Predictive Inference for a Wide Class of Temporal Data

Matheus Silva

October 2022

Please find the latest version here

Abstract

Forecasters usually report point predictions; however, understanding the randomness around such values is of practical importance. An example: a central bank predicts 2% inflation next quarter but is also interested in an interval (0%-4%, for example) that will contain the future realization of this series with a set probability. I show how to construct intervals as such, and I prove their asymptotic validity. I propose a model free method that encompasses, but not limited to, any off-the-shelf machine-learning method including high-dimensional ones. The method is based on a subsampling estimation strategy, consisting of analysing smaller cuts of the original time series. I prove the prediction intervals constructed with the subsampling method remain valid even when the data exhibits nonstationarities of many kinds — such as time-varying parameters, structural breaks, unit roots, and transitions between steady-states. In addition to this theoretical work, I provide simulation studies to show the numerical performance of this method. I also apply the method to a demand dataset and to the forecast of inflation in a high-dimensional setup. The subsampling procedure extends to allow for comparisons of predictive accuracy between different models.

1 Introduction

Economic forecasting is undoubtedly valuable in guiding decision-making, especially for those decisions whose consequences will be felt for a long period of time. Forecasting provides insights that can guide people to choose the right plan of action, such as central banks setting interest rates, investors designing their portfolios, or businesses making inventory decisions. In a random world, however, quantifying the uncertainty around those forecasts to assess the risks involved in each decision is also important. Although many forecasting methods provide point predictions, few provide a method for understanding the uncertainty involved in the forecasting process.

Quantifying uncertainty around forecasts has widespread uses in economics, for practitioners and researchers alike. Monetary authorities bound by some inflation target want to be cautious when making decisions in the presence of uncertainty around the future states of the economy (Svensson, 1997). Understanding the variability around predictions also affects financial investments because one wants to quantify the risk of their portfolio (the value at risk, VaR). A firm planning its inventory would not want to overpay for ultimately unused goods but also wouldn't want to risk understocking – understanding the fluctuations of demand can help them make a contingency plan in a way that they would not be able to if using point predictions alone. A researcher may be interested in comparing predictions from many different economic forecasters; for example, one can study the point predictions reported in in the Survey of Professional Forecasters of the Philadelphia FED (Engelberg et al., 2009).

Economies are in perpetual motion. Policy changes, periods of high uncertainty, and movements in the international economic conjuncture are only a few reasons why economic data are naturally unstable. Models that do not account for those features relegate those dynamics to the error terms. If the properties of the error term are unknown, characterizing uncertainty in forecasts is challenging. Often, one makes assumptions on the stochastic characteristics of the randomness in the forecasting model either explicitly by including more equations or implicitly in the estimation stage. For example, the use of DSGE models in forecasting macroeconomic series invariably suffers from misspecification. If these models are estimated with the Kalman filter, the uncertainty around the forecasts rely on an underlying assumption of gaussian errors. Even slight deviations from those assumptions will result in incorrect size of both prediction intervals and hypotheses tests related to the model's predictions.

In this paper, I propose a procedure to conduct inference on predictions – intervals and hypotheses tests – that is robust to mischaracterization of the dynamics of data during the modeling process. In brief, the procedure consists of studying smaller, sequential, and overlapping blocks of the original data. The robustness properties of the method derive precisely from the using the smaller sets of observations. Intuitively, the robustness comes from not using a fixed model to study an unstable series with time-varying statistical properties in favor of smaller chunks of data similar to each other. Analyzing those smaller cuts of data allows the equations to capture local dependence structures of the series – precisely the instabilities that need to be accounted for.

The method relies on subsampling. Subsampling has been studied in the context of statistical inference on parameters (Politis, Romano, and Wolf (1999) provide a comprehensive review on the topic). However, subsampling had not been studied in the context of predictions and, since the starting point of questions related inference and prediction are different, many insights are not interchangeable. In contrast to statistical inference, the study of forecasts must consider not only the randomness from the sampling and estimation processes, but also the variability of the components not explicitly included in the model. This last random component is what separates forecasting from inference since a structural error will always be present regardless the sample size increases.

A vast literature studies forecasts with different approaches: parametric estimations, prediction bootstraps, and conformal prediction intervals are perhaps the main ones. As discussed earlier, any misspecification will be relegated to the error terms, yielding incorrect intervals[1]. In the semi-parametric front, Cao (1999) reviews different resampling procedures to create prediction intervals for linear models – a prediction version of the bootstrapping procedure. More recently, the machine-learning literature has developed conformal prediction inference, a data-driven approach to construct probabilistic bounds on yet-to-be-realized data. Although each of these methods has their advantages, whether in terms of statistical efficiency or computational ease, they all hinge on strong assumptions on the dependence of the time series. More specifically, the linear and nonlinear dependences of the series must

[1] This includes both standard frequentist and bayesian procedures since the likelihood functions are incorrectly specified.

decay sufficiently fast. An important violation of this requirement is models with conditional heteroskedasticity used to study, among other things, the volatility of stock returns.

The method I present is valid under a wide range of dependence structures in the data and robust to misspecification. First, the effects of model misspecification are included in the error terms of the model, which the subsampling method is able to handle correctly. To illustrate, consider a forecaster using a linear one-lag autoregressive model to forecast future inflation. All other dynamics in the series are lumped into the error term, including moving average components, nonlinearities, and other relevant covariates. By sweeping over the subsamples, I show one can consistently recover the distribution of the total prediction errors, which include not only the structural errors from the data-generating process, but also misspecifications of the functional form and other characteristics of the error terms not explicitly modeled.

Second, the method allows for structural breaks, heteroskedasticity, and other structural shifts in the economy. Stock and Watson (1996) document significant structural instability in 76 monthly time series, including income, production, inflation, interest-rate spreads, and stock prices. McConnell and Perez-Quiros (2000) document that the American economy went through a structural break at the end of the 20th century. This break is known as the "Great Moderation" period, known for the decreased volatility of the time series of GDP growth, inflation, and other macroeconomic aggregates. The subsampling procedure can deal with such breaks and other nonstationarities through the same mechanisms.

Ultimately, for the method to work, the forecast errors (the differences between actual realizations and the forecasts from the statistical model in hand) must satisfy a mixingale condition. Mixingales are a generalization of the concept of martingale differences commonly employed in the study of asset prices and unit-root series, but it allows for short-term dependence. I show many economic models satisfy this condition, including those that do not have straightforward mixing properties, which is the case in conditional heteroskedasticity models. Section 2 provides the formal definitions in depth. I show the subsampling method can be used to study forecasts under this assumption. More specifically, I prove one can consistently estimate distributions of a variety of statistics of interest and, with those, construct prediction intervals and conduct hypotheses tests on parameters related to predictions.

The first application is the construction of unconditional and conditional prediction intervals. Although the construction of unconditional prediction intervals has been widely studied, the method allows for the construction of robust conditional prediction intervals – those that incorporate information provided by the econometrician. On the other hand, Lee and Scholtes (2014) point to a lack of research on conditional prediction intervals. To the best of my knowledge, the issue has not been addressed. In fact, over the last couple of years, empirical papers in the *International Journal of Forecasting* that report prediction intervals present unconditional versions, parametric conditional prediction intervals[2], or bayesian conditional credible intervals. I prove via simulations that the performance of such intervals is, in general, unsatisfactory and the subsampling method I develop here results in correct conditional prediction intervals with robustness properties.

Subsampling also allows for the construction of robust hypotheses tests. Even though the method allows the investigation of a variety of statistical hypotheses, I am particularly interested in the comparison of predictive accuracy between two models. The measure of predictive accuracy relates to the costs of making forecast errors. The statistic employed in this test is the average loss-differential between two prediction-generating models. These models may be nested or not – the first case is the analog of an F-test for the significance of a set of regressors in the prediction context. The subsampling approach allows for testing such hypotheses when the uncertainty characterization of the model is not possible by standard means, and resampling procedures as the bootstrap do not work. Here, the subsampling strategy consists of subsampling from the data, and then calculating the test statistics. This approach contrasts with Ibragimov and Muller's (2010, 2016) take on the problem, who subsample from the series of loss-differentials itself; see Zhu and Timmermann (2022).

I show the performance of the subsampling method in finite samples on both fronts with Monte Carlo simulations. For the prediction intervals, I study conditional and unconditional prediction intervals in the context of conditional heteroskedasticity – a typical violation of mixing properties of time series. To show the accuracy in non-standard contexts, I construct bands around forecasts from a high-dimensional linear regression estimated with the LASSO, a nonlinear method to estimate high-dimension linear regression models. The numerical simulations corroborate the theoretical results by showing the good performance of both conditional and unconditional prediction intervals in terms of coverage.

The Monte Carlo simulations to study the predictive accuracy of two different models revisit McCracken's (2020) counterexample to Giacomini and White's (GW, 2006) work on

[2] The complete list can be found in the Appendix.

this problem. McCracken (2020) shows the GW test statistic is not asymptotically normal in the presence of long memory – even when the observations time series of interest are i.i.d.. By applying the subsampling method to this example, I show how to conduct the test in this case. The simulations indicate subsampling provides robust confidence intervals that have correct nominal coverage.

Finally, I provide two empirical applications. First, I create prediction intervals around inflation forecasts in a big-data setting (McCracken and Ng, 2015), calculated using a high-dimensional linear regression model – a problem present in many business and policy decisions, such companies pricing their products and central banks studying economic activity.

The second empirical application concerns demand forecasting. I apply my method to construct prediction intervals for a real series of sales of a specific tractor part. The data are a time series that adds the total quantity sold to the lost potential sales, that is, the total number of parts quoted and the orders could not be fulfilled due to the lack of stock. In this case, I study an asymmetric loss function where understocking is more costly than overstocking.

The rest of this paper is structured as follows: Section 2 describes the technical requirements on the structural of the data for my method to work; Section 3 presents the subsampling algorithm in the context of the construction of prediction intervals. Sections 4 and 5 provide an illustrations of the method in the context of inflation forecasting in a big-data setup and business forecasting, respectively. Section 6 describes the subsampling algorithm in the context of hypotheses tests and provides simulations proving the numerical efficacy of subsampling in this case.

2 Framework

I start by formally describing the forecasting problem and the data-generating processes considered in this paper.

2.1 The dependence structure of the data

Let $\{y_t\}$ be a stochastic process of which we are interested in predicting. I abstract from covariates $\{X_t\}$ for the time being. I now describe the conditions on $\{y_t\}$ required for the correct quantification of uncertainty.

The potentially multivariate time series of interest $\{y_t\}$ is generated according to an underlying, also potentially multivariate latent stochastic process $\{Z_t\}$ on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. $\{y_t\}$ is a very general function of $\{Z_t\}$,

$$y_t = \varphi_t(\dots, Z_{t-1}(\omega), Z_t(\omega), Z_{t+1}(\omega), \dots) \tag{1}$$

where $\varphi_t(\cdot)$ is an $\mathcal{F}$-measurable function, which may depend on time t and may take infinite values of $\{Z_t\}$ as arguments.

This general model allows for a variety of important data-generating processes. Two highlights are nonseparable models where the error terms are not additive, and the possible dependence of the series on future shocks. This may be the case when the series $\{y_t\}$ is preprocessed by seasonal adjustments of the removal of a trend with the Hodrik-Prescott filter. It is important to remark that the model in (1) defines how the data is generated and not how one will forecast it as there is no knowledge about the future.

There are two technical conditions required moving forward. The first one is on the underlying stochastic process $\{Z_t\}$ which $\{y_t\}$ is a function of. The second is on the functions $\{\varphi_t(\cdot)\}$ that link the space $\{Z_t\}$ lives in and the one $\{y_t\}$ does.

$\{Z_t\}$ must satisfy a notion of asymptotic independence. That is, all dependence between two different blocks of $\{Z_t\}$ vanish the farther those blocks are from each other. How fast their dependence must disappear will depend on the moments of $\{Z_t\}$ and this condition is made explicit later. This concept if formally stated in Definition **??**.

Definition 2.1. *($\alpha-$mixing)* Let $\mathcal{M}(-\infty, t)$ and $\mathcal{M}(t+\tau, \infty)$ be the $\sigma-$fields spanned by $\{\cdots, Z_{t-1}, Z_t\}$ and $\{Z_{t+\tau}, Z_{t+\tau+1}, \cdots\}$, respectively. The mixing coefficient of $\{Z_t\}$, α_Z, is

$$\alpha_Z(\tau) \equiv \sup_{\substack{A \in \mathcal{M}(-\infty,t) \\ B \in \mathcal{M}(t+\tau,\infty)}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|$$

$\{Z_t\}$ is said to be α-mixing if $\alpha_Z(\tau) \to 0$, as $\tau \to \infty$.

The second condition is a restriction on the sequence of functions $\{\phi_t(\cdot)\}$ that generate $\{y_t\}$. This series must be such that the more information one obtains about the time series, the more one learns about it, and in the limit this information is enough to characterize all the series.

Definition 2.2. *(Near-epoch dependence)* Let $\{Z_t\}$ be a sequence of $\alpha-$mixing random variables in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{F}_{t-m}^{t+m}$ be the $\sigma-$algebra spanned by $\{Z_{t-m}, Z_{t-m+1}, \dots, Z_{t+m}\}$ and $\{\mathcal{F}_{t-m}^{t+m}\}_{m=0}^{\infty}$ is a non-decreasing sequence of $\sigma-$fields. Whenever convenient, I abuse this notation and define $\mathcal{F}_n \equiv \mathcal{F}_{-\infty}^{n}$. If $\{y_t\}$ is an integrable random variable and satisfies

$$\|y_t - \mathbb{E}[y_t|\mathcal{F}_{t-m}^{t+m}]\|_2 \leq d_t \nu_m$$

where $\nu_m \to 0$ as $m \to \infty$ and $\{d_t\}$ is a sequence of positive numbers, then $\{y_t\}$ is near-epoch dependent over $\{Z_t\}$.

To the best of my knowledge, this is the first work to provide results regarding uncertainty quantification around forecasts under this class of dependence. Past results rely on some sort of mixing condition which are much stronger than near-epoch dependence. In particular, my derivations do not exclude many time series models of central importance in the analysis of economic and financial data – the leading example explored in this paper are models that feature conditional heteroskedasticity, known to be near-epoch dependent. I show how the coverage properties of other methods are severely affected in this case in Section 3.4.2.

Example 2.1. *(AR(1) - Andrews, 1984) Let $\{y_t\}$ be generated according to*

$$y_t = \rho y_{t-1} + \varepsilon_t \;\; |\rho| < 1$$
$$\varepsilon_t \sim^d Bernoulli(p)$$

Andrews (1984) shows this process is not α-mixing. However, it is near-epoch dependent since

$$y_t = \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j}$$
$$\mathbb{E}[y_t|\mathcal{F}_{t-m}^{t+m}] = \sum_{j=0}^{m} \rho^j \varepsilon_{t-j}$$
$$\|y_t - \mathbb{E}[y_t|\mathcal{F}_{t-m}^{t+m}]\|_2 = \left\| \sum_{j=m+1}^{\infty} \rho^j \varepsilon_{t-j} \right\|_2 \to 0$$

Example 2.2. *(Random walk with fixed initial value) Let $\{y_t\}$ be generated according to*

$$y_t = y_{t-1} + \varepsilon_t, \;\; t \geq 0$$

where ε_t is a white noise with finite variance and y_0 is fixed. $\{y_t\}$ is not strongly mixing since the dependence between realizations in two different periods τ and τ', y_τ and $y_{\tau'}$, does not vanish. However, it is near-epoch dependent since

$$\begin{aligned} y_t &= \sum_{\tau=0}^{t} y_\tau \\ \mathbb{E}[y_t|\mathcal{F}_{t-m}^{t+m}] &= \sum_{\tau=t-m}^{t} y_\tau \\ \|y_t - \mathbb{E}[y_t|\mathcal{F}_{t-m}^{t+m}]\|_2 &= \left\| \sum_{\tau=0}^{t-m-1} y_\tau \right\|_2 \to 0 \end{aligned}$$

as $m \to \infty$.

Example 2.3. *(Conditional heteroskedasticity) Hansen (1991) shows that the GARCH(1,1) process*

$$y_t = \sigma_t \varepsilon_t$$

where ε_t is white noise and σ_t is the conditional heteroskedasticity, modeled as $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha y_{t-1}^2$ is near-epoch dependent when

$$\mathbb{E}\left[\left(\beta + \alpha\varepsilon_t^2\right)^r |\mathcal{F}_{t-1}\right]^{1/5} \leq c < 1, \; \forall t$$

All the definitions presented above are generalizable to the case with covariates. The reason is that one may extend the underlying random process $\{Z_t\}$ to account for the variables in $\{X_t\}$. Let $\{\tilde{Z}_t\}$ be an α-mixing process in a complete probability space $(\tilde{\omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ over which $\{X_t\}$ is near-epoch dependent. Let $\{y_t\}$ be generated as

$$\begin{aligned} y_t = \varphi_t(&\underbrace{\ldots, Z_{t-1}(\omega), Z_t(\omega), Z_{t+1}(\omega), \ldots}_{\text{“Original” underlying process}}; \\ &\underbrace{\ldots, X_{t-1}(\tilde{\omega}), X_t(\tilde{\omega}), X_{t+1}(\tilde{\omega}), \ldots}_{\text{Other predictors}}) \end{aligned}$$

if we can now combine the probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\tilde{\omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ into $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$ with the same mixing properties, then we can write

$$y_t = \varphi_t^* \underbrace{(\ldots, Z_{t-1}^*(\omega^*), Z_t^*(\omega^*), Z_{t+1}^*(\omega^*), \ldots)}_{\text{Extended underlying process}}$$

2.2 Description of the forecasting problem

The econometrician has access to a sample of size n from the process of interest $\{y_t\}$ (and potentially other covariates $\{X_t\}$). The interest relies on forecasting the $(n+h)^{th}$ realization of $\{y_t\}$. Let $\hat{y}_{n+h|n}$ be any horizon-h forecast.

I consider forecasts for y_{n+h} generated using some statistical model – e.g., artificial neural networks, random forests, and regressions that involve shrinkage estimates. Denote this model by $\mathcal{Y}_\theta(\cdot)$, which is indexed by a vector of parameters θ and takes lags of $\{y_t\}$ and other covariates $\{X_t\}$ as inputs. It will be denoted $\mathcal{Y}_\theta(y, X)$.

I mostly consider forecasts that are optimal according to a loss (or cost) function. The loss function describes the cost incurred in making prediction errors, the difference between the forecast and its actual realization. Let e_t denote the forecasting error in date t and $\mathcal{L}(e)$ the loss (or cost) function[3] as a function of the forecast error e.

The optimal loss-based operational forecast h-periods ahead based on the sample of size n is denoted $\hat{y}^*_{n+h|n}$ and is constructed by finding a vector of estimates of the parameter vector θ that indexes $\mathcal{Y}_\theta$ that minimizes the in-sample loss.

$$\hat{\theta} \equiv \underset{\theta|\mathcal{Y}_\theta}{\arg\min} (n-1)^{-1} \sum_{t=2}^{n} \mathcal{L}(y_t - \hat{y}_{t|n}) \tag{2}$$

$$\hat{y}_{t|n} \equiv \mathcal{Y}_\theta(y_{t-1}, X_{t-1})$$

$$\hat{y}^*_{n+h|n} \equiv \mathcal{Y}_{\hat{\theta}}(y_n, X_n)$$

The loss functions are context-dependent. If the direction of the forecast error does not matter, then one may use a symmetric function. However, as it is the case in many business problems, forecasting errors in one direction are most costly than in the other. An example is a company that decides their inventory investments fearing losing sales due to lack of its product.

[3] I consider functions of this form. However, it is also possible to generalize this to $\mathcal{L}(y, \hat{y})$, where y is a realization of the process and $\hat{y}$ its forecast.

Example 2.4. *The mean squared forecast error cost function is the symmetric function* $\mathcal{L}(e) \equiv e^2$.

Example 2.5. *The linex cost function (Varian, 1975) is an asymmetric function with asymmetry parameters* c_1 *and* c_2 *is* $\mathcal{L}(e) \equiv c_1 \exp\{c_2 \times e\} - c_2 \times e - 1$.

I also work under the assumption that there is only one solution to the forecasting problem (2) and that a Weak Law of Large Numbers holds as well – just so the point forecast is unique and the the parameter estimate $\hat{\theta}$ converges in probability to θ.

3 Prediction intervals

I now lay down the problem of constructing prediction intervals for $\{y_t\}$ in some horizon h based on a sample available to the researcher. Then, I describe how to construct such intervals and the circumstances under which they have the correct asymptotic coverage. I end this Section with a discussion on the practical implementation of this method.

For exposition purposes, say we are interested in forecasting a univariate series $\{y_t\}$ one step head[4]. The goal is to construct intervals of the form $[\underline{y}; \bar{y}]$ that cover the realization y_{n+1} with some probability, say $100 \times \gamma\%$. There are two types of prediction intervals of interest. The unconditional

$$\mathbb{P}(\underline{y} \leq y_{n+1} \leq \bar{y}) \rightarrow^{n \rightarrow \infty} \gamma \tag{3}$$

and the conditional

$$\mathbb{P}(\underline{y} \leq y_{n+1} \leq \bar{y} | y_n) \rightarrow^{n \rightarrow \infty} \gamma \tag{4}$$

Semiparametric unconditional prediction intervals like the ones I study here have been analyzed in various contexts. On the other hand, the properties of conditional prediction intervals have not been investigated as thoroughly. The method I propose covers both cases. In practice, prediction intervals are model-dependent in the sense that their construction depends on the forecasting method used – within the class of forecasting problems considered in Section 2.2.

[4] This is the setup for the rest of the paper as well since the extension to multivariate series and other horizons is straightforward does not add further insights. Appendix D provides an illustration of the multivariate prediction problem.

Although prediction intervals such as the one above look similar to traditional parametric confidence intervals, they differ in the objects studied. Confidence intervals describe the uncertainty around estimates of parameters of some statistical model and are used for statistical inference. Prediction intervals concern future realizations of the series being studied and not some population parameter. Ultimately, this means that prediction intervals will contain two sources of uncertainty: one coming from the estimation of the forecasting model itself and the other coming from the model misspecification (the error term). To see this, define the forecast error

$$e_1 \equiv y_{n+1} - \hat{y}_{n+1|n}$$

For simplicity, assume that $\{y_t\}_t$ is generated according to an AR(1) model

$$y_t = \rho_1 y_{t-1} + \varepsilon_t$$

Take the one-period-ahead forecast $\hat{y}_{n+1|n}$ constructed by estimating this model with OLS,

$$\hat{y}_{n+1|n} = \hat{\rho}_1 y_n$$

The forecast error can be written as

$$e_1 = \underbrace{(\rho_1 - \hat{\rho}_1)}_{\text{Estimation error}} y_n + \underbrace{\varepsilon_{n+1}}_{\text{irreducible error}}$$

The first summand consists of the error from the estimation of the forecasting model itself. Under standard conditions, it vanishes are the sample size increases[5]. The other summand is the irreducible error, which consists of the original error term itself in this case. Instead of converging in probability to some number, it converges to some distribution. Therefore, the inference and the forecasting problems are different from each other yet share the same vocabulary. In reality, since the data-generating process is unknown, the irreducible error term ε_{n+1} will contain not only the original error term but also all the dynamics from model misspecification.

[5] In this case it converges to the true parameter in the data-generating process. In general, it converges to the linear projection coefficient when estimating that linear regression.

3.1 Constructing prediction intervals

The object of interest when constructing prediction intervals is the forecast error[6] based on the optimal operational forecast from the model $\mathcal{Y}$. The following manipulation shows that studying the distribution of the forecast error is equivalent to studying the distribution of the future outcome

$$\mathbb{P}(\underline{y} \leq y_{n+1} \leq \bar{y}) = \mathbb{P}(\underline{e} \leq y_{n+1} - \hat{y}_{n+1|n} \leq \bar{e}) = \mathbb{P}(\underline{e} \leq e_1 \leq \bar{e}) \tag{5}$$

That is also true for the conditional counterpart

$$\mathbb{P}(\underline{y} \leq y_{n+1} \leq \bar{y}|y_n) = \mathbb{P}(\underline{e} \leq y_{n+1} - \hat{y}_{n+1|n} \leq \bar{e}|y_n) = \mathbb{P}(\underline{e} \leq e_1 \leq \bar{e}|y_n) \tag{6}$$

The goal is to construct intervals that do not rely on distributional assumptions and that maintain correct asymptotic coverage under conditional and unconditional heteroskedasticity, among other sources of nonstationarities. I propose a subsampling method to tackle both issues simultaneously. Subsampling methods have been studied in the context of statistical inference (see Politis, Romano, and Wolf (1999) for an extensive treatment of this case). Theorem 3.1 argues that the subsampling procedure extends to forecasting problems as well.

The method consists of using small cuts of the time series, which allow for capturing the nonstationarities because each window will capture the structural changes behind the aggregate series. The basic structure of the subsampling algorithm for the construction of prediction intervals is:

1. Split the original time series into smaller, sequential time series.
2. Leave out the last observation in each split.
3. Calculate the subsampled prediction error.
4. Use the distribution from (1)-(3) to construct the intervals.

The subsampling approach has advantages to other methods to construct prediction intervals. Perhaps the simplest competitors are parametric approaches. Those have clear

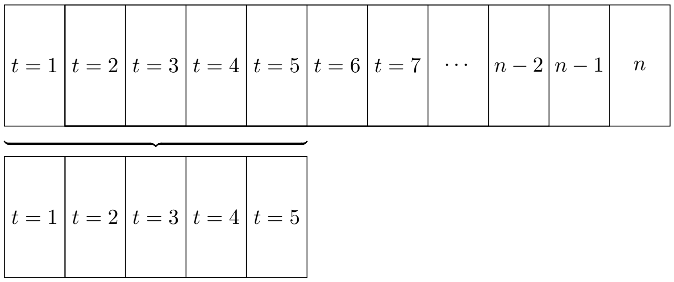
[6] Politis (2013) calls it the predictive root.

advantages when the prediction model is correctly specified, since maximum likelihood estimation will give unbiased and efficient estimates (at least asymptotically). On the other hand, even slight misspecifications hurt those advantages, resulting in prediction intervals that have incorrect coverage. Another strategy, explored in Cao (1999), consists of estimating some forecasting model using the whole sample, then resampling from the residuals to create an artificial time series with the same dependence properties. Related to this bootstrap strategy is the moving-blocks bootstrap, which consists of resampling blocks of the original time series, stitching them together, and then creating prediction intervals from this resampled data. However, any bootstrap procedure depends on the data being stationary, which is not the case of interest in this paper. For example, Cao's (1999) description of his bootstrap method requires the series to be stable over time since the first step consists of estimating a model over the whole sample, and then all the residuals are treated equally in the resampling step. The block bootstrap procedure requires stationary so that two blocks that are far apart from each other carry the same dependence structure, which can then be estimated and used to create prediction intervals.

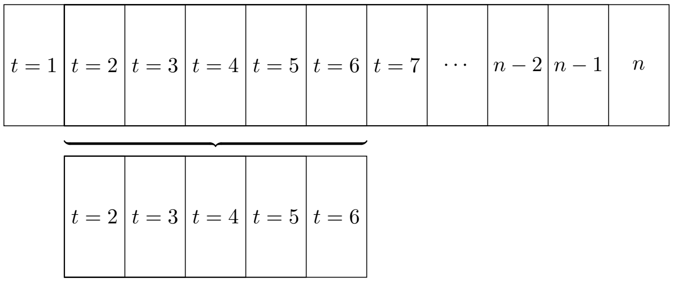
Recently, there has been a revival of Vovk's (2005) work on conformal prediction in the machine learning literature. In simple terms, conformal prediction intervals are constructed based on a measure of "unlikelihood" of conjectured new observations, conditional on the data in hand – the intervals will not contain values of the variable of interest that are far away from the observed set (according to the measure of "unlikelihood" employed). In his book, Vovk (2005) presents a variety of results concerning the correct coverage of such intervals when the data is i.i.d.. With dependent data, the main results come from Chernozhukov et al. (2018) and Xu and Xie (2022). They show that, under quadratic loss and stationarity of the underlying stochastic processes, conformal prediction intervals have asymptotic correct coverage when the data has a dependence structure. The main drawback of the algorithm Chernozhukov et al. (2018) study is its computation cost. One must retrain the forecasting model over a grid[7] of conjectured outcomes whenever a new observation arrives. Xu and Xie (2022) show that one can limit the computational cost by pretraining the forecasting models at the expense of underlying shocks being stationary, which again limits the dependence of the data. In contrast, the subsampling procedure relaxes all those assumptions and, as the algorithm suggests, its computational costs are limited by the number of subsamples.

[7] Deciding the size of the grid and its coarseness may also be problematic.

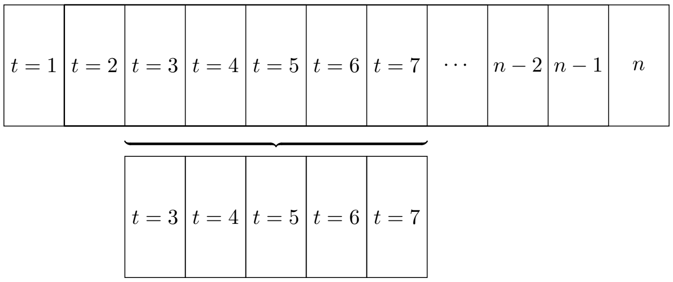
To illustrate the subsampling algorithm, say there is a sample of size n of the series, $\{y_t\}_{t=1}^n$, and the size of the subsample is $b = 5$ periods. Graphically, the first subsample is


$t = 1$ | $t = 2$ | $t = 3$ | $t = 4$ | $t = 5$ | $t = 6$ | $t = 7$ | $\cdots$ | $n - 2$ | $n - 1$ | n
$t = 1$ | $t = 2$ | $t = 3$ | $t = 4$ | $t = 5$

The second,


$t = 1$ | $t = 2$ | $t = 3$ | $t = 4$ | $t = 5$ | $t = 6$ | $t = 7$ | $\cdots$ | $n - 2$ | $n - 1$ | n
$t = 2$ | $t = 3$ | $t = 4$ | $t = 5$ | $t = 6$

The third,


$t = 1$ | $t = 2$ | $t = 3$ | $t = 4$ | $t = 5$ | $t = 6$ | $t = 7$ | $\cdots$ | $n - 2$ | $n - 1$ | n
$t = 3$ | $t = 4$ | $t = 5$ | $t = 6$ | $t = 7$

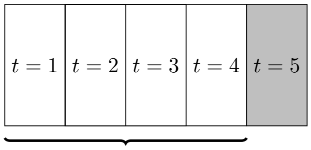
and so on. In terms of notation, denote the subsample of size b starting in period t as $\{y_t, y_{t+1}, \ldots, y_{t+b-1}\}$. Steps 2 and 3 consist of a pseudo-out-of-subsample forecast. By omitting the last observation in each subsample (y_{t+b-1}) and using the remaining ones ($\{y_t, \ldots, y_{t+b-2}\}$) to forecast it, the subsampled prediction error can be calculated as

$$e_{1|t,b} = y_{t+b-1} - \hat{y}_{t+b-1}$$

where $\hat{y}_{t+b-1}$ is the optimal operational forecast constructed using the same method as the original forecast except it uses only the information in the subsample. Mathematically, it is the solution to

$$\hat{y}_{h|t,b} \equiv \underset{\hat{y} \in Y^*}{\arg\min} (b-1)^{-1} \sum_{\tau=t}^{t+b-2} \mathcal{L}(e_\tau) \tag{7}$$

In that graphical example, omit the $t = 5$ observation in the first subsample (in grey), then use observations $\{y_1, y_2, y_3, y_4\}$ to construct a forecast for y_5,



The subsampled prediction error is

$$e_{1|1,5} = y_5 - \hat{y}_5$$

Once this is done for every subsample, one can construct the empirical unconditional and conditional cdfs of the predictive roots,

$$L_{1,n,b}^{unconditional}(e^*) \equiv \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} \mathbf{1}[e_{1|t,b} \leq e^*]$$

$$L_{1,n,b}^{conditional}(e^*, y^*) \equiv \frac{\sum_{t=1}^{n-b+1} \mathbf{1}[e_{1|t,b} \leq e^*] K[h^{-1}(y_{t-b+1} - y^*)]}{\sum_{t=1}^{n-b+1} K[h^{-1}(y_{t-b+1} - y^*)]}$$

where $\mathbf{1}(S)$ equals one when a logical statement S is true and zero otherwise converges to the distribution of the forecasting errors asymptotically. Moreover[8], $K(\cdot)$ is a typical kernel function with $\int K(x)dx = 1$ and $\int xK(x)dx = 0$.

[8] Booth et al. (1992) also mention an alternative estimate for the conditional cdf by averaging the joint cdf over a small window $y^* - h \leq y_n \leq y^* + h$.

```
Algorithm 1: Construction of unconditional prediction intervals via subsampling
Data: A dataset $\{y_t\}$, a statistical model used for forecasting $\mathcal{Y}$, a loss function
      $\mathcal{L}(\cdot)$, and a coverage level $\gamma$.
Result: The interval $[\underline{y}, \bar{y}]$ defined in (3)
Create $n-b+1$ subsamples (rolling windows) of the data;
for each subsample do
    Obtain the forecast $\hat{y}_{1|t,b}$ based on all but the last observation in the subsample
     by solving the cost minimization problem (7);
    Calculate and store the subsampled predictive root, $e_{1|t,b} = y_{t+b-1} - \hat{y}_{1|t,b}$;
end
Construct the forecast for $y_{n+1}$ using the same forecasting method $\mathcal{Y}$;
Obtain $\hat{L}^{-1,unconditional}_{1,b,n}(x)$, the inverse empirical cdf of the centered subsampled
 prediction errors, $e_{1|t,b} - \frac{\sum_t e_{1|t,b}}{n-b+1}$;
Return the following prediction interval (PI):
    $$PI = \left[\hat{y}_{n+1} + L^{-1,unconditional}_{1,b,n}\left(\frac{1-\gamma}{2}\right); \hat{y}_{n+1} + L^{-1,unconditional}_{1,b,n}\left(1-\frac{\gamma}{2}\right)\right]$$
```

```
Algorithm 2: Construction of conditional prediction intervals via subsampling
Data: A dataset $\{y_t\}$, a statistical model used for forecasting $\mathcal{Y}$, a loss function
      $\mathcal{L}(\cdot)$, and a coverage level $\gamma$.
Result: The interval $[\underline{y}, \bar{y}]$ defined in (4)
Create $n-b+1$ subsamples (rolling windows) of the data;
for each subsample do
    Obtain the forecast $\hat{y}_{1|t,b}$ based on all but the last observation in the subsample
     by solving the cost minimization problem (7);
    Store the last value in the subsample $y_{t+b-1}$ ;
    Calculate and store the subsampled predictive root, $e_{1|t,b} = y_{t+b-1} - \hat{y}_{1|t,b}$;
end
Construct the forecast for $y_{n+1}$ using the same forecasting method $\mathcal{Y}$;
Obtain $\hat{L}^{-1,conditional}_{1,b,n}(e^*, y^*)$, the inverse empirical cdf of the centered subsampled
 prediction errors, $e_{1|t,b} - \frac{\sum_t e_{1|t,b}}{n-b+1}$, conditional on $\{y_{t+b-1}\}$;
Return the following prediction interval (PI):
    $$PI = \left[\hat{y}_{n+1} + L^{-1,conditional}_{1,b,n}\left(\frac{1-\gamma}{2}; y_n\right); \hat{y}_{n+1} + L^{-1,conditional}_{1,b,n}\left(1-\frac{\gamma}{2}; y_n\right)\right]$$
```

Since they are calculations with sample realizations, one should interpret them as one of the $100 \times \gamma\%$ prediction intervals constructed this way expected to contain the true realization of y_{n+1}, unconditionally or conditionally on y_n in each case.

3.2 Statistical analysis

Here, I enunciate the results that ensure the correct asymptotic coverage of the prediction intervals constructed in Algorithms 6 and 7. The basic requirement is the existence of the asymptotic distributions of the forecast errors, to which the empirical distributions resulting from the algorithms converge to. If this is the case, then those empirical distributions may be used to construct the asymptotically valid prediction intervals.

In what follows, I omit the dependence of the forecast errors on the forecast-generating model $\mathcal{Y}_\theta(y, X)$. Let

$$J_h(x) = \mathbb{P}[e_h \leq x]$$

be the true distribution of the forecast error in horizon h. Its empirical counterpart in the sense of (2) is

$$J^*_{h,n}(x) = \mathbb{P}[e_{h|n} \leq x]$$

with $e_{n+h|n} \equiv y_{n+h} - \hat{y}^*_{n+h|n}$.

Let the probability distribution of the subsampled predictive roots (based on a subsample of size b) be

$$J^*_{h,b,t,n}(x) = \mathbb{P}[e_{h|t,b,n} \leq x]$$

For the sake of exposition, I relegate all technical conditions related to the series $\{y_t\}$ and its underlying process $\{Z_t\}$ to Appendix B given they can be convoluted and explaining them is not in the scope of this subsection.

Validity of the subsampling procedure

Theorem 3.1. *Let $\{y_t\}$ be a near-epoch dependent series over an $\alpha-$mixing sequence $\{Z_t\}$ and such that the theorems and results in Section B hold. If*

i. The forecast error e_h has an asymptotic distribution, namely $J_h(x) = \mathbb{P}[y_{n+h} - \hat{y}^*_{n+h} \leq x]$. The non-subsampled predictive root $y_{n+h} - \hat{y}^*_{n+h|n}$ has an asymptotic *distribution* $J^*_{h,n}(x) \equiv \mathbb{P}[y_{n+h} - \hat{y}^*_{n+h|n} \leq x]$*. Moreover,* $J^*_{h,n}(x) \to J_h(x)$.

ii. For every continuity point of $J_h(x)$*, and for any sequences* n, b *with* $n \to \infty$*,* $b/n \to 0$*,*

$$\frac{1}{n-b+1} \sum_{t=1}^{n-b+1} J^*_{h,b,t,n}(x) \to J_h(x)$$

then, the subsampled empirical distribution of the predictive root

$$L_{h,b,n}(x) \equiv \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} \mathbf{1}[y_{t+h} - \hat{y}^*_{h|t,b,n} \leq x]$$

is such that

a. $L_{h,b,n}(x) \to J(x)$ *in probability*

b. $\sup_x |L_{h,b,n}(x) - J_h(x)| \to 0$ *in probability*

c. The prediction intervals constructed based on the subsampling distributions above have the correct coverage (asymptotically) for any level $\gamma \in (0,1)$*.*

That is, denoting

$$c_{h,b,n}(\gamma) = \inf\{x : L_{h,b,n}(x) \leq \gamma\}$$

and

$$c(\gamma) = \inf\{x : J(x) \leq \gamma\}$$

If $J(\cdot)$ *is continuous at* $c(\gamma)$*, then*

$$\mathbb{P}[e_h \leq c_{n,b,h}(\gamma)] \to \gamma$$

as $n \to \infty$*.*

Proof. The proof hinges on the properties linking near-epoch dependent and mixingale processes. Then, once those relationships have between the near-epoch dependence of $\{y_t\}$ and the forecast model $\mathcal{Y}_{\hat{\theta}}$ have been established – the forecast error is a mixingale process, – the results in Gallant and White (1984) may be used to prove the convergence. More details can be found in Appendix B.

Theorem 3.2. *Under the same conditions of Theorem 3.1, the results for conditional prediction intervals are analogous.*

Proof. The proof is analogous to that of Theorem 3.1.

It is important to point out that Theorem 3.1 is similar to Theorem 4.2.1 in Politis, Romano, and Wolf (1999). However, there are important differences that make this result unique.

The first is that the objects of interest themselves are not the same. While, Politis, Romano, and Wolf (1999) are interested in characterizing the asymptotic distribution of a general statistical estimator. In such case, under standard regularity conditions for statistical inference, the statistic must be inflated by its convergence rate for the asymptotic distribution to exist[9]. This does not happen here as Theorems 3.1 and 3.2 regard the estimation of the asymptotic distributions (conditional and unconditional) of the forecast error. Because the distribution of the predictive roots combine both the error terms of the original process and estimation and misspecification errors, it does not collapse.

The second aspect is that there is no requirement on the rate of growth of the size of the subsample as long as $b/n \to 0$. This is a consequence of the model-free aspect of the construction of the prediction intervals. However, there are consequences to the choice of the growth of rate of b. As b grows, the uncertainty around the estimators for θ in the forecasting model disappear at the expense of robustness to structural changes. Moreover, in practice the size of the window also presents a choice between the quality of the estimation of forecast errors and the quality of the approximation of their distribution.

3.3 Comments on practical implementation

One of the difficulties in applying any subsampling procedure is deciding the size of each subsample (width of the rolling window) b. Theorem 3.1 is silent about the choice of b in finite samples. Politis, Romano, and Wolf (1999) propose different ways to choose b in the context of the $\alpha-$mixing series and when the goal is inference on parameters. Any of those procedures, however, are not applicable in the case considered here.

[9] The reader will notice that their Theorem 4.2.1 depends on the consistence rates of the estimator analysed with respect to the size of the subsample.

The choice of b aims on striking a balance between conflicting objectives. While larger subsamples include more information in the construction of the forecast, it comes at the cost of both a poorer approximation of the distribution of the predictive roots L since there are fewer blocks and at the expense of robustness to structural changes.

One possibility to choose b is by cross-validation. First, split the sample into training and validation sets – which contains only the last observations. Varying choices of b, run Algorithms 1 and 2 on the training set and check their efficacy on the validation set by means of rejection rates. This is a simple way to select a subsample size that does not depend on any human judgment – as one would pick the b that brings the rejection rates of their tests the closest to the nominal level. On the other hand, the total size of the sample can be a restricting factor since part of it must be used in the validation step and cannot be used in the estimation.

In this case, I suggest exploiting part (a) of Theorem 3.1 by picking two subsample sizes b, b' and running the method over both of them. This will yield two distributions of forecast errors (predictive roots). If those distributions are close enough according to some metric, then stop. If not, make $b \leftarrow b'$ and $b'' > b', b' \leftarrow b''$ and repeat. One can then conduct a Kolmogorov-Smirnoff test by calculating the distance $d(F_b(\cdot), F_{b'}(\cdot))$,

$$d(\hat{F}_b(\cdot), \hat{F}_{b'}(\cdot)) \equiv \sup_x |\hat{F}_b(x) - \hat{F}_{b'}(x)|$$

where $\hat{F}_b(x)$ is the empirical cdf of the predictive root constructed with subsample size b. Another option is to conduct an Anderson-Darling test. In principle, this should work better by starting with a low b and increasing it by one unit each iteration, stopping when the research is satisfied with the approximation. This can be computationally costly and it does not work well in simulations since the difference may be small for small changes in b and larger increases would point to differences between the distributions. All simulations below are conducted under this way to select b, starting with $b = \sqrt[2]{n}$ with $\sqrt[2]{n}$ increases.

3.4 Performance assessment

The goal of this section is to prove the performance of the subsampling method numerically. First, I study and compare unconditional and conditional prediction intervals. I move on to comparing the performance of subsampling with other methods. I will conduct a series of simulations in a toy nonlinear AR-GARCH model. This model features conditional heteroskedasticity – when the variance of the error terms changes depending on the state. The AR-GARCH is the leading example in this section since its wide applicability to financial series[10] and because it has the feature of not being α-mixing and yet being near-epoch dependent (Hansen, 1999).

The model is described by the following equations:

$$y_t = 0.1 y_{t-1} + \sigma_t \varepsilon_t \tag{8}$$

$$\sigma_t^2 = 0.05 + 0.5 * y_{t-1}^2 + 0.4 * \mathbf{1}[|y_{t-1}| > 2]$$

where $\mathbf{1}[\cdot]$ is the indicator function. σ_t is a state-dependent function that gives the model its conditional heteroskedasticity feature. σ_t is increasing in the absolute size of lagged y, continuously in the region $(-2, 2)$ and with an upwards shift outside of that interval. ε_t is a mean-zero variance-one i.i.d. shock. I change the distribution of ε_t in the experiments to study deviations from normality.

3.4.1 Unconditional and conditional intervals

Prediction intervals can be classified as unconditional or conditional depending on the information incorporated into them. The difference relies on studying the conditional distribution of future outcomes, given the information available, or their unconditional probability distribution.

This is particularly relevant in the context of model (8) presented earlier because the conditional variance of the error term is not constant. In what follows, I study the following specification:

$$\varepsilon_t \sim^{i.i.d.} \sqrt{\frac{1}{3}} t_3$$

$$\sigma_t^2 = 0.05 + 0.5 * y_{t-1}^2 + 0.4 * \mathbf{1}[|y_{t-1}| > 2]$$

This case features both heavy tails and conditional heteroskedasticity since the larger y is (in absolute value), the larger variance of the error terms are. Therefore, it is expected that the conditional prediction intervals will be wider as y increases. In practical examples

[10] Studying conditional heteroskedasticity is the basis of estimating the value at risk of a financial portfolio. This measure of risk guides investment decisions since it points out riskier portfolios. Other empirical examples are presented in Engle's (2001) review on applications of conditional heteroskedasticity models.

this can be thought an economy that is more volatile the hotter it runs or a stock that is the target of speculation (investors jump in and out of its market when the more extreme the price changes are).

Theorems 3.1 and 3.2 show theoretically that both subsampled conditional and unconditional prediction intervals have correct unconditional coverage asymptotically. This result is corroborated by Table 3.4.1, which presents simulated the rejection rates of both conditional and unconditional prediction intervals based on the model (8). On the other hand, conditional coverage – in this example, conditioning on the last value of the sample, – is not guaranteed in unconditional prediction intervals. In this case, their incorrect conditional coverage stems directly from the conditional heteroskedasticity of the series. Correct conditional coverage requires the intervals to be wider in the tails of the distribution of y_n than in the middle. To keep the width of the interval constant in the unconditional case, it is necessary to tolerate a higher rejection rate in the tails of the distribution of the predictor and allow for lower rejection around the center. Figure 3.4.1 makes this explicit. Panel (a) shows the contour curves of the conditional distribution of $\{y_n\}$ and the forecast error. This panel shows that the conditional distributions of the forecast errors vary over the distribution of y_n, being more or less concentrated depending on the values of the latter. Panel (b) presents a figure that has y_n on the x-axis, the rejection rates associated with the conditional and the unconditional prediction intervals on the left y-axis, and the density of y_n on the right y-axis.

3.4.2 Comparison with other methodologies

I now show how the subsampling method compares to other alternatives in constructing conditional prediction intervals. The comparison is interesting in misspecified cases since the robustness properties of subsampling imply loss of efficiency under correct specification. The leading example is the following conditional heteroskedastic model AR(1) model from before:

$$y_t = 0.1 y_{t-1} + \sigma_t \varepsilon_t$$

$$\sigma_t^2 = 0.05 + 0.5 * y_{t-1}^2 + 0.4 * \mathbf{1}[|y_{t-1}| > 2]$$

where $\mathbf{1}[\cdot]$ is the indicator function. σ_t is a state-dependent function that gives the model its conditional heteroskedasticity feature. σ_t is increasing in the absolute size of lagged y,

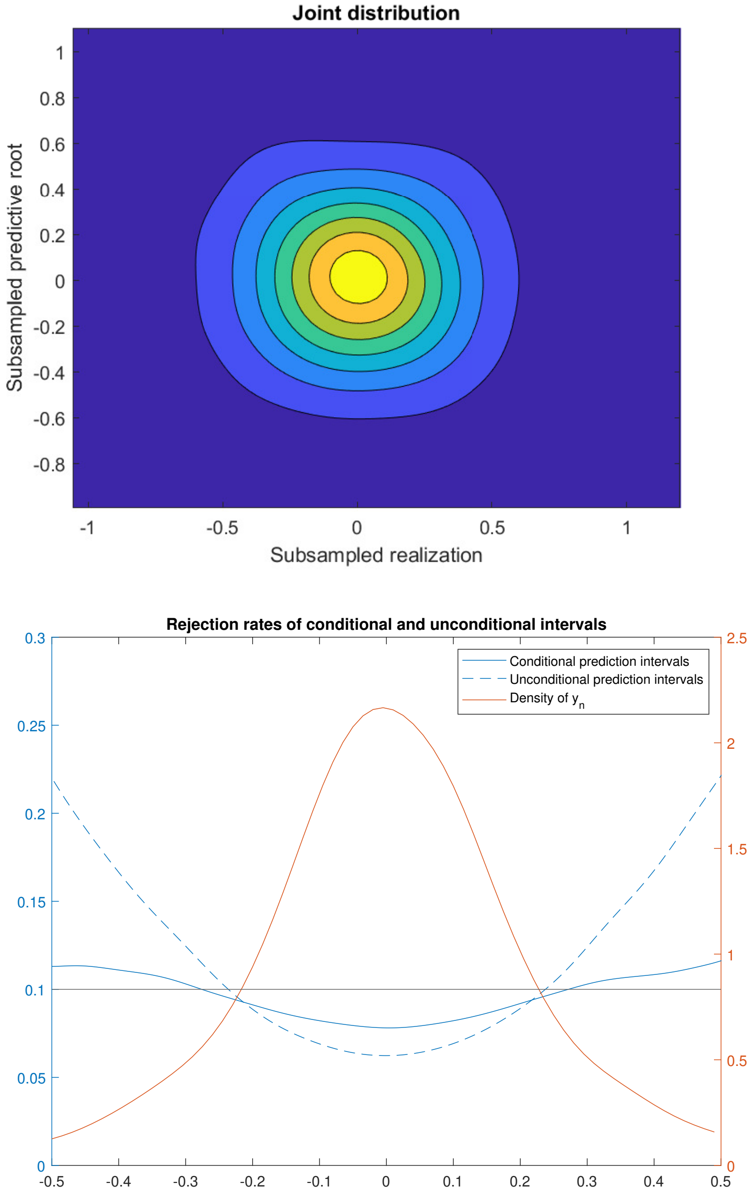


Figure 1: Panel (a, above): Joint distribution of y_n and forecast errors. Panel (b, below): Comparison of the rejection rates of both conditional and unconditional forecast intervals.

$1-\gamma = 10\%$	Unconditional	Conditional
500	9.29	8.64
1,000	10.20	9.54
5,000	10.51	9.40
$1-\gamma = 5\%$	Unconditional	Conditional
500	4.34	4.48
1,000	5.07	4.69
5,000	5.35	4.52
$1-\gamma = 2.5\%$	Unconditional	Conditional
500	2.01	2.53
1,000	2.55	2.79
5,000	2.73	2.42

Table 1: Unconditional rejection rates of both conditional and unconditional prediction intervals.

continuously in the region $(-2, 2)$ and with an upwards shift outside of that interval. ε_t is a mean-zero variance-one i.i.d. shock.

I will focus on five different approaches: the "naive" parametric approach, the predictive posterior from estimating the model with bayesian methods, the bootstrap approach presented in Cao (1997), and two conformal prediction inference methods (Chernozhukov et al's (2018) and Xu and Xie (2022)).

The "naive" approach consists in estimating the model with OLS and using the unconditional variance of the error terms to construct a prediction band around the forecast by adding and subtracting that standard error times a critical value z_α from the normal distribution. That is

$$PI_{naive}(\gamma) = \left[\hat{y}_{n+1|n} - z_{(1-\gamma)/2}\hat{\sigma}_{classic}; \quad \hat{y}_{n+1|n} + z_{(1-\gamma)/2}\hat{\sigma}_{classic}\right]$$

$$\hat{\sigma}^2_{classic} = \sum_{t=1}^{n} \hat{\varepsilon}^2_t / n$$

where $\hat{\sigma}^2_{classic}$ is the estimate of the standard error of the forecast errors, being $\hat{\varepsilon}^2_t$ the squared residual of the regression at time t. This method is only expected to have correct coverage under normality of the error terms and under homoskedasticity, granted these assumptions ensure that intervals of this form.

The bayesian approach consists of sampling from the posterior distribution of the predictions. The model is estimated with the normal-inverse-wishart conjugate priors, for the parameters $P = [\rho_0, \rho_1]$ and the variance of the error term, Σ.

$$\Sigma \sim Inverse\, Wishart(\Omega, \nu)$$

$$P|\Sigma \sim \mathcal{N}(\mu, \Sigma \otimes V)$$

where μ, V, ν, Ω are the hyperparameters of the prior distributions. Once the model is estimated, the credible interval is constructed based on the predictive distribution. This procedure generates incorrect prediction intervals when the model is misspecified – or it approximates poorly the distribution of the data.

Cao et al. (1997) propose a bootstrap procedure based on resampling the residuals of a linear AR forecasting model. It goes as follows

```
Algorithm 3: Cao et al.'s (1997) bootstrap prediction intervals
  Data: A dataset $\{y_t\}_{t=1}^{n}$, a block size $p$, a linear regression model $\mathcal{Y}_\theta$
  Result: The interval $[\underline{y}, \bar{y}]$ defined in (4)
  Estimate the linear regression model with OLS and obtain the residuals $\hat{\varepsilon}$ ;
  Rescale and normalize the residuals from the regression above. Denote them $\varepsilon^*$,
      $\varepsilon_t^* \equiv \sqrt{\frac{n-p}{n-2p}}\,[\hat{\varepsilon}_t - \bar{\hat{\varepsilon}}]$
   Resample $\varepsilon^*$ with replacement ;
  Set $y_n^* = y_n$ and generate artificial series $\{y_t^*\}$ by iterating on
      $y_t^* = \hat{\varphi}_0 + \hat{\varphi}_1 y_{t-1}^* + \varepsilon_t^*$
   Use the percentiles of the bootstrap distribution above to obtain $[\underline{y}, \bar{y}]$
```

This algorithm is proven to be consistent (in the sense it generates asymptotically valid prediction intervals) under the hypotheses that the model is stable, the error terms are i.i.d. with common distribution. Those assumptions are precisely those avoided by the subsampling procedure.

Finally, the two conformal prediction inference procedures. Chernozhukov et al.'s (2018) consists of a generalization of Vovk's extensive[11] work on conformal prediction inference for

[11] The literature on conformal inference dates back to the 1960s. A concise summary can be found in

independent data. The main idea of conformal prediction consists of extending the dataset to allow for a new, conjectured observation and measure how unlikely (according to some measure) it is compared to the dataset in hand[12]. The measure of unlikelyhood is called "nonconformity" in this literature and it is a function that returns a relevant measure of distance from the new observation to the data. I denote this function as $\mathcal{C}(\cdot)$.

Applying Algorithm 1 in Chernozhukov et al.'s (2018) to the leading example in this section yields the following steps:

```
Algorithm 4: Chernozhukov et al.'s (2018) conformal prediction intervals
Data: A dataset $\{y_t\}_{t=1}^{n}$, the linear forecasting model $\mathcal{M}$, a cost function $\mathcal{C}(\cdot)$, a
      set of conjectured values $\{y^{new}\}$, and a miscoverage level $1-\gamma$.
Result: The interval $[\underline{y}, \bar{y}]$ defined in (4)
for each conjectured value $y^{new} \in \{y^{new}\}$ do
    Extend the dataset $\{y_t\}$ with $y^{new}$;
    Estimate the AR(1) model;
    Calculate the residual $e_{y^{new}}$ associated with the conjectured observation and
      store $\mathcal{C}(e_{y^{new}})$;
end
Obtain $[\underline{y}, \bar{y}]$ by selecting the values in $\{y^{new}\}$ associated with the $1-\gamma^{th}$ and $\gamma^{th}$
  percentiles of $\{\mathcal{C}(e_{y^{new}})\}_{y^{new}}$.
```

In their paper, Chernozhukov et al. (2018) prove that this algorithm would yield asymptotically valid prediction intervals if the true data generating process is stable and if the distribution of the error terms is strongly mixing and stationary. The latter condition is violated in the leading example that features conditional heteroskedasticity.

The algorithm proposed in Xu and Xie (2021)[13] is a conformal inference procedure that exploits the stationarity of the data to construct the intervals in a cost-effective way. The main idea is similar to that of Algorithm 4 except that it involves first bootstrapping the data (with the block bootstrap), estimating the model based on those artificial series, and calculating the nonconformity based on the residuals of each of the estimated models.

Vovk's (2009) book, "Algorithmic Learning in a Random World."

[12] For example, one has a sample of size n of independent draws from $\mathcal{N}(0,1)$ that has no outliers. Then, a new value of 100 is more unlikely (according to some measure) than a new value of 0.

[13] Later extended in their 2022 work.

This procedure is called “ensemble batch prediction intervals” due to the model averaging (EnbPi). Algorithm 5 presents the algorithm adapted to the leading example.

Algorithm 5: Simplified Ensemble batch prediction intervals (EnbPi), Xu and Xie (2021)

```
Data: A dataset $\{y_t\}_{t=1}^{n}$, a miscoverage level $1-\gamma$, the number of bootstrap
      replications $B$, and a set of conjectured observations $\{y^{new}\}$.
Result: An ensemble prediction interval $[\underline{y}, \bar{y}]$ as defined in 4
for each bootstrap replication $b$ do
    Resample, with replacement, a set of indices $S_b$ from $(1,\dots,n)$. Each index is
      labeled $b_\tau$. Name it $S_b$ ;
    Estimate the linear regression model
        $$y_t = \rho_0 + \rho_1 y_{t-1} + \varepsilon_t$$
      over the resampled data. For convenience, name it $\hat{f}^b(\{y_{b_\tau}\}_{\tau=1}^{n})$ ;
    Store $\{\hat{f}^b(\cdot), S_b\}$.
end
for each $i = 1,\dots,n$ do
    Compute $\hat{f}^{\phi}_{-\tau}(y_\tau) = \phi(\hat{f}^b(y_\tau), \tau \notin S_b)$ where $\phi(\cdot,\cdot)$ is the average of all forecasts
      computed with the models in the bootstrap step that do not contain
      observation $\tau$. ;
    Compute and store $\hat{\varepsilon}^{\phi}_i \equiv y_\tau - \hat{f}^{\phi}_{-\tau}(y_\tau)$ in a vector $\vec{\varepsilon}$.
end
Compute $\hat{f}^{\phi}_{n+1} = \phi(\hat{f}^b(y_n))$.;
Compute $\hat{\beta} = \arg\min_{\beta\in[0,1-\gamma]}[(\gamma+\beta)^{th}$ percentile of $\vec{\varepsilon}-(\beta)^{th}$ percentile of $\vec{\varepsilon}]$ (This
  last step finds the narrowest prediction interval.)
Return the following interval:
    $$I = \left[\hat{f}^{\phi}_{-(n+1)}(X_{n+1}) - \hat{F}^{-1}\left(\hat{\beta}\right); \hat{f}^{\phi}_{-(n+1)}(X_{n+1}) - \hat{F}^{-1}\left(\gamma+\hat{\beta}\right)\right]$$
  where $\hat{F}^{-1}(\cdot)$ is the empirical cdf of $\vec{\varepsilon}$.
```

EnbPi combines a variety of the techniques presented earlier. It starts with a bootstrap step, which serves as the basis for the conformal prediction step used to finalize the construction of the intervals. By combining both, the algorithm inherits both of their advantages (requiring only a few estimations in the bootstrap step and correct coverage from the conformal inference step). This comes at the expense of being vulnerable in two places. If the bootstrap does not work, then the first part is incorrect. If the conformal inference does not work, then second step is invalid.

To prove the relative performance of the method under misspecification, I will conduct Monte Carlo simulations of four different designs that explore violations of normality and/or the mixing requirements of all the methods previously described. By violating those assumptions, any of other methods ("naive," bayesian, bootstrap, and conformal prediction) are invalid and yield incorrect prediction intervals. The four scenarios are:

1. $\varepsilon_t \sim \mathcal{N}(0,1)$ and $\sigma_t = 1$
2. $\varepsilon_t \sim \mathcal{N}(0,1)$ and $\sigma_t^2 = 0.05 + 0.5 * y_{t-1}^2 + 0.4 * \mathbf{1}[|y_{t-1}| > 2]$
3. $\varepsilon_t \sim \sqrt{\frac{1}{3}} t_3$ and $\sigma_t = 1$
4. $\varepsilon_t \sim \sqrt{\frac{1}{3}} t_3$ and $\sigma_t^2 = 0.05 + 0.5 * y_{t-1}^2 + 0.4 * \mathbf{1}[|y_{t-1}| > 2]$

In each of these scenarios, I estimate a simple AR(1), which correctly describes the dependence of the process (in terms of conditional mean), relegating the misspecification to the volatility component. The estimations for each method are conducted according to the algorithms discussed earlier. I fix $\gamma = 90\%$ and set $n = 1,000$. Table 3.4.2 presents the rejection rate of each method when forecasting the one-step-ahead observation conditional on the last value in the sample. That is, the fraction of times when the true realization of the process was not contained in the intervals constructed.

The results present two lessons. First, the subsampling procedure is slightly conservative when the model is correctly specified, which is the price for paid for the unnecessary robustness in those cases. Second, my procedure shows its good properties compared to all others when faced with the experiments that involve conditional heteroskedasticity.

$\gamma = 90\%$	Experiment #1			
Quartile/Method	Subsampling	Cao et al. (1997)	Chernozhukov et al. (2018)	Xu and Xie (2021)
Bottom 2%	8.24	9.92	9.85	10.87
Q2	8.65	10.13	9.96	10.07
Q3	8.74	10.17	9.94	9.09
Top 2%	8.54	10.13	9.87	9.60

$\gamma = 90\%$	Experiment #2			
Quartile/Method	Subsampling	Cao et al. (1997)	Chernozhukov et al. (2018)	Xu and Xie (2021)
Bottom 2%	8.15	25.67	26.76	25.00
Q2	9.23	7.95	7.14	7.84
Q3	9.24	7.64	7.33	7.64
Top 2%	9.81	25.48	25.52	21.33

$\gamma = 90\%$	Experiment #3			
Quartile/Method	Subsampling	Cao et al. (1997)	Chernozhukov et al. (2018)	Xu and Xie (2021)
Bottom 2%	9.59	9.98	9.89	11.50
Q2	8.58	10.13	9.98	10.12
Q3	8.62	10.14	9.94	10.00
Top 2%	10.00	10.08	9.91	9.33

$\gamma = 90\%$	Experiment #4			
Quartile/Method	Subsampling	Cao et al. (1997)	Chernozhukov et al. (2018)	Xu and Xie (2021)
Bottom 2%	8.68	19.45	19.19	18.50
Q2	9.95	9.22	8.72	8.92
Q3	9.94	8.97	9.02	9.04
Top 2%	9.86	19.86	19.30	19.33

Table 2: Comparison of the coverage of conditional prediction intervals from different methods, $n = 1,000$

$\gamma = 90\%$	Experiment #1		
Quartile/Method	Subsampling	Naive	Bayesian
Bottom 2%	8.24	9.93	10.16
Q2	8.65	10.08	10.07
Q3	8.74	10.09	10.08
Top 2%	8.54	10.11	9.95
$\gamma = 90\%$	Experiment #2		
Quartile/Method	Subsampling	Naive	Bayesian
Bottom 2%	8.15	9.96	7.82
Q2	9.23	8.45	8.12
Q3	9.24	8.54	7.39
Top 2%	9.81	9.48	7.56
$\gamma = 90\%$	Experiment #3		
Quartile/Method	Subsampling	Naive	Bayesian
Bottom 2%	9.59	7.01	7.19
Q2	8.58	7.03	7.00
Q3	8.62	7.04	6.99
Top 2%	10.00	7.14	6.99
$\gamma = 90\%$	Experiment #4		
Quartile/Method	Subsampling	Naive	Bayesian
Bottom 2%	8.68	6.03	5.98
Q2	9.95	6.50	6.15
Q3	9.94	6.36	5.83
Top 2%	9.86	6.18	5.96

Table 3: Comparison of the coverage of conditional prediction intervals from different methods, $n = 1,000$

4 Illustration # 1: LASSO prediction intervals

I now illustrate the construction of prediction intervals using the subsampling procedure laid out in Section 3. I focus on the Least Absolute Shrinkage and Selection Operator (LASSO) regression, a popular way to forecast time series in high-dimensional settings. The LASSO is an estimation procedure for linear models that can handle large datasets by imposing that many regression coefficients are exactly zero, thus shrinking the dataset to relevant predictors and making it possible to estimate. Despite the mixed evidence on its predictive ability in a general context (Dalalyan et al., 2014), empirical evidence in economics suggests that linear regressions estimated with the LASSO method perform well in forecasting tasks compared to other forecasting models such as factor models, random forests, and different bagging procedures (Medeiros et al., 2019).

4.1 Data generation process and model

Let $\{y_t\}$ be some series of interest and $\{\mathbf{X}_t\}$ is a set of d_X covariates $\{X_{1,t}, X_{2,t}, \ldots, X_{d_X,t}\}$. The number of exogenous regressors, d_X, increases linearly with the sample size n according to $d_X = ceil(k_X \times n)$ where $ceil(\cdot)$ is the function that takes any number to the nearest largest integer. The data-generating is linear. All but the first three of the lags of y are irrelevant and all but the first three covariates in $\mathbf{X}$ are irrelevant. The dgp is parameterized as

$$y_t = 0.8y_{t-1} + 0.2y_{t-2} + 0.1y_{t-3} + 0.5X_{1,t-1} + 0.3X_{2,t-1} + \varepsilon_t \tag{9}$$

$$\varepsilon_t \sim^{i.i.d} \mathcal{N}(0,1) \tag{10}$$

$$X_{j,t} \sim^{i.i.d.} \mathcal{N}(0,1) \tag{11}$$

The low-dimensionality is unbeknownst to the econometrician, who estimates a high-dimensional model with an increasing number of lags of y included and all the exogenous regressors. The econometrician bets on sparsity by estimating this expanding linear model, mimicking Giannone et al.'s (2021) setup.

Let d_Y be the number of lags of y included in the estimation. d_Y increases linearly in the sample size according to $d_Y = ceil(k_Y \times n)$. The econometrician also includes all exogenous regressors in the estimation. The linear prediction model used for forecasting is

$$y_t = \rho_0 + \sum_{j=1}^{d_Y} \rho_j y_{t-j} + \sum_{p=1}^{d_Y} \varphi_p X_{p,t-1} + \varepsilon_t \tag{12}$$

where ρ_j and φ_p are the coefficients associated with the $j-th$ lag of y and the $p-th$ exogenous predictor $X_{p,t}$, respectively. ρ_0 is an intercept and ε_t is a stochastic term. The dimensions d_Y and d_X increase linearly in the sample size n, $d_Y = ceil(k_Y \times n)$ and $d_X = ceil(k_X \times n)$.

The estimation is done by using the LASSO. For the sake of notation, let the vector of parameters be denoted β, with $\beta \equiv [\rho_0, \rho_1, \ldots, \rho_{d_Y}, \varphi_1, \ldots, \varphi_p]'$. The estimation problem is

$$\hat{\beta} \equiv \arg\min_{\beta} \sum_{t=1}^{n} (y_t - \rho_0 + \sum_{j=1}^{d_Y} \rho_j y_{n-j} + \sum_{p=1}^{d_Y} \varphi_p X_{p,t-1})^2 + \lambda |\beta|_1 \tag{13}$$

where $|\cdot|_1$ is the 1-norm, and λ is a penalization (tuning) parameter. The solution to this problem, $\hat{\beta}$, is the estimate of β. The optimal operational one-step-ahead forecast according to this criterion is

$$\hat{y}^*_{n+1|n} = \hat{\rho}_0 + \sum_{j=1}^{d_Y} \hat{\rho}_j y_{n-j} + \sum_{p=1}^{d_Y} \hat{\varphi}_p X_{p,n-1}$$

Remark 4.1. The nonlinearity coming from the penalization term $\lambda|\beta|_1$ makes the sampling distribution of the coefficients $\hat{\beta}$ have mass at zero (Knight and Fu, 2000). In this case, the construction of prediction intervals by calculating the standard errors of the coefficients is not available.

■

4.2 Simulation and results

I now show numerically that the subsampling method provides accurate prediction intervals. The object of interest is the one-period-ahead realization of y, y_{n+1}. The forecast for y_{n+1} is constructed with the high-dimensional model (4.2.1) described earlier. Here, I present the results for different values of k_Y and $k_X = 1$. Under these conditions, this model is written as

$$y_t = \rho_0 + \sum_{j=1}^{ceil(k_Y n)} \rho_j y_{t-1-j} + \sum_{p=1}^{n} \varphi_p X_{p,t-1} + \varepsilon_t \tag{14}$$

Notice that not only the set of regressors $(1, \underbrace{y_{t-1}, y_{t-2}, \dots, y_{t-1-d_Y}}_{\text{Lags of } y}, \underbrace{X_{t,1}, \dots, X_{t,k_X}}_{\text{Other features}})'$ grows over time.

Each simulation is conducted in the following manner: first, I generate data according to (9); second, I estimate the LASSO regression described in (14) and do fifteen-fold cross-validation to select the tuning parameter that minimizes the mean squared error[14]; third, I construct unconditional prediction intervals according to Algorithm 1; finally, I compare those intervals with the actual realization of the process.

Tables 4.2 reports the rejection rates associated with the unconditional prediction intervals (relative number of times that the new observation is not contained in the prediction interval constructed in the third step). Those numbers are obtained based on $1,000$ simulations of different setups. I report numbers for different sample sizes $n \in \{200; 400\}$. I also vary the number of lags of y included in the estimation, $k_Y \in \{0.1, 0.2, 0.3\}$. In total, the number of regressors is $ceil(k_Y \times n) + n + 1$ including the intercept – larger than the sample size itself.

[14] In practice, the researcher should select the penalty parameter according to their procedure of choice. The choice here is arbitrary since it does not impact the properties of the subsampled prediction intervals as long as the same procedure is repeated across subsamples.

$\gamma = 80\%$	Rate of growth of dimensionality		
n	$k_Y = 10\%$	$k_Y = 20\%$	$k_Y = 30\%$
200	22.6	21.3	24.25
400	22.38	23.37	23.46

$\gamma = 90\%$	Rate of growth of dimensionality		
n	$k_Y = 10\%$	$k_Y = 20\%$	$k_Y = 30\%$
200	12.5	10.45	11.82
400	13.43	14.67	13.34

$\gamma = 95\%$	Rate of growth of dimensionality		
n	$k_Y = 10\%$	$k_Y = 20\%$	$k_Y = 30\%$
200	6.9	5.13	5.00
400	6.5	5.89	4.59

$\gamma = 97.5\%$	Rate of growth of dimensionality		
n	$k_Y = 10\%$	$k_Y = 20\%$	$k_Y = 30\%$
200	3.33	2.9	3.05
400	2.21	2.68	2.89

Table 4: Rejection rates of the LASSO unconditional prediction intervals, varying the number of lags in the estimation and the sample size.

4.2.1 Application: Inflation forecasting in a data-rich environment

Recently, the increased availability of macroeconomic data together with the developments in nonparametric estimation led to the exploration of modern machine learning methods in forecasting economic aggregates. One challenge faced in this venture is that there is a wide range of macroeconomic series which are relatively short. A variety of measurements of employment available, such as workers in agriculture, industry, and services; different measures of inflation, by region, by product, by level of aggregation; different measures of monetary aggregates; different measures of interest rates, such as interests on mortgages by their length; among many others.

Medeiros et al. (2019) are interested in forecasting inflation in this big-data setting. They compare different forecast models, among which are many machine-learning based, including shrinkage methods that allow for a considerable number of predictors. Their work shows the good performance of those nonparametric methods in a pseudo-out-of-sample exercise by comparing the root mean squared forecast errors of a many different estimation methods. I now revisit their exercise by means of the LASSO sparse linear regression (which performs well in Medeiros et al.'s (2019) work), but I also create prediction intervals – which they do not provide.

For completeness of this discussion, it is important to point out that any of those big-data methods relies on a "bet on sparsity" (Hastie et al. (2001)), the assumption that the underlying structure of the data is sparse (with many irrelevant covariates). Giannone et al.'s (2021) recent work on the study of sparsity in macroeconomic series by estimating a bayesian shrinkage model with a prior distribution on the degree of sparsity (number of predictors with coefficient exactly zero). Their estimations show that there is little evidence of a dense structure in macroeconomic data – which goes in the same line as the Medeiros et al.'s (2019) high-dimensional approaches.

Using the subsampling method and drawing on the LASSO procedure described earlier, I will estimate a linear model to forecast inflation. I use the monthly database of macroeconomic series FRED-MD (McCracken and Ng (2015)), publicly available on the FRED website. The dataset contains monthly information on prices, economic activity, stock markets, output and income, labor market, housing, consumption, inventories, and money and credit. The period considered in the estimation is goes from January 1959 up until June

2022 (totaling 740 observations). I present annualized values but do not do any sort of adjustment to the data. As for predictors, I include in the estimations are 8 lags of all variables available and AR terms for inflation (totaling 984 variables).

The estimation problem is

$$\hat{\beta}^{LASSO} \equiv argmin_{\beta} \sum_{t=9}^{n} (y_t - \mathbf{Z}_t \beta)^2 + \lambda |\beta|_1$$

where $\mathbf{Z}_t$ is a vector that contains all regressors associated with date t (the 984 variables and an intercept, notice that the first eight observations are dropped out because of the lags included). The forecast is:

$$\hat{y}^*_{n+1|n} = \mathbf{Z}_n \hat{\beta}^{LASSO}$$

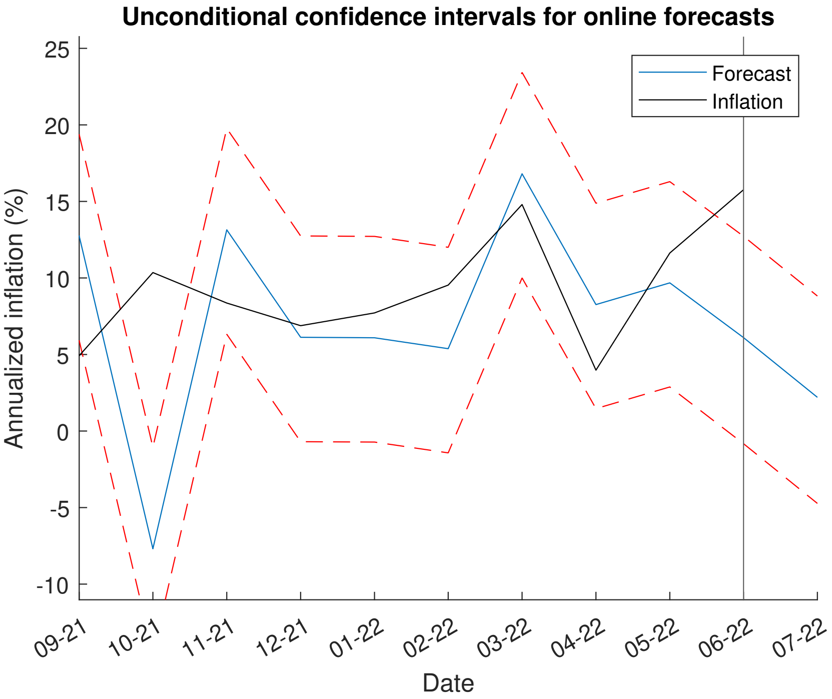


Figure 2: Prediction intervals for one quarter ahead inflation based on the LASSO linear regression.

Figure 4.2.1 presents the result of a sequential exercise. I start by omitting all observations prior to September 2021, then forecast this value and create the subsample prediction bands around it. Then, I reveal this observation and include it in the graph. I then repeat the same procedure for October, November, and so on. The last observation included in the total estimation is June 2022, which is used to forecast the value of inflation in July 2022. For the most part, the prediction intervals contain the true realization with only a couple of exceptions in this period. Here, the prediction intervals have about the same size even though the estimations are done sequentially. This is due to the fact that it is unlikely that one extra observation increases the size of the prediction intervals dramatically.

5 Illustration # 2: Asymmetric loss functions

The next example concerns asymmetric loss functions. This means that forecasting errors in one direction are worse for the decision maker than those in the other direction. For example, a company making inventory decisions must decide how many goods to have in stock. This is clearly an important decision which must be made in an uncertain environment due to demand fluctuations. The decision must balance the risk of losing sales due to the products not being in stock and transportation costs when ordering more units. When the concern is understocking, the loss function can be modeled as a asymmetric function. A common choice is Varian's (1975) linear-exponential (linex) cost function. The leading example in this Section is the study of the demand and the forecasting stage involved in planning the inventory of a company. In the application, I use real demand data for a tractor parts company that includes not only data on sales but also on a measure of lost sales due to lack of inventory.

First, let's set up the environment by defining the loss function. Let c_1 and c_2 be real numbers, the linex loss function is

$$\mathcal{L}(e) \equiv c_1 \exp\{c_2 \times e\} - c_2 \times e - 1 \tag{15}$$

where e is a forecasting error. Parameters c_1 and c_2 determine the shape and the asymmetry of the cost function. In practice, these parameters are calibrated to measure the costs of losing sales and extra stocking costs. The optimal operational forecast solves the cost minimization problem

$$\hat{\theta} \equiv \arg\min_{\theta|\mathcal{Y}_\theta} (n-1)^{-1} \sum_{t=2}^{n} \mathcal{L}(y_t - \hat{y}_{t|n}) \tag{16}$$

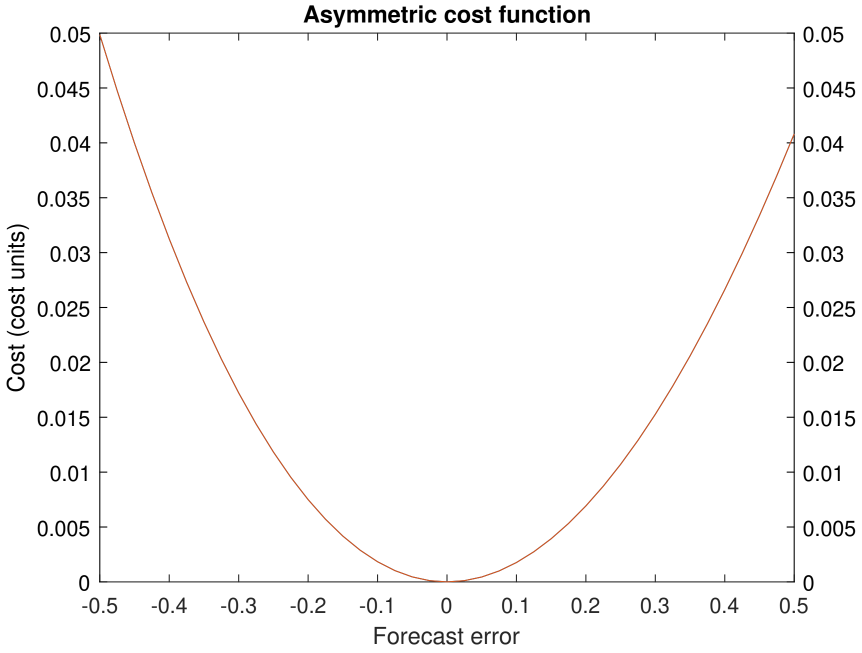


Figure 3: Linex cost function with $c_1 = 1$ and $c_2 = -0.6$

$$\hat{y}_{t|n} \equiv \mathcal{Y}_{\theta}(y_{t-1})$$
$$\hat{y}^*_{n+h|n} \equiv \mathcal{Y}_{\hat{\theta}}(y_n)$$

where the cost function $\mathcal{L}(\cdot)$ is the linex function in (15).

For the purposes of this example, let $c_1 = 1$ and $c_2 = -0.6$. The cost function is such that the cost increases faster when the forecast errors are negative – making understocking costlier than overstocking. The cost function with this parameterization has the following form:

5.1 Data generation process

Let $\{y_t\}$ be some series of interest generated according to the following time-varying process

$$\begin{cases} y_t = \rho_t y_{t-1} + \eta_t \\ \eta_t \sim^{i.i.d.} \mathcal{N}(0,1) \end{cases} \tag{17}$$

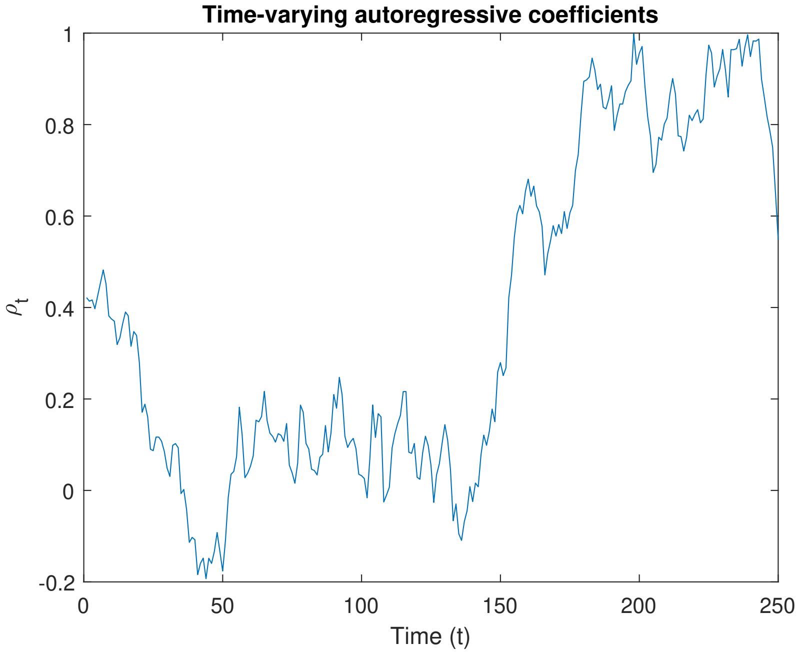


Figure 4: One realization of $\{\rho_t\}$, $T = 250$

where $\{\rho_t\}$ is a series of autoregressive parameters generated according to a "bounded random walk,"

$$\rho_t = \rho_{t-1} + \varepsilon_t$$

and ε_t follows a folded normal distribution (mean zero and variance 0.0025) dependent on ρ_{t-1} in a way that guarantees ρ_t to always be contained in $[-1, 1]$. Figure 5.1 presents a realization of the process of the autoregressive coefficients in the structural model.

5.2 Simulation and results

The simulation study here consists of constructing prediction intervals based on the asymmetric cost function and a data generating process that is both time-dependent. I will show that my method yields prediction intervals with approximate correct coverage in finite samples in this case when the econometrician also misspecifies the functional form.

The forecaster uses a simple AR(1) with constant coefficients.

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \nu_t$$

In each simulation study, I generate a sample from the data-generating process in (17). To construct the forecasts, I solve the cost-minimization problem in (16) via generalized method of moments, with the AR(1) above as $\mathcal{Y}_\theta(\cdot)$, with $\theta = (\varphi_0, \varphi_1)'$. Then, I use the plug-in version of the forecast. When there is aversion to understocking, the cost function is asymmetric and the resulting forecasts exhibit an upward bias.

To show the numerical performance in the context of asymmetric cost functions, model misspecification, and time-varying parameters, I design a following Monte Carlo experiment that highlights those features. In broad strokes, the procedure here is basically the one described in Section 4.2, with the exception that the forecasts are calculated based on the solution of the asymmetric problem defined in 16 and its plug-in estimator. An extra step is necessary to deal with the asymmetric nature of the problem. The process to construct such intervals requires first characterizing the distribution of the subsampled forecast errors, then using their percentiles together with the cost function to obtain the bounds of the intervals.

Moreover, the construction of conditional prediction intervals in this case presents an additional challenge because it involves analysing the conditional distribution. I deal with this by first estimating the conditional distribution of the subsampled predictive roots as in Algorithm 2 and then sample from that distribution by numerically integrating that estimate and using its inverse together with an uniformly distributed random variable[15].

Figure 5.2 presents some of the unconditional prediction intervals from this simulation. The main feature is the asymmetry implied by the linex costs, which end up being quite pronounced, even when the errors in the data generating process are gaussian. These prediction intervals result from the simulations with sample size $n = 750$.

Table 5.2 reports the rejection rates of the conditional and unconditional prediction intervals – the amount of times that the interval calculated did not contain the future observation. As one can see, the simulations show that the intervals constructed have (approximately) the correct coverage – as Theorems 3.1 and 3.2 suggest.

Algorithms 6 and 7 summarize the construction of both types of prediction intervals.

[15] Alternatively, one could also use an accept-reject algorithm to sample from it.

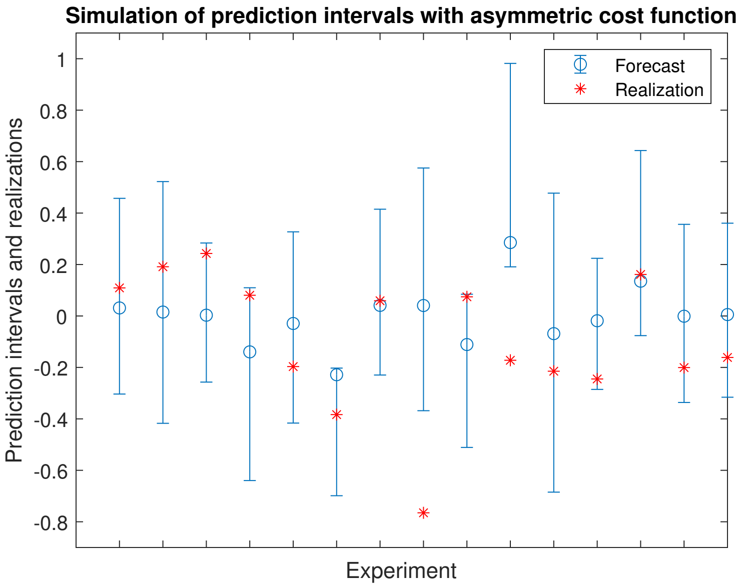


Figure 5: Simulated prediction intervals with the linex cost function

Unconditional prediction intervals				
Sample size	2.5%	5%	10%	20%
250	2.8	4.98	10.4	20.1
500	2.4	5.1	9.85	20.4
750	2.54	5.2	11.5	19.9
1000	2.51	4.9	9.54	19.8
Conditional prediction intervals				
Sample size	2.5%	5%	10%	20%
250	2.25	4.62	9.36	18.29
500	2.35	4.82	9.05	17.6
750	2.62	5.07	9.69	19.6
1000	2.45	4.77	9.81	18.9

Table 5: Rejection rates of conditional and unconditional prediction intervals under asymmetric loss

Algorithm 6: Unconditional prediction intervals for asymmetric errors, asymmetric cost function, and model misspecification (Monte Carlo)

```
Data: A dataset $\{y_t\}$, a statistical model used for forecasting $\mathcal{Y}$, a loss function
      $\mathcal{L}(\cdot)$, and a coverage level $\gamma$, the total number of Monte Carlo simulations $B$.
Result: The interval $[\underline{y}, \bar{y}]$ defined in (3)
for each $b = 1, 2, \ldots, B$ do
    Generate a sample of the process in (17);
    Create $n - b + 1$ subsamples of the data;
    for each subsample do
        Obtain the forecast $\hat{y}_{1|t,b}$ based on all but the last observation in the
          subsample by solving the cost minimization problem;
        Calculate and store the subsampled predictive root, $e_{1|t,b} = y_{t+b-1} - \hat{y}_{1|t,b}$;
    end
    Center the subsampled forecast errors around their average, define them as $\tilde{e}_{1|t,b}$
    $\tilde{e}_{1|t,b} \equiv e_{1|t,b} - \frac{\sum_t e_{1|t,b}}{n-b+1}$ ;
    Obtain the $100 \times \gamma\%$ and $100(1 - \gamma/2)\%$ percentiles of $\tilde{e}_{1|t,b}$;
    Obtain the smallest value $x$ such that $\mathcal{L}(x)$ is equal to the $100 \times \gamma\%$ percentile
      of $\tilde{e}_{1|t,b}$. Name this number $\underline{y}$ ;
    Obtain the largest value $x$ such that $\mathcal{L}(x)$ is equal to the $100(1 - \gamma/2)\%$
      percentile of $\tilde{e}_{1|t,b}$. Name this number $\bar{y}$ ;
    Return the following prediction interval (PI):

        $$PI = \left[\underline{y};\ \bar{y}\right]$$
end
```

Algorithm 7: Conditional prediction intervals for asymmetric errors, asymmetric cost function, and model misspecification (Monte Carlo)

```
Data: A dataset $\{y_t\}$, a statistical model used for forecasting $\mathcal{Y}$, a loss function $\mathcal{L}(\cdot)$, and a coverage level $\gamma$, the total number of Monte Carlo simulations $B$.
Result: The interval $[\underline{y}, \bar{y}]$ defined in (3)
for each $b = 1, 2, \dots, B$ do
    Generate a sample of the process in (17);
    Create $n - b + 1$ subsamples of the data;
    for each subsample do
        Obtain the forecast $\hat{y}_{1|t,b}$ based on all but the last observation in the subsample by solving the cost minimization problem;
        Calculate and store the subsampled predictive root, $e_{1|t,b} = y_{t+b-1} - \hat{y}_{1|t,b}$;
        Store the last value in the subsample $y_{t+b-1}$ ;
    end
    Center the subsampled forecast errors around their average, define them as $\tilde{e}_{1|t,b}$
    $\tilde{e}_{1|t,b} \equiv e_{1|t,b} - \frac{\sum_t e_{1|t,b}}{n-b+1}$ ;
    Obtain $\hat{L}^{-1,conditional}_{1,b,n}(e^*, y^*)$, the inverse empirical cdf of the centered subsampled prediction errors $\tilde{e}_{1|t,b}$, conditional on $\{y_{t+b-1}\}$;
    Sample from the conditional distribution (conditioning on the last value) with an adequate procedure then center the outcomes around their average. Denote them $\tilde{e}^*_{1|t,b}$.;
    Obtain the $100 \times \gamma\%$ and $100(1 - \gamma/2)\%$ percentiles of $\tilde{e}^*_{1|t,b}$;
    Obtain the smallest value $x$ such that $\mathcal{L}(x)$ is equal to the $100 \times \gamma\%$ percentile of $\tilde{e}^*_{1|t,b}$. Name this number $\underline{y}$ ;
    Obtain the largest value $x$ such that $\mathcal{L}(x)$ is equal to the $100(1 - \gamma/2)\%$ percentile of $\tilde{e}^*_{1|t,b}$. Name this number $\bar{y}$ ;
    Return the following prediction interval (PI):
        $PI = [\underline{y};\ \bar{y}]$
end
```

5.3 Application: Demand forecasting

I apply the method to forecast the demand for a tractor part. The data[16] comes from a tractor parts company focused on construction equipment. The data studied here is for a specific pin that holds together the teeth of an excavator – pins should be replaced as soon as they bend from daily use to prevent more serious damage to the tractor and are also a complimentary good with relation to other, more expensive, parts.

The data comes from a small branch of the company currently in expansion to a new market. Planning inventory for this branch is especially important because the shipping costs to that location are high. The data available contains the amount of the part sold, but the firm also collects information on lost sales due to lack of stock – quotes made by potential customers are recorded and contain information on the amount inquired and the reason why the order could not be fulfilled. The period of analysis extends from July 2010 to September 2022.

Figure 5.3 presents the monthly time series of total sales and total lost sales combined. It is important to point out some characteristics of the data. First, there is no graphical evidence that the demand is seasonal[17], because the climate conditions in the region are relatively constant year-round. There does not seem to be a trend, which may change in the future if the business grows. It is also important to point out the spike in 2022 – close to the electoral period.

Management points out that planning carefully the inventory of this store is very important to maintain its profitability. This is mostly due to the very high shipping costs to that location and the limited amount of cargo that trucks can carry. Internally, demand forecast is made with the Croston (1972) model, a choice justified by the relatively intermittent nature of the demand.

The Croston (1972) model is appealing due to its simplicity – it is ultimately a exponential smoothing procedure that allows for periods with zero demand. The model in its traditional form has only one parameter α that describes the degree of the smoothing. Applying this method consists in creating two artificial series from the original data, one that measures the amount of periods without any demand and another modeling the amount of

[16] Due to privacy reasons, this data is not publicly available.

[17] The managers of the store also do not have such perception.

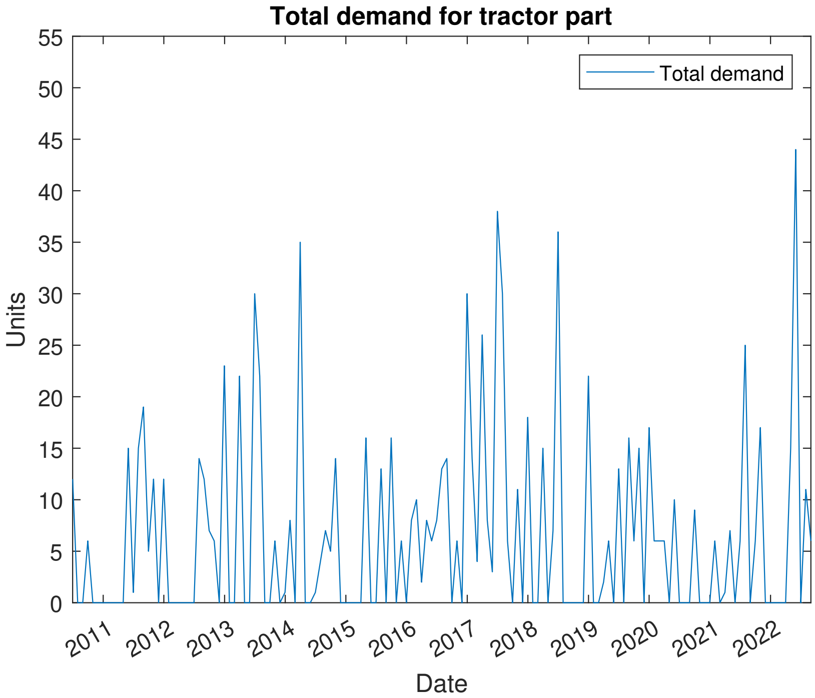


Figure 6: Time series of the total demand for a tractor from July 2010 to September 2022

quantities demanded, when there is any. Let[18] q_i be the i-th nonzero demand and a_i be the duration of the last zero-sales period. The forecast of q_{i+1} and a_{i+1} conditional on the i-th observation are, respectively,

$$\begin{cases} \hat{q}_{i+1|i} = (1-\alpha)\hat{q}_{i|i-1} + \alpha q_i \\ \hat{a}_{i+1|i} = (1-\alpha)\hat{a}_{i|i-1} + \alpha a_i \end{cases}$$

The model is initialized at the initial values of the series. Let j be the last period where the demand is not zero. The final forecast for next period's demand is

$$\hat{y}_{n+1|n} = \frac{\hat{q}_{j+1|j}}{\hat{a}_{j+1|j}}$$

In terms of the mechanical procedure, it remains to choose the value of the smoothing parameter α. I do so by minimizing the linex cost function in each time period. The value of c_2 was set to -0.015 from a back-of-the-envelope estimate of the cost of losing a sale of

[18] I follow Hyndman and Athanasopoulos's (2022) notation here.

that specific part[19].

At the expense of its simplicity, there is no automatic way to create prediction intervals within this framework[20]. It is possible to circumvent this issue by explicitly modelling the process with a latent state space model, for example. However, the subsampling method I propose allows for the calculation of prediction intervals even with the Croston model since it is completely model-free. The manager can now apply Algorithm 7 to construct conditional prediction intervals based on this model. Figure 5.3 presents the results of this estimation by comparing the vintage forecasts with the actual realizations of the process (in an online setting).

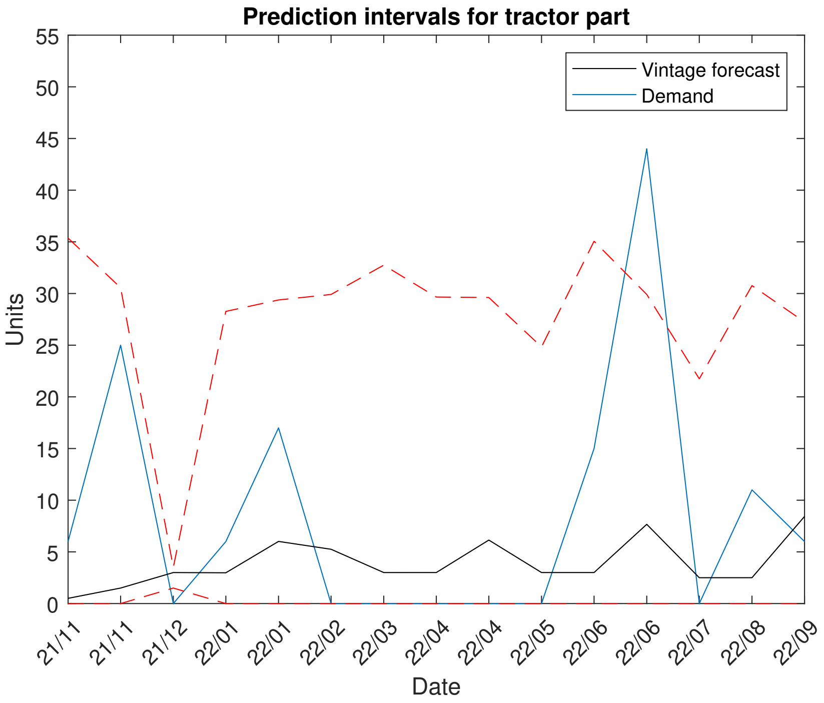


Figure 7: Conditional prediction intervals for the demand of a tractor part

[19] From an absolute near-sighted perspective. This number takes into account the losses of the goods that usually complement with this one. The longer-term cost related to customer trust and the company's image are not accounted for in this simple study but should be measured when making careful demand and forecasting studies.

[20] In the original paper, Croston argues that this model is derived from a gaussian structure. This turned out to be incorrect.

6 Testing hypotheses

I will now show how the subsampling method can be applied to hypotheses tests when it is hard to quantify uncertainty around the predictions.

6.1 Model selection

Gauging the prediction capabilities of a forecasting model or a learning method is important in practice when selecting a parsimonious and interpretable model. The goal of this section is to formalize a way to test different forecasting models, regardless of their functional forms, in a stochastic environment and choose among two alternatives[21]. Although the method is flexible regarding the choice of models, one case of interest is testing the significance of one subset of predictors in a given specification (the equivalent of an F-test in model selection in the linear regression context). In Hastie et al.'s (2016) systematic review of methods to assess the predictive quality of a model, the uncertainty aspect is dealt with heuristically with the "one-standard-deviation rule." The subsampling strategy provides a formal strategy to do so.

While similar to the bootstrap and cross-validation, subsampling provides a robust test for unconditional (and conditional) tests of predictive ability between two different models. For exposition purposes, we shall focus on the Giacomini and White (2006) test for predictive accuracy.

Take the problem of comparing the conditional predictive ability of two different forecasting models $\mathcal{Y}_m^*$, $m \in \{1, 2\}$ presented in Giacomini and White (2006). We are interested in predicting $\{y_t\}$ h steps ahead. Based on a cost function $\mathcal{L}(\cdot)$, we construct the optimal operational forecasts $\hat{y}^*_{n+h|n,m}$ from both prediction models using information up to date n. The hypothesis to be tested is that of equality of the expected loss of both models conditional on the information available

$$H_0 : \mathbb{E}[\mathcal{L}(y_{n+h} - \hat{y}^*_{n+h|n,1}) - \mathcal{L}(y_{n+h} - \hat{y}^*_{n+h|n,2})|\mathcal{F}_t] = 0$$

We will proceed to test this hypothesis with the subsampling procedure. It is important to point out two differences between the subsampling approach considered here and that

[21] This is generalizable with the methods developed in Silvapulle and Sen (2004).

of Giacomini and White (2006). First, they conduct this test with a fixed-length rolling window to prevent the uncertainty around the estimators to vanish[22]. On the other hand, the subsampling idea explicitly allows for expanding the sample size in the estimation, as long as it is sufficiently slow. The second aspect is that the class of temporal dependence allowed is significantly wider, from the discussion in Section 2.

I will now illustrate the use of the subsampling procedure by revisiting McCracken's (2020) counterexample to Giacomini and White (2006). The setup in question is the following: the data generating process of $\{y_t\}$ is

$$y_t = \mu + \varepsilon_t$$

where ε_t are i.i.d. with mean zero and variance σ^2. The econometrician has access to a sample of size $n + 1$ from this process and is interested in testing the "no-change in one-step-ahead point forecast" hypothesis of a model with only a constant and one that is pure noise. The loss function is quadratic and strategy is to use a fixed window of size R over which the average is calculated. Under these conditions, the test statistic is

$$\hat{d}_{t+1} = (y_{t+1} - 0)^2 - (y_{t+1} - \bar{y}_R)^2$$

where $\bar{y}_R$ is the sample average of y in the finite, fixed window considered. If $\mu = \sigma/\sqrt{R}$, the null hypothesis of equal predictive accuracy ($H_0 : \mathbb{E}(\hat{d}_{t+1}) = 0$) is always true.

McCracken (2020) considers the usual statistic to conclude this test of equal predictive accuracy

$$\Theta_n = \sqrt{n-R}\frac{\sum_{t=R+1}^{n} \hat{d}_{t+1}}{\sum_{t=R+1}^{n} \hat{d}_{t+1}^2}$$

which under the assumptions in Diebold and Mariano (1995), Giacomini and White (2005), and Diebold (2015) is asymptotically normally distributed. This, however, is not true when the window is fixed and finite because the statistic $\hat{d}_{t+1}$ exhibits long memory – in this case, the statistic Θ_n diverges with probability one. In the end, the statistic diverges because its convergence rate is not $\sqrt{n-R}$. In fact, further manipulating McCracken's (2020) equation

[22] It is also brought up in their paper that some exponentially fast weighting scheme will also render their results valid.

shows that the statistic Θ_n^*

$$\Theta_n^* = \frac{\sum_{t=R+1}^{n} \hat{d}_{t+1}}{\sum_{t=R+1}^{n} \hat{d}_{t+1}^2} \rightarrow^d \mathcal{D}$$

where $\mathcal{D}$ is some probability distribution.

By construction $\hat{d}_{t+1}$ is not $\alpha-$mixing because two $\hat{d}_{\tau_1}$ and $\hat{d}_{\tau_2}$ will always share (at least) the $\bar{y}_R$ term above no matter how large $|\tau_1 - \tau_2|$ is. This violates Giacomini and White's (2005) and Diebold and Mariano's (1995) assumptions. On the other hand, $\{d_{t+1}\}$ is NED over an $\alpha-$mixing process (since ε_t are i.i.d.), making it a candidate for the methods developed in this work.

By "naive" method I mean that one calculates Θ_n^* and compares it to a standard normal distribution hoping it provides an accurate approximation of $\mathcal{D}$. The "bootstrap" strategy consists of resampling with replacement from the data (for each bootstrap repetition, resample $n + 1$ observations, then collect R of them in a window and calculate Θ_n^* for that bootstrap repetition). Finally, the "subsampling" approach is just the one presented in Algorithm **??**. In this case, the strategy is to resample $b + R$ observations without replacement for each simulation, collect the first R observations as the fixed window and calculate Θ_b^*, the subsampled version of Θ_n^*.

Figure 8 shows a typical realization of the distribution of Θ_n^*, a standard normal distribution, and the distributions obtained from the bootstrap and subsampling procedures. As expected, the normal distribution is not a good approximation for the true distribution of Θ_n^*. Even though the bootstrap distribution gets relatively close to the distribution of Θ_n^*, the approximation is still poor because of the long memory of the series. Finally, as expected from Section **??**, the subsampling procedure yields accurate results in simulation. This proves its efficacy in this case.

6.2 Implication for i.i.d. data

McCracken's (2020) counterexample is presented in the context of time series. However, it has direct implications to quantifying uncertainty when the data is i.i.d., the usual case studied in much of the machine learning literature.

As the example shows, one must be careful when conducting cross-validation procedures to compare different models. Simple, off-the-shelf procedures are not usually correct due to

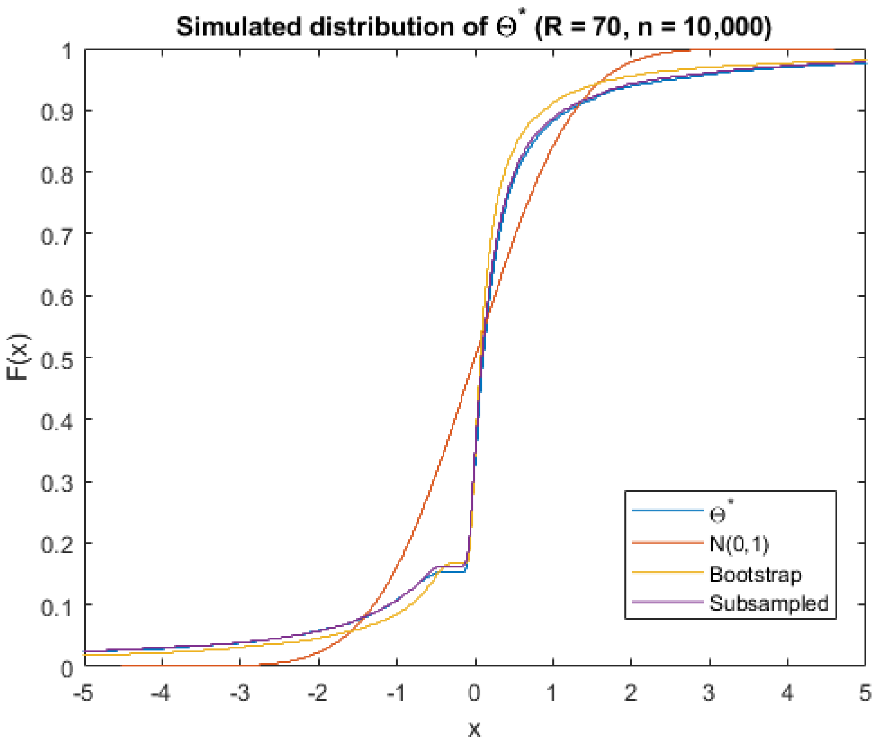


Figure 8: Typical simulation of McCracken's (2020) exercise with $n = 10,000$ and $R = 70$

the long memory induced, even when the data observed is i.i.d.. This issue, as shown above, is also not avoided by bootstrapping the data.

The correct procedure, as the evidence points, is to subsample in the lines described earlier: set the size of the training sample and then conduct what would be regular cross-validation if not for using only a subset of the remaining observations. Repeating this procedure multiple times estimates the true distribution of the cross-validation statistic, as would regular subsampling.

The result that provides technical grounds for this argument is enunciated in Theorem 6.1 below. The main takeaway is that even mild deviations from the traditional Berry-Esseen assumptions may slow down the convergence rate of any symmetric statistic to a rate slower than $\sqrt{n}$.

That is an important point in favor of subsampling. As Politis, Romano, and Wolf (1999) discuss, subsampling is robust to slow and potentially unknown convergence rates of the estimators involved. This is exactly the case covered by Theorem 6.1. Moreover, this is particularly relevant in the nonparametric context I study in this work because the convergence rate of those estimators can be slow and can be only characterized in some particular cases. Chen et al. (2019), for example, demonstrate that estimation using neural networks can achieve the rate of convergence $\sqrt[4]{n}$ under restrictive smoothness conditions. Meaning that subsampling is a viable candidate for a hypothesis testing procedure.

The convergence characteristics of the statistics used, however, are not the only issue. Another implication of Theorem 6.1 may be slower than $\sqrt{n}$ depending on the dependence characteristics of the series – the tension between moment finiteness and temporal dependence, however, remains.

Berry-Esseen bound for U-Statistics of $\alpha-$mixing random variables

Theorem 6.1. *Let $\{Z_t\}_{t=1}^{n}$ be a potentially nonstationary in the strong sense $\alpha-$mixing series with mixing coefficients $\alpha_Z(\tau)$. Let $\psi_{n,b}(Z_\tau, Z_{\tau+1}, \cdots, Z_{\tau+b-1})$ be some function of b sequential entries of $\{Z_t\}$, for $\tau \in \{1, 2, \cdots, n-b+1\}$, that inherits the mixing properties of $\{Z_t\}$. Moreover, $\mathbb{E}[\psi_{n,b}(\cdot)] = \Psi$. Assume that, for some $s \in (2, 3]$, $\sup_\tau \mathbb{E}[|Z_\tau|^s] < \infty$. If $\{Z_t\}$ and $\psi_{n,b}(\cdot)$ satisfy:*

a. For some $\delta > 0$,

$$\sum_{\tau=1}^{\infty}(\tau+1)^2\alpha_Z(\tau)^{\delta/(4+\delta)} < \infty$$

subject to the following restrictions:

$$(\alpha_Z(\tau+1))^{1/s}k^{3/2}4^k \leq 1$$

$$k \geq log(n)/(2\log(16))$$

$$2k\tau + 1 < n$$

b.

$$\mathbb{E}\left[\left|\sum_{\tau=1}^{n-b+1}\frac{\psi_{n,b}(Z_\tau, Z_{\tau+1}, \cdots, Z_{\tau+b-1})}{n-b+1} - \Psi\right|^{2+\delta}\right] < \infty$$

then

$$\sup_z |F_n(z) - \Phi(z)| \leq O(\sqrt{n}^{-1} + \varepsilon(s, \alpha_Z(\cdot)))$$

where $F_n(z) \equiv \mathbb{P}\left[\hat{\sigma}_\psi^{-1} \sum_{\tau=1}^{n-b+1} \frac{\psi_{n,b}(Z_\tau, Z_{\tau+1}, \cdots, Z_{\tau+b-1}) - \Psi}{n-b+1} \leq z\right]$ *with* $\hat{\sigma}_\psi^2$ *is the estimated variance of* $\psi_{n,b}(\cdot)$. $\varepsilon(s, \alpha_Z(\cdot))$ *is a (not strictly) positive quantity that depends on the moments of the series and its temporal dependence.*

Remark 6.1. One interpretation of s is a "worst-case scenario," the maximum finite moment of the whole series. The fact that s belongs to the interval $(2, 3]$ means that the series must have at least finite variance in order for this Central Limit Theorem to hold. The value of ε depends on s and on the mixing coefficients of $\{Z_t\}$. In summary, $F_n(z)$ converges faster to the normal distribution the larger s is and the faster the memory vanishes. The limiting case with i.i.d. data ($\alpha_Z(\tau) = 0, \forall \tau$) with finite third-moment ($s = 3$) goes back to a standard Berry-Esseen bound. The full relation between those quantities is given in the Appendix.

■

This is similar to the central limit theorem presented by White and Domowitz (1984). One important difference is that the variance of the sum S_n is allowed to grow in n, avoiding one of their restrictions on the nonstationarity of the series.

Example 6.1. *White and Domowitz (1984) require*

$$\mathbb{E}\left[n^{-1}\left(\sum_{t=a+1}^{n+a} Z_t\right)^2\right] \to V$$

uniformly in the time shift a *and* V *is a constant. This condition rules out cases where the variance of the series grows indefinitely and is dependent on the period that the window* $\{a+1, \ldots, a+n\}$ *begins. Take* $Z_t \sim^d \mathcal{N}(0, t)$ *with all* Z*'s mutually independent. This situation violates White and Domowitz's requirements since*

$$\mathbb{E}\left[n^{-1}\left(\sum_{t=a+1}^{n+a} Z_t\right)^2\right] = \frac{n+1+2a}{2}$$

However, it still satisfies the conditions for Theorem 6.1 and therefore

$$\sqrt{n}\frac{\bar{Z}_n - 0}{\hat{\sigma}_{Z,n}} \to \mathcal{N}(0, 1)$$

■

$\gamma = 0.30$	$R = 70$	
T	Subsampling	Bootstrap
500	0.2722	0.3905
$1,000$	0.2831	0.3747
$5,000$	0.3005	0.3641
$10,000$	0.2993	0.357
$\gamma = 0.20$	$R = 70$	
T	Subsampling	Bootstrap
500	0.186	0.2864
$1,000$	0.1964	0.2669
$5,000$	0.2025	0.2606
$10,000$	0.202	0.2586
$\gamma = 0.10\%$	$R = 70$	
T	Subsampling	Bootstrap
500	0.0939	0.1507
$1,000$	0.0991	0.1409
$5,000$	0.1024	0.138
$10,000$	0.1034	0.1333
$\gamma = 0.05$	$R = 70$	
T	Subsampling	Bootstrap
500	0.048	0.075
$1,000$	0.0507	0.0723
$5,000$	0.0512	0.0695
$10,000$	0.0503	0.069
$\gamma = 0.025$	$R = 70$	
T	Subsampling	Bootstrap
500	0.026	0.0393
$1,000$	0.0247	0.037
$5,000$	0.0258	0.035
$10,000$	0.026	0.0341

Table 6: Simulations of the rejection rate of McCracken's (2020) setup comparing subsampling and the bootstrap

The Monte Carlo simulation of this particular example yields the following cdf:

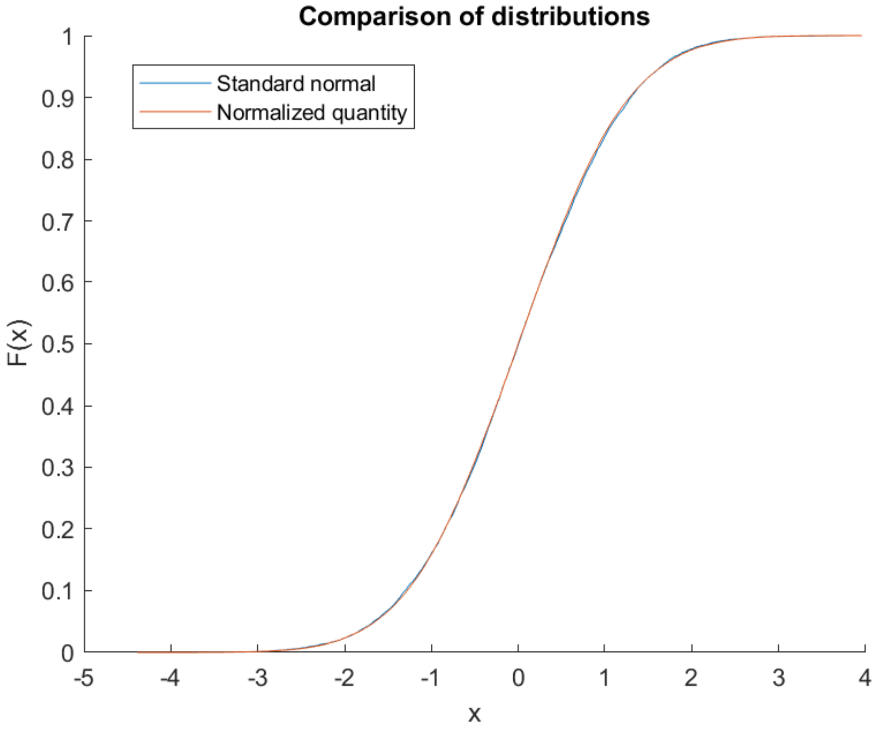


Figure 9: Distribution of the standardized quantity in the ever-growing variance case

7 Conclusion

In this paper, I propose a way to construct robust prediction intervals and test hypotheses involving predictions. I show their validity in a wide class of data-generating processes and models. The basic procedure consists of subsampling from the original series and constructing the empirical distribution of the statistics of interest.

Along with theoretical results, I show the numerical performance of the prediction intervals by Monte Carlo experiments. The simulations show the good coverage of both conditional and unconditional prediction intervals when the data violates standard mixing properties. I apply the method to two important forecasting problems: inflation and demand forecasting.

In the hypotheses tests aspect of the paper, I show the good performance of the subsampling method in a pathological case where both a traditional Central Limit Theorem and the traditional bootstrapping strategy fails.

Extensions of this paper include applying the subsampling method in other economic problems that involve studying i.i.d. data with machine learning models. An example is to test what set of predictors are relevant to forecast consumer default (as in Albanesi and Vamossy, 2019). On a more theoretical front, it is also possible to revisit Giacomini and White's (2006) test for comparing predictive accuracy of two models. The theory developed here allows for testing a mixingale condition analog to GW's martingale difference hypothesis. In practice, I interpret as a long-run equivalence between the forecasts but allowing for short-term deviations from asymptotic relation.

References

[1] Bengt von Bahr and Carl-Gustav Esseen. "Inequalities for the \$r\$th Absolute Moment of a Sum of Random Variables, \$1 \leqq r \leqq 2\$". en. In: *The Annals of Mathematical Statistics* 36.1 (Feb. 1965), pp. 299–303. ISSN: 0003-4851. DOI: 10.1214/aoms/1177700291. URL: http://projecteuclid.org/euclid.aoms/1177700291 (visited on 09/13/2022).

[2] I. A. Ibragimov, IU V. Linnik, and J. F. C. Kingman. *Independent and stationary sequences of random variables*. Groningen: Wolters-Noordhoff, 1971. ISBN: 978-90-01-41885-4.

[3] Herman Callaert and Paul Janssen. "The Berry-Esseen Theorem for U-Statistics". en. In: *The Annals of Statistics* 6.2 (1978), pp. 417–421. URL: http://www.jstor.org/stable/2958885.

[4] Andrew Patton and Allan Timmermann. "Properties of Optimal Forecasts". en. In: (1980), p. 54.

[5] Olaf Bunke and Bernd Droge. "Bootstrap and Cross-Validation Estimates of the Prediction Error for Linear Regression Models". In: *The Annals of Statistics* 12.4 (Dec. 1984). ISSN: 0090-5364. DOI: 10.1214/aos/1176346800. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-12/issue-4/Bootstrap-and-Cross-Validation-Estimates-of-the-Prediction-Error-for/10.1214/aos/1176346800.full (visited on 09/13/2022).

[6] Sangit Chatterjee. "Bootstrapping ARMA Models: Some Simulations". en. In: *IEEE Transactions on Systems, Man, and Cybernetics* 16.2 (1986), pp. 294–299. ISSN: 0018-9472. DOI: 10.1109/TSMC.1986.4308952. URL: http://ieeexplore.ieee.org/document/4308952/ (visited on 09/13/2022).

[7] Thomas Birkel. "On the Convergence Rate in the Central Limit Theorem for Associated Processes". en. In: *The Annals of Probability* 16.4 (Oct. 1988). ISSN: 0091-1798. DOI: 10.1214/aop/1176991591. URL: https://projecteuclid.org/journals/annals-of-probability/volume-16/issue-4/On-the-Convergence-Rate-in-the-Central-Limit-Theorem-for/10.1214/aop/1176991591.full (visited on 09/13/2022).

[8] Jeffrey M. Wooldridge and Halbert White. “Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes”. en. In: *Econometric Theory* 4.2 (Aug. 1988), pp. 210–230. ISSN: 0266-4666, 1469-4360. DOI: 10.1017/S0266466600012032. URL: https://www.cambridge.org/core/product/identifier/S0266466600012032/type/journal_article (visited on 09/13/2022).

[9] Peter Hackl and Anders Holger Westlund, eds. *Economic Structural Change*. en. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991. ISBN: 978-3-662-06826-7 978-3-662-06824-3. DOI: 10.1007/978-3-662-06824-3. URL: http://link.springer.com/10.1007/978-3-662-06824-3 (visited on 09/13/2022).

[10] Bruce E. Hansen. “GARCH(1, 1) processes are near epoch dependent”. en. In: *Economics Letters* 36.2 (June 1991), pp. 181–186. ISSN: 01651765. DOI: 10.1016/0165-1765(91)90186-O. URL: https://linkinghub.elsevier.com/retrieve/pii/016517659190186O (visited on 09/13/2022).

[11] Karl-Heinz Jöckel et al., eds. *Bootstrapping and Related Techniques*. en. Vol. 376. Lecture Notes in Economics and Mathematical Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992. ISBN: 978-3-540-55003-7 978-3-642-48850-4. DOI: 10.1007/978-3-642-48850-4. URL: http://link.springer.com/10.1007/978-3-642-48850-4 (visited on 09/13/2022).

[12] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. en. Boston, MA: Springer US, 1993. ISBN: 978-0-412-04231-7 978-1-4899-4541-9. DOI: 10.1007/978-1-4899-4541-9. URL: http://link.springer.com/10.1007/978-1-4899-4541-9 (visited on 09/13/2022).

[13] V. S. Koroljuk and Yu. V. Borovskich. *Theory of U-Statistics*. en. Dordrecht: Springer Netherlands, 1994. ISBN: 978-90-481-4346-7 978-94-017-3515-5. DOI: 10.1007/978-94-017-3515-5. URL: http://link.springer.com/10.1007/978-94-017-3515-5 (visited on 09/13/2022).

[14] Esther Ruiz. “Quasi-maximum likelihood estimation of stochastic volatility models”. en. In: *Journal of Econometrics* 63.1 (July 1994), pp. 289–306. ISSN: 03044076. DOI: 10.1016/0304-4076(93)01569-8. URL: https://linkinghub.elsevier.com/retrieve/pii/0304407693015698 (visited on 09/13/2022).

[15] F. Jay Breidt, Richard A. Davis, and William T. M. Dunsmuir. "IMPROVED BOOTSTRAP PREDICTION INTERVALS FOR AUTOREGRESSIONS". en. In: *Journal of Time Series Analysis* 16.2 (Mar. 1995), pp. 177–200. ISSN: 0143-9782, 1467-9892. DOI: 10.1111/j.1467-9892.1995.tb00229.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9892.1995.tb00229.x (visited on 09/13/2022).

[16] Peter J. Brockwell and Richard A. Davis. *Time series: theory and methods*. en. 2nd ed. Springer series in statistics. New York: Springer, 1996. ISBN: 978-0-387-97429-3 978-3-540-97429-1.

[17] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge ; New York, NY, USA: Cambridge University Press, 1997. ISBN: 978-0-521-57391-7 978-0-521-57471-6.

[18] Ricardo Cao. "An overview of bootstrap methods for estimating and predicting in time series". en. In: *Test* 8.1 (June 1999), pp. 95–116. ISSN: 1133-0686, 1863-8260. DOI: 10.1007/BF02595864. URL: http://link.springer.com/10.1007/BF02595864 (visited on 09/13/2022).

[19] Robert F. Engle and Aaron D. Smith. "Stochastic Permanent Breaks". en. In: *Review of Economics and Statistics* 81.4 (Nov. 1999), pp. 553–574. ISSN: 0034-6535, 1530-9142. DOI: 10.1162/003465399558382. URL: https://direct.mit.edu/rest/article/81/4/553-574/57160 (visited on 09/13/2022).

[20] James W. Taylor. "Evaluating volatility and interval forecasts". en. In: *Journal of Forecasting* 18.2 (Mar. 1999), pp. 111–128. ISSN: 0277-6693, 1099-131X. DOI: 10.1002/(SICI)1099-131X(199903)18:2⟨111::AID-FOR713⟩3.0.CO;2-C. URL: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1099-131X(199903)18:2%3C111::AID-FOR713%3E3.0.CO;2-C (visited on 09/13/2022).

[21] Keith Knight and Wenjiang Fu. "Asymptotics for Lasso-Type Estimators". In: *The Annals of Statistics* 28.5 (2000). Publisher: Institute of Mathematical Statistics, pp. 1356–1378. ISSN: 00905364. URL: http://www.jstor.org/stable/2674097 (visited on 10/05/2022).

[22] Simone Borra, ed. *Advances in data analysis and classification*. en. Studies in classification, data analysis, and knowledge organization. Berlin ; New York: Springer, 2001. ISBN: 978-3-540-41488-9.

[23] Graham Elliott and Allan Timmermann. “Optimal forecast combinations under general loss functions and forecast error distributions”. en. In: *Journal of Econometrics* 122.1 (Sept. 2004), pp. 47–79. ISSN: 03044076. DOI: 10.1016/j.jeconom.2003.10.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407603002690 (visited on 09/13/2022).

[24] Mervyn J. Silvapulle and Pranab Kumar Sen. *Constrained statistical inference: inequality, order, and shape restrictions*. Wiley series in probability and statistics. Hoboken, N.J: Wiley-Interscience, 2005. ISBN: 978-0-471-20827-3.

[25] John Staudenmayer and John P Buonaccorsi. “Measurement Error in Linear Autoregressive Models”. en. In: *Journal of the American Statistical Association* 100.471 (Sept. 2005), pp. 841–852. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214504000001871. URL: http://www.tandfonline.com/doi/abs/10.1198/016214504000001871 (visited on 09/13/2022).

[26] Graham Elliott, C. W. J. Granger, and Allan Timmermann, eds. *Handbook of economic forecasting*. en. 1st ed. Handbooks in economics 24. OCLC: ocm70063087. Amsterdam ; Boston: Elsevier North-Holland, 2006. ISBN: 978-0-444-62732-2 978-0-444-51395-3 978-0-444-53683-9 978-0-444-62731-5.

[27] Matei Demetrescu. “Optimal forecast intervals under asymmetric loss”. en. In: *Journal of Forecasting* 26.4 (July 2007), pp. 227–238. ISSN: 02776693, 1099131X. DOI: 10.1002/for.1019. URL: https://onlinelibrary.wiley.com/doi/10.1002/for.1019 (visited on 09/13/2022).

[28] Zudi Lu and Oliver Linton. “LOCAL LINEAR FITTING UNDER NEAR EPOCH DEPENDENCE”. en. In: *Econometric Theory* 23.01 (Feb. 2007). ISSN: 0266-4666, 1469-4360. DOI: 10.1017/S0266466607070028. URL: http://www.journals.cambridge.org/abstract_S0266466607070028 (visited on 09/13/2022).

[29] Andrew J. Patton and Allan Timmermann. “Properties of optimal forecasts under asymmetric loss and nonlinearity”. en. In: *Journal of Econometrics* 140.2 (Oct. 2007), pp. 884–918. ISSN: 03044076. DOI: 10.1016/j.jeconom.2006.07.018. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407606001606 (visited on 09/13/2022).

[30] Rafal Synowiecki. “Consistency and application of moving block bootstrap for non-stationary time series with periodic and almost periodic structure”. en. In: *Bernoulli* 13.4 (Nov. 2007). ISSN: 1350-7265. DOI: 10.3150/07-BEJ102. URL: https://projecteuclid.org/journals/bernoulli/volume-13/issue-4/Consistency-and-application-of-moving-block-bootstrap-for-non-stationary/10.3150/07-BEJ102.full (visited on 09/13/2022).

[31] Wei Biao Wu. “M-estimation of linear models with dependent errors”. en. In: *The Annals of Statistics* 35.2 (Apr. 2007). arXiv:math/0412268. ISSN: 0090-5364. DOI: 10.1214/009053606000001406. URL: http://arxiv.org/abs/math/0412268 (visited on 09/13/2022).

[32] Graham Elliott, Ivana Komunjer, and Allan Timmermann. “Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss?” en. In: *Journal of the European Economic Association* 6.1 (Mar. 2008), pp. 122–157. ISSN: 1542-4766, 1542-4774. DOI: 10.1162/JEEA.2008.6.1.122. URL: https://academic.oup.com/jeea/article-lookup/doi/10.1162/JEEA.2008.6.1.122 (visited on 09/13/2022).

[33] Peter Hall and Jiashun Jin. “Properties of higher criticism under strong dependence”. en. In: *The Annals of Statistics* 36.1 (Feb. 2008). ISSN: 0090-5364. DOI: 10.1214/009053607000000767. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-36/issue-1/Properties-of-higher-criticism-under-strong-dependence/10.1214/009053607000000767.full (visited on 09/13/2022).

[34] Joseph Engelberg, Charles F. Manski, and Jared Williams. “Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters”. en. In: *Journal of Business & Economic Statistics* 27.1 (Jan. 2009), pp. 30–41. ISSN: 0735-0015, 1537-2707. DOI: 10.1198/jbes.2009.0003. URL: http://www.tandfonline.com/doi/abs/10.1198/jbes.2009.0003 (visited on 09/13/2022).

[35] Arthur Berg, Timothy L. McMurry, and Dimitris N. Politis. “Subsampling -values”. en. In: *Statistics & Probability Letters* 80.17-18 (Sept. 2010), pp. 1358–1364. ISSN: 01677152. DOI: 10.1016/j.spl.2010.04.018. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167715210001276 (visited on 09/13/2022).

[36] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. en. Springer Series in Statistics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

isbn: 978-3-642-20191-2 978-3-642-20192-9. doi: 10.1007/978-3-642-20192-9. url: http://link.springer.com/10.1007/978-3-642-20192-9 (visited on 09/13/2022).

[37] Olimjon Sh Sharipov and Martin Wendler. *Bootstrap for the Sample Mean and for U-Statistics of Mixing and Near Epoch Dependent Processes*. en. Tech. rep. arXiv:0911.3083. arXiv:0911.3083 [math, stat] type: article. arXiv, July 2011. url: http://arxiv.org/abs/0911.3083 (visited on 09/13/2022).

[38] Charles F. Manski. "Interpreting and Combining Heterogeneous Survey Forecasts". en. In: *The Oxford Handbook of Economic Forecasting*. Ed. by Michael P. Clements and David F. Hendry. 1st ed. Oxford University Press, Sept. 2012, pp. 457–472. isbn: 978-0-19-539864-9 978-0-19-994032-5. doi: 10.1093/oxfordhb/9780195398649.013.0017. url: https://academic.oup.com/edited-volume/28323/chapter/215077919 (visited on 09/13/2022).

[39] Wen-Chi Kuo, Jessica Joy Vardy, and Bruce Alastair Watson. "Mixingales on Riesz spaces". en. In: *Journal of Mathematical Analysis and Applications* 402.2 (June 2013). arXiv:1707.01019 [math, stat], pp. 731–738. issn: 0022247X. doi: 10.1016/j.jmaa.2013.02.001. url: http://arxiv.org/abs/1707.01019 (visited on 09/13/2022).

[40] Thomas Lux and Leonardo Morales-Arias. "Relative forecasting performance of volatility models: Monte Carlo evidence". en. In: *Quantitative Finance* 13.9 (Sept. 2013), pp. 1375–1394. issn: 1469-7688, 1469-7696. doi: 10.1080/14697688.2013.795675. url: http://www.tandfonline.com/doi/abs/10.1080/14697688.2013.795675 (visited on 09/13/2022).

[41] Dimitris N. Politis. "Model-free model-fitting and predictive distributions". en. In: *TEST* 22.2 (June 2013), pp. 183–221. issn: 1133-0686, 1863-8260. doi: 10.1007/s11749-013-0317-7. url: http://link.springer.com/10.1007/s11749-013-0317-7 (visited on 09/13/2022).

[42] Qing Wang and Bruce G. Lindsay. "Variance estimation of a general u-statistic with appllication to cross-validation". en. In: *Statistica Sinica* (2014). issn: 10170405. doi: 10.5705/ss.2012.215. url: http://www3.stat.sinica.edu.tw/statistica/J24N3/J24N34/J24N34.html (visited on 09/13/2022).

[43] Qiying Wang and Nigel Chan. “Uniform convergence rates for a class of martingales with application in non-linear cointegrating regression”. en. In: *Bernoulli* 20.1 (Feb. 2014). arXiv:1402.0966 [math, stat]. ISSN: 1350-7265. DOI: 10.3150/12-BEJ482. URL: http://arxiv.org/abs/1402.0966 (visited on 09/13/2022).

[44] Francis X. Diebold. “Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests”. en. In: *Journal of Business & Economic Statistics* 33.1 (Jan. 2015), pp. 1–1. ISSN: 0735-0015, 1537-2707. DOI: 10.1080/07350015.2014.983236. URL: http://www.tandfonline.com/doi/abs/10.1080/07350015.2014.983236 (visited on 09/13/2022).

[45] Michael Wolf and Dan Wunderli. “Bootstrap Joint Prediction Regions: BOOTSTRAP JOINT PREDICTION REGIONS”. en. In: *Journal of Time Series Analysis* 36.3 (May 2015), pp. 352–376. ISSN: 01439782. DOI: 10.1111/jtsa.12099. URL: https://onlinelibrary.wiley.com/doi/10.1111/jtsa.12099 (visited on 09/13/2022).

[46] George E. P. Box et al. *Time series analysis: forecasting and control*. Fifth edition. Wiley series in probability and statistics. Hoboken, New Jersey: John Wiley & Sons, Inc, 2016. ISBN: 978-1-118-67502-1.

[47] Joshua C.C. Chan and Angelia L. Grant. “Modeling energy price dynamics: GARCH versus stochastic volatility”. en. In: *Energy Economics* 54 (Feb. 2016), pp. 182–189. ISSN: 01409883. DOI: 10.1016/j.eneco.2015.12.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140988315003539 (visited on 09/13/2022).

[48] Graham Elliott and Allan Timmermann. *Economic forecasting*. en. OCLC: ocn928115332. Princeton ; Oxford: Princeton University Press, 2016. ISBN: 978-0-691-14013-1.

[49] Graham Elliott and Allan Timmermann. *Economic forecasting*. OCLC: ocn928115332. Princeton ; Oxford: Princeton University Press, 2016. ISBN: 978-0-691-14013-1.

[50] Graham Elliott and Allan Timmermann. “Forecasting in Economics and Finance”. en. In: *Annual Review of Economics* 8.1 (Oct. 2016), pp. 81–110. ISSN: 1941-1383, 1941-1391. DOI: 10.1146/annurev-economics-080315-015346. URL: https://www.annualreviews.org/doi/10.1146/annurev-economics-080315-015346 (visited on 09/13/2022).

[51] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. en. Cambridge series in statistical and probabilistic mathematics. New York, NY: Cambridge University Press, 2016. ISBN: 978-1-107-04316-9.

[52] Li Pan and Dimitris N. Politis. “Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions”. en. In: *Journal of Statistical Planning and Inference* 177 (Oct. 2016), pp. 1–27. ISSN: 03783758. DOI: 10.1016/j.jspi.2014.10.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S037837581400175X (visited on 09/13/2022).

[53] Li Pan and Dimitris N. Politis. “Bootstrap prediction intervals for Markov processes”. en. In: *Computational Statistics & Data Analysis* 100 (Aug. 2016), pp. 467–494. ISSN: 01679473. DOI: 10.1016/j.csda.2015.05.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167947315001371 (visited on 09/13/2022).

[54] Atsushi Inoue, Lu Jin, and Barbara Rossi. “Rolling window selection for out-of-sample forecasting with time-varying parameters”. en. In: *Journal of Econometrics* 196.1 (Jan. 2017), pp. 55–67. ISSN: 03044076. DOI: 10.1016/j.jeconom.2016.03.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407616301713 (visited on 09/13/2022).

[55] Emmanuel Rio. *Asymptotic Theory of Weakly Dependent Random Processes*. en. Vol. 80. Probability Theory and Stochastic Modelling. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017. ISBN: 978-3-662-54322-1 978-3-662-54323-8. DOI: 10.1007/978-3-662-54323-8. URL: http://link.springer.com/10.1007/978-3-662-54323-8 (visited on 09/13/2022).

[56] Federico M. Bandi and Roberto Renò. “NONPARAMETRIC STOCHASTIC VOLATILITY”. en. In: *Econometric Theory* 34.6 (Dec. 2018), pp. 1207–1255. ISSN: 0266-4666, 1469-4360. DOI: 10.1017/S0266466617000457. URL: https://www.cambridge.org/core/product/identifier/S0266466617000457/type/journal_article (visited on 09/13/2022).

[57] Edgar Dobriban and Stefan Wager. “High-dimensional asymptotics of prediction: Ridge regression and classification”. en. In: *The Annals of Statistics* 46.1 (Feb. 2018). ISSN: 0090-5364. DOI: 10.1214/17-AOS1549. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-46/issue-1/High-dimensional-asymptotics-of-prediction-Ridge-regression-and-classification/10.1214/17-AOS1549.full (visited on 09/13/2022).

[58] Eric Ghysels and Massimiliano Marcellino. *Applied economic forecasting using time series methods*. en. New York: Oxford University Press, 2018. ISBN: 978-0-19-062201-5.

[59] Jing Lei et al. "Distribution-Free Predictive Inference for Regression". en. In: *Journal of the American Statistical Association* 113.523 (July 2018), pp. 1094–1111. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2017.1307116. URL: https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1307116 (visited on 10/05/2022).

[60] Stefania Albanesi and Domonkos F Vamossy. *Predicting Consumer Default: A Deep Learning Approach*. Working Paper 26165. National Bureau of Economic Research, Aug. 2019. DOI: 10.3386/w26165. URL: http://www.nber.org/papers/w26165.

[61] Stefania Albanesi and Domonkos F. Vamossy. *Predicting Consumer Default: A Deep Learning Approach*. en. Tech. rep. arXiv:1908.11498. arXiv:1908.11498 [cs, econ, q-fin] type: article. arXiv, Oct. 2019. URL: http://arxiv.org/abs/1908.11498 (visited on 09/13/2022).

[62] Gabriel Paes Herrera et al. "Long-term forecast of energy commodities price using machine learning". en. In: *Energy* 179 (July 2019), pp. 214–221. ISSN: 03605442. DOI: 10.1016/j.energy.2019.04.077. URL: https://linkinghub.elsevier.com/retrieve/pii/S036054421930708X (visited on 09/13/2022).

[63] Germán Aneiros et al., eds. *Functional and High-Dimensional Statistics and Related Fields*. en. Contributions to Statistics. Cham: Springer International Publishing, 2020. ISBN: 978-3-030-47755-4 978-3-030-47756-1. DOI: 10.1007/978-3-030-47756-1. URL: http://link.springer.com/10.1007/978-3-030-47756-1 (visited on 09/13/2022).

[64] Michael W. McCracken. "Diverging Tests of Equal Predictive Ability". en. In: *Econometrica* 88.4 (2020), pp. 1753–1754. ISSN: 0012-9682. DOI: 10.3982/ECTA17523. URL: https://www.econometricsociety.org/doi/10.3982/ECTA17523 (visited on 09/13/2022).

[65] James Davidson. "Near-Epoch Dependence". en. In: *Stochastic Limit Theory*. Oxford University Press, Nov. 2021, pp. 368–398. ISBN: 978-0-19-284450-7 978-0-19-192720-1. DOI: 10.1093/oso/9780192844507.003.0018. URL: https://academic.oup.com/book/42080/chapter/355980633 (visited on 09/13/2022).

[66] James Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. en. 2nd ed. Oxford University Press, Nov. 2021. ISBN: 978-0-19-284450-7 978-0-19-192720-1. DOI: 10.1093/oso/9780192844507.001.0001. URL: https://academic.oup.com/book/42080 (visited on 09/13/2022).

[67] Ricardo P. Masini, Marcelo C. Medeiros, and Eduardo F. Mendes. "Machine learning advances for time series forecasting". en. In: *Journal of Economic Surveys* (July 2021), joes.12429. ISSN: 0950-0804, 1467-6419. DOI: 10.1111/joes.12429. URL: https://onlinelibrary.wiley.com/doi/10.1111/joes.12429 (visited on 09/13/2022).

[68] Erindi Allaj and Simona Sanfelici. "Early Warning Systems for identifying financial instability". en. In: *International Journal of Forecasting* (Oct. 2022), S0169207022001133. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.08.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022001133 (visited on 11/11/2022).

[69] Karim Barigou et al. "Bayesian model averaging for mortality forecasting using leave-future-out validation". en. In: *International Journal of Forecasting* (Mar. 2022), S0169207022000243. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.01.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022000243 (visited on 11/11/2022).

[70] W Gonzalez-Manteiga and R Cao. "Predicting Using Box-Jenkins, Nonparametric, and Bootstrap Techniques". en. In: (2022), p. 9.

[71] Edward S. Knotek and Saeed Zaman. "Real-time density nowcasts of US inflation: A model combination approach". en. In: *International Journal of Forecasting* (Oct. 2022), S0169207022000589. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.04.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022000589 (visited on 11/11/2022).

[72] Johannes Lederer. *Fundamentals of High-Dimensional Statistics: With Exercises and R Labs*. en. Springer Texts in Statistics. Cham: Springer International Publishing, 2022. ISBN: 978-3-030-73791-7 978-3-030-73792-4. DOI: 10.1007/978-3-030-73792-4. URL: https://link.springer.com/10.1007/978-3-030-73792-4 (visited on 09/13/2022).

[73] E.L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. en. Springer Texts in Statistics. Cham: Springer International Publishing, 2022. ISBN: 978-3-030-70577-0 978-3-030-70578-7. DOI: 10.1007/978-3-030-70578-7. URL: https://link.springer.com/10.1007/978-3-030-70578-7 (visited on 09/13/2022).

[74] Han Li and Hua Chen. "Hierarchical mortality forecasting with EVT tails: An application to solvency capital requirement". en. In: *International Journal of Forecasting* (Sept. 2022), S0169207022001169. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.08.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022001169 (visited on 11/11/2022).

[75] Alfred Müller and Matthias Reuber. "A copula-based time series model for global horizontal irradiation". en. In: *International Journal of Forecasting* (June 2022), S0169207022000401. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.02.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022000401 (visited on 11/11/2022).

[76] Ioannis Nasios and Konstantinos Vogklis. "Blending gradient boosted trees and neural networks for point and probabilistic forecasting of hierarchical time series". en. In: *International Journal of Forecasting* 38.4 (Oct. 2022), pp. 1448–1459. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.01.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022000012 (visited on 11/11/2022).

[77] Kandrika F. Pritularga, Ivan Svetunkov, and Nikolaos Kourentzes. "Shrinkage estimator for exponential smoothing models". en. In: *International Journal of Forecasting* (Aug. 2022), S0169207022001030. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.07.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022001030 (visited on 11/11/2022).

[78] Olivier Sprangers, Sebastian Schelter, and Maarten de Rijke. "Parameter-efficient deep probabilistic forecasting". en. In: *International Journal of Forecasting* (Jan. 2022), S0169207021001850. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2021.11.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207021001850 (visited on 11/11/2022).

[79] Pengjie Wang, Athanasios A. Pantelous, and Farshid Vahid. "Multi-population mortality projection: The augmented common factor model with structural breaks". en. In: *International Journal of Forecasting* (Jan. 2022), S0169207021002144. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2021.12.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207021002144 (visited on 11/11/2022).

[80] Yiren Wang and Dimitris N. Politis. "Model-free bootstrap for a general class of stationary time series". en. In: *Bernoulli* 28.2 (May 2022). ISSN: 1350-7265. DOI: 10.3150/21-BEJ1352. URL: https://projecteuclid.org/journals/bernoulli/volume-28/

issue-2/Model-free-bootstrap-for-a-general-class-of-stationary-time/10.3150/21-BEJ1352.full (visited on 09/13/2022).

[81] Chen Xu and Yao Xie. *Conformal prediction for time series*. en. Tech. rep. arXiv:2010.09107. arXiv:2010.09107 [stat] type: article. arXiv, July 2022. URL: http://arxiv.org/abs/2010.09107 (visited on 09/13/2022).

[82] Chen Xu and Yao Xie. *Conformal prediction set for time-series*. en. Tech. rep. arXiv:2206.07851. arXiv:2206.07851 [cs, stat] type: article. arXiv, June 2022. URL: http://arxiv.org/abs/2206.07851 (visited on 09/13/2022).

[83] Yang Yang, Han Lin Shang, and James Raymer. "Forecasting Australian fertility by age, region, and birthplace". en. In: *International Journal of Forecasting* (Aug. 2022), S0169207022001108. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.08.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207022001108 (visited on 11/11/2022).

[84] Ibukun Amusan and Dr Ewald. "Parameter Estimation for a Stochastic Volatility Model with Additive and Multiplicative Noise". en. In: (), p. 25.

[85] Torben G Andersen and Luca Benzoni. "Stochastic Volatility;" en. In: (), p. 63.

[86] Joshua C C Chan. "Estimation of Stochastic Volatility Models with Heavy Tails and Serial Dependence". en. In: (), p. 21.

[87] Diebold Christoffersen. "Optimal Prediction Under Asymmetric Loss". en. In: (), p. 14.

[88] Russell Davidson and James G MacKinnon. "Bootstrap Methods in Econometrics". en. In: (), p. 28.

[89] Robert F Engle and Andrew J Patton. "What good is a volatility model?" en. In: *QUANTITATIVE FINANCE* (), p. 9.

[90] Tom Heskes. "Practical Confidence and Prediction Intervals". en. In: (), p. 7.

[91] Lutz Kilian. "Recent Developmentsin BootstrappingTime Series". en. In: (), p. 48.

[92] Mark M Meerschaert and Hans-Peter Scheffler. "LIMIT THEOREMS FOR CONTINUOUS-TIME RANDOM WALKS WITH INFINITE MEAN WAITING TIMES". en. In: (), p. 16.

[93] Oscar Nilsson. "On Stochastic Volatility Models as an Alternative to GARCH Type Models". en. In: (), p. 26.

[94] Michael Wolf and Dan Wunderli. “Bootstrap Joint Prediction Regions”. en. In: (), p. 39.

Appendices

A A list of papers that present prediction intervals or quantify uncertainty

1. Allaj and Sanfelici (2022): “Early Warning Systems for identifying financial instability.” - Parametric approach.

2. Knotek II and Zaman (2022): “Real-time density nowcasts of US inflation: A model combination approach.” - Unconditional empirical prediction intervals, requires stationarity.

3. Yang et al. (2022): “Forecasting Australian fertility by age, region, and birthplace.” - A bootstrap approach to prediction intervals, requires stationarity.

4. Li and Hue (2022): “Hierarchical mortality forecasting with EVT tails: An application to solvency capital requirement.” - Parametric approach.

5. Pritularga et al. (2022): “Shrinkage estimator for exponential smoothing models.” - Parametric and nonparametric unconditional prediction intervals.

6. Qi et al. (2022): “fETSmcs: Feature-based ETS model component selection.” - Parametric approach with exponential smoothing.

7. Li et al. (2022): “Bayesian forecast combination using time-varying features.” - Bayesian approach.

8. Sprangers el al. (2022): “Parameter-efficient deep probabilistic forecasting.” - Parametric approach.

9. Wang et al (2022): “Multi-population mortality projection: The augmented common factor model with structural breaks.” - Bayesian approach.

10. Nasios and Vogklis (2022): “Blending gradient boosted trees and neural networks for point and probabilistic forecasting of hierarchical time series.” - Unconditional intervals.

11. Barigou et al. (2022): "Bayesian model averaging for mortality forecasting using leave-future-out validation." - Bayesian approach.

12. Sbrana and Silvestrini (2022): "The RWDAR model: A novel state-space approach to forecasting." - Parametric approach.

13. Müller and Reuber (2022): "A copula-based time series model for global horizontal irradiation." - Impose Gumbel and BB1 copulas, which allow for prediction intervals.

14. Wang et al. (2022): "Distributed ARIMA models for ultra-long time series." - Parametric approach.

B Proof of Theorem 3.1

This proof exploits the definitions listed in Section 2. In particular, the assumption that the series of interest is near-epoch dependent on an underlying α-mixing properties ensures the "mixingality" of the forecast errors. Once that is shown, we can invoke classical results associated with this type of stochastic process to prove the convergence. In what follows, I will use X_t to denote general stochastic process. I will approach the proof constructively. Much of it follows the exposition in Davidson (2021) and Elliott and Timmermann (2016). First, let's start with the definition of a mixingale.

Definition .1. *(Mixingale)* Let $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be a sequence in a filtered probability space $(\Omega, \mathcal{F}, F, \mathbb{P})$. If $\{X_t\}$ are integrable, the sequence is called mixingale if there exist sequences of non-negative numbers $\{c_t\}_{t=-\infty}^{\infty}$ and $\{\xi_m\}_{m=0}^{\infty}$ ($\xi_m \to 0$ as $m \to \infty$) and

$$||\mathbb{E}[X_t|\mathcal{F}_{t-m}]||_2 \leq c_t\xi_m$$
$$||X_t - \mathbb{E}[X_t|\mathcal{F}_{t+m}]||_2 \leq c_t\xi_{m+1}$$

This definition is important because it ensures that many asymptotic results (Weak and Strong Laws of Large Numbers) hold – the reasons for this are left for the next subsection since the goal of this part is to explain the technical details. All the results in this paper will boil down to theorems and lemmas for mixingale processes and the reason are the two following theorems in Davidson (2021), which I enunciate in simplified forms[23] below.

[23] The complete proofs can be found in Chapter 18 of Davidson's book.

Lemma .1. *(Theorem 18.13 in Davidson (2021)) Let y_t be a near-epoch dependent process of size $-a$ on $\{Z_t\}$ with constants c_t. Suppose that $|\varphi_t(y_t)|$ is bounded and that $\varphi_t(\cdot)$ is Lipschitz-continuous for each t (Lipschitz constants are time-dependent). Then, $\{\phi_t(y_t)\}$ is also near-epoch dependent.*

Lemma .2. *(Theorem 18.6 in Davidson (2021)) Let $\{X_t\}_{-\infty}^{\infty}$ be an L_r-bounded sequence with $r = 2 + \delta$ on a near-epoch dependent process $\{Z_t\}$, with constants c_t. Then, if $\{Z_t\}$ is $\alpha-$mixing of size $-a$, $\{X_t - \mathbb{E}(X_t), \mathcal{F}_{-\infty}^{t}\}$ is a mixingale.*

The argument relies on results pertaining WLLNs for mixingales and some regularity conditions, which I enunciate below as Theorems[24].

Theorem .1. *(Theorem 20.17 in Davidson (2021)) Let $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be a mixingale with indices ν_m and d_t. If d_t is constant satisfying $\sup_t d_t < \infty$ and $d_t \geq \sqrt{\mathbb{E}[X_t^2]}$ and ν_m is such that $n^{-1}\sum_{m=1}^{n}\nu_m \to 0$, then*

$$n^{-1}\sum_{t=1}^{n} X_t \to^{L_2} 0$$

where "$\to^{L_2}$" denotes convergence in L_2 norm.

■

This requires that the series is stationary in a wide sense. Alternatively, one may relax the conditions in the previous theorem and combine the two following results:

Theorem .2. *(Theorem 21.18 in Davidson (2021)) Let $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be a mixingale of size $-1/2$ with respect to the sequence c_t and $\{a_t\}_{t=1}^{\infty}$ be a sequence of positive numbers such that*

$$\sum_{t=1}^{\infty} c_t^2/a_t^2 < \infty$$

Then, $a_n^{-1}\sum_{t=1}^{n} X_t \to^{a.s.} 0$.

■

Theorem .3. *(Theorem 19.5 in Davidson (2021)) Almost sure converge implies convergence in probability.*

■

[24] I do not provide proofs for them, but they can be found in Wooldridge (1994) and in Davidson (2021).

The final result is:

Theorem .4. *(Wooldridge, 1994) If either Theorem 20.17 or Theorem 21.15 in Davidson (2021) hold and*

a. $\theta \in \Theta \subset \mathbb{R}^k$*, where* k *is the dimension of* θ *and* Θ *is a compact set.*

b. $\mathcal{L}(\cdot)$ *and* $\mathcal{Y}_\theta(\cdot,\cdot)$ *are measurable in* θ *and continuous with respect to their arguments over the parameter set* Θ*.*

c. If $\mathcal{L}(\cdot)$ *is "Lipschitz continuous" in the following sense:*

$$|\mathcal{L}(y_{t+h} - \mathcal{Y}_{\theta_1}(y_t, X_t) - \mathcal{L}(y_{t+h} - \mathcal{Y}_{\theta_2}(y_t, X_t)| \leq c_t(y_{t+h}, y_t, X_t)||\theta_1 - \theta_2||$$
$$\forall \theta_1, \theta_2 \in \Theta$$

Then,

$$\max_{\theta \in \Theta} \left| (n-1)^{-1} \sum_{t=2}^{n} \mathcal{L}(y_t - \hat{y}_{t|n}) - (n-1)^{-1} \sum_{t=2}^{n} \mathbb{E}\left[\mathcal{L}(y_t - \hat{y}_{t|n})\right] \right| \to^p 0$$

That is, a uniform Weak Law of Large Numbers holds for the objective function in (2). ■

This is an important result. It shows that the average empirical cost converges in probability to its population counterpart. In practice, it means that the empirical solution comes close to the population solution as the sample size increases which is, ultimately, the true optimal forecast (as opposed to the operational optimal forecast). All the theoretical concepts in this Section are the constitute the basis of the results in Sections 3 and **??**.

T

Let $\{y_t\}$ be a near-epoch dependent over an underlying $\alpha-$mixing process. Then, if the assumptions for Theorem .4 hold,

$$\{y_{n+h} - \hat{y}^*_{n+h|n}; \mathcal{F}_t\} \quad \textit{and} \quad \{y_{n+h} - \hat{y}^*_{n+h|n}; \mathcal{F}_t, y^f\}$$

(for some conditioning value y^f) are mixingales.

$$\begin{aligned} \|y_{n+h} - \hat{y}^*_{n+h|n} - \mathbb{E}[y_{n+h} - \hat{y}^*_{n+h|n}]\|_p \leq & \|y_{n+h} - \mathbb{E}[y_{n+h}]\|_p + \\ & + \|\hat{y}^*_{n+h|n} - \mathbb{E}[\hat{y}^*_{n+h|n}]\|_p \end{aligned}$$

and

$$\begin{aligned}\|y_{n+h} - \hat{y}^*_{n+h|n} - \mathbb{E}[y_{n+h} - \hat{y}^*_{n+h|n}|y^f]\|_p \leq &\|y_{n+h} - \mathbb{E}[y_{n+h}|y^f]\|_p + \\ &+ \|\hat{y}^*_{n+h|n} - \mathbb{E}[\hat{y}^*_{n+h|n}|y^f]\|_p\end{aligned}$$

The first term is near-epoch dependent by assumption. The second term is mixing since it involves finite functions of $\{y_t\}$. This second part is hinges on the Assumptions of Theorem .4 since it requires the prediction function to exist. Then, apply Minkowki's and Jensen's inequalities to 1st term (analogous to Theorem 18.6 in Davidson (2021)).

These mixingale processes inherit the good properties of the "asymptotic independence" of the underlying process $\{Z_t\}$, and now we can show the remainder of the claim. Define

$$\begin{aligned}U_{n,h}(x) &\equiv \frac{1}{n-b+1}\sum_t \mathbf{1}[e_{h|n} \leq x] \\ \to \mathbb{E}[U_{n,h}(x)] &= (n-b+1)^{-1}\sum_t J^*_{h,t,n}(x)\end{aligned}$$

Now, to show that $Var(U_{n,h}(x)) \to 0$ as $n \to \infty$, we must first notice that it will involve the covariances of

$$I_{h,b,t} = \mathbf{I}[y_{t+h} - \hat{y}^* \leq x]$$

for which Laws of Large Numbers are immediately available in Gallant and White (1984). I omit the proof of the other results since they are the analogues of the proofs of Theorem 3.2.1 in Politis, Romano, and Wolf (1999).

C Proof of Theorem 6.1

This proof combines the techniques presented in Callert and Janssen (1978), Tikhomirov (1980), and Sunklodas (1984). For simplicity, I restrict it to the rank-2 U-Statistic $\psi(\cdot,\cdot)$ – the extension to a U-Statistic with rank-b kernel can be obtained by increasing the amount of terms and adjusting the definitions accordingly and is therefore omitted. The goal is to compute a bound for

$$\Delta_n \equiv \sup_z |F_n(z) - \Phi(z)|$$

Where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution and $F_n(Z)$ is the c.d.f. of the random variable $S_n = \hat{\sigma}_n^{-1}\hat{U}_n$, the Hajek projection of the U-Statistic U_n. The strategy now consists of working with the following quantity:

$$\hat{\sigma}_n^{-1}(U_n - \hat{U}_n) = \delta_n$$

which – with some abuse of notation – may be split into two parts, $\delta_j^{(\nu)}$ and $\varepsilon_j^{(\nu)}$, one of which

$$\delta_j^{(\nu)} \equiv \hat{\sigma}_n^{-1} \sum_{k_1, k_2:\ 0 \leq k_1 < k_2 \leq m\nu + j} [\psi(Z_{k_1}, Z_{k_2}) - \psi^{(1)}(Z_{k_1}) - \psi^{(1)}(Z_{k_2})]$$

Where m is a number that depends on the sample size[25] and the characteristics of the dependence of $\{Z_t\}_t$. The reason for this notation will become clear when manipulating the characteristic function of interest and it is homeomorphic to Callert and Janssen's (1978) approach to the problem. From here, the proof consists into using δ_n and $\delta_j^{(\nu)}$ and their characteristic functions to compute a Berry-Esseen bound using standard techniques. Using the definitions above, we can write

$$\sup_z \left| \mathbb{P}[\hat{\sigma}^{-1} U_n \leq z] - \Phi(z) \right| \leq \sup_z \left| \mathbb{P}[S_n + \delta_j^{(\nu)} \leq z] - \Phi(z) \right| + \mathbb{P}(\varepsilon_j^{(\nu)} \leq a_n) + O(a_n)$$

Some extra notation is needed for the rest of the proof. Define an order – the lexicographic order – between pairs of the form (k_1, k_2):

$$(k_1, k_2) \succcurlyeq (k_1', k_2') \ \ if \ \ k_1 > k_1' \ \ or \ \ both \ \ k_1 = k_1' \ \ and \ \ k_2 > k_2'$$

Let $\Omega(Q)$ be the following set:

$$\Omega(Q) \equiv \{k_1, k_2 : 1 \leq k_1 < k_2 \leq Q\}$$

Let $\omega_{(q)}$ be the q^{th} element of $\Omega(Q)$ according to the $\succcurlyeq$ order defined previously. It is also worth noticing that the number of elements in $\Omega(Q)$, which I denote $\#\Omega(Q)$, is

$$\#\Omega(Q) = \frac{Q(Q-1)}{2}$$

Finally, define the complementary set of $\Omega(Q)$ as $\Omega^c(Q) \cup \Omega(Q) \equiv \Omega(n)$ and $\Omega^c(Q) \cap \Omega(Q) = \emptyset$. To match with the notation presented earlier,

$$\delta_j^{(\nu)} = \hat{\sigma}_n^{-1} \sum_{q \in \{1, \cdots, \#\Omega(\nu+j)\} \ s.t. \ \omega_{(q)} \in \Omega(\nu+j)} [\psi(Z_{\omega_{(q),1}}, Z_{\omega_{(q),2}}) - \psi^{(1)}(Z_{\omega_{(q),1}}) - \psi^{(1)}(Z_{\omega_{(q),2}})]$$

$$\varepsilon_j^{(\nu)} = \hat{\sigma}_n^{-1} \sum_{q \in \{1, \cdots, \#\Omega^c(\nu+j)\} \ s.t. \ \omega_{(q)} \in \Omega^c(\nu+j)} [\psi(Z_{\omega_{(q),1}}, Z_{\omega_{(q),2}}) - \psi^{(1)}(Z_{\omega_{(q),1}}) - \psi^{(1)}(Z_{\omega_{(q),2}})]$$

[25] m gives a notion of distance between the $\sigma-$algebras induced by $\delta_j^{(\nu)}$ and $\varepsilon_j^{(\nu)}$. This is equivalent to Callert and Janssen's (1978) bound c_n.

We are now ready to apply the method presented in Tikhomirov (1980) and Sunklodas (1984). Define:

$$X_q = \psi(Z_{\omega_{(q),1}}, Z_{\omega_{(q),2}}) - \psi^{(1)}(Z_{\omega_{(q),1}}) - \psi^{(1)}(Z_{\omega_{(q),2}})$$

$$\xi_j^{(l)} = \exp\{it(\varepsilon_j^{(l-1)} - \varepsilon_j^{(l)})\}$$

$$A_q = \hat{\sigma}_n^{-1}[\psi(Z_{k_1}, Z_{k_2}) - \psi^{(1)}(Z_{k_1}) - \psi^{(1)}(Z_{k_2})],\ q = (k_1, k_2) \in \Omega(Q)$$

$$\eta_j^{(r)} = \exp\{-it\delta_j^{(r)}\} - 1$$

$$a_q^{(r-1)} = \mathbb{E}\left[A_q \prod_{l=1}^{r-1} \xi_j^l\right]$$

$$\varphi(t) = \mathbb{E}[\exp\left\{it\delta_j^{(\nu)}\right\}]$$

$$\lambda \equiv \sum_{\tau=1}^{\#\Omega} \alpha(\tau)^{(s-2)/s}$$

$$\aleph \equiv d^{1/s}/\sqrt{c_0}$$

$$\tilde{\sigma}_\delta^2 \equiv Var(\delta_n)/c_0$$

$$k \geq \log(\#\Omega)/(2\log(16))$$

$$k^{\frac{3}{2}} 4^k (\alpha(h+1))^{\frac{1}{s}} \leq 1$$

$$d = \max_{1<j<\#\Omega(\#\Omega-1)/2} |A_j|^s < \infty, \text{ for } s \in (2, 3]$$

$$2kh + 1 < \#\Omega$$

Where $\varphi(t)$ is the characteristic function of $\delta_j^{(\nu)}$. From now on, we will work with its derivative,

$$\varphi'(t) = \hat{\sigma}_n^{-1} \sum_{q \in \{1,\cdots,\#\Omega(\nu+j)\}\ s.t.\ \omega_{(q)} \in \Omega(\nu+j)} \mathbb{E}[A_q \exp\{it\delta_q^{(0)}\}]$$

Add and subtract $\exp\{it\delta_q^{(1)}\}$,

$$\varphi'(t) = \hat{\sigma}_n^{-1} \sum_{q \in \{1,\cdots,\#\Omega(\nu+j)\}\ s.t.\ \omega_{(q)} \in \Omega(\nu+j)} \mathbb{E}\left[A_q \exp\{it\delta_q^{(1)}\}(\exp\{it(\delta_q^{(0)} - \delta_q^{(1)})\} - 1)\right] +$$

$$+\ \hat{\sigma}_n^{-1} \sum_{q \in \{1,\cdots,\#\Omega(\nu+j)\}\ s.t.\ \omega_{(q)} \in \Omega(\nu+j)} \mathbb{E}\left[A_q \exp\{it\delta_q^{(1)}\}\right]$$

Together with the following differential relation

$$\mathbb{E}[\exp\{it\delta_j^{(r)}\}] = \mathbb{E}[\eta_j^{(r)} + 1]\varphi_n(t) + \mathbb{E}\left[(\eta_j^{(r)} - \mathbb{E}(\eta_j^{(r)}))\exp\left\{it\delta_j^{(0)}\right\}\right]$$

this process can be repeated to obtain:

$$\begin{aligned}\varphi'(t) =& i\left[\sum_{j=1}^{\#\Omega} a_j^{(1)}\mathbb{E}\left(\eta_j^{(2)}+1\right)+\sum_{r=3}^{k}\sum_{j=1}^{\#\Omega} a_j^{(r-1)}\mathbb{E}\left(\eta_j^{(r)}+1\right)\right]\varphi(t)+\\ &+i\sum_{j=1}^{\#\Omega}\mathbb{E}\left(A_j e^{i\delta_j^{(1)}}\right)+\\ &+i\sum_{r=2}^{k}\sum_{j=1}^{\#\Omega}\left[\mathbb{E}\left(A_j\prod_{l=1}^{r-1}\xi_j^{(l)}e^{it\delta_j^{(r)}}\right)-\mathbb{E}\left(A_j\prod_{l=1}^{r-1}\xi_j^{(l)}\right)\mathbb{E}e^{it\delta_j^{(r)}}\right]+\\ &+i\sum_{r=2}^{k}\sum_{j=1}^{\#\Omega} a_j^{(r-1)}\mathbb{E}\left[\left(\eta_j^{(r)}-\mathbb{E}\eta_{ij}^{(r)}\right)e^{it\delta_j^{(0)}}+i\sum_{j=1}^{\#\Omega}\mathbb{E}\left(A_j\prod_{l=1}^{k}\xi_j^{(l)}e^{it\delta_j^{(k)}}\right)\right]\end{aligned}$$

It is worth to point out that all newly defined variables inherit the mixing properties of the original series. The strategy of proof now consists in evaluating each of the terms on the right-hand side of this expression then applying Esseen's inequality (see Esseen (1945) or Ibragimov and Linnik (1971)). In what follows, I will impose weaker conditions than those required for the subsampling procedure to work.

Part 1

Let $\mathcal{I}$ be a discrete uniform variable on the set $\{1, ..., \#\Omega\}$, independent of X_q. By applying the Bruss-Robertson-Steele and the Cauchy-Schwarz inequalities we see that for each $i \in \mathcal{I}$,

$$\mathbb{E}[|\delta_{\mathcal{I}}^{(i)}|^{\alpha}] \leq (2hi+1)^{\alpha}(\#\Omega)^{-1}\sum_{j=1}^{\#\Omega}\mathbb{E}[|A_q|^{\alpha}]$$

By applying Holder's inequality,

$$\sum_{q=1}^{\#\Omega}\mathbb{E}\left[|A_q|\times|\delta_q^{(1)}|^{s-1}\right] \leq (\#\Omega)\left(\mathbb{E}[|A_{\mathcal{I}}|^s]\right)^{1/s}\left(\mathbb{E}[|\delta_{\mathcal{I}}^{(1)}|^s]\right)^{(s-1)/s} \leq (2h+1)^{s-1}\sum_{q=1}^{\#\Omega}\mathbb{E}[|A_q|^s]$$

We now apply a result from Chipp (1979)[26] to obtain:

$$\sum_{j=1}^{\#\Omega}\mathbb{E}[A_j\delta_j^{(1)}] = 1+12\Theta\sqrt{1+12\lambda}d^{2/s}c_0^{-3/2}\tilde{\sigma}_{\delta}\left(\alpha(h+1)\right)^{(2-s)/(2s)}$$

Where Θ is some function bounded, in absolute value, by 1.

[26] Equation 4.6 in page 60. This result can be proven further splitting $\delta_j^{(1)}$ into two smaller sums and then applying, in sequence, Holder's and Markov's inequalities to the expectation of each of those.

We can also see that,

$$i\sum_{j=1}^{\#\Omega} a_j^{(1)} = -\sum_{j=1}^{\#\Omega} \mathbb{E}\left[\left(A_j\delta_j^{(1)}\right)\right] t + \Theta\frac{2^{3-s}}{s-1}\sum_{j=1}^{\#\Omega} \mathbb{E}\left[|A_j|\,|\delta_j^{(1)}|^{s-1}\right]|t|^{s-1}$$

Putting those together we arrive at the final evaluation of this first term,

$$i\sum_{j=1}^{\#\Omega} a_j^{(1)} = -t + 12\Theta\sqrt{1+12\lambda}d^{2/s}c_0^{3/2}\tilde{\sigma}_\delta(\alpha(h+1))^{(s-2)/(2s)}|t| + \Theta\frac{2^{3-s}}{s-1}\frac{d}{c_0}\frac{(2h+1)^{s-1}}{\tilde{\sigma}_\delta^{s-2}}|t|^{s-1}$$

Part 2

Leveraging from Tikhomirov (1980), we can show that:

$$|a_j^{(r-1)}| \leq 2d^{2/s}\frac{h+1}{\tilde{\sigma}_\delta^2}|t|$$

For any t such that $|t| \leq \tilde{\sigma}_\delta d^{1/s}(h+1)/32$. The proof also consists in applying a sequence of inequalities for strongly mixing variables. I point to the original paper for a detailed proof of this claim.

Part 3

We now evaluate the first term in the sum,

$$\left|\sum_{j=1}^{\#\Omega} a_j^{(1)}\mathbb{E}\left(\eta_j^{(2)}+1\right) + \sum_{r=3}^{k}\sum_{j=1}^{\#\Omega} a_j^{(r-1)}\mathbb{E}\left(\eta_j^{(r)}+1\right)\right|$$

By using the second part, this boils down to

$$\left|\sum_{j=1}^{\#\Omega} a_j^{(1)}\mathbb{E}\eta_j^{(2)} + \sum_{r=3}^{k}\sum_{j=1}^{\#\Omega} a_j^{(r-1)}\mathbb{E}\left(\eta_j^{(r)}+1\right)\right|$$

By putting together

$$|a_j^{(r-1)}| \leq 2d^{2/s}\frac{h+1}{\tilde{\sigma}_\delta^2}|t|$$

$$k^{3/2}4^k\left(\alpha(h+1)\right)^{1/s} \leq 1$$

and the assumption that the variance of the numerator of the U-Statistic grows at least linearly in the sample size, we show that the summand of interest is bounded by:

$$\sum_{j=1}^{\#\Omega} |a_j^{(1)}| \times \mathbb{E}[|\eta_j^{(2)}|] + \sum_{r=3}^{k}\sum_{j=1}^{\#\Omega} |a_j^{(r-1)}|$$

From Chipp (1979),

$$\left|\sum_{j=1}^{\#\Omega} a_j^{(r-1)} \mathbb{E}\left[\left(\eta_j^{(r)} - \mathbb{E}[\eta_j^{(r)}]\right)\right] \exp\{itZ_{\#\Omega}\}\right| \leq$$

$$\left\{\left(\sum_{j=1}^{\#\Omega} \sum_{|p-j|\leqslant 2rh} + \sum_{j=1}^{\#\Omega} \sum_{|p-j|>2rh}\right) a_j^{(r-1)} \overline{a_p^{(r-1)}} Cov\left(\eta_j^{(r)}, \eta_p^{(r)}\right)\right\}^{1/2} \leq$$

$$\leq \left\{\sum_{j=1}^{\#\Omega} \sum_{|p-j|\leqslant 2rh} \left|a_j^{(r-1)}\right| \left|a_p^{(r-1)}\right| \left(\mathbb{E}\left|\eta_j^{(r)}\right|^2\right)^{1/2} \left(\mathbb{E}\left|\eta_p^{(r)}\right|^2\right)^{1/2}\right\} +$$

$$+ \left\{24 \sum_{j=1}^{\#\Omega} \sum_{|p-j|>2rh} \left|a_j^{(r-1)}\right| \left|a_p^{(r-1)}\right| (\alpha(|p-j| - 2rh))^{(s-2)/s} \quad \left(\mathbb{E}\left[\left|\eta_j^{(r)}\right|^s\right]\right)^{1/s} \left(\mathbb{E}\left[\left|\eta_p^{(r)}\right|^s\right]\right)^{1/s}\right\}$$

This quantity is bounded by

$$2a^{(r-1)} \left\{\sum_{j=1}^{\#\Omega} \sum_{|p-j|\leqslant 2/h} 1\right\}^{1/2} + 4\sqrt{6} a^{(r-1)} \times$$

$$\times \left\{\sum_{j=1}^{\#\Omega} \sum_{\{p-j|>2rh} (\alpha(|p-j| - 2rh))^{(3-2)/3}\right\}^{1/2} \leqslant C c_0^{-1/2} B_{\#\Omega} \left[r^{1/2}(h+1)^{1/2} + \alpha^{1/2}\right] a^{(r-1)}.$$

Part 4

The goal now is to evaluate:

$$\sum_{j=1}^{\#\Omega} |\mathbb{E}(A_j \exp\{it\eta_j^{(1)}\})|$$

Part 5

Putting together Theorem 17.2.1 in Ibraginov and Linnik (1971) and Chipp's (1979) Equation 4.5, we obtain:

$$\begin{aligned}
&\left|\mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(l)} \exp\{itz_j^{(r)}\}\right) - \mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(l)}\right) \mathbb{E}\exp\{itz_j^{(r)}\}\right| \\
&\leqslant \left|\mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(l)} \exp\{itz_j^{(r)}\}\right) - \mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(l)} \exp\{itz_j^{(r)}\}\right) \mathbb{E}\exp\{it_j^{(r)}\}\right| + \\
&+ \left|\mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(l)} e^{it_j^{(r)}}\right) - \mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(t)}\right) \mathbb{E}e^{it_j^{(r)}|}\right| + \\
&+ \left|\mathbb{E}e^{it\hat{\delta}_j^{(r)}\mathbb{E}} e^{i\hat{t}_j^{(r)}} - \mathbb{E}e^{it_2(r)}\right| \left|\mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(\#\Omega)}\right)\right| \\
&\leqslant 48\frac{d^{1/s}}{B_u}(\alpha(h+1))^{(s-1)/s}2^{r-1} \\
&+ \left|\mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(l)} e^{it_j^{(r)}}\right) - \mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(t)}\right) \mathbb{E}e^{it_j^{(r)}|}\right| + \\
&+ \left|\mathbb{E}e^{it\hat{\delta}_j^{(r)}\mathbb{E}} e^{i\hat{t}_j^{(r)}} - \mathbb{E}e^{it_2(r)}\right| \left|\mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(\#\Omega)}\right)\right| \\
&\leqslant 48\frac{d^{1/s}}{B_u}(\alpha(h+1))^{(s-1)/s}2^{r-1}
\end{aligned}$$

If we combine this last inequality with the condition that the variance increases at least linearly – as previously done – and the restriction that $k^{\frac{3}{2}}4^k(\alpha(h+1))^{\frac{1}{s}} \leq 1$,

$$\sum_{r=2}^{k}\sum_{j=1}^{\#\Omega}\left|\mathbb{E}\left(A_j \prod_{l=1}^{r-1} \xi_j^{(l)} e^{iz_j^{(r)}}\right) - \mathbb{E}\left(A_j \sum_{l=1}^{r-1} \xi_j^{(l)}\right) \mathbb{E}e^{iz_j^{(r)}}\right| \leqslant 48c_0^{-1}d^{1/s}B_{\#\Omega}(\alpha(h+1))^{(s-2)/s}$$

Part 6

The goal of this last part is to characterize the behavior of the characteristic function $\varphi(\cdot)$ and its derivative as the proof of the theorem relies on the application of the Esseen inequality

(Esseen (1945)). For the sake of conciseness, define:

$$\begin{aligned}
\psi_0 &= C\frac{d^{1/s}}{c_0}B_{\#\Omega}(\alpha(h+1))^{(s-2)/s}, \\
\psi_1 &= C\frac{d^{2/s}}{c_0^{3/2}}(1+\lambda)^{1/2}B_{\#\Omega}(\alpha(h+1))^{(s-2)/2s} \\
\psi_2 &= C\frac{d^{3/s}}{c_0}\frac{(h+1)^2}{B_{\#\Omega}} \\
\psi_3 &= C\frac{d}{c_0}\frac{(h+1)^{s-1}}{B_{\#\Omega}^{s-2}} \\
\tilde{\psi}_0 &= C\frac{d^{1/s}}{c_0^{1/2}}\left((h+1)^{1/2}+\lambda^{1/2}\right)(\alpha(h+1))^{(s-2)/s}+\psi_0 \\
\tilde{\psi}_2 &= C\frac{d^{3/s}}{c_0^{1/2}}\left((h+1)^{1/2}+\alpha^{1/2}\right)\frac{(h+1)^3}{B_{\#\Omega}^2} \\
T_2 &= \min\left\{\frac{1}{\psi_0};\frac{1}{6\psi_2};\left(\frac{1}{6\psi_3}\right)^{1/(s-2)}\right\}
\end{aligned}$$

Under strong mixing, the differential equation presented in the beginning of this proof can be written as:

$$\varphi'(t) = (-t+\Theta\Psi(t))\varphi(t)+\Theta\tilde{\Psi}(t)$$

Where $\Psi(t) \equiv \psi_0 + \psi_1|t| + \psi_2 t^2 + \psi_3|t|^{s-1}$ and $\tilde{\Psi}(t) \equiv \tilde{\psi}_0 + \tilde{\psi}_2 t^2$. This can be shown by applying the results derived in Parts 1-6 to the equation for $\varphi'(t)$ obtained with Tikhomirov's (1980) procedure. By solving it, we obtain:

$$\begin{aligned}
\left|\varphi(t) - \exp\left\{-t^2/2\right\}\right| \leq |x_0|\exp\left\{-\sqrt{t}/2+|x_0|\right\}+ \\
+\exp\left\{-t^2/2\right\}\int_0^{|t|}\tilde{\Psi}(u)\exp\left\{u^2/2+\int_{|u|}^{|t|}\Psi(v)dv\right\}du
\end{aligned}$$

With $x_0 = \int_0^t \Theta\Psi(u)du$. For any $u \in [0,t]$, $t \leq T_2$, and $\psi_1 \leq 1/6$, $\int_u^t \Psi(v)dv \leq \frac{t^2-u^2}{4}+1$ which, in turn, means that $\int_0^{|t|} u^2\exp\{u^2/4\}du \leq 2|t|\exp\{t^2/4\}$ and $\int_0^{|t|}\exp\{u^2/4\}du \leq \min\left(4/|t|,|t|\right)\exp\{t^2/4\}$. Combining the solution of the differential equation from above with these last inequalities we obtain the main result

$$\begin{aligned}
\left|\varphi(t)-\exp\left\{-t^2/2\right\}\right| \leq \left(\psi_0|t|+\psi_1 t^2/2+\psi_2|t|^3/3+\psi_3|t|^s/s\right)\exp\{-t^2/4+1\}+ \\
+\exp\{1\}\,\tilde{\psi}_0\min\left(4/|t|,|t|\right)+2\exp\{1\}\,\psi_2|t|
\end{aligned}$$

Finally, we can apply Esseen's inequality to obtain the convergence rate:

$$\begin{aligned}\Delta_n \leqslant C \Bigg[& \aleph^s \frac{(h+1)^{s-1}}{\tilde{\sigma}_\delta^{(s-2)}} + \aleph^3 \frac{(h+1)^2}{\tilde{\sigma}_\delta} + \aleph^2 \left((h+1)^{1/2} + \lambda^{1/2} \right) \frac{h+1}{\tilde{\sigma}_\delta} + \\ & + \aleph^2 (1+\lambda)^{1/2} \tilde{\sigma}_\delta (\alpha(h+1))^{(s-2)/2s} + \aleph \left((h+1)^{1/2} + \lambda^{1/2} \right) (\alpha(h+1))^{(s-2)/s} \Bigg]\end{aligned}$$

Concluding the proof of the Theorem.

D Joint macroeconomic forecasts

In this exercise, I construct prediction intervals for simulated macroeconomic data generated from the medium-scale dynamic stochastic general equilibrium model in Smets and Wouters (2007). The model is estimated with US data from 1966Q1 to 2004Q4.

The data generating process is the vector autoregressive model (VAR) that describes the equilibrium of Smets and Wouter's (2007) model. This model consists of utility-maximizing infinitely-lived households, profit-maximizing intermediate and final goods firms, and fiscal and monetary authorities. The prices of the goods and the wages are sticky, market imperfections that generate the rigid behavior in the time-series. The monetary authority changes the current interest rate in this economy by following a Taylor rule and the fiscal authority equilibrates the government's budget in every period. I refer the reader to the paper for more details on the derivation of this solution and the estimation procedure.

Figure 10 shows a typical realization of an inflation series drawn from this model. Figure 11 shows the constructed prediction intervals at the 5% significance level.

The empirical follow-up of this simulation is presented in Appendix E.

It is also easy to construct joint prediction intervals (see Wolf, 2015) by subsampling since the algorithm is the same except for having a vector of variables on the left hand side of the statistical model. In this exercise, I estimate a VAR(4) given by

$$\begin{bmatrix} y_t \\ \pi_t \end{bmatrix} = \begin{bmatrix} \gamma_0^y \\ \gamma_0^\pi \end{bmatrix} + \sum_{j=1}^{4} \begin{bmatrix} y_{t-j} \\ \pi_{t-j} \end{bmatrix} \Gamma_j + \varepsilon_t$$

where y_t and π_t are the output gap and the inflation rate in date t, respectively. γ_0^y and γ_0^π are constants, each Γ_j is a 2-by-2 parameter matrix, and ε_t is a 2-by-1 random vector. Running Algorithm 2 on this data, we can construct prediction regions. One of those is presented in Figure 12 below.

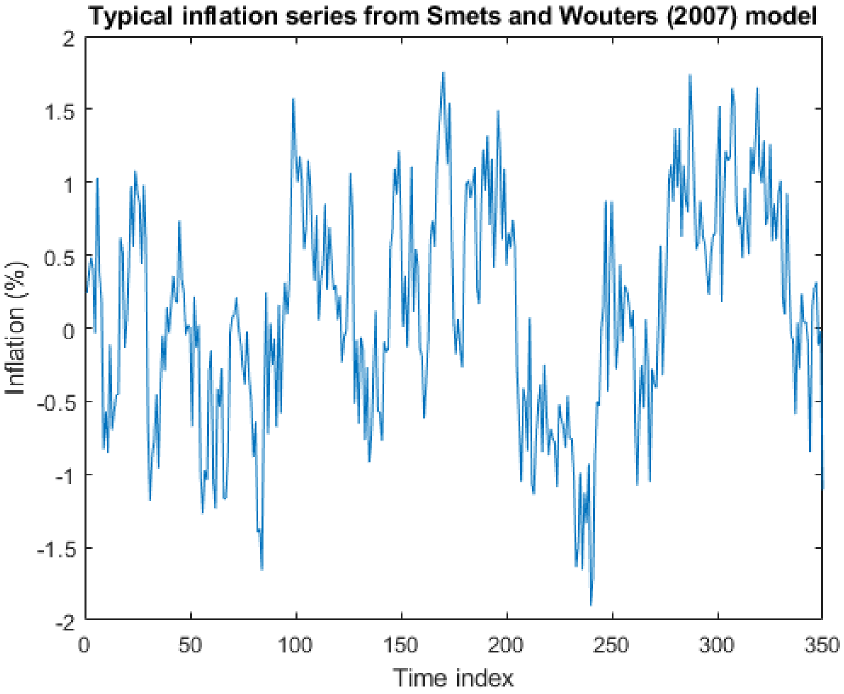


Figure 10: Simulated series of inflation using the Smets and Wouters (2007) model

E Inflation forecasting - Nakamura (2005)

Nakamura (2005) shows that neural networks can provide more accurate macroeconomic predictions. She shows that an "early stopping" training strategy (stop training the network as soon as the mean square prediction error starts increasing in the validation set). The "early stopping" procedure tackles the tendency of big nonparametric models to overfit the data.

Here, I revisit her paper by reestimating her model and computing prediction intervals around the forecasts. The goal is to predict inflation (denoted π) using a parsimonious neural network. The forecasts are constructed based on the fully-connected model

$$\hat{\pi}_t = \alpha_0 + \alpha_1 \tanh(\beta_0^1 + \beta_1^1 \pi_{t-1} + \beta_2^1 \pi_{t-2}) + \alpha_2 \tanh(\beta_0^2 + \beta_1^2 \pi_{t-1} + \beta_2^2 \pi_{t-2})$$

where $\tanh(\cdot)$ is the hyperbolic tangent function. Moreover, let $\hat{\pi}_{t+h}$ be the forecast of inflation h periods ahead. I present results of two different exercises. The first one, which I name "online prediction," consists of re-estimating the model as soon as the new data point

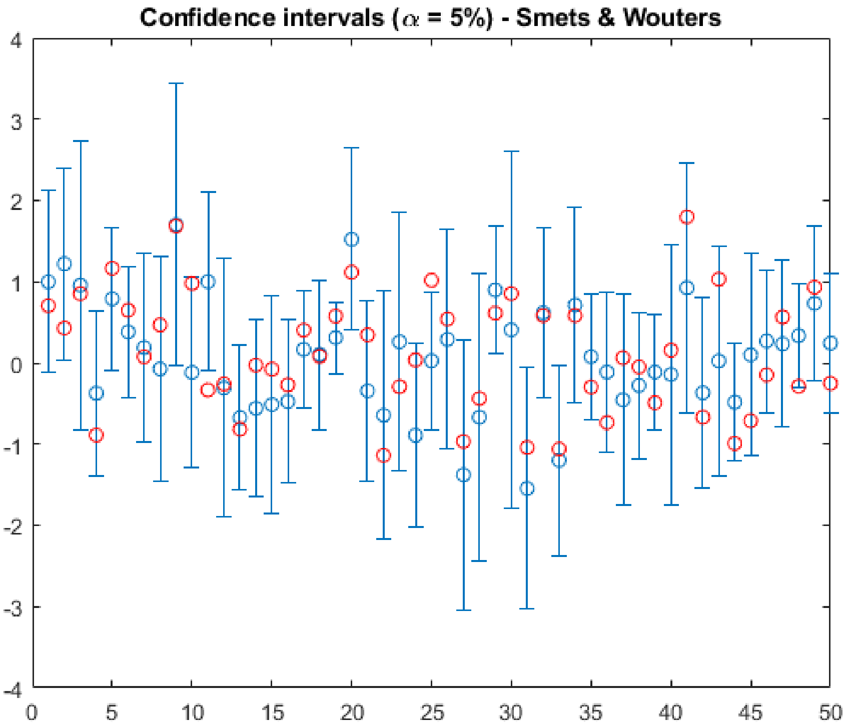


Figure 11: Prediction intervals for one-period-ahead inflation.

is revealed. I then forecast inflation one period ahead and repeat the estimation when a new data point is revealed. Figure 13 presents the results for forecasts constructed with the neural network described previously.

These graphs present the series of realized inflation in blue to the left of 2017Q3, and the series of 1-step-ahead forecasts constructed with older data in red and what was later realized in black. The dashed red lines are the 95% prediction intervals.

The second exercise, named "offline prediction," concerns forecasts for longer horizons based on the information available up until the date the model is estimated (construct $\hat{\pi}_{n+h}$ with $h \geq 1$ based on a sample of size n). One way to approach this is by training the neural network for values of $h \geq 1$

$$\hat{\pi}_{t+h} = \alpha_0 + \alpha_1 \tanh(\beta_0^1 + \beta_1^1 \pi_t + \beta_2^1 \pi_{t-1}) + \alpha_2 \tanh(\beta_0^2 + \beta_1^2 \pi_t + \beta_2^2 \pi_{t-1})$$

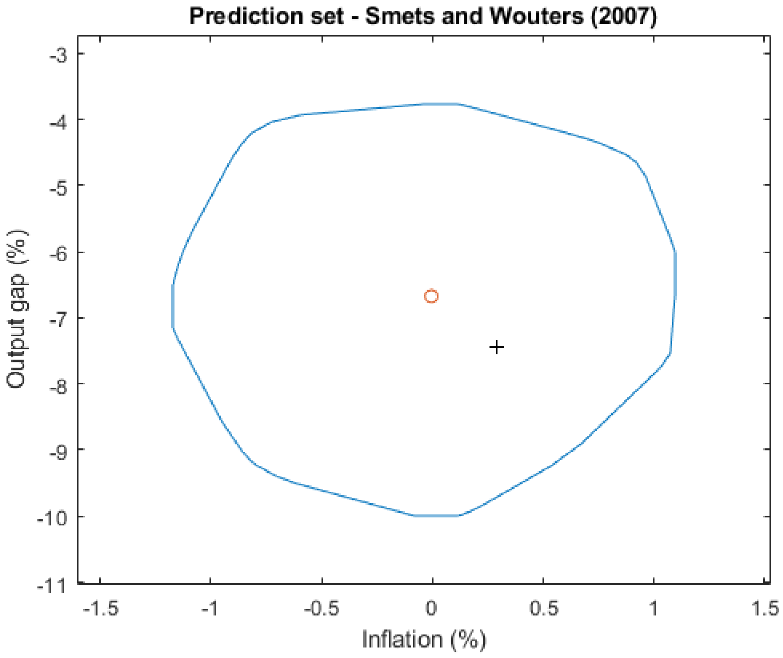


Figure 12: Joint prediction region in the Smets and Wouters's (2007) simulations

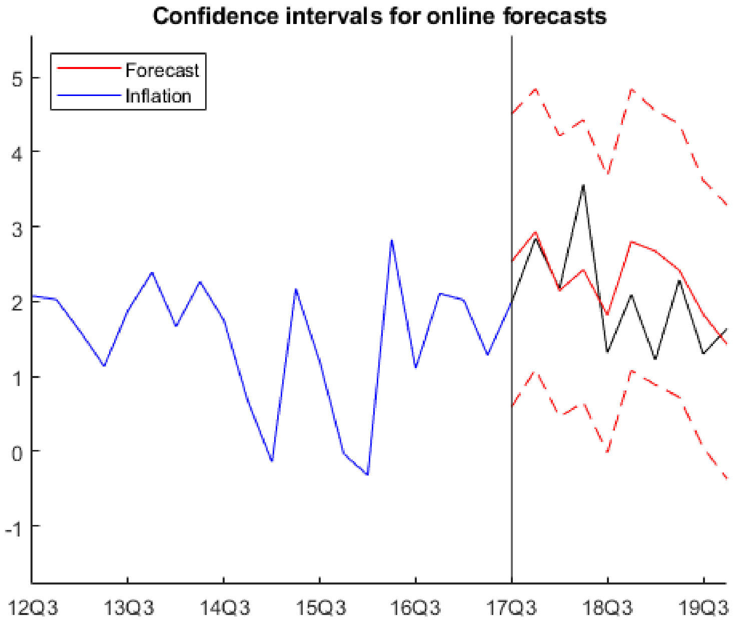


Figure 13: Online prediction in Nakamura (2005) using NN

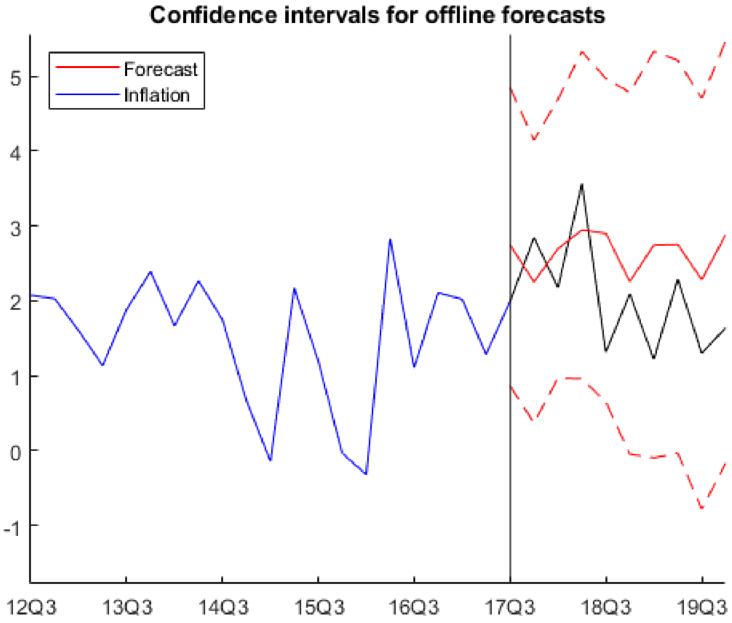


Figure 14: Offline prediction in Nakamura (2005) using NN