Prediction Intervals in High-Dimensional Regression

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

[a] *Department of Statistics, University of Chicago, 5747 S. Ellis Avenue, Chicago, IL 60637, USA*
[b] *Department of Statistics and Operations Research, University of Vienna,*
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

Abstract

We construct prediction intervals for future time-aggregates of an univariate response time series. This may depend on (potentially infinitely) many predictors. We chose to use LASSO but our results can be easily extended to other estimators. Allowing for general stationary error processes including long-memory, heavy tailed and non-linear, we provide consistency results for prediction intervals using a sharp tail probability inequality. Finally, we construct prediction intervals for hourly electricity prices over horizons spanning 17 weeks and compare them to selected Bayes and model-implied alternatives.

Keywords: consistency, LAD, LASSO, electricity prices, bootstrap, time-aggregation

[*] Corresponding author

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

1. Introduction

Prediction intervals (p.i.'s) help forecaster to access the uncertainty about future values of time series. Although, there are situations, where point-forecasts are preferred, here we consider p.i.'s as our final goal. P.i.'s provide more information than simple point-forecasts and it is theoretically and practically challenging to show their validity (Chatfield, 1993; Clements and Taylor, 2003), since not only the coverage probability but also their width matters. Suppose that univariate target time series $y_i, i = 1, \dots, n$ follows a regression model

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

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 $S_m = y_{n+1} + \dots + y_{n+m}$, when both $n, m \to \infty$. Zhou et al. (2010) utilize LAD[1] estimator for the $\boldsymbol{\beta}$. However, if $p > n$ this and other conventional estimators, such as OLS, fail. Even if $p < n$ but large and $\boldsymbol{\beta}$ has many zero-elements, there are more efficient estimators, which lead to different p.i.'s $[L, U]^{\hat{\boldsymbol{\beta}}}$ such that for given α

$$P\left([L, U]^{\hat{\boldsymbol{\beta}}} \ni S_m\right) = 1 - \alpha.$$

In the current paper, we therefore extend the theoretical properties of the p.i.'s proposed by Zhou et al. (2010) to case of (potentially infinitely) many predictors. Although we present the results specifically for the LASSO estimator, they are applicable to other similar estimators as well. As a second contribution, we extend the theory of Zhou et al. (2010) to non-linear error process (e_i). In order to do so, we follow the functional dependence framework as proposed in a seminal paper by Wu (2005). Our error processes form much larger class including some popular examples such as smooth transition autoregression (Lundbergh et al., 2003).

Our main motivation comes from fields such as economics and finance, particularly energy industry, where the target generally depends on a large number of predictors. In addition, many of these time series are subject to structural breaks and other sources of complications for the forecaster (Koop and Potter, 2001; Cheng et al., 2016). It is generally accepted that inclusion of many (disaggregated) predictors provides some additional forecasting power over conventional univariate resp. low-dimensional approaches[2] (see Elliott et al., 2013; Kim and Swanson, 2014; Ludwig

[1] Least absolute deviation

[2] Yet there are empirical studies which do not corroborate these beliefs (Stock and Watson, 2012).

et al., 2015). But most empirical studies utilizing economic big data provide evidence based on short horizons and point-forecasts. This is partially because many series in finance (realized volatility of returns) or macroeconomics (GNP, inflation, interest rates, population, productivity and unemployment) have joint long-run characteristics. These characteristics can overthrow any transitory behaviour and cannot be ignored. But they also depend on nuisance parameters (Elliott et al., 2015). Dealing with them is beyond our scope and remains as future challenge for any sensible frequentist approach. Instead, we found interesting applications of our regression framework in electricity price forecasting (EPF). The electricity prices generally depend on exogenous variables including weather conditions, local economy and environmental policy[3] (Knittel and Roberts, 2005; Huurman et al., 2012). Additionally, EPF is challenging also due to complex seasonality (daily, weekly, yearly), heteroscedasticity, heavy-tails and sudden price spikes. There is a substantial amount of literature about EPF (see Weron, 2014, for recent review). We focus on long- and medium-horizon p.i.'s, which are essential for power portfolio risk management, derivatives pricing, medium-and long-term contracts evaluation and maintenance scheduling. Recently, Ludwig et al. (2015) found that inclusion of local (disaggregated) wind speed and temperature measured at 151 weather stations across Germany leads to forecasting improvements for EPEX SPOT electricity market. Since, they are focused on short-term point-forecasts, we try to verify if their findings hold for p.i.'s over longer horizons spanning up to 17 weeks. For this, we adopt their data set. Additionally, we include deterministic seasonal predictors and day of week indicators. The model and implementation details are described in the empirical part. For visual out-of-sample comparison of our p.i.'s we include alternative p.i.'s such as 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, which can be easily computed with automatic forecasting R-package "forecast" (see Hyndman and Khandakar, 2008).

The rest of the article is organized as follows: In Section 2, we construct the p.i.'s of Zhou et al. (2010) under different scenarios for number of regressors and for the error process (e_i). Section 3 and Section 4 summarize the asymptotic results for the cases without and with covariates. Section 5 shows simulation results and our real data analysis. Section 6 concludes.

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

2. Construction of prediction intervals

In this paper, we cover forecasting in a regression-set-up for a large number of scenarios. We are able to capture both linear and non-linear errors, short-range or long-range dependence, light-tailed or heavy-tailed behavior of the noise process, linear or robust regression in case of finitely many regressors and LASSO or others in case of infinitely many regressors under proper sparsity condition. Thus organization and presentation of these results in a concise manner is essential. The two methods we use in this paper for forecasting are from Zhou et al. (2010). However we provide some data-driven adjustments to arrive at better forecasting performance. The first of these is based on a quenched central limit theorem for short-range dependence and light-tailed error processes. The second one is more generally applicable since it is based on empirical quantiles. We first discuss, as a primer, how to forecast if there is no covariates present. Then we introduce finitely many covariates and finally conclude with infinitely many covariates.

2.1. Primary model: Without covariates

For this set-up, we have $y_i = e_i$ where e_i are zero-mean noise processes. Depending on the nature of dependence and tail behavior we can have the following two methods to estimate the quantiles. Note that, these methods are similar to those reported in Zhou et al. (2010) and Chudy et al. (2017).

2.1.1. Quenched CLT inspired method

For predicting m-step ahead aggregated response i.e $e_{n+1} + \ldots + e_{n+m}$, one can estimate the long-run variance σ^2 of the e_i process [HERE SUBSTITUTE THE OTHER ESTIMATOR?]

$$\hat{\sigma}^2 = \sum_{|i| \leq k_n} \hat{\gamma}_k = \sum_{|i| \leq k_n} \frac{1}{n} \sum_{j=1}^{n-|i|} (e_j - \bar{e})(e_{j+|i|} - \bar{e}),$$

and use the following version of the $100(1-\alpha)\%$ p.i.

$$[L, U] = \pm \hat{\sigma} t_{df,\alpha/2} \sqrt{m},$$

as desired prediction interval. The justification behind such an interval will be discussed in details as part of our asymptotic results where we show that under mild conditions the mean zero process $e_{n+1} + \ldots + e_{n+m}$ converge to an asymptotic normal distribution.

2.1.2. Empirical method based on quantiles

This is a substantially more general method that can account for long-range dependence or heavy-tailed behavior of the error process. The prediction intervals in this case will be

$$[L, U] = Q_{\alpha/2}, Q_{(1-\alpha)/2},$$

where Q_u is the u-th empirical quantiles of $\sum_{j=i-m+1}^{i} e_i; i = m, \ldots, n$. This simple prediction interval enjoys the advantage of interpretability, general applicability and still provides reasonable coverage for approprirate rate of growth of m compared to the size of observed sample n.

2.2. Finitely many covariates

We discuss two possible types of regression.

- *Least square regression* Assume the following model

  $$y_i = x_i^T \beta + e_i, \quad i = 1, \ldots, n,$$

  where β is a p-dimensional parameter vector. We wish to construct prediction interval for $y_{n+1} + \ldots + y_{n+m}$ after observing $(y_i, x_i); i = 1, \ldots, n$. If the error process show light tailed behavior and short-range dependence the popular least square regression estimates of β can lead to a good prediction interval. We estimate β by $\hat{\beta} = \arg\min \sum_i (y_i - x_i^T \beta)^2$ and then construct p.i. as

  $$\sum_{i=n+1}^{n+m} x_i^T \hat{\beta} + \text{ p.i. for } \sum_{i=n+1}^{n+m} \hat{e}_i. \tag{2.1}$$

  where $\hat{e}_i = y_i - x_i^T \hat{\beta}$ are the residuals from the estimation. Thus it suffices to discuss the construction of the p.i. for $\sum_{i=n+1}^{n+m} \hat{e}_i$ after observing $(y_i, x_i)_{i=1,\ldots,n}$. Since one of the major part of this paper also focuses on the scenario without the covariates we start discussing only how to construct the CI for $\sum_{i=n+1}^{n+m} e_i$ where $e_1, \ldots, e_n$ are observed. One can then replicate the same ideas by replacing e_i by $\hat{e}_i$. Theorem 4.5 shows the consistency properties for the case with the covariates.

- *Robust regression* For heavy-tailed or long-range dependent it is better (Huber and Ronchetti, 2009) to use robust regression to estimate the regression coefficient. In this case, the final prediction interval for the response y_i remains the same as (2.1) with $\hat{\beta}$ being estimated by a more general distance ρ

$$\hat{\beta} = \arg\min \sum_i \rho(y_i - x_i^T \beta).$$

Examples of such robust regression includes the $\mathbb{L}^q$ regression for $1 \leq q \leq 2$, quantile regressions $\rho(x) = qx^+ + (1-q)(-x)^{,}0 < q < 1$, where $x^+ = max(x, 0)$ and Huber's estimate $(x^2 \mathbf{1}_{|x| \leq c})/2 + (c|x|c^2/2)\mathbf{1}_{|x|>c}, c > 0$ etc.

2.3. Infinitely many covariates

Consider the case where the number of covariates $p \gg n$. We use LASSO to find the estimates of β

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{R}^p} \frac{1}{n}(y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.2}$$

where λ is the penalty parameter. We have a prediction interval of the form (2.1) with $\hat{\beta}$ replaced by the LASSO estimator. It is important to note that, there are other regression estimates in the scenario $p \gg n$ that can work here as well. However, we keep the discussion concise by discussing just the LASSO.

3. Asymptotic results without covariates

3.1. Linear Process

Assume

$$e_i = \sum_{j=0}^{\infty} a_j \epsilon_{i-j}. \tag{3.1}$$

Assume $E(|e_i|^p) < \infty$ for some $p > 2$. Wu and Woodroofe (2004) proved that, if for some $q > 5/2$,

$$\|E(S_m|\mathcal{F}_0)\| = O\left(\frac{\sqrt{m}}{\log^q m}\right), \tag{3.2}$$

then we have the a.s. convergence

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

where Δ denotes the Levy distance, $m \to \infty$ and $\sigma^2 = \lim_{m\to\infty} \|S_m\|^2/m$ is the long-run variance. Now, under linearity,

$$\|E(S_m|\mathcal{F}_0)\|^2 = \|(a_1 + \ldots + a_m)\epsilon_0 + (a_2 + \ldots + a_m)\epsilon_{-1} + \ldots\|^2 = \sum_{i=1}^{m} b_i^2, \tag{3.3}$$

where $b_i = a_i + \ldots + a_m$.

3.2. Non-linear Process

In this subsection, we propose extension of the results from Zhou et al. (2010) to non-linear processes. For the two methods mentioned above, we define a functional dependence measure of the non-linear process in two different ways. These assumptions are quite mild and easily verifiable compared to the more popularly used strong mixing conditions.

3.2.1. CLT for non-linear processes:Dependence structure

In this subsection, we relax the linearity assumption of e_i and assume a much more general set-up following Wu (2005)'s framework to formulate dependence through coupling. Assume e_i is a stationary process that admits the following representation

$$e_i = H(\mathcal{F}_i) = H(\epsilon_i, \epsilon_{i-1}, \ldots), \tag{3.4}$$

where H is such that e_i are well-defined random variable and $\epsilon_i, \epsilon_{i-1}, \ldots$ are independent innovations. One can see that it is a vast generalization from the linear structure of e_i assumed in (3.1). We need to define the dependence between (e_i) process. Define the following functional dependence measure

$$\delta_{j,p} = \sup_i \|e_i - e_{i,(i-j)}\|_p = \sup_i \|H_i(\mathcal{F}_i) - H_i(\mathcal{F}_{i,(i-j)}\|_p, \tag{3.5}$$

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,

$$\mathcal{F}_{i,k} = (\epsilon_i, \epsilon_{i-1}, \ldots, \epsilon'_k, \epsilon_{k-1}, \ldots) \tag{3.6}$$

and $e_{i,\{i-j\}} = H(\mathcal{F}_{i,\{i-j\}})$. Clearly, $\mathcal{F}_{i,k} = \mathcal{F}_i$ is $k > i$. As Wu (2005) suggests, $\|H(\mathcal{F}_i) - H(\mathcal{F}_{i,(i-j)})\|_p$ measures the dependence of X_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 and thus improves upon the usual strong mixing conditions which are often difficult to verify. Define the cumulative dependence measure

$$\Theta_{j,p} = \sum_{i=j}^{\infty} \delta_{i,p}, \tag{3.7}$$

which can be thought as cumulative dependence of $(X_j)_{j \geq k}$ on ϵ_k. For the quenched CLT in (3.2), we assume the following rate for $\Theta_{j,p}$.

$$\Theta_{j,p} = j^{-\chi}(\log j)^{-A} \text{ where } = \begin{cases} A > 0 \text{ for } 1 < \chi < 3/2, \\ A > 5/2 \text{ for } \chi \geq 3/2, \end{cases} \tag{3.8}$$

.

The m-dependence approximation is a key idea for the proof for the non-linear case,

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

where $\tilde{S}_m = \sum_{i=1}^m \tilde{X}_i = \sum_{i=1}^m E(X_i|\epsilon_i, \dots \epsilon_{i-m})$. The proof of (3.2) follows along the line of (3.3) from the facts $\|P_j(\tilde{X}_i)\|_2 \leq \delta_{i-j,2}$,

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

However, one important limitation of the quenched CLT based method is that one cannot apply this if the tail behavior of the error process is heavy and the error process is long-range dependent. The quantile based method is more generally applicable.

3.2.2. Quantile estimation: Dependence structure

For the non-linear case however, one does not have such decomposition of the error process. Since the coefficients a_j in the decomposition measures how much e_i depend on ϵ_{i-j}, it will be beneficial to somehow control this dependence. With this motivation, we use the predictive density-based dependence measure. We assume e_i admits the following causal representation

$$e_i = H(\epsilon_i, \epsilon_{i-1}, \dots), \tag{3.9}$$

where ϵ_i are i.i.d. Let $\mathcal{F}_k$ denote the σ-field generated by $(\epsilon_k, \epsilon_{k-1}, \dots)$. Let (ϵ'_i) be an i.i.d. copy of (ϵ_i) and

$$\mathcal{F}'_k = (\dots, \epsilon_{-1}, \epsilon'_0, \epsilon_1, \dots, \epsilon_k),$$

be the coupled shift process. Let $F_1(u, t|\mathcal{F}_k) = P\{G(t; \mathcal{F}_{k+1}) \leq u|\mathcal{F}_k\}$ 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

$$\psi_{k,q} = \sup_{t\in[0,1]} \sup_{u\in\mathbb{R}} \|f_1(u,t|\mathcal{F}_k) - f_1(u,t|\mathcal{F}'_k)\|_q. \tag{3.10}$$

Quantity (3.10) 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:

1. Smoothness (third order continuous differentiability): $f, m, \sigma \in C^3(\mathbb{R}[0,1])$;
2. For short-range dependence: $\Psi_{0,2} < \infty$ where $\Psi_{m,q} = \sum_{k=m}^{\infty} \psi_{m,q}$

   For long-range dependence: $\Psi_{0,2}$ can possibly be infinite.
3. (DEN) condition: There exists a constant $c_0 < \infty$ such that almost surely,

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

The (DEN) Condition (3) implies that the marginal density $f(u,t) = Ef_1(u,t|\mathcal{F}_0) \leq c_0$. Next define dependence adjusted norm

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

3.2.3. Quantile estimation consistency for non-linear process

Using the dependence measure on predictive densities, we will be able to extend the results from Zhou et al. (2010) to a more general non-linear set-up. Recall the sufficient conditions for the linear cases were based on the coefficients of the linear process. Here, however, the conditions will be based on the dependence measures. Recall the functional dependence measure defined at (3.10). Assume

$$\begin{aligned} \text{(SRD)} &: \sum_{j=0}^{\infty} |\psi_{j,q}| < \infty, \qquad (3.11) \\ \text{(LRD)} &: \psi_{j,q} = j^{-\gamma} l(j), \gamma < 1, l(\cdot) \text{ is slowly varying function (s. v. f.) .} \end{aligned}$$

We propose the following result as our new contribution in this paper for the non-linear processes.

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

$$\tilde{Y}_i = \frac{\sum_{j=i-m+1}^{i} e_j}{H_m}, \quad i = m, m+1, \ldots \tag{3.12}$$

and

$$H_m = \begin{cases} \sqrt{m}, & \text{if (SRD) holds and } \mathbb{E}(\epsilon_j^2) < \infty, \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x) \leq \frac{1}{m}\} & \text{if (SRD) holds and } \mathbb{E}(\epsilon_j^2) = \infty, \\ m^{3/2-\gamma} l(m) & \text{if (LRD) holds and } \mathbb{E}(\epsilon_j^2) < \infty, \\ \inf\{x : \mathbb{P}(|\epsilon_i| > x) m^{1-\gamma} l(m) & \text{if (LRD) holds and } \mathbb{E}(\epsilon_j^2) = \infty. \end{cases} \tag{3.13}$$

We have the following different rates of convergence of quantiles based on the nature of tail or dependence:

Theorem 3.1. *[Quantile consistency result:Non-linear error]*

- *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 $0 < q < 1$,*

$$|\hat{Q}(q) - \tilde{Q}(q)| = O_{\mathbb{P}}(m/\sqrt{n}). \tag{3.14}$$

- *Light tailed (LRD): Suppose (LRD) and (DEN) hold with γ and $l(\cdot)$ in (3.11). If $m^{5/2-\gamma} n^{1/2-\gamma} l^2(n) \to 0$, then for any fixed $0 < q < 1$,*

$$|\hat{Q}(q) - \tilde{Q}(q)| = O_{\mathbb{P}}(mT^{1/2-\gamma}|l(n)|). \tag{3.15}$$

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

$$|\hat{Q}(q) - \tilde{Q}(q)| = O_{\mathbb{P}}(mn^\nu) \text{ for all } \nu > 1/\alpha - 1. \tag{3.16}$$

—

- *Heavy-tailed (LRD): Suppose (LRD) hold with γ and $l(\cdot)$ in (3.11). If $m = O(n^k)$ for some $k < (\alpha\gamma - 1)/(2\alpha + 1 - \alpha\gamma)$, then for any fixed $0 < q < 1$,*

$$|\hat{Q}(q) - \tilde{Q}(q)| = O_{\mathbb{P}}(mn^\nu) \text{ for all } \nu > 1/\alpha - \gamma. \tag{3.17}$$

4. Asymptotic results in presence of covariates

We divide our asymptotic results based on the estimation of regression coefficient β. The usual linear or robust regression is more straight-forward whereas one would need sparsity conditions imposed for the LASSO estimation for presence of infinitely many regressors.

4.1. Regression with finitely many regressors

Theorem 4.1 (Residual consistency for regression)**.**

$$\sum_i |\hat{e}_i - e_i| = o_P(\Pi(n)).$$

where the error bound $\Pi(n)$ will be different for different behavior of the error process e_i.

4.2. Regression with infinitely many regressors

4.2.1. Tail Probability inequality

We discuss a key tail probability inequality for the different settings as this can be of independent interest. Let $S_{n,b} = \sum b_i e_i$.

Theorem 4.2 (Nagaev inequality for linear processes)**.** *We have the following tail probability bounds of $S_{n,b}$ for the four different settings.*

- *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,*

$$P(|S_{n,b}| \geq x) \leq (1 + 2/q)^q \frac{|b|_q^q (\sum_j |a_j|)^q \|\epsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{c_q x^2}{n(\sum_j |a_j|)^2 \|\epsilon_0\|_2^2}\right), \tag{4.1}$$

- *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,*

$$P(|S_{n,b}| \geq x) \leq C_1 \frac{K^q |b|_q^q \|n^{q(1-\beta)} \epsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{C_2 x^2}{n^{3-2\beta} \|\epsilon_0\|_2^2 K^2}\right), \tag{4.2}$$

- *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*

$$P(|S_{n,b}| \geq x) \leq (1+2/q)^q \frac{|b|_q^q (\sum_j |a_j|)^q \|\epsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{c_q x^2}{n(\sum_j |a_j|)^2 \|\epsilon_0\|_2^2}\right), \quad (4.3)$$

- *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,*

$$P(|S_{n,b}| \geq x) \leq C_1 \frac{K^q |b|_q^q \|n^{q(1-\beta)}\epsilon_0\|_q^q}{x^q} + 2\exp\left(-\frac{C_2 x^2}{n^{3-2\beta}\|\epsilon_0\|_2^2 K^2}\right), \quad (4.4)$$

For non-linear error process, however, it is difficult to discuss long-range dependence as one needs an appropriate model for that. For definiteness, we stick to short range dependence which means $\Theta_{0,q} < \infty$ where q is less or more than 2 depending on the tail-behavior of the error process

Theorem 4.3 (Nagaev inequality for non-linear processes)**.** *For short-range dependent processes, we have the following two versions of Nagaev inequality*

- *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,*

$$P(|S_{n,b}| \geq x) \leq C_1 \frac{r_n}{(\sum_j |b_j|)^q \|e.\|_{q,\alpha}^q} x^q + C_2 \exp\left(-\frac{C_3 x^2}{n\|e.\|_{2,\alpha}^2}\right), \quad (4.5)$$

- *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, C_2, C_3 that depend on only q and α,*

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

Our main result for this section will be Theorem 4.5 and it will say that the error bounds obtained in Theorem 3.1 remain intact if we make a proper choice of the sparsity condition. Before that, we state a crucial lemma from Bickel et al. (2009).

Lemma 4.4. *Let $\lambda = 2r$ in (2.2). Also assume,*

$$r = \max(A(n^{-1}\log p)^{1/2}\|e_.\|_{2,\alpha}, B\|e_.\|_{q,\alpha}|X|_q/n) \tag{4.7}$$

On the event

$$\mathcal{A} = \bigcup_{j=1}^{p}\{2|V_j| \le r\}, \text{ where } V_j = \frac{1}{n}\sum_{i=1}^{n} e_i x_{ij},$$

we have,

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

Remark This allows us to use Nagaev inequality from Theorem 4.2 and 4.3 to V_j.

Let $\bar{Q}_n(q)$ be the qth empirical quantile of $(\tilde{\hat{Y}}_i)_m^n$.

Theorem 4.5. *[Quantile consistency for LASSO] Assume s, the number of non-zero coordinates in β satisfies the following CHECK MATH:*

$$\begin{aligned} s^2\log p &\ll n & (4.8)\\ s &\ll \frac{n^{1-\max\{0,1/2-1/q-\alpha\}}}{|X|_q}, & (4.9)\\ s &\ll \frac{n^{\frac{2}{\alpha}+\frac{\alpha-1}{\alpha+1}}L_1(n)}{|X|_q}, & (4.10)\end{aligned}$$

Then the conclusions in Theorem 3.1, and Theorem 4.3 hold with $Q_n(q)$ replaced by $\bar{Q}_n(q)$.

5. Simulation and real data evaluation

5.1. Simulation

The focus is on evaluation of p.i.'s discussed in previous section based on their coverage probability. We start by generating the error process (e_i) as:

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

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

(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 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 selected based on respective models fitted to the electricity prices used later in 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) heavy-tail and short-memory error-process (b) heavy-tail and long-memory error-process and (c) non-linear and heavy tail error-process. The latter process is know as logistic smooth transition autoregression (LSTAR). Eventually, we add the exogenous covariates to the error process, obtaining:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i, i = 1, \ldots, m + n. \tag{5.1}$$

The length of the simulated data is $n + m$. The first n values are for estimation and last m values for evaluation, i.e., we compute the p.i.'s based on $(y_1, \boldsymbol{x}_1) \ldots, (y_n, \boldsymbol{x}_n)$ in order to predict the target $\bar{y}_{+1:m} = 1/m \sum_{i=1}^{m} y_{n+i}$. We forecasts averages instead of sums of future values, 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 the three estimators of $\boldsymbol{\beta}$, i.e. OLS, LAD and LASSO. In case (ii), we only have the LASSO, since the other two estimators are not 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 described in the empirical part. The elements of $\boldsymbol{\beta} \in \mathbb{R}^p$ in (5.1) are obtained as i.i.d draws from either uniform distribution $U[-1, 1]$ or Cauchy distribution. Finally, properties of LASSO estimator depend on sparsity assumption, hence we exploit different scenarios for sparsity of $\boldsymbol{\beta}$, namely for $s = 100(1 - ||\beta||_0/p) = 99\%, 90\%, 70\%, 50\%, 20\%$. For the LASSO, this can be seen as robustness check towards the sparsity assumption. Once computed, $\boldsymbol{\beta}$ is fixed for all 1000 repetitions of the experiment. The effects of different sparsity-scenarios on p.i.'s based on OLS and LAD seem negligible, therefore, we report the results for OLS and LAD separately from LASSO.

For sake of brevity, only p.i.'s with nominal coverage (n.c.) $100(1 - \alpha) = 90\%$ are reported in this paper[4]. We compute the coverage probability (c.p.) $\widehat{(1 - \alpha)} =$

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

$\frac{1}{1000}\sum_{i=1}^{1000}\mathbb{I}\left([L,U]_i^{\hat{\boldsymbol{\beta}}} \ni \bar{y}_{i,+1:m}\right)$, where $\mathbb{I}$ for the i-th trial is 1 when $\bar{y}_{i,+1:m}$ is covered by the $[L,U]_i$ and 0 otherwise.

Starting with the case $p < n$ 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 $p = 319$, which is close to the set-up of our empirical application. The major difference, however is in the maximal horizon, which will reach 17 weeks later in the empirical part. First thing, there is barely a difference in c.p. based on possibility that $\boldsymbol{\beta}$ elements are drawn from uniform or Cauchy distributions for the OLS-QTL and LAD-QTL. For both LASSO-CLT and -QTL, the c.p.'s are generally higher when coefficients are uniformly distributed. Under Cauchy, the CLT p.i.'s exhibits very low c.p.'s when sparsity is low and the error process has heavy tails. But in case the error process has also long-memory, which provides a counter effect, the CLT's p.i.'s have higher c.p. Despite this fact, the CLT may still have only half the c.p. of QTL. The effect of non-linearity seems small compared to the effect of long-memory. Generally, we see that c.p.'s decrease with growing horizon. The effect of horizon on LASSO-CLT and OLS-QLT is higher than the effect on LASSO- and LAD-QTL. For uniformly distributed coefficients, the LASSO-QTL almost always provides the highest c.p. close to the n.c. (90%). Even if the sparsity of $\boldsymbol{\beta}$ decreases, the c.p. of LASSO-QLT is highest save for case $s = 10\%$ and $m = 4$ weeks. But for Cauchy distribution, the winner is not clear, because if $s < 80\%$ and $m > 2$ weeks, LAD-QTL often gives higher c.p. with exception of long-memory errors, where the LASSO-QTL always wins.

For the case $p > n$ 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 much shorter. Notice however, that the maximal proportion $m/n > 1/4$. The effect of such large m to n proportion on the QTL is clearly negative. As a remedy for this obvious shortcoming, we exploit a data-driven adjustment to the QTL method. The adjustments is based on replication of the residual $\hat{e}_i = y_i - \hat{y}_i$ using stationary bootstrap (Politis and Romano, 1994) and estimation of p.i.'s by kernel quantile estimator (Falk, 1984) instead of the sample version suggested by Zhou et al. (2010). We employ this adjusted QTL (ADJ) p.i.'s also in the empirical part of this section where we forecast the spot electricity prices. Although the sample size will be large, $m/n \approx 1/3$, which is similar to the current case. For details about these adjustments, see the implementation in the following empirical part. As already noted, in this case we have only the LASSO estimator. Hence our focus is on comparison of CLT, QTL and ADJ. First thing to notice when comparing with previous case (Table 1i) is a decrease of c.p. for QTL. Surprisingly, it turns to opposite for the CLT. Generally in this setup, the CLT gives higher c.p.

than QLT, particularly if the process has long-memory or if the sparsity is low. On the other hand, we emphasize the improvement of c.p. provided by ADJ. Although the c.p. of the ADJ is lower for case $m = 1$ day, it does not decrease so rapidly when horizon grows. In fact, ADJ can outperform both QTL and CLT by more than 10pp for longer horizons. Despite this improvement, the c.p. may still fall down to 60% when the error process shows long-memory.

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
	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-week	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 are have short-memory & heavy tails or long-memory & heavy tails or are non-linear & have heavy tails. The elements of regression coefficient $\boldsymbol{\beta}$ are i.i.d. from uniform distribution $U[-1, 1]$ or Cauchy distribution. The sparsity of $\boldsymbol{\beta}$'s varies form 99% to 20%.

5.2. 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 in Sections 4 and 5.1,

(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 auto-regression (Hyndman and Athanasopoulos, 2013, sec. 9.3).

Data. We forecast $\bar{y}_{+1:m} = \sum_{t=1}^{m} y_{n+t}$, i.e. the average[5] 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 market operating 24 hours a day, we have 11640 observations between 01/01/2013 00:00:00 UTC[6] 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. Such complex 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 Figureoa, 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.

As mentioned in Section 1, the local weather variables are also used as predictors. 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 over period of 5 years (2009-2013) i.e. including the training period but not the evaluation period (see above). In order

[5] 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 common basis for the comparison.

[6] Coordinated Universal Time.

to approximate some missing in-sample data and unobserved values for evaluation period, we take hourly-specific-averages[7] of each weather time series over these 5 years.

In total, we have 168 trigonometric predictors, 151 weather predictors and 2 dummies which gives a full set of 321 predictors. We denote these predictors

$$x_t^T = (d_{\text{sa}}, d_{\text{su}}, \sin(\omega_1 t), \cos(\omega_1 t), \dots, , \sin(\omega_{84} t), \cos(\omega_{84} t), w_{1,t}, \dots, w_{73,t}, \tau_{1,t}, \dots, \tau_{78,t}), \tag{5.2}$$

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

Methods. In Figure 1B, we see a drop of the price level during December 2013. The forecasts based on the whole training period would therefore suffer from bias. On the contrary, 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 $v_{n-t+1} = \delta^{t-1}((1-\delta))/(1-\delta^t)$, $t = 1, \dots, n$ 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:
ADJ:

1. Estimate regression $y_t \sim x_t^T$, $t = 1 \dots, n$ with LASSO (Friedman et al., 2010).
2. Replicate residuals $\hat{e}_t = y_t - \hat{y}_t$, B times obtaining $\hat{e}_t^b, t = 1, \dots, n$, $b = 1, \dots, B$.
3. Compute $(\bar{e}_{t(m)}^b) = m^{-1} \sum_{i=1}^m e_{t-i+1}^b$, $t = m, \dots n$ from every replicated series.
4. 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 $\bar{e}_{n(m)}^b$, $b = 1, \dots, B$.
5. 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 $h = 1, \dots, m$.

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

RBS:

We prefer to give the reader an intuition instead of detailed implementation steps for this method. Additional details about the implementation can be found in the supplementary Appendix of Müller and Watson (2016). The p.i's are specifically designed for long-horizon predictions, in particular for case $m/n \approx 1/2$. It is a univariate approch, i.e., the p.i.'s are obtained only from past values of y_t. First, the high-frequency noise is partialed out from y_t by low-frequency cosine transformations. Projecting $\bar{y}_{+1:m}$ on the space spanned by these transformations is the key to obtain conditional distribution of $\bar{y}_{+1:m}$. In order to expand the class of process 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.

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

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

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

4. 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:

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

2. 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 predictors (see the supplementary Appendix for details).

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

4. 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 competitors, we explore the benefits from using disaggregated weather data. Therefore, we compute the p.i.'s (i) using no regressors as in Figure 2IA, (ii) using only deterministic regressors as in Figure 2IB, (iii) 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, (iv) using all 321 predictors as in Figure 2ID. Generally, we see only little difference between the first three plots as all respective p.i.'s are obviously biased. A striking improvement is achieved by including disaggregated weather series.

Finally getting to the comparison with alternative methods, which we see in Figure 2II, the only methods, providing sensible p.i.'s are RBS and ETS. RBS works well over the whole 17-weeks-long evaluation period (Figure 2IIA). However, when compared to the ADJ, the p.i's seem a bit too wide, hence lacking some precision. ETS become too wide as the horizon grows. The NAR is even more biased than the ADJ without predictors, especially for large m. Not so surprisingly, the ARX perform worst of all methods. It might be that the down-weighting provided by autoregression is simply too mild. Besides, the narrow p.i.'s are result of ignoring the parameter (among other types of) uncertainty.

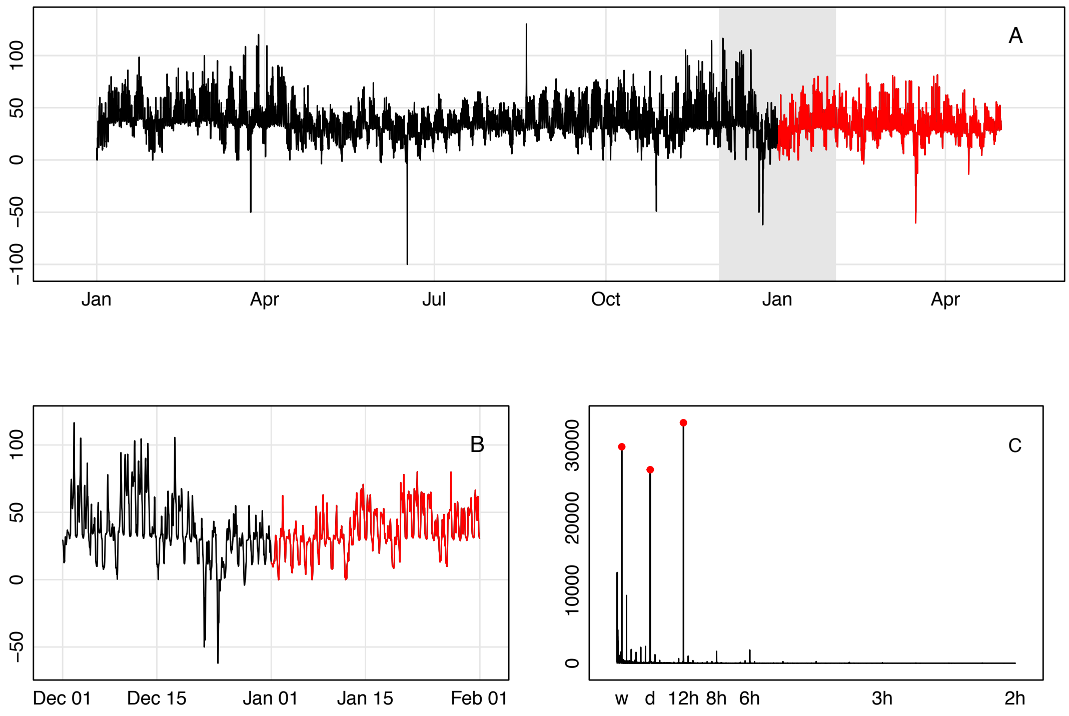
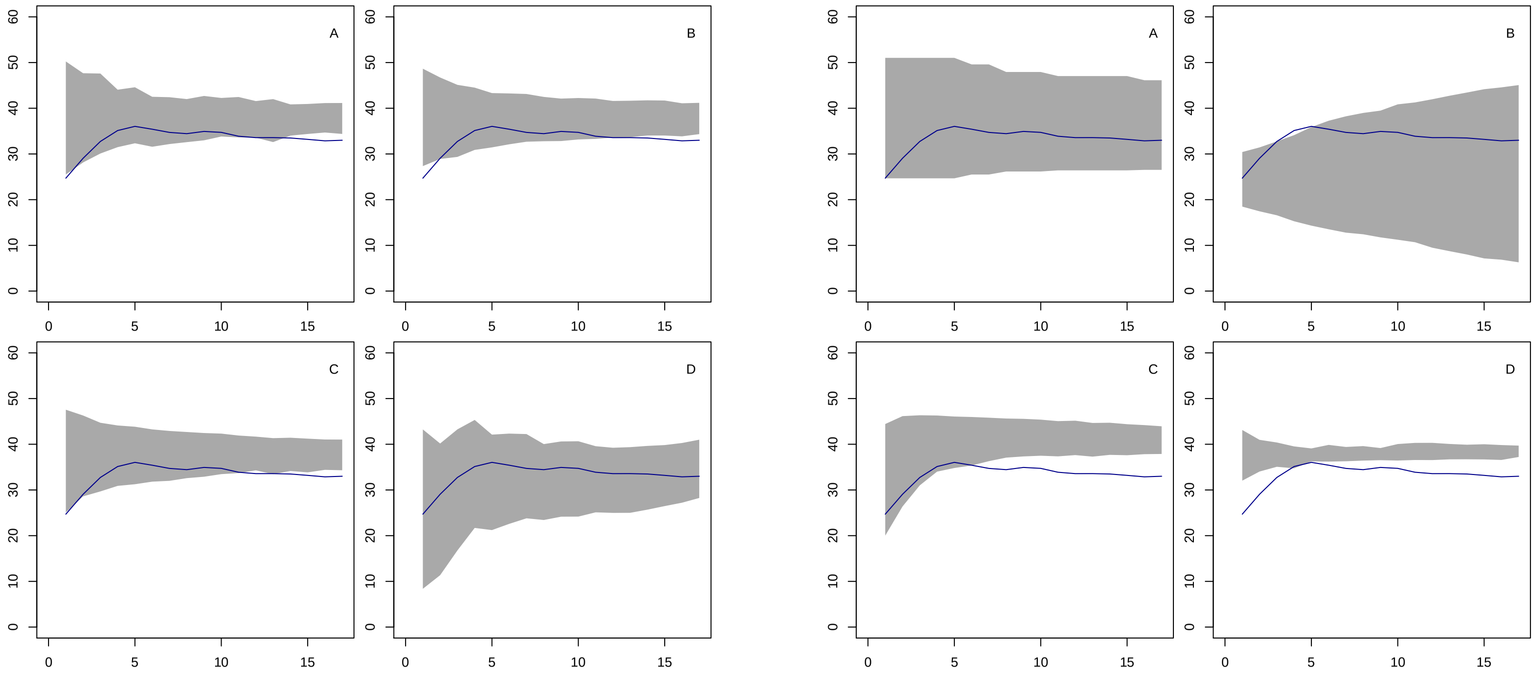


Figure 1: Electricity spot prices, A) Full sample, B) Drop in price level, C) Periodogram with peaks at periods 1 week, 1 day and 12 hours. According to European convention, the term spot refers to day-ahead rather than real-time unlike in the US, where term forward is common.



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

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

6. Discussion

We considered constructing empirically valid prediction intervals for high-dimensional regression. From the theoretical perspective, we have extended the results of Zhou et al. (2010) into high-dimensional set-up. For low dimensional set-up, we extended the results also to case of non-linear error process.

The quantile method was successfully applied to predict spot electricity prices for Germany and Austria using large set of local weather time series. The results proved superiority their over conventional exponential smoothing and neural network bootstrapped prediction intervals as well as recently proposed low-frequency approach of Müller and Watson (2016). Regarding our application to electricity price forecasting, it would be interesting to consider even larger set of predictors, e.g., augmented by macroeconomic predictors like fuel prices and GDP.

Possible extensions to the current paper 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).

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.

Chudy, M., S. Karmakar, and W. Wu (2017). Long-term prediction intervals for economic time series. *preprint*.

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.

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

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

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

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

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.

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.

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.

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

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. *International Journal of Forecasting 30*, 4.

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

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

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

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 - Proofs
Proof of theorem 1
Lemma B.1 from 2009 lasso by bickel
Nagaev on V_j
Proof of lemma
Quantile consistency from Xiao, Xu and Wu
Proof for the non-linear case.
Define $\tilde{Z}_i$ as follows

$$\tilde{Z}_{i-1} = \frac{\sum_{j=1}^{\infty} \tilde{b}_j \epsilon_{i-j}}{H_m} \tag{6.1}$$

where $\tilde{b}_j = a_0 + a_1 + \ldots + a_j$ if $1 \leq j \leq m-1$ and $\tilde{b}_j = a_{j-m+1} + a_{j-m+2} + \ldots + a_j$ if $j \geq m$.

Define

$$\tilde{F}_n^*(x) = \frac{1}{n-m+1} \sum_{i=m}^{n} F_\epsilon(H_m(x - \tilde{Z}_{i-1})),$$

where $F_\epsilon(\dot{})$ is the distribution function of ϵ. Let $tildeF(x) = P(\tilde{Y}_i \leq x)$. We write

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

Define $P_i(Y) = E(Y|\mathcal{F}_i) - E(Y|\mathcal{F}_{i-1})$. Using this, one can write $M_n(x)$ as follows

$$M_n(x) = \frac{1}{n-m+1} \sum_{i=m}^{n} P_i(I(Y_i \leq x)) \tag{6.2}$$

Lemma 6.1. *Under conditions of Theorem 4.2 and Theorem 4.3,*

$$\sup_{|u| \leq b_n} |M_n(x+u) - M_n(x)| = O_p\left(\sqrt{\frac{H_m b_n}{n}} \log^{1/2} n + n^{-3}\right), \tag{6.3}$$

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

Proof. Let $c_0 = \sup_x |f_\epsilon(x)| < \infty$. Since $P(\tilde{Y}_i \leq x|\mathbb{F}_{i-1}) = F_\epsilon(H_m(x - \tilde{Z}_{i-1}))$, we have $P(x \leq \tilde{Y}_i \leq x + u|\mathbb{F}_{i-1}) \leq H_m c_0 u$ for all $u > 0$. Therefore for any $u \in [-b_n, b_n]$, we have

$$\sum_{i=m} n[E(V) - E^2(V)] \leq c_0(n-m+1)H_m b_n \quad \text{where } V = I(x \leq \tilde{Y}_i \leq x+u|\mathbb{F}_{i-1}) \tag{6.4}$$

Applying Freedman's martingale inequality and a chaining argument, we have (6.3). Since the chaining argument is essentially similar to Lemma 5 in , Lemma 4 in and Lemma 6 in we skip the details □

Lemma 6.2. *Under conditions of SRD, DEN and light-tailed*

$$\| \sup_{|u| \leq b_n} |N_n(x+u) - N_n(x)| \| = O\left(\frac{b_n m^{3/2}}{\sqrt{n}}\right) \tag{6.5}$$

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

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

. Since , we are left to prove

Hence,

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

.

Let $(\epsilon'_i)_{-\infty}^{\infty}$ be an i.i.d. copy of $(\epsilon_i)_{-\infty}^{\infty}$ and $\tilde{Z}^*_{i-1,k} = \tilde{Z}_{i-1} - \tilde{b}_k \epsilon_{i-k}/\sqrt{m} + \tilde{b}_k \epsilon'_{i-k}/\sqrt{m}$. Note that for $k \geq 1$,

$$\begin{aligned} \|\mathcal{P}_{i-k} f_\epsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\| &\leq \|f_\epsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1})) - f_\epsilon(\sqrt{m}(x+u-\tilde{Z}^*_{i-1,k}))\| && (6.6) \\ &\leq \sup_{v \in \mathbb{R}} |f'_\epsilon(v)| \sqrt{m}(\tilde{Z}_{i-1} - \tilde{Z}^*_{i-1.k})\|\|\| \leq c_1 \tilde{b}_k && (6.7) \end{aligned}$$

where $c_1 = \sup_{vin\mathbb{R}} \|\|\|\epsilon_0 - \epsilon'_0\| < \infty$. 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 $\mathcal{P}_{i-k}, i = m, \ldots, n$

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

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

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

Recall the functional dependence measure. The proofs for the linear cases will go through if we replace f_ϵ by f_1, a_j by $\delta_{j,2}$.

□

Lemma 6.3. *Under conditions of LRD, DEN and heavy-tailed, we have for any* $\rho\in(1/\gamma,\alpha)$

$$\|\sup_{|u|\le b_n}|N_n(x+u)-N_n(x)|\|_\rho=O\left(H_m b_n m n^{1/\rho-\gamma}|l(n)|\right)\tag{6.8}$$

Proof. Similar to the proof of Lemma 6.2, it suffices to prove, for some $0<C<\infty$,

$$\|R_n(x+u)\|_\rho\le Cmn^{1/\rho+1-\gamma}|l(n)|\text{ for all }u\in[-b_n,1-b_n]\tag{6.9}$$

Since $1<\rho<2$, by Burkholder's inequality of martingales, we have, with $C_\rho=[18\rho^{3/2}(\rho-1)^{-1/2}]^\rho$.

$$\begin{aligned}\|R_n(x+u)\|_\rho^\rho &= \|\sum_{k=-\infty}^{n-1}\mathcal{P}_k\sum_{i=m}^{n}f_\epsilon(H_m(x-\tilde{Z}_{i-1}))\|_\rho^\rho \qquad (6.10)\\ &\le C_\rho\sum_{k=-\infty}^{n-1}\|\mathcal{P}_k\sum_{i=m}^{n}f_\epsilon(H_m(x-\tilde{Z}_{i-1}))\|_\rho^\rho\\ &\le C_\rho\sum_{k=-\infty}^{n-1}(\sum_{i=m}^{n}\|\mathcal{P}_kf_\epsilon(H_m(x-\tilde{Z}_{i-1}))\|_\rho)^\rho\\ &\le C_\rho(\sum_{k=-\infty}^{-n}+\sum_{k=-n+1}^{0}+\sum_{k=1}^{n-1})(\sum_{i=m}^{n}\|\mathcal{P}_kf_\epsilon(H_m(x-\tilde{Z}_{i-1}))\|_\rho)^\rho\\ &\le C_\rho(I+II+III),\end{aligned}$$

Since $E(|\epsilon_i|^\rho) < \infty$, similarly as (6.6), we have for $k \le i-1$ that

$$\|\mathcal{P}_k f_\epsilon(H_m(x - Z_{i-1}))\|_\rho \le c_1 |\tilde{b}_{i-k}|, \tag{6.11}$$

where $c_1 = \sup_{v\in\mathbb{R}} |f'_\epsilon(v)| \|\epsilon_0 - \epsilon'_0\|_\rho < \infty$. Thus using Karamata's theorem for the term I, we have

$$\begin{aligned} I \le c_1^\rho \sum_{k=-\infty}^{-n} (\sum_{i=m}^{n} |\tilde{b}_{i-k}|)^\rho &\le c_1^\rho \sum_{k=n}^{\infty} (m \sum_{i=1}^{n} |a_{k+i}|)^\rho \\ &\le c_1^\rho m^\rho n^{\rho-1} \sum_{k=n}^{\infty} \sum_{i=1}^{n} |a_{k+i}|^\rho \\ &= O[m^\rho n^{1+\rho(1-\gamma)} |l(n)|^\rho] \end{aligned} \tag{6.12}$$

Since $\rho > 1$ and $\rho\gamma > 1$, we use H"o"lder inequality to manipulate term III as follows:

$$\begin{aligned} III \le c_1^\rho \sum_{k=1}^{n-1} (\sum_{i=\max(m,k+1)}^{n} |\tilde{b}_{i-k}|)^\rho &\le c_1^\rho \sum_{k=1}^{n-1} (m \sum_{i=0}^{n-k} |a_i|)^\rho \\ &= m^\rho \sum_{k=1}^{n-1} O[(n-k)^{1-\gamma} |l(n-k)|]^\rho \\ &= O[m^\rho n^{1+\rho(1-\gamma)} |l(n)|^\rho]. \end{aligned} \tag{6.13}$$

Similarly for term II we have, $II = O[m^\rho n^{1+\rho(1-\gamma)} |l(n)|^\rho]$. Combining this with (6.12) and (6.13), we finish the proof of the lemma. □

of Theorem 4.2. 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 $\tilde{Q}_q$ is well-defined and it converges to qth quantile of a $N(0, \sigma^2)$ distribution as $m \to \infty$. Furthermore, note that e_i is a weighted sum of i.i.d. random variables and the density $f_\epsilon(\cdot)$ is bounded. Hence a standard characteristic function argument yields

$$\sup_x |f_m(x) - \phi(x/\sigma)/\sigma| \to 0, \tag{6.14}$$

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 6.1 and 6.2 imply that

$$\begin{aligned} |\tilde{F}_n(\tilde{Q}_q + B_n) - \tilde{F}(\tilde{Q}_q + B_n) - [F_n(\tilde{Q}_q) - \tilde{F}(\tilde{Q}_q)]| &= O_P(\frac{B_n m^{3/2}}{\sqrt{n}} + m^{1/4}\sqrt{\frac{B_n}{n}}(\log n)^{1/2}) \\ &= o_p(B_n), \end{aligned} \tag{6.15}$$

where $B_n = \bar{c}_n m/\sqrt{n}$. Furthermore, similar arguments as those in Lemma 6.1 and 6.2 imply

$$|\tilde{F}_n(\tilde{Q}_q) - \tilde{F}(\tilde{Q}_q)| = O_p(\frac{m}{\sqrt{n}}) = o_P(B_n). \tag{6.16}$$

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

$$\tilde{F}(\tilde{Q}_q + B_n) - \tilde{F}(\tilde{Q}_q) = B_n f_m(\tilde{Q}_q) + O(B_n)^2. \tag{6.17}$$

By (6.14), $f_m(\tilde{Q}_q) > 0$ for sufficiently large n. Plugging in (6.16) and (6.17) into (6.15), we have $P(\tilde{F}_n(\tilde{Q}_q + B_n) > q) \to 1$. Hence $P(\hat{Q}_n(q) > \tilde{Q}_q + B_n) \to 0$ by the monotonicity of $\tilde{F}_n(\cdot)$. Similar arguments yield $P(\hat{Q}_n(q) < \tilde{Q}_q - B_n) \to 0$. Using the fact that c_n can approach infinity arbitrarily slowly, we finish the proof of Theorem 4.2.

□

Proof of Theorem 4.5. From Lemma 4.4, we have

$$\sup_{m \le i \le n} |\sum_{k=i-m+1}^{i} (\hat{e}_i - e_i)| = O_p(\pi(n)), \tag{6.18}$$

for a suitable π depending on the λ and sparsity s. Thus

$$\bar{Q}_n(q) - \hat{Q}_n(q) = O_p\left(\frac{\pi(n)}{H_m}\right). \tag{6.19}$$

□

Appendix B: Additional information for section 5

Additional notes on implementation of emp-LASSO

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

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 predictors, 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 predictor observations. We use the same exponential down-weighting scheme as for the emp-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:
 alpha = 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 predictors)/2. In this case, we use aggregated wind speed and temperatures, and dummies for weekend so the number of exogenous predictors is 4. In order to get fair comparison with the emp-LASSO, we also use exponential downweighting on the exogenous predictors, 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
```