Long-term prediction intervals with many covariates

Sayar Karmakar*, Marek Chudý, and Wei Biao Wu

University of Florida, University of Vienna, University of Chicago

September 30, 2021

Abstract

Accurate forecasting is one of the fundamental focuses in the literature of econometric time-series. Often practitioners and policymakers want to predict outcomes of an entire time horizon in the future instead of just a single k-step ahead prediction. These series, apart from their own possible nonlinear dependence, are often also influenced by many external predictors. In this paper, we construct prediction intervals of time-aggregated forecasts in a high-dimensional regression setting. Our approach is based on quantiles of residuals obtained by the popular LASSO routine. We allow for general heavy-tailed, long-memory, and nonlinear stationary error processes and stochastic predictors. Through a series of systematically arranged consistency results, we provide theoretical guarantees of our proposed quantile-based method in all of these scenarios. After validating our approach using simulations we also propose a novel bootstrap-based method that can boost the coverage of the theoretical intervals. Finally analyzing the EPEX Spot data, we construct prediction intervals for hourly electricity prices over horizons spanning 17 weeks and contrast them to selected Bayesian and bootstrap interval forecasts.

Keywords: Forecasting, Heavy-tailed distribution, Long-range dependence, Electricity prices, Bootstrap, Time-aggregation

*Corresponding author, email: sayarkarmakar@ufl.edu, Address: 230 Newell Drive, Gainesville, FL- 32601, USA. Phone: +1(352)273-1839.

1   Introduction

Prediction intervals (PI hereafter) help forecasters to access the uncertainty concerning the future values of time series. This has been a central topic of interest in the analysis of time series from both theoretical and algorithmic point of view. The k-step ahead prediction of a time-series after observing $X_1, \ldots, X_t$ has been discussed from many different perspectives. While estimation of the conditional mean $\mathbb{E}(X_{n+k}|X_1, \ldots, X_n)$ has received most of the attention since Cover (1975); other approaches were also developed significantly in the last few decades, see Gyorfi and Ottucsak (2007), Morvai (2003), Gyorfi et al. (1998), Morvai et al. (1996), Belyaev (1968) Morvai et al. (1997), Brockwell et al. (1991), Györfi et al. (2013), Pourahmadi (2001). Nonetheless in the eye of an applied scientist, evaluation of prediction intervals (PIs henceforth) is challenging (e.g., Chatfield, 1993; Clements and Taylor, 2003), since a higher empirical coverage probability often comes with the cost of larger width, thus less precision. In this paper, we focus on the construction and evaluation of PIs for a single univariate series using information from many predictors under a possibly high-dimensional linear model with a general error process. Moreover, the target for which we construct the PIs is not a single future value of the univariate series, but a sum of its values over the entire forecasting horizon. This setup has practical applications in telecommunication and data services, energy, and finance. The characteristic feature of these sectors is that data about sales, prices, or returns are collected with hourly frequency while the management is interested in predictions of aggregated volumes over horizon spanning several weeks or months. In the volatility forecasting of stock returns using popular ARCH/GARCH models, often aggregated (over time) mean squared error (AMSE) is an evaluation metric (See (Starica, 2003; Fryzlewicz et al., 2008; Karmakar et al., 2020) etc.). When it comes to predicting energy or electricity consumption, policy makers want to predict future usages for entire one month or several months ahead. The aggregated predictions help them in setting the price of such commodities. See some of the applications mentioned in Zhou et al. (2010). Time-aggregation is also useful to arrive at strategic decisions made by trust funds, pension management and insurance companies, portfolio management of specific derivatives (Kitsul and Wright, 2013) and assets (see Bansal et al., 2016) among others.

Towards including external regressors, Ludwig et al. (2015) found that the inclusion of disaggregated wind speed and temperature (measured at more than 70 weather stations across Germany) leads to improvements of forecasts for EPEX SPOT hourly day-ahead electricity prices.[1]. We adapt their framework to allow for long-

---

[1]In electricity price forecasting (EPF), one has to control for weather conditions, local economy,

2

run dependence which is important for long-horizon and medium-horizon forecasting. Similar applications can be found in in power portfolio risk management, derivatives pricing, medium-term and long-term contract evaluation, and maintenance scheduling. To challenge our forecasts, we provide out-of-sample comparison with several forecasting methods, which include Bayes PIs of Müller and Watson (2016) and bootstrap PIs obtained from methods such as exponential smoothing, neural networks, and regression with auto-correlated errors implemented in the R-package `forecast` (see Hyndman and Khandakar, 2008).

An interesting advantage of predicting a sum of future values instead of just a single k-step ahead prediction is it allows for some weak law of large numbers to kick in if the horizon length of prediction is large. Towards the methodological development and corresponding theoretical advancement, we explore the properties of the quantile-based PIs of Zhou et al. (2010) in a high-dimensional regression setting. First we identify that Zhou et al. (2010) uses linear processes in their modelling of innovation processes. This limits the scope of applicability. Using the idea of predictive density and the functional dependence framework as proposed in a seminal paper by Wu (2005a), we extend the quantile consistency results to a large class of nonlinear processes which includes thresholded/transition autoregressive (Lundbergh et al., 2003) processes and some nonlinear version of GARCH processes.

On the regression front, the quantile consistency results obtained so far were limited to only a low-dimension regression setting. This builds a strong motivation to explore a high-dimensional regression setting as this would allow us to incorporate several predictors that better explain the variability in the data. In particular, we put a special focus on the scenario where the number of predictors grows much faster than the sample size ($\log p = o(n)$). Using the popular LASSO estimator, we were able to provide theoretical guarantees for these PIs. One interesting contribution of this paper is how our consistency rates explicitly depict the price of having short or long-range dependence and lighter or heavier tails. We also show sharp consistency results for stochastic design in the high-dimensional regime which can be seen as a contribution that is important on its own. Thus on one hand, this paper advances the theory of future-aggregated prediction by enlarging the scope of time-series; at the same time, it allows for many covariates in a high-dimensional regression setting. In particular, this paper can be seen to provide some theoretical justification for how a simple LASSO-based prediction routine can sustain such general time-series dependence.

---

and environmental policy (Knittel and Roberts, 2005; Huurman et al., 2012). Additionally, EPF is challenging due to complex seasonality (daily, weekly, and yearly), heteroscedasticity, heavy-tails, and sudden price spikes (see Weron, 2014, for a recent review).

Besides the theoretical contributions, we tackle the practical validity of PIs in the scenario of long-horizon forecasting with only a relatively short sample. One new discovery was that the original quantile-based PIs fail when the length of time-horizon to be aggregated for forecasting grows compared to the length of the past data. In light of this shortcoming, we employ a bootstrap-assisted step to improve the out-of-sample coverage probability. Apart from a conjectural viewpoint justifying our bootstrap procedure, these simulation results support our choice of PIs for the final out-of-sample experiment using real-world data on hourly electricity prices.

The rest of the paper is organized as follows: Section 2 shows the construction of quantile-based PIs and also details a novel bootstrap adjustment to improve coverage. Section 3 states the quantile consistency results in an exhaustive number of cases: the error/innovation process can be linear or non-linear, short-range or long-range and can have light or heavy tail. Building on the consistency for just the error process, we use Section 4 to consider consistency results for LASSO-fitted residuals. Apart from the traditional fixed design, we also provide discussion on stochastic design in this section. Moving on, the simulations are shown in details in Section 5 where we compare OLS, LAD and LASSO for both low-dimension and high-dimension regression setup. Section 6 is used to analyze EPEX spot electricity using several methods including ours. For contrasting our method with the competing ones in literature we use a Pseudo-out-of-sample (POOS hereafter) approach. Finally Section 7 provides concluding remarks. We defer some theoretical results and all proofs of theorems to the appendix.

We now introduce some notation. For a random vector Y, write $Y \in \mathcal{L}_p$, for $p > 0$, if $\|Y\|_p := E(|Y|^p)^{1/p} < \infty$. For the $\mathcal{L}_2$ norm write $\|\cdot\| = \|\cdot\|_2$. Throughout the text, c_p denotes a constant that depends only on p and c denotes a universal constants. These might take different values in different lines, unless otherwise specified. Then, $x^+ = \max(x, 0)$ and $x^- = -\min(x, 0)$. For two positive sequences a_n and b_n, if $a_n/b_n \to 0$, write $a_n = O(b_n)$. Write $a_n \lesssim b_n$ if $a_n \leq cb_n$, for some $c < \infty$. For a random variable sequence X_n and a possibly random or non-random sequence Y_n, we say $X_n = o_{\mathbb{P}}(Y_n)$ if $X_n/Y_n \to 0$ in probability. Respectively if X_n/Y_n is a tight random variable we say $X_n = O_{\mathbb{P}}(Y_n)$. The d-variate normal distribution with mean μ and covariance matrix Σ is denoted by $N(\mu, \Sigma)$. Denote by I_d the $d \times d$ identity matrix. For a matrix $A = (a_{ij})$, we define its element-wise $\ell - \infty$ norm as $|A|_\infty = max_{ij}|a_{ij}|$ and its ℓ_1 norm as $|A|_1 = max_j \sum_i |a_{ij}|$.

2 Methods: Construction of prediction intervals

Suppose an univariate target time series $y_i, i = 1, \ldots, n$ follows a regression model

$$y_i = \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta} + e_i, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \tag{2.1}$$

where e_i stands for the mean-zero temporally-dependent error process. Assuming model (2.1), we wish to construct PI for $y_{n+1} + \ldots + y_{n+m}$ after observing $(y_i, \boldsymbol{x}_i); i = 1, \ldots, n$. We first discuss the scenario $\boldsymbol{\beta} = 0$, i.e., $y_i = e_i$ is a zero-mean noise process. This serves as a primer to the high-dimensional regression problem and introduces a basic quantile-based method of constructing the prediction interval.

2.1 Without covariates

We will appropriately define our notion of short-range and long-range and light-tailed and heavy-tailed distribution later in Section 3.2. Loosely speaking, short-range dependence stands for faster decay of dependence as the lag increases where long-range dependent process kills dependence slowly. Heavy-tailed refers to the scenario where the error process does not possess even the finite second moment. Depending on the strength of dependence and the tail behavior, Zhou et al. (2010) proposed two different types of PIs for m-step ahead aggregated response $e_{n+1} + \ldots + e_{n+m}$. We present them here, with some suggestive modifications especially for the first approach where we estimate the long-run variance differently.

2.1.1 Quenched CLT method

If the process e_t shows short-range dependence and light-tailed behavior, then in the light of a quenched central limit theorem, Zhou et al. (2010) proposed the following PI for $\frac{1}{m}(e_{n+1} + \cdots + e_{n+m})$,

$$[L, U] = [Q^N(\alpha/2), Q^N(1 - \alpha/2)] \frac{\sigma}{\sqrt{m}}, \tag{2.2}$$

where σ is the long-run standard deviation (sd) of e_t. However, since σ is unknown, it must be estimated. One can estimate the long-run variance σ^2 of the e_i process using the sub-sampling block estimator (see eq. (2) in Dehling et al., 2013)

$$\tilde{\sigma} = \frac{\sqrt{\pi l/2}}{n} \sum_{k=1}^{\kappa} \left| \sum_{i=(k-1)l+1}^{kl} e_i \right|, \tag{2.3}$$

5

with block length l and number of blocks $\kappa = \lceil n/l \rceil$ in order to obtain $100(1-\alpha)\%$ asymptotic PI

$$[L, U] = \pm \hat{\sigma} Q_{\kappa-1}^t(\alpha/2)\sqrt{m},$$

where $Q_{\kappa-1}^t$ is student-t quantile with $\kappa - 1$ degrees of freedom.

2.1.2 Empirical method based on quantiles

A substantially more general method that can account for long-range dependence or heavy-tailed behavior of the error process uses the following quantile

$$\hat{Q}(u), \text{ the } u\text{-th empirical quantile of } \sum_{j=i-m+1}^{i} e_i; i = m, \ldots, n.$$

The PI in this case is

$$[L, U] = \left[\hat{Q}(\alpha/2), \hat{Q}(1-\alpha)/2) \right]. \tag{2.4}$$

This approach enjoys reasonable coverage for moderate rate of growth for m compared to sample size n.

2.2 Low dimensional regression

Now assume the scenario where $\boldsymbol{\beta}$ is possibly non-zero in (2.1). For $p < n$, with dependent but linear error process, after estimating $\boldsymbol{\beta}$, Zhou et al. (2010) constructed PI as

$$\sum_{i=n+1}^{n+m} \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \text{ PI for } \sum_{i=n+1}^{n+m} \hat{e}_i, \tag{2.5}$$

where $\hat{e}_i = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$ are the regression residuals. Regarding the choice of estimator for $\boldsymbol{\beta}$, if the error process shows light tailed behavior and short-range dependence, we typically use OLS estimator $\hat{\boldsymbol{\beta}} = \operatorname{argmin} \sum_i (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$. For heavy-tailed or long-range dependent errors (see Huber and Ronchetti, 2009), it is better to use robust regression with general distance $\rho(\cdot)$ and $\hat{\boldsymbol{\beta}} = \operatorname{argmin} \sum_i \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$. Examples of distance include the $\mathcal{L}^q$ regression for $1 \le q \le 2$. In this paper our focus is on a LASSO-type least square estimation (cf. 2.6) in presence of possibly dependent and nonlinear errors where the number of covariates p is much larger than the sample size n.

2.3 High dimensional regression

Consider the high dimensional regression situation i.e. $p \gg n$. Here we use the popular LASSO estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \qquad (2.6)$$

with the penalty coefficient λ. Then we get PIs (2.5) with $\hat{\boldsymbol{\beta}}$ replaced by the LASSO estimator. One of the key contributions of our paper lies in the fact that the intuitive and computationally fast lasso estimator adapts well for the time-aggregated response scenario but still provides theoretical consistency guarantee as we show in Section 4.

2.4 Future predictors

Note that, the PI in (2.5) requires the future values of covariates $\boldsymbol{x}_i$ namely $\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+m}$. Zhou et al. (2010) discussed the scenario where these predictors are trigonometric and thus perfectly predictable. For practical implementation in econometrics, it is more appealing to consider stochastic design so that one can allow for the covariates $\boldsymbol{x}_i$ to be observed random variables as well. This naturally raises the question of how to predict the future values of the covariates. For the theoretical part, we shed some light on consistency results for the stochastic design in Section 4. However, in the implementation in Section 6, we choose to simply use a year-over-year mean forecast for the weather covariates.

2.5 Bootstrap adjustment

In our actual implementation for real data, we propose a bootstrap adjusted version of the algorithm in Section 2.2. In the stationary bootstrap, one randomly draws a sequence of starting point uniformly from $\{1, 2, \ldots, n\}$ and a sequence of Geometric random variable $L_1, L_2, \cdots$. Depending on the values of L_i and the starting value, we draw a consecutive L_i-length block and then finally concatenate these blocks together to arrive at a replicated series of length n. Then finally we look at the final m for each such series, whereas it is understandable that it was not particularly important to only look at the last m since the starting point is uniform. We conjecture that this ensures consistency of the quantiles despite providing some more dispersion to the average. This extra dispersion arises as this method concatenates blocks containing same elements with a nontrivial probability and thus adding those covariances in the dispersion of the overall average. We postpone a rigorous theoretical justification of

this innovative bootstrap technique to a future work since this paper focuses more on the exploration of the performance of LASSO fitted residuals. For the convenience of the readers we summarize the algorithm down below:

Algorithm 2.1 (Bootstrap PIs ADJ).   *i. Estimate regression $y_t \sim \boldsymbol{x}_t^{\mathsf{T}}$, $t = 1 \ldots, n$ with LASSO.*

ii. Using stationary bootstrap (See Politis and Romano (1994)), replicate residuals $\hat{e}_t = y_t - \hat{y}_t$, B times obtaining $\hat{e}_t^b, t = 1, \ldots, n$, $b = 1, \ldots, B$.

iii. Compute $(\bar{e}_{t(m)}^b) = m^{-1} \sum_{i=1}^m e_{t-i+1}^b$, $t = m, \ldots n$ from every replicated series.

iv. Estimate the $\alpha/2$th and $(1 - \alpha/2)$th quantile $\hat{Q}(\alpha/2)$ and $\hat{Q}(1 - \alpha/2)$ using Gaussian kernel density estimator from $\bar{e}_{n(m)}^b$, $b = 1, \ldots, B$.

v. The PI for $\bar{y}_{+1:m}$ is $[L, U] = \bar{\hat{y}}_{n,1:m} + [\hat{Q}(\alpha/2), \hat{Q}(1 - \alpha/2)]$, where $\bar{\hat{y}}_{n,1:m}$ is the average of h-step-ahead forecasts for $h = 1, \ldots, m$.

Our theoretical results are concerned with the consistency of the usual quantiles from the original series. But we conjecture that the stationary bootstrap technique to obtain the replicated series retains the asymptotic dependence structure of the original series and thus the quantiles of the m-length average from the original series and that from the final m of the replicated series are close to each other. Additionally, using the Gaussian kernel density to obtain the kernel quantile estimator (See Sheather and Marron (1990)) further improves the performance in prediction. These adjustments are supported by the empirical evidence given in Chudý et al. (2020) in a univariate setup.

Next, we exhibit some theoretical consistency results addressing various different dependency cases, tail-decay and high dimensional regression in the next two sections.

3   Prediction interval for error process

Before moving on to a more general discussion with a large number of covariates compared to sample size, we use this section to discuss a primer without covariates i.e. our response here is just a mean-zero error process. This section also elaborately describes the model specifications on the error process i.e. whether the process is short/long-range dependent or has light/heavy tails. Under short-range dependence, if window size m is long enough then the dependence of $y_{n+1} + \ldots + y_{n+m}$ on $y_1, y_2, \ldots, y_n$ diminishes and the conditional distribution of $(y_{n+1} + \ldots + y_{n+m})/\sqrt{m}$

given $y_1, \ldots, y_n$ is almost similar to the unconditional distribution and thus one can obtain a simple central limit theorem to quantify the uncertainty surrounding the prediction. Consider (e_i) a mean-zero stationary process and let $S_m = e_1 + \ldots + e_m$. Wu and Woodroofe (2004) proved that for $q > 5/2$, the condition

$$\|\mathbb{E}(S_m|\mathcal{F}_0)\|_2 = O\left(\frac{\sqrt{m}}{\log^q m}\right), \tag{3.1}$$

gives the a.s. convergence

$$\Delta(\mathbb{P}(S_m/\sqrt{m} \leq \cdot|\mathcal{F}_0), N(0, \sigma^2)) = 0 \text{ a.s.,} \tag{3.2}$$

where $\|\cdot\|_2$ denotes the L_2-norm, Δ denotes the Levy distance, $\mathcal{F}_i$ is the $\sigma-$field $\sigma(\ldots, e_{i-1}, e_i)$, $m \to \infty$ and $\sigma^2 = \lim_{m\to\infty} \|S_m\|_2^2/m$ is the long-run variance. We start by collecting an asymptotic normality result from Chudý et al. (2020). When the error process has possibly long-range dependence such asymptotic normality fails. Keeping the linear structure intact, we provide some empirical consistency result for the quantile-based methods from Section 2. These results depict all possible scenarios of heaviness of tail and range of dependence. Finally we extend these results to the more practically applicable non-linear case under the functional dependence framework. Throughout this section, the observed process e_i are either direct linear sum of independent and identically distributed (i.i.d. henceforth) innovations that are unobserved or we assume e_i to be a general non-linear function of these. For the latter, we define some tractable moment-based coupling measure which we call functional dependence measure.

3.1 Linear error process: Theoretical results

Assuming linearity of the mean-zero noise process e_i in the following manner

$$e_i = \sum_{j=0}^{\infty} a_j \epsilon_{i-j}, \text{ with } \epsilon_i \text{ i.i.d., } \mathbb{E}(|\epsilon_1|^p) < \infty \tag{3.3}$$

for some $p > 2$, it is easy to derive the following conditions on a_i to ensure the convergence in (3.2). The proof can be found in the appendix of Chudý et al. (2020).

Theorem 3.1 (Theorem 1 from Chudý et al. (2020)). *Assume the process e_t admits the representation (3.3) where a_i satisfies*

$$a_i = O(i^{-\chi}(\log i)^{-A}), \quad \chi > 1, A > 0, \tag{3.4}$$

9

where larger χ and A means fast decay rate of dependence. Further assume, $A > 5/2$ if $1 < \chi < 3/2$. Then the sufficient condition (3.1) implies that the convergence (3.2) to the normal distribution holds.

Remark The central limit theorem described in (3.2) does not hold if the sequence a_i is not absolutely summable or if the moment assumption in (3.3) is relaxed.

Next, as outlined above we proceed for a quantile consistency result for linear processes to validate our proposed methods. We start by formally defining short-range and long-range dependence for linear processes with the representation in (3.3). For short-range dependence we assume

$$\text{(SRDL): } \sum |a_i| < \infty.$$

For long range dependence, we first revisit the definition of slowly varying function (s.v.f.). A function $g(\cdot)$ is called a s.v.f. if for all $a > 0$, $\lim_{x\to\infty} g(ax)/g(x) = 1$. Assume

$$\text{(LRDL}(\gamma)): l^*(i) = a_i i^{-\gamma} \text{ is a slowly varying function (s.v.f.)}$$

for some $q < \gamma < 1$ where $1/q = \sup\{t : \mathbb{E}(|\epsilon_j|^t) < \infty\}$. Note that, the definition of $(\text{LRDL}(\cdot))$ also takes into account the possible heavy-tailed distribution of ϵ_j. We also assume ϵ_i admits a density f_ϵ and

$$\text{(DENL)}: \sup_{x\in\mathbb{R}} (f_\epsilon(x) + |f'_\epsilon(x)|) < \infty.$$

For a fixed $0 < u < 1$, let $\hat{Q}(u)$ and $\tilde{Q}(u)$ denote the u-th sample quantile and the actual quantile of $\tilde{S}_i$, $i = m, \ldots, n$, respectively, where

$$\tilde{S}_i = \frac{\sum_{j=i-m+1}^{i} e_j}{H_m^*}, \quad i = m, m+1, \ldots \tag{3.5}$$

and

$$H_m^* = \begin{cases} \sqrt{m}, & \text{for Case1,} \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x) \le \frac{1}{m}\} & \text{for Case2,} \\ m^{3/2-\gamma} l^*(m) & \text{for Case3,} \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x) m^{1-\gamma} l^*(m) & \text{for Case4,} \end{cases}$$

10

where Case 1-2 denotes SRDL holds and $\mathbb{E}(\epsilon_j^2) < \infty$ or $\mathbb{E}(\epsilon_j^2) = \infty$ respectively whereas Case 3-4 stands for LRDL(γ) holds and $\mathbb{E}(\epsilon_j^2) < \infty$ or $\mathbb{E}(\epsilon_j^2) = \infty$ respectively. We collect the following theorem from Zhou et al. (2010) for the rates of convergence of quantiles depending on the nature of the error process in terms of tail behaviour and dependence:

Theorem 3.2. *[Empirical quantile consistency: linear error process] Assume (DENL) holds. Additionally*
- *Light tailed (SRDL): Suppose (SRDL) holds and $\mathbb{E}(\epsilon_j^2) < \infty$. If $m^3/n \to 0$, then for any fixed $0 < u < 1$,*

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_\mathbb{P}(m/\sqrt{n}). \tag{3.6}$$

- *Light tailed (LRDL): Suppose (LRDL) holds with γ. If $m^{5/2-\gamma}n^{1/2-\gamma}l^2(n) \to 0$, then for any fixed $0 < u < 1$,*

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_\mathbb{P}(mn^{1/2-\gamma}|l^*(n)|). \tag{3.7}$$

- *Heavy-tailed (SRDL): Suppose (SRDL) holds and $\mathbb{E}(|\epsilon_j|^q) < \infty$ for some $1 < q < 2$. If $m = O(n^k)$ for some $k < (q-1)/(q+1)$, then for any fixed $0 < u < 1$,*

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_\mathbb{P}(mn^\nu) \text{ for all } \nu > 1/q - 1. \tag{3.8}$$

- *Heavy-tailed (LRDL): Suppose (LRDL) holds with γ. If $m = O(n^k)$ for some $k < (q\gamma - 1)/(2q + 1 - q\gamma)$, then for any fixed $0 < u < 1$,*

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_\mathbb{P}(mn^\nu) \text{ for all } \nu > 1/q - \gamma. \tag{3.9}$$

A technical fact about these results is that their proofs are heavily dependent on the linear structure. We believe that it is an important task to extend these to the more general non-linear scenarios but it comes at the cost of some more abstraction in how these nonlinear processes are posited. We use a very general framework of functional dependence form Wu (2005a) to first describe the class of non-linear process and then provide analogs of central limit theorem and quantile consistency results.

3.2 Central limit theory for nonlinear dependence

Economic and financial time series are often subject to structural changes and thus the linear models often fail to capture the diverse range of data generating process. Comparatively possible non-linear class of time-series models is significantly richer in its scope. Useful nonlinear time-series models, among many, include the

regime-switching autoregressive processes, which assume that the series change their dynamics when passing from one regime to another and neural-network models. We provide extension of the asymptotic normality result (3.2) to nonlinear error process. We lay down the definition of non-linearity through the following structure: Let e_i be a stationary process that admits the following representation

$$e_i = H(\mathcal{F}_i) = H(\epsilon_i, \epsilon_{i-1}, \ldots), \tag{3.10}$$

where H is such that e_i is a well-defined random variable, $\epsilon_i, \epsilon_{i-1}, \ldots$ are i.i.d. innovations and $\mathcal{F}_i$ denotes the σ-field generated by $(\epsilon_i, \epsilon_{i-1}, \ldots)$. One can see that it is a vast generalization from the linear structure in H. In order to derive a result similar to (3.2) but in the non-linear regime, we define the following functional dependence measure for e_i in (3.10), by which we follow Wu (2005a)'s framework to formulate dependence through coupling:

$$\delta_{j,p} = \|e_i - e_{i,(i-j)}\|_p = \|H(\mathcal{F}_i) - H(\mathcal{F}_{i,(i-j)}\|_p, \tag{3.11}$$

where $\mathcal{F}_{i,k}$ is the coupled version of $\mathcal{F}_i$ with ϵ_k in $\mathcal{F}_i$ replaced by an i.i.d. copy ϵ'_k, $\mathcal{F}_{i,k} = (\epsilon_i, \epsilon_{i-1}, \ldots, \epsilon'_k, \epsilon_{k-1}, \ldots)$ and $e_{i,(i-j)} = H(\mathcal{F}_{i,(i-j)})$. Clearly, $\mathcal{F}_{i,k} = \mathcal{F}_i$ is $k > i$. As Wu (2005a) suggests, $\|H(\mathcal{F}_i) - H(\mathcal{F}_{i,(i-j)})\|_p$ measures the dependence of e_i on ϵ_{i-j}. This dependence measure can be seen as an input-output system and a natural analogue of the linear coefficient at a certain lag. It facilitates mild moment conditions on the dependence of the process which are easily verifiable compared to the more popular strong mixing conditions. Define the cumulative dependence measure

$$\Theta_{j,p} = \sum_{i=j}^{\infty} \delta_{i,p}, \tag{3.12}$$

which can be thought as cumulative dependence of $(e_j)_{j \geq k}$ on ϵ_k.

Theorem 3.3. *Assume e_i admits the representation in (3.10). Also assume under the functional dependence formulation in 3.11, the following rate holds for the cumulative dependence $\Theta_{j,p}$:*

$$\Theta_{j,p} = O(j^{-\chi}(\log j)^{-A}) \text{ where } \begin{cases} A > 0 \text{ for } 1 < \chi < 3/2, \\ A > 5/2 \text{ for } \chi \geq 3/2, \end{cases} \tag{3.13}$$

then the convergence in (3.2) holds.

Apart from the asymptotic normality result from Theorem 3.3, we will next show that the estimated quantiles from the moving blocks are consistent. However, the

12

same notion of non-linear dependence falls short of establishing quantile consistency due to a technical reason and thus we resort to first define some form of predictive dependence.

3.3 Quantile consistency for nonlinear process: predictive dependence

For a general nonlinear process with possibly heavy tails and long-range dependence, the central limit theorem fails. We present one of the main results in this paper by showing the empirical quantile consistency that validates PIs of the form (2.4). Towards that, we need to control the latent dependence of e_i on ϵ_{i-j} keeping in mind a representation like (3.10). Therefore, we introduce the predictive density-based dependence measure. Let $\mathcal{F}'_k = (\epsilon_k, \ldots, \epsilon_1, \epsilon'_0, \epsilon_{-1}, \ldots,)$, be the coupled shift process derived from $\mathcal{F}_k$ by substitution of ϵ_0 by its i.i.d. copy ϵ'_0. Let $F_1(u|\mathcal{F}_k) = P\{H(\mathcal{F}_{k+1}) \le u|\mathcal{F}_k\}$ be the one-step ahead predictive or conditional distribution function and $f_1(u|\mathcal{F}_k) = dF_1(u|\mathcal{F}_k)/du$, be the corresponding conditional density. We define the predictive dependence measure

$$\psi_{k,q} = \sup_{u \in \mathbb{R}} \|f_1(u|\mathcal{F}_k) - f_1(u|\mathcal{F}'_k)\|_q. \tag{3.14}$$

The quantity $\psi_{k,q}$ (3.14) measures the contribution (read change in q-th norm) of ϵ_0, the innovation at step 0, on the conditional or predictive density at step k. We shall make the following assumptions:

i. For short-range dependence: $\Psi_{0,2} < \infty$ where $\Psi_{m,q} = \sum_{k=m}^{\infty} \psi_{k,q}$; for long-range dependence: $\Psi_{0,2}$ can possibly be infinite;

ii. (DEN) There exists a constant $c_0 < \infty$ such that almost surely,

$$\sup_{u \in \mathbb{R}}\{f_1(u|\mathcal{F}_0) + |df_1(u|\mathcal{F}_0)/du|\} \le c_0.$$

The (DEN) implies that the marginal density $f(u) = \mathbb{E}f_1(u|\mathcal{F}_0) \le c_0$. Recall the sufficient conditions for the linear cases in Zhou et al. (2010) were based on the coefficients of the linear process. Here, the conditions for both short-range and long-range dependent errors here were transferred onto predictive dependence measure. We assume:

$$\text{(SRD)} \quad : \quad \sum_{j=0}^{\infty} |\psi_{j,q}| < \infty, \tag{3.15}$$

$$\text{(LRD}(\gamma)) \quad : \quad \psi_{j,q} = j^{-\gamma}l(j), q < \gamma < 1, l(\cdot) \text{ is a slowly varying function (s.v.f.) }.$$

13

where $1/q = \sup\{t : \mathbb{E}(|\epsilon_j|^t) < \infty\}$. For a fixed $0 < u < 1$, let $\hat{Q}(u)$ and $\tilde{Q}(u)$ denote the u-th sample quantile and actual quantile of $\tilde{S}_i$; $i = m, \ldots, n$, where

$$\tilde{S}_i = \frac{\sum_{j=i-m+1}^{i} e_j}{H_m}, \quad i = m, m+1, \ldots \tag{3.16}$$

and

$$\tag{3.17}$$

$$H_m = \begin{cases} \sqrt{m}, & \text{for Case1,} \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x) \leq \frac{1}{m}\} & \text{for Case2,} \\ m^{3/2-\gamma}l(m) & \text{for Case3,} \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x)m^{1-\gamma}l(m) & \text{for Case4,} \end{cases}$$

Here, Case 1-2 denotes SRD holds and $\mathbb{E}(\epsilon_j^2) < \infty$ or $\mathbb{E}(\epsilon_j^2) = \infty$ respectively whereas Case 3-4 stands for $LRD(\gamma)$ holds and $\mathbb{E}(\epsilon_j^2) < \infty$ or $\mathbb{E}(\epsilon_j^2) = \infty$ respectively.
Then we have following rates of convergence of quantiles depending on the nature of the error process in terms of tail behaviour and dependence:

Theorem 3.4 (Empirical quantile consistency: nonlinear error process). - *Light tailed (SRD): Suppose (DEN) and (SRD) hold and $\mathbb{E}(\epsilon_j^2) < \infty$. If $m^3/n \to 0$, then for any fixed $0 < u < 1$,*
$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(m/\sqrt{n}). \tag{3.18}$$

- *Light tailed (LRD): Suppose (LRD) and (DEN) hold with γ and $l(\cdot)$ in (3.15). If $m^{5/2-\gamma}n^{1/2-\gamma}l^2(n) \to 0$, then for any fixed $0 < u < 1$,*

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(mn^{1/2-\gamma}|l(n)|). \tag{3.19}$$

- *Heavy-tailed (SRD): Suppose (DEN) and (SRD) hold and $\mathbb{E}(|\epsilon_j|^q) < \infty$ for some $1 < q < 2$. If $m = O(n^k)$ for some $k < (q-1)/(q+1)$, then for any fixed $0 < u < 1$,*

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(mn^\nu) \text{ for all } \nu > 1/q - 1. \tag{3.20}$$

- *Heavy-tailed (LRD): Suppose (LRD) hold with γ and $l(\cdot)$ in (3.15). If $m = O(n^k)$ for some $k < (q\gamma - 1)/(2q + 1 - q\gamma)$, then for any fixed $0 < u < 1$,*

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(mn^\nu) \text{ for all } \nu > 1/q - \gamma. \tag{3.21}$$

Note that the results presented above talks about the scenario without any predictor. For the low dimensional regression case, the proofs follow exactly similar to how (Zhou et al., 2010) proceeded by showing

$$\sup_{i \leq n} |\hat{e}_i - e_i| = O_{\mathbb{P}}(\Pi(n))$$

for some suitably chosen small $\Pi(n)$. We skip the details here and directly move on to the scenario where the number of predictors far outgrows the number of past time-points available.

4 High dimensional regression regime

In this section, we focus on results concerning theoretical guarantees of the prediction intervals for high-dimensional regression where the error process is possibly nonlinear and shows temporal dependence. Note that all the results from this section can be easily extended to situations where the error process has exponentially decaying tails such as sub-exponential and sub-gaussian, but we restrict ourselves to conditions of only finitely many moments which is a significantly weaker assumption than that is present in the literature.

The results for this subsection heavily depend on optimal concentration inequalities for $S_{n,b} = \sum_{i=1}^{n} b_i e_i$ that have not been established for possibly nonlinear dependence before this. Since this is just a proof technique tool we postpone the concentration results here. However, as one of our final aim is to point out the price of dependence in a wide range of general scenarios, it is important to define dependence-adjusted norm to be able to state the concentration and consistency results. Recall (3.12) and assume the short-range dependence, $\Theta_{0,q} < \infty$ holds with q being less or more than 2 depending on the tail-behavior of the error process. Further, we define dependence adjusted norm, for $\alpha > 0$,

$$\|e.\|_{q,\alpha} = \sup_{t \geq 0}(t + 1)^{\alpha} \sum_{i=t}^{\infty} \delta_{i,q}. \tag{4.1}$$

It is easy to note that, finiteness of the dependence-adjusted measure is a stronger ask than the finiteness of $\Theta_{0,q}$. Next, we show that for short-range dependent nonlinear error processes the error bounds obtained in Theorem 3.4 remain intact under a proper choice of the sparsity condition.

4.1 Lasso with fixed design

For the model in (2.1), we first assume that the $\boldsymbol{x}_i$'s are fixed and the future $\boldsymbol{x}_i$'s are known. Under this setting, we next show the quantile consistency for the nonlinear process. Note that a very similar result can be shown for the linear process with the error process admitting a simpler representation (3.3), however we skip writing that as a separate theorem here to avoid repetitiveness.

Theorem 4.1. *(Empirical quantile consistency for LASSO-nonlinear)*
Assume the covariates are so scaled such that $\|X\|_2 = (np)^{1/2}$. Denote $\lambda = 2r$ in the criterion function (2.6) where

$$r = \max\{A\sqrt{n^{-1}\log p}\|e_.\|_{2,\alpha}, B\|e_.\|_{q,\alpha}\|X\|_q n^{-1+\min\{0,1/2-1/q-\alpha\}}\}.$$

We assume that the restricted eigenvalue assumption $RE(s, \kappa)$ in Bickel et al. (2009) holds with constant $\kappa = \kappa(s, 3)$, where s is the number of non-zero entries in true parameter vector $\boldsymbol{\beta}$ and

$$\kappa(s, c) = \min_{J \subset \{1, \cdots, p\}, |J| \leq s, } \min_{|u_{J^c}|_1 \leq c|u_J|_1} \frac{\|Xu\|_2}{\sqrt{n}\|u_J\|_2}. \tag{4.2}$$

Here u_J stands for modified u by setting its elements outside J to zero. Let $\bar{Q}_n(u)$ be the u-th empirical quantile of $(\tilde{\hat{S}}_i)_m^n$. Assume that (SRD) holds and for r defined in (4.2),

$$\begin{aligned}
(for\ q \geq 2) \quad s &= o\left(\frac{m}{r^2 n}\right), \\
(for\ 1 < q \leq 2), \quad s &= o\left(\frac{H_m^2 |l(n)|^2}{r^2 n^{2\gamma-1}}\right),
\end{aligned} \tag{4.3}$$

where γ and $l(\cdot)$ are defined in (3.15), H_m in (3.17) and α in the definition of r is in the context of the dependence adjusted norm defined in (4.1), then the (SRD) specific conclusions of Theorem 3.4 hold with $Q_n(u)$ replaced by $\bar{Q}_n(u)$.

One can note the $\sqrt{\log p/n}$ term we have in our definitions for $r = \lambda/2$. This allows us to capture the ultra-high dimensional scenario where $\log p = o(n)$, the usual benchmark in the high-dimensional literature. The additional terms involving $\|e_.\|_{.,\alpha}$ are due to the dependence present in the error process. The sparsity condition for the light tail case, i.e. $q \geq 2$, in (4.3) in the view of the choice of r in (4.2) can be written as: For $q \geq 2$

$$s \ll \min \left(\frac{n^{4/3}}{\log p \|e.\|_{2,\alpha}^2}, \frac{n^{7/3 - 2\max\{0, 1/2 - 1/q - \alpha\}}}{|X|_q^2 \|e.\|_{q,\alpha}^2} \right)$$

for the choice of $m = o(n^{1/3})$. Thus we can allow p to grow a bit faster than the usual ultra-high dimensional benchmark $e^{O(n)}$ rate. This is an interesting result and our conjecture is that this added advantage is due to considering future aggregation instead of just k-step ahead forecast for a fixed k. In other words, if prediction horizon m is allowed to grow to ∞, the m-length average of residuals can automatically provide some concentration. Thus it can allow for scenarios where estimation of $\boldsymbol{\beta}$ is not very precise. We believe that this is an interesting exploration of the relaxation of the sparsity condition compared to the usual LASSO literature.

For the special case of the linear process (See 3.3), the conditions in (4.3) will remain identical and one can provide some more specifications in the definition of r in (4.2) using the linear coefficients a_i from (3.3) in the view of the Nagaev-type concentration inequalities derived in Result 9.1. Moreover, for the linear process, one can also state the corresponding results for the (LRDL) case, however since the condition on sparsity is unaffected by this nature of dependence, we do not state them separately here.

4.2 Lasso with stochastic design: covariate prediction issue

For the high-dimensional regression scenario, it is very natural to ask whether one could relax the fixed predictor settings to random design as it is practically impossible to find many perfectly predictable covariates. We show that under a mild condition on the possibly stochastic covariate process, the prediction intervals based on LASSO-optimized residuals also work in a stochastic design set-up. This is a particularly interesting extension from Zhou et al. (2010) since in the high-dimensional regime it is impractical to assume an exponentially growing number of covariates to be perfectly predictable for its future values. Moreover, apart from allowing a random design, we also allow the covariates to be temporally dependent. For this subsection, we restrict ourselves to only nonlinear processes for the sake of clarity. We assume

$$\boldsymbol{x}_t = \mathbf{G}_x(\epsilon_t^x, \epsilon_{t-1}^x, \ldots) = (g_1(\epsilon_t^x, \epsilon_{t-1}^x, \ldots), g_2(), \ldots, g_p(\epsilon_t^x, \epsilon_{t-1}^x, \ldots))^{\mathsf{T}} \qquad (4.4)$$

where ϵ_i^x are i.i.d., $\mathbf{G}_x$ is a measurable function and $\boldsymbol{x}_t$ is $p \times 1$ high-dimensional mean-zero stationary process. Let $\{\epsilon_i^{x\prime}\}$ be an i.i.d. copy of $\{\epsilon_i^x\}$. Due to the

high dimension of $\boldsymbol{x}_t$ and the dependence as a stochastic time-series, we opt for a dependence-adjusted uniform (over co-ordinates) functional dependence measure. For a fixed co-ordinate $1 \leq j \leq p$ and $q \geq 1$, define

$$\delta^{\boldsymbol{x}}_{k,q,j} = \|g_j((\epsilon^x_i, \epsilon^x_{i-1}, \ldots, \epsilon^x_{i-k}, \ldots)) - g_j((\epsilon^x_i, \epsilon^x_{i-1}, \ldots, \epsilon^{x'}_{i-k}, \ldots))\|_q. \tag{4.5}$$

Use this to define the dependence-adjusted norm as

$$\Phi_{q,\alpha} = \sup_{1 \leq j \leq p} \|X_{.j}\|_{q,\alpha} \text{ where } \|X_{.j}\|_{q,\alpha} = \sup_{m \geq 0}(m+1)^\alpha \sum_{t=m}^{\infty} \delta_{t,q,j}. \tag{4.6}$$

The quantity $\Phi_{q,\alpha}$ provides a concise and natural measure of dependence which can effectively account for high dimensionality and temporal dependence. Let also the error process (e_i) admit the following representation

$$e_i = G_e(\epsilon^e_i, \epsilon^e_{i-1}, \ldots), \tag{4.7}$$

where ϵ^x_i are i.i.d. and G_e is a measurable function. One can then define the cumulative dependence using this functional dependence measure. However, thanks to the structure of the regression problem we can skip defining that for the e_i process. Instead, for the quantile consistency in the case of stochastic design, we will need a notion of functional dependence on the cross-product process $x_{.j}e_{.}$ as follows

$$\delta^{xe}_{k,q} = \max_{j \leq p} \|x_{lj}e_l - x^*_{k,lj}e^*_{k,l}\|_q, \tag{4.8}$$

where $e^*_{k,l} = G_e(\epsilon^e_l, \epsilon^e_{i-1}, \ldots, \epsilon^e_{i-k}{}', \ldots)$ and $\{\epsilon^e_i\}'$ is an i.i.d. copy of $\{\epsilon^e_i\}$. Using (4.8), we define the dependence adjusted norm as (4.1) for the $x_{.j}e_{.}$ process uniformly over $1 \leq j \leq p$ as follows:

$$\max_{j \leq p} \|x_{.j}e_{.}\|_{q,\alpha} = \sup_{t \geq 0}(t+1)^\alpha \sum_{i=t}^{\infty} \delta^{xe}_{i,q}. \tag{4.9}$$

With this background on the stochastic covariate and the error process we now state the quantile consistency result below.

18

Theorem 4.2. *(Empirical quantile consistency for LASSO-stochastic)*
Assume $\Phi_{q,\alpha} < \infty$ and $\max_{j\le p} \|x_{.j}e_.\|_{q,\alpha} < \infty$ for some $\alpha > 0$. Let $\bar{Q}_n(u)$ be the u-th empirical quantile of $(\tilde{\hat{S}}_i)_m^n$. We assume that the restricted eigenvalue assumption $RE_{stoch}(s,\kappa)$ holds with constant $\kappa_{stoch} := \kappa(s,3)$, where s is the number of non-zero entries in true parameter vector $\boldsymbol{\beta}$ and

$$\kappa(s,c) = \sqrt{\min_{J\subset\{1,\cdots,p\},|J|\le s,\ |u_{J^c}|_1\le c|u_J|_1} \min \frac{u^\mathsf{T}\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t^\mathsf{T})u}{\|u_J\|_2^2}}. \tag{4.10}$$

Assume that the sparsity conditions in (4.3) hold with the choice of $\lambda = 2r$ where

$$r = \max\{A\sqrt{n^{-1}\log p}\max_{j\le p}\|x_{.j}e_.\|_{2,\alpha}, B\max_{j\le p}\|x_{.j}e_.\|_{q,\alpha}n^{-1+\min\{0,1/2-1/q-\alpha\}}\}.$$

Additionally s satisfies $s\sqrt{\log p}/\sqrt{n} \to 0$. Then the SRD specific conclusions of Theorem 3.4 hold with $Q_n(u)$ replaced by $\bar{Q}_n(u)$.

Remark Note that, the uniform functional dependence measure on the cross-product space $x_{.j}e_.$ can often be simplified using Hölder inequalities and the usual triangle inequality technique. For some examples and calculations of the functional dependence measure for the nonlinear covariate processes, see Wu and Wu (2016).

The prediction intervals based along the line of (2.5) would need the future values of $\boldsymbol{x}_i$ which we do not observe. One possible solution to this is to fit a vector-autoregressive (VAR) model with appropriate lags and then estimate the k-step ahead predictions for $1 \le k \le m$ using the estimated matrix coefficients of the VAR process. It is also possible to lay down assumptions on the $\boldsymbol{x}_i$ and e_i process and handle this in a much more rigorous way. But since our focus is on the relatively easier but asymptotically valid methods of estimating the prediction intervals based on quantiles, we omit that discussion. Instead, for practical implementation, (cf. our data analysis from Section 6 where, for covariates, we used wind and temperature data that are stochastic in nature) we just use the past year-over-year mean as to substitute for $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}$ for $i = n+1, \ldots, n+m$. Since m is also growing, it is natural that the values of aggregated $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}$ would concentrate around aggregated past year-over-year mean. We plan to discuss the issue of simultaneously estimating the future $\boldsymbol{x}_i$'s in a future work.

5 Simulation

In this section, we compare the predictive performance of methods discussed above using OLS, LAD and LASSO estimator in both low-dimensional and high-dimensional setup. In the low-dimension setup the comparison between CLT based methods (Quenched CLT as described in Section 2) and QTL- Quantile based methods will be evaluated.

5.1 Simulation set-up

The focus here is on evaluation of PIs discussed in the previous section based on their coverage probability. We start by generating the error process (e_t) as:

(a) $e_i = \phi_1 e_{i-1} + \sigma \epsilon_i,$

(b) $e_i = \sigma \sum_{j=0}^{\infty} (j+1)^\gamma \epsilon_{i-j},$

(c) $e_i = \phi_1 e_{i-1} + G(e_{i-1}; \delta, T)(\phi_2 e_{i-1}) + \sigma \epsilon_i,$

with ϵ_i i.i.d. from an α^*-stable distribution. The heavy-tails index $\alpha^* = 1.5$, autocovariance decay parameter $\gamma = -0.8$, speed-of-transition parameter $\delta = 0.05$, autoregressive coefficients $\phi_1 = 0.6$ and $\phi_2 = -0.3$, the noise standard deviation $\sigma = 54.1$, and threshold $T = 0$, were all selected based on the autoregressive models fitted to the electricity prices used later in the empirical part. The logistic transition function is given by $G(e_{i-1}; \delta, T) = (1 + \exp(-\delta(e_{i-1} - T)))^{-1}$. These three specifications represent

(a) a heavy-tail and short-memory error-process,

(b) a heavy-tail and long-memory error-process, and

(c) a nonlinear error-process know as the logistic smooth transition autoregression (LSTAR) with heavy-tailed innovations.

respectively. Eventually, we add a large number of exogenous covariates to the error process, obtaining $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i, i = 1, \dots, n + m$. We compute our PIs based on $(y_1, \boldsymbol{x}_1) \dots, (y_n, \boldsymbol{x}_n)$ and evaluate them on $\bar{y}_{+1:m} = 1/m \sum_{i=1}^{m} y_{n+i}$. Note that we predict the averages instead of sums. This is motivated by easier comparison of predictive performance across different forecast horizons, and also turns out as more appropriate in the following empirical part.

Regarding covariates, we consider two scenarios (i) $p < n$ and (ii) $p > n$. In scenario (i) (See Table 1i), we compare PIs based on OLS, LAD, and LASSO estimators. We set $n = 8736$ ($\approx$ 1 year of hourly data), $m = 168, 336, 504, 672$ (1,2,3,4 weeks of hourly data), and $p = 319$ (151 weather variables and 168 periodic variables), similarly to our empirical application (except that the horizon there spans up to 17 weeks) described in Section 6. In scenario (ii) (See Table 1ii), we only use the LASSO as the other two estimators are not uniquely identified. We set[2] $n = 336$ (2 weeks of hourly data), $m = 24, 48, 72, 96$ (1,2,3,4 days of hourly data) and $p = 487$ (151 weather variables and 336 periodic variables).

The elements of $\boldsymbol{\beta} \in \mathbb{R}^p$ are i.i.d. from the uniform distribution[3] $U[-1, 1]$. Moreover, as properties of the LASSO estimator depend on the sparsity of $\boldsymbol{\beta}$, we assume $s = (1 - \|\boldsymbol{\beta}\|_0/p) = 50\%$. Supplementary online material contains results for sparsity $s = 90\%$ and 20%, i.e., for high and low sparsity set-up. Throughout the experiment, we keep the (sparse) $\boldsymbol{\beta}$ fixed for all 1000 repetitions. We compute PI for nominal coverage $(1 - \alpha) = 60\%, 80\%, 90\%, 95\%$ and compare the quenched CLT method and QTL method proposed in Section 2, based on their coverage probabilities (CP hereafter with the plural being CPs)

$$\widehat{(1 - \alpha)} = \frac{1}{1000} \sum_{j=1}^{1000} \mathbb{I}\left([L, U]_{j, \hat{\boldsymbol{\beta}}} \ni \bar{y}_{j, +1:m}\right),$$

and based on the average Winkler loss (Winkler, 1972)

$$\mathcal{L} = |[L, U]_{j, \hat{\boldsymbol{\beta}}}| + \frac{2}{\alpha} \inf_{z \in [L, U]_{j, \hat{\boldsymbol{\beta}}}} |\bar{y}_{j, +1:m} - z|, \quad j = 1, \ldots, 1000,$$

where $\mathbb{I}$ for the j-th trial is 1 when $\bar{y}_{j, +1:m}$ is covered by the interval $[L, U]_{j, \hat{\boldsymbol{\beta}}}$ and 0 otherwise. The Winkler loss is a commonly used quantile-type loss, which penalizes the width of the PI and the size of misses, thus being appropriate for comparing PIs based on known quantiles (see Askanazi et al., 2018, for a recent discussion). Notably, the higher is the nominal coverage $1 - \alpha$ the more weight is assigned to the misses via the factor $2/\alpha$.

5.2 Simulation results

Two general conclusions may be drawn from the experiment, independent of whether the dimensionality of covariates is high or low:

---

[2]We reduce n for computational convenience so that the $p > n$ does not have to be very large.

[3]Previous version of this paper also contained results based on Cauchy distribution. Leading to the same general conclusions, we have omitted them for the sake of brevity.

- For all methods, their CP is always below the nominal coverage for which they were calibrated.

- Moreover, CP sinks with the growing forecast horizon. By contrast, the Winkler loss does not increase monotonically with the growing horizon, which indicates stability in terms of the trade-off between CP and "sharpness" across these forecast horizons.

Further results require the distinction of the low-dimensional and high-dimensional scenarios. In the low dimensional scenario (see Table 1i), in general, the following holds:

- PIs have the lowest CP when the underlying process has a long memory.

- QTL PIs dominate the quenched CLT across all series both in terms of CP and Winkler loss. Hence, CLT cannot compensate out the CP by "sharpness". The CP can become lower than half the nominal coverage for the long horizon.

- QTL PIs perform best when based on LASSO and LAD estimators. While LASSO QTL dominates in terms of CP, the LAD CTL has better "sharpness", leading to a slight preference of LAD over LASSO based on Winkler loss. The exception from this rule is when series exhibit long memory. Allowing higher sparsity in $\boldsymbol{\beta}$ would make LASSO the Winkler loss winner, while lower sparsity would empower the LAD.

For the high-dimensional scenario, some of the previous statements do not hold. Similar to the design in Chudý et al. (2020), the horizon/sample ratios become very high, i.e., $m/n > 1/4$. Moreover, we must face the curse of dimensionality for which we use LASSO[4]. Still, this set-up has a largely negative impact on QTL's performance (see Table 1ii), and therefore we exploit a data-driven adjustment based on replication of the residual $\hat{e}_i = y_i - \hat{y}_i$ using stationary bootstrap. We denote the adjusted QTL by ADJ. Details concerning implementation are in Section 6, where ADJ is used under similar a set-up as here. The simulations provide us with the following results:

- QTL PIs do not dominate the quenched CLT across all series and horizons. In fact, for the shortest horizon, CLT wins in terms of CP of Winkler loss (at least for the large nominal coverage). Moreover, for long-memory series, CLT wins across all horizons.

---

[4]Details concerning the selection of tuning parameter λ for LASSO can be found in Appendix C

- In general, the CP of QTL is worse than in the previous set-up, especially if the forecast horizon is long.

- However, the bootstrap adjustment introduced in Section 2.5 leads to major improvement across all series and all horizons (except the shortest one). In terms of Winkler loss, the improvement is most visible in the case of non-linear series.

Additionally, we may wonder what the impact of the sparsity and the generating distribution of $\boldsymbol{\beta}$ is. In general, the CPs are slightly higher when $\boldsymbol{\beta}$ is very sparse. In turn, $\boldsymbol{\beta}$ drawn from the Cauchy distribution leads to slightly smaller CPss. Still, both alternative set-ups lead to the same general conclusions with differences between CPss of identical methods (for identical series and horizons) within the range of two percentage points.

Table (i): Scenario $n > p$. QTL implemented using each of the estimators OLS, LAD or LASSO and CLT using LASSO only.

nominal	(e_i)	m-weeks	Coverage short-heavy 1	2	3	4	Coverage long-heavy 1	2	3	4	Coverage non-lin-heavy 1	2	3	4	Winkler short-heavy 1	2	3	4	Winkler long-heavy 1	2	3	4	Winkler non-lin-heavy 1	2	3	4
60%		ols-qtl	47.0	41.0	38.0	37.3	38.9	33.1	28.8	25.9	48.6	42.0	41.0	37.9	2.1	1.7	1.5	1.4	7.0	6.4	6.5	6.6	1.5	1.2	1.1	1.0
		lad-qtl	51.5	**51.1**	48.5	**45.9**	38.0	33.4	32.1	29.2	50.7	**51.8**	**49.1**	**50.0**	**1.5**	**1.2**	**1.1**	**1.0**	6.0	5.5	5.6	5.6	**1.0**	**0.8**	**0.8**	**0.7**
		lss-qtl	**53.4**	50.3	**49.8**	36.7	**55.4**	**53.0**	**54.6**	**48.8**	**55.0**	49.5	42.6	43.1	1.6	1.4	1.2	1.7	**6.0**	**5.2**	**5.2**	**5.1**	1.2	1.0	1.2	1.1
		lss-clt	48.2	35.9	35.2	19.3	43.9	32.3	28.4	23.7	48.8	33.7	23.0	23.1	1.7	1.6	1.4	2.1	6.1	5.6	5.6	5.7	1.3	1.2	1.4	1.3
80%		ols-qtl	65.9	60.5	55.6	52.2	56.5	50.1	44.6	39.3	67.4	62.5	59.5	54.2	3.0	2.5	2.2	2.0	10.3	9.7	10.0	10.4	2.2	1.8	1.6	1.5
		lad-qtl	74.5	71.5	71.0	**68.7**	61.4	53.9	53.2	49.0	73.7	72.3	**72.4**	**69.7**	**2.1**	**1.7**	**1.6**	**1.5**	8.8	8.0	8.6	8.6	**1.5**	**1.2**	**1.1**	**1.1**
		lss-qtl	**77.0**	**74.6**	**73.5**	60.4	**75.1**	**74.9**	**73.0**	**70.2**	**75.3**	**74.3**	68.3	66.6	2.3	1.9	1.8	2.2	**8.5**	**7.4**	**7.7**	**7.6**	1.6	1.4	1.5	1.4
		lss-clt	70.1	54.1	51.4	29.3	60.5	47.5	43.7	35.4	65.1	51.0	36.6	35.9	2.4	2.2	2.0	3.2	8.9	8.3	8.7	9.0	1.8	1.7	2.1	1.9
90%		ols-qtl	78.4	73.8	70.2	65.6	70.1	63.7	60.8	52.6	78.7	74.4	71.9	68.1	4.3	3.6	3.3	3.2	14.6	14.1	15.2	16.2	2.9	2.6	2.4	2.3
		lad-qtl	87.3	84.1	84.1	**83.2**	76.9	73.0	71.9	65.6	87.4	83.9	**85.8**	81.5	**3.1**	**2.5**	2.9	**2.5**	12.2	11.5	13.4	13.5	**2.1**	**1.8**	**2.1**	**1.9**
		lss-qtl	**87.7**	**87.5**	**85.7**	75.2	**86.5**	**86.1**	**83.5**	**81.6**	**88.3**	**85.4**	82.9	**81.9**	3.2	2.6	3.0	3.2	**11.7**	**10.8**	**11.8**	**11.6**	2.2	1.9	2.4	2.2
		lss-clt	79.7	68.2	63.8	39.2	71.0	59.2	54.5	46.8	75.8	64.5	45.7	45.4	3.3	3.1	**2.8**	5.0	12.7	12.4	13.6	14.2	2.5	2.4	3.2	3.0
95%		ols-qtl	86.3	82.4	76.4	72.1	81.9	75.1	69.1	60.7	86.8	82.2	78.6	73.3	6.1	5.5	5.0	4.9	20.5	21.2	24.1	26.2	3.9	4.1	3.7	3.6
		lad-qtl	92.7	91.2	88.1	**87.1**	87.0	83.6	77.2	71.4	92.7	90.9	**89.5**	**86.4**	**4.5**	4.5	**3.9**	**3.5**	16.8	17.7	20.0	20.4	**3.1**	**3.1**	**2.9**	**2.7**
		lss-qtl	**93.5**	**92.3**	**90.1**	80.9	**92.6**	**90.6**	**87.6**	**86.8**	**92.8**	**91.7**	87.8	**86.4**	4.6	4.6	4.0	4.3	**16.5**	**16.6**	**16.8**	**16.2**	**3.1**	**3.1**	3.2	3.0
		lss-clt	87.5	76.6	73.8	46.3	78.7	67.0	62.2	53.7	82.4	71.5	54.5	55.6	4.8	**4.4**	4.2	7.9	18.7	18.9	22.0	23.0	3.4	3.4	4.9	4.7

Table (ii): Scenario $p > n$. QTL and CLT as above. ADJ is a bootstrap version QTL for better performance under short-sample.

nominal	β / (e_i)	m-days	Coverage short-heavy 1	2	3	4	Coverage long-heavy 1	2	3	4	Coverage non-lin-heavy 1	2	3	4	Winkler short-heavy 1	2	3	4	Winkler long-heavy 1	2	3	4	Winkler non-lin-heavy 1	2	3	4
60%		lss-qtl	50.5	**49.4**	47.0	41.3	52.7	47.1	**43.9**	38.7	**51.3**	47.7	47.2	40.2	**4.8**	**3.8**	3.8	3.4	8.6	7.7	7.8	7.8	**3.1**	**2.5**	2.5	2.3
		lss-clt	**54.6**	39.6	37.2	32.2	**68.4**	**49.5**	42.2	33.8	48.3	35.3	33.6	27.8	5.5	4.9	4.9	4.3	**8.0**	**7.1**	7.4	7.5	3.9	3.6	3.6	3.5
		lss-adj	50.2	48.0	**48.0**	**47.9**	50.9	46.5	43.0	**40.0**	51.1	**48.6**	**48.9**	**48.6**	4.9	3.9	**3.7**	**3.3**	8.4	7.2	**6.9**	**6.8**	**3.1**	**2.5**	**2.4**	**2.1**
80%		lss-qtl	71.8	**66.6**	61.2	54.1	70.0	64.2	56.6	50.0	71.2	65.7	62.4	55.0	**7.4**	6.2	**5.4**	5.1	13.5	12.0	11.8	11.7	4.6	4.0	3.6	3.3
		lss-clt	**73.9**	56.8	51.7	48.9	**81.6**	**68.0**	**59.3**	51.2	70.3	55.7	47.8	40.7	7.8	6.7	7.1	6.3	**12.5**	**10.5**	10.8	11.1	5.3	5.2	5.2	5.3
		lss-adj	71.4	65.5	**63.5**	**62.6**	68.9	63.7	58.0	**54.7**	**72.0**	**67.7**	**67.4**	**64.9**	7.5	**6.1**	5.6	**4.9**	12.6	10.9	**10.7**	**10.5**	4.8	**4.0**	**3.5**	**3.1**
90%		lss-qtl	82.9	74.2	67.7	60.6	80.6	73.1	62.5	55.2	**82.1**	73.9	68.1	62.0	12.4	9.0	**8.0**	7.9	20.7	17.4	17.9	18.7	7.9	5.9	5.3	5.0
		lss-clt	**84.0**	69.7	63.5	59.9	**88.3**	**78.2**	**70.4**	62.9	81.1	68.7	59.6	52.3	**11.5**	**8.8**	9.9	8.9	**18.8**	**15.7**	**16.1**	16.7	7.4	6.7	7.4	7.8
		lss-adj	80.8	**75.8**	**74.2**	**72.5**	79.3	73.7	67.6	**64.5**	81.0	**77.9**	**76.2**	**73.9**	12.0	**8.8**	**8.0**	**7.3**	19.5	16.5	16.4	**16.5**	**7.3**	**5.7**	**5.0**	**4.5**
95%		lss-qtl	86.2	77.7	72.4	64.6	85.1	76.3	66.6	57.7	85.2	78.1	72.5	66.6	19.3	14.0	12.9	13.2	32.9	27.9	30.0	32.0	11.3	9.1	8.4	8.2
		lss-clt	**90.1**	80.1	74.0	70.4	**91.8**	**83.5**	**76.9**	**71.3**	85.8	77.8	70.6	62.6	**17.4**	**12.4**	13.6	12.5	**29.2**	**24.2**	**25.0**	**26.0**	**10.5**	9.0	10.4	11.5
		lss-adj	86.3	**81.9**	**79.8**	**78.1**	85.8	79.2	74.5	68.5	**86.0**	**84.2**	**81.6**	**80.7**	18.7	13.0	**11.8**	**11.3**	30.8	25.9	26.3	26.8	10.9	**8.4**	**7.5**	**6.8**

Table 1: Simulated out-of-sample forecasting experiment. The reported values are coverage probabilities, i.e., relative (%) counts of out-of-sample values covered in 1000 trials (left part) and average Winkler loss values (right part). The nominal coverage (the first column) ranks from is 95% to 60%. Simulated error processes heavy tails and either short memory, long memory or are nonlinear. The elements of regression coefficient β are drawn independently from uniform distribution $U[-1, 1]$. The sparsity of β's is fixed to 50%. For convenience, the best value for each nominal coverage and horizon is marked in bold.

6 Real data: EPEX Spot electricity prices

Next we compare and contrast forecasts obtained by our methods with the existing ones through POOS in a real-life data. Following is a list of competing methods we will explore:

(ADJ) Adjusted QTL-LASSO method described below in the Methods subsection 2.5.

(RBS) Robust Bayes (Müller and Watson (2016)),

(ARX) Bootstrap path simulation from ARMAX models,

(ETS) Exponential smoothing state-space model (Hyndman et al., 2008),

(NAR) Neural network autoregression (Hyndman and Athanasopoulos, 2013, sec. 9.3).

We first give a quick overview of the last three methods above so that they are comparable to the ADJ as specified in Section 2.5.

Robust Bayes PIs RBS:

For this sophisticated univariate approach, we focus on intuition and refer to the supplementary Appendix of Müller and Watson (2016) for more details about the implementation. The *robust Bayes* PIs are specifically designed for long-horizon predictions, e.g., when $m/n \approx 1/2$. First, the high-frequency noise is extracted out from y_t using low-frequency cosine transformation. Projecting $\bar{y}_{+1:m}$ on the space spanned by the first q frequencies is the key to obtaining the conditional distribution of $\bar{y}_{+1:m}$. In order to expand the class of processes for which this method can be used while keeping track of parameter uncertainty, Müller and Watson (2016) employed a Bayesian approach. In addition, the resulting PIs are further enhanced to attain the frequentist coverage using the least favorable distribution. This requires advanced algorithmic search for quantiles of non-standard distributions, which is its main drawback in terms of implementation. On the other hand, their supporting online materials provide some pre-computed inputs which make the computation faster.

 (i) For q small, compute the cosine transformations $\boldsymbol{x}^{\mathsf{T}} = (x_1, \ldots, x_q)$ of series y_t.

 (ii) Approximate the covariance matrix of $(\bar{y}_{+1:m}, \boldsymbol{x}^{\mathsf{T}})$.

(iii) Solve the minimization problem (14) in (Müller and Watson, 2016, page 1721) to get robust quantiles having uniform coverage.

25

(iv) The PIs are given by $[L, U] = \bar{y} + [Q_q^{\text{robust}}(\alpha/2), Q_q^{\text{robust}}(1 - \alpha/2)]$.

Bootstrap PIs for ARX, ETS and NAR:

(i) Adjust y_t for weekly periodicity using, e.g., seasonal and trend decomposition method proposed by Cleveland et al. (1990).

(ii) Perform automatic model selection based on AIC and fit the respective model to adjusted y_t. For ARX and NAR, we also use aggregated weather data defined as $\bar{w}_t = \sum_{k=1}^{73} w_{k,t}$, $\bar{\tau}_t = \sum_{l=1}^{78} \tau_{l,t}$ and the weekend-dummy variables as exogenous covariates (see the supplementary Appendix C for details).

(iii) Simulate $b = 1, \ldots, B$ future paths $\hat{y}_{n,t}^b$ of length m from the estimated model.

(iv) Obtain respective quantiles from set of averages $\bar{\hat{y}}_{+,1:m}^b, b = 1, \ldots, B$.

6.1 Data description and goal

We forecast $\bar{y}_{+1:m} = 1/m \sum_{t=1}^{m} y_{n+t}$, i.e. the average of m future hourly day-ahead spot electricity prices for Germany and Austria - the largest market in the European Power Exchange (EPEX SPOT). One of the reasons why we decided to forecast future averages was that the Bayes approach of Müller and Watson (2016) is designed specifically for the means. Since all other methods are flexible, we used the means as a common basis for the comparison.

The prices arise from day-ahead hourly auctions where traders trade for specific hours of the next day. With the market operating 24 hours a day, we have 11640 observations between 01/01/2013 00:00:00 UTC[5] and 04/30/2014 23:00:00 UTC. We split the data into a training period spanning from 01/01/2013 00:00:00 UTC to 12/31/2013 23:00:00 UTC and an evaluation period spanning from 01/01/2014 00:00:00 UTC to 04/30/2014 23:00:00 UTC (see Figure 1A). The forecasting horizon is $m = 1, \ldots, 17$ weeks ($168, \ldots, 2856$ hours).

Inspection of the periodogram for the prices in Figure 1C reveals peaks at periods 1 week, 1 day and 1/2 day. The mixed seasonality is difficult to model by SARIMA or ETS models which are suitable for monthly and quarterly data or by dummy variables. Instead, we use sums of sinusoids

$$g_t^k = R\sin(\omega_k t) + \phi) = \boldsymbol{\beta}_k^{(s)}(R, \phi)\sin(\omega_k t) + \boldsymbol{\beta}_k^{(c)}(R, \phi)\cos(\omega_k t)$$

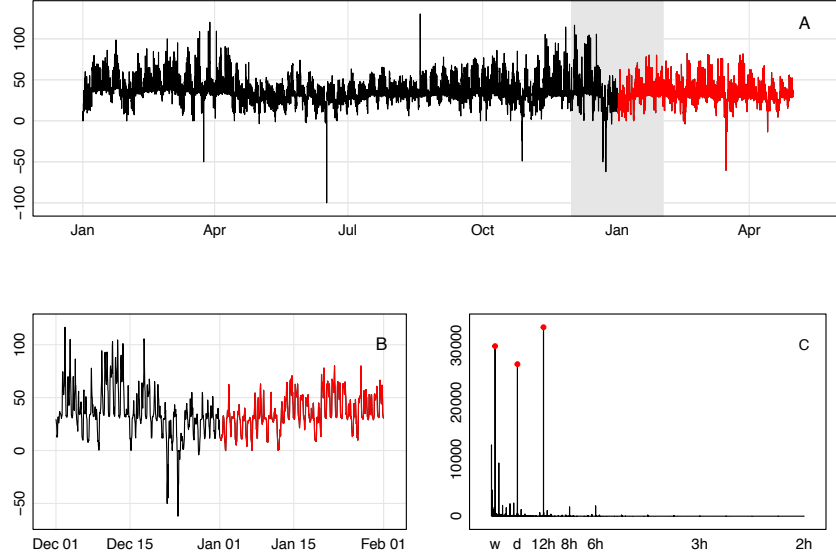
[5]Coordinated Universal Time.

26

Figure 1: Electricity spot prices, A) Full sample, B) Drop in price level, C) Periodogram with peaks at periods 1 week, 1 day and 12 hours.

with seasonal Fourier frequencies $\omega_k = 2\pi k/168$, $k = 1, 2, \ldots, \frac{168}{2}$ corresponding to periods 1 week, 1/2 week, ... , 2 hours (see Bierbauer et al., 2007; Weron and Misiorek, 2008; Cartea and Figureoa, 2005). The coefficients of linear combination $\boldsymbol{\beta}_k^{(s)}, \boldsymbol{\beta}_k^{(c)}$ can be estimated by least squares. In addition, we use 2 dummy variables as indicators for all weekends.

As mentioned in Section 1, the local weather variables are also used as covariates. The weather conditions implicitly capture seasonal patterns longer than a week, which is very important when forecasting long horizons. Local weather is represented by 151 hourly wind speed, and temperature series is observed throughout 5 years (2009-2013), i.e., including the training period but not the evaluation period (see above). In order to approximate some missing in-sample data and unobserved values for the evaluation period, we take hourly-specific-averages[6] of each weather series over these 5 years. In total, we have 168 trigonometric covariates, 151 weather covariates and 2 dummies which gives a full set of 321 covariates. We denote the covariates

$$\boldsymbol{x}_t^\mathsf{T} = (d_{\mathrm{sa}}, d_{\mathrm{su}}, \sin(\omega_1 t), \cos(\omega_1 t), \ldots \sin(\omega_{84} t), \cos(\omega_{84} t), w_{1,t}, \ldots, w_{73,t}, \tau_{1,t}, \ldots, \tau_{78,t}),$$

for $t = 1, \ldots, n$, with d as dummies for weekend, w_k, and τ_l as the wind speed and temperature measured at the k-th, and l-th weather stations.

---

[6]See alternative approximation of future values by bootstrap (in Hyndman and Fan, 2010)

27

6.2 Discussion on the performance of different methods

Before we compare the ADJ with the other competitors, we will address a few issues that are usual with analysis of any real-life datasets. In Figure 1B, we see a drop in the price level during December 2013. The forecasts based on the whole training period would therefore suffer from bias. By contrast, using only the post-break December data would mean a loss of potentially valuable information. An optimal trade-off in such situations can be achieved by down-weighing older observations (see Pesaran et al., 2013), also called exponentially weighted regression (Taylor, 2010). In order to achieve better forecasting performance, we use the exponentially weighted regression with standardized exponential weights $v_{n-t+1} = \delta^{t-1}((1-\delta))/(1-\delta^t)$, $t = 1, \ldots, n$ and with $\delta = 0.8$. This applies to ADJ and NAR methods. The ETS and ARX models provide exponential down-weighing implicitly, but with optimally selected weights. Müller and Watson (2016) showed that the RBS is robust to structural changes. We would also like to see if there are actual benefits from using disaggregated weather data instead of weather data aggregated across the weather stations. Therefore, we compute the ADJ PIs using no regressors as in Figure 2A, using only deterministic regressors as in Figure 2B, using deterministic regressors and aggregated weather variables defined as $\bar{w}_t = \sum_{k=1}^{73} w_{k,t}$, $\bar{\tau}_t = \sum_{l=1}^{78} \tau_{l,t}$ as in Figure 2C and finally, using all 321 covariates as in Figure 2D. As we can see, there is only very little difference between the first three plots, which means that using only deterministic regressors with or without the aggregated weather data does not prevent the bias at the end of the evaluation period. On the other hand, if we use the disentangled local weather data, significant improvement is achieved.

Finally, we get to the comparison with the alternative PIs denoted as RBS, ETS, NAR, and ARX. All these PIs are given in Figure 3. Of the four methods, only RBS gives sensible PIs. RBS works consistently well over the whole 17-weeks-long evaluation period (Figure 3A). However, when compared to the ADJ, the p.i's seem too conservative. Hence the ADJ provides more precision on top of decent coverage. Prediction intervals by ETS get too conservative as the horizon grows and do not provide a valid alternative to ADJ. The NAR is even more biased than the ADJ without covariates, especially for large m. Not so surprisingly, the ARX perform worst of all methods, presumably because the exponential down-weighing implied by the simple autoregression is too mild. Besides, the narrow PIs are the result of ignoring the parameter (among other types of) uncertainty.
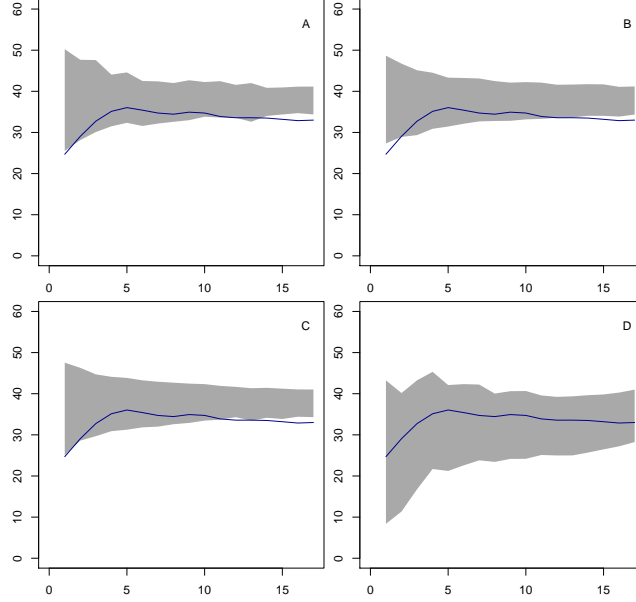
Figure 2: A) ADJ (without covariates), B) ADJ with 170 deterministic covariates, C) ADJ with 170 deterministic covariates & 2 aggregated (across stations) weather series, D) ADJ with 170 deterministic covariates& 151 disaggregated weather series.

7 Conclusion

We constructed quantile-based prediction intervals in a regression framework. From a theoretical perspective, we have extended the results of Zhou et al. (2010) to high-dimensional set-up and also to the case of the nonlinear error process. We showed the consistency of fitted residuals for the ultra-high dimensional case $\log p = o(n)$. Under some mild conditions on the fixed or stochastic covariate process, we were able to establish quantile consistency for the normalized average of fitted residuals and thus provide a significant extension to theoretical validity of the non-parametric and simple quantile-based prediction intervals. The quantile method has been additionally adjusted for short sample and long horizon and was successfully applied to predict spot electricity prices for Germany and Austria using a large set of local weather time series. The results have shown the superiority of the adjusted method over selected conventional methods and approaches such as exponential smoothing, neural networks as well as the recently proposed low-frequency approach of Müller and Watson (2016).

Regarding future work, some interesting extensions can include multivariate target series and subsequent construction of simultaneous prediction intervals. Appli-
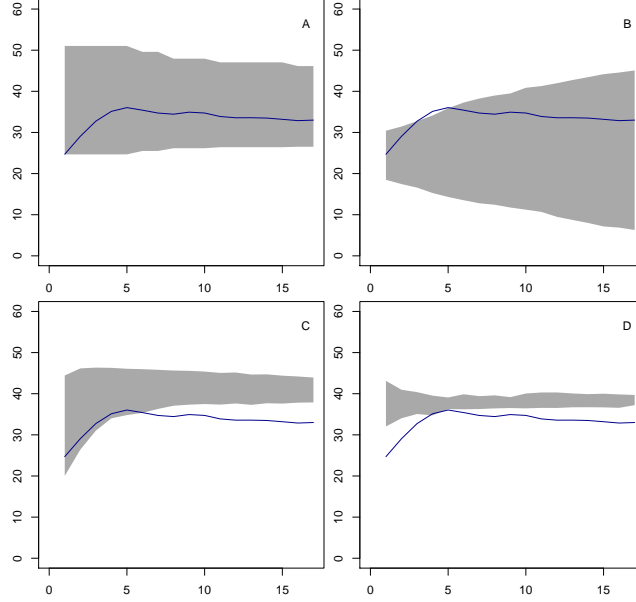
Figure 3: A) robust Bayes method, B) exponential smoothing state space model (A,N,N) with tuning parameter 0.0446, C) neural network autoregression (38,22) with one hidden layer D) ARMAX. For C) and D) weekend dummies & aggregated weather are used as exogenous covariates.

cations of such simultaneous intervals could include the prediction of spot electricity prices for each hour simultaneously in the spirit of Raviv et al. (2015).

8 Data Availability Statement

We are thankful to Stefan Feuerriegel for providing data from their paper Ludwig et al. (2015) for the analysis in a direct communication.

9 Acknowledgement

We are thankful to two anonymous referees and the editor for their helpful comments and corrections that helped improve the presentation of this paper significantly.

References

Askanazi, R., F. X. Diebold, F. Schorfheide, and M. Shin (2018). On the comparison of interval forecasts. *Journal of Time Series Analysis 39*(6), 953–965.

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

Belyaev, Y. (1968). K., on the number of exits across the boundary of a region by a vector stochastic process, theor. *Probab. Appl 13*, 320–324.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist. 37*(4), 1705–1732.

Bierbauer, M., C. Menn, S. Rachev, and S. Trück (2007). Spot and derivative pricing in the eex power market. *Journal of Banking & Finance 31*(11), 3462–3485.

Brockwell, P. J., R. A. Davis, and S. E. Fienberg (1991). *Time series: theory and methods: theory and methods*. Springer Science & Business Media.

Cartea, A. and M. Figureoa (2005). Pricing in electricity markets: a mean reverting jump diffusion model with seasonality. *Applied Mathematical Finance 12*(4), 313–335.

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

Chudý, M., S. Karmakar, and W. B. Wu (2020). Long-term prediction intervals of economic time series. *Empirical Economics 58*(1), 191–222.

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

Cleveland, R. B., W. S. Cleveland, M. J. E., and I. Terpenning (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics 6*, 3–73.

Cover, T. M. (1975). Open problems in information theory. In *1975 IEEE Joint Workshop on Information Theory*, pp. 35–36. IEEE Press.

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.

Fryzlewicz, P., T. Sapatinas, and S. Subba Rao (2008). Normalized least-squares estimation in time-varying ARCH models. *Ann. Statist. 36*(2), 742–786.

Györfi, L., W. Härdle, P. Sarda, and P. Vieu (2013). *Nonparametric curve estimation from time series*, Volume 60. Springer.

Gyorfi, L., G. Lugosi, and G. Morvai (1998). A simple randomized algorithm for consistent sequential prediction of ergodic time series.

Gyorfi, L. and G. Ottucsak (2007). Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory 53*(5), 1866–1872.

Hannan, E. (1979). The central limit theorem for time series regression. *Stochastic Processes and their Applications 9*(3), 281–289.

Huber, P. J. and E. M. Ronchetti (2009). *Robust statistics* (Second ed.). Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ.

Huurman, C., F. Ravazzolo, and C. Zhou (2012). The power of weather. *Computational Statistics & Data Analysis 56*(11), 3793–3807.

Hyndman, R. J. and G. Athanasopoulos (2013). *Forecasting: principles and practice.* OTexts: Melbourne, Australia. Accessed on 12/12/2017.

Hyndman, R. J. and S. Fan (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems 25*(2), 1142–1153.

Hyndman, R. J. and Y. Khandakar (2008). Automatic time series forecasting: the forecast package for r. *Journal of Statistical Software 27*(1), 1–22.

Hyndman, R. J., A. B. Koehler, J. K. Ord, and R. D. Snyder (2008). *Forecasting with Exponential Smoothing: The State Space Approach.* Berlin: Springer-Verlag Berlin Heidelberg.

Karmakar, S., S. Richter, and W. B. Wu (2020+). Simultaneous inference for time-varying models. *In revision,* *https://sayarkarmakar.github.io/publications/sayar1.pdf*.

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

Knittel, C. R. and M. R. Roberts (2005). An empirical examination of restructured electricity prices. *Energy Economics 27*(5), 791–817.

Ludwig, N., S. Feuerriegel, and D. Neumann (2015). Putting big data analytics to work: Feature selection for forecasting electricity prices using the lasso and random forests. *Journal of Decision Systems 24*(1), 19–36.

Lundbergh, S., T. Teräsvirta, and D. van Dijk (2003). Time-varying smooth transition autoregressive models. *Journal of Business & Economic Statistics 21*(1), 104–121.

Morvai, G. (2003). Guessing the output of a stationary binary time series. In *Foundations of Statistical Inference*, pp. 207–215. Springer.

Morvai, G., S. Yakowitz, L. Györfi, et al. (1996). Nonparametric inference for ergodic, stationary time series. *The Annals of Statistics 24*(1), 370–379.

Morvai, G., S. J. Yakowitz, and P. Algoet (1997). Weakly convergent nonparametric forecasting of stationary time series. *IEEE Transactions on Information Theory 43*(2), 483–498.

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

Nagaev, S. V. (1979). Large deviations of sums of independent random variables. *Ann. Probab. 7*(5), 745–789.

Pesaran, M. H., A. Pick, and M. Pranovich (2013). Optimal forecasts in the presence of structural breaks. *Journal of Econometrics 177*(2), 134–152.

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

Pourahmadi, M. (2001). *Foundations of time series analysis and prediction theory*, Volume 379. John Wiley & Sons.

Raviv, E., K. E. Bouwman, and D. van Dijk (2015). Forecasting day-ahead electricity prices: Utilizing hourly prices. *Energy Economics 50*, 227–239.

Rio, E. (2009). Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability 22*(1), 146–163.

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

Starica, C. (2003). Is garch (1, 1) as good a model as the accolades of the nobel prize would imply? *Available at SSRN 637322*.

Taylor, J. W. (2010). Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles. *International Journal of Forecasting 26*(4), 627–646.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting 30*(4), 1030 – 1081.

Weron, R. and A. Misiorek (2008). Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *International Journal of Forecasting 24*(4), 744–763. Energy Forecasting.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association 67*(337), 187–191.

Wu, W. B. (2005a). Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA 102*(40), 14150–14154 (electronic).

Wu, W. B. (2005b). On the Bahadur representation of sample quantiles for dependent sequences. *Ann. Statist. 33*(4), 1934–1963.

Wu, W. B. (2007). *M*-estimation of linear models with dependent errors. *Ann. Statist. 35*(2), 495–521.

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

Wu, W. B. and Y. N. Wu (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat. 10*(1), 352–379.

Zhang, D. and W. B. Wu (2021). Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *The Annals of Statistics 49*(1), 233–254.

Zhou, Z. and W. B. Wu (2009). Local linear quantile estimation for nonstationary time series. *Ann. Statist. 37*(5B), 2696–2729.

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:-Some useful concentration inequalities

Throughout, we use properties of the following projection operator $\mathcal{P}_\rangle$ defined for a sequence of random variable Y_i adapted to a σ-field filtration $\mathcal{F}_i$

$$\mathcal{P}_i Y = \mathbb{E}(Y|\mathcal{F}_i) - \mathbb{E}(Y|\mathcal{F}_{i-1}), \quad Y \in \mathcal{L}_1. \tag{9.1}$$

In particular, one particular inequality that allows an easy transfer from a linear case to a corresponding non-linear case is

$$\|\mathcal{P}_i Y_j\|_p \leq \delta_{i,p}^Y$$

with $\delta_{i,p}^Y$ standing for the dependence measure as defined in (3.11) for the mean-zero stationary process Y_j. Similar inequality holds for other functional dependence measures we defined in the paper. For a proof, the interested readers can look at the seminal text by Wu (2005a) where functional dependence measure was formally introduced. Apart from this, we have also repeatedly used the orthogonality of the projection operator $\mathcal{P}_i$ defined in (9.1) i.e. $\mathbb{E}(\mathcal{P}_i(X_k)\mathcal{P}_j(X_k)) = 0$ for $i \neq j$. With these tools in hand, we proceed to describe the proofs in this paper.

Let $S_{n,b} = \sum_{i=1}^n b_i e_i$. Only the light-tailed versions of the Results 9.1 and 9.2 were proposed and proved in Wu and Wu (2016). In this exposition, we clearly point out the corresponding results for the long-range and short-range dependence along with the new addition of what happens for light-tailed or heavy-tailed scenarios. These results can be of independent interest since this is first such in the literature that takes care of heaviness of tails under dependence.

Result 9.1. *(Nagaev inequality for linear processes) Assume that the error process e_i admits the representation (3.3). Then we have the following concentration results for $S_{n,b} = \sum_{i=1}^n b_i e_i$,*
- Light-tailed SRD: If $\sum_j |a_j| < \infty$ and $\epsilon_j \in \mathcal{L}^q$ for some $q > 2$, then, for some constant c_q,

$$\mathbb{P}(|S_{n,b}| \geq x) \leq (1 + 2/q)^q \frac{\|b\|_q^q (\sum_j |a_j|)^q \|\epsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{c_q x^2}{n(\sum_j |a_j|)^2 \|\epsilon_0\|_2^2}\right),$$

- Light-tailed LRD: If $K = \sum_j |a_j|(1+j)^{\boldsymbol{\beta}} < \infty$ for $0 < \boldsymbol{\beta} < 1$ and $\epsilon_j \in \mathcal{L}^q$ for some $q > 2$, then, for some constant C_1, C_2 depending on only q and $\boldsymbol{\beta}$,

$$\mathbb{P}(|S_{n,b}| \geq x) \leq C_1 \frac{K^q \|b\|_q^q \|n^{q(1-\boldsymbol{\beta})}\epsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{C_2 x^2}{n^{3-2\boldsymbol{\beta}}\|\epsilon_0\|_2^2 K^2}\right), \tag{9.2}$$

- *Heavy-tailed SRD: If $\sum_j |a_j| < \infty$ and $\epsilon_j \in \mathcal{L}^q$ for some $1 < q \leq 2$, then, for some constant c_q*

$$\mathbb{P}(|S_{n,b}| \geq x) \leq c_q \frac{\|b\|_q^q (\sum_j |a_j|)^q \|\epsilon_0\|_q^q}{x^q}, \tag{9.3}$$

- *Heavy-tailed LRD: If $K = \sum_j |a_j|(1+j)^\beta < \infty$ for $0 < \beta < 1$ and $\epsilon_j \in \mathcal{L}^q$ for some $q > 2$, then, for some constants C_1, C_2 depending only on q and β,*

$$\mathbb{P}(|S_{n,b}| \geq x) \leq C_1 \frac{K^q \|b\|_q^q \|n^{q(1-\beta)} \epsilon_0\|_q^q}{x^q}. \tag{9.4}$$

The proofs of the Theorem 3.1 and Theorem 3.2 heavily rely on the representation in (3.3). Therefore it is important to explore a new approach to prove the analogous result for nonlinear processes.

Result 9.2. *(Nagaev inequality for nonlinear processes) Assume $\|e.\|_{q,\alpha} < \infty$ for some $\alpha > 0$,*
- *Light-tailed SRD:- Assume that $\|e.\|_{q,\alpha} < \infty$ where $q > 2$ and $\alpha > 0$ and $\sum_{i=1}^n b_i^2 = n$. Let $r_n = 1$(resp. $(\log n)^{1+2q}$ or $n^{q/2-1-\alpha q}$) if $\alpha > 1/2 - 1/q$ (resp. $\alpha = 0$ or $\alpha < 1/2 - 1/q$). Then for all $x > 0$, for constants C_1, C_2, C_3 that depend on only q and α,*

$$\mathbb{P}(|S_{n,b}| \geq x) \leq C_1 \frac{r_n}{(\sum_j |b_j|)^q \|e.\|_{q,\alpha}^q} x^q + C_2 \exp\left(-\frac{C_3 x^2}{n\|e.\|_{2,\alpha}^2}\right), \tag{9.5}$$

- *Heavy-tailed SRD:- Assume that $\|e.\|_{q,\alpha} < \infty$ where $1 < q < 2$ and $\alpha > 0$ and $\sum_{i=1}^n b_i^2 = n$. Let $r_n = 1$(resp. $(\log n)^{1+2q}$ or $n^{q/2-1-\alpha q}$) if $\alpha > 1/2 - 1/q$ (resp. $\alpha = 0$ or $\alpha < 1/2 - 1/q$). Then for all $x > 0$, for constants C_1 that depend on only q and α,*

$$\mathbb{P}(|S_{n,b}| \geq x) \leq C_1 \frac{r_n}{(\sum_j |b_j|)^q \|e.\|_{q,\alpha}^q} x^q. \tag{9.6}$$

Sketch for the proof of Result 9.1 and 9.2. The details are omitted for these proofs since the steps are similar to that of Theorem 1 and Theorem 2 in Wu and Wu (2016). Very briefly, the basic strategy lies in writing $S_{n,b}$ with the following decomposition

$$S_{n,b} = S_{n,b,0} + (S_{n,b} - S_{n,b,n}) + \sum_{l=1}^{\lceil \log n/\log 2) \rceil} \sum_{k=1}^{n} b_k (e_{k,\min(n,2^l)} - e_{k,2^{l-1}}),$$

where $e_{i,\tau} = \mathbb{E}(e_i | \epsilon_{i-\tau}, \ldots, \epsilon_i)$ and $S_{n,b,m} = \sum_{k=1}^n b_k e_{k,m}$.

36

With this decomposition, the summands in the first term are independent and hence Nagaev inequality in Corollary 1.7 from Nagaev (1979) can be applied. For the second term Burkholder and Hölder inequalities are used whereas for the third term due to 2^l-dependence we divide the summands in independent odd and even parts and proceed with Corollary 1.7 from Nagaev (1979) again. For the heavy-tailed scenarios, note that the exponential terms from the light-tailed are not presented separately for the interest of space since one just needs to use corollary 1.6 instead of corollary 1.7 from Nagaev (1979).

$\square$

The proof for the quantile consistency results for the lasso-fitted residuals in Section 4 requires the following important result.

Lemma 9.3. *Assume the the restricted eigenvalue condition (4.2) with κ as presented in Theorem 4.1. Then on the event $\mathcal{A} = \bigcap_{j=1}^p \{2|V_j| \le r\}$, where $V_j = \frac{1}{n}\sum_{i=1}^n e_i x_{ij}$, we have,*

$$\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2/n \le 16 s r^2/\kappa^2, \quad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \le 4 s r/\kappa^2. \tag{9.7}$$

Proof of Lemma 9.3. Since $\hat{\boldsymbol{\beta}}$ minimizes the LASSO-objective in (2.6), plugging in $y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + e_i$ and $\lambda = 2r$, we have

$$2r\|\hat{\boldsymbol{\beta}}\|_1 + \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + e_i - x_i^\mathsf{T}\hat{\boldsymbol{\beta}})^2 \;\le\; 2r\|\boldsymbol{\beta}\|_1 + \frac{1}{n}\sum_{i=1}^n e_i^2$$

$$2r\|\hat{\boldsymbol{\beta}}\|_1 + \|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2/n \;\le\; 2r\|\boldsymbol{\beta}\|_1 + 2e^\mathsf{T} X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/n$$

$$r\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + \|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2/n \;\le\; 2r(\|\boldsymbol{\beta}\|_1 - |\hat{\boldsymbol{\beta}}\|_1) + r\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + 2e^\mathsf{T} X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/n.$$

These entail, on the event $\{|V_j| = |(e^\mathsf{T} X)_j|/n \le r/2 \text{ for all } j\}$

$$r\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + \|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2/n \le 4r\|\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J\|_1 \le 4r\sqrt{s}\|\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J\|_2, \tag{9.8}$$

where $J = \{j : \beta_j \ne 0\}$ as for $j \in J^c$, $|\hat{\beta}_j - \beta_j| + |\beta|_j - |\hat{\beta}|_j = 0$. However, also note that from the first inequality in (9.8), we have $\|\hat{\boldsymbol{\beta}}_{J^c} - \boldsymbol{\beta}_{J^c}\|_1 \le 3\|\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J\|_1$. Thus by (4.2), $\kappa\|\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J\|_2 \le \|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2/\sqrt{n}$. Plugging this in (9.8) immediately yields the first inequality of (9.7) whereas the second is derived by square completion.

$\square$

Lemma 9.4. *Recall the dependence adjusted and dimension incorporated functional dependence measure as defined in the context of Theorem 4.2. Assume $\Phi_{q,\alpha} < \infty$ for some $q \geq 1$ and $\alpha > 0$. For a fixed $\delta > 0$, assume $p, n \to \infty$ with $2C_\alpha \Phi_{4,\alpha}^4 s^2 \log p < n\delta^2$. Then,*

$$\mathbb{P}(|X^\mathsf{T} X/n - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})|_\infty > \delta/s) \to 0.$$

Proof of Lemma 9.4. We prove this by applying Hansen-Wright inequality as presented in Zhang and Wu (2021) on quadratic forms. Although they proved it for a more general locally stationary process and defined their functional dependence measure accordingly, it also holds for the simpler (not accounting for local stationarity) functional dependence measure defined in the context of Theorem 4.2. Apply Theorem 6.4 in Zhang and Wu (2021) for $B = 0$, $x = n\delta/s$, $a_t = 1$ to obtain

$$\mathbb{P}(|X^\mathsf{T} X - n\mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})|_\infty > n\delta/s) \quad \leq \quad Cp^2 \exp\left(-\frac{n\delta^2}{s^2 C_\alpha \Phi_{4,\alpha}^4}\right)$$

which goes to 0 due to our assumption on how p, s and n scale with each other.

$\square$

10 Appendix B:- Proofs of Theorems

Proof of Theorem 3.3. The m-dependence approximation is the key idea for the proof for the nonlinear case,

$$\|\mathbb{E}(\tilde{S}_m|\mathcal{F}_0) - \mathbb{E}(S_m|\mathcal{F}_0)\| \leq \|S_m - \tilde{S}_m\| \leq m^{1/2}\Theta_{m,p} \ll m^{1/2}/(\log m)^{5/2},$$

where $\tilde{S}_m = \sum_{i=1}^m \tilde{e}_i = \sum_{i=1}^m \mathbb{E}(e_i|\epsilon_i, \ldots \epsilon_{i-m})$. Note the following decomposition

$$\mathbb{E}(\tilde{S}_m|\mathcal{F}_0) \quad = \quad \sum_{j=-m}^0 \mathcal{P}_j \tilde{S}_m = \sum_{j=-\infty}^0 (\mathbb{E}(\tilde{S}_m|\mathcal{F}_j) - \mathbb{E}(\tilde{S}_m|\mathcal{F}_{j-1})).$$

The proof of the sufficient condition for convergence (3.1) then follows from $\|\mathcal{P}_j \tilde{e}_i\|_2 \leq \delta_{i-j,2}$ and the condition on $\Theta_{m,p}$.

$\square$

For the rest of the proofs we introduce some notations. Define

$$\tilde{Y}_i = H_m^{-1} \sum_{j=i-m+1}^i e_j \text{ for } i = m, m+1, \ldots, \tag{10.1}$$

and let $\tilde{Z}_i = \tilde{Y}_i - \mathbb{E}(\tilde{Y}_i|\mathcal{F}_{i-1})$. Define

$$\tilde{F}_n^*(x) = \frac{1}{n-m+1}\sum_{i=m}^{n}\mathbb{P}(\tilde{Y}_i \leq x).$$

Let $\tilde{F}(x) = \mathbb{P}(\tilde{Y}_i \leq x)$. Let $\tilde{F}_n(x)$ denote the empirical distribution function of $\tilde{Y}_i, i = m, \ldots n$. We use the following decomposition:

$$\tilde{F}_n(x) - \tilde{F}(x) = M_n(x) + N_n(x) \tag{10.2}$$

where $M_n(x) = \tilde{F}_n(x) - \tilde{F}_n^*(x)$ and $N_n(x) = \tilde{F}_n^*(x) - \tilde{F}(x)$. In the next series of lemma, we prove some consistency results involving $M_n(\cdot)$ and $N_n(\cdot)$. These are fundamental in the eventual proof of quantile consistency results from Section 3.2. The proof techniques are similar to that used in Zhou and Wu (2009), Zhou et al. (2010) adapted to the nonlinear setting and specifically exploiting the predictive density dependence defined in (3.14). Using the projection operator defined in (9.1), one can write $M_n(x)$ as follows

$$M_n(x) = \frac{1}{n-m+1}\sum_{i=m}^{n}\mathcal{P}_i I(\tilde{Y}_i \leq x). \tag{10.3}$$

Next we present two important lemmas concerning local equicontinuity of the two terms $M_n(\cdot)$ and $N_n(\cdot)$. Let f_ϵ is the density of the conditional distribution of $\tilde{Y}_i$ given $\mathcal{F}_{i-1}$.

Lemma 10.1. *Under conditions of Theorem 9.1 and Theorem 9.2,*

$$\sup_{|u|\leq b_n} |M_n(x+u) - M_n(x)| = O_{\mathbb{P}}\left(\sqrt{\frac{H_m b_n}{n}}\log^{1/2} n + n^{-3}\right),$$

where b_n is a positive bounded sequence with $\log n = o(H_m n b_n)$.

Proof. Note that, $\mathbb{P}(x \leq \tilde{Y}_i \leq x + u|\mathcal{F}_{i-1}) \leq H_m c_0 u$ for all $u > 0$ where $c_0 = \sup_x |f_\epsilon(x)| < \infty$. Therefore, the result in (10.4) follows by Freedman's martingale inequality and a chaining argument as for any $u \in [-b_n, b_n]$, we have

$$\sum_{i=m}^{n}[\mathbb{E}(V_i) - \mathbb{E}(V_i)^2] \leq c_0(n-m+1)H_m b_n \tag{10.4}$$

39

where $V_i = I(x \leq \tilde{Y}_i \leq x + u | \mathcal{F}_{i-1})$. We skip the details and direct interested readers to Lemma 5 in Wu (2005b), Lemma 4 in Wu (2007) and Lemma 6 in Zhou and Wu (2009).

$\square$

Lemma 10.2. *Under conditions of SRD, DEN and light-tailed*

$$\| \sup_{|u| \leq b_n} |N_n(x + u) - N_n(x)| \| = O\left(\frac{b_n m^{3/2}}{\sqrt{n}}\right). \tag{10.5}$$

Proof. From the definition of $N_n(x)$ in (10.2), we have

$$N_n(x + u) - N_n(x) = \sqrt{m} \frac{\int_0^u R_n(x + t) dt}{n - m + 1}, \tag{10.6}$$

where, for $x \in \mathbb{R}$,

$$R_n(x) = \sum_{i=m}^{n} [f_\epsilon(H_m(x - \tilde{Z}_{i-1})) - \mathbb{E}(f_\epsilon(H_m(x - \tilde{Z}_{i-1})))]. \tag{10.7}$$

Let $(\epsilon_i')_{-\infty}^{\infty}$ be an i.i.d. copy of $(\epsilon_i)_{-\infty}^{\infty}$. Let $Z_{i-1,k}^* = H(\epsilon_i, \epsilon_{i-1}, \ldots)$. Denote $\tilde{Z}_{i-1,k}^* = H(\epsilon_i, \epsilon_{i-1}, \ldots, \epsilon_{i-k}', \ldots)$. Also, introduce the coefficients $\tilde{b}_{j,q}$ as follows

$$\tilde{b}_{j,q} = \begin{cases} \psi_{0,q} + \psi_{1,q} + \ldots + \psi_{j,q} & \text{if } 1 \leq j \leq m - 1 \\ \psi_{j-m+1,q} + \psi_{j-m+2,q} + \ldots + \psi_{j,q} & \text{if } j \geq m. \end{cases} \tag{10.8}$$

Write $\tilde{b}_j$ for $\tilde{b}_{j,2}$. Then

$$\begin{aligned} \|\mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x + u - \tilde{Z}_{i-1}))\| &\leq \|f_\epsilon(\sqrt{m}(x + u - \tilde{Z}_{i-1})) - f_\epsilon(\sqrt{m}(x + u - \tilde{Z}_{i-1,k}^*))\| \\ &\leq \sup_{v \in \mathbb{R}} |f_\epsilon'(v)| \|\sqrt{m}(\tilde{Z}_{i-1} - \tilde{Z}_{i-1,k}^*)\| \leq c_1 \tilde{b}_k, \end{aligned} \tag{10.9}$$

for some $c_1 < \infty$. Further note that

$$R_n(x + u) = \sum_{k=1}^{\infty} \sum_{i=m}^{n} \mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x + u - \tilde{Z}_{i-1}))$$

and by the orthogonality of $\mathcal{P}_{i-k}, i = m, \ldots, n$

40

$$\| \sum_{i=m}^{n} \mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x + u - \tilde{Z}_{i-1})) \|^2 = \sum_{i=m}^{n} \| \mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x + u - \tilde{Z}_{i-1})) \|^2$$

$$\leq c_1^2 (n - m + 1) \tilde{b}_k^2.$$

Therefore, for all $u \in [-b_n, b_n]$,

$$\|R_n(x + u)\| \leq \sum_{k=1}^{\infty} \| \sum_{i=m}^{n} \mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x + u - \tilde{Z}_{i-1})) \|$$

$$\leq c_1 \sqrt{n} \sum_{k=1}^{\infty} |\tilde{b}_k| \leq c_1 m \sqrt{n} \sum_{j=0}^{\infty} |\psi_{j,2}| = O(m\sqrt{n})$$

in the view of the SRD condition in (3.15). This concludes the proof of (10.5) in the light of (10.6) and (10.7).

$\square$

Lemma 10.3. *Under conditions of LRD, DEN and heavy-tailed, we have for any* $\rho \in (1/\gamma, \alpha)$

$$\| \sup_{|u| \leq b_n} |N_n(x + u) - N_n(x)| \|_\rho = O\left(H_m b_n m n^{1/\rho - \gamma} |l(n)|\right). \tag{10.10}$$

Proof. Similar to the proof of Lemma 10.2, it suffices to prove, for some $0 < C < \infty$,

$$\|R_n(x + u)\|_\rho \leq C m n^{1/\rho + 1 - \gamma} |l(n)| \text{ for all } u \in [-b_n, 1 - b_n] \tag{10.11}$$

Since $1 < \rho < 2$, by (Rio (2009)) Burkholder type inequality of martingales, we have, with $C_\rho = (\rho - 1)^{-1}$.

$$\|R_n(x + u)\|_\rho^\rho = \| \sum_{k=-\infty}^{n-1} \mathcal{P}_k \sum_{i=m}^{n} f_\epsilon(H_m(x - \tilde{Z}_{i-1})) \|_\rho^\rho$$

$$\leq C_\rho \sum_{k=-\infty}^{n-1} \| \mathcal{P}_k \sum_{i=m}^{n} f_\epsilon(H_m(x - \tilde{Z}_{i-1})) \|_\rho^\rho$$

$$\leq C_\rho \sum_{k=-\infty}^{n-1} (\sum_{i=m}^{n} \| \mathcal{P}_k f_\epsilon(H_m(x - \tilde{Z}_{i-1})) \|_\rho)^\rho$$

$$\leq C_\rho \sum_{k=-\infty}^{-n} + \sum_{k=-n+1}^{0} + \sum_{k=1}^{n-1} (\sum_{i=m}^{n} \| \mathcal{P}_k f_\epsilon(H_m(x - \tilde{Z}_{i-1})) \|_\rho)^\rho$$

$$\leq C_\rho (I + II + III).$$

41

Since $\mathbb{E}|\epsilon_i|^\rho < \infty$, similarly as (10.9), we have for $k \le i - 1$ that

$$\|\mathcal{P}_k f_\epsilon(H_m(x - Z_{i-1}))\|_\rho \le c_1|\tilde{b}_{i-k}|, \tag{10.12}$$

for some $c_1 < \infty$. Thus using Karamata's theorem for the term I, we have

$$
\begin{aligned}
I &\le c_1^\rho \sum_{k=-\infty}^{-n} (\sum_{i=m}^{n} |\tilde{b}_{i-k}|)^\rho \le c_1^\rho \sum_{k=n}^{\infty} (m \sum_{i=1}^{n} |\psi_{k+i,\rho}|)^\rho \tag{10.13} \\
&\le c_1^\rho m^\rho n^{\rho-1} \sum_{k=n}^{\infty} \sum_{i=1}^{n} |\psi_{k+i,\rho}|^\rho \\
&= O[m^\rho n^{1+\rho(1-\gamma)}|l(n)|^\rho].
\end{aligned}
$$

Since $\rho > 1$ and $\rho\gamma > 1$, we use Hölder inequality to manipulate term III as follows:

$$
\begin{aligned}
III &\le c_1^\rho \sum_{k=1}^{n-1} (\sum_{i=\max(m,k+1)}^{n} |\tilde{b}_{i-k}|)^\rho \le c_1^\rho \sum_{k=1}^{n-1} (m \sum_{i=0}^{n-k} |\psi_{i,\rho}|)^\rho \tag{10.14} \\
&= m^\rho \sum_{k=1}^{n-1} O[(n-k)^{1-\gamma}|l(n-k)|]^\rho \\
&= O[m^\rho n^{1+\rho(1-\gamma)}|l(n)|^\rho].
\end{aligned}
$$

Similarly for term II we have, $II = O[m^\rho n^{1+\rho(1-\gamma)}|l(n)|^\rho]$. Combining this with (10.13) and (10.14), we finish the proof of the lemma. $\qquad\square$

Proof of Theorem 3.4. Recall the definition of $\tilde{Y}_i$ from (10.1). As $m \to \infty$, $\tilde{Q}(u)$ is well-defined and it converges to uth quantile of a $N(0, \sigma^2)$ distribution since the central limit theorem of Hannan (1979) entails $\tilde{Y}_i \xrightarrow{D} N(0, \sigma^2)$, where $\sigma = \|\sum_{i=0}^{\infty} \mathcal{P}_0 e_i\| < \infty$. A standard characteristic function argument yields

$$\sup_x |\sigma f_m(x) - \phi(x/\sigma)| \to 0, \tag{10.15}$$

where $f_m(\cdot)$ is the density of $\tilde{Y}_i$ and $\phi(x)$ is the density of a standard normal random variable. Let (c_n) be an arbitrary sequence of positive numbers that goes to infinity. Let $\bar{c}_n = \min(c_n, n^{1/4}/m^{3/4})$. Then $\bar{c}_n \to \infty$. For $T_n = \bar{c}_n m/\sqrt{n}$, Lemma 10.1 and Lemma 10.2 imply that

$$|\tilde{F}_n(\tilde{Q}(u) + T_n) - \tilde{F}(\tilde{Q}(u) + T_n) - [F_n(\tilde{Q}(u)) - \tilde{F}(\tilde{Q}(u))]|$$

$$= O_{\mathbb{P}}\left(\frac{T_n m^{3/2}}{\sqrt{n}} + m^{1/4}\sqrt{\frac{T_n}{n}}(\log n)^{1/2}\right) = o_{\mathbb{P}}(T_n). \qquad (10.16)$$

Similar arguments as those in Lemma 10.1 and Lemma 10.2 imply

$$|\tilde{F}_n(\tilde{Q}(u)) - \tilde{F}(\tilde{Q}(u))| = O_{\mathbb{P}}(\frac{m}{\sqrt{n}}) = o_{\mathbb{P}}(T_n). \qquad (10.17)$$

Using Taylor's expansion of $\tilde{F}(\cdot)$, we have

$$\tilde{F}(\tilde{Q}(u) + T_n) - \tilde{F}(\tilde{Q}(u)) = T_n f_m(\tilde{Q}(u)) + O(T_n)^2. \qquad (10.18)$$

By (10.15), $f_m(\tilde{Q}(u)) > 0$ for sufficiently large n. Plugging in (10.17) and (10.18) into (10.16), we have $\mathbb{P}(\tilde{F}_n(\tilde{Q}(u) + T_n) > u) \to 1$. Hence $\mathbb{P}(\hat{Q}_n(u) > \tilde{Q}(u) + T_n) \to 0$ by the monotonicity of $\tilde{F}_n(\cdot)$. Similar arguments yield $\mathbb{P}(\hat{Q}_n(u) < \tilde{Q}(u) - T_n) \to 0$. Since $f_m(\tilde{Q}(u)) > 0$ and $T_n \to 0$ arbitrarily slow, Theorem 3.4 is proved.

$\square$

Next we use Lemma 9.3 along with Nagaev inequalities from Section 3 to prove Theorem 4.1 and Theorem 4.2.

Proof of Theorem 4.1. In the view of Lemma 9.3 and appropriate Nagaev inequality from Result 9.2, we have,

$$\mathbb{P}\left(\frac{1}{n}\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 \geq 16sr^2/\kappa^2\right) \leq \sum_{j=1}^{p}\mathbb{P}(2|V_j| > r) = o(1)$$

due to the choice of r in (4.2). Thus,

$$\sup_{m \leq i \leq n}|\sum_{k=i-m+1}^{i}(\hat{e}_i - e_i)| \leq m\sup_{1 \leq i \leq n}|\hat{e}_i - e_i| \leq m\sqrt{\frac{1}{n}\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2} = O_{\mathbb{P}}(m\sqrt{s}r).$$

The rest of the proof follows from the following observation; for any fixed $0 \leq u \leq 1$,

$$|\bar{Q}_n(u) - \hat{Q}_n(u)| = O_{\mathbb{P}}\left(\frac{m\sqrt{s}r}{H_m}\right), \qquad (10.19)$$

where H_m is properly chosen for the heavy or light tails as described in (3.17). Then the right hand side of (10.19) by the choice of s as specified in (4.3) are smaller than the upper bounds of the SRD specific cases mentioned in Theorem 3.4.

$\square$

Proof of Theorem 4.2. Consider the events

$$B_1 = \{n^{-1}|e^\mathsf{T} X|_\infty < r/2\} \text{ and } B_2 = \{|X^\mathsf{T} X/n - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})|_\infty < \kappa_{stoch}^2/(32s)\}.$$

Under B_2, the following holds for a vector v with $|v_{J^c}|_1 \leq 3|v_J|_1$,

$$
\begin{aligned}
\frac{v^\mathsf{T} X^\mathsf{T} X v}{n\|v_J\|_2^2} &= \frac{v^\mathsf{T}(X^\mathsf{T} X/n - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T}))v}{\|v_J\|_2^2} + \frac{v^\mathsf{T}\mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})v}{\|v_J\|_2^2} \\
&\geq -\frac{|v|_1|(X^\mathsf{T} X/n - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T}))v|_\infty}{\|v_J\|_2^2} + + \frac{v^\mathsf{T}\mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})v}{\|v_J\|_2^2} \\
&\geq -\frac{|v|_1^2|X^\mathsf{T} X/n - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})|_\infty}{\|v_J\|_2^2} + + \frac{v^\mathsf{T}\mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})v}{\|v_J\|_2^2} \\
&\geq -\frac{16|v_J|_1^2|X^\mathsf{T} X/n - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})|_\infty}{\|v_J\|_2^2} + + \frac{v^\mathsf{T}\mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})v}{\|v_J\|_2^2} \\
&\geq -\frac{16\left(\frac{\kappa_{stoch}^2}{32s}\right)|v_J|_1^2}{\|v_J\|_2^2} + \frac{v^\mathsf{T}\mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})v}{\|v_J\|_2^2} \\
&\geq -\frac{\kappa_{stoch}^2}{2} + \frac{v^\mathsf{T}\mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^\mathsf{T})v}{\|v_J\|_2^2}.
\end{aligned}
$$

In the light of (4.10), this implies

$$\min_{J \subset \{1,\cdots,p\}, |J| \leq s, \, |u_{J^c}|_1 \leq 3|u_J|_1} \frac{u^\mathsf{T} X^\mathsf{T} X u}{n\|u_J\|_2^2} > \frac{\kappa_{stoch}^2}{2}.$$

Following the (9.8) and the derivation leading onto it, we have under $B_1 \cap B_2$,

$$\kappa_{stoch}\|\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J\|_2 \leq \frac{\sqrt{2}}{\sqrt{n}}\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2. \tag{10.20}$$

which in turn means

$$\mathbb{P}\left(\frac{1}{n}\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 \geq 32sr^2/\kappa_{stoch}^2\right) \leq \mathbb{P}(B_1^c \cup B_2^c).$$

Thus the proof is concluded by estimating the probabilities of B_1^c and B_2^c to be small. For B_1 the proof consists essentially of same steps as Theorem 4.1. In particular, note that we no longer have the additional constraint on X that diagonals of $X^\mathsf{T} X/n$ is 1 and thus we need to apply the Nagaev type concentration inequalities from Result 9.2 on the mean-zero process $\sum_{i=1}^n x_{i,j} e_i$ directly. Note that, in this scenario, while applying the Nagaev-type inequalities on $S_{n,b}$, we choose $b_k = 1$ for $1 \leq k \leq n$ and thus $\|b\|_2^2 = n$, $\|b\|_q^q = n$.

Thus, in the view of the choice of r (refined for the stochastic case) in (4.11), $\mathbb{P}(B_1) \to 1$. Lemma 9.4 ensures $\mathbb{P}(B_2) \to 1$ under the condition on the dependence adjusted functional dependence measure. This in turn yields $\mathbb{P}(B_1 \cap B_2) \to 1$. This finishes the proof of the Theorem 4.2.

$\square$

11   Appendix C:- Additional notes on implementation of methods

Additional notes on implementation of QTL-LASSO

We use LASSO implementation in R-package `glmnet` with tuning parameter λ chosen by cross validation and with weights argument $(v_1 \ldots, v_n) = ((1 - \delta)\delta^{(T-1)}/(1 - \delta^T), \ldots, 1)$ to account for the structural change in coefficients. $\delta = 0.8$.

Additional notes and outputs on implementation of ets, nnar and armax with software output

The ETS(A,N,N) with tuning parameter $= 0.0446$, NNAR$(38, 22)$ with one hidden layer and ARMA$(2, 1)$ were selected by AIC and estimated by R-package `forecast`. NNAR and ARMAX allow for exogenous covariates, therefore we include aggregated weather series $\bar{w}_t = \sum_{k=1}^{73} w_{k,t}$, $\bar{\tau}_t = \sum_{l=1}^{78} \tau_{l,t}$, and weekend dummies as well. For NNAR, we can provide weights for the covariate observations. We use the same exponential down-weighing scheme as for the QTL-LASSO, but with $\alpha = 0.98$, which gave better results.

First the price series y_t is seasonally adjusted using STL decomposition (R-core function). The seasonally adjusted prices z_t is used as input for the models implemented in R-package `forecast`. The models are specified as follows:

ETS The model is selected according to AIC criterion. We restrict the model without trend component, because the prices do not show any trend pattern (see 1). However, probably due to breaks in price-level, the AIC would select a trend component. This results in future paths that vary a lot. For optimization criterion, for, we use Average MSFE, over maximal possible horizon=30 hours. This results into model with tuning parameter 0.0446 selected by AIC. This gives better forecasting results than minimizing in-sample MSE which would result in tuning parameter 0.99 and huge PIs.

```
ETS (A, N, N)
# means additive model, without
trend and seasonal components.
Call: ets(y = y, model = "ZNZ", opt.crit = "amse", nmse = 30)

Smoothing parameters:
 delta = 0.0446

Initial states:
 l = 34.1139

 sigma: 8.4748

  AIC     AICc    BIC
116970.9 116970.9 116992.1
```

NNAR The model is selected according to AIC criterion. The model is restricted in that it allows only one hidden layer. The number of nodes in this layer is by default given as (#AR lags+#exogenous covariates)/2. In this case, we use aggregated wind speed and temperatures, and dummies for weekends so the number of exogenous covariates is 4. In order to get fair comparison with the QTL-LASSO, we also use exponential downweighting on the exogenous covariates, this time with tuning parameter 0.95.

```
NNAR (38,22)
# means AR order is 38 and there are 22 nodes
in the hidden layer
Call:  nnetar(y = y, xreg =  cbind(Weather_agg, dummy_12)
,weights = expWeights(alpha=0.95)))

Average of 20 networks, each of which is a 42-22-1 network with 969 weights
options were - linear output units

sigma^2 estimated as 15.34
```

ARMA The model is selected according to AIC criterion. we use aggregated wind speed and temperatures, and dummies for the weekend.

```
Regression with ARIMA(2,0,1) errors

Coefficients:
    ar1   ar2   ma1   intc.
    0.506 0.328 0.491 59.378
s.e.0.060 0.055 0.057  1.644
    xreg1  xreg2 xreg3 xreg4
    -4.551 -0.489 0.481 0.133
s.e.0.404  0.060 0.538 0.538
sigma^2 estimated as 24.35:
log likelihood=-26410.34
AIC=52838.68   AICc=52838.7   BIC=52902.38
```

12 Appendix D:- Additional simulation results

(i) Case $n > p$. QTL implemented using each of the estimators OLS, LAD or LASSO and CLT using LASSO only.

nominal	(e_i) m-weeks	Coverage short-heavy 1	2	3	4	long-heavy 1	2	3	4	non-lin-heavy 1	2	3	4	Winkler loss short-heavy 1	2	3	4	long-heavy 1	2	3	4	non-lin-heavy 1	2	3	4
60%	ols-qtl	47.0	41.0	38.0	37.3	38.9	33.1	28.8	25.9	48.6	42.0	41.0	37.9	2.1	1.7	1.5	1.4	7.0	6.4	6.5	6.6	1.5	1.2	1.1	1.0
	lad-qtl	51.5	51.1	48.5	45.9	38.0	33.4	32.1	29.2	50.7	51.8	49.1	50.0	1.5	1.2	1.1	1.0	6.0	5.5	5.6	5.6	1.0	0.8	0.8	0.7
	lss-qtl	56.8	55.6	54.2	49.7	54.9	53.6	53.0	50.0	57.4	55.8	51.8	49.6	1.5	1.2	1.1	1.1	5.9	5.2	5.2	5.1	1.1	0.9	0.9	0.8
	lss-clt	54.3	46.5	43.1	37.5	43.3	30.8	28.0	23.6	51.8	42.4	35.0	33.2	1.5	1.3	1.2	1.2	6.0	5.5	5.6	5.6	1.1	1.0	0.9	0.9
80%	ols-qtl	65.9	60.5	55.6	52.2	56.5	50.1	44.6	39.3	67.4	62.5	59.5	54.2	3.0	2.5	2.2	2.0	10.3	9.7	10.0	10.4	2.2	1.8	1.6	1.5
	lad-qtl	74.5	71.5	71.0	68.7	61.4	53.9	53.2	49.0	73.7	72.3	72.4	69.7	2.1	1.7	1.6	1.5	8.8	8.0	8.6	8.6	1.5	1.2	1.1	1.1
	lss-qtl	80.1	77.9	74.1	71.2	74.8	75.4	73.5	70.8	77.2	75.4	73.7	71.4	2.1	1.7	1.6	1.6	8.4	7.3	7.7	7.6	1.5	1.3	1.2	1.2
	lss-clt	74.1	65.8	62.1	53.2	62.3	49.9	42.7	36.8	71.1	61.3	54.1	50.6	2.2	1.7	1.7	1.8	8.7	8.2	8.7	8.8	1.6	1.4	1.4	1.3
90%	ols-qtl	78.4	73.8	70.2	65.6	70.1	63.7	60.8	52.6	78.7	74.4	71.9	68.1	4.3	3.6	3.3	3.2	14.6	14.1	15.2	16.2	2.9	2.6	2.4	2.3
	lad-qtl	87.3	84.1	84.1	83.2	76.9	73.0	71.9	65.6	87.4	83.9	85.8	81.5	3.1	2.5	2.9	2.5	12.2	11.5	13.4	13.5	2.1	1.8	2.1	1.9
	lss-qtl	88.9	87.8	87.0	83.9	87.5	86.4	83.1	81.2	88.9	85.9	85.9	83.1	3.1	2.5	2.8	2.6	11.7	10.8	11.8	11.6	2.1	1.8	2.1	1.9
	lss-clt	85.1	78.7	73.7	64.7	72.1	59.4	54.5	47.5	81.3	72.7	65.4	63.4	3.1	2.5	2.5	2.6	12.5	12.2	13.5	14.0	2.2	1.9	2.0	2.0
95%	ols-qtl	86.3	82.4	76.4	72.1	81.9	75.1	69.1	60.7	86.8	82.2	78.6	73.3	6.1	5.5	5.0	4.9	20.5	21.2	24.1	26.2	3.9	4.1	3.7	3.6
	lad-qtl	92.7	91.2	88.1	87.1	87.0	83.6	77.2	71.4	92.7	90.9	89.5	86.4	4.5	4.5	3.9	3.5	16.8	17.7	20.0	20.4	3.1	3.1	2.9	2.7
	lss-qtl	93.9	93.3	90.0	87.5	92.4	91.3	88.2	86.7	93.5	92.5	90.7	88.0	4.5	4.4	3.8	3.5	16.5	16.5	16.7	16.3	3.0	3.1	2.9	2.7
	lss-clt	89.0	85.7	80.0	74.1	78.4	67.6	62.1	55.1	87.0	79.5	75.0	72.2	4.5	3.6	3.7	3.8	18.5	18.6	21.8	22.8	3.1	2.8	3.0	3.0

(ii) Case $p > n$. QTL and CLT as above. ADJ is a bootstrap version QTL for better performance under short-sample.

nominal	(e_i) m-days	Coverage short-heavy 1	2	3	4	long-heavy 1	2	3	4	non-lin-heavy 1	2	3	4	Winkler loss short-heavy 1	2	3	4	long-heavy 1	2	3	4	non-lin-heavy 1	2	3	4
60%	lss-qtl	50.9	48.4	47.2	41.7	53.0	46.6	44.0	38.6	49.9	48.5	46.3	39.4	4.8	3.8	3.8	3.5	8.6	7.7	7.8	7.8	3.1	2.5	2.5	2.2
	lss-clt	54.6	39.3	37.4	32.7	68.8	48.9	41.8	33.4	48.2	33.9	34.0	27.7	5.5	4.9	4.9	4.3	8.0	7.1	7.4	7.5	3.9	3.6	3.6	3.6
	lss-adj	50.8	47.5	48.1	47.1	51.3	46.9	43.4	40.3	51.2	48.8	49.0	48.6	4.8	3.9	3.7	3.3	8.4	7.2	7.0	6.8	3.1	2.5	2.3	2.1
80%	lss-qtl	71.7	66.3	61.2	53.4	70.5	64.3	56.0	50.6	69.7	64.8	62.2	54.4	7.4	6.2	5.4	5.1	13.5	11.9	11.9	11.7	4.6	4.1	3.6	3.3
	lss-clt	73.4	57.6	51.8	49.1	82.2	68.2	59.7	51.3	69.6	54.4	47.7	40.5	7.9	6.7	7.1	6.3	12.5	10.5	10.8	11.2	5.3	5.0	5.2	5.3
	lss-adj	71.7	65.1	65.0	63.0	69.2	64.7	57.9	54.5	71.6	67.1	66.9	65.3	7.5	6.1	5.6	5.0	12.5	10.8	10.7	10.5	4.9	4.0	3.5	3.1
90%	lss-qtl	82.7	74.0	68.6	61.3	81.2	73.1	62.5	55.2	81.8	73.4	67.9	62.2	12.5	9.0	8.1	7.9	20.7	17.4	18.0	18.7	7.9	6.0	5.3	5.0
	lss-clt	83.7	69.3	63.6	59.3	88.9	78.0	70.2	62.9	80.9	68.7	58.8	52.4	11.5	8.8	9.8	8.9	18.8	15.7	16.0	16.7	7.4	6.7	7.3	7.8
	lss-adj	81.4	75.3	73.9	72.1	79.3	73.3	67.7	63.8	81.5	77.8	74.9	74.4	12.0	8.8	8.1	7.4	19.4	16.5	16.4	16.5	7.5	5.7	5.0	4.5
95%	lss-qtl	86.1	77.6	73.3	65.8	85.3	76.8	66.4	57.2	85.5	77.9	71.7	66.0	19.4	13.9	13.0	13.3	32.9	27.9	30.0	32.0	11.3	9.2	8.4	8.3
	lss-clt	90.3	80.0	74.3	70.4	91.6	83.7	76.5	71.3	86.1	77.9	70.6	62.5	17.5	12.4	13.6	12.5	29.2	24.1	24.9	26.0	10.6	9.1	10.2	11.5
	lss-adj	86.2	82.1	80.6	78.4	85.8	79.3	75.0	68.3	86.4	83.0	81.7	79.6	18.6	13.0	11.9	11.4	31.1	25.8	26.3	26.9	11.0	8.5	7.4	6.9

Table 2: Simulated out-of-sample forecasting experiment. The reported values are coverage probabilities, i.e., relative (%) counts of out-of-sample values covered in 1000 trials (left part) and average Winkler loss values (right part). The nominal coverage (the first column) ranks from is 95% to 60%. Simulated error processes heavy tails and either short memory, long memory or are nonlinear. The elements of regression coefficient β are drawn independently from uniform distribution $U[-1, 1]$. The sparsity of β's is fixed to 90%.

(i) Case $n > p$. QTL implemented using each of the estimators OLS, LAD or LASSO and CLT using LASSO only.

nominal	(e_i) / m-weeks	Coverage short-heavy 1	2	3	4	long-heavy 1	2	3	4	non-lin-heavy 1	2	3	4	Winkler loss short-heavy 1	2	3	4	long-heavy 1	2	3	4	non-lin-heavy 1	2	3	4
60%	ols-qtl	47.0	41.0	38.0	37.3	38.9	33.1	28.8	25.9	48.6	42.0	41.0	37.9	2.1	1.7	1.5	1.4	7.0	6.4	6.5	6.6	1.5	1.2	1.1	1.0
	lad-qtl	51.5	51.1	48.5	45.9	38.0	33.4	32.1	29.2	50.7	51.8	49.1	50.0	1.5	1.2	1.1	1.0	6.0	5.5	5.6	5.6	1.0	0.8	0.8	0.7
	lss-qtl	54.8	47.5	49.9	28.2	54.7	52.5	53.4	50.3	55.7	52.1	40.1	38.8	1.7	1.6	1.3	2.2	6.2	5.3	5.3	5.1	1.1	1.0	1.3	1.2
	lss-clt	50.1	28.8	34.6	9.9	43.0	32.0	26.1	22.8	49.6	36.5	17.3	17.7	1.8	1.8	1.5	2.7	6.4	5.7	5.7	5.6	1.2	1.2	1.6	1.5
80%	ols-qtl	65.9	60.5	55.6	52.2	56.5	50.1	44.6	39.3	67.4	62.5	59.5	54.2	3.0	2.5	2.2	2.0	10.3	9.7	10.0	10.4	2.2	1.8	1.6	1.5
	lad-qtl	74.5	71.5	71.0	68.7	61.4	53.9	53.2	49.0	73.7	72.3	72.4	69.7	2.1	1.7	1.6	1.5	8.8	8.0	8.6	8.6	1.5	1.2	1.1	1.1
	lss-qtl	75.6	71.2	72.2	55.0	74.4	73.8	74.0	70.4	77.9	74.0	66.3	65.5	2.3	2.1	1.8	2.6	8.7	7.6	7.7	7.6	1.6	1.4	1.7	1.5
	lss-clt	65.7	46.9	50.1	16.6	59.0	46.6	40.9	36.8	70.3	53.9	28.8	26.9	2.5	2.6	2.1	4.3	9.2	8.5	8.9	8.9	1.6	1.7	2.5	2.4
90%	ols-qtl	78.4	73.8	70.2	65.6	70.1	63.7	60.8	52.6	78.7	74.4	71.9	68.1	4.3	3.6	3.3	3.2	14.6	14.1	15.2	16.2	2.9	2.6	2.4	2.3
	lad-qtl	87.3	84.1	84.1	83.2	76.9	73.0	71.9	65.6	87.4	83.9	85.8	81.5	3.1	2.5	2.9	2.5	12.2	11.5	13.4	13.5	2.1	1.8	2.1	1.9
	lss-qtl	88.4	85.2	85.6	72.3	86.2	85.0	83.9	81.6	89.2	86.4	80.4	80.3	3.1	2.9	3.0	3.5	11.9	10.9	11.8	11.7	2.2	1.9	2.5	2.3
	lss-clt	78.2	60.5	61.9	23.2	67.7	58.0	52.2	47.7	80.4	64.6	38.9	36.6	3.4	3.6	3.1	6.9	13.2	12.7	13.8	14.0	2.3	2.4	3.9	3.7
95%	ols-qtl	86.3	82.4	76.4	72.1	81.9	75.1	69.1	60.7	86.8	82.2	78.6	73.3	6.1	5.5	5.0	4.9	20.5	21.2	24.1	26.2	3.9	4.1	3.7	3.6
	lad-qtl	92.7	91.2	88.1	87.1	87.0	83.6	77.2	71.4	92.7	90.9	89.5	86.4	4.5	4.5	3.9	3.5	16.8	17.7	20.0	20.4	3.1	3.1	2.9	2.7
	lss-qtl	92.6	91.3	89.5	79.1	92.3	90.1	88.0	86.4	92.9	92.2	86.7	85.7	4.7	4.8	4.1	4.6	16.7	16.5	16.8	16.3	3.1	3.2	3.2	3.0
	lss-clt	86.1	71.4	72.8	29.3	77.7	66.6	59.9	55.2	85.6	72.8	47.5	44.9	4.9	5.2	4.5	11.1	19.3	19.3	22.2	22.9	3.3	3.5	6.2	5.9

(ii) Case $p > n$. QTL and CLT as above. ADJ is a bootstrap version QTL for better performance under short-sample.

nominal	β (e_i) / m-days	Coverage short-heavy 1	2	3	4	long-heavy 1	2	3	4	non-lin-heavy 1	2	3	4	Winkler loss short-heavy 1	2	3	4	long-heavy 1	2	3	4	non-lin-heavy 1	2	3	4
60%	lss-qtl	51.7	48.9	46.8	42.2	51.9	47.3	44.9	39.3	50.4	48.6	46.3	40.7	4.8	3.8	3.9	3.4	8.6	7.7	7.8	7.8	3.1	2.5	2.5	2.3
	lss-clt	54.9	39.5	37.0	32.5	68.6	49.1	42.0	34.7	48.6	34.8	33.4	27.6	5.5	4.9	4.9	4.3	8.0	7.1	7.3	7.5	4.0	3.6	3.6	3.5
	lss-adj	51.3	47.0	47.9	47.2	50.4	46.0	44.2	40.6	49.5	48.9	48.8	48.7	4.8	3.9	3.8	3.2	8.4	7.2	6.9	6.8	3.2	2.5	2.3	2.1
80%	lss-qtl	72.1	66.2	61.8	54.3	70.7	64.2	56.9	50.5	70.7	66.3	62.1	55.1	7.4	6.2	5.5	5.0	13.5	12.0	11.8	11.7	4.7	4.1	3.6	3.3
	lss-clt	74.5	57.7	51.8	49.2	82.1	67.7	59.5	51.2	69.4	53.8	46.6	40.5	7.8	6.7	7.1	6.3	12.5	10.5	10.8	11.2	5.3	5.0	5.2	5.3
	lss-adj	71.0	64.6	64.6	64.6	69.7	64.2	57.7	55.2	70.4	67.7	67.0	64.1	7.5	6.1	5.7	4.8	12.4	10.9	10.7	10.5	4.9	4.0	3.5	3.1
90%	lss-qtl	82.8	74.3	68.5	61.0	81.1	73.1	62.6	55.4	81.5	74.2	67.9	62.9	12.4	9.0	8.2	7.8	20.7	17.4	17.9	18.7	8.0	6.0	5.2	5.0
	lss-clt	84.1	70.0	64.6	59.8	88.5	78.3	70.6	62.8	80.7	67.2	59.7	52.3	11.5	8.8	9.9	8.9	18.8	15.7	16.0	16.7	7.4	6.7	7.3	7.8
	lss-adj	81.3	76.8	74.2	71.7	80.4	73.7	68.6	65.0	80.4	78.1	76.3	73.7	12.0	8.7	8.1	7.2	19.4	16.6	16.3	16.5	7.5	5.8	5.0	4.5
95%	lss-qtl	86.4	77.4	73.0	64.9	85.4	76.5	66.7	57.6	85.1	77.3	72.6	66.6	19.3	14.0	13.2	13.0	32.9	28.0	30.0	32.0	11.4	9.3	8.3	8.3
	lss-clt	90.4	80.0	74.0	70.2	92.1	83.2	77.3	70.3	86.2	77.2	69.6	62.8	17.5	12.4	13.6	12.4	29.1	24.2	24.9	26.1	10.6	9.1	10.3	11.5
	lss-adj	86.1	82.9	80.7	78.8	85.9	78.6	75.3	68.7	86.2	83.1	82.3	80.6	18.8	13.0	12.1	11.2	31.0	25.9	26.2	26.9	11.2	8.6	7.5	6.9

Table 3: Simulated out-of-sample forecasting experiment. The reported values are coverage probabilities, i.e., relative (%) counts of out-of-sample values covered in 1000 trials (left part) and average Winkler loss values (right part). The nominal coverage (the first column) ranks from is 95% to 60%. Simulated error processes heavy tails and either short memory, long memory or are nonlinear. The elements of regression coefficient β are drawn independently from uniform distribution $U[-1, 1]$. The sparsity of β's is fixed to 20%.