Prediction Intervals in high-dimensional regression

Sayar Karmakar, Marek Chudy, Wei Biao Wu

University of Chicago & University of Vienna

We construct prediction interval for future sums of response variables that depend on potentially infinitely many predictors. Starting the discussion for finitely many regressors with different estimation techniques, such as least squares, quantile or robust, we move on to use LASSO for the estimation of the regression parameters for infinitely many regressor variables. We deal with a large class of stationary mean-zero noise process; specifically we allow long-range dependence, heavy-tailed behavior and non-linearity in them. The consistency properties of our intervals are shown using a sharp tail probability inequality. We also provide substantiation of our theory through an application in spot electricity price forecasting and compare it to selected existing methods such as exponential smoothing and neural networks.

Key Words and Phrases: Consistency, high-dimensional regression, prediction intervals, long-run prediction, LASSO, electricity prices

1. Introduction

Forecasting of response variables is one of the key motivations in time-series analysis since many practitioners are interested in knowing about the future. This helps them decide strategies, pricing policies etc. In this paper we consider a linear regression model and forecast time-aggregated values of the response variable. Consider the following model

$$y_i = x_i^T \beta + e_i, \quad \beta \in \mathbb{R}^p \tag{1.1}$$

For finite p, this was previously discussed by Zhou et al. (2010) for linear error process e_i. They used a robust least-absolute-deviation (LAD) estimation of the regression parameter. However, if p grows to ∞, that framework would fail. In many sectors of the economics including macro-economy, finance or energy, the response variables depend on a large number of covariates. It is also widely accepted that big data can

provide forecasting advantages over univariate approaches (see Elliott et al., 2013, Cheng et al., 2016, Ludwig et al., 2015) although some empirical studies cannot corroborate these beliefs (see Stock and Watson, 2012, Kim and Swanson, 2014) and some provide opposite evidence (Chudy and Reschenhofer). This was one of the key-motivations to extend the prediction intervals proposed by Zhou et al. (2010) to accommodate infinitely many predictors. We use Lasso to estimate the β parameter but our theory is generalizable to other possible estimations as well.

The second significant contribution we provide in this paper is to extend the class of noise process to allow non-linearity. We follow the functional dependence framework in a seminal paper by Wu (2005) to allow the following general class of error process

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

where ϵ_i are independent and identically distributed (i. i. d.) random variables. This is a much larger class than just the linear class considered in Zhou et al. (2010) and thus can capture many popular time-series models.

The empirical studies utilizing economic big-data focus mainly on forecasting over short-horizons. This is partially because many series in finance (realized volatility of returns) or macroeconomics (GNP, inflation, interest rates, population, productivity and unemployment) exhibit long range dependence. The joined effect of these long-term characteristics dominates any short-run relationship. On the other hand, it depends on nuisance parameters and the theory becomes very involved. In order to find suitable application for the high-dimensional extension of Zhou et al. (2010) we turn to electricity price forecasting (EPF). The electricity prices depend on many exogenous inputs including weather conditions, local economy and environmental policy (Knittel and Roberts, 2005, Huurman et al., 2012). Generally speaking, EPF is challenging because of complex seasonality with daily, weekly and yearly patterns, dependence on short and long-term weather conditions, day-of-week effect, macroeconomic development etc. The prices exhibit heteroscedasticity, heavy tails and sudden price spikes (see an exhaustive review in Weron, 2014). Long-horizon EPF is of major interest for power portfolio risk management, derivatives pricing, medium-and long-term contracts evaluation and maintenance scheduling. Recently, Ludwig et al. (2015) found that disaggregated data on wind speed and temperature have additional forecasting power and lead to better forecast for EPEX SPOT energy market. Their local (disaggregated) set of covariates consist of wind speed measured at 73 weather stations and of hourly temperatures measured at 78 weather stations. However, they are focused on short-term point-forecasts. In order to extend their findings, we adopt their dataset and provide PI's for hourly spot electricity prices. Our choice of predictors is based on periodicity, day of week effect and weather conditions. For visual

out-of-sample comparison we include alternative PI's obtained from methods such as exponential smoothing, neural networks and regression with auto-correlated errors[1]. The comparison shows that the Zhou et al. (2010), combined with LASSO, provide far best PI's among these competitors.

The rest of the article is organized as follows: We discuss construction of prediction interval in section 2 under different scenarios such as number of regressors, light or heavy-tailed nature etc. Section 3 discusses some existing results for the linear error processes and provides similar results for a class of non-linear processes. Section 4 discusses a sharp tail probability inequality which in turn helps us to show consistency properties for LASSO regression. We conclude the paper through substantiation sought through simulation and real data analysis in Section 5. In particular, we construct prediction intervals for hourly spot electricity prices and compare them with existing methods.

2. Construction of prediction intervals

In this section we first focus on two methods discussed by Zhou et al. (2010) for the construction of the prediction interval in the special case of $\beta = 0$, i.e. absence of any covariates. The case for the regressors then naturally follows by constructing the PI for the residuals and adding back the estimated effect. We enlarge the scope of regression from just the least-absolute-deviation (LAD) to a general class of robust regression. For the case of infinitely many regressors, we use LASSO residuals.

2.1. Without covariates

2.1.1. CLT based

Method 1: (CLT based) If the process e_i show short-range dependency or light-tailed behavior, one can use the long-run covariance to construct the PI

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

where σ is the long run covariance. However, since σ is unknown, one can use a lag-window estimate

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

[1]The respective PI's can easily be computed with automatic forecasting R-package "forecast" (see Hyndman and Khandakar, 2008)

and use the following version of the $100(1-\alpha)\%$ PI

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

2.1.2. Quantile based

Method 2: (Quantile based) This method does not necessarily require short-range of light-tailed behavior and has more general applicability. Construct for $i = m, \ldots, n$,

$$\tilde{Y}_i = \frac{\sum_{j=i-m+1}^{i} e_i}{H_m}.$$

Then we have the following $100(1-\alpha)\%$ PI

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

where Q_u is the u-th empirical quantiles of $\tilde{Y}_i, i = m, \ldots, n$.

2.1.3. CLT for linear process

We first propose this method for the special case $x_i = 0$ as a premier. Also this is applicable only for linear processes with light-tailed innovations and short-range dependence. Assume $E(|e_i|^p) < \infty$ for some $p > 2$. Wu and Woodroofe (2004) proved that, if for some $q > 5/2$,

$$\|E(S_m|\mathcal{F}_0)\| = O\left(\frac{\sqrt{m}}{\log^q m}\right), \tag{2.1}$$

then we have the a.s. convergence

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

where $m \to \infty$ and $\sigma^2 = \lim_{m\to\infty} \|S_m\|^2/m$ is the long-run covariance. For the linear case, evaluating (2.1) is particularly easy. Assume

$$e_i = \sum_{j=0}^{\infty} a_j \epsilon_{i-j}. \tag{2.2}$$

$$\|E(S_m|\mathcal{F}_0)\|^2 = \|(a_1 + \ldots + a_m)\epsilon_0 + (a_2 + \ldots + a_m)\epsilon_{-1} + \ldots\|^2 = \sum_{i=1}^{m} b_i^2, \tag{2.3}$$

where $b_i = a_i + \ldots + a_m$. We assume the special formulation of a_i

$$a_i = i^{-\chi}(\log i)^{-A}, \quad \chi > 1, A > 0, \tag{2.4}$$

as $\chi > 1$ ensures short-range dependence in this case. Note that, $\sum_{i=1}^{m} b_i^2$ assumes the following value depending on $\chi > 3/2$ or not.

$$\sum_{i=1}^{m} b_i^2 = \begin{cases} m^{3-2\chi}(\log m)^{-2A}, & \text{for } 3-2\chi > 0 \\ O(1) & \text{for } 3-2\chi \leq 0. \end{cases} \tag{2.5}$$

Thus, (2.1) holds, for the following a_j in the short-range dependent linear case.

$$a_j = \begin{cases} j^{-\chi}(\log j)^{-A}, & \text{for } 1 < \chi < 3/2, A > 5/2 \\ j^{-\chi}(\log j)^{-A} & \text{for } \chi \geq 3/2, A > 0. \end{cases} \tag{2.6}$$

.
However it remains to estimate the long-run covariance σ. We can use a lag window estimate

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

where $\bar{e} = n^{-1} \sum_{i=1} e_i$. If m is large, then $S_m/(\sigma\sqrt{m})$ can be approximated by t_{df} with appropriate df needed to estimate σ. Thus we can use L and U as $\hat{\sigma} t_{df,\alpha/2}\sqrt{m}$ for the quantile of $S_{n+m} - S_{n+1} = \sum_{i=n+1}^{n+m} e_i$.

2.1.4. Advantages and drawbacks of CLT based method

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

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

2.1.5. Data-based adjustment

The two methods for constructing prediction intervals provide nice theoretical results such as consistency and Bahadur representations. See our results in the following subsection or those from Zhou et al. (2010). However, in real life, one has to see whether such properties translate also into practice using real-data evaluation. Chudy et al. shows that bootstrapping and kernel quantile estimation can significantly enhance the forecasting performance for the univariate case. Their results show that as the forecasting horizon grows, *Method 1* tends to lose the coverage probability. Employing the data-driven adjustments based on replication of the series using stationary bootstrap (Politis and Romano, 1994) and kernel quantile estimator (Falk, 1984) instead of the sample version suggested by Zhou et al. (2010) increases the coverage of *Method 1*. In particular, when $m/n \approx 1/2$ the increase of coverage by 20 percent points can be achieved. In our current application to spot electricity prices, the horizon reaches $m \approx 1/3n$ which is close to the latter scenario. Therefore, we apply these adjustments (see section 5 for details on implementation).

2.2. With covariates

2.2.1. Linear regression

Assume the following model

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

where β is a p-dimensional parameter vector with $p \gg n$. We use LASSO to find the estimates of β

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{R}^p} \frac{1}{n}(y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.7}$$

where λ is the penalty parameter. After obtaining $\hat{\beta}$, we obtain the residuals $\hat{e}_i = y_i - x_i^T \hat{\beta}$. The PI for y_i will be

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

Thus it suffices to discuss the construction of the PI for $\sum_{i=n+1}^{n+m} \hat{e}_i$ after observing $(y_i, x_i)_{i=1,\dots,n}$. Since one of the major part of this paper also focuses on the scenario

without the covariates we keep the discussion concise by 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 3.3 shows the consistency properties for the case with the covariates.

2.2.2. Robust regression

2.3. Quantile Based Method

Before we discuss consistency properties for this method, we would like to describe precisely the dependence structure and our assumptions. Our theory combines all possible combinations of the following of the error process.

- Short-range or long-range dependent process
- Linear or non-linear process
- Light-tailed or heavy tailed innovations for the linear processes.

2.3.1. Description of the error process

In this subsection we provide a systematic description of the error process and the dependence structure we impose. For linear error process e_i, we assume the following decomposition

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

where the i.i.d. innovations ϵ_j is allowed to have both light-tail i.e. $E(|\epsilon_i|^2) < \infty$ or heavy-tailed i.e. $\alpha = \sup_{t>0}\{t : E(|\epsilon_i|^t) < \infty\} < 2$.

$$\begin{aligned} \text{(SRD)} \quad &: \quad \sum_{j=0}^{\infty} |a_j| < \infty \\ \text{(DEN)} \quad &: \quad \sup_{x \in \mathbb{R}} f_\epsilon(x) + |f'_\epsilon(x)| < \infty \\ \text{(LRD)} \quad &: \quad a_j = j^{-\gamma} l_j \text{ where } \gamma < 1, l_j \text{ is slowly varying function} \end{aligned} \tag{2.9}$$

2.4. Quantile estimation consistency for linear process

Let $\hat{Q}_n(\alpha/2)$ and $\hat{Q}_n(100(1-\alpha/2))$ be the $\alpha/2$ and $(100(1-\alpha/2))$ th quantiles for

$$\tilde{Y}_i = \frac{\sum_{j=i-m+1}^{i} e_j}{H_m}, \quad i = m, m+1, \ldots \tag{2.10}$$

where H_m is a suitable normalizing constant. We have $H_m = m^{1/2}$ and $H_m = n^{1/\alpha} L_1(n)$ for light-tailed and heavy-tailed distributions where $\alpha = \max_t\{t : E(|\epsilon_i|^t) < \infty\}$ and $L_1(n)$ is a s. v. f. Also, Let $Q_n(q)$ denote the true quantiles of $\tilde{Y}_i$.

Theorem 2.1 (Quantile consistency for linear processes)**.** *We have the following different rates of convergence of quantiles based on the different cases:*

- *Linear light tailed SRD: Suppose DEN holds and* $E(\epsilon_j^2) < \infty$*. If SRD holds and* $m^3/n \to 0$*, then for any fixed* $0 < q < 1$*,* $|Q_n(q) - \tilde{Q}_q| = O_p(m/\sqrt{n})$*.*
- *Linear light tailed LRD: LRD holds with* γ *and* $l(n)$ *in (**??**) and* $m^{5/2-\gamma} n^{1/2-\gamma} l^2(n) \to 0$*, then, for any fixed* $0 < q < 1$*,* $|Q_n(q) - \tilde{Q}_q| = O_p(mn^{1/2-\gamma}|l(n)|)$*.*
- *Linear heavy-tailed SRD: Suppose DEN holds and* $E(\epsilon_j^\alpha) < \infty$ *for some* $1 < \alpha < 2$*. If SRD holds and* $m = O(n^k)$ *for some* $k < (\alpha - 1)/(\alpha + 1)$*. Then for any fixed* $0 < q < 1$*,* $|Q_n(q) - \tilde{Q}_q| = O_p(mn^\nu)$ *for all* $\nu > 1/\alpha - 1$*.*
- *Linear heavy-tailed LRD: If LRD holds with* γ *and* $l(n)$ *in (**??**) and* $m = O(n^k)$ *for some* $k < (\alpha\gamma - 1)/(2\alpha + 1 - \alpha\gamma)$*, then, for any fixed* $0 < q < 1$*,* $|Q_n(q) - \tilde{Q}_q| = O_p(mn^\nu)$ *for all* $\nu > 1/\alpha - \gamma$*.*

3. Finitely many regressors

3.1. Linear case

Old results

3.1.1. Assumptions

3.1.2. Results

3.2. Non-linear case

New contribution

3.2.1. CLT for non-linear processes

3.2.2. 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

$$e_i = H(\mathcal{F}_i) = H(\epsilon_i, \epsilon_{i-1}, \ldots), \tag{3.1}$$

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 (2.2). We need to define the dependence between (e_i) process. Define the following functional dependence measure

$$\delta_{j,p} = \sup_i \|e_i - e_{i,(i-j)}\|_p = \sup_i \|H_i(\mathcal{F}_i) - H_i(\mathcal{F}_{i,(i-j)}\|_p, \tag{3.2}$$

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,

$$\mathcal{F}_{i,k} = (\epsilon_i, \epsilon_{i-1}, \ldots, \epsilon'_k, \epsilon_{k-1}, \ldots) \tag{3.3}$$

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

$$\Theta_{j,p} = \sum_{i=j}^{\infty} \delta_{i,p}, \tag{3.4}$$

which can be thought as cumulative dependence of $(X_j)_{j\geq k}$ on ϵ_k. For the quenched CLT in (2.1), we assume the following rate for $\Theta_{j,p}$.

$$\Theta_{j,p} = j^{-\chi}(\log j)^{-A} \text{ where } = \begin{cases} A > 0 \text{ for } 1 < \chi < 3/2, \\ A > 5/2 \text{ for } \chi \geq 3/2, \end{cases} \tag{3.5}$$

.

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 $\tilde{S}_m = \sum_{i=1}^m \tilde{X}_i = \sum_{i=1}^m E(X_i|\epsilon_i, \ldots \epsilon_{i-m})$. The proof of (2.1) follows along the line of (2.3) from the facts $\|P_j(\tilde{X}_i)\|_2 \leq \delta_{i-j,2}$,

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

3.2.3. 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

$$e_i = H(\epsilon_i, \epsilon_{i-1}, \ldots), \tag{3.6}$$

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

$$\mathcal{F}_k' = (\ldots, \epsilon_{-1}, \epsilon_0', \epsilon_1, \ldots, \epsilon_k),$$

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

$$\psi_{k,q} = \sup_{t\in[0,1]} \sup_{u\in\mathbb{R}} \|f_1(u, t|\mathcal{F}_k) - f_1(u, t|\mathcal{F}_k'\|_q. \tag{3.7}$$

Quantity (3.7) 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

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

3.2.4. 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 (**??**).

Theorem 3.1. *We have the following different rates of convergence of quantiles based on the different cases:*

- *Suppose DEN holds and* $E(\epsilon_j^2) < \infty$. *If SRD holds and* $m^3/n \to 0$, *then for any fixed* $0 < q < 1$, $|Q_n(q) - \tilde{Q}_q| = O_p(m/\sqrt{n})$.
- *If LRD holds with* γ *and* $l(n)$ *in (**??**) and* $m^{5/2-\gamma}n^{1/2-\gamma}l^2(n) \to 0$, *then, for any fixed* $0 < q < 1$, $|Q_n(q) - \tilde{Q}_q| = O_p(mn^{1/2-\gamma}|l(n)|)$.
- *Suppose DEN holds and* $E(\epsilon_j^\alpha) < \infty$ *for some* $1 < \alpha < 2$. *If SRD holds and* $m = O(n^k)$ *for some* $k < (\alpha - 1)/(\alpha + 1)$. *Then for any fixed* $0 < q < 1$, $|Q_n(q) - \tilde{Q}_q| = O_p(mn^\nu)$ *for all* $\nu > 1/\alpha - 1$.
- *If LRD holds with* γ *and* $l(n)$ *in (**??**) and* $m = O(n^k)$ *for some* $k < (\alpha\gamma - 1)/(2\alpha + 1 - \alpha\gamma)$, *then, for any fixed* $0 < q < 1$, $|Q_n(q) - \tilde{Q}_q| = O_p(mn^\nu)$ *for all* $\nu > 1/\alpha - \gamma$.

3.3. Quantile estimation consistency in presence of covariates

Our main result for this section will be Theorem 3.3 and it will say that the error bounds obtained in Theorem 2.1, and 3.1 remains intact if we make a proper choice of the sparsity condition. Before that, we state a crucial lemma from Bickel, Ritov and Tsybakov (2009 , Bickel et al. (2009)).

Lemma 3.2. *Let* $\lambda = 2r$ *in (2.7). Also assume,*

$$r = \max(A(n^{-1}\log p)^{1/2}\|e_\cdot\|_{2,\alpha}, B\|e_\cdot\|_{q,\alpha}|X|_q/n) \tag{3.8}$$

On the event

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

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 4.1 and 4.2 to V_j.

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

Theorem 3.3. *Assume s, the number of non-zero coordinates in β satisfies the following*

$$s \tag{3.9}$$

Then the conclusions in Theorem 2.1, ??, 3.1 and ?? hold with $Q_n(q)$ replaced by $\bar{Q}_n(q)$.

4. Infinitely many regressors

4.1. 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 4.1 (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*

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

- *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*,*

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

Theorem 4.2 (Nagaev inequality for non-linear processes)**.** *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*,*

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

4.2. Consistency of Prediction Interval

Sparsity condition needed

5. Prediction intervals for European Power Exchange spot electricity prices: regression approach with many covariates

5.1. Simulation results

5.2. Real data analysis

In this section, we conduct a graphical POOS comparison of following methods:

(zxw) PI's of Zhou et al. (2010),
(mw) PI's of Müller and Watson (2016),
(armax) PI's implied by ARMAX type models.
(ets) PI's implied by Exponential smoothing state space model (Hyndman et al., 2008),
(nnar) PI's implied by Neural network auto-regression (Venables and Ripley, 2002, Hyndman and Athanasopoulos, 2013, pp. 279).

Data We forecast hourly day-ahead spot electricity prices for Germany and Austria - the largest market at the European Power Exchange (EPEX SPOT). The prices arise from a 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[2] 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 is $m = 1, 2, \ldots, 17$ weeks $(168, 336, \ldots, 2856$ hours).

Inspection of the periodogram for the prices (see Figure 1C) reveals peaks at periods 1 week, 1 day and 1/2 day. Such seasonality is far too complex to be handled by seasonal differencing, i.e., SARIMA or ETS models, which are rather designed for monthly and quarterly data or by dummy variables. Instead, it is convenient to use sums of sinusoids $g_t^k = R\sin(\omega_k t + \phi) = \alpha_k(R, \phi)\sin(\omega_k t) + \beta_k(R, \phi)\cos(\omega_k t)$ 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 α_k, β_k are estimated from the prices by (penalized and weighted) least squares. In addition, we use 2 dummy predictors as indicators for weekend.

As mentioned in Section 1, we use local weather condition as predictors too. Local weather is represented by 151 hourly wind speed and temperature series observed over period of 5 years (2009-2013) including the training period (see above). In order to approximate the missing in-sample data and unobserved values during evaluation period, we take hourly-specific-averages[3] 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

$$X_t = (d_{\text{sa}}, d_{\text{su}}, \sin(\omega_1 t), \cos(\omega_1 t), \ldots, , \sin(\omega_{84} t), \cos(\omega_{84} t), w_{1,t}, \ldots, w_{73,t}, \tau_{1,t}, \ldots, \tau_{78,t)}, \tag{5.1}$$

for $t = 1, \ldots, T$, 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 electricity price level during December 2013. Although the prices eventually raise during January 2014 the forecasts based on the whole training period are likely to have a bias. However, using only the post-break December data would decrease their efficiency. 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). We use standardized exponential weights $v_{T-t+1} = \alpha^{t-1}((1-\alpha))/(1-\alpha^t)$, $t = 1, \ldots, T$, with

[2]Coordinated Universal Time.

[3]An alternative bootstrap approximation of unknown future observations was proposed by Hyndman and Fan (2010).

$\alpha = 0.8$. The implementation of the competing methods follows these steps:
Empirical quantile method modified for case of many predictors (emp-LASSO):

1. Use LASSO (Tibshirani, 1996, Friedman et al., 2010) e.g. from R-package glmnet to estimate coefficients of regression model $y_t \sim X_t$, $t = 1 \ldots, T$ (see supplementary appendix for further details).
2. Extract the regression residuals $\hat{e}_t = y_t - \hat{y}_t$, $t = 1, \ldots, T$ and apply the steps 1,2 and 3 from the implementation steps for *emp* to these residuals.
3. The PI is $[\bar{\hat{y}}_{T,1:m} + \hat{Q}((1-\alpha)/2), \bar{\hat{y}}_{T,1:m} + \hat{Q}((1+\alpha)/2)]$, where $\bar{\hat{y}}_{T,1:m}$ is the average of h-step-ahead forecasts for $h = 1, \ldots, m$. These forecasts are obtained from the estimated model using computed and approximated future values of the predictors.

The implementation of *MN* follows steps described in Section **??**. We use AIC to select the appropriate orders for *armax, ets and nnar* (see appendix for further details). The computation of implied PI's follows these steps:

1. Adjust spot price series y_t for weekly periodicity using seasonal and trend decomposition suggested by Cleveland et al. (1990).
2. Fit the models to seasonally adjusted spot prices y_t using aggregated weather data and dummies as exogenous predictors (see details in the supplementary appendix).
3. Simulate $b = 1, \ldots, B$ future paths $\hat{y}^b_{T,t}$, of length m from the estimated model.
4. Obtain bootstrap PI's as sample quantiles from set of averages $\bar{\hat{y}}^b_{T,T+1:T+m}$, $b = 1, \ldots, B$.

POOS results Before we compare the *emp-LASSO* with competitors, we explore the benefits from incorporating disaggregated weather data. To do this, we compute the PI's using (i) no regressors in Figure 2IA, (ii) using only deterministic regressors in Figure 2IB, (iii) using both deterministic regressors and aggregated weather defined as $\bar{w}_t = \sum_{k=1}^{73} w_{k,t}$, $\bar{\tau}_t = \sum_{l=1}^{78} \tau_{l,t}$ and in Figure 2IC and finally, (iv) using all 321 predictors in Figure 2ID. We see only little difference between the first three plots as all three PI's seem to be upward-biased. Clearly, a striking improvement is achieved by including disaggregated weather series.

From the four other methods only *mw* and *ets* provide useful PI's. Figure 2IIA shows that *MN* works well over the whole 17-weeks-long evaluation period. However, when compared to *emp-LASSO*, we see a that the latter gives advantage in terms of precision. *ets* becomes too wide as the horizon grows. The *nnar* is clearly biased for large m. Surprisingly, the *armax*, which performs exponential smoothing by definition, perform worst of all. It might be that the down-weighting is simply

too mild.

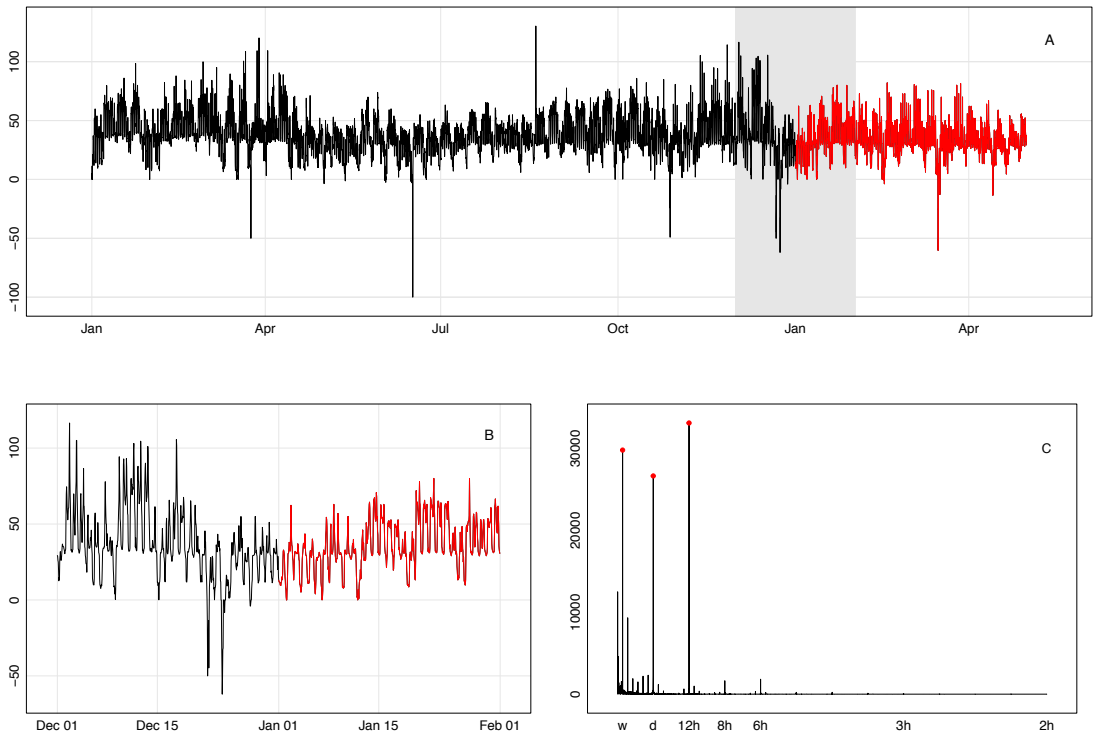
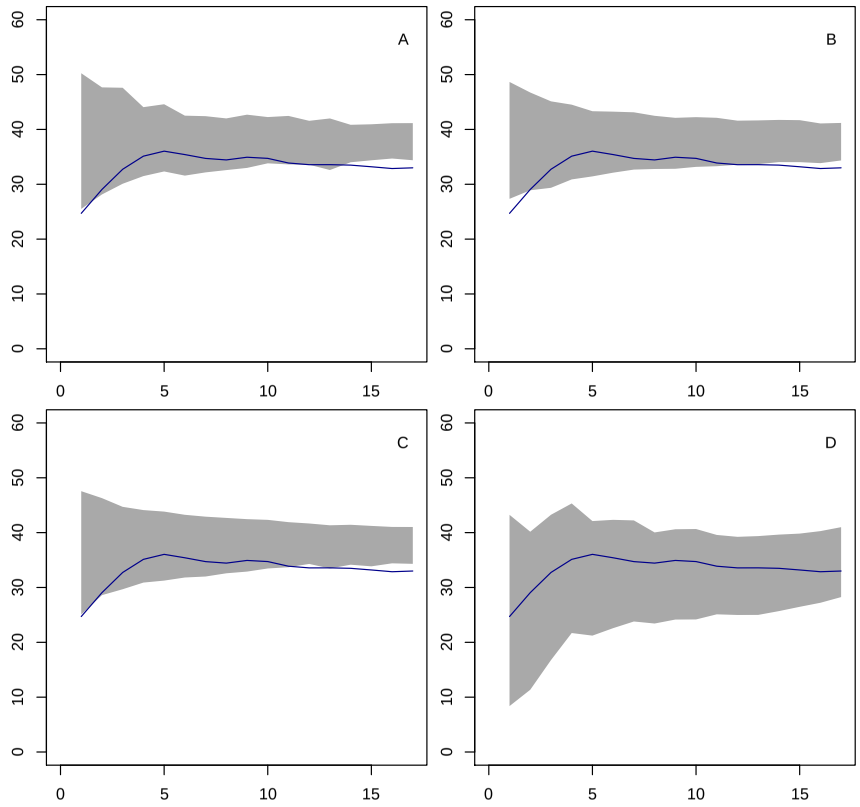
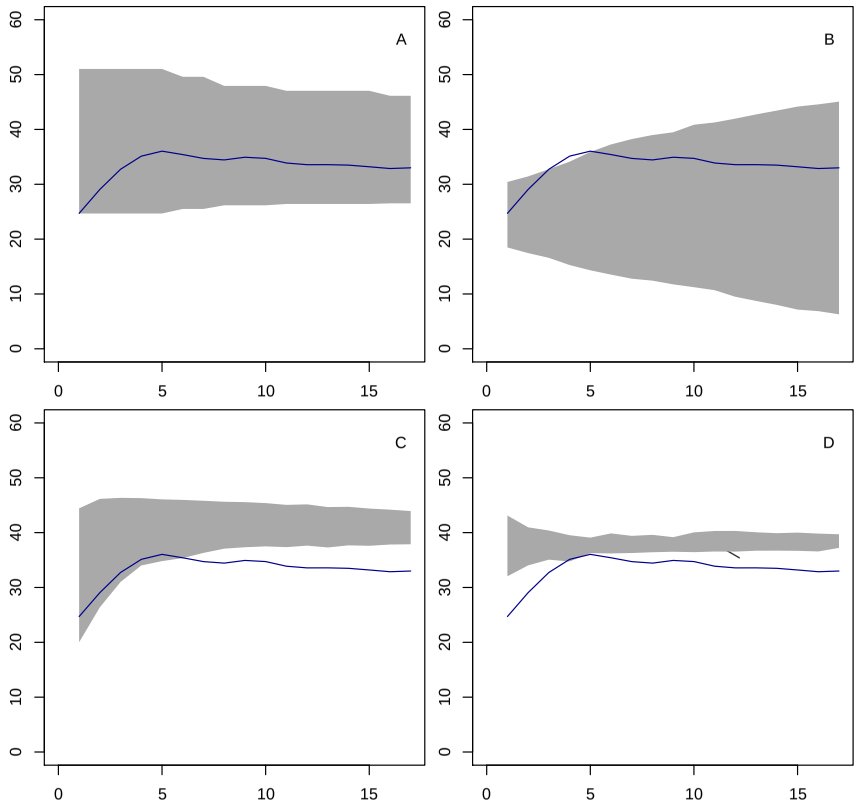


Fig 1: Electricity spot prices, A) Full sample, B) Drop in price level, C) Periodogram with peaks at periods 1 week, 1 day and 12 hours. According to European convention, the term spot refers to day-ahead rather than real-time unlike in the US, where term forward is common.



(I) A) emp (without predictors), B) emp-LASSO with 170 deterministic predictors, C) emp-LASSO with 170 deterministic predictors & 2 aggregated weather time series, D) emp-LASSO with 170 deterministic predictors & 151 disaggregated weather series.



(II) A) MN, B) ETS (A,N,N) with tuning parameter 0.0446, C) NNAR (38,22) with one hidden layer and with weekend dummies & aggregated weather as predictors, D) ARMA with weekend dummies & aggregated weather as exogenous predictors.

Fig 2: PI's (gray) for average spot electricity prices (blue) over forecasting horizon $m = 1, \ldots, 17$ weeks.

6. Discussion

We have considered problem of 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 by utilizing the LASSO estimator.

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 of conventional exponential smoothing and neural network approach 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, 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).

References

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. ISSN 0090-5364. . URL http://dx.doi.org/10.1214/08-AOS620.

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

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

X. Cheng, Z. Liao, and F. Schorfheide. Shrinkage estimation of high-dimensional factor models with structural instabilities. *The Review of Economic Studies*, 83 (4):1511–1543, 2016. . URL +http://dx.doi.org/10.1093/restud/rdw005.

M. Chudy and E. Reschenhofer. Forecasting with dynamic factor models. URL http://homepage.univie.ac.at/marek.chudy/Research/forecasting_with_dynamic_factor_models.pdf.

M. Chudy, S. Karmakar, and W. Wu. Long-term prediction intervals for economic time series. URL http://homepage.univie.ac.at/marek.chudy/Research/Long_term_PI_s_for_economic_time_series.pdf.

R. B. Cleveland, W. S. Cleveland, M. J. E., and I. Terpenning. Stl: A seasonal-trend

decomposition procedure based on loess. *Journal of Official Statistics*, 6:3–73, 1990.

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

M. Falk. Relative deficiency of kernel type estimators of quantiles. *Ann. Statist.*, 12 (1):261–268, 1984. . URL https://doi.org/10.1214/aos/1176346405.

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

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

R. J. Hyndman and G. Athanasopoulos. Forecasting: principles and practice, 2013.

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

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

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

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

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

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

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

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

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

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

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

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

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:1, 1996.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag New York, New York, 2002.

R. Weron. Electricity price forecasting: A review of the state-of. *International Journal of Forecasting*, 30:4, 2014.

R. Weron and A. Misiorek. Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *Accessed*, 9(2):2017, 2008. URL https://mpra.ub.uni-muenchen.de/10428.

W. B. Wu. Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154 (electronic), 2005. ISSN 1091-6490. . URL http://dx.doi.org/10.1073/pnas.0506715102.

W. B. Wu and M. Woodroofe. Martingale approximations for sums of stationary processes. *Ann. Probab.*, 32(2):1674–1690, 2004. ISSN 0091-1798. . URL http://dx.doi.org/10.1214/009117904000000351.

Z. Zhou, Z. Xu, and W. B. Wu. Long-term prediction intervals of time series. *IEEE Trans. Inform. Theory*, 56(3):1436–1446, 2010. ISSN 0018-9448. . URL http://dx.doi.org/10.1109/TIT.2009.2039158.

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 4.1* **??**, *4.2 and* **??**,

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

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

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

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*

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

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 $k \geq 1$,

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

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

$$\begin{aligned}\|R_n(x+u)\| &\le \sum_{k=1}^{\infty}\|\sum_{i=m}^{n}\mathcal{P}_{i-k}f_\epsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\|\\ &\le c_1\sqrt{n}\sum_{k=1}^{\infty}|\tilde{b}_k|\le c_1m\sqrt{n}\sum_{j=0}^{\infty}|a_j|.\end{aligned}$$

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* $\rho\in(1/\gamma,\alpha)$

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

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 $C_\rho=[18\rho^{3/2}(\rho-1)^{-1/2}]^\rho$.

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

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

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

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

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

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

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

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

Proof. .OF THEOREM 4.1 By central limit theorem of **?**, 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 4.1.

□

Proof of Theorem 3.3. From Lemma 3.2, 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 PI'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=0.9

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
```