University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 64

Bootstrap Joint Prediction Regions

Michael Wolf and Dan Wunderli

February 2012

Bootstrap Joint Prediction Regions*

Michael Wolf†
Department Economics
University of Zurich
CH-8032 Zurich, Switzerland
michael.wolf@econ.uzh.ch

Dan Wunderli
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
dan.wunderli@econ.uzh.ch

February, 2012

Abstract

Many economic and financial 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 to some previous proposals via Monte Carlo simulations. An empirical application to a real data set is also provided.

KEY WORDS: Generalized error rates; path-forecast; simultaneous prediction intervals.

JEL CLASSIFICATION NOS: C14, C32, C53.

*We thank Christian Kascha and participants of the Zurich Workshop on Economics 2011 for helpful comments.

†Research has been supported by the NCCR Finrisk project "New Methods in Theoretical and Empirical Asset Pricing"

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.

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.[1] Therefore, standard normal-theory prediction intervals are only valid, even asymptotically, 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. But 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 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. But, 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

[1] For example, such an assumption is made by Jordà and Marcellino (2010).

one would like to know how high inflation might rise over the next $H = 12$ months, with probability $1-\alpha$, one needs to construct a joint prediction region 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 joint prediction region 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 h periods 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 joint prediction region that will contain the entire future path with probability $(1-\alpha)^H$ only.[2]

Unfortunately, the method of stringing together marginal prediction intervals for horizons $h = 1$ up to $h = H$ is still widely in use, such as in the fan charts for GDP growth and CPI inflation published by the Bank of England and the Central Bank of Norway.[3]

The construction of joint prediction regions for future paths of a random variable of interest has been rather neglected in the forecasting literature so far. Two notable exceptions are Jordà and Marcellino (2010) and Staszewska-Bystrova (2010). 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. Therefore, neither of the proposed methods appears entirely safe to use in practice.

In this paper, 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 joint prediction regions that will only contain all elements of future paths up to a small number $k-1$ of them with probability $1-\alpha$. If the maximum forecast horizon H is large, the applied researcher may deem the criterion 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 joint prediction region; 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 joint prediction region with probability $1-\alpha$. The choice of k must be made by the applied researcher, not by the statistician. But 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' joint prediction region that must

[2] 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 joint prediction region that will contain the entire future path with probability somewhere between $(1-\alpha)^H$ and $1-\alpha$. The exact probability is a function of the dependence structure of the events $\{E_h\}_{h=1}^{H}$.

[3] Several examples can be found at http://www.bankofengland.co.uk/publications/inflationreport/ and http://www.norges-bank.no/english/inflationreport/.

contain all elements of a future path with probability $1-\alpha$.

The remainder of the paper 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 to 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 concludes. All mathematical proofs and some further background results are collected in an appendix.

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 joint prediction region 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 $\mu \equiv \mathbb{E}(y)$. 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 times, the two are actually the same, that is $\widehat{y} = \widehat{\mu}$; for example, in the context of linear regression models. Therefore, in terms of a (point) forecast of y compared a (point) estimate of μ, there often is no difference at all.

But what if one desires an 'uncertainty interval' in addition? Such an interval should contain the random variable y or its mean μ, respectively, with a prespecified probability $1-\alpha$. (To be careful, this probability only exists before computing the interval from a frequentist view point, at least for the mean μ.) Now the two solutions are fundamentally different and the former interval will have to be wider due to the additional randomness contained in the random variable y compared to 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.

It is also useful to point out that there is a duality between a confidence interval for μ and a hypothesis test for μ. For concreteness, consider the two-sided hypothesis testing problem

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0 \ . \tag{1}$$

If a level α test is available for this problem, for every value of μ_0, then a two-sided confidence interval CI for μ at level $1-\alpha$ can be constructed by *inverting* the hypothesis test as follows:

$$\text{CI} \equiv \{\mu_0 : \mu_0 \text{ is not rejected by the hypothesis test}\} \ . \tag{2}$$

That is, the collection of values μ_0 not rejected by the test at significance level α constitutes a confidence interval with confidence level $1-\alpha$. Conversely, a hypothesis test for problem (1) can be carried by *inverting* a two-sided confidence interval for μ: one simply rejects H_0 at significance level α if and only if (iff) the value μ_0 is not contained in the confidence interval with confidence level $1-\alpha$.

Analogously, there is a duality between a one-sided confidence interval for μ and a one-sided hypothesis test for μ; the details are straightforward.

On the other hand, there is no duality between a prediction interval for y and a hypothesis test for y. This is because y is a random variable and not a (non-random) parameter and hypothesis tests on such random quantities do not exist. In particular, the testing problem

$$H_0 : \widehat{y} - y = 0 \quad \text{versus} \quad H_1 : \widehat{y} - y \neq 0 \tag{3}$$

is nonsensical. The quantity $\widehat{y} - y$ is a random variable. If its distribution is continuous, then $\widehat{y} - y$ will be different from zero with probability one, irrespective of the 'quality' of the forecast $\widehat{y}$.[4]

2.2 Path-Forecast

More generally, consider a random vector $Y \equiv (y_1, \ldots, y_H)'$ of interest with mean $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_H)' = \mathbb{E}(Y)$. For the purposes of this paper, Y will typically correspond to the values of a random variable one to H periods into the future; that is, to a future *path* of a random variable. But the discussion below 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 $\widehat{Y}$ and the estimator of $\boldsymbol{\mu}$ by $\widehat{\boldsymbol{\mu}}$. (When Y corresponds to a future path of a random variable, $\widehat{Y}$ is also called a *path-forecast*.) Again, often times, the two are actually the same, that is, $\widehat{Y} = \widehat{\boldsymbol{\mu}}$; for example, in the context of linear regression models. Therefore, again, in terms of a (point) forecast of Y compared to 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 $\boldsymbol{\mu}$, respectively, with a prespecified

[4] Nevertheless, a testing problem of this sort is considered by Jordà et al. (2010, Subsection 2.1).

probability $1-\alpha$. Again, the two solutions are fundamentally different and the former region will have to be larger (in volume) due to the additional randomness contained in Y compared to its mean $\boldsymbol{\mu}$.

Again, there is a duality between a joint confidence region for the parameter $\boldsymbol{\mu}$ and a hypothesis test for $\boldsymbol{\mu}$. In the multivariate setting, the testing problem is inherently of a two-sided nature:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 \ . \tag{4}$$

If a level α test is available for this problem, for every value of $\boldsymbol{\mu}_0$, then a joint confidence region JCR for $\boldsymbol{\mu}$ at level $1-\alpha$ can be constructed by *inverting* the hypothesis test as follows:

$$\text{JCR} = \{\boldsymbol{\mu}_0 : \boldsymbol{\mu}_0 \text{ is not rejected by the hypothesis test}\} \ . \tag{5}$$

That is, the collection of values not rejected by the test at significance level α constitutes a joint confidence region with confidence level $1-\alpha$. Conversely, a hypothesis test for problem (4) can be carried out by *inverting* a joint confidence region for $\boldsymbol{\mu}$: one simply rejects H_0 at significance level α iff the value $\boldsymbol{\mu}_0$ is not contained in the joint confidence region with confidence level $1-\alpha$.

Again, on the other hand, there is no duality between a joint prediction region for Y and a hypothesis test for Y. In particular, the testing problem

$$H_0 : \widehat{Y} - Y = \mathbf{0} \quad \text{versus} \quad H_1 : \widehat{Y} - Y \neq \mathbf{0} \ , \tag{6}$$

where $\mathbf{0} \equiv (0, 0, \ldots, 0)'$, is nonsensical. The quantity $\widehat{Y} - Y$ is a random vector. If its distribution is continuous, then $\widehat{Y} - Y$ will be different from the vector zero with probability one, irrespective of the 'quality' of the forecast $\widehat{Y}$.[5]

A potential complication with joint regions arises when uncertainty statements concerning the individual components y_h or μ_h, respectively, are desired. For example, this is typically the case when a joint prediction region for Y is to be constructed in addition to a path-forecast $\widehat{Y}$. One desires lower and upper bounds for each component y_h in such a manner that the entire vector Y be contained in the implied rectangle with probability $1-\alpha$. This is a trivial task if the underlying joint prediction region is already of rectangular form. But this is not true for all methods to compute joint regions; many 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 joint confidence region for $\boldsymbol{\mu}$ is obtained by inverting the classical F-test. Let $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\mu}})$ denote an estimated covariance matrix of $\widehat{\boldsymbol{\mu}}$. Then the joint confidence region is given by

$$\text{JCR} \equiv \left\{\boldsymbol{\mu}_0 : (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)' \left[\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\mu}})\right]^{-1} (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \leq \chi^2_{H,1-\alpha}\right\} \ , \tag{7}$$

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 joint confidence region is usually justified by a central limit theorem

[5] Nevertheless, a testing problem of this sort is considered by Jordà et al. (2010, Subsection 2.2).

implying an approximate multivariate normal distribution of $\widehat{\boldsymbol{\mu}}$ with mean $\boldsymbol{\mu}$. Such a central limit theorem will hold under mild regularity conditions; for example, see White (2001).

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

$$\text{JPR} \equiv \left\{X : (\widehat{Y} - X)' \left[\widehat{\boldsymbol{\Sigma}}(\widehat{Y})\right]^{-1} (\widehat{Y} - X) \leq \chi^2_{H,1-\alpha}\right\} . \tag{8}$$

The use of this joint prediction region is only justified if $\widehat{U}$ 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. But 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 3.2 below.

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, with sides parallel to the axes of $\mathbb{R}^H$, 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 generally overly conservative. If statements concerning the individual components are desired, it is better to construct 'direct' rectangular joint regions instead; that is, joint regions that are designed to be of rectangular form to begin with.

Remark 2.1. It will be useful to illustrate these concepts in simple, parametric setup. Assume $Y \equiv (Y_1, Y_2)' \sim N(\boldsymbol{\mu}, \boldsymbol{I}_2)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and $\boldsymbol{I}_2$ is the identity matrix of dimension two. Therefore, Y_1 and Y_2 are independent with $Y_h \sim N(\mu_h, 1)$. 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, $\widehat{\boldsymbol{\mu}} \equiv Y$.

The Scheffé joint confidence region is obtained by inverting the classical F-test. It is a circle centered 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 $\sqrt{5.99} = 2.45$. The implied rectangular joint confidence region, obtained by projecting the circle on the two axes, is a square with center Y and half length 2.45.

On the other hand, a 'direct' rectangular joint confidence region is given by

$$\left[Y_1 \pm d_{2,1-\alpha}\right] \times \left[Y_1 \pm d_{2,1-\alpha}\right] ,$$

where $d_{2,1-\alpha}$ 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 $d_{2,0.95} = 2.24$.

The 'direct' rectangular joint confidence region is thus a square with center Y and half length 2.24. Therefore, it is smaller than the implied rectangular joint confidence region by the Scheffé method.

The Scheffé region itself has a smaller volume than the 'direct' rectangular region when $\alpha = 0.05$, namely

$$2.45^2 \cdot \pi = 18.86 < 20.07 = (2 \cdot 2.24)^2 \ .$$

But when a rectangular region is needed in the end, projecting the Scheffé region on the axes results in a larger region compared to the 'direct' rectangular region. An illustration is provided in Figure 1. ■

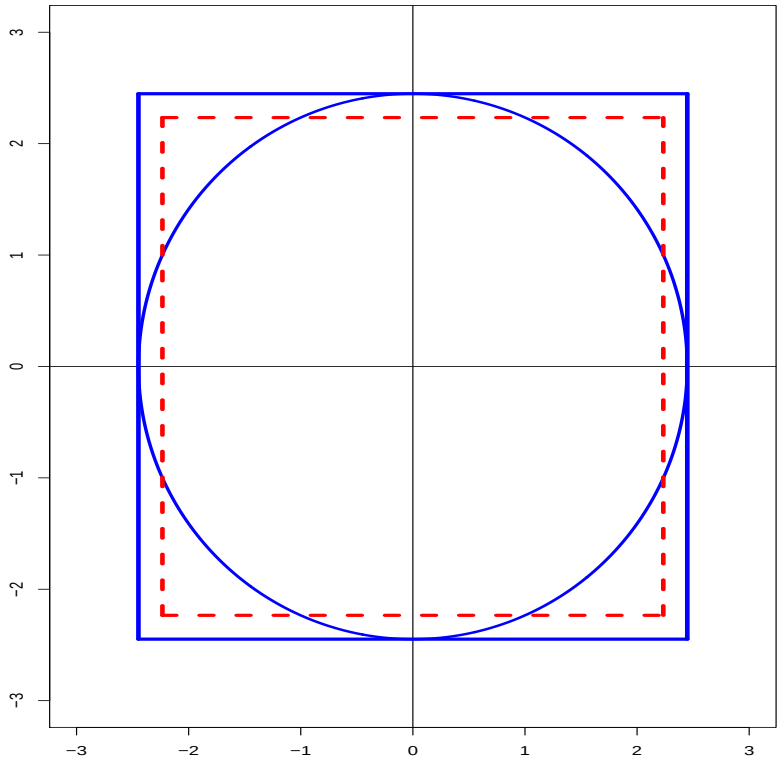

Figure 1: An illustration of Remark 2.1. One observes $\widehat{\boldsymbol{\mu}} = Y = (0.0, 0.0)$ and wishes to construct a joint confidence region for $\boldsymbol{\mu}$ with confidence level $1 - \alpha = 0.95$. The solid ellipse is the Scheffé joint confidence region: a circle with radius 2.45. The solid rectangle is the implied (that is, projected on the axes) rectangular joint confidence region: a square with half length 2.45. The dashed rectangle is the 'direct' rectangular joint confidence region: a square with half length 2.24.

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 joint confidence region (JCR) for $\boldsymbol{\mu}$,

$$\text{FWE} \equiv \mathbb{P}\{\text{At least one of the } \mu_h \text{ not contained in the JCR}\} \ , \tag{9}$$

whereas for a joint prediction region (JPR) for Y,

$$\text{FWE} \equiv \mathbb{P}\{\text{At least one of the } y_h \text{ not contained in the JPR}\} \ . \tag{10}$$

Jordà et al. (2010, Section 2.2) argue that controlling the FWE can be too strict:

> 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.

The decision whether the FWE is too strict or not in a given application has to be made by the applied researcher, not by the econometrician. It is the job of the econometrician to provide the applied researcher with an alternative tool in case his decision is against control of the FWE. Jordà et al. (2010) propose as an alternative to control the *false discovery rate* (FDR). Unfortunately, this proposal is actually equivalent to control of the familywise error rate in the context of joint confidence regions and joint prediction regions. The explanation of this fact is a bit lengthy and can be found in Appendix B.

Although control of the FDR is not a meaningful alternative, it is possible to construct joint confidence regions as well as joint prediction regions based on a generalized error rate that is meaningful in the context of joint regions. The solution is to use the *generalized familywise error rate* (k-FWE).

For a joint confidence region (JCR) for $\boldsymbol{\mu}$,

$$k\text{-FWE} \equiv \mathbb{P}\{\text{At least } k \text{ of the } \mu_h \text{ not contained in the JCR}\} \ , \tag{11}$$

whereas for a joint prediction region (JPR) for Y,

$$k\text{-FWE} \equiv \mathbb{P}\{\text{At least } k \text{ of the } y_h \text{ not contained in the JPR}\} \ . \tag{12}$$

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 two 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 analogous in the context of a joint confidence region for $\boldsymbol{\mu}$ and is detailed in Romano and Wolf (2005, 2007) already.

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

Remark 2.2. If the cumulative distribution function of a random variable is not invertible, then its quantiles are not necessarily uniquely defined. For concreteness, we adopt the following definition for quantiles in this paper.

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 $\inf\{x : F(x) \geq \lambda\}$. ■

3 Joint Prediction Regions Based on k-FWE Control

The goal is to construct a joint prediction region for a future path 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 $Y_{T,H} \equiv (y_{T+1}, \ldots, y_{T+H})'$. At time t, denote a forecast h periods ahead by $\widehat{y}_t(h)$. Then a path-forecast for $Y_{T,H}$ is given by $\widehat{Y}_T(H) \equiv (\widehat{y}_T(1), \ldots, \widehat{y}_T(H))'$. Denote the vector of prediction errors by $\widehat{U}_T(H) \equiv (\widehat{u}_T(1), \ldots, \widehat{u}_T(H))' \equiv \widehat{Y}_T(H) - Y_{T,H}$. Finally, $\widehat{\sigma}_T(h)$ denotes a prediction standard error, that is, a standard error for $\widehat{u}_T(h)$: it is an estimator of the unknown standard deviation of the random variable $\widehat{u}_T(h)$.

We further assume a generic method to compute a vector of bootstrap prediction errors $\widehat{U}^*_T(H) \equiv (\widehat{u}^*_T(1), \ldots, \widehat{u}^*_T(H))'$, based on artificial bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$ generated from an estimated probability mechanism $\widehat{\mathbb{P}}_T$.[6] Such bootstrap forecast errors can

[6]The estimated probability mechanism has subscript T because it is a function of the observed data $\{y_1, \ldots, y_T\}$.

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, $\widehat{\sigma}^*_T(h)$ denotes a bootstrap prediction standard error, that is, a standard error for $\widehat{u}^*_T(h)$.

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 $\widehat{y}_T(h)$, for $h = 1, \ldots, H$. To be concrete, assume he uses an ARIMA model. The fitted model also provides prediction standard errors $\widehat{\sigma}_T(h)$. Next, the applied researchers generates bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$. To this end, he can use a parametric bootstrap, based on the ARIMA model fitted from the original data; this would be a suitable approach if he believes that his null model is correctly specified. Alternatively, he can use a nonparametric time series bootstrap (say a blocks bootstrap or a sieve bootstrap); this would be a suitable approach if he believes that his null model might be misspecified.[7] Not making use of the stretch $\{y^*_{T+1}, \ldots, y^*_{T+H}\}$, he computes forecasts $\widehat{y}^*_T(h)$ and prediction standard errors $\widehat{\sigma}^*_T(h)$. Finally, he computes $\widehat{u}^*_T(h) \equiv \widehat{y}^*_T(h) - y^*_{T+h}$.

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

Assumption 3.1. *$\widehat{J}_T$ converges in distribution to a non-random continuous limit law $\widehat{J}$. Furthermore, $\widehat{J}^*_T$ consistently estimates this limit law: $\rho(\widehat{J}_T, \widehat{J}^*_T) \to 0$ in probability, for any metric ρ metrizing weak convergence.*

Expressed in words, Assumption 3.1 states that, as the sample size T increases, the conditional distribution of the vector of standardized bootstrap prediction errors $\widehat{S}^*_T(H)$ becomes a more and more reliable approximation to the (unknown) conditional distribution of the vector of true standardized prediction errors $\widehat{S}_T(H)$.

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

Some further notation is required. Suppose $X \equiv (x_1, \ldots, x_H)'$ is a vector with H components. First, for $k \in \{1, \ldots, H\}$, k-max(X) returns the k^{th}-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$(X) \equiv x_{(H-k+1)}$. Second, for $k \in \{1, \ldots, H\}$, k-min(X) returns the k^{th}-smallest value of the x_h; that is, k-min$(X) \equiv x_{(k)}$. Third, $|X|$ denotes the vector $\big(|x_1|, \ldots, |x_H|\big)'$.

Let $d^{max}_{|\cdot|,1-\alpha}(k)$ denote the $1-\alpha$ quantile of the random variable k-max$\big(|\widehat{S}_T(H)|\big)$. Then a two-sided joint prediction region for $Y_{T,H}$ that exactly controls the k-FWE is given by

$$\big[\widehat{y}_T(1) \pm d^{max}_{|\cdot|,1-\alpha}(k) \cdot \widehat{\sigma}_T(1)\big] \times \cdots \times \big[\widehat{y}_T(H) \pm d^{max}_{|\cdot|,1-\alpha}(k) \cdot \widehat{\sigma}_T(H)\big] \ . \tag{13}$$

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

The implication is that the probability that the region (13) will contain at least $H-k+1$ elements of $Y_{T,H}$ is (at least) equal to $1-\alpha$ in finite samples. This property follows immediately from the definition of $d^{max}_{|\cdot|,1-\alpha}(k)$.

The problem is that this ideal region is not feasible, since the constant $d^{max}_{|\cdot|,1-\alpha}(k)$ is unknown. It has to be estimated in practice by $d^{max,*}_{|\cdot|,1-\alpha}(k)$, which is defined as the $1-\alpha$ quantile of the random variable $k\text{-max}\big(|\widehat{S}^*_T(H)|\big)$. This quantile can typically not be derived analytically, but it can be simulated to arbitrary precision from a sufficiently large number of bootstrap samples; see Algorithm 3.1 below.

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

$$\big[\widehat{y}_T(1) \pm d^{max,*}_{|\cdot|,1-\alpha}(k)\cdot\widehat{\sigma}_T(1)\big] \times \cdots \times \big[\widehat{y}_T(H) \pm d^{max,*}_{|\cdot|,1-\alpha}(k)\cdot\widehat{\sigma}_T(H)\big] \ . \tag{14}$$

The implication is that the probability that the region (14) will contain at least $H-k+1$ elements of $Y_{T,H}$ is (at least) equal to $1-\alpha$ asymptotically.

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

Let $d^{max,*}_{1-\alpha}(k)$ denote the $1-\alpha$ quantile of the random variable $k\text{-max}\big(\widehat{S}^*_T(H)\big)$. Then a one-sided lower joint prediction region for $Y_{T,H}$ that asymptotically controls the k-FWE is given by

$$\big[\widehat{y}_T(1) - d^{max,*}_{1-\alpha}(k)\cdot\widehat{\sigma}_T(1), \infty\big) \times \cdots \times \big[\widehat{y}_T(H) - d^{max,*}_{1-\alpha}(k)\cdot\widehat{\sigma}_T(H), \infty\big) \ . \tag{15}$$

Let $d^{min,*}_{\alpha}(k)$ denote the α quantile of the random variable $k\text{-min}\big(\widehat{S}^*_T(H)\big)$. Then a one-sided upper joint prediction region for $Y_{T,H}$ that asymptotically controls the k-FWE is given by

$$\big(-\infty, \widehat{y}_T(1) - d^{min,*}_{\alpha}(k)\cdot\widehat{\sigma}_T(1)\big] \times \cdots \times \big(-\infty, \widehat{y}_T(H) - d^{min,*}_{\alpha}(k)\cdot\widehat{\sigma}_T(H)\big] \ . \tag{16}$$

Note here that $d^{min,*}_{\alpha}(k)$ is generally a negative number so that, for each component h, the upper end of the corresponding interval is indeed larger than the prediction $\widehat{y}_T(h)$.

As is clear from the definitions, the multipliers $d^{max,*}_{|\cdot|,1-\alpha}(k)$, $d^{max,*}_{1-\alpha}(k)$, are both monotonically decreasing in k, while the multiplier $d^{min,*}_{\alpha}(k)$ is monotonically increasing in k. Consequently, the larger the value of k, the smaller in volume are the regions (14)–(16); for an illustration, see Subsection 5.1. (When we speak of 'volume' for the one-sided regions (15)–(16), 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 joint prediction regions.

Proposition 3.1. *Under Assumption 3.1, each of the joint prediction regions (JPRs) (14)–(16) for $Y_{T,H}$ satisfies*

$$\limsup_{T\to\infty} \ k\text{-}FWE \leq \alpha \ , \tag{17}$$

where

$$k\text{-}FWE \equiv \mathbb{P}\{\textit{At least } k \textit{ of the } y_{T+h} \textit{ not contained in the JPR}\} \ . \tag{18}$$

The following algorithm details how to compute the three multipliers $d^{max,*}_{|\cdot|,1-\alpha}(k)$, $d^{max,*}_{1-\alpha}(k)$, and $d^{min,*}_{\alpha}(k)$ 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 $\widehat{\mathbb{P}}_T$.

Algorithm 3.1 (Computation of the JPR Multipliers; Univariate Case)**.**

1. Generate bootstrap data $\{y^*_1, \ldots, y^*_T, y^*_{T+1}, \ldots, y^*_{T+H}\}$ from $\widehat{\mathbb{P}}_T$.
2. Not making use of the stretch $\{y^*_{T+1}, \ldots, y^*_{T+H}\}$, compute forecasts $\widehat{y}^*_T(h)$ and prediction standard errors $\widehat{\sigma}^*_T(h)$.
3. Compute bootstrap prediction errors $\widehat{u}^*_T(h) \equiv \widehat{y}^*_T(h) - y^*_{T+h}$.
4. Compute standardized bootstrap prediction errors $\widehat{s}^*_T(h) \equiv \widehat{u}^*_T(h)/\widehat{\sigma}^*_T(h)$ and let $\widehat{S}^*_T(H) \equiv \big(\widehat{s}^*_T(1), \ldots, \widehat{s}^*_T(H)\big)'$.
5. Compute $k\text{-max}^*_{|\cdot|} \equiv k\text{-max}\big(\big|\widehat{S}^*_T(H)\big|\big)$, $k\text{-max}^* \equiv k\text{-max}\big(\widehat{S}^*_T(H)\big)$, and $k\text{-min}^* \equiv k\text{-min}\big(\widehat{S}^*_T(H)\big)$.
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 $d^{max,*}_{|\cdot|,1-\alpha}(k)$ is the empirical $1-\alpha$ quantile of the statistics $\{k\text{-max}^*_{|\cdot|,1}, \ldots, k\text{-max}^*_{|\cdot|,B}\}$.
   - 7.2 $d^{max,*}_{1-\alpha}(k)$ is the empirical $1-\alpha$ quantile of the statistics $\{k\text{-max}^*_1, \ldots, k\text{-max}^*_B\}$.
   - 7.3 $d^{min,*}_{\alpha}(k)$ is the empirical α quantile of the statistics $\{k\text{-min}^*_1, \ldots, k\text{-min}^*_B\}$.

In an application, the number of bootstrap samples, B, should be chosen as large as possible; at the very least $B \geq 1,000$.

Remark 3.1. Proposition 3.1 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 $\widehat{J}^*_T$ as close as possible to the true distribution $\widehat{J}_T$. How this is to be done in detail depends on the particular bootstrap method chosen by the applied researcher. Many papers 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 which have to be estimated from the original data $\{y_1, \ldots, y_T\}$ to compute the forecasts $\widehat{y}_T(h)$ and the prediction standard errors $\widehat{\sigma}_T(h)$ should be re-estimated from the bootstrap data $\{y^*_1, \ldots, y^*_T\}$ to compute the forecasts $\widehat{y}^*_T(h)$ and the prediction standard errors $\widehat{\sigma}^*_T(h)$. 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 $\widehat{y}^*_T(h)$ and the prediction standard errors $\widehat{\sigma}^*_T(h)$. But 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 to 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 K-variate time series $\{Z_1, \ldots, Z_T\}$ 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 $Z_t \equiv (y_t, z_{2,t}, \ldots, z_{K,t})'$.

In this more general case, the forecast of y_{T+h}, denoted by $\widehat{y}_T(h)$ 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 $\widehat{\sigma}_T(h)$.

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

Denote $Z_t^* \equiv (y_t^*, z_{2,t}^*, \ldots, z_{K,t}^*)'$. The forecast of y_{T+h}^*, denoted by $\widehat{y}_T^*(h)$ 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 $\widehat{\sigma}_T^*(h)$.

Assumption 3.1 continues to be based on the two vectors of standardized prediction errors $\widehat{S}_T(H) \equiv (\widehat{u}_T(1)/\widehat{\sigma}_T(1), \ldots, \widehat{u}_T(H)/\widehat{\sigma}_T(H))'$ and $\widehat{S}_T^*(H) \equiv (\widehat{u}_T^*(1)/\widehat{\sigma}_T^*(1), \ldots, \widehat{u}_T^*(H)/\widehat{\sigma}_T^*(H))'$, respectively. Only that now, more generally, $\widehat{J}_T$ denotes the probability law under $\mathbb{P}$ of $\widehat{S}_T(H)|Z_T, Z_{T-1}, \ldots$; and $\widehat{J}_T^*$ denotes the probability law under $\widehat{\mathbb{P}}_T$ of $\widehat{S}_T^*(H)|Z_T^*, Z_{T-1}^*, \ldots$

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 joint prediction regions are still given by (14)–(16) and Proposition 3.1 continues to hold.

The following algorithm details how to compute the three multipliers $d_{|\cdot|,1-\alpha}^{max,*}(k)$, $d_{1-\alpha}^{max,*}(k)$, and $d_{\alpha}^{min,*}(k)$ 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 $\widehat{\mathbb{P}}_T$.

Algorithm 3.2 (Computation of the JPR Multipliers; Multivariate Case)**.**

1. Generate bootstrap data $\{Z_1^*, \ldots, Z_T^*, Z_{T+1}^*, \ldots, Z_{T+H}^*\}$ from $\widehat{\mathbb{P}}_T$.
2. Not making use of the stretch $\{Z_{H+1}^*, \ldots, Z_{T+H}^*\}$, compute forecasts $\widehat{y}_T^*(h)$ and prediction standard errors $\widehat{\sigma}_T^*(h)$.
3. Identical to Algorithm 3.1.

⋮

7. Identical to Algorithm 3.1.

3.3 Comparison with Previous Methods

Jordà and Marcellino (2010) propose an alternative 'asymptotic' method to construct a joint prediction region for $Y_{T,H}$ that controls the FWE.[8] It is based on the assumption that

$$\sqrt{T}\big(\widehat{Y}_T(H) - Y_{T,H}|Z_T, Z_{T-1}, \ldots\big) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Xi}_H) \ , \tag{19}$$

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

The proposed joint prediction region is given by

$$\widehat{Y}_T(H) \pm P \left[\sqrt{\frac{\chi^2_{h,1-\alpha}}{h}} \right]_{h=1}^{H} \ , \tag{20}$$

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

First, assumption (19) implies that the conditional distribution of the vector of prediction errors $\widehat{U}_T(H) \equiv \widehat{Y}_T(H) - Y_{T,H}$ is approximately multivariate normal with mean zero, at least for large T. Such a result appears overly strict. 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 (19) implies in addition that the conditional covariance matrix of the vector of prediction errors $\widehat{U}_T(H) \equiv \widehat{Y}_T(H) - Y_{T,H}$ vanishes asymptotically. This appears unrealistic. While, under mild regularity conditions, the variance of an estimator of a population parameter vanishes asymptotically, the same is not true for the variance of a prediction error. Even if all model parameters are known, a future observation cannot be predicted perfectly because of its random nature.

Remark 3.2. To illustrate the first two points, consider the simple AR(1) model

$$y_t = \nu + \rho y_{t-1} + \epsilon_t \ , \tag{21}$$

where $|\rho| < 1$ and the errors $\{\epsilon_t\}$ are independent and identically distributed (i.i.d.) with mean zero and finite variance σ^2_ϵ. At time T, the forecast of y_{T+1} is given by

$$\widehat{y}_T(1) \equiv \widehat{\nu} + \widehat{\rho} y_T \ , \tag{22}$$

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

$$\widehat{u}_T(1) = \widehat{\nu} + \widehat{\rho} y_T - y_{T+1} \ . \tag{23}$$

[8] They use the term *joint confidence region* instead of *joint prediction region*.

As T tends to infinity, the conditional distribution of $\widehat{u}_T(1)$ converges weakly to the unconditional distribution of $-\epsilon_{T+1}$ (which does not depend on T). This distribution is neither necessarily normal nor does its variance vanish. As a result, assumption (19) does not hold in this simple example. ■

Third, Jordà and Marcellino (2010) initially consider the following rectangular joint prediction region:

$$\widehat{Y}_T(H) \pm P\left[\sqrt{\frac{\chi^2_{H,1-\alpha}}{H}}\,\mathbf{1}_H\right] , \tag{24}$$

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

$$\left\{\widetilde{Y} : T(\widehat{Y}_T(H) - \widetilde{Y})'\widehat{\mathbf{\Xi}}_H^{-1}(\widehat{Y}_T(H) - \widetilde{Y}) \leq \chi^2_{H,1-\alpha}\right\} . \tag{25}$$

As we have explained above, deriving a rectangular joint confidence region from an initial joint confidence region of elliptical form is suboptimal in terms of the volume of the rectangular joint confidence region.

Fourth, Jordà and Marcellino (2010) arrive at their final joint prediction region (20) by 'refining' the initial joint prediction region (24) by a step-down recursive procedure that is entirely ad-hoc and lacks a theoretical justification.

Fifth, a counter-intuitive feature of the joint prediction region (20) is that its width is not necessarily (weakly) monotonically increasing in the forecast horizon h; for an example, see Subsection 5.1. The reason is that the multipliers $\sqrt{\chi^2_{h,1-\alpha}/h}$ can be strictly monotonically decreasing in h, at least for commonly used values of α, as illustrated in Figure 2.

Since there is no proof of asymptotic validity, under realistic conditions, of the method proposed by Jordà and Marcellino (2010), the method is not trustworthy to use in practice.

Staszewska-Bystrova (2010) proposes an alternative bootstrap method to construct a joint prediction region 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 $\widehat{Y}_T^{*,b}(H)$, for $b = 1, \ldots, B$. One then discards αB of these bootstrap path-forecasts: namely those $\widehat{Y}_T^{*,b}(H)$ that are 'furthest' away from the original path-forecast $\widehat{Y}_T(H)$, where the distance between two $H \times 1$ vectors is measured by the Euclidian distance.[9] Finally, the joint prediction region is defined as the envelope of the remaining $(1-\alpha)B$ bootstrap path-forecasts; as a result. Although this *neighboring paths (NP)* method seems to perform well in some simulation studies, there are several concerns.

First, the method is purely 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

[9] Staszewska-Bystrova (2010) also considers other distance measures, but concludes that the Euclidean distance seems to work best.

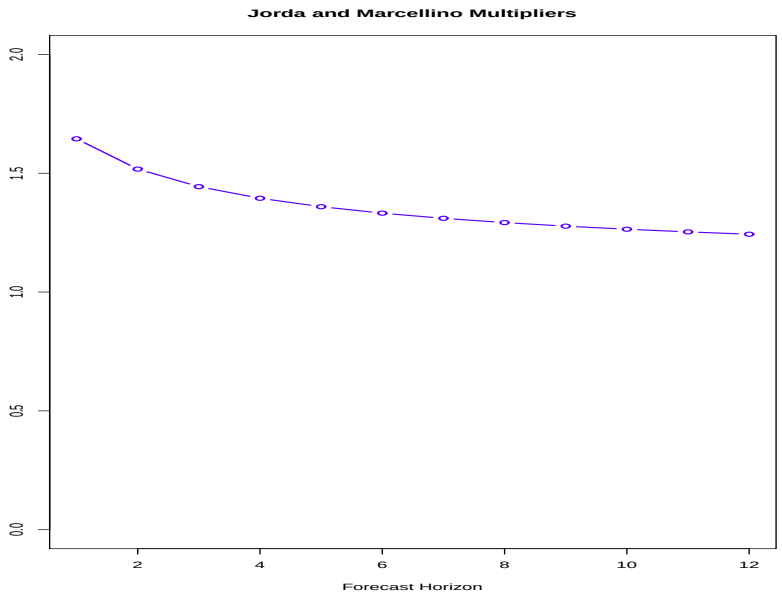

Figure 2: Plot of the Jordà and Marcellino (2010) multipliers $\sqrt{\chi^2_{h,1-\alpha}/h}$ used in their Scheffé joint prediction region (20), for $H = 12$ and $\alpha = 0.1$.

(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 uncorrelataed; 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 (2010), may not be generally valid.

Third, the method is in the spirit of Efron's (1979) percentile method that 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 can 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 (2010) only considers normal errors with mean zero in the data generating processes (DGPs) of her simulation study. On the other hand, the joint prediction regions we propose in Subsections 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 joint prediction region is given by the envelope of the $(1-\alpha)B$ not-discarded bootstrap path-forecasts $\widehat{Y}_T^{*,b}(H)$, the region typically has a jagged shape, which is unattractive; for an example, see Subsection 5.1.

Last but not least, it is not clear whether the methods of Jordà and Marcellino (2010) and

Staszewska-Bystrova (2010) can be generalized to construct a joint prediction region for $Y_{T,H}$ that controls the k-FWE for $k \geq 2$; see (12). By offering a method to construct rectangular joint prediction regions 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 3.3 (Property of Balance)**.** Under a mild additional assumption not covered by Assumption 3.1, our bootstrap joint prediction regions (14)–(16) can be easily seen to have the desirable property of being *balanced*, at least asymptotically.

A rectangular joint prediction region 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 $h = 1, \ldots, H$.[10]

For concreteness, focus on the joint prediction region (14) whose implied prediction interval for y_{T+h} is given by $[\widehat{y}_T(h) \pm d^{max,*}_{|\cdot|,1-\alpha}(k) \cdot \widehat{\sigma}_T(h)]$. Then the probability

$$\mathbb{P}\Big\{ y_{T+h} \in \big[\widehat{y}_T(h) \pm d^{max,*}_{|\cdot|,1-\alpha}(k) \cdot \widehat{\sigma}_T(h)\big] \Big\} \tag{26}$$

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

$$\frac{\widehat{y}_T(h) - y_{T+h}}{\widehat{\sigma}_T(h)} \tag{27}$$

is the same for all $h = 1, \ldots, H$, asymptotically.[11]

A joint prediction region that is balanced implicitly treats all forecasts $\widehat{y}_T(h)$ 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 joint prediction region for $Y_{T,H}$:

$$\mathrm{PI}_T(1) \times (-\infty, \infty) \times \ldots \times (-\infty, \infty) \ , \tag{28}$$

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

It appears at least doubtful whether the property of balance could be established for the joint prediction regions proposed by Jordà and Marcellino (2010) and Staszewska-Bystrova (2010). ■

[10] For a discussion of the concept of balance in the alternative contexts of joint confidence regions and multiple testing, the reader is referred to Beran (1988a,b) and Romano and Wolf (2010).

[11] 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 way to compute the forecasts $\widehat{y}_T(h)$ and the prediction standard errors $\widehat{\sigma}_T(h)$.

4 Monte Carlo Simulations

This section compares the finite-sample performance of various methods to construct joint prediction regions. 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.

Before we present the details of the Monte Carlo setup, we need to be specific about how we estimate the model, compute the prediction standard errors, and generate the bootstrap data.

4.1 Preliminaries

The general AR(p) model is given by

$$y_t = \nu + \rho_1 y_{t-1} + \ldots + \rho_p y_{t-p} + \epsilon_t \ , \tag{29}$$

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

$$y_t = \nu + \rho y_{t-1} + \psi_1 \Delta y_{t-1} + \ldots + \psi_{p-1} \Delta y_{t-p+1} + \epsilon_t \ , \tag{30}$$

to bring out the role of the largest autoregressive root $\rho \equiv \rho_1 + \ldots + \rho_p$. Here, Δ is the first-difference operator. The parameters of formulations (29) and (30) are related by

$$\rho_1 = \rho + \psi_1 \ , \quad \rho_j = -\psi_{j-1} + \psi_j \quad \text{for } 2 \le j \le p-1 \ , \quad \rho_p = -\psi_{p-1} \ . \tag{31}$$

The usefulness of by bias-corrected estimators when making forecasts based on AR(p) models has been long recognized and goes back to Kilian (1998).[12]

Let $\widehat{\rho}_{OLS}$ denote the usual OLS estimator of ρ based on formulation (30). We employ the following bias-corrected estimator of ρ:

$$\widehat{\rho}_{BC} \equiv \widehat{\rho}_{OLS} + \frac{1 + 3\,\widehat{\rho}_{OLS}}{T} \ ; \tag{32}$$

for example, see White (1961). The corresponding bias-corrected estimators of $(\nu, \psi_1, \ldots, \psi_{p-1})$ are obtained by regressing $y_t - \widehat{\rho}_{BC} y_{t-1}$ on $(1, \Delta y_{t-1}, \ldots, \Delta y_{t-p-1})$ via OLS. By relation (31), we obtain in turn the bias-corrected estimators of formulation (29), denoted by $(\widehat{\nu}_{BC}, \widehat{\rho}_{1,BC}, \ldots, \widehat{\rho}_{p,BC})$.[13]

The corresponding, centered residuals $\widehat{\epsilon}_t$, for $p+1 \le t \le T$, are obtained as follows:

$$\widehat{\epsilon}_t \equiv \widehat{\epsilon}_{t,BC} - \frac{1}{T-p} \sum_{l=p+1}^{T} \widehat{\epsilon}_{l,BC} \quad \text{with} \quad \widehat{\epsilon}_{t,BC} \equiv y_t - \widehat{\nu}_{BC} - \widehat{\rho}_{1,BC} \cdot y_{t-1} - \ldots - \widehat{\rho}_{p,BC} \cdot y_{t-p} \ . \tag{33}$$

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

[13] 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), though the reader is referred to http://www.math.umbc.edu/∼anindya/errata.pdf for an errata concerning the latter reference.

The residual variance is

$$\widehat{\sigma}_{\epsilon}^2 \equiv \frac{1}{T-2p-1} \sum_{t=p+1}^{T} \widehat{\epsilon}_t^2 \ , \tag{34}$$

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

The prediction standard errors $\widehat{\sigma}_T(h)$ 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 $(\widehat{\nu}_{BC}, \widehat{\rho}_{1.BC}, \ldots, \widehat{\rho}_{p,BC})$, and denote the parameters of this MA(∞) model by $(\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\theta}_2, \ldots)$, with $\widehat{\theta}_0 \equiv 1$. Then compute

$$\widehat{\sigma}_T(h) \equiv \widehat{\sigma}_{\epsilon} \sqrt{\widehat{\theta}_0^2 + \ldots + \widehat{\theta}_{h-1}^2} \ . \tag{35}$$

Remark 4.1. It is well-known that the usual Box-Jenkins prediction standard errors (35) are somewhat too small in magnitude in finite samples, since they do not account for the estimation uncertainty in the model parameters $(\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\theta}_2, \ldots)$. 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 (37). Since the bias contained in the prediction standard errors is, approximately, the same in the real world compared to the bootstrap world, the resulting mistakes, approximately, cancel out and one still obtains joint prediction regions with very good finite-sample properties; see Subsection 4.3. ■

Bootstrap data $\{y_1^*, \ldots, y_T^*, y_{T+1}^*, \ldots, y_{T+H}^*\}$ 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 $\widehat{\epsilon}_{p+1}, \ldots, \widehat{\epsilon}_T$.

Second, let $y_t^* = y_t$, for $1 \leq t \leq p$ and then

$$y_t^* = \widehat{\nu}_{BC} + \widehat{\rho}_{1,BC} \cdot y_{t-1}^* + \ldots + \widehat{\rho}_{p,BC} \cdot y_{t-p}^* + \epsilon_t^* \ , \quad \text{for } p+1 \leq t \leq T \ . \tag{36}$$

Third, generate y_t^*, for $T+1 \leq t \leq T+H$, analogously to (36), but conditional on $\{y_{t-p+1}, \ldots, y_T\}$ rather than on $\{y_{t-p+1}^*, \ldots, y_T^*\}$.

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 4.2. 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 $(\widehat{\nu}^*, \widehat{\rho}_{1,BC}^*, \ldots, \widehat{\rho}_{p,BC}^*)$.

The bootstrap residual variance $\widehat{\sigma}_\epsilon^{2,*}$ is computed analogously to (34).

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

$$\widehat{\sigma}_T^*(h) \equiv \widehat{\sigma}_\epsilon^* \sqrt{(\widehat{\theta}_0^*)^2 + \ldots + (\widehat{\theta}_{h-1}^*)^2} \; . \tag{37}$$

4.2 Monte Carlo Setup

First, we consider an AR(1) model with $\nu = 0$ and with $\rho \equiv \rho_1 \in \{0.9, 0.5, -0.5, -0.9\}$. 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:

- $(\boldsymbol{\epsilon_t \sim N(0,1)})$ Standard normal.
- $(\boldsymbol{\epsilon_t \sim t_3})$ A t-distribution with 3 degrees of freedom, standardized to have variance one.
- $(\boldsymbol{\epsilon_t \sim \chi_3^2})$ A chi-square distribution with 3 degrees of freedom, centered to have mean zero and standardized to have variance one.

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 above three distributions.

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 $\{1, 2, \ldots, 5\}$.[14] This is the case both in the 'real' world and in the bootstrap world.[15] The sample size is $T \in \{100, 400\}$. For compactness, we only consider errors ϵ_t that are i.i.d. standard normal.

[14] The BIC is known to be a consistent information criterion, unlike the Akaike information criterion (AIC), say. Therefore, in terms of Assumption 3.1, using the BIC to estimate the order of an AR(p) model is asymptotically equally valid as using the true order.

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

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 coverage level $1-\alpha$.
- **(Scheffé)** The 'asymptotic' Scheffé JPR (20) of Jordà and Marcellino (2010).
- **(NP Heuristic)** The neighboring-paths heuristic bootstrap JPR of Staszewska-Bystrova (2010).
- **(k-FWE JPR)** Our two-sided bootstrap JPR (14).

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 $B = 1,000$ always.

All empirical coverages are computed from 1,000 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.

4.3 Results

In each case, we report the proportion of times that the k-FWE criterion is not violated. The thus obtained empirical coverages are then compared to the nominal coverage level given by $1 - 0.1 = 0.9 = 90\%$.

The results for the AR(1) model with $p = 1$ known are presented in Tables 1 and 2. The results for the AR(2) model with $p = 2$ known are presented in Tables 3 and 4. The results for the AR(2) model with $p = 2$ unknown and estimated using the BIC are presented in Table 5.

The various results can be summarized as follows:

- Joint Marginals always undercovers and its performance gets worse as the maximum forecast horizon H increases. This behavior is as expected and has been demonstrated before by Jordà and Marcellino (2010) and Staszewska-Bystrova (2010) already.
  Nevertheless, it is worth repeating the underlying message one more time: stringing together marginal prediction intervals does not result in a valid joint prediction region for the entire future path.
- The performance of Scheffé ranges from acceptable to horrible. For example, in the AR(1) model, the performance is acceptable for $\rho = 0.9$, where the empirical coverage is (reasonably) close to 90%; on the other hand, the performance is horrible for $\rho = -0.9$, where the coverage is, basically, equal to 0%.
  In general, the performance of Scheffé seems to decrease both in the largest autoregressive root — given by ρ in the AR(1) model and by $\rho_1 + \rho_2$ in the AR(2) model — and in the maximum forecast horizon H. In the vast majority of cases, the empirical coverage is unacceptably far away from the nominal level.
  As a consequence, Scheffé cannot be recommended for application in practice.

- The performance of NP Heuristic is quite good when the largest autoregressive root is close to one. Otherwise, the performance is acceptable: the empirical coverage is generally somewhat less than the nominal level and it 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 good to excellent. 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.
- There does not appear to be a 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 quite close to the corresponding coverages for known p, even for $T = 100$ already.

Remark 4.3. Our simulation results for Scheffé in the context of the AR(1) model are in general agreement with corresponding results reported by Jordà and Marcellino (2010, Table III).

They consider an AR(1) model with autoregressive coefficient $\rho \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, without stating the sample size T, the distribution of the errors, whether the order $p = 1$ is known or estimated from the data, and by which method the model parameters are estimated.

At any rate, Jordà and Marcellino (2010) also find that the performance of Scheffé decreases with the value of ρ and is poor when ρ is not close to one. For example, for $H = 12$ and $\rho = 0.5$, they report an empirical coverage of 33.3% for a nominal coverage level of 68% and an empirical coverage of 80.2% for a nominal coverage level of 95%.

Jordà and Marcellino (2010) do not consider any negative values of ρ, where we observe the worst performance of Scheffé; nor do they consider any values $H > 12$. ■

Nominal Coverage $1-\alpha = 90\%$

	$\epsilon_t \sim N(0,1)$			$\epsilon_t \sim t_3$			$\epsilon_t \sim \chi^2_3$		
$\rho = 0.9$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	66.7	53.3	38.2	71.1	58.6	43.6	69.3	56.2	41.3
Scheffé	86.6	85.9	84.9	86.9	85.5	83.6	88.7	87.7	86.5
NP Heuristic	90.1	90.6	90.9	88.9	88.1	86.6	90.6	90.7	90.4
1-FWE JPR	90.3	90.0	89.7	90.1	89.5	88.8	89.8	90.3	90.1
2-FWE JPR	90.2	89.6	89.4	90.3	89.4	88.2	90.1	90.0	89.9
3-FWE JPR	89.8	89.3	89.2	89.9	89.7	88.3	90.3	89.3	90.2
$\rho = 0.5$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	57.8	35.6	14.6	61.5	40.3	18.7	62.1	41.1	19.5
Scheffé	78.2	68.0	54.6	77.8	65.1	47.7	80.1	68.4	51.9
NP Heuristic	88.2	86.9	84.2	86.3	82.9	75.1	89.1	87.7	84.2
1-FWE JPR	89.8	89.0	88.0	89.3	87.8	84.0	89.8	88.2	85.8
2-FWE JPR	90.1	89.2	88.9	90.2	88.9	87.2	90.0	89.8	89.4
3-FWE JPR	89.9	89.5	89.3	90.0	90.2	88.5	90.3	89.3	90.2
$\rho = -0.5$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	58.6	36.6	15.3	62.2	41.1	19.7	62.6	41.4	19.2
Scheffé	08.1	00.1	00.0	20.3	05.4	01.0	15.7	02.6	00.1
NP Heuristic	87.6	85.7	82.3	85.8	82.2	75.0	87.9	85.9	81.1
1-FWE JPR	89.8	89.1	88.6	89.3	87.9	84.0	88.7	87.8	84.9
2-FWE JPR	89.9	89.4	89.3	90.2	89.4	87.3	89.5	89.3	88.7
3-FWE JPR	89.7	89.8	89.5	90.4	90.2	88.5	90.3	89.9	89.4
$\rho = -0.9$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	67.9	55.2	40.7	71.9	60.4	46.4	71.1	58.8	44.1
Scheffé	00.1	00.0	00.0	00.3	00.1	00.0	00.4	00.1	00.0
NP Heuristic	87.9	87.6	88.1	87.0	85.9	84.6	87.9	87.3	86.7
1-FWE JPR	90.0	89.8	90.4	89.8	89.3	88.3	89.5	89.4	89.2
2-FWE JPR	90.0	89.6	89.4	90.2	89.5	88.0	90.0	89.5	89.9
3-FWE JPR	89.9	89.7	89.3	90.0	89.5	88.1	90.1	89.8	89.8

Table 1: AR(1), Known Order, $T = 100$: Empirical Coverages.

Nominal Coverage $1-\alpha = 90\%$

	$\epsilon_t \sim N(0,1)$			$\epsilon_t \sim t_3$			$\epsilon_t \sim \chi^2_3$		
$\rho = 0.9$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	68.5	55.2	38.3	72.2	60.5	44.6	69.8	57.0	40.5
Scheffé	88.1	88.0	87.9	89.0	88.3	87.5	90.0	89.7	89.4
NP Heuristic	89.2	88.5	87.9	88.6	87.8	86.5	89.3	88.9	88.5
1-FWE JPR	90.0	90.0	89.9	90.1	90.1	89.8	90.0	90.2	89.8
2-FWE JPR	89.9	89.9	89.6	90.3	90.2	89.7	90.0	89.9	89.9
3-FWE JPR	90.0	90.0	89.9	90.1	90.1	89.9	90.1	90.0	90.0
$\rho = 0.5$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	58.1	34.8	12.8	60.9	38.6	16.0	60.2	37.9	15.3
Scheffé	80.8	70.5	51.3	80.7	68.3	47.8	81.3	69.3	49.3
NP Heuristic	89.0	87.7	85.5	88.3	86.4	82.6	89.2	88.1	86.2
1-FWE JPR	89.8	89.8	89.6	90.1	89.9	88.8	89.9	89.7	88.8
2-FWE JPR	89.9	89.7	89.5	90.3	90.3	89.4	89.8	89.9	89.6
3-FWE JPR	89.9	89.7	89.9	90.2	90.2	90.0	90.1	90.0	90.1
$\rho = -0.5$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	57.8	34.8	12.9	61.2	38.8	15.9	61.5	39.6	16.6
Scheffé	07.1	00.4	00.0	20.0	04.1	00.4	14.0	01.6	00.1
NP Heuristic	88.5	87.3	84.6	88.3	86.2	82.3	88.7	87.3	85.0
1-FWE JPR	89.9	89.9	88.8	90.0	89.8	89.0	89.8	89.7	88.7
2-FWE JPR	89.9	89.9	89.8	90.3	89.9	89.6	89.9	89.8	89.4
3-FWE JPR	90.0	90.1	89.9	90.2	90.1	90.2	90.2	90.1	89.7
$\rho = -0.9$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	68.4	55.4	38.8	72.6	60.9	45.3	71.4	59.0	42.6
Scheffé	00.1	00.0	00.0	00.1	00.0	00.0	00.1	00.0	00.0
NP Heuristic	88.6	87.9	86.8	88.4	87.4	85.8	88.8	87.8	86.8
1-FWE JPR	89.9	90.0	90.2	90.2	90.2	89.8	90.0	90.2	89.7
2-FWE JPR	89.8	90.1	90.0	90.3	90.3	89.7	89.9	89.9	89.8
3-FWE JPR	89.9	90.0	89.8	90.0	90.1	90.0	90.0	90.0	89.7

Table 2: AR(1), Known Order, $T = 400$: Empirical Coverages.

Nominal Coverage $1-\alpha = 90\%$

	$\epsilon_t \sim N(0,1)$			$\epsilon_t \sim t_3$			$\epsilon_t \sim \chi^2_3$		
$(\rho_1, \rho_2) = (1.75, -0.85)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	73.6	60.8	45.7	76.9	64.9	50.3	75.7	63.4	49.0
Scheffé	79.0	73.8	55.8	84.0	79.4	61.2	86.7	82.3	63.8
NP Heuristic	89.1	90.2	91.4	87.9	87.9	86.8	89.6	90.1	90.4
1-FWE JPR	89.0	88.2	87.2	89.0	87.6	86.6	88.6	88.6	87.2
2-FWE JPR	88.9	88.3	87.3	89.3	88.0	86.4	88.9	88.7	87.8
3-FWE JPR	88.8	88.8	88.2	89.5	88.4	86.6	89.3	88.9	88.3
$(\rho_1, \rho_2) = (1.25, -0.75)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	63.9	46.9	28.4	69.1	53.9	35.9	68.1	52.3	33.5
Scheffé	60.7	21.9	05.6	67.9	32.4	13.3	68.9	29.9	10.1
NP Heuristic	88.6	88.2	87.8	87.1	85.5	81.9	88.7	87.8	86.5
1-FWE JPR	90.0	89.6	89.5	89.7	88.5	86.5	89.3	89.3	88.1
2-FWE JPR	90.1	89.4	89.3	90.0	88.7	86.7	89.8	89.5	88.7
3-FWE JPR	89.8	89.6	89.2	90.2	89.6	87.0	90.2	89.0	89.2
$(\rho_1, \rho_2) = (-0.65, 0.15)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	63.9	48.2	30.6	67.6	52.9	35.5	67.6	52.1	33.8
Scheffé	03.6	00.2	00.0	09.1	02.0	00.3	07.2	00.8	00.1
NP Heuristic	88.1	87.7	87.6	86.4	84.6	81.1	88.2	87.2	85.3
1-FWE JPR	90.1	89.7	89.8	89.5	88.7	86.3	89.1	88.9	87.6
2-FWE JPR	89.8	89.2	89.2	90.2	89.4	87.5	89.8	89.3	89.2
3-FWE JPR	89.6	89.4	88.5	90.1	89.3	87.3	89.9	89.6	89.0
$(\rho_1, \rho_2) = (-0.7, -0.2)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	60.7	39.9	18.1	64.6	45.2	23.8	64.5	44.6	22.3
Scheffé	01.8	00.1	00.0	06.9	01.0	00.2	04.1	00.2	00.0
NP Heuristic	88.2	87.3	85.3	86.2	83.8	78.2	87.9	86.2	82.1
1-FWE JPR	88.8	89.4	88.9	89.2	87.8	84.3	88.8	88.0	85.3
2-FWE JPR	89.7	89.4	89.4	90.0	88.9	86.0	89.3	88.6	87.4
3-FWE JPR	89.7	89.7	89.5	90.3	90.0	88.0	90.5	89.9	88.9

Table 3: AR(2), Known Order, $T = 100$: Empirical Coverages.

Nominal Coverage $1-\alpha = 90\%$

	$\epsilon_t \sim N(0,1)$			$\epsilon_t \sim t_3$			$\epsilon_t \sim \chi^2$		
$(\rho_1, \rho_2) = (1.75, -0.85)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	75.9	63.2	46.4	78.9	67.5	51.7	76.9	64.5	48.2
Scheffé	80.2	76.1	56.5	86.2	82.7	65.5	88.4	85.5	89.9
NP Heuristic	89.2	89.0	88.6	89.1	88.4	87.2	89.3	89.1	88.3
1-FWE JPR	89.8	89.8	89.7	90.2	89.9	89.6	89.9	89.7	89.4
2-FWE JPR	89.9	89.6	89.5	90.3	90.1	89.3	89.8	89.7	89.4
3-FWE JPR	90.0	89.7	89.7	90.0	90.0	89.6	90.0	89.9	89.0
$(\rho_1, \rho_2) = (1.25, -0.75)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	65.3	47.2	25.8	69.8	53.8	33.5	68.5	51.9	30.9
Scheffé	62.5	20.0	03.2	72.0	33.2	10.7	72.0	28.4	06.3
NP Heuristic	88.8	87.9	86.5	88.5	87.4	84.9	89.0	87.9	86.0
1-FWE JPR	89.9	89.9	90.0	90.1	90.1	89.5	89.8	89.9	89.6
2-FWE JPR	89.9	89.8	89.8	90.3	90.2	89.5	89.8	89.8	89.4
3-FWE JPR	90.0	89.7	89.7	90.0	90.1	89.6	90.0	90.0	89.8
$(\rho_1, \rho_2) = (-0.65, 0.15)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	64.1	47.2	26.6	67.8	52.1	31.8	67.3	51.0	30.6
Scheffé	03.2	00.1	00.0	09.8	01.3	00.1	06.7	00.5	00.0
NP Heuristic	88.8	88.0	86.1	88.2	87.0	84.2	88.7	87.7	86.3
1-FWE JPR	89.9	89.7	90.1	90.2	90.1	89.3	89.8	90.0	89.5
2-FWE JPR	89.9	89.8	89.9	90.3	90.2	89.7	90.0	89.8	89.6
3-FWE JPR	90.0	89.9	89.7	90.0	90.1	90.0	90.0	89.9	89.8
$(\rho_1, \rho_2) = (-0.7, -0.2)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	59.8	37.3	14.8	64.0	42.6	19.2	63.5	42.1	18.9
Scheffé	01.7	00.1	00.0	06.7	00.5	00.1	03.7	00.1	00.0
NP Heuristic	88.7	87.8	85.4	88.2	86.6	83.4	88.6	87.3	85.2
1-FWE JPR	89.9	89.8	89.9	89.9	89.9	88.9	89.8	89.8	88.7
2-FWE JPR	90.1	89.8	89.8	90.1	89.9	89.1	89.9	89.5	89.1
3-FWE JPR	90.0	90.0	89.8	90.0	90.2	90.1	90.2	90.0	89.6

Table 4: AR(2), Known Order, $T = 400$: Empirical Coverages.

Nominal Coverage $1-\alpha = 90\%$

	$T=100$, $\epsilon_t \sim N(0,1)$			$T=400$, $\epsilon_t \sim N(0,1)$		
$(\rho_1, \rho_2) = (1.75, -0.85)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	72.1	61.8	49.2	76.2	64.5	48.0
Scheffé	87.9	86.0	64.4	89.2	88.8	66.1
NP Heuristic	89.2	91.5	93.1	89.8	90.7	90.5
1-FWE JPR	90.4	90.5	89.6	89.8	89.7	87.6
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
$(\rho_1, \rho_2) = (1.25, -0.75)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	63.6	46.1	27.0	65.3	47.1	25.5
Scheffé	63.7	23.2	07.5	66.5	21.6	04.2
NP Heuristic	87.9	86.7	85.8	88.8	87.8	86.0
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
$(\rho_1, \rho_2) = (-0.65, 0.15)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	65.1	48.9	30.4	64.5	47.2	26.2
Scheffé	02.6	00.2	00.0	02.9	00.1	00.0
NP Heuristic	88.8	87.9	86.8	89.1	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
$(\rho_1, \rho_2) = (-0.7, -0.2)$	$H=6$	$H=12$	$H=24$	$H=6$	$H=12$	$H=24$
Joint Marginals	59.9	39.5	18.2	59.6	37.3	14.9
Scheffé	03.0	00.1	00.0	01.9	00.1	00.0
NP Heuristic	87.8	86.9	85.3	88.7	87.7	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

Table 5: AR(2), BIC Order Selection: Empirical Coverages.

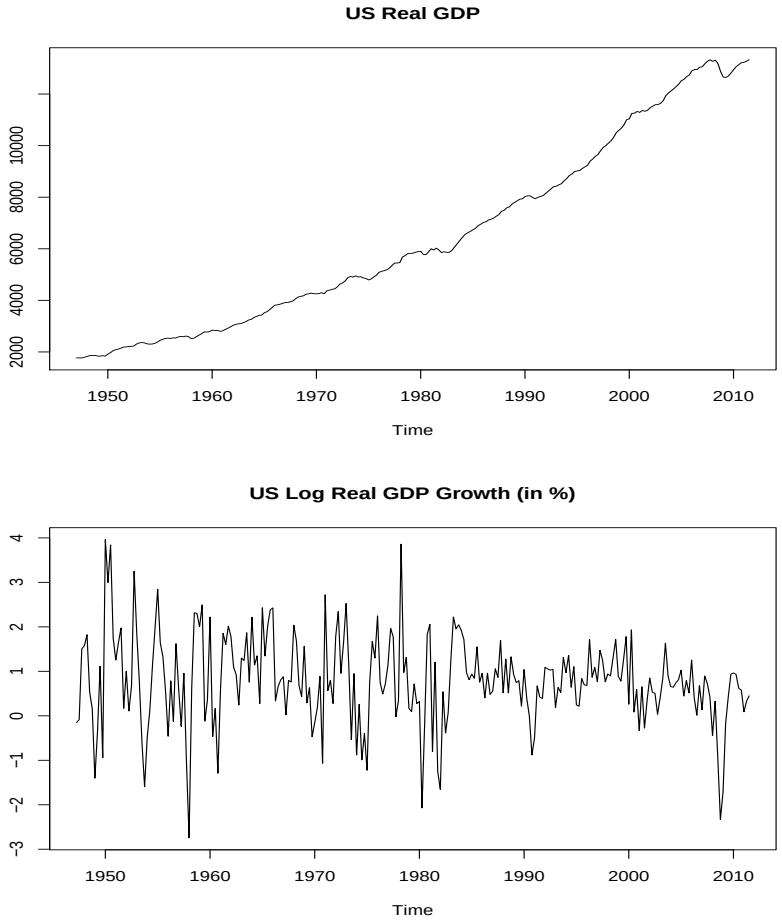

Figure 3: Quarterly data on US real gross domestic product (in 2005 chained dollars) from Q1/1947 until Q3/2001. The upper panel displays the raw data and the lower panel displays the first differences of the logarithmic data (in percent).

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.[16] The data are seasonally adjusted and expressed in billions of chained 2005 dollars. Figure 3 displays the raw data as well as the first differences of the logarithmic data (in percent). 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 three years. The nominal coverage is given by $1 - \alpha = 90\%$.

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

We use the AR(p) methodology described in Section 4 to compute bootstrap joint prediction regions, 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 variables (for example, see Stock and Watson, 2001) or a nonlinear forecasting model (for example, see 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 joint prediction regions 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 joint prediction regions 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 $\widehat{p} = 1$. The initial model fitted via OLS is given by

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

Using the bias correction (32) yields the following final model used for forecasting purposes:

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

Figure 4 compares Scheffé, NP-Heuristic, and 1-FWE JPR.[17] 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.[18]
- 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; see Subsection 3.3.

Figure 5 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 joint prediction region (with prespecified probability 90%), he obtains a smaller and more informative region in return.

[17] The number of bootstrap samples for NP-Heuristic and k-FWE JPR is $B = 10,000$.

[18] The width of Scheffé at forecast horizon h can be decreasing in h, for large values of h, since the multipliers $\sqrt{\chi^2_{h,1-\alpha}/h}$ used in the Scheffé joint prediction region (20) are strictly monotonically decreasing in h for $\alpha = 0.1$; see Figure 2.

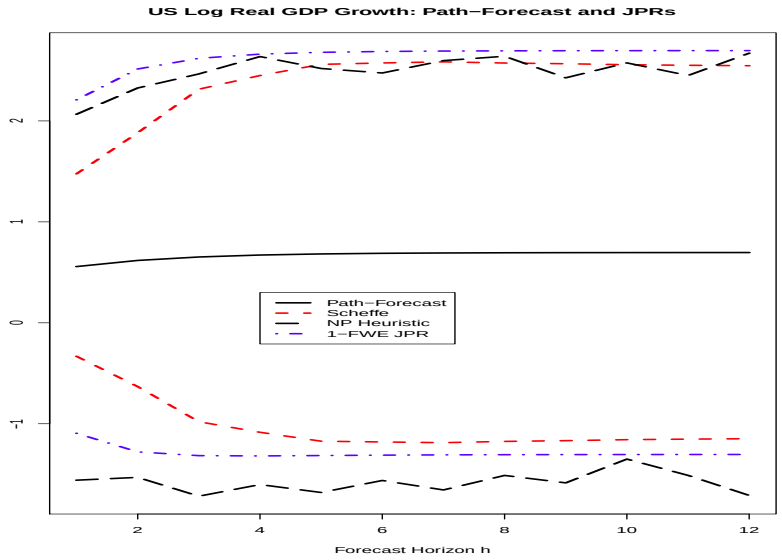

Figure 4: Path-forecast and various joint prediction regions for US log real GDP growth. The forecast period ranges from Q4/2011 until Q3/2014. The nominal coverage is given by $1 - \alpha = 90\%$.

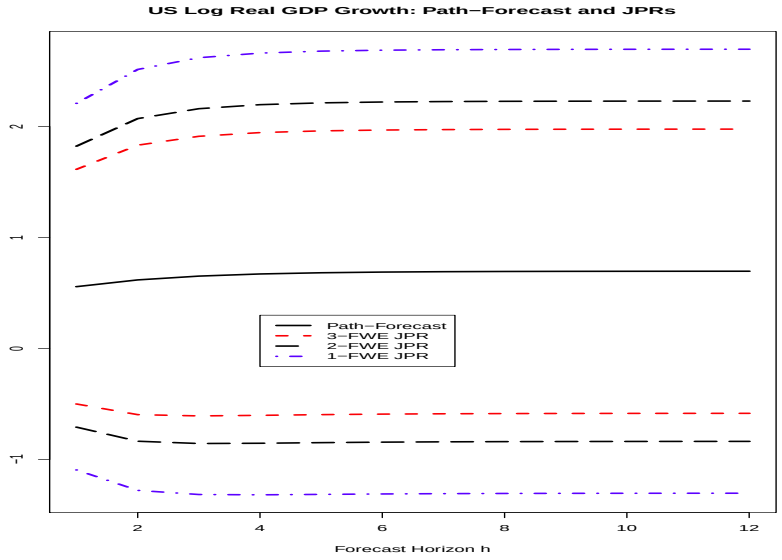

Figure 5: Path-forecast and various joint prediction regions for US log real GDP growth. The forecast period ranges from Q4/2011 until Q3/2014. The nominal coverage is given by $1 - \alpha = 90\%$.

5.2 Backtest Exercise

Although the previous exercise serves to illustrate the salient features of the various joint prediction regions, it does not address their performance in terms of coverage. First, the future data ranging from Q4/2011 until Q3/2014 have not been observed yet (at the time of writing this paper). Second, even when these data become eventually known, they only correspond to a single instance of a future 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 joint prediction region (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 get 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.

The results are presented in Table 6.[19] 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.

Nominal Coverage $1 - \alpha = 90\%$

Method	Empirical Coverage
Joint Marginals	64.6
Scheffé	73.2
NP Heuristic	89.7
1-FWE JPR	89.9
2-FWE JPR	85.1
3-FWE JPR	87.3

Table 6: Empirical Out-Of-Sample Coverages for US Log Real GDP Growth.

[19] The number of bootstrap samples for NP-Heuristic and k-FWE JPR is $B = 5,000$.

6 Conclusions

Many economic and financial 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 joint prediction region. Such a region is supposed to contain the entire future path with a prespecified probability $1 - \alpha$.

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

Compared to two previous proposals in the literature, our bootstrap joint prediction regions 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 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 (with prespecified probability $1 - \alpha$) 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 researchers that a small number, like one or two, of elements of the future path fall outside the joint prediction region. In return, he will then obtain a smaller and more informative region.

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.

Beran, R. (1988b). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83:687–697.

Bowden, D. C. (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, M. P. and Taylor, N. (2001). Bootstrapping prediction intervals for autoregressive models. *International Journal of Forecasting*, 17:247–267.

De Gooijer, J. G. and Hyndman, R. J. (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*. Springer, New York.

Jordà, Ò., Knüppel, M., and Marcellino, M. G. (2010). Empirical simultaneous confidence regions for path-forecasts. Discussion Paper No. DP7797, CEPR. Available at SSRN: `http://ssrn.com/abstract=1611493`.

Jordà, Ò. and Marcellino, M. G. (2010). Path-forecast evaluation. *Journal of Applied Econometrics*, 25:635–662.

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics*, 80:218–230.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Pascual, L., Romo, J., and Ruiz, E. (2001). Effects of parameter estimation on prediction densities: a bootstrap approach. *International Journal of Forecasting*, 17(1):83–103.

Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science*, 18:219–230.

Potter, S. M. (1995). A nonlinear approach to US GNP. *Journal of Applied Econometrics*, 2:109–125.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008). Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2):404–447.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Annals of Statistics*, 35(4):1378–1408.

Romano, J. P. and Wolf, M. (2010). Balanced control of generalized error rates. *Annals of Statistics*, 38(1):598–633.

Roy, A. and Fuller, W. A. (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*. John Wiley & Sons, New York.

Staszewska-Bystrova, A. (2010). Bootstrap prediction bands for forecast paths from vector autoregressive models. *Journal of Forecasting*, Online Version, DOI:10.1002/for.1205.

Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115.

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

White, H. L. (2001). *Asymptotic Theory for Econometricians*. Academic Press, New York, revised edition.

White, J. (1961). Asymptotic expansions for the mean and variance of the serial correlation coefficient. *Biometrika*, 48:85–95.

A Proofs

Proof of Proposition 3.1: We prove the stated result for the joint prediction region (15). The proofs for the joint prediction regions (14) and (16) are completely analogous.

Let $\widehat{L}_T$ denote a random variable with distribution $\widehat{J}_T$ and let $\widehat{L}$ denote a random variable with distribution $\widehat{J}$. By Assumption 3.1 and the continuous mapping theorem, k-max$(\widehat{L}_T)$ converges weakly to k-max$(\widehat{L})$, whose distribution is continuous. Our notation implies that the conditional sampling distribution under $\mathbb{P}$ of k-max$(\widehat{S}_T(H))$ is identical to the distribution of k-max$(\widehat{L}_T)$. By similar reasoning, the conditional sampling distribution under $\widehat{\mathbb{P}}_T$ of k-max$(\widehat{S}^*_T(H))$ also converges weakly to the distribution of k-max$(\widehat{L})$. To then show that

$$\mathbb{P}\big\{k\text{-max}(\widehat{S}_T(H)) \leq d^*_{1-\alpha}(k)\big\} \to 1-\alpha \tag{40}$$

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 (15),

$$k\text{-FWE} = 1 - \mathbb{P}\big\{k\text{-max}(\widehat{S}_T(H)) \leq d^*_{1-\alpha}(k)\big\} \ , \tag{41}$$

the proof that the stated result (17) holds for the joint prediction region (15) now follows immediately from (40). ■

B Generalized Error Rates, Multiple Testing, and Joint Confidence/Prediction Regions

The goal of this appendix is to explain why control of the *false discovery rate* (FDR) is actually equivalent to control of the *familywise error rate* (FWE) in the context of joint confidence regions and joint prediction regions.

In doing so, we first need to discuss some concepts from the literature on *multiple testing*. In a multiple testing problem one considers H individual hypotheses of the kind

$$H_{0,h} : \mu_h = \mu_{0,h} \quad \text{vs.} \quad H_{1,h} : \mu_h \neq \mu_{0,h} \ . \tag{42}$$

(For concreteness, we consider two-sided hypotheses here; one could also consider one-sided hypotheses instead.) The goal is to make individual decisions, in terms of rejecting or not, concerning each $H_{0,h}$ while controlling a prespecified error rate.

Denote by $I(\mathbb{P})$ the set of true null hypotheses, that is,

$$I(\mathbb{P}) \equiv \{h : H_{0,h} \text{ is true}\} \ . \tag{43}$$

The most stringent error rate is the *familywise error rate* (FWE), defined as the probability of rejecting at least one true null hypothesis:

$$\text{FWE} \equiv \mathbb{P}\{\text{Reject at least one of the } H_{0,h} : h \in I(\mathbb{P})\} \ . \tag{44}$$

It is worth to pause a moment here and to note that the FWE in the context of multiple testing only depends on the set of true null hypotheses. This is in contrast to definition (9) where the FWE depends on all components μ_h. The reasons is that in the context of constructing a joint confidence region for $\boldsymbol{\mu}$, there are no true and false parameters μ_h; they are all 'true' and of interest. Similarly for the definition (10) of the FWE in the context of constructing joint prediction regions: all components y_h are 'true' and of interest.

When control of the FWE is deemed too stringent in the context of multiple testing, one can control *generalized error rates* instead. Such generalized error rates are more liberal in terms of rejecting true null hypotheses and, in return, offer a greater ability to reject false null hypotheses.

The most popular generalized error rate, to date, is the *false discovery rate* (FDR). It is the expected value of the *false discovery proportion* (FDP). When applying a multiple testing procedure there will be a (random) total number of R rejections out of the H individual decision problems. Out of these R total rejections, there will be F false rejections (that is, rejections of true null hypotheses). Then one defines

$$\text{FDP} \equiv \frac{F}{R} \quad \text{and} \quad \text{FDR} \equiv \mathbb{E}(\text{FDP}) \ , \tag{45}$$

with $\text{FDP} \equiv 0$ in case there are no rejections at all. Control of the FDR amounts to ensuring that $\text{FDR} \leq \gamma$, for some prespecified (small) value $\gamma \in (0, 1)$.

Crucially, the definitions of the FDP and the FDR in the context of multiple testing rely on the notion of a subset of true hypotheses out of the universe of all H hypotheses. But the equivalent of such a subset does not exist in the context of a joint confidence region for $\boldsymbol{\mu}$: all components μ_h are 'true' and of interest. Therefore, controlling the FDR is nonsensical in such a context. In particular, whenever there are any components μ_h at all not contained in the joint confidence region, the FDP is automatically equal to one. And so control of the FDR is actually equivalent to control of the FWE. The reason is that ensuring that $\mathbb{E}(\text{FDP}) \leq \gamma$ is equivalent to ensuring that

$$\mathbb{P}\{\text{At least one } \mu_h \text{ not contained in the JCR}\} \leq \gamma \ . \tag{46}$$

For the same reason, control of the FDR is equivalent to control of the FWE also in the context of constructing a joint prediction region for Y.

Remark B.1. As an aside, the FDR is, arguably, more popular than it deserves to be. Many applied researchers do not really understand this error rate and the implications when it is applied to a set of data. Since the FDR is the *expected value* of the FDP, little can be said about the realized value of the FDP after applying a multiple testing method which controls the FDR to a set of data. On the other hand, many applied researchers seem to believe that the realized FDP can be at most γ. But such belief is just as valid as believing that the realization of random variable drawn from the standard normal distribution can be at most zero (since the standard normal distribution has expected value zero).

If a statement concerning the realized FDP is the goal, one should control the FDP instead in the sense of ensuring that

$$\mathbb{P}\{\text{FDP} > \gamma\} \leq \alpha \ . \tag{47}$$

In this way one can be $1 - \alpha$ confident that the realized FDP is at most γ. The reader is referred to Korn et al. (2004) and Romano et al. (2008) for a more detailed discussion. ■

Although control of the FDR is not a meaningful alternative, it is possible to construct joint confidence regions as well as joint prediction regions based on a generalized error rate that is meaningful in these contexts. The solution is to use the *generalized familywise error rate* (k-FWE). Start with the context of multiple testing. For an integer $k \geq 1$, the definition is

$$k\text{-FWE} \equiv \mathbb{P}\{\text{Reject at least } k \text{ of the } H_{0,h}\colon\ h \in I(\mathbb{P})\} \ . \tag{48}$$

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.

Realizing that in the contexts of estimating and forecasting, all components are 'true' and of interest, the definition of the k-FWE can easily be adapted as already described in (11)–(12). For a joint confidence region (JCR) for $\boldsymbol{\mu}$,

$$k\text{-FWE} \equiv \mathbb{P}\{\text{At least } k \text{ of the } \mu_h \text{ not contained in the JCR}\} \ ,$$

whereas for a joint prediction region (JPR) for Y,

$$k\text{-FWE} \equiv \mathbb{P}\{\text{At least } k \text{ of the } y_h \text{ not contained in the JPR}\} \ .$$