JOURNAL OF TIME SERIES ANALYSIS
J. Time. Ser. Anal. **36**: 352–376 (2015)
Published online 14 October 2014 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12099

SPECIAL ISSUE

BOOTSTRAP JOINT PREDICTION REGIONS

MICHAEL WOLF[a]* AND DAN WUNDERLI[b]

[a] *Department of Economics, University of Zurich, Zurich, Switzerland*
[b] *Swiss National Bank, Bern, Switzerland*

Many statistical applications require the forecast of a random variable of interest over several periods into the future. The sequence of individual forecasts, one period at a time, is called a path forecast, where the term *path* refers to the sequence of individual future realizations of the random variable. The problem of constructing a corresponding joint prediction region has been rather neglected in the literature so far: such a region is supposed to contain the entire future path with a prespecified probability. We develop bootstrap methods to construct joint prediction regions. The resulting regions are proven to be asymptotically consistent under a mild high-level assumption. We compare the finite-sample performance of our joint prediction regions with some previous proposals via Monte Carlo simulations. An empirical application to a real data set is also provided.

Received 30 January 2014; Revised 26 August 2014; Accepted 26 August 2014

Keywords: Generalized error rates; path forecast; simultaneous prediction intervals.

1. INTRODUCTION

When predicting a random variable, a point forecast alone is often considered insufficient. In addition, a statement about the uncertainty contained in the point forecast, as expressed by a *prediction interval*, may also be desired. This is similar to the situation where a point estimator of a population parameter alone is considered insufficient and where a statement about the uncertainty contained in the point estimate, as expressed by a *confidence interval*, is also desired. However, constructing a prediction interval for a random variable is inherently more difficult than constructing a confidence interval for a population parameter.

In the latter problem, typically, a central limit theorem can be applied to argue that an estimator of the parameter has, approximately, a normal distribution for large sample sizes. This allows for the construction of standard, normal-theory confidence intervals described in any basic statistics text book. The use of bootstrap methods as an alternative is 'only' motivated by higher-order considerations: standard methods already result in confidence intervals that are consistent, that is, have coverage probability equal to the nominal level $1 - \alpha$ asymptotically.

In the former problem, no central limit theorem can be applied to argue that the difference between a point forecast and the random variable of interest has, approximately, a normal distribution for large sample sizes. Therefore, standard normal-theory prediction intervals are only valid under restrictive parametric assumptions. The use of bootstrap methods as an alternative is motivated by first-order considerations already: they result in prediction intervals that are consistent under very general assumptions where standard, normal-theory prediction intervals fail. How to apply the bootstrap to construct prediction intervals that are not only asymptotically consistent but also have good finite-sample properties is not a trivial problem. However, it can be considered solved by now to a satisfactory degree; for example, see De Gooijer and Hyndman (2006, Section 12) for an overview.

The discussion up to this point only applies to a single (future) random variable. In many applications, however, a random variable of interest is predicted up to H periods into the future. For example, one might predict future

* Correspondence to: Michael Wolf, Department of Economics, University of Zurich, 8032 Zurich, Switzerland. E-mail: michael.wolf@econ.uzh.ch

inflation for the next $H = 12$ months. A *path* refers to the sequence of future realizations 1 to H periods into the future. A *path forecast* refers to the sequence of corresponding forecasts 1 to H periods into the future.

On the one hand, one can construct H marginal prediction intervals by using a given method to construct a prediction interval repeatedly, one period at a time. However, by design, probability statements then only apply marginally, one period at a time: the prediction interval at a specific horizon h, for some $1 \leq h \leq H$, will contain the random variable h periods into the future with prespecified probability $1 - \alpha$.

On the other hand, a more general problem is the construction of a *joint prediction region* that will contain the entire future path with the desired probability $1-\alpha$. For example, if one would like to know with probability $1-\alpha$ how high inflation might rise over the next $H = 12$ months, one needs to construct a joint prediction region (JPR) for the future path at level $1 - \alpha$ as opposed to stringing together 12 marginal prediction intervals, each one at level $1 - \alpha$.

It should be clear that stringing together marginal prediction intervals for horizons $h = 1$ up to $h = H$, each one at level $1 - \alpha$, will not result in a JPR that contains the entire future path with probability $1 - \alpha$. Instead, apart from pathological cases, the joint coverage probability will be strictly less than $1 - \alpha$, and decreasing in H. Denote by E_h the event that the random variable at period h in the future will fall into its prediction interval. If the events $\{E_h\}_{h=1}^{H}$ are independent of each other, then stringing together marginal prediction intervals results in a JPR that will contain the entire future path with probability $(1 - \alpha)^H$ only. In practice, the events $\{E_h\}_{h=1}^{H}$ are typically not independent of each other. Stringing together marginal prediction intervals then results in a JPR that will contain the entire future path with probability somewhere between $\max\{0, 1-H\cdot\alpha\}$ and $1-\alpha$, where the lower bound is obtained by Bonferroni's inequality. The exact probability is a function of the dependence structure of the events $\{E_h\}_{h=1}^{H}$.

By using Bonferroni's inequality, a conservative JPR can be constructed by stringing together marginal prediction intervals at level $1 - \alpha/H$ instead of at level $1 - \alpha$; for example, such an approach is already mentioned by Lütkepohl (1991, Section 2.2.3). However, since Bonferroni's inequality is crude, such an approach results in joint confidence regions (JCRs) whose coverage probability is generally (much) above $1 - \alpha$ and that are thus unnecessarily wide, leading to a loss of information.

The construction of JPRs for future paths of a random variable of interest with coverage probability $1 - \alpha$ has been rather neglected in the forecasting literature so far. Two notable exceptions are Jordà and Marcellino (2010) and Staszewska-Bystrova (2011). The former work proposes an 'asymptotic' method that relies on the overly strong assumption that forecast errors have, approximately, a normal distribution. The latter work proposes a bootstrap method that is of heuristic nature only.

In this article, we propose a bootstrap method to construct joint predictions regions that are proven to contain future paths of a random variable of interest with probability $1 - \alpha$, at least asymptotically, under a mild high-level assumption.

In addition, we also consider the more general problem of constructing JPRs that will contain all elements of future paths but a small number of them with probability $1 - \alpha$; this small number will be denoted by $k - 1$. If the maximum forecast horizon H is large, the applied researcher may deem the original criterion (namely, that all elements of the future path must be contained in the joint prediction region with probability $1 - \alpha$) as too strict. For example, when $H = 24$, it may be deemed acceptable that up to $k - 1 = 2$ elements of the future path may fall outside the JPR; thus requiring that 'only' at least 22 of the 24 elements – or at least 90% of the 24 elements – of the future path be contained in the JPR with probability $1 - \alpha$. The choice of k must be made by the applied researcher, not by the statistician. However, it will be useful to the applied researcher to have a method available that can handle any desired value of k. In particular, the choice $k = 1$ yields a 'standard' JPR that must contain all elements of a future path with probability $1 - \alpha$.

The remainder of the article is organized as follows. Section 2 contains some background results that are useful for setting the stage. Section 3 describes our method to construct joint prediction regions and compares it with some previous proposals in the literature. Section 4 studies finite-sample performance via Monte Carlo simulations. Section 5 provides an empirical application to real data. Finally, Section 6 contains concluding remarks. Some additional figures and tables can be found in the online Supporting Information.

DOI: 10.1111/jtsa.12099

2. BACKGROUND RESULTS

Our motivating problem is the construction of a joint prediction region for a future path of a random variable of interest. However, the proposed methodology applies more generally to the construction of a JPR of an arbitrary random vector that has not been observed yet. In explaining the methodology, it will be convenient to start with the special case of a single random variable that has not been observed yet.

2.1. Single Forecast

First, consider a single random variable y with mean $\mu := \mathbb{E}(y)$. This special case makes it easier to explain some fundamental concepts before considering the more general case of a random vector with H elements.

One may wish to predict y or to estimate μ. Denote the forecast of y by $\widehat{y}$ and the estimator of μ by $\widehat{\mu}$. Often, the two are actually the same; that is, $\widehat{y} = \widehat{\mu}$, for example, in the context of linear regression models under quadratic loss. Therefore, in terms of a (point) forecast of y compared with a (point) estimate of μ, there often is no difference at all.

However, what if one desires an 'uncertainty interval' in addition? Such an interval should contain the random variable y or its mean μ, with a prespecified probability $1-\alpha$. Now the two solutions are fundamentally different, and the former interval will have to be wider because of the additional randomness contained in the random variable y compared with its mean μ. To make this distinction apparent in the notation, we prefer to call the solution to the former problem a *prediction interval* and the solution to the latter problem a *confidence interval*. In doing so, we are in agreement with De Gooijer and Hyndman (2006, p. 460):

> Unfortunately, there is still some confusion in terminology with many authors by 'confidence interval' instead of 'prediction interval'. A confidence interval is for a model parameter, whereas a prediction interval is for a random variable. Almost always, forecasters will want prediction intervals – intervals which contain the true values of future observations with [a] specified probability.

2.2. Path Forecast

More generally, consider a random vector $Y := (y_1, \ldots, y_H)'$ of interest with mean $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_H)' = \mathbb{E}(Y)$. For the purposes of this article, Y will typically correspond to the values of a random variable 1 to H periods into the future, that is, to a future *path* of a random variable. However, the succeeding discussion applies to any random vector. The underlying probability mechanism is denoted by $\mathbb{P}$.

One can wish to predict Y or to estimate $\boldsymbol{\mu}$. Denote the forecast of Y by $\hat{Y}$ and the estimator of $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}}$. (When Y corresponds to a future path of a random variable, $\hat{Y}$ is also called a *path forecast*.) Again, often, the two are actually the same; that is, $\hat{Y} = \hat{\boldsymbol{\mu}}$, for example, in the context of linear regression models under quadratic loss. Therefore, again, in terms of a (point) forecast of Y compared with a (point) estimate of $\boldsymbol{\mu}$, there often is no difference at all.

What if one desires the extension of an 'uncertainty interval' for a univariate quantity to a '(joint) uncertainty region' for a multivariate quantity? In the most stringent case, such a region should contain the *entire* random vector Y or its mean $\boldsymbol{\mu}$, with a prespecified probability $1-\alpha$. Again, the two solutions are fundamentally different, and the former region will have to be larger (in volume) because of the additional randomness contained in Y compared with its mean $\boldsymbol{\mu}$.

A potential complication with joint regions arises when uncertainty statements concerning the individual components y_h or μ_h are desired. For example, this is typically the case when a JPR for Y is to be constructed in addition to a path forecast $\hat{Y}$. One desires lower and upper bounds for each component y_h in such a manner that the entire vector Y is contained in the implied rectangle with probability $1-\alpha$. This is a trivial task if the underlying JPR is already of rectangular form. However, this is not true for all methods to compute joint regions; many

DOI: 10.1111/jtsa.12099

methods result in regions of elliptical form instead. The most prominent example is the Scheffé joint region, dating back to Scheffé (1953, 1959).

The Scheffé JCR for $\boldsymbol{\mu}$ is obtained by inverting the classical F-test. Let $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}})$ denote an estimated covariance matrix of $\hat{\boldsymbol{\mu}}$. Then the JCR is given by

$$\text{JCR} := \left\{ \boldsymbol{\mu}_0 : (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)' \left[\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\mu}}) \right]^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \leq \chi^2_{H,1-\alpha} \right\}, \tag{1}$$

where $\chi^2_{H,1-\alpha}$ denotes the $1-\alpha$ quantile of the chi-square distribution with H degrees of freedom. The use of this JCR is usually justified by a central limit theorem implying an approximate multivariate normal distribution of $\hat{\boldsymbol{\mu}}$ with mean $\boldsymbol{\mu}$. Such a central limit theorem will hold under mild regularity conditions; for example, see White (2001).

The Scheffé JPR for Y is obtained similarly. Define the vector of prediction errors by $\hat{U} := \hat{Y} - Y$ and let $\hat{\boldsymbol{\Sigma}}\left(\hat{U}\right)$ denote an estimated covariance matrix of this vector. Then the joint prediction region is given by

$$\text{JPR} := \left\{ X : \left(\hat{Y} - X\right)' \left[\hat{\boldsymbol{\Sigma}}\left(\hat{U}\right) \right]^{-1} (\hat{Y} - X) \leq \chi^2_{H,1-\alpha} \right\}. \tag{2}$$

The use of this JPR is only justified if $\hat{U}$ has approximately a multivariate normal distribution with mean zero. This is a strong additional assumption, which is often violated in practice. A central limit theorem can typically be applied to argue that an *estimator* has, approximately, a normal distribution for large sample sizes. However, a central limit theorem can never be applied to argue that a *forecast error has*, approximately, a normal distribution for large sample sizes. This point is illustrated via a simple example in Remark 4.

If the joint region is of elliptical form and statements concerning the individual components are desired, the joint region has to be 'projected' on the axes of $\mathbb{R}^H$. This action implies a *larger* rectangular joint region, namely, the smallest rectangle that contains the original elliptical region. As a result, if the elliptical region has joint coverage probability $1-\alpha$, then the implied rectangular region has joint coverage probability larger than $1-\alpha$. Therefore, such a projection method is conservative. If statements concerning the individual components are desired, it is advantageous to construct 'direct' rectangular joint regions instead.

Remark 1. It is useful to illustrate these concepts in a simple, parametric set-up. Assume $Y := (Y_1, Y_2)' \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{I}_2)$, where $\boldsymbol{\mu} := (\mu_1, \mu_2)'$ and $\boldsymbol{I}_2$ is the identity matrix of dimension 2. Therefore, Y_1 and Y_2 are independent with $Y_h \sim \text{N}(\mu_h, 1)$. The goal is to construct a joint confidence region for $\boldsymbol{\mu}$. The point estimator for $\boldsymbol{\mu}$ is simply given by the observed random vector; that is, $\hat{\boldsymbol{\mu}} := Y$.

The Scheffé JCR is obtained by inverting the classical F-test. It is a circle centred at Y with radius $\sqrt{\chi^2_{2,1-\alpha}}$, where $\chi^2_{2,1-\alpha}$ denotes the $1-\alpha$ quantile of the chi-square distribution with two degrees of freedom. For example, when $\alpha = 0.05$, the radius is $\sqrt{5.99} \approx 2.45$. The implied rectangular joint confidence region, obtained by projecting the circle on the two axes, is a square with centre Y and an (approximate) half length of 2.45.

On the other hand, a 'direct' rectangular JCR is given by

$$\left[Y_1 \pm d^{\max}_{|\cdot|,1-\alpha}\right] \times \left[Y_2 \pm d^{\max}_{|\cdot|,1-\alpha}\right],$$

where $d^{\max}_{|\cdot|,1-\alpha}$ is the $1-\alpha$ quantile of the random variable $\max\{|Y_1 - \mu_1|, |Y_2 - \mu_2|\}$. These quantiles are not commonly tabulated but can be easily simulated to arbitrary precision. For example, when $\alpha = 0.05$, then $d^{\max}_{|\cdot|,0.95} \approx 2.24$.

The 'direct' rectangular joint confidence region is thus a square with centre Y and an (approximate) half length of 2.24. Therefore, it is smaller than the implied rectangular joint confidence region by the Scheffé method. An illustration is provided in Figure A.1 in the online Supporting Information.

DOI: 10.1111/jtsa.12099

The stringent joint regions discussed so far control the probability of containing the entire vector of interest to be (at least) equal to $1-\alpha$. Equivalently, they control the probability of missing at least one component of the vector to be (at most) equal to α. Borrowing from the multiple testing literature, the latter probability can be termed the *familywise error rate* (FWE); for example, see Romano *et al.* (2008). So for a JCR for $\boldsymbol{\mu}$,

$$\text{FWE} := \mathbb{P}\{\text{at least one of the } \mu_h \text{ is not contained in the JCR}\}, \tag{3}$$

whereas for a JPR for Y,

$$\text{FWE} := \mathbb{P}\{\text{at least one of the } y_h \text{ is not contained in the JPR}\}. \tag{4}$$

When the maximum forecast horizon H is large, control of the FWE may be deemed to strict.[1] The decision on whether the FWE is too strict or not in a given application has to be made by the applied researcher, not by the statistician. It is the job of the statistician to provide the applied researcher with an alternative tool in case his decision is against control of the FWE. In such a case, we suggest to use the *generalized FWE* (k-FWE).

For a JCR for $\boldsymbol{\mu}$,

$$k\text{-FWE} := \mathbb{P}\{\text{at least } k \text{ of the } \mu_h \text{ is not contained in the JCR}\}, \tag{5}$$

whereas for a JPR for Y,

$$k\text{-FWE} := \mathbb{P}\{\text{at least } k \text{ of the } y_h \text{ is not contained in the JPR}\}. \tag{6}$$

As a special case, the choice $k = 1$ gives back the FWE. On the other hand, any choice $k \geq 2$ results in a less stringent error rate.

As will be discussed in Section 3, the larger the value of k, the smaller the resulting joint region. Consequently, by being willing to miss a small number of components in the joint region, the applied researcher can obtain more precise bounds in return.

Since the number of components, H, is known, control of the k-FWE immediately gives control on the probability of the proportion of components not contained in the joint region. Take the example of a path forecast with $H = 24$ components, as when predicting monthly inflation for the next 2 years. Then the choice $k = 3$ allows for a proportion of missed components up to 10%. This is because one or two missed components, out of the $H = 24$, do not constitute a violation of the 3-FWE criterion, but three or more missed components do.

The next section details how the k-FWE, which includes the FWE as a special case, can be controlled in practice. It only does this in the context of a joint prediction region for Y. The method is similar in the context of a joint confidence region for $\boldsymbol{\mu}$ and is detailed by Romano and Wolf (2005, 2007) already.

Since the method is based on quantiles of random variables whose cumulative distribution function may not be invertible, the following remark is in order.

Remark 2. If the cumulative distribution function of a random variable is not invertible, then its quantiles are not necessarily uniquely defined. To be specific, we adopt the following definition for quantiles in this article.

Let X be a random variable with cumulative distribution function $F(\cdot)$. Then, for $\lambda \in (0,1)$, the λ quantile of (the distribution of) X is defined as $\inf\{x : F(x) \geq \lambda\}$.

[1] Jordà *et al.* (2010, Section 2.2) state: 'For example, in a prediction of a path of monthly inflation over the next two years, control of the FWE would result in rejection of such paths as when the trajectory of inflation is [almost] correctly predicted for 23 periods but the prediction of the last month is particularly poor.'

DOI: 10.1111/jtsa.12099

3. JPRs BASED ON K-FWE CONTROL

The goal is to construct a JPR for a future path that controls the k-FWE, for an arbitrary integer $1 \leq k < H$. In particular, the special choice $k = 1$ corresponds to control of the FWE.

Any formal analysis has to be put into a suitable framework. To this end, we borrow some notation from Jordà *et al.* (2010). We start out by discussing the case of a univariate time series, which simplifies the notation and makes it easier to focus on the methodology.

3.1. Univariate Time Series

One observes a univariate time series $\{y_1, \ldots, y_T\}$ generated from a true probability mechanism $\mathbb{P}$ and wishes to predict the future path $Y_{T,H} := (y_{T+1}, \ldots, y_{T+H})'$. At time t, denote a forecast h periods ahead by $\hat{y}_t(h)$. Then a path forecast for $Y_{T,H}$ is given by $\hat{Y}_T(H) := (\hat{y}_T(1), \ldots, \hat{y}_T(H))'$. Denote the vector of prediction errors by $\hat{U}_T(H) := (\hat{u}_T(1), \ldots, \hat{u}_T(H))' := \hat{Y}_T(H) - Y_{T,H}$. Finally, $\hat{\sigma}_T(h)$ denotes a prediction standard error, that is, a standard error for $\hat{u}_T(h)$: it is an estimator of the unknown standard deviation of the random variable $\hat{u}_T(h)$.

We further assume a generic method to compute a vector of bootstrap prediction errors $\hat{U}^*_T(H) := \left(\hat{u}^*_T(1), \ldots, \hat{u}^*_T(H)\right)'$, based on artificial bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$ generated from an estimated probability mechanism $\hat{\mathbb{P}}_T$. Such bootstrap forecast errors can be computed in many different ways. We shall not enter this debate here; the goal is to provide a generic procedure to construct a joint prediction region where application-specific details are up to the applied researcher. Finally, $\hat{\sigma}^*_T(h)$ denotes a bootstrap prediction standard error, that is, a standard error for $\hat{u}^*_T(h)$.

We now briefly illustrate these concepts. The observed data are $\{y_1, \ldots, y_T\}$. The applied researcher selects a suitable 'null' model, fits it to the data, and then uses the fitted model to make the predictions $\hat{y}_T(h)$, for $h = 1, \ldots, H$. To be specific, assume he uses an ARIMA model. The fitted model also provides prediction standard errors $\hat{\sigma}_T(h)$. Next, the applied researcher generates bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$. To this end, he or she can use a parametric bootstrap, based on the ARIMA model fitted from the original data; this might be a preferred approach if he or she believes that his or her null model is correctly specified. Alternatively, he or she can use a non-parametric time-series bootstrap (say a block bootstrap or a sieve bootstrap); this would be a suitable approach if he or she believes that his or her null model might be misspecified.[2] Not making use of the stretch $\{y^*_{T+1}, \ldots, y^*_{T+H}\}$, he or she computes forecasts $\hat{y}^*_T(h)$ and prediction standard errors $\hat{\sigma}^*_T(h)$. Finally, he or she computes $\hat{u}^*_T(h) := \hat{y}^*_T(h) - y^*_{T+h}$.

Our high-level assumption in the following is based on the two vectors of *standardized prediction errors* $\hat{S}_T(H) := (\hat{u}_T(1)/\hat{\sigma}_T(1), \ldots, \hat{u}_T(H)/\hat{\sigma}_T(H))'$ and $\hat{S}^*_T(H) := \left(\hat{u}^*_T(1)/\hat{\sigma}^*_T(1), \ldots, \hat{u}^*_T(H)/\hat{\sigma}^*_T(H)\right)'$. Denote the probability law under $\mathbb{P}$ of $\hat{S}_T(H)|y_T, y_{T-1}, \ldots$ by $\hat{J}_T$. Also denote the probability law under $\hat{\mathbb{P}}_T$ of $\hat{S}^*_T(H)|y^*_T, y^*_{T-1}, \ldots$ by $\hat{J}^*_T$. In the asymptotic framework, T tends to infinity, whereas H remains fixed.

Assumption 1. $\hat{J}_T$ converges in distribution to a non-random continuous limit law $\hat{J}$. Furthermore, $\hat{J}^*_T$ consistently estimates this limit law: $\rho\left(\hat{J}_T, \hat{J}^*_T\right) \to 0$ in probability, for any metric ρ metrizing weak convergence.

Expressed in words, Assumption 1 states that, as the sample size T increases, the conditional distribution of the vector of standardized bootstrap prediction errors $\hat{S}^*_T(H)$ becomes a more and more reliable approximation to the (unknown) conditional distribution of the vector of true standardized prediction errors $\hat{S}_T(H)$.

We next specify the forms of the JPRs for $Y_{T,H}$, first for the two-sided case and then for the one-sided case.

[2] For an overview of non-parametric time-series bootstrap methods, the reader is referred to Bühlmann (2002), Lahiri (2003), and Politis (2003).

DOI: 10.1111/jtsa.12099

Some further notation is required. Suppose $X := (x_1, \ldots, x_H)'$ is a vector with H components. First, for $k \in \{1, \ldots, H\}$, k-max(X) returns the kth-largest value of the x_h. So, if the elements x_h, $1 \leq h \leq H$, are ordered as $x_{(1)} \leq \cdots \leq x_{(H)}$, then k-max$(X) := x_{(H-k+1)}$. Second, for $k \in \{1, \ldots, H\}$, k-min(X) returns the kth smallest value of the x_h; that is, k-min$(X) := x_{(k)}$. Third, $|X|$ denotes the vector $(|x_1|, \ldots, |x_H|)'$.

Let $d_{|\cdot|,1-\alpha}^{k\text{-max}}$ denote the $1-\alpha$ quantile of the random variable k-max $\left(|\hat{S}_T(H)|\right)$. Then a two-sided JPR for $Y_{T,H}$ that controls the k-FWE in finite samples is given by

$$\left[\hat{y}_T(1) \pm d_{|\cdot|,1-\alpha}^{k\text{-max}} \cdot \hat{\sigma}_T(1)\right] \times \cdots \times \left[\hat{y}_T(H) \pm d_{|\cdot|,1-\alpha}^{k\text{-max}} \cdot \hat{\sigma}_T(H)\right]. \tag{7}$$

The implication is that the probability that the region (7) will contain at least $H-k+1$ elements of $Y_{T,H}$ is equal to (at least) $1-\alpha$ in finite samples. This property follows immediately from the definition of the multiplier $d_{|\cdot|,1-\alpha}^{\max}(k)$.

The problem is that the ideal region (7) is not feasible, since the multiplier $d_{|\cdot|,1-\alpha}^{k\text{-max}}$ is unknown. This multiplier has to be estimated in practice by $d_{|\cdot|,1-\alpha}^{k\text{-max},*}$, which is defined as the $1-\alpha$ quantile of the random variable k-max $\left(|\hat{S}_T^*(H)|\right)$. This quantile can typically not be derived analytically, but it can be simulated to arbitrary precision from a sufficiently large number of bootstrap samples (Algorithm 1).

Then a two-sided JPR for $Y_{T,H}$ that controls the k-FWE asymptotically is given by

$$\left[\hat{y}_T(1) \pm d_{|\cdot|,1-\alpha}^{k\text{-max},*} \cdot \hat{\sigma}_T(1)\right] \times \cdots \times \left[\hat{y}_T(H) \pm d_{|\cdot|,1-\alpha}^{k\text{-max},*} \cdot \hat{\sigma}_T(H)\right]. \tag{8}$$

The implication is that the probability that the region (8) will contain at least $H-k+1$ elements of $Y_{T,H}$ is equal to (at least) $1-\alpha$ asymptotically.

The modifications to the one-sided case are as follows; we only present the feasible regions.

Let $d_{1-\alpha}^{k\text{-max},*}$ denote the $1-\alpha$ quantile of the random variable k-max $\left(\hat{S}_T^*(H)\right)$. Then a one-sided lower JPR for $Y_{T,H}$ that controls the k-FWE asymptotically is given by

$$\left[\hat{y}_T(1) - d_{1-\alpha}^{k\text{-max},*} \cdot \hat{\sigma}_T(1), \infty\right) \times \cdots \times \left[\hat{y}_T(H) - d_{1-\alpha}^{k\text{-max},*} \cdot \hat{\sigma}_T(H), \infty\right). \tag{9}$$

Let $d_{\alpha}^{k\text{-min},*}$ denote the α quantile of the random variable k-min $\left(\hat{S}_T^*(H)\right)$. Then a one-sided upper JPR for $Y_{T,H}$ that controls the k-FWE asymptotically is given by

$$\left(-\infty,\ \hat{y}_T(1) - d_{\alpha}^{k\text{-min},*} \cdot \hat{\sigma}_T(1)\right] \times \cdots \times \left(-\infty,\ \hat{y}_T(H) - d_{\alpha}^{k\text{-min},*} \cdot \hat{\sigma}_T(H)\right]. \tag{10}$$

Note here that $d_{\alpha}^{k\text{-min},*}$ is generally a negative number so that, for each horizon h, the upper end of the corresponding interval is indeed larger than the forecast $\hat{y}_T(h)$.

As is clear from the definitions, both the multipliers $d_{|\cdot|,1-\alpha}^{k\text{-max},*}$ and $d_{1-\alpha}^{k\text{-max},*}$ are monotonically decreasing in k, while the multiplier $d_{\alpha}^{k\text{-min},*}$ is monotonically increasing in k. Consequently, the larger the value of k, the smaller in volume are the regions (8)–(10); for an illustration, see Section 5.1. (When we speak of 'volume' for the one-sided regions (9) and (10), we implicitly refer to the relevant lower or upper 'half volumes', since the entire volume is always infinite, of course.)

The following proposition formally establishes the asymptotic validity of these feasible bootstrap JPRs.

Proposition 1. Under Assumption 1, each of the JPRs (8)–(10) for $Y_{T,H}$ satisfies

$$\limsup_{T \to \infty} \ k\text{-FWE} \leq \alpha, \tag{11}$$

DOI: 10.1111/jtsa.12099

where

$$k\text{-FWE} := \mathbb{P}\{\text{at least } k \text{ of the } y_{T+h} \text{ is not contained in the JPR}\}. \tag{12}$$

Proof
We prove the stated result for the JPR (9). The proofs for the JPRs (8) and (10) are completely analogous.

Let $\hat{L}_T$ denote a random variable with distribution $\hat{J}_T$, and let $\hat{L}$ denote a random variable with distribution $\hat{J}$. By Assumption 1 and the continuous mapping theorem, k-max $\left(\hat{L}_T\right)$ converges weakly to k-max $\left(\hat{L}\right)$, whose distribution is continuous. Our notation implies that the conditional sampling distribution under $\mathbb{P}$ of k-max $\left(\hat{S}_T(H)\right)$ is identical to the distribution of k-max $\left(\hat{L}_T\right)$. By similar reasoning, the conditional sampling distribution under $\hat{\mathbb{P}}_T$ of k-max $\left(\hat{S}^*_T(H)\right)$ also converges weakly to the distribution of k-max $\left(\hat{L}\right)$. We then show that

$$\mathbb{P}\left\{k\text{-max}\left(\hat{S}_T(H)\right) \leq d^{k\text{-max},*}_{1-\alpha}\right\} \to 1-\alpha \tag{13}$$

is similar to the proof of Theorem 1 of Beran (1984).

Since by definition of the k-FWE and the construction of the joint prediction region (9),

$$k\text{-FWE} = 1 - \mathbb{P}\left\{k\text{-max}\left(\hat{S}_T(H)\right) \leq d^{k\text{-max},*}_{1-\alpha}\right\}, \tag{14}$$

the proof that the stated result (11) holds for the JPR (9) now follows immediately from (13). □

The following algorithm details how to compute the three multipliers $d^{k\text{-max},*}_{|\cdot|,1-\alpha}$, $d^{k\text{-max},*}_{1-\alpha}$, and $d^{k\text{-min},*}_{\alpha}$ in practice. The algorithm assumes a generic bootstrap method, chosen by the applied researcher, to generate bootstrap data and standardized bootstrap prediction errors. In particular, such a bootstrap method is based on an estimated probability mechanism $\hat{\mathbb{P}}_T$.

Algorithm 1. Computation of the JPR multipliers, univariate case:

(1) Generate bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$ from $\hat{\mathbb{P}}_T$.
(2) Not making use of the stretch $\{y^*_{T+1}, \ldots, y^*_{T+H}\}$, compute forecasts $\hat{y}^*_T(h)$ and prediction standard errors $\hat{\sigma}^*_T(h)$.
(3) Compute bootstrap prediction errors $\hat{u}^*_T(h) := \hat{y}^*_T(h) - y^*_{T+h}$.
(4) Compute standardized bootstrap prediction errors $\hat{s}^*_T(h) := \hat{u}^*_T(h)/\hat{\sigma}^*_T(h)$ and let $\hat{S}^*_T(H) := \left(\hat{s}^*_T(1), \ldots, \hat{s}^*_T(H)\right)'$.
(5) Compute $k\text{-max}^*_{|\cdot|} := k\text{-max}\left(|\hat{S}^*_T(H)|\right)$, $k\text{-max}^* := k\text{-max}\left(\hat{S}^*_T(H)\right)$, and $k\text{-min}^* := k\text{-min}\left(\hat{S}^*_T(H)\right)$.
(6) Repeat this process B times, resulting in statistics $\{k\text{-max}^*_{|\cdot|,1}, \ldots, k\text{-max}^*_{|\cdot|,B}\}$, $\{k\text{-max}^*_1, \ldots, k\text{-max}^*_B\}$, and $\{k\text{-min}^*_1, \ldots, k\text{-min}^*_B\}$.
(7) Compute the corresponding empirical quantiles:

(7.1) $d^{k\text{-max},*}_{|\cdot|,1-\alpha}$ is the empirical $1-\alpha$ quantile of the statistics $\{k\text{-max}^*_{|\cdot|,1}, \ldots, k\text{-max}^*_{|\cdot|,B}\}$.
(7.2) $d^{k\text{-max},*}_{1-\alpha}$ is the empirical $1-\alpha$ quantile of the statistics $\{k\text{-max}^*_1, \ldots, k\text{-max}^*_B\}$.
(7.3) $d^{k\text{-min},*}_{\alpha}$ is the empirical α quantile of the statistics $\{k\text{-min}^*_1, \ldots, k\text{-min}^*_B\}$.

In an application, the number B of bootstrap samples should be chosen as large as possible, at the very least $B \geq 1000$.

DOI: 10.1111/jtsa.12099

Remark 3. Proposition 1 only addresses asymptotic consistency. It does not address finite-sample performance. To ensure best-possible finite-sample performance, the applied researcher should make an effort to match the bootstrap distribution $\hat{J}_T^*$ as close as possible to the true distribution $\hat{J}_T$. How this is to be carried out in detail depends on the particular bootstrap method chosen by the applied researcher. Many articles have been written on this problem already; for example, see De Gooijer and Hyndman (2006, Section 12).

We confine ourselves to the general statement that model parameters that have to be estimated from the original data $\{y_1, \ldots, y_T\}$ to compute the forecasts $\hat{y}_T(h)$ and the prediction standard errors $\hat{\sigma}_T(h)$ should be re-estimated from the bootstrap data $\{y_1^*, \ldots, y_T^*\}$ to compute the forecasts $\hat{y}_T^*(h)$ and the prediction standard errors $\hat{\sigma}_T^*(h)$. It may be tempting, say in order to save computing time, to simply use the estimated model parameters from the original data $\{y_1, \ldots, y_T\}$ to compute the forecasts $\hat{y}_T^*(h)$ and the prediction standard errors $\hat{\sigma}_T^*(h)$. However, such an approach does not reflect the fact that the true model parameters are unknown and generally leads to bootstrap prediction errors that are too small in magnitude.

3.2. Multivariate Time Series

Compared with the special case of a univariate time series, the methodology does not change in any fundamental way in the general case of a multivariate time series, as in the case of VAR forecasting. Mainly, the notation becomes more complex.

One observes a time series $\{Z_1, \ldots, Z_T\}$, where $Z_t := (z_{1,t}, \ldots, z_{K,t})'$, generated from a true probability mechanism $\mathbb{P}$ and wishes to predict the next stretch of H observations for a particular component of Z_t. Assume without loss of generality that one wishes to predict the first component of Z_t and write $Z_t := (y_t, z_{2,t}, \ldots, z_{K,t})'$.

In this more general case, the forecast of y_{T+h}, denoted by $\hat{y}_T(h)$ again, will be a function of $\{Z_1, \ldots, Z_T\}$ instead of a function of $\{y_1, \ldots, y_T\}$ only, and similarly for the corresponding prediction standard error $\hat{\sigma}_T(h)$.

Artificial bootstrap data $\{Z_1^*, \ldots, Z_T^*, Z_{T+1}^*, \ldots, Z_{T+H}^*\}$ are generated from an estimated probability mechanism $\hat{\mathbb{P}}_T$. In particular, K-variate VAR models appear a popular choice to this end with applied researchers; more generally, structural VAR, vector error correction model (VECM), or structural VECM models can also be used; for example, see Lütkepohl (2005).

Denote $Z_t^* := \left(y_t^*, z_{2,t}^*, \ldots, z_{K,t}^*\right)'$. The forecast of y_{T+h}^*, denoted by $\hat{y}_T^*(h)$ again, will be a function of $\{Z_1^*, \ldots, Z_T^*\}$ instead of a function of $\{y_1^*, \ldots, y_T^*\}$ only, and similarly for the corresponding prediction standard error $\hat{\sigma}_T^*(h)$.

Assumption 1 continues to be based on the two vectors of standardized prediction errors $\hat{S}_T(H) := (\hat{u}_T(1)/\hat{\sigma}_T(1), \ldots, \hat{u}_T(H)/\hat{\sigma}_T(H))'$ and $\hat{S}_T^*(H) := \left(\hat{u}_T^*(1)/\hat{\sigma}_T^*(1), \ldots, \hat{u}_T^*(H)/\hat{\sigma}_T^*(H)\right)'$. Only that now, more generally, $\hat{J}_T$ denotes the probability law under $\mathbb{P}$ of $\hat{S}_T(H)|Z_T, Z_{T-1}, \ldots$; and $\hat{J}_T^*$ denotes the probability law under $\hat{\mathbb{P}}_T$ of $\hat{S}_T^*(H)|Z_T^*, Z_{T-1}^*, \ldots$.

Having detailed how the quantities of interest are defined and computed in the more general case, the methodology outlined in the case of a univariate time series applies verbatim.

The various forms of the JPRs are still given by (8)–(10), and Proposition 1 continues to hold.

The following algorithm details how to compute the three multipliers $d_{|\cdot|,1-\alpha}^{k\text{-max},*}$, $d_{1-\alpha}^{k\text{-max},*}$, and $d_{\alpha}^{k\text{-min},*}$ in practice. The algorithm assumes a generic bootstrap method, chosen by the applied researcher, to generate bootstrap data and standardized bootstrap prediction errors. In particular, such a bootstrap method is based on an estimated probability mechanism $\hat{\mathbb{P}}_T$.

Algorithm 2. Computation of the JPR multipliers, multivariate case:

(1) Generate bootstrap data $\{Z_1^*, \ldots, Z_T^*, Z_{T+1}^*, \ldots, Z_{T+H}^*\}$ from $\hat{\mathbb{P}}_T$.
(2) Not making use of the stretch $\{Z_{H+1}^*, \ldots, Z_{T+H}^*\}$, compute forecasts $\hat{y}_T^*(h)$ and prediction standard errors $\hat{\sigma}_T^*(h)$.

DOI: 10.1111/jtsa.12099

(3) Identical to Algorithm 1.
⋮
(7) Identical to Algorithm 1.

3.3. Comparison with Two Previous Methods

Jordà and Marcellino (2010) propose an 'asymptotic' method to construct a JPR for $Y_{T,H}$ that controls the FWE.[3] It is based on the assumption that

$$\sqrt{T}\left(\hat{Y}_T(H) - Y_{T,H}|Z_T, Z_{T-1}, \ldots\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{\Xi}_H\right), \tag{15}$$

where $\xrightarrow{d}$ denotes convergence in distribution, and on the availability of a consistent estimator $\hat{\mathbf{\Xi}}_H \xrightarrow{\mathbb{P}} \mathbf{\Xi}_H$, where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability.

The proposed JPR is given by

$$\hat{Y}_T(H) \pm P\left[\sqrt{\frac{\chi^2_{h,1-\alpha}}{h}}\right]_{h=1}^{H}, \tag{16}$$

where P is the lower-triangular Cholesky decomposition of $\hat{\mathbf{\Xi}}_H/T$, satisfying $PP' = \hat{\mathbf{\Xi}}_H/T$, and the quantity to the right of P is an $H \times 1$ vector whose hth entry is given by $\sqrt{\chi^2_{h,1-\alpha}/h}$. This approach is problematic for several reasons.

First, assumption (15) implies that the conditional distribution of the vector of prediction errors $\hat{U}_T(H) := \hat{Y}_T(H) - Y_{T,H}$ is approximately multivariate normal with mean zero, at least for large T. This is unrealistic. The conditional distribution of a prediction error depends on the conditional distribution of the random variable to be predicted. If the latter distribution is non-normal, which is the case in many applications, then the former distribution is generally non-normal as well.

Second, assumption (15) implies in addition that the conditional distribution of the vector of (unscaled) prediction errors, $\hat{U}_T(H) = \hat{Y}_T(H) - Y_{T,H}$, converges weakly to a point mass at zero. This is unrealistic. Although, under mild regularity conditions, the difference between an estimator and the population parameter it estimates converges weakly to zero (i.e., the estimator is consistent), the same is not true for a vector of prediction errors. Even if all model parameters are known, a future observation cannot be predicted perfectly because of its random nature.

Remark 4. To illustrate the first two points, consider the simple AR(1) model

$$y_t = \nu + \rho y_{t-1} + \epsilon_t, \tag{17}$$

where $|\rho| < 1$ and the errors $\{\epsilon_t\}$ are i.i.d. with mean zero and finite variance $\sigma^2_\epsilon > 0$. At time T, the forecast of y_{T+1} is given by

$$\hat{y}_T(1) := \hat{\nu} + \hat{\rho} y_T, \tag{18}$$

[3] They use the term *JCR* instead of *JPR*.

DOI: 10.1111/jtsa.12099

where $\hat{\nu}$ and $\hat{\rho}$ are suitable, consistent estimators of ν and ρ, respectively. The forecast error is given by

$$\hat{u}_T(1) = \hat{\nu} + \hat{\rho} y_T - y_{T+1}. \tag{19}$$

As T tends to infinity, the conditional distribution of $\hat{u}_T(1)$ converges weakly to the unconditional distribution of $-\epsilon_{T+1}$ (which does not depend on T). This distribution is neither necessarily normal, nor is it a point mass at zero. As a result, assumption (15) does not hold in this simple example.

Third, Jordà and Marcellino (2010) initially consider the following rectangular JPR:

$$\hat{Y}_T(H) \pm P \left[\sqrt{\frac{\chi^2_{H,1-\alpha}}{H}} \cdot \mathbf{1}_H \right], \tag{20}$$

where $\mathbf{1}_H$ is an $H \times 1$ vector of 1s. It is derived by an application of Bowden's (1970) lemma to an elliptical JPR based on Scheffé's (1953, 1959) method:

$$\left\{ \tilde{Y} : T \left(\hat{Y}_T(H) - \widetilde{Y} \right)' \hat{\boldsymbol{\Xi}}_H^{-1} \left(\hat{Y}_T(H) - \tilde{Y} \right) \leq \chi^2_{H,1-\alpha} \right\}. \tag{21}$$

As we have explained earlier, deriving a rectangular joint confidence region from an initial JCR of elliptical form is suboptimal in terms of the volume of the rectangular joint confidence region. Furthermore, it would appear that the application of Bowden's (1970) lemma is incorrect. For the special case when $\boldsymbol{\Xi}_H$ is a multiple of the H-dimensional identity matrix, it can be seen that a projection of the elliptical region (21) on the axes yields the rectangular region

$$\hat{Y}_T(H) \pm P \left[\sqrt{\chi^2_{H,1-\alpha}} \cdot \mathbf{1}_H \right], \tag{22}$$

so that the division of $\chi^2_{H,1-\alpha}$ by H in the region (20) is erroneous; for example, see Figure A.1.

Fourth, Jordà and Marcellino (2010) arrive at their final JPR (16) by 'refining' the initial JPR (20) by a step-down recursive procedure that is heuristic and lacks a theoretical justification.

Fifth, a counter-intuitive feature of the JPR (16) is that its width is not necessarily (weakly) monotonically increasing in the forecast horizon h; for an example, see Section 5.1. The reason is that the multipliers $\sqrt{\chi^2_{h,1-\alpha}/h}$ are strictly monotonically decreasing in h, at least for commonly used values of α, as illustrated in Figure A.1 of the online Supporting Information.

Sixth, Staszewska-Bystrova (2013) shows that if the matrix P contains negative entries, it must be replaced by $|P|$ in the various joint prediction regions studied by Jordà and Marcellino (2010), and in particular in their final proposal, namely in the region (16). In Monte Carlo simulations, we will therefore consider the *modified* Scheffé JPR, given by

$$\hat{Y}_T(H) \pm |P| \left[\sqrt{\frac{\chi^2_{h,1-\alpha}}{h}} \right]_{h=1}^{H}. \tag{23}$$

This modified JPR is also promoted by Jordà *et al.* (2014).

Staszewska-Bystrova (2011) proposes an alternative bootstrap method to construct a JPR for $Y_{T,H}$ that controls the FWE. In a nutshell, the method works as follows. Conditional on the observed data, one generates B bootstrap path forecasts $\hat{Y}_T^{*,b}(H)$, for $b = 1, \ldots, B$. One then discards αB of these bootstrap path forecasts: namely those $\hat{Y}_T^{*,b}(H)$ that are 'furthest' away from the original path forecast $\hat{Y}_T(H)$, where the distance between two $H \times 1$

DOI: 10.1111/jtsa.12099

vectors is measured by the Euclidian distance.[4] Finally, the JPR is defined as the envelope of the remaining $(1-\alpha)B$ bootstrap path forecasts, where the term *envelope* refers to the smallest region containing all remaining bootstrap path forecasts. Although this *neighbouring paths (NP)* method seems to perform quite well in simulation studies, there are several concerns.

First, the method is heuristic. No proof of asymptotic validity, under some suitable high-level assumption, is provided.

Second, the method seems restricted to (V)AR models, since it uses the backward representation of a (V)AR model to generate the bootstrap path forecasts; see Thombs and Schucany (1990) for an early use of this representation in AR models. As an additional restriction, a problem of the backward representation when the forward errors are non-normal is that even if the forward errors are independent, the backward errors are not independent, but merely uncorrelated; Pascual *et al.* (2001) point this out already. Hence, using Efron's (1979) bootstrap on the residuals in the backward representation, as proposed by Staszewska-Bystrova (2011), may not be generally valid.

Third, the method is in the spirit of Efron's (1979) percentile method, which amounts to 'looking up the wrong tails of a distribution'; see Hall (1992, Sections 1.3 and 3.4) for a discussion. Theoretical arguments suggest that such a method may only work well when the conditional distribution of the vector of forecast errors is symmetric around zero, as would be the case for a multivariate normal distribution. The performance of the method may suffer when prediction errors are, conditionally, skewed or have non-zero mean. Staszewska-Bystrova (2011) only considers normal errors with mean zero in the DGPs of her simulation study. On the other hand, the JPRs we propose in Sections 3.1 and 3.2 are based on Hall's percentile-t method, which has a sound theoretical foundation and is more generally valid than Efron's percentile method; again, see Hall (1992, Sections 1.3 and 3.4).

Fourth, since the JPR is given by the envelope of the $(1-\alpha)B$ non-discarded bootstrap path forecasts $\hat{Y}_T^{*,b}(H)$, the region typically has a jagged shape, which is unattractive; for an example, see Section 5.1.

Last but not least, it is not clear whether the methods of Jordà and Marcellino (2010) and Staszewska-Bystrova (2011) can be generalized to construct a JPR for $Y_{T,H}$ that controls the k-FWE for $k \geq 2$; see (6). By offering a method to construct rectangular JPRs for $Y_{T,H}$ that control the k-FWE for arbitrary $k \geq 1$, we provide applied researchers with a more flexible and versatile tool.

Remark 5 (Property of balance)**.** Under a mild additional assumption not covered by Assumption 1, our bootstrap joint prediction regions (8)–(10) can be easily seen to have the desirable property of being *balanced*, at least asymptotically.

A rectangular JPR for the future path $Y_{T,H}$ is *balanced* if the probability that y_{T+h} will be contained in its implied (simultaneous) prediction interval is the same for all $h = 1, \ldots, H$.[5]

To be specific, focus on the joint prediction region (8) whose implied prediction interval for y_{T+h} is given by $\left[\hat{y}_T(h) \pm d^{k\text{-max},*}_{|\cdot|,1-\alpha} \cdot \hat{\sigma}_T(h)\right]$. Then the probability

$$\mathbb{P}\left\{y_{T+h} \in \left[\hat{y}_T(h) \pm d^{k\text{-max},*}_{|\cdot|,1-\alpha} \cdot \hat{\sigma}_T(h)\right]\right\} \tag{24}$$

is the same for all $h = 1, \ldots, H$, asymptotically, under the additional assumption that the marginal distribution of

$$\frac{\hat{y}_T(h) - y_{T+h}}{\hat{\sigma}_T(h)} \tag{25}$$

[4] Staszewska-Bystrova (2011) also considers other distance measures but concludes that the Euclidean distance seems to work best.
[5] For a discussion of the concept of balance in the alternative contexts of JCRs and multiple testing, the reader is referred to Beran (1988a,1988b) and Romano and Wolf (2010).

DOI: 10.1111/jtsa.12099

is the same for all $h = 1, \ldots, H$, asymptotically. For example, this additional assumption holds if the time series $\{y_1, \ldots, y_T, y_{T+1}, \ldots, y_{T+H}\}$ is generated by an ARIMA model with i.i.d. errors, for any reasonable model-based way to compute the forecasts $\hat{y}_T(h)$ and the prediction standard errors $\hat{\sigma}_T(h)$.

A JPR that is balanced implicitly treats all forecasts $\hat{y}_T(h)$ as equally important, since the probability that the k-FWE criterion will be violated is evenly spread out over all forecast horizons h.

Another way to argue that balance is a desirable property is to consider the following (extremely) unbalanced JPR for $Y_{T,H}$:

$$\mathrm{PI}_T(1) \times (-\infty, \infty) \times \ldots \times (-\infty, \infty), \tag{26}$$

where $\mathrm{PI}_T(1)$ is a marginal prediction interval for y_{T+1} with level $1 - \alpha$. Although this JPR is clearly perverse, it nevertheless has the property of containing the entire future path $Y_{T,H}$ with the desired probability $1 - \alpha$, asymptotically [as long as the prediction interval $\mathrm{PI}_T(1)$ has the property of containing y_{T+1} with probability $1 - \alpha$, asymptotically].

It is not clear whether the property of balance can be established for the JPRs proposed by Jordà and Marcellino (2010) and Staszewska-Bystrova (2011).

4. MONTE CARLO SIMULATIONS

This section compares the finite-sample performance of various methods to construct JPRs. We restrict ourselves to univariate forecast procedures. To this end, we use AR(p) models with various lag lengths p that are first assumed to be known. Later, this assumption is relaxed, and p is chosen in a data-dependent fashion. Importantly, we also generate data from non-AR(p) models, thereby allowing for model misspecification, which is very relevant to practical work.

Before we present the details of the Monte Carlo set-up, we need to be specific about how we estimate the model, compute the prediction standard errors, and generate the bootstrap data in the calculation of our bootstrap JPRs.

4.1. Preliminaries

The general AR(p) model is given by

$$y_t = \nu + \rho_1 y_{t-1} + \ldots + \rho_p y_{t-p} + \epsilon_t, \tag{27}$$

where the errors $\{\epsilon_t\}$ are i.i.d. with mean zero and finite variance σ_ϵ^2. It can be alternatively expressed as

$$y_t = \nu + \rho y_{t-1} + \psi_1 \Delta y_{t-1} + \ldots + \psi_{p-1} \Delta y_{t-p+1} + \epsilon_t, \tag{28}$$

to bring out the role of the largest AR root $\rho := \rho_1 + \ldots + \rho_p$. Here, Δ is the first-difference operator; that is, $\Delta y_t := y_t - y_{t-1}$. The parameters of formulations (27) and (28) are related by

$$\rho_1 = \rho + \psi_1, \quad \rho_j = -\psi_{j-1} + \psi_j \quad \text{for} \quad 2 \le j \le p-1, \quad \rho_p = -\psi_{p-1}. \tag{29}$$

The usefulness of bias-corrected estimators when making forecasts based on AR(p) models has been long recognized and goes back to Kilian (1998).[6]

[6] Kilian (1998) considers the construction of confidence intervals for impulse response functions, not the construction of prediction intervals for future observations. However, his bias correction has since been successfully applied to the latter problem as well; for example, see Clements and Taylor (2001).

DOI: 10.1111/jtsa.12099

Let $\hat{\rho}_{\mathrm{OLS}}$ denote the usual OLS estimator of ρ based on formulation (28). We employ the following bias-corrected estimator of ρ.

$$\hat{\rho}_{\mathrm{BC}} := \hat{\rho}_{\mathrm{OLS}} + \frac{1 + 3\,\hat{\rho}_{\mathrm{OLS}}}{T}; \tag{30}$$

for example, see White (1961). The corresponding bias-corrected estimators of $(\nu, \psi_1, \ldots, \psi_{p-1})$ are obtained by regressing $y_t - \hat{\rho}_{\mathrm{BC}} y_{t-1}$ on $(1, \Delta y_{t-1}, \ldots, \Delta y_{t-p-1})$ via OLS. By relation (29), we obtain in turn the bias-corrected estimators of formulation (27), denoted by $(\hat{\nu}_{\mathrm{BC}}, \hat{\rho}_{1,\mathrm{BC}}, \ldots, \hat{\rho}_{p,\mathrm{BC}})$.[7]

The corresponding, centred residuals $\hat{\epsilon}_t$, for $p + 1 \leq t \leq T$, are obtained as follows.

$$\hat{\epsilon}_t := \hat{\epsilon}_{t,\mathrm{BC}} - \frac{1}{T-p} \sum_{l=p+1}^{T} \hat{\epsilon}_{l,\mathrm{BC}} \quad \text{with} \quad \hat{\epsilon}_{t,BC} := y_t - \hat{\nu}_{BC} - \hat{\rho}_{1,\mathrm{BC}} \cdot y_{t-1} - \ldots - \hat{\rho}_{p,\mathrm{BC}} \cdot y_{t-p}. \tag{31}$$

The residual variance is

$$\hat{\sigma}_{\epsilon}^2 := \frac{1}{T - 2p - 1} \sum_{t=p+1}^{T} \hat{\epsilon}_t^2, \tag{32}$$

where the number of estimated parameters, $p + 1$, is subtracted from the 'sample size' of the residuals, $T - p$, in the numerator in the spirit of the usual definition of the residual variance in a linear regression model.

The forecasts $\hat{y}_T(h)$ are computed in the usual fashion.

The prediction standard errors $\hat{\sigma}_T(h)$ are computed in the usual Box–Jenkins fashion. To this end, consider the MA(∞) representation that is equivalent to the AR(p) model with parameters $(\hat{\nu}_{\mathrm{BC}}, \hat{\rho}_{1.\mathrm{BC}}, \ldots, \hat{\rho}_{p,\mathrm{BC}})$, and denote the parameters of this MA(∞) model by $\left(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \ldots\right)$, with $\hat{\theta}_0 := 1$. Then compute

$$\hat{\sigma}_T(h) := \hat{\sigma}_{\epsilon} \sqrt{\hat{\theta}_0^2 + \ldots + \hat{\theta}_{h-1}^2}\,. \tag{33}$$

Remark 6. It is well known that the usual Box–Jenkins prediction standard errors (33) are somewhat too small in magnitude in finite samples, as they do not account for the estimation uncertainty in the model parameters $\left(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \ldots\right)$. However, this is not really a problem for our bootstrap approach as long as we use the same method to compute the bootstrap prediction standard errors; see equation (35). Since the bias contained in the prediction standard errors is, approximately, the same in the real world compared with the bootstrap world, the resulting mistakes, approximately, cancel out, and one still obtains JPRs with very good finite-sample properties (Section 4.3).

Bootstrap data $\left\{y_1^*, \ldots, y_T^*, y_{T+1}^*, \ldots, y_{T+H}^*\right\}$ are generated according to Pascual *et al.* (2001) as follows.
First, draw $\epsilon_{p+1}^*, \ldots, \epsilon_{T+H}^*$ i.i.d. from the empirical distribution of $\hat{\epsilon}_{p+1}, \ldots, \hat{\epsilon}_T$.
Second, let $y_t^* := y_t$, for $1 \leq t \leq p$, and then

$$y_t^* := \hat{\nu}_{\mathrm{BC}} + \hat{\rho}_{1,\mathrm{BC}} \cdot y_{t-1}^* + \ldots + \hat{\rho}_{p,\mathrm{BC}} \cdot y_{t-p}^* + \epsilon_t^*, \quad \text{for} \quad p + 1 \leq t \leq T. \tag{34}$$

Third, generate y_t^*, for $T + 1 \leq t \leq T + H$, similarly to (34), but conditional on $\{y_{t-p+1}, \ldots, y_T\}$ rather than on $\{y_{t-p+1}^*, \ldots, y_T^*\}$.

[7] Of course, other bias corrections can be employed as well, such as the bootstrap bias correction of Kilian (1998) or the analytic bias correction of Roy and Fuller (2001), although the reader is referred to an errata concerning the latter reference.

DOI: 10.1111/jtsa.12099

An implication of the method of Pascual *et al.* (2001) is that the stretch $\{y^*_{T+1}, \ldots, y^*_{T+H}\}$ is not a continuation of the stretch $\{y^*_1, \ldots, y^*_T\}$. This feature appears counter-intuitive at first, but it allows for bootstrap forecasts conditional on the (relevant) past of the original data rather than on the (relevant) past of the bootstrap data, which is clearly desirable.

Remark 7. Thombs and Schucany (1990) propose an alternative method to generate bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$, based on the backward representation of an AR(p) model. Their method ensures that $y^*_t = y_t$, for $t - p + 1 \leq t \leq p$, so that the stretch $\{y^*_{T+1}, \ldots, y^*_{T+H}\}$ is also a continuation of the stretch $\{y^*_1, \ldots, y^*_T\}$. However, it only applies to AR(p) models with normal forward errors ϵ_t. The method of Pascual *et al.* (2001) applies much more widely, in particular to AR(p) models with possibly non-normal forward errors ϵ_t. Since the assumption of normal forward errors ϵ_t is often violated in practice, we opt for the method of Pascual *et al.* (2001).

Denote the bias-corrected estimators of $(\nu, \rho_1, \ldots, \rho_p)$ computed from the stretch $\{y^*_1, \ldots, y^*_T\}$ by $\left(\hat{\nu}^*, \hat{\rho}^*_{1,\mathrm{BC}}, \ldots, \hat{\rho}^*_{p,\mathrm{BC}}\right)$.

The bootstrap residual variance $\hat{\sigma}^{2,*}_\epsilon$ is computed similarly to (32).

The bootstrap forecasts $\hat{y}^*_T(h)$ are computed conditional on $\{y_{t-p+1}, \ldots, y_T\}$ rather than on $\{y^*_{t-p+1}, \ldots, y^*_T\}$.

The bootstrap prediction standard errors $\hat{\sigma}^*_T(h)$ are computed in the same way as the 'original' prediction standard errors $\hat{\sigma}^*_T(h)$. To this end, consider the MA(∞) representation that is equivalent to the AR(p) model with parameters $\left(\hat{\nu}^*_{\mathrm{BC}}, \hat{\rho}^*_{1,\mathrm{BC}}, \ldots, \hat{\rho}^*_{p,\mathrm{BC}}\right)$, and denote the parameters of this MA(∞) model by $\left(\hat{\theta}^*_0, \hat{\theta}^*_1, \hat{\theta}^*_2, \ldots\right)$, with $\hat{\theta}^*_0 := 1$. Then compute

$$\hat{\sigma}^*_T(h) := \hat{\sigma}^*_\epsilon \sqrt{\left(\hat{\theta}^*_0\right)^2 + \ldots + \left(\hat{\theta}^*_{h-1}\right)^2}. \tag{35}$$

Remark 8. Pan and Politis (2014) propose a number of backward and forward bootstrap procedures for prediction intervals under different AR models and base their prediction intervals on the concept of prediction roots (similar to confidence roots) to be distinguished from the so-called percentile method. They also give extensive references to the current literature. In particular, Pan and Politis (2014) – as well as Politis (2013) – make a strong case for the use of predictive residuals (as opposed to fitted residuals $\hat{\epsilon}_t$) in resampling. However, the Monte Carlo evidence of Pan and Politis (2014) indicates that noticeable benefits are only realized for non-studentized roots. Since our methodology corresponds to using a studentized root, we stick with the fitted residuals for simplicity.

4.2. Monte Carlo Design

First, we consider an AR(1) model with $\nu = 0$ and with $\rho := \rho_1 \in \{0.9, 0.5, -0.5, -0.9\}$. The order $p = 1$ is assumed to be known. The sample size is $T \in \{100, 400\}$. The errors ϵ_t are i.i.d. according to one of the following three distributions.

- $(\epsilon_t \sim N(0, 1))$, standard normal.
- $(\epsilon_t \sim t_3)$, a t-distribution with three degrees of freedom, standardized to have a variance 1.
- $(\epsilon_t \sim \chi^2_3)$, a chi-square distribution with three degrees of freedom, centred to have mean 0 and standardized to have variance 1.

Second, we consider an AR(2) model with $\nu = 0$ and $(\rho_1, \rho_2) \in \{(1.85, -0.75), (1.25, -0.75), (-0.65, 0.15), (-0.7, -0.2)\}$. The order $p = 2$ is assumed to be known. The sample size is $T \in \{100, 400\}$. The errors ϵ_t are i.i.d. according to one of the three listed distributions.

DOI: 10.1111/jtsa.12099

Third, we consider an AR(2) model with $\nu = 0$ and $(\rho_1, \rho_2) \in \{(1.85, -0.75), (1.25, -0.75), (-0.65, 0.15), (-0.7, -0.2)\}$. The order $p = 2$ is assumed to be unknown and is estimated from the data using the Bayesian information criterion (BIC) optimizing over the set $\{1, 2, \ldots, 9, 10\}$.[8] This is the case both in the 'real' world and in the bootstrap world.[9] The sample size is $T \in \{100, 400\}$. For compactness, we only consider errors ϵ_t that are i.i.d. standard normal.

Fourth, we consider an MA(1) model

$$y_t = \mu + \epsilon_1 + \theta \epsilon_{t-1}, \tag{36}$$

with $\mu = 0$ and $\theta \in \{0.9, -0.5, 0.5, 0.9\}$. For compactness, we only consider errors ϵ_t that are i.i.d. standard normal.

Fifth, we consider the nonlinear threshold AR model used by Montgomery *et al.* (1998, Section 2.3) in modelling the US unemployment rate. This model is given by

$$y_t = \begin{cases} 0.01 + 0.73\, y_{t-1} + 0.10\, y_{t-2} + 0.28\, \epsilon_t & \text{if } y_{t-2} \leq 0.1 \\ 0.18 + 0.80\, y_{t-1} - 0.56\, y_{t-2} + 0.41\, \epsilon_t & \text{otherwise} \end{cases}, \tag{37}$$

where the errors ϵ_t are white noise with mean 0 and variance 1. For compactness, we only consider errors ϵ_t that are i.i.d. standard normal.

Importantly, in the last two cases, we employ the same methodology to compute JPRs as in the third case: based on an AR(p) model with the lag order estimated using the BIC. Therefore, the model to compute JPRs is actually misspecified, which happens often in applied work.

The following four methods to construct JPRs are compared.

- **Joint marginals**: String together H marginal, two-sided symmetric bootstrap prediction intervals for y_{T+h}, each with nominal coverage level $1 - \alpha$
- **Scheffé**: The modified 'asymptotic' Scheffé JPR (23) of Jordà and Marcellino (2010), employing the absolute-value correction of Staszewska-Bystrova (2013)
- **NP heuristic**: The NP heuristic bootstrap JPR of Staszewska-Bystrova (2011)
- **k-FWE JPR**: Our two-sided bootstrap JPR (8).

For a fair comparison, we use the bias correction described in Section 4.1 with all methods.

The nominal k-FWE level is $\alpha = 0.1$. We consider $k \in \{1, 2, 3\}$ for k-FWE JPR. All other methods only use $k = 1$. The forecast horizon is $H \in \{6, 12, 24\}$. The number of bootstrap samples for k-FWE JPR and NP heuristic is $B = 1000$ always.

All empirical coverages are computed from 1000 generated data sets $\{y_1, \ldots, y_T\}$, each with 100 corresponding, independent continuations $\{y_{T+1}, \ldots, y_{T+H}\}$. As a result, the empirical coverages are based on a total of 100,000 repetitions each and are thus highly accurate. For each scenario, we report the proportion of times that the k-FWE criterion is not violated. The thus-obtained empirical coverages should be compared with the nominal coverage level given by $1 - 0.1 = 0.9 = 90\%$.

For the cases where the lag order p is estimated using the BIC, which are the most relevant cases for applied work, we also compute empirical geometric-average widths of the various JPRs to compare their 'volumes'. More specifically, any rectangular JPR is the Cartesian product of H simultaneous prediction intervals $\mathrm{PI}_1, \ldots, \mathrm{PI}_H$, with corresponding widths $w_1, \ldots, w_H$. The geometric average of the H widths is given by

[8] The BIC is known to be a consistent information criterion, unlike the Akaike information criterion, say. Therefore, in terms of Assumption 1, using the BIC to estimate the order of an AR(p) model is asymptotically equally valid as using the true order.

[9] As a result, it is possible that a different order is used in the 'real' world compared with the bootstrap world.

DOI: 10.1111/jtsa.12099

$$\bar{w}_{\text{geo}} := \left(\prod_{h=1}^{H} w_h\right)^{\frac{1}{H}}$$

and is a one-to-one function of the volume of the joint prediction region, $\prod_{i=1}^{H} w_i$. (The reason that we record the geometric average instead is that this makes it easier to study a 'typical' width as we increase H, keeping everything else the same.) For a given scenario, we report the sample mean of the $\bar{w}_{\text{geo}}$ over all 1000 simulations as the empirical geometric-average width.

4.3. Results

The results for the AR(1) model with $p = 1$ known are presented in Tables A.1 and A.2 of the online Supporting Information. The results for the AR(2) model with $p = 2$ known are presented in Tables A.3 and A.4 of the online Supporting Information. The results for the AR(2) model with $p = 2$ unknown and estimated using the BIC are presented in Tables I and II. The results for the MA(1) model are presented in Tables III and IV. Finally, the results for the threshold AR model are presented in Tables V and VI.

The various results concerning empirical coverage can be summarized as follows.

- Joint marginals always undercovers, and its performance becomes worse as the maximum forecast horizon H increases. This behaviour is as expected and has been demonstrated before by Jordà and Marcellino (2010) and Staszewska-Bystrova (2011) already.

Table I. AR(2) model, Bayesian information criterion order selection: empirical coverages

	Nominal coverage $1-\alpha = 90\%$					
	$T = 100, \epsilon_t \sim \text{N}(0,1)$			$T = 400, \epsilon_t \sim \text{N}(0,1)$		
	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
$(\rho_1, \rho_2) = (1.75, -0.85)$						
Joint marginals	72.9	62.5	49.5	76.1	64.0	47.2
Scheffé	88.2	87.9	87.5	89.2	89.2	89.1
NP heuristic	89.5	93.2	95.0	89.7	90.6	90.4
1-FWE JPR	90.4	90.5	89.6	89.8	89.7	89.7
2-FWE JPR	90.4	89.8	89.7	89.9	89.8	89.7
3-FWE JPR	90.0	90.3	89.0	90.0	89.7	89.6
$(\rho_1, \rho_2) = (1.25, -0.75)$						
Joint marginals	64.2	46.5	27.2	65.0	46.7	25.0
Scheffé	86.1	85.3	84.2	87.3	87.0	86.7
NP heuristic	88.2	87.4	86.0	88.8	87.4	85.5
1-FWE JPR	90.0	89.4	89.3	89.9	89.8	89.9
2-FWE JPR	90.2	89.5	89.5	89.9	89.9	89.8
3-FWE JPR	89.8	89.5	89.3	89.9	89.8	89.7
$(\rho_1, \rho_2) = (-0.65, 0.15)$						
Joint marginals	65.2	49.2	30.2	64.7	47.3	26.3
Scheffé	85.6	83.8	80.1	87.0	86.6	85.7
NP heuristic	89.2	88.2	86.9	89.2	88.0	86.1
1-FWE JPR	90.4	90.1	89.7	90.0	90.0	89.7
2-FWE JPR	90.5	89.9	89.8	90.1	90.0	90.0
3-FWE JPR	89.7	89.7	89.6	90.0	89.8	89.8
$(\rho_1, \rho_2) = (-0.7, -0.2)$						
Joint marginals	59.7	39.3	17.8	60.0	37.3	14.6
Scheffé	81.0	72.9	57.1	82.1	71.0	48.9
NP heuristic	88.1	87.4	85.5	88.7	87.6	85.5
1-FWE JPR	89.4	89.3	88.7	89.9	89.8	89.8
2-FWE JPR	89.2	89.4	89.8	90.0	90.0	90.0
3-FWE JPR	89.4	89.7	89.8	90.0	90.1	89.9

FWE, familywise error rate; JPR, joint prediction region; NP, neighbouring paths.

DOI: 10.1111/jtsa.12099

Table II. AR(2) model, Bayesian information criterion order selection: empirical geometric-average widths

	Nominal coverage $1-\alpha = 90\%$					
	$T = 100, \epsilon_t \sim \mathrm{N}(0,1)$			$T = 400, \epsilon_t \sim \mathrm{N}(0,1)$		
	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
$(\rho_1, \rho_2) = (1.75, -0.85)$						
Joint marginals	9.5	13.7	17.2	9.3	12.9	15.7
Scheffé	14.7	24.3	35.1	14.6	24.1	35.3
NP heuristic	13.2	23.0	35.5	12.4	18.8	25.1
1-FWE JPR	12.5	19.5	26.1	11.8	17.8	24.1
2-FWE JPR	11.3	18.1	24.8	10.6	16.6	22.9
3-FWE JPR	10.1	17.1	23.1	9.5	15.3	21.4
$(\rho_1, \rho_2) = (1.25, -0.75)$						
Joint marginals	5.4	6.1	6.7	5.4	6.1	6.6
Scheffé	8.0	10.8	13.4	8.0	10.7	13.3
NP heuristic	7.6	9.48	11.2	7.48	9.24	10.8
1-FWE JPR	7.7	9.6	11.4	7.5	9.4	11.1
2-FWE JPR	6.4	8.3	10.1	6.2	8.1	9.7
3-FWE JPR	5.1	7.1	9.0	4.9	6.9	8.6
$(\rho_1, \rho_2) = (-0.65, 0.15)$						
Joint marginals	4.4	4.8	5.1	4.3	4.7	4.9
Scheffé	6.6	8.3	9.5	6.6	8.4	9.9
NP heuristic	6.2	7.4	8.5	6.0	7.1	8.1
1-FWE JPR	6.2	7.4	8.6	6.0	7.2	8.2
2-FWE JPR	4.9	6.2	7.5	4.7	6.0	7.2
3-FWE JPR	4.0	5.4	6.7	3.9	5.3	6.5
$(\rho_1, \rho_2) = (-0.7, -0.2)$						
Joint marginals	4.0	4.2	4.2	4.0	4.1	4.1
Scheffé	5.4	5.7	5.7	5.3	5.4	5.3
NP heuristic	5.7	6.4	7.1	5.6	6.3	6.9
1-FWE JPR	5.7	6.5	7.1	5.6	6.4	7.0
2-FWE JPR	4.4	5.3	6.1	4.3	5.2	5.9
3-FWE JPR	3.4	4.4	5.3	3.7	4.3	5.2

FWE, familywise error rate; JPR, joint prediction region; NP, neighbouring paths.

Nevertheless, it is worth repeating the underlying message one more time: stringing together marginal prediction intervals does not result in a valid JPR for the entire future path.

- The performance of Scheffé ranges from acceptable to poor. In general, the performance decreases in the maximum forecast horizon H. For many scenarios, the empirical coverage is unacceptably far away from the nominal level and can even fall below 50%.

  As a consequence, Scheffé cannot be recommended for practical application.

  It should be pointed out that the original Scheffé method (16), without the absolute-value correction of Staszewska-Bystrova (2013), performs even worse and can have coverage probability near 0 when the matrix P contains negative entries.[10]
- The performance of NP heuristic is good when the largest AR root is close to 0. Otherwise, the performance is acceptable: the empirical coverage generally is somewhat less than the nominal level and decreases in the maximum forecast horizon H, even for $T = 400$.
- The performance of k-FWE JPR is the best of all methods; it ranges from very good to good. There can be some mild undercoverage when $T = 100$ and $H = 24$; but in almost all cases, the empirical coverage is very close to the nominal level. In particular, the performance is remarkably stable, both over the maximum forecast horizon H and over the value of k in the k-FWE criterion.

[10] Corresponding results are available from the authors on request.

DOI: 10.1111/jtsa.12099

Table III. MA(1) model, Bayesian information criterion order selection: empirical coverages

	Nominal coverage $1-\alpha = 90\%$					
	$T = 100, \epsilon_t \sim \mathrm{N}(0,1)$			$T = 400, \epsilon_t \sim \mathrm{N}(0,1)$		
	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
$\theta = 0.9$						
Joint marginals	52.1	32.3	13.2	50.9	30.1	10.7
Scheffé	70.5	62.0	50.4	71.5	62.1	44.4
NP heuristic	85.3	84.6	82.6	85.0	84.7	82.8
1-FWE JPR	88.1	89.1	88.3	89.3	89.6	89.3
2-FWE JPR	89.0	89.2	89.0	89.4	89.5	89.7
3-FWE JPR	89.4	89.6	89.7	89.8	89.7	89.8
$\theta = 0.5$						
Joint marginals	52.9	31.5	12.0	52.2	29.9	9.9
Scheffé	68.9	54.8	38.0	70.1	53.6	29.1
NP heuristic	86.4	84.9	81.5	86.5	85.8	83.3
1-FWE JPR	88.7	89.0	87.5	89.6	89.9	89.1
2-FWE JPR	89.5	89.3	89.3	89.8	89.7	89.0
3-FWE JPR	89.9	90.0	89.6	89.0	89.9	89.7
$\theta = -0.5$						
Joint marginals	53.4	32.1	12.1	52.4	30.0	10.1
Scheffé	68.6	52.3	31.0	70.3	53.2	27.3
NP heuristic	85.9	84.0	79.6	86.5	85.8	83.2
1-FWE JPR	88.7	88.9	87.5	89.5	89.8	89.1
2-FWE JPR	90.0	89.1	89.3	89.8	89.7	90.0
3-FWE JPR	90.2	89.9	89.6	90.0	90.0	89.7
$\theta = -0.9$						
Joint marginals	52.3	32.6	13.5	51.4	30.7	11.1
Scheffé	70.3	60.0	43.1	71.9	62.0	42.6
NP heuristic	84.4	83.5	80.6	84.7	84.7	82.7
1-FWE JPR	88.1	89.0	88.5	89.4	89.6	89.2
2-FWE JPR	89.3	89.1	89.6	89.8	89.6	89.8
3-FWE JPR	89.8	89.5	89.9	89.9	89.8	89.8

FWE, familywise error rate; JPR, joint prediction region; NP, neighbouring paths.

- When the distribution of the errors is heavy tailed or skewed, as opposed to normal, the performance of all methods generally suffers somewhat. However, it suffers less for k-FWE JPR compared with NP heuristic, which is the only competitor with acceptable coverage properties.
- There is no noticeable penalty to not knowing the AR model order p. When p is estimated from the data using the BIC, the empirical coverages are generally close to the corresponding coverages for known p, even for $T = 100$ already.
- When the AR(p) forecasting model is misspecified, the ranking of the various methods remains unchanged. Furthermore, k-FWE JPR is more robust to model misspecification than NP heuristic. This is an important finding for applied researchers: k-FWE JPR based on the AR(p) forecasting model has the potential to work well for general stationary time series, even for nonlinear ones.

The various results concerning empirical geometric-average widths can be summarized as follows.

- As expected, joint marginals has the smallest geometric-average width throughout.
- Despite its poor coverage properties, Scheffé has larger geometric-average width than NP heuristic or 1-FWE JPR in some scenarios.
- NP heuristic has generally somewhat smaller geometric-average width than 1-FWE JPR, which is perfectly in line with its generally somewhat smaller coverage probabilities. In other words, adjusted for coverage, there does not appear any noticeable difference in geometric-average width between the two methods.
- As expected, the geometric-average width of k-FWE JPR decreases in k, and the differences are quite large. As a result, there really is a pay-off in terms of 'volume' of the JPR if the applied researcher is willing to choose a value of k larger than 1.

DOI: 10.1111/jtsa.12099

Table IV. MA(1) model, Bayesian information criterion order selection: empirical geometric-average widths

	Nominal coverage $1-\alpha = 90\%$					
	$T = 100, \epsilon_t \sim \mathrm{N}(0,1)$			$T = 400, \epsilon_t \sim \mathrm{N}(0,1)$		
	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
$\theta = 0.9$						
Joint marginals	4.4	4.5	4.5	4.3	4.4	4.4
Scheffé	5.7	6.0	6.2	5.6	5.8	5.8
NP heuristic	6.3	7.1	7.7	6.1	6.9	7.5
1-FWE JPR	6.2	7.1	7.8	6.0	6.9	7.6
2-FWE JPR	4.8	5.7	6.5	4.6	5.5	6.3
3-FWE JPR	3.7	4.8	5.7	3.5	4.6	5.5
$\theta = 0.5$						
Joint marginals	3.7	3.7	3.7	3.6	3.7	3.7
Scheffé	4.4	4.6	4.7	4.4	4.4	4.4
NP heuristic	5.3	5.8	6.3	5.2	5.7	6.2
1-FWE JPR	5.3	5.9	6.4	5.1	5.8	6.3
2-FWE JPR	3.9	4.7	5.3	3.8	4.5	5.2
3-FWE JPR	3.0	3.9	4.7	2.9	3.8	4.5
$\theta = -0.5$						
Joint marginals	3.7	3.7	3.7	3.6	3.7	3.7
Scheffé	4.4	4.4	4.4	4.4	4.4	4.3
NP heuristic	5.2	5.7	6.2	5.2	5.7	6.2
1-FWE JPR	5.2	5.9	6.4	5.1	5.8	6.3
2-FWE JPR	3.9	4.6	5.3	3.8	4.5	5.2
3-FWE JPR	3.0	3.9	4.6	2.9	3.8	4.5
$\theta = -0.9$						
Joint marginals	4.4	4.5	4.5	4.3	4.4	4.4
Scheffé	5.6	5.8	5.8	5.6	5.8	5.7
NP heuristic	6.2	6.9	7.5	6.1	6.8	7.4
1-FWE JPR	6.2	7.0	7.7	6.0	6.9	7.5
2-FWE JPR	4.7	5.6	6.4	4.6	5.5	6.3
3-FWE JPR	3.6	4.7	5.6	3.5	4.6	5.5

FWE, familywise error rate; JPR, joint prediction region; NP, neighbouring paths.

Table V. Threshold AR model, Bayesian information criterion order selection: empirical coverages

	Nominal coverage $1-\alpha = 90\%$					
	$T = 100, \epsilon_t \sim \mathrm{N}(0,1)$			$T = 400, \epsilon_t \sim \mathrm{N}(0,1)$		
	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint marginals	63.6	44.8	23.7	63.3	43.4	20.8
Scheffé	84.3	80.9	76.0	86.0	83.9	79.5
NP heuristic	88.3	87.4	85.7	87.9	86.7	84.3
1-FWE JPR	89.1	88.5	88.8	89.7	89.2	89.0
2-FWE JPR	89.5	89.1	89.3	89.9	89.9	89.6
3-FWE JPR	90.0	90.2	90.9	90.2	90.4	90.5

FWE, familywise error rate; JPR, joint prediction region; NP, neighbouring paths.

5. EMPIRICAL APPLICATION

The goal of this section is to compare the various joint prediction regions for a set of real data. To this end, we downloaded quarterly data on US real gross domestic product from Q1/1947 until Q3/2011, made freely available by the Federal Reserve Bank of St Louis.[11] The data are seasonally adjusted and expressed in billions of chained 2005 dollars. Figure 1 displays the raw data as well as the first differences of the logarithmic data (in per cent).

[11] The data can be downloaded at http://research.stlouisfed.org/fred2/series/GDPC1/

DOI: 10.1111/jtsa.12099

Table VI. Threshold AR model, Bayesian information criterion order selection: empirical geometric-average widths

	Nominal coverage $1-\alpha = 90\%$					
	$T = 100, \epsilon_t \sim \mathrm{N}(0,1)$			$T = 400, \epsilon_t \sim \mathrm{N}(0,1)$		
	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint marginals	1.5	1.6	1.6	1.4	1.5	1.6
Scheffé	2.2	2.5	2.8	2.1	2.5	2.7
NP heuristic	2.1	2.5	2.9	2.1	2.4	2.7
1-FWE JPR	2.3	2.7	3.1	2.2	2.6	3.0
2-FWE JPR	1.8	2.3	2.7	1.7	2.2	2.5
3-FWE JPR	1.4	1.9	2.4	1.4	1.8	2.2

FWE, familywise error rate; JPR, joint prediction region; NP, neighbouring paths.

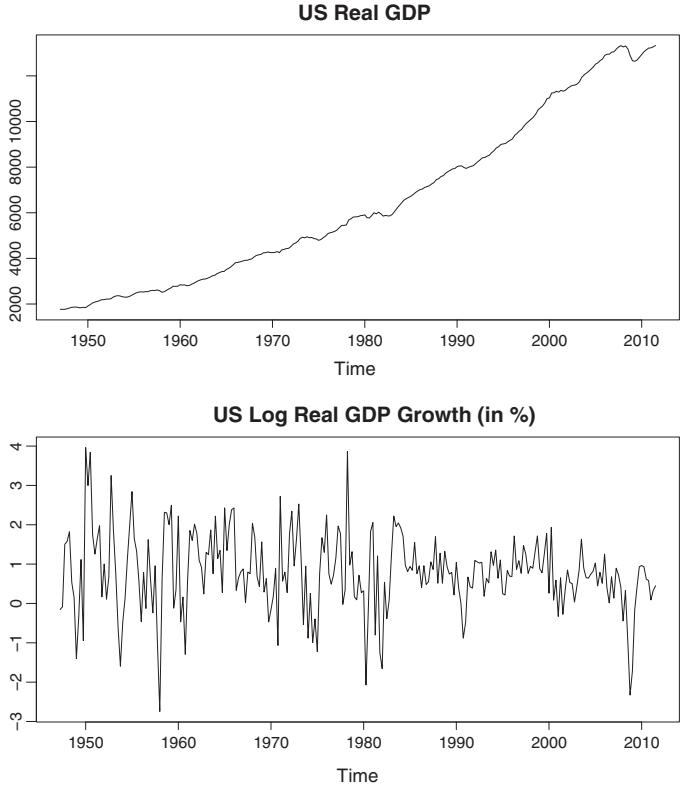

Figure 1. Quarterly data on US real gross domestic product (GDP; in 2005 chained dollars) from Q1/1947 until Q3/2001. The upper panel displays the raw data, and the lower panel displays the first differences of the logarithmic data (in per cent)

We take the latter series as our series of interest with a total of 258 observations. The task then is to forecast log quarter-to-quarter growth for the next H quarters and to compute corresponding joint prediction regions. We choose $H = 12$, which corresponds to a maximum forecast horizon of 3 years. The nominal coverage is given by $1-\alpha = 90\%$.

We use the AR(p) methodology described in Section 4 to compute bootstrap JPRs, where the lag order p is assumed to be unknown and estimated from the (bootstrap) data using the BIC. Of course, a more 'complex' methodology could be used instead, such as a multivariate forecasting model based on additional macroeconomic

DOI: 10.1111/jtsa.12099

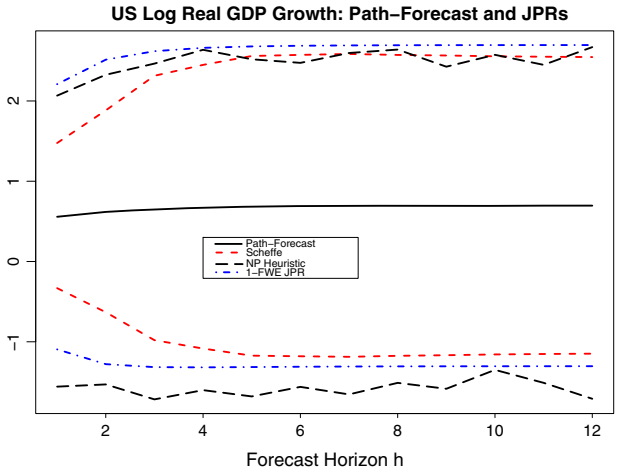

Figure 2. Path forecast and various joint prediction regions (JPRs) for US log real gross domestic product (GDP) growth. The forecast period ranges from Q4/2011 until Q3/2014. The nominal coverage is given by $1 - \alpha = 90\%$. FWE, familywise error rate; NP, neighbouring paths

variables (e.g., Stock and Watson, 2001) or a nonlinear forecasting model (e.g., Potter, 1995). The goal of this section, however, is not necessarily to find the single best forecasting model for the given data set but to see how the various JPRs behave relative to each other for a common, simple, and reasonable forecasting model, such as the AR(p) model.

5.1. Illustration Exercise

We first illustrate the salient features of the various JPRs by using the last $T = 120$ quarters (or 30 years) to forecast the not-yet-observed future path ranging from Q4/2011 until Q3/2014. We do not use the entire data set, since the assumption of stationarity is doubtful, given that the overall volatility seems to have decreased after 1980.

The lag order for the original data estimated by the BIC is $\hat{p} = 1$. The initial model fitted via OLS is given by

$$\hat{y}_{t+1} = 0.318 + 0.542 \cdot y_t. \tag{38}$$

Using the bias correction (30) yields the following final model used for forecasting purposes:

$$\hat{y}_{t+1} = 0.304 + 0.564 \cdot y_t. \tag{39}$$

Figure 2 compares Scheffé, NP heuristic, and 1-FWE JPR.[12] The main findings are as follows:

- Scheffé has a substantially smaller volume than the other two regions: this is not surprising given the simulation results of the previous section, where it was seen that Scheffé typically undercovers by a substantial amount.
- A further, counter-intuitive feature of Scheffé is that its width is non-monotonic in the forecasting horizon h: the width is largest for $h = 7$ and monotonically decreases after that, if only slightly. The theoretical reason for this counter-intuitive feature was discussed in Section 3.3.

[12] The number of bootstrap samples for NP heuristic and k-FWE JPR is $B = 10{,}000$.

DOI: 10.1111/jtsa.12099

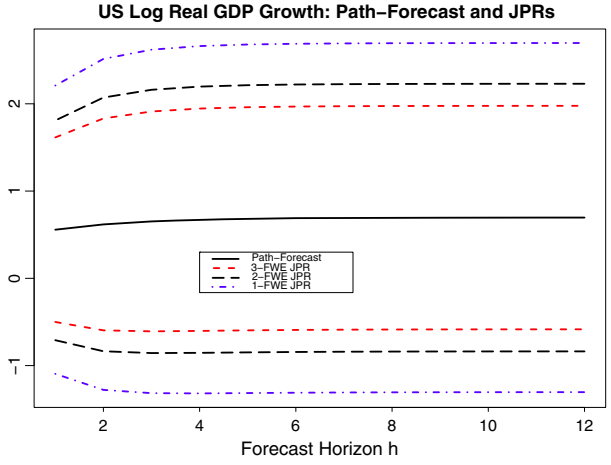

Figure 3. Path forecast and various joint prediction regions (JPRs) for US log real gross domestic product (GDP) growth. The forecast period ranges from Q4/2011 until Q3/2014. The nominal coverage is given by $1 - \alpha = 90\%$. FWE, familywise error rate; NP, neighbouring paths

- Although NP heuristic and 1-FWE JPR are comparable in terms of their volume, an unattractive feature of NP heuristic is its jagged shape, which is a result of the underlying methodology (Section 3.3).

Figure 3 compares 1-FWE JPR, 2-FWE JPR, and 3-FWE JPR. As implied by theory, the volume of k-FWE JPR decreases in the value of k. Therefore, if the applied researcher is willing to miss up to one (or two) elements of the future path in the JPR (with a prespecified probability of 90%), she or he obtains a smaller and more informative region in return.

5.2. Backtest Exercise

Although the previous exercise serves to illustrate the salient features of the various JPRs, it does not address their performance in terms of coverage. First, the data ranging from Q4/2011 until Q3/2014 have not been entirely observed yet (at the time of writing this article). Second, even when these data become eventually known, they only correspond to a single instance of a path; to compute meaningful empirical coverages, a large number of such paths are needed.

Therefore, we resort to the following backtest exercise, for a given method to construct a JPR designed to control the k-FWE:

- Using the stretch $\{y_t, \ldots, y_{t+119}\}$ only, compute the JPR for the next $H = 12$ periods.
- Compare the computed JPR against the path $(y_{t+120}, \ldots, y_{t+131})'$ to check whether all but at most $k - 1$ elements of the path are contained in the JPR. If the answer is yes, call the outcome a 'success'.
- Do this for $t = 1, \ldots, 258 - 120 - 12 = 126$.
- Report the empirical coverage as the fraction of 'successes' out of these 126 'trials'.

This means that we use a rolling window of 120 quarters to compute a JPR for the next path of $H = 12$ quarters. Since only 'past and present' information is used to forecast the 'future', we obtain a fair assessment of a method's out-of-sample performance in this way. Although the assessment is fair, it is not overly accurate, since the empirical coverage is based on 126 out-of-sample 'trials' only, which are not even independent of each other.

DOI: 10.1111/jtsa.12099

Table VII. Empirical out-of-sample coverages for US log real gross domestic product growth

Nominal coverage $1-\alpha = 90\%$	
Method	Empirical coverage
Joint marginals	64.3
Scheffé	73.0
NP heuristic	88.1
1-FWE JPR	89.7
2-FWE JPR	85.7
3-FWE JPR	87.3

The results are presented in Table VII.[13] It is seen that joint marginals and Scheffé undercover by a substantial amount, while NP heuristic and k-FWE JPR perform very well to well. These findings are line with those of the Monte Carlo simulations of the previous section.

6. CONCLUDING REMARKS

Many statistical applications require the forecast of a random variable of interest over several periods into the future; that is, one needs to forecast an entire future path. In addition to the resulting path forecast, one often would also like to compute a corresponding JPR. Such a region is supposed to contain the entire future path with a prespecified probability $1-\alpha$.

In this article, we have proposed bootstrap JPRs of three different shapes: one-sided lower, one-sided upper, and two-sided. In this way, the applied researcher can choose the most suitable shape for the task at hand. Furthermore, the JPRs are completely generic in that they allow the applied researcher to select whichever methods are deemed most appropriate to make forecasts, compute prediction standard errors, and generate bootstrap data.

Compared with two previous proposals in the literature, our bootstrap JPRs have two important advantages. First, they are proven to be asymptotically consistent under a realistic, mild high-level assumption. Second, they enjoy superior finite-sample properties, as demonstrated via extensive Monte Carlo simulations.

As an additional bonus, we also offer generalized joint prediction regions obtained by the bootstrap. Such regions are not required to contain the entire future path but only the entire future path up to a small, user-defined number of elements, with prespecified probability $1-\alpha$. If the maximum forecast horizon is large, it may be deemed acceptable by the applied researcher that a small number, such as one or two, of elements of the future path fall outside the JPR. In return, he or she will then obtain a smaller and more informative region.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

REFERENCES

Beran R. 1984. Bootstrap methods in statistics. *Jahresbericht der Deutschen Mathematiker-Vereinigung* **86**(1): 14–30.
Beran R. 1988a. Balanced simultaneous confidence sets. *Journal of the American Statistical Association* **83**: 679–686.

[13] The number of bootstrap samples for NP heuristic and k-FWE JPR is $B = 5000$.

DOI: 10.1111/jtsa.12099

Beran R. 1988b. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* **83**: 687–697.
Bowden DC. 1970. Simultaneous confidence bands for linear regression models. *Journal of the American Statistical Association* **65**(329): 413–421.
Bühlmann P. 2002. Bootstrap for time series. *Statistical Science* **17**: 52–72.
Clements MP, Taylor N. 2001. Bootstrapping prediction intervals for autoregressive models. *International Journal of Forecasting* **17**: 247–267.
De Gooijer JG, Hyndman RJ. 2006. 25 years of time series forecasting. *International Journal of Forecasting* **22**: 443–473.
Efron B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**: 1–26.
Hall P. 1992. *The Bootstrap and Edgeworth Expansion*. New York: Springer.
Jordà Ò, Marcellino MG. 2010. Path-forecast evaluation. *Journal of Applied Econometrics* **25**: 635–662.
Jordà Ò, Knüppel M, Marcellino MG. 2010. Empirical simultaneous confidence regions for path-forecasts. *No. DP7797*, *CEPR*. Available at SSRN: http://ssrn.com/abstract=1611493.
Jordà Ò, Knüppel M, Marcellino MG. 2014. Empirical simultaneous prediction regions for path-forecasts. *International Journal of Forecasting* **29**(3): 456–468.
Kilian L. 1998. Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics* **80**: 218–230.
Lahiri SN. 2003. *Resampling Methods for Dependent Data*. New York: Springer.
Lütkepohl H. 1991. *Introduction to Multiple Time Series Analysis*. Berlin: Springer.
Lütkepohl H. 2005. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.
Montgomery AL, Zarnowitz V, Tsay RS, Tiao GT. 1998. Forecasting the U.S. unemployment rate. *Journal of the American Statistical Association* **93**(442): 478–493.
Pan L, Politis DN. 2014. Bootstrap prediction intervals for linear, nonlinear, and nonparametric autoregressions, *Department of Economics, UCSD*. Available at http://www.escholarship.org/uc/item/67h5s74t.
Pascual L, Romo J, Ruiz E. 2001. Effects of parameter estimation on prediction densities: a bootstrap approach. *International Journal of Forecasting* **17**(1): 83–103.
Politis DN. 2003. The impact of bootstrap methods on time series analysis. *Statistical Science* **18**: 219–230.
Politis DN. 2013. Model-free model-fitting and predictive distributions. *Test* **22**(2): 183–250.
Potter SM. 1995. A nonlinear approach to US GNP. *Journal of Applied Econometrics* **2**: 109–125.
Romano JP, Wolf M. 2005. Stepwise multiple testing as formalized data snooping. *Econometrica* **73**(4): 1237–1282.
Romano JP, Wolf M. 2007. Control of generalized error rates in multiple testing. *Annals of Statistics* **35**(4): 1378–1408.
Romano JP, Wolf M. 2010. Balanced control of generalized error rates. *Annals of Statistics* **38**(1): 598–633.
Romano JP, Shaikh AM, Wolf M. 2008. Formalized data snooping based on generalized error rates. *Econometric Theory* **24**(2): 404–447.
Roy A, Fuller WA. 2001. Estimation for autoregressive time series with a root near 1. *Journal of Business and Economics Statistics* **19**(4): 482–493.
Scheffé H. 1953. A method for judging all contrasts in the analysis of variance. *Biometrika* **40**: 87–104.
Scheffé H. 1959. *The Analysis of Variance*. New York: John Wiley & Sons.
Staszewska-Bystrova A. 2011. Bootstrap prediction bands for forecast paths from vector autoregressive models. *Journal of Forecasting* **30**(8): 721–735.
Staszewska-Bystrova A. 2013. Modified Scheffé's prediction bands. *Jahrbücher für Nationalökonomie und Statistik* **233**(5+6): 680–690.
Stock JH, Watson MW. 2001. Vector autoregressions. *Journal of Economic Perspectives* **15**(4): 101–115.
Thombs L, Schucany WR. 1990. Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association* **85**(410): 486–492.
White J. 1961. Asymptotic expansions for the mean and variance of the serial correlation coefficient. *Biometrika* **48**: 85–95.
White HL. 2001. *Asymptotic Theory for Econometricians*, Revised Edition. New York: Academic Press.

DOI: 10.1111/jtsa.12099