Prediction Intervals in High-Dimensional Regression

S. Karmakar[a,*], M. Chudý[b], W.B. Wu[c]

[a] *Department of Statistics, University of Florida, 230 Newell Drive, Gainesville, FL 32601, USA*
[b] *Institute for Financial Policy, Ministry of Finance, Stefanovicova 5, 81105 Bratislava, Slovakia*
[c] *Department of Statistics, University of Chicago, 5747 S Ellis Ave, Chicago, IL 60615, USA*

Abstract

We construct prediction intervals for the time-aggregated univariate response time series in a high-dimensional regression regime. The consistency of our approach is shown for those cases when the number of observations is less than the number of covariates, particularly for the popular LASSO estimator. We allow for general heavy-tailed, long-memory, and non-linear stationary process and validate our approach using simulations. Finally, we construct prediction intervals for hourly electricity prices over horizons spanning 17 weeks and compare them to selected Bayesian and bootstrap interval forecasts.

Keywords: consistency, Non-linear, LASSO, electricity prices, bootstrap, time-aggregation

*Corresponding author
Email addresses: sayarkarmakar@ufl.edu (S. Karmakar), marek.chudy@univie.ac.at (M. Chudý), wbwu@galton.uchicago.edu (W.B. Wu)

Submission status: submitted to ISF conference *Last update: March 19, 2019*

1. Introduction

Prediction intervals (p.i.'s) help the forecasters to access the uncertainty concerning the future values of time series. The benefit of interval forecast compared to point forecasts comes for the cost of a more challenging evaluation (Chatfield, 1993; Clements and Taylor, 2003), since a higher empirical coverage probability often comes with the cost of a larger width, thus less precision. Suppose an univariate target time series $y_i, i = 1, \ldots, n$ follows a regression model

$$y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + e_i, \quad \boldsymbol{\beta} \in \mathbb{R}^p. \tag{1.1}$$

For convenience, we scale the diagonal entries of the Gram matrix $X^\mathsf{T}X/n$ to be 1. For finite $p < n$, Zhou et al. (2010) showed that empirical quantiles obtained from rolling sums of past residuals provide theoretically valid asymptotic p.i.'s for the aggregated future value $S_{+m} = y_{n+1} + \ldots + y_{n+m}$, when both $n, m \to \infty$. Moreover, the specific choice of estimator for $\boldsymbol{\beta}$ leads to particular p.i.'s $[L, U]_{\hat{\boldsymbol{\beta}}}$ such that for given α

$$P\left([L, U]_{\hat{\boldsymbol{\beta}}} \ni S_m\right) \underset{n,m\to\infty}{\to} 1 - \alpha.$$

Zhou et al. (2010) utilized the LAD[1] estimator for $\boldsymbol{\beta}$; however, if $p > n$ this and other conventional estimators, such as OLS, would fail. Even if $p < n$, and there is a reason to assume $\boldsymbol{\beta}$ is sparse, using some alternative estimators, such as LASSO, may be more efficient for forecasting. Our motivation comes from fields such as economics and energy, where the targets supposedly depend on a large number of covariates. Furthermore, many economic or energy time series are subject to structural breaks, which makes their future values and the uncertainty surrounding them even harder to assess (Koop and Potter, 2001; Cheng et al., 2016). Generally, most forecasters accept that the inclusion of many disaggregated covariates provides some additional forecasting power over conventional univariate and low-dimensional multivariate approaches (see Stock and Watson, 2012; Elliott et al., 2013; Kim and Swanson, 2014). In their case study for EPEX SPOT hourly day-ahead electricity prices, Ludwig et al. (2015) found that inclusion of wind speed and temperature measured at local weather stations across Germany leads to improvements of day-ahead point forecasts over the approach when the temperature resp. wind are aggregated across the stations. We found this application interesting for our regression framework, but we focus on long- and medium-horizon (spanning up to 17 weeks ahead) p.i.'s, which are essential for

at this point it seems as a technical detail, does not seem to belong into introduciton, besides, X is not defined.

[1] Least absolute deviation

power portfolio risk management, derivatives pricing, medium- and long-term contract evaluation, and maintenance scheduling rather than for day-to-day operations. Furthermore, we use some alternative methods to provide a visual out-of-sample comparison with our p.i.'s. The alternative approaches include Bayes p.i.'s of Müller and Watson (2016) and bootstrap p.i.'s 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).

Electricity price forecasting (EPF) generally uses exogenous variables including weather conditions, local economy, and environmental policy[2] (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. There is a substantial amount of literature about EPF (see Weron, 2014, for a recent review). These complex nature can often be also explained through different covariates which can be both deterministic and stochastic. We explore the problem of time-aggregated forecasting for a high dimensional linear regression model with a possibly non-linear and dependent error process.

Our contributions in this paper are multifold. From a theoretical perspective, it was important to explore whether the simple quantile-based method proposed by Zhou et al. (2010) would extend to non-linear processes because of its vast generality. Using the idea of predictive density and the functional dependence framework as proposed in a seminal paper by Wu (2005a), we were able to extend the quantile consistency results for the error process to non-linear processes which includes smooth transition autoregression (Lundbergh et al., 2003) processes or very commonly used ARCH-GARCH type processes.

Our second contribution lies in exploring the prediction performance in a high-dimensional regression situation, especially allowing for the number of predictors to grow faster than the sample size ($\log p = o(n)$) and thus hitting the popular benchmark for the high-dimensional literature. We extended the theoretical properties for the LASSO estimation in this high-dimensional regime. Moreover, we explicitly discuss the price of having short or long-range dependence and lighter or heavier tails. Note that all these results can be easily extended to situations where the error process has exponentially decaying tails such as sub-exponential and sub-gaussian, but since similar discussion with independent processes are already prevalent in the literature we wanted to restrict ourselves to condition of only finitely many moments. It is important to note that we specifically exploited a particular corollary from Na-

[2]In 2005, Germany launched a program aiming at reducing emissions by increasing the share of renewable energy. The share was 25% during 2013-2014 and reached 36% in 2017.

gaev (1979) to extend the non-linear process results from Wu and Wu (2016) to the scenario where the second moment of the error process does not exist.

As a third contribution, we discuss how one can allow stochastic design in the lasso program and still obtain sharp consistency results. This was particularly appealing since for the small $p < n$ case as described in Zhou et al. (2010), it makes sense to only allow for trigonometric covariates whose future values are known. However, it is more practical and challenging to allow for stochastic design. Our application to electricity price forecasting uses both the trigonometric fixed design and the stochastic weather variables. We utilize the consistency of the LASSO estimator in this set-up and show the consistency of the quantiles.

Finally, we tackle the scenario of short-sample long-horizon forecasting of electricity prices with a simple bootstrap approach. Apart from a conjectural viewpoint justifying our bootstrap procedure, we were able to empirically validate our method by doing a pseudo out of sample (POOS) comparison. We also

The rest of the article is organized as follows: In Section 2, we summarize the construction of p.i.'s by methods inspired from Zhou et al. (2010) focusing on the particular cases: no covariates, fixed number of covariates and where the number of covariates is much larger than the sample size. Section 3 discusses the error process specifications. The discussion of how we model non-linearity in two different ways and the quantile consistency results are the main results of this section. We discuss three scenarios: short- and long-range dependence, and under light-tailed and heavy-tailed distribution. We provide the discussion of quantile consistency of the original response using the LASSO-fitted residuals in Section 4. This covers both deterministic design and the more statistically meaningful stochastic design. Section5 shows simulation results and some data-driven adjustments to arrive at better forecasting performance when facing the curse of dimensionality. This section also presents our real data analysis. Section 6 gives concluding remarks. The proofs of theorems can be found in appendix section 7.

2. Construction of prediction intervals

We first discuss the scenario $\beta = 0$ i.e. $y_i = e_i$ in (1.1) is a zero-mean noise process as a primer.

2.1. Without covariates

Depending on its memory and tail behavior, Zhou et al. (2010) proposed two different types of p.i. of predicting 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.

Quenched CLT method. 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)

$$\tilde{\sigma} = \frac{\sqrt{\pi l/2}}{n} \sum_{k=1}^{\kappa} \left| \sum_{i=(k-1)l+1}^{kl} e_i \right|, \tag{2.1}$$

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

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

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

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 u-th empirical quantiles $\hat{Q}(u)$ of $\sum_{j=i-m+1}^{i} e_i; i = m, \ldots, n$. The p.i.'s in this case are

$$[L, U] = \left[\hat{Q}(\alpha/2), \hat{Q}(1-\alpha)/2) \right]. \tag{2.2}$$

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

2.2. High dimensional regime

Assuming model (1.1), we wish to construct p.i. for $y_{n+1} + \ldots + y_{n+m}$ after observing $(y_i, x_i); i = 1, \ldots, n$. After estimating $\boldsymbol{\beta}$, Zhou et al. (2010) construct p.i. as

$$\sum_{i=n+1}^{n+m} \boldsymbol{x}_i^{\mathsf{T}} \hat{\boldsymbol{\beta}} + \text{ p.i. for } \sum_{i=n+1}^{n+m} \hat{e}_i, \tag{2.3}$$

where $\hat{e}_i = y_i - \boldsymbol{x}_i^{\mathsf{T}} \hat{\boldsymbol{\beta}}$ 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 $\hat{\boldsymbol{\beta}} = \text{argmin} \sum_i (y_i - \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta})^2$. For heavy-tailed or long-range dependent errors (see Huber and Ronchetti, 2009), it is better to use robust regression with general distance ρ and $\hat{\boldsymbol{\beta}} = \text{argmin} \sum_i \rho(y_i - \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta})$. Examples of distance include the $\mathcal{L}^q$ regression for $1 \leq q \leq 2$.

Note that, the p.i. in (2.3) requires the future covariate values of x_i namely $x_{n+1}, \ldots, x_{n+m}$. Zhou et al. (2010) discusses the scenario where these predictors are trigonometric and are completely deterministic. One easily understands that for practical purposes this might not be true since in real life it is more appealing or statistically interesting to use some real life covariate. This naturally rises the

question of predicting the future covariate values. Since our approach here is not focused on predicting the predictors we do not address forecasting of the future predictor values and for practical examples as in the electricity price forecast as described in Section 5 we provide a simple way to forecast for the future values for the set of predictors that are not well-known apriori.

3. Prediction interval for the error process

Before moving on to a more general discussion with exponentially many covariates, we use this section to discuss a primer without covariates i.e. our response here is just a mean-0 error process. This section elaborately describes the model specifications on the error process. We collect the results for the linear processes from Zhou et al. (2010) and Chudý et al. (2019) without proof and provide our proofs for the non-linear case.

3.1. Asymptotic normality for linear process

Using the simple idea that under short-range dependence if m is long enough then the dependence of $y_{T+1} + \ldots + y_{T+m}$ on $y_1, y_2, \ldots, y_T$ diminishes and the conditional distribution of $(y_{T+1} + \ldots + y_{T+m})/\sqrt{m}$ given $y_1, \ldots, y_T$ is almost similar to the unconditional distribution and thus one can obtain a simple central limit theorem to quantify the uncertainty in prediction. 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 Δ denotes the Levy distance, $m \to \infty$ and $\sigma^2 = \lim_{m\to\infty} \|S_m\|_2^2/m$ is the long-run variance. Assuming linearity of the mean-zero noise process e_i in the following manner

$$e_i = \sum_{j=0}^{\infty} a_j \epsilon_{i-j}, \quad \text{such that} \quad e_i \text{ i.i.d.}, \mathbb{E}(|e_1|^p) < \infty; p > 2, \tag{3.3}$$

it is easy to derive the following conditions on a_i to ensure the convergence in (3.2). The proof is skipped here as it was proved in appendix of the forthcoming Chudý et al. (2019).

Theorem 3.1 (Theorem 1 from Chudý et al. (2019)). *Assume the process e_t admits the representation (3.3) where a_i satisfies*

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

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 asymptotic normal distribution holds.

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.

3.2. About empirical quantile method

In this subsection, we discuss an intuitive quantile-based method motivated to allow the relaxation of short range dependence or moment conditions. We rewrite the short-range dependence assumption for a linear process with the representation in (3.3) as

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

For long range dependence in linear process, we assume the same representation as (3.3) and

(LRDL) with γ case we assume $l^*(i) = a_i i^{-\gamma}$ is a slowly varying function where $q < \gamma < 1$ and $1/q = \sup\{t : \mathbb{E}(|\epsilon_j|^t) < \infty\}$. Note that, the definition of (LRDL) also takes into account the possible heavy-tailed distribution of ϵ_j. Additionally assume ϵ_i admits a density f_ϵ and

(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 actual quantile of $\tilde{S}_i$, $i = m, \ldots, n$, where

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

and

$$H_m = \begin{cases} \sqrt{m}, & \text{if (SRDL) holds and } \mathbb{E}(\epsilon_j^2) < \infty, \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x) \le \frac{1}{m}\} & \text{if (SRDL) holds and } \mathbb{E}(\epsilon_j^2) = \infty, \\ m^{3/2-\gamma} l^*(m) & \text{if (LRDL) holds and } \mathbb{E}(\epsilon_j^2) < \infty, \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x) m^{1-\gamma} l^*(m) & \text{if (LRDL) holds and } \mathbb{E}(\epsilon_j^2) = \infty. \end{cases} \tag{3.6}$$

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* $0 < u < 1$,

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(m/\sqrt{n}). \tag{3.7}$$

- *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* $0 < u < 1$,

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(mn^{1/2-\gamma}|l^*(n)|). \tag{3.8}$$

- *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* $0 < u < 1$,

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

- *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* $0 < u < 1$,

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(mn^{\nu}) \text{ for all } \nu > 1/q - \gamma. \tag{3.10}$$

Here heavy-tailed refers to the scenario where the innovation process does not possess finite second moment. The proof of Theorem 3.1 and Theorem 3.4 heavily rely on the representation in (3.3) and thus it was important to explore from both a theoretical and application perspective to prove analogous result for non-linear processes.

3.3. Non-linear error process: functional dependence

Economic and financial time series are often subject to structural changes and thus the linear models do not provide proper approximation for their data-generating process. Useful non-linear time-series models 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 (3.2) to non-linear error process assuming e_i is a stationary process that admits the following representation

$$e_i = H(\mathcal{F}_i) = H(\epsilon_i, \epsilon_{i-1}, \ldots), \tag{3.11}$$

where H is such that e_i are 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 of e_i. In order to get (3.2), we define

a functional dependence measure for e_i in (3.11), by which we follow Wu (2005a)'s framework to formulate dependence through coupling. We define the following functional dependence measure

$$\delta_{j,p} = \|e_i - e_{i,(i-j)}\|_p = \|H_i(\mathcal{F}_i) - H_i(\mathcal{F}_{i,(i-j)}\|_p, \tag{3.12}$$

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 thought as an input-output system. It facilitates easily verifiable and 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

$$\Theta_{j,p} = \sum_{i=j}^{\infty} \delta_{i,p}, \tag{3.13}$$

which can be thought as cumulative dependence of $(e_j)_{j\geq k}$ on ϵ_k. Further, we define dependence adjusted norm, for $\alpha > 0$,

$$\|e.\|_{q,\alpha} = \sup_{t\geq 0}(t+1)^{\alpha} \sum_{i=t}^{\infty} \delta_{i,q}. \tag{3.14}$$

Theorem 3.3. *Assume e_i admits the representation in (3.11) the following rate holds for $\Theta_{j,p}$.*

$$\Theta_{j,p} = j^{-\chi}(\log j)^{-A} \text{ where } = \begin{cases} A > 0 \text{ for } 1 < \chi < 3/2, \\ A > 5/2 \text{ for } \chi \geq 3/2, \end{cases} \tag{3.15}$$

then the convergence in (3.2) holds.

Proof. The m-dependence approximation is a 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 $\tilde{S}_m = \sum_{i=1}^m \tilde{e}_i = \sum_{i=1}^m \mathbb{E}(e_i|\epsilon_i, \ldots \epsilon_{i-m})$. The proof of (3.2) then follows from $\|P_j(\tilde{e}_i)\|_2 \leq \delta_{i-j,2}$, since

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

□

3.4. Non-linear error process: predictive dependence

In order to show the empirical quantile consistency that validates p.i.'s of the form (2.2), we need to control the latent dependence of e_i on ϵ_{i-j}. Therefore, we introduce the predictive density-based dependence measure. Let $\mathcal{F}'_k = (\ldots, \epsilon_{-1}, \epsilon'_0, \epsilon_1, \ldots, \epsilon_k)$, be the coupled shift process derived from $\mathcal{F}_k$ by substitution of ϵ_0 by its i.i.d. copy ϵ'_0. Let $F_1(u,t|\mathcal{F}_k) = P\{G(t;\mathcal{F}_{k+1}) \le u|\mathcal{F}_k\}$ be the one-step ahead predictive or conditional distribution function and $f_1(u,t|\mathcal{F}_k) = \delta F_1(u,t|\mathcal{F}_k)/\delta u$, be the corresponding conditional density. We define the predictive dependence measure

$$\psi_{k,q} = \sup_{t\in[0,1]} \sup_{u\in\mathbb{R}} \|f_1(u,t|\mathcal{F}_k) - f_1(u,t|\mathcal{F}'_k)\|_q. \tag{3.16}$$

Quantity (3.16) measures the contribution of ϵ_0, the innovation at step 0, on the conditional or predictive distribution 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,

$$\sup_{t\in[0,1]} \sup_{u\in\mathbb{R}} \{f_1(u,t|\mathcal{F}_0) + |\delta f_1(u,t|\mathcal{F}_0)/\delta u|\} \le c_0.$$

The (DEN) implies that the marginal density $f(u,t) = Ef_1(u,t|\mathcal{F}_0) \le c_0$. Recall the sufficient conditions for the linear cases in Zhou et al. (2010) were based on the coefficients of the linear process. Here, however, the conditions for both short- and long-range dependent errors are based on the predictive dependence measure. We assume:

$$\begin{aligned} &\text{(SRD)} \quad : \quad \sum_{j=0}^{\infty} |\psi_{j,q}| < \infty, \\ &\text{(LRD)} \quad : \quad \psi_{j,q} = j^{-\gamma} l(j), \gamma < 1, l(\cdot) \text{ is slowly varying function (s. v. f.)} . \end{aligned} \tag{3.17}$$

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

$$\tilde{S}_i = \frac{\sum_{j=i-m+1}^{i} e_j}{H_m}, \quad i = m, m+1, \ldots \tag{3.18}$$

and

$$H_m = \begin{cases} \sqrt{m}, & \text{if (SRD) holds and } \mathbb{E}(\epsilon_j^2) < \infty, \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x) \le \frac{1}{m}\} & \text{if (SRD) holds and } \mathbb{E}(\epsilon_j^2) = \infty, \\ m^{3/2-\gamma} l(m) & \text{if (LRD) holds and } \mathbb{E}(\epsilon_j^2) < \infty, \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x) m^{1-\gamma} l(m) & \text{if (LRD) holds and } \mathbb{E}(\epsilon_j^2) = \infty. \end{cases} \tag{3.19}$$

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: non-linear 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* $0 < u < 1$,

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(m/\sqrt{n}). \tag{3.20}$$

- *Light tailed (LRD): Suppose (LRD) and (DEN) hold with* γ *and* $l(\cdot)$ *in (3.17). If* $m^{5/2-\gamma}n^{1/2-\gamma}l^2(n) \to 0$, *then for any fixed* $0 < u < 1$,

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(mn^{1/2-\gamma}|l(n)|). \tag{3.21}$$

- *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* $0 < u < 1$,

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(mn^{\nu}) \text{ for all } \nu > 1/q - 1. \tag{3.22}$$

- *Heavy-tailed (LRD): Suppose (LRD) hold with* γ *and* $l(\cdot)$ *in (3.17). If* $m = O(n^k)$ *for some* $k < (q\gamma - 1)/(2q + 1 - q\gamma)$, *then for any fixed* $0 < u < 1$,

$$|\hat{Q}(u) - \tilde{Q}(u)| = O_{\mathbb{P}}(mn^{\nu}) \text{ for all } \nu > 1/q - \gamma. \tag{3.23}$$

We discussed quantile consistency results for the error process so far in our exposition since this forms the foundation of our proposed method for constructing the prediction interval. In the next section, we discuss the estimation in presence of covariates and show some similar consistency results.

4. High dimensional regression

Consider the case $p \gg n$ with LASSO estimator with l_1-penalty

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{R}^p}{\operatorname{argmin}} \frac{1}{n}(y_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|\,, \tag{4.1}$$

with penalty coefficient λ. Then we get p.i.'s (2.3) with $\hat{\boldsymbol{\beta}}$ replaced by the LASSO estimator. For the non-stochastic covariate design, the effect of the covariates remain constant no matter whether it is observed in the past or to be observed in the time. However, for the stochastic design, this could be a potential concern. We can assume that the 'deterministic' part $\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ would still capture the general mean effect, and

then we employ uncertainty over the rest of the error process and then append the prediction interval just for the residual or estimated noise process to the estimated effect of these covariates.

Before we present the quantile consistency results for the case where the number of predictors grows much faster, we also state key optimal tail-probability inequalities needed for the two cases, linear and non-linear error process. Let $S_{n,b} = \sum_{i=1}^{n} b_i e_i$. We have the following tail probability bounds for $S_{n,b}$ under the short- and long-range dependence and light and heavy tails for the error process. Only the light-tailed versions of 4.1 and Result 4.2 were proposed and proved in Wu and Wu (2016). We skip the proof since one needs to use corollary 1.6 instead of corollary 1.7 from Nagaev (1979) and proceed according to Wu and Wu (2016). However, we still state the results differently distinguishing between the cases for short or long-range dependence and light or heavy tails.

Result 4.1. *(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,*

$$\mathbb{P}(|S_{n,b}| \geq x) \leq (1+2/q)^q \frac{|b|_q^q (\sum_j |a_j|)^q \|\epsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{c_q x^2}{n(\sum_j |a_j|)^2 \|\epsilon_0\|_2^2}\right), \quad (4.2)$$

- *Light-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 constant C_1, C_2 depending on only q and β,*

$$\mathbb{P}(|S_{n,b}| \geq x) \leq C_1 \frac{K^q |b|_q^q \|n^{q(1-\beta)}\epsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{C_2 x^2}{n^{3-2\beta}\|\epsilon_0\|_2^2 K^2}\right), \quad (4.3)$$

- *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*

$$\mathbb{P}(|S_{n,b}| \geq x) \leq c_q \frac{|b|_q^q (\sum_j |a_j|)^q \|\epsilon_0\|_q^q}{x^q}, \quad (4.4)$$

- *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,*

$$\mathbb{P}(|S_{n,b}| \geq x) \leq C_1 \frac{K^q |b|_q^q \|n^{q(1-\beta)}\epsilon_0\|_q^q}{x^q}. \quad (4.5)$$

Next, we discuss a Nagaev inequality for the non-linear process using the dependence adjusted norm from (3.14). Under the short-range dependence, $\Theta_{0,q} < \infty$ where q is less or more than 2 depending on the tail-behavior of the error process.

Result 4.2. *(Nagaev inequality for non-linear 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,

$$\mathbb{P}|S_{n,b}| \geq x) \leq C_1 \frac{r_n}{(\sum_j |b_j|)^q \|e.\|_{q,\alpha}^q} x^q + C_2 \exp\left(-\frac{C_3 x^2}{n\|e.\|_{2,\alpha}^2}\right), \tag{4.6}$$

- *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,

$$\mathbb{P}|S_{n,b}| \geq x) \leq C_1 \frac{r_n}{(\sum_j |b_j|)^q \|e.\|_{q,\alpha}^q} x^q. \tag{4.7}$$

Note that, we do not discuss the long-range dependence case for the non-linear process since definition wise it is very technical and thus of little practical interest. Next we show that for short-range dependent non-linear error process states that the error bounds obtained in Theorem 3.4 remain intact if we make a proper choice of the sparsity condition.

4.1. Lasso with fixed design

For the model in (1.1), we first assume that the x_i's are fixed and the future x_i's are completely known. Under this setting, we next show the quantile consistency for the non-linear process. Note that a very similar result can be proved for the linear process with the error process admitting a simpler representation (3.3) however we skip writing that as a separate theorem here for avoiding repetitiveness.

Theorem 4.3. *(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 (4.1) where*

$$r = \max\{A\sqrt{n^{-1}\log p}\|e.\|_{2,\alpha}, B\|e.\|_{q,\alpha}|X|_q n^{-1+\min\{0,1/2-1/q-\alpha\}}\}. \tag{4.8}$$

We assume that the restricted eigenvalue assumption RE(s, κ) in Bickel et al. (2009) holds with constant κ = κ(s, 3), where s is the number of non-zero entries in true parameter vector β and

$$\kappa_{s,c} = \min_{J \subset \{1,\cdots,p\}, |J| \leq s,\, |u_{J^c}|_1 \leq c|u_J|_1} \frac{|Xu|_2}{\sqrt{n}|u_J|_2}. \tag{4.9}$$

Let $\bar{Q}_n(u)$ be the u-th empirical quantile of $(\tilde{\tilde{S}}_i)_m^n$. Assume that (SRD) holds and for r defined in (4.8),

$$\begin{aligned} \text{(for light tails, i.e. } q \geq 2) \quad s &= o\left(\frac{m}{r^2 n}\right), \\ \text{(for heavy tails, i.e. } 1 < q \leq 2), \quad s &= o\left(\frac{H_m^2 |l(n)|^2}{r^2 n^{2\gamma-1}}\right), \end{aligned} \tag{4.10}$$

where γ and $l(\cdot)$ are defined in (3.17), H_m in (3.6) and α is in the context of the dependence adjusted norm defined in (3.14), 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 $\|e_\cdot\|_{\cdot,\alpha}$ are due to the dependence present in the error process. The sparsity condition for the light tail case, i.e. $q \geq 2$, in 4.10 in the view of the choice of r in (4.8) can be written as

$$\text{(for light tails)} \quad s \ll \min\left(\frac{n^{4/3}}{\log p \|e_\cdot\|_{2,\alpha}^2}, \frac{n^{7/3-2\max\{0,1/2-1/q-\alpha\}}}{|X|_q^2 \|e_\cdot\|_{q,\alpha}^2}\right) \tag{4.11}$$

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 surprising result makes sense in the context of this paper, we are only interested in the long-term predictions rather than in the immediate future point. Thus some relaxation can be done using the fact that if m is allowed to grow to ∞, the m-length average of residuals can automatically provide some concentration. 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 linear process (See 3.3), the conditions in (4.10) will remain identical and one can provided some more specifications in the definition of r in (4.8) using the linear coefficients a_i from (3.3) in the view of the Nagaev-type concentration

inequalities derived in Result 4.1. 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

We show that under a mild condition on the covariate process the same optimization routine LASSO also works for the quantile consistency results in a regime where the covariates are stochastic. 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 the covariates to vary as the random variable we also allow the covariates to be dependent and thus allowing for the scenario where the covariates could evolve. For this subsection, we restrict ourselves to only non-linear processes for a clearer exposition. We assume

$$\mathbf{x}_i = G_x(\epsilon_i^x, \epsilon_{i-1}^x, \ldots), \tag{4.12}$$

where ϵ_i^x are i.i.d and G_x is a measurable function. Let $\{\epsilon_i^x\}'$ be an i.i.d. copy of $\{\epsilon_i^x\}$. Define the functional dependence measure on the stationary process $\mathbf{x}_i$ as follows

$$\delta_{k,q}^{\mathbf{x}} = \|\mathbf{x}_i - \mathbf{x}_{k,i}^*\|_q, \tag{4.13}$$

where $\mathbf{x}_{k,i}^* = G_x(\epsilon_i^x, \epsilon_{i-1}^x, \ldots, \epsilon_{i-k}^x{}', \ldots)$. Also let the error process e admit the following representation

$$e_i = G_e(\epsilon_i^e, \epsilon_{i-1}^e, \ldots), \tag{4.14}$$

where ϵ_i^x are i.i.d. and G_x is a measurable function. One can then define the cumulative dependence using this functional dependence measure. For the quantile consistency in the case of stochastic design we will need a notion of functional dependence on the cross-product process $x_{\cdot j}e_{\cdot}$ as follows

$$\delta_{k,q}^{xe} = \max_{j \le p} \|x_{lj}e_l - x_{k,lj}^* e_{k,l}^*\|_q, \tag{4.15}$$

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 (4.15), we define the dependence adjusted norm as (3.14) for the $x_{\cdot j}e_{\cdot}$ process uniformly over $1 \le j \le p$ as follows:

$$\max_{j\le p} \|x_{\cdot j}e_{\cdot}\|_{q,\alpha} = \sup_{t\ge 0}(t+1)^{\alpha}\sum_{i=t}^{\infty}\delta^{xe}_{i,q}. \tag{4.16}$$

Theorem 4.4. *(Empirical quantile consistency for LASSO-stochastic)*
Assume $\max_{j\le p}\|x_{\cdot j}e_{\cdot}\|_{q,\alpha} < \infty$ *for some* $\alpha > 0$. *Let* $\bar{Q}_n(u)$ *be the* u*-th empirical quantile of* $(\hat{\tilde{S}}_i)_m^n$. *Denote the number of non-zero elements of* β *by* s *and assume that the sparsity conditions in 4.10 hold with the choice of* $r = \lambda/2$ *as follows:*

$$r = \max\{A\sqrt{n^{-1}\log p}\|e_{\cdot}\|_{2,\alpha}, B\|e_{\cdot}\|_{q,\alpha}n^{\min\{0,1/2-1/q-\alpha\}}\}. \tag{4.17}$$

Then the SRD specific conclusions of Theorem 3.4 with $Q_n(u)$ *replaced by* $\bar{Q}_n(u)$.

Note that, the uniform functional dependence measure on the cross-product space $x_{\cdot j}e_{\cdot}$ 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).

5. Simulation and real data evaluation

In this section, we discuss the set-up, the low and high dimensional cases in the first three subsections and use the last one for discussing some extensive data analysis.

5.1. Simulation set-up

The focus is on evaluation of p.i.'s 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 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 non-linear error-process know as the logistic smooth transition autoregression (LSTAR) with heavy tailed innovations.

Eventually, we add a large number of exogenous covariates to the error process, obtaining $y_t = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + e_i, i = 1, \ldots, n+m$. We compute our p.i.'s based on $(y_1, \boldsymbol{x}_1) \ldots, (y_n, \boldsymbol{x}_n)$ 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, which will be motivated in following empirical part.

Regarding covariates, we consider two scenarios (i) $p < n$ and (ii) $p > n$. In scenario (i), we compare p.i.'s based on OLS, LAD and LASSO estimators. In case (ii), we only have the LASSO, since the other two estimators are not uniquely identified. The vector $\boldsymbol{x}_i \in \mathbb{R}^p$ consists of the same 151 weather variables and 168 (resp. 336 for case ii) deterministic variables, which will be described in Section 5.4. The elements of $\boldsymbol{\beta} \in \mathbb{R}^p$ are obtained as i.i.d draws from the uniform distribution $U[-1, 1]$ or Cauchy distribution. Unlike for the OLS and LAD, the theoretical properties of LASSO depend on the sparsity assumption, hence we exploit the robustness check with different scenarios for the sparsity of $\boldsymbol{\beta}$, i.e., $s = 100(1 - \|\beta\|_0/p) = 99\%, 90\%, 70\%, 50\%, 20\%$. We keep the (sparse) $\boldsymbol{\beta}$ fixed for all 1000 repetitions of our experiment.

I failed to verify this claim for the LAD estimator

For the sake of brevity, only p.i.'s with nominal coverage (n.c.) $100(1-\alpha) = 90\%$ are reported[3]. For the two methods (quenched CLT method and QTL, i.e., empirical quantile method) proposed in Section 2, we compute the coverage probabilities (c.p.'s) $\widehat{(1-\alpha)} = \frac{1}{1000}\sum_{j=1}^{1000} \mathbb{I}\left([L, U]_{j,\hat{\boldsymbol{\beta}}} \ni \bar{y}_{j,+1:m}\right)$, 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.

5.2. Low-dimensional case

Results for the case $p < n$ are given in Table 1i. We set $n = 8736$ ($\approx$ 1 year of hourly data), $m = 168, 336, 504, 672$ (1,2,3,4 weeks of hourly data), and $p = 319$. It is close to the set-up of our empirical application (except that the horizon spans up to 17 weeks in the empirical part). Overall results suggest that for LASSO the empirical quantile method provides better results than the quenched CLT method. The latter is much more sensitive to cases of long memory and non-linearity, especially for the longer horizons. Comparison of LASSO-QTL to LAD-QTL and OLS-QTL suggests the following:

[3] P.i.'s for n.c. 67% and respective lengths of all p.i.'s can be obtained from the authors upon request.

Uniform vs. Cauchy β's. It does not seem to make any difference whether elements of $\boldsymbol{\beta}$ are drawn from the uniform or the Cauchy distribution for the OLS-QTL and LAD-QTL, whereas the LASSO performs better under the uniform distribution. The LASSO provides (almost always) the highest c.p.'s close to 90%. Under Cauchy-distributed coefficients, the winner is not clear, because if $s < 80\%$ and $m > 2$ weeks, LAD often gives a higher c.p. except for the long-memory errors, where the LASSO always wins.

Short memory vs. long memory vs. non-linearity. The curse of non-linearity is rather small compared to the curse of the long memory on the OLS and LAD. For instance with the OLS, c.p.'s drop by 8 percent points (pp) between the short- and long-memory cases, whereas for the LASSO c.p.'s decreased too but only by 3pp.

Horizon. Independent of which scenario we look at, the growing horizon makes QLT deteriorate quickly. The OLS is more sensitive to the horizon than the LASSO or LAD. For instance, in the long-memory scenario, the c.p. falls by 17pp between 1 week and 4 weeks horizon for the OLS, by 11pp for the LAD and by only 6pp for the LASSO.

5.3. High-dimensional case

The case $p > n$ is given in Table 1ii. We set $n = 336$ (2 weeks of hourly data), $m = 24, 48, 72, 96$ (1,2,3,4 days of hourly data) and $p = 487$. Both the sample size n and the forecasting horizon m become shorter but the horizon/sample proportions reach $m/n > 1/4$. The reason behind such set-up is that for p to be larger than n, it is convenient to keep n small so that the p does not have to be very large. This is practical since otherwise, the computation becomes very slow. We do use a much longer sample in the low-dimensional case though since it gives us more insight into the empirical performance of the methods under such two different scenarios. It turns out that the sample size has a relatively large impact on performance. The combination of the curse of dimensionality and the large horizon/sample ratio deteriorates the performance of LASSO-QTL. For a remedy, we exploit a data-driven adjustment of the QTL based on replication of the residual $\hat{e}_i = y_i - \hat{y}_i$ using stationary bootstrap (Politis and Romano, 1994a) and kernel quantile estimator (Falk, 1984). We denote the adjusted QTL by ADJ and provide the computation steps for this adjusted method later in Section 5.4. The ADJ is also used instead of QTL in Section 5.4 because of the large $m/n \approx 1/3$.

Unlike the low-dimensional case, the CLT dominates QLT, especially if the process has a long memory or if the sparsity is low. On the other hand, ADJ, which had lower c.p.'s initially, outperforms both QTL and CLT by more than 10pp for

longer horizons. Despite the improvement, the c.p.'s get close to 60% when the error process shows long memory.

5.4. Prediction intervals for European Power Exchange spot electricity prices

The focus is on graphical comparison of out-of-sample p.i.'s obtained by:

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

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

(ARX) Bootstrap path simulations from ARMAX models,

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

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

Data. We forecast $\bar{y}_{+1:m} = 1/m \sum_{t=1}^{m} y_{n+t}$, i.e. the average[4] of m future hourly day-ahead spot electricity prices for Germany and Austria - the largest market at the European Power Exchange (EPEX SPOT). 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 till 12/31/2013 23:00:00 UTC and an evaluation period spanning from 01/01/2014 00:00:00 UTC till 04/30/2014 23:00:00 UTC (see Figure 1A). The forecasting horizon $m = 1, 2, \ldots, 17$ weeks $(168, 336, \ldots, 2856$ 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 $g_t^k = R\sin(\omega_k t) + \phi) = \beta_k^{(s)}(R, \phi)\sin(\omega_k t) + \beta_k^{(c)}(R, \phi)\cos(\omega_k t)$ 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 Figueroa, 2005; Hyndman and Athanasopoulos, 2013). The coefficients of linear combination $\beta_k^{(s)}, \beta_k^{(c)}$ can be estimated by least squares. In addition, we use 2 dummy variables as indicators for weekend.

[4]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.

[5]Coordinated Universal Time.

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 for long horizons. Local weather is represented by 151 hourly wind speed, and temperature series 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 these covariates

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

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.

Methods. 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-weighting 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-weighting implicitly, but with optimally selected weights. Müller and Watson (2016) showed that the RBS is robust to structural changes. What follows are the main implementation steps for the methods used in this section:
Bootstrap p.i. for 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 (1994b)), 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 $(\bar{e}^b_{t(m)}) = m^{-1} \sum_{i=1}^{m} e^b_{t-i+1}$, $t = m, \ldots n$ 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 $\bar{e}^b_{n(m)}$, $b = 1, \ldots, B$.

(v) The p.i. 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 $h = 1, \ldots, m$.

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

Note that, 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 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 kernelized quantile estimator (See Sheather and Marron (1990)) further improves the performance in prediction. These improvements are supported by the empirical pseudo-out-of-sample validation where we roll the window of available and to be predicted data through the entire time horizon we have in the data.

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 however providing some more dispersion to the average as with a nontrivial probability it concatenates blocks containing same elements and thus adding those covariances in the dispersion of the overall average. However, 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.

Bootstrap p.i. for RBS:

For this sophisticated univariate approach, we focus on the intuition and refer to the supplementary Appendix of Müller and Watson (2016) for more details about the implementation. The *robust Bayes* (RBS) p.i's are specifically designed for long-horizon predictions, e.g., when $m/n \approx 1/2$. First, the high-frequency noise is partialed 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) employ Bayes approach. In addition, the resulting p.i.'s are further enhanced to attain the frequentist coverage using 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.
(iv) The p.i.'s are given by $[L, U] = \bar{y} + [Q_q^{\text{robust}}(\alpha/2), Q_q^{\text{robust}}(1-\alpha/2)]$.

Bootstrap p.i.'s 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 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 $\bar{\hat{y}}_{+,1:m}^b, b = 1, \ldots, B$.

POOS results. Before we compare the ADJ to the other competitors, we would 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 p.i.'s using no regressors as in Figure 2IA, using only deterministic regressors as in Figure 2IB, 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 2IC and finally, using all 321 covariates as in Figure 2ID. 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.

Why is it POOS?

Finally, we get to the comparison with the alternative p.i.'s denoted as RBS, ETS, NAR, and ARX. All these p.i.'s are given in Figure 2II. Of the four methods, only RBS gives sensible p.i.'s. RBS works consistently well over the whole 17-weeks-long evaluation period (Figure 2IIA). 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-weighting implied by the simple autoregression is too mild. Besides, the narrow p.i.'s are the result of ignoring the parameter (among other types of) uncertainty.

6. Discussion

We have constructed valid prediction intervals based on the high-dimensional regression model. From a theoretical perspective, we have extended the results of Zhou et al. (2010) into the high-dimensional set-up and also to the case of the non-linear error process. Through a thorough evaluation of a strong Nagaev type inequality, we showed the consistency of fitted residuals for the ultra-high dimensional case $\log p = o(n)$. Another significant contribution of this paper is to discuss the stochastic design matrix since if there are exponentially many covariates, it is unnatural to assume fixed or perfectly predictable covariates. Under some mild condition on the 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). However, it is not possible to draw any general conclusions from this specific case study.

Regarding our application to electricity price forecasting, it would be interesting to consider a larger set of covariates, e.g., augmented by macroeconomic covariates like the fuel prices and the GDP growth. Some interesting extensions of the current paper would include multivariate target series and subsequent construction of simultaneous prediction intervals. Applications of such simultaneous intervals could include prediction of spot electricity prices for each hour simultaneously in the spirit of Raviv et al. (2015).

Acknowledgements

We would like to thank Stefan Feuerriegel for kindly sharing the data with us, Erhard Reschenhofer for useful comments and Eric Laas-Nesbitt for proofreading the draft. We also acknowledge the computational resources provided by the Vienna Scientific Cluster. M. Chudý gratefully acknowledges financial support from J.W. Fulbright Commission for Educational Exchange in the Slovak Republic, The Ministry of Education, Science, Research and Sport of the Slovak Republic and by the Stevanovich Center for Financial Mathematics.

References

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.

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.

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

Chudý, M., S. Karmakar, and W. B. Wu (2019+). Long-term prediction intervals of economic time series. *revised for Empirical economics*.

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.

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.

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

Falk, M. (1984). Relative deficiency of kernel type estimators of quantiles. *Ann. Statist. 12*(1), 261–268.

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.

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

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

Koop, G. and S. M. Potter (2001). Are apparent findings of nonlinearity due to structural instability in economic time series? *Econometrics Journal 4*(1), 37–55.

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.

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 (1994a). The stationary bootstrap. *Journal of the American Statistical Association 89*, 1303–1313.

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

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.

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

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.

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.

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.

sparsity	β	Uniform												Cauchy											
	(e_i)	short-heavy				long-heavy				non-lin-heavy				short-heavy				long-heavy				non-lin-heavy			
	m-weeks	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	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	78.4	73.8	70.2	65.6	70.1	63.7	60.8	52.6	78.7	74.4	71.9	68.1
	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	87.3	84.1	84.1	83.2	76.9	73.0	71.9	65.6	87.4	83.9	85.8	81.5
99%	lss-qtl	89.0	88.2	87.7	84.7	87.1	86.4	83.3	81.4	89.5	87.2	88.7	85.5	88.3	87.7	85.3	83.7	86.2	85.9	82.5	80.8	88.9	86.3	86.1	80.8
	lss-clt	85.7	81.0	76.8	72.7	71.9	59.9	53.0	47.2	83.6	77.7	74.0	70.4	78.0	72.7	66.5	65.4	71.7	59.9	53.0	47.4	71.5	69.4	66.8	56.2
90%	lss-qtl	89.5	88.0	86.8	84.7	86.7	85.9	83.7	81.3	88.4	85.9	85.8	83.5	87.7	83.0	81.2	77.9	85.7	85.2	80.8	78.7	89.3	85.6	83.1	74.2
	lss-clt	84.0	77.0	72.9	68.2	71.4	60.5	52.7	47.2	80.1	74.4	64.7	61.6	48.5	30.8	28.0	34.4	69.8	56.3	49.8	41.5	34.4	32.5	30.2	14.2
70%	lss-qtl	88.6	86.4	86.2	82.8	86.6	85.6	83.4	81.1	87.7	84.1	82.9	81.4	86.0	78.1	77.0	79.0	83.8	82.2	79.7	74.2	88.5	90.0	84.5	75.3
	lss-clt	82.5	71.9	67.3	64.9	71.4	59.9	51.8	47.8	76.5	61.5	46.5	53.8	26.4	16.9	12.3	17.1	65.5	52.9	48.4	33.5	20.5	24.6	16.7	4.3
50%	lss-qtl	88.8	84.1	84.5	81.9	86.8	85.5	83.5	80.6	86.9	83.8	79.5	81.5	81.6	77.4	70.1	81.9	82.6	82.8	78.7	73.9	87.7	94.6	83.6	76.4
	lss-clt	80.6	63.8	64.7	62.4	71.2	60.0	51.8	47.1	74.2	49.8	32.0	48.6	21.3	16.9	9.0	10.6	61.2	50.9	44.8	33.6	17.7	25.4	17.3	8.6
20%	lss-qtl	88.7	80.9	84.2	82.8	86.4	85.7	83.6	81.5	87.9	81.0	76.0	80.8	72.7	74.6	52.6	80.6	78.5	83.3	77.3	74.8	82.3	98.7	74.9	81.3
	lss-clt	83.4	54.4	65.0	65.4	72.7	59.6	53.2	45.8	75.0	33.1	15.0	49.5	11.9	11.5	5.3	7.9	53.9	49.4	40.0	35.8	19.7	39.8	17.7	1.8

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

sparsity	β	Uniform												Cauchy											
	(e_i)	short-heavy				long-heavy				non-lin-heavy				short-heavy				long-heavy				non-lin-heavy			
	m-days	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
99%	lss-qlt	81.7	74.2	68.1	61.9	80.4	73.8	63.4	56.2	80.6	75.7	75.2	67.8	82.2	74.7	67.9	62.5	80.6	73.5	63.5	56.1	80.1	72.6	68.6	61.9
	lss-clt	84.0	71.1	66.1	59.7	88.0	77.2	69.4	62.6	80.7	67.6	59.1	53.9	84.2	71.2	66.3	60.7	87.8	77.3	69.8	63.1	81.2	68.3	59.6	53.7
	lss-adj	80.1	75.5	74.6	73.3	79.6	73.6	69.0	64.6	80.6	76.6	75.7	74.7	80.8	75.5	74.9	72.8	79.6	74.3	68.7	64.6	79.9	77.0	75.1	74.7
90%	lss-qlt	82.0	74.0	68.2	61.5	80.5	73.3	63.2	56.1	80.5	72.6	68.3	61.8	81.5	74.5	70.7	62.2	80.5	73.1	63.3	59.8	80.8	70.8	68.3	59.9
	lss-clt	84.0	71.5	66.2	60.2	88.0	77.2	69.6	62.7	81.5	68.4	59.5	53.5	85.8	72.1	67.7	63.9	88.2	77.5	71.3	65.7	80.6	70.5	64.0	54.7
	lss-adj	81.3	76.2	74.7	73.6	79.3	74.4	67.9	64.7	80.5	76.8	76.3	74.1	80.0	76.7	74.2	73.2	80.5	73.2	68.3	67.3	80.7	73.0	72.5	71.8
70%	lss-qlt	80.7	74.1	69.2	61.7	80.4	73.4	63.0	56.4	80.2	73.1	68.7	62.9	78.3	71.5	68.9	64.3	80.9	72.1	64.3	64.3	79.8	70.8	66.0	59.0
	lss-clt	83.8	71.5	66.0	59.9	88.2	77.4	69.9	62.8	81.1	68.6	59.7	53.9	88.1	73.0	69.2	69.2	88.2	78.2	73.1	72.9	81.5	77.1	69.7	54.9
	lss-adj	79.9	75.7	75.0	73.4	79.7	73.9	68.6	64.8	80.9	77.6	75.4	75.1	78.4	72.3	72.0	74.5	79.5	74.4	69.9	71.8	80.1	71.3	71.8	71.0
50%	lss-qlt	80.7	73.5	69.3	62.3	80.5	73.2	62.5	56.4	80.0	73.7	68.6	61.8	72.9	71.7	70.6	62.8	80.9	72.7	63.2	71.7	79.1	73.6	70.3	59.2
	lss-clt	83.8	71.8	65.9	59.9	88.3	77.2	70.0	63.1	80.9	68.7	59.9	54.3	90.3	73.8	70.5	72.7	89.1	78.6	75.4	79.1	81.8	77.6	70.8	55.0
	lss-adj	80.8	75.4	74.3	74.0	79.8	74.5	68.9	64.8	81.2	77.2	75.6	74.5	74.1	71.8	74.6	73.0	80.1	74.7	70.0	76.9	78.5	73.2	75.3	72.3
20%	lss-qlt	81.2	73.2	68.4	61.5	80.1	73.6	63.0	55.6	79.5	73.4	68.9	61.7	67.3	68.8	75.6	65.4	79.5	72.2	59.3	76.9	80.5	78.6	74.7	61.2
	lss-clt	84.3	71.4	66.4	59.8	88.3	78.4	70.1	62.9	80.7	68.2	59.6	53.7	91.7	75.0	71.7	74.9	88.5	79.0	78.5	83.8	80.1	80.9	79.2	52.2
	lss-adj	80.6	75.5	74.5	72.5	79.4	74.7	68.9	64.9	80.0	75.8	77.4	74.1	65.9	67.8	77.8	76.8	77.4	75.0	70.1	80.5	79.9	77.8	78.5	72.5

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

Table 1: Simulated pseudo-out-of-sample forecasting experiment. The reported values are relative (%) counts of out-of-sample values covered in 1000 trials. The nominal coverage is 90%. Simulated error processes have short memory & heavy tails or long memory & heavy tails or are non-linear with heavy-tailed innovations. The elements of regression coefficient $\boldsymbol{\beta}$ are drawn as i.i.d. from uniform distribution $U[-1, 1]$ or Cauchy distribution. The sparsity of $\boldsymbol{\beta}$'s varies from 99% to 20%.

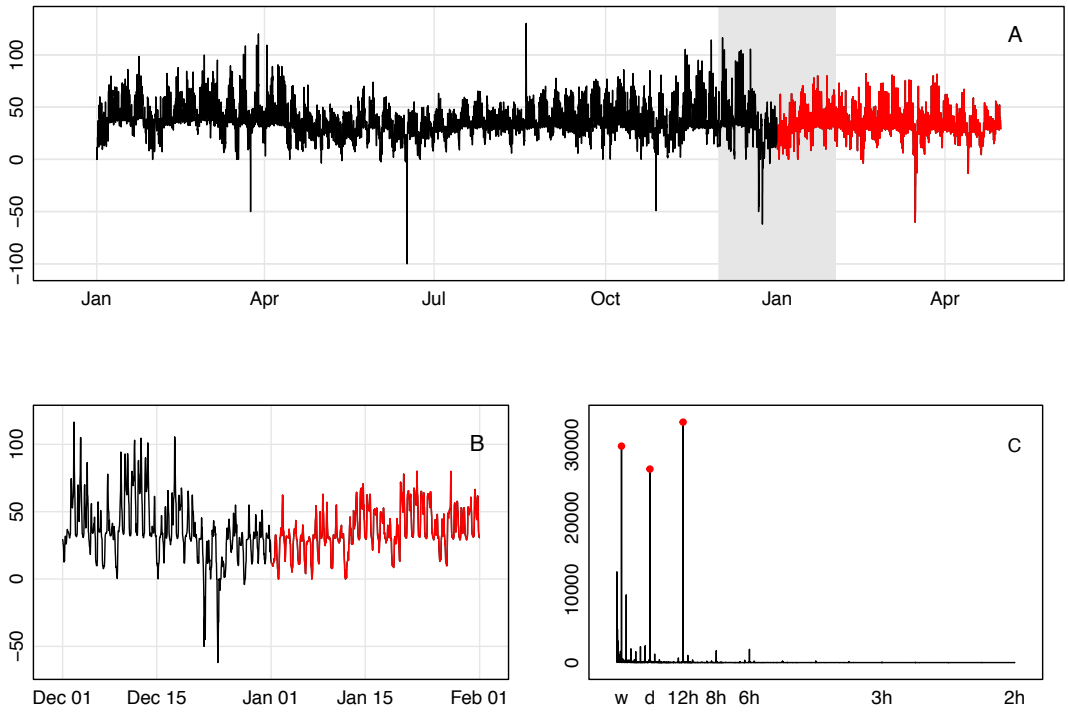
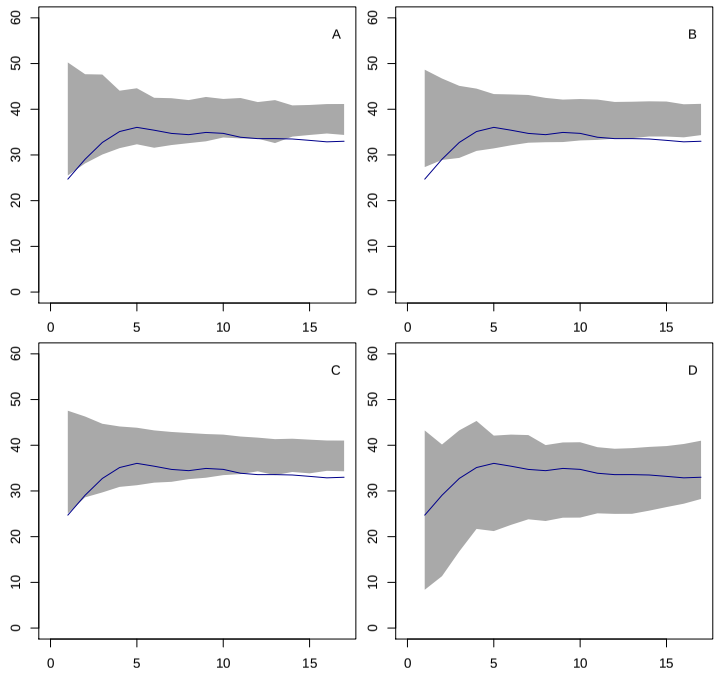
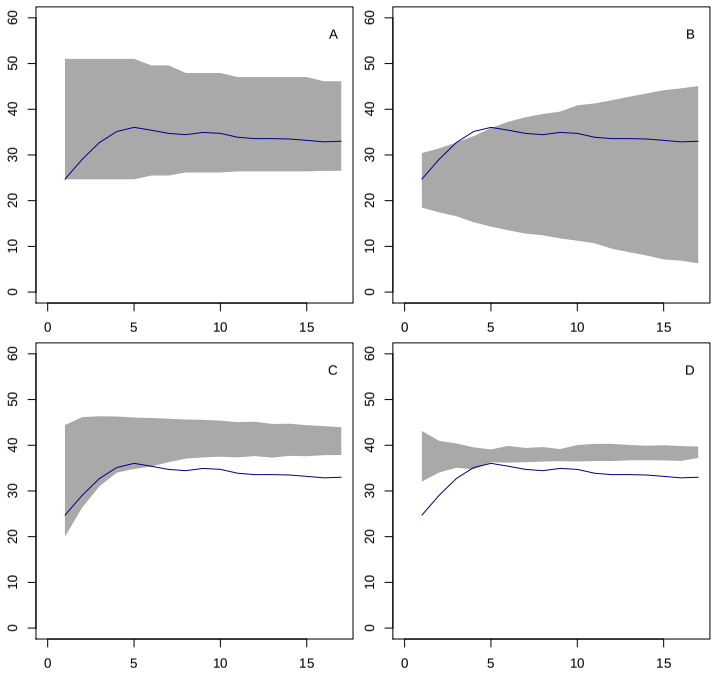


Figure 1: Electricity spot prices, A) Full sample, B) Drop in price level, C) Periodogram with peaks at periods 1 week, 1 day and 12 hours.



(I) A) ADJ (without covariates), B) ADJ with 170 deterministic covariates, C) ADJ with 170 deterministic covariates & 2 aggregated (across stations) weather time series, D) ADJ with 170 deterministic covariates & 151 disaggregated weather series.



(II) 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.

Figure 2: Prediction intervals (gray) for average spot electricity prices (blue) over forecasting horizon $m = 1, \ldots, 17$ weeks.

Appendix A: Additional information for section 5

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 $(v_1 \ldots, v_T) = ((1-\delta)\delta^{(T-1)}/(1-\delta^T), \ldots, 1)$ to account for the structural change in coefficients. $\delta = 0.8$.

Additional notes 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-weighting 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 in that we dont use 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 too varying future paths. 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 is gives better forecasting results than minimizing in-sample MSE which would result in tuning parameter 0.99 and huge p.i.'s.

```
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 restrited 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 weekend 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 that order of AR component is 38 and there are 22 nodes in the hidden layer
Call:  nnetar(y = y, xreg = cbind(Weather_agg, dummy_12), weights = expWeights(alpha=

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 weekend.

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

Coefficients:
         ar1     ar2     ma1  intercept    xreg1    xreg2   xreg3   xreg4
      0.5062  0.3284  0.4908    59.3783  -4.5514  -0.4894  0.4811  0.1333
s.e.  0.0601  0.0548  0.0568     1.6441   0.4037   0.0602  0.5382  0.5382

sigma^2 estimated as 24.35:  log likelihood=-26410.34
AIC=52838.68   AICc=52838.7   BIC=52902.38
```

7. Appendix A - Proofs

Define

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

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

$$\tilde{F}_n(x) - \tilde{F}(x) = (\tilde{F}_n(x) - \tilde{F}_n^*(x)) + (\tilde{F}_n^*(x) - \tilde{F}(x)) = M_n(x) + N_n(x)$$

Define, for a random variable $Z \in \mathcal{L}^1$, $P_i(Z) = \mathbb{E}(Z|\mathcal{F}_i) - \mathbb{E}(Z|\mathcal{F}_{i-1})$. Using this, one can write $M_n(x)$ as follows

$$M_n(x) = \frac{1}{n-m+1} \sum_{i=m}^{n} P_i(I(\tilde{Y}_i \leq x)). \tag{7.1}$$

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 7.1. *Under conditions of Theorem 4.1 and Theorem 4.2,*

$$\sup_{|u| \leq b_n} |M_n(x+u) - M_n(x)| = O_{\mathbb{P}}\left(\sqrt{\frac{H_m b_n}{n}} \log^{1/2} n + n^{-3}\right), \tag{7.2}$$

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 for any $u \in [-b_n, b_n]$, we have

$$\sum_{i=m}^{n} [\mathbb{E}(V_i) - \mathbb{E}(V_i)^2] \leq c_0(n-m+1)H_m b_n \quad \text{where } V_i = I(x \leq \tilde{Y}_i \leq x+u|\mathcal{F}_{i-1}). \tag{7.3}$$

The result in (7.2) follows by applying Freedman's martingale inequality and a chaining argument. We skip the details as this chaining argument is lengthy and essentially is very similar to that presented in Lemma 5 in Wu (2005b), Lemma 4 in Wu (2007) and Lemma 6 in Zhou and Wu (2009).

□

Lemma 7.2. *Under conditions of SRD, DEN and light-tailed*

$$\| \sup_{|u|\le b_n} |N_n(x+u) - N_n(x)| \| = O\left(\frac{b_n m^{3/2}}{\sqrt{n}}\right). \tag{7.4}$$

Proof. Since $N_n(x) = \tilde{F}_n^*(x) - \tilde{F}(x)$, we have

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

where

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

The proof of (7.4) is complete pending we prove the following

$$\|R_n(x+u)\| \le Cm\sqrt{n} \text{ for all } u \in [-b_n, b_n].$$

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

$$\tilde{b}_{j,q} = \begin{cases} \psi_{0,q} + \psi_{1,q} + \ldots + \psi_{j,q} & \text{if } 1 \le j \le m-1 \\ \psi_{j-m+1,q} + \psi_{j-m+2,q} + \ldots + \psi_{j,q} & \text{if } j \ge m. \end{cases} \tag{7.5}$$

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

$$\begin{aligned} \|\mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\| &\le \|f_\epsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1})) - f_\epsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1,k}^*))\| \\ &\le \sup_{v\in\mathbb{R}} |f_\epsilon'(v)|\sqrt{m}(\tilde{Z}_{i-1} - \tilde{Z}_{i-1.k}^*)\|\| \le c_1\tilde{b}_k, \end{aligned} \tag{7.6}$$

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 $\mathcal{P}_{i-k}, i = m, \ldots, n$

$$\|\sum_{i=m}^{n} \mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\|^2 = \sum_{i=m}^{n} \|\mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\|^2 \leq c_1^2 (n-m+1)\tilde{b}_k^2.$$

Therefore, for all $u \in [-b_n, b_n]$, by the short-range dependence condition as

$$\begin{aligned} \|R_n(x+u)\| &\leq \sum_{k=1}^{\infty} \|\sum_{i=m}^{n} \mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\| \\ &\leq c_1\sqrt{n}\sum_{k=1}^{\infty} |\tilde{b}_k| \leq c_1 m\sqrt{n}\sum_{j=0}^{\infty} |\psi_{j,2}|. \end{aligned}$$

□

Lemma 7.3. *Under conditions of LRD, DEN and heavy-tailed, we have for any* $\rho \in (1/\gamma, \alpha)$

$$\|\sup_{|u|\leq b_n} |N_n(x+u) - N_n(x)|\|_\rho = O\left(H_m b_n m n^{1/\rho-\gamma} |l(n)|\right). \tag{7.7}$$

Proof. Similar to the proof of Lemma 7.2, it suffices to prove, for some $0 < C < \infty$,

$$\|R_n(x+u)\|_\rho \leq C m n^{1/\rho+1-\gamma} |l(n)| \text{ for all } u \in [-b_n, 1-b_n] \tag{7.8}$$

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

$$\begin{aligned} \|R_n(x+u)\|_\rho^\rho &= \|\sum_{k=-\infty}^{n-1} \mathcal{P}_k \sum_{i=m}^{n} f_\epsilon(H_m(x-\tilde{Z}_{i-1}))\|_\rho^\rho \qquad (7.9) \\ &\leq C_\rho \sum_{k=-\infty}^{n-1} \|\mathcal{P}_k \sum_{i=m}^{n} f_\epsilon(H_m(x-\tilde{Z}_{i-1}))\|_\rho^\rho \\ &\leq C_\rho \sum_{k=-\infty}^{n-1} (\sum_{i=m}^{n} \|\mathcal{P}_k f_\epsilon(H_m(x-\tilde{Z}_{i-1}))\|_\rho)^\rho \\ &\leq C_\rho (\sum_{k=-\infty}^{-n} + \sum_{k=-n+1}^{0} + \sum_{k=1}^{n-1})(\sum_{i=m}^{n} \|\mathcal{P}_k f_\epsilon(H_m(x-\tilde{Z}_{i-1}))\|_\rho)^\rho \\ &\leq C_\rho(I + II + III). \end{aligned}$$

Since $\mathbb{E}|\epsilon_i|^\rho) < \infty$, similarly as (7.6), we have for $k \leq i-1$ that

$$\|\mathcal{P}_k f_\epsilon(H_m(x - Z_{i-1}))\|_\rho \leq c_1 |\tilde{b}_{i-k}|, \tag{7.10}$$

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

$$\begin{aligned} I \leq c_1^\rho \sum_{k=-\infty}^{-n} (\sum_{i=m}^{n} |\tilde{b}_{i-k}|)^\rho &\leq c_1^\rho \sum_{k=n}^{\infty} (m \sum_{i=1}^{n} |\psi_{k+i,\rho}|)^\rho \\ &\leq c_1^\rho m^\rho n^{\rho-1} \sum_{k=n}^{\infty} \sum_{i=1}^{n} |\psi_{k+i,\rho}|^\rho \\ &= O[m^\rho n^{1+\rho(1-\gamma)} |l(n)|^\rho]. \end{aligned} \tag{7.11}$$

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

$$\begin{aligned} III \leq c_1^\rho \sum_{k=1}^{n-1} (\sum_{i=\max(m,k+1)}^{n} |\tilde{b}_{i-k}|)^\rho &\leq c_1^\rho \sum_{k=1}^{n-1} (m \sum_{i=0}^{n-k} |\psi_{i,\rho}|)^\rho \\ &= m^\rho \sum_{k=1}^{n-1} O[(n-k)^{1-\gamma} |l(n-k)|]^\rho \\ &= O[m^\rho n^{1+\rho(1-\gamma)} |l(n)|^\rho]. \end{aligned} \tag{7.12}$$

Similarly for term II we have, $II = O[m^\rho n^{1+\rho(1-\gamma)} |l(n)|^\rho]$. Combining this with (7.11) and (7.12), we finish the proof of the lemma. □

Proof of Theorem 4.1. By central limit theorem of Hannan (1979), we have $\tilde{Y}_i \xrightarrow{D} N(0, \sigma^2)$, where $\sigma = \|\sum_{i=0}^{\infty} \mathcal{P}_0 e_i\| < \infty$. Hence $\tilde{Q}(u)$ is well-defined and it converges to uth quantile of a $N(0, \sigma^2)$ distribution as $m \to \infty$. A standard characteristic function argument yields

$$\sup_x |f_m(x) - \phi(x/\sigma)/\sigma| \to 0, \tag{7.13}$$

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 $\bar{c}_n \to \infty$. Lemma 7.1 and 7.2 imply that

$$
\begin{aligned}
& |\tilde{F}_n(\tilde{Q}(u)+B_n)-\tilde{F}(\tilde{Q}(u)+B_n)-[F_n(\tilde{Q}(u))-\tilde{F}(\tilde{Q}(u))]| \\
= \; & O_{\mathbb{P}}\frac{B_n m^{3/2}}{\sqrt{n}}+m^{1/4}\sqrt{\frac{B_n}{n}}(\log n)^{1/2}) \\
= \; & o_{\mathbb{P}}(B_n), \qquad (7.14)
\end{aligned}
$$

where $B_n = \bar{c}_n m/\sqrt{n}$. Furthermore, similar arguments as those in Lemma 7.1 and 7.2 imply

$$
|\tilde{F}_n(\tilde{Q}(u))-\tilde{F}(\tilde{Q}(u))| = O_{\mathbb{P}}(\frac{m}{\sqrt{n}}) = o_{\mathbb{P}}(B_n). \qquad (7.15)
$$

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

$$
\tilde{F}(\tilde{Q}(u)+B_n)-\tilde{F}(\tilde{Q}(u)) = B_n f_m(\tilde{Q}(u))+O(B_n)^2. \qquad (7.16)
$$

By (7.13), $f_m(\tilde{Q}(u)) > 0$ for sufficiently large n. Plugging in (7.15) and (7.16) into (7.14), we have $\mathbb{P}(\tilde{F}_n(\tilde{Q}(u)+B_n) > u) \to 1$. Hence $\mathbb{P}(\hat{Q}_n(u) > \tilde{Q}(u)+B_n) \to 0$ by the monotonicity of $\tilde{F}_n(\cdot)$. Similar arguments yield $\mathbb{P}(\hat{Q}_n(u) < \tilde{Q}(u)-B_n) \to 0$. Using the fact that c_n can approach infinity arbitrarily slowly, we finish the proof of Theorem 4.1.

□

The proof for the quantile consistency results for the lasso fitted residuals in Section 4 requires the following important result from Bickel et al. (2009).

Lemma 7.4. *(Bickel et al., 2009) Assume the conditions on sparsity s and the restricted eigenvalue condition as presented in Theorem 4.3. Write*

$$
r = \max(A(n^{-1}\log p)^{1/2}\|e_\cdot\|_{2,\alpha}, B\|e_\cdot\|_{q,\alpha}\|X\|_q n^{\max(-1,-1/2-1/q-\alpha)}).
$$

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,

$$
r\|\hat{\beta}-\beta\|_1 + \|X(\hat{\beta}-\beta)\|_2^2/n \le 4r\|\hat{\beta}_J-\beta_J\|_1 \le 4r\sqrt{s}\|\hat{\beta}_J-\beta_J\|_2, \qquad (7.17)
$$

where $s = \#J$ with $J = \{j : \beta_j \neq 0\}$.

Proof of Theorem 4.3. In the view that Lemma 7.4 we have

$$
\mathbb{P}(\frac{1}{n}\|X(\hat{\beta}-\beta)\|_2^2 \ge 16sr/\kappa^2) \le \sum_{j=1}^p \mathbb{P}(|V_j| > r).
$$

Thus, applying the appropriate Nagaev inequality from Theorem 4.2

$$\sup_{m\leq i\leq n} \mid \sum_{k=i-m+1}^{i} (\hat{e}_i - e_i)| \leq m\|\hat{e} - e\|_\infty \leq m\sqrt{\frac{1}{n}\|X(\hat{\beta} - \beta)\|_2^2} = O_{\mathbb{P}}(m\sqrt{sr}). \quad (7.18)$$

since s satisfies the conditions in (4.10). Then the rest of the proof follows from the following observation; for any fixed $0 \leq u \leq 1$,

$$\bar{Q}_n(u) - \hat{Q}_n(u) = O_{\mathbb{P}}\left(\frac{m\sqrt{sr}}{H_m}\right), \quad (7.19)$$

where H_m is properly chosen for the heavy or light tails as described in (3.6). Then the right hand side by the choice of s as specified in (4.10) are smaller than the right hand side of the SRD specific cases mentioned in Theorem 3.4.

□

Proof of Theorem 4.4. Consider the event $B = \{n^{-1}|X^\mathsf{T}e|_\infty < r\}$. Since this is equivalent to the event $|V_j| < r$ for all j, the proof consists essentially of same steps as Theorem 4.3. In particular, note that we do not 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 on $\sum_{j=1}^{n} x_{l,j}e_j$ directly. Thus, in the view of the choice of r in (4.17), Theorem 4.4 follows.

□