean-based conformal prediction intervals either substantially under- or overcover.

To motivate DCP, note that a conditionally valid prediction interval is given by

$$\left[Q\left(\frac{\alpha}{2}, x\right), Q\left(1-\frac{\alpha}{2}, x\right)\right], \qquad \textbf{[3]}$$

where $Q(\tau, x)$ is the τ quantile of Y_t given $X_t = x$. To implement the prediction interval Eq. **3**, a plug-in approach would replace Q with a consistent estimator $\hat{Q}$:

$$\left[\hat{Q}\left(\frac{\alpha}{2}, x\right), \hat{Q}\left(1-\frac{\alpha}{2}, x\right)\right]. \qquad \textbf{[4]}$$

This approach exhibits two well-known drawbacks. First, it will often exhibit undercoverage in finite samples (e.g., ref. 16). Second, it is neither conditionally nor unconditionally valid under misspecification.

We build upon conformal prediction (17, 18) and use the conditional ranking as a conformity score. This choice is particularly useful when working with regression models for conditional distributions such as QR and DR.[†] Our method is conditionally valid under correct specification, while the construction of the procedure as a conformal prediction method guarantees the unconditional validity under misspecification. Let $F(y, x) = P(Y_t \leq y \mid X_t = x)$ denote the conditional cumulative distribution function (CDF) of Y_t given $X_t = x$. Throughout the paper, we assume that $F(\cdot, X_t)$ is a continuous function almost surely. Our method is based on the probability integral transform, which states that the conditional rank, $U_t := F(Y_t, X_t)$, has the uniform distribution on $(0, 1)$ and is independent of X_t.

To construct the prediction interval, we test the plausibility of each $y \in \mathbb{R}$. By the probability integral transform, conditional on X_{T+1}, $F(Y_{T+1}, X_{T+1})$ belongs to $[\alpha/2, 1-\alpha/2]$ with probability $1-\alpha$. Thus, collecting all values $y \in \mathbb{R}$ satisfying $F(y, X_{T+1}) \in [\alpha/2, 1-\alpha/2]$ yields a conditionally valid prediction interval in the sense of Eq. **2**. We operationalize this idea by proposing a conformal prediction procedure based on the estimated ranks, $\hat{U}_t^{(y)} := \hat{F}^{(y)}(Y_t, X_t)$. For each $y \in \mathbb{R}$, $\hat{F}^{(y)}$ is an estimator of F obtained based on the augmented data, $\{(Y_t, X_t)\}_{t=1}^{T+1}$, where $Y_{T+1} = y$. Data augmentation is a key feature of conformal prediction. It implies the model-free unconditional exact finite-sample validity with iid (or exchangeable) data and thus, guards against model misspecification and overfitting. Without data augmentation, the resulting prediction intervals are not exactly valid, not even with correct specification and iid data.

Our baseline method asymptotically coincides with the oracle interval in Eq. **3**. This oracle interval may not be the shortest possible prediction interval in general. Therefore, we also develop a simple and easy to implement adjustment of our baseline method for improving efficiency, which we refer to as optimal DCP. In *Predicting Wages Using CPS Data*, we show empirically that optimal DCP yields shorter prediction intervals than baseline DCP when the conditional distribution is skewed.

We establish the following theoretical performance guarantees for the baseline and optimal DCP.

1) Asymptotic conditional validity under consistent estimation of the conditional CDF
2) Unconditional validity under model misspecification:
   - Finite-sample validity with iid (or exchangeable) data
   - Asymptotic validity with time series data
3) For optimal DCP:
   - Under weak conditions: asymptotic conditional validity and optimality (shortest length)
   - Under strong conditions: asymptotic convergence to the optimal prediction interval

Motivating Example

We illustrate the advantages of DCP relative to mean-based conformal prediction (e.g., ref. 20) based on the following simple analytical example:

$$Y_t = X_t + X_t\varepsilon_t, \quad X_t \overset{iid}{\sim} \text{Uniform}(0,1), \quad \varepsilon_t \overset{iid}{\sim} N(0,1). \qquad \textbf{[5]}$$

Our motivating example draws on refs. 16 and 20–22 that illustrate the importance of accounting for heteroscedasticity. We focus on the population conformal prediction (or oracle) problem under correct specification and abstract from finite-sample issues.

Mean-based conformal prediction is based on the residuals $R_t = Y_t - E(Y_t \mid X_t) = Y_t - X_t = X_t\varepsilon_t$. The mean-based prediction interval is

$$C^{\text{reg}}_{(1-\alpha)}(x) = \left[x - Q_{|R|}(1-\alpha), x + Q_{|R|}(1-\alpha)\right], \qquad \textbf{[6]}$$

where $Q_{|R|}(1-\alpha)$ is the $(1-\alpha)$ quantile of the distribution of $|R_t|$. An important property and drawback of $C^{\text{reg}}_{(1-\alpha)}$ is that its length, $2 \cdot Q_{|R|}(1-\alpha)$, is fixed and does not depend on $X_{T+1} = x$ (16, 20). This feature implies that $C^{\text{reg}}_{(1-\alpha)}$ is not adaptive to the heteroskedasticity in the location-scale model Eq. **5** and not conditionally valid.

[†] This transformation is also very useful in other prediction problems (e.g., ref. 19).

DCP is based on the ranks $U_t = \Phi(\varepsilon_t)$, where $\Phi(\cdot)$ is the CDF of $N(0,1)$. The DCP prediction interval is

$$\mathcal{C}^{\text{dcp}}_{(1-\alpha)}(x) = \left[x - x \cdot Q_{|\varepsilon|}(1-\alpha),\, x + x \cdot Q_{|\varepsilon|}(1-\alpha)\right], \quad \textbf{[7]}$$

where $Q_{|\varepsilon|}(1-\alpha) = \Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha)$ quantile of $|\varepsilon_t|$. Unlike $\mathcal{C}^{\text{reg}}_{(1-\alpha)}$, the length of $\mathcal{C}^{\text{dcp}}_{(1-\alpha)}$, $2x \cdot Q_{|\varepsilon|}(1-\alpha)$, depends on $X_{T+1} = x$. Our construction automatically adapts to the heteroskedasticity in model Eq. **5** and is conditionally valid.

Fig. 1 provides an illustration. Fig. 1*A* shows that the conditional length of $\mathcal{C}^{\text{reg}}_{(0.9)}$ is constant, whereas the length of $\mathcal{C}^{\text{dcp}}_{(0.9)}$ varies as a function of x. $\mathcal{C}^{\text{dcp}}_{(0.9)}$ is shorter than $\mathcal{C}^{\text{reg}}_{(0.9)}$ for low values and wider for high values of x. Fig. 1*B* shows that $\mathcal{C}^{\text{dcp}}_{(0.9)}$ is valid for all x, whereas $\mathcal{C}^{\text{reg}}_{(0.9)}$ overcovers for low values and undercovers for high values of x. Fig. 1 illustrates the advantage of our method.

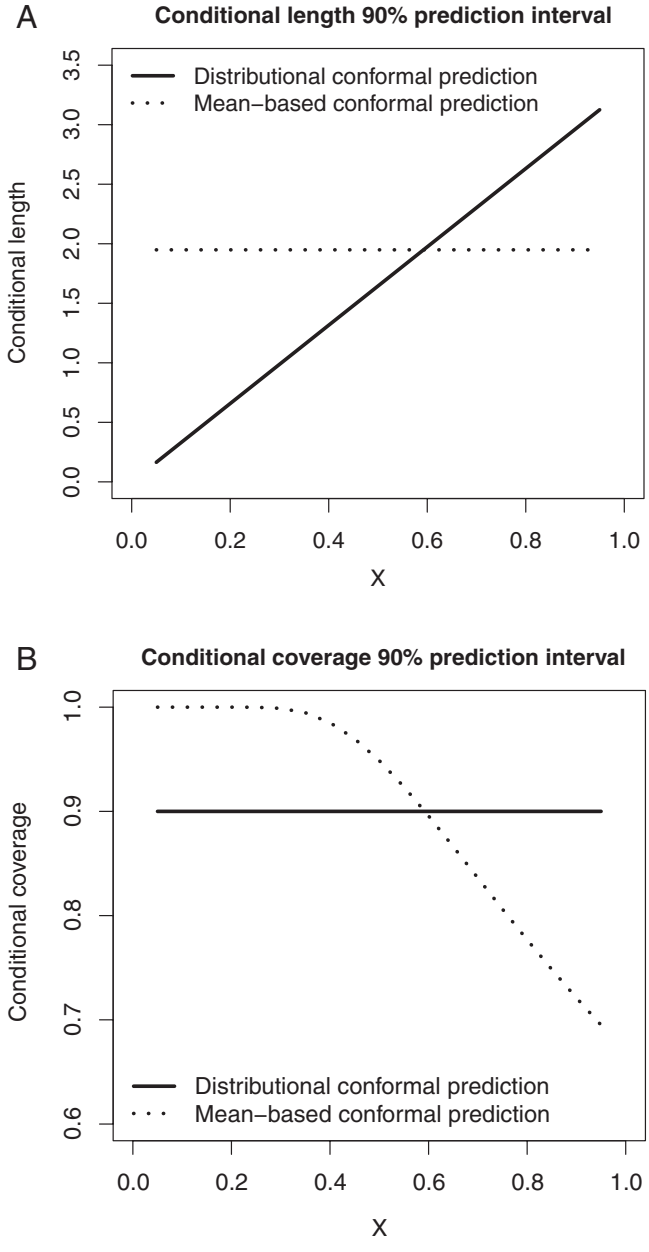

Fig. 1. Motivating example. (*A*) Conditional length 90% prediction interval. (*B*) Conditional coverage 90% prediction interval.

For predictor values where the conditional variance is low, it yields shorter prediction intervals while ensuring conditional coverage for values where the conditional dispersion is large by suitably enlarging the prediction interval.

Related Literature

We build on and contribute to the literature on conformal prediction (e.g., refs. 13, 15–18, 20, 23, and 24) and the literature on model-free prediction (19, 25), as well as the literature on quantile prediction methods (e.g., ref. 26 has a review).

Within the conformal prediction literature, our paper is most closely related to refs. 13, 16, and 20. Ref. 13 proposes conditionally valid and asymptotically efficient conformal prediction intervals based on estimators of the conditional density. We take a different and complementary approach, allowing researchers to leverage powerful regression methods for modeling conditional distributions, including QR and DR approaches. Ref. 20 develops conformal prediction methods based on regression models for conditional expectations. However, as discussed in *Motivating Example*, this approach is not conditionally valid under heteroskedasticity. They also propose a locally weighted conformal prediction approach, where the regression residuals are weighted by the inverse of a measure of their variability. This approach can alleviate some of the limitations of mean-based conformal prediction but is motivated by and based on restrictive locations-scale models. By contrast, our approach is generic and exploits flexible and substantially more general models for the whole conditional distribution.

Ref. 16 proposes a split conformal approach based on QR models, which they call conformalized quantile regression (CQR). Refs. 14 and 27 have related approaches, and ref. 28 has a general approach to adaptive conformal prediction. CQR is based on splitting the data into two subsets, $\mathcal{T}_1$ and $\mathcal{T}_2$. Based on $\mathcal{T}_1$, they estimate two separate quantile functions $\hat{Q}(\alpha/2, x)$ and $\hat{Q}(1-\alpha/2, x)$ and construct the prediction intervals as

$$\left[\hat{Q}(\alpha/2, x) - Q_E(1-\alpha),\, \hat{Q}(1-\alpha/2, x) + Q_E(1-\alpha)\right],$$

where $Q_E(1-\alpha)$ is the $(1-\alpha)(1+1/|\mathcal{T}_2|)$ th empirical quantile of

$$E_t = \max\left\{\hat{Q}(\alpha/2, X_t) - Y_t,\, Y_t - \hat{Q}(1-\alpha/2, X_t)\right\}$$

in $\mathcal{T}_2$. Constructing prediction intervals based on deviations from quantile estimates is similar to working with deviations from mean estimates, as the deviations are measured in absolute levels. By contrast, exploiting the probability integral transform, our approach is generic and relies on permuting ranks, which naturally have the same scaling on $(0,1)$. Note, however, that our paper was inspired by ref. 16, and we view our proposal as a (fully quantile rank–based) refinement of ref. 16. The value of this refinement is especially apparent in the second empirical example. In addition, we also give quantile-based optimal prediction intervals.

Our adjustment for constructing efficient prediction intervals is related to and inspired by conformal prediction literature on minimum-volume prediction sets based on density estimators (e.g., refs. 13, 23, and 29–31) and nearest-neighbor estimators (32). It is most closely related and can be viewed as an alternative to conformal histogram regression (33). The main differences between our approach and conformal histogram regression are the following. First, our method is based on an optimization problem formulated in terms of estimated quantile functions and does not require estimating a conditional density or histogram. Second, we do not work with nested sets but instead, use a simple adjustment of our baseline conformity score. Finally, our approach works for general outcome distributions and does not rely on assuming unimodal distributions.

Conceptually, our paper is further related to the transformation-based model-free prediction approach developed in refs. 19 and 25 in that we rely on transformations of the original setup into one that is easier to work with (i.e., ranks that are uniformly distributed) and study the properties of our approach in a model-free setting. An important difference is the implementation of the resulting procedure. The transformation-based approach is based on the bootstrap, whereas our approach is based on permuting ranks. Permuting ranks estimated based on the augmented data guarantees the model-free finite-sample validity of our method with exchangeable data. To our knowledge, no exact finite-sample validity results have been developed for the bootstrap-based approach.

DCP

Here, we introduce DCP. We present a full and a split-sample version of our method.

Full DCP. Let y denote a test value for Y_{T+1}. We test the plausibility of each value $y \in \mathbb{R}$, collect all plausible values, and report them as the prediction set. In practice, we consider a grid of test values $\mathcal{Y}_{\text{trial}}$.[‡] Define the augmented data $Z^{(y)} = \{Z_t^{(y)}\}_{t=1}^{T+1}$, where

$$Z_t^{(y)} = \begin{cases} (Y_t, X_t) & \text{if } 1 \le t \le T \\ (y, X_t) & \text{if } t = T+1 \end{cases}. \quad [8]$$

Based on the augmented dataset $Z^{(y)}$, we estimate the conditional CDF using a suitable method such as QR and DR, which are discussed in more detail in *SI Appendix*. Let $\hat{F}^{(y)}$ denote the estimator for F based on the augmented sample. If the original estimate is not monotonic, we rearrange it (e.g., refs. 35 and 36) so that $\hat{F}^{(y)}(\cdot, x)$ is always monotonic. To simplify the exposition, we keep these rearrangements implicit.

We compute the ranks $\{\hat{U}_t^{(y)}\}_{t=1}^{T+1}$, where

$$\hat{U}_t^{(y)} = \begin{cases} \hat{F}^{(y)}(Y_t, X_t) & \text{if } 1 \le t \le T \\ \hat{F}^{(y)}(y, X_t) & \text{if } t = T+1 \end{cases}, \quad [9]$$

and obtain P values as

$$\hat{p}(y) = \frac{1}{T+1} \sum_{t=1}^{T+1} \mathbf{1}\left\{\hat{V}_t^{(y)} \ge \hat{V}_{T+1}^{(y)}\right\}, \quad [10]$$

where $\hat{V}_t^{(y)} := \psi(\hat{U}_t^{(y)})$ and $\psi(\cdot)$ is a deterministic function. For our baseline method, we use $\psi(x) = |x - 1/2|$. In *Extension: Optimal DCP*, we show how to choose ψ optimally to ensure efficiency. Prediction intervals are computed as $\widehat{\mathcal{C}}^{\text{full}}_{(1-\alpha)}(X_{T+1}) = \{y \in \mathcal{Y}_{\text{trial}} : \hat{p}(y) > \alpha\}$.[§] We summarize our approach in *Algorithm 1*.

Algorithm 1: (Full DCP).

Input: Data $\{(Y_t, X_t)\}_{t=1}^{T}$, miscoverage level $\alpha \in (0,1)$, a point X_{T+1}, test values $\mathcal{Y}_{\text{trial}}$

Process: For $y \in \mathcal{Y}_{\text{trial}}$,

1) define the augmented data $Z^{(y)}$ as in Eq. **9**
2) compute $\hat{p}(y)$ as in Eq. **10**

Output: Return $(1-\alpha)$ prediction set $\widehat{\mathcal{C}}^{\text{full}}_{(1-\alpha)}(X_{T+1}) = \{y \in \mathcal{Y}_{\text{trial}} : \hat{p}(y) > \alpha\}$

Split DCP. An important drawback of full DCP (*Algorithm 1*) is its computational burden due to the grid search. Since $\hat{F}^{(y)}$ is obtained based on the augmented data, one has to choose $\mathcal{Y}_{\text{trial}}$ and reestimate the entire conditional distribution for all $y \in \mathcal{Y}_{\text{trial}}$. Therefore, we propose a split conformal procedure that exploits sample splitting, avoids grid search, and only requires estimating F once. Sample splitting is a popular approach for improving the computational performance of conformal prediction methods (e.g., refs. 16 and 20).

Algorithm 2: (Split DCP).

Input: Data $\{(Y_t, X_t)\}_{t=1}^{T}$, miscoverage level $\alpha \in (0,1)$, point X_{T+1}

Process:

1) Split $\{1, \dots, T\}$ into $\mathcal{T}_1 := \{1, \dots, T_0\}$ and $\mathcal{T}_2 := \{T_0 + 1, \dots, T\}$
2) Obtain $\hat{F}$ based on $\{Z_t\}_{t \in \mathcal{T}_1}$
3) Compute $\{\hat{V}_t\}_{t \in \mathcal{T}_2} = \{\psi(\hat{U}_t)\}_{t \in \mathcal{T}_2}$, where $\hat{U}_t = \hat{F}(Y_t, X_t)$
4) Compute $\hat{Q}_{\mathcal{T}_2}$, the $(1-\alpha)(1 + 1/|\mathcal{T}_2|)$ empirical quantile of $\{\hat{V}_t\}_{t \in \mathcal{T}_2}$

Output: Return $(1-\alpha)$ prediction set $\widehat{\mathcal{C}}^{\text{split}}_{(1-\alpha)}(X_{T+1}) = \left\{y : \psi\left(\hat{F}(y, X_{T+1})\right) \le \hat{Q}_{\mathcal{T}_2}\right\}$

(Since $\hat{F}(\cdot, X_{T+1})$ is monotonic, $\widehat{\mathcal{C}}^{\text{split}}_{(1-\alpha)}(X_{T+1})$ is an interval)

In *Algorithm 2*, we split $\{1, \dots, T\}$ into $\{1, \dots, T_0\}$ and $\{T_0 + 1, \dots, T\}$. With iid data, one can also consider random splits.

Split DCP lends itself naturally to simple in-sample validity checks with both cross-sectional and time series data as illustrated in *Empirical Applications*.

Theoretical Performance Guarantees

In this section, we establish the theoretical properties of our procedure. We focus on full-sample DCP (*Algorithm 1*). For the split-sample approach (*Algorithm 2*), we provide a modified version (*SI Appendix*, *Algorithm S1*) in *SI Appendix* and present its theoretical properties in *Extension: Optimal DCP*.

When the data are iid (or exchangeable), our method achieves finite-sample unconditional validity in a model-free manner, as a consequence of general results on conformal inference and permutation inference more generally (e.g., refs. 17 and 37).

Theorem 1 (Finite-sample unconditional validity). *Suppose that the data are iid or exchangeable and that the estimator of the conditional distribution is invariant to permutations of the data. Then*,

$$P\left(Y_{T+1} \in \widehat{\mathcal{C}}^{\text{full}}_{(1-\alpha)}(X_{T+1})\right) \ge 1 - \alpha.$$

The proof of *Theorem 1* is standard and omitted. *Theorem 1* highlights the strengths and drawbacks of conformal prediction methods. Most commonly used estimators of the conditional CDF such as QR and DR are invariant to permutations of the data. As a result, *Theorem 1* provides a model-free unconditional performance guarantee in finite samples, allowing for arbitrary misspecification of the model of the conditional CDF. On the other hand, it has a major theoretical drawback. Even with iid data, it provides no guarantee at all on conditional validity.

Our next theoretical results provide a remedy. We impose the following weak regularity conditions.

Assumption 1. *Suppose that there exists a nonrandom function $F^*(\cdot, \cdot)$ such that the following conditions hold as $T \to \infty$. Define $V_t := \psi(F^*(Y_t, X_t))$ for $1 \le t \le T+1$.*

[‡] For example, we can choose $\mathcal{Y}_{\text{trial}}$ to be a fine grid between $-\max_{1 \le t \le T} |Y_t|$ and $\max_{1 \le t \le T} |Y_t|$. This choice has a theoretical justification since under exchangeability, $P\left(|Y_{T+1}| > \max_{1 \le t \le T} |Y_t|\right) \le 1/(1+T)$ (34) (a discussion is in the conformal Inference R-package; https://github.com/ryantibs/conformal).

[§] Instead of $\widehat{\mathcal{C}}^{\text{full}}_{(1-\alpha)}(X_{T+1})$, we typically report the closed interval $\tilde{\mathcal{C}}^{\text{full}}_{(1-\alpha)}(X_{T+1}) = \left[\min\left(\widehat{\mathcal{C}}^{\text{full}}_{(1-\alpha)}(X_{T+1})\right), \max\left(\widehat{\mathcal{C}}^{\text{full}}_{(1-\alpha)}(X_{T+1})\right)\right]$.

1) *There exists a strictly increasing continuous function* $\phi : [0, \infty) \to [0, \infty)$ *such that* $\phi(0) = 0$ *and* $(T + 1)^{-1} \sum_{t=1}^{T+1} \phi(|\hat{V}_t - V_t|) = o_P(1)$ *and* $\hat{V}_{T+1} = V_{T+1} + o_P(1)$, *where* $\hat{V}_t := \hat{V}_t^{(Y_{T+1})} = \psi(\hat{F}^{(Y_{T+1})}(Y_t, X_t))$ *for* $1 \le t \le T + 1$.
2) $\sup_{v \in \mathbb{R}} |\tilde{G}(v) - G(v)| = o_P(1)$, *where* $\tilde{G}(v) = (T + 1)^{-1} \sum_{t=1}^{T+1} \mathbf{1}\{V_t < v\}$ *and* $G(\cdot)$ *is the distribution function of* V_{T+1}.
3) $\sup_{x_1 \neq x_2} |G(x_1) - G(x_2)|/|x_1 - x_2|$ *is bounded*.

Assumption 1 allows for some flexibility with respect to the model estimator. Here, we only require F^* to be a nonrandom function, which may or may not be F. The interpretation is straightforward when $F^* = F$ since this simply means that the estimator $\hat{F}$ is consistent for F. We discuss the case of $F^* \neq F$ after *Theorem 2* below. Note that we can replace the consistency requirement in *Assumption 1* with a stronger uniform consistency requirement, $\sup_{x,y} |\hat{F}(y, x) - F^*(y, x)| = o_P(1)$.

We also notice that the quantities $\hat{V}_t$ and V_t are defined under the true Y_{T+1}. This means that $\hat{F}^{(y)}$ uses $y = Y_{T+1}$. In other words, the estimator $\hat{F}$ based on the sample $\{(X_t, Y_t)\}_{t=1}^{T+1}$ would be consistent for some F^* if Y_{T+1} were observed.[¶] Since the goal of *Assumption 1* is to guarantee the coverage probability for Y_{T+1}, the conditions in *Assumption 1* only need to hold for $y = Y_{T+1}$.

Notice that $\hat{F}$ is consistent for F^* under a very weak norm, and no rate condition is required. When $\psi(x) = |x - 1/2|$, a simple example of $\phi(\cdot)$ in *Assumption 1* is $\phi(x) = x^q$ for some $q > 0$; in other words, a sufficient condition is $(T + 1)^{-1} \sum_{t=1}^{T+1} |\hat{F}(Y_t, X_t) - F^*(Y_t, X_t)|^q = o_P(1)$, which can be verified for many existing estimators with $q = 2$.

The following lemma gives the basic consistency result.

Lemma 1. *Let Assumption 1 hold. Then,* $\hat{G}(\hat{V}_{T+1}) = G(V_{T+1}) + o_P(1)$, *where* $\hat{G}(v) = (T + 1)^{-1} \sum_{t=1}^{T+1} \mathbf{1}\{\hat{V}_t < v\}$.

By *Assumption 1*, $G(\cdot)$ is uniformly continuous and thus, continuous. Since $G(\cdot)$ is the distribution function of V_{T+1}, we have that $G(V_{T+1})$ has the uniform distribution on $(0, 1)$ [i.e., $P(G(V_{T+1}) \le \alpha) = \alpha$]. This implies the unconditional asymptotic validity.

Theorem 2 (Asymptotic unconditional validity). *Let Assumption 1 hold. Then,*

$$P\left(Y_{T+1} \in \widehat{\mathcal{C}}^{\text{full}}_{(1-\alpha)}(X_{T+1})\right) = 1 - \alpha + o(1).$$

Theorem 2 establishes the asymptotic unconditional validity of the procedure. Since *Theorem 1* already establishes the unconditional validity in finite samples for iid or exchangeable data without assuming any consistency of $\hat{F}$, the main purpose of *Theorem 2* is to address the case of nonexchangeable data (e.g., time series data with ergodicity), especially when the model is misspecified (i.e., if $F^* \neq F$) .

To illustrate model misspecification, consider the popular linear QR model, which assumes $Q(\tau, x) = x^\top \beta(\tau)$, and thus, $F(y, x) = F(y, x; \beta) = \int_0^1 \mathbf{1}\{x^\top \beta(\tau) \le y\} d\tau$. This model is typically estimated by $\hat{\beta}(\tau) = \arg\min_\beta \sum_{t=1}^{T+1} \rho_\tau(Y_t - X_t^\top \beta)$ with $\rho_\tau(a) = a(\tau - \mathbf{1}\{a < 0\})$. Under misspecification [$Q(\tau, x) \neq x^\top \beta(\tau)$], $\hat{\beta}(\tau)$ is still estimating $\beta^*(\tau) = \arg\min_\beta \sum_{t=1}^{T+1} E\rho_\tau(Y_t - X_t^\top \beta)$, and F^* is defined using $\beta^*(\cdot)$ [e.g., $F^*(y, x) = \int_0^1 \mathbf{1}\{x^\top \beta^*(\tau) \le y\} d\tau$]. For parametric models, F^* is usually the probability limit of $\hat{F}$. In general, we can consider a model $\mathcal{F}$ and minimize the empirical risk $\hat{F} = \arg\min_{g \in \mathcal{F}} \sum_{t=1}^{T+1} L(Y_t, X_t, g)$ for some loss function L. Even if the model is misspecified ($F \notin \mathcal{F}$) , it is still possible to show that $\hat{F}$ is close (in some norm) to $F^* = \arg\min_{g \in \mathcal{F}} \sum_{t=1}^{T+1} E[L(Y_t, X_t, g)]$. In *SI Appendix*, we provide a more detailed discussion of this and some theoretical results verifying the consistency requirement in *Assumption 1* for the time series case; ref. 24 has a general discussion of conformal prediction in time series settings.

The cost of allowing for misspecification is that one cannot guarantee conditional validity when $F^* \neq F$. On the other hand, *Lemma 1* implies that the prediction intervals are conditionally valid when $F^* = F$.

Theorem 3 (Asymptotic conditional validity). *Let Assumption 1 hold with* $F^* = F$. *Then,*

$$P\left(Y_{T+1} \in \widehat{\mathcal{C}}^{\text{full}}_{(1-\alpha)}(X_{T+1}) \mid X_{T+1}\right) = 1 - \alpha + o_P(1).$$

Theorems 2 and *3* establish the asymptotic validity of our procedure under weak and easy to verify conditions. They formalize the key intuition that conditional validity hinges on the quality of the estimator $\hat{F}$ of the conditional CDF.[#]

Extension: Optimal DCP

In *Theoretical Performance Guarantees*, we have seen that a generic conformity score $\psi(y, x) = |F(y, x) - 1/2|$ leads to conditional validity if the conditional distribution F can be estimated consistently. We now characterize an optimal choice of conformity score that results in the shortest prediction interval. Detailed implementation algorithms, technical assumptions, and proofs are provided in *SI Appendix*.

Let $\mathcal{Z}$ and $\mathcal{X}$ denote the support of $Z_t = (Y_t, X_t)$ and X_t, respectively. The optimal prediction interval is

$$\mathcal{C}^{\text{opt}}_{(1-\alpha)}(x) = [r_1(x, \alpha),\ r_2(x, \alpha)], \qquad [11]$$

where the functions $r_1(\cdot, \cdot), r_2(\cdot, \cdot)$ satisfy that for any $x \in \mathcal{X}$,

$$r_2(x, \alpha) - r_1(x, \alpha) = \min_{F(z_2, x) - F(z_1, x) \ge 1 - \alpha} z_2 - z_1. \qquad [12]$$

The question is whether it is possible to design a conformity score that achieves the above optimal prediction interval. To answer this question formally, we consider a generic conformity score $\psi(y, x)$, which might contain components that need to be estimated.

Permuting a large number of values of $\{\psi(Y_t, X_t)\}$ in conformal predictions leads to taking the sample $(1 - \alpha)$ quantile of $\psi(Y_t, X_t)$ as the output. For example, following *Algorithm 2*, one would output the $(1 - \alpha)(1 + 1/|\mathcal{I}_2|)$ empirical quantile of $\{\psi(Y_t, X_t)\}$. Assuming a law of large numbers, this empirical quantile would be close to the population $(1 - \alpha)$ quantile of $\psi(Y_t, X_t)$, leading to the asymptotic conformal prediction interval for Y_{T+1}:

$$\mathcal{C}^{\text{conf}}_{(1-\alpha)}(X_{T+1}) = \{y :\ \psi(y, X_{T+1}) \le Q_\psi(1 - \alpha)\}, \qquad [13]$$

where $Q_\psi(1 - \alpha)$ is the $(1 - \alpha)$ quantile of $\psi(Y_t, X_t)$. The following result shows how to construct the optimal conformity score ψ.

Lemma 2. *Let* $\psi_*(y, x) = |F(y, x) - b(x, \alpha) - (1 - \alpha)/2|$, *where* $b(\cdot, \cdot)$ *is a function satisfying that for any* $x \in \mathcal{X}$,

$$b(x, \alpha) \in \arg\min_{z \in [0, \alpha]} Q(z + 1 - \alpha, x) - Q(z, x). \qquad [14]$$

[¶] This is not really much different from assuming that $\hat{F}$ based on the sample $\{(X_t, Y_t)\}_{t=1}^{T}$ is consistent for some F^*.

[#] In *Theorem 3*, we assume $F^* = F$. Since the first version of this paper was posted, ref. 38 has provided more general results where $F^* \approx F$.

Let $\mathcal{C}^{\text{conf}}_{(1-\alpha)}(X_{T+1})$ *be defined as in Eq.* **13** *with the above conformity score* ψ_*. *Assume that* $F(\cdot, x)$ *is a continuous function for any* $x \in \mathcal{X}$. *Then,* $Q_\psi(1-\alpha) = (1-\alpha)/2$ *and*

$$\mu\left(\mathcal{C}^{\text{opt}}_{(1-\alpha)}(X_{T+1})\right) = \mu\left(\mathcal{C}^{\text{conf}}_{(1-\alpha)}(X_{T+1})\right) \text{ almost surely},$$

where $\mu(\cdot)$ *denotes the Lebesgue measure. If the optimization problem in Eq.* **11** *has a unique solution for any* $x \in \mathcal{X}$, *then*

$$\mathcal{C}^{\text{opt}}_{(1-\alpha)}(X_{T+1}) = \mathcal{C}^{\text{conf}}_{(1-\alpha)}(X_{T+1}) \text{ almost surely}.$$

Lemma 2 motivates conformity scores of the form $\psi_*(y,x) = |F(y,x) - [b(x,\alpha) + (1-\alpha)/2]|$, where $b(\cdot,\cdot)$ solves Eq. **14**. Compared with the choice of $\psi(y,x) = |F(y,x) - 1/2|$ mentioned in *Theoretical Performance Guarantees*, we can view ψ_* as having a "shape" adjustment $b(x,\alpha) - \alpha/2$. Since $F(Y_t, X_t)$ is independent of X_t, the optimal conformity score measures the distance between two independent components: $F(Y_t, X_t)$ and $1/2 + (b(X_t,\alpha) - \alpha/2)$. Hence, by *Lemma 2*, in order to take into account the shape of the conditional distribution $F(\cdot, x)$, it suffices to consider the scalar quantity $1/2 + (b(x,\alpha) - \alpha/2)$.

In some special cases, the shape adjustment can be shown to be zero [i.e., $b(x,\alpha) = \alpha/2$]. One typical example is when $F(\cdot, x)$ is a symmetric unimodal distribution with a well-defined conditional density.[‖] Therefore, the choice of $\psi(y,x) = |F(y,x) - 1/2|$ mentioned in *Theoretical Performance Guarantees* is optimal in these cases. However, *Lemma 2* provides a construction that achieves optimality more generally. By the definition of ψ_* and $Q_\psi(1-\alpha) = (1-\alpha)/2$, the prediction interval is

$$\mathcal{C}^{\text{conf}}_{(1-\alpha)}(x) = [Q(b(x,\alpha), x),\ Q(b(x,\alpha) + 1 - \alpha, x)]. \qquad \textbf{[15]}$$

We illustrate this in Fig. 2 with $\alpha = 0.1$. Eq. **15** implies that $b(x,\alpha)$ is the quantile index of the lower bound of the interval. For the symmetric distribution in Fig. 2, *Top*, we see $b(x,\alpha) = 0.05$, which is $\alpha/2$. For the asymmetric distribution in Fig. 2, *Bottom*, we see that $b(x,\alpha) = 0.007$, which is far away from $\alpha/2 = 0.05$.

The first result in *Lemma 2* is general and allows for the lack of uniqueness of the optimal prediction interval. For example, if F is the uniform distribution on a certain interval, then all conditionally valid prediction intervals have the same length. Clearly, in this case, achieving the optimal length is the only goal one can hope for. When we can uniquely define the optimal prediction interval, *Lemma 2* implies that the conformal procedure can recover the uniquely defined optimal interval, not just achieving the optimal length.

Lemma 2 also confirms the insight of ref. 13; the optimal confidence set for $X_{T+1} = x$ should take the form $\{y : f(y,x) \geq c(x)\}$ for some $c(x) > 0$, where $f(y,x) = \partial F(y,x)/\partial y$. Assume that $F(\cdot, x)$ is a unimodal distribution and $f(\cdot, x)$ is a continuous function for any $x \in \mathcal{X}$. Then, this confidence set is an interval. This means that $\{y : f(y,x) \geq c(x)\} = [c_1(x), c_2(x)]$ and $f(c_1(x), x) = f(c_2(x), x) = c(x)$. We notice that $c_1(x), c_2(x)$ are related to our results in that $c_1(x) = Q(b(x,\alpha), x)$ and $c_2(x) = Q(b(x,\alpha) + 1 - \alpha, x)$. To see this, simply observe that the first-order condition of the optimization problem in Eq. **14** is $1/f(Q(z + 1 - \alpha, x), x) - 1/f(Q(z, x)) = 0$, which implies that

$$f(Q(b(x,\alpha) + 1 - \alpha, x)) = f(Q(b(x,\alpha), x)).$$

To make the procedure operational, we provide the conformal prediction interval $\widehat{\mathcal{C}}^{\text{conf}}_{(1-\alpha)}(X_{T+1})$ defined in *SI Appendix*, *Algorithm S1*. We can provide the following guarantee.

Theorem 4. *Let SI Appendix, Assumption S1 hold. Then,*

$$P\left(Y_{T+1} \in \widehat{\mathcal{C}}^{\text{conf}}_{(1-\alpha)}(X_{T+1}) \mid X_{T+1}\right) = 1 - \alpha + o_P(1)$$

and

$$\mu\left(\widehat{\mathcal{C}}^{\text{conf}}_{(1-\alpha)}(X_{T+1})\right) \leq \mu\left(\mathcal{C}^{\text{opt}}_{(1-\alpha)}(X_{T+1})\right) + o_P(1).$$

The main requirements in *SI Appendix, Assumption S1* are consistency of $\hat{F}$ and that the density f bounded below on its

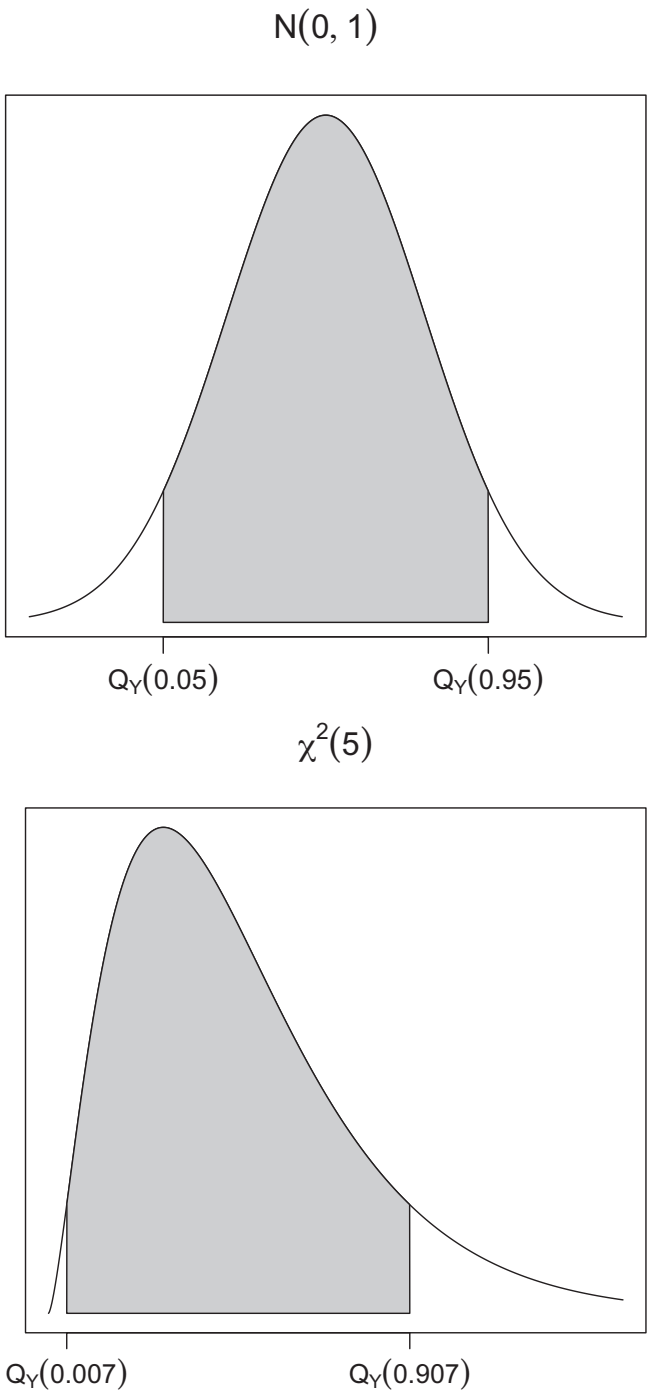

Fig. 2. Optimal prediction intervals. (*Top*) Symmetric distribution. (*Bottom*) Asymmetric distribution.

[‖] In this case, $Q(1/2+\delta, x) - Q(1/2, x) = Q(1/2, x) - Q(1/2-\delta, x)$, and the conditional density is increasing on $(-\infty, Q(1/2, x))$ and decreasing on $(Q(1/2, x), \infty)$. One can show $b(x, \alpha) = \alpha/2$ by taking the first-order derivative for the optimization problem in Eq. **14** and setting it to zero.

support. This is quite mild in the sense that it does not imply that the optimal prediction interval in Eq. **11** is uniquely defined. For example, it allows f to be a uniform distribution. Therefore, as discussed above, the conformal prediction interval would have approximately the shortest length but might not converge to $\mathcal{C}^{\text{opt}}_{(1-\alpha)}(X_{T+1})$ in Eq. **11**.

The following theorem provides a stronger result about $\widehat{\mathcal{C}}^{\text{conf}}_{(1-\alpha)}(X_{T+1})$ based on stronger assumptions.

Theorem 5. *Let SI Appendix, Assumption S2 hold. Consider the conformal prediction interval* $\widehat{\mathcal{C}}^{\text{conf}}_{(1-\alpha)}(X_{T+1})$ *defined in SI Appendix, Algorithm S1. Then,*

$$\mu\left(\widehat{\mathcal{C}}^{\text{conf}}_{(1-\alpha)}(X_{T+1}) \triangle \mathcal{C}^{\text{opt}}_{(1-\alpha)}(X_{T+1})\right) = o_P(1),$$

where $\triangle$ *denotes the symmetric difference of sets* [*i.e.*, $A \triangle B = (A \backslash B) \bigcup (B \backslash A)$], $\mathcal{C}^{\text{opt}}_{(1-\alpha)}(X_{T+1})$ *is defined in Eq.* **11**.

The key component of *SI Appendix, Assumption S2* is consistent estimation of b. *Theorem 5* shows that $\widehat{\mathcal{C}}^{\text{conf}}_{(1-\alpha)}(X_{T+1})$ is close to $\mathcal{C}^{\text{opt}}_{(1-\alpha)}(X_{T+1})$ in the sense that the symmetric difference between these two sets has vanishing Lebesgue measure.

Empirical Applications

We illustrate the performance of DCP in two empirical applications and provide a comparison with alternative approaches. These examples, especially the second, illustrate the value of our proposal. We consider eight different conformal prediction methods.

1) **DCP-QR:** DCP with QR (*Algorithm 2*)
2) **DCP-QR*:** Optimal DCP with QR (*SI Appendix, Algorithm S1*)
3) **DCP-DR:** DCP with DR (*Algorithm 2*)
4) **CQR:** CQR with QR (16)
5) **CQR-m:** CQR variant (14, 27) with QR
6) **CQR-r:** CQR variant (14) with QR
7) **CP-OLS:** Mean-based split conformal prediction (CP) with Ordinary Least Squares (OLS)
8) **CP-loc:** Locally weighted conformal prediction (20) with OLS

All computations were carried out in R (39). Code and data for replicating the empirical results are deposited in GitHub (https://github.com/kwuthrich/Replication_DCP).

Predicting Stock Market Returns. Here, we consider the problem of predicting stock market returns, which are known to exhibit substantial heteroskedasticity (a recent review is in chapter 13 in ref. 40 and references therein). We use data on daily returns of the market portfolio (Center for Research in Security Prices value-weighted portfolio) from 1 July 1926 to 30 June 2021.** We use lagged realized volatility X_t to predict the present return Y_t.†† Daily returns are not iid and exhibit time series dependence. In *SI Appendix*, we show that the key conditions underlying our theoretical results hold when the data are β-mixing. Several stochastic volatility models for asset returns, including the popular generalized autoregressive conditional heteroskedasticity models, can be shown to be β-mixing (e.g., refs. 42–44).

We evaluate the performance of the different methods by splitting the data into a training and a test sample. To account for the dependence in the data, we present results averaged over five consecutive prediction exercises. In the first exercise, we apply split conformal prediction with an equal split ($|\mathcal{T}_1| = |\mathcal{T}_2|$) to the first 50% of observations and use the next 10% for testing. In the second exercise, we drop the first 10% of the observations, apply split conformal prediction to the next 50% of observations, and use the next 10% for testing and so on.

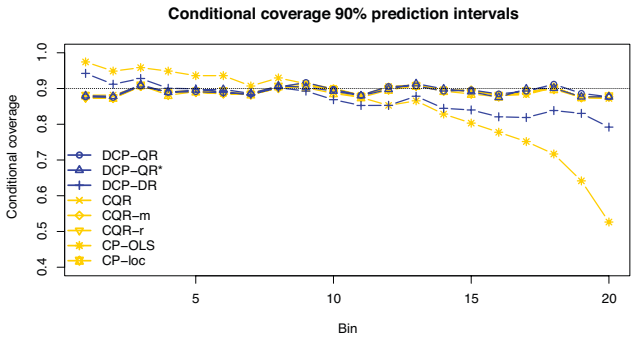

Fig. 3. Conditional coverage 90% prediction intervals by realized volatility.

Fig. 3 plots the empirical coverage probabilities for 20 bins obtained by dividing up the support of X_t based on equally spaced quantiles. DCP-QR and DCP-QR* yield prediction intervals with coverage levels that are almost constant across all bins and close to the nominal level. They outperform DCP-DR, which undercovers in high-volatility regimes. The conditional coverage properties of DCP-QR and DCP-QR* are very similar to CQR, CQR-m, CQR-r, and CP-loc. This suggest that location-scale models, which are nested by QR, provide a good approximation of the conditional distribution. CP-OLS exhibits overcoverage under low-volatility regimes and substantial undercoverage under high-volatility regimes. This finding has important practical implications since the volatility tends to be high during periods of crisis, which is precisely when accurate risk assessments are most needed.

Fig. 4 shows the conditional length of the prediction intervals. DCP-QR, DCP-QR*, CQR, CQR-m, CQR-r, and CP-loc yield prediction intervals of similar length. The DCP-DR prediction intervals are somewhat shorter than those of the QR-based methods at the upper tail. Finally, CP-OLS yields prediction intervals that are almost constant across all values of realized volatility; they are longer than the DCP intervals at the lower tail and shorter at the upper tail.‡‡

Predicting Wages Using CPS Data. We consider the problem of predicting wages using individual characteristics. We use the 2012 Current Population Survey (CPS) data provided in the R package hdm (45), which contain information on $N = 29{,}217$ observations. Here, we use the index i instead of t. To illustrate the impact of skewness on the performance of the different prediction methods, we use the hourly wage as our dependent

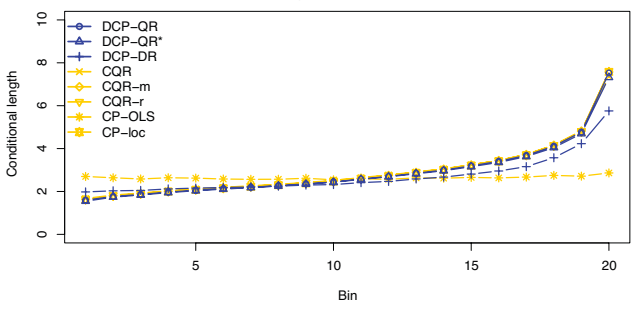

Fig. 4. Conditional length 90% prediction intervals by realized volatility.

**The data are constructed from the Fama/French Three Factors data (41) available from Kenneth R. French's data library (accessed 17 August 2021).

††We compute realized volatility as the square root of the sum of squared returns over the last 22 d.

‡‡The CP-OLS prediction intervals are not exactly constant because we are reporting results averaged over five experiments.

Table 1. Coverage 90% prediction intervals

	DCP-QR	DCP-QR*	DCP-DR	CQR	CQR-m	CQR-r	CP-OLS	CP-loc
Unconditional coverage	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
SD of predicted conditional coverage ($\times 100$)	1.80	1.71	3.08	2.21	2.36	2.30	11.13	4.11

variable Y_i.[###] Predictors X_i include indicators for gender, marital status, educational attainment, region, experience, experience squared, and all two-way interactions such that $\dim(X_i) = 100$ after removing constant variables.

Following refs. 14 and 16, we evaluate the performance of the different methods by randomly holding out 20% of the data for testing, $\mathcal{I}_{\text{test}}$, and applying split conformal prediction with an equal split to the remaining 80% of the data. We repeat the whole experiment 20 times.

Table 1 shows that all conformal prediction methods exhibit excellent unconditional coverage properties, confirming the theoretical finite-sample guarantees. To assess and compare the conditional coverage properties, for each method, we compute conditional coverage probabilities as the predictions from logistic regressions of $\{Y_i \in \widehat{C}^{\text{split}}_{(1-\alpha)}(X_i)\}_{i\in\mathcal{I}_{\text{test}}}$ on $\{X_i\}_{i\in\mathcal{I}_{\text{test}}}$, where $\widehat{C}^{\text{split}}_{(1-\alpha)}$ is the split conformal prediction interval obtained by the corresponding method. The less dispersed the predicted coverage probabilities are around the nominal level $1-\alpha = 0.9$, the better the overall conditional coverage properties of a method. Table 1 plots the SD of the predicted coverage probabilities.[¶¶] DCP-QR* yields the lowest dispersion of all methods. The predicted coverage probabilities based on DCP-QR are less dispersed than those obtained from CQR, CQR-m, and CQR-r. CP-loc yields a higher dispersion than the methods based on QR and DR, which demonstrates the value added of using flexible models of the conditional distribution. Overall, DCP performs much better than CP-OLS, for which the predicted coverage probabilities exhibit a very high dispersion. *SI Appendix*, Fig. S1 plots histograms of the predicted coverage probabilities.

Table 2. Average length 90% prediction intervals

DCP-QR	DCP-QR*	DCP-DR	CQR	CQR-m	CQR-r	CP-OLS	CP-loc
34.22	29.61	33.69	34.52	34.84	34.63	33.84	32.66

Table 2 shows the average length of the prediction intervals. DCP-QR* produces the shortest prediction intervals among of all methods. This demonstrates the practical advantage of the shape adjustment when the conditional distribution is skewed. The results also suggest a trade-off between conditional coverage accuracy and average length. For example, CP-OLS and CP-loc, which both exhibit poor conditional coverage properties, yield shorter prediction intervals than DCP-QR.

Data Availability. Data and computer codes to replicate all the results in this paper have been deposited in GitHub (https://github.com/kwuthrich/Replication_DCP). All data are referenced in the main text.

ACKNOWLEDGMENTS. We thank the editor, two anonymous referees, Dimitris Politis, and Allan Timmermann for valuable comments. V.C. acknowledges funding from the NSF. All remaining errors are our own.

[###] We obtain the hourly wage by exponentiating the log hourly wage provided in the dataset.

[¶¶] Using $\sqrt{1/|\mathcal{I}_{\text{test}}|\sum_{i\in\mathcal{I}_{\text{test}}}(\widehat{\text{Coverage}}_i - 0.9)^2}$, where $\widehat{\text{Coverage}}_i$ is the predicted coverage probability, instead of the SD yields very similar results.

1. R. Koenker, G. Bassett, Regression quantiles. *Econometrica* **46**, 33–50 (1978).
2. S. Foresi, F. Peracchi, The conditional distribution of excess returns: An empirical analysis. *J. Am. Stat. Assoc.* **90**, 451–466 (1995).
3. V. Chernozhukov, I. Fernandez-Val, B. Melly, Inference on counterfactual distributions. *Econometrica* **81**, 2205–2268 (2013).
4. J. W. Taylor, A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *J. Forecast.* **19**, 299–311 (2000).
5. P. Chaudhuri, W. Y. Loh, Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli* **8**, 561–576 (2002).
6. N. Meinshausen, Quantile regression forests. *J. Mach. Learn. Res.* **7**, 983–999 (2006).
7. M. D. Cattaneo, Y. Feng, R. Titiunik, Prediction intervals for synthetic control methods. arXiv [Preprint] (2019). https://arxiv.org/abs/1912.07120 (Accessed 20 August 2021).
8. V. Chernozhukov, K. Wüthrich, Y. Zhu, An exact and robust conformal inference method for counterfactual and synthetic controls. *J. Am. Stat. Assoc.*, 10.1080/01621459.2021.1920957 (2021).
9. D. Kivaranovic, R. Ristl, M. Posch, H. L. Leeb, Conformal prediction intervals for the individual treatment effect. arXiv [Preprint] (2020). https://arxiv.org/abs/2006.01474 (Accessed 20 August 2020).
10. L. Lei, E. J. Candès, Conformal inference of counterfactuals and individual treatment effects. arXiv [Preprint] (2020). https://arxiv.org/abs/2006.06138 (Accessed 20 August 2021).
11. Y. Romano, R. F. Barber, C. Sabatti, E. J. Candes, With malice towards none: Assessing uncertainty via equalized coverage. arXiv [Preprint] (2019). https://arxiv.org/abs/1908.05428 (Accessed 20 August 2021).
12. R. Foygel Barber, E. J. Candès, A. Ramdas, R. J. Tibshirani, The limits of distribution-free conditional predictive inference. *J. IMA* **10**, 455–482 (2021).
13. J. Lei, L. Wasserman, Distribution-free prediction bands for non-parametric regression. *J. Royal Stat. Soc. Ser. B. Stat. Methodol.* **76**, 71–96 (2014).
14. M. Sesia, E. J. Candes, A comparison of some conformal quantile regression methods. *Stat* **9**, e261 (2020).
15. V. Vovk, "Conditional validity of inductive conformal predictors" in *Proceedings of the Asian Conference on Machine Learning*, S. C. H. Hoi, W. Buntine, Eds. (PMLR, Singapore Management University, Singapore), vol. 25, pp. 475–490 (2012).
16. Y. Romano, E. Patterson, E. Candes, Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.*, **32**, 3543–3553 (2019).
17. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic Learning in a Random World* (Springer Science & Business Media, 2005).
18. V. Vovk, I. Nouretdinov, A. Gammerman, On-line predictive linear regression. *Ann. Stat.* **37**, 1566–1590 (2009).
19. D. N. Politis, *Model-Free Prediction and Regression: A Transformation-Based Approach to Inference* (Springer, New York, NY, 2015).
20. J. Lei, M. G. Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **113**, 1094–1111 (2018).
21. R. Koenker, G. Bassett, Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* **50**, 43–61 (1982).
22. R. Koenker, *Quantile Regression, Econometric Society Monographs* (Cambridge University Press, 2005).
23. J. Lei, J. Robins, L. Wasserman, Distribution free prediction sets. *J. Am. Stat. Assoc.* **108**, 278–287 (2013).
24. V. Chernozhukov, K. Wüthrich, Z. Yinchu, "Exact and robust conformal inference methods for predictive machine learning with dependent data" in *Proceedings of the 31st Conference on Learning Theory*, S. Bubeck, V. Perchet, P. Rigollet, Eds. (PMLR, Cambridge, MA, 2018), vol. 75, pp. 732–749.
25. D. N. Politis, Model-free model-fitting and predictive distributions. *Test* **22**, 183–221 (2013).
26. I. Komunjer, "Chapter 17 - Quantile prediction" in *Handbook of Economic Forecasting*, G. Elliott, A. Timmermann, Eds. (Elsevier, 2013), pp. 961–994.
27. D. Kivaranovic, K. D. Johnson, H. Leeb, "Adaptive, distribution-free prediction intervals for deep networks" in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, S. Chiappa, R. Calandra, Eds. (PMLR, Cambridge, MA, 2020), vol. 108, pp. 4346–4356.
28. V. Vovk *et al.*, "Conformal calibrators" in *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, A. Gammerman, V. Vovk, Z. Luo, E. Smirnov, G. Cherubin, Eds. (PMLR, Cambridge, MA, 2020), vol. 128, pp. 84–99.
29. D. J. Eck, F. W. Crawford, Efficient and minimal length parametric conformal prediction regions. arXiv [Preprint] (2019). https://arxiv.org/abs/1905.03657 (Accessed 20 August 2021).
30. R. Izbicki, G. T. Shimizu, R. B. Stern, Flexible distribution-free conditional predictive bands using density estimators. arXiv [Preprint] (2019). https://arxiv.org/abs/1910.05575 (Accessed 20 August 2021).
31. R. Izbicki, G. Shimizu, R. B. Stern, CD-split and HPD-split: Efficient conformal regions in high dimensions. arXiv [Preprint] (2020). https://arxiv.org/abs/2007.12778 (Accessed 20 August 2021).
32. L. Gyorfi, H. Walk, "Nearest neighbor based conformal prediction" (Rep. Stuttgarter Mathematische Berichte 2020-002, Universität Stuttgart, Stuttgart, Germany, 2020).
33. M. Sesia, Y. Romano, Conformal histogram regression. arXiv [Preprint] (2021). https://arxiv.org/abs/2105.08747 (Accessed 20 August 2021).

34. W. Chen, Z. Wang, W. Ha, R. F. Barber, Trimmed conformal prediction for high-dimensional models. arXiv [Preprint] (2016). https://arxiv.org/abs/1611.09933 (Accessed 20 August 2021).
35. V. Chernozhukov, I. Fernandez-Val, A. Galichon, Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* **96**, 559–575 (2009).
36. V. Chernozhukov, I. Fernández-Val, A. Galichon, Quantile and probability curves without crossing. *Econometrica* **78**, 1093–1125 (2010).
37. W. Hoeffding, The large-sample power of tests based on permutations of observations. *Ann. Math. Stat.* **23**, 169–192 (1952).
38. E. J. Candès, L. Lei, Z. Ren, Conformalized survival analysis. arXiv [Preprint] (2021). https://arxiv.org/abs/2103.09763 (Accessed 20 August 2021).
39. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2021).
40. G. Elliott, A. Timmermann, *Economic Forecasting* (Princeton University Press, 2016).
41. R. K. French, Kenneth French Data Library. Fama/French 3 Factors [Daily] Data (2021). http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. Accessed 17 August 2021.
42. F. Boussama, "Ergodicité, mélange et estimation dans les modeles GARCH," PhD thesis, Paris 7, Paris, France (1998).
43. M. Carrasco, X. Chen, Mixing and moment properties of various GARCH and stochastic volatility models. *Econom. Theory* **18**, 17–39 (2002).
44. C. Francq, J. M. Zakoïan, Mixing properties of a general class of GARCH (1,1) models without moment assumptions on the observed process. *Econom. Theory* **22**, 815–834 (2006).
45. V. Chernozhukov, C. Hansen, M. Spindler, hdm: High-dimensional metrics. *R J.* **8**, 185–199 (2016).