International Journal of Forecasting 30 (2014) 217–234

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Empirical prediction intervals revisited

Yun Shin Lee [a,*], Stefan Scholtes [b]

[a] KAIST College of Business, Korea Advanced Institute of Science and Technology, 85 Hoegi-Ro, Dongdaemoon-Gu, Seoul 130-722, Republic of Korea

[b] Judge Business School, University of Cambridge, Trumpington Street, Cambridge, CB2 1AG, United Kingdom

ARTICLE INFO

Keywords:
Interval forecasting
Probabilistic forecasting
Out-of-sample forecast error
Model uncertainty
Non-Gaussian distribution

ABSTRACT

Empirical prediction intervals are constructed based on the distribution of previous out-of-sample forecast errors. Given historical data, a sample of such forecast errors is generated by successively applying a chosen point forecasting model to a sequence of fixed windows of past observations and recording the associated deviations of the model predictions from the actual observations out-of-sample. The suitable quantiles of the distribution of these forecast errors are then used along with the point forecast made by the selected model to construct an empirical prediction interval. This paper re-examines the properties of the empirical prediction interval. Specifically, we provide conditions for its asymptotic validity, evaluate its small sample performance and discuss its limitations.

1. Introduction

Prediction intervals are valuable complements to point forecasts, as they indicate the precision of the forecasts: future realizations will fall within a prediction interval with a prescribed probability. The problem of constructing prediction intervals has traditionally been studied using a theoretical (model-based) approach, which assumes that the applied forecasting model specifies the underlying stochastic process correctly and that the forecast errors follow a specific distribution (Chatfield, 1993). It is assumed that the chosen forecasting model makes unbiased point forecasts, i.e., the mean of the forecast error is zero. The variance of the forecast error is found using theoretical formulae derived from the chosen forecasting model (see for example Box, Jenkins, & Reinsel, 1994, for ARMA models). Although in principle other error distributions are also possible, it is often assumed that the error distribution is Gaussian, as this facilitates the derivation of theoretical formulae. It has long been known, however, that such theoretical prediction intervals tend to be too narrow if the forecasting model is misspecified, i.e., if the forecast errors have a non-zero mean or if the error distribution is non-normal, see e.g. Chatfield (1993, 1995). If there are doubts about model assumptions, empirically based approaches offer a useful alternative.

The literature on empirical approaches to estimating prediction intervals can be divided into two strands. The first strand has explored the use of *empirical residual errors*, in order to avoid assumptions regarding the spread and shape of the error distribution. They compute the residual errors of a fitted forecasting model at different forecast lead times and apply non-parametric methods, such as Chebyshev's inequality (Gardner, 1988) and kernel density estimators (Wu, 2010), and semi-parametric methods, such as quantile regression (Taylor & Bunn, 1999), to construct prediction intervals. Whilst these approaches relax assumptions on the spread and shape of the error distribution, they remain based on residual errors rather than out-of-sample forecast errors. It is well known, however, that true post-sample forecast errors tend to be larger than the fitted residuals (Makridakis & Winkler, 1989). The fitted residuals – the differences between the observed and fitted values in-sample – measure how well

* Corresponding author. Tel.: +82 2 958 3339; fax: +82 2 958 3359.
E-mail addresses: yunshin@business.kaist.ac.kr (Y.S. Lee), s.scholtes@jbs.cam.ac.uk (S. Scholtes).

0169-2070/$ – see front matter
http://dx.doi.org/10.1016/j.ijforecast.2013.07.018

the chosen model fits the data. Out-of-sample forecast errors – the differences between the realizations (which are not included in the fitting process) and the predictions of the model – indicate the chosen model's true predictive performance. They incorporate all causes of errors in the model predictions simultaneously, including random variations in the data-generating process, parameter estimation errors, and errors due to incorrect model specifications.

The second strand of the literature therefore employs *empirical out-of-sample forecast errors* to construct prediction intervals. This approach is based on the generation of a sample of out-of-sample forecast errors by fitting a chosen point forecasting model successively to a sequence of windows of past observations, and recording the associated deviations of the model predictions from the actual observations out-of-sample. Given a desired nominal coverage rate – the prespecified probability that the interval should contain future observations – the relevant quantiles of the distribution of these empirical forecast errors are used with the point forecasts made by the selected model to calculate an empirical prediction interval. This concept was introduced by Williams and Goodman (1971), and is increasingly applied as an alternative to traditional approaches (see e.g. Cohen, 1986; Jogensen & Sjoerg, 2003; Rayer, Smith, & Tayman, 2009; Isengildina-Massa, Irwin, Good, & Massa, 2011). However, little is known about the theoretical underpinnings of the approach, and some important questions remain unanswered: under what conditions is this empirical approach robust under model uncertainty? What is the finite sample performance of the approach? When is the approach preferable to the alternatives? The purpose of this paper is to focus on the empirical approach that uses out-of-sample forecast errors, and give this approach a full re-examination. Specifically, we consider two sources of model misspecification:

1. incorrect assumptions on the forecast error distribution;
2. incorrect assumptions on the functional form of the point forecasting model, leading to a biased point forecast;

and examine the robustness of the empirical approach against these two types of model uncertainty using asymptotic results, and simulation and empirical studies. We also discuss its limitations.

To illustrate the benefits of using out-of-sample forecast errors to construct prediction intervals, consider the process $Y_t = \mu + u_t$, where $u_t \sim N(0, \sigma_u^2)$. Suppose that the chosen point forecasting model is biased and produces one-step-ahead point forecasts at time t by $\hat{Y}_{t,1} = \hat{\mu}_t = \mu + b_t$, where $b_t \sim N(b, \sigma_b^2)$. This leads to out-of-sample forecast errors $E_{t,1} = Y_t - \hat{Y}_{t,1} = \mu - \hat{\mu}_t + u_t = -b_t + u_t$, and implies that $\mathrm{E}(E_{t,1}) = -b$ and $\mathrm{Var}(E_{t,1}) > \sigma_u^2$. Therefore, we can use the mean of the forecast error to re-center the prediction interval in order to correct for the forecast bias, and also use the larger variance of the forecast error to widen the interval so as to incorporate model uncertainty in addition to the true random variation u_t of the process.

Our asymptotic results show that when the data-generating process is *stationary ergodic*, the mean and variance of the out-of-sample forecast errors can be estimated consistently, and therefore the empirical prediction intervals have asymptotically correct coverages, regardless of the point forecasting model selected. Furthermore, the assumption of Gaussian errors can be avoided by applying the empirical quantiles of the forecast errors when calculating the interval endpoints. Therefore, empirical prediction intervals avoid the assumptions of a correctly specified forecasting model and Gaussian forecast errors. Since empirical prediction intervals are valid for arbitrary point forecasting models, their use also extends to forecasting models that include judgemental aspects that cannot be subsumed in the theoretical approach to estimating prediction intervals.

We evaluate the finite sample performance of the empirical prediction intervals using Monte Carlo experiments, and provide an empirical study of real exchange rate forecasts. The focus of the simulation and empirical studies is on an examination of the robustness of the approach in the face of model misspecification, in comparison with an alternative theoretical (model-based) approach and a purely non-parametric approach. Both simulation and empirical studies indicate that empirical prediction intervals are particularly robust for time series that are nearly non-stationary. In addition, given that the empirical approach relies on the generation of empirical forecast errors, it necessitates the availability of sufficient data. We find that the empirical prediction intervals for up to 10-step-ahead forecasts are fairly robust for sample sizes above 120.

The major limitation of the empirical approach is that the estimated intervals are not conditional on past observations or other predictors. If the point forecasting model contains predictors and produces biased conditional point forecasts, then the empirical approach will not produce asymptotically correct conditional intervals, as the approach widens the intervals by incorporating unconditional model uncertainty. This unconditional aspect of the approach does not cause its performance to deteriorate on average (Chatfield, 1993), but may lead to larger standard deviations of the interval estimates in practical situations, compared to alternative approaches that are conditional on previous observations. This points to a crucial trade-off in applications: the benefit of robustness against the unbiasedness of the point forecasting model must be traded off against the loss in efficiency resulting from the unconditional nature of the approach. However, if the point forecasting model employed is known to produce unbiased point forecasts conditional on predictors, the empirical approach will construct consistent conditional intervals as well.

This paper is organized as follows. In Section 2, we describe the main approaches for obtaining theoretical and empirical prediction intervals. Section 3 specifies assumptions for the asymptotic validity of the empirical approach. Section 4 contains a small-sample Monte Carlo study that compares the relative performances of the theoretical and empirical prediction intervals. An application to real data is presented in Section 5, and Section 6 provides a conclusion. The Appendix contains the main proof of the asymptotic analysis in Section 3.

2. Constructing prediction intervals: theoretical and empirical approaches

We consider a stochastic process $\{Z_t : \Omega \to \mathbb{R}^{s+1},\ s \in \mathbb{N},\ t = 1, 2, \ldots\}$ on a probability space $(\Omega, \mathcal{F}, P)$, and define the observed vector Z_t as $(Y_t, X_t')'$, where $\{Y_t : \Omega \to \mathbb{R}\}$ is the variable of interest and $\{X_t : \Omega \to \mathbb{R}^s\}$ is a vector of covariates. We let $\mathcal{F}_t$ be the filtration generated by $(Z_1', \ldots Z_t')'$.

Suppose that a forecasting model g is chosen for making τ-step-ahead point forecasts $\hat{Y}_{t,\tau} = g(Z_t, Z_{t-1}, \ldots, Z_{t-w+1})$ at time t. Here, w is a window size, i.e., the size of a subsample used to make point forecasts, and g is a measurable function. Note that this setup allows the incorporation of various different point forecasting models, including univariate, where $\hat{Y}_{t,\tau}$ may depend on $Y_t, Y_{t-1}, \ldots$; multivariate, where $\hat{Y}_{t,\tau}$ may also depend on covariates $X_t, X_{t-1}, \ldots$; and judgemental models, where $\hat{Y}_{t,\tau}$ is generated by expert judgment conditional on $\mathcal{F}_t$ (c.f. Giacomini & White, 2006).

Let $E_{t,\tau}$ be the out-of-sample forecast errors associated with the estimated point forecasts $\hat{Y}_{t,\tau}$,

$$E_{t,\tau} = Y_{t+\tau} - \hat{Y}_{t,\tau} = Y_{t+\tau} - g(Z_t, Z_{t-1}, \ldots, Z_{t-w+1}).$$

We assume that the forecast errors $E_{t,\tau}$ have an (unknown) cumulative distribution function $F_\tau(e) = \Pr(E_{t,\tau} \le e)$. The quantiles of the forecast error distribution, $Q_\tau(p) = \min\{e : F_\tau(e) \ge p\}$, are then used to compute a $100\alpha\%$ prediction interval around the point forecast $\hat{Y}_{t,\tau}$. Specifically, the interval endpoints for $Y_{t+\tau}$ are

$$[L_{t,\tau}, U_{t,\tau}] = [\hat{Y}_{t,\tau} + Q_\tau((1-\alpha)/2), \hat{Y}_{t,\tau} + Q_\tau((1+\alpha)/2)].$$

Since the true forecast error quantiles $Q_\tau(p)$ are unknown, they must be estimated in order to calculate the interval endpoints. Next, we describe the theoretical and empirical approaches to estimating Q_τ.

2.1. Theoretical approach

The prevalent theoretical approach constructs prediction intervals by assuming that the applied forecasting model is specified correctly for the underlying stochastic process and that the forecast errors are normally distributed. Specifically, the forecast errors are assumed to have a zero mean, and their variance is estimated based on the analytical formulae derived from the chosen forecasting model. Given a series of n realizations of Z_t, i.e., $\{z_t : t = 1, 2, \ldots, n\}$, and a chosen forecasting model g, denote the estimated point forecast and error variance for lead time τ by $\hat{y}_{n,\tau}$ and $\hat{\sigma}_\tau^2$, respectively. A theoretical $100\alpha\%$ prediction interval for $Y_{n+\tau}$ is then given by

$$[\hat{L}_{n,\tau}, \hat{U}_{n,\tau}] = [\hat{y}_{n,\tau} \pm z_{(1-\alpha)/2}\hat{\sigma}_\tau],$$

where $z_{(1-\alpha)/2} = \Phi^{-1}(\frac{1-\alpha}{2})$ and Φ is a standard normal distribution function.

Theoretical formulae for estimating the τ-step-ahead forecast error variance $\hat{\sigma}_\tau^2$ are available for many classes of models and are a function of the residual errors of the fitted model. For example, if the forecasting model is ARIMA, specified in the infinite-moving average form of $Y_t = u_t + \psi_1 u_{t-1} + \psi_2 u_{t-2} + \cdots$, $u_t \sim N(0, \sigma_u^2)$, it can be shown that $\hat{\sigma}_\tau^2 = \hat{\sigma}_u^2[1 + \hat{\psi}_1^2 + \hat{\psi}_2^2 + \cdots]$, where $\hat{\sigma}_u^2$ and $\hat{\psi}_i$ are the estimated residual error variance and model parameters at $t = n$.

It is important to note that the conditional validity of the theoretical approach requires the forecasting model to be specified correctly, with true parameter values. When the parameter values are estimated with errors, the conditional distribution of forecast errors may not be normally distributed, even for the Gaussian data-generating process, and also the conditional mean of the errors is equal not to zero but to the forecast bias (Phillips, 1979). Bootstrap approaches (e.g. Kim, 2001; Reeves, 2005; Stine, 1985; Thombs & Schucany, 1990) have been used to address this problem of parameter uncertainty. In the face of model misspecification, the situation becomes much worse. Theoretical prediction intervals and other model-dependent bootstrap intervals become asymptotically invalid, even unconditionally. Specifically, the intervals tend to be too narrow to encompass the required proportion of future observations (Chatfield, 1993, 1995).

2.2. Empirical approach

The empirical approach to prediction interval estimation does not assume that the chosen forecasting model is specified correctly. Instead, it is based on the empirical analysis of past forecast errors that would have been made by the chosen model. Empirical forecast errors are generated systematically by applying the chosen point forecasting model g iteratively to subsamples of past observations and recording the deviations of the forecasts from the known out-of-sample realizations. Given a series of n realizations, the process starts at $t = w < n - \tau$. At every time t with $w \le t \le n - \tau = l$, the τ-step-ahead point forecast $\hat{y}_{t,\tau}$ is calculated based on the last w observations. This gives rise to corresponding empirical forecast errors by

$$\hat{e}_{t,\tau} = y_{t+\tau} - \hat{y}_{t,\tau} = y_{t+\tau} - g(z_t, z_{t-1}, \ldots, z_{t-w+1}).$$

In this way, we obtain a sample of the $k = n - \tau - w + 1 = l - w + 1$ out-of-sample forecast errors. These empirical errors act as a proxy for the true post-sample forecast errors. It is important to note that a fixed window size of w is used to generate the empirical forecast errors. We will discuss the choice of window size in more detail in Section 3.

Given k sampled forecast errors $\{\hat{e}_{t,\tau} : t = w, w+1, \ldots, l\}$, denote the estimated $p\%$ forecast error quantile by $\hat{Q}_\tau(p)$. We consider both parametric and non-parametric approaches when estimating $\hat{Q}_\tau$. For the parametric approach, we assume that the τ-step-ahead forecast errors are normally distributed with finite mean μ_τ and variance σ_τ^2, and estimate the sample mean $\hat{\mu}_\tau = k^{-1}\sum_{t=w}^{l} \hat{e}_{t,\tau}$ and the sample variance $\hat{\sigma}_\tau^2 = k^{-1}\sum_{t=w}^{l}(\hat{e}_{t,\tau} - \hat{\mu}_{n,\tau})^2$. The parametric empirical (P-empirical) forecast error quantile is then calculated as $\hat{Q}_\tau(p) = \hat{\mu}_\tau + z_p\hat{\sigma}_\tau$, and the P-empirical prediction interval with the nominal $100\alpha\%$ coverage is

$$[\hat{L}_{n,\tau}, \hat{U}_{n,\tau}] = [\hat{y}_{n,\tau} + \hat{\mu}_\tau \pm z_{(1-\alpha)/2}\hat{\sigma}_\tau].$$

Both the theoretical and P-empirical intervals assume that the forecast errors are normally distributed. The main difference is that the P-empirical intervals are based on estimating the forecast error variance for a lead time τ from the sample variance of τ-step-ahead forecast errors directly, while the theoretical intervals are computed using theoretical formulae that are based on one-step-ahead forecast errors and the properties of the forecasting model, assuming that the latter is a correct specification of the underlying data-generating process.

The non-parametric approach to constructing empirical prediction intervals works with the empirical distribution of the generated forecast errors $\hat{F}_\tau(e) = k^{-1}\sum_{t=w}^{l}\mathbb{I}(\hat{e}_{t,\tau} \leq e)$ directly, where $\mathbb{I}(S)$ is the indicator function of a set S. Denote the rth order statistic of the k empirical forecast errors for a given lead time τ by $\hat{o}(r)_{k,\tau}$. The non-parametric empirical (NP-empirical) forecast error quantile is then $\hat{Q}_\tau(p) = \hat{o}(r)_{k,\tau}$, where $r = \lfloor kp \rfloor + 1$ and $\lfloor s \rfloor$ denotes the largest integer m such that $m \leq s$. Therefore, the NP-empirical prediction interval is given by

$$[\hat{L}_{n,\tau}, \hat{U}_{n,\tau}] = [\hat{y}_{n,\tau} + \hat{o}(r_L)_{k,\tau}, \hat{y}_{n,\tau} + \hat{o}(r_U)_{k,\tau}].$$

Here, $r_L = \lfloor k(1-\alpha)/2 \rfloor + 1$ and $r_U = \lfloor k(1+\alpha)/2 \rfloor + 1$.

3. Asymptotic justification for the empirical approach

Recall that the true interval endpoints for $Y_{n+\tau}$ with the nominal coverage $100\alpha\%$ are $[L_{n,\tau}, U_{n,\tau}] = [\hat{Y}_{n,\tau} + Q_\tau((1-\alpha)/2), \hat{Y}_{n,\tau} + Q_\tau((1+\alpha)/2)]$. Given the last realized values at $t = n$, $\mathbf{z}_n = (z_n, z_{n-1}, \ldots, z_{n-w+1})'$, the point forecast $\hat{y}_{n,\tau}$ is fixed. The large-sample validity of empirical prediction intervals therefore depends entirely on the limiting behavior of the forecast error quantile Q_τ. Since the forecast horizon $\tau \geq 1$ is fixed, we drop the subscript τ in this section.

A critical assumption for the consistency of both the parametric and non-parametric quantile estimates is the ergodic stationarity of the underlying data-generating process.

Assumption 1. The observed stochastic process Z_t is stationary ergodic.

Lemma 1. *If Assumption 1 holds and a forecasting model is given as $g(Z_t, Z_{t-1}, \ldots, Z_{t-w+1})$ with a fixed window size w, the forecast error E_t is also stationary ergodic.*

Assumption 2. The forecast errors $\{E_t\}$ have a finite mean $\mu = \mathrm{E}[E_t]$ and variance $\sigma^2 = \mathrm{E}[E_t - \mu]^2$, and a cumulative distribution function of the form $F(e) = \Phi(e, \mu, \sigma^2)$, where Φ is a standard normal distribution function.

Lemma 2. *If a point forecasting model is misspecified such that $\mathrm{E}(Y|Z) \neq g(Z)$, the mean of forecast errors $\mu = \mathrm{E}[E_t]$ is not zero and is equal to the forecast bias. The variance $\sigma^2 = \mathrm{E}[E_t - \mu]^2$ is larger than the variance of forecast errors when an unbiased forecasting model is used. Furthermore, if Assumptions 1 and 2 hold, we have $\hat{\mu}_n \xrightarrow{a.s} \mu$ and $\hat{\sigma}_n^2 \xrightarrow{a.s} \sigma^2$.*

Theorem 1. *If Assumptions 1 and 2 hold and $p \in (0, 1)$, then the parametric sample quantile satisfies $\hat{Q}(p) \xrightarrow{a.s} Q(p) = F^{-1}(p)$.*

Lemma 2 implies that when the point forecasting model is misspecified, the mean forecast error measures the bias in the point forecasts, which can be used to re-center the prediction interval. At the same time, the variance of the forecast errors becomes larger than the variance of the true random variations of the underlying process, which leads to the associated prediction intervals being wider, in order to accommodate additional errors due to the biased point forecasting model. By estimating the mean and variance of the forecast error consistently, we obtain asymptotically valid parametric empirical intervals. We note that Theorem 1 is not restricted to the case where the forecast errors have a normal distribution but can be generalized, by the continuous mapping theorem, to any distribution that is continuous in its first and second moments, such as the exponential distribution.

The non-parametric approach drops the Gaussian error assumption and requires only mild conditions on the forecast error distribution, namely continuity and bounded density.

Assumption 3. The cumulative distribution function $F(e) = \mathrm{E}[\mathbb{I}(E_t \leq e)]$ of the forecast errors is continuously differentiable, with a positive and finite density $f(e) = F'(e)$ in the neighborhood of $Q(p) = F^{-1}(p)$.

Theorem 2. *If Assumptions 1 and 3 hold and $p \in (0, 1)$, then the non-parametric sample quantile satisfies $\hat{Q}(p) \xrightarrow{a.s} Q(p) = F^{-1}(p)$.*

The above two theorems provide conditions under which the quantile of the forecast error distribution associated with a chosen forecasting model is estimated consistently. As was discussed above, with consistent quantile estimates, one can calculate asymptotically correct interval endpoints. It is remarkable that the assumptions do not include a direct assumption on the fitted model g, but only an assumption on the true data-generating process. If the specified assumptions are satisfied, empirical prediction intervals with *any* point forecasting model will have a correct coverage on average as $n \to \infty$, and are therefore robust under model uncertainty; it is not necessary to make any assumptions on either the predictors or the predicted. This is in contrast to the critical importance of a correct model specification for alternative model-based approaches, in order to achieve asymptotically correct coverage.

Note that Assumption 1 holds for a wide range of time series models (e.g., ARMA models). The assumption suggests that the observed data need to be made stationary prior to applying the empirical approach, e.g., through appropriate deseasonalizing and differencing, in order to achieve an asymptotically correct coverage of the empirical prediction intervals. Note also that the normality assumption of forecast errors in Assumption 2 may be true asymptotically when a Gaussian data-generating model is identified correctly (Chatfield, 1993). However, the normality assumption is often invalid in practical applications. In this case, the use of non-parametric sample quantiles results in more robust prediction intervals.

As was noted earlier, it is critical for the asymptotic validity of the empirical approach that limited memory

predictors be used. Specifically, we use a rolling scheme which fixes the size of the fitting sample to w and drops distant observations as more recent ones are added (e.g. Giacomini & White, 2006). Expanding memory predictors, such as a recursive scheme that uses all of the available data at time t, are not permitted. The empirical forecast errors collected from the recursive scheme are not stationary, and cannot be used to estimate the unknown properties of the underlying forecast errors consistently. Take the example of estimating the forecast error distribution of a correctly specified model, but with parameter uncertainty. When the forecast errors are sampled using the recursive scheme, the portion of the forecast error that is due to the parameter estimation error reduces over time, as more and more data are used to fit the model. Therefore, the parameter estimation error cannot be estimated consistently using empirical forecast errors of expanding memory predictors. In contrast, limited memory estimators with a fixed window size generate asymptotically nonvanishing estimation errors and do not suffer from these inconsistencies. Consequently, the resulting forecast errors are stationary (Lemma 1) and their quantiles can be estimated consistently. Note that the forecast errors generated by the empirical approach are serially correlated. This serial correlation, however, does not affect the consistency of the quantile estimates.

It is also important to note that, despite its robustness, the asymptotic validity of the empirical approach is shown unconditional on the last realized values, $\mathbf{z}_n = (z_n, z_{n-1}, \ldots, z_{n-w+1})'$. The empirical intervals are unconditional because both the forecast error distribution and the forecast error quantile are estimated unconditionally. This unconditionality of the approach does not cause its performance to deteriorate on average, as is indicated by Theorems 1 and 2, but will lead to the interval estimates having a larger standard deviation than the conditional interval estimates. There is a clear trade-off between the benefit of robustness against misspecifying a point forecasting model on the one hand and the benefit of conditionality when the model is correct on the other. In order to analyze the asymptotic validity of the intervals conditional on predictors, we have to assume that the correct conditioning function and set of predictors are known (i.e., $\mathrm{E}(Y|Z) = g(Z)$). Under this assumption, the empirical approach can be shown to deliver asymptotically correct intervals conditional on the predictors (see the Appendix).

Next, we show the asymptotic normality result for the non-parametric sample quantile of forecast errors. For this, we impose stronger mixing conditions on the memory of the observed stochastic process Z_t. Definitions of ϕ-mixing and α-mixing can be found in Appendix B.

Assumption 4. The observed stochastic process Z_t is stationary and either (i) ϕ-mixing such that $\phi(m) = O(m^{-2})$ as $n \to \infty$, or (ii) α-mixing where there exists a $\Delta(> 0)$ such that $\alpha(m) = O(m^{-(5/2)-\Delta})$.

Note that, under general conditions, finite autoregressive moving average (ARMA) processes have exponentially decaying memories, and therefore satisfy Assumption 4. Also, we define

$$\begin{aligned}\nu^2 &= \lim_{k\to\infty} \{k\mathrm{Var}\, \hat{F}_k(Q(p))\} \\ &= \lim_{k\to\infty} k\mathrm{Var}\{[\mathbb{I}(E_w \le Q(p)) + \cdots + \mathbb{I}(E_l \le Q(p))]/k\}.\end{aligned}$$

Then, we make the following final assumption on ν^2.

Assumption 5. $0 < \nu^2 < \infty$.

Theorem 3. *If Assumptions 3–5 hold and $p \in (0, 1)$, then the non-parametric sample quantile satisfies*

$$\frac{n^{1/2} f(Q(p))}{\nu} (\hat{Q}(p) - Q(p)) \to^D N(0, 1)$$

as $n \to \infty$.

4. Monte Carlo analysis

We design simulation experiments in order to provide an empirical illustration of the asymptotic theory. We illustrate that the conditions for the large-sample validity of the empirical approach do not include a direct assumption on the fitted forecasting model. Also, for the NP-empirical prediction intervals, no specific parametric assumption on the error distribution is required. To test this, we designed our simulation experiments to investigate the validity of the empirical intervals when facing two sources of model uncertainty. The first source of model uncertainty comes from the distribution of the forecast errors, assuming a correct point forecasting model (Section 4.1), while the second source is the specification of the point forecasting model itself (Section 4.2). We report the performance of the empirical approach for different values of the sample size, to illustrate its asymptotic validity. Here, we illustrate the empirical approach using pure time series models. It is important, however, to remember that the empirical approach applies to any arbitrary point forecasting mechanism, including multivariate models and judgemental forecasts.

The simulation experiment is designed as follows. We first assume that the underlying data-generating process is described by the following ARMA models:

Model 1. $Y_t = 0.85Y_{t-1} + u_t$
Model 2. $Y_t = 0.75Y_{t-1} - 0.40Y_{t-2} + 0.20Y_{t-3} + u_t$
Model 3. $Y_t = 0.75Y_{t-1} + u_t - 0.20u_{t-1}$,

where u_t has mean zero and standard deviation $\sigma_u = 1$.

For each simulation run, we generate a single series of $n = 120$ consecutive observations using the assumed data-generating model. Based on the generated sample, $100\alpha\%$ prediction intervals are calculated for each lead time τ. Specifically, we use the fixed window size $w = 30$ and generate a sample of $k = n - \tau - w + 1 = 91 - \tau$ out-of-sample forecast errors. Based on these empirical forecast errors, parametric and non-parametric empirical prediction intervals are calculated for a chosen coverage percentage $100\alpha\%$. 1000 realized out-of-sample observations are generated for each post-sample period, $n + 1, n + 2, \ldots, n + \tau$, conditional on the last n observations. The realizations for $t = n + \tau$ are then compared with each estimated prediction interval, with

a lead time of τ, to calculate its coverage rate, i.e., the frequency with which the prediction interval contains out-of-sample realizations. Ideally, the intervals should have a $100\alpha\%$ coverage rate; deviations from $100\alpha\%$ indicate inaccurate interval estimates. We repeat this for 1000 simulation runs and report the average coverage rate and standard error (se). We consider the nominal coverage rates $\alpha = 0.80$ and 0.95, and lead times τ ranging from 1 to 10. For brevity, we only report the results for $\alpha = 0.80$, as the results associated with $\alpha = 0.95$ provide qualitatively similar results. Similarly, we only report the results for lead times $\tau = 1, 3, 5$, and 10.

We use two benchmarks for our study: theoretical prediction intervals and purely non-parametric prediction intervals. The theoretical approach constructs prediction intervals as if the fitted model with the Gaussian assumption would describe the true data-generating model fully, and ignores model uncertainty. This theoretical interval provides a reference point for investigating the effect of ignoring model uncertainty over coverage accuracy, in comparison with the empirical prediction intervals, which are designed to account for this uncertainty, at least asymptotically. The purely non-parametric prediction intervals are obtained by calculating the quantiles of the empirical distribution of Y_t itself. Under a general dependence assumption on the underlying process, these intervals are known to be consistent (Yoshihara, 1995). However, the non-parametric approach is only applicable for constructing one-step-ahead prediction intervals, and is problematic for multi-period horizons. All of the computations are conducted using the R statistical package (version 2.13.1). The R code used in this study can be provided on request.

4.1. The case of making an incorrect distributional assumption

In order to investigate the effect of non-Gaussian error distributions on the coverage rate, we consider three alternative error distributions for each data-generating model (Models 1–3), namely Gaussian, exponential, and a contaminated normal distribution $0.9F_1 + 0.1F_2$, where $F_1 \sim N(-1, 1)$ and $F_2 \sim N(9, 1)$. Each distribution has been centered to have a zero mean. These distributions represent the ideal, skewed, and bimodal skewed alternatives, respectively. The Gaussian distribution has been chosen as a benchmark for purposes of comparison, given that the theoretical and P-empirical intervals have been derived under this assumption. In this section, we assume that the functional form of the point forecasting model is identified correctly, and focus on the impact of an incorrect assumption on the coverage rate.

Table 1 suggests that both theoretical and empirical prediction intervals underestimate the nominal coverage rate, even when the data-generating model is identified correctly and the forecast errors are normally distributed. Both the theoretical and empirical prediction intervals are too narrow, because they ignore the uncertainty in parameter estimation. For the theoretical intervals, the coverage is particularly underestimated when the data-generating model is an AR(1) model. This results from the large bias in autoregressive estimation in small samples, especially for highly autocorrelated processes (Phillips, 1979). The coverage of the P-empirical intervals is also underestimated, because the mean and variance of τ-step-ahead forecast errors are estimated with some errors in a finite sample. It is also evident that the P-empirical intervals outperform the NP-empirical intervals. This is because, under Gaussian forecast errors, the NP-empirical intervals require a larger sample of empirical forecast errors in order to estimate quantiles non-parametrically for a given accuracy level, particularly for extreme quantiles near 0 or 1. Since both approaches are asymptotically optimal under this condition, we observe improvements as the sample size increases.

When the error distribution is not Gaussian but is exponential or mixed, the theoretical and P-empirical intervals that assume Gaussian errors tend to estimate coverages which are greater than the nominal coverage, especially for shorter lead times. The degree of this tendency illustrates the effect on the error distribution of making incorrect assumptions. In contrast, the NP-empirical intervals avoid making any distributional assumption, and their average coverages are therefore less sensitive to the choice of the error distribution. More importantly, unlike theoretical and P-empirical intervals, the coverage of the NP-empirical intervals becomes closer to the nominal value as the sample size increases (see Fig. 1).

We also tested the intervals built from the empirical distribution of Y_t itself. This purely non-parametric approach also avoids a distributional assumption, and thus its coverage rate is robust to the choice of the error distribution. When the underlying process is highly correlated (for example Model 1), we find that the NP-empirical intervals outperform the non-parametric intervals. This is due to the use of a point forecast as a center point of the empirical prediction interval, which can capture autocorrelation in the underlying process. Unlike the NP-empirical intervals, the non-parametric approach requires a larger sample size to obtain a reasonable degree of precision when the underlying process is highly positively correlated (see Fig. 1).

The performances of the empirical prediction intervals depend on the accuracy of the estimated point forecast and the accuracy of the sample quantile estimate of the forecast errors, both of which are determined by the chosen window length w. Recall that the number of empirical forecast errors generated is $k = n - \tau - w + 1$. Therefore, there is a tradeoff to be made between the accuracy of the sample quantiles, which improves with a smaller w, and the accuracy of the estimated point forecast made at $t = n$, which improves with a larger w. The appropriate window length will naturally depend on the sample size. Based on the simulation setups used in Table 1, we find that $20 \leq w \leq 30$ gives a credible coverage accuracy when $n = 120$ (see Fig. 2).

4.2. The case of using an incorrect point forecasting model

We design seven Monte Carlo experiments in order to consider particular cases of model misspecification resulting from the incorrectly identified structure of the point forecasting model. In order to isolate the effect of model uncertainty due to employing an incorrect point

Table 1
Percentage coverage by 80% prediction intervals for Models 1–3 with three alternative error distributions and a correct point forecasting model ($n = 120$).

Model	Distribution	Lead	*Theoretical*		*P-empirical*		NP-*empirical*		*Non-parametric*	
			Average	(se)	Average	(se)	Average	(se)	Average	(se)
1	Normal	1	78.73	(0.11)	79.15	(0.19)	77.90	(0.21)	75.61	(0.82)
		3	76.68	(0.15)	77.41	(0.31)	76.44	(0.33)	75.02	(0.47)
		5	75.37	(0.19)	75.97	(0.39)	75.10	(0.40)	74.70	(0.33)
		10	73.96	(0.23)	73.56	(0.50)	72.99	(0.51)	74.16	(0.27)
	Exponential	1	89.14	(0.11)	86.70	(0.25)	78.10	(0.43)	76.02	(0.86)
		3	81.22	(0.21)	80.43	(0.37)	76.68	(0.40)	75.41	(0.51)
		5	78.26	(0.23)	78.21	(0.42)	75.67	(0.45)	74.94	(0.37)
		10	76.07	(0.25)	75.74	(0.53)	73.48	(0.54)	74.39	(0.28)
	Mixture	1	89.72	(0.05)	88.00	(0.28)	76.69	(0.55)	74.97	(0.93)
		3	82.21	(0.21)	80.54	(0.42)	75.72	(0.52)	74.69	(0.57)
		5	81.62	(0.24)	79.02	(0.47)	74.90	(0.53)	74.28	(0.43)
		10	77.47	(0.26)	76.75	(0.56)	73.51	(0.56)	74.00	(0.30)
2	Normal	1	77.94	(0.12)	79.22	(0.25)	78.01	(0.26)	78.19	(0.41)
		3	77.48	(0.14)	78.21	(0.29)	76.96	(0.31)	77.87	(0.15)
		5	77.96	(0.13)	78.34	(0.24)	77.12	(0.26)	77.61	(0.15)
		10	78.09	(0.13)	77.59	(0.27)	76.23	(0.29)	77.65	(0.15)
	Exponential	1	88.18	(0.15)	85.71	(0.37)	77.77	(0.49)	77.04	(0.58)
		3	83.82	(0.18)	82.41	(0.40)	77.17	(0.45)	77.71	(0.18)
		5	84.59	(0.17)	82.63	(0.36)	76.96	(0.41)	77.65	(0.17)
		10	84.57	(0.16)	82.32	(0.36)	76.65	(0.40)	77.60	(0.15)
	Mixture	1	89.49	(0.06)	86.87	(0.38)	76.79	(0.66)	79.42	(0.48)
		3	83.21	(0.12)	81.74	(0.42)	75.73	(0.61)	77.94	(0.18)
		5	83.70	(0.10)	82.70	(0.38)	75.48	(0.56)	77.75	(0.17)
		10	83.93	(0.10)	82.97	(0.41)	74.66	(0.57)	77.73	(0.15)
3	Normal	1	78.44	(0.11)	79.75	(0.19)	78.55	(0.21)	77.65	(0.45)
		3	77.56	(0.14)	79.03	(0.22)	77.97	(0.24)	77.37	(0.23)
		5	77.08	(0.15)	78.34	(0.25)	77.61	(0.26)	77.13	(0.19)
		10	76.96	(0.16)	77.12	(0.30)	76.23	(0.32)	77.03	(0.18)
	Exponential	1	88.56	(0.12)	86.29	(0.30)	77.15	(0.45)	74.13	(0.65)
		3	82.75	(0.19)	81.53	(0.31)	76.85	(0.36)	74.42	(0.26)
		5	80.97	(0.19)	80.11	(0.32)	76.26	(0.37)	74.62	(0.13)
		10	80.32	(0.19)	78.34	(0.36)	74.95	(0.41)	74.75	(0.05)
	Mixture	1	89.42	(0.08)	87.68	(0.29)	76.99	(0.57)	73.41	(0.71)
		3	85.09	(0.14)	83.14	(0.31)	76.53	(0.50)	72.60	(0.26)
		5	83.75	(0.15)	81.62	(0.34)	76.10	(0.48)	72.47	(0.15)
		10	83.16	(0.15)	80.16	(0.37)	75.11	(0.49)	72.50	(0.06)

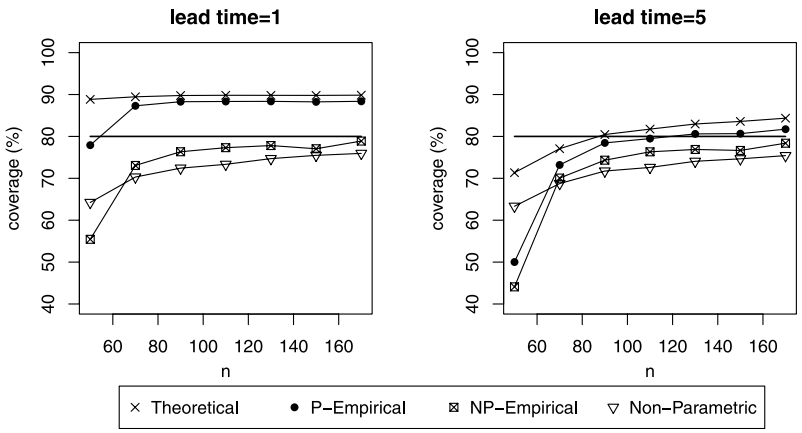

Fig. 1. Impact of sample size n on the percentage coverage by 80% prediction intervals when the data-generating model is described by Model 1 with a contaminated error distribution. The solid horizontal line indicates the nominal coverage rate of 80%.

forecasting model from that due to assuming an incorrect error distribution for the theoretical and P-empirical intervals, we assume Gaussian forecast errors. Our experimental setting is summarized in Table 2. Experiments 1 and 2 correspond to model misspecification cases where the lag orders of the autoregressive models are identified incorrectly. Experiments 3 and 4 illustrate cases where a unit root in the underlying process is assessed incorrectly. Experiments 5 and 6 are based on two common point forecasting methods, the simple moving average and

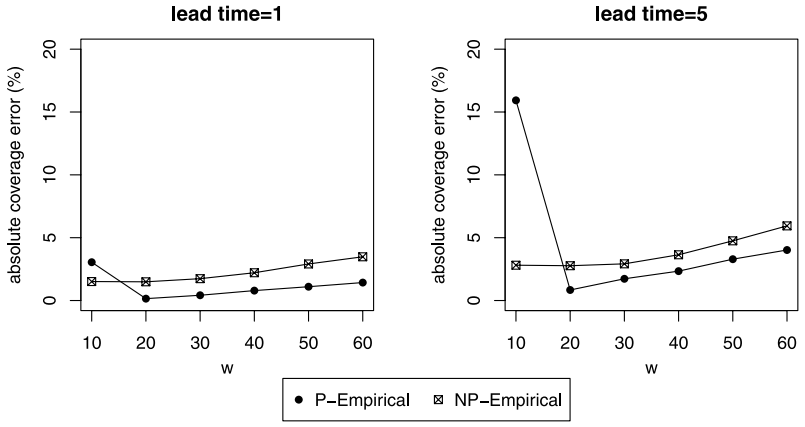

Fig. 2. Impact of window size w on the absolute percentage coverage error made by 80% prediction intervals when the data-generating model is described by Model 2 with a Gaussian error distribution.

Table 2
Monte Carlo experiment setups for the case of using an incorrect point forecasting model. The underlying process is assumed to be Gaussian.

Experiment	True data-generating model		Selected point forecasting model
1	$Y_t = 0.85Y_{t-1} + u_t$	AR(1)	AR(3)
2	$Y_t = 0.75Y_{t-1} - 0.40Y_{t-2} + 0.20Y_{t-3} + u_t$	AR(3)	AR(1)
3	$Y_t = 0.85Y_{t-1} + u_t$	AR(1)	Random walk
4	$Y_t = Y_{t-1} + u_t$	IMA(1, 0)	AR(1)
5	$Y_t = 0.75Y_{t-1} + u_t - 0.20u_{t-1}$	ARMA(1, 1)	Moving average
6	$Y_t = 0.75Y_{t-1} + u_t - 0.20u_{t-1}$	ARMA(1, 1)	Exponential smoothing
7	$Y_t = 0.85Y_{t-1} + u_t$	AR(1)	AR(1)+1

exponential smoothing, which misspecify the underlying data-generating process. Experiment 7 introduces a deterministic bias equal to the standard deviation of the residual error to the estimated AR(1) model.

Before we report and discuss the performances of the estimated prediction intervals, we illustrate that the assumed model misspecifications are not unlikely to occur. To this end, we use the specified data-generating model to generate 1000 samples of length $n = 120$. We then apply common model selection methods to choose a model specification for each of the generated series, and calculate the probability of specifying the model correctly over 1000 simulated series.

For Experiments 1 and 2, we use the Akaike Information Criterion (AIC) to determine the appropriate lag order q of an AR(q) model. The Monte Carlo simulations estimate that the probabilities of identifying the correct models, AR(1) for Experiment 1 and AR(3) for Experiment 2, are only 72% and 53% respectively.

To assess the likelihood of model misspecification for Experiments 3 and 4, we employ the Dickey–Fuller test (Dickey & Fuller, 1979) at a 5% significance level. This unit root test is known to have a low statistical power over stable autoregressive alternatives, particularly with roots near unity (DeJong, Nankervis, Savin, & Whiteman, 1992; Diebold & Rudebusch, 1991). For the stationary data-generating model assumed in Experiment 3, the unit root test falsely fails to reject the null hypothesis of a unit root in approximately 70% of cases. For the non-stationary data-generating process in Experiment 4, the test rejects the hypothesis of a unit root with a 5% probability, which is consistent with the chosen significance level.

Makridakis, Wheelwright, and Hyndman (1998) report that simple forecasting methods, such as moving average and exponential smoothing, applied in Experiments 5 and 6, are used most frequently in practice, often without a justification of their suitability, due to their ease of use and flexibility. Moving averages and exponential smoothing, however, are optimal only when the underlying processes are i.i.d. and ARIMA(0, 1, 1), respectively. The use of these models for other data-generating processes amounts to model misspecification. Note that Experiment 6 could also represent the case of misspecification of a unit root.

Table 3 compares the average coverage rates of the theoretical and empirical prediction intervals. The simulation results suggest that, in the case of using the incorrect point forecasting model, the empirical prediction intervals are more robust than the theoretical intervals, in the sense that estimation outliers of coverage rates are rare. In particular, the empirical prediction intervals are robust to the misspecification of a unit root in the data series (Experiments 3, 4 and 6), for which the correct specification between a stationary and a non-stationary model within the framework of the theoretical approach is critical (Chatfield, 1993). In this case, the size of the improvement generated by the empirical approach generally increases as the lead time increases. Also, when a deterministic bias is introduced to a point forecasting model, as in Experiment 7, the empirical approach re-centers the prediction interval by consistently estimating this bias using the mean of forecast errors (Lemma 2). The purely non-parametric approach is

Table 3
Percentage coverage by 80% prediction intervals for Experiments 1–7, which illustrate the case where a misspecified point forecasting model is used and the underlying process is Gaussian, as summarized in Table 2 ($n = 120$).

Experiment	Lead	*Theoretical*		P-*empirical*		NP-*empirical*		*Non-parametric*	
		Average	(se)	Average	(se)	Average	(se)	Average	(se)
1	1	78.05	(0.12)	79.11	(0.27)	77.68	(0.29)	76.89	(0.79)
	3	76.63	(0.17)	77.25	(0.40)	76.10	(0.42)	75.88	(0.45)
	5	75.48	(0.20)	76.23	(0.46)	75.15	(0.47)	75.23	(0.32)
	10	74.26	(0.23)	75.33	(0.56)	73.80	(0.56)	74.56	(0.26)
2	1	78.48	(0.16)	79.40	(0.24)	78.13	(0.27)	78.36	(0.43)
	3	77.29	(0.13)	78.41	(0.23)	77.35	(0.25)	77.91	(0.16)
	5	78.08	(0.13)	78.18	(0.22)	77.05	(0.25)	77.83	(0.15)
	10	77.97	(0.13)	77.53	(0.24)	76.43	(0.26)	77.75	(0.16)
3	1	79.62	(0.13)	79.70	(0.14)	78.23	(0.18)	76.54	(0.79)
	3	82.94	(0.22)	79.24	(0.27)	78.15	(0.29)	75.89	(0.46)
	5	85.91	(0.28)	78.69	(0.37)	77.96	(0.38)	75.52	(0.33)
	10	91.56	(0.31)	76.84	(0.52)	76.34	(0.54)	74.98	(0.26)
4	1	78.32	(0.11)	78.82	(0.19)	77.54	(0.21)	57.85	(1.29)
	3	75.45	(0.15)	76.84	(0.31)	75.63	(0.34)	56.83	(1.10)
	5	72.91	(0.20)	75.35	(0.39)	74.29	(0.40)	55.73	(0.98)
	10	67.45	(0.29)	72.84	(0.54)	71.22	(0.54)	53.16	(0.79)
5	1	78.11	(0.22)	83.88	(0.30)	82.94	(0.32)	78.03	(0.43)
	3	74.76	(0.20)	77.63	(0.31)	76.74	(0.33)	77.47	(0.22)
	5	74.63	(0.16)	77.10	(0.29)	76.33	(0.31)	77.37	(0.19)
	10	74.35	(0.15)	75.89	(0.30)	75.19	(0.33)	77.05	(0.19)
6	1	79.16	(0.15)	79.92	(0.20)	78.60	(0.23)	78.57	(0.39)
	3	79.28	(0.24)	80.00	(0.27)	78.88	(0.28)	77.41	(0.21)
	5	82.00	(0.30)	79.70	(0.34)	78.90	(0.35)	76.91	(0.18)
	10	88.97	(0.30)	78.68	(0.43)	77.77	(0.44)	76.68	(0.18)
7	1	59.45	(0.19)	79.28	(0.18)	78.15	(0.20)	76.37	(0.78)
	3	68.04	(0.22)	78.01	(0.30)	76.82	(0.31)	75.70	(0.44)
	5	68.84	(0.25)	76.86	(0.37)	75.86	(0.38)	75.30	(0.32)
	10	68.70	(0.28)	74.70	(0.48)	74.13	(0.49)	74.70	(0.26)

based on using the empirical distribution of Y_t, and is thus robust to model misspecification. However, as was noted earlier, its performance deteriorates when the underlying process is highly autocorrelated (see for example Experiments 3 and 4). Similar results were obtained for 95% prediction intervals, except that the superiority of the empirical approach is reduced somewhat.

Fig. 3 shows a comparison of the average length of the estimated prediction intervals with the true interval length. The true interval length is calculated using the true data-generating model in Table 2, and is appropriate for quantifying the levels of the underlying random variations. We find that the empirical approach constructs wider prediction intervals than the true intervals. These wider intervals are necessary for the accommodation of errors resulting from model misspecification, in addition to random variations, which alone determine the width of the true interval. The empirical approach constructs systematically widened intervals based on empirical forecast errors that are collected out-of-sample and have a larger variance than the underlying data generating process (Lemma 2). It therefore incorporates all causes of errors in the model predictions. The larger the effect of uncertainty in the point forecasting model, the wider the intervals generated by the empirical approach. Due to their asymptotic properties, the length of the empirical intervals becomes just sufficient to cover the desired future realizations on average as the sample size increases (see Fig. 4). The purely non-parametric approach is robust, but is only applicable for one-step-ahead forecasts. As a result, the width of the non-parametric intervals remains constant as the lead time changes.

Regarding the theoretical approach, Fig. 3 shows that the corresponding prediction intervals can be either wider or narrower than the true interval. This is a result of ignoring the model uncertainty and making false assumptions about the forecast error variances using the theoretical error variance formulae of a fitted model. Specifically, the use of a stationary model when the underlying process is indeed stationary (Experiments 1, 2 and 5) tends to result in overly narrow intervals because it ignores the additional uncertainty arising from model misspecification. However, if non-stationary models, such as random walks and exponential smoothing, are fitted to an underlying stationary process (Experiments 3 and 6), the theoretical prediction intervals become too wide and critically overestimate the nominal coverage rates, particularly for longer lead times (see Table 3). This is because non-stationary forecasting models theoretically assume that the variance of the forecast errors increases linearly with the lead time, whereas the underlying stationary processes have gradually increasing forecast error variances to a finite upper bound. In contrast, fitting a stationary model to a non-stationary process (Experiment 4) underestimates the future uncertainty and results in intervals that are too narrow. Being able to distinguish between stationary and non-stationary models is therefore critical for the theoretical approach. It is important to note

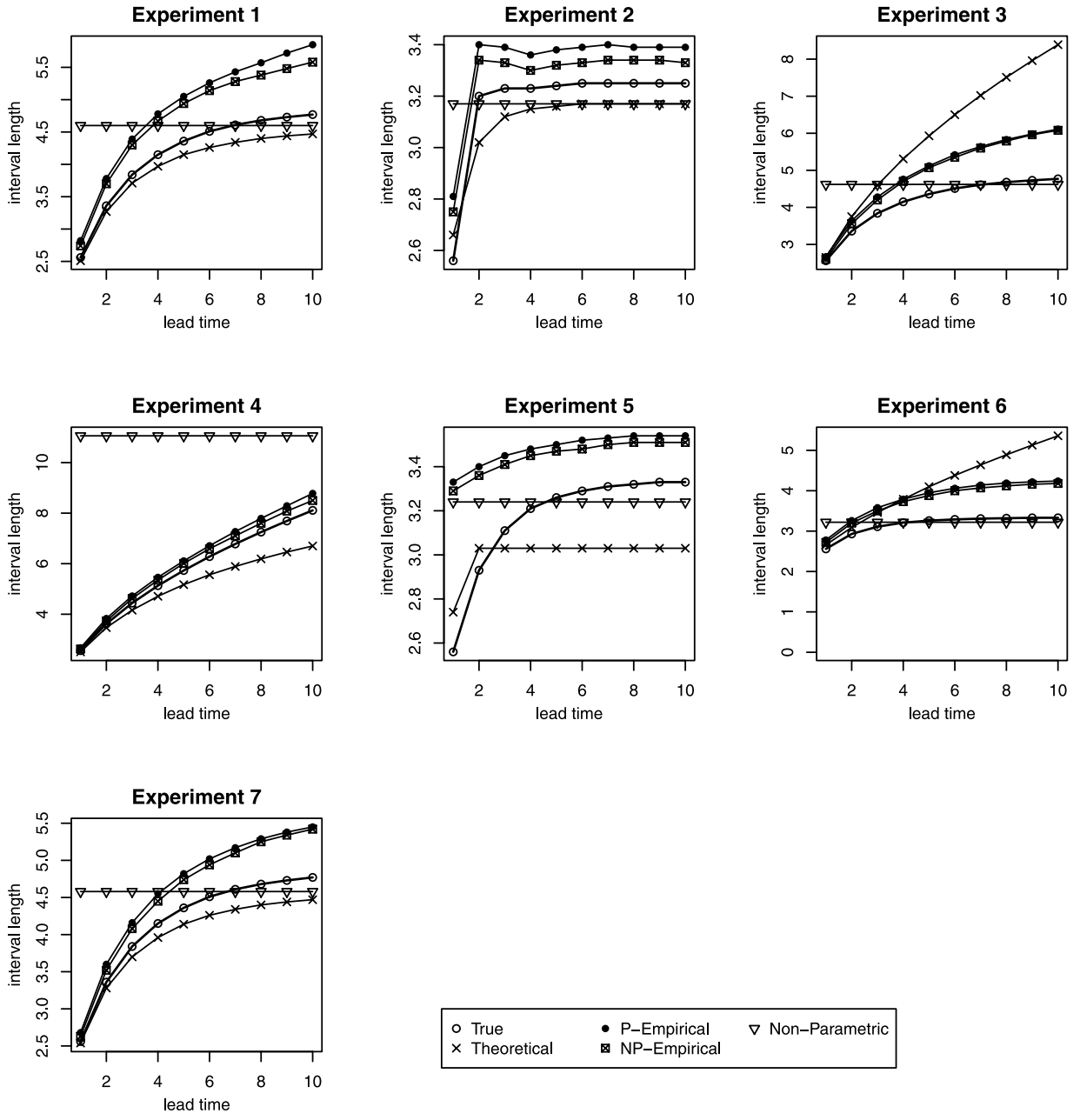

Fig. 3. Average interval length of 80% prediction intervals for Experiments 1–7, summarized in Table 2.

that model uncertainty does not always lead to a lower-than-nominal coverage of the theoretical intervals, as has previously been reported in the literature.

Fig. 4 examines the coverage rate of the empirical approach as a function of the sample size under model uncertainty and compares it with the theoretical approach. It demonstrates, as would be expected from the asymptotic results of Section 3, that the average coverage rates of the empirical intervals converge to the nominal coverage as the sample size increases, even when the point forecasting model is misspecified. In contrast, the coverage rates of the theoretical approach using an incorrect model do not improve with larger sample sizes.

5. An empirical example

In this section, we examine the robustness of the empirical prediction intervals, using real time series. We focus on the case of model misspecification resulting from uncertainty about the presence of a unit root in the process, for which the empirical approach was particularly useful in the simulation study. The time series we consider contain real exchange rates, as there is long-standing academic debate about whether or not such series contain a unit root.

The study of real exchange rates is concerned with the relative price of a basket of goods across countries,

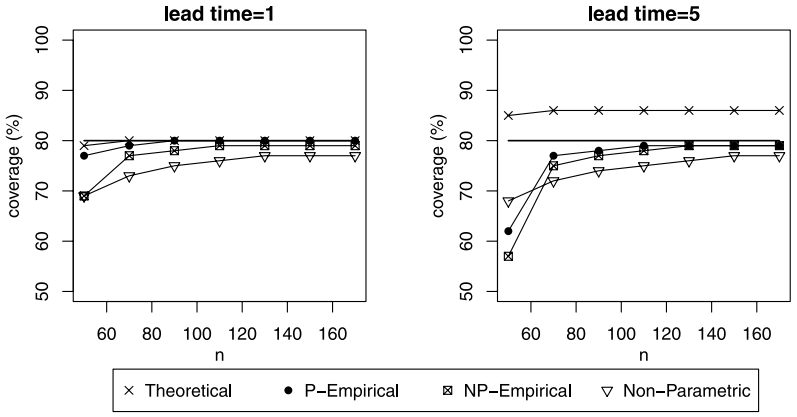

Fig. 4. Impact of the sample size n on the percentage coverage by 80% prediction intervals for Experiment 3 in Table 2. The solid horizontal line indicates a nominal coverage rate of 80%.

expressed in a common currency (Imbs, Mumtaz, Ravn, & Rey, 2005). These rates are important for the prices of tradable goods, foreign-exchange futures or options, and portfolios of international assets. An extensive body of literature in this field studies the validity of the theory of purchasing power parity, which posits that there is a real exchange rate equilibrium. If the theory is valid, real exchange rates are mean-reverting, stationary processes. Taylor (2006) provides a review of the literature on long-run purchasing power parity and the stability of real exchange rates. He notes that the idea of the validity of purchasing power parity has been highly controversial over the past three decades. One strand of the literature confirms the validity of the purchasing power parity condition by rejecting the null hypothesis of a unit root, and argues that a simple stationary autoregressive process, like AR(1), describes the behavior of real exchange rates accurately. In contrast, a second strand finds evidence of a unit root in the process and argues that shocks to real exchange rates accumulate, and rates will not exhibit mean reversion behavior. They find that a simple random walk can both fit and predict real exchange rates satisfactorily.

There is therefore a natural degree of uncertainty about appropriate forecasting models for real exchange rates: both AR(1) and random walk models appear to be sensible choices. These two models are, however, very different in their implied forecast error variances. As we discussed in Section 4.2, a stationary AR(1) model leads to a forecast error variance that converges to a finite upper bound as the lead time increases, whereas a non-stationary random walk implies a linearly increasing forecast error variance without bounds. Consequently, theoretical prediction intervals using an incorrect model assumption may be too wide or too narrow and have inaccurate coverage rates, particularly for longer lead times. This is therefore an ideal setting for the empirical approach, as it produces robust prediction intervals independently of the assumed forecasting model. To illustrate this robustness, we conduct experiments on the coverage accuracies of estimated prediction intervals with both an AR(1) and a random walk as point forecasting models. As in the simulation study, we use theoretical prediction intervals and purely non-parametric prediction intervals as benchmarks.

We select nine real exchange rate series between the US dollar and the currencies of Canada, Japan, Norway, Switzerland, the United Kingdom, France, Italy, The Netherlands and Spain. The raw data are monthly time series of nominal exchange rates and consumer price indices obtained from the IMF's International Financial Statistics (series AE and 64, May 2010 edition). All variables are transformed to logarithms. Real exchange rates at time t are computed as $q_t = e_t - p_t + p_t^*$, where e_t is the log nominal exchange rate, expressed as the domestic price of one unit of foreign currency (US \$), and p_t and p_t^* are the logarithms of the consumer price index of the domestic and foreign (US) country, respectively (Taylor, 2006). Fig. 5 shows the nine real exchange rates series we studied. The real exchange rate series of Canada, Japan, Norway, Switzerland and the UK have 446 monthly observations (January 1974–February 2010), and those of France, Italy, The Netherlands and Spain have 312 monthly observations (January 1974–December 1998).

We set the sample size $n = 120$, window size $w = 30$, lead times $\tau = 1, 2, \ldots, 10$, and nominal coverage $\alpha = 0.80$ and 0.95, as in the simulation study. We tested various different window sizes, and found that the preferred window size lies between 20 and 30, as in the simulation study. For each real exchange rate series, we use the first 120 months of the series to construct prediction intervals for each month of the following 10-month period. The actual observations over this 10-month period are then compared to the estimated prediction intervals, giving one coverage sample. We then roll forward one month and use the next 120 observations to compute the next set of prediction intervals over the following 10-month period. Again, the realizations are compared with the prediction intervals to obtain a second coverage sample. This process continues until we reach the end of the data set. In total, for each forecast horizon, from 1 to 10, we have obtained 317 prediction intervals for the data for Canada, Japan, Norway, Switzerland and the UK, and 183 prediction intervals for the data for France, Italy, The Netherlands and Spain. The average coverage rate is then estimated as the proportion of iterations for which the observed value lies within the correspondingly constructed prediction intervals.

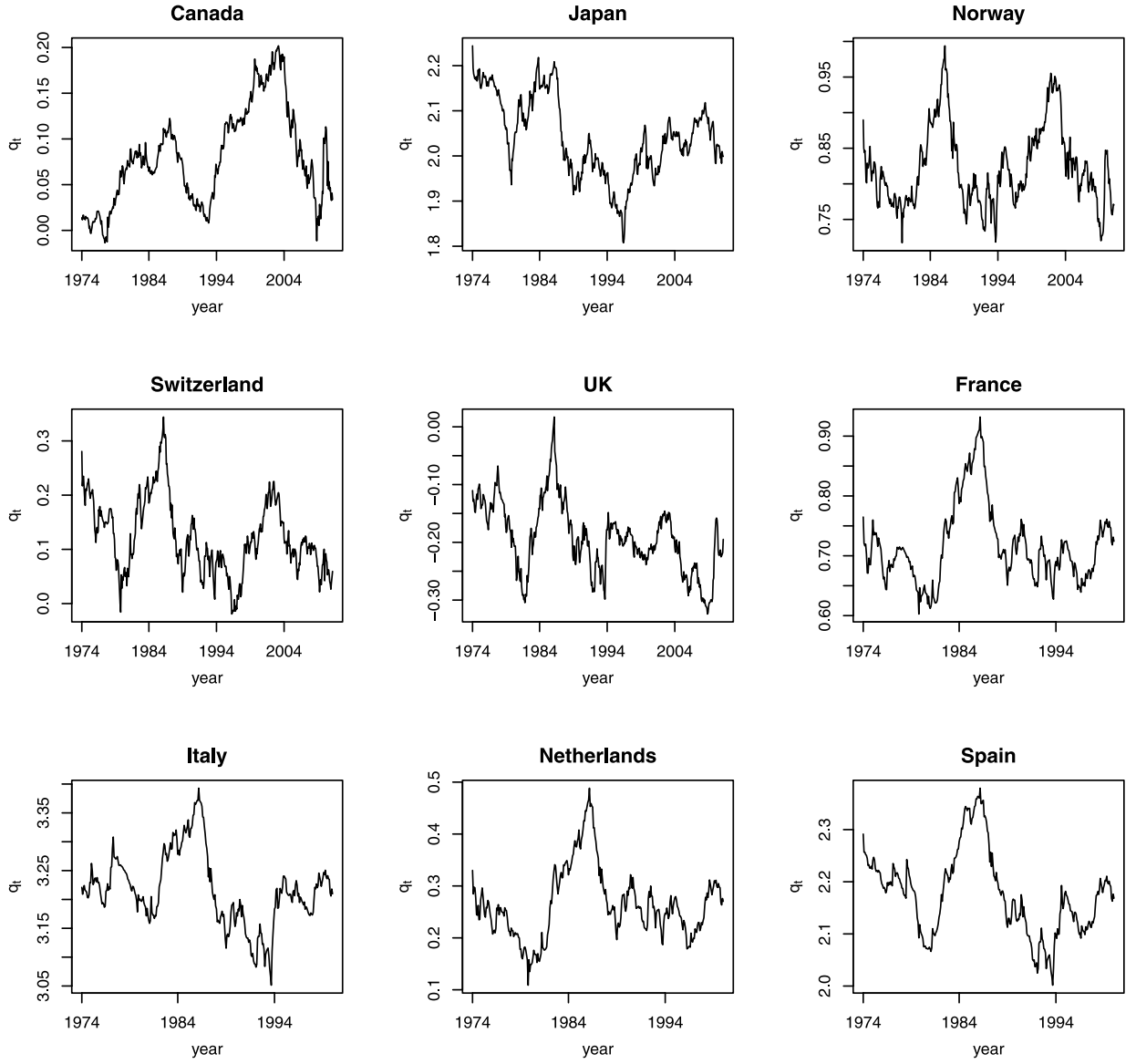

Fig. 5. Monthly real exchange rates series between the US dollar and the currencies of Canada, Japan, Norway, Switzerland, the United Kingdom, France, Italy, The Netherlands and Spain.

Figs. 6 and 7 compare the theoretical and empirical prediction intervals when the chosen point forecasting models are an AR(1) and a random walk, respectively. The figures suggest that the empirical prediction intervals are remarkably robust to the choice of forecasting model. In contrast, the performances of the theoretical prediction intervals depend heavily on the model chosen, with the performances of any one model varying significantly across countries. When one relies on the theory of purchasing power parity and uses a theoretical approach with an AR(1) forecasting model to construct prediction intervals, the estimated intervals can be too narrow to quantify the future uncertainty. It is also evident that the average coverages of the P-empirical and NP-empirical are similar, which indicates that the Gaussian assumption on forecast errors is not the major source of model uncertainty for forecasting real exchange rates. The purely non-parametric approach performs poorly in this context as a consequence of the near non-stationarity of the real exchange rate series. Similar comments apply to the 95% prediction intervals, except that the superiority of the empirical approach is not quite as pronounced.

The test of correct conditional coverage of Christoffersen (1998) is also calculated. The Christoffersen test for correct conditional coverage is a combination of the tests for unconditional coverage and independence. Table 4 reports the average coverages and p-values of the Christoffersen test for one-step-ahead prediction intervals when an AR(1) model is the chosen point forecasting model. We find that the average coverage rate of the empirical intervals is close to the nominal value. However, there are a few occasions where the Christoffersen test rejects the null of

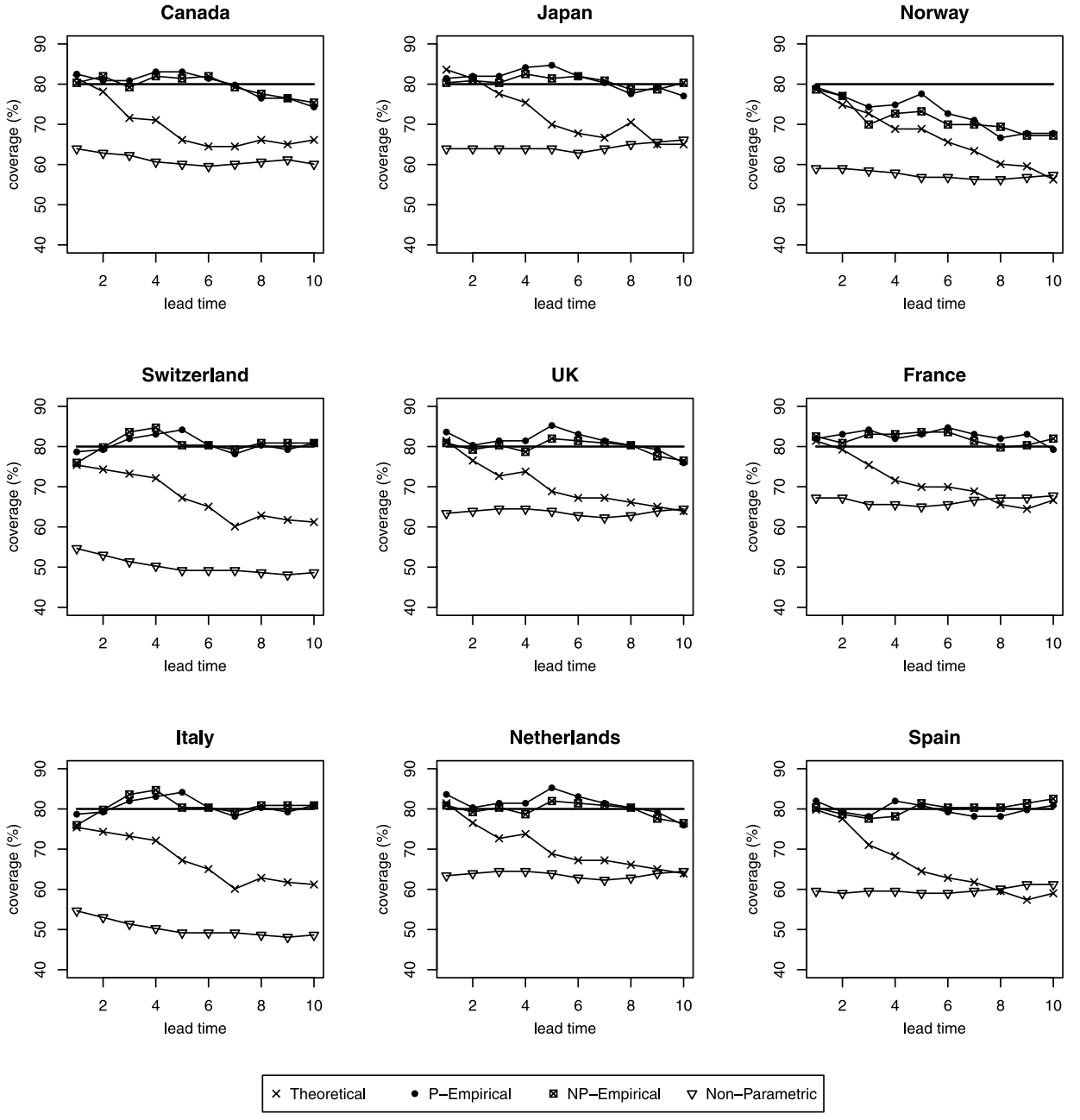

Fig. 6. Percentage coverage by 80% prediction intervals using an AR(1) model as a point forecasting model for real exchange rates series.

correct conditional coverage at the 5% significance level. This illustrates a trade-off between the improved average and the reduced conditional performance of the empirical approach. Similar conclusions can be drawn when the random walk is the chosen point forecasting model.

6. Conclusion

This paper investigates the robustness of using an empirical approach to construct prediction intervals. The empirical approach is based on the generation of a sample of empirical forecast errors, based on moving a fixed time window over the data, predicting on the basis of the data within that window, and collecting out-of-sample prediction errors at the desired lead time. This sequence of forecast errors directly captures not only random variation in the data-generating process, but also uncertainty in parameter estimation, as well as, importantly, any uncertainty associated with a point forecasting model. If the distribution of these forecast errors is used to construct prediction intervals in the face of parameter and model uncertainty, we can systematically widen the interval

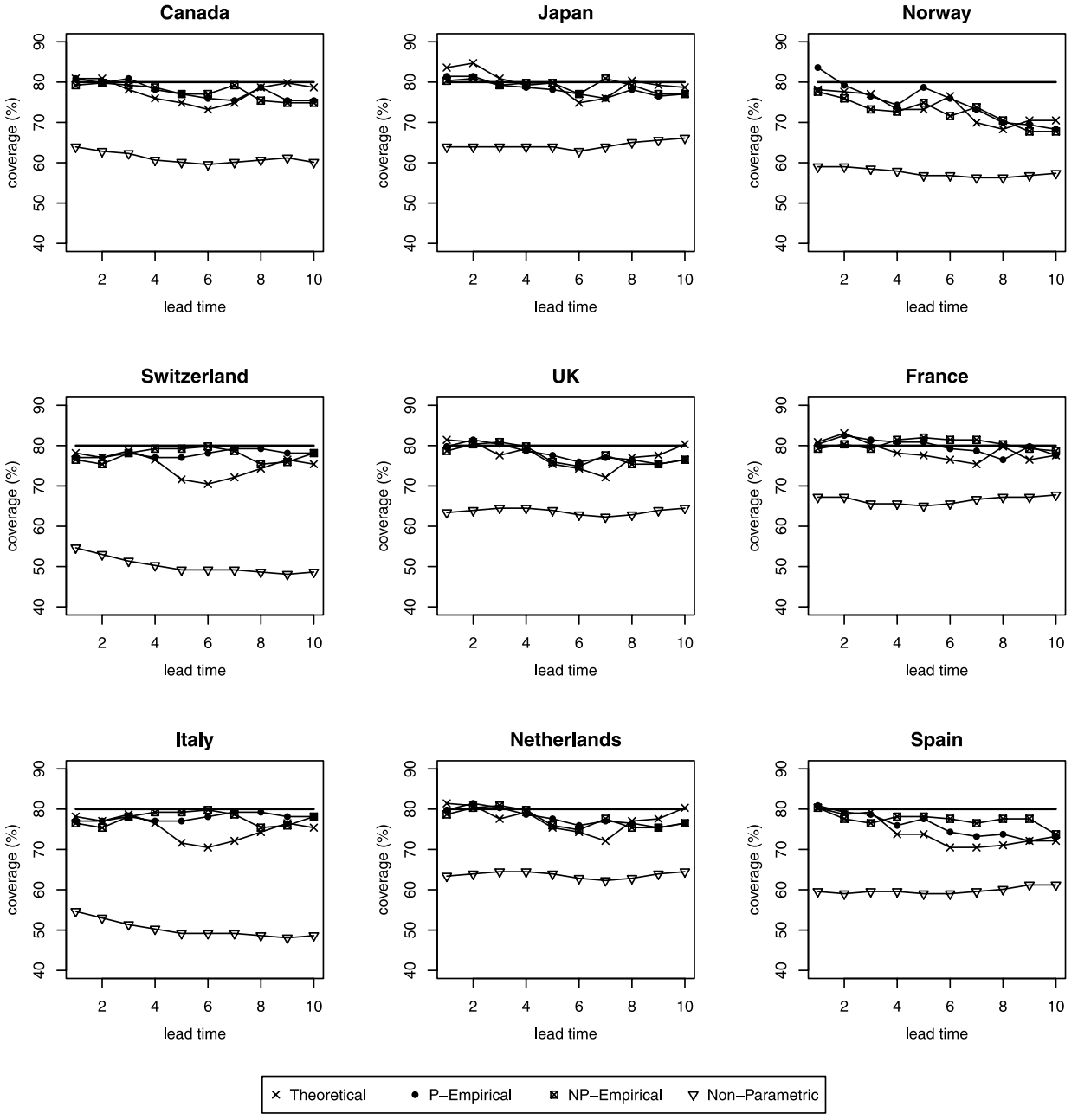

Fig. 7. Percentage coverage by 80% prediction intervals using a random walk model as a point forecasting model for real exchange rates series.

widths that lead to correct coverage rates. The asymptotic analysis suggests that, for a data-generating process that is stationary ergodic, the empirical prediction intervals can perform credibly with an arbitrary point forecasting model (i.e., the model may be misspecified), and is therefore robust to model uncertainty. Simulations and empirical studies confirm this claim.

In most real-life scenarios, it is difficult to characterize the underlying process completely. The empirical prediction intervals are therefore useful in practice for the probabilistic forecasting of economic and business time series. We find that the empirical approach is particularly beneficial for near unit root processes. The empirical approach is also particularly useful in reality, as it is widely applicable various point-forecasting models, including the prevalent moving averages and exponential smoothing models, and also judgemental forecasting models, for which no theoretical formulae for constructing prediction intervals are available.

Two caveats are required, however. First, in order to obtain an adequate coverage accuracy, the observed data need to be made stationary, through appropriate differencing or deseasonalizing, prior to applying the empirical approach. Second, an adequate coverage performance of

Table 4
Percentage coverage and p-value of the Christoffersen test for 80% prediction intervals when an AR(1) model is the chosen point forecasting model and $\tau = 1$.

Country	Approach	Desired coverage	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
Canada	Theoretical	Coverage	0.08	0.18	0.27	0.38	0.49	0.60	0.67	0.76	0.85	0.89
		p-value	0.30	0.40	0.20	0.43	0.78	0.91	0.21	0.06	**0.00**	**0.00**
	P-empirical	Coverage	0.12	0.18	0.28	0.40	0.48	0.59	0.68	0.78	0.87	0.90
		p-value	0.32	0.48	0.43	0.91	0.41	0.65	0.37	0.27	**0.04**	**0.00**
	NP-empirical	Coverage	0.10	0.19	0.27	0.35	0.44	0.55	0.65	0.76	0.87	0.91
		p-value	0.78	0.78	0.30	0.07	**0.03**	0.06	**0.05**	0.10	0.06	**0.00**
	Non-parametric	Coverage	0.04	0.10	0.15	0.18	0.20	0.22	0.27	0.38	0.53	0.64
		p-value	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Japan	Theoretical	Coverage	0.12	0.21	0.34	0.45	0.55	0.62	0.75	0.83	0.91	0.95
		p-value	0.24	0.58	0.11	0.07	0.07	0.43	**0.04**	0.16	0.51	0.95
	P-empirical	Coverage	0.12	0.24	0.36	0.45	0.54	0.64	0.75	0.83	0.92	0.96
		p-value	0.18	0.06	**0.04**	0.06	0.20	0.14	0.07	0.12	0.30	0.56
	NP-empirical	Coverage	0.09	0.20	0.30	0.41	0.51	0.61	0.73	0.84	0.91	0.95
		p-value	0.40	0.89	0.86	0.73	0.70	0.65	0.30	0.09	0.64	0.75
	Non-parametric	Coverage	0.04	0.09	0.18	0.28	0.36	0.44	0.48	0.58	0.75	0.83
		p-value	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Norway	Theoretical	Coverage	0.11	0.24	0.35	0.45	0.56	0.64	0.71	0.78	0.88	0.93
		p-value	0.78	0.13	0.08	0.06	**0.03**	0.11	0.67	0.34	0.24	0.10
	P-empirical	Coverage	0.12	0.24	0.35	0.47	0.58	0.67	0.74	0.82	0.88	0.93
		p-value	0.24	0.08	**0.05**	**0.02**	**0.00**	**0.01**	0.13	0.48	0.32	0.10
	NP-empirical	Coverage	0.09	0.19	0.29	0.40	0.48	0.58	0.71	0.82	0.88	0.94
		p-value	0.51	0.68	0.58	0.82	0.54	0.50	0.76	0.40	0.32	0.36
	Non-parametric	Coverage	0.04	0.13	0.24	0.33	0.42	0.47	0.52	0.60	0.70	0.76
		p-value	**0.00**	**0.00**	**0.02**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Switzerland	Theoretical	Coverage	0.10	0.19	0.30	0.41	0.50	0.58	0.69	0.82	0.93	0.97
		p-value	0.78	0.78	0.95	0.65	0.96	0.43	0.67	0.32	**0.04**	0.09
	P-empirical	Coverage	0.10	0.21	0.32	0.42	0.53	0.60	0.72	0.83	0.94	0.96
		p-value	0.78	0.58	0.51	0.50	0.35	0.82	0.36	0.16	**0.02**	0.39
	NP-empirical	Coverage	0.11	0.20	0.29	0.39	0.51	0.62	0.74	0.83	0.91	0.93
		p-value	0.52	1.00	0.67	0.65	0.87	0.57	0.16	0.20	0.40	0.16
	Non-parametric	Coverage	0.05	0.13	0.17	0.26	0.36	0.44	0.53	0.64	0.74	0.81
		p-value	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
UK	Theoretical	Coverage	0.13	0.25	0.36	0.46	0.54	0.65	0.74	0.83	0.88	0.94
		p-value	0.09	**0.04**	**0.02**	**0.02**	0.13	**0.05**	0.13	0.12	0.32	0.66
	P-empirical	Coverage	0.10	0.25	0.35	0.44	0.55	0.65	0.75	0.83	0.90	0.93
		p-value	0.93	**0.04**	**0.05**	0.11	0.07	**0.05**	**0.03**	0.16	0.78	0.16
	NP-empirical	Coverage	0.11	0.19	0.30	0.37	0.50	0.61	0.71	0.83	0.90	0.93
		p-value	0.52	0.78	0.95	0.36	0.87	0.65	0.67	0.12	0.78	0.10
	Non-parametric	Coverage	0.12	0.24	0.35	0.41	0.48	0.54	0.61	0.68	0.79	0.83
		p-value	0.18	0.10	**0.05**	0.65	0.54	**0.02**	**0.00**	**0.00**	**0.00**	**0.00**
France	Theoretical	Coverage	0.13	0.24	0.35	0.46	0.56	0.65	0.71	0.82	0.92	0.95
		p-value	0.17	0.17	0.13	0.09	0.13	0.21	0.84	0.56	0.44	0.88
	P-empirical	Coverage	0.12	0.21	0.36	0.46	0.58	0.66	0.74	0.83	0.94	0.96
		p-value	0.36	0.75	**0.05**	0.06	**0.02**	0.12	0.24	0.34	0.07	0.60
	NP-empirical	Coverage	0.10	0.17	0.26	0.34	0.48	0.60	0.72	0.84	0.94	0.95
		p-value	0.98	0.25	0.24	0.09	0.51	0.93	0.49	0.13	0.07	0.88
	Non-parametric	Coverage	0.08	0.20	0.27	0.35	0.43	0.48	0.58	0.66	0.75	0.82
		p-value	0.44	0.97	0.31	0.16	0.07	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Italy	Theoretical	Coverage	0.14	0.27	0.38	0.49	0.56	0.64	0.70	0.76	0.89	0.95
		p-value	0.07	**0.02**	**0.02**	**0.01**	0.10	0.34	0.91	0.17	0.49	0.85
	P-empirical	Coverage	0.09	0.25	0.39	0.48	0.55	0.62	0.72	0.81	0.91	0.95
		p-value	0.61	0.09	**0.01**	**0.02**	0.13	0.51	0.60	0.69	0.61	0.85
	NP-empirical	Coverage	0.11	0.19	0.31	0.39	0.52	0.62	0.68	0.82	0.91	0.95
		p-value	0.49	0.69	0.79	0.72	0.61	0.51	0.56	0.44	0.79	0.85
	Non-parametric	Coverage	0.05	0.10	0.16	0.29	0.33	0.42	0.45	0.55	0.69	0.74
		p-value	**0.02**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Netherlands	Theoretical	Coverage	0.11	0.19	0.34	0.44	0.54	0.65	0.72	0.82	0.90	0.95
		p-value	0.49	0.83	0.23	0.33	0.28	0.21	0.60	0.56	0.83	0.85
	P-empirical	Coverage	0.09	0.20	0.31	0.46	0.52	0.62	0.77	0.84	0.91	0.95
		p-value	0.79	0.97	0.67	0.12	0.72	0.51	**0.03**	0.13	0.61	0.85
	NP-empirical	Coverage	0.09	0.17	0.29	0.36	0.48	0.58	0.72	0.83	0.91	0.95
		p-value	0.61	0.34	0.72	0.21	0.61	0.60	0.49	0.25	0.61	0.85
	Non-parametric	Coverage	0.10	0.21	0.27	0.36	0.42	0.51	0.58	0.63	0.77	0.81
		p-value	0.98	0.75	0.40	0.21	**0.02**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**

(*continued on next page*)

Table 4 (*continued*)

Country	Approach	Desired coverage	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
Spain	Theoretical	Coverage	0.16	0.27	0.39	0.49	0.58	0.66	0.74	0.80	0.87	0.92
		p-value	**0.01**	**0.03**	**0.01**	**0.01**	**0.02**	0.12	0.18	0.97	0.17	0.09
	P-empirical	Coverage	0.14	0.24	0.34	0.46	0.58	0.67	0.78	0.83	0.88	0.93
		p-value	0.07	0.17	0.23	0.09	**0.04**	**0.05**	**0.01**	0.34	0.36	0.28
	NP-empirical	Coverage	0.07	0.17	0.28	0.37	0.47	0.60	0.72	0.82	0.92	0.93
		p-value	0.20	0.25	0.60	0.42	0.35	0.93	0.49	0.44	0.44	0.28
	Non-parametric	Coverage	0.11	0.17	0.26	0.35	0.44	0.53	0.56	0.59	0.70	0.76
		p-value	0.65	0.25	0.24	0.16	0.10	**0.03**	**0.00**	**0.00**	**0.00**	**0.00**

Note: boldface indicates that the null hypothesis of correct conditional coverage is rejected at the 5% significance level.

empirical intervals requires appropriate sample sizes. We find that a sample of at least 120 observations is generally required in order to collect a sufficient number of empirical forecast errors for up to 10 forecast horizons. In view of the present rate of technological development, we feel that the computational intensity of the approach is unlikely to be a significant constraint.

The major limitation of empirical prediction intervals lies in the fact that they are not conditional on past observations. The unconditionality of the approach does not generally cause its performance to deteriorate, as is indicated by the asymptotic and simulation results; however, it will lead to larger standard deviations of the interval estimates than alternative approaches that compute conditional prediction intervals given the recent state of the system. Future work should extend the empirical approach to incorporate the conditionality of forecast errors. Exponential weighting schemes, such as those discussed by Taylor (2007), and the conditional autoregressive value at risk of Engle and Manganelli (2004), could possibly be applied to empirical forecast errors in order to make the quantile estimation adaptive and conditional.

Acknowledgments

This work was partially supported by a Mellon Sawyer Dissertation Fellowship in "Modeling Futures: Understanding Risk and Uncertainty", Center for Research in the Arts, Social Sciences and Humanities (CRASSH), University of Cambridge. The authors thank Philip Dawid, Paul Kattuman, David Spiegelhalter, James Taylor and Danny Ralph for helpful discussions and for their comments. The authors would also like to thank the Handling Editor, the Associate Editor and two anonymous referees for helpful comments and suggestions. Any remaining errors are our responsibility.

Appendix A. Asymptotic justification

Proof of Lemma 1. Note that $E_{t,\tau} = Y_{t+\tau} - \hat{Y}_{t,\tau} = Y_{t+\tau} - g(Z_t, Z_{t-1}, \ldots, Z_{t-w+1}) = h(Z_{t+\tau}, Z_t, Z_{t-1}, \ldots, Z_{t-w+1})$, where $h(.)$ is a measurable function and w is fixed. Therefore, it follows from Stout (1974, pp. 182–183) that if Z_t is stationary ergodic, E_t is also stationary ergodic. □

Proof of Lemma 2. Let an unbiased forecasting model $g^*(Z_t) = \mathrm{E}(Y_t|Z_t)$ and let u_t be the forecast error (residual error) of this model, i.e., $Y_t = g^*(Z_t) + u_t$ and $\mathrm{E}(u_t) = 0$. Then, we can write the resulting forecast errors of the chosen forecasting model $g(Z_t) \neq g^*(Z_t)$ as $E_t = Y_t - \hat{Y}_t = g^*(Z_t) - g(Z_t) + u_t$. Therefore, $\mu = \mathrm{E}(E_t) = g^*(Z_t) - g(Z_t) \neq 0$. Also, $\sigma^2 = \mathrm{Var}(E_t) = \mathrm{Var}(g^*(Z_t)) + \mathrm{Var}(g(Z_t)) + \mathrm{Cov}(g^*(Z_t), g(Z_t)) + \mathrm{Cov}(g(Z_t), u_t) + \mathrm{Var}(u_t) > \mathrm{Var}(u_t)$.

To see $\hat{\mu}_n \xrightarrow{a.s} \mu$, note that $\{E_t\}$ is stationary ergodic by Lemma 1. Assumption 2 states that $\mathrm{E}|E_t| < \infty$, and thus $k^{-1}\sum_{t=w}^{l} \hat{e}_t \xrightarrow{a.s} \mu$ follows from the ergodic theorem as in Theorem 3.34 of White (2001).

To show that $\hat{\sigma}_n^2 \xrightarrow{a.s} \sigma^2$, we expand $\hat{\sigma}_n^2 = k^{-1}\sum_{t=w}^{l}(\hat{e}_t - \hat{\mu}_n)^2 = k^{-1}\sum_{t=w}^{l}\hat{e}_t^2 - \hat{\mu}_n^2$. We have already shown that $\hat{\mu}_n \xrightarrow{a.s} \mu$, and hence $\hat{\mu}_n^2 \xrightarrow{a.s} \mu^2$ by the continuous mapping theorem. It remains to be shown that $k^{-1}\sum_{t=w}^{l}\hat{e}_t^2 \xrightarrow{a.s} \mathrm{E}(E_t^2)$. Since $\{E_t\}$ is stationary ergodic, $\{E_t^2\}$ is also stationary ergodic by Stout (1974, pp. 182–183). In view of Assumption 2, the ergodic theorem implies $k^{-1}\sum_{t=w}^{l}\hat{e}_t^2 \xrightarrow{a.s} \mathrm{E}(E_t^2)$, which completes the proof. □

Proof of Theorem 1. In view of Lemma 2, this is a direct result of the continuous mapping theorem. □

Proof of Theorem 2. Fix p with $0 < p < 1$. We will show that for every $\varepsilon > 0$, there exists $N = N(\varepsilon)$ such that, for all $n > N$, $|\hat{Q}_n(p) - Q(p)| < \varepsilon$. It will suffice to show this for a sufficiently small ε. By a Taylor expansion,

$$F(Q(p) + \varepsilon) = F(Q(p)) + \varepsilon f(Q(p)) + o(\varepsilon). \tag{1}$$

Since $f(Q(p)) > 0$ by Assumption 3, we may assume that ε is sufficiently small to guarantee that

$$\varepsilon f(Q(p)) + o(\varepsilon) > 0. \tag{2}$$

Next, we show that $\hat{F}_n(e) - F(e) = R_n(e)$, where $R_n(e)$ converges to zero almost surely. By Lemma 1, $\{E_t\}$ is stationary ergodic and $\{\mathbb{I}(E_t \leq e)\}$ is also stationary ergodic. Because $\mathrm{E}|\mathbb{I}(E_t \leq e)| < \infty$, the ergodic theorem implies $k^{-1}\sum_{t=w}^{l}\mathbb{I}(\hat{e}_t \leq e) \xrightarrow{a.s} \mathrm{E}(\mathbb{I}(E_t \leq e))$. Thus, we have

$$\hat{F}_n(e) - F(e) = R_n(e) \xrightarrow{a.s} 0. \tag{3}$$

If we replace $F(Q(p) + \varepsilon)$ by $\hat{F}_n(Q(p) + \varepsilon) - R_n(Q(p) + \varepsilon)$ in Eq. (1), we obtain

$$\hat{F}_n(Q(p) + \varepsilon) = F(Q(p)) + \varepsilon f(Q(p)) + o(\varepsilon) + R_n(Q(p) + \varepsilon).$$

In view of Eqs. (2) and (3), we can choose N_1 so that $\varepsilon f(Q(p)) + o(\varepsilon) + R_n(Q(p) + \varepsilon) > 0$ for all $n > N_1$. Because $F(Q(p)) = p$, we obtain $\hat{F}_n(Q(p) + \varepsilon) > p$ for all $n > N_1$.

In the same way, we can show that $p > \hat{F}_n(Q(p) - \varepsilon)$ for all $n > N_2$, provided that ε is small enough. Therefore, $\hat{F}_n(Q(p) + \varepsilon) > p > \hat{F}_n(Q(p) - \varepsilon)$ for all $n > \max(N_1, N_2)$. Since $\hat{F}_n$ is nondecreasing, $Q(p) + \varepsilon > \hat{Q}_n(p) > Q(p) - \varepsilon$, as desired. □

Definition 1. Let $\{Z_t\}_{t=-\infty}^{\infty}$ be a stationary random process in $\mathbb{R}$. Denote by $\mathfrak{A}_a^b$ the σ-algebra generated by events of the form $\{(Z_{i_1}, \ldots, Z_{i_n}) \in E\}$, where $a \le i_1 < i_2 < \cdots < i_n \le b$ and E is a n-dimensional Borel set. For all $A \in \mathfrak{A}_{-\infty}^n$ and $B \in \mathfrak{A}_{n+m}^{\infty}$, we define the mixing coefficients

$$\phi(m) = \sup |\mathrm{P}(B|A) - \mathrm{P}(B)|,$$
$$\alpha(m) = \sup |\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)|,$$
$$\beta(m) = \mathrm{E}(\sup |\mathrm{P}(B|A) - \mathrm{P}(B)|).$$

If, for the sequence $\{Z_t\}$, $\phi(m) \to 0$ ($\alpha(m) \to 0$, $\beta(m) \to 0$) as $m \to \infty$, $\{Z_t\}$ is called ϕ-mixing (α-mixing, β-mixing).

If $\phi(m)$ ($\alpha(m)$, $\beta(m)$) $= O(m^{-a-\Delta})$ for some $\Delta > 0$, then ϕ (α, β) is of size $-a$.

Lemma 3. *Let h be a measurable function $h : \mathbb{R}^{w+1} \to \mathbb{R}$ and define $U_t = h(V_t, \ldots, V_{t-w})$, where w is finite. If $\{V_t\}$ is ϕ-mixing (α-mixing, β-mixing) of size $-a$, $a > 0$, then $\{U_t\}$ is ϕ-mixing (α-mixing, β-mixing) of size $-a$.*

Proof. See Theorems 3.35 and 3.49 of White (2001). □

Proof of Theorem 3. As $\{Z_t\}$ is stationary and mixing, it follows from Lemma 3 that $\{E_t\}$ and $\{\mathbb{I}(E_t \le e)\}$ are also stationary and mixing of the same size as $\{Z_t\}$. In view of Assumption 5, the rest of the proof is then an immediate consequence of Theorem 2 of Yoshihara (1995).

Appendix B. Conditional quantile estimation with correct model g^*

We consider a special case where a functional form of the correct point forecasting model g^* is known, and show that the *conditional* forecast error quantile is estimated consistently. For this, we limit our forecasting models to a parametric model and assume that the chosen parametric model delivers the right conditional mean in a population such that $g^*(Z_t, \beta) = \mathrm{E}(Y_t|Z_t)$, where the parameter $\beta \in \mathcal{B}$ is a vector of unknown model parameters for $\mathcal{B}$, a real and compact set. Therefore, we have

$$Y_t = g^*(Z_t, \beta) + u_t,$$

where u_t is the residual error of the underlying process with mean zero and its pth quantile is denoted by q. In finite samples, the chosen model makes a point forecast by $\hat{Y}_t = g^*(Z_t, \hat{\beta})$, where $\hat{\beta}$ is the estimated parameter. We can then write

$$\begin{aligned} E_t &= Y_t - \hat{Y}_t = g^*(Z_t, \beta) - g^*(Z_t, \hat{\beta}) + u_t \\ &= g^*(Z_t, \beta_1) + u_t. \end{aligned} \tag{5}$$

Note that $g^*(Z_t, \beta_1)$ is the conditional bias in the point forecast due to parameter estimation. By estimating this bias parametrically by $g^*(Z_t, \hat{\beta}_1)$ and estimating the empirical quantile of u_t non-parametrically, we can estimate the sample conditional quantile $\hat{Q}_{|z}(p)$ given Z_t. More specifically, let $\hat{u}_t = E_t - g^*(Z_t, \hat{\beta}_1)$, and let $\hat{q}$ be the pth empirical quantile of $\hat{u}_t$ using order statistics. Then, we write the sample quantile *conditional* on Z_t as

$$\hat{Q}_{|z}(p) = g^*(Z_t, \hat{\beta}_1) + \hat{q}.$$

To show the asymptotic normality of this estimator $\hat{Q}_{|z}(p)$, we make the following regularity conditions, which are a standard set of conditions for $\hat{\beta}_1 \to_p \beta_1^*$ and $\hat{q} \to_p q^*$, where β_1^* and q^* are the unknown true values.

Assumption 6. The observed stochastic process Z_t is stationary and β-mixing such that $\sum_{m=1}^{\infty} m^{1/(r-1)}\beta(m) < \infty$ for $r > 1$.

Assumption 7. The residual error u_t is independent of Z_t and τ-dependent, with a marginal distribution function $H(u)$ and continuously differentiable density $h(u)$, where $h(u) > 0$. The conditional density of $Y_{t+\tau}$ given $Z_t = z$ is bounded: $\eta(y|z) \le \bar{\eta} < \infty$.

Assumption 8. Let $\hat{\beta}_1$ be an estimator of the parameter β_1 in Eq. (5), which can be written as an approximate method of moments estimator, i.e., for some function $l_t(\beta_1)$, $\hat{\beta}_1$ satisfies $\frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} l_t(\hat{\beta}_1) = o(1)$, and the function $l_t(\beta_1)$ is continuously differentiable in β_1.

Assumption 9. For all β_1, $\mathrm{E}|l_t(\beta_1)|^{2r} < \infty$.

Assumption 10. $\mathrm{E}l_t(\beta_1) = 0$ only if $\beta_1 = \beta_1^*$.

Assumption 11. $\mathrm{E}l_t l_t' = L > 0$ and $\mathrm{rank}(l_{\beta_1}) = d$, where $l_t = l_t(\beta_1^*)$ and $l_{\beta_1} = (\partial/\partial\beta_1')\mathrm{E}l_t(\beta_1^*)$.

Assumption 12. For some $C < \infty$ and all β_1, $\mathrm{E}\sup_{\beta_2:|\beta_1-\beta_2|<\delta} |l_t(\beta_1) - l_t(\beta_2)|^{2r} \le C\delta$.

Assumption 13. The function $g^*(Z_t, \beta_1)$ satisfies

$$\sup_{\beta_2:|\beta_1-\beta_2|\le\delta} |g(Z_t, \beta_1) - g(Z_t, \beta_2)| \le a(Z_t)\delta,$$

with $\mathrm{E}a(Z_t) < \infty$.

Theorem 4. *If Assumptions 6–13 hold and $p \in (0, 1)$, then the sample quantile conditional on Z_t satisfies*

$$\frac{n^{1/2}}{\sigma^2_{Q(p)_{|z}}}(\hat{Q}_{|z}(p) - Q_{|z}(p)) \to^D N(0, 1).$$

The exact expression for $\sigma^2_{Q(p)_{|z}}$ can be found in Eq. (9) of Hansen (2006).

Proof. This is a variation of Theorem 1 by Hansen (2006), where the quantile estimation is applied to the out-of-sample forecast errors E_t (instead of being applied to Y_t directly). Hence, given Theorem 1 of Hansen (2006), and under our Assumptions 6–13, it is sufficient to show that E_t is stationary and β-mixing, such that $\sum_{m=1}^{\infty} m^{1/(r-1)}\beta(m) < \infty$ for $r > 1$ and $\eta(y|z) = f(e|z)$, where $f(e|z)$ is the conditional density of E_t.

Under our Assumption 6, it follows from Lemma 3 that E_t is also stationary and absolutely regular. Further, $E_t = Y_t - \hat{Y}_t$, and $\hat{Y}_t = g(Z_t, \hat{\beta})$ is fixed conditional on Z_t, and therefore we have $\eta(y|z) = f(e|z)$. □

References

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: forecasting and control*. Englewood Cliffs, NJ: Prentice-Hall.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics*, *11*(2), 121–135.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, *158*(3), 419–466.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841–862.

Cohen, J. E. (1986). Population forecasts and confidence intervals for Sweden: A comparison of model-based and empirical approaches. *Demography*, *23*(1), 105–126.

DeJong, D. N., Nankervis, J. C., Savin, N. E., & Whiteman, C. H. (1992). Integration versus trend stationary in time series. *Econometrica*, *60*(2), 423–433.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *American Statistical Association*, *74*(366), 427–431.

Diebold, F. X., & Rudebusch, G. D. (1991). On the power of Dickey–Fuller tests against fractional alternatives. *Economics Letters*, *35*(2), 155–160.

Engle, R. F., & Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics*, *22*(4), 367–381.

Gardner, E. S., Jr. (1988). A simple method of computing prediction intervals for time series forecasts. *Management Science*, *34*(4), 541–546.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.

Hansen, B. E. (2006). Interval forecasts and parameter uncertainty. *Journal of Econometrics*, *135*(2), 377–398.

Imbs, J., Mumtaz, H., Ravn, M. O., & Rey, H. (2005). PPP strikes back: aggregation and the real exchange rate. *Quarterly Journal of Economics*, *120*(1), 1–43.

Isengildina-Massa, O., Irwin, S., Good, D. L., & Massa, L. (2011). Empirical confidence intervals for USDA commodity price forecasts. *Applied Economics*, *43*(26), 3789–3803.

Jogensen, M., & Sjoerg, D. I. K. (2003). An effort prediction interval approach based on the empirical distribution of previous estimation accuracy. *Information and Software Technology*, *45*(3), 123–136.

Kim, J. H. (2001). Bootstrap-after-bootstrap prediction intervals for autoregressive models. *Journal of Business and Economic Statistics*, *19*(1), 117–128.

Makridakis, S., Wheelwright, S., & Hyndman, R. (1998). *Forecasting methods and applications*. New York: John Wiley & Sons.

Makridakis, S., & Winkler, R. L. (1989). Sampling distributions of post-sample forecasting errors. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, *38*(2), 331–342.

Phillips, P. C. B. (1979). The sampling distribution of forecasts from a first-order autoregression. *Journal of Econometrics*, *9*(3), 241–261.

Rayer, S., Smith, S., & Tayman, J. (2009). Empirical prediction intervals for county population forecasts. *Population Research and Policy Review*, *28*(6), 773–793.

Reeves, J. J. (2005). Bootstrap prediction intervals for ARCH models. *International Journal of Forecasting*, *21*(2), 237–248.

Stine, R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, *80*(392), 1026–1031.

Stout, W. F. (1974). *Almost sure convergence*. New York: Academic Press.

Taylor, M. P. (2006). Real exchange rates and purchasing power parity: Mean-reversion in economic thought. *Applied Financial Economics*, *16*(1), 1–17.

Taylor, J. W. (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research*, *178*(1), 154–167.

Taylor, J. W., & Bunn, D. W. (1999). A quantile regression approach to generating prediction intervals. *Management Science*, *45*(2), 225–237.

Thombs, L. A., & Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, *85*(410), 486–492.

White, H. (2001). *Asymptotic theory for econometricians*. San Diego: Academic Press.

Williams, W. H., & Goodman, M. L. (1971). A simple method for the construction of empirical confidence limits for economic forecasts. *Journal of the American Statistical Association*, *66*(336), 752–754.

Wu, J. J. (2010). Semiparametric forecast intervals. *Journal of Forecasting*, *31*(3), 189–228.

Yoshihara, K. (1995). The Bahadur representation of sample quantiles for sequences of strongly mixing random variables. *Statistics and Probability Letters*, *24*, 299–304.

Yun Shin Lee is an Assistant Professor of Operations Strategy and Management Science at KAIST College of Business, Korea Advanced Institute of Science and Technology. She received an M.Phil. in Management Science and a Ph.D. in Management Studies from Judge Business School, University of Cambridge. Her research focus is on time series forecasting and inventory planning, particularly when the data-generating model is uncertain.

Stefan Scholtes is Dennis Gillings Professor of Health Management at Judge Business School, University of Cambridge. He is an associate editor of *Mathematics of Operations Research*. He uses patient-level data to analyze the effect of organizational risk factors on the quality and efficiency of hospital services.