Distributional conformal prediction*

Victor Chernozhukov† Kaspar Wüthrich‡ Yinchu Zhu§

September 17, 2019

Abstract

We propose a robust method for constructing conditionally valid prediction intervals based on regression models for conditional distributions such as quantile and distribution regression. Our approach exploits the probability integral transform and relies on permuting estimated "ranks" instead of regression residuals. Unlike residuals, these ranks are independent of the covariates, which allows us to establish the conditional validity of the resulting prediction intervals under consistent estimation of the conditional distributions. We also establish theoretical performance guarantees under arbitrary model misspecification. The usefulness of the proposed method is illustrated based on two applications. First, we study the problem of predicting daily returns using realized volatility. Second, we consider a synthetic control setting where the goal is to predict a country's counterfactual GDP growth rate based on the contemporaneous GDP growth rates of other countries.

*We thank Allan Timmermann for valuable comments. Victor Chernozhukov gratefully acknowledges funding by the National Science Foundation. The usual disclaimer applies.

†Department of Economics and Center for Statistics and Data Science, Massachusetts Institute of Technology, email: vchern@mit.edu

‡Department of Economics, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA, email: kwuthrich@ucsd.edu

§Lundquist College of Business, University of Oregon, email: yzhu6@uoregon.edu

1 Introduction

In this paper, we develop an approach for constructing prediction intervals based on learning conditional distributions. The proposed framework is generic and can be implemented using a great variety of flexible and powerful methods, including quantile regression (QR) (e.g., Koenker and Bassett, 1978), distribution regression (DR) (e.g., Foresi and Peracchi, 1995; Chernozhukov et al., 2013), as well as high-dimensional methods such as quantile neural networks (e.g., Taylor, 2000) and quantile trees and random forests (e.g., Chaudhuri and Loh, 2002; Meinshausen, 2006).

We observe data $\{(Y_t, X_t)\}_{t=1}^T$, where Y_t is the continuous outcome of interest and X_t is a $p \times 1$ vector of predictors. This paper proposes a method for constructing prediction intervals for Y_{T+1} given knowledge of X_{T+1}. Our setting is very general and encompasses many different prediction problems, including cross-sectional prediction (where X_t is a vector of covariates), k-step-ahead forecasts (where X_t contains lagged predictors), autoregressions (where X_t contains lags of Y_t), and synthetic control and counterfactual prediction problems (where X_t contains control outcomes).

For a given miscoverage level $\alpha \in (0, 1)$, our goal is to construct a prediction interval for Y_{T+1} given X_{T+1}, with the property that

$$P\left(Y_{T+1} \in \mathcal{C}_{(1-\alpha)}\left(X_{T+1}\right) \mid X_{T+1}\right) \geq 1 - \alpha. \tag{1}$$

We refer to a prediction set satisfying property (1) as *conditionally* valid. Conditional validity is stronger than *unconditional validity*:

$$P\left(Y_{T+1} \in \mathcal{C}_{(1-\alpha)}\left(X_{T+1}\right)\right) \geq 1 - \alpha. \tag{2}$$

Unconditional validity only requires the prediction set to exhibit correct coverage on average as opposed to requiring accurate coverage for every value of $X_{T+1} = x$. In fact, unconditional validity is a weak requirement and can be achieved by only using $\{Y_t\}_{t=1}^T$ without any covariates.

Since X_{T+1} is known, the more relevant requirement is to have validity for this given value of X_{T+1}, rather than for other (perhaps average) values of X_{T+1}. For example, consider the problem of predicting stock returns given the realized volatility. Since the distribution is more dispersed when the variance is higher, a natural prediction algorithm should give a wider prediction interval for higher values of volatility. Many of the existing conformal prediction methods and prediction methods more generally are based on a model for the conditional mean and thus fail to have this basic property, despite being unconditionally

valid. In this paper, we therefore propose a prediction method that is based on modeling the entire conditional distribution (rather than the conditional mean) and thereby generates conditionally valid prediction sets that fully utilize the information in the covariates.

As we shall see in our empirical application on predicting daily stock returns, the usual 90%-prediction set based on modeling the mean has unconditional coverage of 90%, but the coverage probability can drop to 50% when the volatility is high. In contrast, the proposed method provides a coverage probability close to 90% for all values of volatility. 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.

To motivate the our approach, suppose that we have a model $q(\tau, x)$ for the conditional τ-quantile of Y_t given $X_t = x$, $Q_{Y_t|X_t}(\tau \mid x)$. Under correct specification, $q(\tau, x) = Q_{Y_t|X_t}(\tau \mid x)$ for all (τ, x) and a conditionally valid prediction interval is given by

$$\left[q\left(\frac{\alpha}{2}, x\right), q\left(1 - \frac{\alpha}{2}, x\right)\right]. \tag{3}$$

To implement the prediction interval (3), a plug-in approach would replace $q(\tau, x)$ with a consistent estimator $\hat{q}(\tau, x)$

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

An important drawback of this plug-in approach is the lack of validity under misspecification of $q(\tau, x)$. Misspecification means that $q(\tau, x) \neq Q_{Y_t|X_t}(\tau \mid x)$ for at least some (τ, x) such that the prediction interval (3) will no longer be valid, neither conditionally nor unconditionally.

In this paper, we build upon the idea of conformal prediction and introduce the conditional ranking as a natural choice of conformity score. The proposal is conditionally valid under correct specification, while the construction of the procedure as a permutation test guarantees unconditional validity under misspecification.[1] Let $F_{Y_t|X_t}(y \mid x)$ denote the conditional cumulative distribution function (CDF) of Y_t given $X_t = x$. Our method is based on the probability integral transform, which states that the conditional "rank"

$$U_t := F_{Y_t|X_t}(Y_t \mid X_t) \quad \text{has the uniform distribution on } (0, 1) \text{ and is independent of } X_t.$$

We construct prediction intervals via test inversion. To decide whether a $Y_{T+1} = y$ should be included in $\mathcal{C}_{(1-\alpha)}(X_{T+1})$, we test the null hypothesis that $Y_{T+1} = y$. Let $f(y, x)$ denote

[1] Conformal prediction was invented by Vovk et al. (2005). More recent work includes: Vovk et al. (2009), Lei et al. (2013), Lei and Wasserman (2014) Lei et al. (2018), Chernozhukov et al. (2018b), and Romano et al. (2019). On a broader level, conformal inference is related to the literature on permutation tests (e.g., Romano, 1990; Lehmann and Romano, 2005).

a model for $F_{Y_t|X_t}(y \mid x)$. If the model is correctly specified (i.e., if $f(y,x) = F_{Y_t|X_t}(y \mid x)$ for all (y,x)), the probability integral transform implies that $U_t := f(Y_t, X_t)$ is uniformly distributed and is independent of X_t. Therefore, $f(Y_{T+1}, X_{T+1})$ belongs to $[\alpha/2, 1-\alpha/2]$ with probability $1-\alpha$, conditional on X_{T+1}. As a result, collecting y satisfying $f(y, X_{T+1}) \in [\alpha/2, 1-\alpha/2]$ yields a conditionally valid prediction interval in the sense of (1). We operationalize this idea by proposing a conformal inference procedure which is based on the estimated ranks, $\hat{U}_t := \hat{f}(Y_t, X_t)$, where $\hat{f}(y,x)$ is a suitable estimator of $f(y,x)$. We follow the practice in conformal prediction and use the entire sample $\{(Y_t, X_t)\}_{t=1}^{T+1}$ with $Y_{T+1} = y$ to compute the estimate $\hat{f}(\cdot,\cdot)$. Doing so allows us to obtain a baseline unconditional validity and thus guards against potential model misspecification. If one does not impose the null hypothesis that $Y_{T+1} = y$ for estimation, the resulting procedures might not be valid even in the ideal case of i.i.d. data.

We establish two different theoretical performance guarantees for our procedure.

(i) Conditional asymptotic validity under consistent estimation of conditional distribution[2]
(ii) Unconditional validity under arbitrary model misspecification: Finite-sample validity under i.i.d. (or exchangeable) data as well as asymptotic validity under dependent data

We illustrate the usefulness of our method with two applications. First, we study the problem of predicting daily returns using realized volatility. Second, we consider a synthetic control setting where the goal is to predict a country's counterfactual GDP growth rate based on the contemporaneous GDP growth rates of other countries. In both applications, the proposed distributional conformal inference approach exhibits better conditional coverage properties than regression-based conformal inference approaches.

1.1 Motivating example

We now illustrate the nature of conditional and unconditional validity based on a simple example:

$$Y_t = X_t + X_t \varepsilon_t, \quad X_t \overset{iid}{\sim} \text{Unif}(0,1), \quad \varepsilon_t \overset{iid}{\sim} N(0,1). \tag{5}$$

For simplicity, we focus on the population conformal prediction problem under correct specification and abstract from finite sample issues. Classical conformal prediction methods (e.g., Lei et al., 2018) are based on

$$\tilde{U}_t = Y_t - m(X_t),$$

[2]It is theoretically impossible to obtain meaningful distribution-free prediction intervals with conditional validity (e.g., Lei and Wasserman, 2014); see also Foygel Barber et al. (2019) for further discussions.

where $m(x)$ is a model for the conditional expectation of Y_t given $X_t = x$, $E(Y_t \mid x)$. Under correct specification (if $m(x) = E(Y_t \mid x) = x$), $\tilde{U}_t = X_t \varepsilon_t$. Then, the null hypothesis $Y_{T+1} = y$ is rejected if $|\tilde{U}_{T+1}|$ is larger than the $1-\alpha$ quantile of the distribution of $|\tilde{U}_{T+1}|$, $Q_{|\tilde{U}|}(1-\alpha)$. The resulting prediction interval is

$$\mathcal{C}^{\text{reg}}_{(1-\alpha)}(x) = \left[x - Q_{|\tilde{U}|}(1-\alpha), x + Q_{|\tilde{U}|}(1-\alpha)\right] \tag{6}$$

The key property of the prediction interval $\mathcal{C}^{\text{reg}}_{(1-\alpha)}$ is that its length, $2 \cdot Q_{|\tilde{U}|}(1-\alpha)$, is fixed and does not depend on the value of the predictor $X_{T+1} = x$. This feature implies that the band is not adaptive to the heteroscedasticity present in the location-scale model (5) and thus is not conditionally valid.

By contrast, under correct specification, our approach is based on the ranks $U_t = \Phi(\varepsilon_t)$. The null that $Y_{T+1} = y$ is rejected if $|U_{T+1} - 0.5|$ is larger than the $(1-\alpha)$ quantile of $|U_t - 0.5|$, which is equivalent to $|\varepsilon_{T+1}|$ being larger than the $(1-\alpha)$ quantile of $|\varepsilon_t|$, $Q_{|\varepsilon|}(1-\alpha)$. The resulting prediction interval is

$$\begin{aligned} \mathcal{C}_{(1-\alpha)}(x) &= \left[Q_{Y_t|X_t}\left(\frac{\alpha}{2} \mid x\right), Q_{Y_t|X_t}\left(1-\frac{\alpha}{2} \mid x\right)\right] \\ &= \left[x - x \cdot Q_{|\varepsilon|}(1-\alpha), x + x \cdot Q_{|\varepsilon|}(1-\alpha)\right] \end{aligned} \tag{7}$$

Unlike $\mathcal{C}^{\text{reg}}_{(1-\alpha)}$, the length of the prediction interval $\mathcal{C}_{(1-\alpha)}$ is $2x \cdot Q_{|\varepsilon|}(1-\alpha)$ and depends on $X_{T+1} = x$. Our construction is fully adaptive to the heteroscedasticity in model (5) and conditionally valid.

Figure 1 illustrates the difference between the classical prediction interval and our approach. Panel A plots the conditional length when $(1-\alpha) = 0.9$. The length of $\mathcal{C}^{\text{reg}}_{(0.9)}$ is constant, whereas the length of $\mathcal{C}_{(0.9)}$ varies as a function of the predictor value $X_{T+1} = x$. $\mathcal{C}_{(0.9)}$ is shorter than $\mathcal{C}^{\text{reg}}_{(0.9)}$ for low values of x and longer $\mathcal{C}^{\text{reg}}_{(0.9)}$ for high values. Panel B shows the conditional coverage. The prediction interval $\mathcal{C}_{(0.9)}$ is valid for all x, whereas the $\mathcal{C}^{\text{reg}}_{(0.9)}$ overcovers for low values and undercovers for high values. Taken together, Panels A and B illustrate the advantage of our method. For predictor values where the conditional variance is low, it yields shorter confidence intervals, while ensuring conditional coverage for predictor values where the conditional dispersion is large by suitably enlarging the prediction interval.

1.2 Related literature

We build on and contribute to the literature on conformal prediction and permutation tests as well as the literature on quantile prediction methods (see e.g., Komunjer, 2013, for a

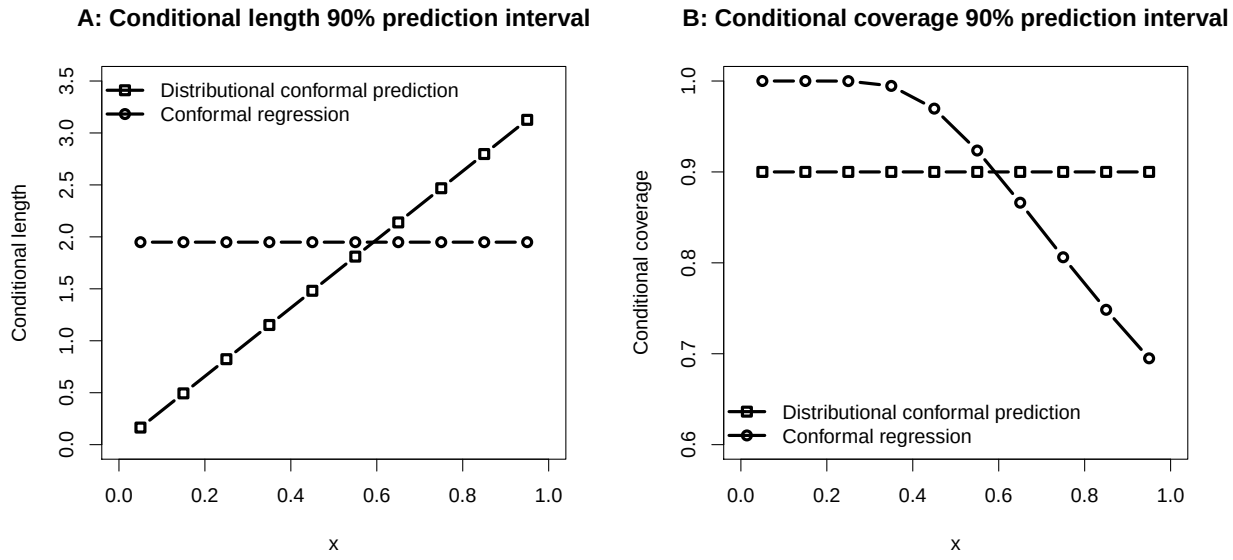

Figure 1: Motivating example

review). Our paper is most closely related to Lei and Wasserman (2014), Lei et al. (2018), and in particular Romano et al. (2019). Lei and Wasserman (2014) propose conditionally valid and asymptotically efficient conformal prediction intervals based on estimators of the conditional density. Here we take a different and complementary approach, which allows researchers to leverage powerful regression methods for conditional CDFs, including quantile and distribution regression. Lei et al. (2018) develop conformal prediction methods based on regression models for conditional expectations which are based on permuting regression residuals. A key feature and disadvantage of this method is that it yields prediction intervals whose width is roughly constant for all values of the predictors. This property is desirable in linear models with homoscedasticity but can lead to poor conditional coverage properties under heteroscedasticity. To make the prediction method adaptive to heteroscedasticity, Lei et al. (2018) 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 the mean regression-based conformal inference, but is motivated by and based on restrictive locations-scale-type models. By contrast, our approach is generic and exploits flexible and substantially more general models for the whole conditional distribution. For example, linear QR nests location-scale models as a special case.

Romano et al. (2019) propose a split conformal approach based on QR models. Their approach is based splitting the data into two subsets, $\mathcal{I}_1$ and $\mathcal{I}_2$. Based on $\mathcal{I}_1$, they estimate two separate quantile functions $\hat{q}_{\alpha/2}(x)$ and $\hat{q}_{1-\alpha/2}(x)$ and construct the prediction intervals

as

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

where $Q(1-\alpha)$ is the $(1-\alpha)(1+1/|\mathcal{I}_2|)$-th empirical quantile of $\max\left\{\hat{q}_{\alpha/2}(X_t) - Y_t, Y_t - \hat{q}_{1-\alpha/2}(x)\right\}$ in $\mathcal{I}_2$. Constructing prediction intervals based on deviations of quantile estimates is similar in spirit to working with deviations from of mean estimates as deviations are measured in absolute levels. By contrast, exploiting the probability integral transform, our approach is generic and relies on permuting deviations in ranks, which naturally have the same scaling on $(0, 1)$.

2 A distributional conformal prediction method

Here we introduce a generic conformal prediction approach for constructing prediction intervals. In Section 3, we show that the prediction intervals are conditionally valid under correct specification and unconditionally valid under arbitrary misspecification.

2.1 Method and algorithm

The prediction intervals are constructed by inverting a test. Let y be denote a test value for Y_{T+1}. We test plausibility of each value $y \in \mathbb{R}$, collect all plausible values, and report them as the confidence set. In practice, we consider a grid of candidate values $y \in \{y^1, \ldots, y^H\} =: \mathcal{Y}_{\text{trial}}$. Define the augmented data set as $Z^{(y)} = \{Z_t\}_{t=1}^{T+1}$, where

$$Z_t = \begin{cases} (Y_t, X_t) & \text{if } 1 \leq t \leq T \\ (y, X_t) & \text{if } t = T+1 \end{cases} \tag{8}$$

Based on the augmented dataset $Z^{(y)}$ estimate the conditional CDF, $F_{Y_t|X_t}(y \mid x)$, using a suitable model $f(y, x)$; see Section 2.2 for examples. Denote the estimator by $\hat{f}(y, x)$. We then compute the ranks $\{\hat{U}_t\}_{t=1}^{T+1}$, where

$$\hat{U}_t = \begin{cases} \hat{f}(Y_t, X_t) & \text{if } 1 \leq t \leq T \\ \hat{f}(y, X_t) & \text{if } t = T+1 \end{cases} \tag{9}$$

P-values are computed as

$$\hat{p}(y) = \frac{1}{T+1} \sum_{t=1}^{T+1} 1\left\{|\hat{U}_t - 0.5| \geq |\hat{U}_{T+1} - 0.5|\right\}. \tag{10}$$

The distributional conformal prediction approach is summarized in Algorithm 1.

Algorithm 1 (Distributional conformal prediction)**.**

Input: Data $\{(Y_t, X_t)\}_{t=1}^T$, miscoverage level $\alpha \in (0,1)$, point X_{T+1} at which to construct the prediction interval, trial values $\mathcal{Y}_{\text{trial}}$

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

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

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

2.2 Regression models for conditional distributions

Here we discuss popular regression approaches for modeling and estimating the conditional CDF. We focus on semiparametric (and potentially penalized) quantile and distribution regression models because of their relevance in applied research, but we emphasize that our method is generic and also works in conjunction with nonparametric quantile and distribution regression estimators (e.g., Chaudhuri, 1991; Koenker et al., 1994; He et al., 1998) as well as methods based on trees and random forests (e.g., Chaudhuri and Loh, 2002; Meinshausen, 2006) and neural networks (e.g., Taylor, 2000).

2.2.1 Quantile regression models

QR models postulate models for the conditional quantile of Y_t given X_t:

$$P(Y_t \leq q(\tau, X_t) \mid X_t) = \tau \tag{11}$$

These models imply models for the conditional distribution, $F_{Y_t|X_t}$, through the relation

$$F_{Y_t|X_t}(y \mid x) = \int_{(0,1)} 1\left\{Q_{Y_t|X_t}(\tau \mid x) \leq y\right\} d\tau. \tag{12}$$

Consider the leading example when the conditional quantile function is assumed to be linear:

$$q(\tau, x) = x'\beta(\tau), \tag{13}$$

If X_t is low dimensional, the parameter of interest $\beta(\tau)$ can be estimated using linear QR (Koenker and Bassett, 1978) as the solution to a convex program

$$\hat{\beta}(\tau) \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \rho_\tau\left(Y_t - X_t'\beta\right), \tag{14}$$

where $\rho_\tau(u) := u(\tau - 1\{u < 0\})$ is the "check function". In high-dimensional problems, it may be convenient to augment the program (14) with an appropriate penalization $\mathcal{P}(\beta)$:

$$\hat{\beta}(\tau) \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} \rho_\tau \left(Y_t - X_t'\beta\right) + \mathcal{P}(\beta) \tag{15}$$

Examples of $\mathcal{P}(\beta)$ include ℓ_1-penalties (e.g., Koenker, 2004; Li and Zhu, 2008; Belloni and Chernozhukov, 2011) and SCAD (e.g., Wu and Liu, 2009). Based on estimators of $\{\hat{\beta}(\tau)\}$, the conditional distribution can be estimated as

$$\hat{f}(y, x) = \epsilon + \int_{\epsilon}^{1-\epsilon} 1\left\{x'\hat{\beta}(\tau) \leq y\right\} d\tau,$$

for some small constant $\epsilon > 0$. The trimming by ϵ avoids the estimation of tail quantiles.

2.2.2 Distribution regression models

Instead of modeling the quantile function, one can directly model and estimate the conditional CDF using DR (e.g., Foresi and Peracchi, 1995; Chernozhukov et al., 2013, 2019). DR methods impose a partially linear model for the CDF:

$$F_{Y_t|X_t}(y \mid x) = \Lambda\left(x'\beta(y)\right),$$

where $\beta(y)$ is the parameter of interest which may depend on y and Λ is a known link function, for example, the Probit or Logit link.

If X_t is low dimensional, the parameters $\beta(y)$ can be estimated as

$$\hat{\beta}(y) \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} -\left[1\left\{Y_t \leq y\right\} \log\left(\Lambda\left(X_t'\beta(y)\right)\right) + 1\left\{Y_t > y\right\} \log\left(1 - \Lambda\left(X_t'\beta(y)\right)\right)\right] \tag{16}$$

When Λ is the Probit (Logit) link, this is nothing else than a Probit (Logit) regression of the indicator $1\{Y_t \leq y\}$ on X_t. In high dimensional settings, one can add a penalty $\mathcal{P}(\beta)$ to the problem (16) (e.g., an ℓ_1-penalty or an elastic net penality):

$$\hat{\beta}(y) \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{t=1}^{T} -\left[1\left\{Y_t \leq y\right\} \log\left(\Lambda\left(X_t'\beta(y)\right)\right) + 1\left\{Y_t > y\right\} \log\left(1 - \Lambda\left(X_t'\beta(y)\right)\right)\right] + \mathcal{P}(\beta) \tag{17}$$

Given an estimator $\{\hat{\beta}(y)\}$, the conditional distribution can be estimated as

$$\hat{f}(y, x) = \Lambda\left(x'\hat{\beta}(y)\right).$$

The finite sample properties of the resulting estimator $\hat{f}$ can be improved using rearrangement (Chernozhukov et al., 2009).

3 Theoretical performance guarantees

In settings where the data are i.i.d. (or exchangeable), our method achieves finite-sample unconditional validity in a model-free manner.

Theorem 1 (Finite sample unconditional validity)**.** *Suppose that the data are i.i.d. or exchangeable and that the estimator $\hat{f}$ is invariant to permutations of the data. Then*

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

Theorem 1 is built upon classical arguments and 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 under the very strong i.i.d assumption, it provides no guarantee at all on the conditional validity.

Our next theoretical results provide a remedy. Recall that we compute $\hat{U}_t = \hat{f}(Y_t; X_t)$ and then permute $\hat{V}_t = \psi(\hat{U}_t)$, where $\psi(\cdot)$ is a deterministic function ($\psi(x) = |x - 1/2|$ in Section 2.1). Let f denote the probability limit of $\hat{f}$ and define $\xi_t = f(Y_t; X_t)$ as well as $V_t = \psi(\xi_t)$. The procedure is based on the permutation p-value, which can be written as $\hat{p} = \hat{p}(Y_{T+1}) = 1 - \hat{G}(\hat{V}_{T+1})$ with $\hat{G}(v) = (T+1)^{-1} \sum_{t=1}^{T+1} \mathbf{1}\{\hat{V}_t < v\}$. Our theoretical analysis of conditional validity proceeds as follows. We show that the permutation p-value $\hat{p}(Y_{T+1})$ is approximately the ranking of V_{T+1}. Since this ranking always has the uniform distribution on (0,1) unconditionally, the question of conditional validity becomes whether or not V_{T+1} is independent of X_{T+1}. If f is equal to $F_{Y_t|X_t}$ (or its transformation), then V_{T+1} is independent of X_{T+1} and the procedure is conditionally valid; otherwise, conditional validity will not hold in general. We impose the following weak regularity conditions.

Assumption 1. *Suppose that the following conditions hold.*

1. *There exists a strictly increasing continuous function $\phi : \mathbb{R} \to [0, \infty)$ such that $\phi(0) = 0$ and $(T+1)^{-1} \sum_{t=1}^{T+1} \phi(\hat{V}_t - V_t) = o_P(1)$ and $\hat{V}_{T+1} = V_{T+1} + o_P(1)$.*
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.*

The first part of Assumption 1 is very weak as it merely says that $\hat{f}$ is consistent to its probability limit under a very "weak" norm and no rate condition is required. When

$\psi(x) = |x - 1/2|$ as in Section 2.1, a sufficient condition is $(T+1)^{-1}\sum_{t=1}^{T+1}|\hat{f}(Y_t, X_t) - f(Y_t, X_t)|^q = o_P(1)$ for some $q > 0$, which can be verified for many existing estimators. Arguably the most important conceptual advantage is that Assumption 1 allows $f \neq F_{Y_t|X_t}$, i.e., $\xi_t \neq U_t$. A leading example is when the model is misspecified such that $\hat{f}$ is not consistent for $F_{Y_t|X_t}$. Since we are only requiring $\hat{f}$ to converge to its probability limit f, the analysis of conditional validity reduces to the interpretation of this probability limit. The rest of Assumption 1 states that the empirical distribution of the target variable V_t converges to its population distribution function, which is Lipschitz. Under $V_t = |\xi_t - 1/2|$, a sufficient condition is that ξ_t is ergodic and has a bounded density. Ergodicity is much weaker than the i.i.d (or exchangeability) assumption and can still be verified for a large class of stochastic processes. The following lemma gives the basic consistency result.

Lemma 1. *Let Assumption 1 hold. Then*

$$\hat{G}(\hat{V}_{T+1}) = G(V_{T+1}) + o_P(1).$$

Without additional assumptions, we only know that $G(\cdot)$ is the distribution function of V_{T+1}, which means that $G(V_{T+1})$ has the uniform distribution on $(0,1)$, i.e., $P(G(V_{T+1}) \leq \alpha) = \alpha$. This leads to asymptotic unconditional validity.

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

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

Theorem 2 establishes the unconditional validity of the procedure under very weak conditions. This result allows for general forms of dependent data and does not require $\hat{f}$ to be a consistent estimator of $F_{Y_t|X_t}$. This is a reassuring back-up since one might not always be able to guarantee accurate estimation of the true conditional distribution. Although both Theorems 1 and 2 provide unconditional guarantee, the former assumes i.i.d (or exchangeable) data, whereas the latter only assumes ergodicity and consistency of $\hat{f}$ to its probability limit f and thus allows for dependent data. Another implication of Lemma 1 is that if f is $F_{Y_t|X_t}$, then the prediction intervals are conditionally valid.

Theorem 3 (Asymptotic conditional validity)**.** *Let Assumption 1 hold with $V_t = \psi(U_t)$, where $U_t = F_{Y_t|X_t}(Y_t \mid X_t)$. Then*

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

Theorems 2 and 3 establish the 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.

4 Extensions

4.1 Split distributional conformal prediction

The possible drawback of Algorithm 1 is the computational burden. For example, when $f(y,x)$ is a DR model, estimating $f(y,x)$ requires estimating the whole DR process, which involves estimating many (potentially penalized) Logit or Probit regressions. Since $f(y,x)$ is estimated under the null hypothesis based on the augmented data, one has to estimate the whole distribution regression process for all test values $y \in \mathcal{Y}_{\text{trial}}$. To avoid this we propose a split conformal procedure that exploits sample splitting and only requires estimating the conditional distribution once. Sample splitting is a popular approach to alleviate the computational burden in the conformal prediction literature (e.g., Lei et al., 2018; Romano et al., 2019). Under the conditions in Section 3, Algorithm 2 yields theoretically valid prediction intervals.

Algorithm 2 (Split distributional conformal prediction)**.**

Input: Data $\{(Y_t, X_t)\}_{t=1}^{T}$, miscoverage level $\alpha \in (0,1)$, point X_{T+1} at which to construct the prediction interval, trial values $\mathcal{Y}_{\text{trial}}$

Process:

1. Split $\{1,\dots,T+1\}$ into $\mathcal{I}_1 := \{1,\dots,T_0\}$ and $\mathcal{I}_2 := \{T_0+1,\dots,T+1\}$
2. Obtain an estimator of the conditional CDF, $\hat{f}(y,x)$, based on $\{Z_t\}_{t\in\mathcal{I}_1}$
3. For $y \in \mathcal{Y}_{\text{trial}}$,
   (a) define the augmented data $Z^{(y)} = \{Z_t\}_{t\in\mathcal{I}_2}$
   (b) compute $\hat{p}(y)$ as in (10), where $\hat{U}_t$ are obtained using $\hat{f}(y,x)$

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

4.2 In-sample validation

The proposed procedure lends itself to a simple in-sample validity check based on the following Algorithm 3. The output of Algorithm 3 can be used to assess the conditional coverage properties as illustrated in the empirical applications of Section 5. For example, one can compute the empirical coverage rate conditional on different values of X_t or assess independence of $p(X_t)$ and X_t.

Algorithm 3 (In-sample validation)**.**

Input: Data $\{(Y_t, X_t)\}_{t=1}^T$

Process:

1. Obtain an estimator of the conditional CDF, $\hat{f}(y, x)$, based on $\{(Y_t, X_t)\}_{t=1}^T$.
2. For $\tilde{t} \in \{1, \ldots, T\}$, compute

$$\hat{p}(Y_{\tilde{t}}) = \frac{1}{T} \sum_{t=1}^{T} 1\left\{|\hat{U}_t - 0.5| \geq |\hat{U}_{\tilde{t}} - 0.5|\right\}.$$

Output: P-values $\hat{p}\left(Y_1\right), \ldots, \hat{p}\left(Y_T\right)$

5 Applications

5.1 Predicting market returns

We apply our method to the problem of predicting stock market returns, which are known to exhibit substantial heteroscedasticity. We use daily returns on the market portfolio (CRSP value-weighted portfolio) from July 1, 1926 to June 30, 2019.[3] We use lagged realized volatility X_t to predict the present return Y_t.[4] We compare three different methods: QR, DR, and classical conformal prediction based on OLS.

To examine the conditional coverage properties, we apply the in-sample validation Algorithm 3. Table 1 shows the correlation between $\hat{p}(Y_t)$ on X_t. Conditional coverage implies that p-values are independent of X_t and thus a true correlation of zero. QR yields a smaller correlation than DR and both distributional conformal prediction approaches yields much lower correlations than classical conformal prediction based on OLS. The negative correlation highlights the undercoverage in high volatility states. When X_t is high, the p-value tends to be low, indicating a lower probability of covering the true value.

Table 1: $\text{Corr}(\hat{p}\left(Y_t\right), X_t)$

QR	DR	OLS
0.00	-0.09	-0.28

[3]The CRSP data is constructed from the Fama/French 3 Factors available at this link.

[4]We compute realized volatility as the square root of the sum of squared returns over the last 22 days.

In Figure 2, we plot the empirical coverage probabilities for 40 bins obtained by dividing up the support of X_t based on equally spaced quantiles. QR exhibits the best overall performance. It yields prediction intervals with coverage levels that are almost constant across all bins and close to the nominal level. QR outperforms DR, which overcovers in low volatility regimes and undercovers in high volatility regimes. A comparison to classical conformal prediction based on OLS clearly demonstrates the advantage of our method. OLS performs much worse and exhibits substantial undercoverage in high-volatility regimes, where conditional performance guarantees are arguably most important.

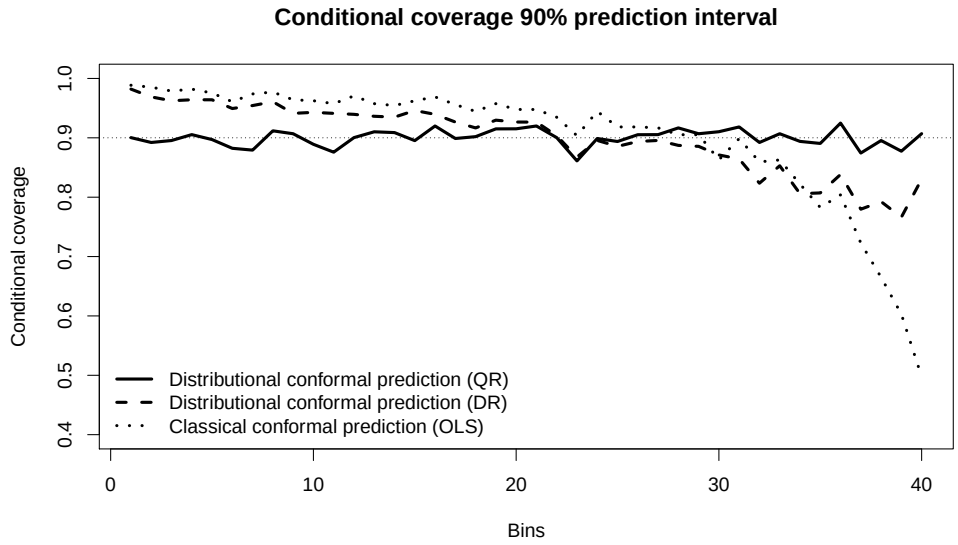

Figure 2: Conditional coverage properties based on daily returns data

5.2 Synthetic control and counterfactual prediction

Here we illustrate our approach in a panel data setting where the goal is test hypotheses about the causal effects of a policy intervention by predicting the counterfactual outcome of the treated unit in the absence of the treatment. Such problems are very common in applied research and there are many different approaches for predicting counterfactuals, including difference-in-differences (e.g., Ashenfelter and Card, 1985; Card and Krueger, 1994; Bertrand et al., 2004; Athey and Imbens, 2006), synthetic control (e.g, Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015), as well as factor, matrix completion, and interactive fixed effects models (e.g., Hsiao et al., 2012; Gobillon and Magnac, 2016; Xu, 2017; Athey et al., 2017).

Consider a setting with one treated unit $i = 0$ and N control units $i = 1, \ldots, N$. The treated unit is untreated for the first T periods, and treated in period $T + 1$. Denote by

Y_{it}^I and Y_{it}^N potential outcomes with and without the treatment. Observed outcomes Y_{it} are related to potential outcomes as $Y_{it} = D_{it}Y_{it}^I + (1 - D_{it})Y_{it}^N$, where the treatment indicator is defined as $D_{it} = \mathbf{1}\{t > T, i = 0\}$. The object of interest is the treatment effect for the treated unit in the post treatment period, $Y_{0T+1}^I - Y_{0T+1}^N$. By the definition of potential outcomes

$$Y_{0T+1}^I - Y_{0T+1}^N = Y_{0T+1} - Y_{0T+1}^N$$

Hence, the key object of interest is Y_{0T+1}^N, the counterfactual outcome of the treated unit in the absence of the treatment. Here we illustrate our approach by constructing prediction intervals Y_{0T+1}^N. Tests for sharp null hypotheses about $Y_{0T+1}^I - Y_{0T+1}^N$ can be performed as in Chernozhukov et al. (2018a).

We apply our approach to predict the country-level GDP growth rate Y_{0T+1}^N based on GDP growth rates in the other countries $(Y_{1T+1}^N, \ldots, Y_{NT+1}^N)$ (all of which are observed since none of the controls is being treated). We use data on quarterly GDP growth rates from 1960q2 to 2019q1 for 23 OECD countries obtained from the OECD database (OECD, 2019).[5] Country-level GDP data has been used in synthetic control settings for instance to estimate the effect of the German reunification by Abadie et al. (2015).

Control outcomes are modeled using a factor model as in Hsiao et al. (2012), Gobillon and Magnac (2016), and Xu (2017) among many others:

$$Y_{it}^N = \lambda_i' F_t + \varepsilon_{it}, \quad 1 \leq i \leq N,\ 1 \leq t \leq T + 1,$$

where λ_i are the country-specific loadings and F_t are the time-varying common factors. To illustrate, we assume that the number of factors is equal to one. We obtain estimates of F_t, $\hat{F}_t$, using singular value decomposition. We then use $X_t := \hat{F}_t$ to predict the counterfactual $Y_t := Y_{0t}^N$. We repeat this exercise for all 23 countries, treating all other countries as control units. We compare three different methods: QR, DR, and classical conformal prediction based on OLS.

To investigate the conditional coverage probabilities, we apply the in-sample validation Algorithm 3. Table 2 displays the correlation of the p-values $\hat{p}(Y_t)$ on X_t. Conditional coverage implies that p-values are independent of X_t and thus a zero correlation. Both QR and DR both perform very well, yielding small correlations for most countries. By contrast, classical conformal prediction can exhibit substantial correlations that are larger than for QR and DR in most cases.

Figure 5 plots the empirical coverage probabilities for five bins obtained by dividing up the support of X_t based on equally spaced quantiles. While there are countries for which all

[5] We omit countries for which the data are not available for all years from our analysis.

Table 2: Absolute values of $\text{Corr}(\hat{p}(Y_t), X_t)$

	QR	DR	OLS
KOR	**0.03**	**0.03**	0.15
AUS	**0.02**	0.04	0.20
ISL	**0.00**	0.04	0.15
GBR	**0.00**	0.02	0.20
DNK	**0.02**	0.03	0.04
ESP	0.05	**0.04**	0.19
JPN	**0.01**	0.04	0.10
NOR	0.01	**0.00**	0.03
FIN	**0.02**	0.07	0.14
NLD	**0.00**	0.07	0.19
USA	**0.01**	**0.01**	0.07
MEX	**0.01**	**0.01**	0.04
BEL	**0.02**	0.07	**0.02**
DEU	0.02	**0.01**	0.07
LUX	**0.01**	0.04	0.02
PRT	**0.01**	0.06	0.16
ITA	**0.02**	0.04	0.04
AUT	**0.00**	0.02	0.11
IRL	**0.00**	0.03	0.04
CHE	0.07	**0.02**	0.28
SWE	**0.01**	0.03	0.03
FRA	**0.00**	0.07	0.15
GRC	0.02	**0.00**	0.06

three methods perform well, QR and DR tend to exhibit better coverage properties for low values of X_t, where OLS can exhibit notable undercoverage.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of californias tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *The American Economic Review*, 93(1):113–132.

Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–660.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2017). Matrix completion methods for causal panel data models.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Belloni, A. and Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, pages 82–130.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84(4):772–793.

Chaudhuri, P. (1991). Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of Multivariate Analysis*, 39(2):246 – 269.

Chaudhuri, P. and Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 8(5):561–576.

Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):pp. 2205–2268.

Chernozhukov, V., Fernandez-Val, I., Melly, B., and Wuthrich, K. (2019). Generic inference on quantile and quantile effect functions for discrete outcomes. *Journal of the American Statistical Association*, 0(0):1–24.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2018a). An exact and robust conformal inference method for counterfactual and synthetic controls. arXiv:1712.09089.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2018b). Exact and robust conformal inference methods for predictive machine learning with dependent data. *arXiv:1802.06300*.

Foresi, S. and Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, 90(430):pp. 451–466.

Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *arXiv:1903.04684*.

Gobillon, L. and Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *The Review of Economics and Statistics*, 98(3):535–551.

He, X., Ng, P., and Portnoy, S. (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):537–550.

Hsiao, C., Steve Ching, H., and Ki Wan, S. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 27(5):705–740.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74 – 89. Special Issue on Semiparametric and Nonparametric Mixed Models.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 81(4):673–680.

Komunjer, I. (2013). Chapter 17 - quantile prediction. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2 of *Handbook of Economic Forecasting*, pages 961 – 994. Elsevier.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Science & Business Media.

Lei, J., GSell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287.

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

Li, Y. and Zhu, J. (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185.

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

OECD (2019). Quarterly gdp (indicator) (accessed on 15 august 2019).

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692.

Romano, Y., Patterson, E., and Candes, E. J. (2019). Conformalized quantile regression. arXiv:1905.03222.

Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2009). On-line predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, 19(2):801–817.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

Appendix

A Proofs

Proof of Lemma 1. The argument is similar to Lemma 12 of Chernozhukov et al. (2018a). Let $\delta > 0$ be a constant to be chosen later. Define quantities $R_T = \sup_{v\in\mathbb{R}} |\tilde{G}(v) - G(v)|$ and $W = \sup_{x_1\neq x_2} |G(x_1) - G(x_2)|/|x_1 - x_2|$.

Let $A = \{t : \ |\hat{V}_t - V_t| \geq \delta\}$. Fix $x \in \mathbb{R}$. Then

$$\begin{aligned}
&(T+1)\left|\hat{G}(x) - \tilde{G}(x)\right| \\
&\leq \left|\sum_{t\in A}\left(\mathbf{1}\{\hat{V}_t < x\} - \mathbf{1}\{V_t < x\}\right)\right| + \left|\sum_{t\in A^c}\left(\mathbf{1}\{\hat{V}_t < x\} - \mathbf{1}\{V_t < x\}\right)\right| \\
&\overset{\text{(i)}}{\leq} 2|A| + \left|\left(\sum_{t\in A^c}\mathbf{1}\{\hat{V}_t < x\}\right) - \left(\sum_{t\in A^c}\mathbf{1}\{V_t < x\}\right)\right| \qquad (18)
\end{aligned}$$

where (i) follows by the boundedness of $\mathbf{1}\{\cdot\}$. We notice that for $t \in A^c$, $V_t - \delta < \hat{V}_t < V_t + \delta$. Therefore,

$$\sum_{t\in A^c}\mathbf{1}\{V_t < x - \delta\} \leq \sum_{t\in A^c}\mathbf{1}\{\hat{V}_t < x\} \leq \sum_{t\in A^c}\mathbf{1}\{V_t < x + \delta\}.$$

Since $\sum_{t\in A^c}\mathbf{1}\{V_t \leq x\}$ is also between $\sum_{t\in A^c}\mathbf{1}\{V_t \leq x - \delta\}$ and $\sum_{t\in A^c}\mathbf{1}\{V_t \leq x + \delta\}$, it follows that

$$\begin{aligned}
&\left|\left(\sum_{t\in A^c}\mathbf{1}\{\hat{V}_t < x\}\right) - \left(\sum_{t\in A^c}\mathbf{1}\{V_t < x\}\right)\right| \\
&\leq \left|\left(\sum_{t\in A^c}\mathbf{1}\{V_t \leq x + \delta\}\right) - \left(\sum_{t\in A^c}\mathbf{1}\{V_t \leq x - \delta\}\right)\right| \\
&= \left|(T+1)\left[\tilde{G}(x+\delta) - \tilde{G}(x-\delta)\right] - \left(\sum_{t\in A}\mathbf{1}\{V_t \leq x + \delta\}\right) + \left(\sum_{t\in A}\mathbf{1}\{V_t \leq x - \delta\}\right)\right| \\
&\leq (T+1)\left[\tilde{G}(x+\delta) - \tilde{G}(x-\delta)\right] + \left(\sum_{t\in A}\mathbf{1}\{V_t \leq x + \delta\}\right) + \left(\sum_{t\in A}\mathbf{1}\{V_t \leq x - \delta\}\right) \\
&\leq (T+1)\left(G(x+\delta) - G(x-\delta) + 2R_T\right) + 2|A| \\
&\leq (T+1)\left(2\delta W + 2R_T\right) + 2|A|.
\end{aligned}$$

Combine the above display with (18), we obtain that

$$(T+1)\left|\hat{G}(x) - \tilde{G}(x)\right| \leq 4|A| + (T+1)\left(2\delta W + 2R_T\right).$$

Since the right-hand side does not depend on x, we have that

$$\sup_{x\in\mathbb{R}}\left|\hat{G}(x)-\tilde{G}(x)\right| \leq 4\frac{|A|}{T+1}+2\delta W+2R_T.$$

To bound $|A|$, we notice that

$$|A|\phi(\delta) \leq \sum_{t\in A}\phi(\hat{V}_t-V_t) \leq \sum_{t=1}^{T+1}\phi(\hat{V}_t-V_t) \leq o_P(T+1).$$

Hence, the above two displays imply that

$$\sup_{x\in\mathbb{R}}\left|\hat{G}(x)-G(x)\right| \leq \sup_{x\in\mathbb{R}}\left|\hat{G}(x)-\tilde{G}(x)\right|+R_T \leq o_P(1/\phi(\delta))+2\delta W+3R_T. \tag{19}$$

Now we fix an arbitrary $\eta\in(0,1)$. Choose $\delta=\eta/(6W)$. Since $1/\phi(\delta)$ is a constant and $R_T=o_P(1)$ by assumption, (19) implies that

$$\begin{aligned}&\limsup_{T\to\infty} P\left(\sup_{x\in\mathbb{R}}\left|\hat{G}(x)-\tilde{G}(x)\right|>\eta\right)\\ &\leq \limsup_{T\to\infty} P\left(|o_P(1/\phi(\delta))|>\eta/3\right)+\limsup_{T\to\infty} P\left(|2\delta W|>\eta/3\right)+\limsup_{T\to\infty} P\left(|R_T|>\eta/9\right)=0.\end{aligned}$$

Since $\eta>0$ is arbitrary, we have

$$\sup_{x\in\mathbb{R}}\left|\hat{G}(x)-G(x)\right| = o_P(1).$$

Thus,

$$\hat{G}(\hat{V}_{T+1}) = G(\hat{V}_{T+1})+o_P(1) \overset{\text{(i)}}{=} G(V_{T+1})+o_P(1),$$

where (i) follows by $|G(\hat{V}_{T+1})-G(V_{T+1})|\leq W|\hat{V}_{T+1}-V_{T+1}|$. The proof is complete. □

Proof of Theorem 1. The result follows standard arguments (e.g., Chernozhukov et al., 2018a,b). □

Proof of Theorem 2. The result follows directly from Lemma 1. □

Proof of Theorem 3. Notice that U_t is independent of X_t. Since $V_{T+1}=\psi(U_{T+1})$, V_{T+1} is also independent of X_{T+1}. This means that

$$P(G(V_{T+1})\leq\alpha \mid X_{T+1}) = P(G(V_{T+1})\leq\alpha).$$

Since $G(\cdot)$ is the distribution function of V_{T+1}, we have that $P(G(V_{T+1})\leq\alpha)=\alpha$. The desired result follows by Lemma 1. □

B Monte Carlo study

Here we present simulation evidence comparing our method to classical conformal prediction based on OLS and conformalized QR (CQR) (Romano et al., 2019). We consider four different data generating processes (DGPs). For all DGPs, we generate X_t from a uniform distribution on $[0, 5]$, $X_t \overset{iid}{\sim} \text{Unif}(0, 5)$. For a random variable A, let σ_A denote its standard deviation.

DGP1:

$$Y_t = X_t + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, 1)$$

DGP3:

$$Y_t = X_t + \sigma^{-1}_{X_t\varepsilon_t} X_t \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, 1)$$

DGP3:

$$Y_t = X_t + \sigma^{-1}_{X_t^2\varepsilon_t} X_t^2 \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, 1)$$

DGP4:

$$Y_t \overset{iid}{\sim} \text{Pois}\left(\sin^2(X_t) + 0.1\right) + 0.03 X_i \varepsilon_{1,t} + 25 \cdot 1\left\{U_t < 0.01\right\} \varepsilon_{2,t},$$

where $\text{Pois}(\lambda)$ is the Poission distribution with mean λ, $\varepsilon_{1,t} \overset{iid}{\sim} N(0, 1)$, $\varepsilon_{2,t} \overset{iid}{\sim} N(0, 1)$, and $U_i \overset{iid}{\sim} \text{Unif}(0, 1)$. This DGP is used in Romano et al. (2019, Appendix B).

We consider two different sample sizes $T \in \{100, 500\}$ and compare four different methods:

1. Algorithm 1 based on QR (**QR**)
2. Algorithm 1 based on DR (**DR**)
3. Conformal regression based on OLS (**OLS**)
4. Conformalized QR (Romano et al., 2019, Algorithm 1 based on QR) (**CQR**)

Because the data are i.i.d. all four methods are unconditionally valid in finite samples. We therefore focus on conditional validity. We perform the following exercise. In every simulation iteration, we collect X_{T+1} and record whether or not Y_{T+1} was covered by the prediction interval. We then compute the conditional coverage rate within equally spaced bins on $[0, 5]$ (the support of X_{T+1}).

Figures 3–4 present the results. We find that the proposed distributional conformal prediction approach performs well in terms of conditional coverage under correct specification and, when combined with QR, consistently outperforms classical conformal prediction under misspecification. It is interesting to compare our method based on QR to the CQR approach

of Romano et al. (2019) as both methods rely on the same linear QR model. The two approaches exhibit similar conditional coverage properties under DGP1 and DGP4, but there are notable differences under DGP2 and DGP3. Under DGP2 (where linear QR is correctly specified), distributional conformal prediction yields accurate conditional coverage, while CQR undercovers for low values of X_{T+1}. Under DGP3 (where linear QR is misspecified), while both methods undercover for some values of X_{T+1}, our approach based on QR yields a better worst case performance.

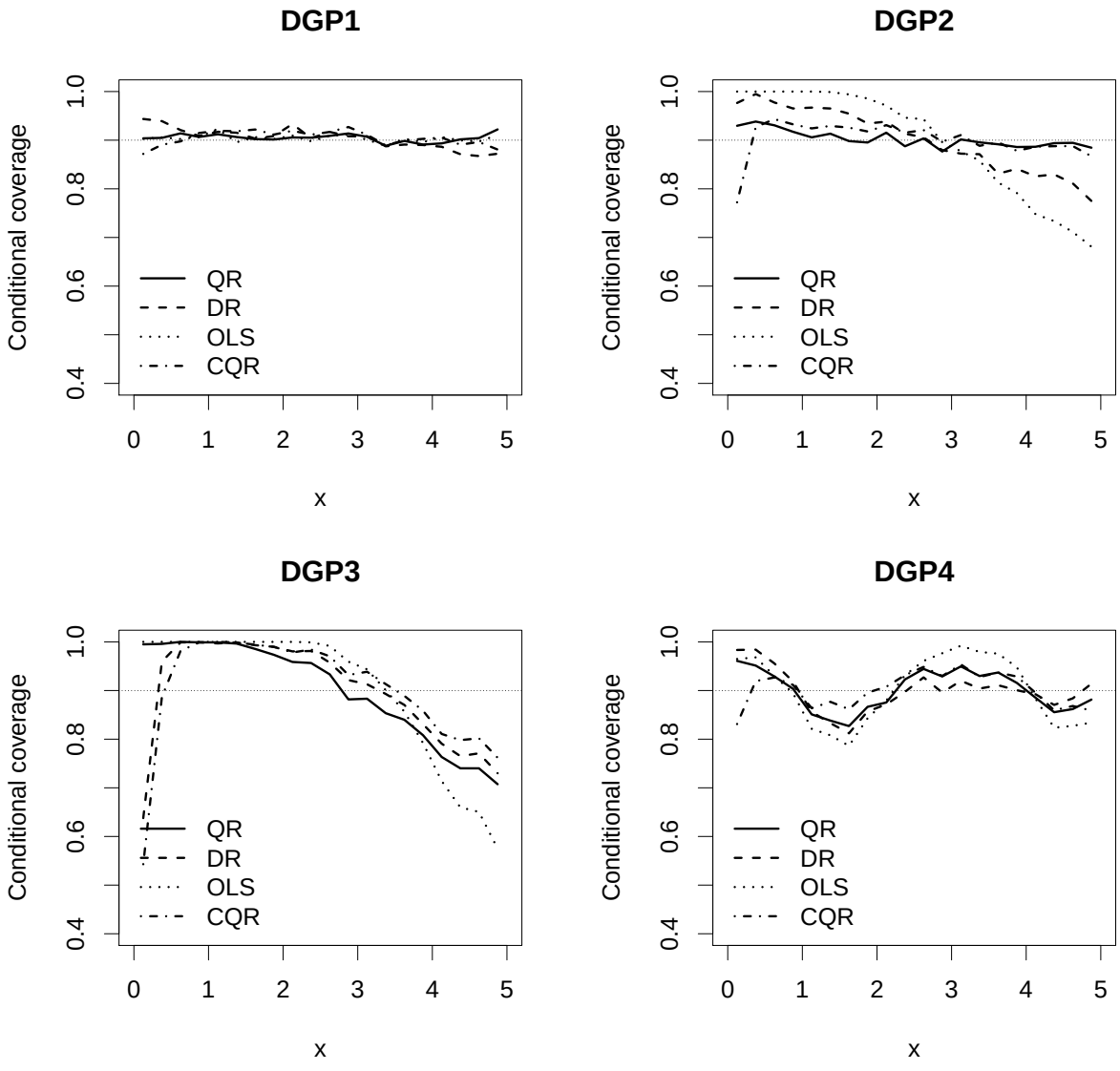

Figure 3: Simulation results for $T = 100$

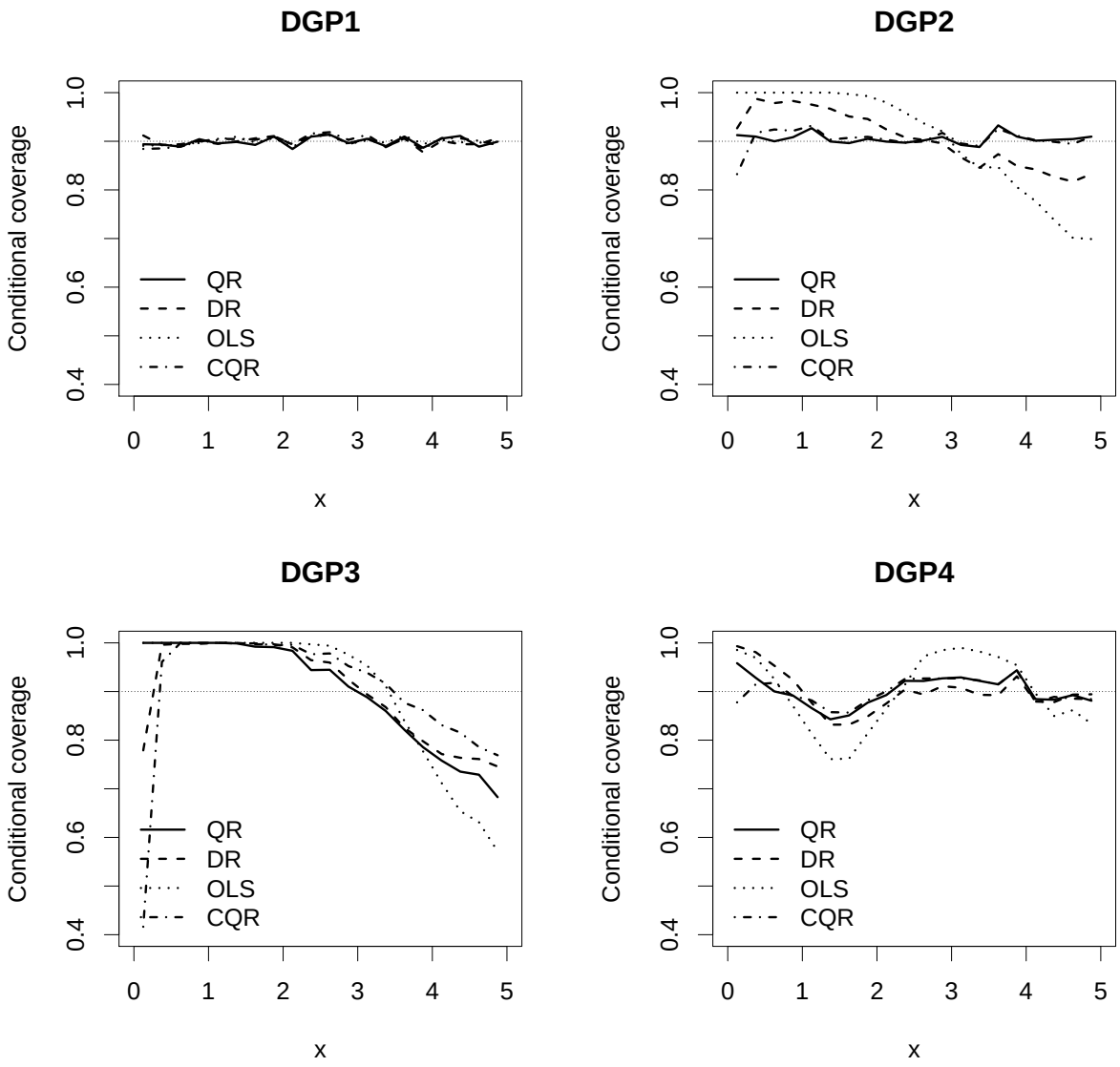

Figure 4: Simulation results for $T = 500$

C Additional figures

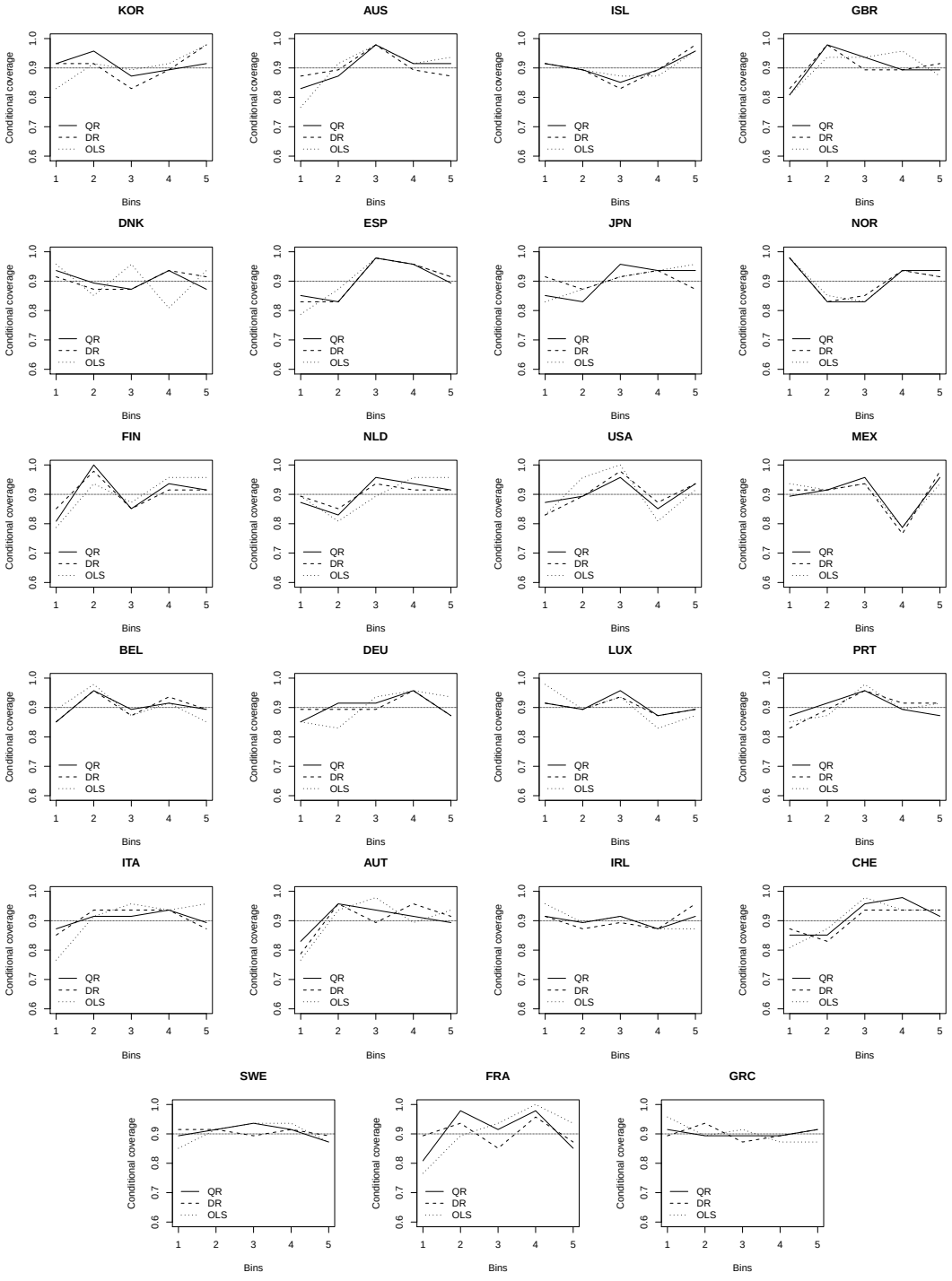

Figure 5: Conditional coverage of 90% prediction intervals by country