Submitted to Journal of Forecasting

Long-Term Prediction Intervals of Economic Time Series

Marek Chudy, Sayar Karmakar, Wei Biao Wu

University of Vienna & University of Chicago, November 22, 2017

We construct long-term prediction intervals for time-aggregated future values of univariate economic time series. We provide data-driven adjustments for estimation procedure in order to improve coverage probability under short sample constraint. Pseudo-out-of-sample evaluation shows that our methods perform at least as well as the selected alternative methods, which are based on Bayesian approach, model-implied analytic formulas and bootstrapping. The best of our methods is used for prediction of eight macroeconomic indicators over horizon spanning several decades.

Key Words and Phrases: Prediction intervals, long horizon, time-aggregation, stationary bootstrap, kernel quantile estimator, pseudo-out-of-sample evaluation

1. Introduction

Many economic and financial time series are driven by both short and long-term dynamics. Their importance for forecasting varies from series to series and depends on forecast origin and horizon. In this paper, we deal with situation when forecasting horizon is long with respect to the sample size. Such long-term forecasts are published every year by the U.S. Congressional Budget Office (CBO) as Long-Term Budget Outlook[1]. CBO predicts US federal spending and revenue growth for several decades ahead under the assumption of stable tax and spending policies. Such conditions are never met in practice. Hence CBO argues, that the predictions are not reflecting their beliefs in what will actually happen, but help measuring budgetary effects of proposed changes in federal revenues or spending. But legislation is not the only source of uncertainty about future economy. Structural changes happen in the long run with respect to other variables as well as to own dynamics.

For decades, economists and econometricians spoke about Great Moderation, i.e., period when key economic indicators like GDP exhibited low volatility. Recently, economies around the world went trough turbulent periods of high volatility in both private and public sector. Hardly, anybody could have predicted that 30 years ago. Well, the Budget and Economic Outlook from January 2000 said that the baseline projections allow for an average recession sometime in the next 10 years (2000-2010). Today, we already know that the recession which was about to come was far bigger than average. We could and should help the lawmakers to prevent or at least damper sever recessions or crisis in the future by giving them realistic estimates about the true state of the economy for the next decades (see Christoffersen and Diebold, 1998). Trust funds, various entities

[1] Available from https://www.cbo.gov/publication/52480

which operate with derivatives and experts which are pricing pension insurance, they all rely on long-run forecasts of economic time series and/or their aggregates (Kitsul and Wright, 2013). From financial perspective, long run volatility of asset returns is crucial for portfolio allocation (see Pastor and Stambaugh, 2012, Bansal et al., 2016, who address the issue of time-aggregation).

Decision-makers in both public and private sector nowadays call for predictions in form of boundaries $[L, U]$ covering the future values of interest with high probability instead of point-forecast of questionable accuracy. Such boundaries are usually called prediction intervals (PI's). The main advantage of PI's over point-forecasts is their ability to assess the uncertainty about the target. The major difficulty with PI's is about their width. Often, PI's implied by (parametric) autoregressive models are too narrow. The main reason is that most analytic formulas ignore parameter and model uncertainty. Further issues are related to false distributional assumption on innovation process and restrictive assumption about stationarity including structural stability. Some of these can be solved using bootstrapping (Clements and Taylor, 2001, 2003, Clements and Kim, 2007). In contrast, methods which work well for non-stationary series in short-horizon, such as exponential smoothing, produce too wide, thus impractical PI's for the long-run. A comprehensive assessment of this topic is provided in Chatfield (1993). Otherwise, the literature on long-run PI's is relatively sparse.

The uncertainty about predicting long-run time-aggregates of the US economic growth, inflation, population, etc. has been recently elaborated in Müller and Watson (2016). Their target is construction of Bayesian prediction intervals for averages of growth rates over 10 to 75 years ahead. Their basic idea is to extract information from specific frequency band. Similar idea in band-regression framework goes back to Engle (1974) and was used for forecasting in Reschenhofer and Chudy (2015). Meanwhile Müller and Watson (2016) provide asymptotically valid PI's under rich class of models for data generating processes (DGP's) with long memory. In their set-up, forecasting horizon grows proportionally with the sample size. Significant contributions regarding the problem of model and parameter uncertainty are provided too. Their approach is Bayesian, but their PI's have frequentist's coverage thanks to utilizing so called least favorable distribution for enhanced robustness.

A simple but theoretically valid method for PI's of long-run aggregated future time series values was proposed in Zhou et al. (2010). However, the latter paper evaluates the PI's only based on simulated data. We find it therefore necessary to verify their results on real data. In particular, economic time series are know for relatively short sample (given that most post WWII economic indicators are reported on monthly/quarterly bases), high-persistence[2] (see Diebold and Rudebusch, 1989, Baillie, 1996, Diebold and Linder, 1996, who also give PI's), heteroscedasticity and structural changes (Cheng et al., 2016). Problems such as the latter are inevitable in the long-run (Stock and Watson, 2005). We therefore provide valid data-driven adjustment of Zhou et al. (2010) motivated by these characteristics with a special focus on predictive performance under short samples. Our adjustments employ stationary bootstrap (Politis and Romano, 1994) and kernel quantile estimator (Sheather and Marron, 1990). Probing into the involved derivations related to bootstrapping and the quantile consistency, we also provide some intuitive justifications for these empirically-motivated adjustments.

Since neither Zhou et al. (2010) or Müller and Watson (2016) compare their PI's to any sort of benchmark, we take over this responsibility and conduct an extensive pseudo-out-of-sample (POOS) comparison study. Long daily time series of SP 500 returns and US 3-months TB interest rates provide basis for statistical assessment of coverage probability and precision (measured by the

[2] Nevertheless, anti-persistence can be observed here as well, often as result of (over-)differencing.

median width of PI). We augment the comparison with PI's implied by ARFIMA-GARCH models. In this case,the PI's are obtained from (i) forecasts for time-aggregated series or (ii) from time-aggregated forecasts of disaggregated series. To compute (i) and (ii) we use both analytic formulas and bootstrap path-simulations. Results of POOS exercise suggest that the adjusted methods of Zhou et al. (2010) perform at least as well as their competitors in terms of coverage probability and show some advantages in terms of precision.

Throughout this paper, we focus on PI's estimated from historical data on the predicted series. A multivariate or even high-dimensional extension would of course be attractive as it is widely believed that big data contain additional forecasting power. Unfortunately, in the economic literature the boom of forecasting with many predictors (e.g. Stock and Watson, 2012, Elliott et al., 2013, Kim and Swanson, 2014) is mainly focused on short horizons and point-forecasting (for exception see Bai and Ng, 2006). This is not just by coincidence. Many economic time series exhibit persistence (of various degrees). This is their key property in the long-run. These long-term effects, combined over many series, are difficult to understand. Not only but also due to their dependence on unknown nuisance parameters (see Elliott et al., 2015).

The article continues as follows: In Section 2 we summarize methods of Zhou et al. (2010). We then justify the data-driven adjustments and illustrate their performance using simulations. Section 3 provides details on implementation of the adjusted methods and their competitors and gives results of the POOS comparison. Section 4 provides PI's for eight macro-indicators over horizon of up to seven decades from now. Section 5 contains concluding remarks.

2. Methods and simulation results

In this section we briefly discuss the basic prediction intervals proposed in Zhou et al. (2010) and some modifications we are proposing to adjust for short sample size and real data evaluation. Assume we observe $y_1, \ldots, y_T$ and we want to provide PI for $\bar{y}_{T+1:T+m} = (y_{T+1} + \ldots + y_{T+m})/m$. As Zhou et al. (2010) suggests, for a weakly dependent series y_i, predicting the m step ahead future-aggregated value is easier for larger m than smaller m since with a large m and short range dependence the dependence of $y_{T+1} + \ldots + y_{T+m}$ on $y_1, \ldots y_T$ diminishes. However, for the case of long range dependent series the above is not true and constructing the PI is tougher. To this end, we first very briefly discuss the two suggestions from Zhou et al. (2010) followed by their merits and demerits. Then we discuss our proposals of variants of these methods to achieve good forecasting performance for real data.

2.1. Summary of the two methods from Zhou et al. (2010)

CLT based method (asp) If the demeaned process ($z_t = y_t - \bar{y}$) show short-range dependency and light-tailed behavior, then Zhou et al. (2010) suggested using the long-run covariance σ to construct the following PI for the (z_t) process

$$[L, U] = [\sigma Q^N(\alpha/2)/\sqrt{m}, \sigma Q^N(1-\alpha/2)/\sqrt{m}], \tag{2.1}$$

in the light of a quenched CLT. However, since σ is unknown, one can use e.g., sub-sampling block estimator (see eq. (2) in Dehling et al., 2013)

$$\tilde{\sigma} = \frac{\sqrt{\pi/2l}}{T} \sum_{i=1}^{\kappa} \left| \sum_{t=(i-1)l+1}^{il} z_t \right|, \tag{2.2}$$

with the block length l and number of blocks $\kappa = \lceil \frac{T}{l} \rceil$. Then the PI for the original data (y_t) with nominal coverage $100(1-\alpha)\%$ is given by[3]

$$[L, U] = [\bar{y} + \tilde{\sigma} Q^t_{\kappa-1}(\alpha/2)/\sqrt{m}, \bar{y} + \tilde{\sigma} Q^t_{\kappa-1}(1-\alpha/2)/\sqrt{m}]. \tag{2.3}$$

Note that the df is related to the estimation of σ.

Quantile based method (emp) This method does not necessarily require short-range or light-tailed behavior and has more general applicability. Given data $y_t, t = 1, \ldots, T$, construct for $t = m, \ldots, T$,

$$\tilde{y}_t = \bar{y}_{t-m+1:t} = \frac{1}{m} \sum_{j=1}^{m} y_{t-j+1}. \tag{2.4}$$

Then the PI with nominal coverage $100(1-\alpha)\%$ is given by

$$[L, U] = [\hat{Q}(\alpha/2), \hat{Q}(1-\alpha/2)], \tag{2.5}$$

where $\hat{Q}$ is the respective empirical quantile of $\tilde{y}_t, t = m, \ldots, T$.

2.1.1. CLT-based method for light-tailed processes

This is applicable only for linear processes with light-tailed innovations and short-range dependence. We also write the discussion only for the zero-mean process z_t instead of the observed process y_t. Under stationarity the problem of predicting $\bar{z}_{T+1:T+m} = (S_{T+m} - S_T)/m = (z_{T+1} + \ldots + z_{T+m})/m$ after observing $z_1, \ldots, z_T$ can be equally thought as predicting $S_m/\sqrt{m}$ after observing $\ldots, z_{-1}, z_0$. Let $\mathcal{F}_0$ be the σ-field generated by $\ldots, z_{-1}, z_0$. Assume $\mathbb{E}(|z_i|^p) < \infty$ for some $p > 2$. Wu and Woodroofe (2004) proved that, if for some $q > 5/2$,

$$\|E(S_m|\mathcal{F}_0)\| = O\left(\frac{\sqrt{m}}{\log^q m}\right), \tag{2.6}$$

then we have the a.s. convergence

$$\Delta(\mathbb{P}(S_m/\sqrt{m} \leq \cdot|\mathcal{F}_0), N(0, \sigma^2)) = 0 \text{ a.s.}, \tag{2.7}$$

where Δ denotes the Levy distance, $m \to \infty$ and $\sigma^2 = \lim_{m\to\infty} \|S_m\|^2/m$ is the long-run covariance. Under the assumption of z_i assuming linearity of the following form,

$$z_i = \sum_{j=0}^{\infty} a_j \epsilon_{i-j}. \tag{2.8}$$

where ϵ_i are mean-zero, i. i. d. with finite second moment, evaluating (2.6) is easy. We assume the special formulation of a_i

$$a_i = i^{-\chi} (\log i)^{-A}, \quad \chi > 0, A > 0, \tag{2.9}$$

[3] We denote $Q^t_{\kappa-1}$ the quantile of Student t distribution with $\kappa - 1$ degrees of freedom.

where larger χ and A means fast decay rate of dependence. Further assume

$$a_j = \begin{cases} j^{-\chi}(\log j)^{-A}, & \text{for } 1 < \chi < 3/2, A > 0 \\ j^{-\chi}(\log j)^{-A} & \text{for } \chi \geq 3/2, A > 5/2, \end{cases} \tag{2.10}$$

then (2.6) holds.

Proof.

$$\|\mathbb{E}(S_m|\mathcal{F}_0)\|^2 = \|(a_1 + \ldots + a_m)\epsilon_0 + (a_2 + \ldots + a_m)\epsilon_{-1} + \ldots\|^2 = \sum_{i=1}^{m} b_i^2, \tag{2.11}$$

where $b_i = a_i + \ldots + a_m$. Note that, $\sum_{i=1}^m b_i^2$ assumes the following value depending on $\chi > 3/2$ or not.

$$\sum_{i=1}^{m} b_i^2 = \begin{cases} m^{3-2\chi}(\log m)^{-2A}, & \text{for } 3 - 2\chi > 0 \\ O(1) & \text{for } 3 - 2\chi \leq 0. \end{cases} \tag{2.12}$$

Thus, (2.6) holds, for the following a_j in the short-range dependent linear case.

□

However it remains to estimate the long-run variance σ. One popular choice is to use a lag window estimate as suggested in Zhou et al. (2010). For this paper, we stick to use estimator of the form (2.2).

2.1.2. *Advantages and drawbacks of CLT-based method*

Note that, this result does not require the rate of growth of m compared to the sample size T. If m is large, this is a good intuitive method. However, the following could be pointed out as drawbacks.

- It is not tuning parameter free. The predictive performance heavily depends on estimation quality of σ and the choice of *df* of the t-distribution for computing the quantiles.
- For small m, the approximation in the limit theorem does not work well.
- For heavy-tailed innovations or long-range dependence, the notion of σ, the long-run variance does not exists and thus the result is not applicable.

2.1.3. *Quantile-based method for possibly heavy-tailed processes*

We look at the empirical quantiles of the partial sums of length m. Since we are not predicting single or average of a few observations for the future, one can intuitively expect, in the light of weak dependence,

$$\mathbb{P}(a \leq \frac{y_{T+1} + \ldots y_{T+m}}{m} \leq b | y_1, \ldots, y_T) \approx \mathbb{P}(a \leq \frac{y_{T+1} + \ldots y_{T+m}}{m} \leq b), \tag{2.13}$$

if m/T is not too small. Thus it suffices to estimate the quantiles of $(y_{T+1} + \ldots + y_{T+m})/m$ which can then be done by empirical quantiles of $\tilde{y}_i$ where

$$\tilde{y}_i = \frac{1}{m}(y_{i-m+1} + \ldots y_i), i = m, \ldots, T. \tag{2.14}$$

The attractive feature of this quantile estimation is its simplicity and interpretability. In this paper, we extend the consistency properties proved in Zhou et al. (2010) for the non-linear cases.

The second method based on quantiles is more generally applicable. Heavy-tailed nature is common in many financial datasets and thus one cannot used a CLT based method depending on a apparently non-existent notion of long-run variance. We will be able to extend Zhou et al. (2010)'s result for non-linear processes in this case as well. For the demeaned process z_t, we assume the following decomposition

$$z_i = \sum_j a_j \epsilon_{i-j}, \tag{2.15}$$

where the i.i.d. innovations ϵ_j is allowed to have both light-tail i.e. $\mathbb{E}(|\epsilon_i|^2) < \infty$ or heavy-tailed i.e. $\alpha = \sup_{t>0}\{t : \mathbb{E}(|\epsilon_i|^t) < \infty\} < 2$. We will impose the following conditions on the coefficients for short or long range dependence and also assume boundedness of the density of ϵ in the following sense:

$$\begin{aligned}
(\text{SRD}) \quad &: \quad \sum_{j=0}^{\infty} |a_j| < \infty \\
(\text{DEN}) \quad &: \quad \sup_{x\in\mathbb{R}} f_\epsilon(x) + |f'_\epsilon(x)| < \infty \\
(\text{LRD}) \quad &: \quad a_j = j^{-\gamma} l(j) \text{ where } \gamma < 1, l(j) \text{ is slowly varying function (s. v. f.)}
\end{aligned} \tag{2.16}$$

where an s. v. f. is a function $g(x)$ such that $\lim_{x\to\infty} g(tx)/g(x) = 1$ for any t. The condition (SRD) is a classic short range dependent condition. See Box et al. (2015) for more discussion on this. (LRD) refers to the long-memory of the time series and it is satisfied by a large class of models such as ARFIMA. (DEN) is also a mild condition since by inversion theorem, all symmetric α-stable distribution falls under this. We borrow the following result from Zhou et al. (2010) for linear process. It is worth noting that one can extend this to non-linear processes as well by defining the coupling-based dependence on predictive density of ϵ_i as done in Zhang et al. (2015) but we postpone that discussion for a future paper.

2.1.4. Quantile consistency results for the quantile-based method

For a fixed $0 < q < 1$, let $\hat{Q}(q)$ and $\tilde{Q}(q)$ denote the q th sample quantile and actual quantile of $\tilde{y}_t$ from (2.4). We have the following different rates of convergence of quantiles based on the different cases:

Theorem 2.1 (Quantile consistency for linear processes)**.** *[Zhou et al. (2010) Theorem 1:4]*

- *Linear light tailed SRD: Suppose DEN holds and $\mathbb{E}(\epsilon_j^2) < \infty$. If SRD holds and $m^3/T \to 0$, then for any fixed $0 < q < 1$,*

$$|\hat{Q}(q) - \tilde{Q}(q)| = O_{\mathbb{P}}(m/\sqrt{T}). \tag{2.17}$$

- *Linear light tailed LRD: If LRD holds with γ and $l(\cdot)$ in (2.16) and $m^{5/2-\gamma}T^{1/2-\gamma}l^2(T) \to 0$, then, for any fixed $0 < q < 1$,*

$$|\hat{Q}(q) - \tilde{Q}(q)| = O_{\mathbb{P}}(mT^{1/2-\gamma}|l(T)|). \tag{2.18}$$

- *Linear heavy-tailed SRD: Suppose DEN holds and $\mathbb{E}(|\epsilon_j|^\alpha) < \infty$ for some $1 < \alpha < 2$. If SRD holds and $m = O(T^k)$ for some $k < (\alpha - 1)/(\alpha + 1)$. Then for any fixed $0 < q < 1$,*

$$|\hat{Q}(q) - \tilde{Q}(q)| = O_{\mathbb{P}}(mT^\nu) \text{ for all } \nu > 1/\alpha - 1. \tag{2.19}$$

—

- *Linear heavy-tailed LRD: If LRD holds with γ and $l(\cdot)$ in (2.16) and $m = O(T^k)$ for some $k < (\alpha\gamma - 1)/(2\alpha + 1 - \alpha\gamma)$, then, for any fixed $0 < q < 1$,*

$$|\hat{Q}(q) - \tilde{Q}(q)| = O_{\mathbb{P}}(mT^\nu) \text{ for all } \nu > 1/\alpha - \gamma. \tag{2.20}$$

2.2. Adjustments for the economic time series forecasting -

The simulation setup in (Zhou et al., 2010, page 1440) counts on T=6000 observations. This is hardly plausible for economic data. Hence, we adjust the methods of Zhou et al. (2010) according to this new economic forecasting set-up. We have relatively bigger prediction horizon i.e m might not be very small compared to T whereas T itself is often much smaller compared to what Zhou et al. (2010) had. Often theses series are observed for only 260 working days in an year and thus it is very important to modify our methods that can accommodate such a small sample size without losing its predictive performance. In this article we propose some novel modifications to both the methods mentioned above that seems to have good predictive performance based on our experience from both simulation studies and POOS evaluation in macroeconomic real data. We use stationary bootstrap (Politis and Romano, 1994) with optimal window width as proposed by Politis and White (2004), Patton et al. (2009) to compute the replicated series. Next kernel quantile estimators (see Silverman, 1986, Sheather and Marron, 1990) are used instead of simple sample quantiles. These two modifications, in our experience, betters the performance from those in Zhou et al. (2010).

2.2.1. Stationary bootstrap

For this part, we first demean the data and then resample from it using a stationary bootstrap. As proposed by Politis and Romano (1994) in their seminal paper, this method retains the stationarity of the distribution and is less sensitive to the choice of block size as in moving block bootstrap (Kunsch, 1989). It also retains the dependence structure asymptotically. Under mixing conditions, consistency of stationary bootstrap mean has been studied in literature. Gonçalves and de Jong (2003) shows under mixing conditions and some mild moment conditions

$$\sup_x |\mathbb{P}^*(\sqrt{T}(\bar{X}_T^* - \bar{X}_T) \leq x) - \mathbb{P}(\sqrt{T}(\bar{X}_T - \mu) \leq x)| = o_{\mathbb{P}}(c_T), \tag{2.21}$$

holds for some suitable $c_T \to 0$. It is easy to show similar results for only the average of m consecutive y_i under the assumption of linearity. That can in turn be used, along the same line of proof by Zhou et al. (2010) to show consistency results for the bootstrapped versions similar to those mentioned in Theorem 2.1. To keep our discussion concise and focused on empirical evaluations, we do not state them here. Interested readers can also look at the arguments by Sun and Lahiri (2006) for moving block bootstrap and the corresponding changes as suggested in Lahiri (2013) to get an idea how to show quantile consistency result. For time-series forecasting and quantile regression stationary bootstrap has been used by White (2000), Han et al. (2016) among others.

2.2.2. Kernel quantile estimation

The efficiency of kernel sample quantile estimators over the usual sample quantiles has been proved in Falk (1984) and was extended to several variants by Sheather and Marron (1990). As proposed in the latter, the improvement in MSE is of the order $\int uK(u)K^{-1}(u)du$ for the used symmetric kernel K. Also, the theorems mentioned in Section 2 are easily extendable to these kernel quantile estimators as well. One can use the Bahadur type representations for the kernel quantile estimators as shown in Falk et al. (1985) and obtain similar results of consistency. We do not discuss the proof in detail here.

2.3. Simulation Results

In order to assess their out-of-sample forecasting performance, we conduct a POOS comparison of coverage probabilities between original methods of Zhou et al. (2010) and the data-driven modifications proposed in 2.2. From large proportion, we adopt the simulation framework of Zhou et al. (2010) with following scenarios for z_t:

(i) $z_t = 0.6z_{t-1} + \sigma\epsilon_t$, with iid mixture-normal noise $\epsilon_t \sim \frac{1}{2}N(0,1) + \frac{1}{2}N(0,1.25)$,
(ii) $z_t = \sigma\sum_{j=0}^{\infty}(j+1)^{-0.8}\epsilon_{t-j}$, with noise as in (i),
(iii) $z_t = 0.6z_{t-1} + \sigma\epsilon_t$, with α-stable noise (with heavy-tail index =1.5 and scale parameter =1)
(iv) $z_t = \sigma\sum_{j=0}^{\infty}(j+1)^{-0.8}\epsilon_{t-j}$, with noise as in (iii).

These scenarios correspond to (i) light-tail and short memory (ii) light-tail and long-memory (iii) heavy-tail and short memory and (iv) heavy-tail and long memory DGP's. For each scenario, we generate pseudo-data of length $T+m$. We use the first T of them for estimation and last m for evaluation. The experiment is repeated $N_{\text{trials}} = 10000$ times for each scenario.

The choice of parameters[4] $T = 260, m = 20, 30, 40, 60, 90, 130$ and $\sigma = 1.31$ corresponds with the set-up for the real-data experiment in the following section. Following Müller and Watson (2016), we estimate the PI's for nominal coverage probabilities $100(1-\alpha) = 90\%$ and 67%. We compute the coverage probability $100\widehat{(1-\alpha)} = \frac{100}{N_{\text{trials}}}\sum_{i=1}^{N_{\text{trials}}}\mathbb{I}(\hat{\text{PI}}_i \ni \bar{z}_{i,T+1:T+m})$, where $\mathbb{I}$ is 1 when $\bar{z}_{i,T+1:T+m}$ is covered by the $\hat{\text{PI}}_i$ and 0 otherwise. Furthermore, we report the median PI width $\hat{w} = \text{median}(|\hat{\text{PI}}|_1, \ldots, |\hat{\text{PI}}|_{N_{\text{trials}}})$.

We start by summarizing the results for scenario (i). For short horizon, both original methods of Zhou et al. (2010) provide high coverage probabilities just below the nominal level. But as the horizon grows, coverage prob. of empirical method rapidly falls. e.g., for $m = 130$ only $\approx 48\%$ of the future averages fall between its bounds. Employing the kernel quantile estimates increases this number by 4pp which is still very low. On the other hand, the adjustment based on bootstrapping

[4]The value of σ was estimated by fitting an $AR(1)$ model to the full data set of SP 500 returns (see next section).

manages to keep relatively high coverage despite of being slightly below original empirical PI's for short horizons. In fact the improvement of coverage is by striking 26pp for $m = 130$. This gets even better if the latter two adjustments are combined. According to Zhou et al. (2010), the original *asp* PI's are asymptotically inferior, compared to the *emp* PI's. Yet we see that for short samples, former PI's work quite well even for the longest horizon. Nevertheless their coverage probability improves when accounting of the uncertainty about σ, i.e., using the t-quantiles instead of Gaussian Zhou et al. (used by 2010).

The second scenario introduces long memory into the competition. It indeed lowers the coverage probability across all the methods. However, most strikingly affected is the original empirical method for which the coverage probability decreases from 81% for $m = 20$ to only 37% for $m = 130$. The combined kernel-bootstrap adjustment increases the latter by 20pp. Despite of such improvement, the overall coverage is still very low. The performance of asymptotic methods is also negatively affected, but the t-quantile adjustment leads to at least modest improvement.

Employment of heavy-tailed noise in (iii) has particularly negative effect on the asymptotic PI's (coverage probability falls by 13pp) whereas empirical seem to be robust against such violation of assumptions (fall by 4-6pp).

Scenario (iv) combines the effect of (ii) and (iii). This causes fall of coverage probability for both original methods below 45%. Proposed adjustments are able to increase the coverage by at most 10pp.

2.4. Alternative methods

Bayes PI's based on low-frequency information (mw) Assume that the set of predictors for $\bar{y}_{T+1:T+m}$ consists of low-frequency cosine transformations $X = (X_1, \ldots, X_q)$ of y_t. The covariance matrix $\Sigma(S)$ of $(X_1, \ldots, X_q, \bar{y}_{T+1:T+m})$ is a function of (pseudo) spectra $S(m/T, q, \theta)$ (see Müller and Watson, 2016, eq. (20)). The parametrization $\theta = (b, c, d)$ is valid for broad family of processes. e.g., fractionally integrated with parameter $-0.4 \leq d \leq 1$, local- level effects with $b \geq 0$ and local-to-unity effects with $c \geq 0$. For fixed θ, a conditional distribution for $\bar{y}_{T+1:T+m}$ turns out to be Student-t with q degrees of freedom. In case y_t is $I(0)$ (i.e. $\theta = 0$) the PI's take the form

$$[L, U] = [\bar{y}_{1:T} + Q_q^t(\alpha/2)\sqrt{\frac{m+T}{mq}X\text{'}X}, \bar{y}_{1:T} + Q_q^t(1-\alpha/2)\sqrt{\frac{m+T}{mq}X\text{'}X}]. \tag{2.22}$$

As an upgrade of this *naive* PI (2.22) Müller and Watson (2016) suggest setting uniform prior on θ and they get *(Bayes)* PI. The robust Bayes version, denoted *(MN)*, is guaranteed to have uniform coverage across parameter space Θ.

PI's implied by ARFIMA-GARCH type models (afg) Univariate ARMA models and their extensions are often used as benchmark in POOS forecasting with economic time series (see Kim and Swanson, 2014, Stock and Watson, 2012). However, the implied PI's are criticized for being generally too narrow. This is mainly because the available analytic formulas for these PI's ignore various sources of uncertainty Chatfield (1993). Strictly speaking, the PI's for conditional mean have different target from PI's by Zhou et al. (2010), Müller and Watson (2016). Nevertheless, we take them into account, as we think that forecasting practitioners would use them. One advantage is that the PI's are already build into many software packages, thus easily available. In order to get the PI's for future averages, we can either (i) use averages of the in-sample observations y_t

		coverage probability $100(\widehat{1-\alpha})$						median length $\hat{w}$					
	horizon (days)	20	30	40	60	90	130	20	30	40	60	90	130
short-normal	emp-original	85.25	82.33	78.85	72.85	62.19	47.97	2.45	1.95	1.61	1.18	0.80	0.50
	emp-kernel	87.53	84.82	81.81	76.28	66.46	51.91	2.59	2.07	1.73	1.28	0.88	0.55
	emp-boot	82.56	80.84	80.45	79.19	76.70	74.70	2.24	1.84	1.59	1.31	1.07	0.89
	kernel-boot	85.82	84.28	83.81	82.47	80.38	78.06	2.43	2.00	1.73	1.42	1.16	0.97
	asp-original	85.05	83.50	82.52	80.89	78.80	76.44	2.37	1.93	1.67	1.37	1.12	0.93
	asp-tdist	86.11	84.03	83.50	82.17	79.97	77.51	2.42	1.97	1.71	1.39	1.14	0.95
long-normal	emp-original	80.88	75.96	71.72	63.45	51.20	37.31	3.86	3.31	2.89	2.24	1.58	0.97
	emp-kernel	83.45	79.06	74.99	67.89	56.40	41.39	4.08	3.53	3.11	2.47	1.76	1.09
	emp-boot	76.42	72.20	69.25	64.21	58.54	53.78	3.42	2.94	2.59	2.14	1.75	1.46
	kernel-boot	80.58	76.02	73.25	68.43	62.87	57.63	3.71	3.19	2.81	2.34	1.91	1.60
	asp-original	77.33	70.98	66.96	61.12	55.01	50.18	3.41	2.78	2.41	1.97	1.61	1.34
	asp-tdist	84.44	78.50	74.80	68.74	63.26	57.96	4.09	3.34	2.89	2.36	1.93	1.60
short-heavy	emp-original	84.39	80.21	76.71	70.12	58.54	44.45	7.08	5.51	4.51	3.30	2.28	1.42
	emp-kernel	85.93	82.15	79.13	73.43	63.37	48.51	7.27	5.75	4.79	3.60	2.55	1.58
	emp-boot	82.40	79.38	78.04	74.96	70.53	66.74	6.27	5.15	4.45	3.59	2.93	2.46
	kernel-boot	84.59	82.10	80.70	78.40	74.78	71.46	6.60	5.47	4.75	3.88	3.18	2.67
	asp-original	83.43	80.27	78.31	74.76	69.49	64.32	5.76	4.70	4.08	3.32	2.72	2.26
	asp-tdist	83.92	80.55	78.72	74.99	69.62	64.44	5.82	4.75	4.13	3.36	2.75	2.29
long-heavy	emp-original	80.64	76.27	71.27	62.87	49.82	33.62	11.03	9.39	8.19	6.38	4.58	2.84
	emp-kernel	82.53	78.51	74.32	66.80	55.13	37.49	11.47	9.92	8.78	7.01	5.21	3.23
	emp-boot	78.23	74.59	70.16	64.31	57.06	50.64	9.62	8.33	7.37	6.08	5.03	4.21
	kernel-boot	80.96	77.22	73.38	68.14	61.08	54.23	10.23	8.90	7.93	6.60	5.49	4.59
	asp-original	77.73	71.64	66.70	59.51	51.70	44.34	9.01	7.37	6.40	5.22	4.26	3.55
	asp-tdist	83.24	78.34	73.81	67.49	60.01	52.46	11.17	9.16	7.91	6.45	5.30	4.38

(A) Results of simulated forecasting experiment for nominal coverage probability $100(1-\alpha) = 90\%$.

		coverage probability $100(\widehat{1-\alpha})$						median length $\hat{w}$					
	horizon (days)	20	30	40	60	90	130	20	30	40	60	90	130
short-normal	emp-original	62.79	60.68	59.19	54.49	45.88	33.48	1.46	1.18	1.01	0.78	0.53	0.33
	emp-kernel	65.28	63.12	61.20	56.16	47.76	34.91	1.55	1.25	1.06	0.81	0.56	0.34
	emp-boot	58.82	57.74	56.81	55.63	53.22	50.49	1.33	1.09	0.95	0.78	0.64	0.53
	kernel-boot	62.48	61.55	60.63	59.12	56.89	54.13	1.44	1.18	1.03	0.85	0.69	0.57
	asp-original	61.52	59.88	58.93	57.21	55.14	52.14	1.40	1.14	0.99	0.81	0.66	0.55
	asp-tdist	61.81	60.53	59.47	57.61	55.47	52.29	1.41	1.15	1.00	0.81	0.66	0.55
long-normal	emp-original	57.98	56.14	53.68	48.54	38.94	27.71	2.35	2.07	1.89	1.57	1.12	0.69
	emp-kernel	60.30	58.05	55.88	49.70	40.39	28.75	2.47	2.18	1.97	1.62	1.16	0.71
	emp-boot	52.78	50.01	47.72	43.92	38.84	35.42	2.04	1.77	1.56	1.31	1.07	0.89
	kernel-boot	56.27	53.11	50.60	46.86	41.46	37.82	2.20	1.90	1.68	1.41	1.16	0.96
	asp-original	53.22	47.78	44.50	39.71	34.97	31.64	2.02	1.65	1.43	1.17	0.95	0.79
	asp-tdist	59.29	54.20	50.21	45.16	40.08	36.17	2.34	1.91	1.65	1.35	1.10	0.92
short-heavy	emp-original	62.74	60.31	59.95	54.78	44.39	31.04	3.34	2.90	2.79	2.17	1.54	0.93
	emp-kernel	65.36	63.16	61.99	56.50	46.58	32.43	3.55	3.09	2.88	2.27	1.62	0.98
	emp-boot	60.15	57.80	57.52	55.54	51.06	46.67	3.08	2.71	2.48	2.10	1.74	1.46
	kernel-boot	63.89	61.67	60.90	58.78	54.62	50.24	3.38	2.94	2.68	2.26	1.88	1.58
	asp-original	65.12	59.64	56.94	51.68	45.07	40.71	3.41	2.78	2.42	1.97	1.61	1.34
	asp-tdist	64.49	59.31	56.29	51.00	44.63	40.45	3.39	2.77	2.40	1.95	1.60	1.33
long-heavy	emp-original	59.02	55.99	54.33	48.08	37.93	24.88	5.91	5.48	5.24	4.44	3.27	2.02
	emp-kernel	61.52	58.50	56.04	49.65	39.27	25.96	6.22	5.79	5.42	4.57	3.40	2.10
	emp-boot	55.02	51.52	49.55	45.38	39.54	33.50	5.20	4.67	4.26	3.68	3.07	2.56
	kernel-boot	58.55	55.03	52.49	48.24	41.97	35.92	5.60	5.04	4.59	3.94	3.31	2.76
	asp-original	57.44	49.86	45.48	39.22	32.53	27.15	5.33	4.36	3.79	3.09	2.52	2.10
	asp-tdist	64.32	57.08	51.97	45.33	38.57	32.41	6.36	5.22	4.51	3.67	3.02	2.49

(B) Results of simulated forecasting experiment for nominal coverage probability $100(1-\alpha) = 67\%$.

Table 1: Comparison of methods as proposed by Zhou et al. (2010) with data-driven adjustments thereof. We simulate following time series: *short memory & normal tail, long memory & normal tail, short memory & heavy tail and long memory & heavy tail.* The reported values are coverage probability and median width of the respective PI's.

(ii) average the forecasts of y_t, over $t = T+1, \dots, T+m$. There are pros and cons about each approach regarding the implementation and effective use of our relatively short sample. Our search for temporal aggregation in the empirical literature (e.g., page 302 in Lütkepohl, 2006, Marcellino, 1999) did not lead us to any conclusion about superiority of (i) over (ii) or vice-versa. Therefore, we use both (i) and (ii) in the POOS exercise for sake of comparison. In both cases, we fit ARFIMA(p, d, q)-GARCH(P, Q) models to the (transformed) data with the rugarch R-package (see Ghalanos, 2015). The fractional parameter $d \in [0, 0.5]$ is either fixed to 0 (for spret) or estimated by maximum likelihood. The ARMA orders $p, q \in \{1, \dots, 4\}$ are selected by AIC. The GARCH orders are fixed $P, Q \in \{0, 1\}$. For sake of saving space, we report the best $100(\widehat{1-\alpha})$ and $\hat{w}$ among all alternative choices of d, P and Q.

3. Empirical pseudo-out-of-sample exercise with long financial time series

In this section, we conduct a statistical POOS comparison of following PI's:

(zxw) adjusted PI's of Zhou et al. (2010),
(mw) Bayesian PI's of Müller and Watson (2016),
(afg) PI's implied by ARFIMA-GARCH type models.

Data and set-up for POOS exercise Data on the univariate time series y_t are sampled at equidistant times $t = 1, \dots, T$. We are forecasting the average of future m values $\bar{y}_{T+1:T+m} = \sum_{t=1}^{m} y_{T+t}$. The POOS comparison is based on these time series[5]

(spret) SP 500 value weighted daily returns incl. dividends available from 1/2/1926 till 12/31/2014 with total of 23 535 observations,
(spret2) squared daily returns, with the same time span and
(tb3m) nominal interest rates for 3-month U.S. Treasury Bills available from 4/1/1954 till 8/13/2015 with total of 15 396 observations.

The sample size for post WWII quarterly macroeconomic time series is $4 \times 68 = 272$ observations. In order to get a similar set up in the POOS exercise, we use a rolling sample estimation scheme with sample size T=260 days, (i.e., one year of daily data), and forecasting horizon $m = 20, 30, 40, 60, 90$ and 130 days. The rolling samples are overlapping in last (resp. first) $T-m$ observations, so that e.g. for $m = 130$, the consecutive samples share half the observations. Hence for the returns time series and for $m = 130$ (resp. $m = 20$), we get $N_{\text{trials}} = 178$ (resp. 1163) non-overlapping in-or-out POOS trials. Same as in 2.3, we evaluate the coverage probability $100(\widehat{1-\alpha}) = \frac{100}{N_{\text{trials}}} \sum_{i=1}^{N_{\text{trials}}} \mathbb{I}(\hat{\mathrm{PI}}_i \ni \bar{y}_{i,T+1:T+m})$, for nominal coverage probabilities $100(1-\alpha) = 90\%$ and 67% and report the median PI width $\hat{w} = \text{median}(|\hat{\mathrm{PI}}|_1, \dots, |\hat{\mathrm{PI}}|_{N_{\text{trials}}})$ All models are selected and parameters estimated anew at each forecast origin. The methods are implemented as follows:
zxw: *Empirical quantile method (emp):*

1. Replicate mean adjusted series $z_t = y_t - \bar{y}_{1:T}$, $t = 1, \dots, T$ B times getting $z_t^b, t = 1, \dots, T$, $b = 1, \dots, B$.
2. Compute the series of overlapping rolling means $\bar{z}_t^b = \frac{1}{m} \sum_{i=1}^{m} z_{t-i}^b$, $t = m, \dots T$ from every replicated series.

[5] Replication files for all our results can be downloaded from http://homepage.univie.ac.at/marek.chudy/.

3. Estimate the $\alpha/2$th and $(1-\alpha/2)$th quantile $\hat{Q}(\alpha/2)$ and $\hat{Q}(1-\alpha/2)$ using kernel density estimator (see Silverman, 1986) from $\bar{z}_T^b$, $b = 1, \ldots, B$ with $T = 260$.
4. The PI for $\bar{y}_{T+1:T+m}$ is $[L, U] = [\bar{y}_{1:T} + \hat{Q}(\alpha/2), \bar{y}_{1:T} + \hat{Q}(1-\alpha/2)]$.

Asymptotic method based on quenched CLT approximation (asp):

1. Estimate the long-run standard deviation from z_t, $t = 1, \ldots, T$ using the sub-sampling estimator (2.2) with block-length proposed by Carlstein (1986).
2. The PI is given by $[L, U] = [\bar{y}_{1:T} + Q^t_{\kappa-1}(\alpha/2)\tilde{\sigma}/\sqrt{m}, \bar{y}_{1:T} + Q^t_{\kappa-1}(1-\alpha/2)\tilde{\sigma}/\sqrt{m}]$.

For these two methods, we consider *spret* to be $I(0)$, *spret2* to be fractionally integrated $I(d)$ with $d = 0.5$ (see Andersen et al., 2003) and *tb3m* is assumed to be $I(1)$. The methods, as shown in the previous simulation, work well specifically for $I(0)$. If convenient, we replace z_t by respective differences $dz_t = (1-L)^d z_t$, $t = 2, \ldots, T$, where L denotes lag operator (detailed steps are given in the online appendix).

mw: For simplicity, we give implementation steps for *Bayes* only. Further steps necessary for computing *MN* are in supplementary appendix.
Bayes PI's (Bayes):

1. Set q small and compute the cosine transformations $X = (X_1, , X_q)$ of the mean adjusted series z_t, $t = 1, \ldots, T$. Standardize them as $Z = (Z_1, , Z_q) = X/\sqrt{X`X}$.
2. For a grid of parameter values $\theta = (b, c, d)$ satisfying, $0.4 \leq d \leq 1$; $b, c \geq 0$, compute the matrix $\Sigma(\theta, q, m/T)$ using e.g. a numerical integration algorithm (for details see the supplementary appendix of Müller and Watson (2016)).
3. Choose a prior for $\theta = (b, c, d)$ and compute the posterior distribution.
4. Decompose the covariance matrix as $\Sigma = \left(\begin{smallmatrix}\Sigma_{ZZ} & \Sigma_{Z\bar{z}} \\ \Sigma_{Z\bar{z}} & \Sigma_{\bar{z}\bar{z}}\end{smallmatrix}\right)$ and obtain covariance matrix of residuals $\Sigma_{UU} = \Sigma_{\bar{z}\bar{z}} - \Sigma'_{Z\bar{z}}(\Sigma^{-1}_{ZZ})\Sigma_{Z\bar{z}}$.
5. Compute the quantiles $Q_q^{\mathrm{tmix}}(\alpha/2), Q_q^{\mathrm{tmix}}(1-\alpha/2)$ of the conditional distribution (mixture-t) of $\bar{z}_{T+1:T+m}$ using sequential bisection approximation.
6. The PI's are given by $[L, U] = [\bar{y}_{1:T} + Q_q^{\mathrm{tmix}}(\alpha/2)\sqrt{X`X}, \bar{y}_{1:T} + Q_q^{\mathrm{tmix}}(1-\alpha/2)\sqrt{X`X}]$.

The practical availability of the latter PI's is limited due to the advanced numerical approximations required for estimation. The authors provide pre-estimated inputs for the parameters $0.4 \leq d \leq 1$, $b, c \geq 0$, $q = 12$ and $0.075 \leq m/T \leq 1.5$ which we use in the POOS exercise.

afg: *Fitting model to averaged series (avg-series):*

1. Compute the series of overlapping rolling means $\bar{y}_t = \frac{1}{m}\sum_{i=0}^{m-1} y_{t-i}$, for $t = m, \ldots, T$.
2. Fit selected ARFIMA-GARCH to $\bar{y}_t$.
3. Obtain the PI's using

   (anal) forecasts for conditional mean $\bar{y}_{T,T+m} = \mathbb{E}[\bar{y}_{T+m}|\bar{y}_T, \bar{y}_{T-1}, \ldots]$ and prediction error sd $\mathrm{s}_{T,T+m}$. The PI is $[L, U] = [\hat{\bar{y}}_{T,T+m} + Q_t(\alpha/2)\hat{\mathrm{s}}_{T,T+m}, \hat{\bar{y}}_{T,T+m} + Q_t(1-\alpha/2)\hat{\mathrm{s}}_{T,T+m}]$.

   (boot) bootstrap for simulating $b = 1, \ldots, B$ future paths $\hat{\bar{y}}^b_{T,t}, t = T = 1, \ldots, T+m$, from the estimated model (see Ghalanos, 2015). The PI's are obtained as sample-quantiles of $\hat{\bar{y}}^b_{T,T+m}$ over $b = 1, \ldots, B$.

Fitting model to non-averaged series & averaging forecasts (avg-forecasts):

1. Fit selected ARFIMA-GARCH to $y_t, t = 1, \dots, T$.
2. Obtain the PI's using
   (anal) averaged forecast $\bar{\hat{y}}_{T,T+1:T+m} = \frac{1}{m}\sum_{i=1}^{m} \hat{y}_{T,T+i}$, with $y_{T,T+i} = \mathbb{E}[y_{T+i}|y_T, y_{T-1}, \dots]$ and respective prediction error sd[6] $\hat{\text{as}}_{T,T+1:T+m}$.
   The PI is $[L, U] = [\bar{\hat{y}}_{T,T+1:T+m} + Q_t(\alpha/2)\hat{\text{as}}_{T+m,\bar{f}}, \bar{\hat{y}}_{T,T+1:T+m} + Q_t(1-\alpha/2)\hat{\text{as}}_{T+m,\bar{f}}]$.

   (boot) bootstrap for simulating $b = 1, \dots, B$ future paths $\hat{y}^b_{T,t}, t = T+1, \dots, T+m$ from the estimated model. The PI's are obtained as quantiles from set of B averages $\bar{\hat{y}}^b_{T,T+1:T+m}$, $b = 1, \dots, B$.

POOS results: The results are summarized in Tables 2A and 2B. For the *spret* both *Bayes* and *MN* exceed the nominal coverage while both *zxw* PI's stay slightly below. Additionally, for *zxw* the coverage probability is decreasing as the horizon grows. For $m = 130$ the difference in coverage probability between best *mw* and *zxw* reaches 15 percent points (pp). However, the *zxw* provide shorter, hence more precise PI's. In fact, *MN* is almost twice the size of *emp* for $m = 130$. Comparing the two *afg* methods, we see that *avg-forecast* dominate over *avg-series*. Additionally, we see that bootstrap PI's have slightly higher coverage probability than analytic PI. Overall the most conservative PI's for *spret* are provided by *mw* followed by *zxw* and *afg*.

The first two methods switch their places for the *spret2* series. The *zxw* gives higher coverage. But the median width is also generally higher than for *mw*. This means that we pay for higher coverage probability with precision. For *anal* PI's the results have changed in favor of *avg-series* version. For *boot* PI's, however, there is no clear winner. Overall, among *afg* methods, the bootstrap is better than analytic forecasts at least for *spret2*. With the growing horizon all *afg* methods suffer significant fall in coverage probability caused by huge reduction of their width.

For the *tb3m*, we see *zxw* generally give higher coverage probability while their width is mostly below the width of *mw*. The coverage probability for *afg* falls much below the nominal level as the horizon grows. In contrary to the cases above, the *anal* PI's dominate the *boot* PI's, still, their coverage probability for $m = 130$ reaches only half of the nominal coverage.

Overall comparison of the two *zxw* methods corroborates the results obtained by simulations in Zhou et al. (2010). With the exception of *spret asp* gives slightly higher coverage probability than *emp*. Yet, the latter has a significant advantage in terms of width which is why we give preference to the *emp* method over the *asp* in the following predictions for the economic time series.

4. Prediction intervals for eight macroeconomic time series and SP 500 returns

Finally, we provide the long-run PI's for eight quarterly post WWII US macro-economic time series and quarterly returns. Thus provide alternative to PI's in Table 5 in Müller and Watson (2016, pages 1731-1732). The eight time series are: real per capita GDP, real per capita consumption expenditures, total factor productivity, labor productivity, population, prices (PCE[7]), inflation (CPI[8]) and Japanese inflation (CPI). The data are available from 1Q-1947 till 4Q-2014 and we forecast them over next $m = 10, 25$ and 50 years. For a subset of these series, we report results based on longer (yearly) sample starting in 1Q-1920. We also add the horizon $m = 75$ years for

[6] The formula for the prediction error sd of averages can be found in the supplementary appendix.
[7] Personal consumption expenditure deflator.
[8] Consumer price index.

this subset. The plots for all series used in the current and previous section are presented in the supplementary appendix.

The *zxw* methods perform well in the POOS exercise. Tables 3 and 4 in the current paper therefore provide valid alternative to Bayes PI's of(Müller and Watson, 2016). In the latter paper, authors compare these PI's to predictions provided by CBO. They come to the conclusion that the similarity between the PI's for series such as GDP is due to combination of (i) CBOs ignorance for parameter uncertainty and (ii) CBOs ignorance of possible anti-persistence of GDP during Great moderation. As (i) and (ii) have rather opposite effects on the PI's, they eventually seem to cancel out.

For *per capita real GDP*, *per capita consumption* and *productivity*, *emp* estimates are wider than in Müller and Watson (2016), especially for GDP. Wide PI's are often considered as failure of the forecasting method or model. On the other hand, it can also reflect the higher uncertainty about the series future. The width of PI's for GDP is not surprising given that similar as CBO, we also do not account for the possible anti-persistence during the Great moderation. With the longer yearly sample our PI's get wider, as a result of higher volatility in early 20th century. Interestingly, the growth in Labor production seems to be higher in general than according to *mw*.

Consumption, population, inflation and *prices* are known to be persistent, therefore, we would expect that, similarly as in case of interest rates, *emp* could give better coverage and possibly narrower PI's. Growth of population and prices as well as inflation have approx. same amount of uncertainty according to both our *emp* and *mw*. The location of the series is generally lower according to *emp*, especially for inflation, where the shift is about -2pp compared to *mw*.

Finally, for the *quarterly returns*, we might expect *emp* to give less conservative thus shorter estimates, and we see this happening with discrepancy growing along with horizons. Yet, it is clear that *mw* is especially conservative in uncertainty about positive returns, where differences from *emp* reach 11pp. Employing the longer time series makes the difference fall to 3pp.

5. Discussion

We have considered problem of constructing empirically valid prediction intervals for univariate time series. In an extensive POOS forecasting exercise, we have shown that the methods previously suggested by Zhou et al. (2010) can compete with sophisticated alternative prediction intervals designed specifically for economic framework Müller and Watson (see 2016). However, the methods need to be adjusted under short-sample constraint, which is common in practice. Based on the comparison results, we provided alternative and possibly accurate long-run prediction intervals for leading macroeconomic indicators.

Comparison of the empirical quantile method in regression set-up with the quantile (auto-) regression (Koenker, 2005) was beyond the scope of this paper and should be considered in the future. In addition, employment of statistical tests for performance comparison (Clements and Taylor, 2003, Gneiting and Raftery, 2007, Weron and Misiorek, 2008) should be considered in the future.

From the theoretical perspective, extension of Zhou et al. (2010) into high-dimensional regression framework by utilizing e.g. LASSO estimator is already being developed Karmakar et al. (2018). Even more challenging is case of multivariate target series and subsequent construction of simultaneous prediction intervals which can have interesting implications for market trading strategies. (Han et al. (2018))

	horizon (days)	coverage probability $100(\widehat{1-\alpha})$ 20	30	40	60	90	130	median length $\hat{w}$ 20	30	40	60	90	130
SP 500 returns	emp	89.94	89.16	88.30	88.89	87.60	81.56	0.67	0.56	0.48	0.39	0.32	0.26
	asp	87.70	87.35	85.89	85.53	83.72	78.21	0.62	0.51	0.44	0.36	0.29	0.24
	mn	92.00	94.58	94.32	94.83	96.90	96.09	0.82	0.81	0.64	0.65	0.58	0.49
	bayes	87.96	92.39	90.71	93.28	92.64	93.30	0.71	0.71	0.54	0.54	0.47	0.41
	i0	86.07	90.19	88.47	92.25	92.25	84.92	0.64	0.64	0.47	0.47	0.40	0.34
	avg-series-anal	83.32	80.39	81.41	77.78	68.22	68.72	0.68	0.53	0.45	0.34	0.25	0.20
	avg-series-boot	82.46	80.26	81.76	72.09	67.44	62.57	0.66	0.52	0.44	0.34	0.26	0.20
	avg-4cast-anal	85.21	84.52	84.51	81.14	79.84	75.98	0.61	0.50	0.44	0.36	0.29	0.24
	avg-4cast-boot	85.12	85.16	83.30	81.40	79.84	73.74	0.60	0.49	0.43	0.35	0.28	0.23
SP 500 returns^2	emp	92.61	91.87	91.05	90.44	90.31	87.15	1.84	1.72	1.73	1.68	1.70	1.68
	asp	91.49	93.03	92.08	91.99	91.86	89.39	2.01	2.00	2.01	1.99	2.04	2.00
	mn	89.51	89.42	87.78	87.60	87.21	86.03	1.61	1.62	1.43	1.45	1.40	1.40
	bayes	88.39	88.26	85.20	86.30	84.50	83.24	1.48	1.49	1.31	1.35	1.35	1.28
	i0	87.79	88.77	81.76	83.46	78.29	68.16	1.41	1.40	1.04	1.04	0.90	0.72
	avg-series-anal	72.74	70.45	66.95	59.59	57.75	44.13	0.90	0.82	0.70	0.60	0.56	0.46
	avg-series-boot	78.66	77.18	70.78	65.08	58.67	51.67	1.23	0.90	0.82	0.71	0.65	0.48
	avg-4cast-anal	41.44	33.55	28.92	23.77	16.67	13.97	0.45	0.38	0.34	0.29	0.27	0.23
	avg-4cast-boot	75.00	75.00	70.89	67.00	62.69	52.27	1.01	0.82	0.73	0.70	0.62	0.44
TB3M interest rate	emp	89.29	88.69	85.45	85.32	82.74	75.86	0.48	0.58	0.67	0.81	1.01	1.18
	asp	92.33	92.46	89.95	89.68	86.31	81.90	0.54	0.65	0.75	0.91	1.11	1.34
	mn	90.34	85.91	86.24	80.16	73.21	72.41	0.77	0.77	0.92	0.92	1.10	1.24
	bayes	90.34	85.52	85.45	79.37	73.21	72.41	0.76	0.77	0.91	0.92	1.08	1.22
	i0	64.42	61.11	45.50	42.86	32.14	25.00	1.44	1.42	1.07	1.07	0.86	0.70
	avg-series-anal	83.07	80.95	71.69	61.11	52.98	36.21	0.48	0.59	0.64	0.74	0.90	0.95
	avg-series-boot	42.46	41.67	41.80	47.22	45.83	45.69	0.81	0.97	0.95	1.28	1.42	1.64
	avg-4cast-anal	84.13	81.94	79.10	75.40	76.19	68.10	0.45	0.56	0.65	0.82	1.04	1.33
	avg-4cast-boot	63.93	68.38	60.49	64.41	57.89	47.06	0.52	0.61	0.68	0.87	0.94	0.93

(A) Results of POOS forecasting experiment for nominal coverage probability $100(1-\alpha) = 90\%$.

	horizon (days)	coverage probability $100(\widehat{1-\alpha})$ 20	30	40	60	90	130	median length $\hat{w}$ 20	30	40	60	90	1
SP 500 returns	emp	67.50	68.52	67.30	66.67	63.95	61.45	0.39	0.33	0.28	0.23	0.19	0.
	asp	65.43	65.03	64.37	61.50	58.91	60.34	0.37	0.30	0.26	0.21	0.17	0.
	mn	71.28	79.61	80.21	84.75	84.11	82.12	0.45	0.45	0.39	0.40	0.36	0.
	bayes	66.38	74.32	68.50	73.13	75.19	72.63	0.40	0.40	0.30	0.30	0.25	0.
	i0	63.89	72.65	64.20	70.03	70.93	68.16	0.36	0.36	0.27	0.27	0.23	0.
	avg-series-anal	56.92	55.74	56.80	53.49	49.61	45.25	0.37	0.29	0.25	0.19	0.14	0.
	avg-series-boot	59.07	55.61	56.28	51.16	46.90	38.55	0.39	0.31	0.25	0.20	0.15	0.
	avg-4cast-anal	60.19	59.74	58.69	56.59	52.33	58.10	0.32	0.27	0.23	0.19	0.16	0.
	avg-4cast-boot	62.51	61.03	61.45	57.11	53.49	57.54	0.34	0.28	0.25	0.20	0.16	0.
SP 500 returns^2	emp	76.61	78.45	78.31	79.33	78.68	77.65	0.87	0.90	0.92	0.91	0.95	0.
	asp	81.77	81.29	82.44	82.43	82.17	79.89	1.14	1.12	1.13	1.11	1.14	1.
	mn	75.84	77.03	74.35	74.94	71.32	66.48	0.89	0.91	0.84	0.85	0.84	0.
	bayes	73.95	74.58	70.91	70.28	67.44	62.57	0.84	0.84	0.74	0.75	0.70	0.
	i0	71.97	73.42	57.31	60.98	51.94	44.69	0.80	0.80	0.59	0.59	0.52	0.
	avg-series-anal	45.06	41.55	39.41	36.01	33.72	25.14	0.40	0.37	0.31	0.27	0.25	0.
	avg-series-boot	58.84	55.19	51.95	42.06	36.00	31.67	0.65	0.49	0.48	0.43	0.37	0.
	avg-4cast-anal	21.32	14.97	12.74	8.79	6.59	6.15	0.20	0.17	0.15	0.13	0.12	0.
	avg-4cast-boot	50.31	46.82	48.73	45.00	32.84	31.82	0.52	0.44	0.40	0.41	0.37	0.
TB3M interest rate	emp	69.18	67.86	66.93	61.11	57.14	59.48	0.27	0.33	0.38	0.47	0.59	0.
	asp	76.32	74.40	71.16	67.86	63.10	62.93	0.32	0.38	0.44	0.54	0.66	0.
	mn	71.83	65.87	67.46	56.75	50.00	56.03	0.43	0.44	0.52	0.53	0.62	0.
	bayes	71.56	65.67	66.93	56.75	50.00	56.03	0.43	0.44	0.52	0.53	0.62	0.
	i0	34.39	32.54	22.49	18.65	18.45	13.79	0.82	0.81	0.61	0.61	0.49	0.
	avg-series-anal	55.82	56.55	49.47	42.06	28.57	20.69	0.25	0.31	0.34	0.39	0.46	0.
	avg-series-boot	37.17	35.12	34.66	40.48	35.12	36.21	0.48	0.56	0.56	0.76	0.84	1.
	avg-4cast-anal	57.67	59.13	55.03	55.16	51.19	51.72	0.23	0.28	0.33	0.41	0.55	0.
	avg-4cast-boot	48.09	50.43	49.38	49.15	28.95	29.41	0.28	0.34	0.35	0.52	0.54	0.

(B) Results of POOS forecasting experiment for nominal coverage probability $100(1-\alpha) = 67\%$.

Table 2: Comparison of *zxw*, *mw* and *afg* on each of the three time series: *spret, spret2, tb3m*. The reported values are coverage probability and median width of the respective PI's.

	nominal coverage $100(1-\alpha) = 67\%$			nominal coverage $100(1-\alpha) = 90\%$		
horizon (years)	10	25	50	10	25	50
GDP/Pop	[-0.88 , 4.65]	[-1.11 , 4.76]	[-0.87 , 4.68]	[-2.97 , 6.78]	[-2.96 , 6.59]	[-2.86 , 6.48]
Cons/Pop	[0.56 , 3.45]	[0.60 , 3.45]	[0.59 , 3.49]	[-0.54 , 4.53]	[-0.43 , 4.38]	[-0.40 , 4.55]
TF prod.	[-0.46 , 2.92]	[-0.37 , 2.96]	[-0.40 , 2.76]	[-1.61 , 4.16]	[-1.55 , 4.02]	[-1.49 , 3.83]
Labour prod.	[0.89 , 3.42]	[0.84 , 3.24]	[0.90 , 3.37]	[-0.11 , 4.35]	[0.06 , 4.11]	[0.08 , 4.15]
Population	[0.44 , 0.95]	[0.25 , 1.00]	[-0.11 , 0.90]	[0.24 , 1.17]	[-0.06 , 1.35]	[-0.50 , 1.35]
PCE infl.	[-4.06 , 2.32]	[-6.01 , 3.83]	[-9.50 , 5.15]	[-7.39 , 4.70]	[-9.74 , 7.40]	[-14.38 , 9.86]
CPI infl.	[-4.75 , 1.61]	[-6.32 , 2.04]	[-9.43 , 3.29]	[-9.00 , 4.03]	[-10.54 , 6.57]	[-14.95 , 7.46]
Jap. CPI infl.	[-5.20 , 2.79]	[-7.12 , 4.18]	[-8.72 , 5.85]	[-8.10 , 7.63]	[-11.38 , 10.07]	[-14.51 , 12.19]
Returns	[2.20 , 12.20]	[3.50 , 10.75]	[4.78 , 10.22]	[-1.93 , 15.69]	[0.88 , 12.95]	[2.90 , 11.71]

Table 3

Prediction intervals for long-run averages of quarterly post WWII growth rates. The results for 8 macroeconomic time series and quarterly SP 500 returns provide alternative to Table 5 of Müller and Watson (2016). Latter paper reports PI's for horizon $m = 75$ years also for these short post WWII quarterly series. However, as this horizon would exceed the sample size, we cant provide the emp as alternative. Hence, we present results for this horizon only for the longer annual data in the next table.

	horizon (years)	10	25	50	75
67%	GDP/Pop	[-1.43 , 5.61]	[-1.59 , 5.68]	[-1.85 , 5.65]	[-1.72 , 5.36]
	Cons/Pop	[-1.07 , 4.27]	[-1.15 , 4.41]	[-0.96 , 4.33]	[-1.08 , 4.26]
	Population	[0.33 , 0.99]	[0.08 , 1.11]	[-0.21 , 1.16]	[-0.54 , 1.15]
	CPI infl.	[-2.72 , 6.02]	[-2.80 , 6.21]	[-3.19 , 6.69]	[-5.27 , 9.46]
	Returns	[0.38 , 13.61]	[3.74 , 10.68]	[3.60 , 9.67]	[4.44 , 8.29]
90%	GDP/Pop	[-5.00 , 8.44]	[-4.30 , 8.47]	[-4.92 , 8.24]	[-4.49 , 7.96]
	Cons/Pop	[-3.12 , 6.21]	[-3.03 , 6.22]	[-2.80 , 6.03]	[-2.90 , 6.27]
	Population	[0.13 , 1.23]	[-0.24 , 1.51]	[-0.63 , 1.74]	[-1.13 , 1.81]
	CPI infl.	[-6.02 , 12.65]	[-9.00 , 12.13]	[-8.13 , 12.87]	[-11.26 , 16.13]
	Returns	[-3.64 , 17.50]	[0.45 , 12.62]	[1.61 , 11.77]	[2.82 , 9.49]

Table 4

Prediction intervals for long-run averages of annual growth rates and annual SP 500 returns.

References

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71*(2), 579–625.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*, 1133–1150.

Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics 73*, 5–59.

Bansal, R., D. Kiku, and A. Yaron (2016). Risks for the long run: Estimation with time aggregation. *Journal of Monetary Economics 82*, 52–69.

Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics 14*, 1171–1179.

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

Cheng, X., Z. Liao, and F. Schorfheide (2016). Shrinkage estimation of high-dimensional factor models with structural instabilities. *The Review of Economic Studies 83*(4), 1511–1543.

Christoffersen, P. F. and F. X. Diebold (1998). Cointegration and long-horizon forecasting. *Journal of Business & Economic Statistics 16*, 450–458.

Clements, M. P. and J. H. Kim (2007). Bootstrap prediction intervals for autoregressive time series. *Computational Statistics & Data Analysis 51*, 3580–3594.

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

Clements, M. P. and N. Taylor (2003). Evaluating interval forecasts of high-frequency financial data. *Applied Econometrics 18*, 445–456.

Dehling, H., R. Fried, O. Shapirov, D. Vogel, and M. Wornowizki (2013). Estimation of the variance of partial sums of dependent processes. *Statistics & Probability Letters 83*(1), 141–147.

Diebold, F. X. and P. Linder (1996). Fractional integration and interval prediction. *Economic Letters 50*, 305–313.

Diebold, F. X. and G. D. Rudebusch (1989). Long memory and persistence in aggregate output. *Journal of Monetary Economics 24*, 189–209.

Elliott, G., A. Gargano, and A. Timmermann (2013). Complete subset regressions. *Journal of Econometrics 177*(2), 357–373.

Elliott, G., U. K. Mller, and M. W. Watson (2015). Nearly optimal tests when a nuisance parameter is present under the null hypothesis. *Econometrica 83*(2), 771–811.

Engle, R. F. (1974). Band spectrum regression. *International Economic Review 15*, 1–11.

Falk, M. (1984). Relative deficiency of kernel type estimators of quantiles. *The Annals of Statistics*, 261–268.

Falk, M. et al. (1985). Asymptotic normality of the kernel quantile estimator. *The Annals of Statistics 13*(1), 428–433.

Ghalanos, A. (2015). rugarch: Univariate garch models. 1 1. *Accessed 2015.*

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Gonçalves, S. and R. de Jong (2003). Consistency of the stationary bootstrap under weak moment conditions. *Economics Letters 81*(2), 273–278.

Han, H., O. Linton, T. Oka, and Y.-J. Whang (2016). The cross-quantilogram: measuring quantile dependence and testing directional predictability between time series. *Journal of Econometrics 193*(1), 251–270.

Han, Y., M. Chudy, and W. B. Wu (2018+). Long term prediction for high-dimensional time series. *preprint*.

Karmakar, S., M. Chudy, and W. B. Wu (2018+). Long term prediction in high-dimensional regression. *preprint*.

Kim, H. and N. Swanson (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics 178*, 352–367.

Kitsul, Y. and J. Wright (2013). The economics of options-implied inflation probability density functions. *Journal of Financial Economics 110*, 696–711.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, 1217–1241.

Lahiri, S. N. (2013). *Resampling methods for dependent data.* Springer Science & Business Media.

Lütkepohl, H. (2006). Forecasting with varma models. vol. 1. In *Handbook of Economic Forecasting*, pp. 287–325. edited by G. Granger, C.W.J. and Timmermann, A. Elliott,. Elsevier B.V: by -.

Marcellino, M. (1999). Some consequences of temporal aggregation in empirical analysis. *Journal of Business & Economic Statistics 17*(1), 129–136.

Müller, U. and M. Watson (2016). Measuring uncertainty about long-run predictions. *Review of Economic Studies 83*(4), 1711–1740.

Pastor, L. and R. F. Stambaugh (2012). Are stocks really less volatile in the long run. *Journal of Finance 67*(2), 431–478.

Patton, A., D. N. Politis, and H. White (2009). Correction to "automatic block-length selection for the dependent bootstrap" by d. politis and h. white. *Econometric Reviews 28*(4), 372–375.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical Association 89*, 1303–1313.

Politis, D. N. and H. White (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews 23*(1), 53–70.

Reschenhofer, E. and M. Chudy (2015). Adjusting band-regression estimators for prediction: Shrinkage and downweighting. *International Journal of Econometrics and Financial Management 3*(3), 121–130.

Sheather, S. J. and J. S. Marron (1990). Kernel quantile estimators. *Journal of the American Statistical Association 85*, 410–416.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall/CRC.

Stock, J. and M. Watson (2012, October). Generalised shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics 30*(4), 482–493.

Stock, J. H. and M. W. Watson (2005). Understanding changes in international business cycle dynamics. *Journal of the European Economic Association 3*, 968–1006.

Sun, S. and S. N. Lahiri (2006). Bootstrapping the sample quantile of a weakly dependent sequence. *Sankhyā: The Indian Journal of Statistics*, 130–166.

Weron, R. and A. Misiorek (2008). Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *Accessed 9*(2), 2017.

White, H. (2000). A reality check for data snooping. *Econometrica 68*(5), 1097–1126.

Wu, W. B. and M. Woodroofe (2004). Martingale approximations for sums of stationary processes. *Ann. Probab. 32*(2), 1674–1690.

Zhang, T., W. B. Wu, et al. (2015). Time-varying nonlinear regression models: Nonparametric estimation and model selection. *The Annals of Statistics 43*(2), 741–768.

Zhou, Z., Z. Xu, and W. B. Wu (2010). Long-term prediction intervals of time series. *IEEE Trans. Inform. Theory 56*(3), 1436–1446.

Appendix A - Figures for time series used in Section 3

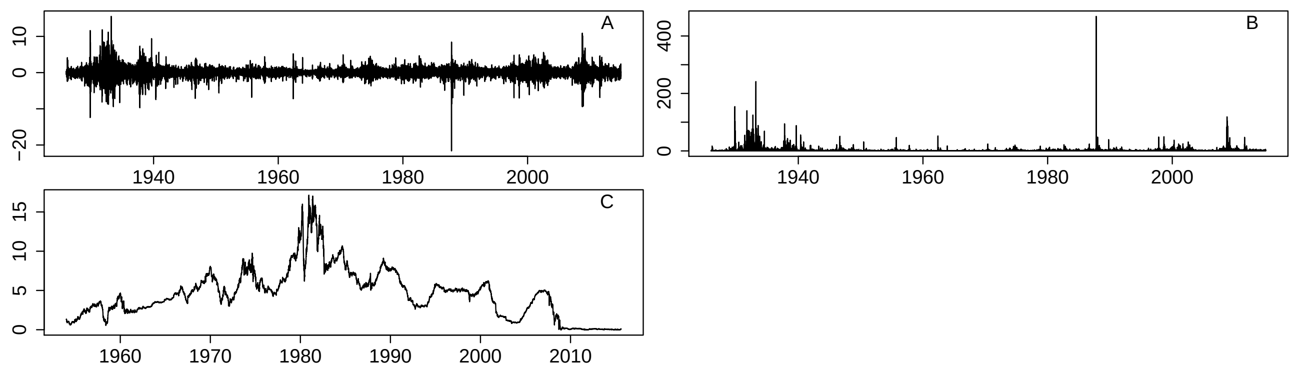

Fig 1: Daily time series: A) SP 500 value weighted daily returns incl. dividend, B) squared returns, C) nominal interest rates for 3-month U.S. Treasury Bills.

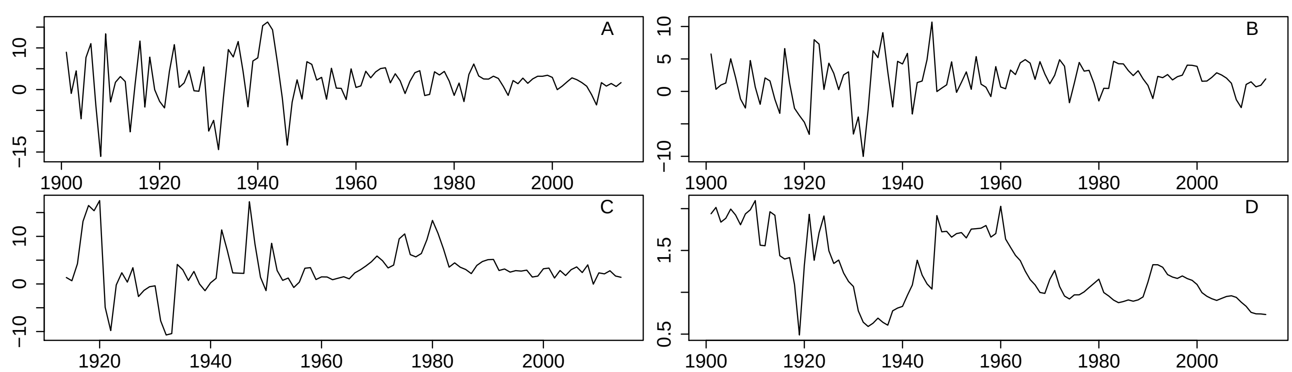

Fig 2: Annual time series - growth rates: A) real per capita GDP, B) real per capita consumption expenditures, C) Inflation (CPI), D) population.

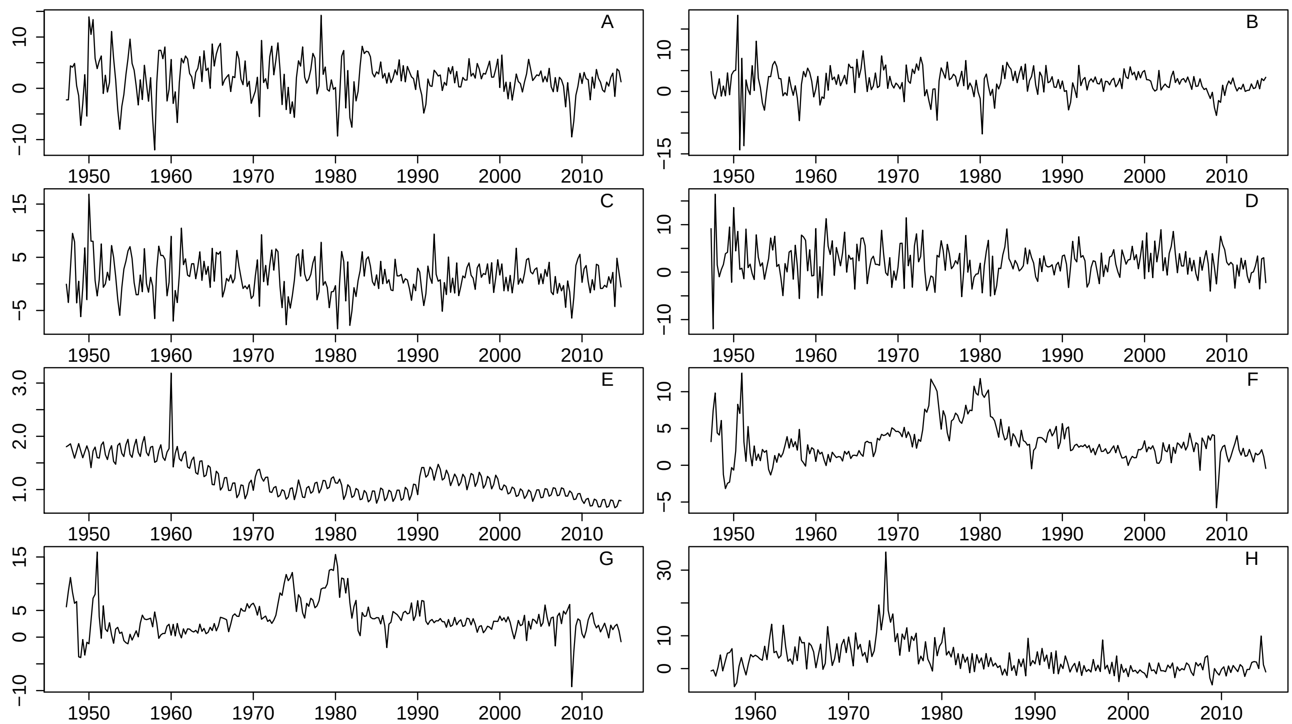

Fig 3: Quarterly time series - growth rates: A) real per capita GDP, B) real per capita consumption expenditures, C) total factor productivity, D) labor productivity, E) population, F) prices (PCE), G) Inflation (CPI), H) Japanese Inflation.

Appendix B - Additional steps for implementation of zxw and mw

All macroeconomic time series considered in section 3 are log-differences. As pointed out by Müller and Watson (2016), this can not rule out the presence of long memory dynamics or even a unit root (e.g. for population of inflation, we can still observe significant autocorrelations in the series for large lags). Note that for the interest rates series, we assume a deterministic trend component rather than a constant level, hence the respective PIs formula, the location is $\frac{m+1}{2}\overline{\Delta y}_{1:T}$ instead of $\bar{y}_{1:T}$).

emp:

1. compute the mean adjusted series $z_t = y_t - \bar{y}_{1:T}$, $t = 1, \ldots, T$.
2. Fix $d = 0.5$ or $d = 1$ and compute the difference series $dz_t = (1-L)^d$, $t = 2, \ldots, T$, where L denotes lag operator.
3. Replicate $dz_t, t = 2, \ldots, T$ B times getting $dz_t^b, t = 2, \ldots, T$, $b = 1, \ldots, B$.
4. Compute the series of overlapping rolling means $\bar{dz}_t^b = \frac{1}{m}\sum_{i=1}^m dz_{t-i}^b$, $t = m, \ldots T$ from every replicated series.
5. Estimate quantiles $\hat{Q}(\alpha/2)$ and $\hat{Q}(1-\alpha/2)$ from $\bar{z}_T^b$, $b = 1, \ldots, B$ with $T = 260$ obtained as $\bar{z}_T^b = \frac{1}{m}\sum_{i=1}^m (1-L)^{-d} dz_{T-i}^b$.
6. The PI is given by $[L, U] = [\bar{y}_{1:T} + Q_{\kappa-1}^t(\alpha/2)\tilde{\sigma}/\sqrt{m}, \bar{y}_{1:T} + Q_{\kappa-1}^t(1-\alpha/2)\tilde{\sigma}/\sqrt{m}]$.

asp:

1. compute the mean adjusted series $z_t = y_t - \bar{y}_{1:T}$, $t = 1, \ldots, T$.
2. Fix $d = 0.5$ or $d = 1$ and compute the difference series $dz_t = (1-L)^d$, $t = 2, \ldots, T$, where L denotes lag operator.
3. Estimate the long-run standard deviation $\tilde{\sigma}$ of dz_t, $t = 2, \ldots, T$.
4. Compute the long-run standard deviation of z_t: for $d = 1$, $\sigma_z(\tilde{\sigma}) = \tilde{\sigma}\sqrt{(m+1)/2}$ and for[9] $d = 0.5$, $\sigma_z(\tilde{\sigma}) = \frac{\tilde{\sigma}}{m}\sqrt{\sum_{i=1}^m (\sum_{j=0}^{m-i}(-1)^j\binom{-0.5}{j})^2}$.
5. The PI is given by $[\bar{y}_{1:T} + Q_{\kappa-1}^t(\alpha/2)\sigma_z, \bar{y}_{1:T} + Q_{\kappa-1}^t(1-\alpha/2)\sigma_z]$.

MN:

compute steps 1-4.

5.1 Compute weights for specific choice of q and m/T and the prior from step 3.

5.2 Numerically approximate s. c. least favorable distribution (LFD) of θ for specific choice of q and m/T (see the supplementary appendix of Müller and Watson (2016)).

5.3 Using the weights and the LFD solve the minimization problem (14) on page 1721 in Müller and Watson (2016) to get quantiles which give uniform coverage and minimize the expected PIs width.

6. same as in *Bayes* with the robust quantiles.

[9] see http://mathworld.wolfram.com/BinomialSeries.html