High-Dimensional Linear Regression for Dependent Data with Applications to Nowcasting

Booth School of Business, University of Chicago
(Joint with Yuefeng Han)

IMA, April 23, 2018

Outline

1. Forecasting in a data-rich environment
2. What is nowcasting? Why is it useful?
3. LASSO regressions for dependent data. How?
4. Some examples
5. Properties of LASSO estimators for dependent data

What is nowcasting?

Using high-frequency data to update predictions of a low-frequency variable of interest
Consider the prediction of U.S. quarterly GDP growth rate.

1. Traditional approach: Using quarterly data with either time series or macro-economic models
2. Many new economic data become available once the quarter starts, e.g. monthly unemployment rate, imports and exports, PMI index.
3. How to take advantages of the newly available information?

Mixed-frequency data sampling (MIDAS) and others

A simple illustration

Monthly data do contain useful information.

Consider the quarterly U.S. GDP growth rates from 1987.II to 2016.III.

1. GDP series alone has 117 observations.
   One can build a time series model for prediction, e.g. AR models.
2. Many monthly macroeconomic variables become available:
   I used 13, including unemployment rate, total non-farm payrolls, capacity ultilization of manufacturing, all employments of manufacturing sector, etc.

Linear regression

1. y_t: quarterly GDP growth rate
2. $\boldsymbol{x}_t$: one month into the quarter monthly data
3. Model 1:
$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \epsilon_t$$
4. Let r_t be the residuals of an AR(2) model of the GDP series
5. Model 2:
$$r_t = \boldsymbol{x}_t'\boldsymbol{\gamma} + \epsilon_t.$$

May use data of two months into a quarter.

Empirical results

1. Model 1(y_t): R^2 = 52.69%; Adj-R^2 = 46.72%
2. Model 2(r_t): R^2 = 34.24%; Adj-R^2 = 25.94%

If the 2nd monthly data are included,

1. Model 1(y_t): R^2 = 64.40%; Adj-R^2 = 54.12%
2. Model 2(r_t): R^2 = 52.23%; Adj-R^2 = 38.43%

Monthly data can be used to update the prediction

Why stop at 13 monthly variables?

1. Many monthly data available
2. Weekly data are also available, e.g. initial jobless claims, etc.
3. Even daily data are available, e.g. interest rates, financial indexes, etc.

In short, we have many variables available. The number may exceed the data points available.

How to handle them? How to identify the helpful ones in predicting GDP?
$\Rightarrow$ big data environment. Data are serially dependent.

Lasso may fail for dependent data

1. Data generating process: scalar Gaussian autoregressive, AR(3), model

$$x_t = 1.9x_{t-1} - 0.8x_{t-2} - 0.1x_{t-3} + a_t, \quad a_t \sim N(0,1).$$

   Generate 2000 observations. See Figure 1.
2. Big data setup
   - Dependent x_t: $t = 11, \ldots, 2000$
   - Regressors: $X_t = [x_{t-1}, x_{t-2}, \ldots, x_{t-10}, z_{1t}, \ldots, z_{10,t}]$, where z_{it} are iid $N(0,1)$.
   - Dimension = 20, sample size 1990.
3. Run the Lasso regression via the **lars** package of R. See Figure 2 for results. Lag 3, x_{t-3} was not selected.

Lasso fails in the presence of strong serial dependence.

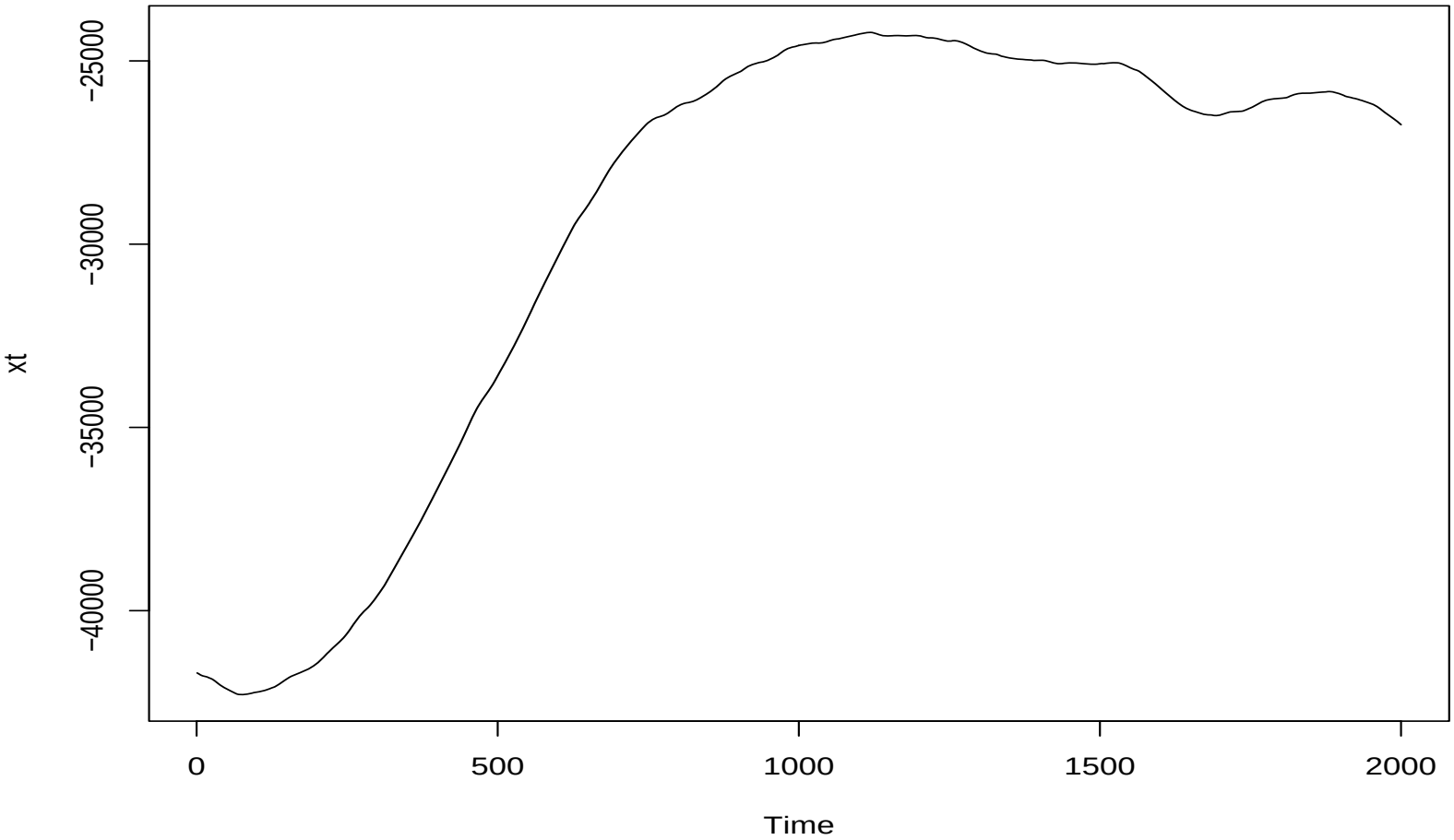


Figure: Time plot of simulated AR(3) time series with 2000 observations

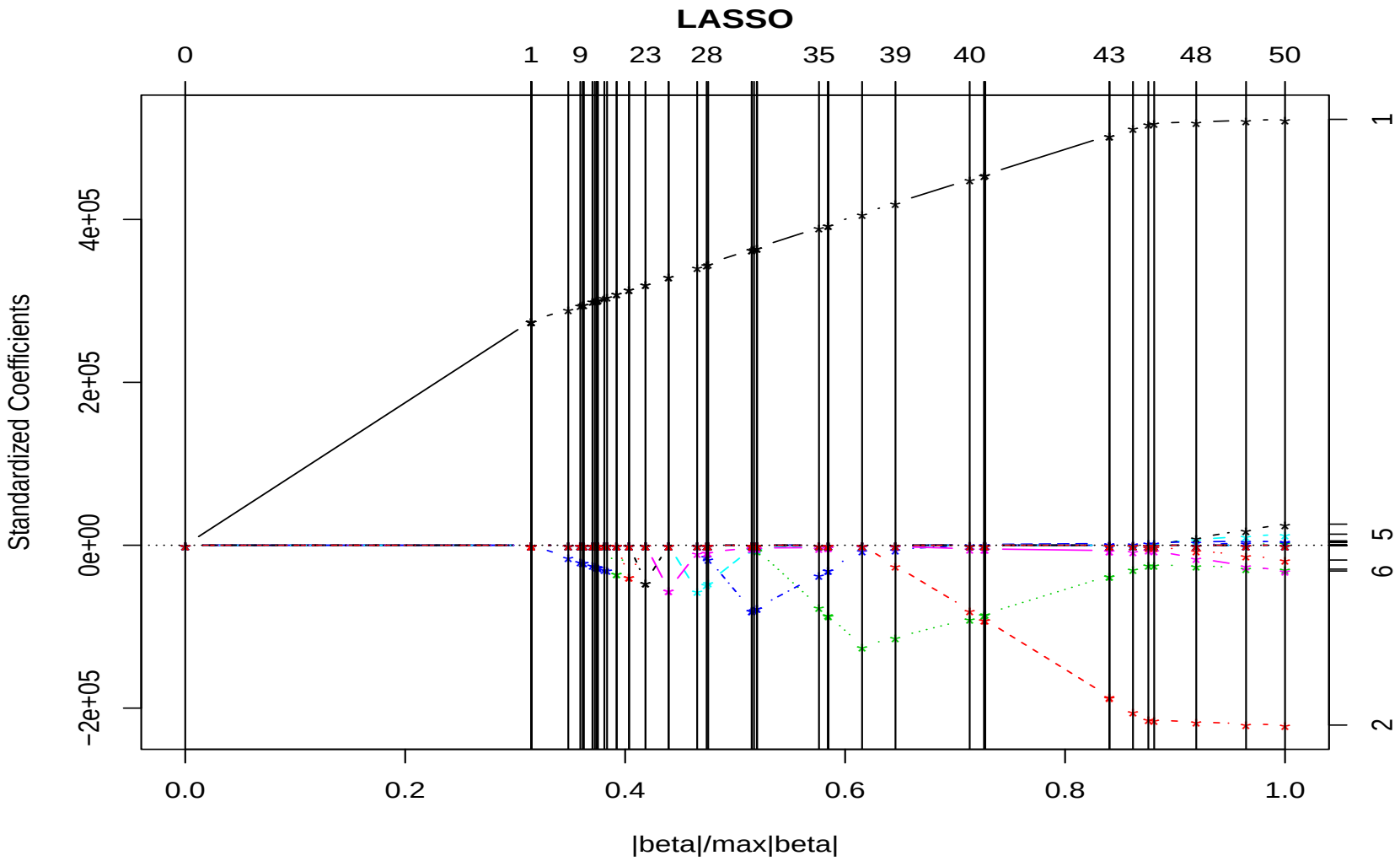


Figure: Results of Lasso regression for the AR(3) series

OLS works fine for AR models

Run the linear regression using the first three variables of X_t.

- Fitted model

$$x_t = 1.902x_{t-1} - 0.807x_{t-2} - 0.095x_{t-3} + \epsilon_t, \quad \sigma_\epsilon = 1.01.$$

- All estimates are statistically significant with p-value less than 2.22×10^{-5}.
- The residuals are well behaved, e.g. $Q(10) = 12.23$ with p-value 0.20 (after adjusting the df).

Simple time series method works for dependent data. (No surprise!)

Why does lasso fail?

Two possibilities:

1. Scaling effect: Lasso standardizes each variable in X_t. For unit-root non-stationary time series, standardization might wash out the dependence in the stationary part
2. Multicollinearity: Unit-root time series have strong serial correlations. [ACF approach 1 for all lags.]

This artificial example highlights the difference between independent and dependent data.

Need to develop methods for dependent big data! Or study conditions under which the traditional methods continue to hold.

Possible solutions

1. Re-parameterization using time series properties
2. Use different penalties for different parameters

The first approach is easier.
For the particular time series, we can define $\Delta x_t = (1 - B)x_t$ and $\Delta^2 x_t = (1 - B)^2 x_t$. Then,

$$\begin{aligned} x_t &= 1.9x_{t-1} - 0.8x_{t-2} - 0.1x_{t-3} + a_t \\ &= x_{t-1} + \Delta x_{t-1} - 0.1\Delta^2 x_{t-1} + a_t \\ &= \text{double} + \text{single} + \text{stationary} + a_t. \end{aligned}$$

The coefficients of $x_{t-1}, \Delta x_{t-1}, \Delta^2 x_{t-1}$ are 1, 1, an -0.1, respectively.

Different frameworks for LASSO

The X-matrix of conventional LASSO consists of

$$(x_{t-1}, x_{t-2}, \ldots, x_{t-10}, z_{1t}, \ldots, z_{10,t}),$$

where z_{it} are iid $N(0,1)$.
Under the re-parameterization, the X-matrix becomes

$$(x_{t-1}, \Delta x_{t-1}, \Delta^2 x_{t-1}, \ldots, \Delta^2 x_{t-8}, z_{1t}, \ldots, z_{10,t}).$$

These two X-matrices provide theoretically the same information. However, the first one has high multicollinearity, but the 2nd one does not, especially after standardization.

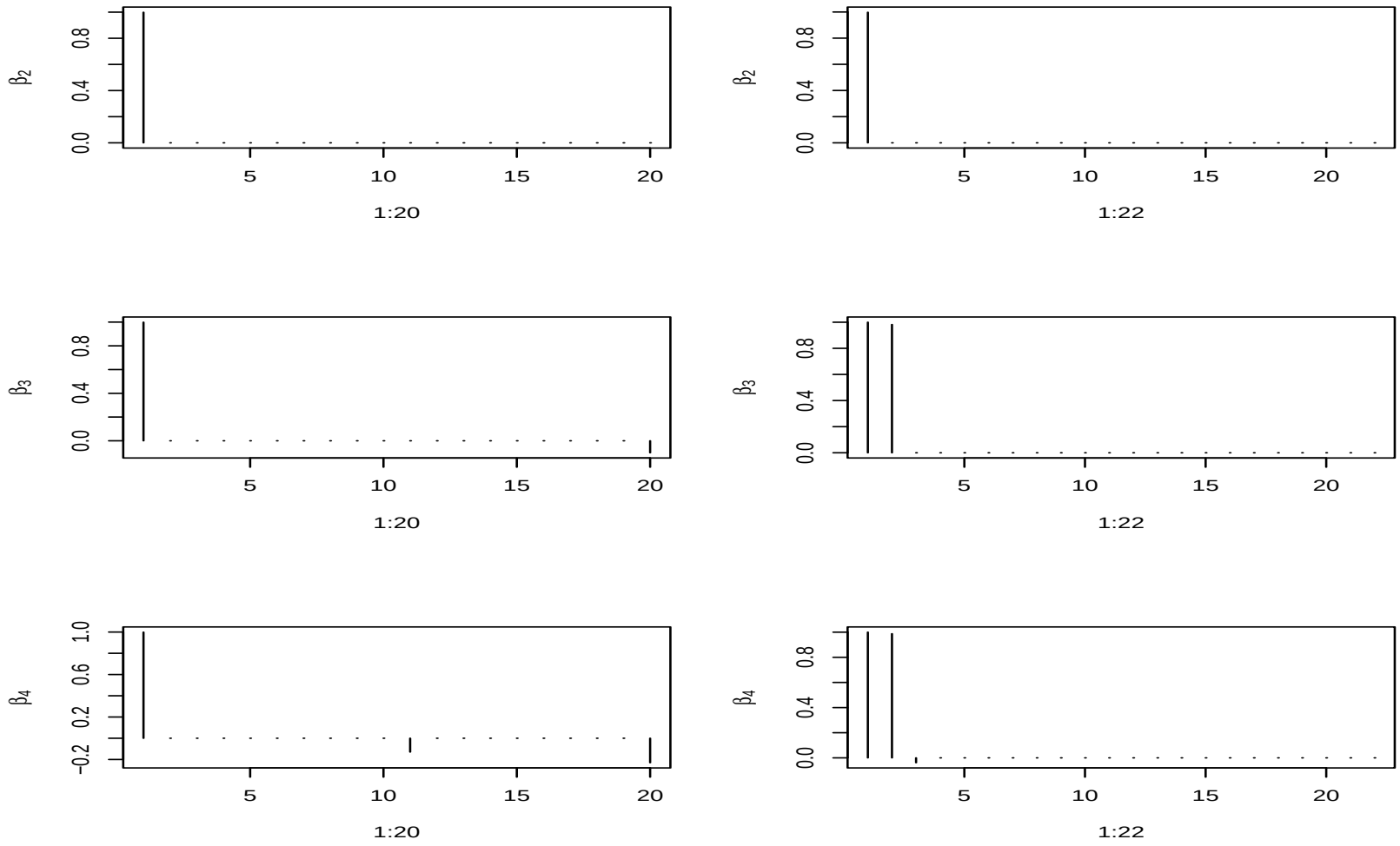


Figure: Comparison of β-estimates of **lars** results

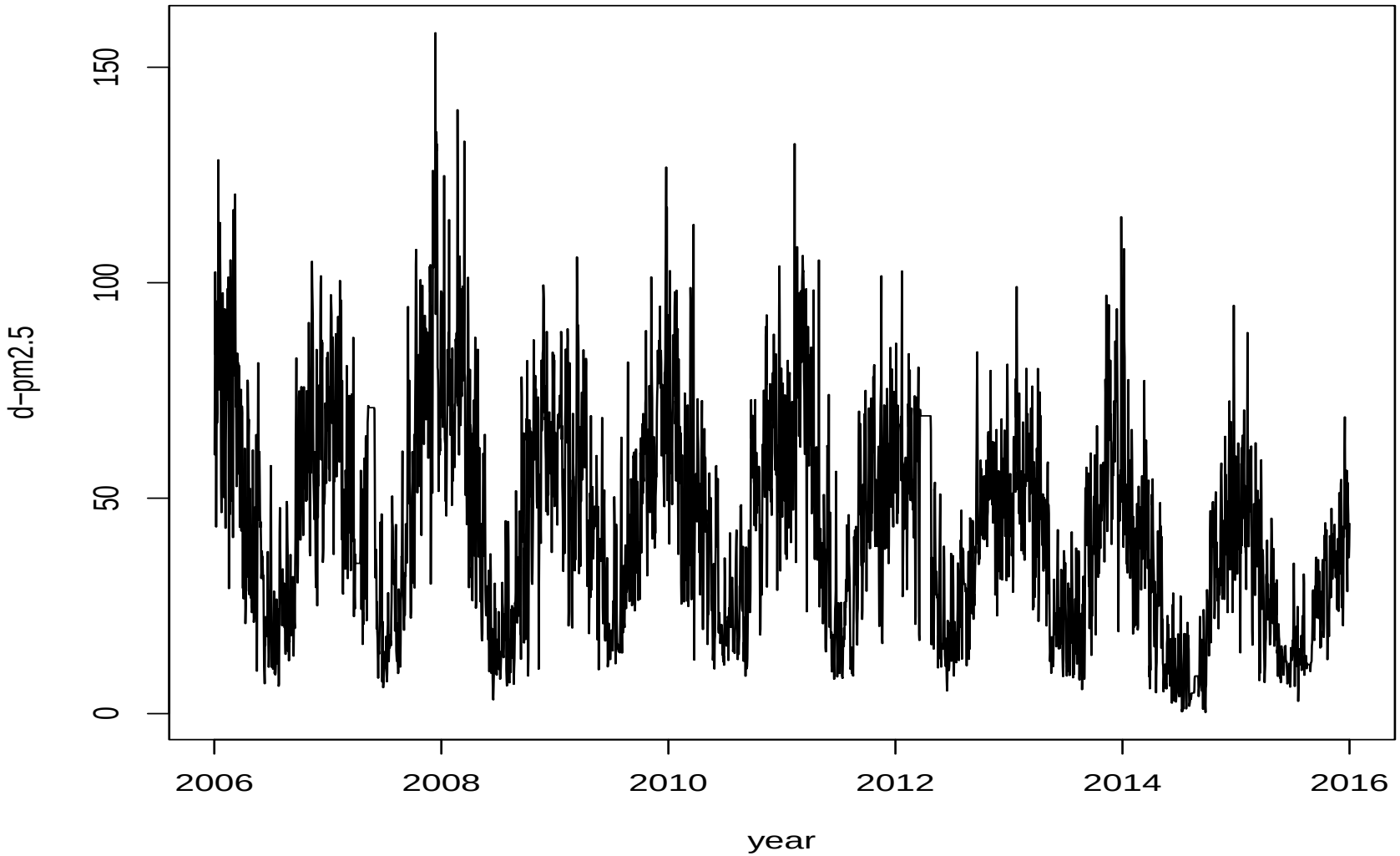
Theoretical justification

Focus on the particular series x_t used. Some properties of the series are

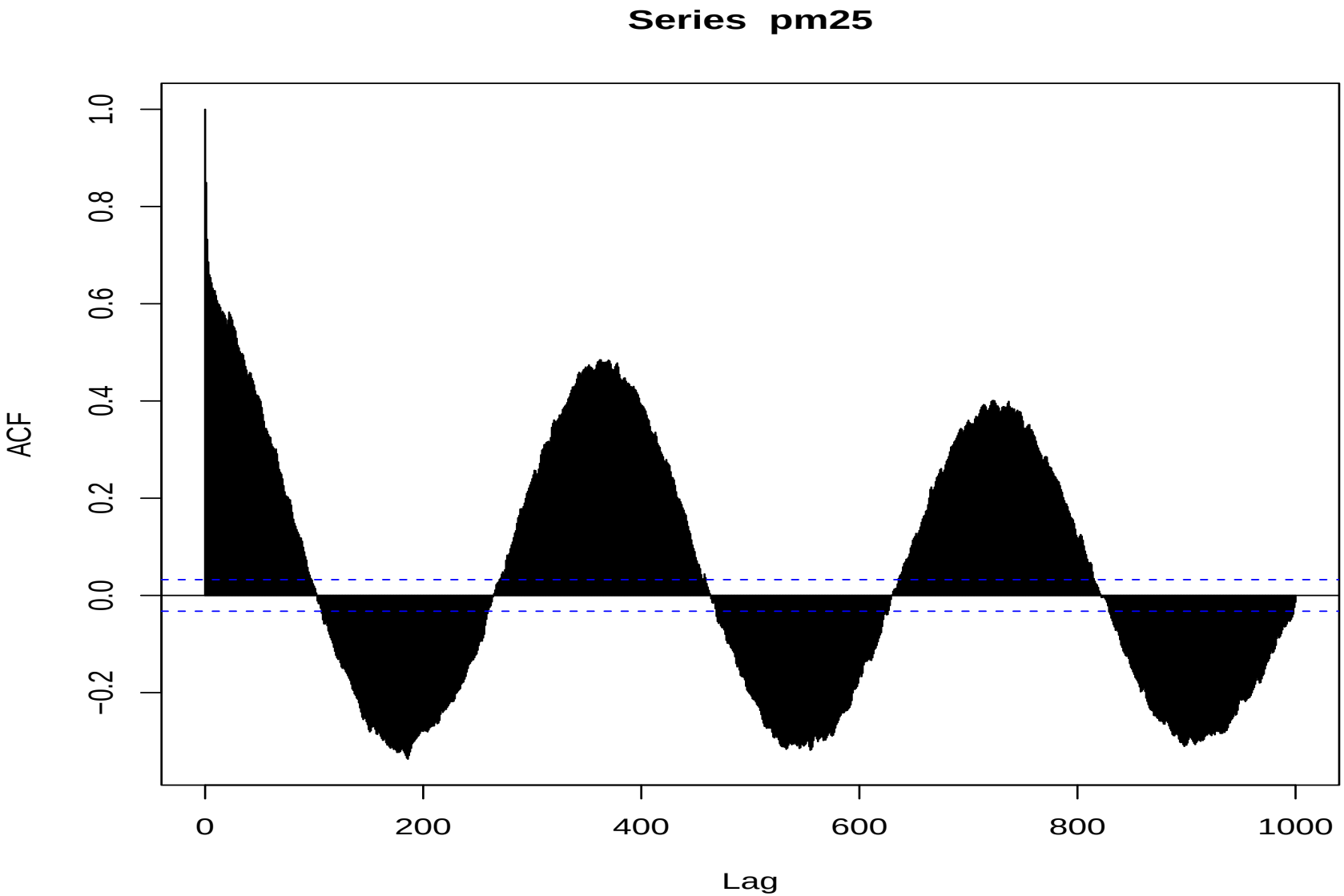
1. $n^{-4}\sum_{t=1}^{n} x_t^2 \Rightarrow \int_0^1 \bar{W}^2$, where $\bar{W} = \int_0^1 W(s)ds$ with $W(s)$ the standard Brownian motion.
2. $n^{-5/2}\sum_{t=1}^{n} x_t \Rightarrow \int_0^1 \bar{W}$
3. $n^{-3}\sum_{t=1}^{n} x_t \Delta x_t \Rightarrow \int_0^1 \bar{W}W$
4. $n^{-2}\sum_{t=1}^{n} (\Delta x_t)^2 \Rightarrow \int_0^1 W^2$

Standardization may wash out the Δx_{t-1} and $\Delta^2 x_{t-1}$ parts.

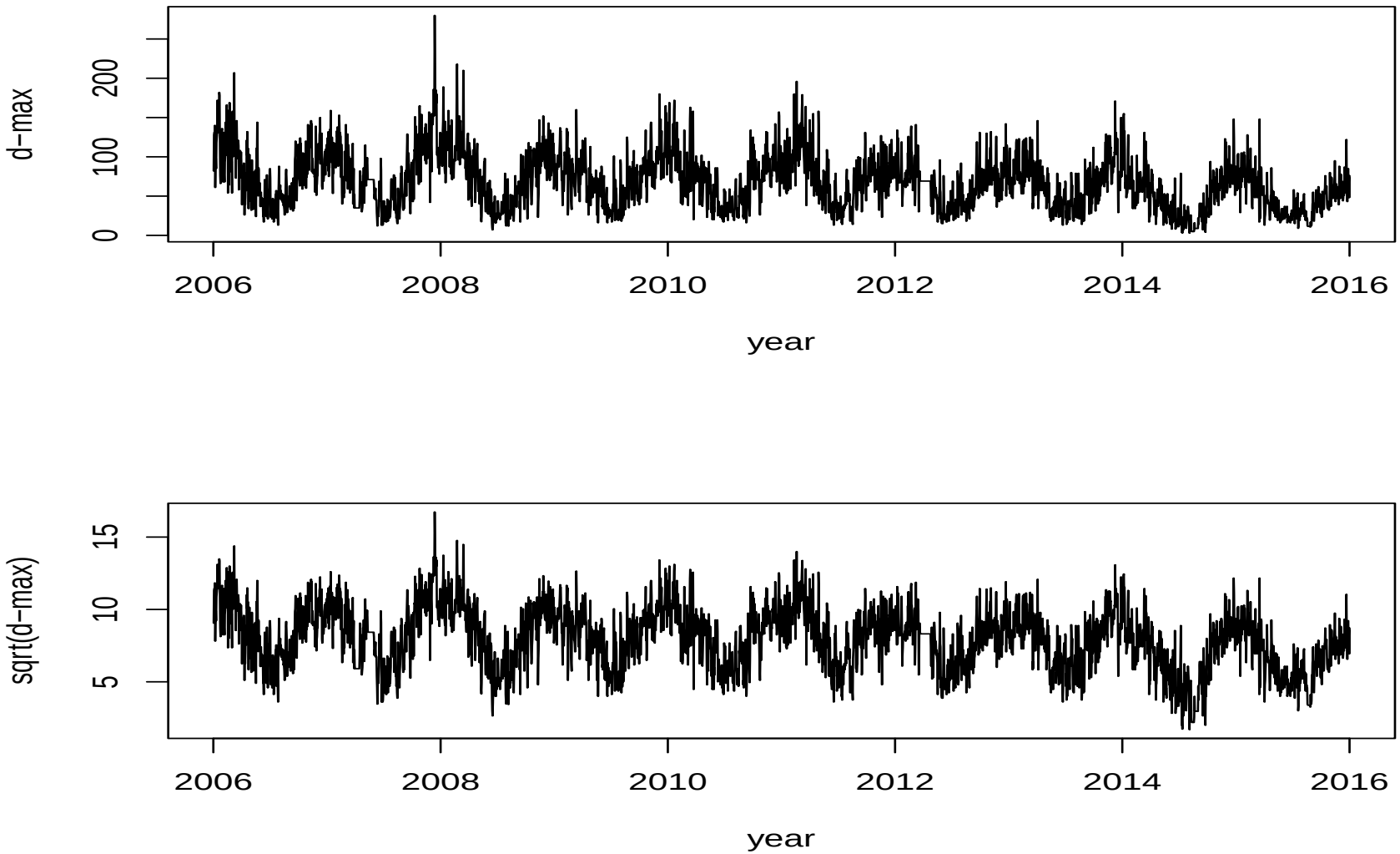
Daily $PM_{2.5}$: a mixed-frequency example



Sample autocorrelations of daily $PM_{2.5}$



Daily Maximum $PM_{2.5}$: a mixed-frequency example



Empirical analysis

1. Has strong seasonal pattern and dynamic dependence
2. Response variable: square-root of daily maximum $PM_{2.5}$ (y_t)
3. Hourly data are available. Use the first 6 hourly data as possible regressors (z_{it})
4. Regressors: $\boldsymbol{x}_t = (y_{t-1}, \ldots, y_{t-22})'$ or $\boldsymbol{X}_t = (\boldsymbol{x}_t', z_{1t}, \ldots, z_{6t})'$
5. In our case, it means that at 6 am, one can update the forecast of the maximum value of $PM_{2.5}$
6. Benchmark: AR(22) model selected by AIC
7. Forecasting subsample: the last two years with 730 observations

Empircal results of daily $\sqrt{\max PM_{2.5}}$

Method	Bias	RMSE	MAE
Benchmark	-0.099	1.145	0.886
Backtest(2,1,1)	-0.012	1.141	0.874
nstaFore(10)	-0.095	1.146	0.889
glmnet(0.1)	-0.107	1.141	0.886
PLS(22)	0.101	1.154	0.930
PCR(22)	0.114	1.192	0.886
RandomForest	-0.330	1.254	0.982
Backtest(hourly)	-0.008	0.887	0.682
nstaFore(10,hourly)	-0.088	0.915	0.698
PLS(25,hourly)	-0.138	0.943	0.720
PCR(25,hourly)	-0.163	0.972	0.737
glmnet (0.6,hourly)	-0.130	0.927	0.706
RandomForest(hourly)	-0.216	0.914	0.703
RandomForest-sel(hourly)	-0.202	0.887	0.694

- Random Forest-sel: Class 1: hourly data + (y_{t-1}, y_{t-2})