Machine Learning using Nonstationary Data*

Jin Xi†

November 4, 2023

Abstract

Machine learning offers a promising set of tools for forecasting. However, some of the well-known properties do not apply to nonstationary data. This paper uses a simple procedure to extend machine learning methods to nonstationary data that does not require the researcher to have prior knowledge of which variables are nonstationary or the nature of the nonstationarity. I illustrate theoretically that using this procedure with LASSO or adaptive LASSO generates consistent variable selection on a mix of stationary and nonstationary explanatory variables. In an empirical exercise, I examine the success of this approach at forecasting U.S. inflation rates and the industrial production index using a number of different machine learning methods. I find that the proposed method either significantly improves prediction accuracy over traditional practices or delivers comparable performance, making it a reliable choice for obtaining stationary components of high-dimensional data.

Keywords: Nonstationarity, High-Dimensional Data, Machine Learning, Forecasting, LASSO, Adaptive LASSO.

*The author is indebted to her advisor James Hamilton, and her committee members - Yixiao Sun, Allan Timmermann, Xinwei Ma, and Alexis Toda - for their constant support and helpful comments. The author also thanks Kaspar Wüthrich, Ying Zhu, and participants at UCSD, the Macro Forecasting Seminar, and the Trends in Macroeconometrics workshop.

†Department of Economics, University of California, San Diego. Address: 9500 Gilman Dr. La Jolla, CA 92093. Email: x5jin@ucsd.edu.

1 Introduction

Machine learning refers to a variety of computational methods for forming forecasts using large numbers of potential predictors. Medeiros and Vasconcelos (2016) and Goulet Coulombe et al. (2022) found that machine learning delivers promising forecasts for many economic variables. Good forecasts of inflation are particularly important for policy makers and business planners, but are particularly hard to obtain using OLS, as demonstrated by Atkeson and Ohanian (2001), Fisher et al. (2002), and Stock and Watson (2007). Machine learning methods may materially improve inflation forecasts (Inoue and Kilian, 2008; Medeiros et al., 2021). Stock returns are another variable that is notorious for being difficult to forecast, for which machine learning has shown promise (Koo et al., 2020; Lee et al., 2020).

One challenge in applying these methods is that many of the variables in economics and finance appear to be nonstationary. Stationarity is important for two reasons. First, initial normalization is common for approaches such as LASSO, ridge regression, elastic net, and principal component analysis. This normalization entails subtracting the sample mean from each series and dividing it by the sample standard deviation. However, if a variable is nonstationary, the sample mean and sample standard deviation diverge as $T \to \infty$.

Second, the presence of highly persistent series can exert a dominant influence during estimation and result in misleading outcomes. Onatski and Wang (2021) provide theoretical evidence demonstrating that nonstationary series tend to absorb a significant portion of data variation, creating the illusion of a few influential factors even in the absence of underlying structural factors. Additionally, integrated variables, such as those exhibiting unit roots, typically exhibit weak sample correlations with stationary variables that have low persistence. Consequently, regression techniques such as LASSO and its variants may encounter difficulties in identifying relevant predictors through the regularization technique.

Isolating the stationary component of each series poses an important challenge. The common practice involves transforming each series into a stationary form, which often involves rigorous analysis of the underlying trend mechanism and multiple intricate

tests. Determining the appropriate transformation is not always trivial. Furthermore, when dealing with a large set of predictors, it becomes important to employ an efficient and easily replicable technique that does not necessitate individual analysis of each series.

For many stationary or nonstationary variables, one can characterize the error made when attempting to predict the variable h periods in advance using a linear function of p of its own lagged values. Hamilton (2018) showed that this error remains stationary for a large class of both stationary and nonstationary processes, including those with one or more unit roots or deterministic polynomial time trends. Hamilton proposed that the forecast error over a two-year horizon provides a practical interpretation of what we could mean by the stationary cyclical component of a potentially nonstationary variable. One appealing feature of this definition is that the forecast error can be consistently estimated using OLS regression without the need for prior knowledge of the trend mechanism. As a result, this detrending procedure is fully automatic and treats each series in a uniform manner. I review these results in Section 2 of this paper.

My proposal is to use this automatic detrending prior to applying any of the popular machine learning methods. That is, for each explanatory variable z_{it} in the original data, we first estimate the following regression by OLS,

$$z_{it} = \hat{\alpha}_{i0} + \hat{\alpha}_{i1} z_{i,t-h} + \hat{\alpha}_{i2} z_{i,t-h-1} + \cdots + \hat{\alpha}_{ip} z_{i,t-h-p+1} + \hat{x}_{it}, \tag{1}$$

and then use the estimated OLS residuals $\{\hat{x}_{it}\}_{i=1}^{n}$ as potential predictors to forecast the future values of some stationary variable of interest y_t. In Section 3, I verify theoretically that this approach results in consistent variable selection when used with two popular machine learning methods, LASSO and adaptive LASSO. LASSO was first proposed by Tibshirani (1996) and extended to adaptive LASSO by Zou (2006). The asymptotic properties when the potential predictors are all stationary were analyzed by Zhao and Yu (2006) and Medeiros and Mendes (2016). I demonstrate that with some mild additional conditions, these results can be extended to nonstationary predictor variables using the method proposed here.

Section 4 conducts an empirical study to evaluate the performance of various ma-

chine learning methods in forecasting inflation rates and industrial production. To that end, I first reproduce the results in Medeiros et al. (2021), forecasting inflation and industrial production over their original sample period of 1990-2015 using a large data set and a wide range of machine learning techniques and using their original proposed transformations to make each individual variable stationary. Medeiros et al. (2021) find that machine learning methods, particularly the random forests, provide good forecasting performance over this sample period. I then analyze the same data over the same period using the automatic detrending procedure proposed here in place of the individually-selected transformations used in the original study. The results indicate that the forecasts obtained using my proposed method are comparable to the original findings. I further extended the dataset up to December 2022. During this extended period, prediction is considerably more challenging. I discovered that our automatic approach for machine learning estimates significantly reduces prediction errors, achieving up to a 50% reduction when compared to using the selected transformations employed in Medeiros et al. (2021). Notably, the application of the random forests method to automatically detrended data consistently produces outstanding forecasts, irrespective of the sample period. Overall, I conclude that regardless of the specific machine learning method or sample period used, basing estimates on automatically detrended series leads at worst to comparable results and in many cases to significantly more accurate predictions.

2 Proposed Method for Detrending Variables

In an important paper, Hamilton (2018) defined the cyclical component of any variable to be the error we would make in trying to forecast the value at date t as a linear function of p of its own values observed as of date $t-h$. Let z_{it} denote the level of a potentially nonstationary variable that is proposed as a potential predictor of some

stationary variable of interest. Hamilton's cyclical component x_{it} is defined as

$$\begin{aligned} x_{it} &= z_{it} - \mathbb{P}(z_{it}|1, z_{i,t-h}, z_{i,t-h-1}, \ldots, z_{i,t-h-p+1}) \\ &= z_{it} - \alpha_{i0} - \alpha_{i1} z_{i,t-h} - \alpha_{i2} z_{i,t-h-1} - \cdots - \alpha_{ip} z_{i,t-h-p+1}, \end{aligned} \tag{2}$$

where $\mathbb{P}(y|x)$ denotes population linear projection of y on x, h is the forecasting horizon for detrending, and p is the number of lags used for prediction. Note that x_{it} is a population concept that is determined by the true data-generating process for z_{it}.

This definition of the cyclical component is related to the definition of trend proposed by Beveridge and Nelson (1981). Their decomposition defines the permanent component δ_{it} to be the long-run forecast of the future value:

$$\delta_{it} = \lim_{h \to \infty} \lim_{p \to \infty} E[z_{i,t+h}|z_{it}, z_{i,t-1}, \ldots, z_{i,t-p+1}]. \tag{3}$$

However, their results only apply to unit root processes and require the researcher to know the population parameters of the process in order to be able to calculate the double limits in (3).

This paper follows Hamilton (2018) and Hamilton and Xi (2023) in taking both h and p to be fixed. They argue that $h = 24$ for monthly data or $h = 8$ for quarterly data is a good choice, because the error in making a two-year-ahead forecast comes primarily from cyclical factors such as whether a recession occurs in the next two years, whereas something that can be predicted more than two years ahead results from broader and slower-moving economic forces.

As demonstrated by Hamilton (2018) and Hamilton and Xi (2023), the cyclical component defined by (2) is attractive for three important reasons. First, regardless of the choice of h and p, the population object defined in (2) is stationary for a broad class of stationary and nonstationary processes that could have generated z_{it}. This property is fundamental for the processed data to be used in machine learning approaches. Second, the population linear projection coefficients α_{ij} can be consistently estimated by OLS estimation of (1) without needing to know the nature of the nonstationarity. We can apply a unified linear projection without considering whether a series is stationary or

not. Third, the cumulative squared difference between the OLS regression estimate of the cyclical component $\hat{x}_{it}$ and the true cyclical component x_{it} defined in (2) is bounded in probability,

$$\sum_{t=1}^{T} v_{it}^2 = O_p(1), \tag{4}$$

for $v_{it} = \hat{x}_{it} - x_{it}$. Condition (4) turns out to be more than needed to extend many machine learning methods to nonstationary data, as I demonstrate in the next section.

I now illustrate the above three claims with some leading examples.

Consider first a stationary AR(1) process,

$$z_{it} = \phi_i z_{i,t-1} + \epsilon_{it},$$

where ϵ_{it} is white noise and $|\phi_i| < 1$. The linear projection coefficients in (2) for this case turn out to be $\alpha_{i1} = \phi_i^h$ and $\alpha_{ij} = 0$ for $j \neq 1$. The cyclical component is $x_{it} = \sum_{s=0}^{h-1} \phi_i^s \epsilon_{i,t-s}$ which is stationary with variance $V_i = \sum_{s=0}^{h-1} \phi_i^{2s} \sigma_\epsilon^2$. The OLS estimate is

$$\hat{\boldsymbol{\alpha}}_i = \left(\sum_{t=1}^{T} \boldsymbol{w}_{it} \boldsymbol{w}_{it}' \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{w}_{it} z_{it} \right)$$

for $\boldsymbol{w}_{it} = (1, z_{i,t-h}, z_{i,t-h-1}, ..., z_{i,t-h-p+1})'$. The difference between the OLS regression coefficient and the population linear projection coefficient is

$$\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i = \left(T^{-1} \sum_{t=1}^{T} \boldsymbol{w}_{it} \boldsymbol{w}_{it}' \right)^{-1} \left(T^{-1} \sum_{t=1}^{T} \boldsymbol{w}_{it} x_{it} \right) \xrightarrow{p} [E(\boldsymbol{w}_{it} \boldsymbol{w}_{it}')]^{-1} [E(\boldsymbol{w}_{it} x_{it})] = \mathbf{0}$$

since $E(\boldsymbol{w}_{it} x_{it}) = \mathbf{0}$. This confirms that OLS regression gives consistent estimates of the population coefficients. The cumulative squared difference between the estimated and true cyclical components is

$$\begin{aligned} \sum_{t=1}^{T} v_{it}^2 &= \sum_{t=1}^{T} (\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i)' \boldsymbol{w}_{it} \boldsymbol{w}_{it}' (\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i) \\ &= \left[T^{1/2} (\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i) \right]' \left(T^{-1} \sum_{t=1}^{T} \boldsymbol{w}_{it} \boldsymbol{w}_{it}' \right) \left[T^{1/2} (\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i) \right]. \end{aligned}$$

But $T^{-1}\sum_{t=1}^{T}\boldsymbol{w}_{it}\boldsymbol{w}_{it}' \xrightarrow{p} E(\boldsymbol{w}_{it}\boldsymbol{w}_{it}')$ and $T^{1/2}(\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i) = O_p(1)$, confirming (4).

For a more general process z_{it} that is strictly stationary and ergodic, x_{it} defined by (2) is a linear combination of z_{it} and its past values, each of which is stationary. Hence, the forecast error x_{it} is stationary, and it's not hard to show that we still have $\hat{\boldsymbol{\alpha}}_i \xrightarrow{p} \boldsymbol{\alpha}_i$ and $\sum_{t=1}^{T} v_{it}^2 = O_p(1)$.

Consider next a random walk: $z_{it} = z_{i,t-1} + \epsilon_{it}$. This equation implies that $z_{it} = z_{i,t-h} + x_{it}$ for $x_{it} = \sum_{s=0}^{h-1}\epsilon_{i,t-s}$. The population linear projection coefficients are $\alpha_{ij} = 1$ for $j = 1$ and zero otherwise. The term x_{it} is the true cyclical component, which is stationary with variance $h\sigma_\epsilon^2$. Taking for illustration the simple case when $h = p = 1$ and ϵ_{it} is a martingale-difference sequence with variance σ_i^2 and finite fourth moments, we have the well known unit-root results that

$$T^{-2}\sum_{t=1}^{T} z_{i,t-1}^2 \xrightarrow{d} \sigma_i^2 \int_0^1 [W_i(r)]^2 dr$$

$$T(\hat{\alpha}_i - \alpha_i) \xrightarrow{d} \left[\int_0^1 [W_i(r)]^2 dr\right]^{-1}\left[\int_0^1 W_i(r)]dW_i(r)\right]$$

for $W_i(r)$ standard Brownian motion. These imply $\hat{\alpha}_i \xrightarrow{p} \alpha_i$ and

$$\sum_{t=1}^{T} v_{it}^2 = \left[T(\hat{\alpha}_i - \alpha_i)\right]^2\left[T^{-2}\sum_{t=1}^{T} z_{i,t-1}^2\right] \xrightarrow{d} \frac{\left[\int_0^1 W_i(r)]dW_i(r)\right]^2}{\sigma_i^2\left[\int_0^1 [W_i(r)]^2 dr\right]}$$

which again is $O_p(1)$.

More generally, for any variable, we have the accounting identity

$$z_{it} = z_{i,t-h} + \sum_{j=0}^{h-1}\Delta z_{i,t-j}.$$

If z_{it} is any unit-root process, then Δz_{it} is stationary meaning z_{it} differs from $z_{i,t-h}$ by a stationary term. The optimal forecast of z_{it} as a linear function of $z_{i,t-h}, z_{i,t-h-1}, ..., z_{i,t-h-p+1}$ is then equal to $z_{i,t-h}$ plus the forecast of $\sum_{j=0}^{h-1}\Delta z_{i,t-j}$. The latter can be calculated

using a linear projection on $p-1$ past changes of z_i:

$$\tilde{\boldsymbol{w}}_{i,t-h}^{p} = (1, \Delta z_{i,t-h}, \Delta z_{i,t-h-1}, \ldots, \Delta z_{i,t-h-p+2})'.$$

Let

$$\pi_{i,j-h} = \mathbb{E}[(\tilde{\boldsymbol{w}}_{i,t-h}^{p})'\tilde{\boldsymbol{w}}_{i,t-h}^{p}]^{-1})\mathbb{E}[(\tilde{\boldsymbol{w}}_{i,t-h}^{p})'\Delta z_{i,t-j}]$$

be the coefficient from projecting $\Delta z_{i,t-j}$ on $\tilde{\boldsymbol{w}}_{i,t-h}^{p}$. It follows that,

$$\begin{aligned}\mathbb{P}(z_{it}|1, z_{i,t-h}, z_{i,t-h-1}, \ldots, z_{i,t-h-p+1}) &= z_{i,t-h} + \sum_{j=0}^{h-1} \pi_{i,j-h}' \tilde{\boldsymbol{w}}_{i,t-h}^{p} \\ &= \alpha_{i0} + \alpha_{i1} z_{i,t-h} + \alpha_{i2} z_{i,t-h-1} + \cdots + \alpha_{ip} z_{i,t-h-p+1}.\end{aligned}$$

This identity allows us to derive α_{ij} from $\pi_{i,j-h}$. One can also generalize the unit-root results above as discussed in Hamilton (2018) and Hamilton and Xi (2023).

The above results can also be extended to more general forms of nonstationarity. Specifically, the population cyclical component x_{it} defined by (2) is stationary and can be consistently estimated by the corresponding OLS regression if either: (1) z_{it} is stationary around a polynomial deterministic trend of time with order of $d_i \leq p$, and satisfies

$$T^{-\frac{1}{2}} \sum_{s=1}^{[Tr]} (z_{it} - \delta_{i0} - \delta_{i1} t - \delta_{i2} t^2 \cdots - \delta_{i,d_i} t^{d_i}) \xrightarrow{d} \omega_i W_i(r),$$

where $[Tr]$ is the largest integer no greater than Tr, W_i is a Brownian motion; or alternatively (2) if d_i differences of z_{it} are stationary for some $d_i \leq p$, and satisfy

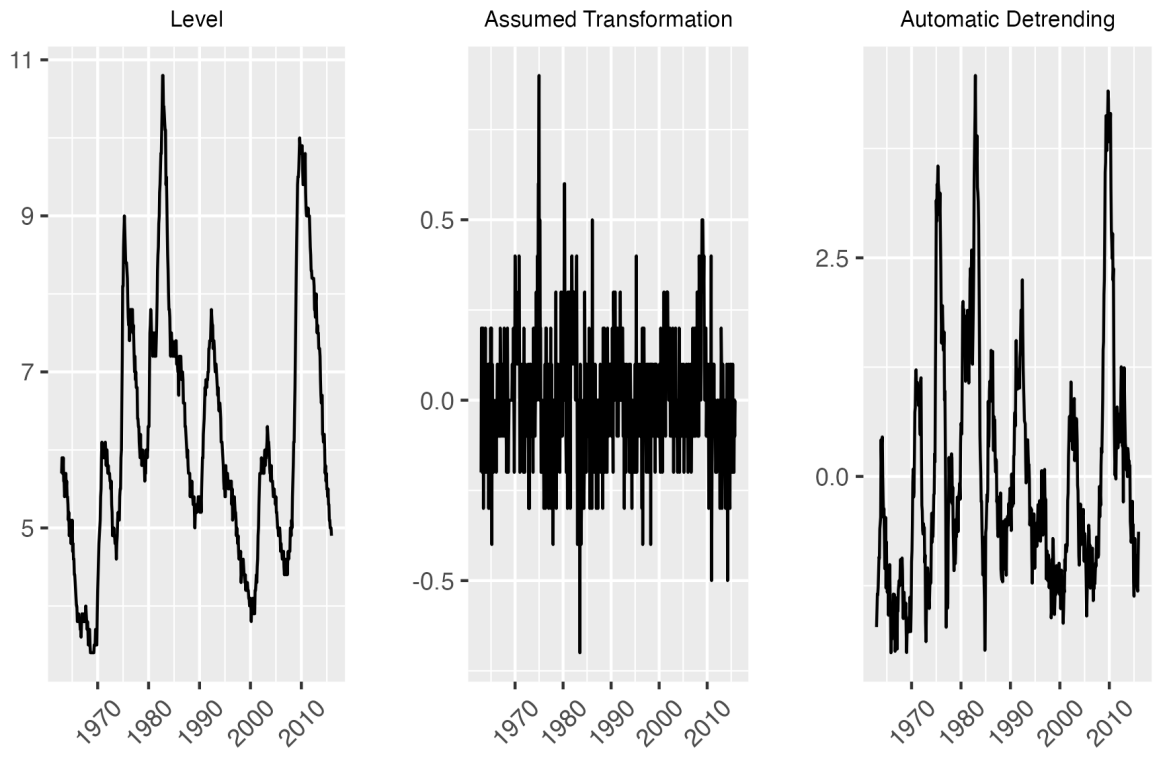
$$T^{-\frac{1}{2}} \sum_{s=1}^{[Tr]} (\Delta^{d_i} z_{it} - \mu_i) \xrightarrow{d} \omega_i W_i(r),$$

where μ_i represents the population mean of $\Delta^{d_i} z_{it}$.

To summarize, for a broad class of stationary and nonstationary processes, the cyclical component defined by (2) is stationary and can be consistently estimated with an OLS regression. We do not have to use any a priori knowledge about whether potential predictor variables are stationary, the value of d_i, or what transformations we need to

make in order to get a stationary variable. We can always use OLS to get a consistent estimate of the population cyclical component x_{it} without knowing d_i or the type of nonstationarity.

Figure 1: Level, first-differenced, and automatically detrended unemployment rate, 1990:01 - 2015:12

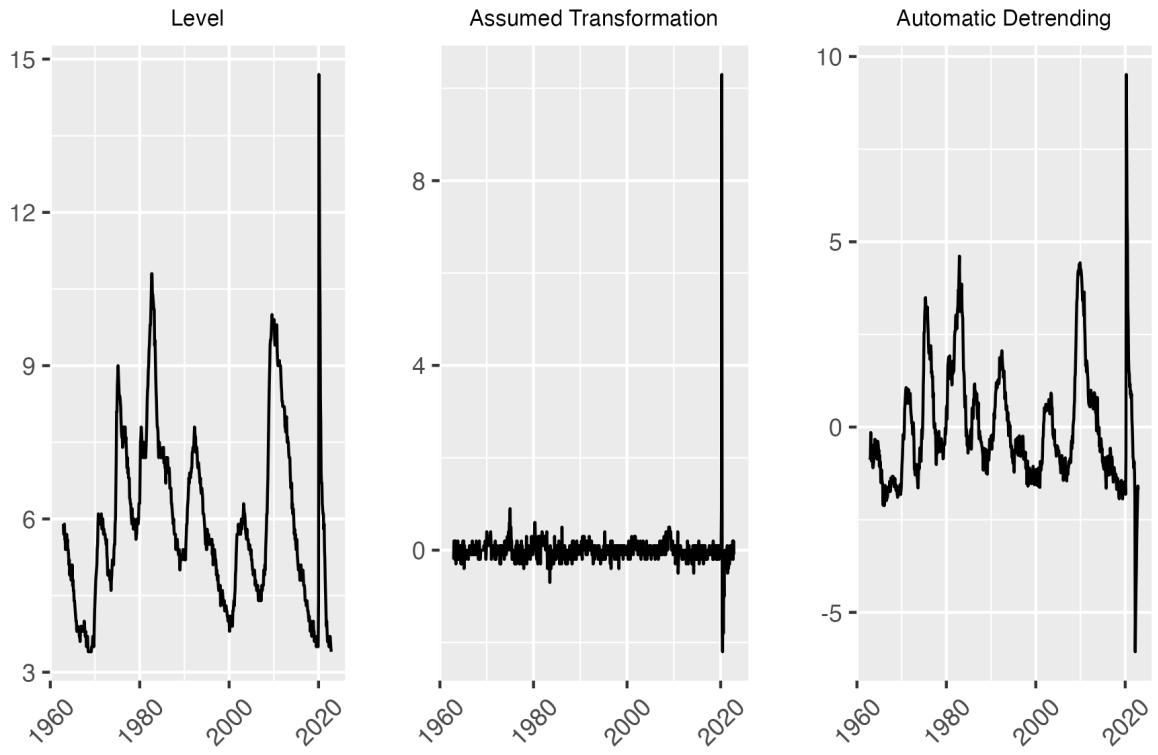


Lastly, we use the unemployment rate as an example to demonstrate this automatic detrending method. On the left of Figure 1 I plot the level of unemployment rate from January 1990 to December 2015. Based solely on the dynamics, it is not straightforward to determine whether the series is stationary or not. McCracken and Ng (2016) propose to use the first differences (FD) of unemployment, depicted in the middle of Figure 1, as a stationary series. This transformation is a reasonable way to generate a stationary series; however, the output behaves significantly differently from the original series. On the right, I plot the unemployment rate processed by the automatic detrending (AD) method. This output appears to be more informative in displaying the cyclical

dynamics of unemployment while showing no obvious trend.

Figure 2 presents the same set of three plots from January 1960 to December 2022. Due to the COVID-19 shock, at the beginning of year 2020, unemployment increases rapidly to 14.7% and dropped below 8% within 5 months. Such drastic movements lead to significant changes in first differences of unemployment. As depicted in the middle plot, first differences after 2020 are on a different scale compared to previous values. On the right, the automatically detrended series is behaves more closely to the original series, exhibiting a spike but remaining within the same scale as previous values. These figures, therefore, provide evidence that the automatic detrending method is capable of capturing the key information in the original series with the trend eliminated.

Figure 2: Level, first-differenced, and automatically detrended unemployment rate, 1990:01 - 2022:12



3 Theoretical Results

In this section I analyze the consequences of applying two popular machine learning tools, LASSO and adaptive LASSO, to residuals from the n first-stage regressions described in equation (1).

Statistical learning has two primary objectives: achieving robust prediction accuracy and uncovering relevant predictive factors. Variable selection is especially important when the underlying model has a sparse representation. With the boom of big data, the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996) received a lot of attention. The LASSO estimator uses a l_1 regularization technique that limits prediction variance in two ways: firstly, it tends to select a small set of variables with strong predictive power, and secondly, it tends to shrinks the non-zero coefficients towards zero. It has also been proved that the l_1 approach is able to discover the right sparse representation of the model under certain condition (Meinshausen and Bühlmann, 2006).

However, for LASSO to achieve selection consistency, a nontrivial assumption known as the "irrepresentable condition" is required. This condition serves to restrict the total number of irrelevant predictors represented by those that are relevant (Zhao and Yu (2006); Zou (2006)). Consequently, there are scenarios in which LASSO selection cannot be consistent. For that reason, Zou (2006) introduces the adaptive LASSO estimator as a variant of LASSO. The adaptive LASSO circumvents this condition by incorporating a set of data-dependent weights to penalize each coefficient individually. Zou (2006) establishes its oracle property with i.i.d. variables. Subsequently, Medeiros and Mendes (2016) investigate adaptive LASSO in a time-series context with stationary covariates. They provide sufficient conditions under which selection consistency and oracle property are achieved.

Most theoretical analyses of LASSO and adaptive LASSO have relied on the assumption of i.i.d or stationary data, yet in practice many variables we encounter are nonstationary. In this section, I explore these two models when applying the automatic detrending approach to the raw data prior to estimation with LASSO or adaptive LASSO. I prove that, under certain additional conditions applicable to a wide range

of data generating processes, both models can consistently reveal the correct sparse representation.

For illustrative purposes, I begin by examining properties of LASSO with a fixed number of covariates and relevant covariates, as the assumptions are more straightforward to interpret. The results can be naturally extended to allow for large number of predictors. Subsequently, I study the behavior of adaptive LASSO, allowing both the number of covariates and relevant covariates to approach infinity.

3.1 LASSO with Fixed n and Large T

Consider modeling a stationary variable y_t with a linear function of an $n \times 1$ vector $\boldsymbol{x}_t$ of potential predictors:

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta_0} + u_t. \tag{5}$$

Existing methods often require all predictors in $\boldsymbol{x}_t$ to be stationary. Consequently, researchers typically form an assessment of what transformations are needed, if any, to induce stationarity in each variable. In a common scenario, $\boldsymbol{x}_t$ is a mix of levels, first differences, and second differences of the original raw data z_t. Researchers assume that these transformed $\boldsymbol{x}_t$ are the predictors that appear in equation (5).

Instead, I propose using the stationary cyclical component x_{it} defined in equation (2) as a potential predictor of y_t. The postulation is that the ideal forecasting equation would resemble the form of (5) if we knew the true cyclical component x_{it} for each predictor. The data input feasible to us would be the OLS residuals $\hat{\boldsymbol{x}}_t$. The objective here is to examine whether LASSO can identify the relevant predictors in $\boldsymbol{x}_t$ if one could only use $\hat{\boldsymbol{x}}_t$.

Following common theory and practice, I assume that each potential predictor has zero mean and unit standard deviation. Note that this would be problematic if x_{it} was nonstationary, since the population variance is undefined and the sample standard deviation diverges to infinity. However, the true cyclical component is stationary by construction, and the sample variance of $\hat{x}_{it}$ converges in probability to the unit population variance.

The proposed feasible LASSO estimator is defined as

$$\hat{\boldsymbol{\beta}}^{LASSO} = \arg\min_{\beta} \|\boldsymbol{Y} - \hat{\boldsymbol{X}}\boldsymbol{\beta_0}\|_2^2 + \lambda_T \sum_{i=1}^{n} |\beta_j|, \tag{6}$$

where $\hat{\boldsymbol{X}}$ is a $T \times n$ matrix of residuals from the n OLS regressions in (1), and $\|\cdot\|_2^2$ denotes the Euclidean norm. The second term in equation (6) is the so-called "l_1 penalty ". λ_T is a non-negative regularization parameter that diverges with the sample size T. It allows LASSO to continuously shrink coefficients toward zero, or sometimes to exactly zero.

3.1.1 Assumptions for LASSO

In this section, I first outline the standard conditions that are typically imposed when we observe either i.i.d data or the stationary component $\boldsymbol{x}_t$. These assumptions pertain to the data generating process of $\boldsymbol{x}_t$, convergence of the empirical covariance matrix, and the "irrepresentability" or irrelevant variables. Additionally, it is crucial to have a reasonably good estimator $\hat{\boldsymbol{X}}$ such that the key information in $\boldsymbol{x}_t$ is not "filtered away". In response to that, I specify a sufficient assumption required for the estimator $\hat{\boldsymbol{X}}$.

The first set of assumptions restricts the data generating processes of $\boldsymbol{x_t}$ and u_t.

Assumption (DGP). *We make the following assumptions about* $\{\boldsymbol{x_t}, u_t\}$.

(i) $\{\boldsymbol{x_t}, u_t\}$ *is a zero-mean weakly stationary process with finite second moment.*

(ii) $E(\boldsymbol{x}_t\boldsymbol{x}_t') = \boldsymbol{\Omega}$ *is nonsingular with ones along the principal diagonal.*

Assumption DGP is standard. DGP (i) assumes that all variables that generate y_t are stationary with finite second moments. DGP (ii) is a standard normalization.

Before the next assumption, we need to first introduce some notations. Assume a subset of the predictors are useful for explaining y_t, whose indices are denoted by $S = \{i : \beta_{0,i} \neq 0\} \subseteq \{1, \ldots, n\}$. Let $s = |S|$ denote the number of relevant variables, and the remaining $n - s$ are considered as irrelevant variables. Usually it is assumed that the model has a sparse representation, i.e., $s \ll n$. Without loss of generality,

write $\boldsymbol{X}$ as $[\ \boldsymbol{X}(1)\ \ \boldsymbol{X}(2)\ ]$, where $\boldsymbol{X}(1)$ is a $T \times s$ matrix with the relevant variables, and $\boldsymbol{X}(2)$ is a $T \times (n-s)$ matrix with the irrelevant variables. Let $\boldsymbol{x}_t(1)'$ denote the tth row of $\boldsymbol{X}(1)$ and $\boldsymbol{x}_t(2)'$ the tth row of $\boldsymbol{X}(2)$. Similarly, let $\boldsymbol{\beta}_0 = [\boldsymbol{\beta}_0(1)', \mathbf{0}'_{n-s}]'$, where $\boldsymbol{\beta}_0(1) \in \mathbb{R}^s$ represents the active coefficients associated with $\boldsymbol{X}(1)$, and $\mathbf{0}'_{n-s}$ is a $(n-s) \times 1$ vector of zeros. We partition the variance matrix of $\boldsymbol{x}_t$ as

$$E\left[\boldsymbol{x}_t \boldsymbol{x}_t'\right] = \begin{bmatrix} E[\boldsymbol{x}_t(1)\boldsymbol{x}_t(1)'] & E[\boldsymbol{x}_t(1)\boldsymbol{x}_t(2)'] \\ E[\boldsymbol{x}_t(2)\boldsymbol{x}_t(1)'] & E[\boldsymbol{x}_t(2)\boldsymbol{x}_t(2)'] \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix} := \boldsymbol{\Omega}.$$

Accordingly, we partition the empirical covariance matrix as

$$T^{-1}\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} T^{-1}\boldsymbol{X}(1)'\boldsymbol{X}(1) & T^{-1}\boldsymbol{X}(1)'\boldsymbol{X}(2) \\ T^{-1}\boldsymbol{X}(2)'\boldsymbol{X}(1) & T^{-1}\boldsymbol{X}(2)'\boldsymbol{X}(2) \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{\Omega}}_{11} & \tilde{\boldsymbol{\Omega}}_{12} \\ \tilde{\boldsymbol{\Omega}}_{21} & \tilde{\boldsymbol{\Omega}}_{22} \end{bmatrix} := \tilde{\boldsymbol{\Omega}}.$$

Note that this empirical covariance matrix is not feasible since $\boldsymbol{X}$ is not directly observed.

Assumption (DESIGN). *The following conditions hold jointly.*

(i) $\boldsymbol{\beta}_0$ *is an element of an open subset* $\boldsymbol{\Theta}_n \subseteq \mathcal{R}^n$ *that contains* $\mathbf{0}$.

(ii) $\tilde{\boldsymbol{\Omega}} = T^{-1}\boldsymbol{X}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{\Omega}$ *as* $T \to \infty$.

Next, we introduce a condition on the correlation between the relevant and irrelevant variables.

Assumption (IC). *There exists a positive* $(n-s) \times 1$ *vector* $\boldsymbol{\eta} < \mathbf{1}$ *such that with probability approaching one*

$$|\boldsymbol{\Omega}_{21}(\boldsymbol{\Omega}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1))| \leq \mathbf{1} - \boldsymbol{\eta},$$

where the inequality holds element-by-element. Here $\boldsymbol{sign}(\cdot)$ *is a sign function that maps positive elements to 1, negative elements to -1 and zeros to zeros, i.e.,* $\boldsymbol{sign}(\boldsymbol{\beta}_0(1)) = \mathbb{1}\{\boldsymbol{\beta}_0(1) > 0\} - \mathbb{1}\{\boldsymbol{\beta}_0(1) < 0\}$.

The above condition is referred to as the "Irrepresentable Condition" (Zhao and Yu (2006); Zou (2006)). Note that $\mathbf{\Omega}_{21}(\mathbf{\Omega}_{11})^{-1}$ is the coefficient from a population linear projection of $\boldsymbol{x}_{2t}$ on $\boldsymbol{x}_{1t}$. The condition bounds the size of this coefficient, and thereby limits the extent to which the irrelavant variables could be correlated with the active variables. In particular, the Irrepresentable Condition would hold for any $\boldsymbol{sign}(\boldsymbol{\beta}_0(1))$ if

$$\max_{j \in S^c} \|(\boldsymbol{X}(1)'\boldsymbol{X}(1))^{-1}\boldsymbol{X}(1)'[\boldsymbol{X}(2)]_j\|_1 \leq 1 - \eta, \tag{7}$$

where $[\boldsymbol{X}(2)]_j$ is the jth column of $\boldsymbol{X}(2)$. Here $\|\boldsymbol{w}\|_1 = \mathbf{E}[\sum_{i=1}^n |w_i|]$ represents the population l_1-norm of an $n \times 1$ vector $\boldsymbol{w}$. Equation (7) is also referred to as the Mutual Incoherence condition (Wainwright, 2019). It explicitly restricts the maximum amount of any irrelevant variable $\boldsymbol{X}_{2j}$ could be explained by $\boldsymbol{X}(1)$. In the ideal case, if $\boldsymbol{X}_{2j}$ is orthogonal to the space spanned by $\boldsymbol{X}(1)$ for all $j = 1, \ldots, n-s$, then LASSO can differentiate the relevant and irrelevant variables perfectly. As we cannot hope for such orthogonality, the Irrepresentable Condition imposes certain level of orthogonality to exist. For instance, Nardi and Rinaldo (2011) demonstrate that if model (5) is a simple autoregression with large n, then condition IC can be satisfied with exponentially decaying autocovariances.

Lastly, I describe the condition required for the estimator $\hat{\boldsymbol{X}}$.

Assumption (ESTIMATION). *There exists $\epsilon > 0$ such that for all $i = 1, \ldots, n$*

$$\sum_{t=1}^{T} v_{it}^2 = O_p(T^{1/2-\epsilon}),$$

where $v_{it} = \hat{x}_{it} - x_{it}$ is the estimation error.

Note this assumption is strictly weaker than the condition (4), allowing $\sum_{t=1}^T v_{it}^2$ to grow with the sample size, as long as it does so slower than $\sqrt{T}$.

3.1.2 Main Results for LASSO

Theorem 1 (Sign Consistency for LASSO). *Suppose Assumptions DGP, DESIGN, IC and ESTIMATION hold. Let* $S = \{i : \beta_{0,i} \neq 0\}$ *denote the set of indices of nonzero coefficients, and* $S_T^{LASSO} = \{i : \hat{\beta}_i^{LASSO} \neq 0\}$. *If* $\frac{\lambda_T}{T} \to 0$ *and* $\frac{\lambda_T}{T^{\frac{1+c}{2}}} \to \infty$ *for some* $0 \leq c < 1$, *then*

(i) $P(S_T^{LASSO} = S) \to 1$.

(ii) $P[\boldsymbol{sign}(\hat{\boldsymbol{\beta}}^{LASSO}(1)) = \boldsymbol{sign}(\boldsymbol{\beta}_0(1))] \to 1 \quad$ *as* $T \to \infty$, *where* $\hat{\boldsymbol{\beta}}^{LASSO}(1)$ *represents the coefficient estimates associated with* $\boldsymbol{X}(1)$.

The first part of Theorem 1 establishes that LASSO can consistently choose relevant predictors with automatically detrended data. The second part shows that LASSO can correctly choose the signs of those relevant variables asymptotically.

3.2 Adaptive LASSO with Growing n and T

In this section, we establish selection consistency for the adaptive LASSO estimator, allowing both the number of candidate predictors and relevant predictors to approach infinity.

Similar to the previous section, we first layout assumptions imposed when the stationary component $\boldsymbol{x}_t$ was observed directly. These conditions were first established by Medeiros and Mendes (2016). They solve the standard minimization problem for adaptive LASSO:

$$\tilde{\boldsymbol{\beta}}^{MM} = \arg\min_{\beta} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda_{n,T} \sum_{i=1}^{n} w_i |\beta_i| \tag{8}$$

$\boldsymbol{w} := \{w_1, w_2, \ldots, w_n\}$ is a weights vector that allows each coefficient to be penalized individually. These weight parameters are data-driven and are crucial to achieve selection consistency without the Irrepresentable Condition.

The adaptive LASSO estimator is

$$\hat{\boldsymbol{\beta}}^{adalasso} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{Y} - \hat{\boldsymbol{X}}\boldsymbol{\beta}||_2^2 + \lambda_{n,T} \sum_{i=1}^{n} w_i |\beta_i|. \tag{9}$$

We adhere to the notation convention in Section 3.1.1. Let $\tilde{\boldsymbol{\beta}}^{MM}(1)$ and $\tilde{\boldsymbol{\beta}}^{MM}(2)$ denote the coefficient estimates associated with the relevant variables and irrelevant variables obtained from model (8), and let $\hat{\boldsymbol{\beta}}^{adalasso}(1)$ and $\hat{\boldsymbol{\beta}}^{adalasso}(2)$ be the estimates obtained from (9). To emphasize that the number of active predictors may increase with n, we use the notation $S_n = \{i : \beta_{0,i} \neq 0\} \subseteq \{1, \dots, n\}$ to denote the potentially growing set of indices of relevant covariates, and $s_n = |S_n|$ to denote its cardinality.

3.2.1 Assumptions

Assumption (DGP'). *In addition to DGP, the following assumptions hold for* $\{\boldsymbol{x}_t, u_t\}$.

(i) $\max_{1 \leq i \leq n} T^{-1} \sum_{t=1}^{T} x_{it}^2 \xrightarrow{p} 1$

(ii) For some finite, positive constant c_m *and* $m \geq 2$, $E|x_{it} u_t|^m \leq c_m$ *for* $\forall i = 1, \dots, n$ *and* $\forall t$.

(iii) $E[u_t | \mathcal{F}_t] = 0$, *where* $\mathcal{F}_t = \sigma(\boldsymbol{x_t}, u_{t-1}, \boldsymbol{x_{t-1}}, u_{t-2}, \boldsymbol{x_{t-2}}, u_{t-3}, \dots)$.

Assumption DGP' is also assumed in Medeiros and Mendes (2016). DGP' (i) requires the sample variances of all variables converges uniformly to their unit variances. This condition implicitly restricts the growing rates of n. We refer readers to Appendix A in Medeiros and Mendes (2016) for an in-depth discussion of potential data generating processes of x_{it} and rates of n. DGP' (iii) assumes u_t to be a martingale difference process and allows for conditional heteroskedasticity.

Assumption (DESIGN'). *In additional to DESIGN, the following conditions hold jointly.*

(i) There exists $\beta_{min} > 0$ *such that* $\min_{i=1,\dots s_n} |\beta_{0,i}| > \beta_{min}$.

(ii) *There exists a constant* $\beta_{max} < \infty$ *such that* $\max_{i=1,\dots s_n} |\beta_{0,i}| < \beta_{max}$.

(iii) *There exists a constant* $0 < \phi_{\min} < 1$ *such that* $\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\boldsymbol{\Omega_{11}}\boldsymbol{\alpha} > 2\phi_{\min}$.

(iv) $\max_{1\leq i,j\leq s_n} [|\tilde{\boldsymbol{\Omega}}_{\boldsymbol{11}} - \boldsymbol{\Omega_{11}}|]_{i,j} \leq \frac{\phi_{\min}}{s_n}$ *with probability converging to one as* $T \to \infty$.

DESIGN' (i) and (ii) define a lower bound and an upper bound of the non-zero coefficients $\{\beta_{0,1}, \dots, \beta_{0,s_n}\}$. Note that $\beta_{\min}$ is allowed to decrease with the sample size, and the rate is specified in the main theorem. DESIGN (iii) requires the matrix $\boldsymbol{\Omega_{11}}$ to be non-degenerate, in the sense that the minimal eigenvalue is bounded. DESIGN (iv) imposes uniform convergence on the sub-matrix $\tilde{\boldsymbol{\Omega}}_{\boldsymbol{11}}$, which places constraints on the dependence and tail structure of relevant predictors.

Assumption (WEIGHTS). *As* $T \to \infty$, *the weights* $w_1, \dots, w_n$ *satisfy*

(i) *There exists* $0 < a < 1$ *and a sufficiently large positive constant* c_w *such that*

$$\min_{i=s_n+1,\dots,n} T^{-a/2} w_i > c_w \sqrt{\frac{s_n}{\phi_{\min}}}$$

with probability approaching one.

(ii) *There exists* $w_{max} < T^{\frac{a}{2}}$ *such that* $\sum_{i=1}^{s_n} w_i^2 < s_n w_{max}^2$ *with probability converging to one.*

As discussed by Zou (2006), the weights vector is the key to achieve the oracle properties without imposing condition IC. The weights put distinct degrees of penalization to each coefficient. In the ideal case, we hope the irrelevant predictors are assigned with greater penalization, forcing their coefficient estimates towards zero values. For the same reason, weights imposed on relevant predictors should be limited. WEIGHTS (i) sets a lower bound for the rate of divergence of weights for irrelevant covariates, and WEIGHTS (ii) restricts the amount of weights put on relevant covariates.

With fixed number of covariates, Zou (2006) proposed to take $w_i = 1/|\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}|^\gamma$, where $\hat{\boldsymbol{\beta}}$ is a root-n-consistent estimator (for example, $\hat{\boldsymbol{\beta}}^{OLS}$) and γ is a positive constant. In that case, WEIGHTS (i) is satisfied for $\gamma > a$, and WEIGHTS (ii) automatically holds

for large T. When $n \to \infty$, a root-n-consistent estimator still satisfies WEIGHTS (i) if s_n increases slowly.

Assumption (REG). *As $T \to \infty$, the regularization parameter $\lambda_{n,T}$ satisfies*

(i)

$$\frac{n^{1/m} T^{(1-a)/2}}{\lambda_{n,T}} \to 0.$$

(ii)

$$\frac{s_n^{1/2} w_{max}}{\phi_{\min}} \frac{\lambda_{n,T}}{\sqrt{T}} \to 0.$$

The size of $\lambda_{n,T}$ depends on both n and T. If n is fixed, we can simply choose $\lambda_{n,T} = o(T^{1/2})$. For growing n, conditions (i) and (ii) implicitly impose constraints on the relationship between n, s, and T. This is because both conditions can only be satisfied jointly if $\frac{w_{\max}}{\phi_{\min}} n^{\frac{1}{m}} s_n^{\frac{1}{2}}$ increases slower than $T^{\frac{a}{2}}$. Medeiros and Mendes (2016) illustrate that $\lambda_{n,T} = n^{\frac{1}{m}} T^{\frac{1}{2} - a(\frac{1}{2} - \frac{1}{m})}$ is a viable choice, given that $s_n^{\frac{1}{2}} w_{\max} / \phi_{\min} = O(n^{\frac{b}{m}})$ for some $b > 0$ and that n is a specified polynomial function in T.

Lastly, we investigate conditions required for the estimator $\hat{\boldsymbol{X}}$. We impose two conditions on the order of estimation error $v_{it} := \hat{x}_{it} - x_{it}$.

Assumption (ESTIMATION'). *As $T \to \infty$,*

(i)

$$\max_{1 \leq i \leq n} \sum_{t=1}^{T} v_{it}^2 = o_p\left(\lambda_{n,T}^2 T^{a-1}\right).$$

(ii)

$$\sum_{t=1}^{T} \Big(\sum_{i=1}^{s_n} v_{it} \Big)^2 = o_p\left(\lambda_{n,T}^2 T^{a-1}\right).$$

The first part of assumption ESTIMATION' requires the maximum of sum of squared estimation error to increase at a rate slower than $\lambda_{n,T}^2 T^{a-1}$. Note that according to REG (i), the rate of $\lambda_{n,T}^2 T^{a-1}$ exceeds $n^{\frac{2}{m}}$. Suppose we take our previous example of $\lambda_{n,T} = n^{\frac{1}{m}} T^{\frac{1}{2}-a(\frac{1}{2}-\frac{1}{m})}$, then we have $\lambda_{n,T}^2 T^{a-1} = n^{\frac{2}{m}} T^{\frac{2}{m}}$. Since we've shown that for a wide class of data generating processes $\sum_{t=1}^T v_{it}^2 = O_p(1)$, restricting the maximum of this variable then implicitly restricts the number of candidate predictors. Condition (ii) can be satisfied if $\max_{1\leq i\leq s_n} \sum_{t=1}^T v_{it}^2 = o_p\left(s_n^{-1}\lambda_{n,T}^2 T^{a-1}\right)$, although this condition is much stronger than (ii). In the example of $s = n^{\frac{1}{m}}$, (ii) is satisfied if $\max_{1\leq i\leq s_n} \sum_{t=1}^T v_{it}^2 = o_p\left(n^{\frac{1}{m}} T^{\frac{2}{m}}\right)$. Therefore, Estimation' (ii) restricts the number of active predictors.

3.2.2 Main Results for Adaptive LASSO

The following theorem reproduces the first main result established in Medeiros and Mendes (2016). This result shows that the adaptive LASSO estimator consistently selects the relevant variables and their signs when $\boldsymbol{x}_t$ is feasible to the researcher.

Theorem 2. *(Medeiros and Mendes, 2016) Suppose Assumption DGP', DESIGN', WEIGHTS and REG hold. Let* $S_{n,T}^{MM} = \{i : \tilde{\beta}_i^{MM} \neq 0\}$. *If* $\beta_{min} > \frac{\lambda_{n,T}}{T^{1-a/2}} \frac{s_n^{1/2}}{\phi_{\min}}$, *then*

(i) $P(S_n = S_{n,T}^{MM}) \to 1$.

(ii) $P[sign(\tilde{\boldsymbol{\beta}}^{MM}(1)) = sign(\boldsymbol{\beta_0}(1))] \to 1 \quad as\ T \to \infty$.

The next theorem demonstrates that the previous result also applies to $\hat{\boldsymbol{\beta}}^{adalasso}$ when $\boldsymbol{x}_t$ has to be estimated, provided that condition ESTIMATION' holds.

Theorem 3. *Assume DGP', DESIGN', WEIGHTS, REG, and ESTIMATION' hold. Let* $S_{n,T}^{adalasso} = \{i : \hat{\beta}_i^{adalasso} \neq 0\}$. *If* $\beta_{min} > \frac{\lambda_{n,T}}{T^{1-a/2}} \frac{s_n^{1/2}}{\phi_{\min}}$, *then*

(i) $P(S_n = S_{n,T}^{adalasso}) \to 1$.

(ii) $P[sign(\tilde{\boldsymbol{\beta}}^{adalasso}(1)) = sign(\boldsymbol{\beta_0}(1))] \to 1 \quad as\ T \to \infty$.

4 Forecasting Inflation and Industrial Production with Large Data

A fundamental goal of statistical learning is to enhance forecasting performance when many potential predictors are available. Inflation rates are particularly important for policymakers and business planners. Factor model first became prevalent in forecasting inflation. Stock and Watson (1999) pioneer the use of the first principal component derived from a pool of 168 macroeconomic variables in forecasting inflation. They find the inclusion of measures of aggregate activities consistently improve upon Phillip curve forecasts over the period 1959:01 - 1997:09. Their findings lead to the establishment of the Chicago Fed National Activity Index, the first principal component of 85 existing monthly macroeconomic indicators. Later on, more evidence emerged to support the usage of other machine learning methods. For instance, Inoue and Kilian (2008) find that bagging consistently outperforms the univariate benchmark in forecasting inflation over the period 1971:04 - 2003:07. The reduction in Mean Squared Error (MSE) can exceed 35% when forecasting inflation values in 1 year. They also find similar improvements when employing the Bayesian shrinkage predictor, the ridge regression predictor, the iterated LASSO predictor, and the Bayesian model average predictor based on random subsets of extra predictors.

A recent study deserving special attention is the one conducted by Medeiros et al. (2021). Using the FRED-MD data developed by McCracken and Ng (2016), they undertake a comprehensive investigation in search of the best machine learning approach for predicting inflation. They provide strong evidence that machine learning models are are systematically more accurate than the benchmarks when forecasting inflation during 1990:01 - 2015:12, achieving reductions in MSE up to 30% in some scenarios. They identify the random forests model as the winning model, attributing its success to its nonlinear nature and outstanding variable selection capabilities.

In most application we encounter, it is common to see researchers apply transformations that are assumed to remove trends. Each series is assigned with an individually hand-picked transformation, and the resulting data input usually consists of levels,

first-differences, second-differences, etc., of the original series. This is undoubtedly a reasonable way to generate stationary series. Yet choosing the right transformation can be a difficult task, especially for high-dimensional data. Moreover, as we illustrated in Section 2, transformed data tend to exhibit quite different properties.

In this application, I investigate the predictability of two variables - inflation rates and industrial production - using high-dimensional data that has been detrended using the proposed automatic detrending approach as opposed to using assumed transformations. To that end, I undertake a replication of the forecasting exercise in Medeiros et al. (2021) and introduce two important extensions to address our research question. Firstly, to determine which method yields more informative series for prediction, I employ the same models in Medeiros et al. (2021) as benchmarks and compare them with models using data detrended through the proposed automatic detrending procedure. This exercise aims to assess the predictive content of data applied with assumed transformation versus data detrended with the proposed automatic procedure for the sample period 1990:01 to 2015:12, which is the out-of-sample window studied in Medeiros et al. (2021).

Secondly, We update the above exercise to the most recent data. In addition to serving as a robustness check, the second extension is of practical importance, especially in light of the significant shocks experienced by many variables during the COVID-19 pandemic era. According to the April 2022 vintage of FRED-MD data, 40 out of 127 variables should be considered as outliers when compared to ten times their interquartile range. Therefore, selecting the appropriate method to isolate the stationary component becomes even more crucial during the time of COVID. Our results demonstrate that the proposed automatic detrending method significantly improves the predictability of both variables of interest.

4.1 Methodology and Data

Let y_t represent the outcome variables. In particular, we are interested in two variables:

$$y_t^{CPI} = 1200 \left[\log\left(CPI_t\right) - \log\left(CPI_{t-1}\right)\right] \tag{10}$$

and

$$y_t^{IP} = 1200 \left[\log\left(IP_t\right) - \log\left(IP_{t-1}\right)\right], \tag{11}$$

where CPI_t and IP_t are the consumer price index and industrial production index at time t.

Consider the following forecasting model

$$y_{t+h} = g^{(h)}(\boldsymbol{x}_t) + u_{t+h}, \tag{12}$$

where $\boldsymbol{x}_t$ is a $n \times 1$ vector of stationary predictors. Given that the raw data may exhibit nonstationarity, we assume $\boldsymbol{x}_t$ is only estimable, through either the proposed automatic detrending method or via the assumed transformation. $g^{(h)}(\cdot)$ is a (potentially non-linear) function that maps today's covariates to future values of inflation or industrial production h periods ahead. Here we use h to denote the horizon of the ***forecasting*** exercise. Note that this forecasting horizon differs from the one in the automatic detrending procedure defined in equation (2).

The forecasting model in Medeiros et al. (2021) is characterized by

$$\hat{y}_{t+h|t} = \hat{g}^{(h)}_{t-R_h+1:t}(\hat{\boldsymbol{x}}_t^T),$$

where $\hat{\boldsymbol{x}}_t^T$ denotes series processed by selected transformations, and $\hat{g}^{(h)}_{t-R_h+1:t}(\cdot)$ is the estimated forecasting function given data from time $t - R_h + 1$ to time t. The estimation of function $\hat{g}^{h}_{t-R_h+1:t}$ is based on rolling-windows of a fixed-length, and R_h represents the window size given the forecasting horizon h.

The forecasting model using data processed by our automatic detrending method is given by

$$\hat{y}_{t+h|t} = \hat{g}^{(h)}_{t-R_h+1:t}(\hat{\boldsymbol{x}}_t^{AD}).$$

The FRED-MD data used in this application is a large monthly macroeconomic database established by McCracken and Ng (2016). This dataset is updated in real-

time and can be accessed from McCraken's webpage[1]. FRED-MD covers a wide range of economic activities. Variables are divided into 8 categories: (1) output and income; (2) labor market; (3) housing; (4) consumption, orders, and inventories; (5) money and credit; (6) interest and exchange rates; (7) prices; and (8) stock market.

To facilitate the use of this database, McCracken and Ng (2016) provide a suggested transformation for each individual series. These transformations are indicated by a transformation code ("tcode ") in their database. The analyses conducted in Medeiros et al. (2021) are based on $\hat{x}_t^T$ derived using the specified transformation code.

Following Medeiros et al. (2021), the first forecasting exercise uses data from January 1960 to December 2015, and the out-of-sample period is January 1990 to December 2015. For the second exercise, I use data from January 1960 to December 2022, and the out-of-sample window is January 2016 to December 2022. Series with missing values are dropped, which leaves us 122 variables. Additionally, the first four principal component estimates are also included as potential predictors. For each of the 126 predictors, we include four of its lags, resulting in a total of 508 predictors.

4.2 Machine Learning Methods

In this section, I provide a brief overview of the machine learning approaches utilized in each forecasting exercise.

We employ two benchmark models: the random walk model (RW) and the autoregressive model (AR). In the RW model, prediction for h periods ahead is simply the current value, i.e.,

$$\hat{y}_{t+h|t} = y_t.$$

For the AR model, the prediction would be

$$\hat{y}_{t+h|t} = \hat{\phi}_0^h + \hat{\phi}_1^h y_t + \ldots, \hat{\phi}_p^h y_{t-p+1},$$

[1] https://research.stlouisfed.org/econ/mccracken/fred-databases/.

where the number of lags p is determined by BIC.

The traditional OLS approach tends to overfit high-dimensional data, and it achieves a perfect fit when $n \geq T$. However, this perfect in-sample fit merely means that OLS attempts to fit all information, including the noises. Consequently, the out-of-sample forecasts typically perform poorly and are sensitive to variations in the data values.

To reduce this forecasting variance, three classes of statistical learning methods are commonly used. The first class uses the shrinkage technique. Shrinkage methods tend to force coefficients toward zeros, and sometimes shrink to exactly zeros. The second class consists of the factor model and its variants. The factor model assumes that there exists a few common factors that drives the important dynamics in all variables. Each common factor is a weighted average of all variables, and usually only a few factors are included in predictions. The last class employs forecast combinations, where each forecast is a weighted average of multiple forecasts. Forecast combination could be especially helpful when the optimal model is uncertain or when many proxies are available to measure the same economic activity.

In what follows, we delve into the specifications of each class of models.

4.2.1 Shrinkage

The shrinkage technique has an attractive bias-variance trade-off that helps with the over-fitting issue. It reduces, or discard, the coefficient estimate a variable when its information is weak. The shrinkage technique limits the size of coefficient estimates by adding a penalty term, which is represented by $R(\beta_i, \lambda, w_i)$. in the following definition:

$$\hat{\beta} = \arg\min_{\beta} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \sum_{i=1}^{n} R(\beta_i, \lambda_{n,T}, w_i).$$

I consider the following three shrinkage methods.

1. Least Absolute Shrinkage and Selection Operator (LASSO)

As we described in Section 3.1.1, Lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} ||\boldsymbol{Y} - \hat{\boldsymbol{X}}\boldsymbol{\beta}||_2^2 + \lambda_{n,T} \sum_{i=1}^{n} |\beta_i|.$$

The penalty term $\lambda_{n,T} \sum_{i=1}^{n} |\beta_j|$ forces many coefficient estimates to be exactly zero. LASSO often offers good forecast performance because it reduces the variance component largely by discarding many predictors whose information is weak. It further reduces variance by limiting the size of non-zero coefficients estimates. The regularization parameter $\lambda_{n,T}$ is chosen by BIC.

2. Adaptive LASSO (adaLasso)

   Adaptive LASSO reduces forecast variances in a similar way to LASSO. It is defines as

$$\hat{\boldsymbol{\beta}}^{adalasso} = \arg\min_{\beta} ||\boldsymbol{Y} - \hat{\boldsymbol{X}}\boldsymbol{\beta_0}||_2^2 + \lambda_{n,T} \sum_{i=1}^{n} w_i|\beta_i|.$$

   Compared to LASSO, the adaptive LASSO introduces a set of data-dependent weight parameters $w_i = |\hat{\beta}^{FS}|^{-1}$, where $\hat{\beta}^{FS}$ comes from a first-step estimation[2].

3. Ridge Regression (RR)

   Ridge regression is proposed by Hoerl and Kennard (1970). It takes the following l_2 penalization form:

$$\hat{\boldsymbol{\beta}}^{RR} = \arg\min_{\beta} ||\boldsymbol{Y} - \hat{\boldsymbol{X}}\boldsymbol{\beta_0}||_2^2 + \lambda_{n,T} \sum_{i=1}^{N} \beta_i^2.$$

   Ridge regression does not discard elements; instead, it assigns a small weight to each predictor, and downweighs those that are less informative.

4.2.2 Common Factor Methods

1. Factor Model

[2] A viable choice of $\hat{\beta}^{FS}$ is the inverse of OLS estimate. I follow Medeiros et al (2021) and use $w_i = \frac{1}{\hat{\beta}_i^{lasso} + T^{-1/2}}$.

The factor model assumes that the underlying dynamics of a large number of predictors can be explained by a few common latent factors. It assumes

$$x_{it} = \lambda_i' f_t + u_{it},$$

where λ_i is a $r \times 1$ vector of factor loadings that are invariant over time, and f_t represents the r latent common factors. One needs to first estimate the common factors, and then use them to do forecasts. Factors can be computed by principal component analysis, whose solution is analytically available.

2. Target Factor Model

   The Target Factor Model, proposed by Bai and Ng (2008), is designed for scenarios with many weak predictors. It computes principal components with a subset of informative predictors. A first-stage linear regression is employed to select variables that are significant given a significance level.

4.2.3 Forecast Combinations

The forecast combination method takes average over several model predictions. Aggregating multiple predictive models helps to improve prediction accuracy because a single predictive model can be sensitive to small changes in the data or unstable decision rules.

1. Bagging

   In a series of influential technical reports, Breiman was among the earliest to demonstrate, both theoretically and empirically, that aggregating multiple versions of an estimator into an ensemble can give substantial gains in accuracy. Breiman (1996) proposes the bootstrap aggregation idea and named it bagging. For each bootstrapped subsample, the algorithm selects the variables with high t-statistics, and estimates the model again with only those selected variables. This set of coefficients are then used to get one prediction of the outcome of interests. The final prediction is calculated as the average of predictions from all bootstraps.

2. Complete Subset Regression (CSR)

While bagging averages predictions from sub-samples, CSR averages forecasts from subsets of covariates. For a fixed number of $k < n$ predictors, complete subset regression runs through all models with k predictors and take average of their predictions as the final prediction. Typically it is reasonable to choose k far below n. Combining diverse models often produces more stable forecasts. It was shown theoretically by Elliott et al. (2015) that CSR exhibits attractive bias-variance trade-off.

However, when the total number of predictors n is large, this algorithm is still computationally difficult even for small k. Medeiros et al. (2021) adopt a pre-selection procedure to eliminate weak predictors in the first stage: a linear regression with all predictors is fitted to select those with large t-statistics. This pre-selection procedure is also used in my exercise.

3. Random Forests (RF)

   The RF model, introduced by Breiman (2001), is established upon the bootstrap aggregation technique used in the bagging method. RF constructs a tree using each individual sub-sample and achieves variance reduction by averaging predictions from many sub-samples. Given that trees are notoriously noisy, RF greatly benefits from the aggregating procedure.

   Additionally, Random Forest (RF) employs only a subset of predictors to build each tree. At each split in the tree, the algorithm is restricted from considering the majority of predictors. Consequently, each tree uses a very different subset of predictors. Such design aims to reduce the correlation among trees, thereby further reduce variation stems from noisy predictors.

   Another notable feature of RF is its nonlinearity. While this feature increases computational complexity, it also allows RF to capture complex interaction structures in the data and reduces bias.

4.3 Results

The models are evaluated based on the root mean squared errors (RMSE). To facilitate illustration, we present their RMSEs relative to those of the random walk model. A RMSE ratio lower than one indicates that the model outperforms the random walk.

4.3.1 Forecasting CPI and IP from 1990 to 2015

Our first exercise uses data since 1960:01 to form out-of-sample predictions from 1990:01 to 2015:12. The primary objective is to compare the prediction power of two data inputs: data processed with assumed transformations versus data detrended using the proposed automatic method.

Table 1 reports the RMSE ratio for our first forecasting exercise, with forecasting horizon h ranges from 1 month to 12 months. The outcome variable here is y_t^{CPI} defined in (10). Each column in Table 1 displays the results for a particular forecasting horizon. To provide an overall performance summary, the last column in Table 1 displays the average RMSE ratio across all 12 forecast horizons.

For each machine learning method, the results are presented in two rows, with white-colored rows displaying outcomes obtained using data processed with assumed transformations ($\boldsymbol{x}_t^T$), and grey-colored rows presenting results from series detrended using the automatic method ($\boldsymbol{x}_t^{AD}$). For comparison purposes, bold text indicates scenarios where one data input results in an RMSE that is 5% lower than the other for a specific forecasting horizon and machine learning model. These numbers highlight cases where using $\hat{\boldsymbol{x}}_t^T$ and $\hat{\boldsymbol{x}}_t^{AD}$ can potentially lead to significantly different forecasts. Additionally, cells highlighted in yellow are used to emphasize scenarios where the RMSE difference exceeds 20%. Our results displayed in white-colored rows differ slightly from those in Medeiros et al. (2021) because our rolling window size R_h is 3-years shorter, as a result of a prior detrending step with linear projections.

We find that noticeable differences emerge when employing the ridge regression, CSR and bagging. For ridge regression and CSR, using transformed series consistently yields better forecasting performance, with reductions in RMSE of up to 20% for ridge regression and 11% for CSR. Taking the average over RMSE ratios, the differences are

11% and 7%, respectively. In contrast, the bagging method generally favors the use of our automatic detrending for most forecasting horizons. Notably, when the forecasting horizon is set to 3, 10, or 11 months, the improvements exceed 20% and may reach up to 23%. For the other machine learning methods, $\hat{x}_t^T$ and $\hat{x}_t^{AD}$ generate comparable performance with only marginal differences.

Table 2 presents the same forecasting exercise conducted for predicting industrial production between 1990:01 - 2015:12. We observe that adaptive LASSO and ridge regression favors the use of transformed series, especially for longer forecasting horizon. Using $\hat{x}_t^T$ improves RMSE by 11%, 12.5%, and 25% when h equals to 8, 11, or 12 months. For the other horizons the difference is small. In the case of ridge regression, the improvement ranges from 8% to 11% for forecasting horizons between 8 and 12 months. The bagging method once again prefer the use of $\hat{x}_t^{AD}$, particularly when h ranges from 7 to 9 months. The reduction can be as substantial as 47%, which is notably significant when compared to other scenarios.

It's worth noting that the random forests method exhibits a preference for $\hat{x}_t^{AD}$. This method stands out in Medeiros et al. (2021) for its exceptional forecasting performance across all exercises. Here we observe some significant improvements when h is 3, 6, or 7 months, and moderate improvements for the other forecasting horizons.

4.3.2 Forecasting CPI and IP from 2016 to 2022

In this subsection we conduct the same forecasting exercise for period 2016:01 to 2022:12.

Table 3 reports predictions of CPI. Notably, we observe a shift in results: the ridge regression method now strongly favors the of $\hat{x}_t^{AD}$ obtained from our automatic detrending method, marking a reversal of the outcomes seen in Table 1. The gain in terms of RMSE reaches up to 53.2%, and exceeds 20% for almost all horizons. We also observe strong performance using $\hat{x}_t^{AD}$ in the case of adaptive LASSO and factor model, with improvements up to 20% and 25%. Although when h is between 3 to 5 months, LASSO and adaptive LASSO appears to favor $\hat{x}_t^T$. Lastly, CSR once again favor $\hat{x}_t^T$, with improvements up to 15%.

The random forests method repeatedly delivers outstanding forecasting performance,

Table 1: Forecasting errors for CPI from 1990 to 2015

	forecasting horizon												
	1	2	3	4	5	6	7	8	9	10	11	12	Average
RW	1	1	1	1	1	1	1	1	1	1	1	1	1
AR	0.88	0.78	0.75	0.76	0.74	0.75	0.74	0.72	0.72	0.78	0.80	0.73	0.76
Lasso	**0.82**	0.74	0.72	0.75	0.73	0.73	0.72	0.72	0.73	0.77	0.80	0.72	0.75
	0.90	0.74	0.74	0.77	0.75	0.75	0.74	0.75	0.76	0.80	0.82	0.72	0.77
adaLasso	0.89	0.75	0.75	0.77	0.73	0.74	0.73	0.77	**0.78**	0.85	0.84	0.77	0.78
	0.89	0.75	0.75	0.78	0.76	0.76	0.74	0.75	0.85	**0.80**	0.86	0.74	0.79
Ridge Reg	**0.86**	**0.82**	**0.78**	**0.80**	**0.77**	**0.79**	**0.77**	0.84	0.83	**0.91**	**0.88**	**0.74**	**0.82**
	0.96	0.89	0.93	0.95	0.90	0.98	0.86	0.85	0.86	0.98	0.95	0.93	0.92
Factor	0.88	0.77	0.74	0.76	0.74	0.74	0.75	0.74	0.75	0.79	0.79	0.73	0.77
	0.90	0.77	0.74	0.76	0.73	0.73	0.74	0.73	0.73	0.78	0.79	0.72	0.76
Target Factor	0.85	0.77	0.76	0.78	0.75	0.76	0.78	0.74	0.74	0.78	0.80	0.74	0.77
	0.88	0.79	0.76	0.79	0.76	0.74	0.75	0.75	0.77	0.81	0.82	0.76	0.78
CSR	**0.88**	0.75	**0.72**	**0.75**	**0.73**	**0.74**	**0.73**	**0.73**	**0.73**	**0.78**	**0.80**	**0.72**	**0.76**
	0.94	0.77	0.78	0.81	0.79	0.79	0.79	0.78	0.79	0.86	0.90	0.81	0.82
Bagging	0.85	1.03	1.04	0.94	0.98	**1.07**	**0.86**	1.06	0.93	1.05	1.38	**0.98**	1.01
	0.88	**0.88**	**0.80**	**0.85**	1.02	1.17	0.93	1.10	0.97	**0.83**	**1.07**	1.12	0.97
Random Forests	0.86	0.77	0.73	0.76	0.74	0.74	0.75	0.73	0.73	0.77	0.80	0.73	0.75
	0.90	0.80	0.76	0.77	0.75	0.76	0.78	0.76	0.75	0.78	0.82	0.75	0.78

Note: The table shows the root mean squared error (RMSE) relative to the random walk's (RW) RMSE from predicting inflation rate of 1990:01-2015:12 sample. White-colored rows concerns data detrended by transformations. Grey-colored rows concerns data detrended by linear projection. Each column shows RMSE ratio given a forecasting horizon. The last column shows the average RMSE ratio across all forecast horizons.

Table 2: Forecasting errors for IP from 1990 to 2015

	forecasting horizon												
	1	2	3	4	5	6	7	8	9	10	11	12	Average
RW	1	1	1	1	1	1	1	1	1	1	1	1	1
AR	0.76	0.77	0.81	0.81	0.78	0.78	0.73	0.75	0.76	0.73	0.72	0.71	0.76
Lasso	**0.82**	0.74	0.72	0.75	0.73	0.73	0.72	0.72	0.73	0.77	0.80	0.72	0.75
	0.90	0.74	0.74	0.77	0.75	0.75	0.74	0.75	0.76	0.80	0.82	0.72	0.77
adaLasso	0.77	0.78	0.82	0.86	0.82	0.79	0.74	**0.73**	0.76	0.75	**0.77**	**0.77**	0.78
	0.72	0.79	0.86	0.84	0.82	0.82	0.77	0.82	0.81	0.78	1.02	0.88	0.83
Ridge Reg	0.72	0.75	0.81	0.82	0.85	0.83	0.81	**0.93**	**0.91**	**0.87**	**0.89**	**0.77**	**0.83**
	0.75	0.79	0.85	0.80	**0.80**	0.83	0.77	1.01	1.03	0.98	1.06	0.91	0.88
Factor	0.74	**0.74**	0.83	0.83	0.82	0.78	0.74	0.76	0.77	0.72	0.72	0.72	0.76
	0.77	0.83	0.85	0.83	0.83	0.81	0.76	0.78	0.78	0.74	0.72	0.71	0.79
Target Factor	0.72	0.73	0.82	0.80	0.78	0.75	0.74	0.76	0.76	0.71	0.71	0.71	0.75
	0.75	0.75	0.85	0.83	0.82	0.78	0.76	0.78	0.78	0.74	0.73	0.72	0.77
CSR	0.72	0.76	0.84	0.84	0.83	0.82	0.76	0.77	0.77	0.73	0.72	0.71	0.77
	0.75	0.79	0.86	0.85	0.83	0.84	0.78	0.82	0.81	0.75	0.76	0.74	0.79
Bagging	0.78	0.82	0.92	0.93	**0.86**	0.93	1.75	2.12	2.03	1.11	1.43	1.03	1.23
	0.79	0.85	0.94	0.95	1.03	0.90	**0.93**	**1.48**	**1.88**	1.14	1.43	1.21	**1.12**
Random Forests	0.72	0.78	0.84	0.82	0.82	0.82	0.77	0.80	0.79	0.77	0.76	0.75	0.79
	0.73	0.77	**0.79**	0.79	0.77	**0.78**	**0.75**	0.79	0.80	0.76	0.74	0.74	0.77

Note: The table shows the root mean squared error (RMSE) relative to the random walk's (RW) RMSE from predicting log difference of industrial production of 1990:01-2015:12 sample. White-colored rows concerns data detrended by transformations. Grey-colored rows concerns data detrended by linear projection. Each column shows RMSE ratio given a forecasting horizon. The last column shows the average RMSE ratio across all forecast horizons.

especially when using $\hat{x}_t^{AD}$. Remarkably, the RMSE ratios can be as low as 0.68, a rather substantial improvement over the benchmark. When compared to using transformed series $\hat{x}_t^T$, the improvement reaches up to 18%.

Finally, in Table 4 we observe that ridge regression continues to strongly favor $\hat{x}_t^{AD}$ over $\hat{x}_t^T$ when forecasting industrial production from 2016:01 to 2022:12, potentially achieving a 49% improvement in RMSE. Factor model also displays significant preference of $\hat{x}_t^{AD}$ over $\hat{x}_t^T$, showing improvements up to 38%. For nearly all forecasting horizons, the improvements are larger than 20%. Moreover, the random forests method once again consistently outperforms with the use of $\hat{x}_t^{AD}$ across forecasting horizons. Target factor, CSR and bagging also provide evidence that employing $\hat{x}_t^{AD}$ results in significantly more accurate predictions, particularly when the forecasting horizon is short. On the other hand, the LASSO estimator favors $\hat{x}_t^T$ when h is short, with differences in RMSE remaining below 15%.

The two forecasting exercises present strong evidence that the automatic detrending method handles drastic shocks better than transformations. To provide some intuition, let's consider a simple random walk process. In this scenario, the errors for a two-year-ahead forecast are the accumulation of 24 different one-month-ahead forecast errors. From the perspective of the Central Limit Theorem, a single shock would not significantly influence the value of this cumulative error. the authors discovered only two outliers among 134 variables in the April 2020 vintage when using the automatic detrending method, as opposed to 40 out of 123 when using transformations.

Lastly, in Appendix B we conduct a robustness check where the detrending coefficients in equation (2) are calculated with sample prior to year 2016. This exercise estimates the cyclical components with no information during the out-of-sample period, including the COVID shock. In general, we find no notable changes to our conclusions here.

4.3.3 Summary

In summary, our findings indicate that the automatic detrending method has the potential to significantly enhance forecasting performance across various machine learning

Table 3: Forecasting errors for CPI from 2016 to 2022

	forecasting horizon												
	1	2	3	4	5	6	7	8	9	10	11	12	Average
RW	1	1	1	1	1	1	1	1	1	1	1	1	1
AR	0.90	0.79	0.78	0.77	0.79	0.80	0.83	0.84	0.82	0.87	0.85	0.78	0.82
Lasso	0.85	0.83	**0.77**	**0.73**	**0.77**	0.76	0.79	0.80	0.77	0.84	0.82	0.78	0.79
	0.89	0.86	0.90	0.90	0.87	0.79	0.78	0.78	0.75	0.82	0.80	0.75	0.82
adaLasso	0.92	0.93	**0.83**	**0.76**	**0.78**	0.79	0.80	0.82	0.83	0.87	0.98	0.91	0.85
	0.91	0.87	0.92	0.94	0.94	0.80	0.78	**0.77**	**0.76**	**0.82**	**0.78**	**0.72**	0.84
Ridge Reg	1.46	1.46	1.16	1.91	2.04	1.67	2.22	1.83	1.70	2.00	1.75	1.68	1.74
	0.91	**0.93**	1.12	**1.11**	**1.14**	**1.06**	**1.04**	**0.93**	**1.04**	**1.17**	**1.38**	**1.29**	**1.09**
Factor	1.14	0.92	1.02	0.95	0.83	0.84	0.85	0.88	0.82	0.84	0.85	0.97	0.91
	0.90	**0.80**	**0.83**	**0.81**	0.82	0.80	0.83	**0.84**	0.84	0.85	**0.79**	**0.72**	**0.82**
Target Factor	0.98	**0.85**	**0.90**	0.86	0.80	0.84	0.89	1.18	0.89	**0.90**	0.93	0.95	0.91
	0.94	1.01	1.03	0.85	0.83	0.86	0.90	**0.88**	0.90	0.98	**0.88**	**0.84**	0.91
CSR	**0.86**	**0.75**	**0.76**	**0.75**	**0.79**	**0.78**	**0.82**	0.83	0.81	**0.86**	**0.80**	**0.80**	**0.80**
	0.99	0.84	0.86	0.86	0.93	0.88	0.90	0.87	0.84	0.96	0.86	0.80	0.88
Bagging	**0.82**	1.12	1.15	1.30	1.39	1.07	**1.15**	1.08	1.00	**1.00**	**0.98**	**1.03**	1.09
	1.13	1.12	**0.88**	**1.14**	**1.13**	1.12	1.26	1.10	**0.79**	1.11	1.32	1.24	1.11
Random Forests	0.85	0.83	0.84	0.80	0.82	0.79	0.82	0.84	0.83	0.90	0.87	0.83	0.83
	0.88	**0.71**	**0.69**	**0.70**	**0.72**	**0.74**	0.78	**0.76**	**0.76**	**0.81**	**0.74**	**0.68**	**0.75**

Note: The table shows the root mean squared error (RMSE) relative to the random walk's (RW) RMSE from predicting inflation rate of 2016:01-2022:12 sample. White-colored rows concerns data detrended by transformations. Grey-colored rows concerns data detrended by linear projection. Each column shows RMSE ratio given a forecasting horizon. The last column shows the average RMSE ratio across all forecast horizons.

Table 4: Forecasting errors for IP from 2016 to 2022

	forecasting horizon												
	1	2	3	4	5	6	7	8	9	10	11	12	Average
RW	1	1	1	1	1	1	1	1	1	1	1	1	1
AR	0.87	0.73	0.71	0.71	0.72	0.73	0.71	0.67	0.71	0.77	0.70	0.71	0.73
Lasso	**0.84**	0.83	**0.77**	**0.73**	**0.77**	0.76	0.79	0.80	0.77	0.84	0.82	0.78	0.79
	0.89	0.86	0.90	0.90	0.87	0.79	0.78	0.78	0.75	0.82	0.80	0.75	0.82
adaLasso	1.04	0.78	0.68	0.71	0.73	0.73	0.73	0.67	0.73	0.79	0.69	0.71	0.75
	0.79	**0.65**	0.68	0.69	0.72	0.73	0.71	0.67	0.70	0.77	0.68	0.70	**0.71**
Ridge Reg	1.12	0.95	1.05	1.48	1.56	0.86	0.95	1.22	1.01	1.26	1.16	1.40	1.17
	0.95	**0.75**	**0.77**	**0.77**	**0.79**	**0.77**	**0.75**	**0.72**	**0.83**	**0.90**	**0.87**	**0.80**	**0.80**
Factor	1.31	0.94	0.85	0.84	0.80	0.78	0.72	0.71	0.71	0.78	0.71	0.71	0.82
	0.81	**0.66**	**0.68**	**0.70**	**0.71**	**0.72**	0.70	**0.66**	0.70	0.78	0.69	0.71	0.71
Target Factor	1.30	0.88	0.78	0.79	0.76	0.74	0.71	0.68	0.70	0.79	0.70	0.72	0.79
	0.80	**0.66**	**0.68**	**0.70**	**0.71**	0.72	0.70	0.66	0.70	0.78	0.69	0.71	**0.71**
CSR	1.14	0.75	0.70	0.71	0.72	0.72	0.72	0.67	0.70	0.78	0.70	0.71	0.75
	0.80	**0.64**	0.66	0.69	0.71	0.72	0.70	0.67	0.70	0.77	0.69	0.71	**0.71**
Bagging	0.98	0.74	0.75	0.76	0.76	0.85	0.83	0.74	0.82	0.88	0.82	0.78	0.81
	0.89	**0.67**	**0.68**	**0.72**	0.75	**0.73**	**0.73**	0.71	**0.75**	0.86	0.80	0.75	**0.75**
Random Forests	**0.86**	0.70	0.71	0.76	0.76	0.79	0.75	0.69	0.74	0.82	0.75	0.73	0.76
	0.92	**0.64**	**0.67**	**0.71**	0.73	**0.74**	**0.70**	0.67	**0.70**	**0.77**	**0.69**	0.70	**0.72**

Note: The table shows the root mean squared error (RMSE) relative to the random walk's (RW) RMSE from predicting log difference of industrial production of 2016:01-2022:12 sample. White-colored rows concerns data detrended by transformations. Grey-colored rows concerns data detrended by linear projection. Each column shows RMSE ratio given a forecast horizon. The last column shows the average RMSE ratio across all forecasting horizons.

models. When predicting CPI and IP from 2016 to 2022, RMSE improvements can reach up to 53.2% and 49% respectively. Notably, the random forests method consistently delivers outstanding prediction performance among all machine learning methods and exhibits a consistent preference for $\hat{x}_t^{AD}$.

For the other models, using the automatic detrending method displays comparable performance across most model specifications. This suggests that our method is a reliable and efficient option for trend removal while preserving important information in the original data.

5 Conclusion

As machine learning methods have gained prominence in forecasting, an effective method for isolating stationary components from nonstationary raw data becomes essential. In this paper, I propose an automatic detrending method as a reliable and computationally efficient approach to remove nonstationarity in high-dimensional data. This method could generate data inputs that is statistically appealing for predicting macroeconomic variables with various machine learning methods. I show theoretically that LASSO and adaptive LASSO are able to recover the true model representation with cyclical estimates derived from the automatic detrending method. Our empirical evidence further supports that this detrending procedure results in informative predictors, achieving significantly low RMSE when predicting CPI and IP during period 2016-2022. This finding is robust to the specific machine learning method being used or the out-of-sample forecasting period.

Appendix

A Proofs

A.1 Proof of Theorem 1

This section first provides some useful lemmas that guide us to the results in Theorem 1. All inequalities hold element-by-element. $[\cdot]_{ij}$ denotes the element in the ith row and jth column of a matrix, and $[\cdot]_i$ denotes the ith element of a vector. We use $\boldsymbol{X}_j$ to denote the jth column of matrix $\boldsymbol{X}$. We define the $T \times n$ matrix $\boldsymbol{V} = \hat{\boldsymbol{X}} - \boldsymbol{X}$.

Lemma 1. *Under Condition DGP and Estimation,*

$$\frac{1}{T}\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}} \xrightarrow{p} \boldsymbol{\Omega}$$

Proof of Lemma 1. We write

$$\begin{aligned}
&\Big[\frac{1}{T}\sum_{t=1}^{T}\hat{x}_t\hat{x}_t'\Big]_{ij} \\
=&\frac{1}{T}\sum_{t=1}^{T}(x_{it}+v_{it})(x_{jt}+v_{jt}) \\
=&\frac{1}{T}\sum_{t=1}^{T}x_{it}x_{jt}+\frac{1}{T}\sum_{t=1}^{T}x_{it}v_{jt}+\frac{1}{T}\sum_{t=1}^{T}v_{it}x_{jt}+\frac{1}{T}\sum_{t=1}^{T}v_{it}v_{jt} \\
\leq&\frac{1}{T}\sum_{t=1}^{T}x_{it}x_{jt}+\Bigg\{\Big(\frac{1}{T}\sum_{t=1}^{T}x_{it}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{T}\sum_{t=1}^{T}v_{jt}^2\Big)^{\frac{1}{2}} \\
&+\Big(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{T}\sum_{t=1}^{T}x_{jt}^2\Big)^{\frac{1}{2}}+\Big(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{T}\sum_{t=1}^{T}v_{jt}^2\Big)^{\frac{1}{2}}\Bigg\} \xrightarrow{p} [\boldsymbol{\Omega}]_{ij}\,.
\end{aligned}$$

The convergence follows from DGP(i), DESIGN (ii) and ESTIMATION.

We have the similar convergence result for the lower bound:

$$\Big[\frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{x}}_t\hat{\boldsymbol{x}}_t'\Big]_{ij} \geq \frac{1}{T}\sum_{t=1}^{T}x_{it}x_{jt} - \Bigg\{\Big(\frac{1}{T}\sum_{t=1}^{T}x_{it}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{T}\sum_{t=1}^{T}v_{jt}^2\Big)^{\frac{1}{2}} + \Big(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{T}\sum_{t=1}^{T}x_{jt}^2\Big)^{\frac{1}{2}} + \Big(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{T}\sum_{t=1}^{T}v_{jt}^2\Big)^{\frac{1}{2}}\Bigg\} \xrightarrow{p} [\boldsymbol{\Omega}]_{ij}\,.$$

□

Lemma 2. *For any $\boldsymbol{\eta}' \in (\boldsymbol{\eta}, \mathbf{1})$,*

$$P\left\{|\hat{\boldsymbol{\Omega}}_{21}(\hat{\boldsymbol{\Omega}}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1))| \leq \mathbf{1} - \boldsymbol{\eta}'\right\} \to 1 \text{ as } T \to \infty.$$

Lemma 2 gives us an empirical version of Condition IC.

Proof of Lemma 2. By Lemma 1, we have $\hat{\boldsymbol{\Omega}}_{21} \xrightarrow{p} \boldsymbol{\Omega}_{21}$ and $\hat{\boldsymbol{\Omega}}_{11} \xrightarrow{p} \boldsymbol{\Omega}_{11}$. Hence,

$$\hat{\boldsymbol{\Omega}}_{21}(\hat{\boldsymbol{\Omega}}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1)) - \boldsymbol{\Omega}_{21}(\boldsymbol{\Omega}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1)) \xrightarrow{p} 0.$$

Therefore,

$$\begin{aligned}&|\hat{\boldsymbol{\Omega}}_{21}(\hat{\boldsymbol{\Omega}}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1))| - |\boldsymbol{\Omega}_{21}(\boldsymbol{\Omega}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1))| \\ \leq &|\hat{\boldsymbol{\Omega}}_{21}(\hat{\boldsymbol{\Omega}}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1)) - \boldsymbol{\Omega}_{21}(\boldsymbol{\Omega}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1))| \xrightarrow{p} 0\end{aligned}$$

Lastly,

$$\begin{aligned}&P(|\hat{\boldsymbol{\Omega}}_{21}(\hat{\boldsymbol{\Omega}}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1))| > \mathbf{1} - \boldsymbol{\eta}') \\ =&P(|\hat{\boldsymbol{\Omega}}_{21}(\hat{\boldsymbol{\Omega}}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1))| - |\boldsymbol{\Omega}_{21}(\boldsymbol{\Omega}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0(1))| > \boldsymbol{\eta} - \boldsymbol{\eta}') \to 0\end{aligned}$$

□

Lemma 3. *With probability converging to one,*

$$\left\{\mathcal{A}_T \cap \mathcal{B}_T\right\} \subseteq \left\{\boldsymbol{sign}(\hat{\boldsymbol{\beta}}^{LASSO}(\lambda_T)) = \boldsymbol{sign}(\boldsymbol{\beta}_0)\right\}$$

for

$$\mathcal{A}_T = \cap_{i=1}^{s}\{|(\hat{\mathbf{\Omega}}_{11})^{-1}\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}'(1)\hat{\boldsymbol{U}}|_i < \sqrt{T}|\boldsymbol{\beta}_{0,i}| - \frac{\lambda_T}{2\sqrt{T}}|(\hat{\mathbf{\Omega}}_{11})^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0)(1)|_i\},$$
$$\mathcal{B}_T = \cap_{i=1}^{n-s}\{|\hat{\mathbf{\Omega}}_{21}(\hat{\mathbf{\Omega}}_{11})^{-1}\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}'(1)\hat{\boldsymbol{U}} - \frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}'(2)\hat{\boldsymbol{U}}|_i \leq \frac{\lambda_T}{2\sqrt{T}}[\boldsymbol{\eta}']_i\},$$

where $\hat{\boldsymbol{U}} = \boldsymbol{Y} - \hat{\boldsymbol{X}}\boldsymbol{\beta_0} = \boldsymbol{U} + (\boldsymbol{X} - \hat{\boldsymbol{X}})\boldsymbol{\beta_0}$ *is the error term* $\boldsymbol{U}$ *plus the estimation error, and* $[\eta']_i$ *denotes the* i*th element of vector* $\boldsymbol{\eta'}$*.*

Proof of Lemma 3. The proof is a direct application of Zhao and Yu (2006). We replace $\boldsymbol{X}$ and $\boldsymbol{U}$ with $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{U}}$ in our application. □

Proof of Theorem 1. The folloing argument closely follows Theorem 1 in Zhao and Yu (2006).

$$\begin{aligned}
&1 - P(\mathcal{A}_T \cap \mathcal{B}_T)\\
\leq& P(\mathcal{A}_T^c) + P(\mathcal{B}_T^c)\\
\leq& \sum_{i=1}^{s} P\Big(\frac{1}{\sqrt{T}}\Big|\Big[\hat{\mathbf{\Omega}}_{\mathbf{11}}^{-1}\hat{\boldsymbol{X}}(1)'\hat{\boldsymbol{U}}\Big]_i\Big| \geq \sqrt{T}\Big|\beta_{0,i}\Big| - \frac{\lambda_T}{2\sqrt{T}}\Big|\Big[\hat{\mathbf{\Omega}}_{\mathbf{11}}^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_0)(1)\Big]_i\Big|\Big)\\
&+ \sum_{i=1}^{n-s} P\Big(|\hat{\mathbf{\Omega}}_{21}(\hat{\mathbf{\Omega}}_{11})^{-1}\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}'(1)\hat{\boldsymbol{U}} - \frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}'(2)\hat{\boldsymbol{U}}|_i > \frac{\lambda_T}{2\sqrt{T}}[\boldsymbol{\eta}']_i\Big)
\end{aligned}$$

By ESTIMATION and stationarity of $\{\boldsymbol{x}_t, u_t\}$, the ith element of $\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}(1)\hat{\boldsymbol{U}}$ is

$$\begin{aligned}
&\Big[\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}'(1)\hat{\boldsymbol{U}}\Big]_i\\
=&\Big[\frac{1}{\sqrt{T}}\boldsymbol{X}'(1)\boldsymbol{U}\Big]_i + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}u_{it} + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_{it}v_{it} + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}^2\\
\leq&\Big[\frac{1}{\sqrt{T}}\boldsymbol{X}'(1)\boldsymbol{U}\Big]_i + \Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_{it}^2\Big)^{\frac{1}{2}} + \Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_{it}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}^2\Big)^{\frac{1}{2}} + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}^2\\
=&\Big[\frac{1}{\sqrt{T}}\boldsymbol{X}'(1)\boldsymbol{U}\Big]_i + o_p(1)
\end{aligned}$$

Then by Lemma 1, we have the standard convergence results:

$$\hat{\mathbf{\Omega}}_{11}^{-1}\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}'(1)\hat{\boldsymbol{U}} \xrightarrow{d} N(0, \mathbf{\Omega}_{11}^{-1}\boldsymbol{V}_1)$$
$$\hat{\mathbf{\Omega}}_{21}(\hat{\mathbf{\Omega}}_{11})^{-1}\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}'(1)\hat{\boldsymbol{U}} - \frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}'(2)\hat{\boldsymbol{U}} \xrightarrow{d} N(0, \boldsymbol{V}_2)$$

for some matrices $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$. Lastly, because $\frac{\lambda_T}{T} \to 0$ and $\frac{\lambda_T}{T^{\frac{1+c}{2}}} \to \infty$, it is clear that $P(\mathcal{A}_T^c) + P(\mathcal{B}_T^c) \to 0$, which concludes the proof. □

A.2 Proof of Theorem 3

We start with some technical lemmas and propositions. The following lemma bounds the asymptotic difference in the eigenvalues of $\hat{\mathbf{\Omega}}$ and $\mathbf{\Omega}$.

Lemma 4. *Under condition DESIGN' and ESTIMATION',*

$$\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\hat{\mathbf{\Omega}}_{\mathbf{11}}\boldsymbol{\alpha} > \phi_{min} \quad w.p.a\ 1.$$

Proof of Lemma 4. First, we will show the following statement:

$$\max_{1\le i,j\le s_n} [|\hat{\mathbf{\Omega}}_{\mathbf{11}} - \mathbf{\Omega}_{\mathbf{11}}|]_{ij} \le \frac{\phi_{\min}}{s_n} \text{ w.p.a 1.} \tag{13}$$

Proving statement (13) uses the triangle inequality and Assumption DESIGN (iv):

$$\begin{aligned}\max_{1\le i,j\le s_n} [|\hat{\mathbf{\Omega}}_{\mathbf{11}} - \mathbf{\Omega}_{\mathbf{11}}|]_{ij} &\le \max_{1\le i,j\le s_n} [|\hat{\mathbf{\Omega}}_{\mathbf{11}} - \tilde{\mathbf{\Omega}}_{\mathbf{11}}|]_{ij} + \max_{1\le i,j\le s_n} [|\tilde{\mathbf{\Omega}}_{\mathbf{11}} - \mathbf{\Omega}_{\mathbf{11}}|]_{ij} \\ &\le \max_{1\le i,j\le s_n} [|\hat{\mathbf{\Omega}}_{\mathbf{11}} - \tilde{\mathbf{\Omega}}_{\mathbf{11}}|]_{ij} + \frac{\phi_{\min}}{s_n}.\end{aligned}$$

Write out the element at ith row and jth column of $\hat{\mathbf{\Omega}}_{\mathbf{11}}$:

$$\begin{aligned}
[\hat{\mathbf{\Omega}}_{\mathbf{11}}]_{ij} &= \frac{\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_j}{T} \\
&= \frac{1}{T}\sum_{t=1}^{T}\hat{x}_{it}\hat{x}_{jt} \\
&= \frac{1}{T}\sum_{t=1}^{T}[(x_{it}+v_{it})(x_{jt}+v_{jt})] \\
&= \frac{1}{T}\sum_{t=1}^{T}x_{it}x_{jt} + \frac{1}{T}\sum_{t=1}^{T}v_{it}x_{jt} + \frac{1}{T}\sum_{t=1}^{T}x_{it}v_{jt} + \frac{1}{T}\sum_{t=1}^{T}v_{it}v_{jt} \\
&= [\mathbf{\Omega}_{\mathbf{11}}]_{ij} + \frac{1}{T}\sum_{t=1}^{T}v_{it}x_{jt} + \frac{1}{T}\sum_{t=1}^{T}x_{it}v_{jt} + \frac{1}{T}\sum_{t=1}^{T}v_{it}v_{jt}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\max_{1\le i,j\le s_n}[|\hat{\mathbf{\Omega}}_{\mathbf{11}} - \mathbf{\Omega}_{\mathbf{11}}|]_{ij} \\
\le &\max_{1\le i,j\le s_n}\left\{\left|\frac{1}{T}\sum_{t=1}^{T}v_{it}x_{jt}\right| + \left|\frac{1}{T}\sum_{t=1}^{T}x_{it}v_{jt}\right| + \left|\frac{1}{T}\sum_{t=1}^{T}v_{it}v_{jt}\right|\right\} \\
\le &\max_{1\le i,j\le s_n}\left\{(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2)^{\frac{1}{2}}(\frac{1}{T}\sum_{t=1}^{T}x_{jt}^2)^{\frac{1}{2}} + (\frac{1}{T}\sum_{t=1}^{T}x_{it}^2)^{\frac{1}{2}}(\frac{1}{T}\sum_{t=1}^{T}v_{jt}^2)^{\frac{1}{2}} + (\frac{1}{T}\sum_{t=1}^{T}v_{it}^2)^{\frac{1}{2}}(\frac{1}{T}\sum_{t=1}^{T}v_{jt}^2)^{\frac{1}{2}}\right\} \\
\le &2\max_{1\le i\le s_n}(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2)^{\frac{1}{2}}\max_{1\le i\le s_n}(\frac{1}{T}\sum_{t=1}^{T}x_{it}^2)^{\frac{1}{2}} + \max_{1\le i\le s_n}(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2)
\end{aligned}$$

By condition REG and ESTIMATION', it can be shown that $\max_{1\le i\le s_n}\frac{1}{T}\sum_{t=1}^{T}v_{it}^2 \xrightarrow{p} 0$, and by Condition DGP' $\max_{1\le i\le s_n}\frac{1}{T}\sum_{t=1}^{T}x_{it}^2 = O_p(1)$. Therefore

$$\max_{1\le i\le n}[|\hat{\mathbf{\Omega}}_{\mathbf{11}} - \mathbf{\Omega}_{\mathbf{11}}|]_{ij} \xrightarrow{p} 0,$$

and this concludes our proof of (13).

Next, for any vector $\boldsymbol{\alpha} \in \mathcal{R}^s\backslash\{0\}$,

$$\begin{aligned}
\boldsymbol{\alpha}'\mathbf{\Omega}_{\mathbf{11}}\boldsymbol{\alpha} - \boldsymbol{\alpha}'\hat{\mathbf{\Omega}}_{\mathbf{11}}\boldsymbol{\alpha} \le |\boldsymbol{\alpha}'(\mathbf{\Omega}_{\mathbf{11}} - \hat{\mathbf{\Omega}}_{\mathbf{11}})\boldsymbol{\alpha}| &\le |\boldsymbol{\alpha}|_1|(\mathbf{\Omega}_{\mathbf{11}} - \hat{\mathbf{\Omega}}_{\mathbf{11}})\boldsymbol{\alpha}|_\infty \\
&\le |\boldsymbol{\alpha}|_1^2\frac{\phi_{min}}{s_n} \le s_n\boldsymbol{\alpha}'\boldsymbol{\alpha}\frac{\phi_{min}}{s_n} = |\boldsymbol{\alpha}|_2^2\phi_{min}.
\end{aligned}$$

where $|\cdot|_1$, $|\cdot|_2$ and $|\cdot|_\infty$ are l_1 norm, l_2 norm and supremum norm. Therefore,

$$\frac{\boldsymbol{\alpha}'\hat{\boldsymbol{\Omega}}_{\mathbf{11}}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\boldsymbol{\alpha}} \geq \frac{\boldsymbol{\alpha}'\boldsymbol{\Omega}_{\mathbf{11}}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\boldsymbol{\alpha}} - \phi_{min}.$$

We minimize both sides and use condition DESIGN (4) to get

$$\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\hat{\boldsymbol{\Omega}}_{\mathbf{11}}\boldsymbol{\alpha} > \inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\boldsymbol{\Omega}_{\mathbf{11}}\boldsymbol{\alpha} - \phi_{min} > \phi_{min} \text{ w.p.a 1.}$$

□

In the next part, we bound the term $T^{-\frac{1}{2}}\left|\sum_{t=1}^{T} x_{it}u_t\right|$.

Lemma 5. *Define the event* $\mathcal{E}_{n,T}(\lambda_0) = \left\{2\max_{i=1,\ldots,n} T^{-\frac{1}{2}}\left|\sum_{t=1}^{T} x_{it}u_t\right| < \lambda_0\right\}$. *Event* $\mathcal{E}_{n,T}(\lambda_{n,T}T^{-(1-a)/2})$ *has probability approaching one. More specifically,*

$$P\left(\max_{i=1,\ldots,n} T^{-\frac{1}{2}}\left|\sum_{t=1}^{T} x_{it}u_t\right| > \frac{\lambda_{n,T}}{2\sqrt{T}}T^{\frac{a}{2}}\right) \leq c\frac{n}{\lambda_{n,T}^{m}}T^{m(1-a)/2}$$

for some constant $c > 0$.

Proof. See Lemma 4 in Medeiros & Mendes (2016). □

Lemma 6. *Let* $\boldsymbol{W}(1) = diag(w_1, \ldots, w_s)$ *be the diagonal matrix consists of weights of active predictors, and* $\boldsymbol{\nu_0} = sign[\boldsymbol{\beta_0}(1)]$ *be the vector of signs of those active predictors. The probability that the adaptive LASSO estimator correctly selects the sign of all predictors has the following lower bound:*

$$P[\boldsymbol{sign}(\hat{\boldsymbol{\beta}}) = \boldsymbol{sign}(\boldsymbol{\beta_0})] \geq P(\mathcal{A}_{n,T} \cap \mathcal{B}_{n,T}),$$

where

$$\mathcal{A}_{n,T} = \cap_{i=1}^{s_n}\left\{\frac{1}{\sqrt{T}}\left|\left[\hat{\boldsymbol{\Omega}}_{\mathbf{11}}^{-1}\hat{\boldsymbol{X}}(1)'\hat{\boldsymbol{U}}\right]_i\right| < \sqrt{T}|\boldsymbol{\beta}_{\mathbf{0},i}| - \frac{\lambda}{2\sqrt{T}}\left|\left[\hat{\boldsymbol{\Omega}}_{\mathbf{11}}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu_0}\right]_i\right|\right\}$$

$$\mathcal{B}_{n,T} = \cap_{i=s_n+1}^{n}\left\{2\frac{1}{\sqrt{T}}\left|\hat{\boldsymbol{X}}_{\boldsymbol{i}}'\boldsymbol{M}(1)\hat{\boldsymbol{U}}\right| < \frac{1}{\sqrt{T}}\lambda\left[w_i - \left|T^{-1}\hat{\boldsymbol{X}}_{\boldsymbol{i}}'\hat{\boldsymbol{X}}(1)\hat{\boldsymbol{\Omega}}_{\mathbf{11}}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu_0}\right|\right]\right\},$$

where $\hat{\boldsymbol{U}} = \boldsymbol{Y} - \hat{\boldsymbol{X}}\boldsymbol{\beta_0} = \boldsymbol{U} + (\boldsymbol{X} - \hat{\boldsymbol{X}})\boldsymbol{\beta_0}$ *is the error term* $\boldsymbol{U}$ *plus the estimation error, and* $\boldsymbol{M}(1) = \boldsymbol{I_T} - \hat{\boldsymbol{X}}(1)(\hat{\boldsymbol{X}}(1)'\hat{\boldsymbol{X}}(1))^{-1}\hat{\boldsymbol{X}}(1)'$.

Lemma 6 is similar to Lemma 3 and incorporate the weight parameters in bounding events $\mathcal{A}_{n,T}$ and $\mathcal{B}_{n,T}$. The two events can roughly be described as "including all relevant predictors" and "excluding irrelevant predictors".

Proof. The proof is a direct application of Proposition 1 of Zhao and Yu (2006). □

Lastly, the following two lemmas relate the two events in Lemma 6 to the event in Lemma 5.

Lemma 7. *Assume DESIGN', WEIGHTS, and ESTIMATION' hold jointly, and that* $\beta_{\min} > \frac{\lambda_{n,T}}{T^{1-a/2}} \frac{s_n^{\frac{1}{2}}}{\phi_{\min}}$. *Then* $\mathcal{E}_{n,T}(\lambda_{n,T} T^{-(1-a)/2}) \subseteq \mathcal{A}_{n,T}$.

Proof of Lemma 7. By definition, we can write

$$\mathcal{A}_{n,T}^c = \cup_{i=1}^{s_n} \Big\{ \frac{1}{\sqrt{T}} \Big| \Big[\hat{\boldsymbol{\Omega}}_{\boldsymbol{11}}^{-1} \hat{\boldsymbol{X}}(1)' \hat{\boldsymbol{U}} \Big]_i \Big| \geq \sqrt{T} \big| \beta_{0,i} \big| - \frac{\lambda}{2\sqrt{T}} \Big| \Big[\hat{\boldsymbol{\Omega}}_{\boldsymbol{11}}^{-1} \boldsymbol{W}(1) \boldsymbol{\nu_0} \Big]_i \Big| \Big\}. \tag{14}$$

For the LHS of (14),

$$\begin{aligned} \frac{1}{\sqrt{T}} \Big| \Big[\hat{\boldsymbol{\Omega}}_{\boldsymbol{11}}^{-1} \hat{\boldsymbol{X}}(1)' \hat{\boldsymbol{U}} \Big]_i \Big| &\leq \Big(\inf_{\boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}' \hat{\boldsymbol{\Omega}}_{\boldsymbol{11}} \boldsymbol{\alpha} \Big)^{-1} \Big[\sum_{j=1}^{s_n} (T^{-1/2} \hat{\boldsymbol{X}}_j' \hat{\boldsymbol{U}})^2 \Big]^{1/2} \\ &\leq \phi_{\min}^{-1} \Big[\sum_{j=1}^{s_n} (T^{-1/2} \hat{\boldsymbol{X}}_j' \hat{\boldsymbol{U}})^2 \Big]^{1/2} \text{ w.p.a. 1.} \end{aligned}$$

Similarly, with the Cauchy-Schwarz inequality and Assumption WEIGHTS we can show that

$$\Big| \Big[\hat{\boldsymbol{\Omega}}_{\boldsymbol{11}}^{-1} \boldsymbol{W}(1) \boldsymbol{\nu_0} \Big]_i \Big| \leq \phi_{\min}^{-1} s_n^{\frac{1}{2}} w_{\max} \leq \frac{s_n^{\frac{1}{2}} T^{\frac{a}{2}}}{\phi_{\min}} \text{ w.p.a. 1.}$$

For $\beta_{\min} > \frac{\lambda}{T^{1-a/2}} \frac{s_n^{\frac{1}{2}}}{\phi_{\min}}$,

$$\sqrt{T} |\beta_{0,i}| - \frac{\lambda}{2\sqrt{T}} \Big| \Big[\hat{\boldsymbol{\Omega}}_{\boldsymbol{11}}^{-1} \boldsymbol{W}(1) \boldsymbol{\nu_0} \Big]_i \Big| \geq \frac{\lambda}{2\sqrt{T}} \frac{s_n^{\frac{1}{2}} T^{\frac{a}{2}}}{\phi_{\min}} \text{ w.p.a. 1.}$$

Put the above results together, we have

$$\begin{aligned}\mathcal{A}_{n,T}^c &\subseteq \left\{ \sum_{j=1}^{s_n} (T^{-1/2}\hat{\boldsymbol{X}}_j'\hat{\boldsymbol{U}})^2 \geq (\frac{\lambda_{n,T}}{2\sqrt{T}} {s_n}^{\frac{1}{2}} T^{\frac{a}{2}})^2 \right\} \\ &\subseteq \left\{ \max_{i=1,\dots s_n} 2|T^{-1/2}\hat{\boldsymbol{X}}_j'\hat{\boldsymbol{U}}| \geq \frac{\lambda_{n,T}}{2T^{(1-a)/2}} \right\} \\ &\subseteq \left\{ \max_{j=1,\dots s_n} 2|T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U}| \geq \frac{\lambda_{n,T}}{2T^{(1-a)/2}} \right. \\ &\quad \left. + T^{-1/2} \max_{j=1,\dots s_n} 2\left|\boldsymbol{X}_j'(\boldsymbol{X}-\hat{\boldsymbol{X}})\boldsymbol{\beta_0} + \boldsymbol{V}_j'\boldsymbol{U} + \boldsymbol{V}_j'(\boldsymbol{X}-\hat{\boldsymbol{X}})\boldsymbol{\beta_0}\right| \right\} \text{ w.p.a 1,}\end{aligned}$$

where the $T \times 1$ vector $\boldsymbol{V}_j = \hat{\boldsymbol{X}}_j - \boldsymbol{X}_j$ is $T \times 1$ is the jth column of matrix $\boldsymbol{V} = \hat{\boldsymbol{X}} - \boldsymbol{X}$.

Recall that

$$\mathcal{E}_{n,T}^c(\lambda_{n,T}T^{-(1-a)/2}) = \left\{ \max_{i=1,\dots,n} T^{-\frac{1}{2}} \Big| \sum_{t=1}^{T} x_{it}u_t \Big| > \frac{\lambda_{n,T}}{2\sqrt{T}} T^{\frac{a}{2}} \right\}.$$

To show that $\mathcal{A}_{n,T}^c \subseteq \mathcal{E}_{n,T}^c(\lambda_{n,T}T^{-(1-a)/2})$ w.p.a 1, it suffices to show that

$$\begin{aligned}&T^{-1/2} \max_{j=1,\dots s_n} \left|\boldsymbol{X}_j'(\boldsymbol{X}-\hat{\boldsymbol{X}})\boldsymbol{\beta_0} + \boldsymbol{V}_j'\boldsymbol{U} + \boldsymbol{V}_j'(\boldsymbol{X}-\hat{\boldsymbol{X}})\boldsymbol{\beta_0}\right| \\ \leq & T^{-1/2}\Big\{ \max_{j=1,\dots s_n} \left|\boldsymbol{X}_j'(\boldsymbol{X}-\hat{\boldsymbol{X}})\boldsymbol{\beta_0}\right| + \max_{j=1,\dots s_n} \left|\boldsymbol{V}_j'\boldsymbol{U}\right| + \max_{j=1,\dots s_n} \left|\boldsymbol{V}_j'(\boldsymbol{X}-\hat{\boldsymbol{X}})\boldsymbol{\beta_0}\right| \Big\} \\ = & o_p(\frac{\lambda_{n,T}}{T^{(1-a)/2}}).\end{aligned}$$

We write the first term as

$$\begin{aligned}\max_{j=1,\dots s_n} \Big|\frac{1}{\sqrt{T}}\boldsymbol{X}_j'(\boldsymbol{X}-\hat{\boldsymbol{X}})\boldsymbol{\beta_0}\Big| &= \max_{j=1,\dots s_n} \Big|\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_{jt} \sum_{i=1}^{n} v_{it}\beta_i\Big| \\ &\leq \max_{j=1,\dots s_n} \Big(\frac{1}{T}\sum_{t=1}^{T} x_{jt}^2\Big)^{\frac{1}{2}} \Big|\beta_{max}\Big| \Big(\sum_{t=1}^{T}(\sum_{i=1}^{s_n} v_{it})^2\Big)^{\frac{1}{2}} \\ &= o_p\Big(\frac{\lambda_{n,T}}{T^{(1-a)/2}}\Big),\end{aligned}$$

where $\max_{j=1,\dots s_n} \frac{1}{T}\sum_{t=1}^{T} x_{jt}^2 = O_p(1)$ by Assumption DGP'.

Similarly, for the second term we have

$$\begin{aligned}\max_{j=1,\dots s_n}\Big|\frac{1}{\sqrt{T}}\boldsymbol{V}_j'\boldsymbol{U}\Big| &= \max_{j=1,\dots s_n}\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{jt}u_t\Big| \\ &\le \max_{j=1,\dots s_n}\Big(\sum_{t=1}^{T} v_{jt}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{T}\sum_{t=1}^{T} u_t^2\Big)^{\frac{1}{2}} \\ &= o_p\Big(\frac{\lambda_{n,T}}{T^{(1-a)/2}}\Big)\end{aligned}$$

Lastly, for the third term,

$$\begin{aligned}\max_{j=1,\dots s_n}\Big|\frac{1}{\sqrt{T}}\boldsymbol{V}_j'(\boldsymbol{X}-\hat{\boldsymbol{X}})\boldsymbol{\beta_0}\Big| &= \max_{j=1,\dots s_n}\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{jt}\sum_{i=1}^{n} v_{it}\beta_{0,i}\Big| \\ &\le \max_{j=1,\dots s_n}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{jt}^2\Big)^{\frac{1}{2}}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\big(\sum_{i=1}^{s_n} v_{it}\beta_{0,i}\big)^2\Big)^{\frac{1}{2}} \\ &\le \max_{j=1,\dots s_n}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{jt}^2\Big)^{\frac{1}{2}}\Big|\beta_{max}\Big|\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\big(\sum_{i=1}^{s_n} v_{it}\big)^2\Big)^{\frac{1}{2}} \\ &= o_p\Big(\frac{\lambda_{n,T}}{T^{(2-a)/2}}\Big).\end{aligned}$$

Therefore,

$$\mathcal{A}_{n,T}^c \subseteq \left\{\max_{j=1,\dots s} 2|T^{-1/2}\boldsymbol{X}_j'\boldsymbol{U}| \ge \frac{\lambda_{n,T}}{2T^{(1-a)/2}}\right\} \text{ w.p.a 1,}$$

which concludes that $\mathcal{E}_{n,T}(\lambda_{n,T}T^{-(1-a)/2}) \subseteq \mathcal{A}_{n,T}$. □

Lemma 8. *If DGP', DESIGN', WEIGHTS, and ESTIMATION' hold jointly, then* $\mathcal{E}_{n,T}(\lambda_{n,T}T^{-\frac{1-a}{2}}) \subseteq \mathcal{B}_{n,T}$ *w.p.a. 1.*

Proof of Lemma 8.

$$\mathcal{B}_{n,T}^c = \cup_{i=s_n+1}^{n}\left\{2\Big|\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}_i'\boldsymbol{M}(1)\hat{\boldsymbol{U}}\Big| \ge \frac{1}{\sqrt{T}}\lambda_{n,T}\Big[w_i - \Big|T^{-1}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}(1)\hat{\boldsymbol{\Omega}}_{\boldsymbol{11}}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu_0}\Big|\Big]\right\}. \quad (15)$$

We plug in $\boldsymbol{M}(1)$ to the LHS of (15):

$$\begin{aligned}\hat{\boldsymbol{X}}_i'\boldsymbol{M}(1)\hat{\boldsymbol{U}} &= \hat{\boldsymbol{X}}_i'\hat{\boldsymbol{U}} - \hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i(1)[\hat{\boldsymbol{X}}_i(1)'\hat{\boldsymbol{X}}_i(1)]^{-1}\hat{\boldsymbol{X}}_i(1)'\hat{\boldsymbol{U}} \\ &:= \hat{\boldsymbol{A}}_i + \hat{\boldsymbol{B}}_i.\end{aligned}$$

We leave $\hat{\boldsymbol{A}}_i$ as it is. Apply the Cauchy-Schwarz inequality to $\hat{\boldsymbol{B}}_i$ to get

$$|\hat{\boldsymbol{B}}_i| \leq \big(\sum_{t=1}^{T}\hat{x}_{it}^2\big)^{\frac{1}{2}}|\hat{\boldsymbol{U}}'\hat{\boldsymbol{X}}_i(1)[\hat{\boldsymbol{X}}_i(1)'\hat{\boldsymbol{X}}_i(1)]^{-1}\hat{\boldsymbol{X}}_i(1)'\hat{\boldsymbol{U}}|^{\frac{1}{2}},$$

where

$$\begin{aligned}\Big|\sum_{t=1}^{T}\hat{x}_{it}^2\Big|^{\frac{1}{2}} &= \sqrt{T}\Big|\frac{1}{T}\sum_{t=1}^{T}x_{it}^2 + \frac{2}{T}\sum_{t=1}^{T}x_{it}v_{it} + \frac{1}{T}\sum_{t=1}^{T}v_{it}^2\Big|^{\frac{1}{2}} \\ &\leq \sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}x_{it}^2 + \frac{1}{T}\sum_{t=1}^{T}(x_{it}^2+v_{it}^2) + \frac{1}{T}\sum_{t=1}^{T}v_{it}^2\right)^{\frac{1}{2}} \\ &= \sqrt{T}\left(\frac{2}{T}\sum_{t=1}^{T}(x_{it}^2-1) + 2 + \frac{2}{T}\sum_{t=1}^{T}v_{it}^2\right)^{\frac{1}{2}} \\ &= 2\sqrt{T} + o_p(\sqrt{T}) + o_p(\lambda_{n,T}T^{\frac{a-1}{2}}),\end{aligned}$$

and

$$\Big|\hat{\boldsymbol{U}}'\hat{\boldsymbol{X}}_i(1)\Big[\hat{\boldsymbol{X}}_i(1)'\hat{\boldsymbol{X}}_i(1)\Big]^{-1}\hat{\boldsymbol{X}}_i(1)'\hat{\boldsymbol{U}}\Big|^{\frac{1}{2}} \leq \left[\frac{\sum_{i=1}^{s}(\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{U}})^2}{\phi_{\min}}\right]^{\frac{1}{2}}$$

by Assumption DESIGN'. Therefore, w.p.a. 1

$$\begin{aligned}|\hat{\boldsymbol{B}}_i| &\leq \Big(2\sqrt{T} + o_p(\lambda_{n,T}T^{\frac{a-1}{2}})\Big)\left[\frac{\sum_{i=1}^{s}\Big(\frac{1}{\sqrt{T}}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{U}}\Big)^2}{\phi_{\min}}\right]^{\frac{1}{2}} \\ &\leq \frac{(2+o_p(\lambda T^{\frac{a-2}{2}}))s^{\frac{1}{2}}}{\phi_{\min}^{\frac{1}{2}}}\left(\max_{i=1,\ldots,s_n}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{U}}\right).\end{aligned}$$

For the RHS of (15):

$$\begin{aligned}\left[T^{-1}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}(1)\hat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\boldsymbol{\nu_0}\right]^2 &= \left\{T^{-1}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}(1)[T^{-1}\hat{\boldsymbol{X}}(1)'\hat{\boldsymbol{X}}(1)]^{-1}\boldsymbol{W}(1)\boldsymbol{\nu_0}\right\}^2 \\ &\leq \boldsymbol{\nu_0}'\boldsymbol{W}(1)[T^{-1}\hat{\boldsymbol{X}}(1)'\hat{\boldsymbol{X}}(1)]^{-1}\boldsymbol{W}(1)\boldsymbol{\nu_0} \times T^{-1}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}_i \\ &\leq \frac{\sum_{j=1}^{s_n} w_i^2}{\inf_{\alpha'\alpha=1}\alpha'\hat{\Omega}_{11}\alpha}\frac{\sum_{t=1}^{T}\hat{x}_{it}^2}{T} \\ &\leq \frac{s_n w_{\max}^2}{\phi_{\min}}\frac{\sum_{t=1}^{T}\{2x_{it}^2+2v_{it}^2\}}{T} \text{ w.p.a 1} \\ &\leq \frac{s_n T^a}{\phi_{\min}}(2+o_p(\lambda_{n,T}^2 T^{a-2})) \text{ w.p.a 1} \\ &= O_p(\frac{s_n T^a}{\phi_{\min}}).\end{aligned}$$

Note that the final equality follows from Condition REG:

$$\lambda_{n,T}^2 T^{a-2} = o_p(T^{a-1}\frac{\phi_{\min}}{s_n w_{\max}^2}),$$

where $a<1$, $\phi_{\min}$ is non-increasing, and s and $w_{\max}$ are non-decreasing. Together with condition WEIGHT, the RHS of (15) has the following bound

$$\begin{aligned}&\frac{1}{\sqrt{T}}\lambda_{n,T}\left[w_i - \left|T^{-1}\hat{\boldsymbol{X}}_i'\hat{\boldsymbol{X}}(1)\hat{\boldsymbol{\Omega}}_{11}^{-1}\boldsymbol{W}(1)\nu_0\right|\right] \\ \geq &\frac{1}{\sqrt{T}}\lambda_{n,T}\left[T^{\frac{a}{2}} - O_p\left(\frac{{s_n}^{\frac{1}{2}}T^{\frac{a}{2}}}{\phi_{\min}^{\frac{1}{2}}}\right)\right] \\ = &O_p\left(\frac{\lambda_{n,T}}{\sqrt{T}}\sqrt{\frac{s_n T^a}{\phi_{min}}}\right).\end{aligned}$$

Hence,

$$\begin{aligned}\mathcal{B}_T^c &\subseteq \left\{ \max_{i=1,\dots,s_n} \frac{2}{\sqrt{T}} \hat{\boldsymbol{X}}_i' \hat{\boldsymbol{U}} \Big(1 + \frac{(2 + o_p(\lambda_{n,T} T^{\frac{a-2}{2}})) s_n^{\frac{1}{2}}}{\phi_{\min}^{\frac{1}{2}}}\Big) \geq O_p\left(\frac{\lambda_{n,T}}{\sqrt{T}} \sqrt{\frac{s_n T^a}{\phi_{min}}} \right) \right\} \\ &= \left\{ \max_{i=1,\dots,s} \frac{2}{\sqrt{T}} \hat{\boldsymbol{X}}_i' \hat{\boldsymbol{U}} O_p\left(\frac{s_n^{\frac{1}{2}}}{\phi_{\min}^{\frac{1}{2}}} \right) \geq O_p\left(\frac{\lambda_{n,T}}{\sqrt{T}} \sqrt{\frac{s_n T^a}{\phi_{min}}} \right) \right\} \quad \text{w.p.a 1} \\ &\subseteq \left\{ \max_{i=1,\dots,s_n} \frac{2}{\sqrt{T}} \hat{\boldsymbol{X}}_i' \hat{\boldsymbol{U}} \geq \frac{\lambda_{n,T}}{2T^{(1-a)/2}} \right\} \quad \text{w.p.a 1.}\end{aligned}$$

Lastly, we use the same argument as Lemma 7 to conclude that

$$\mathcal{B}_T{}^c \subseteq \left\{ \max_{i=1,\dots,s_n} \frac{1}{\sqrt{T}} \boldsymbol{X}_i' \boldsymbol{U} \geq \frac{\lambda_{n,T}}{2T^{(1-a)/2}} \right\} \quad \text{w.p.a 1.}$$

□

Proof of Theorem (3). By Lemma 7 and Lemma 8, w.p.a. 1

$$\mathcal{E}_{n,T}(\lambda_{n,T} T^{-\frac{1-a}{2}}) \subseteq \mathcal{A}_{n,T} \cap \mathcal{B}_{n,T}.$$

Together with Lemma 5 and Lemma 6, we have

$$\begin{aligned}Pr[\boldsymbol{sign}(\hat{\boldsymbol{\beta}}^{adalasso}) = \boldsymbol{sign}(\boldsymbol{\beta_0})] &\geq P(\mathcal{A}_{n,T} \cap \mathcal{B}_{n,T}) \\ &\geq P[\mathcal{E}_{n,T}(\lambda_{n,T} T^{-(1-a)/2})] \to 1.\end{aligned}$$

□

B Robustness Check

We conduct a robustness check for the forecasting exercise in Section 4.3.2. In particular, we estimate the coefficients in equation (2) with sample from 1960:01 to 2015:12. This set of coefficients was then used to detrend all observations from 1960:01 to 2022:12. This forecasting exercise intends to use no data of the out-of-sample forecasting period to conduct the detrending procedure, including the upcoming COVID shock. Because this exercise modifies the calculation of the detrending coefficients, only results displayed in the grey-colored cells, i.e., those concerns the automatic detrending method, differs from the results in Table 3 and 4.

In short, we find no significant changes to our all main takeaways in Section 4.3.2. Table 5 and Table 6 display the forecasting results of CPI and IP during 2016:01 to 2022:12 with the new detrending coefficients. We find that the performances of LASSO and adaptive LASSO are slightly worse when h is below 6 months, although there are also some improvement for adaptive LASSO when h is longer than 10 months. On the contrary, the bagging method performs slightly worse when h is longer than 10 months.

Table 5: Forecasting errors for CPI from 2016 to 2022

	forecasting horizon												
	1	2	3	4	5	6	7	8	9	10	11	12	Average
RW	1	1	1	1	1	1	1	1	1	1	1	1	1
AR	0.89	0.79	0.78	0.77	0.79	0.80	0.83	0.84	0.82	0.87	0.85	078	0.82
Lasso	**0.84**	0.83	**0.77**	**0.73**	**0.77**	0.76	0.79	0.80	**0.77**	0.84	0.82	0.78	0.79
	0.90	**0.78**	0.81	0.79	0.81	0.76	0.78	0.84	0.83	0.86	0.83	0.76	0.81
adaLasso	0.92	0.93	**0.83**	**0.76**	**0.78**	0.79	0.80	**0.82**	0.83	0.87	0.98	0.91	0.85
	0.89	**0.81**	0.82	0.78	0.82	0.79	0.80	0.87	0.85	0.86	0.82	0.76	0.82
Ridge Reg	1.46	1.46	1.16	1.91	2.04	1.67	2.22	1.83	1.70	2.00	1.75	1.68	1.74
	0.92	**1.07**	**1.01**	**1.13**	**1.28**	**1.12**	**1.17**	**1.02**	**0.97**	**0.98**	**1.06**	**0.96**	**1.06**
Factor	1.14	0.92	1.02	0.95	0.83	0.84	0.85	0.88	0.82	0.84	0.85	0.97	0.91
	0.92	**0.85**	**0.78**	**0.79**	0.81	**0.79**	0.81	**0.82**	0.81	0.85	**0.80**	**0.74**	**0.81**
Target Factor	0.98	0.85	0.90	0.86	0.80	0.84	0.89	1.18	0.89	0.90	0.93	0.95	0.91
	0.92	0.85	**0.78**	**0.79**	0.81	**0.79**	**0.81**	**0.82**	**0.81**	**0.85**	**0.80**	**0.74**	0.81
CSR	**0.86**	**0.75**	**0.76**	**0.75**	**0.79**	**0.78**	**0.82**	0.83	0.81	**0.86**	**0.80**	**0.80**	**0.80**
	1.02	0.86	0.87	0.87	0.90	0.87	0.91	0.87	0.81	0.90	0.92	0.86	0.89
Bagging	**0.81**	**1.12**	1.15	1.51	1.47	1.07	**1.15**	1.11	1.04	1.02	**0.98**	1.03	1.12
	1.07	1.31	**0.98**	**1.03**	**1.18**	1.15	1.32	1.15	**0.90**	0.98	1.04	1.02	1.10
Random Forests	0.85	0.83	0.84	0.80	0.82	0.79	0.82	0.84	0.83	0.90	0.87	0.83	0.83
	0.88	**0.74**	**0.72**	**0.72**	**0.75**	**0.75**	**0.78**	**0.78**	**0.75**	**0.80**	**0.74**	**0.68**	**0.76**

Note: The table shows the root mean squared error (RMSE) relative to the random walk's (RW) RMSE from predicting inflation rate of 2016:01-2022:12 sample. White-colored rows concerns data detrended by transformations. Grey-colored rows concerns data detrended by linear projection. Each column shows RMSE ratio given a forecasting horizon. The last column shows the average RMSE ratio across all forecast horizons.

Table 6: Forecasting errors for IP from 2016 to 2022

	forecasting horizon												
	1	2	3	4	5	6	7	8	9	10	11	12	Average
RW	1	1	1	1	1	1	1	1	1	1	1	1	1
AR	0.87	0.73	0.71	0.71	0.72	0.73	0.71	0.67	0.71	0.77	0.70	0.71	0.73
Lasso	0.88	0.67	0.66	0.70	0.72	0.72	0.71	0.67	0.71	0.78	0.69	0.71	0.71
	0.88	0.65	0.67	0.69	0.72	0.72	0.71	0.67	0.70	0.78	0.68	0.70	0.71
adaLasso	1.04	0.78	0.68	0.71	0.73	0.73	0.73	0.67	0.73	0.79	0.69	0.71	0.75
	0.84	**0.67**	0.68	0.70	0.72	0.73	0.71	0.67	0.70	0.77	0.68	0.70	0.71
Ridge Reg	1.12	0.95	1.05	1.48	1.56	0.86	0.95	1.22	1.01	1.26	1.16	1.40	1.17
	1.15	**0.85**	**0.79**	**0.76**	**0.84**	**0.80**	**0.78**	**0.75**	**0.87**	**1.14**	**1.00**	**0.90**	**0.88**
Factor	1.31	0.94	0.85	0.84	0.80	0.78	0.72	0.71	0.71	0.78	0.71	0.71	0.82
	0.81	0.67	0.69	**0.70**	**0.71**	**0.72**	0.70	0.66	0.70	0.77	0.69	0.71	0.71
Target Factor	1.30	0.88	0.78	0.79	0.76	0.74	0.71	0.68	0.70	0.79	0.70	0.72	0.79
	0.81	**0.67**	**0.69**	**0.70**	**0.71**	0.72	0.70	0.66	0.70	0.77	0.69	0.71	**0.71**
CSR	1.14	0.75	0.70	0.71	0.72	0.72	0.72	0.67	0.70	0.78	0.70	0.71	0.75
	0.83	**0.66**	**0.67**	0.69	0.71	0.72	0.70	0.67	0.70	0.77	0.68	0.70	**0.71**
Bagging	0.98	0.74	0.75	0.76	0.76	0.85	0.83	0.74	0.82	0.88	0.82	0.78	0.81
	0.94	**0.68**	**0.69**	0.74	0.74	**0.76**	**0.73**	**0.70**	**0.75**	**0.83**	**0.74**	0.75	**0.76**
Random Forests	**0.86**	0.70	0.71	0.76	0.76	0.79	0.75	0.69	0.74	0.82	0.75	0.73	0.76
	0.92	**0.64**	**0.66**	**0.69**	**0.71**	**0.71**	**0.70**	0.67	**0.71**	**0.77**	**0.69**	0.74	**0.72**

Note: The table shows the root mean squared error (RMSE) relative to the random walk's (RW) RMSE from predicting log difference of industrial production of 2016:01-2022:12 sample. White-colored rows concerns data detrended by transformations. Grey-colored rows concerns data detrended by linear projection. Each column shows RMSE ratio given a forecast horizon. The last column shows the average RMSE ratio across all forecasting horizons.

References

Atkeson, A. E. and Ohanian, L. (2001). Are phillips curves useful for forecasting inflation? *Quarterly review - Federal Reserve Bank of Minneapolis*, 25(1):2–11.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Beveridge, S. and Nelson, C. R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle. *Journal of Monetary Economics*, 7:151–174.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Elliott, G., Gargano, A., and Timmermann, A. (2015). Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics Control*, 54:86–110.

Fisher, J. D., Liu, C. T., and Zhou, R. (2002). When can we forecast inflation? *Economic Perspectives*, 26(1):32–44.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Jounal of Applied Econometrics*, 37(5):920–964.

Hamilton, J. D. (2018). Why you should never use the hodrick-prescott filter. *The Review of Economics and Statistics*, 100(5):831–843.

Hamilton, J. D. and Xi, J. (2023). Principal component analysis for nonstationary series. Working paper, University of California at San Diego.

Inoue, A. and Kilian, L. (2008). How useful is bagging in forecasting economic time series? a case study of u.s. consumer price inflation. *Journal of the American Statistical Association*, 103(482):511–522.

Koo, B., Anderson, H. M., Seo, M. H., and Yao, W. (2020). High-dimensional predictive regression in the presence of cointegration. *Journal of Econometrics*, 219(2):456–477.

Lee, J. H., Shi, Z., and Gao, Z. (2020). On lasso for predictive regression. *Journal of Econometrics*, 229(2):322–349.

McCracken, M. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business Economic Statistics*, 34(4):574–589.

Medeiros, M. C. and Mendes, E. F. (2016). l_1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, 191:255–271.

Medeiros, M. C. and Vasconcelos, G. F. (2016). Forecasting macroeconomic variables in data-rich environments. *Economics Letters*, 138:50–52.

Medeiros, M. C., Vasconcelos, G. F., Veiga, , and Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business Economic Statistics*, 39(1):98–119.

Meinshausen, N. and Bühlmann, P. (2006). Consistent neighbourhood selection for high-dimensional graphs with the lasso. *Annals of Statistics*, 34(3):1436–1462.

Nardi, Y. and Rinaldo, A. (2011). Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, 102(3):528–549.

Onatski, A. and Wang, C. (2021). Spurious factor analysis. *Econometrica*, 89(2):591–614.

Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44:293–335.

Stock, J. H. and Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39(1):3–33.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, New York.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.