CONFORMAL PREDICTIONS FOR LONGITUDINAL DATA

Devesh Batra[1], Salvatore Mercuri[1,⋆], and Raad Khraishi[1,2]

[1]Data Science & Innovation, NatWest Group, London, United Kingdom[†]
[2]Institute of Finance and Technology, UCL, London, United Kingdom[‡]

ABSTRACT

We introduce Longitudinal Predictive Conformal Inference (LPCI), a novel distribution-free conformal prediction algorithm for longitudinal data. Current conformal prediction approaches for time series data predominantly focus on the univariate setting, and thus lack cross-sectional coverage when applied individually to each time series in a longitudinal dataset. The current state-of-the-art for longitudinal data relies on creating infinitely-wide prediction intervals to guarantee both cross-sectional and asymptotic longitudinal coverage. The proposed LPCI method addresses this by ensuring that both longitudinal and cross-sectional coverages are guaranteed without resorting to infinitely wide intervals. In our approach, we model the residual data as a quantile fixed-effects regression problem, constructing prediction intervals with a trained quantile regressor. Our extensive experiments demonstrate that LPCI achieves valid cross-sectional coverage and outperforms existing benchmarks in terms of longitudinal coverage rates. Theoretically, we establish LPCI's asymptotic coverage guarantees for both dimensions, with finite-width intervals. The robust performance of LPCI in generating reliable prediction intervals for longitudinal data underscores its potential for broad applications, including in medicine, finance, and supply chain management.

Keywords Conformal predictions · longitudinal data

1 Introduction

The improvement in predictive performance of machine learning models over the last two decades have made them essential components of decision-making pipelines across high-stake domains such as medicine and finance. However, the point estimates yielded by these predictive models are insufficient in these critical application domains, where uncertainty estimates are of particular interest for informed decision-making (see Harries et al. (1999); Díaz-González et al. (2012); Mears et al. (2015) for uncertainty quantification in these domains). While post-hoc methods such as bootstrapping, jackknife and other ensembling procedures (see Alaa et al. (2020); van der Schaar et al. (2020); Xu & Xie (2020)) are popularly used for uncertainty estimation of a particular statistic (such as model metrics), they are only able to provide theoretical guarantees under additional assumptions on the underlying model and data distribution. This limitation, however, is addressed by the conformal prediction framework, which provides a principled way to perform model-agnostic and distribution-free uncertainty quantification of complex machine learning models.

Conformal predictions are a powerful tool for constructing prediction sets or intervals that provide reliable coverage guarantees for the true value. These guarantees are typically based on the assumption of data exchangeability, which is often violated in time series data due to temporal dependencies and non-stationarity. However, accurate uncertainty quantification is crucial for time series data, which is central to applications ranging from medical diagnosis to energy demand and stock market forecasting. Consequently, there has been a growing interest in developing conformal prediction methods that can handle non-exchangeable data. Most of the focus for time series data has been on univariate

[⋆] No longer at institute[1]
[†] Correspondence to: devesh.batra@natwest.com
[‡] Correspondence to: raad.khraishi@ucl.ac.uk

data, with Xu & Xie (2022) recently developing the Sequential Predictive Conformal Inference (SPCI) method and showing its asymptotic conditional coverage.

In this work, we study conformal predictions for longitudinal data, which consists of a number of uniquely identified time series. This is often the format of time series data seen in applications in medicine and finance, where we might have a time series for each patient, stock, or customer. In this context, predictions can be made longitudinally, forecasting future values for each time series, or cross-sectionally, predicting values for new, as yet unseen, time series. Consequently, when developing a conformal prediction framework for longitudinal data, two types of coverage need to be addressed: longitudinal coverage, which measures coverage over time for each series, and cross-sectional coverage, which measures coverage over the population for each time-point. Whilst existing methods designed for univariate time series can be applied to this setting (e.g., by modelling each time series individually), they fail to provide cross-sectional coverage guarantees. Moreover, longitudinal coverage guarantees that are asymptotic in the length of the time series, as in the SPCI method (Xu & Xie, 2022), can be suboptimal for longitudinal data with a large population and relatively short time series. Hence, it is imperative to develop conformal prediction frameworks that leverage data from across both longitudinal and cross-sectional axes to construct prediction intervals. Previous work for longitudinal data is limited, with Stankeviciute et al. (2021) and Lin et al. (2022) being notable examples. Stankeviciute et al. (2021) showed only cross-sectional coverage guarantees, while Lin et al. (2022) recognised the need to consider both cross-sectional and longitudinal coverage. The TQA-B and TQA-E methods (Lin et al., 2022) improved longitudinal coverage empirically, with TQA-E even providing a theoretical guarantee. However TQA-E suffered from the creation of infinitely-wide intervals in up to 4% of time-points in experiments (Lin et al., 2022, Table 7).

To overcome this, we propose Longitudinal Predictive Conformal Inference (LPCI), extending the SPCI framework (Xu & Xie, 2022) to longitudinal data, which obtained asy mptotic longitudinal coverage of univariate time series without creating infinitely-wide intervals. This involves modelling the conformal scores (in this case the residuals of predictions) as a quantile fixed-effects regression problem, and using this trained quantile regressor to construct prediction intervals. We make further modifications to the underlying approach in order to improve empirical performance, for example by taking exponentially weighted averages in the quantile regressors training data, rather than simple lags as is done in Xu & Xie (2022). We prove asymptotically that LPCI achieves theoretical cross-sectional and longitudinal coverage without the creation of infinitely-wide intervals. Our experiments show that LPCI observes the expected cross-sectional coverage, and longitudinal coverage is greater than or competitive with TQA-B and TQA-E. We also compare to using SPCI separately on each time series as well as to Conformalized Quantile Regression (CQR) of Romano et al. (2019), with LPCI outperforming both. Finally, we also show that the widths of the prediction intervals of LPCI are more adaptive than CQR, by which we mean that LPCI produces narrower intervals where the model is more certain and wider intervals where the model is less certain, making it more suitable as a measure of uncertainty for longitudinal data.

1.1 Related work

The conformal prediction framework has gained significant traction and has become increasingly popular in the last few years for distribution-free and model-agnostic uncertainty quantification Angelopoulos & Bates (2022). Originating from work by Vovk et al. (2005), Shafer & Vovk (2008), and Papadopoulos (2008), the Split conformal method of Lei et al. (2015) pioneered the type of distribution-free and model-agnostic conformal predictions that are currently prevalent (Gupta et al., 2022; Angelopoulos et al., 2022; Kivaranovic et al., 2020). In Split conformal methods, conformal scores are computed on a calibration set and used to measure score quantiles and create intervals. The only assumption on the data is exchangeability, which ensures that the scores calculated on the random calibration set for determining the quantile of the non-conformity scores are transferable to unseen data.

There has been relatively little work on conformal predictions for the cases where exchangeability does not hold, however this is an increasing area of focus. Much of the earlier relevant methods in this area addressed the more general problem of distributional shift, which may be used to make conformal predictions in an online learning setting. Tibshirani et al. (2020) considered the covariate shift problem. Gibbs & Candès (2021) developed a conformal predictive framework to apply to distributional shifts in an online setting, and Feldman et al. (2022) extended this methodology to apply to arbitrary distributional shifts, making no assumptions on the data distribution. As noted by several researchers, the general framework of Gibbs & Candès (2021) may be used for time series data, and the specific temporal dependence of such data can be leveraged to fine-tune and improve this approach (Zaffran et al., 2022). A common approach to improve the coverage of prediction intervals is to perform leave-one-out (LOO) ensemble predictions to feed into the non-conformal scores, as is done by Xu & Xie (2020) and Jensen et al. (2022). Whilst these methods may empirically improve longitudinal coverage, it can be difficult to prove asymptotic longitudinal coverage theoretically without additional distributional assumptions. The SPCI algorithm of Xu & Xie (2022) was shown to achieve longitudinal asymptotic coverage for a single *univariate* time series without any additional distributional assumptions. This was achieved through the introduction of a quantile random forest regressor to predict the required quantile of the non-conformal scores. Moreover, the SPCI algorithm improved on the width of the prediction intervals over the former EnbPI method by the same authors (Xu & Xie, 2020). The ideas of SPCI in using quantile regressors built on previous examples, such as Romano et al. (2019), where the authors applied quantile regression to construct intervals for exchangeable data, a method known as Conformalized Quantile Regression (CQR).

Research for multivariate time series data is limited. Nonetheless, longitudinal time series data is considered by Stankeviciute et al. (2021); Lin et al. (2022). Stankeviciute et al. (2021) focused on multi-step forecasting of future time-points for each group. They proved a cross-sectional coverage guarantee, but did not introduce any guarantees on the conditional longitudinal coverage of each individual time series. Lin et al. (2022) considered the cross-sectional setting, where an estimator is trained on a longitudinal dataset, and developed a conformal prediction algorithm that produces prediction intervals for an *unseen* group. Lin et al. (2022) showed cross-sectional coverage, however longitudinal coverage was only guaranteed with the creation of infinitely-wide intervals at certain time-points. Our work differs by introducing a conformal prediction framework for longitudinal data that guarantees coverage with finite-width intervals.

2 Methodology

Longitudinal datasets consist of temporal sequences of observations $\{(X_t^{(g)}, Y_t^{(g)})\}_{t=1}^T$ of length $T > 1$, for each group $g \in G = \{1, \dots, |G|\}$, where $Y_t^{(g)} \in \mathbb{R}$ is continuous scalar and $X_t^{(g)} \in \mathbb{R}^d$ consists of features which may contain exogenous features, dates, group identifiers or lags and moving averages of $Y_t^{(g)}$. We are assuming that the dataset is balanced in the sense that there exist observations for each combination of g and t

We distinguish two types of study for such datasets: **cross-sectional** and **longitudinal**. In **cross-sectional studies** we partition the groups $G = G_{\text{train}} \sqcup G_{\text{test}}$ and train an estimator $\hat{f}$ on $\{(X_t^{(g)}, Y_t^{(g)}) \mid 1 \leq t \leq T, g \in G_{\text{train}}\}$; point predictions $\hat{Y}_t^{(g)} := \hat{f}(X_t^{(g)})$ are made on the test groups $g \in G_{\text{test}}$ for the same time-points $1 \leq t \leq T$ seen in the training set. In **longitudinal studies** $G = G_{\text{train}} = G_{\text{test}}$ and we train estimator $\hat{f}$ on all groups, $\{(X_t^{(g)}, Y_t^{(g)}) \mid 1 \leq t \leq T, g \in G\}$; point predictions $\hat{Y}_{T+k}^{(g)} := \hat{f}(X_t^{(g)})$ are made across all groups $g \in G$ at future time-points $k \geq 1$.

In conformal predictions, point predictions are accompanied by prediction intervals $\widehat{C}_{t-1}(X_t^{(g)})$ which should ideally contain the true value $Y_t^{(g)}$ with probability $1 - \alpha$ for each group, where $\alpha \in (0, 1)$ is a pre-specified significance level. Such intervals are constructed by measuring (non-)conformity of predictions with the true value. A common scoring function to do so in regression is the residual:

$$\hat{\epsilon}_t^{(g)} := Y_t^{(g)} - \hat{Y}_t^{(g)} \in \mathbb{R}. \tag{1}$$

In traditional conformal predictions, a theoretical guarantee on the coverage level is provided to the effect that the probability of the true values lying in prediction intervals is $1 - \alpha$. However, due to the lack of exchangeability in time series data it is generally not possible to obtain such strong guarantees. As discussed in Lin et al. (2022), for longitudinal data there are two types of coverage. *Cross-sectional coverage* concerns the level of coverage over the group dimension g, whereas *longitudinal coverage* concerns the level of coverage of the temporal dimension t.

Definition 2.1 (Asymptotic cross-sectional coverage)**.** *We say that the conformal intervals* $\widehat{C}_{t-1}(X_t^{(g)})$ *have* asymptotic cross-sectional coverage *if, for all* $\varepsilon > 0$*, there exists* T_0 *such that*

$$\Pr(Y_t^{(g)} \in \widehat{C}_{t-1}(X_t^{(g)})) > 1 - \alpha - \varepsilon,$$

for all $t > T_0$ *and* $g \in G$*.*

Definition 2.2 (Asymptotic longitudinal coverage)**.** *We say that the conformal intervals* $\widehat{C}_{t-1}(X_t^{(g)})$ *have* asymptotic longitudinal coverage *for a group* g*, if*

$$\Pr(Y_t^{(g)} \in \widehat{C}_{t-1}(X_t^{(g)}) \mid X_t^{(g)}) \to 1 - \alpha \textit{ uniformly as } T \to \infty.$$

Cross-sectional coverage is marginal over the groups for a fixed given (large enough) time-point; on the other hand, longitudinal coverage is asymptotic in t and conditional over the temporal dimension. As in Lin et al. (2022), we make the reasonable assumption that the groups are exchangeable. As a result, we are able to obtain in Theorem 2.5 an asymptotic cross-sectional coverage guarantee; note however that this is weaker than the corresponding cross-sectional coverage guarantee in (Lin et al., 2022, Theorem 3.1) which is finite-sample. We also prove asymptotic longitudinal coverage in Theorem 2.4 by adapting the argument of (Xu & Xie, 2022, Theorem 4.4).

2.1 LPCI algorithm

Longitudinal Predictive Conformal Inference (`LPCI`) is a conformal prediction method for longitudinal data that extends the SPCI algorithm of Xu & Xie (2022). We model the longitudinal data as

$$Y_t^{(g)} = \hat{f}(X_t^{(g)}) + \hat{\epsilon}_t^{(g)},$$

where $\hat{f}$ is a trained regression model. We assume that $X_t^{(g)}$ contains a group identifier feature alongside other features. This setup is equivalent to a fixed-effects model. The point predictor $\hat{f}$ is being trained on the entire dataset so as to leverage inter-group dependencies and data more efficiently. This distinguishes the method from training SPCI independently on G univariate time series. In such an approach, each independent SPCI model has data available only for a single group's time series; this may be quite small and, as we show in Section 3, can lead to poor performance of SPCI.

Like SPCI we make use of a past window of residual errors to make quantile predictions and form future prediction intervals. In contrast to SPCI we create exponentially-weighted means of these residuals. Suppose we are at time t and, for a fixed window size $w \geq 1$, we have observed w values for group g so far. We let

$$\mathcal{E}_{t,w}^{(g)} := (\bar{\epsilon}_{t-1}^{(g)}, \ldots, \bar{\epsilon}_{t-w}^{(g)}) \in \mathbb{R}^w,$$

denote the lagged window of exponentially-weighted mean residuals for group g, where

$$\bar{\epsilon}_k^{(g)} = k^{-1} \sum_{i=1}^{k} \gamma^{k-i} \hat{\epsilon}_i^{(g)},$$

and $\gamma \in [0, 1]$ is the smoothing parameter. Let $F(z \mid \mathcal{E}_{t,w}^{(g)}) = \Pr(\hat{\epsilon_t}^{(g)} \leq z \mid \mathcal{E}_{t,w}^{(g)})$ be the unknown distribution function of the current residual. The true pth quantile of the residual is defined by

$$Q_{t,p}^{(g)} := \inf\{e \in \mathbb{R} \mid F(e \mid \mathcal{E}_{t,w}^{(g)}) \geq p\}, \tag{2}$$

where $p \in (0, 1)$; as discussed in Meinshausen (2006), for continuous distribution functions, the quantiles are defined by the property $F(Q_{t,p}^{(g)} \mid \mathcal{E}_{t,w}^{(g)}) = p$ and so we have

$$F(Q_{t,1-\alpha/2}^{(g)} \mid \mathcal{E}_{t,w}^{(g)}) - F(Q_{t,\alpha/2}^{(g)} \mid \mathcal{E}_{t,w}^{(g)}) = 1 - \alpha, \tag{3}$$

where $\alpha \in (0, 1)$. The main idea behind both SPCI and LPCI is to train a model whose quantile estimates $\widehat{Q}_{t,p}$ converge uniformly to the true quantile as the training data increases. In the limit, intervals constructed using $\widehat{Q}_{t,1-\alpha/2}$ and $\widehat{Q}_{t,\alpha/2}$ will contain the true value with probability $1 - \alpha$ by virtue of Eq. (3), this is shown in more detail in Section 2.2. The prediction interval in LPCI is constructed as

$$\widehat{C}_{t-1}(X_t^{(g)}) := \left[\hat{Y}_t^{(g)} + \widehat{Q}_{t,\beta}^{(g)}, \hat{Y}_t^{(g)} + \widehat{Q}_{t,1-\alpha+\beta}^{(g)}\right], \tag{4}$$

where $\beta \in [0, \alpha]$ minimises the interval width of $(1 - \alpha)$-quantiles:

$$\beta = \underset{p \in [0,\alpha]}{\operatorname{argmin}} \left\{ \widehat{Q}_{t,1-\alpha+p}^{(g)} - \widehat{Q}_{t,p}^{(g)} \right\}. \tag{5}$$

To estimate the quantiles we model the past residuals as a fixed-effects model by fitting a quantile random forest regressor on features containing $\mathcal{E}_{t,w}^{(g)}$ and group identifiers, with current residual $\hat{\epsilon}_t$ as the target. At the first date in the testing time-point (either $t = 0$ or $t = T + 1$ depending on the setting), there exist T past residuals across groups in G_{train}. We create a training dataset of size $(\tilde{T}|G_{\text{train}}|) \times (w + 1)$, where $\tilde{T} := T - w$, by defining samples and labels, respectively, for $t' = 1, \ldots, \tilde{T}$, as

$$\begin{aligned} \widetilde{X}_t^{(g)} &: = \{(\mathcal{E}_{t+w,w}^{(g)}, g)\} \in \mathbb{R}^{w+1}, \\ \widetilde{Y}_t^{(g)} &: = \hat{\epsilon}_{t+w}^{(g)} \in \mathbb{R}. \end{aligned} \tag{6}$$

After each new time-point in the testing data, $|G_{\text{test}}|$ new residuals are observed and the quantile random forest is subsequently retrained on data including the new observations from G_{test}. In the cross-sectional setting, for the first w time-points of the testing period we do not have observed residuals for the test groups. In this case we create dummy residuals of zeros. The full algorithm for LPCI is given in Algorithm 1.

Algorithm 1 Longitudinal Predictive Conformal Inference (LPCI).

```
**Require:** Training data $\{X_t^{(g)}, Y_t^{(g)} \mid g \in G_{\text{train}}, 1 \le t \le T\}$, where $X_t^{(g)} \in \mathbb{R}^d$ and $Y_t^{(g)} \in \mathbb{R}$; regression training algorithm $\mathcal{A}$; significance level $\alpha$; quantile regression training algorithm $\mathcal{Q}$; window size $w$; prediction groups $G_{\text{test}}$ and prediction time-frame $T_{\text{test}}$ ($1 \le t \le T$ if cross-sectional; $t > T$ if longitudinal).
**Output:** Point predictions $\hat{Y}_t^{(g)} \in \mathbb{R}$ and prediction intervals $\hat{C}_{t-1}(X_t^{(g)}) \in \mathbb{R}^2$, for $g \in G_{\text{test}}$ and all $t \in T_{\text{test}}$.
Train a point predictor $\hat{f} = \mathcal{A}(\{X_t^{(g)}, Y_t^{(g)}\}_{g,t})$.
Use $\hat{f}$ to obtain a panel dataset of residuals $\{\hat{\epsilon}_t^{(g)}\}_{g,t}$ as described in Eq. (1).
Form quantile regression training data $\{\widetilde{X}_t^{(g)}, \widetilde{Y}_t^{(g)} \mid g \in G_{\text{train}}, 1 \le t \le \tilde{T}\}$ as described in Eq. (6), where $\tilde{T} = T - w, \widetilde{X}_t^{(g)} \in \mathbb{R}_{w+1}, \widetilde{Y}_t^{(g)} \in \mathbb{R}$.
**for** $t' \in T_{\text{test}}$ **do**
    **for** $h \in G_{\text{test}}$ **do**
        Train the quantile regressor $\widehat{\mathcal{Q}} = \mathcal{Q}(\{\widetilde{X}_t^{(g)}, \widetilde{Y}_t^{(g)}\}_{g,t})$ on the latest residual data.
        Obtain the point prediction $\hat{Y}_t^{(h)} = \hat{f}(X_{t'}^{(h)})$.
        Form the prediction intervals $\widehat{C}_{t'-1}(X_{t'}^{(h)})$ as described in Eq. (4) by taking quantile estimates on the latest residual data $\widehat{Q}_{t',p}^{(h)} = \widehat{\mathcal{Q}}(\widetilde{X}_{t'-w}^{(h)})$.
        Obtain the latest residuals after observation of true value $\hat{\epsilon}_{t'}^{(h)} = Y_{t'}^{(h)} - \hat{Y}_{t'}^{(h)}$.
        Update the training data for the quantile random forest by adding the new samples $\widetilde{X}_{t'-w}^{(h)}$ and $\widetilde{Y}_{t'-w}^{(h)}$.
    **end for**
**end for**
**return** $\{\hat{Y}_{t'}^{(h)}, \widehat{C}_{t'-1}(X_{t'}^{(h)}) \mid t' \in T_{\text{test}}, h \in G_{\text{test}}\}$.
```

2.2 Theory

We show that LPCI obtains asymptotic marginal cross-sectional and conditional longitudinal coverage as the number of training samples of the quantile random forest tends to ∞. This limit can be achieved by either taking the number of time-points $T \to \infty$ or the number groups $G \to \infty$ corresponding. However, since the number of groups (in both the cross-sectional and longitudinal case) is generally kept fixed, we only consider the limit in the number of time-points. All proofs can be found in Appendix A.1.

First we prove asymptotic longitudinal conditional coverage (Definition 2.2). Our argument follows that of Xu & Xie (2022) with some changes and additional assumptions. The aim is to prove that

$$\Pr\left(Y_t^{(g)} \in \widehat{C}_{t-1}(X_t^{(g)}) \;\middle|\; X_t^{(g)}\right) \to 1 - \alpha \text{ uniformly as } T \to \infty. \tag{7}$$

By definition of the interval, Eq. (4), we have

$$\Pr(Y_t^{(g)} \in \widehat{C}_{t-1}(X_t^{(g)}) \mid X_t^{(g)}) = F(\widehat{Q}_{t,1-\alpha+\beta}^{(g)} \mid \mathcal{E}_{t,w}^{(g)}) - F(\widehat{Q}_{t,\beta}^{(g)} \mid \mathcal{E}_{t,w}^{(g)}). \tag{8}$$

If we have uniform convergence of the longitudinal quantile random forest to the true quantile:

$$\sup_{p \in [0,1]} |\widehat{Q}_{t,p}^{(g)} - Q_{t,p}^{(g)}| \to 0 \text{ as } T \to \infty, \tag{9}$$

then through Eq. (3), we arrive at the asymptotic guarantee of Eq. (7). So it suffices to prove Eq. (9).

First, consider a *random forest regressor* (non-quantile) trained on the training data $\{(\widetilde{X}_t^{(g)}, \widetilde{Y}_t^{(g)}) \mid g \in G_{\text{train}}, 1 \le t \le \tilde{T}\}$ containing the residuals. This random forest estimates a conditional cumulative distribution function $\widehat{F}(z \mid x)$ of the residuals. The idea is that if the estimated distribution $\widehat{F}(z \mid x)$ converges uniformly to the true (unknown) distribution of the residuals $F(z \mid x)$, then actually the estimated quantiles of the corresponding *quantile* random forest converge uniformly to the true quantiles – this is (Xu & Xie, 2022, Proposition 4.2) and it holds for our setting as well. So we need only focus on the random forest's distribution function.

We assume that, for each t and each g, the support of the features are contained in a compact space $\text{Supp}(\widetilde{X}_t^{(g)}) \subseteq \mathbb{B} \subseteq \mathbb{R}^d$. Each parameterised tree $T(\theta)$ in the random forest is trained such that each leaf of the tree is associated with a rectangular subspace of $\mathbb{B}$. These subspaces are disjoint and cover $\mathbb{B}$ so that, for every $x \in \mathbb{B}$ there exists a unique leaf $\ell(x, \theta)$ such that $x \in R_{\ell(x,\theta)}$. Assume that the random forest has parameterised trees $\{T(\theta_k)\}_{k=1}^K$ which have been trained on the training data. Given $x \in \mathbb{B}$, we can give an explicit form of the conditional distribution function of the quantile random forest through weighting past observations as follows:

$$L(x,\theta) := \#\{(g,t) \in \{1,\dots,|G_{\text{train}}|\} \times \{1,\dots,\tilde{T}\} \mid \widetilde{X}_t^{(g)} \in R_{\ell(x,\theta)}\}, \tag{10}$$
$$w_t(x,\theta) := L(x,\theta)^{-1}\mathbb{1}(\widetilde{X}_t^{(g)} \in R_{\ell(x,\theta)}), \tag{11}$$
$$w_t(x) := K^{-1}\sum_{k=1}^{K} w_t(x,\theta_k). \tag{12}$$

The value $L(x,\theta)$ counts the total node size of the leaf $\ell(x,\theta)$ across all observed groups and time-steps. The value $w_t(x,\theta)$ weights a relevant leaf by its node-size and finally $w_t(x)$ takes the average across ensemble trees. We have that

$$\sum_{g\in G_{\text{train}}}\sum_{t=1}^{\tilde{T}} w_t(x) = 1.$$

The cumulative distribution function of the estimates of the *random forest* trained on $\{(\widetilde{X}_t^{(g)}, \widetilde{Y}_t^{(g)})\}_{g,t}$, conditioned on some input $\widetilde{X}_k^{(h)}$, is given (Meinshausen, 2006) by

$$\widehat{F}(z \mid \widetilde{X}_k^{(h)} = x) := \sum_{g\in G_{\text{train}}}\sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{1}(\widetilde{Y}_t^{(g)} \leq z). \tag{13}$$

Proposition 2.3. *Under Assumptions A.1–A.5 of Xu & Xie (2022) and our additional Assumption A.1 the estimated conditional distribution function $\widehat{F}(z \mid x)$ (Eq. (13)) for the random forest trained on $|G_{train}| \cdot \tilde{T}$ samples converges uniformly in z for each $x \in \mathbb{B}$:*

$$\sup_{z\in\mathbb{R}} |\widehat{F}(z \mid x) - F(z \mid x)| \to 0 \textit{ in probability as } \tilde{T} \to \infty.$$

Theorem 2.4 (Asymptotic longitudinal conditional coverage of `LPCI`)**.** *Under the same assumptions of Proposition 2.3, we have, for any $\alpha \in (0,1)$ and any $g \in G$, that*

$$|\Pr(Y_t^{(g)} \in \widehat{C}_{t-1}(X_t^{(g)}) \mid X_t^{(g)}) - (1-\alpha)| \to 0 \textit{ in probability as } T \to \infty. \tag{14}$$

Theorem 2.5 (Asymptotic cross-sectional marginal coverage of LPCI)**.** *Let $\varepsilon > 0$ be given. Under the same assumptions of Proposition 2.3 and the additional assumption that the data $\{(X_t^{(g)}, Y_t^{(g)})\}_{g,t}$ is exchangeable in g, there exists T_0 such that for any $t > T_0$ and any $\alpha \in (0,1)$, we have*

$$\Pr(Y_t^{(h)} \in \widehat{C}_{t-1}(X_t^{(h)})) > 1-\alpha-\varepsilon,$$

for any $h \in G_{test}$.

3 Results

In this section, we give results in two experimental settings. In Section 3.1 we provide results in a cross-sectional experimental setting and in Section 3.2 we provide results in a longitudinal experimental setting. We aim to give empirical results that evidence the theoretical analysis in the previous section. To this end, we verify the following:

1. LPCI observes expected marginal cross-sectional coverage.
2. LPCI improves longitudinal coverage rates over literature benchmarks.
3. LPCI exhibits high adaptivity of interval width.

With respect to adaptivity of interval widths, the first aim is to construct prediction intervals that are not simply wider than benchmark methods at every data point, as this would artificially improve coverage. Hence, we need to ensure that average interval width is not too large. Additionally, always producing similar interval widths across the board is also generally undesirable – we want conformal prediction intervals to correctly reflect the uncertainty related to the prediction. Therefore, we usually expect to have a broad range of interval widths widths, see Angelopoulos & Bates (2022, Section 3.1). We measure this by measuring the standard deviation of interval widths predicted at each time-point in the test set. Overall, we evaluate LPCI using the following metrics:

Table 1: Experimental details for the two tested datasets. n_x and G_x refer to the number of samples and groups in the $x \in \{\text{train}, \text{test}\}$ datasets respectively. T refers to the length of each time series.

	n_{train}	n_{test}	G_{train}	G_{test}	T
COVID	9,000	2,400	300	80	30
EEG	25,600	12,800	400	200	64

- **Marginal coverage:** the coverage rate averaged across all groups and all time-points. This measures *cross-sectional coverage*.
- **Tail coverage:** the average coverage across the lowest 10% covered groups. This measures *longitudinal coverage*.
- **Width coefficient of variation (CoV):** measures the standard deviation divided by the mean width, taken across all groups and time-points. This metric penalises widths that have a large mean and narrow distribution, and favours widths with a lower average and broader distribution. This helps to measure *width adaptivity*.

3.1 Cross-sectional Experiments

Experimental Setup. In the cross-sectional setting, we divide our dataset into train and test sets by randomly selecting groups to be entirely contained in either the train or test set, with the temporal dimension being kept the same across both sets. This approach allows us to evaluate the generalisation performance of our model on unseen group categories while maintaining the integrity of the temporal structure in the data.

Baselines. To benchmark LPCI, we compare against the TQA-B and TQA-E methods from Lin et al. (2022). We follow their experimental methodology as mentioned in their paper in order to compare results. We report TQA-B and TQA-E results from Lin et al. (2022). We also compare against Conformalized Quantile Regression (CQR, Romano et al. (2019)) and a traditional split conformal method (Split) with a tuned random forest regressor as base estimator. We use implementations of CQR and Split from the MAPIE Python package.[1]

Datasets. We perform experiments on two datasets considered by Lin et al. (2022): COVID contains daily cases across 380 local council authorities in the United Kingdom across the month of March 2022; EEG contains 64 downsampled electroencephalogram readings across one second for 600 patient-trials. Table 1 contains experimental details of these datasets.

Experimental Details. Our point predictor is a random forest regression model. This is trained as an 5-fold ensemble model with individual estimators, with groups contained entirely within folds and predictions being made by mean aggregation of the individual learners. We use a single lagged target $Y_{t-1}^{(g)}$ as a feature and label-encoded group identifiers as features. Both the target and lagged target columns are standardised across groups. We tune the hyperparameters of the random forest using a randomised grid search across 5-fold group cross validation splits.

The significance level is $\alpha = 0.1$. For training the quantile random forest, we use a window size of 20, yielding a training data consisting of 20 lagged exponentially-weighted averages of past residuals, as described in Section 2. We also tune the hyperparameters of the quantile random forest using a randomised grid search across 5-fold cross validation splits. As in Lin et al. (2022), we report metric values only for the last 20 time-points for each group. A final point to mention in the cross-sectional setting is the need to create dummy residuals for the QRF on the first date, since no historical residuals are available at that point for the test groups; since we are only computing metrics on the final 20 time-points for each group this has limited impact on the results.

Table 2: Marginal coverage of LPCI on the COVID and EEG datasets compared to baselines in the cross-sectional setting. Marginal coverages in italics do not satisfy expected coverage rates. Experiments are repeated over five seeds. We report the mean ± standard deviation across experiments.

Dataset	LPCI	TQA-B	TQA-E	CQR	Split
COVID	0.964 ± 0.006	0.908 ± 0.015	0.917 ± 0.009	*0.897 ± 0.010*	0.900 ± 0.018
EEG	0.924 ± 0.067	0.907 ± 0.015	0.906 ± 0.008	0.901 ± 0.008	0.908 ± 0.004

[1] https://pypi.org/project/MAPIE/

Table 3: Tail coverage on the COVID and EEG datasets for LPCI compared to baselines in the cross-sectional setting. Higher values are preferred. The best overall results are **bolded**; the best results for methods that do not create infinitely-wide intervals are underlined. Experiments are repeated over five seeds and we report the mean ± standard deviation across experiments.

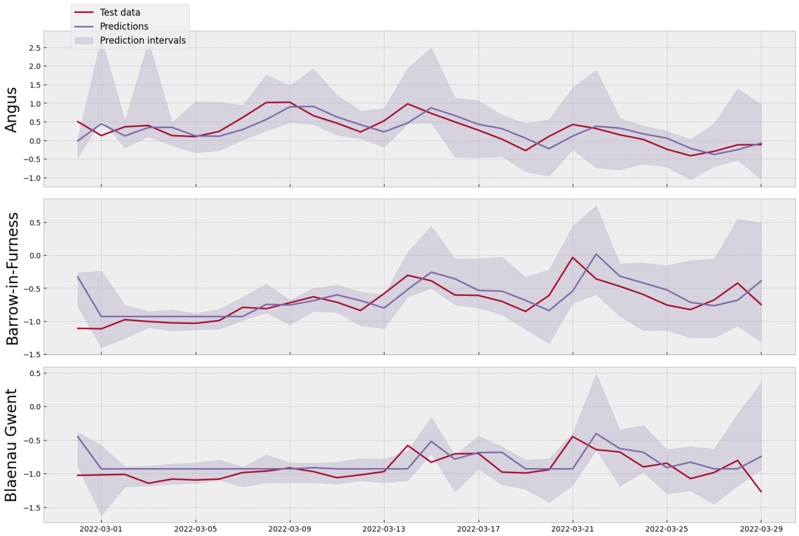
Dataset	LPCI	TQA-B	TQA-E	CQR	Split
COVID	**0.874 ± 0.015**	0.700 ± 0.045	0.824 ± 0.013	0.779 ± 0.015	0.607 ± 0.048
EEG	**0.792 ± 0.007**	0.710 ± 0.022	**0.790 ± 0.012**	0.711 ± 0.009	0.737 ± 0.009

Table 4: Width CoV for COVID and EEG datasets. Higher values are preferred. We do not include the TQA-B or TQA-E methods in this table, since standard deviations are not available from Lin et al. (2022), however a comparison of widths alone can be found in Appendix A.2. The best results are bolded. Experiments are repeated over five seeds and we report the mean ± standard deviation across experiments.

Dataset	LPCI	CQR	Split
COVID	**0.536 ± 0.023**	0.391 ± 0.046	0.056 ± 0.030
EEG	**0.275 ± 0.047**	0.075 ± 0.008	0.037 ± 0.014

Discussion. Cross-sectional results of LPCI on COVID and EEG datasets for the three metrics described can be found in Tables 2, 3, 4. We observe LPCI obtains expected marginal coverage of greater than 0.9 in both cases and gets higher marginal coverage than both TQA-B and TQA-E (Table 2). On the COVID dataset, the same can be said for tail coverage rate (Table 3). For the EEG dataset, LPCI improves tail coverage over TQA-B and is competitive with TQA-E, however note that this is achieved without the creation of infinitely-wide intervals which occurs in TQA-E. Table 4 demonstrates that the widths of LPCI have a greater spread relative to the average width than compared to the CQR and Split methods. Combined with the improved tail coverage rates from Table 3, this shows that LPCI's intervals are more capable of adapting to harder predictions and suggests that LPCI exhibits higher width adaptivity than CQR and Split.

In Figure 1 we show some examples of interval widths created by LPCI for the cross-sectional experiment on the COVID dataset. For LPCI, we saw longitudinal (tail) coverage of 0.874 (Table 3). Note how, for Barrow-In-Furness and Blenau Gwent, the first half of the month corresponds to relatively flat true and predicted values. Correspondingly LPCI produces narrow intervals, reflecting low uncertainty. In the second half of March 2022, for both of those councils the true values become more volatile, and greater model errors occur. As we would hope to see, there is a widening of LPCI intervals accompanying this phenomenon.



Coverage	Width RSD
93%	0.39
93%	0.54
83%	0.5

Figure 1: Examples of LPCI intervals for 3 councils in the COVID test set in the cross-sectional experiment. The red line is the observed data, the purple line is the predictions of the underlying model, and the shaded regions are the LPCI intervals. Note that observed values have been normalised.

3.2 Longitudinal Experiments

Experimental Setup. In the longitudinal setting, we partitioned our dataset by dividing the data along the time dimension. This allowed us to evaluate the generalisation performance of our model on unseen time steps while ensuring that each group was represented in both the train and test sets.

Baselines. In this case, we are unable to benchmark against Lin et al. (2022), since they do not measure results in such a setting. Instead we benchmark against using SPCI separately on each group as well as CQR and Split.

Datasets. We use the COVID dataset in which we have 380 groups corresponding to local council authorities in the UK. We train on all groups in the month of February 2022 and test on all groups in the month of March 2022.

Experimental Details. The experiments are almost identical to those in the cross-sectional setting of Section 3.1, except we have no need to create dummy prediction data for the quantile random forest on the first testing date as historical data exists for all groups.

Table 5: LPCI results on the COVID dataset in the longitudinal experiment setup. Marginal coverages that do not satisfy expected coverage rates are in italics. Up and down arrows indicate whether higher or lower values are preferred for that metric, respectively; best tail coverage is bolded; best width CoV for valid marginal coverages is bolded. Experiments are repeated over five seeds and we report the mean ± standard deviation across experiments.

Metric	LPCI	SPCI	CQR	Split
Marginal coverage	0.936 ± 0.004	*0.657 ± 0.007*	0.918 ± 0.003	*0.892 ± 0.001*
Tail coverage ↑	**0.830 ± 0.005**	0.539 ± 0.003	0.760 ± 0.004	0.625 ± 0.001
Width CoV ↑	**0.500 ± 0.009**	0.582 ± 0.007	0.304 ± 0.016	0.103 ± 0.002

Discussion. Table 5 contains LPCI results on the COVID dataset for the longitudinal setting. Only LPCI and CQR obtain expected marginal coverage rates, with LPCI obtaining slightly better coverage than CQR. The tail coverage of LPCI significantly outperforms other methods.

The width coefficient of variation is higher for LPCI than it is for CQR and Split which, as per the discussion in the previous section, indicates a greater width adaptivity for LPCI. The greater width adaptivity of LPCI over CQR is visualised in Figure 2; the left plot shows the wider spread of widths for intervals created by LPCI compared to CQR, the right plot shows that the group coverage of LPCI intervals are more concentrated towards higher coverage rates than those of CQR. This indicates that, for the COVID dataset, the greater spread of widths for LPCI is accompanied by improved coverage rates over the worst-covered groups, and therefore that LPCI is better able to reflect the uncertainty of the model.

The Split method is expected to obtain poor tail coverage, since it does not readily apply to time series data. However, it is interesting to observe the poor performance of SPCI in this scenario, which does come with longitudinal guarantees for univariate time series. This is the result of two key factors. Firstly we only have 30 training and test points available for each time series and at the final testing date each SPCI has only seen 60 datapoints in total. This lack of data leads to poor performance on each individual time series. Secondly, SPCI used in this way is not able to leverage the exchangeability of groups in the dataset. This shows how the LPCI methodology of modelling a quantile random forest over the entire dataset can help to both model inter-group dependencies and overcome data limitations.

4 Conclusion

In this paper, we proposed Longitudinal Predictive Conformal Inference (LPCI) as a distribution-free and model-agnostic uncertainty quantification approach for longitudinal time series data. Our method provides asymptotic theoretical guarantees across both cross-sectional and temporal dimensions. This is done without creating infinitely-wide prediction intervals at any point, which is a first in conformal predictions for longitudinal data. In experiments, we observed that LPCI was able to improve longitudinal coverage over the current state-of-the-art, whilst also exhibiting a greater adaptivity of widths. Theoretically, however, our framework lacks finite-sample cross-sectional coverage and is only able to guarantee such coverage asymptotically; this could be fixed by combining some ideas from the approach of Lin et al. (2022). Moreover, we may be able to leverage more predictive power by looking at alternative quantile regression models, which may help to improve empirical longitudinal coverage even further.

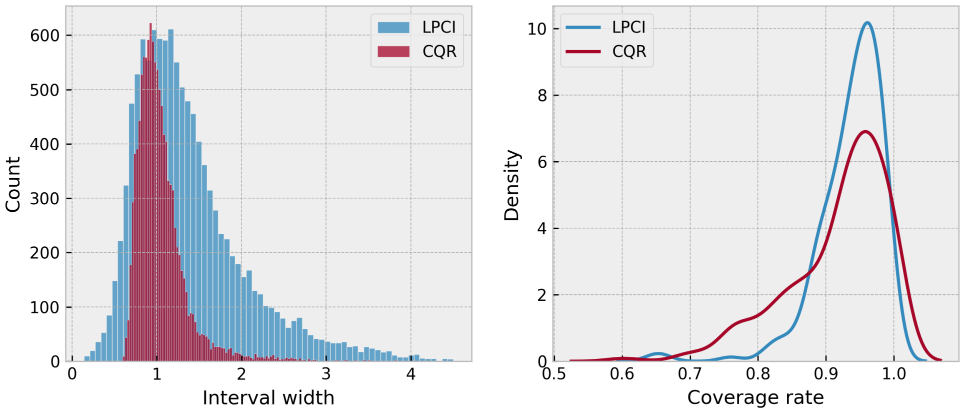


Figure 2: (left) Distribution of LPCI widths compared to CQR widths on the COVID dataset. We observe a significantly larger spread in both directions for LPCI than for CQR. (right) Distribution of coverage rates for each local council authority in the COVID datasets for LPCI compared to CQR.

Acknowledgments

The authors would like to thank Greig Cowan and Graham Smith of NatWest Group's Data Science & Innovation team for the time and support needed to develop this research paper. We would also like to express our sincere gratitude to Greig Cowan for his insightful feedback and constructive suggestions that improved the quality of this paper.

References

Alaa, Ahmed, Gurdasani, Deepti, Harris, Adrian, Rashbass, Jem, and Schaar, Mihaela. Machine learning to guide the use of adjuvant therapies for breast cancer, 08 2020.

Angelopoulos, Anastasios, Bates, Stephen, Malik, Jitendra, and Jordan, Michael I. Uncertainty sets for image classifiers using conformal prediction, 2022.

Angelopoulos, Anastasios N. and Bates, Stephen. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022.

Díaz-González, Francisco, Sumper, Andreas, Gomis-Bellmunt, Oriol, and Villafáfila-Robles, Roberto. A review of energy storage technologies for wind power applications. *Renewable & Sustainable Energy Reviews*, 16:2154–2171, 2012.

Feldman, Shai, Ringel, Liran, Bates, Stephen, and Romano, Yaniv. Risk control for online learning models, 2022. URL `https://arxiv.org/abs/2205.09095`.

Gibbs, Isaac and Candès, Emmanuel. Adaptive conformal inference under distribution shift, 2021. URL `https://arxiv.org/abs/2106.00170`.

Gupta, Chirag, Kuchibhotla, Arun K., and Ramdas, Aaditya. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022. ISSN 0031-3203. doi:https://doi.org/10.1016/j.patcog.2021.108496. URL `https://www.sciencedirect.com/science/article/pii/S0031320321006725`.

Harries, Michael, Wales, New South, et al. Splice-2 comparative evaluation: Electricity pricing. 1999.

Jensen, Vilde, Bianchi, Filippo Maria, and Anfinsen, Stian Normann. Ensemble conformalized quantile regression for probabilistic time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022. doi:10.1109/tnnls.2022.3217694. URL `https://doi.org/10.1109%2Ftnnls.2022.3217694`.

Kivaranovic, Danijel, Johnson, Kory D., and Leeb, Hannes. Adaptive, distribution-free prediction intervals for deep networks, 2020.

Lei, Jing, Rinaldo, Alessandro, and Wasserman, Larry. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1–2):29–43, jun 2015. ISSN 1012-2443. doi:10.1007/s10472-013-9366-6. URL `https://doi.org/10.1007/s10472-013-9366-6`.

Lin, Zhen, Trivedi, Shubhendu, and Sun, Jimeng. Conformal prediction with temporal quantile adjustments, 2022. URL `https://arxiv.org/abs/2205.09940`.

Mears, Daniel, Cochran, Joshua, and Lindsey, Andrea. Offending and racial and ethnic disparities in criminal justice: A conceptual framework for guiding theory and research and informing policy. *Journal of Contemporary Criminal Justice*, 32, 10 2015. doi:10.1177/1043986215607252.

Meinshausen, Nicolai. Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999, dec 2006. ISSN 1532-4435.

Papadopoulos, Harris. *Inductive Conformal Prediction: Theory and Application to Neural Networks*. 08 2008. ISBN 978-953-7619-03-9. doi:10.5772/6078.

Romano, Yaniv, Patterson, Evan, and Candès, Emmanuel J. Conformalized quantile regression, 2019.

Shafer, Glenn and Vovk, Vladimir. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12): 371–421, 2008. URL `http://jmlr.org/papers/v9/shafer08a.html`.

Stankeviciute, Kamile, M. Alaa, Ahmed, and van der Schaar, Mihaela. Conformal time-series forecasting. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., and Vaughan, J. Wortman (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6216–6228. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/312f1ba2a72318edaaa995a67835fad5-Paper.pdf`.

Tibshirani, Ryan J., Barber, Rina Foygel, Candes, Emmanuel J., and Ramdas, Aaditya. Conformal prediction under covariate shift, 2020.

van der Schaar, Mihaela, Alaa, Ahmed M., Floto, R. Andres, Gimson, Alexander, Scholtes, Stefan, Wood, Angela M., McKinney, Eoin F., Jarrett, Daniel, Lio', Pietro, and Ercole, Ari. How artificial intelligence and machine learning can help healthcare systems respond to covid-19. *Machine Learning*, 110:1 – 14, 2020.

Vovk, Vladimir, Gammerman, Alex, and Shafer, Glenn. *Algorithmic Learning in a Random World*. 01 2005. doi:10.1007/b106715.

Xu, Chen and Xie, Yao. Conformal prediction for time series, 2020. URL `https://arxiv.org/abs/2010.09107`.

Xu, Chen and Xie, Yao. Sequential predictive conformal inference for time series, 2022. URL `https://arxiv.org/abs/2212.03463`.

Zaffran, Margaux, Dieuleveut, Aymeric, Féron, Olivier, Goude, Yannig, and Josse, Julie. Adaptive conformal predictions for time series, 2022. URL `https://arxiv.org/abs/2202.07282`.

A Supplementary Material

A.1 Assumptions and proofs

Assumption A.1. *Define $U_t^{(g)} := F(\widetilde{Y}_t^{(g)} \mid \widetilde{X}_t^{(g)})$ as the quantile of the tth observation of the gth group, condition on the feature $\widetilde{X}_t^{(g)}$ with the property that $U_t \sim U[0,1]$. If $x \in$ Supp$(\{X_t^{(g)}\}_{t\geq 1})$, then we assume that there exists a function $\tilde{g}$ such that*

$$\mathit{Cov}(\mathbb{1}(U_t^{(g)} \leq F(z \mid x)), \mathbb{1}(U_t^{(h)} \leq F(z \mid x))) \leq \tilde{g}(t),$$

for all $t \geq 1$ and all $g \neq h$. Moreover, $\tilde{g}$ has the following bounded growth

$$\lim_{\tilde{T}\to\infty} \left[\int_1^{\tilde{T}} \tilde{g}(u)du\right] / \tilde{T}^2 = 0.$$

Note that Assumption A.1 is similar to (Xu & Xie, 2022, Assumption A.1), however here we are ranging over the group axis. It essentially assumes subquadratic growth of the cross-sectional covariance as the time-points tend to ∞. Note that if the groups are exchangeable, then it is enough to check this assumption once, since all the covariances of pairwise groups are equal.

Proof of Proposition 2.3. We follow the proof of (Xu & Xie, 2022, Proposition 4.3) and (Meinshausen, 2006, Theorem 1) with minor changes due to the slight difference in the form of the estimated conditional distribution function. Note that (Meinshausen, 2006, Theorem 1) assumes i.i.d observations, which neither we nor Xu & Xie (2022) do. We denote $U_t^{(g)} := F(\widetilde{Y}_t^{(g)} | \widetilde{X}_t^{(g)})$ as the quantile of the tth observation of the gth group; we have $U_t^{(g)} \sim U[0,1]$. Following the calculation of Xu & Xie (2022), we get

$$|\widehat{F}(z \mid x) - F(z \mid x)| \leq \overbrace{\left| \sum_{g \in \tilde{G}} \sum_{t=1}^{\tilde{T}} w_t(x) \mathbb{1}(U_t^{(g)} \leq F(z \mid x)) - F(z \mid x) \right|}^{(a)} + \\ + \sum_{g \in \tilde{G}} \left(\underbrace{\left| \sum_{t=1}^{\tilde{T}} w_t(x) [\mathbb{1}(U_t^{(g)} \leq F(z \mid \widetilde{X}_t^{(g)})) - \mathbb{1}(U_t^{(g)} \leq F(z \mid x))] \right|}_{(b)} \right).$$

The aim is to show that this converges to zero as $\tilde{T} \to \infty$. Term (b) is the same as it was in Xu & Xie (2022), where they showed convergence to zero. Hence we need only show that term (a) converges to zero.

Define $U^{(g)} := \sum_g \sum_t w_t(x) \mathbb{1}(U_t^{(g)} \leq F(z \mid x))$; since $\sum_g \sum_t w_t(x) = 1$ we have $\mathbb{E}[U^{(g)}] = F(z \mid x)$ and, by Chebyshev's inequality, that

$$\Pr\left(\left| \sum_{g \in \tilde{G}} \sum_{t=1}^{\tilde{T}} w_t(x) \mathbb{1}(U_t^{(g)} \leq F(z \mid x)) - F(z \mid x) \right| \geq \varepsilon \right) \leq \text{Var}(U^{(g)})/\varepsilon^2,$$

for any $\varepsilon > 0$. We show that $\text{Var}(U^{(g)}) \to 0$ as $\tilde{T} \to \infty$. By properties of variance and covariance on linear combinations we obtain

$$\text{Var}(U^{(g)}) = \sum_{g \in \tilde{G}} \text{Var}\left(\sum_{t=1}^{\tilde{T}} w_t(x) \mathbb{1}(U_t^{(g)} \leq F(z \mid x)) \right) \tag{15}$$

$$+ \sum_{g \neq h} \sum_{i \neq j} w_i(x) w_j(x) \,\text{Cov}(\mathbb{1}(U_i^{(g)} \leq F(z \mid x)), \mathbb{1}(U_j^{(h)} \leq F(z \mid x))) \tag{16}$$

$$+ \sum_{g \neq h} \sum_{t=1}^{\tilde{T}} w_t^2(x) \,\text{Cov}(\mathbb{1}(U_t^{(g)} \leq F(z \mid x)), \mathbb{1}(U_t^{(h)} \leq F(z \mid x))). \tag{17}$$

The summand of $\sum_g$ in (15) is shown to converge to zero in Xu & Xie (2022). Likewise, the summand of $\sum_{g \neq h}$ in (16) also converges to zero as shown in Xu & Xie (2022) (this requires (Xu & Xie, 2022, Assumption A.1)). Under our additional Assumption A.1 and exchangeability of the groups, the third term converges to zero as follows:

$$\sum_{t=1}^{\tilde{T}} w_t^2(x) \,\text{Cov}(\mathbb{1}(U_t^{(1)} \leq F(z \mid x)), \mathbb{1}(U_t^{(2)} \leq F(z \mid x)) \leq \int_1^{\tilde{T}} \mathcal{O}\left(\frac{1}{\tilde{T}^2}\right) \tilde{g}(u) du \to 0.$$

□

Proof of Theorem 2.4. As $T \to \infty$, the number of samples in the training data for the quantile random forest tends to ∞ as well due to the continual residual updates. In this limit, the conditional distribution function $\widehat{F}(z \mid \mathcal{E}_{t,w}^{(g)})$ of the random forest converges uniformly to the true distribution by Proposition 2.3. Hence we get uniform convergence $\widehat{Q}_{t,p} \to Q_{t,p}$ of the estimated quantiles of the quantile random forest to the true quantile by (Xu & Xie, 2022, Proposition 4.3). By (Xu & Xie, 2022, Assumption A.4) the points of discontinuity of the distribution function F have measure zero, the proof then follows by the continuous mapping theorem. □

Proof of Theorem 2.5. By exchangeability of groups, $F(\widehat{Q}_{t,1-\alpha+\beta}^{(h)} \mid \mathcal{E}_{t,w}^{(h)}) = F(\widehat{Q}_{t,1-\alpha+\beta}^{(g)} \mid \mathcal{E}_{t,w}^{(g)})$ for any group g; likewise for the other quantile $\widehat{Q}_{\beta}^{(h)}$. Using this and Eq. (8) we obtain

$$\begin{aligned} \Pr(\hat{\epsilon}_t^{(h)} \leq \widehat{Q}_{1-\alpha+\beta}^{(h)}) &= \mathbb{E}_h \Pr(Y_t^{(h)} \in C_{t-1}(X_t^{(h)}) \mid X_t^{(h)}) \\ &= \frac{1}{|G|} \sum_{g \in G} \left[F(\widehat{Q}_{t,1-\alpha+\beta}^{(g)} \mid \mathcal{E}_{t,w}^{(g)}) - F(\widehat{Q}_{t,\beta}^{(g)} \mid \mathcal{E}_{t,w}^{(g)}) \right] \\ &= F(\widehat{Q}_{t,1-\alpha+\beta}^{(h)} \mid \mathcal{E}_{t,w}^{(h)}) - F(\widehat{Q}_{t,\beta}^{(h)} \mid \mathcal{E}_{t,w}^{(h)}). \end{aligned}$$

Now given $\varepsilon > 0$, let $\sigma_t > 0$ be such that

$$\begin{aligned}\Pr(Q^{(h)}_{t,1-\alpha+\beta} - \sigma_t < \hat{\epsilon}^{(h)}_t \leq Q^{(h)}_{t,1-\alpha+\beta}) < \varepsilon/2,\\ \Pr(Q^{(h)}_{t,\beta} < \hat{\epsilon}^{(h)}_t \leq Q^{(h)}_{t,\beta} + \sigma_t) < \varepsilon/2.\end{aligned}$$

By Proposition 2.3 and (Xu & Xie, 2022, Proposition 4.3), there exists T_0 such that for all $t > T_0$, we have $|\widehat{Q}^{(h)}_{t,p} - Q^{(h)}_{t,p}| < \sigma_t$. Since the cumulative distribution is non-decreasing, we have $F(Q_{t,p} - \sigma_t) \leq F(\widehat{Q}_{t,p}) \leq F(Q_{t,p} + \sigma_t)$. Hence, for $t > T_0$ we get

$$F(\widehat{Q}^{(h)}_{t,1-\alpha+\beta} \mid \mathcal{E}^{(h)}_{t,w}) - F(\widehat{Q}^{(h)}_{t,\beta} \mid \mathcal{E}^{(h)}_{t,w}) \geq F(Q^{(h)}_{t,1-\alpha+\beta} - \sigma_t \mid \mathcal{E}^{(h)}_{t,w}) - F(Q^{(h)}_{t,\beta} + \sigma_t \mid \mathcal{E}^{(h)}_{t,w}).$$

Now we have

$$\begin{aligned}F(Q^{(h)}_{t,1-\alpha+\beta} - \sigma_t \mid \mathcal{E}^{(h)}_{t,w}) &= F(Q^{(h)}_{t,1-\alpha+\beta} \mid \mathcal{E}^{(h)}_{t,w}) - \Pr(Q^{(h)}_{t,1-\alpha+\beta} - \sigma_t < \hat{\epsilon}^{(h)}_t \leq Q^{(h)}_{t,1-\alpha+\beta})\\ &> 1 - \alpha + \beta - \varepsilon/2,\\ F(Q^{(h)}_{t,\beta} + \sigma_t \mid \mathcal{E}^{(h)}_{t,w}) &= F(Q^{(h)}_{t,\beta} \mid \mathcal{E}^{(h)}_{t,w}) + \Pr(Q^{(h)}_{t,\beta} < \hat{\epsilon}^{(h)}_t \leq Q^{(h)}_{t,\beta} + \sigma_t)\\ &< \beta + \varepsilon/2.\end{aligned}$$

Therefore

$$F(Q^{(h)}_{t,1-\alpha+\beta} - \sigma_t \mid \mathcal{E}^{(h)}_{t,w}) - F(Q^{(h)}_{t,\beta} + \sigma_t \mid \mathcal{E}^{(h)}_{t,w}) > 1 - \alpha - \varepsilon.$$

□

A.2 Additional Results

Tables 6 and 7 give values for width mean and standard deviation of the various methods considered in Section 3.

Table 6: The mean and standard deviation of interval widths for COVID and EEG datasets. Up and down arrows indicate whether higher or lower values are preferred for that metric, respectively. The best results are bolded. Experiments are repeated over five seeds and we report the mean ± standard deviation across experiments. *Since TQA-E can create infinitely-wide intervals, these numbers are calculated by replacing infinite widths with 2x the maximum finite width (Lin et al., 2022).

Dataset	Metric	LPCI	TQA-B	TQA-E	CQR	Split
COVID	Mean ↓	0.980 ± 0.048	**0.755 ± 0.033**	1.070* ± 0.308	0.947 ± 0.069	0.930 ± 0.024
	St. dev. ↑	**0.526 ± 0.044**	–	–	0.372 ± 0.065	0.053 ± 0.029
EEG	Mean ↓	1.592 ± 0.116	1.315 ± 0.039	1.585* ± 0.111	**1.395 ± 0.084**	1.398 ± 0.067
	St. dev. ↑	**0.41 ± 0.046**	–	–	0.106 ± 0.014	0.053 ± 0.021

Table 7: LPCI mean and standard deviation of widths on the COVID dataset in the longitudinal experiment setup. Experiments are repeated over five seeds and we report the mean ± standard deviation across experiments.

Metric	LPCI	SPCI	CQR	Split
Mean width ↓	1.379 ± 0.004	0.703 ± 0.012	**1.05 ± 0.014**	1.124 ± 0.003
St. dev. of width ↑	**0.689 ± 0.012**	0.409 ± 0.005	0.30 ± 0.016	0.116 ± 0.002