Essays in High Dimensional Time Series Analysis

Kashif Yousuf

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Essays in High Dimensional Time Series Analysis

Kashif Yousuf

Due to the rapid improvements in the information technology, high dimensional time series datasets are frequently encountered in a variety of fields such as macroeconomics, finance, neuroscience, and meteorology. Some examples in economics and finance include forecasting low frequency macroeconomic indicators, such as GDP or inflation rate, or financial asset returns using a large number of macroeconomic and financial time series and their lags as possible covariates. In these settings, the number of candidate predictors (p_T) can be much larger than the number of samples (T), and accurate estimation and prediction is made possible by relying on some form of dimension reduction. Given this ubiquity of time series data, it is surprising that few works on high dimensional statistics discuss the time series setting, and even fewer works have developed methods which utilize the unique features of time series data. This chapter consists of three chapters, and each one is self contained.

The first chapter deals with high dimensional predictive regressions which are widely used in economics and finance. However, the theory and methodology is mainly developed assuming that the model is stationary with time invariant parameters. This is at odds with the prevalent evidence for parameter instability in economic time series. To remedy this, we present two L_2 boosting algorithms for estimating high dimensional models in which the coefficients are modeled as functions evolving smoothly over time and the predictors are locally stationary. The first method uses componentwise local constant estimators as base learner, while the second relies on componentwise local linear estimators. We establish consistency of both methods, and address the practical issues of choosing the bandwidth for the base learners and

the number of boosting iterations. In an extensive application to macroeconomic forecasting with many potential predictors, we find that the benefits to modeling time variation are substantial and are present across a wide range of economic series. Furthermore, these benefits increase with the forecast horizon and with the length of the time series available for estimation. This chapter is jointly written with Serena Ng.

The second chapter deals with high dimensional non-linear time series models, and deals with the topic of variable screening/targeting predictors. Rather than assume a specific parametric model a priori, this chapter introduces several model free screening methods based on the partial distance correlation and developed specifically to deal with time dependent data. Methods are developed both for univariate models, such as nonlinear autoregressive models with exogenous predictors (NARX), and multivariate models such as linear or nonlinear VAR models. Sure screening properties are proved for our methods, which depend on the moment conditions, and the strength of dependence in the response and covariate processes, amongst other factors. Finite sample performance of our methods is shown through extensive simulation studies, and we show the effectiveness of our algorithms at forecasting US market returns. This chapter is jointly written with Yang Feng.

The third chapter deals with variable selection for high dimensional linear stationary time series models. This chapter analyzes the theoretical properties of Sure Independence Screening (SIS), and its two stage combination with the adaptive Lasso, for high dimensional linear models with dependent and/or heavy tailed covariates and errors. We also introduce a generalized least squares screening (GLSS) procedure which utilizes the serial correlation present in the data. By utilizing this serial correlation when estimating our marginal effects, GLSS is shown to outperform SIS in many cases. For both procedures we prove two stage variable selection consistency

when combined with the adaptive Lasso.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank a number of key individuals whose thoughtful guidance and advising supported the completion of this dissertation. First, I would like to acknowledge the training and feedback provided by my advisor Yang Feng. I would also like to thank Serena Ng, who has unofficially been a second advisor to me, for all the intellectual support and direction on the application of my work to economic contexts. Without their patience and guidance this thesis would not have been completed. I would also like to thank Professors Cindy Rush, Victor De La Pena, and Qingfeng Liu for agreeing to be a part of my commitee.

This dissertation also benefited from countless colleagues in the Department of Statistics at Columbia. I would like to thank Florian Stebbegg, Jon Auerbach, Adji Dieng, Wenda Zhou and others for their friendship and intellectual discussions motivating some of the completion of this work. I would also like to thank Professor Arian Maleki for his friendship and academic support. I had the opportunity to interact with many individuals across our department - from administrative staff such as Dood and Anthony who were able to handle any troubles encountered along the way, to professors, all of whom played an important role in the growth of my career and education.

Along the way my support system and many friendships have sustained the energy which facilitated the completion of this PhD. I would also like to thank my parents for their support throughout my life. Their positivity and faith in me has helped me

push through some of the difficult roadblocks one endures in the research process. This work could not have been written without their blessings and encouragement.

To My Family.

Chapter 1

Boosting High Dimensional Predictive Regressions with Time Varying Parameters

1.1 Introduction

Due to the rapid improvements in the information technology, high dimensional time series datasets are frequently encountered in a variety of fields in economics and finance (see Fan et al. (2011c); Shapiro (2017) for examples). In these settings, the number of candidate predictors (p_T) is much larger than the number of samples (T), and accurate estimation and prediction is made possible by relying on some form of dimension reduction. Ng (2013) puts the methods used in high dimension predictive regressions into two classes: a dense class which assumes that the covariates have a low rank representation that can be exploited for subsequent modeling, and a sparse class which assumes that the number of relevant predictors is far smaller than the number of predictors available. Research within the first class usually assumes a linear latent factor model which is estimated by principal components or partial least squares.[1] The second class treats the problem as one of variable selection in high dimension. Prominent methods in this class include screening, penalized likelihood,

[1] Stock and Watson (2002b); Bai and Ng (2002) and Kelly and Pruitt (2015).

lasso, and boosting methods.

This paper contributes to the literature in the second class. A key assumption made in the vast majority of works on sparse modeling is of a stationary underlying model with time invariant parameters.[2] The assumption is very restrictive in practice, as empirical evidence of parameter instability and time varying effects have been well documented in macroeconomics.[3] Parameter instability can be driven by structural changes in technological advancements, government or monetary policy changes, and preference shifts at the individual level (Chen and Hong, 2012). Ignoring these instabilities can lead to large forecasting errors, with Clements and Hendry (1996) and others even arguing that these instabilities are the main source of error for forecasting models.

Consider a high dimensional linear time varying parameter (TVP) model:

$$Y_t = \boldsymbol{\beta}_t \boldsymbol{x}_{t-h} + \epsilon_t \text{ for } t = 1, \ldots, T, \tag{1.1}$$

where Y_t is the response, $\boldsymbol{x}_{t-h} = (X_{1,t-h}, \ldots, X_{p_T,t-h})$ is a p_T-dimensional vector of predictors (with $p_T >> T$), $\boldsymbol{\beta} = (\beta_{1,t}, \ldots, \beta_{p_T,t})$ is a vector of time varying parameters, and ϵ_t are errors; the precise assumptions on the model will be stated in section 1.3. Given the evidence for parameter instability, the question remains on how to best represent and model this change, especially when dealing with high dimensional predictors. Parameter instability is most commonly represented in the econometrics

[2]Examples include Medeiros and Mendes (2016), Kock and Callot (2015), Han and Tsay (2017), and Basu and Michailidis (2015) which focus on the Lasso or the adaptive Lasso, and Lutz and Bühlmann (2006) which focuses on L_2 boosting for stationary VAR models.

[3]See (Stock and Watson, 1996; Rossi, 2013; Hamilton, 1989), asset pricing (Goyal and Welch, 2003; Paye and Timmermann, 2006; Rapach et al., 2010; Dangl and Halling, 2012), and exchange rate prediction (Schinasi and Swamy, 1989).

literature by random walks or by one or more discrete structural breaks.[4] Modeling variations by random walks can be quite restrictive as it imposes a specific structure on the evolution of the parameters. Discrete breaks require knowledge of the break dates, and not all time variations are well characterized by discrete shifts. Technology and taste shifts are arguably evolving slowly over time. Smooth transition models as in Terasvirta (1994) are still tightly parameterized. Furthemore, these methods are mainly designed for a fixed p_T. A third approach is to use rolling-window estimation to capture the smooth change in the parameters. As will soon be clear, rolling-window estimation is a special case of our proposed approach with a particular choice of kernel and bandwidth.

In this paper, we model these high-dimensional parameters as smooth functions of time whose functional forms are unknown and are estimated non-parametrically. We present two L_2 boosting algorithms which differ in their choice of base learners; the first uses componentwise local constant estimators as base learners, while the second relies on componentwise local linear estimators as base learners. We consider the use of local linear estimators since they have been shown to be a superior estimator theoretically, with smaller asymptotic bias at the boundaries of the sample (Cai, 2007). We establish consistency of both our methods when dealing with high dimensional locally stationary predictors and errors with only polynomially decaying tails. Although we focus on linear time varying parameter models, L_2 boosting methods can easily be adapted to fit more general non-linear models by considering alternative base learners such as regression trees with varying degrees of depth. This makes the L_2 boosting framework more flexible than the often used ℓ_1 penalized likelihood

[4]The first approach has a long history in macroeconomics, some examples include Cogley and Sargent (2001); Primiceri (2005); Koop and Korobilis (2013). For the literature on structural breaks, see Perron et al. (2006); Casini and Perron (2018) for surveys.

approaches.

The smooth TVP model considered in this paper has been studied in the econometrics literature for the case when the number of predictors is fixed and assumed known. Under this assumption, Robinson (1989, 1991) studied the asymptotic properties of the local constant estimator of the coefficient functions. The theory was further developed in several directions.[5] To our knowledge, there were only two attempts at modeling sparse high dimensional smooth TVP models, both dealing with locally stationary sub-Gaussian predictors, and rely on l_1 regularization methods along with kernel smoothing to estimate the coefficient functions. In particular, Ding et al. (2017) deals with locally stationary sparse VAR processes, and proposes a hybrid estimator which combines l_1 regularization with local constant estimation. Lee et al. (2016) deals with models where the set of non-zero coefficient functions does not change with over time, and proposes a computationally intensive penalized local linear estimation method. Our work adds to this line of research by proposing L_2 boosting algorithms for high dimensional smooth TVP models characterized by (1.1).

Our methods compare favorably to more commonly used alternatives for modeling time varying parameters such as assuming the coefficients are stochastic and generated by a random walk, or using a rolling window estimator with a fixed window length. These models are typically estimated via MCMC, or other computationally intensive methods, which excludes the use of high dimensional datasets. Rolling window fore-

[5] Some examples include: Orbe et al. (2005, 2006) considered shape restricted estimation. Cai (2007) analyzed the asymptotic properties of the local linear estimator. Inoue et al. (2017) considered the question of optimal bandwidth selection for the local constant estimator when using the uniform kernel. Zhang et al. (2015), Hu et al. (2018), and Vogt et al. (2012) allow for non-stationary predictors and non-linear time varying functions of these predictors. Zhou and Wu (2009); Zhou (2010) considered local linear quantile estimation, Phillips et al. (2017) obtained results for cointegration models, and Chen (2015) dealt with models with endogenous predictors.

casts, although they are usually not presented this way, are actually equivalent to using a local constant estimator using a uniform kernel and a fixed bandwidth. This choice of fixed bandwidth is arbitrary and can lead to larger forecast errors vs using the optimal bandwidth (Inoue et al., 2017). Additionally, local constant estimators have higher asymptotic bias at the boundary of the sample vs local linear estimators. In contrast, our L_2 boosting algorithms are capable of variable selection and estimation simultaneously at a very low computational cost even for very high dimensional data. Also, using non-parametric methods to estimate the time varying coefficient functions allows our method to perform well even under model misspecifications such as discrete breaks, stochastic coefficients generated by a random walk, and time invariant coefficients; see Giraitis et al. (2013); Inoue et al. (2017) and our simulations section for more details.

On the empirical side we include an extensive application to macroeconomic forecasting. Although parameter instability has long been established in the econometrics literature (Stock and Watson, 2003, 2009; Breitung and Eickmeier, 2011), the question of whether one can exploit this instability to improve macroeconomic forecasts is far less clear (see section 1.7 or Rossi (2013) for more details). Some issues which have hindered the utility of modeling time variation are: 1) the bias-variance tradeoff encountered when using a reduced sample for modeling, 2) misspecification and/or estimation error incurred when trying to estimate the nature of time variation, and 3) computational constraints restricting the use of high dimensional predictors when estimating traditional TVP models with stochastic coefficients.

To analyze the effectiveness of modeling time variation with our methods, we use a panel of 123 monthly series from the FRED-MD database and focus on forecasting 8 major macroeconomic series over a range of forecast horizons. Using an out of sample period of over 47 years, we find that: **1)** the benefits of modeling time variation with

our methods are substantial, especially when considering longer forecast horizons, **2)** the benefits of using our time varying boosting models vs their time invariant counterparts increases as the length of the available sample increases, and **3)** the benefits of modeling time variation appear to be confined to the high dimensional setting, as we confirm the results in Stock and Watson (1996) that modeling time variation in AR models offers little to no benefits for the majority of series.

The rest of the paper is organized as follows. Section 1.2 reviews the locally stationary framework, along with the functional dependence measure which will be used to quantify dependence. We also discuss the assumptions placed on the structure of the covariate and response processes; these assumptions are very mild, allowing us to represent a wide variety of stochastic processes which arise in practice. Section 1.3 introduces our boosting algorithms for both local constant or local linear least squares base learners, and studies the asymptotic properties of these procedures. The asymptotic properties, and the number of predictors allowed depend on the strength of dependence, and the moment conditions of the underlying processes. Section 1.6 presents results from Monte Carlo simulations, and section 1.7 contains our application to macroeconomic forecasting. Lastly, concluding remarks are in section 1.9.

1.2 The Econometric Framework

We first start with a review of locally stationary processes which were first introduced by Dahlhaus (1996); Dahlhaus et al. (1997) using a time varying spectral representation. This was expanded in Dahlhaus et al. (2018) to a more general definition which facilitated theoretical results for a large class of non-linear processes; see Dahlhaus (2012) for a partial survey of the results pertaining to locally stationary processes. Heuristically speaking, a locally stationary process is a non-stationary

process which can be well approximated by a stationary process locally in time. This is a convenient framework to model non-stationarity induced by smooth time varying parameters. Consider the model (1.1), with $\boldsymbol{\beta}_t$ being a vector of unknown deterministic smooth functions of time, as a consequence Y_t in (1.1) is clearly non-stationary. Due to this non-stationarity, letting $T \to \infty$ will not lead to consistent estimates of $\boldsymbol{\beta}_t$, since future observations may not contain any information about the probabilistic structure of the process at the present time t. Therefore, it is common to work in the infill asymptotics framework with rescaled time $t/T \in [0,1]$, with $\boldsymbol{\beta}_t = \boldsymbol{\beta}(t/T)$ (Dahlhaus et al., 1997; Robinson, 1989; Cai, 2007). Letting $T \to \infty$ now implies that we observe $\boldsymbol{\beta}(t/T)$ on a finer grid within the same interval, thereby increasing the amount of local information available. Although this setting is not commonly seen in forecasting time series, a prediction theory is still possible. For example, we can view our data as having been observed for $t = 1, \ldots, T/2$ (i.e. on the interval $[0, 1/2]$), and we are forecasting the next few observations (see Dahlhaus et al. (1997); Dahlhaus (1996)).

For a formal description of locally stationary processes we use the definition and assumptions stated in Dahlhaus et al. (2018) and Richter and Dahlhaus (2018):

Definition 1.2.1. Let $q > 0$, and $||W||_q = (E|W|^q)^{1/q}$. Let $Y_{t,T}, t = 1, \ldots, T$ be a triangular array of stochastic processes. For each $u \in [0,1]$, let $\tilde{Y}_t(u)$ be a stationary and ergodic process satisfying:

1. $D_q = \max\{\sup_{u\in[0,1]} ||\tilde{Y}_t(u)||_q, \sup_{T\in N} \sup_{t=1,\ldots,T} ||Y_{t,T}||_q\} < \infty$
2. There exists $C_B > 0$ such that uniformly in $t = 1, \ldots, T$ and $u, v \in [0,1]$:

$$||\tilde{Y}_t(u) - \tilde{Y}_t(v)||_q \leq C_B|u - v|, \;\; ||Y_{t,T} - \tilde{Y}_t(t/T)||_q \leq C_B T^{-1} \tag{1.2}$$

From the second assumption we obtain: $||Y_{t,T}-\tilde{Y}_t(u)||_q \leq O(|t/T-u|+T^{-1})$, thus for rescaled time points t/T near u, the process $Y_{t,T}$ can be approximated by a stationary process $\tilde{Y}_t(u)$ with asymptotically negligible error. Consider the model used in Robinson (1989); Cai (2007): $Y_{t,T} = \boldsymbol{\beta}(t/T)\boldsymbol{X}_t + \epsilon_t$, where $\boldsymbol{X}_t, \epsilon_t$ are stationary processes, and $\boldsymbol{\beta}(\cdot)$ is a lipschitz continuous function. Under these conditions $Y_{t,T}$ is a locally stationary process, with stationary approximation: $\tilde{Y}_t(u) = \boldsymbol{\beta}(u)\boldsymbol{X}_t + \epsilon_t$. A slightly more complicated example is a tvAR(1) process: $Y_{t,T} = \alpha(t/T)Y_{t-1,T} + \epsilon_t = \sum_{j=0}^{\infty}[\prod_{k=1}^{j-1}\alpha(\frac{t-k}{T})]\epsilon_{t-j}$. Intuitively one can see that if we assume $\alpha(\cdot)$ is lipschitz continuous then the process is locally stationary with stationary approximation: $\tilde{Y}_t(u) = \alpha(u)\tilde{Y}_{t-1}(u) + \epsilon_t$, and $||Y_{t,T} - \tilde{Y}_t(u)||_q \leq O(|t/T - u| + T^{-1})$.[6] The stationary approximation is the key to estimation and formulating an asymptotic theory when dealing with locally stationary processes. Estimation of parameters such as $\alpha(u)$ and local covariances is carried out by assuming, for each rescaled time point u, that the process is essentially stationary on a small window around u. We then carry out estimation via stationary methods using observations within this window.[7]

In order to establish asymptotic properties of our L_2 boosting procedures, we rely on the functional dependence measure used in the context of locally stationary processes in Dahlhaus et al. (2018); Richter and Dahlhaus (2018). We first introduce the following notation: Let $\{e_t\}_{t\in\mathcal{Z}}$ be a sequence of iid random variables, and let $\mathcal{F}_t = (e_t, e_{t-1}, \ldots)$, $\mathcal{F}_t^* = (e_t, e_{t-1}, \ldots, e_0^*, e_{-1}, \ldots)$ with $e_0^*, e_t, t \in \mathbb{Z}$ being iid. Additionally, let $\mathcal{H}_t = (\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t-1}, \ldots)$, $\mathcal{H}_t^* = (\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t-1}, \ldots, \boldsymbol{\eta}_0^*, \boldsymbol{\eta}_{-1}, \ldots)$ with $\boldsymbol{\eta}_0^*, \boldsymbol{\eta}_t, t \in \mathbb{Z}$ being iid random vectors. Throughout this paper, we assume the following structure for the

[6] Under appropriate conditions, more general non-linear time varying processes which satisfy the recursion: $Y_{t,T} = G_{\epsilon_t}(Y_{t-1,T}, \ldots, Y_{t-p,T}, \max(t/T, 0))$, for $t \leq T$, can be shown to be locally stationary (Dahlhaus et al., 2018). Examples of such processes include time varying ARMA, time varying GARCH, time varying VAR, and time varying random coefficient processes.

[7] We note that assuming approximate stationarity on a small window is essentially the justification of the commonly used rolling window estimators.

stationary approximation for univariate processes (such as the response and error processes), and multivariate processes (such as the covariate process) respectively:

$$\tilde{Y}_t(u) = g(u, \mathcal{F}_t) \text{ and } \tilde{\boldsymbol{x}}_t(u) = \boldsymbol{h}(u, \mathcal{H}_t) = (h_1(u, \mathcal{H}_t), \ldots, h_{p_T}(u, \mathcal{H}_t)), \tag{1.3}$$

where $g(\cdot, \cdot)$, and $\boldsymbol{h}(\cdot, \cdot)$ are real valued measurable functions. These representations allow us to define the functional dependence measure as: $\delta_q^{\tilde{Y}(u)}(t) = ||\tilde{Y}_t(u) - g(u, \mathcal{F}_t^*)||_q$, and $\delta_q^{\tilde{X}_j(u)}(t) = ||\tilde{X}_{j,t}(u) - h_j(u, \mathcal{H}_t^*)||_q$. Additionally, we assume short range dependence of the form:

$$\Delta_{0,q}^{\tilde{Y}} = \sum_{k=0}^{\infty} \sup_{u\in[0,1]} \delta_q^{\tilde{Y}(u)}(k) \leq \infty, \text{ and } \Phi_{0,q}^{\tilde{\boldsymbol{x}}} = \max_{j\leq p_T} \sum_{k=0}^{\infty} \sup_{u\in[0,1]} \delta_q^{\tilde{X}_j(u)}(k) \leq \infty, \tag{1.4}$$

for some $q > 2$ to be specified in the next section.

We place assumptions on the stationary approximation rather than directly on the process itself. This leads to results using weaker assumptions, and to more interpretable dependence measures. For an intuitive explanation of this measure, we consider the stationary approximation at time u_0 ($\tilde{Y}_t(u_0)$) and we obtain $\delta_q^{\tilde{Y}(u_0)}(k) = ||\tilde{Y}_k(u_0) - g(u_0, \mathcal{F}_k^*)||_q$. We can view $\delta_q^{\tilde{Y}(u_0)}(k)$ as measuring the dependence of $\tilde{Y}_k(u_0)$ on the innovation ϵ_0, which for weakly dependent processes decreases suitably quickly as $k \to \infty$. For a concrete example, consider a stationary AR(1) process $\tilde{Y}_t(u_0) = \sum_{j=0}^{\infty} a(u_0)^j e_{t-j}$ with e_i iid, then $\delta_q^{\tilde{Y}(u_0)}(k) = |a(u_0)^k|||e_0 - e_0^*||_q$, and $\Delta_{0,q}^{\tilde{Y}(u_0)} = ||e_0 - e_0^*||_q \sum_{k=0}^{\infty} |a(u_0)^k|$. Now in the locally stationary setting, we take the supremum over the rescaled time interval to account for the non-stationarity of the processes, thereby obtaining $\Delta_{0,q}^{\tilde{Y}} = ||e_0 - e_0^*||_q \sup_{u\in[0,1]} \sum_{k=0}^{\infty} |a(u)^k|$. A very wide variety of locally stationary processes encountered in practice including time varying linear processes, tv-ARMA, tv-GARCH, tv-TAR, and tv-VAR, and time varying random coefficient

processes have stationary approximations which satisfy (1.4), and have geometrically decaying functional dependence measures (see Dahlhaus et al. (2018)).

1.3 Boosting High Dimensional TVP Models

Ever since the introduction of AdaBoost in the 1990's (Freund and Schapire, 1997), boosting algorithms have been one of the most successful and widely utilized machine learning methods (Friedman et al., 2001). AdaBoost, which was developed for classification, consisted of iteratively fitting a series of weak classifiers or learners onto reweighted data and taking a weighted average of the predictions from each of these simple models. The success of AdaBoost was originally thought to originate from averaging many weak classifiers and from a reweighting scheme which placed large weights on heavily misclassified observations. Later work by Friedman (2001), and Friedman et al. (2000) established AdaBoost as a gradient descent algorithm in function space using an exponential loss function. This functional gradient descent view connected boosting to the common optimization view of statistical inference, and led to extensions of boosting beyond the realm of classification. Friedman (2001) proposed several new boosting algorithms using alternative base learners and loss functions including squared error loss, leading to L_2 boosting. Additionally, Efron et al. (2004) and Friedman et al. (2001), made connections for linear models between L_2 boosting and common statistical procedures such as the Lasso and forward stage-wise regression.[8] [9] These insights shed light on L_2 boosting as a method which performs variable selection and shrinkage leading to sparse models. For an excellent

[8]For theoretical connections one can consult Chapter 16.2 of Friedman et al. (2001), and additional works such as Hastie et al. (2007); Rosset et al. (2004)

[9]Empirical comparisons between boosting with linear least squares learners and the lasso have shown close performance with boosting performing slightly better in the case of high correlated predictors Hastie et al. (2007); Hepp et al. (2016).

survey of the statistical view of boosting and results pertaining to several common boosting algorithms, one can consult Buhlmann and Hothorn (2007). [10]

We are interested in estimating the following model:

$$Y_{t,T} = \boldsymbol{\beta}'(t/T)\boldsymbol{x}_{t-h,T} + \epsilon_{t,T} \text{ for } t = 1, \ldots, T, \tag{1.5}$$

where $Y_{t,T}$ is the response, $\boldsymbol{x}_{t-h,T} = (X_{1,t-h,T}, \ldots, X_{p_T,t-h,T})'$ is a p_T-dimensional vector of locally stationary predictors (with $p_T >> T$), $\boldsymbol{\beta}'(t/T) = (\beta_1(t/T), \ldots, \beta_{p_T}(t/T))$ is a vector of unknown functions of time defined on the grid $[0, 1]$, which becomes finer as $T \to \infty$, and $\epsilon_{t,T}$ denotes the locally stationary error process with $E(\epsilon_{t,T}\boldsymbol{x}_{t-h,T}) = 0 \; \forall \; t, T$. We denote the stationary approximation of the response as $\tilde{Y}_t(u) = \boldsymbol{\beta}'(u)\tilde{\boldsymbol{x}}_{t-h}(u) + \tilde{\epsilon}_t(u)$. To simplify notation, we discuss estimation at the boundary point $u = T/T = 1$. Before we introduce our boosting algorithms, it helps to first introduce the population version of componentwise L_2 boosting with linear base learners as applied to the stationary approximations $(\tilde{Y}_T(u), \tilde{\boldsymbol{x}}_{T-h}(u))$:

Algorithm: Population level L_2 Boosting

1. Set $F^{(0)}(u, \tilde{\boldsymbol{x}}_{T-h}(u)) = E(\tilde{Y}_T(u))$

2. For $m = 1, \ldots, M_T$, where M_T is some stopping iteration, do:

   (a) Compute $\tilde{U}_T^{(m)}(u) = \tilde{Y}_T(u) - F^{(m-1)}(u, \tilde{\boldsymbol{x}}_{T-h}(u))$.

   (b) Let $\mathcal{S}_m = \text{argmin}_{j \leq p_T} E(\tilde{U}_T^{(m)}(u) - \alpha_j^{(m)}(u)\tilde{X}_{j,T-h}(u))^2$,
   where $\alpha_j^{(m)}(u) = E(\tilde{X}_{j,T-h}(u)\tilde{U}_T^{(m)}(u))/E(\tilde{X}_{j,T-h}^2(u))$.

[10] Additionally, one can consult Buhlmann (2006) for extensions of boosting to stationary VAR processes, and Bai and Ng (2009); Ng (2014) for applications to macroeconomic forecasting and recession classification respectively.

(c) Update $F^{(m)}(u, \tilde{\boldsymbol{x}}_{T-h}(u)) = F^{(m-1)}(u, \tilde{\boldsymbol{x}}_{T-h}(u)) + \upsilon \cdot \alpha^{(m)}_{\mathcal{S}_m}(u)\tilde{X}_{\mathcal{S}_m,T-h}(u)$, where $\upsilon \in (0, 1]$ is a step length factor.

3. Output $F^{(M_T)}(u, \tilde{\boldsymbol{x}}_{T-h}(u)) = F^{(0)}(u, \tilde{\boldsymbol{x}}_{T-h}(u)) + \upsilon \sum_{m=1}^{M_T} \alpha^{(m)}_{\mathcal{S}_m}(u)\tilde{X}_{\mathcal{S}_m,T-h}(u)$

Although we use linear base learners, we note that our methods can be extended to a broader class of models by using a more general base learner, such as $g_j(u, \tilde{X}_{j,T-h}(u)) = E(\tilde{Y}_T(u)|\tilde{X}_{j,T-h}(u))$, and estimating using kernel regressions or smoothing splines. For the corresponding sample version of L_2 boosting with linear base learners, it is informative to consider the case of stationary response and predictor processes. In the stationary setting, we can remove the dependence on T and the sample version of our algorithm simplifies to $\hat{\mathcal{S}}_m = \text{argmin}_j \sum_{t=1}^{T}(U_t^{(m)} - \hat{\alpha}_j^{(m)} X_{t,j})^2$, where $\hat{\alpha}_j^{(m)} = T^{-1}\sum_{t=1}^{T} X_{j,t-h}U_t^{(m)}$, assuming $E(X_t), E(Y_t) = 0$, and $E(X_t^2) = 1$. For the case of locally stationary response and predictor processes the situation is more complicated as the above estimator is inconsistent for $\alpha_j^{(m)}(u)$. Intuitively, this inconsistency arises since observations "far" from rescaled time u contain little information about the probabilistic structure of the processes at time u.

To proceed with estimation in the locally stationary setting, $\forall m$ and $j \leq p_T$, we have $U_{t,T}^{(m)} = \alpha_j^{(m)}(t/T)X_{j,t-h,T} + \epsilon_{j,t,T}$,
where $\alpha_j^{(m)}(t/T) = E(\tilde{X}_{j,t-h}(t/T)\tilde{U}_T^{(m)}(t/T))/E(\tilde{X}^2_{j,t-h}(t/T))$.[11] By local stationarity and assuming appropriate smoothness conditions, we have the following expansion:

$$\alpha_j^{(m)}(t/T) = \alpha_j^{(m)}(u) + \dot{\alpha}_j^{(m)}(u)(t/T - u) + \ddot{\alpha}_j^{(m)}(c)(t/T - u)^2, \tag{1.6}$$

where $\dot{\alpha}(\cdot), \ddot{\alpha}(\cdot)$ denote the first and second derivative respectively of the function, with c between u and t/T. To compute the local constant estimate for $\alpha_j^{(m)}(u)$, we

[11] Recall that $E(X_{j,t-h,T}U_{t,T}^{(m)})/E(X^2_{j,t-h,T}) = \alpha_j^{(m)}(t/T) + O(T^{-1})$ by local stationarity.

ignore the linear term in the Taylor expansion to obtain the following approximation: $U_{t,T}^{(m)} \approx \alpha_j^{(m)}(u)X_{j,t-h,T}+\epsilon_{j,t,T}$ for t/T near u. The local constant estimator for $\alpha_j^{(m)}(u)$ is then

$$\hat{\alpha}_{lc,j}^{(m)}(u) = \frac{\sum_{t=1}^{T} K_b(t/T-u)X_{j,t-h,T}U_{t,T}^{(m)}}{\sum_{t=1}^{T} K_b(t/T-u)X_{j,t-h,T}^2}, \tag{1.7}$$

where $K_b(x) = b^{-1}K(x/b)$, is a kernel function and b is the bandwidth. Therefore, $\hat{\alpha}_{lc,j}^{(m)}(u)$ is a weighted least squares estimate, with the weights given by the kernel values. For now, one can think of this estimator as aiming to use information from observations "near" time T, while discounting information from distant points. A simple example of the local constant estimate is the rolling window estimate: using the uniform kernel $K(x) = \mathbb{1}_{|x|\leq 1}$, with a fixed bandwidth $b = b_0$, we obtain a rolling window estimate which uses the last b_0T observations in our sample.

The local constant estimate is widely used for estimating time varying effects, however the Taylor expansion of $\alpha_j^{(m)}(t/T)$ suggests we can obtain a better approximation by using the linear term in the expansion (1.6). This was analyzed rigorously in Cai (2007), which showed that for boundary points the local linear estimator is theoretically superior to the local constant estimator. Using the expansion (1.6), we obtain: $U_{t,T}^{(m)} \approx \alpha_j^{(m)}(u)X_{j,t-h,T} + \dot{\alpha}_j^{(m)}(u)X_{j,t-h,T}(t/T-u) + \epsilon_{j,t,T}$, for t/T near u. Let $\boldsymbol{Z}_{j,t-h,T}\boldsymbol{\theta}_j^{(m)}(u)$ where $\boldsymbol{Z}_{j,t-h,T} = (X_{j,t-h,T}, X_{j,t-h,T}(t/T-u))$, $\boldsymbol{\theta}_j^{(m)}(u) = (\alpha_j^{(m)}(u), \dot{\alpha}_j^{(m)}(u))'$. The local linear estimate is obtained by minimizing a weighted least squares criterion:

$$\hat{\boldsymbol{\theta}}_j^{(m)}(u) = (\hat{\alpha}_{ll,j}^{(m)}(u), \hat{\dot{\alpha}}_{ll,j}^{(m)}(u)) = \text{argmin}_{\boldsymbol{\theta}_j^{(m)}(u)} \sum_{t=1}^{T} K_b(t/T-u)(U_{t,T}^{(m)} - \boldsymbol{Z}_{t-h,T}\boldsymbol{\theta}_j^{(m)}(u))^2 \tag{1.8}$$

Using these estimators we can formulate our L_2 boosting algorithm for (1.5) using local constant, and local linear estimators as base learners. We first start with our first algorithm which uses local constant estimators:

Algorithm 1: Local Constant L_2 Boosting (LC-Boost)

1. Set $\hat{F}_{lc}^{(0)}(u, \boldsymbol{x}_{t,T}) = T^{-1}\sum_{i=h+1}^{T} K_b(i/T - u)Y_{i,T}$, for $t = 1, \ldots, T-h$
2. For $m = 1, \ldots, M_T$, where M_T is some stopping iteration, do:
   (a) Compute the residuals $\hat{U}_{i,T}^{(m)} = Y_{i,T} - \hat{F}_{lc}^{(m-1)}(u, \boldsymbol{x}_{i-h,T})$ for $i = h+1, \ldots, T$.
   (b) Let $\hat{\mathcal{S}}_m = \text{argmin}_{j \leq p_T} \sum_{i=h+1}^{T} K_b(i/T - u)(\hat{U}_{i,T}^{(m)} - \hat{\alpha}_{lc,j}^{(m)}(u)X_{j,i-h,T})^2$
   (c) Update $\hat{F}_{lc}^{(m)}(u, \boldsymbol{x}_{i-h,T}) = \hat{F}_{lc}^{(m-1)}(u, \boldsymbol{x}_{i-h,T}) + \upsilon\hat{\alpha}_{lc,\mathcal{S}_m}^{(m)}(u)X_{\mathcal{S}_m,i-h,T}$, where $\upsilon \in (0, 1]$ is a step length factor.
3. Output $\hat{F}_{lc}^{(M_T)}(u, \boldsymbol{x}_{T-h,T}) = \hat{F}_{lc}^{(0)}(u, \boldsymbol{x}_{t,T}) + \upsilon\sum_{m=1}^{M_T} \hat{\alpha}_{lc,\mathcal{S}_m}^{(m)}(u)X_{\mathcal{S}_m,T-h,T}$

Let $\boldsymbol{z}_{t,T} = (\boldsymbol{x}_{t,T}, \boldsymbol{x}_{t,T}(t/T - u))$, our boosting algorithm using local linear estimates as base learners is:

Algorithm 2: Local Linear L_2 Boosting (LL-Boost)

1. Set $\hat{F}_{ll}^{(0)}(u, \boldsymbol{x}_{i-h,T}) = T^{-1}\sum_{i=h+1}^{T} K_b(i/T - u)Y_{i,T}$, for $i = h+1, \ldots, T$
2. For $m = 1, \ldots, M_T$, where M_T is some stopping iteration, do:
   (a) Compute the residuals $\hat{U}_{i,T}^{(m)} = Y_{i,T} - \hat{F}_{ll}^{(m-1)}(\boldsymbol{x}_{i-h,T})$ for $i = h+1, \ldots, T$.
   (b) Let $\hat{\mathcal{S}}_m = \text{argmin}_{j \leq p_T} \sum_{i=1}^{T} K_b(i/T - u)(\hat{U}_{i,T}^{(m)} - \boldsymbol{Z}_{j,i-h,T}\hat{\boldsymbol{\theta}}_j^{(m)}(u))^2$.

(c) Update $\hat{F}_{ll}^{(m)}(u, \boldsymbol{x}_{i-h,T}) = \hat{F}_{ll}^{(m-1)}(u, \boldsymbol{z}_{i-h,T}) + \upsilon \cdot \boldsymbol{Z}_{\mathcal{S}_m, i-h,T} \hat{\boldsymbol{\theta}}_{\mathcal{S}_m}^{(m)}(u)$, where $\upsilon \in (0, 1]$ is a step length factor.

3. Output $\hat{F}_{ll}^{(M_T)}(u, \boldsymbol{x}_{T-h,T}) = \hat{F}_{ll}^{(0)}(u, \boldsymbol{x}_{i-h,T}) + \upsilon \sum_{m=1}^{M_T} \boldsymbol{Z}_{\mathcal{S}_m, T-h,T} \hat{\boldsymbol{\theta}}_{\mathcal{S}_m}^{(m)}(u)$

We see that boosting is a stagewise estimation procedure, where at each stage only one learner is updated and the previously selected terms are unchanged. This stagewise fitting procedure induces regularization through limiting the number of steps (M_T), and the step length factor (υ). We usually fix the the step-length factor (υ) to a low number such as $\upsilon = .1$, making the stopping iteration (M_T) akin to the regularization parameter of the Lasso.[12] In light of this, boosting can be thought of as a close relative of the lasso, with the advantage of being able to approximate the ℓ_1 penalized solution in situations where it is impossible or computationally burdensome to compute the Lasso solution (Friedman et al., 2004).

By viewing boosting as a general regularized function estimation procedure, we can formulate a generic local constant boosting procedure which can be easily be computed for a wide variety of base learners and (almost everywhere) differentiable loss functions ($L(\cdot, \cdot)$).

Algorithm 3: Generic Local Constant Boosting

1. Set $\hat{F}_G^{(0)}(u, \boldsymbol{x}_{t,T}) = \text{argmin}_c T^{-1} \sum_{i=h+1}^{T} K_b(i/T - u) L(Y_{i,T}, c)$, for $t = 1, \ldots, T-h$

2. For $m = 1, \ldots, M_T$, where M_T is some stopping iteration, do:

   (a) Compute the pointwise negative gradient:

   $U_{i,T}^{(m)} = \frac{d}{dF} L(Y_{i,T}, F)\Big|_{F=\hat{F}_G^{(m-1)}(u, \boldsymbol{x}_{i-h,T})}$ evaluated at $i = h+1, \ldots, T$.

[12] Given that each predictor can be selected multiple times, especially for low values of υ, the number of predictors in the estimated model is $\leq M_T$, and all predictors which have not been selected by step M_T have an effect of zero.

(b) Let $\hat{\mathcal{S}}_m = \text{argmin}_{j \leq p_T} \sum_{i=h+1}^{T} K_b(i/T - u)(\hat{U}_{i,T}^{(m)} - \widehat{g}_j^{(m)}(u, X_{j,i-h,T}))^2$

(c) Update $\hat{F}_G^{(m)}(u, \boldsymbol{x}_{i-h,T}) = \hat{F}_G^{(m-1)}(u, \boldsymbol{x}_{i-h,T}) + \upsilon \widehat{g}_{S_m}^{(m)}(u, X_{S_m,i-h,T})$, where $\upsilon \in (0, 1]$ is a step length factor.

3. Output $\hat{F}_G^{(M_T)}(u, \boldsymbol{x}_{T-h,T}) = \hat{F}_G^{(0)}(u, \boldsymbol{x}_{t,T}) + \upsilon \sum_{m=1}^{M_T} \widehat{g}_{S_m}^{(m)}(u, X_{S_m,T-h,T})$

The algorithm can be modified to allow $g_j(u, \cdot)$ to be a function of several variables e.g. a predictor along with a number of its lags.

1.4 Implementation

Implementation of these algorithms is very simple and can be carried out using existing software packages. We first discuss the choice of the kernel function $K(\cdot)$, bandwidth (b), stopping iteration (M_T), and step length factor (υ). We set $\upsilon = .1$, which is the default choice in statistical software packages and applied work (Buhlmann and Hothorn, 2007; Friedman, 2001; Hofner et al., 2014). In non-parametric statistics and machine learning the most commonly used kernels are the Gaussian Kernel and the Epanechnikov kernel $K(u) = .75(1 - u^2)\mathbb{1}_{|u| \leq 1}$, while in econometrics the uniform kernel $\mathbb{1}_{|u| \leq 1}$ is more widely used. Both the uniform kernel and the Epanechnikov Kernel use a subset of the sample, with the Epanechnikov kernel also downweighting more distant observations within this subset. The Gaussian kernel does not truncate the sample, instead it smoothly downweights more distant observations. It has a much smoother downweighting scheme than the Epanechnikov kernel, which can be beneficial in many applications.[13] In general however, the choice of a kernel does not have much impact on the performance, as opposed to the selection of the bandwidth parameter which is crucial.

[13] We decide to use the uniform kernel in our applications due to its close connections with the rolling window estimator. Using the Gaussian Kernel gave us similar results.

We first discuss bandwidth selection for an out of sample forecasting exercise. To help with exposition, we use a concrete example: assume we have monthly data ranging from 1960:1 to 2018:8, giving us about ~ 700 observations. We begin our forecasts on 1970:1 and move forward utilizing an expanding window framework. We use one-sided kernels to avoid looking into the future. We choose our bandwidth parameter using a cross validation approach. We first form a grid of values $B = (b_1, \ldots, b_n)$ from which to select the bandwidth parameter. For each forecast, our cross validation procedure uses the last ω (where ω is chosen by the researcher) observations of our sample for an out of sample forecasting exercise. We then choose the bandwidth which minimizes the MSFE over this sub-sample. Therefore, the selected bandwidth is:

$$b^*_{T_0} = \text{argmin}_{b_i \in B} \omega^{-1} \sum_{\tau=T_0-\omega}^{T_0-h} (Y_{\tau,T} - \hat{F}^{(M_T)}_{\tau,b_i}(\tau/T, \boldsymbol{x}_{\tau-h,T}))^2,$$

where $\hat{F}^{(M_T)}_{\tau,b_i}(\tau/T, \boldsymbol{x}_{\tau-h,T})$ refers to the LC-Boost or LL-Boost estimate of $\boldsymbol{x}_{\tau-h,T}\beta(\tau/T)$ using only observations until time τ, and the bandwidth b_i. For our first out of sample forecast we set $T_0 = 120$, which is the length of the sample available at the time, and for each additional forecast we increment T_0 by 1 until we reach the end of the sample. In the special case of using LC-Boost with a one sided uniform kernel, we are selecting the optimal window size at each time point, via cross validation, for a rolling window forecast. With the bandwidths representing the fraction of the sample we are using for estimation.

For in-sample estimation problems, two sided kernels are used in our algorithms with a weighted leave one out cross validation procedure to select the bandwidth.

The procedure is as follows:

$$b^*_{T_0} = \text{argmin}_{b_i \in B} T^{-1} \sum_{\tau=h}^{T} (Y_{\tau,T} - \hat{F}^{(M_T)}_{lc,-\tau,b_i}(\tau/T, \boldsymbol{x}_{\tau-h,T}))^2 K_{b_i}(\tau/T - T_0/T),$$

where $\hat{F}^{(M_T)}_{lc,-\tau,b_i}(\tau/T, \boldsymbol{x}_{\tau-h,T})$ refers to the estimate of $\boldsymbol{x}_{\tau-h}\beta(\tau/T)$, which uses all observations except $(Y_{\tau,T}, \boldsymbol{x}_{\tau-h,T})$. The kernel in the above equation discounts errors far away from the time point t_0 when selecting the optimal bandwidth. This procedure gives us a bandwidth for each time point in the sample, and if one wants a single bandwidth for all time points, the kernel can be removed.

To select the stopping iteration M_T, we specify an upper bound for the number of iterations M_{upp} (we set $M_{upp} = 100$), where $M_T \leq M_{upp}$. The stopping iteration is then selected using the corrected AIC (AIC_c) statistic given in Buhlmann (2006):

$$M_T = \text{argmin}_{m \leq M_{upp}} AIC_c(m),$$

where $AIC_c(m)$ is the AIC of the model using m iterations.[14]

Our methods can be computed extremely quickly using the existing R package **mboost**. Our base learners are univariate or bivariate weighted least squares estimates which can be implemented through existing functions in the package once we specify the kernel values as weights. We can also implement the generic local constant boosting algorithm for wide a variety of base learners and loss functions such as absolute loss, Huber loss and quantile loss.[15] As an example, to obtain quantiles

[14] Alternatively, we can jointly select M_T and the bandwidth $b^*_{T_0}$ by forming a two dimensional grid and selecting the optimal combination using the cross validation procedure described earlier. We decide to use the AIC_c statistic in this work. We note that when dealing with very large sample sizes and/or more complicated base learners which are a function of more than one variable, using cross validation to select M_T, using a moderately sized grid, can often be quicker since calculation of the corrected AIC requires computing the trace of the Hat matrix.

[15] We refer the reader to Hofner et al. (2014) which provides an excellent introduction and tutorial

for our forecasts, we specify the quantile loss for a given quantile[16], and compute the optimal bandwidth for our base learners by using the cross validation procedure mentioned above. A density forecast can be obtained from these estimated quantiles by using the procedure outlined in **?**.

1.5 Asymptotic Theory

In order to prove our asymptotic results, we need the following assumptions:

Condition 1.5.1. Assume $\sup_{u\in[0,1]} |\boldsymbol{\beta}(u)|_1 < \infty$

Condition 1.5.2. Assume the error and the covariate processes are locally stationary and have representations given in (1.3). Additionally, we assume the following decay rates $\Phi^{\boldsymbol{x}}_{m,r} = O(m^{-\alpha_x}), \Delta^{\epsilon}_{m,q} = O(m^{-\alpha_\epsilon})$, for some $\alpha_x, \alpha_\epsilon > 0$, $q > 2, r > 4$ and $\tau = \frac{qr}{q+r} > 2$.

Condition 1.5.3. Let $\Sigma_{\tilde{x}}(u) = E(\tilde{\boldsymbol{x}}'_t(u)\tilde{\boldsymbol{x}}_t(u))$ be the covariance matrix function. For $u \in [0,1]$, assume that $\boldsymbol{\beta}(u), \Sigma_{\tilde{x}}(u) \in \mathcal{C}^2[0,1]$, where $\mathcal{C}^2[0,1]$ denotes the class of functions defined on $[0,1]$ that are twice differentiable with bounded derivatives.

Condition 1.5.4. The kernel function $K(u)$ is bounded and symmetric, and of bounded variation with compact support. Additionally, the bandwidth (b) satisfies $bT = R_T = O(T^{\psi})$, where $\psi \in (0,1)$.

Condition 1.5.1 requires ℓ_1 sparsity of the time varying coefficients, and allows the active set of predictors to change over time. Our asymptotic results do not require sparsity in the number of non-zero coefficients (ℓ_0 sparsity). Condition 1.5.2 assumes

to the **mboost** package. It also lists the wide variety of base learners and loss functions supported by the package.

[16]See **?** for more details on the quantile boosting algorithm.

the covariate and error processes are locally stationary, and presents the dependence and moment conditions on these processes, where higher values of $\alpha_x, \alpha_\epsilon$ indicate weaker temporal dependence. We assume our predictor and error processes have at least $r > 4$ and $q > 2$ finite moments respectively. Examples of processes satisfying condition 1.5.2 were given in section 1.2.

Given that $\boldsymbol{x}_{t-h}$ can contain lags of $Y_{t,T}$, an example of a model which satisfies the above conditions is as follows: Let $\boldsymbol{W}_{t,T} = (Y_{t,T}, \boldsymbol{z}_{t,T})$, where $\boldsymbol{z}_{t,T}$ represents our exogenous series, and $\boldsymbol{W}_{t,T} = \sum_{i=1}^{\ell} \boldsymbol{A}_i(t/T)\boldsymbol{W}_{t-i,T} + \boldsymbol{\eta}_t$. Then the stationary approximation is $\widetilde{\boldsymbol{W}}_t(t/T) = \sum_{i=1}^{\ell} \boldsymbol{A}_i(t/T)\widetilde{\boldsymbol{W}}_{t-i}(t/T) + \boldsymbol{\eta}_t$, with cumulative functional dependence measure $\Phi_{0,r}^{\tilde{\boldsymbol{W}}} = \sup_{u\in[0,1]} \sum_{k=0}^{\infty} O(\lambda_{\max}(\boldsymbol{A}^*(u))^k)$(Chen et al., 2013), where $\boldsymbol{A}^*(u)$ is the companion matrix. We can then define $\boldsymbol{x}_{t-1,T} = (\boldsymbol{W}_{t-1,T}, \ldots, \boldsymbol{W}_{t-l,T})$, and $\boldsymbol{\beta}(t/T)$ as the first row of the companion matrix $\boldsymbol{A}^*(u)$. We weaken the assumptions placed in the works Cai (2007); Robinson (1989); Chen and Hong (2012) which restricted the predictors and errors to be stationary, thus ruling out models with lagged dependent variables. Compared to previous works on high dimensional TVP models, such as Ding et al. (2017); Lee et al. (2016), we use a different dependence framework, and allow the predictors and errors to have polynomially decaying tails.

Condition 1.5.3 is a sufficient condition to guarantee that the expansion (1.6) exists, i.e: $\alpha_j^{(m)}(u) \in \mathcal{C}_2[0,1], \forall m$ and $j \leq p_T$. Sufficient conditions needed for smoothness of the covariance matrix function were given in Ding et al. (2017) for the case of locally stationary VAR processes, and one can consult Dahlhaus et al. (2018) for sufficient conditions for more general processes. Condition 1.5.4 is a standard condition and it includes the commonly used Epanechnikov ($K(u) = .75(1-u^2)\mathbb{1}_{|u|\leq 1}$) and uniform ($K(u) = \mathbb{1}_{|u|\leq 1}$) kernels. It also places the standard conditions on the

effective sample size R_T. Let

$$a_T = \left[R_T^{-\tau+\tau\kappa+1} + p_T R_T^{-r/2+r\kappa+1} + p_T \exp\left(-R_T^{1-2\kappa}\right) + \exp\left(-R_T^{1-2\kappa}\right) \right]$$

The following two theorems presents the consistency of LC-Boost and LL-Boost.

Theorem 1. *Let $\boldsymbol{x}^*_{T-h,T}$ denote a new predictor variable, independent of and with the same distribution as $\boldsymbol{x}_{T-h,T}$. Let $\kappa \in (0, 1/2)$ be such that $\kappa < \psi^{-1} - 1$, Suppose that conditions 1.5.1, 1.5.2, 1.5.3, and 1.5.4 hold. Then*

a. on a set with probability at least $1 - O(p_T a_T)$, our LC-Boost estimate $\hat{F}_{lc}^{(M_T)}(\cdot,\cdot)$ satisfies: $E(|\hat{F}_{lc}^{(M_T)}(u, \boldsymbol{x}^*_{uT-h,T}) - \boldsymbol{\beta}'(u)\boldsymbol{x}^*_{uT-h,T}|^2) = o_p(1) \quad (T \to \infty)$ for some sequence $M_T \to \infty$ sufficiently slowly,

b. on a set with probability at least $1 - O(p_T a_T)$, our LL-Boost estimate $\hat{F}_{ll}^{(M_T)}(\cdot,\cdot)$ satisfies $E(|\hat{F}_{ll}^{(M_T)}(u, \boldsymbol{x}^*_{uT-h,T}) - \boldsymbol{\beta}'(u)\boldsymbol{x}^*_{uT-h,T}|^2) = o_p(1) \quad (T \to \infty)$ for some sequence $M_T \to \infty$ sufficiently slowly,

This is an extension of theorem 1 in Buhlmann (2006) to the locally stationary time series setting with local constant or local linear base learners. From the above theorems, we see the range of p_T depends primarily on the moment conditions, the effective sample size R_T, and κ. For example, if we assume only finite polynomial moments with $r = q$, $\alpha \geq 1/2 - 2/r$ then, $p_T = o(R_T^{r/4-r\kappa/4-1/2})$ for our estimates to be consistent. If we assume, subgaussian or subexponential predictors for example we have $p_T = o(T^{\phi})$ for arbitrary $\phi > 0$. This is the same range Buhlmann (2006) obtained for iid sub-Gaussian predictors and errors. Given the $O(T^{-1})$ encountered when approximating a locally stationary process by a stationary distribution, we are unable to extend the theory to the ultra-high dimensional setting i.e $p_T = o(\exp(n^c))$ for $c < 1$.

We also provide results for the stationary time series with time invariant parameters. In this setting, we use the linear least squares base learner and use the entire sample for estimation. For the case of only a finite number of moments, the results in theorem 1 easily carry over to the stationary time invariant setting (i.e $\boldsymbol{\beta}(t/T) = \boldsymbol{\beta}\ \forall t, T$), by letting $R_T = T$, and computing the relevant functional dependence measures. However, we can obtain a larger range for p_T, if we assume a stronger moment condition such as:

Condition 1.5.5. Assume the response and the covariate processes are stationary and have representations given in (1.3). Additionally, assume $\upsilon_x = \sup_{q\geq 2} q^{-\tilde{\alpha}_x}\Phi^{\boldsymbol{x}}_{0,q} < \infty$ and $\upsilon_\epsilon = \sup_{q\geq 2} q^{-\tilde{\alpha}_\epsilon}\Delta^{\epsilon}_{0,q} < \infty$, for some $\tilde{\alpha}_x, \tilde{\alpha}_\epsilon \geq 0$.

Condition 1.5.5 strengthens the moment condition 1.5.2, and requires that all moments of the covariate and response processes are finite. To illustrate the role of the constants $\tilde{\alpha}_x$ and $\tilde{\alpha}_\epsilon$, consider the example where $\epsilon_t = \sum_{j=0}^{\infty} a_j e_{t-j}$ with e_i iid, and $\sum_{j=0}^{\infty}|a_j| < \infty$. Then $\Delta^{\epsilon}_{0,q} = ||e_0 - e_0^*||_q \sum_{j=0}^{\infty}|a_j|$. Now if we assume e_0 is sub-Gaussian, then $\tilde{\alpha}_\epsilon = 1/2$, since $||e_0||_q = O(\sqrt{q})$, and if e_i is sub-exponential, we have $\tilde{\alpha}_\epsilon = 1$.

The following corollary states the corresponding results for the stationary time series setting. We define $\tilde{\psi} = \frac{2}{1+2\tilde{\alpha}_x+2\tilde{\alpha}_\epsilon}, \tilde{\varphi} = \frac{2}{1+4\tilde{\alpha}_x}$, and let

$$b_T = \left[\exp\left(-\frac{T^{1/2-\kappa}}{\upsilon_x\upsilon_\epsilon}\right)^{\tilde{\psi}} + p_T \exp\left(-\frac{T^{1/2-\kappa}}{\upsilon_x^2}\right)^{\tilde{\varphi}}\right],$$

and let $\hat{F}^{(M_T)}(\boldsymbol{x}_t)$ denote our L_2 boosting estimate for Y_t, we then have:

Corollary 2. *Let $\kappa \in (0, 1/2)$, and $\boldsymbol{x}^*_{T-h}$ denote a new predictor variable, independent of and with the same distribution as $\boldsymbol{x}_{T-h}$. Suppose conditions 1.5.1, 1.5.4, and 1.5.5 hold. Then on a set with probability at least $1-O(p_T b_T)$, we have that our L_2 Boosting*

estimate $\hat{F}^{(M_T)}(\cdot)$ *satisfies:*

$$E(|\hat{F}^{(M_T)}(\boldsymbol{x}^*_{T-h}) - \boldsymbol{\beta}\boldsymbol{x}^*_{T-h}|^2) = o_p(1) \quad (T \to \infty).$$

We see that in the stationary setting our theorems improve upon previous results by providing a more detailed and larger range for p_T. For example, assuming sub-Gaussian predictors and errors we obtain $p_T = o(\exp(T^{\frac{1-2\kappa}{3}}))$, and $p_T = o(\exp(T^{\frac{1-2\kappa}{5}}))$ for subexponential predictors and errors. As a comparison Buhlmann (2006) obtained $p_T = o(T^{\phi})$, for arbitrary $\phi > 0$, when applying L_2 boosting for stationary sub-Gaussian time series.

1.6 Simulations

In this section, we evaluate the forecasting performance of our algorithms in a finite sample setting. Let $Y_{t,T}$ denote our response, and let $\boldsymbol{x}_{t-1,T} = (Y_{t-1,T}, \ldots, Y_{t-3,T}, \boldsymbol{z}_{t-1,T}, \ldots, \boldsymbol{z}_{t-3,T})$ represent our potential set of predictors, where $\boldsymbol{z}_{t-1,T} \in \mathcal{R}^{d_T}$ represents our d_T exogenous series at time t. We fix $T = 200$, and $d_T = 100$, giving us $p_T = 303$ potential predictors. We consider 14 DGPs and our general model is, for $t = 1, \ldots T$,

$$\begin{aligned} Y_{t,T} &= \rho Y_{t-1,T} + \sum_{j=1}^{4}(b + \beta_j(t/T))z_{j,t-1,T} + \epsilon_t \\ \boldsymbol{z}_{t,T} &= A(t/T)\boldsymbol{z}_{t-1,T} + \boldsymbol{\eta}_t \end{aligned}$$

and it is assumed that $\rho = .6$, $b = 0.5$. For DGPs 1-12 we let $A(t/T) = \{.4^{|i-j|+1}\}_{i,j \le d_T}$, and for DGPs 13 and 14 we let $A(t/T) = (1 - t/T)A_1 + (t/T)A_2$, where the matrices $A_1 = \{.2^{|i-j|+1}\}_{i,j \le d_T}, A_2 = \{.4^{|i-j|+1}\}_{i,j \le d_T}$. Define $\text{LGT}(\gamma, c, t/T) = (1 +$

$\exp(-\gamma(t/T-c))^{-1}$, time variation in the coefficients is modeled as follows:

DGP	Description	$\beta_1(t/T)$	$\beta_2(t/T)$	$\beta_3(t/T)$	$\beta_4(t/T)$	$\boldsymbol{z}_{t,T}$
1	constant	0	0	0	0	stationary
2	break in error variance	0	0	0	0	stationary
3	early break, $T_b = 50$			$-1(t > T_b)$		stationary
4	mid break, $T_b = 100$			$-1(t > T_b)$		stationary
5	late break, $T_b = 150$			$-1(t > T_b)$		stationary
6	small random walk			$\Delta\beta_j(t/T) \sim N(0, \frac{.5}{\sqrt{T}})$		stationary
7	big random walk			$\Delta\beta_j(t/T) \sim N(0, \frac{1}{\sqrt{T}})$		stationary
8	smooth, $c = .25$	LGT(10,c)	LGT(5,c)	LGT(20,c)	LGT(10,c)	stationary
9	smooth, $c = .75$	LGT(10,c)	LGT(5,c)	LGT(20,c)	LGT(10,c)	stationary
10	smooth, $c = .90$	LGT(10,c)	LGT(5,c)	LGT(20,c)	LGT(10,c)	stationary
11	steep	$-.3(\frac{t}{T})^2$	$(\frac{t}{T})^2$	$-.4(\frac{t}{T})$	$\frac{t}{T}$	stationary
12	exotic	0	0	$3\cos(\frac{2\pi t}{T})$	$2\frac{t}{T}\sin(2\pi\frac{t}{T})$	stationary
13	smooth, $c = .75$	LGT(10,c)	LGT(5,c)	LGT(20,c)	LGT(10,c)	locally stationary
14	late break, $T_b = 150$			$-1(t > T_b)$		locally stationary

For all DGPs, we report results when generating the innovations as $\boldsymbol{\eta}_t \overset{iid}{\sim} N(0, I_{d_T})$ or from a $t_5(0, 3/5 * I_{d_T})$. For DGP 2, we have a break in the error variance: $\epsilon_t = D(0,1)(t < 150) + D(0, 2.5)(t \geq 150)$, where the distribution D is either a normal or t_5 distribution. For the remaining DGPs $\epsilon_t \overset{iid}{\sim} N(0,1)$ or $\overset{iid}{\sim} t_5$.

1.6.1 Methods and Forecast Design

We consider the forecasting performance of the following methods:

- LC-Boost, LL-Boost.
- L_2 Boosting, with time invariant coefficients estimated on the full sample.
- Lasso, AR(3) both estimated on the full sample.

- Rolling window L_2 Boosting, Rolling window AR(3) model both with window length $T/5$.

AR and rolling window AR models are commonly used benchmarks in macroeconomic forecasting. Models estimated on the full sample assume time invariant parameters, or more generally assume the time variation is small. Estimation using LC-Boost, LL-Boost or a rolling window approach involves using a subsample of the data leading to a bias-variance tradeoff. Due to this tradeoff, methods accounting for time variation are not guaranteed to outperform their time invariant counterparts in a finite sample setting.

All boosting models are computed using the R package **mboost**, and the lasso model is computed using the R package **glmnet**. For LC-Boost and LL-Boost, we use the uniform kernel and we estimate the bandwidth via the cross validation procedure described in section 1.4, with $\omega = 20$, and $B = [.3, .4, \ldots, 1]$. The number of steps in all boosting models is determined using AIC with the maximum number of steps set to $M_{upp} = 100$. Lastly, the penalty parameter in the Lasso model is estimated using the BIC statistic.

For each simulation, and for all methods, we forecast $Y_{T,T}$ and compute the out of sample forecast error, which is then averaged over 1000 simulations. Specifically, for a given simulation, let $\hat{Y}_{T,T}^{(k)}$ represent the out of sample forecast of $Y_{T,T}$. We then compute MSFE$=\frac{1}{1000}\sum_{k=1}^{1000}(\hat{Y}_{T,T}^{(k)} - Y_{T,T}^{(k)})^2$, for each method. We report this MSFE relative to the MSFE obtained from L_2 boosting model with time invariant coefficients.

1.6.2 Results

The results for Gaussian innovations are in table 1.1. The results for t_5 innovations are contained in the appendix. We first dicuss results for the Gaussian case. DGP 1 and 2 contain time invariant coefficients, with DGP 2 having a structural break in the variance of the noise. In both these DGPs using the full sample yields the best estimator. LC-Boost only has a minor error inflation compared to using the whole sample, whereas LL-Boost does worse than LC-Boost in this setting. The under performance of LL-Boost vs LC-Boost in these settings is likely due to the bias-variance tradeoff when using local linear vs local constant methods. If the time variation is non-existent or mild, as is the case here, the additional variance incurred by estimating more parameters can cancel out any benefit obtained from bias reduction. DGP $3, 4, 5$ all contain a discrete structural break, and we see that both LC-Boost and LL-Boost outperform other methods. When the structural break occurs near the end of the sample, LL-Boost has large gains over LC-Boost.

DGP 6 has a slowly varying random walk, and we observe that LC-Boost and LL-Boost perform slightly better than using the full sample. DGP 7 has larger time variation in the coefficients, and we see that LL-Boost and LC-Boost easily outperforms the other methods. DGP 8, 9 and 10 have smooth transition logistic functions, where c is the analogous to the breakpoint in a discrete break model, and γ represents smoothness of the transition.[17] Out of the three DGPs, time varying methods perform best when $c = .75$, with the performance deteriorating in the other two cases as the time variation occurs either too close to the forecast date or too far away. DGP 11 and 12 contain coefficient functions which are highly non-linear, and LL-boost shows very large improvements vs LC-Boost. DGP 13 and 14 show that adding lo-

[17] We note that setting γ to infinity results in a discrete break model.

cally stationary predictors leads to only slight change in the results vs DGP 9 and 5 respectively.

When we have t_5 innovations, the results generally follow the conclusions stated earlier, except the improvements are noticeably smaller in many cases. The presence of additional noise in the data likely impacts our method in two ways: due to the additional noise in the data, the bias-variance tradeoff is less favorable to using a subset of the full sample. Additionally, the noise in the data makes the cross validation error estimate less reliable, leading to errors in estimating the optimal bandwidth parameter.[18]

The results suggest the following conclusions:

1. When the time variation in the coefficients is non-existent or minor, using the full sample often gives the best performance. The performance of LC-Boost is only marginally weaker than using the full sample, while the performance of LL-Boost takes a more significant hit.

2. LL-Boost and LC-Boost forecasts both seem to underperform forecasts using the full sample when there is a break of in the conditional variance rather than the conditional mean.

3. Using LL-Boost leads to large improvements in forecasting performance vs LC-Boost when we have significant time variation in the coefficients. This is especially true when the time variation occurs closer to the forecast date and/or the coefficient functions are highly non-linear.

4. Time varying methods are likely to be less useful when we have a low sample

[18] We also repeated each of the simulations using the Gaussian kernel instead of the uniform kernel. In general we found very similar performance between the two kernels. For the case of t_5 innovations and little to no time variation in the coefficients, we found the Gaussian kernel was more effective for LL-Boost. Given the close similarities between the kernels, we omit the results.

Table 1.1: Relative MSFE, Gaussian Innovations

DGP	AR (3)	Rolling AR (3)	Rolling Boost	LC-Boost	LL-Boost	Lasso
1	2.22	2.41	1.92	1.05	1.19	1.06
2	1.79	1.79	1.42	1.08	1.16	1.23
3	1.16	1.24	1.01	.61	.67	1.14
4	.91	.98	.80	.55	.58	1.02
5	.72	.78	.63	.76	.53	.92
6	5.25	5.93	1.62	.91	.90	1.06
7	3.47	3.75	.96	.68	.60	1.06
8	4.92	5.30	1.53	.59	.56	1.13
9	1.94	2.08	.73	.52	.35	1.20
10	1.77	1.88	.95	.79	.53	1.22
11	2.81	3.10	.99	.75	.61	1.15
12	1.04	1.09	.32	.63	.16	1.15
13	2.11	2.20	.83	.63	.39	1.20
14	.73	.78	.67	.80	.52	.97

size coupled with high noise. Some of the difficulties in this setting may be overcome by selecting the bandwidth parameter using a larger validation set along with a finer grid of bandwidth values.

1.7 Application to Macroeconomic Forecasting

As discussed in the introduction, the parameter instability of various macroeconomic series has long been established in the econometrics literature. Some examples include Stock and Watson (1996, 2009); Breitung and Eickmeier (2011), all of which find instability in either the univariate relationship of a large number of series or in the factor loadings of a dynamic factor model of a large panel of macroeconomic series. Similarly, Stock and Watson (2003) and Rossi and Sekhposyan (2010) have found evidence of instability in the predictive ability of various series in forecasting output and inflation. However, the question of whether forecasts can be improved by modeling parameter instability, especially when using high dimensional predictors, is

far less clear.

Proponents of modeling parameter instability include works such as Clements and Hendry (1996) which argue that ignoring these instabilities are the main sources of forecast breakdowns. On the other hand, empirical evidence in favor of ignoring instabilities include Stock and Watson (1996) which had shown there is little benefit to modeling time variation in a wide range of autoregressive and bivariate forecasts, and Kim and Swanson (2014); Koop (2013) which showed forecasts estimated by recursive estimation (using the full sample) performed as well as or better than rolling window forecasts for a range of models estimated from a large panel of macroeconomic series. Additionally, a number of works such as Pettenuzzo and Timmermann (2017); Koop and Korobilis (2013); Eickmeier et al. (2015), have estimated TVP models using Bayesian methods and their results suggest that TVP models offer only minor improvements in the accuracy of point forecasts when compared to low dimensional constant parameter models.[19] Lastly, on the theoretical side, Bates et al. (2013) has shown the standard principal components estimator remains consistent even in the presence of "small" breaks and/or mild time variation in the factor loadings of a dynamic factor model.

To illustrate the difficulty of exploiting parameter instability, consider a simple example where there is a single discrete structural break in the forecasting model. Even if the researcher knew the precise date of the break and decided to use only post break observations for estimation there is a bias-variance trade off in using less data for estimation (Pesaran and Timmermann, 2007). Therefore in the presence of

[19]These works did find TVP models produced larger improvements to density forecasts. We note that works such as Koop and Korobilis (2012); Groen et al. (2013); Chan et al. (2012) have also estimated TVP models using Bayesian methods and found significant improvements to point forecasts when compared to a low dimensional constant parameter benchmark. However, these works are restricted to forecasting inflation with low dimensional predictors.

small instabilities, such as small breaks or very slowly varying coefficients, using the entire sample through recursive estimation can be more beneficial than using only a subset of the data. Due to this bias-variance tradeoff and the uncertainty around the precise nature of time variation, the majority of works on macroeconomic forecasting tend to use the full sample available when forecasting. Furthermore, these issues are more severe when using high dimensional predictors.

Given the above discussion, we use the methods developed in this paper to answer a number of questions such as:

- Does modeling parameter instability improve macroeconomic forecasts?
- Which models are best able to deal with underlying parameter instability?
- Which variables and forecast horizons benefit most from the use of time varying parameter models?
- During which time periods do time varying methods perform best?

To answer these questions, we use the August 2018 (2018:8) vintage of the FRED-MD database which contains 128 monthly macroeconomic series collected from a broad range of categories. See McCracken and Ng (2016) for a more detailed description of each series, as well as transformations needed to achieve approximate stationarity.[20] We remove 5 series which contain large amounts of missing values, leaving us with 123 monthly macroeconomic series which run from January 1960 to August 2018. We focus our analysis on 8 major macroeconomic series: Industrial Production (IP), Total Nonfarm Payroll (PAYEMS), Unemployment Rate (UNRATE), Civilian Labor Force (CLF), Real Personal Income Excluding Transfer Receipts (RPI), Consumer Price Index (CPI), Effective Fed Funds Rate (FF), and Three Month Treasury

[20]We depart from the recommended transformations for the housing series (Group 4) which we treat as I(1) in logs.

Bill (TB3MS). For each series, we compare the out of sample forecasting performance of several models at the $h = 1, 3, 6, 12$ month forecasting horizons.

1.7.1 Methods and Forecast Design

For all the methods we consider, let $Y^h_{t,T}$ denote our h-step ahead target variable to be forecast. As an example, for CPI our target variable is $Y^h_{t,T} = \frac{1200}{h} log(\frac{CPI_t}{CPI_{t-h}})$, and we define the target similarly for the rest of the series except FEDFUNDS and TB3MS which are modeled as $I(1)$ in levels (i.e. $Y^h_{t,T} = \frac{12}{h}(\text{FEDFUNDS}_t - \text{FEDFUNDS}_{t-h})$). Next let $\boldsymbol{z}_{t-h,T}$ denote the rest of our 122 predictor series at time $t-h$, and let $\boldsymbol{x_{t-h}} = (Y_{t-h,T}, \ldots, Y_{t-h-3}, \boldsymbol{z_{t-h,T}}, \ldots, \boldsymbol{z_{t-h-3,T}})$ where $Y_{t-h,T} = Y^1_{t-h,T}$.

For all time varying methods we estimate the bandwidth using the cross validation procedure detailed in section 1.4. For selecting the bandwidth we use a grid of values from .3 to 1 with increments of .025 i.e. $B = [.3, .325, \ldots., 1]$, and we use the last $\omega = 60$ observations as our validation set.[21] Additionally, we estimate all models under consideration using time invariant methods in order to assess the benefits of directly modeling time variation. We evaluate the forecasting performance of the following methods:

Method	Parameter	Predictors considered
AR	time invariant	$(Y_{t-h,T}, \ldots, Y_{t-h-3})$
TV-AR	local constant	$(Y_{t-h,T}, \ldots, Y_{t-h-3})$
Boost	time invariant	$\boldsymbol{x_{t-h}}$
Lasso	time invariant	$\boldsymbol{x_{t-h}}$
LC-Boost	local constant	$\boldsymbol{x_{t-h}}$
LL-Boost	local linear	$\boldsymbol{x_{t-h}}$
LC-Boost-Factor	local constant	$(Y_{t-h,T}, \ldots, Y_{t-h-3}, \boldsymbol{F_{t-h,T}}, \ldots, \boldsymbol{F_{t-h-3,T}})$
LL-Boost-Factor	local linear	$(Y_{t-h,T}, \ldots, Y_{t-h-3}, \boldsymbol{F_{t-h,T}}, \ldots, \boldsymbol{F_{t-h-3,T}})$
DI	time invariant	$(Y_{t-h,T}, \ldots, Y_{t-h-3}, \boldsymbol{F_{t-h,T}})$
Boost Factor	time invariant	$(Y_{t-h,T}, \ldots, Y_{t-h-3}, \boldsymbol{F_{t-h,T}}, \ldots, \boldsymbol{F_{t-h-3,T}})$

[21] For all local constant methods we report results using the uniform kernel, the results are very similar if we use the Gaussian kernel. For local linear methods we use the Gaussian kernel.

The last four methods are of the following form:

$$Y_{t,T}^{h} = \alpha(t/T) + \sum_{j=0}^{3} \alpha_j(t/T) Y_{T-h-j,T} + \sum_{j=0}^{l} \boldsymbol{\beta}_j'(t/T) \boldsymbol{F}_{t-h-j,T} + \epsilon_t, \tag{1.9}$$

where $\boldsymbol{F}_{t-h,T} = (F_{1,t-h,T}, \ldots, F_{k,t-h,T})$ is a k-dimensional vector of factors which are estimated using the principal components of our 122 predictor series $\boldsymbol{z}_{t-h,T}$. We ignore possible time variation in our predictors when estimating our factors, and rely on results showing the consistency of the principal components estimator under mild time variation and structural breaks in the factor loadings (Bates et al., 2013). We instead focus on modeling the time variation in the coefficients of the forecasting equation (1.9). As an example, for LC-Boost Factor we set $k = 8$, $l = 3$ and estimate the model using our LC-Boost algorithm. And for DI we set $k = 4$, $l = 0$ and estimate the model assuming time invariant coefficients and utilizing the full sample.[22]

Remark 1. We note that constant parameter versions of high dimensional methods (e.g. Boost, Boost Factor, Lasso) have greater adaptability to time variation than low dimensional regressions which assume the set of relevant predictors/factors is fixed over time. The idea is that by combining information from a large set of predictors our forecasts are more robust to instabilities which occur in a specific predictor's forecasting ability.[23] When combined with a recursive window forecasting scheme, these methods indirectly capture at least some of the time variation present in the data.

We use an expanding (recursive) window scheme designed to simulate real time

[22] Additionally, Stock and Watson (2009) conjectured, for macroeconomic data, that the time variation in the coefficients $\boldsymbol{\beta}(t/T)$ is far more important than possible time variation in the factors. Their empirical results showed in sample estimates of the factors as well in-sample forecasting results were little changed by allowing for a one time break in the factors.

[23] Empirical evidence of this was provided in Carrasco and Rossi (2016).

forecasting. Our out of sample forecasting period starts in 1971:9 and ends in 2018:8 for a total of 564 months (47 years). To construct the first forecast of time t=1971:9 we estimate the factors, the coefficients, and select the hyperparameters using data available only until time 1971:9-h. We then expand our window by one observation and estimate the forecast of time $t+1$=1971:10 using information available until time $t-h+1$, and so on until we reach the end of our sample.

1.8 Results

Our benchmark model for all series and forecasting horizons is an AR(4) model with time invariant parameters. Due to space considerations we report some of our results in the appendix. We start by giving an overview of the results for the full out sample period, which are reported in table 1.2 for $h = 12$ and in the appendix for $h = 6, 3, 1$, before analyzing how performance varies over time. For the time varying methods we observe the following: the TV-AR model fails to improve upon the benchmark AR model for the vast majority of series and forecast horizons, confirming the results of Stock and Watson (1996) on an expanded sample. Out of our four time varying Boosting methods, LC-Boost Factor appears to perform best. LC-Boost Factor outperforms the benchmark for all series and forecast horizons, it also performs best, out of all models, the majority of times. In contrast, our LL-Boosting methods appear to perform poorly relative to LC-Boost.[24] Given the results in section 1.6, this suggests that the parameters as a function of time may not be sufficiently curvy enough for local linear methods to benefit. For time invariant methods we observe the following: Boost-Factor and Boost performs similarly and generally outperform DI and Lasso models.

[24]We omit the performance of LL-Boost as it was outperformed by both LL-Boost Factor and both LC-Boost methods.

Comparing across forecast horizons: we observe that, for all high dimensional methods, improvements to the benchmark are greater as we increase our forecast horizon. For $h = 1$, many of the methods appear to perform similarly, with Boost Factor and LC-Boost Factor appearing to perform best. For longer forecast horizons, LC-Boost Factor is the best performing model the majority of the time, with the gap between LC-Boost Factor and its competitors widening as we increase the forecast horizon. Additionally, the benefits to modeling time varying parameters are more apparent at longer forecast horizons.

1.8.1 Analyzing Performance Over Time

Relying only on the aggregate performance of a model over the entire out of sample period can hide many important details and lead to misleading conclusions. We rely on two methods to analyze how performance varies over time; the first is to plot the MSFE as a function of the start date for the out of sample forecasting period. More specifically, let T_1 denote the start forecast date, then for a given method i and horizon h, we calculate

$$MSFE^h_{(i)}(T_1, T_2) = \frac{\sum_{t=T_1}^{T_2} \hat{\epsilon}^2_{t,(i)}}{\sum_{t=T_1}^{T_2} \hat{\epsilon}^2_{t,(AR)}}, \qquad \text{with} \quad T_2 = 2018:8. \tag{1.10}$$

The second method is to analyze the forecasting performance over three important subperiods. The first subperiod, which we refer to as “Pre-Great Moderation", consists of 136 observations, 1971:9-1982:12, and corresponds roughly to the period before the start of the “Great Moderation". The second subperiod is from 1983:1-2006:12, and corresponds roughly to the “Great Moderation", a period where the volatility of a large number of macroeconomic series was significantly reduced (Stock and Watson, 2002b). The third subperiod is from 2007:1-2018:8, which we refer to as “Post Great

Moderation", covers the period right before the great recession and takes us to the end of our sample.

For the first method, we let T_1 vary from 1971:9 until 2006:12. We plot the MSFE by T_1 for the top 5 performing methods: LC-Boost Factor, LC-Boost, Boost, Boost Factor, and DI. The figures 1.1-1.3; contain the results for horizons $h = 12, 6, 1$ respectively.[25] Looking at figures 1.1-1.3, we see that LC-Boost Factor is easily the best performing method for horizon $h = 12, 6$, and to a lesser extent $h = 1$. Comparing across all horizons, we notice:

- The performance improvements for LC-Boost factor, relative to its time invariant counterparts, are more apparent as we increase the forecast horizon.
- As we increase T_1, the gap between LC-Boost Factor and the time invariant methods widens. In particular we notice a large separation in performance starting during great moderation period.

Additionally, we also observe that the commonly used DI model loses much of its predictive ability during the Great moderation and performs worse than the benchmark for about half of the series. This result suggests that DI gained most of its predictability vs the benchmark during the "Pre-Great Moderation" period.

Table 1.2 contain the results for each of the subperiods for horizon $h = 12$; the corresponding results for $h = 6, 3, 1$ are found in the appendix. For each subperiod we report the MSFE, relative to the MSFE of the benchmark AR(4) model. We start with the "Pre-Great Moderation" period, and note that with the exception of TV-AR models, all other models strongly outperform the benchmark model the majority of the time during this period. Time invariant methods such as Boost Factor and DI models perform best, and their performance is strongest when forecasting at longer

[25] The corresponding results for $h = 3$ are reported in the appendix.

horizons. LC-Boost factor appears to be slightly lag behind these two methods during this time period. As we enter the “Great Moderation" period, the performance of all models generally declines relative to the AR benchmark. In particular, time invariant methods such as DI and Boost take a large hit and underperform the benchmark in many cases, especially for $h = 12$. LC-Boost Factor undergoes a much smaller decline compared to the rest of the models, and emerges as the best performing model during this time period. Importantly, we also observe that LC-Boost performs at the same level or worse than its time invariant counterpart Boost in the majority of cases. This suggests that although there seems to be a large amount of time variation in this period, the bias variance tradeoff in modeling it is not favorable to a model with a large amount of potential predictors (~ 500 predictors).

During the “Post Great Moderation" period we notice two interesting developments: The performance of LC-Boost methods show large improvements over both the benchmark AR model, and their time invariant counterparts, for all forecast horizons, with the improvement being greatest for longer horizons. On the other hand, time invariant methods experience smaller improvements, and in many cases their performance worsens compared to the Great Moderation period.

1.8.2 Assessing Benefits of Modeling Time Varying Parameters

In order to assess the benefits of *directly* modeling parameter instability, we compare the performance of LC-Boost Factor vs Boost-Factor. These also happen to be the best time varying and time invariant methods respectively. We start our analysis by first plotting the MSFE of LC-Boost Factor, *relative to the MSFE of Boost Factor*, as a function of the start date for the out of sample forecast period, i.e.

$MSFE_{(LCBoostFactor)}(T_1, T_2)/MSFE_{(BoostFactor)}(T_1, T_2)$. The results are in figure 1.4. We observe that for all series and forecast horizons LC-Boost Factor almost never performs worse than Boost Factor, and outperforms it the vast majority of the time. Furthermore, the gap between the two methods widens as we increase the start date of the out of sample period, and as we increase the forecast horizon. For example, if we consider horizon $h = 12$, and we start the out of sample period in the early 1990's, LC Boost offers, on average, over a 20 percent improvement over Boost-Factor. We observe similar patterns, although the improvements are not as large ($\sim$ 10-15 percent on average), for horizons $h = 3, 6$. An exception seems to be for $h = 1$, which shows little improvements for the majority of series, with the exceptions coming from the two interest rate series and EMS.

Next, we attempt to get a finer look at how the benefits of modeling parameter instability vary over time. We first define the *local* MSFE (L-MSFE), of method i at time t_0 as :

$$\text{L-MSFE}_i(t_0) = \frac{\sum_{t=t_0-\Delta}^{t_0+\Delta} \hat{\epsilon}^2_{t,(i)}}{\sum_{t=t_0-70}^{t_0+\Delta} \hat{\epsilon}^2_{t,(AR)}}, \qquad \text{RL-MSFE}_i(t_0) = \frac{\text{L-MSFE}_i(t_0)}{\text{L-MSFE}_{\text{BoostFactor}}(t_0)} \quad (1.11)$$

with the convention that $\hat{\epsilon}_{t,(i)} = 0$ for $t \leq 0, t \geq T$. This amounts to using a uniform kernel to weight the forecast errors with a bandwidth chosen such that the window size has $\Delta = 70$ observations. We then plot RL-MSFE$_i(t_0)$ for $i =$LCBoostFactor, for $t_0 = 1977{:}3, \ldots, 2012{:}10$ for all series and forecast horizons. The endpoints are chosen so that the first and last values in the plot correspond to the RL-MSFE during the "Pre-Great Moderation" and "Post-Great Moderation" periods respectively.

The results are in figure 1.5, and we observe that the first value is usually near or above one for all variables except for CLF (Civilian Labor Force). This suggests that

during the “Pre-Great Moderation" there seems to be little or no benefit to modeling time variation. This can reflect either a lack of underlying parameter instability during this time period, or the relatively low sample size available combined with high volatility made it difficult to exploit the time variation present. During the Great Moderation period, almost all series experience large declines in RL-MSFE, with the exact timing of the decline differing by series. For IP and RPI the benefits to modeling parameter instability appear to decrease from their mid 1990's levels, while for the rest of the series we see further improvements until the end of the sample. These results suggest that there is a large amount of parameter instability which started during the Great moderation period and continued though the sample.

Lastly, we attempt to examine the degree and timing of time variation by examining the bandwidth values selected. We define the local bandwidth of LC-Boost Factor as:

$$\text{L-BW}(t_0) = \sum_{t=t_0-\Delta}^{t_0+\Delta} \hat{b}_{t_0}/(2\Delta), \tag{1.12}$$

with the convention that $\hat{b}_{t_0} = 0$ for $t \leq 0, t \geq T$. Recall that $\hat{b}_{t_0}$ is the bandwidth chosen at time t_0. Since we are using the uniform kernel, $\hat{b}_{t_0}$ represents the fraction of the sample available at *time* t_0 that we are using for estimation. As as example, at time t_0=1977:3 we have a total of 196 observations available for estimation, therefore a value of $\hat{b}_{t_0} = .61$ for t_0=1977:3 implies we are using ≈ 120 observations. We set $\Delta = 70$ observations, and then plot L-BW(t_0) for $t_0 = 1977{:}3, \ldots, 2012{:}10$ for all series and forecast horizons. As an additional comparison we also plot the local bandwidth implied by a rolling window estimator which uses a fixed window length of 120 observations.

The results are seen in figure 1.6, and we notice that for the pre Great Moderation

period the local bandwidths are usually between .7-.8 for most series. As we enter the Great Moderation we notice that the local Bandwidths generally tend to increase initially before declining. However, we notice the timing and degree of declines differs by series. For some series such as IP and RPI, the local BW tends to increase after reaching their lows in the mid 1990s, whereas for other series such as UNRATE, FEDFUNDS, and TB3MS the local BW start their decline in the 1990s. In contrast, we see that using a fixed rolling window of 120 observations implies a monotonically decreasing bandwidth and assumes the same bandwidth regardless of series or horizon. To determine the importance of estimating the optimal bandwidth via cross valiation, we compare the local MSFE of LC-Boost Factor to Boost Factor estimated using a 120 observation rolling window in the appendix. The results show that for the vast majority of series and horizons the rolling window estimator is strongly outperformed by LC-Boost Factor with the largest out performance occurring during the Great Moderation period.

Overall our results suggest the following conclusions:

1) Parameter instability starts to appear around the beginning of the Great Moderation period. This instability seriously deteriorates the relative forecasting performance of time invariant methods; with the effect being more severe for longer horizon forecasts.

2) Due to the large improvements in point forecasts from our methods, it is likely that this instability has a substantial impact on the conditional mean as well as the variance of various economic series.

3) Lastly, there are large benefits to modeling parameter instability if done properly. Given the high bias variance tradeoff encountered in using a reduced sample size, these benefits can easily be missed. For example, models such as

LC-Boost have more difficulty in learning the time variation in the data due to the large amount of potential predictors.

4) The commonly used rolling window estimation method can understate the benefits of modeling parameter instability by failing to account for differences in the degree of parameter instability by series, forecast horizon, and time period.

To elaborate more on point **3)** above, we compare the L-MSFE of the following models in the appendix: LL-Boost vs LC-Boost, LC-Boost vs LC-Boost Factor, and LC-Boost vs Boost. We see from the results that LL-Boost Factor was strongly outperformed by LC-Boost Factor in the earlier parts of the sample, suggesting that there was little time variation during the pre-great Moderation period. As our sample size available for estimation increases we see the performance of LL-Boost factor improve to the point where it does as well as or outperforms LC-Boost factor in about half of the series, especially for longer horizons. Compared to LC-Boost Factor, we observe that the benefits of modeling time variation via LC-Boost are smaller and are realized far later in the sample. As an example, for $h = 12$ we notice that LC-Boost performs worse than Boost during most of the Great moderation period. Additionally, for many of the series, the improvements of LC-Boost over Boost start to occur near the end of the great moderation period. In contrast, LC-Boost Factor is able to adapt to the time variation far earlier as a result of having a more favorable bias variance tradeoff.

1.9 Conclusion

In this work, we have presented two L_2 Boosting algorithms for estimating high dimensional predictive regressions with time varying coefficient. We proved the consistency of both of these methods, and showed their effectiveness in modeling the pa-

rameter instability present in macroeconomic series. Compared to other TVP methods, our methods are very efficient computationally even for high dimensional data; a single LC-Boost forecast, including implementing the cross validation procedure, can be estimated within a matter of seconds. Additionally, they can be implemented by researchers and practitioners using the easy to use R package **mboost**. Furthermore, the boosting framework can be easily adapted to fitting more complex non-linear models.

There are many topics available for further study, one such topic is in selecting the important bandwidth parameter for our models. Although our cross validation procedure seems to perform adequately, we welcome further improvements to this methodology. Lastly, although our empirical example focused on forecasting, our models are applicable in a far broader range of settings.

Table 1.2: Relative MSFE $h = 12$

Full Out of Sample Period 1971:9-2018:8								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.04	.99	1.1	.64	.99	.86	1	1.07
DI	.79	.79	.69	.99	.84	.94	.87	.91
Lasso	.77	.81	.75	.77	.93	.96	.76	.89
Boost	.78	.73	.73	.74	.85	.88	.79	.88
Boost Factor	.75	.81	.62	.96	.80	.89	.78	.85
LC-Boost	.74	.73	.62	.66	.88	.80	.85	.92
LC-Boost Factor	.62	.64	.58	.63	.75	.77	.84	.90
LL-Boost Factor	.74	.85	.70	.76	.76	.82	1.20	1.37
“Pre-Great Moderation" 1971:9-1982:12								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.07	1	1.15	.71	.96	1.11	1.01	1.14
DI	.38	.49	.48	1.51	.61	82	.88	.89
Lasso	.30	.51	.41	1.27	.71	1.13	.76	.77
Boost	.27	.44	.43	1.24	.67	.92	.72	.75
Boost Factor	.32	.50	.43	1.34	.65	.84	.74	.80
LC-Boost	.29	.44	.38	.86	.77	1.03	.70	.79
LC-Boost Factor	.31	.54	.43	.60	.66	.94	.77	.81
LL-Boost Factor	.41	.83	.66	.86	.80	1.04	1.28	1.57
“Great Moderation" 1983:1-2006:12								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.12	1	1.11	.65	1.06	1.07	1.04	1.03
DI	1.18	1.08	.75	.88	1	1	.85	.92
Lasso	1.23	1.32	.97	.85	1.10	.90	.80	1.03
Boost	1.36	1.20	.95	.81	1.08	.91	.85	1
Boost Factor	1.25	1.16	.67	.89	1	.90	.80	.90
LC-Boost	1.43	1.26	.97	.82	1.16	.90	1.07	1.05
LC-Boost Factor	.99	.89	.71	.71	.89	1	.97	1.01
LL-Boost Factor	1.03	1	.79	.90	.86	1.14	1.16	1.22
“Post Great Moderation" 2007:1-2018:8								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	.96	.98	1	.58	.96	.41	.79	.83
DI	1.10	1.02	.91	.76	.88	1.02	.94	1
Lasso	1.15	.76	.99	.36	.95	.79	.89	1.07
Boost	1.13	.72	.91	.33	.81	.81	1.13	1.25
Boost Factor	1.04	.98	.82	.79	.77	.94	1	1.01
LC-Boost	.94	68	.60	.38	.73	.48	1.29	1.21
LC-Boost Factor	.82	.54	.66	.57	.72	.42	.92	.98
LL-Boost Factor	1.01	.69	.66	.55	.66	.33	.56	.67

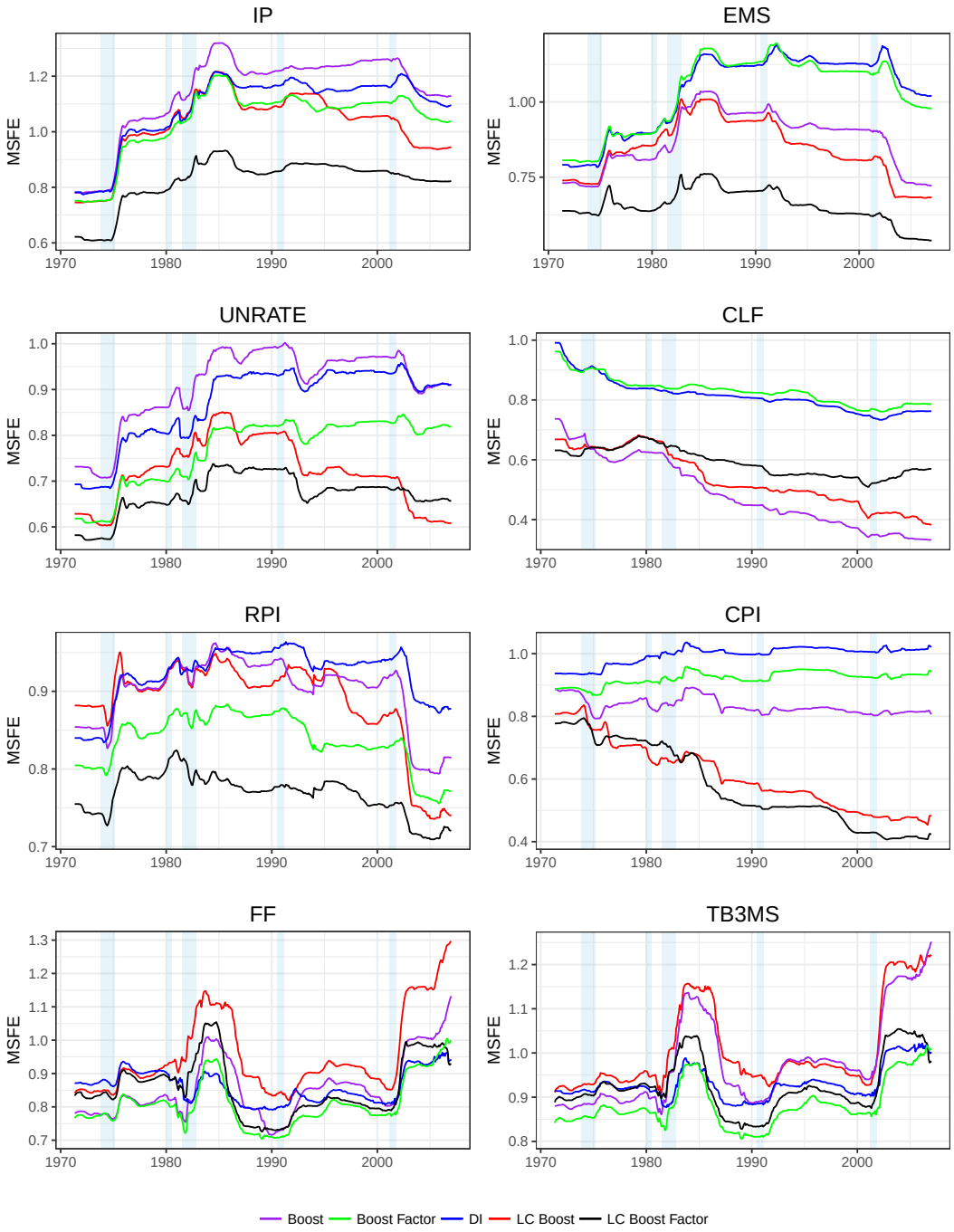

Figure 1.1: **MSFE by start date of out of sample period. Horizon** $h = 12$. More specifically we plot: $MSFE^{12}_{(i)}(T_1, T_2) = \sum_{t=T_1}^{T_2} \hat{\epsilon}^2_{t,(i)} / \sum_{t=T_1}^{T_2} \hat{\epsilon}^2_{t,(AR)}$, where we T_1 vary from 1971:9 until 2006:12, with T_2=2018:8. Shaded regions represent NBER recession dates.

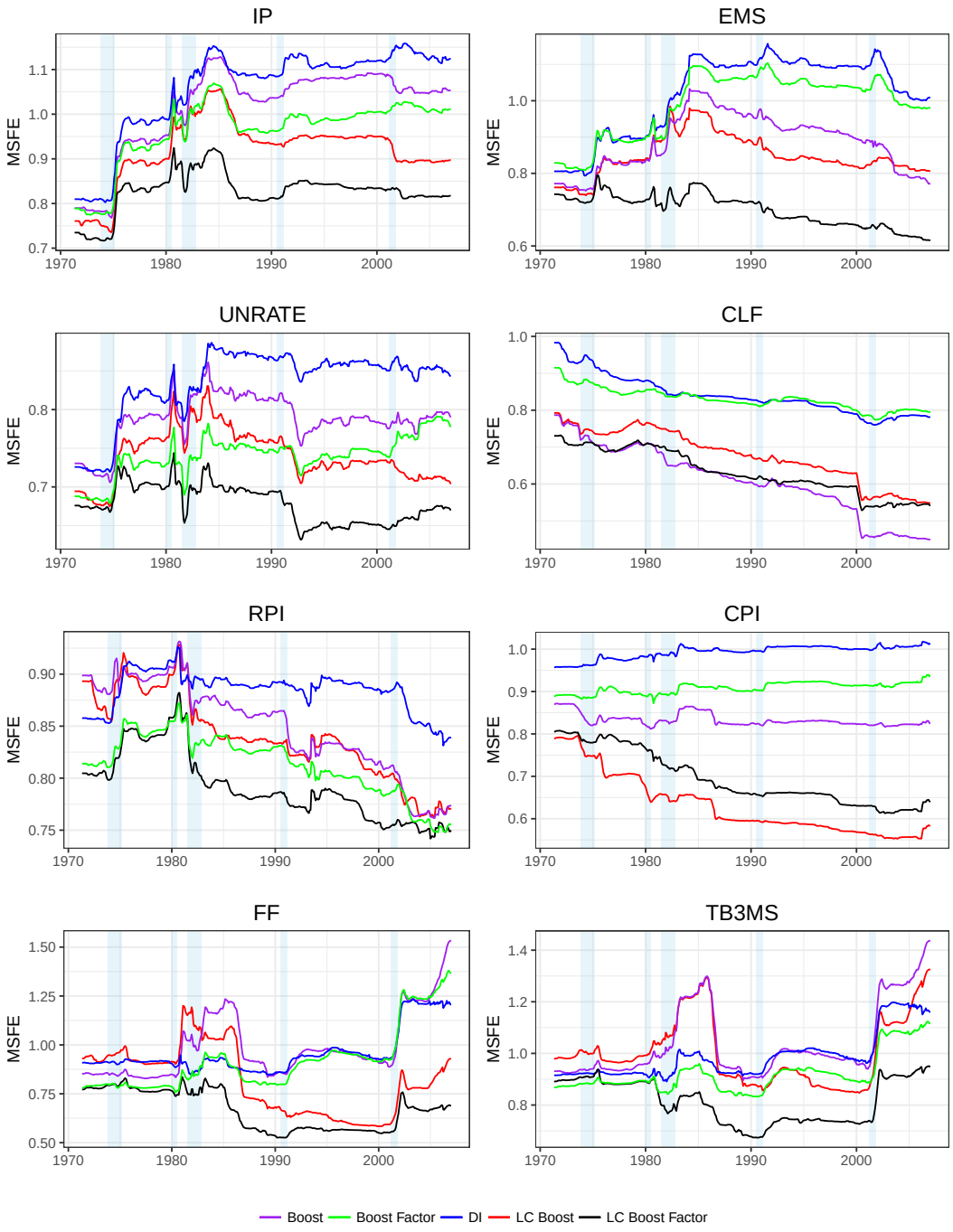

Figure 1.2: **MSFE by start date of Out of sample period. Horizon** $h = 6$. See notes to figure 1.1.

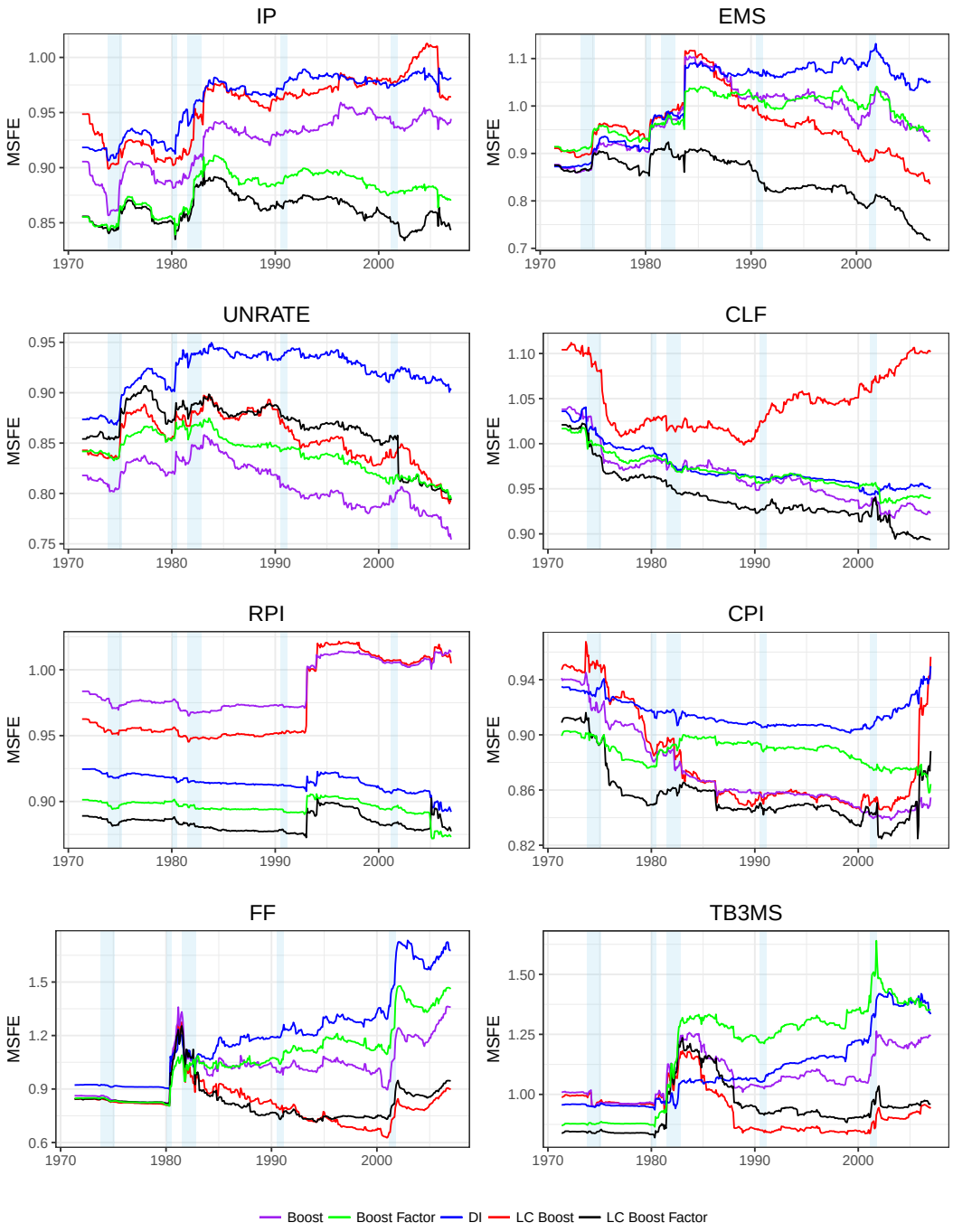

Figure 1.3: **MSFE by start date of out of sample period. Horizon $h = 1$.** See notes to figure 1.1.

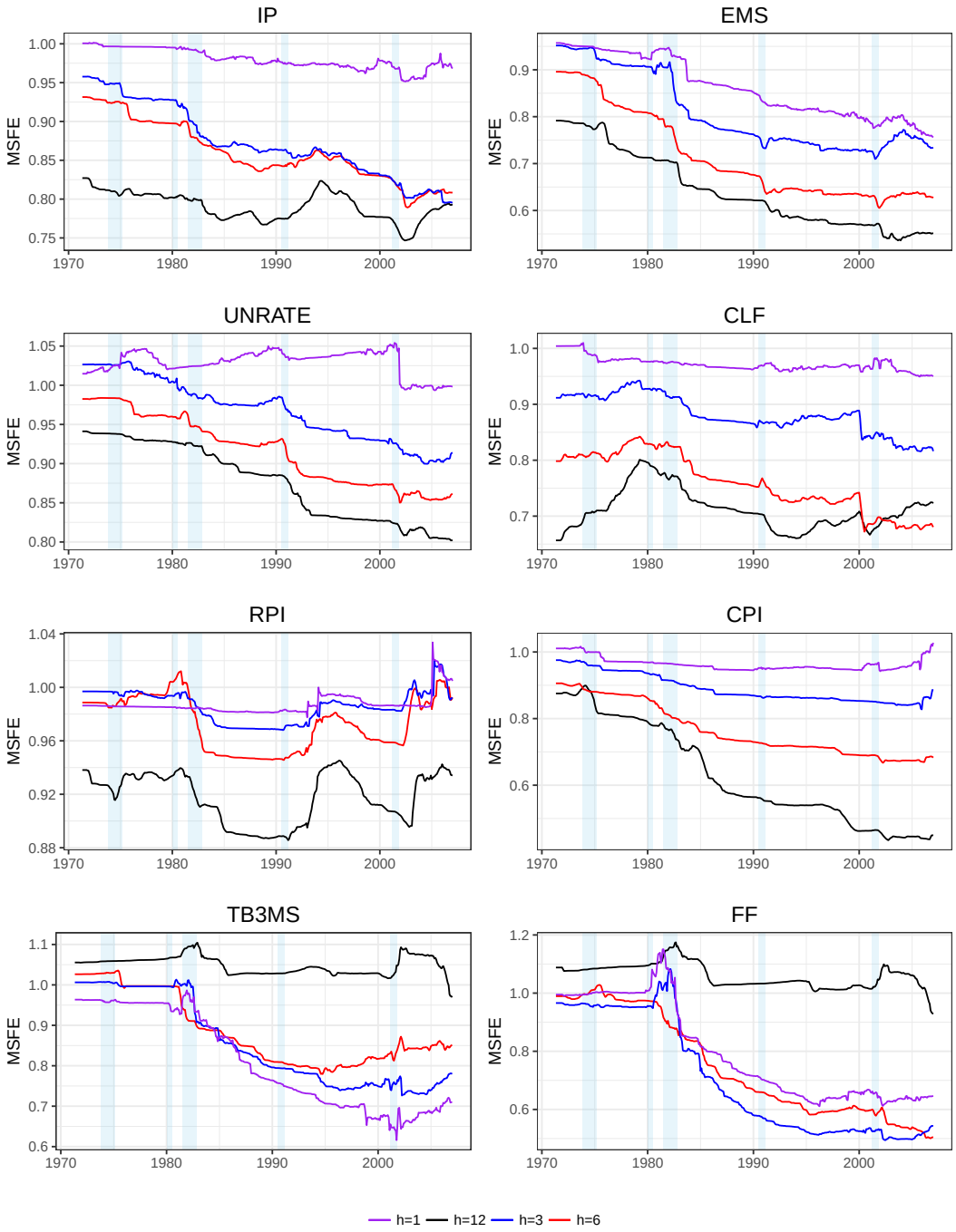

Figure 1.4: **MSFE of LC-Boost Factor (LC-BF) relative to MSFE of Boost Factor (BF) by start date of out of sample period**: See notes to figure 1.1 or equation (1.10) for details. Colored lines represent the different horizons.

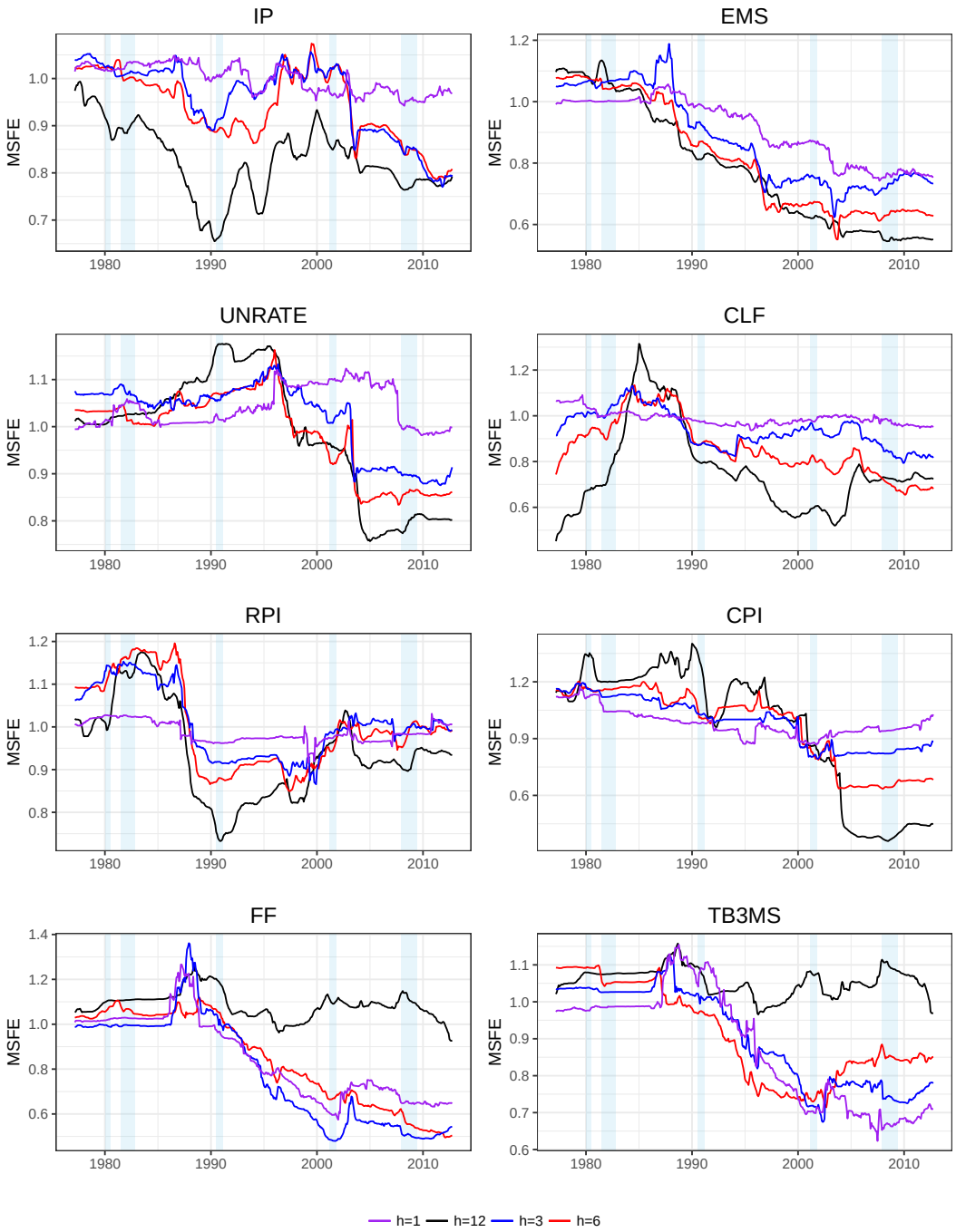

Figure 1.5: **Local MSFE of LC-Boost Factor relative to Local MSFE of Boost Factor**: See (1.11) for details.

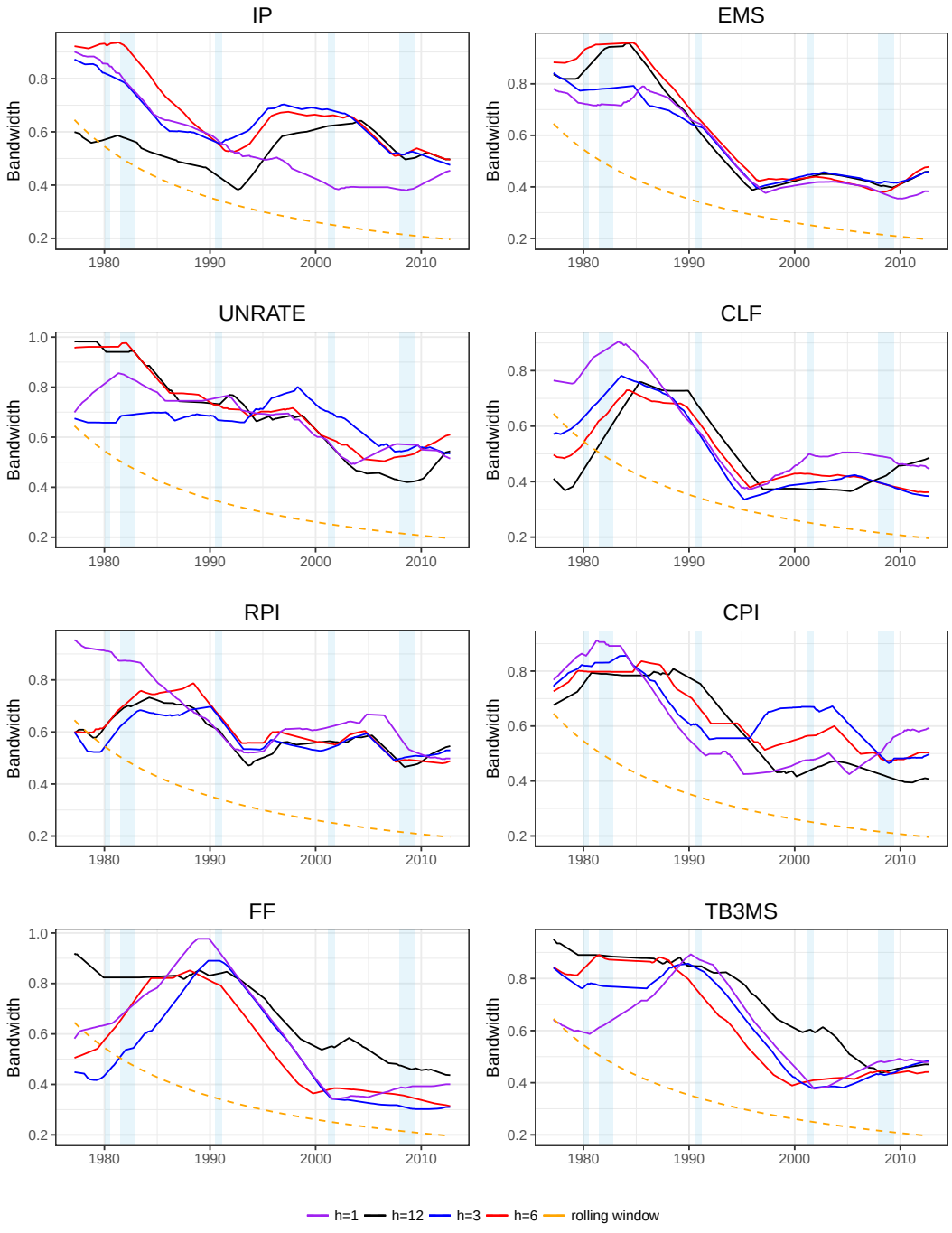

Figure 1.6: **Local Bandwidth of LC-Boost Factor**: See (1.12) for details.

Appendix A: Proofs

Proof of Theorem 1.

The proof follows the framework of Buhlmann (2006), which handled the case of boosting for iid data using linear least squares base learners.[26] The proof depends on an application of Temlyakov's result (Temlyakov, 2000) for the population version of L_2 boosting known as "weak greedy algorithm". To simplify the notation, we work with the uniform kernel and show the proof for the rescaled time point $u = T/T = 1$. The proof is almost exactly the same for more general kernels satisfying condition 3.6, and for other rescaled time points. We start by considering a step size of $\nu = 1$, and smaller step sizes can be handled as in section 6.3 of Buhlmann (2006).

We introduce the following notation: Let $\tilde{\boldsymbol{x}}_{T-h}(u)$ and $\tilde{Y}_T(u)$ be the stationary approximation to $\boldsymbol{x}_{T-h,T}$ and $Y_{T,T}$ respectively with approximation error $O_p(T^{-1}) \to 0$. Let the inner product $\left\langle \tilde{X}_{j,t}(u), \tilde{X}_{k,t}(u) \right\rangle = E(\tilde{X}_j(u)\tilde{X}_k(u))$ with $||\tilde{X}_j(u)||^2 = E(\tilde{X}_j^2(u))$. For ease of presentation let $||\tilde{X}_j(u)||^2 = 1$, $\forall j \leq p_T$. Let $f(u, \tilde{\boldsymbol{x}}_{T-h}(u)) = \tilde{\boldsymbol{x}}_{T-h}(u)\boldsymbol{\beta}(u)$ be the stationary approximation to $f(u, \boldsymbol{x}_{T-h,T})$ with approximation error $O_p(T^{-1})$. For readability, we define, for any rescaled time point $u_0 \in [0, 1]$:

$$f(\tilde{\boldsymbol{x}}_{u_0T-h}(u_0)) \equiv f(u_0, \tilde{\boldsymbol{x}}_{u_0T-h}(u_0)), \text{and } f(\boldsymbol{x}_{u_0T-h,T}) \equiv f(u_0, \boldsymbol{x}_{u_0T-h,T}).$$

We now define a sequence of remainder functions for the population version of L_2

[26] Since we refer to Buhlmann (2006) during our proof, we try, when its possible, to keep the notation consistent with their work.

Boosting:

$$R^0 f(\tilde{\boldsymbol{x}}_{T-h}(u)) = f(\tilde{\boldsymbol{x}}_{T-h}(u)),$$
$$R^m f(\tilde{\boldsymbol{x}}_{T-h}(u)) = R^{m-1} f(\tilde{\boldsymbol{x}}_{T-h}(u)) - \left\langle R^{m-1} f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_{\mathcal{S}_m,T-h}(u)\right\rangle \tilde{X}_{\mathcal{S}_m,T-h}(u),$$

Where $\mathcal{S}_m = \text{argmax}_j |\left\langle R^{m-1} f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_{j,T-h}(u)\right\rangle|$. Given that this criterion is sometimes infeasible to realize in practice, a weaker criterion is: Choose any $\mathcal{S}_m$ which satisfies:

$$|\left\langle R^{m-1} f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_{\mathcal{S}_m}(u)\right\rangle| \geq b * \sup_j |\left\langle R^{m-1} f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_j(u)\right\rangle|, \text{ for some } b \in (0,1] \tag{1.13}$$

We then obtain:

$$f(\tilde{\boldsymbol{x}}_{T-h}(u)) = \sum_{j=0}^{m-1} \left\langle R^j f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_{\mathcal{S}_{j+1}}(u)\right\rangle + R^m f(\tilde{\boldsymbol{x}}_{T-h}(u)),$$
$$||R^m f(\tilde{\boldsymbol{x}}_{T-h}(u))||^2 = ||R^{m-1} f(\tilde{\boldsymbol{x}}_{T-h}(u))||^2 - |\left\langle R^{m-1} f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_{\mathcal{S}_m}(u)\right\rangle|^2$$

If (1.13) is met, then we have the following bound, for the population version of L_2 Boosting, provided by Temlyakov (2000):

$$||R^m f(\tilde{\boldsymbol{x}}_{T-h}(u))||^2 \leq B(1 + mb^2)^{\frac{-b}{4+2b}} \tag{1.14}$$

with as defined in (1.13), and $\sup_{u\in[0,1]} |\boldsymbol{\beta}(u)| \leq B < \infty$.

To analyze the sample version of our LC-Boost algorithm, we introduce the fol-

lowing notation:

$$\left\langle X_{j,\cdot,T}, X_{k,\cdot,T} \right\rangle_{(T)} = \sum_{t=1}^{T} K_b(t/T-u) X_{j,t-h,T} X_{k,t-h,T} = S_T^{-1} \sum_{t=T-S_T}^{T} X_{j,t-h,T} X_{k,t-h,T}$$

$$\left\langle f, X_{k,\cdot,T} \right\rangle_{(T)} = \sum_{t=1}^{T} K_b(t/T-u) f(\boldsymbol{x}_{t-h}) X_{k,t-h,T} = S_T^{-1} \sum_{t=T-S_T}^{T} f(\boldsymbol{x}_{t-h}) X_{k,t-h,T}$$

$$||X_{j,\cdot,T}||^2_{(T)} = S_T^{-1} \sum_{t=T-S_T}^{T} X^2_{j,t-h,T}$$

As previously, we can define the sequence of sample remainder functions as:

$$\hat{R}^0_T f(\boldsymbol{x}_{T-h,T}) = f(\boldsymbol{x}_{T-h,T}),$$
$$\hat{R}^m_T f(\boldsymbol{x}_{T-h,T}) = \hat{R}^{m-1}_T f(\boldsymbol{x}_{T-h,T}) - \left\langle \hat{R}^{m-1}_T f, X_{\widehat{\mathcal{S}}_m,\cdot,T} \right\rangle_{(T)} X_{\widehat{\mathcal{S}}_m,T-h,T}, m = 1, 2, \ldots$$

where: $\hat{\mathcal{S}}_1 = \text{argmax}_j |\langle Y_{\cdot,T}, X_{j,\cdot,T} \rangle|$ and $\hat{\mathcal{S}}_m = \text{argmax}_j |\langle \hat{R}^m_T f, X_{j,\cdot,T} \rangle|$. Therefore, $\hat{R}^m_T f(\boldsymbol{x}_{T-h,T}) = f(\boldsymbol{x}_{T-h,T}) - \hat{F}^{(M_T)}_{lc}(u, \boldsymbol{x}_{T-h,T})$, is the difference between $f(\boldsymbol{x}_{T-h,T})$ and its LC-Boost estimate.

Lastly, to proceed with the proof, we define a sequence of semi-population version remainder functions as:

$$\tilde{R}^0_T f(\tilde{\boldsymbol{x}}_{T-h}(u)) = f(\tilde{\boldsymbol{x}}_{T-h}(u)),$$
$$\tilde{R}^m_T f(\tilde{\boldsymbol{x}}_{T-h}(u)) = \tilde{R}^{m-1}_T f(\tilde{\boldsymbol{x}}_{T-h}(u)) - \left\langle \tilde{R}^{m-1}_T f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_{\widehat{\mathcal{S}}_m,T-h}(u) \right\rangle \tilde{X}_{\widehat{\mathcal{S}}_m,T-h}(u),$$

The difference between the population and the semi-population remainder functions, is that the semi-population version uses selectors $\hat{\mathcal{S}}_m$ estimated from the sample. The strategy of the proof is: first we establish that the selectors $\hat{\mathcal{S}}_m$ satisfy a finite sample analogue of (1.13), which allows us to apply Temlyakov's result (1.14) to the semipopulation version. Lastly, we analyze the difference between the sample and the

semipopulation versions: $\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) - \tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))$.

We need the following lemma:

Lemma 1. *Under conditions 1.5.1, 1.5.2, 1.5.3, 1.5.4, and for κ as defined in Theorem 1, the following hold:*

1. $\sup_{j,k\leq p_T} |S_T^{-1} \sum_{t=T-S_T}^{T} X_{j,t,T} X_{k,t,T} - E(\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u))| = \zeta_{T,1} = O_p(S_T^{-\kappa})$

2. $\sup_{j\leq p_T} |S_T^{-1} \sum_{t=T-S_T}^{T} X_{j,t-h,T}\epsilon_{t,T}| = \zeta_{T,2} = O_p(S_T^{-\kappa})$

3. $\sup_{j\leq p_T} |S_T^{-1} \sum_{t=T-S_T}^{T} X_{j,t,T} f(\boldsymbol{x}_{t,T}) - E(\tilde{X}_{j,T}(u) f(\tilde{\boldsymbol{x}}_T(u)))| = \zeta_{T,4} = O_p(S_T^{-\kappa})$

4. $\sup_{j\leq p_T} |S_T^{-1} \sum_{t=T-S_T}^{T} X_{j,t-h,T} Y_{t,T}) - E(\tilde{X}_{j,T-h}(u) Y_T(u)))| = \zeta_{T,4} = O_p(S_T^{-\kappa})$

Let $\zeta_T = \max(\zeta_{T,1}, \zeta_{T,2}, \zeta_{T,3}, \zeta_{T,4}) = O_p(S_T^{-\kappa})$ and denote by ω a realization of all S_T sample points involved in estimation. The next lemma bounds the difference between the sample and population learners at step m.

Lemma 2. *Suppose conditions 1.5.1, 1.5.2, 1.5.3, and 1.5.4 hold. Then for κ as defined in Theorem 1 and on the set $\mathcal{A}_T = \{\omega : \zeta_T(w) < 1/2\}$, we have:*

$$\sup_{j\leq p_T} |\langle \hat{R}^{m-1} f, X_{j,\cdot,T} \rangle_{(T)} - \langle \tilde{R}^{m-1} f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_{j,T-h}(u) \rangle| \leq C(5/2)^m \zeta_T,$$

where C does not depend on m, T.

It's clear from lemma 1, that $P(\mathcal{A}_T) \to 1$. Which gives us the following lemma:

Lemma 3. *Suppose the conditions needed for lemma 1 hold, then for $m = m_T \to \infty$*

slow enough we have:

$$||\tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))|| = o_p(1)$$

We now analyze the term: $\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) = f(\boldsymbol{x}_{T-h,T}) - \hat{F}_{lc}^{(M_T)}(u, \boldsymbol{x}_{T-h,T})$. By the triangle inequality we obtain:

$$||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T})|| \leq ||\tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))|| + ||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) - \tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))|| \quad (1.15)$$

the first term can be handled with lemma 3. For the second term, let $A_T(m) = ||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) - \tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))||$. Using the definitions of the remainder functions, we then have a recursive relation:

$$\begin{aligned} A_T(m) &\leq A_T(m-1) + |\langle \hat{R}^{m-1} f, X_{\widehat{\mathcal{S}}_{m,\cdot,T}} \rangle_{(T)} - \langle \tilde{R}^{m-1} f, \tilde{X}_{\widehat{\mathcal{S}}_{m,T-h}} \rangle| ||X_{\widehat{\mathcal{S}}_{m,T-h,T}}|| \\ &\quad + ||\langle \tilde{R}^{m-1} f, \tilde{X}_{\widehat{\mathcal{S}}_{m,T-h}} \rangle|| ||X_{\widehat{\mathcal{S}}_{m,T-h,T}} - \tilde{X}_{\widehat{\mathcal{S}}_{m,T-h}}(u)|| \\ &\leq A_T(m-1) + C(5/2)^m \zeta_T + O(1/T) \text{ on the set } \mathcal{A}_T \end{aligned}$$

Where the last inequality follows from local stationarity and lemma 2. By the above recursive equation we obtain: $||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) - \tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))|| \leq 3^m \zeta_T C$. If we choose $m = m_T \to \infty$ slow enough (e.g $m_T = o(log(T))$), then along with lemma 3 and (1.15) we obtain:

$$||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T})|| = ||f(\boldsymbol{x}_{T-h,T}) - \hat{F}_{lc}^{(M_T)}(u, \boldsymbol{x}_{T-h,T})|| = o_p(1)$$

□

Proof of Lemma 1.

We start with (i), and we bound:

$$P(|S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t,T}X_{k,t,T} - E(\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u))| > S_T^{-\kappa}) \tag{1.16}$$

$$\leq P\Bigg(|S_T^{-1}\sum_{t=T-S_T}^{T} [X_{j,t,T}X_{k,t,T} - E(X_{j,t,T}X_{k,t,T})]| \tag{1.17}$$

$$+ |S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t,T}X_{k,t,T}) - E(\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u))| > S_T^{-\kappa}\Bigg) \tag{1.18}$$

We deal with the second term, which can be thought of as the bias. The product process $X_{j,t,T}X_{k,t,T}$ is locally stationary with the stationary approximation at rescaled time t/T being $\tilde{X}_{j,t}(t/T)\tilde{X}_{k,t}(t/T)$. One can see this by noting:

$$\begin{aligned} ||\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u)) - \tilde{X}_{j,T}(v)\tilde{X}_{k,T}(v))||_{r/2} &\leq (||\tilde{X}_{j,T}(u)||_r||\tilde{X}_{k,T}(u) - \tilde{X}_{k,T}(v)||_r \\ &\quad + ||\tilde{X}_{k,T}(u)||_r||\tilde{X}_{j,T}(u) - \tilde{X}_{j,T}(v)||_r) \\ &\leq C(|u-v|) \end{aligned}$$

We can employ the same techniques to that $||X_{j,t,T}X_{k,t,T} - \tilde{X}_{j,t}(t/T)\tilde{X}_{k,t}(t/T)|| \leq C(T^{-1})$.

Therefore by local stationarity we obtain:

$$|S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t,T}X_{k,t,T}) - E(\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u))| \leq O(S_T/T) \tag{1.19}$$

Note that if u is an interior point (i.e $u \in (b_T, 1-b_T)$ where b_T is the bandwidth), the bound improves to $O((S_T/T)^2)$. Let $\tilde{\sigma}_{jk}(u) = E(\tilde{X}_{j,t}(u)\tilde{X}_{k,t}(u))$, then by condition

1.5.3, we obtain:

$$\tilde{\sigma}_{jk}(t/T) = \tilde{\sigma}_{jk}(u) + \tilde{\dot{\sigma}}_{jk}(u)(t/T - u) + O(b_T^2)$$

where $\tilde{\dot{\sigma}}_{jk}(u)$ refers to the derivative of the covariance matrix w.r.t the rescaled time index. This gives us:

$$|S_T^{-1} \sum_{t=Tu-S_T}^{Tu+S_T} E(X_{j,t,T}X_{k,t,T}) - E(\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u))| \leq O((S_T/T)^2) \tag{1.20}$$

Now we deal with the term (1.17), note that the functional dependence measure of the stationary approximation Using this we compute the functional dependence measure of $\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u))$ as:

$$\begin{aligned}\sup_{u\in[0,1]} ||\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u)) - \tilde{X}^*_{j,T}(u)\tilde{X}^*_{k,T}(u))||_{r/2} \leq \sup_{u\in[0,1]} (&||\tilde{X}_{j,T}(u)||_r||\tilde{X}_{k,T}(u) - \tilde{X}_{k,T}(u)^*||_r \\ &+ \sup_{u\in[0,1]} ||\tilde{X}_{k,T}(u)||_r||\tilde{X}_{j,T}(u) - \tilde{X}_{j,T}(u)^*||_r)\end{aligned} \tag{1.21}$$

Therefore, the $\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u))$ has a finite cumulative dependence measure by the weak dependence condition imposed on $\tilde{\boldsymbol{x}}_t(u)$. Taking into account (1.19), and the above we can then apply theorem 2.7 (iii) in Dahlhaus et al. (2018) to obtain:

$$\begin{aligned}&P(|S_T^{-1} \sum_{t=T-S_T}^{T} X_{j,t,T}X_{k,t,T} - E(\tilde{X}_{j,T}(u)\tilde{X}_{k,T}(u))| > S_T^{-\kappa}) \\ &\leq O(S_T^{-r/2+r\kappa/2+1}) + O(\exp(-S_T^{1-2\kappa}))\end{aligned}$$

Applying the union bound then completes the proof.

For (ii), we proceed similarly. We bound:

$$P(|S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}\epsilon_{t,T} - E(\tilde{X}_{j,T-h}(u)\tilde{\epsilon}_T(u))| > S_T^{-\kappa}) \tag{1.22}$$

$$\leq P\left(|S_T^{-1}\sum_{t=T-S_T}^{T}[X_{j,t-h,T}\epsilon_{t,T}| + |O(T^{-1})| > S_T^{-\kappa}\right) \tag{1.23}$$

Given that $E(X_{j,t-h,T}\epsilon_{t,T}) = 0,\ \forall j$. We have that $E(\tilde{X}_{j,T-h}(u)\tilde{\epsilon}_T(u)) = O(T^{-1})$. Now we apply the same procedure as previously. We have that:

$$\sup_{u\in[0,1]} ||\tilde{X}_{j,t}(u)\tilde{\epsilon}_t(u)) - \tilde{X}^*_{j,t}(u)\tilde{\epsilon}^*_t(u))||_\tau \leq \sup_{u\in[0,1]}(||\tilde{X}_{j,t}(u)||_r||\tilde{\epsilon}_t(u) - \tilde{\epsilon}_t(u)^*||_q \tag{1.24}$$

$$+ \sup_{u\in[0,1]} ||\tilde{\epsilon}_t(u)||_q||\tilde{X}_{j,t}(u) - \tilde{X}_{j,t}(u)^*||_r)$$

This has a finite cumulative functional dependence measure by the weak dependence conditions imposed. Once again, by applying theorem 2.7 (iii) in Dahlhaus et al. (2018) we obtain:

$$P(|S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}\epsilon_{t,T} - E(\tilde{X}_{j,T-h}(u)\tilde{\epsilon}_T(u))| > S_T^{-\kappa}) \tag{1.25}$$

$$\leq O(S_T^{-\kappa+\tau\kappa+1}) + O(\exp(S_T^{1-2\kappa})) \tag{1.26}$$

Applying the union bound gives the final result.

For (iii), we note that $f(t/T, \boldsymbol{x}_{t,T}) \equiv f(\boldsymbol{x}_{t,T})$ is a locally stationary process with stationary approximation $f(t/T, \tilde{\boldsymbol{x}}_t(t/T))$. And the stationary approximation has cumulative functional dependence measure $\sup_{u\in[0,1]}|\boldsymbol{\beta}(u)|\Phi^{\boldsymbol{x}}_{0,r}$. We can then compute the cumulative dependence measure of the product process $f(\tilde{\boldsymbol{x}}_t(t/T))\tilde{X}_{j,t}(t/T)$

similarly as for part (i). We then obtain

$$P(|S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t,T}f(\boldsymbol{x}_{t,T}) - E(\tilde{X}_{j,T}(u)f(\tilde{\boldsymbol{x}}_T(u)))| > S_T^{-\kappa})$$
$$\leq O(S_T^{-r/2+r\kappa/2+1}) + O(\exp(-S_T^{1-2\kappa}))$$

We can handle the bias term, in the same way we did for part(i), given that the product process $f(\boldsymbol{x}_{t,T})X_{j,t,T}$ is locally stationary with the stationary approximation being twice differentiable w.r.t to the rescaled time index. Taking the union bound then gives us the result.

The result for (iv) follows immediately from parts (ii) and (iii). □

Proof of Lemma 2.

Recall that:

$$\tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u)) = \tilde{R}^{m-1}f(\tilde{\boldsymbol{x}}_{T-h}(u)) - \left\langle \tilde{R}^{m-1}f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_{\widehat{\mathcal{S}}_m,T-h}(u)\right\rangle \tilde{X}_{\widehat{\mathcal{S}}_m,T-h}(u),$$
$$\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) = \hat{R}^{m-1}f(\boldsymbol{x}_{T-h,T}) - \left\langle \hat{R}^{m-1}f, X_{\widehat{\mathcal{S}}_m,\cdot,T}\right\rangle_{(T)} X_{\widehat{\mathcal{S}}_m,T-h,T}$$

We denote: $A_T(m,j) = \left\langle \hat{R}^{m-1}f, X_{j,\cdot,T}\right\rangle_{(T)} - \left\langle \tilde{R}^{m-1}f, \tilde{X}_{j,T-h}\right\rangle$. We proceed with a recursive analysis. Note that for $m = 0$, the result follows from lemma 1. By using

the above definitions we get the following recursive relation:

$$
\begin{aligned}
&A_T(m,j) \leq A_T(m-1,j) \\
&- (\langle \tilde{R}^{m-1} f, \tilde{X}_{\widehat{\mathcal{S}}_m, T-h} \rangle) \Big(\langle X_{\widehat{\mathcal{S}}_m,\cdot,T}, X_{j,\cdot,T} \rangle_{(T)} - \langle \tilde{X}_{\widehat{\mathcal{S}}_m,T-h}(u), \tilde{X}_{j,T-h}(u) \rangle \Big) \\
&- \Big(\langle \hat{R}^{m-1} f, X_{\widehat{\mathcal{S}}_m,\cdot,T} \rangle_{(T)} - \langle \tilde{R}^{m-1} f, \tilde{X}_{\widehat{\mathcal{S}}_m,T-h} \rangle \Big) \langle X_{\widehat{\mathcal{S}}_m,\cdot,T}, X_{j,\cdot,T} \rangle_{(T)} \\
&= A_T(m-1,j) - I_{T,m}(j) - II_{T,m}(j)
\end{aligned}
$$

Now we have that $\sup_j |I_{T,m}(j)| \leq ||f(\tilde{\boldsymbol{x}}_{T-h})|| \zeta_T$, by lemma 1, and the norm reducing property of the remainder functions. Similarly, $\sup_j |II_{T,m}(j)| \leq (1 + \zeta_T) \sup_j A_T(m-1,j)$. The rest of the proof follows from Buhlmann (2006).

□

Proof of Lemma 3.

The proof closely follows the one laid out in Buhlmann (2006), therefore we omit the details here. □

Proof of Corollary 3.

We only need to change lemma 1, from the proof of theorem 1. The rest of the proof is essentially the same, if we switch we replace the locally stationary variables with stationary ones. We borrow some arguments from the proof of theorem 2 in Yousuf (2018). The main technical tool we use is theorem 3 in Wu and Wu (2016). For that, we first define the *predictive dependence measure* introduced by Wu (2005). The predictive dependence measure for stationary univariate and multivariate processes

is defined respectively as:

$$\theta_q(\epsilon_i) = ||\mathrm{E}\left(\epsilon_i|\mathcal{F}_0\right) - \mathrm{E}\left(\epsilon_i|\mathcal{F}_{-1}\right)||_q,$$
$$\theta_q(X_{j,i}) = ||\mathrm{E}\left(X_{j,i}|\mathcal{H}_0\right) - \mathrm{E}\left(X_{j,i}|\mathcal{H}_{-1}\right)||_q. \tag{1.27}$$

With the cumulative predictive dependence measures defined as:

$$\Theta_{0,q}(\boldsymbol{x}) = \max_{j\leq p_n}\sum_{i=0}^{\infty}\theta_q(X_{ij}),\ \text{and}\ \Theta_{0,q}(\boldsymbol{\epsilon}) = \sum_{i=0}^{\infty}\theta_q(\epsilon_i).$$

By theorem 1 in Wu (2005), we have $\Theta_{0,q}(\boldsymbol{x}) \leq \Phi^{\boldsymbol{x}}_{0,q}$, and similarly $\Theta_{0,q}(\boldsymbol{\epsilon}) \leq \Delta^{\epsilon}_{0,q}$. Where $\Phi^{\boldsymbol{x}}_{0,q}, \Delta^{\epsilon}_{0,q}$ represent the cumulative functional dependence measures. From Section 2 in Wu and Wu (2016): $||X_{j,i}||_q \leq \Phi^{\boldsymbol{x}}_{0,q}$, and $||\epsilon_i||_q \leq \Delta^{\boldsymbol{\epsilon}}_{0,q}$. We only discuss parts (i) and (ii) from lemma 1, the others can be done similarly. We now define $\boldsymbol{G}_{jk} = (G_{1,jk}, \ldots, G_{T,jk})$ where $G_{i,jk} = X_{j,i}X_{k,i}$, and let $\boldsymbol{R}_j = (R_{1,j}, \ldots, R_{T,j})$ where $R_i = X_{j,i}\epsilon_i$. We need to bound the sums: $\sum_{i=1}^{T}(G_{i,jk} - E(G_{i,jk}))/T$ and $\sum_{i=1}^{T} R_{i,j}/T$.

As previously, we have (by Holder's inequality)

$$\sum_{t=0}^{\infty}||X_{j,t}X_{k,t} - X^*_{j,t}X^*_{k,t}||_q \tag{1.28}$$
$$\leq \sum_{t=0}^{\infty}(||X_{j,t}||_{2q}||X_{k,t} - X^*_{k,t}||_{2q} + ||X_{k,t}||_{2q}||X_{j,t} - X^*_{j,t}||_{2q}) \leq 2\Phi^2_{0,2q}(\boldsymbol{x}) \tag{1.29}$$

Using these, along with Condition 4.5, we obtain:

$$\sup_{q\geq 4} q^{-2\tilde{\alpha}_x}\Theta_q(\boldsymbol{G}_{jk}) \leq \sup_{q\geq 4} 2q^{-2\tilde{\alpha}_x}\Phi^2_{0,2q}(\boldsymbol{x}) < \infty \tag{1.30}$$

Combining the above and using Theorem 3 in Wu and Wu (2016), we obtain:

$$P\left(\left|\sum_{t=1}^{T}(G_{t,jk}-E(G_{t,jk}))\right|>\frac{cT^{1-\kappa}}{2}\right)\leq C\exp\left(-\frac{T^{1/2-\kappa}}{\upsilon_x^2}\right)^{\tilde{\alpha}} \tag{1.31}$$

Similarly, using the same procedure we obtain:

$$P\left(\left|\sum_{t=1}^{T}R_{t,j}\right|>\frac{cT^{1-\kappa}}{2}\right)\leq C\exp\left(-\frac{T^{1/2-\kappa}}{\upsilon_x\upsilon_\epsilon}\right)^{\tilde{\alpha}'} \tag{1.32}$$

We can use the same procedure to get the corresponding bounds for the terms in lemma 1 (iii) and (iv). Now using the above bounds and following the steps in the proof of Theorem 1 we obtain the result.

□

Proof of Theorem 2.

The proof of the LL-Boost is more complicated than the LC-Boost case due to the additional linear term. Fortunately, the population version stays the same between both versions. This allows us to use the same framework as previously, where we relied on Temlyakov's result on weak greedy algorithms. We do need to make a number of changes from the proof of theorem 1, and we start by introducing the following notation: let $\boldsymbol{Z}_{j,t,T}=(X_{j,t,T},X_{j,t,T}(t/T-u))$, and let:

$$\begin{aligned}
&\hat{\boldsymbol{h}}(Y_{\cdot,T},X_{j,\cdot,T})=(\hat{h}_1(Y_{\cdot,T},X_{j,\cdot,T}),\hat{h}_2(Y_{\cdot,T},X_{j,\cdot,T}))\\
&=\text{argmin}_{\boldsymbol{h}}S_T^{-1}\sum_{t=T-S_T}^{T}(Y_{t,T}-h_1X_{j,t-h,T}-h_2X_{j,t-h,T}(t/T-u))^2\\
&\hat{\boldsymbol{h}}(X_{k,\cdot,T},X_{j,\cdot,T})=(\hat{h}_1(X_{k,\cdot,T},X_{j,\cdot,T}),\hat{h}_2(X_{k,\cdot,T},X_{j,\cdot,T}))\\
&=\text{argmin}_{\boldsymbol{h}}S_T^{-1}\sum_{t=T-S_T}^{T}(X_{k,t-h,T}-h_1X_{j,t-h,T}-h_2X_{j,t-h,T}(t/T-u))^2
\end{aligned}$$

represent the estimated local linear regression coefficients. The arguments to the function $h(\cdot,\cdot)$ refer to the dependent and independent variables respectively. These functions are linear functions of the first argument. We also let $\boldsymbol{h}(\tilde{Y},\tilde{X}_j),\boldsymbol{h}(\tilde{X}_k,\tilde{X}_j)$ represent the population version of these coefficients.[27] We then define our selectors as:

$$\hat{\mathcal{S}}_1 = \text{argmax}_j||\hat{\boldsymbol{h}}(Y_{\cdot,T},X_{j,\cdot,T})\boldsymbol{Z}_{j,\cdot,T}||_{(T)},\ldots,\hat{\mathcal{S}}_m = \text{argmax}_j||\hat{\boldsymbol{h}}(\hat{R}_T^m f,X_{j,\cdot,T})\boldsymbol{Z}_{j,\cdot,T}||_{(T)}$$

Where the sample remainder functions are defined as:

$$\hat{R}_T^0 f(\boldsymbol{x}_{T-h,T}) = f(\boldsymbol{x}_{T-h,T}),$$
$$\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) = \hat{R}_T^{m-1} f(\boldsymbol{x}_{T-h,T}) - \hat{\boldsymbol{h}}(\hat{R}_T^m f,X_{\hat{\mathcal{S}}_m,\cdot,T})\boldsymbol{Z}_{\hat{\mathcal{S}}_m,\cdot,T}, m=1,2,\ldots$$

Therefore, $\hat{R}_T^m f(\boldsymbol{x}_{T-h}) = f(\boldsymbol{x}_{T-h}) - \hat{F}_{ll}^{(M_T)}(u,\boldsymbol{x}_{T-h,T})$, is the difference between $f(\boldsymbol{x}_{T-h})$ and its LL-Boost estimate. Now the semipopulation version has the same form as in theorem 1 except it uses the selected base learners $\hat{\mathcal{S}}_m$ as defined above. Recall that $Y_{t,T} = \alpha_j^{(m)}(t/T)X_{j,t-h,T} + \epsilon_{j,t,T}$, where
$\alpha_j(t/T) = E(\tilde{X}_{j,t-h}(t/T)\tilde{Y}_T(t/T))/E(\tilde{X}_{j,t-h}^2(t/T))$. We also define the following:

$$X_{j,t,T} = \alpha_{jk}(t/T)X_{k,t-h,T} + \epsilon_{jk,t,T}, \tag{1.33}$$
$$\text{where } \alpha_j^{(m)}(t/T) = E(\tilde{X}_{j,t-h}(t/T)\tilde{X}_{k,t-h,T}(t/T))/E(\tilde{X}_{k,t-h}^2(t/T)) \tag{1.34}$$

We now need the following lemmas:

Lemma 4. *Under conditions 1.5.1, 1.5.2, 1.5.3, 1.5.4, and for κ as defined in Theorem 2, the following hold:*

[27] We note that $\boldsymbol{h}$ is also a function of the rescaled time point u, but we ignore this for now.

1. $\sup_{j,k\leq p_T} |S_T^{-1}\sum_{t=T-S_T}^{T}\Big[X_{j,t,T}X_{k,t,T}(t/T-u)^i - E(X_{j,t,T}X_{k,t,T}(t/T-u)^i)\Big]| = O_p(S_T^{-\kappa+i}/T^i)$ *for* $i=1,2$

2. $\sup_{j\leq p_T} |S_T^{-1}\sum_{t=T-S_T}^{T}\Big[X_{j,t,T}^{i_1}(t/T-u)^{i_2} - E(X_{j,t,T}^{i_1}(t/T-u)^{i_2})\Big]| = \zeta_{T,i_1,i_2} = O_p(S_T^{-\kappa+i_2}/T^{i_2})$ *for* $i_1=1,2$ *and* $i_2=1,2,3$.

3. $\sup_{j,k\leq p_T} |S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}\epsilon_{jk,t,T}(t/T-u)^i - E(S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}\epsilon_{jk,t,T}(t/T-u)^i)| = O_p(S_T^{-\kappa+i}/T^i)$ *for* $i=0,1$

4. $\sup_{j\leq p_T} |S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}\epsilon_{j,t,T}(t/T-u)^i - E(S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}\epsilon_{j,t,T}(t/T-u)^i)| = O_p(S_T^{-\kappa+i}/T^i)$ *for* $i=0,1$

Lemma 5. *Under conditions 1.5.1, 1.5.2, 1.5.3, 1.5.4, and for* κ *as defined in Theorem 2, the following hold:*

1. $\sup_{j,k\leq p_T} |\hat{\boldsymbol{h}}(X_{j,\cdot,T}, X_{k,\cdot,T}) - \boldsymbol{h}(\tilde{X}_j, \tilde{X}_k)| = \zeta_{T,1} = O_p(S_T^{-\kappa})$

2. $\sup_{j\leq p_T} |\hat{\boldsymbol{h}}(\epsilon_{\cdot,T}, X_{j,\cdot,T}) - \boldsymbol{h}(\tilde{\epsilon}, \tilde{X}_j)| = \zeta_{T,2} = O_p(S_T^{-\kappa})$

3. $\sup_{j\leq p_T} |\hat{\boldsymbol{h}}(f, X_{j,\cdot,T}) - \boldsymbol{h}(\tilde{f}, \tilde{X}_j)| = \zeta_{T,3} = O_p(S_T^{-\kappa})$

4. $\sup_{j\leq p_T} |\hat{\boldsymbol{h}}(Y_{\cdot,T}, X_{j,\cdot,T}) - \boldsymbol{h}(\tilde{Y}, \tilde{X}_j)| = \zeta_{T,4} = O_p(S_T^{-\kappa})$

We introduce the following notation for the next lemma. Let

$$\hat{\boldsymbol{h}}(X_{k,\cdot,T}(\cdot/T-u),X_{j,\cdot,T}) = (\hat{h}_1(X_{k,\cdot,T}(\cdot/T-u),X_{j,\cdot,T}),\hat{h}_2(X_{k,\cdot,T}(\cdot/T-u),X_{j,\cdot,T}))$$
$$= \text{argmin}_{\boldsymbol{h}} S_T^{-1}\sum_{t=T-S_T}^{T}(X_{k,t-h,T}(t/T-u)-h_1X_{j,t-h,T}-h_2X_{j,t-h,T}(t/T-u))^2$$

Recall that $\hat{\boldsymbol{h}}(X_{k,\cdot,T}(\cdot/T-u),X_{j,\cdot,T}) = \hat{\boldsymbol{A}} * \hat{\boldsymbol{B}}$, where:

$$\hat{\boldsymbol{A}} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}^2 & S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}^2(t/T-u) \\ S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}^2(t/T-u) & S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}^2(t/T-u)^2 \end{bmatrix}^{-1}$$

$$\hat{\boldsymbol{B}} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}X_{k,t-h,T}(t/T-u) \\ S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}X_{k,t-h,T}(t/T-u)^2 \end{bmatrix}$$

We then let $\boldsymbol{h}(X_{k,\cdot,T}(\cdot/T-u),X_{j,\cdot,T}) = \boldsymbol{A} * \boldsymbol{B}$:

$$\boldsymbol{A} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t-h,T}^2) & S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t-h,T}^2)(t/T-u) \\ S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t-h,T}^2)(t/T-u) & S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t-h,T}^2)(t/T-u)^2 \end{bmatrix}^{-1}$$

$$\boldsymbol{B} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t-h,T}X_{k,t-h,T})(t/T-u) \\ S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t-h,T}X_{k,t-h,T})(t/T-u)^2 \end{bmatrix}$$

Lemma 6. *Under conditions 1.5.1, 1.5.2, 1.5.3, 1.5.4, and for κ as defined in Theorem 2, the following hold:*

1. $\sup_{j,k\leq p_T} |\hat{h}_1(X_{k,\cdot,T}(\cdot/T-u),X_{j,\cdot,T}) - h_1(X_{k,\cdot,T}(\cdot/T-u),X_{j,\cdot,T})|$
   $= \zeta_{T,5} = O_p(S_T^{-\kappa+1}/T)$

2. $\sup_{j,k\leq p_T} |\hat{h}_2(X_{k,\cdot,T}(\cdot/T-u), X_{j,\cdot,T}) - h_2(X_{k,\cdot,T}(\cdot/T-u), X_{j,\cdot,T})|$
   $= \zeta_{T,6} = O_p(S_T^{-\kappa})$

Let $\zeta_T = \max(\zeta_{T,1}, \zeta_{T,2}, \zeta_{T,3}, \zeta_{T,4}, \zeta_{T,6}) = O_p(S_T^{-\kappa})$ and denote by ω a realization of all S_T sample points involved in estimation. The next lemma bounds the difference between the sample and population learners at step m.

Lemma 7. *Suppose conditions 1.5.1, 1.5.2, 1.5.3, and 1.5.4 hold. Then for κ as defined in Theorem 2 and on the set $\mathcal{A}_T = \{\omega : \zeta_T(w) < 1/2, \zeta_{T,5}(\omega) \leq S_T/T, \zeta_{T,2,2}(\omega) \leq S_T^2/T^2\}$, we have:*

$$\sup_{j\leq p_T} ||\boldsymbol{Z}_{j,\cdot,T}\hat{\boldsymbol{h}}(\hat{R}_T^m f, X_j) - \left\langle \tilde{R}^{m-1} f(\tilde{\boldsymbol{x}}_{T-h}(u)), \tilde{X}_{j,T-h}(u)\right\rangle \tilde{X}_{j,T-h}(u)||_{(T)} \leq C(5/2)^m \zeta_T,$$

where C does not depend on m, T.

Lemma 8. *Suppose the conditions needed for Theorem 2 hold, then for $m = m_T \to \infty$ slow enough we have:*

$$||\tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))|| = o_p(1)$$

With the above lemmas we are now ready to analyze the term: $\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) = f(\boldsymbol{x}_{T-h,T}) - \hat{F}_{ll}^{(M_T)}(u, \boldsymbol{x}_{T-h,T})$. By the triangle inequality we obtain:

$$||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T})|| \leq ||\tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))|| + ||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) - \tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))|| \quad (1.35)$$

the first term can be handled with lemma 8. For the second term, let $A_T(m) = ||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) - \tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))||$. Using the definitions of the remainder functions,

we then have a recursive relation:

$$\begin{aligned} A_T(m) &\leq A_T(m-1) + |\langle \hat{R}^{m-1} f, X_{\widehat{\mathcal{S}}_{m,\cdot,T}} \rangle_{(T)} - \langle \tilde{R}^{m-1} f, \tilde{X}_{\widehat{\mathcal{S}}_{m,T-h}} \rangle| ||X_{\widehat{\mathcal{S}}_{m,T-h,T}}|| \\ &+ ||\langle \tilde{R}^{m-1} f, \tilde{X}_{\widehat{\mathcal{S}}_{m,T-h}} \rangle|| ||X_{\widehat{\mathcal{S}}_{m,T-h,T}} - \tilde{X}_{\widehat{\mathcal{S}}_{m,T-h}}(u)|| \\ &\leq A_T(m-1) + C(5/2)^m \zeta_T + O(1/T) \text{ on the set } \mathcal{A}_T \end{aligned}$$

Where the last inequality follows from local stationarity and lemma 7.[28] By the above recursive equation we obtain: $||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T}) - \tilde{R}_T^m f(\tilde{\boldsymbol{x}}_{T-h}(u))|| \leq 3^m \zeta_T C$. If we choose $m = m_T \to \infty$ slow enough (e.g $m_T = o(log(T))$), then along with lemma 3 and (1.35) we obtain:

$$||\hat{R}_T^m f(\boldsymbol{x}_{T-h,T})|| = ||f(\boldsymbol{x}_{T-h,T}) - \hat{F}_{ll}^{(M_T)}(u, \boldsymbol{x}_{T-h,T})|| = o_p(1)$$

□

Proof of Lemma 4.

We start with (i), and we bound:

$$P(|S_T^{-1}\Big[\sum_{t=T-S_T}^{T} X_{j,t,T}X_{k,t,T}(t/T-u)^i - E(\sum_{t=T-S_T}^{T} X_{j,t,T}X_{k,t,T}(t/T-u)^i)\Big]| > S_T^{-\kappa+i}/T^i) \tag{1.36}$$

We have a sum similar to that in lemma 1, except we have weights $(t/T-u)^i$. Now given that $\sum_{t=T-S_T}^{T}(t+1)^i/T^i - t^i/T^i \leq C(S_T^i/T^i)$. Additionally, since we had shown in the proof of lemma 1 that the product process $X_{j,t,T}X_{k,t,T}$ is locally stationary and satisfies the weak dependence condition, we can use theorem 2.7 (iii) in Dahlhaus

[28] Although not the exact statement of 7, the proof handles the specific term we need.

et al. (2018) directly to obtain:

$$(1.36) \leq O(S_T^{-r/2+r\kappa/2+1}) + O(\exp(-S_T^{1-2\kappa}))$$

Part (ii) can be handled similarly. For part (iii), we have that $\epsilon_{jk,t,T} = X_{j,t,T} - \alpha_{jk}(t/T)X_{k,t-h,T}$, therefore is locally stationary and its stationary approximation satisfies the weak dependence condition, and has r finite moments. Therefore, we get the same result as for (iii). For (iv), note that by definition $\epsilon_{j,t,T}$ has $\min(r,q)$ finite moments, and is locally stationary. Now if we let $r_1 = \min(r,q)$ we get the same result as for part (i) with r_i instead of r . □

Proof of Lemma 5.

We mainly discuss the proof for part (i), the rest can be handled similarly. Note that $X_{k,t-h,T} = \boldsymbol{h}(\tilde{X}_j, \tilde{X}_k)\boldsymbol{Z}_{j,t-h,T} + \ddot{h}_1((\tilde{X}_j, \tilde{X}_k))(c)(t/T-u)^2 + \epsilon_{jk,t,T}$. Where $\ddot{h}_1(\tilde{X}_j, \tilde{X}_k)(\cdot)$ refers to the second derivative of $h_1(\tilde{X}_j, \tilde{X}_k)(\cdot)$. Therefore we have:

$$|\hat{\boldsymbol{h}}(X_{j,\cdot,T}, X_{k,\cdot,T}) - \boldsymbol{h}(\tilde{X}_j, \tilde{X}_k)| = \hat{\boldsymbol{A}}^{-1} * \hat{\boldsymbol{B}} + \hat{\boldsymbol{A}}^{-1} * \hat{\boldsymbol{C}}$$

where

$$\hat{\boldsymbol{A}}^{-1} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} X_{k,t-h,T}^2 & S_T^{-1}\sum_{t=T-S_T}^{T} X_{k,t-h,T}^2(t/T-u) \\ S_T^{-1}\sum_{t=T-S_T}^{T} X_{k,t-h,T}^2(t/T-u) & S_T^{-1}\sum_{t=T-S_T}^{T} X_{k,t-h,T}^2(t/T-u)^2 \end{bmatrix}^{-1}$$

$$\hat{\boldsymbol{B}} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} X_{k,t-h,T}(t/T-u)^2 \\ S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}(t/T-u)^3 \end{bmatrix}$$

$$\hat{\boldsymbol{C}} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} X_{k,t-h,T}\epsilon_{jk,t-h,T} \\ S_T^{-1}\sum_{t=T-S_T}^{T} X_{j,t-h,T}\epsilon_{jk,t-h,T}(t/T-u) \end{bmatrix}$$

And let

$$\boldsymbol{A}^{-1} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{k,t-h,T}^2) & S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{k,t-h,T}^2)(t/T-u) \\ S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{k,t-h,T}^2)(t/T-u) & S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{k,t-h,T}^2)(t/T-u)^2 \end{bmatrix}^{-1}$$

$$\boldsymbol{B} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{k,t-h,T})(t/T-u)^2 \\ S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t-h,T})(t/T-u)^3 \end{bmatrix}$$

$$\boldsymbol{C} = \begin{bmatrix} S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{k,t-h,T}\epsilon_{jk,t-h,T}) \\ S_T^{-1}\sum_{t=T-S_T}^{T} E(X_{j,t-h,T}\epsilon_{jk,t-h,T})(t/T-u) \end{bmatrix}$$

Given $\hat{\boldsymbol{A}}$ is a 2 by 2 matrix we can calculate its inverse directly:

$$\hat{\boldsymbol{A}}^{-1} = \hat{a}_0^{-1} \begin{bmatrix} S_T^{-1} \sum_{t=T-S_T}^{T} X_{k,t-h,T}^2 (t/T-u)^2 & S_T^{-1} \sum_{t=T-S_T}^{T} -X_{k,t-h,T}^2 (t/T-u) \\ S_T^{-1} \sum_{t=T-S_T}^{T} -X_{k,t-h,T}^2 (t/T-u) & S_T^{-1} \sum_{t=T-S_T}^{T} X_{k,t-h,T}^2 \end{bmatrix}$$

where $\hat{a}_0 = [\hat{A}_{11}\hat{A}_{22} - \hat{A}_{12}\hat{A}_{21}]$

We first handle $\hat{h}_1(X_{j,\cdot,T}, X_{k,\cdot,T})$, and $\hat{h}_2(X_{j,\cdot,T}, X_{k,\cdot,T})$ can be handled similarly. From the above equations we obtain:

$$\begin{aligned} &P(|\hat{h}_1(X_{j,\cdot,T}, X_{k,\cdot,T}) - h_1(X_{j,\cdot,T}, X_{k,\cdot,T})| > S_T^{-\kappa}) \\ &\leq P(|\hat{A}_{11}^{-1}\hat{B}_1 + \hat{A}_{12}^{-1}\hat{B}_2 - (A_{11}^{-1}B_1 + A_{12}^{-1}B_2)| > cS_T^{-\kappa}) \\ &\leq P(|\hat{A}_{11}^{-1}\hat{B}_1 - A_{11}^{-1}B_1| > cS_T^{-\kappa}) + P(|\hat{A}_{12}^{-1}\hat{B}_2 - A_{12}^{-1}B_2| > cS_T^{-\kappa}) \end{aligned}$$

For the first term, we let:

$$\hat{Q}_1 = \hat{B}_1 * S_T^{-1} \sum_{t=T-S_T}^{T} X_{k,t-h,T}^2 (t/T-u)^2,$$
$$\text{and } Q_1 = B_1 * S_T^{-1} \sum_{t=T-S_T}^{T} E(X_{k,t-h,T}^2)(t/T-u)^2$$

we then have: $|\hat{A}_{11}^{-1}\hat{B}_1 - A_{11}^{-1}B_1| = |\hat{Q}_1\hat{a}_0^{-1} - Q_1 a_0^{-1}| = |(\hat{a}_0^{-1} - a_0^{-1})(\hat{Q}_1 - Q_1) + (\hat{Q}_1 - Q_1)a_0^{-1} + (\hat{a}_0^{-1} - a_0^{-1})Q_1|$. We have that $a_0 = O(S_T^2/T^2)$, and $Q = O(S_T^4/T^4)$.

Therefore:

$$P(|\hat{A}_{11}^{-1}\hat{B}_1 - A_{11}^{-1}B_1| > cS_T^{-\kappa}) \leq P(|(\hat{a}_0^{-1} - a_0^{-1})(\hat{Q}_1 - Q_1)| > c_2 n^{-\kappa}/3) \tag{1.37}$$

$$+ P(|(\hat{Q}_1 - Q_1)a_0^{-1}| > cS_T^{-\kappa}/3|) \tag{1.38}$$

$$+ P(|(\hat{a}_0^{-1} - a_0^{-1})Q_1| > cS_T^{-\kappa}/3). \tag{1.39}$$

For the RHS of (1.37), we obtain:

$$P(|(\hat{a}_0^{-1} - a_0^{-1})(\hat{Q}_1 - Q_1)| > cn^{-\kappa}/3) \leq P(|\hat{Q}_1 - Q_1| > CS_T^{-\kappa/2}) + P(|\hat{a}_0^{-1} - a_0^{-1}| > CS_T^{-\kappa/2}).$$

Therefore we can focus on the terms (1.38),(1.39). We can handle (1.38) directly using the fact that $a_0 = O(S_T^2/T^2)$ along with lemma 5. For (1.39), note that $Q = O(S_T^4/T^4)$, and $|\hat{a}_0^{-1} - a_0^{-1}| = (\hat{a}_0 - a_0)/(\hat{a}_0 a_0)$. We then obtain:

$$(1.39) \leq P(|\hat{a}_0 - a_0| > CS_T^{-\kappa} D_T S_T^2/T^2) + P(|\hat{a}_0| < D_T)$$

We can now choose $D_T \leq \min_{k \leq p_T} a_0 * (log(S_T))^{-1}$. And we obtain the bound by applying lemma 4. For $P(|\hat{A}_{12}^{-1}\hat{B}_2 - A_{12}^{-1}B_2| > cS_T^{-\kappa})$ we obtain a bound in similar fashion. Applying the union bound gives us the result.

□

The proof for lemma 6 can be obtained similarly to the proof of lemma 5. The proofs for lemmas 7 and 8 follow the same steps as in theorem 1, therefore we omit the details.

Appendix B: Additional Empirical Results:

In this section we first include the MSFE by start date of the out of sample period for horizon $h = 3$. We then include results comparing: 1) LC-Boost vs Boost, 2) LC-Boost vs LC-Boost Factor, 3) Boost Factor estimated using a 10 year rolling window vs LC-Boost Factor, 4) LC-Boost Factor vs TV-DI and 5) LL-Boost Factor vs LC-Boost Factor . The TV-DI model uses 4 factors along with lags of $Y_{t,T}$ like the DI model, but estimates the model using local constant estimation. Our method of comparison is to plot the relative local MSFE of the two methods being compared.

We also include the relative MSFE for horizons $h = 6, 3, 1$ for all methods used in the main text. Lastly we include the simulation results using t_5 innovations.

Table 1.3: Relative MSFE $h = 6$

Full Out of Sample Period 1971:9-2018:8								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.1	1.01	1.08	.77	1	.83	1	.99
DI	.81	.81	.73	.98	.86	.96	.91	.92
Lasso	.82	.78	.72	.85	.91	.90	.80	.93
Boost	.79	.77	.73	.79	.90	.87	.86	.93
Boost Factor	.79	.83	.69	.92	.81	.89	.79	.87
LC-Boost	.76	.76	.69	.79	.89	.89	.94	.98
LC-Boost Factor	.73	.74	.67	.73	.80	.80	.78	.90
LL-Boost Factor	.88	.95	.84	.79	.79	.79	.97	1.37
"Pre-Great Moderation" 1971:9-1982:12								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.08	1	1.12	.88	.97	1.07	1.06	1.05
DI	.56	.64	.61	1.40	.76	.88	.91	.88
Lasso	.55	.62	.56	1.20	.80	1.03	.77	.93
Boost	.55	.64	.63	1.20	.78	.95	.79	.83
Boost Factor	.59	.71	.64	1.14	.75	.88	.76	.86
LC-Boost	.55	.64	.59	.93	.98	1.12	.91	.90
LC-Boost Factor	.60	.76	.66	.85	.82	1.02	.78	.94
LL-Boost Factor	.73	.99	.88	.88	.80	.94	.95	1.01
"Great Moderation" 1983:1-2006:12								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.08	.98	1.04	.81	1.02	.93	.86	.88
DI	1.07	1.03	.82	.89	.96	.96	.86	.96
Lasso	1.06	1.11	.88	.90	1.03	.86	.87	.93
Boost	1.08	1.09	.85	.80	1	.86	1.01	1.08
Boost Factor	1.04	.99	.70	.87	.93	.83	.82	.85
LC-Boost	1.14	.99	.85	.89	.95	.74	1.04	1.1
LC-Boost Factor	.98	.80	.71	.80	.85	.83	.80	.76
LL-Boost Factor	1.04	.98	.84	.92	.97	.88	1.06	1.05
"Post Great Moderation" 2007:1-2018:8								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.18	1.09	1.06	.64	.98	.59	.72	.76
DI	1.12	1	.84	.78	.83	1.01	1.21	1.16
Lasso	1.21	.86	.86	.52	.91	.82	1	1.02
Boost	1.05	.77	.79	.45	.77	.82	1.53	1.43
Boost Factor	1.01	.98	.78	.79	.76	.94	1.37	1.12
LC-Boost	.89	.80	.70	.54	.77	.58	.93	1.32
LC-Boost Factor	.82	.61	.67	.54	.75	.64	.69	.95
LL-Boost Factor	1.07	.80	.74	.53	.62	.62	.75	.92

Table 1.4: Relative MSFE $h = 3$

Full Out of Sample Period 1971:9-2018:8								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.04	1.01	1.01	.86	1.01	.91	.98	.98
DI	.84	.81	.78	.98	.85	.96	.90	.94
Lasso	.89	.83	.82	1.03	.87	.97	.84	.96
Boost	.89	.78	.75	.89	.86	.87	.93	1.04
Boost Factor	.80	.83	.76	.96	.81	.90	.82	.89
LC-Boost	.79	.77	.73	.87	.85	.85	.88	1.02
LC-Boost Factor	.77	.78	.77	.87	.80	.87	.79	.89
LL-Boost Factor	.84	.82	.82	.90	.81	.87	.92	1.05
“Pre-Great Moderation" 1971:9-1982:12								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.05	1	1	.97	1.01	1.01	1	1.01
DI	.72	.72	.72	1.29	.87	.87	.89	.88
Lasso	.73	.74	.74	1.44	.85	1.09	.84	.98
Boost	.77	.89	.74	1.14	.95	.91	.92	1
Boost Factor	.71	.78	.78	1.15	.80	.83	.80	.86
LC-Boost	.78	.69	.75	1.02	.95	.88	.89	1.02
LC-Boost Factor	.74	.82	.84	1.03	.85	.97	.78	.89
LL-Boost Factor	.84	.93	.90	1.05	.94	.87	.87	1
“Great Moderation" 1983:1-2006:12								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.06	1.06	1.03	.90	1	.83	.85	.88
DI	.92	.95	.86	.91	.87	.97	.88	1.16
Lasso	.96	1.04	.91	.98	.92	.90	.78	.88
Boost	.97	.96	.84	.91	.90	.90	.89	1.12
Boost Factor	.89	.89	.76	.94	.87	.89	.86	.95
LC-Boost	.94	.96	.81	.94	.88	.83	.86	1
LC-Boost Factor	.86	.78	.79	.92	.84	.83	.84	.89
LL-Boost Factor	.94	.85	.85	.97	.90	.86	1.23	1.28
“Post Great Moderation" 2007:1-2018:8								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1	.98	1.08	.70	1	.92	.74	.86
DI	1.03	.95	.78	.83	.82	1	1.37	1.28
Lasso	1.19	.81	.87	.75	.83	.83	.79	.97
Boost	1.07	.84	.68	.66	.79	.84	1.36	1.41
Boost Factor	.92	.90	.70	.83	.76	.95	1.45	1.20
LC-Boost	.66	.74	.60	.65	.77	.83	.65	1.06
LC-Boost Factor	.73	.66	.64	.68	.75	.84	.78	.93
LL-Boost Factor	.76	.67	.63	.68	.68	.89	.87	1.05

Table 1.5: Relative MSFE $h = 1$

Full Out of Sample Period 1971:9-2018:8								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.04	1.04	1.01	.96	1.06	.96	.99	1.02
DI	.92	.87	.87	1.04	.92	.93	.92	.95
Lasso	1.06	.92	.94	1.26	.94	1	.88	.98
Boost	.91	.88	.82	1.03	.98	.94	.86	1
Boost Factor	.86	.91	.84	1.02	.90	.90	.85	.88
LC-Boost	.94	.91	.85	1.1	.96	.95	.84	.99
LC-Boost Factor	.85	.87	.85	1.02	.89	.91	.84	.84
LL-Boost Factor	.92	.95	.88	1.04	.98	.93	.91	.88
“Pre-Great Moderation" 1971:9-1982:12								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.08	1.02	1.06	.99	1.03	.99	1	1.03
DI	.86	.76	.75	1.18	1.01	.97	.90	.94
Lasso	1.21	.83	1.05	1.95	1.15	1.22	.89	.97
Boost	.90	.78	.77	1.19	1.12	1.10	.84	.97
Boost Factor	.80	.86	.80	1.15	.97	.91	.82	.81
LC-Boost	.96	.82	.77	1.30	1.08	1.14	.82	.96
LC-Boost Factor	.81	.86	.79	1.19	.96	1.02	.82	.78
LL-Boost Factor	.95	.96	.87	1.20	1.12	1.04	.85	.75
“Great Moderation" 1983:1-2006:12								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1	1.1	.99	.98	1.08	.92	.92	.97
DI	.94	.96	.96	.98	.93	.89	1.01	.98
Lasso	.91	1	.93	.98	.92	.94	.82	1.11
Boost	.88	.98	.90	1	.94	.89	1.02	1.25
Boost Factor	.91	.97	.91	.99	.91	.93	1.03	1.31
LC-Boost	.91	1.04	.93	.97	.90	.79	.98	1.24
LC-Boost Factor	.91	.95	.94	.97	.88	.84	1	1.30
LL-Boost Factor	.90	.99	.94	1.01	.96	.84	1.50	1.84
“Post Great Moderation" 2007:1-2018:8								
	IP	PAYEMS	UNRATE	CLF	RPI	CPI	FF	TB3MS
TV-AR	1.04	.94	.99	.87	1.05	.97	.84	.96
DI	.98	1.05	.90	.95	.89	.95	1.67	1.33
Lasso	1.05	1.02	.81	.93	.90	.88	.77	.96
Boost	.94	.93	.75	.92	1.01	.85	1.35	1.24
Boost Factor	.87	.95	.79	.94	.87	.86	1.46	1.35
LC-Boost	.96	.83	.79	1.1	1	.95	.89	.94
LC-Boost Factor	.84	.71	.79	.89	.87	.89	.94	.96
LL-Boost Factor	.88	.75	.78	.88	.98	.94	.85	1.06

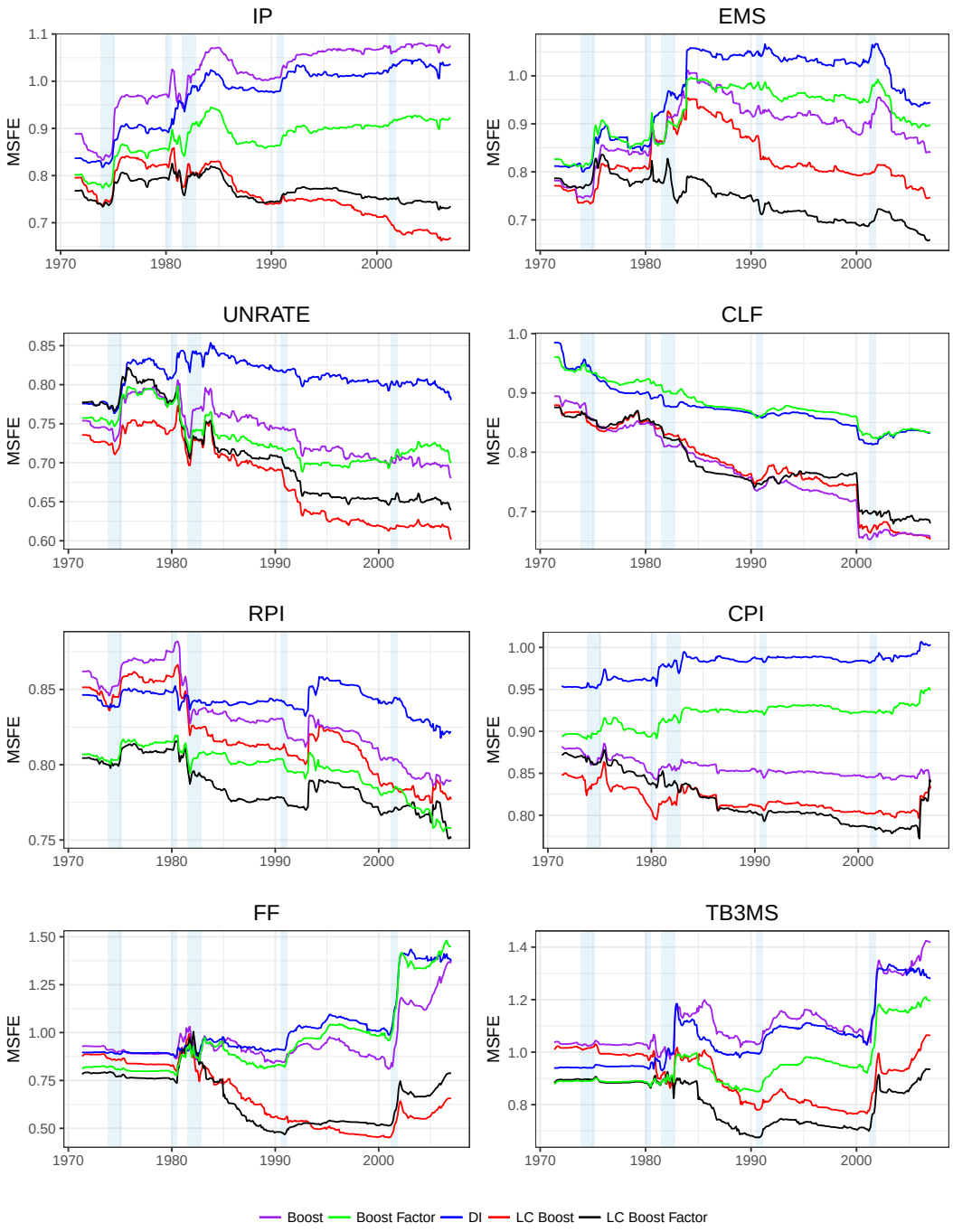

Figure 1.7: **MSFE by start date of out of sample period. Horizon $h = 3$.** See notes to figure 1.1.

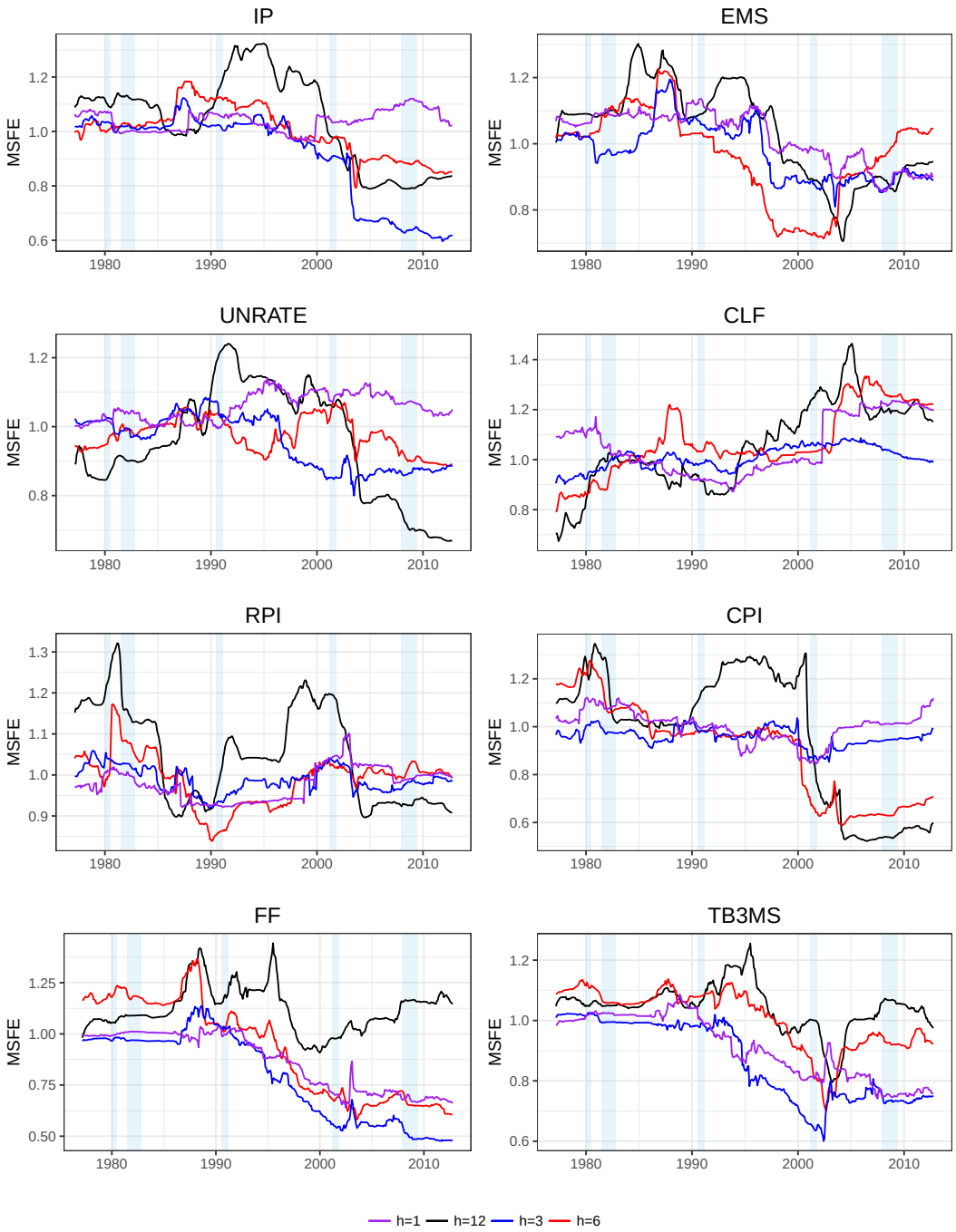

Figure 1.8: **L-MSFE of LC-Boost relative to L-MSFE of Boost**: This figure uses a window size of 70 observations to calculate the rolling MSFE, see section 1.8.2 for details.

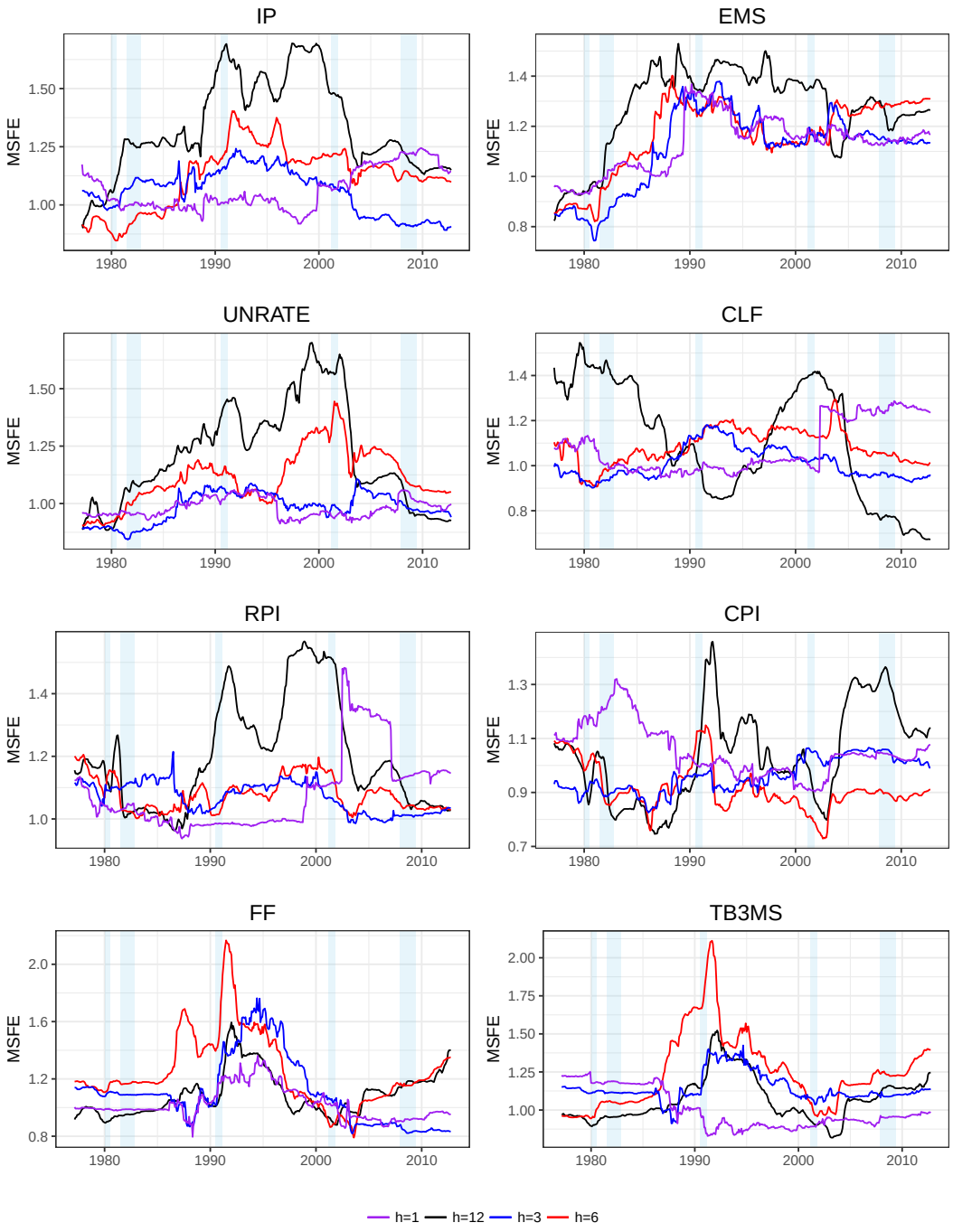

Figure 1.9: **L-MSFE of LC-Boost relative to L-MSFE of LC-Boost Factor**: This figure uses a window size of 70 observations to calculate the rolling MSFE, see section 1.8.2 for details.

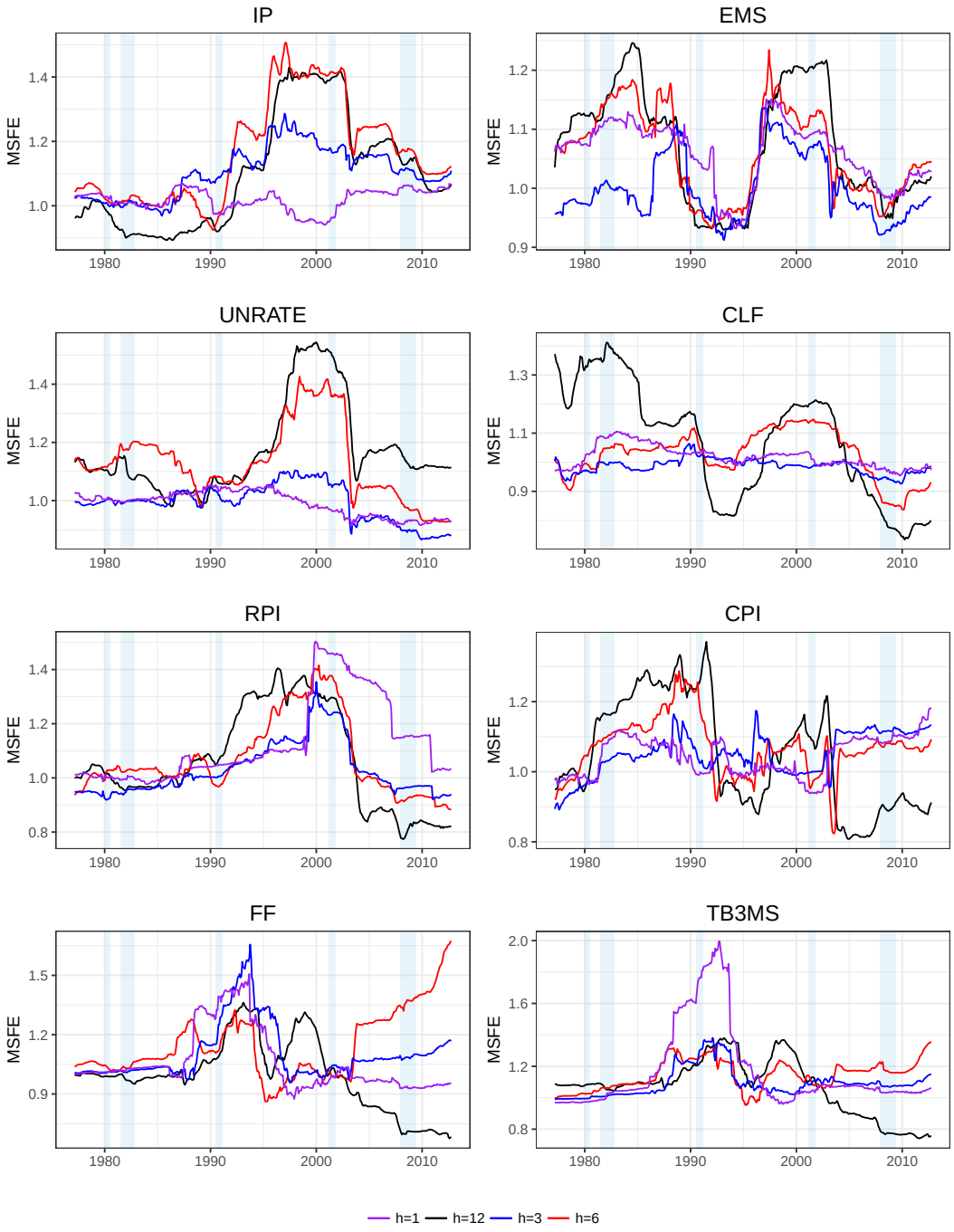

Figure 1.10: **L-MSFE of Boost Factor using 10 year rolling window relative to L-MSFE of LC-Boost Factor**: This figure uses a window size of 70 observations to calculate the rolling MSFE, see section 1.8.2 for details.

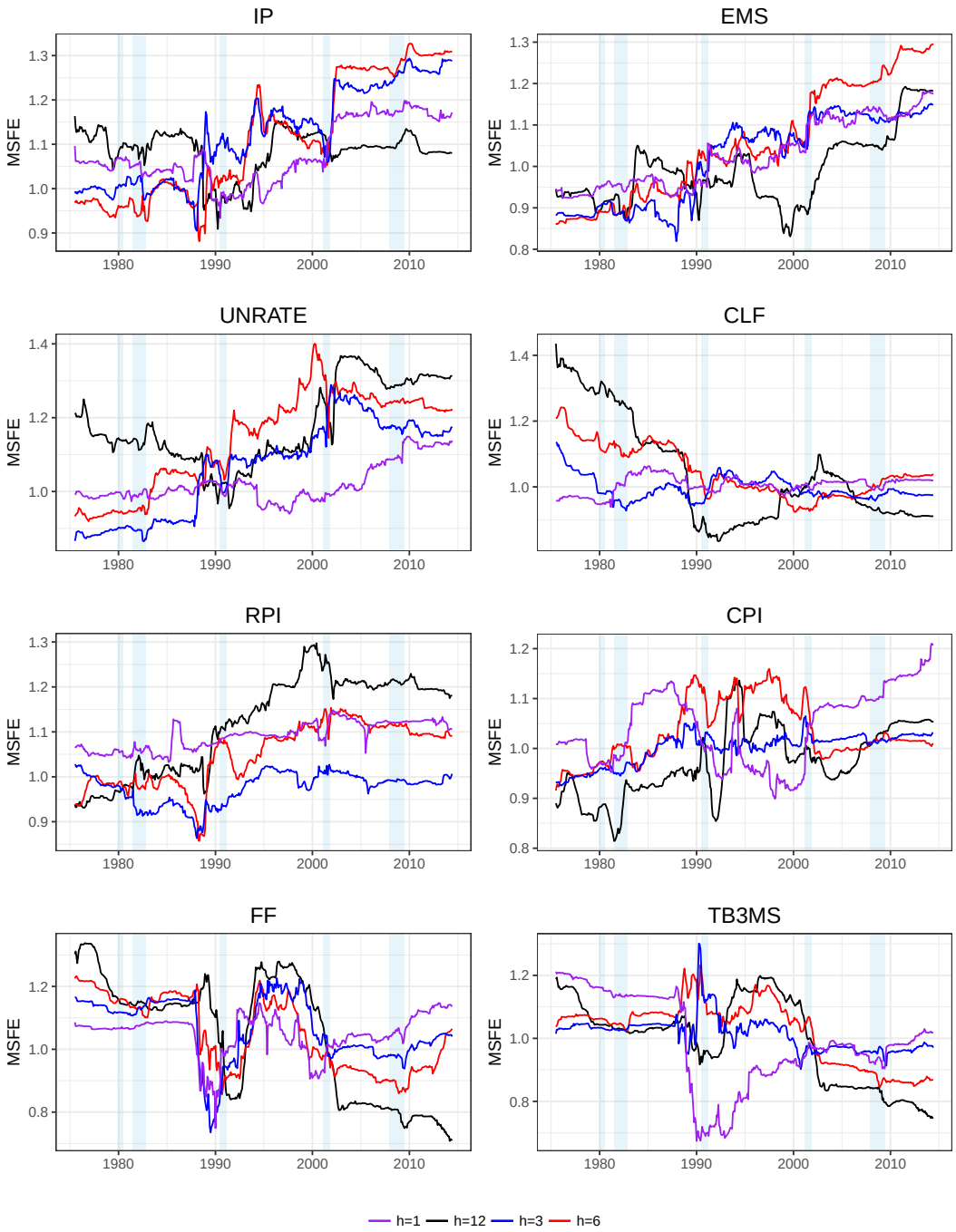

Figure 1.11: **L-MSFE of TV-DI relative to L-MSFE of LC-Boost Factor**: This figure uses a window size of 90 observations to calculate the rolling MSFE, see section 1.8.2 for details.

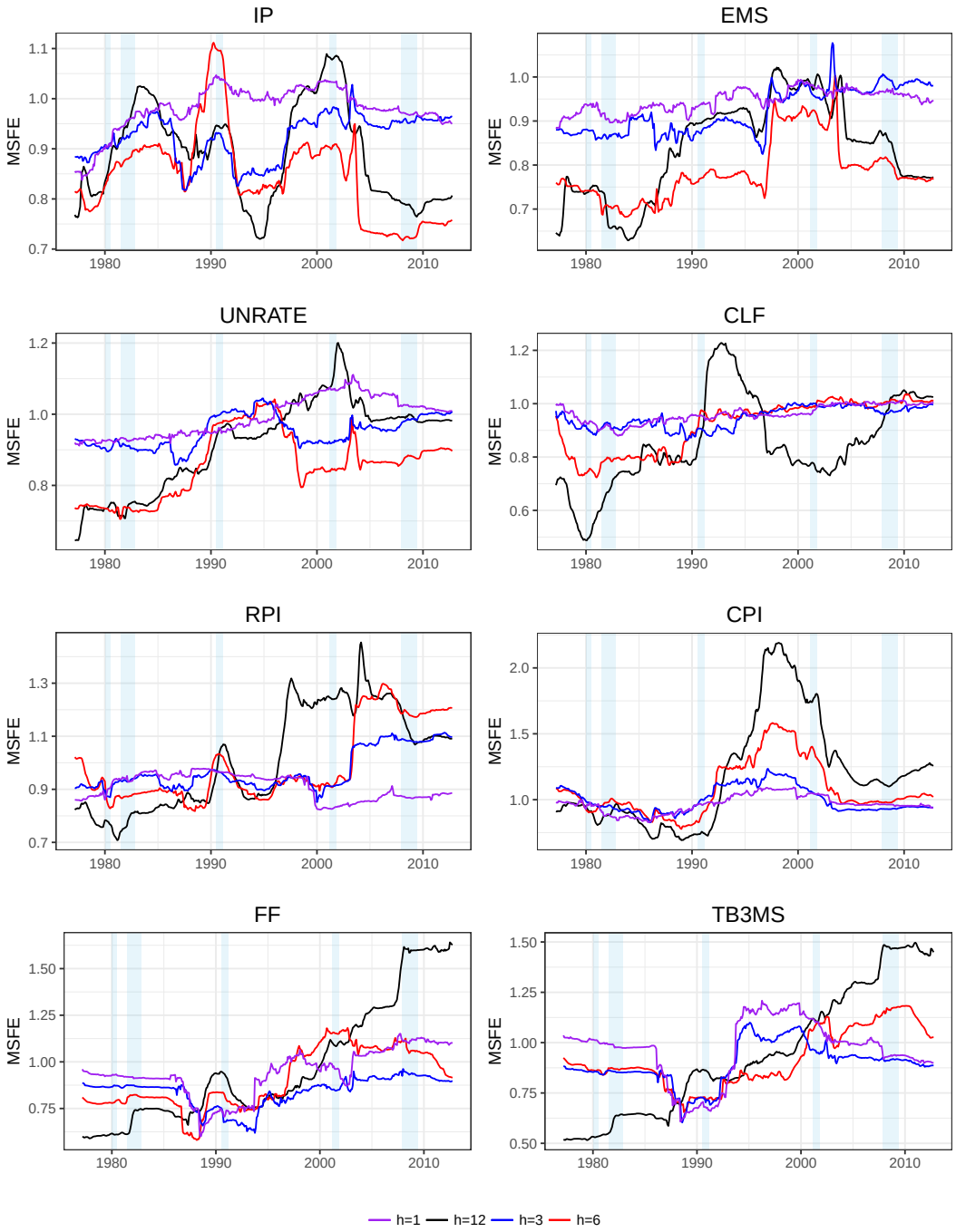

Figure 1.12: **L-MSFE of LC-Boost Factor relative to L-MSFE of LL-Boost Factor**: We use a window size of 70 observations, see notes to figure 1.5 for details. Colored lines represent the different horizons.

Table 1.6: DGP 1-14 : Relative MSFE, t_5 Innovations

DGP	AR (3)	Rolling AR (3)	Rolling Boost	LC-Boost	LL-Boost	Lasso
1	1.67	1.84	1.80	1.06	1.21	1.03
2	1.24	1.26	1.26	1.12	1.12	1.18
3	1.13	1.22	1.10	.76	.77	1.09
4	.96	1.04	.94	.72	.68	1.02
5	.77	.83	.76	.87	.68	.92
6	4.23	4.78	1.85	1.00	1.03	1.05
7	3.58	3.85	1.33	.90	.83	1.09
8	3.73	4.19	1.92	.72	.70	1.11
9	1.81	1.97	1.03	.71	.49	1.18
10	1.67	1.82	1.24	.92	.68	1.20
11	2.20	2.37	1.18	.85	.75	1.10
12	1.08	1.21	.42	.76	.21	1.13
13	1.99	2.05	1.06	.75	.54	1.21
14	.81	.89	.85	.92	.71	1.02

Chapter 2

Targeting Predictors via Partial Distance Correlation with Applications to Financial Forecasting

2.1 Introduction

High dimensionality is an increasingly common characteristic of data being collected in fields as diverse as genetics, neuroscience, astronomy, finance, and macroeconomics. In these fields, we frequently encounter situations in which the number of candidate predictors (p_n) is much larger than the number of samples (n), and one of the common ways statistical inference is made possible is by relying on the assumption of sparsity. The sparsity assumption, which states that only a small number of covariates contributes to the response, has led to a wealth of theoretical results and methods available for identifying important predictors in this high dimensional setting. These methods broadly fall into two classes: screening methods and penalized likelihood methods; we focus on the screening approach in this work. For the case where p_n is much larger than n, variable screening is usually used as a first stage method, which can then be followed by a second stage method, such as penalized likelihood methods, or principal components regression on the set of targeted predictors selected at the screening stage. This two stage procedure in many cases is computa-

tionally more feasible and lowers estimation error, especially if the dimensionality is very high at the first stage.

Fan and Lv (2008) proposed Sure Independence Screening (SIS) for the linear model, and it is based on ranking the magnitudes of the marginal Pearson correlations of the covariates with the response. A large amount of work has been done since then to generalize the procedure to various other types of models including: generalized linear models (Fan and Song, 2010), nonparametric additive models (Fan et al., 2011a), Cox proportional hazards model (Fan et al., 2010), linear quantile models (Ma et al., 2017), and varying coefficient models (Fan et al., 2014) [1]. The main theoretical result of these methods is the so called "sure screening property", which states that under appropriate conditions we can reduce the dimension of the feature space from size $p_n = O\left(\exp\left(n^{\alpha}\right)\right)$ to a far smaller size d_n, while retaining all the relevant predictors with probability approaching 1. We note that variable screening methods are closely related to the targeted predictors framework more commonly used in econometrics. As introduced in Bai and Ng (2008), the targeted predictors framework was focused on selecting predictors using linear dependence measures for the specific setting of forecasting from a second stage principal components regression. This can be thought of as a specific type of variable screening with linear dependence measures.

Although there has been a large amount of interest in developing screening methods, it is surprising to see that almost all of the works operate under the assumption of independent observations. This is even more surprising given the ubiquity of time

[1] In addition, model-free screening methods, which do not assume any particular model *a priori*, have also been developed. Some examples include: a distance correlation based method in Li et al. (2012b), the fused Kolmogorov filter in Mai et al. (2015), a conditional distance correlation method in Liu and Wang (2017), and a smoothing bandwidth based method in Feng et al. (2017). For a partial survey of screening methods, one can consult Liu et al. (2015).

dependent data in many scientific disciplines. Some examples in economics and finance include forecasting low frequency macroeconomic indicators, such as GDP or inflation rate, or financial asset returns using a large number of macroeconomic and financial time series and their lags as possible covariates (Stock and Watson, 2002a; Bai and Ng, 2009; Gu et al., 2018). These examples, amongst others, highlight the importance of developing screening methods specifically for time dependent data.

2.1.1 Our Contributions

In creating a screening method for stationary short range dependent time series, we aim to account for some of the unique features of time series data such as:

- A prior belief that a certain number of lags of the response variable are to be in the model.
- An ordered structure of the covariates, in which lower order lags of covariates are thought to be more informative than higher order lags.
- The frequent occurrence of multivariate response models such linear or nonlinear VAR models.

We also aim to have a model free screening approach which can handle continuous, discrete or grouped time series. Using a model free approach allows us to avoid placing assumptions on the structure of the model (i.e linearity) thereby making our methods robust to model misspecification at the screening stage. This gives us full flexibility to select a non-linear or non-parametric second stage procedure. Our goal is thus to extend the targeted predictors framework to more general non-linear or non-parametric models while accounting for some of the unique features of time series data. This is especially useful given that recent work has shown the benefits of considering non-linear and non-parametric models in forecasting macroeconomic and

financial time series.[2] Lastly, using a non-linear dependence measure is helpful even when we aim to fit a second stage linear model, as the marginal relationship between the predictors and the response can be highly non-linear.

To achieve our goals, we will introduce several distance correlation based screening procedures. Distance correlation (DC) was introduced by Székely et al. (2007), for measuring dependence and testing independence between two random vectors. The consistency, and weak convergence of sample distance correlation has been established for stationary time series in Zhou (2012) and Davis et al. (2016b). DC has a number of useful properties such as:

- The distance correlation of two random vectors equals to zero if and only if these two random vectors are independent.
- Ability to handle discrete time series, as well as grouped predictors.
- An easy to compute partial distance correlation has also been developed, allowing us to control for the effects of a multivariate random vector (Székely and Rizzo, 2014).

The first property allows us to develop a model free screening approach, which is robust to model misspecification. The second property is useful when dealing with linear or nonlinear VAR models for discrete or continuous data. The third property will allow us to account for the first two unique features of time series data mentioned previously.

Broadly speaking, we will be dealing with two types of models: univariate response models, some examples of which include linear or nonlinear autoregressive models with exogenous predictors (NARX), and multivariate response models such as linear

[2]Some examples include Gu et al. (2019, 2018) which showed that non-linear methods such as regression trees and neural networks are the best performing methods at forecasting asset returns. Additionally, in macroeconomics the sufficient forecasting framework (Fan et al., 2017) has shown improvements over using linear principal components regression.

or nonlinear VAR models. In both settings, we rely on partial distance correlation to build our screening procedures. Partial distance correlation produces a rich family of screening methods by taking different choices for the conditioning vector. In many applications, it is usually the case that researchers have prior knowledge that a certain subset of predictors is relevant to the response. Utilizing this prior knowledge usually enhances the screening procedure, as shown in the case of generalized linear models in Barut et al. (2016). Therefore our procedure can be viewed as a model free adaption of this principle to the time series setting. We discuss approaches for choosing the conditioning vector of each predictor, and we usually assume at least a few lags of the response variable are part of the conditioning vector of each predictor. We also discuss ways in which we can leverage the ordered structure of our lagged covariates to add additional variables to our conditioning vectors.

To motivate the multivariate response setting, consider a linear VAR(1) model: $\boldsymbol{x}_t = B_1\boldsymbol{x}_{t-1}+\boldsymbol{\eta}_t$, where $\boldsymbol{x}_t$ is a p-variate random vector. The number of parameters to estimate in this model is p^2, which can quickly become computationally burdensome even for screening procedures. In many cases however, there exists a certain group structure amongst the predictors, which is known to researchers in advance, along with a sparse conditional dependency structure between these groups (Basu et al., 2015). For example, in macroeconomics or finance, different sectors of the economy can be grouped into separate clusters. Using this group structure, we can apply the partial distance correlation to screen relationships at the group level, thereby quickly reducing the number of variables for a second stage procedure.

2.1.2 Comparisons to Existing Work

To the best of our knowledge there have been only two works, Chen et al. (2017) and Yousuf (2018), dealing with the issue in a stationary time series setting, both of which dealt with models with a univariate response. The former work extended the nonparametric independence screening approach used for independent observations to the time series setting. However, the method does not utilize the serial dependence in the data, or account for the unique properties of time series data we outlined. The latter work (Yousuf, 2018) extended the theory of SIS to heavy tailed and/or dependent data as well as proposing a GLS based screening method to correct for serial correlation. However, this work is limited to linear models and the other unique qualities of time series data outlined above are ignored.

Compared to the recent works on screening using distance correlation based methods (Wen et al., 2018; Liu and Wang, 2017), our work differs in a number of ways. First, our work deals with the time series setting, where both the covariates and response are stationary time series, and can have polynomially decaying tails. Second, our screening procedures are developed specifically in order to account for certain unique features in time series data mentioned previously. Lastly, we choose to rely on partial DC, instead of conditional DC (Wang et al., 2015), when controlling for confounding variables.

2.1.3 Organization

The rest of the paper is organized as follows. 2.2 reviews the distance correlation based methods we use for our algorithms. Sections 2.3 and 2.4 introduces our screening procedures for univariate response and multivariate response models respectively, along with their sure screening properties. Section 2.5 covers simulation results. We

present an application to forecasting monthly US market returns in Section 2.6. Section 2.7 covers the asymptotic properties of our methods. The concluding remarks are in Section 2.8. Lastly, the proofs for all theorems, along with additional simulations are placed in the supplementary material.

2.2 Review of Distance Correlation Based Methods

2.2.1 Preliminaries

We start with a brief overview of the distance covariance, distance correlation, and partial distance correlation measures.

Definition 2.2.1. For any random vectors $\boldsymbol{u} \in \mathcal{R}^q, \boldsymbol{v} \in \mathcal{R}^p$, let $\phi_{\boldsymbol{u}}(\boldsymbol{t}), \phi_{\boldsymbol{v}}(\boldsymbol{s})$ be the characteristic function of $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively. The distance covariance between $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as Székely et al. (2007):

$$dcov^2(\boldsymbol{u}, \boldsymbol{v}) = \int_{\mathcal{R}^{p+q}} |\phi_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{t}, \boldsymbol{s}) - \phi_{\boldsymbol{u}}(\boldsymbol{t})\phi_{\boldsymbol{v}}(\boldsymbol{s})|^2 \omega^{-1}(\boldsymbol{t}, \boldsymbol{s}) d\boldsymbol{t} d\boldsymbol{s},$$

where the weight function $\omega(\boldsymbol{t}, \boldsymbol{s}) = c_p c_q |\boldsymbol{t}|_p^{1+p} |\boldsymbol{s}|_q^{1+q}$, where $c_p = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}$. Throughout this article $|\boldsymbol{a}|_p$ stands for the Euclidean norm of $\boldsymbol{a} \in \mathcal{R}^p$.

Given this choice of weight function, by Székely et al. (2007), we have a simpler formula for the distance covariance. Let $(\boldsymbol{u}, \boldsymbol{v}), (\boldsymbol{u'}, \boldsymbol{v'}), (\boldsymbol{u''}, \boldsymbol{v''})$ be iid, each with joint distribution $(\boldsymbol{u}, \boldsymbol{v})$, and let:

$$S_1 = E(|\boldsymbol{u}-\boldsymbol{u'}|_p |\boldsymbol{v}-\boldsymbol{v'}|_q), \; S_2 = E(|\boldsymbol{u}-\boldsymbol{u'}|_p) E(|\boldsymbol{v}-\boldsymbol{v'}|_q), \; S_3 = E(|\boldsymbol{u}-\boldsymbol{u'}|_p) E(|\boldsymbol{v}-\boldsymbol{v''}|_q).$$

Then, provided that second moments exist, we have by remark 3 in Székely et al. (2007), $dcov^2(\boldsymbol{u}, \boldsymbol{v}) = S_1 + S_2 - 2S_3$. We can now estimate this quantity using moment

based methods. Suppose we observe $(\boldsymbol{u}_i, \boldsymbol{v}_i)_{i=1,\ldots,n}$, the sample estimates for the distance covariance and distance correlation are:

$$\hat{dcov}^2(\boldsymbol{u}, \boldsymbol{v}) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3, \text{ and } \widehat{dcor}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\widehat{dcov}(\boldsymbol{u}, \boldsymbol{v})}{\sqrt{\widehat{dcov}(\boldsymbol{u}, \boldsymbol{u})\widehat{dcov}(\boldsymbol{v}, \boldsymbol{v})}},$$

$$\text{where } \hat{S}_1 = n^{-2}\sum_{i,j=1}^{n} |\boldsymbol{u}_i - \boldsymbol{u}_j|_p |\boldsymbol{v}_i - \boldsymbol{v}_j|_q, \quad \hat{S}_2 = n^{-2}\sum_{i,j=1}^{n} |\boldsymbol{u}_i - \boldsymbol{u}_j|_p n^{-2}\sum_{i,j=1}^{n} |\boldsymbol{v}_i - \boldsymbol{v}_j|_q,$$

$$\hat{S}_3 = n^{-3}\sum_{i,j,l=1}^{n} |\boldsymbol{u}_i - \boldsymbol{u}_j|_p |\boldsymbol{v}_i - \boldsymbol{v}_l|_q.$$

As shown in Székely et al. (2007), the distance covariance given above has the property that $dcov(\boldsymbol{u}, \boldsymbol{v}) = 0$ if and only if $\boldsymbol{u}, \boldsymbol{v}$ are independent. Additionally, they prove consistency and weak convergence of the sample distance correlation estimator in the iid setting. These results were extended to strictly stationary α-mixing processes in Zhou (2012); Davis et al. (2016b) and Fokianos and Pitsillou (2017) .

Partial distance correlation (PDC) was introduced in Székely and Rizzo (2014), as a means of measuring nonlinear dependence between two random vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ while controlling for the effects of a third random vector $\boldsymbol{\mathcal{Z}}$. We refer to the vector $\boldsymbol{\mathcal{Z}}$ as the conditioning vector. Székely and Rizzo (2014) showed that the PDC can be evaluated using pairwise distance correlations. Specifically, the PDC between $\boldsymbol{u}$ and $\boldsymbol{v}$, controlling for $\boldsymbol{\mathcal{Z}}$, is defined as:

$$pdcor(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\mathcal{Z}}) = \frac{dcor^2(\boldsymbol{u}, \boldsymbol{v}) - dcor^2(\boldsymbol{u}, \boldsymbol{\mathcal{Z}})dcor^2(\boldsymbol{v}, \boldsymbol{\mathcal{Z}})}{\sqrt{1 - dcor^4(\boldsymbol{u}, \boldsymbol{\mathcal{Z}})}\sqrt{1 - dcor^4(\boldsymbol{v}, \boldsymbol{\mathcal{Z}})}},$$

if $dcor(\boldsymbol{u}, \boldsymbol{\mathcal{Z}}), dcor(\boldsymbol{v}, \boldsymbol{\mathcal{Z}}) \neq 1$, otherwise $pdcor(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\mathcal{Z}}) = 0$. The sample PDC $(\hat{pdcor}(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\mathcal{Z}}))$, is defined by plugging in the sample distance correlation estimators in the above definition. We note that theorem 3 in this work also establishes concentration bounds, in the time series setting, for the sample DC and PDC, which

should be of independent interest. For more details about the PDC measure, one can consult Székely and Rizzo (2014).

2.2.2 Partial DC vs Conditional DC

As we noted in the introduction, we have chosen to use partial DC instead of conditional DC in our screenign algorithms. There are a number of reasons for this: First, partial DC can be easily computed using pairwise distance correlations, and is much more computationally tractable when dealing with large numbers of predictors. Computing conditional DC is more complicated, therefore using conditional DC based screening procedure has a much higher computational burden. More importantly, the computation of conditional DC involves the choice of a bandwidth parameter to compute a kernel density estimate of the conditioning random vector. Selecting this bandwidth matrix is difficult in practice, especially for multivariate conditioning vectors where the curse of dimensionality rapidly deteriorates the quality of our estimates. In order to illustrate these effects, consider the following simple example: We generate $n = 100$ random samples from $Y_t = \sum_{j=1}^{6} \beta_j X_{t-1,j} + \epsilon_t$, where $\epsilon_t \sim N(0, 1)$, and $\beta_j = .75, j = 1, \ldots, 6$ and 0 otherwise. And $\boldsymbol{X}_t \sim N(0, I_p)$, where $p = 500$, and let our conditioning vector $\boldsymbol{Z}_{t-1} \sim N(0, I_r)$ be independent of both $Y_t, \boldsymbol{X}_{t-1}$. To implement a simple conditional DC screening and partial DC screening procedure, we first compute the partial DC and conditional DC between each covariate ($X_{t-1,j}$) and the response (Y_t), using $\boldsymbol{Z}_{t-1}$ as our conditioning vector.

We run 200 simulations and report the median minimum model size, i.e minimum size of the screened subset needed to include all the relevant covariates. The results are displayed in table 1, which show that as we increase r, the dimension of our conditioning vector, the performance of conditional DC screening completely breaks

Table 2.1: Median Minimum Model Size

	$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 5$
PDC	19.5	21	18	20.5	20.5
CDC	26.5	90	308	338	372

down, while partial DC screening retains its performance. From this simple example, we see that using conditional DC in place of partial DC is not an option when dealing with multidimensional conditioning vectors. We note that for time series models it is usually the case that we are dealing with multivariate conditioning vectors or even high dimensional conditioning vectors as in the case of VAR models.

Due to the reasons given above, using conditional DC is not a feasible option for screening when dealing with time series data. On the other hand, we note that unlike conditional DC, partial DC is not a measure of conditional independence, therefore a partial DC of zero does not imply conditional independence. Fortunately, it appears that in practice we rarely encounter the situation in which the variables are conditionally dependent but our partial DC gives us a value statistically indistinguishable from zero. In the supplementary material, we rerun all the simulations, given in the original conditional DC paper (Wang et al., 2015), where conditional dependence exists, and observe that partial DC has at least as much power as conditional DC in detecting this conditional dependence.

2.3 Screening Algorithms

We first review some basic ingredients of screening procedures. Let $\boldsymbol{y} = (Y_1, \ldots, Y_n)^T$ be our response time series, and let $\boldsymbol{x_{t-1}} = (X_{t-1,1}, \ldots, X_{t-1,m_n})$ denote the m_n predictor series at time $t-1$. Given that lags of these predictor series are possible covariates, we let $\boldsymbol{z_{t-1}} = (\boldsymbol{x_{t-1}}, \boldsymbol{x_{t-2}}, \ldots, \boldsymbol{x_{t-h_n}}) = (Z_{t-1,1}, \ldots, Z_{t-1,p_n})$ denote the

length p_n vector of covariates, where $p_n = m_n \times h_n$. Now we denote our set of active covariates as:

$$\mathcal{M}_* = \{j \leq p_n : F(Y_t|Y_{t-1}, \ldots, Y_{t-h_n}, \boldsymbol{z_{t-1}}) \text{ functionally depends on } Z_{t-1,j}\},$$

where $F(Y_t|\cdot)$ is the conditional cumulative distribution function of Y_t. The value h_n represents the maximum lag order we are considering for our response and predictor series. This value can be decided beforehand by the user, or can be selected using a data driven method. We note that we can let the value h_n differ between predictors, however for simplicity of presentation we assume the same value h_n for all predictors. Variable selection methods aim to recover $\mathcal{M}_*$ exactly, which can be a very difficult goal both computationally and theoretically, especially when $p_n \gg n$. In contrast, variable screening methods have a less ambitious goal, and aim to find a set S such that $P(\mathcal{M}_* \subset S) \to 1$ as $n \to \infty$. Ideally we would also hope that $|S| \ll p_n$, thereby significantly reducing the dimension of the feature space for a second stage method.

When developing screening algorithms for time series data, we would like to account for some of its unique properties as mentioned in the introduction. For models with a univariate response, these would be:

- A prior belief that a certain number of lags of the response variable are to be in the model.
- An ordered structuring of the covariates, in which lower order lags of covariates are thought to be more informative than higher order lags.

The first property can be easily accounted for, and there are many different ways to account for the second property. In this section we present two partial distance correlation based screening algorithms, which attempt to account for the ordered structure of our covariates.

2.3.1 Screening Algorithm I: PDC-SIS

In our first algorithm, PDC-SIS, we define the conditioning vector for the l^{th} lag of predictor series k ($X_{t-l,k}$) as:

$$\mathcal{S}_{k,l} = (Y_{t-1}, \ldots, Y_{t-h_n}, X_{t-1,k}, \ldots, X_{t-l+1,k}),$$

where $1 \leq l \leq h_n$. Since we are assuming *a priori* that a certain number of lags of Y_t are to be included in the model, $\{Y_{t-1}, \ldots, Y_{t-h_n}\}$ is part of the conditioning vector for all possible covariates. Our conditioning vector also includes all lower order lags for each lagged covariate we are considering. By including the lower order lags in the conditioning vector, our method tries to shrink towards sub-models with lower order lags. To illustrate this, consider the case where Y_t is strongly dependent on $X_{t-1,j}$ even while controlling for the effects of $Y_{t-1}, \ldots, Y_{t-h_n}$. Under this scenario, if $X_{t-1,j}$ has strong serial dependence, higher order lags of $X_{t-1,j}$ can be mistakenly selected by our screening procedure even if they are not in our active set of covariates.

For convenience, let $\boldsymbol{C} = \{\mathcal{S}_{1,1}, \ldots, \mathcal{S}_{m_n,1}, \mathcal{S}_{1,2}, \ldots, \mathcal{S}_{m_n,h_n}\}$ denote our set of conditioning vectors; where $C_{k+(l-1)*m_n} = \mathcal{S}_{k,l}$ is the conditioning vector for covariate $Z_{t-1,(l-1)*m_n+k}$. Our set of targeted predictors is:

$$\hat{\mathcal{M}}_{\gamma_n} = \left\{ j \in \{1, \ldots, p_n\} : \widehat{|pdcor|}(Y_t, Z_{t-1,j}; C_j) \geq \gamma_n \right\},$$

and we discuss how to select γ_n in section 2.3.3.

2.3.2 Screening Algorithm II: PDC-SIS+

As we have seen, the time ordering of the covariates allows us some additional flexibility in selecting the conditioning vector compared to iid setting. Our previous algorithm attempted to utilize the time series structure of our data by conditioning on previous lags of the covariate. However, rather than simply conditioning only on the previous lags of a covariate, we can condition on additional information available from previous lags of other covariates as well. One way to attempt this, and to potentially improve our algorithm, is to identify strong conditional signals at each lag level and add them to the conditioning vector for all higher order lag levels. By utilizing this conditioning scheme we can pick up on hidden significant variables in more distant lags, and also shrink toward models with lower order lags by controlling for false positives resulting from high autocorrelation, and cross-correlation.

We now give a formal description of PDC-SIS+. The conditioning vector for the first lag level of predictor series k is: $\mathcal{S}_{k,1} = (Y_{t-1}, \ldots, Y_{t-h_n})$, which coincides with the conditioning vector for the first lag level of PDC-SIS. Using the representation $\boldsymbol{z}_{t-1} = (\boldsymbol{x_{t-1}}, \ldots, \boldsymbol{x}_{t-h_n})$, we denote the strong conditional signal set for the first lag level as:

$$\hat{\mathcal{U}}_1^{\Gamma_n} = \left\{ j \in \{1, \ldots, m_n\} : |\widehat{pdcor}(Y_t, Z_{t-1,j}; \mathcal{S}_{j,1})| \geq \Gamma_n \right\}.$$

We then use this information to form our next conditioning vector:

$$\hat{\mathcal{S}}_{k,2} = \left(Y_{t-1}, \ldots, Y_{t-h_n}, X_{t-1,k}, \boldsymbol{z}_{t-1,\hat{\mathcal{U}}_{t-1}^{\Gamma_n}} \right),$$

where $\boldsymbol{z}_{t-1,\hat{\mathcal{U}}_1^{\Gamma_n}}$ is a sub-vector of $\boldsymbol{z}_{t-1}$ which is formed by extracting the indices contained in $\hat{\mathcal{U}}_1^{\Gamma_n}$. We note that any duplicates which result from overlap between $X_{t-1,k}$ and $\boldsymbol{z}_{t-1,\hat{\mathcal{U}}_1^{\Gamma_n}}$ are deleted. For convenience, we define

$\hat{C} = (\hat{\mathcal{S}}_{1,1}, \ldots, \hat{\mathcal{S}}_{m_n,1}, \hat{\mathcal{S}}_{1,2}, \ldots, \hat{\mathcal{S}}_{m_n,h_n})$ as our vector of estimated conditional sets. We then use $(\hat{\mathcal{S}}_{k,2})_{k \leq m_n}$ to compute the strong conditional signal set for the 2^{nd} lag level:

$$\hat{\mathcal{U}}_2^{\Gamma_n} = \left\{ j \in \{m_n + 1, \ldots, 2m_n\} : \widehat{|pdcor|}(Y_t, Z_{t-1,j}; \hat{C}_j) \geq \Gamma_n \right\}.$$

Repeating this procedure we obtain:

$$\hat{\mathcal{S}}_{k,l} = \left(Y_{t-1}, \ldots, Y_{t-h_n}, X_{t-1,k}, \ldots, X_{t-l+1,k}, \boldsymbol{z}_{t-1,\hat{\mathcal{U}}_1^{\Gamma_n}}, \ldots, \boldsymbol{z}_{t-1,\hat{\mathcal{U}}_{l-1}^{\Gamma_n}} \right).$$

We can also vary the threshold Γ_n for each lag level; for simplicity we leave it the same for each of our levels here. Our subset of predictors obtained from this procedure is:

$$\widetilde{\mathcal{M}}_{\gamma_n} = \left\{ j \in \{1, \ldots, p_n\} : \widehat{|pdcor|}(Y_t, Z_{t-1,j}; \hat{C}_j) \geq \gamma_n \right\}.$$

We denote $\mathcal{U}^{\Gamma_n} = \{\mathcal{U}_1^{\Gamma_n}, \ldots, \mathcal{U}_{h_n-1}^{\Gamma_n}\}$ as the population version of the strong conditional signal sets. Although the hope is that $\mathcal{U}^{\Gamma_n} \subset \mathcal{M}_*$, this is not necessary for the success of the algorithm. As seen in Barut et al. (2016) for the case of generalized linear models, conditioning on irrelevant variables could also enhance the power of a screening procedure. We will discuss how to choose the threshold Γ_n for $\hat{\mathcal{U}}^{\Gamma_n}$ in section 2.3.3. In practice we would prefer not to condition on too many variables, therefore the threshold for adding a variable to $\hat{\mathcal{U}}^{\Gamma_n}$ would be high.

Now, we have presented two classes of PDC screening methods. In the first class of methods, the conditional set of each covariate is known as *a priori*, while in the second class the conditional set is estimated from the data. We can easily modify our algorithms for both procedures depending on the situation; for example we can screen groups of lags at a time for certain covariates in PDC-SIS, or allow the lag length h_n to vary by predictor. Additionally, for either procedure we can condition on a small

number of lags of Y_t, and leave the higher order lags of Y_t as possible covariates in our screening procedure.

2.3.3 Threshold Selection

We first discuss how to select the parameter Γ_n for PDC-SIS+. For simplicity we will only use a single threshold for all lag levels. The idea is to create pseudo data $\{(\boldsymbol{x}_t, Y_t^*)\}_{t=1,\ldots,n}$, where $\{Y_t^*\}_{i=1,\ldots,n}$ is formed using a stationary bootstrap. This resampling procedure creates a null model, where $\hat{\omega}_j^* = \hat{pdcor}(Y_t^*, X_{t-1,j}; Y_{t-1}^*, Y_{t-2}^*, \ldots, Y_{t-h_n}^*)$, is a statistical estimate of zero, since asymptotically we have independence between $(Y_t^*, Y_{t-1}^*, \ldots, Y_{t-h_n}^*)$ and $X_{t-1,j}$. We can then choose the $\alpha = .99$ quantile of $\hat{\omega}_1^*, \ldots, \hat{\omega}_{p_n}^*$. Given that this threshold depends on a one resampling, we stabilize this threshold by constructing K (we choose $K = 5$) bootstrap samples.

Our procedure is as follows: we first form K bootstrap samples $\boldsymbol{y}^{(1)*}, \ldots, \boldsymbol{y}^{(K)*}$, and compute $\hat{\omega}_j^{(i)*} = \hat{pdcor}(Y_t^{(i)*}, X_{t-1,j}; Y_{t-1}^{(i)*}, \ldots, Y_{t-h_n}^{(i)*})$ for $i \leq K$, $j \leq p_n$. We then select the $\alpha = .99$ quantile of these values. In order to avoid conditioning on too many variables, an upper bound of $\lceil n^{1/2} \rceil$ variables can be added to our conditioning vector at each lag level. The idea of this procedure is that covariates above this threshold have a partial distance correlation easily distinguishable from zero. This procedure is similar to the random decoupling approach used in Weng et al. (2017) and Barut et al. (2016) for the iid setting.

For both PDC-SIS and PDC-SIS+ we also need to select a threshold γ_n to form our targeted set of predictors. We give three possible methods to select this threshold. The first is to use the bootstrap resampling procedure detailed above, which is a data driven method to select γ_n. Given we used $\alpha = .99$ to select Γ_n, we would want to use

a quantile between .95 and .99 to select γ_n. This is similar in spirit to thresholding by using a cutoff for the t-statistics of each marginal correlation measure used in Bai and Ng (2008). The second approach, which is more commonly used in the literature, is to select the top d_n predictors as ranked by our screening algorithm. When $p_n \gg n$, $d_n = n/log(n)$ or $d_n = n-1$ are common choices used in the literature. Alternatively, d_n can be set by the researcher using prior knowledge of the data.[3] The above two methods can be used without having decided on a second stage procedure beforehand. If one already has decided on the second stage modeling procedure, then γ_n can be selected by cross validation.

2.4 Screening for Multivariate Time Series Models

Multivariate time series models, such as linear VAR models, are commonly used in fields such as macroeconomics (Lütkepohl, 2005), finance, and more recently neuroscience (Valdés-Sosa et al., 2005), and genomics. VAR models provide a convenient framework for forecasting, investigating Granger causality, and modeling the temporal and cross-sectional dependence for large numbers of series. Since the number of parameters grows quadratically with the number of component series, VAR models have traditionally been restricted to situations where the number of component series is small. One way to overcome this limitation is by assuming a sparse structure in our VAR process, and using penalized regression methods such as the Lasso and adaptive Lasso (Zou, 2006b) to estimate the model. Examples of works which pursue this direction include Basu and Michailidis (2015), Basu et al. (2015), Kock and Callot (2015), and Nicholson et al. (2016). However, due to the quadratically increasing nature of the parameter space, penalized regression methods can quickly become computation-

[3]If the number of targeted predictors is selected beforehand, then one can set an upperbound of d_n variables which can added to the conditioning set in PDC-SIS+.

ally burdensome when we have a large panel of component series. For example, in a VAR(k) process: $\boldsymbol{x}_t = \sum_{i=1}^{k} B_i \boldsymbol{x}_{t-i} + \boldsymbol{\eta}_t$, where $\boldsymbol{x}_t \in \mathcal{R}^{m_n}$, $m_n = 1000, k = 5$, the number of parameters to estimate is 5×10^6. Additionally, these methods are restricted to linear VAR models, whereas there is considerable evidence of non-linear effects such as the existence of thresholds, smooth transitions, regime switching, and varying coefficients in fields such as macroeconomics and finance (Kilian and Lütkepohl, 2017).

Screening approaches can be used in this setting, and one option would be to screen separately for each of the m_n series. This can be computationally prohibitive since it requires estimating km_n^2 correlations. However, if we assume a group structure in the component series and a sparse conditional dependency structure between these groups, we can quickly reduce the feature space by screening at the group level using distance correlation based methods. To be more precise, let $\boldsymbol{x}_t$ be a non-linear VAR(k) process:

$$\boldsymbol{x}_t = g(\boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-k}) + \boldsymbol{\eta}_t, \text{ where } \boldsymbol{x}_t \in \mathcal{R}^{m_n}, \boldsymbol{\eta}_t \; iid. \tag{2.1}$$

For simplicity, we let all groups be of size g_n, let $e_n = m_n/g_n$ denote the total number of groups for a given lag level, and denote our groups $(G_{t-1,1}, \ldots, G_{t-k,e_n})$. To get a sense of the computational benefits of screening on the group level, assume for example, $m_n = 500, k = 1$, and we have 25 groups all of size $g_n = 20$. For this linear VAR (1) model, when $n = 200$, we note it takes about 350 times longer to compute all $m_n^2 = 500^2$ pairwise distance correlations $\left\{\hat{dcor}(X_{t,j}, X_{t-1,k})\right\}_{j \le m_n, k \le m_n}$ vs. computing all $e_n^2 = 25^2$ group pairwise distance correlations. After the group screening, examples of second stage procedures include: screening at the individual series level using partial distance correlations, or using a group lasso type procedure (Yuan and Lin, 2006) which can handle sparsity between groups and within groups

for a linear VAR model (Basu et al., 2015).

We now present the details of our group PDC-SIS procedure. We decide to condition on only one lag of the grouped response in our procedure, however this number can also be selected using a data driven procedure. Let $\mathcal{A}^{(i)} = \{(i,k,j) : k \in \{t-1,\ldots,t-h_n\}, j \leq e_n\} \setminus (i,t-1,i)$, refer to the set of possible group connections for $G_{t,i}$. We remove the entry $(i,t-1,i)$ from $\mathcal{A}^{(i)}$, since we are conditioning on $G_{t-1,i}$ and it will not be screened. Let the active group connections for group i be denoted as:

$$\mathcal{M}_*^{(i)} = \left\{ (i,k,j) \in \mathcal{A}^{(i)} : F\left(G_{t,i} | G_{t-1,i}, \bigcup_{r=t-h_n}^{t-1} \{G_{r,l}\}_{l \leq e_n} \right) \right.$$
$$\left. \text{functionally depends on } G_{k,j} \right\}.$$

Now let the overall active group connections set be denoted as $\mathcal{M}_* = \bigcup_{i=1}^{e_n} \mathcal{M}_*^{(i)}$. Similarly, our overall screened set is now:

$$\hat{\mathcal{M}}_{\gamma_n} = \bigcup_{i=1}^{e_n} \hat{\mathcal{M}}_{\gamma_n}^{(i)} = \left\{ (i,k,j) \in \bigcup_{i=1}^{e_n} \mathcal{A}^{(i)} : |\widehat{pdcor}(G_{t,i}, G_{k,j}; G_{t-1,i})| \geq \gamma_n, \right\}.$$

The sure screening properties of our group PDC-SIS procedure are similar to the ones presented in theorem 1, and are presented in the supplementary material. From these results, we can infer the maximum size of the groups is $o(n^{1/2-\kappa})$. Given this bound on the group size, our group PDC-SIS procedure is most advantageous when the number of component series (m_n) increases polynomially with the sample size. This is usually the case in most VAR models seen in practice. A group version of PDC-SIS+ can also be developed similarly to the procedure in section 3, however we do not pursue this direction, as it usually leads to situations where we are conditioning on large

numbers of variables.

2.5 Simulations

We now evaluate the performance of PDC-SIS and PDC-SIS+. We also include the performance of 4 other screening methods whose properties have been investigated in the time series setting, these include: marginal Pearson correlation screening (SIS), nonparametric independence screening (NIS), generalized least squares screening (GLSS), and distance correlation screening (DC-SIS). We also include the performance of a conditional DC screening approach (CDC-SIS), which uses the conditional DC in place of partial DC in our PDC-SIS algorithm.

We use the R package **energy** to compute the partial DC and the R package **cdcsis**, which was used in Wen et al. (2018), to compute the conditional DC. The **cdcsis** package computed the kernel density estimate of the conditioning vector (which is required to estimate the conditional DC) by estimating a diagonal bandwidth matrix using the plug-in method. The NIS estimator is computed using the R package **mgcv**, and the distance and partial distance correlation estimators are computed using the R package **energy**. For computational efficiency, the GLSS estimator is computed using the **nlme** package using an AR(1) approximation for the residual covariance matrix. Simulations for our group PDC-SIS procedure are contained in the supplementary material.

Unless noted otherwise, we fix our sample size $n = 200$, maximum number of lags considered $h_n = 3$, and the conditioning vector always includes three lags of our response. We vary the number of candidate series, m_n, from 500 to 1500, so the number of total covariates, p_n, varies from 1500 to 4500. We repeat each experiment 200 times, and report the median minimum model size needed to include all the

relevant covariates from $\boldsymbol{z}_{t-1} = (\boldsymbol{x_{t-1}}, \ldots, \boldsymbol{x}_{t-3})$. We note that for all procedures being considered, we will not be screening the lags of Y_t. In the supplementary materials, we also report the median rank of our relevant covariates for each procedure. We set $Y_0 = Y_{-1} = \ldots = Y_{-(h_n+1)} = 0$, and generate $n + 200$ samples of our model. We then discard the first $200 - h_n$ samples. To ensure stationarity when generating a nonlinear autoregressive model with exogenous predictors (NARX), we use the sufficient conditions provided in Masry and Tjøstheim (1997).

2.5.1 DGP's

Model 1:

$$Y_t = \sum_{j=1}^{6} \beta_j X_{t-1,j} + \epsilon_t, \text{and } \boldsymbol{x}_t = A_1 \boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t, \tag{2.2}$$

where $A_1 = .6 * I$, and $\boldsymbol{\eta}_t \overset{iid}{\sim} N(0, \Sigma_\eta)$, or $\boldsymbol{\eta}_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$. For this model, we set $\Sigma_\eta = \{.3^{|i-j|}\}_{i,j \leq m_n}$. For the error process, we have an AR(1) process: $\epsilon_t = \alpha \epsilon_{t-1} + e_t$ where $\alpha = .6$, and let $e_t \overset{iid}{\sim} N(0,1)$ or $e_t \overset{iid}{\sim} t_5$.

Model 2:

$$\begin{aligned} Y_t = {} & g_1(Y_{t-1}) + g_2(Y_{t-2}) + g_3(Y_{t-3}) + f_1(X_{t-1,1}) \\ & + f_2(X_{t-2,1}) + f_3(X_{t-1,2}) + f_4(X_{t-2,2}) + \epsilon_t, \end{aligned}$$

where the functions are defined as:

$$g_1(x) = .25x,\ g_2(x) = x\exp(-x^2/2),\ g_3(x) = -.6x + .3x(x > 0),$$
$$f_1(x) = 1.5x + .4x(x > 0),\ f_2(x) = -x,\ f_3(x) = 1.2x + .4x(x > 0),$$
$$f_4(x) = x^2 sin(2\pi x).$$

The covariate process is generated as in (2.2), with $A_1 = \{.4^{|i-j|+1}\}_{i,j\leq m_n}$ and we set $\Sigma_\eta = I_{m_n}$ with $\boldsymbol{\eta}_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\boldsymbol{\eta}_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$. Additionally, we set $\epsilon_t \overset{iid}{\sim} N(0,1)$ or t_5.

Model 3:

$$\begin{aligned} Y_t = {} & g_1(Y_{t-1}) + g_2(Y_{t-2}, Y_{t-1}) + g_3(Y_{t-3}, Y_{t-1}) + f_1(X_{t-1,1}, X_{t-1,4}) \\ & + f_2(X_{t-2,1}, X_{t-1,4}) + f_3(X_{t-1,2}, X_{t-1,4}) + f_4(X_{t-2,2}, X_{t-1,4}) \\ & + f_5(X_{t-1,3}, X_{t-1,4}) + f_6(X_{t-1,4}) + f_7(X_{t-1,3}, X_{t-1,4}) + \epsilon_t, \end{aligned}$$

where the functions are defined as:

$$g_1(x) = .2x + .2x(x > 0),\ g_2(x, y) = .2x + .1x(y > 0),$$
$$g_3(x, y) = x\exp(-y^2/2), f_1(x, y) = f_2(x, y) = f_4(x, y) = x\left(1 + \frac{1}{1 + .5\exp(-y)}\right),$$
$$f_3(x, y) = x\left(2 + \frac{2}{1 + .5\exp(-y)}\right), f_5(x), f_6(x) = 2x.$$
$$f_7(x, y) = x\left(1 + \frac{1}{1 + \exp(-y)}\right)$$

The covariate process is a VAR(2) process: $\boldsymbol{x}_t = A_1\boldsymbol{x}_{t-1} + A_2\boldsymbol{x}_{t-2} + \boldsymbol{\eta}_t$, where $A_1 = \{.3^{|i-j|+1}\}_{i,j\leq m_n}$, $A_2 = \{.2^{|i-j|+1}\}_{i,j\leq m_n}$, and $\Sigma_\eta = \{-.3^{|i-j|}\}_{i,j\leq m_n}$. As before, $\boldsymbol{\eta}_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\boldsymbol{\eta}_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$.

Model 4:

$$\begin{aligned} Y_t = &.25Y_{t-1} + .3Y_{t-2} + .3Y_{t-3} + f_1(X_{t-1,1}) + f_2(X_{t-2,1}) \\ &+ \beta_{1,t} f_3(X_{t-1,2}, X_{t-1,3}) + \beta_{2,t} f_4(X_{t-2,2}, X_{t-2,3}) + \beta_{3,t} f_5(X_{t-1,3}) \\ &+ \beta_{4,t} f_6(X_{t-2,3}) + f_7(X_{t-1,2}) + f_8(X_{t-2,2}, X_{t-1,2}) + \epsilon_t, \end{aligned}$$

where the functions are defined as:

$$\begin{aligned} &f_1(x), f_7(x) = 1.5x + .4x(x > 0), f_2(x) = 1.2x,\ f_3(x, y) = f_4(x, y) = xy, \\ &f_5(x), f_6(x) = x,\ f_8(x, y) = 1.2x + .4x(y > 0),\ \beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t} \overset{iid}{\sim} Unif(.5, 1)\ \forall t. \end{aligned}$$

The covariate process is generated as in (2.2), with $A_1 = \{.4^{|i-j|+1}\}_{i,j \leq m_n}$ and $\Sigma_\eta = \{-.3^{|i-j|}\}_{i,j \leq m_n}$. As in the previous examples, $\boldsymbol{\eta}_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\boldsymbol{\eta}_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$. We also note that the coefficients $\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t}$, are random at each time t.

Model 5: $Y_t = .25Y_{t-1} + .3Y_{t-2} + .3Y_{t-3} + X_{t-1,1} - X_{t-2,1} + .5X_{t-1,2} + .5X_{t-2,2} + \epsilon_t$.

The covariate process is generated as in (2.2), with $A_1 = \{.4^{|i-j|+1}\}_{i,j \leq m_n}$ and we set $\Sigma_\eta = I_{m_n}$ with $\boldsymbol{\eta}_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\boldsymbol{\eta}_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$. Additionally we set $\epsilon_t \overset{iid}{\sim} N(0, 1)$ or t_5.

2.5.2 Results

The results are displayed in table 2.10, and the entries below "Gaussian" correspond to the setting where both e_t and $\boldsymbol{\eta}_t$ are drawn from a Gaussian distribution. Accordingly the entries under "t_5" correspond to the case where e_t and $\boldsymbol{\eta}_t$ are drawn from a t_5 distribution. For DGP 1, We see that all methods besides CD-SIS perform well in this scenario, with GLSS performing best and PDC-SIS following closely even

though the lags of Y_t are not significant variables in this example. The results also show that including heavier tails deteriorates the performance of all methods for this model. As we seen previously, CDC-SIS is not able to hand conditional sets larger than 1 or 2 predictors, and as such performs very poorly for all DGPs. As for the computational burden of CDC-SIS, when $m_n = 500$, PDC-SIS takes about 15 seconds to compute, while CDC-SIS takes nearly 20 minutes to compute.

For DGP 2, the nonlinear transformations used are mainly threshold functions which are popular nonlinear transformations for time series data (Teräsvirta et al., 2010). We see that our method clearly outperforms the other methods across all scenarios. As seen in table 4 of the supplementary file, the covariate $X_{t-2,1}$ seems to be the most difficult to detect for the competing methods, and it appears our conditioning scheme greatly improves the detection of this signal. For DGP 3, we apply a logistic smooth transition function to the covariates, and for the autoregressive terms we mainly employ a hard threshold function. The results are displayed in table 2.10, and the median ranks of each of our significant variables can be found in table 5 of the supplementary file. The variable which appears to be the most difficult to detect seems to be the transition variable, $X_{t-1,4}$. DGP 4 contains a mix of threshold functions, interactions, and random coefficients. Looking at table 6 in the supplementary file, we notice that the covariates $X_{t-1,3}, X_{t-2,3}$, which only appear through random coefficient effects, are the most difficult to predict. Overall, we see that for models 1-5, PDC-SIS+ does as good and in most cases better than PDC-SIS.

2.6 Real Data Application: Forecasting Portfolio Returns

In this section, we present an application to forecasting US monthly equity portfolio returns. We first focus on forecasting market returns as measured by the CRSP

Table 2.2: Median Minimum Model Size

Gaussian, $p_n = 1500$					
	Model 1	Model 2	Model 3	Model 4	Model 5
PDC-SIS	7	61	76	42	24
PDC-SIS+	7	34	41	40	14.5
CDC-SIS	1150	1000	418	439	1010
DC-SIS	11	488	124	306.5	650
NIS	11	488	135	275	709
SIS	10	343.5	92	234.5	630
GLSS	6	179.5	194	800.5	1500
Gaussian, $p_n = 4500$					
	Model 1	Model 2	Model 3	Model 4	Model 5
PDC-SIS	11	149	239	100.5	59
PDC-SIS+	9	84	114.5	79	37
CDC-SIS	3424	2991	1257	1269	1009
DC-SIS	19	1051	444	842.5	1918
NIS	16	861	405	704	1950
SIS	13	722	326	588	1691
GLSS	6	592	1257	2214	4500
t_5, $p_n = 1500$					
	Model 1	Model 2	Model 3	Model 4	Model 5
PDC-SIS	13	79.5	56	51	52
PDC-SIS+	12	77.5	49.5	36.5	28.5
CDC-SIS	1208	522	418	415	3041
DC-SIS	20	408.5	146	306	680
NIS	33	513.5	172	328	681
SIS	21.5	447	148	265	647
GLSS	6	450.5	522	891.5	1500
t_5, $p_n = 4500$					
	Model 1	Model 2	Model 3	Model 4	Model 5
PDC-SIS	36.5	275.5	92	104	162
PDC-SIS+	31.5	121.5	85.5	99	78
CDC-SIS	3548	1930	1249	1237	3010
DC-SIS	68	951.5	314	814.5	1565
NIS	114	1100.5	413	851.5	1732
SIS	66.5	905	350	761	1567
GLSS	7	1386.5	1277	2843.5	4500

value weighted index, and the SP500 index. Additionally, we also focus on forecasting returns from 5 Fama-French portfolios sorted on Market Cap. For our predictor series we use book to market valuation ratios for Fama-French (FF) size and value sorted portfolios, in which U.S. stocks are divided into 25 or 100 portfolios sorted by market cap and book to market ratios. Kelly and Pruitt (2013) build on the present value identity and argue both theoretically and empirically that this cross section of disaggregated valuation ratios is predictive of future market returns. We use the same dataset which was originally analyzed in Kelly and Pruitt (2013), and can be obtained from the second author's website.[4]

Let $\boldsymbol{x}_t$ denote the 100 (or 25) FF portfolios at time t, and let Y_{t+1} denote the portfolio return at time $t+1$. Given that there seems to be some slight autocorrelation in the returns, we treat Y_t as a possible predictor and set it as a conditioning variable in PDC-SIS and PDC-SIS+. We also expand our predictor set to include $\boldsymbol{z}_t = (\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \boldsymbol{x}_{t-3})$. The linear factor augmented autoregression model in which the factors are estimated by principal components is very commonly used in econometrics (Stock and Watson, 2002a). Rather than estimating the principal components over the entire set of predictors $\boldsymbol{z}_t$, Bai and Ng (2008) and Bair et al. (2006) among others have shown that estimating the principal components on a targeted set of predictors can often lead to greater predictive accuracy. This procedure is sometimes known as supervised principal components, especially when marginal correlation screening is used to form our targeted set of predictors. This procedure can possibly be improved by using additional conditioning predictors, and non-linear measures of association, given that the marginal relationship between the response and individual predictors can be non-linear even when the joint relationship is linear.

[4]There were a small number of missing values ($\sim$ 1 percent for 100 portfolio dataset), which we imputed using the cross sectional median of the time period.

Given the above discussion, we report the forecasting performance of 8 different models. The first is a linear AR(1) model: $\hat{Y}_{t+1} = \hat{\alpha}_0 + \hat{\alpha} Y_t$. We then combine each of the six screening methods under consideration (we exclude CDC-SIS due to its poor performance in simulations) with a second stage Factor augmented autoregression (FAAR). For each of the screening methods we select the top $d_n = p_n/10$ predictors of $\boldsymbol{z}_t$, and add Y_t to form the screened subset of predictors. Using this subset of predictors, our forecasts are: $\hat{Y}_{t+1} = \hat{\beta}_0 + \hat{\alpha}_1 Y_t + \hat{\boldsymbol{\gamma}} \hat{\boldsymbol{F}}_t$, where $\hat{\boldsymbol{F}}_t = (\hat{F}_{t,1}, \ldots, \hat{F}_{t,k})$ are k factors which are computed as the first k principal components of the top $d_n = p_n/10$ predictors of $\boldsymbol{z}_t$, as ranked by the screening procedures. We select k using BIC and we allow for values between 2 and 5. Lastly, we include the performance of a factor augmented autoregression estimated using the full set of predictors.

We form expanding window out of sample forecasts , where the first out of sample forecast is for the time period 1980:1 (January 1980), and the last forecast is for time 2010:11. To construct the forecast for 1980:1, we use the observations between 1930:1 to 1979:11 to estimate the factors, and model parameters. Therefore for the models described previously, t =1930:1 to 1979:10. We then use the predictor values at t =1979:11 to form our forecast for 1980:1. The next window uses observations from 1930:1 to 1980:1 to forecast 1980:2. We use the same data split point for our out of sample forecasts as Kelly and Pruitt (2013), which gives us a total of 372 out of sample forecasts. For each of our 8 models, predictive ability is assessed through the out of sample R^2 which is defined as:

$$R^2_{OOS} = 100 * \left(1 - \frac{\sum_{t=T_0}^{2010:11} (\hat{Y}_t - Y_t)^2}{\sum_{t=T_0}^{2010:11} (\bar{Y}_t - Y_t)^2} \right) \tag{2.3}$$

where T_0 =1980:1 and $\bar{Y}_t = \sum_{i=1}^{t} Y_i/t$ is the prevailing mean at time t. The R^2_{OOS} ranges from $(-\infty, 100]$, where 100 indicates a perfect out of sample fit, and negative

values indicating that the method is outperformed by using a simple mean forecast.

We report the results when forecasting the SP500 and CRSP market index using either 25 FF portfolios and 100 FF portfolios in table 2.3. We observe that in both cases that finding targeted predictors via PDC-SIS and PDC-SIS+ easily outperform the alternatives, with the next best model is formed using DC-SIS. On the other hand, linear screening procedures such as SIS and GLSS underperform a factor model estimated on all the predictors, and underperform the mean forecast as well. From the results, we see that even non-linear screening methods all outperform linear methods. Given that DC-SIS does not condition on any predictors, this suggests that accounting for non-linearities in marginal relationships is important even when using linear second stage procedures. In order to assess the predictive ability of the models without the AR(1) term, we estimate our second stage forecasts using only the k factors. This makes our results directly comparable to previous works which ignored the AR(1) term. The results are in table 2.4, and although we have lower R^2_{OOS} values, we observe that our PDC methods easily outperform their competitors.

In table 2.5 we report the results when forecasting the 5 FF size sorted portfolios, and table 2.6 contains the results when excluding the AR(1) term.[5]. The first quintile corresponds to small cap stocks, and we see distance correlation methods strongly outperform other methods for this portfolio. Interestingly, in contrast to Kelly and Pruitt (2013), we obtain the highest predictability for this portfolio. And we generally find portfolios corresponding to smaller cap stocks easier to forecast than larger cap stocks using distance correlation methods, and we observe that the other methods we consider have the opposite trend.

As stated previously, we used a sample split date of 1980:1 for our out of sample

[5]We used the 25 FF portfolios as possible predictors along with their lags. The results were qualitatively similar for the 100 FF portfolio setting, thus we omit its results due to space considerations.

Table 2.3: $R^2_{OOS}(\%)$

	CRSP		SP 500	
	25 FF Portfolios	100 FF Portfolios	25 FF Portfolios	100 FF Portfolios
AR (1)	1.17	1.17	.15	.15
SIS FAAR	.12	-2.17	-3.84	-1.80
PDC-SIS FAAR	**1.95**	**2.02**	**.70**	**.71**
DC-SIS FAAR	.88	1.18	-.54	-.29
PDC-SIS+ FAAR	1.55	1.97	**.70**	**.71**
NIS FAAR	1.42	.85	-1.44	-1.06
GLSS FAAR	.05	-2.01	-3.33	-1.78
FAAR	1.45	-.22	.04	-.07

Table 2.4: $R^2_{OOS}(\%)$, Excluding AR(1) Term

	CRSP		SP 500	
	25 FF Portfolios	100 FF Portfolios	25 FF Portfolios	100 FF Portfolios
SIS Factor	-1.25	-3.75	-2.10	-4.63
PDC-SIS Factor	**.61**	**.77**	**.56**	**.56**
DC-SIS Factor	-.04	.14	-.66	-.85
PDC-SIS+ Factor	.26	.64	**.56**	**.56**
NIS Factor	.26	-.35	-1.41	-1.89
GLSS Factor	-1.19	-3.59	-2.00	-4.01
Factor	.38	-2.46	-.25	-1.82

forecasts. In order to show the robustness of our results to this choice of split date, we plot the R^2_{OOS} for the range of sample split dates between $T_0 = 1960{:}1$ to $T_0 = 1995{:}1$ in figure 2.1. We plot this for both the CRSP index and the SP 500 index, using 100 FF portfolios as predictors. For convenience of presentation we omit the performance of GLSS-FAAR and PDC-SIS+ FAAR models in our plot, given their very close performance with PDC-SIS FAAR and SIS-FAAR respectively. We see from the plot, that PDC-SIS Factor models outperform the alternatives over almost the entire range of sample split points. We also observe using a factor model estimated on all the predictors, along with linear screening rules underperform the historical mean forecast over the range of sample split points. Taken together, the results in this section show the benefits of screening using distance correlation based measures even for linear factor models.

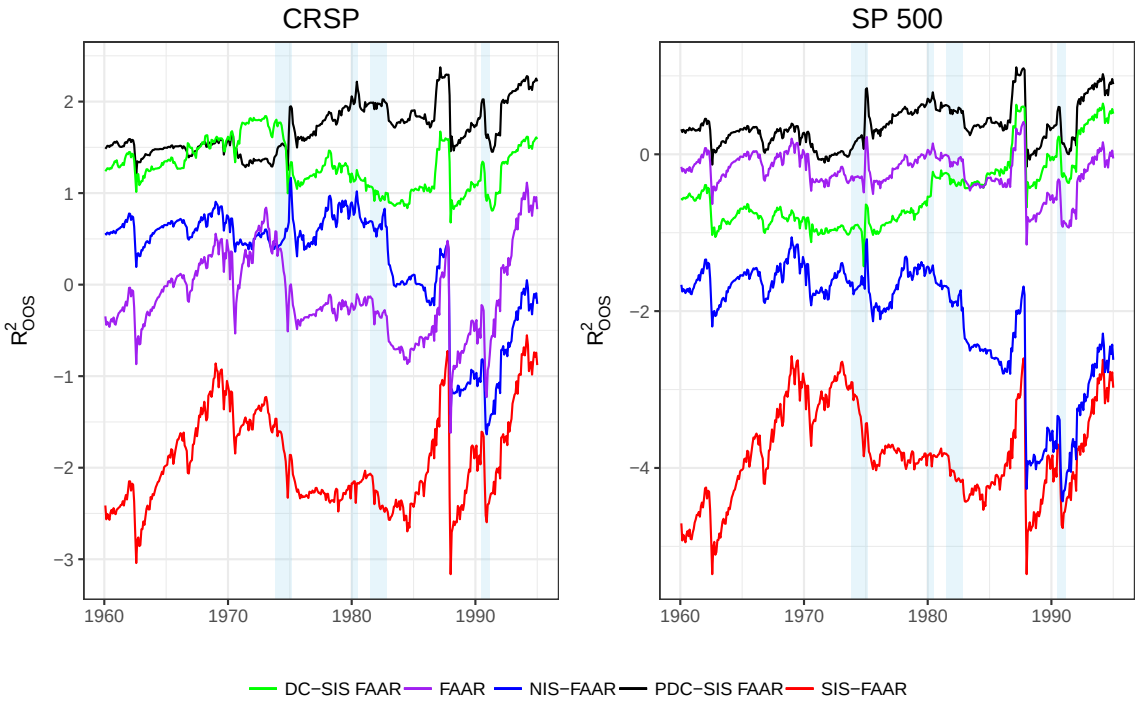

Figure 2.1: R^2_{OOS} **by Sample Split Date.** We select each date between 1960:1-1995:1 as our sample split point and plot the corresponding R^2_{OOS}. We omit the values for GLSS-FAAR and PDC+-FAAR due to having very close results to SIS-FAAR and PDC-FAAR respectively. We used 100 FF portfolios and their lags as possible predictors.

Table 2.5: $R^2_{OOS}(\%)$, Size Sorted Portfolios

	Quintile 1 (Small)	Quintile 2	Quintile 3	Quintile 4	Quintile 5 (Large)
AR(1)	5.12	1.69	1.83	1.45	**.16**
SIS FAAR	3.42	.57	.1	-.86	-1.4
PDC-SIS FAAR	5.42	2.85	2.63	**1.57**	-.6
DC-SIS FAAR	5.74	2.64	**2.73**	1.34	-.63
PDC-SIS+ FAAR	**5.80**	**2.92**	2.54	1.13	-.39
NIS FAAR	2.88	.26	.35	-.66	-.42
GLSS FAAR	4.0	1.26	.58	-.54	-1.47
FAAR	4.11	1.28	1.38	.57	.03

Table 2.6: $R^2_{OOS}(\%)$, Size Sorted Portfolios

	Quintile 1 (Small)	Quintile 2	Quintile 3	Quintile 4	Quintile 5 (Large)
SIS Factor	-2.16	-1.92	-2.4	-3.01	-1.74
PDC-SIS Factor	1.95	-.02	.62	-.05	-.60
DC-SIS Factor	**3.23**	-.09	-.57	-1.02	-.43
PDC-SIS+ Factor	2.48	**.95**	**1.63**	**1.19**	-.49
NIS Factor	-2.52	-2.01	-1.91	-2.61	-.68
GLSS Factor	-2.24	-2.45	-2.28	-2.89	-1.63
Factor	-1.14	-.69	-.57	-1.09	**-.16**

2.7 Asymptotic Properties

2.7.1 Dependence Measures

In order to establish asymptotic properties, we rely on two widely used dependence measures, the functional dependence measure and β-mixing coefficients. We give an overview of the functional dependence measure framework here, and one can consult (Davidson, 1994) for an overview of β-mixing coefficients. For univariate processes, $(Y_i \in \mathcal{R})_{i\in\mathbb{Z}}$, we assume Y_i is a causal, strictly stationary, ergodic process with the following form:

$$Y_i = g\left(\ldots, e_{i-1}, e_i\right), \tag{2.4}$$

where $g(\cdot)$ is a real valued measurable function, and e_i are iid random variables. And for multivariate processes, such as the covariate process $(\boldsymbol{x}_i \in \mathcal{R}^{p_n})_{i\in\mathbb{Z}}$, we assume the

following representation:

$$\boldsymbol{x_i} = \boldsymbol{h}\left(\ldots, \boldsymbol{\eta}_{i-1}, \boldsymbol{\eta}_i\right). \tag{2.5}$$

Where $\boldsymbol{\eta}_i, i \in \mathbb{Z}$, are iid random vectors, $\boldsymbol{h}(\cdot) = (h_1(\cdot)\ldots, h_{p_n}(\cdot))$, $\boldsymbol{x_i} = (X_{i1}, \ldots, X_{ip_n})$, and $X_{ij} = h_j(\ldots, \boldsymbol{\eta}_{i-1}, \boldsymbol{\eta}_i)$.

Processes having these representations are sometimes known as Bernoulli shift processes (Wu, 2009), and include a wide range of stochastic processes such as linear processes with their nonlinear transforms, Volterra processes, Markov chain models, nonlinear autoregressive models such as threshold auto-regressive (TAR), bilinear, GARCH models, among others (Wu, 2011, 2005). These representations allow us to quantify dependence using a functional dependence measure introduced in Wu (2005). The functional dependence measure for a univariate process and multivariate processes is defined respectively as:

$$\begin{aligned}
\delta_q(Y_i) &= ||Y_i - g\left(\mathcal{F}_i^*\right)||_q = \left(E|Y_i - g\left(\mathcal{F}_i^*\right)|^q\right)^{1/q}, \\
\delta_q(X_{ij}) &= ||X_{ij} - h_j\left(\mathcal{H}_i^*\right)||_q = \left(E|X_{ij} - h_j\left(\mathcal{H}_i^*\right)|^q\right)^{1/q},
\end{aligned} \tag{2.6}$$

where $\mathcal{F}_i^* = (\ldots, e_{-1}, e_0^*, e_1, \ldots, e_i)$ with $e_0^*, e_j, j \in \mathbb{Z}$ being iid. And for the multivariate case, $\mathcal{H}_i^* = (\ldots, \boldsymbol{\eta}_{-1}, \boldsymbol{\eta}_0^*, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_i)$ with $\boldsymbol{\eta}_0^*, \boldsymbol{\eta}_j, j \in \mathbb{Z}$ being iid. Since we are replacing e_0 by e_0^*, we can think of this as measuring the dependency of y_i on e_0, since we are keeping all other inputs the same. We assume the cumulative functional dependence measures are finite:

$$\Delta_{0,q}(\boldsymbol{y}) = \sum_{i=0}^{\infty} \delta_q(Y_i) < \infty, \text{ and } \Phi_{m,q}(\boldsymbol{x}) = \max_{j \le p_n} \sum_{i=m}^{\infty} \delta_q(X_{ij}) < \infty. \tag{2.7}$$

This short range dependence condition implies, by the proof of theorem 1 in Wu and

Pourahmadi (2009), the auto-covariances are absolutely summable.

We note that compared to functional dependence measures, β-mixing coefficients can be defined for any stochastic processes, and are not limited to Bernoulli shift processes. On other hand, functional dependence measures are easier to interpret and compute since they are related to the data generating mechanism of the underlying process. In many cases using the functional dependence measure also requires less stringent assumptions (see Wu and Wu (2016), Yousuf (2018) for details). Although there is no direct relationship between these two dependence frameworks, fortunately there are a large number of commonly used time series processes which are β-mixing and satisfy (2.7). For example, under appropriate conditions, linear processes, ARMA, GARCH, ARMA-ARCH, threshold autoregressive, Markov chain models, amongst others, can be shown to be β-mixing (see Pham and Tran (1985), Carrasco and Chen (2002), An and Huang (1996), Lu (1998) for details).

2.7.2 Asymptotic Properties: PDC-SIS

To establish sure screening properties, we introduce the following conditions.

Condition 2.7.1. Assume $|pdcor(Y_t, Z_{t-1,k}; C_k)| \geq c_1 n^{-\kappa}$ for $k \in M_*$ and $\kappa \in (0, 1/2)$.

Condition 2.7.2. Assume the response and the covariate processes have representations (2.4) and (2.8), respectively. Additionally, we assume the following decay rates $\Phi_{m,r}(\boldsymbol{x}) = O(m^{-\alpha_z}), \Delta_{m,q}(\boldsymbol{y}) = O(m^{-\alpha_y})$, for some $\alpha_z, \alpha_y > 0$, $q > 2, r > 4$ and $\tau = \frac{qr}{q+r} > 2$.

Condition 2.7.3. Assume the response and the covariate processes have representations (2.4) and (2.8) respectively. Additionally assume
$\upsilon_z = \sup_{q\geq 2} q^{-\tilde{\alpha}_z} \Phi_{0,q}(\boldsymbol{x}) < \infty$ and $\upsilon_y = \sup_{q\geq 2} q^{-\tilde{\alpha}_y} \Delta_{0,q}(\boldsymbol{y}) < \infty$, for some $\tilde{\alpha}_z, \tilde{\alpha}_y \geq 0$.

Condition 2.7.4. Assume the process $\{(Y_t, \boldsymbol{x}_t)\}$ is β-mixing, with mixing rate $\beta_{xy}(a) = O(\exp(-a^{\lambda_1}))$, for some $\lambda_1 > 0$.

Condition 2.7.1 is a standard population level assumption which allows covariates in the active set to be detected by our screening procedure. Condition 2.7.2 is similar to the one used in Yousuf (2018) and Wu and Wu (2016), and assumes both the response and covariate processes are causal Bernoulli shift processes, and have at least 2 and 4 finite moments respectively. Additionally it presents the dependence conditions on these processes, where higher values of $\alpha_x, \alpha_\epsilon$ indicate weaker temporal dependence. Examples of response processes which satisfy condition 2.7.2 include stationary, causal, finite order ARMA, GARCH, ARMA-GARCH, bilinear, and threshold autoregressive processes, all of which have exponentially decaying functional dependence measures (see Wu (2011) for details). For the covariate process, assume $\boldsymbol{x}_i$ is a vector linear process: $\boldsymbol{x}_i = \sum_{l=0}^{\infty} A_l \boldsymbol{\eta}_{i-l}$. where $\{A_l\}$ are $m_n \times m_n$ coefficient matrices and $\{\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{im_n})\}$ are iid random vectors with $cov(\boldsymbol{\eta_i}) = \Sigma_\eta$. For simplicity, assume $\{\eta_{i,j}, j = 1, \ldots, m_n\}$ are identically distributed, then $\delta_q(X_{ij}) = ||A_{i,j}\boldsymbol{\eta}_0 - A_{i,j}\boldsymbol{\eta}_0^*||_q \leq 2|A_{i,j}|||\eta_{0,1}||_q$, where $A_{i,j}$ is the j^{th} column of A_i. Define $||A_i||_\infty$ as the maximum absolute row sum of A_i, then if $||A_i||_\infty = O(i^{-\beta})$ for $\beta > 1$, we have $\Phi_{m,q}(\boldsymbol{x}) = O(m^{-\beta+1})$. Other examples include stable VAR processes, and multivariate ARCH processes which have exponentially decaying cumulative functional dependence measures (Wu and Wu, 2016; Yousuf, 2018). We note that it is clear that if $\boldsymbol{x}_i$ satisfies condition 2.7.2, then $\boldsymbol{z}_i$ trivially satisfies it as well. Condition 2.7.3 strengthens the moment requirements of condition 2.7.2, and requires that all moments of the covariate and response processes are finite. To illustrate the role of the constants $\tilde{\alpha}_z$ and $\tilde{\alpha}_y$, consider the example where y_i is a linear process: $y_i = \sum_{j=0}^{\infty} f_j e_{i-j}$ with e_i iid and $\sum_{l=0}^{\infty} |f_l| < \infty$, then $\Delta_{0,q}(\boldsymbol{y}) = ||e_0 - e_0^*||_q \sum_{l=0}^{\infty} |f_l|$.

If we assume e_0 is sub-Gaussian, then $\tilde{\alpha}_y = 1/2$, since $||e_0||_q = O(\sqrt{q})$. Similarly, if e_i is sub-exponential, we have $\tilde{\alpha}_y = 1$.

To understand the inclusion of condition 2.7.4, consider the U-statistic:

$$U_r(S_{t_1}, \ldots, S_{t_r}) = \binom{n}{r} \sum_{t_1 \leq t_2 \leq \ldots \leq t_r \leq n} h(S_{t_1}, \ldots, S_{t_r}),$$

which aims to estimate $\theta(h) = \int h(S_{t_1}, \ldots, S_{t_r}) d\mathcal{P}(S_1) \ldots d\mathcal{P}(S_r)$. When $S_1, \ldots, S_n$ are iid, the U-statistic is an unbiased estimator of $\theta(h)$, however for $r > 1$ the U-statistic is no longer unbiased if S_t is serially dependent. Since our sample distance correlation estimate can be written as a sum of U-statistics (Li et al., 2012b), condition 2.7.4 is needed to control the rate at which the above bias vanishes as $n \to \infty$. Conditions 2.7.2 and 2.7.4 are frequently used when dealing with time series data (Wu and Pourahmadi, 2009; Xiao and Wu, 2012; Davis et al., 2016b).

Throughout this paper, let $\alpha = \min(\alpha_x, \alpha_y)$, and $\varrho = 1$, if $\alpha_z > 1/2 - 2/r$, otherwise $\varrho = r/4 - \alpha_z r/2$. Let $\iota = 1$ if $\alpha > 1/2 - 1/\tau$, otherwise $\iota = \tau/2 - \tau\alpha$, and let $\zeta = 1$, if $\alpha_y > 1/2 - 2/q$, otherwise $\zeta = q/4 - \alpha_y q/2$. Additionally, let $K_{y,q} = \sup_{m \geq 0}(m+1)^{\alpha_y} \Delta_{m,q}(\boldsymbol{y})$, and $K_{z,r} = \sup_{m \geq 0}(m+1)^{\alpha_z} \Phi_r(\boldsymbol{x})$. Given condition 2.7.3, it follows that $K_{\epsilon,q}, K_{z,r} < \infty$. Let $t_n = \max_j dim(C_j)$, be the maximum dimension of the conditional vectors. We define $\tilde{\psi} = \frac{2}{1+2\tilde{\alpha}_z+2\tilde{\alpha}_y}$, $\tilde{\varphi} = \frac{2}{1+4\tilde{\alpha}_z}$, $\tilde{\alpha} = \frac{2}{1+4\tilde{\alpha}_y}$. Lastly, for ease of presentation, let $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \ldots, \hat{\omega}_{p_n})$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{p_n})$, where

$\omega_k = pdcor(Y_t, Z_{t-1,k}; C_k)$, $\hat{\omega}_k = \widehat{pdcor}(Y_t, Z_{t-1,k}; C_k)$. In addition, let

$$
\begin{aligned}
a_n &= n^2 \left[\exp\left(-\frac{n^{1/2-\kappa}}{t_n \upsilon_y^2} \right)^{\tilde{\alpha}} + \exp\left(-\frac{n^{1/2-\kappa}}{t_n \upsilon_z \upsilon_y} \right)^{\tilde{\psi}} + \exp\left(-\frac{n^{1/2-\kappa}}{t_n \upsilon_z^2} \right)^{\tilde{\varphi}} \right], \\
b_n &= n^2 \left[\frac{t_n^{r/2} n^{\zeta} K_{y,r}^r}{n^{r/2-r\kappa/2}} + \frac{t_n^{r/2} n^{\iota} K_{z,r}^{r/2} K_{y,r}^{r/2}}{n^{r/2-r/2\kappa}} + \frac{t_n^{r/2} n^{\varrho} K_{z,r}^r}{n^{r/2-r\kappa/2}} \right. \\
&\quad \left. + \exp\left(-\frac{n^{1-2\kappa}}{t_n^2 K_{z,r}^4} \right) + \exp\left(-\frac{n^{1-2\kappa}}{t_n^2 K_{z,r}^2 K_{y,r}^2} \right) + \exp\left(-\frac{n^{1-2\kappa}}{t_n^2 K_{y,r}^4} \right) \right], \\
c_n &= \frac{t_n^{r/2} K_{y,r}^r}{n^{r/4-r\kappa/2}} + \frac{t_n^{r/2} K_{z,r}^{r/2} K_{y,r}^{r/2}}{n^{r/4-r/2\kappa}} + \frac{t_n^{r/2} K_{z,r}^r}{n^{r/4-r\kappa/2}}.
\end{aligned}
$$

For simplicity and convenience of presentation, we assume $q = r$, and one can consult the proof for the general case. The following theorem presents the sure screening properties of PDC-SIS.

Theorem 3. *1. Suppose conditions 2.7.1, 2.7.3, and 2.7.4 hold. For any $c_2 > 0$, we have:*

$$P(\max_{j \le p_n} |\hat{\omega}_k - \omega_k| > c_2 n^{-\kappa}) \le O(p_n a_n).$$

2. Suppose conditions 2.7.1, 2.7.3, and 2.7.4 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \le c_1/2$, we have:

$$P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n} \right) \ge 1 - O(s_n a_n).$$

3. Suppose conditions 2.7.1, 2.7.2, and 2.7.4 hold. For any $c_2 > 0$, we have:

$$\text{if } r < 12, \qquad P(\max_{j \le p_n} |\hat{\omega}_j - \omega_j| > c_2 n^{-\kappa}) \le O(p_n c_n);$$

$$\text{if } r \ge 12, \qquad P(\max_{j \le p_n} |\hat{\omega}_k - \omega_k| > c_2 n^{-\kappa}) \le O(p_n b_n).$$

4. Suppose conditions 2.7.1, 2.7.2, and 2.7.4 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \le c_1/2$,

we have:

$$\text{if } r < 12, \qquad P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(s_n c_n);$$

$$\text{if } r \geq 12, \qquad P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(s_n b_n).$$

From the above theorem, we observe that the range of p_n depends on the temporal dependence in both the covariate and the response processes, the strength of the signal (κ), and the moment conditions. We also have two cases for finite polynomial moments, one for $r < 12$ and one for $r \geq 12$. This is due to our proof technique which relies on both Nagaev and Rosenthal type inequalities. For the case of low moments, we obtain a better bound using a Rosenthal type inequality combined with the Markov inequality, whereas for higher moments Nagaev type inequalities lead to a better bound; more details can be found in the proof which is provided in the supplementary file.

For example, if we assume only finite polynomial moments with $r = q$ and $r < 12$, then $p_n = o(n^{r/4 - r\kappa/2})$. If we assume $\alpha \geq 1/2 - 2/r$ and $r > 12$, $p_n = o(n^{r/2 - r\kappa/2 - 3})$. The constants $K_{z,r}$ and $K_{y,q}$, which are related to the cumulative functional dependence measures, represent the effect of temporal dependence on our bounds when $\alpha \geq 1/2 - 2/r$. However, when using Nagaev type inequalities, there is an additional effect in the case of stronger dependence in the response or covariate process (i.e. $\alpha < 1/2 - 2/r$). For instance, if $\alpha_x = \alpha_\epsilon$ and $q = r$, the range for p_n is reduced by a factor of $n^{r/4 - \alpha r/2}$ in the case of stronger dependence. We observe that if the response and covariates are sub-Gaussian, $p_n = o(\exp(n^{\frac{1-2\kappa}{3}}))$, and if they are sub-exponential, $p_n = o(\exp(n^{\frac{1-2\kappa}{5}}))$.

By choosing an empty conditional set for all the variables, our procedure reduces to the distance correlation screening (DC-SIS) introduced in Li et al. (2012b) for the iid setting. Assuming sub-Gaussian response and covariates, Li et al. (2012b)

obtained $p_n = o(\exp(n^{\frac{1-2\kappa}{3}}))$ for DC-SIS, which matches our rate. In the iid setting with finite polynomial moments, we can use the truncation method in their proof and combined with the Markov inequality to obtain $p_n = o(\exp(n^{r/4-r\kappa/2-1}))$. Our results, which rely on a different proof strategy than the truncation method, provide a better bound even in this setting.

2.7.3 Asymptotic Properties: PDC-SIS+

To show the asymptotic properties associated with PDC-SIS+, we denote

$$\mathcal{S}_{k,l} = \left(Y_{t-1}, \ldots, Y_{t-h_n}, X_{t-1,k}, \ldots, X_{t-l+1,k}, \boldsymbol{z}_{t-1,\mathcal{U}_1^{\lambda_n}}, \ldots, \boldsymbol{z}_{t-1,\mathcal{U}_{l-1}^{\lambda_n}}\right),$$

as the population level counterpart to $\hat{\mathcal{S}}_{k,l}$. In addition, let the threshold $\Gamma_n = \lambda_n + c_1 n^{-\kappa}$, $C = \{\mathcal{S}_{1,1}, \ldots, \mathcal{S}_{m_n,1}, \mathcal{S}_{1,2}, \ldots, \mathcal{S}_{m_n,h_n}\}$, and

$$\mathcal{U}_{l-1}^{\Gamma_n} = \left\{(l-1)m_n + 1 \leq j \leq lm_n : |pdcor(Y_t, Z_{t-1,j}; C_j)| \geq \lambda_n + \tfrac{c_1}{2} n^{-\kappa}\right\},$$

represent the population level strong conditional signal set and the population level set of conditioning vectors, respectively. One of the difficulties in proving uniform convergence of our estimated partial distance correlations in this algorithm is the presence of an estimated conditioning set $\hat{C}$. This issue becomes compounded as we estimate the conditioning vector for higher lag levels, since these rely on estimates of the conditioning vectors for lower ones. To overcome this, we first denote the collection of strong signals from lag 1 to $h_n - 1$ as: $\mathcal{U}^{\Gamma_n} = \left\{\mathcal{U}_1^{\Gamma_n}, \ldots, \mathcal{U}_{h_n-1}^{\Gamma_n}\right\}$. We will assume the following condition:

Condition 2.7.5. For any $j \in \{1, \ldots, (h_n - 1) * m_n\} \setminus \mathcal{U}^{\lambda_n}$, assume $|pdcor(Y_t, Z_{t-1,j}; C_j)| \leq \lambda_n$, where $\lambda_n n^{\kappa} \to \infty$.

Condition 2.7.5 assumes the variables in the strong conditional signal set, $\mathcal{U}^{\Gamma_n}$, are easily identifiable from the rest of the covariates. This separation in the signal strength will allow us to ensure with high probability that our estimated conditional sets match their population level counterparts. The assumption $\lambda_n n^{\kappa} \to \infty$, is introduced to ensure $d_n = |\widetilde{\mathcal{M}}_{\gamma_n}| \gg |\mathcal{U}^{\lambda_n}|$. Although the hope is that $\mathcal{U}^{\lambda_n} \subset \mathcal{M}_*$, this is not required to prove sure screening properties of our algorithm. The sure screening properties for PDC-SIS+ are similar to PDC-SIS, but for the sake of completeness, we state the theorem in full.

Theorem 4. *1. Suppose conditions 2.7.1, 2.7.3, 2.7.4, and 2.7.5 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have*

$$P\left(\mathcal{M}_* \subset \widetilde{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(s_n a_n).$$

2. Suppose conditions 2.7.1, 2.7.2, 2.7.4, and 2.7.5 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have

$$\text{if } r < 12, \qquad P\left(\mathcal{M}_* \subset \widetilde{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(s_n c_n);$$

$$\text{if } r \geq 12, \qquad P\left(\mathcal{M}_* \subset \widetilde{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(s_n b_n).$$

2.8 Discussion

In this work, we have introduced two classes of partial distance correlation based screening procedures, which are applicable to univariate or multivariate time series models. These methods aim to utilize the unique features of time series data as an additional source of information, rather than treating temporal dependence as

a nuisance. The methods introduced can be easily utilized by researchers, given that distance correlation methods are easily computable at low computational cost by existing statistical packages. Lastly, by using a model free first stage procedure we are able to expand the choice of models which can be considered for a second stage procedure. This is especially helpful for the case of nonlinear or non-parametric models where estimation in high dimensions can be computationally challenging.

There are many opportunities for further research, such as developing a theoretical or data driven approach to selecting the number of lags considered in our algorithms. Additionally, we can develop screening algorithms for time series data using measures which are more robust to heavy tailed distributions. Lastly, our procedures were developed under the assumption that the underlying processes are weakly dependent and stationary. Although these assumptions are satisfied for a very wide range of applications, there are many instances where they are violated. For example, non-stationarity is commonly induced by time varying parameters, structural breaks, and cointegrated processes, all of which are common in the fields of macroeconomics and finance. Therefore, developing new methodologies for certain classes of non-stationary processes, such as locally stationary processes, would be particularly welcome.

2.9 Appendix A

The Appendix is organized as follows: Section 2.9.1 compares the empirical power of the Partial DC and Conditional DC measures, section 2.10 contains the sure screening properties, simulations, as well as a real data application of our group PDC-SIS procedure. Section 2.11 contains the proofs of theorems 1 and 2 found in our main paper. Lastly section 2.12 provide more detailed results of the simulations in section 5.1 of our main paper.

2.9.1 Comparing Partial DC vs Conditional DC

In this subsection we will compare the power of Partial DC vs Conditional DC in detecting conditional dependencies. We repeat examples 5-12 in Wang et al. (2015), as these were the examples in their work in which there existed a non-zero conditional distance correlation. For each example, we run 500 simulations, and for each simulation we test the significance of $\widehat{pdcor}(Y, \boldsymbol{X}|\boldsymbol{Z})$ at the .05 level using the R package **energy**. The empirical power for Conditional DC for each of the examples is obtained directly from the results in Wang et al. (2015). We report the empirical power of Partial DC and Conditional DC for each example in Table 2.7. We reprint the details of examples 5-12 for completeness , and to avoid confusion we use the notation used in the original work by Wang et al. (2015). For more details on Conditional DC, we refer readers to Wang et al. (2015).

Example 5:

X, Y, Z have a multivariate normal distribution with mean vector 0, and covariance matrix:

$$\Sigma = \begin{bmatrix} 1 & .7 & .6 \\ .7 & 1 & .6 \\ .6 & .6 & 1 \end{bmatrix}$$

Therefore, the Y, X are conditionally dependent given Z.

Example 6:

$X_1, Z \sim_{iid}$ Binomial$(10, .5)$, and define $X = X_1 + Z$, $Y = (X_1 - 5)^4 + Z$

Example 7:

$X_1, Y_1, Z, \epsilon \sim_{iid} N(0,1)$, and define:

$$Z_1 = .5(Z^3/7 + Z/2), Z_2 = (Z^3/2 + Z)/3, X_2 = Z_1 + tanh(X_1), X_3 = X_2 + X_2^3/3$$
$$Y_2 = Z_2 + Y_1, Y_3 = Y_2 + tanh(Y_2/3)$$

We then standardize X_3, Y_3 and define $X = X_3 = cosh(\epsilon)$, and $Y = Y_3 + cosh(\epsilon^2)$. Therefore, X and Y are not conditionally independent given Z.

Example 8:

$X_1, Z_1, Z_2 \sim_{iid}$ Binomial$(10, .5)$, and define $X = X_1 + Z_1 + Z_2$, $Y = (X_1 - 5)^4 + Z_1 + Z_2$, $\boldsymbol{Z} = (Z_1, Z_2)$.

Example 9:

Suppose $Z_1, \ldots, Z_6 \sim_{iid} t(1)$, the t distribution with 1 degree of freedom, and define:

$$X_i = Z_i, i = 1, 2, 3,\ X_4 = Z_4 + Z_5, Y = \sum_{i=1}^{6} Z_i$$

Therefore $\boldsymbol{X} = (X_1, \ldots, X_5)$, and Y are not conditionally independent given $Z = Z_5$

Example 10:

Suppose $Z_1, \ldots, Z_{13} \sim_{iid} t(1)$, and define:

$$X_i = Z_i, i = 1, \ldots, 9.\ X_{10} = Z_{10} + Z_{11}, Y_1 = Z_1Z_2 + Z_3Z_4 + Z_5Z_11 + Z_{12},$$
$$Y_2 = Z_6Z_7 + Z_8Z_9 + Z_{10}Z_{11} + Z_{13}$$

Therefore $\boldsymbol{X} = (X_1, \ldots, X_{10})$ and $\boldsymbol{Y} = (Y_1, Y_2)$ are not conditionally independent given $Z = Z_1 1$

Example 11:

Suppose $Z_1, \ldots, Z_4 \sim_{iid} t(2)$, and define:

$$X_i = Z_i, i = 1, \ldots, 4. \; Y_1 = \sin(Z_1) + \cos(Z_2) + Z_3^2 + Z_4^2,$$
$$Y_2 = Z_1^2 + Z_2^2 + Z_3 + Z_4$$

Therefore, $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ and $\boldsymbol{Y} = (Y_1, Y_2)$ are not conditionally independent given $\boldsymbol{Z} = (Z_1, Z_2)$.

Example 12:

Suppose $Z_1, \ldots, Z_4 \sim_{iid} t(2)$, and define:

$$X_i = Z_i, i = 1, \ldots, 4. \; Y_1 = Z_1 Z_2 + Z_3^3 + Z_4^2,$$
$$Y_2 = Z_1^3 + Z_2^2 + Z_3 Z_4$$

Therefore, $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ and $\boldsymbol{Y} = (Y_1, Y_2)$ are not conditionally independent given $\boldsymbol{Z} = (Z_1, Z_2)$.

We see from the results in Table 2.7, that partial DC is very effective at detecting the conditional relationship in all the examples given. Additionally, for Examples 5,9,10,11, and 12 partial DC has more power to detect the conditional relationship, whereas only for example 6 does conditional DC outperform partial DC.

Table 2.7: Partial DC (PDC) vs Conditional DC (CDC): Empirical Power

	$n = 50$		$n = 100$		$n = 150$		$n = 200$	
	PDC	CDC	PDC	CDC	PDC	CDC	PDC	CDC
Ex 5	1	.898	1	.993	1	1	1	1
Ex 6	.405	.752	.795	.995	.975	1	1	1
Ex 7	.99	.918	1	.998	1	1	1	1
Ex 8	.365	.361	.71	.731	.925	.949	.995	.977
Ex 9	1	.802	1	.955	1	.975	1	.983
Ex 10	1	.355	1	.789	1	.912	1	.935
Ex 11	.95	.768	.995	.973	1	.994	1	.995
Ex 12	.99	.812	1	.956	1	.976	1	.995

2.10 Appendix B: Group PDC-SIS

2.10.1 Sure Screening Properties for Group PDC-SIS

As in our main paper, we assume the multivariate response process has the representation:

$$\boldsymbol{x_i} = \boldsymbol{h}\left(\ldots, \boldsymbol{\eta}_{i-1}, \boldsymbol{\eta}_i\right). \tag{2.8}$$

Where $\boldsymbol{\eta}_i, i \in \mathbb{Z}$, are iid random vectors. To prove sure screening properties of our group PDC-SIS procedure, we need the following conditions:

Condition 2.10.1. Assume $|pdcor(G_{t,i}, G_{k,j}; G_{t-1,i})| \geq c_1 n^{-\kappa}$ for $(i,k,j) \in M_*$, $\kappa \in (0, 1/2)$.

Condition 2.10.2. Assume our multivariate response process has the representation (2.8). Additionally, we assume the following decay rate $\Phi_{m,r}(\boldsymbol{x}) = O(m^{-\alpha_x})$, for some $\alpha_x > 0$, $r > 4$.

Condition 2.10.3. Assume our multivariate response process $\boldsymbol{x_t}$ has the representation (2.8). Additionally assume $\upsilon_z = \sup_{q \geq 2} q^{-\tilde{\alpha}_x} \Phi_{0,q}(\boldsymbol{x}) < \infty$, for some $\tilde{\alpha}_x \geq 0$.

Condition 2.10.4. Assume the process $\{\boldsymbol{x}_t\}$ is β-mixing, with mixing rate $\beta_x(a) = O(\exp(-a^{\lambda_1}))$, for some $\lambda_1 > 0$.

Let $\varrho = 1$, if $\alpha_x > 1/2 - 2/r$, otherwise $\varrho = r/4 - \alpha_x r/2$. And let $K_{x,r} = \sup_{m\geq 0}(m+1)^{\alpha_x}\Phi_r(\boldsymbol{x})$. Recall that $t_n = \max_j dim(C_j)$ is the maximum dimension of the conditional vectors. Lastly let $\tilde{\varphi} = \frac{2}{1+4\tilde{\alpha}_x}$. The results are similar to those in theorem 1, but for the sake of completeness we present them here as well:

Corollary 5. *1. Suppose conditions 2.10.1, 2.10.3, 2.10.4 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have:*

$$P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O\left(s_n n^2 \exp\left(-\frac{n^{1/2-\kappa}}{t_n \upsilon_z^2}\right)^{\tilde{\varphi}}\right).$$

2. Suppose conditions 2.10.1, 2.10.2, 2.10.4 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have

$$\text{if } r < 12, \quad P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(s_n \frac{t_n^{r/2} K_{x,r}^r}{n^{r/4-r\kappa/2}});$$

$$\text{if } r \geq 12, \quad P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O\left(s_n n^2 \left[\frac{t_n^{r/2} n^{\varrho} K_{x,r}^r}{n^{r/2-r\kappa/2}} + \exp\left(-\frac{n^{1-2\kappa}}{t_n^2 K_{x,r}^4}\right)\right]\right).$$

From the above results we can infer the maximum size of the groups is $o(n^{1/2-\kappa})$. The proof for this corollary is very similar to the proof of theorem 1, therefore we omit the details.

2.10.2 Simulations for group PDC-SIS

We consider the following VAR(1) process,

Model 6:

$$\boldsymbol{x}_t = A_1 \boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t, \tag{2.9}$$

and assume we have 25 groups at each lag level ($e_n = 25$) with equal size $g_n = 20$. We assume a block upper triangular structure for A_1, with two scenarios.

$$A_1 = \begin{bmatrix} B & 0 & C & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & C \\ & & & \ddots & 0 \\ 0 & & & & B \end{bmatrix}. \tag{2.10}$$

We set the number of lags considered, $h = 2$, therefore we have to compute 1225 group distance and partial distance correlations for each scenario. In the first scenario we set the main diagonal blocks to $B = \{.3^{|i-j|+1}\}_{i,j \leq g_n}$, the second upper diagonal blocks to $C = \{.2^{|i-j|+1}\}_{i,j \leq g_n}$, and the rest of the matrix to zero. In the second scenario, we assume the same number of groups and group size, but we set the diagonal group $B = \{.3^{|i-j|+1}\}_{i,j \leq 10}$, and the second upper diagonal block to $C = \{.2^{|i-j|+1}\}_{i,j \leq 10}$. We can view this scenario as one in which we have misspecified the groups (Basu et al., 2015), or one in which we have sparsity within each group. We set $\Sigma_\eta = \{.4^{|i-j|+1}\}_{i,j \leq m_n}$ or $\Sigma_\eta = \{-.4^{|i-j|+1}\}_{i,j \leq m_n}$. And lastly, $\boldsymbol{\eta}_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\boldsymbol{\eta}_t \overset{iid}{\sim} t_3(0, 1/3 * \Sigma_\eta)$.

Since we are assuming the first lag for each group is in the model, we have 23 off-diagonal group connections we want to detect for each scenario. As in our main paper, the sample size is $n = 200$, and we report the median MMS for group DC-SIS, and group PDC-SIS procedure for each scenario in Table 2.8. The MMS in this case is defined as the minimum number of group connections which need to be selected for $\mathcal{M}_*$ to be captured. In order to ensure a fair comparison, we do not evaluate $dcor(G_{t,i}, G_{t-1,i})$ for each group i when using group DC-SIS. The results show that the procedures are robust to the level of sparsity within each group, and our group

Table 2.8: Model 6

	Scenario 1 PDC-SIS	Scenario 1 DC-SIS	Scenario 2 PDC-SIS	Scenario 2 DC-SIS
$N(0, \Sigma_\eta = \{.4^{\lvert i-j\rvert+1}\})$	33	53	32	52
$N(0, \Sigma_\eta = \{-.4^{\lvert i-j\rvert+1}\})$	68	139.5	66	140
t_3, $\Sigma_\eta = \{.4^{\lvert i-j\rvert+1}\})$	38	46.5	37	45
t_3, $\Sigma_\eta = \{-.4^{\lvert i-j\rvert+1}\})$	89	159.5	83.5	145.5

PDC-SIS procedure significantly outperforms the group DC-SIS for all scenarios.

2.10.3 Real data application: Group PDC-SIS

For the multivariate response setting, we focus on the group selection performance. We partition the 132 economic series into 8 broad economic groups: 1) Output and income (17 series) 2) Labor Market (32 series) 3) Housing (10 series) 4) Consumption, Orders, and Inventories (14 series) 5) Money and Credit (11 series) 6) Bonds and Exchange rates (22 series) 7) Prices (21 series) 8) Stock market (4 series). We then supplement this with 300 additional exogenous series ($\boldsymbol{v}_t$) partitioned into groups of size 10. Where $\boldsymbol{v}_t = A_1\boldsymbol{v}_{t-1} + \boldsymbol{\eta}_t$, $A_1 = \alpha * I$, where we vary α from .4 to .8, and we $\boldsymbol{\eta}_t \overset{iid}{\sim} N(0, I)$ or $\boldsymbol{\eta}_t \overset{iid}{\sim} t_3(1/3 * I)$. We have 38 groups for each lag level, and we set the number of lags considered, $h = 2$, giving us about 2900 group comparisons to compute. Let $\boldsymbol{x}_t$ represent our 132 economic series, and let $\boldsymbol{z}_t = (\boldsymbol{x}_t, \boldsymbol{v}_t)$ with $\boldsymbol{v}_t$ being independent of $\boldsymbol{x}_t$. We assume the following one step ahead forecasting strategy:

$$\boldsymbol{z}_t = \boldsymbol{f}(\boldsymbol{z}_{t-1}, \boldsymbol{z}_{t-2}) + \boldsymbol{\epsilon}_t. \tag{2.11}$$

We utilize a rolling window scheme similar to the one described previously, except we are not computing out of sample forecasts. For the first window we use data from t =1984:3 to t =1999:12 to compute our correlations. We then move the window for-

Table 2.9: Group Selection

	PDC-SIS	DC-SIS
Gaussian, $\alpha = .4$	37	34
Gaussian, $\alpha = .6$	32	25
Gaussian, $\alpha = .8$	22	9
t_3, $\alpha = .4$	36	31
t_3, $\alpha = .6$	31	21.5
t_3, $\alpha = .8$	23	8

ward by one month, which gives us 144 windows in total and 191 observations for each window. As discussed in section 4 of our main paper, for each group $G_{t,i}$ we condition on the first lag $G_{t-1,i}$ for PDC-SIS. Let $\{G_{t,j}\}_{j\leq 8}$ represents the 8 economic groups at time t, and let $\mathcal{B} = \{(i,k,j) : i,j \leq 8, k \in \{t-1, t-2\}\} \setminus \{(i,t-1,i) : i \leq 8\}$ denotes the set of possible group connections between the 8 economic groups minus the connection between a group and its first lag. For each window, we select the top $\lceil n/\log(n)\rceil = 37$ group connections, and record the number of group connections which belong to $\mathcal{B}$. We note that all group connections which are to be screened and do not belong to $\mathcal{B}$ are spurious connections by construction.

The results are in Table 2.9, and we report the median number of group connections which belong to $\mathcal{B}$ over the 144 windows. In order to ensure a fair comparison between group DC-SIS and group PDC-SIS, we do not evaluate $dcor(G_{t,i}, G_{t-1,i})$ for each group i when using group DC-SIS. We see that when $\alpha = .4$ and the noise is Gaussian, both group PDC-SIS and group DC-SIS are very effective at selecting connections between economic groups. When the dependence increases and heavy tailed variables are introduced, the performance of group DC-SIS greatly deteriorates with many spurious group connections selected, whereas group PDC-SIS remains effective.

2.11 Appendix C: Proofs of Theorems 3 and 4

Proof of Theorem 3.

We start with part (iii) first. The population version of the partial distance correlation is defined as:

$$pdcor(Y_t, Z_{t-1,k}; C_k) = \frac{dcor^2(Y_t, Z_{t-1,k}) - dcor^2(Y_t, C_k)dcor^2(Z_{t-1,k}, C_k)}{\sqrt{1 - dcor^4(Y_t, C_k)}\sqrt{1 - dcor^4(Z_{t-1,k}, C_k)}}. \quad (2.12)$$

To estimate this quantity, Székely and Rizzo (2014) proposed an unbiased estimator of the distance correlation to serve as the plug-in estimate. This estimate is different from the estimator proposed for the distance correlation in Székely et al. (2007), which is a biased but consistent estimate. In proving asymptotic properties we can use either estimate, and we will use the original estimator given in Székely et al. (2007).

To obtain a bound for $|\widehat{pdcor}(Y_t, Z_{t-1,k}; C_k) - pdcor(Y_t, Z_{t-1,k}; C_k)|$, we start with $|\widehat{dcor}^2(Y_t, Z_{t-1,k}) - dcor^2(Y_t, Z_{t-1,k})|$ in the numerator of (2.12). Recall that:

$$\widehat{dcor}^2(Y_t, Z_{t-1,k}) = \frac{\widehat{dcov}^2(Y_t, Z_{t-1,k})}{\widehat{dcov}(Y_t, Y_t)\widehat{dcov}(Z_{t-1,k}, Z_{t-1,k})}. \quad (2.13)$$

Let $\hat{T}_1 = \widehat{dcov}^2(Y_t, Z_{t-1,k})$,$\hat{T}_2 = \widehat{dcov}(Y_t, Y_t)\widehat{dcov}(Z_{t-1,k}, Z_{t-1,k})$,
and $T_1 = dcov^2(Y_t, Z_{t-1,k})$, $T_2 = dcov(Y_t, Y_t)dcov(Z_{t-1,k}, Z_{t-1,k})$, then

$$|\widehat{dcor}^2(Y_t, Z_{t-1,k}) - dcor^2(Y_t, Z_{t-1,k})| = |\frac{\hat{T}_1}{\hat{T}_2} - \frac{T_1}{T_2}|$$

$$= |(\hat{T}_2^{-1} - T_2^{-1})(\hat{T}_1 - T_1) + (\hat{T}_1 - T_1)/T_2 + (\hat{T}_2^{-1} - T_2^{-1})T_1|. \quad (2.14)$$

Therefore

$$P(|\frac{\hat{T}_1}{\hat{T}_2} - \frac{T_1}{T_2}| > cn^{-\kappa}) \leq P(|(\hat{T}_2^{-1} - T_2^{-1})(\hat{T}_1 - T_1)| > c_2 n^{-\kappa}/3) \tag{2.15}$$

$$+ P(|(\hat{T}_1 - T_1)/T_2| > c_2 n^{-\kappa}/3|) \tag{2.16}$$

$$+ P(|(\hat{T}_2^{-1} - T_2^{-1})T_1| > c_2 n^{-\kappa}/3). \tag{2.17}$$

For the RHS of (2.15), we obtain:

$$P(|(\hat{T}_2^{-1} - T_2^{-1})(\hat{T}_1 - T_1)| > c_2 n^{-\kappa}/3) \leq P(|\hat{T}_1 - E(T_1)| > Cn^{-\kappa/2}) + P(|\hat{T}_2^{-1} - E(T_2)^{-1}| > Cn^{-\kappa/2}).$$

So we focus on terms (2.16) and (2.17). For (2.16), recall that:

$$\hat{dcov}^2(Y_t, Z_{t-1,k}) = \hat{S}_{k1} + \hat{S}_{k2} - 2\hat{S}_{k3}, \tag{2.18}$$

where

$$\begin{aligned} \hat{S}_{k1} &= n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} |Y_i - Y_j||Z_{i,k} - Z_{j,k}|, \\ \hat{S}_{k2} &= n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} |Y_i - Y_j| n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} |Z_{i,k} - Z_{j,k}|, \\ \hat{S}_{k3} &= n^{-3} \sum_{j=1}^{n} \sum_{i=1}^{n} \sum_{l=1}^{n} |Y_i - Y_j||Z_{i,k} - Z_{l,k}|. \end{aligned} \tag{2.19}$$

We begin with the term $|\hat{S}_{k1} - S_{k1}|$, let

$$\hat{S}^*_{k1} = [n(n-1)]^{-1} \sum_{i \neq j} |Y_i - Y_j||Z_{i,k} - Z_{j,k}|,$$

then by equation (B.1) in Li et al. (2012b):

$$P(|\hat{S}_{k1} - S_{k1}| > Cn^{-\kappa}) \leq P(|\hat{S}^*_{k1} - S_{k1}| > Cn^{-\kappa}). \tag{2.20}$$

We also have the following decomposition:

$$|\hat{S}^*_{k1} - S_{k1}| \leq |\hat{S}^*_{k1} - E(\hat{S}^*_{k1})| + |E(\hat{S}^*_{k1}) - S_{k1}|. \tag{2.21}$$

Observe that $\hat{S}^*_{k1}$ is a U-statistic, and is a biased estimate of S_{k1} due to temporal dependence. By condition 3.4, we can control this bias, and we have $|E(\hat{S}^*_{k1} - S_{k1})| = O(n^{-\frac{1}{2}})$ by Yoshihara (1976). Obtaining a bound on $P(|\hat{S}^*_{k1} - S_{k1}| > Cn^{-\kappa})$ is difficult in a time series setting. Borisov and Volodko (2009) and Han (2016) introduced exponential inequalities for U-statistics in a time series setting under uniform mixing type conditions, in addition to restrictions on the kernel function. These restrictions are often too strict and rule out most commonly used time series. For example, even AR(1) processes where the innovations have unbounded support are not uniform mixing (see example 14.8 in Davidson (1994)).

As a result, we will instead rely on Nagaev and Rosenthal type inequalities (Wu and Wu, 2016; Liu et al., 2013) to obtain our bounds. We first show the bounds obtained by using Nagaev inequalities, and then we show the results obtained using Rosenthal type inequalities. Let $\psi_i = (e_i, \boldsymbol{\eta}_i)$ and $H_{i,j} = |Y_i - Y_j||Z_{i,k} - Z_{j,k}|$. We have

$$H_{i,j} = f(\ldots, \psi_0, \ldots, \psi_{\max(i,j)}) \text{ and } \hat{S}^*_{k1} = 2[n(n-1)]^{-1} \sum_{l=1}^{n-1} \sum_{i=1}^{n-l} H_{i,i+l}. \tag{2.22}$$

We can then write:

$$P(|\sum_{l=1}^{n-1}\sum_{i=1}^{n-l}(H_{i,i+l}-E(H_{i,i+l}))|>Cn^{2-\kappa}) \tag{2.23}$$

$$\leq \sum_{l=1}^{n-1} P(|\sum_{i=1}^{n-l}(H_{i,i+l}-E(H_{i,i+l}))|>Cn^{1-\kappa}). \tag{2.24}$$

Note that for any fixed l, $\{H_{i,i+l}\}_{i\in\mathcal{Z}}$ is a Bernoulli shift process, and we can compute the cumulative functional dependence measure as:

$$\begin{aligned}
&\sum_{i=m}^{\infty}|||Y_i-Y_{i+l}||Z_{i,k}-Z_{i+l,k}|-|Y_i^*-Y_{i+l}^*||Z_{i,k}^*-Z_{i+l,k}^*|||_\tau\\
\leq\ &\sum_{i=m}^{\infty}||Y_i-Y_{i+l}||_r|||Z_{i,k}-Z_{i+l,k}|-|Z_{i,k}^*-Z_{i+l,k}^*|||_q\\
&+\sum_{i=m}^{\infty}||Z_{i,k}^*-Z_{i+l,k}^*||_q|||Y_i-Y_{i+l}|-|Y_i^*-Y_{i+l}^*|||_r\\
\leq\ &\sum_{i=m}^{\infty}||Y_i-Y_{i+l}||_r|||Z_{i,k}-Z_{i,k}^*|+|Z_{i+l,k}-Z_{i+l,k}^*|||_q\\
&+\sum_{i=m}^{\infty}||Z_{i,k}^*-Z_{i+l,k}^*||_q|||Y_i-Y_i^*|+|Y_{i+l}-Y_{i+l}^*|||_r\\
\leq\ &2\Delta_{0,q}(\boldsymbol{y})\Phi_{m,r}(\boldsymbol{x})+2\Delta_{m,q}(\boldsymbol{y})\Phi_{0,r}(\boldsymbol{x})=O(m^{-\alpha}).
\end{aligned} \tag{2.25}$$

The last inequality holds since $||Z_{ik}||_r\leq\Phi_{0,r}(\boldsymbol{x})$, by section 2 in Wu and Wu (2016). Therefore,

$$\sup_m(m+1)^\alpha\sum_{i=m}^{\infty}|||Y_i-Y_{i+l}||Z_{i,k}-Z_{i+l,k}|-|Y_i^*-Y_{i+l}^*||Z_{i,k}^*-Z_{i+l,k}^*|||_\tau\leq 4K_{z,r}K_{y,q}. \tag{2.26}$$

Using the above result, and theorem 2 in Wu and Wu (2016), we obtain:

$$P(|\sum_{i=1}^{n-l}(H_{i,i+l}-E(H_{i,i+l}))| > Cn^{1-\kappa}) \leq O(\frac{n^{\iota}K_{z,r}^{\tau}K_{y,q}^{\tau}}{n^{\tau-\tau\kappa}})+O(\exp(-\frac{n^{1-2\kappa}}{K_{z,r}^{2}K_{y,q}^{2}})). \quad (2.27)$$

Using condition 3.4 along with (2.20),(2.21),(2.24), and (2.27), we obtain:

$$P(|\hat{S}_{k1} - S_{k1}| > Cn^{-\kappa}) \leq O(n\frac{n^{\iota}K_{z,r}^{\tau}K_{y,q}^{\tau}}{n^{\tau-\tau\kappa}}) + O(n\exp(-\frac{n^{1-2\kappa}}{K_{z,r}^{2}K_{y,q}^{2}})). \quad (2.28)$$

Next let $\hat{S}_{k2} = \hat{S}_{k2,1}\hat{S}_{k2,2}$, where $\hat{S}_{k2,1} = n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n}|Y_i - Y_j|$ and $\hat{S}_{k2,2} = n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n}|Z_i - Z_j|$. Using this representation we obtain:

$$\begin{aligned} P(|\hat{S}_{k2} - S_{k2}| > Cn^{-\kappa}) \leq\ & P(|(\hat{S}_{k2,1} - S_{k2,1})S_{k2,2}| > Cn^{-\kappa}) \\ & + P(|(\hat{S}_{k2,2} - S_{k2,2})S_{k2,1}| > Cn^{-\kappa}) \\ & + P(|(\hat{S}_{k2,1} - S_{k2,1})(\hat{S}_{k2,2} - S_{k2,2})| > Cn^{-\kappa}). \end{aligned} \quad (2.29)$$

Using the same methods as used for $\hat{S}_{k1}$, we obtain:

$$\begin{aligned} P(|\hat{S}_{k2} - S_{k2}| > Cn^{-\kappa}) \leq\ & O(n\frac{n^{\zeta}K_{z,r}^{r}}{n^{r-r\kappa}}) + O(n\exp(-\frac{n^{1-2\kappa}}{K_{z,r}^{2}})) \\ & + O(n\frac{n^{\varrho}K_{y,q}^{q}}{n^{q-q\kappa}}) + O(n\exp(-\frac{n^{1-2\kappa}}{K_{y,q}^{2}})). \end{aligned} \quad (2.30)$$

We now proceed to $\hat{S}_{k3}$. As in Li et al. (2012b), we define:

$$\begin{aligned}\hat{S}^*_{k3} = [n(n-1)(n-2)]^{-1} \sum_{i<j<l} [&|Z_{ik} - Z_{jk}||Y_j - Y_l| + |Z_{ik} - Z_{lk}||Y_j - Y_l| \\ &+ |Z_{ik} - Z_{jk}||Y_i - Y_l| + |Z_{lk} - Z_{jk}||Y_i - Y_l| \\ &+ |Z_{lk} - Z_{jk}||Y_i - Y_j| + |Z_{lk} - Z_{ik}||Y_i - Y_j|]. \end{aligned} \tag{2.31}$$

Note that $\hat{S}^*_{k3}$ is a U-statistic. Using condition 3.4 and Yoshihara (1976), we can control its bias: $|E(\hat{S}^*_{k3} - S_{k3})| = O(n^{-\frac{1}{2}})$. By equation (A.15) in Li et al. (2012b):

$$P(|\hat{S}_{k3} - S_{k3}| > Cn^{-\kappa}) \leq P(|\hat{S}^*_{k3} - S_{k3}| > Cn^{-\kappa}) \tag{2.32}$$

$$+ P(|\hat{S}^*_{k1} - S_{k1}| > Cn^{-\kappa}). \tag{2.33}$$

We have already dealt with (2.33), so we will proceed to (2.32). It suffices to deal with the first term in (2.31), since the rest can be bounded similarly. Let $H_{i,j,l} = |Z_{ik} - Z_{jk}||Y_j - Y_l| = f(\ldots, \psi_0, \ldots, \psi_{\max(i,j,l)})$. We can then represent

$$\sum_{i<j<l} |Z_{ik} - Z_{jk}||Y_j - Y_l| = \sum_{l=1}^{n-2} \sum_{j=1}^{n-l-1} \sum_{i=1}^{n-j-l} H_{i,i+j,i+j+l}. \tag{2.34}$$

Note that for fixed j, l, $\{H_{i,i+j,i+j+l}\}_{i \in \mathcal{Z}}$ is a Bernoulli shift process, whose cumulative functional dependence measure is the same as (2.25). We can then write:

$$\begin{aligned} &P(|\sum_{l=1}^{n-2} \sum_{j=1}^{n-l-1} \sum_{i=1}^{n-j-l} [H_{i,i+j,i+j+l} - E(H_{i,i+j,i+j+l})]| > Cn^{3-\kappa}) \\ \leq &\sum_{l=1}^{n-2} \sum_{j=1}^{n-l-1} P(|\sum_{i=1}^{n-j-l} [H_{i,i+j,i+j+l} - E(H_{i,i+j,i+j+l})]| > Cn^{1-\kappa}). \end{aligned} \tag{2.35}$$

Using condition 3.4, along with (2.28),(2.31),(2.32),(2.33),(2.35), and theorem 2 in

Wu and Wu (2016), we obtain:

$$P(|\hat{S}_{k3} - S_{k3}| > Cn^{-\kappa}) \leq O(n^2 \frac{n^{\iota} K_{z,r}^{\tau} K_{y,q}^{\tau}}{n^{\tau - \tau\kappa}}) + O(n^2 \exp(-\frac{n^{1-2\kappa}}{K_{z,r}^2 K_{y,q}^2})). \tag{2.36}$$

This gives us a bound for (2.16). For (2.17): $|\hat{T}_2^{-1} - T_2^{-1}| = |\frac{\hat{T}_2 - T_2}{T_2 \hat{T}_2}|$ and T_2 is finite by condition 3.4. Using this, we obtain:

$$\begin{aligned} P(|\hat{T}_2^{-1} - T_2^{-1}| > Cn^{-\kappa}) &\leq P(|\hat{T}_2 - T_2| > |\hat{T}_2| Cn^{-\kappa}) \\ &\leq P(|\hat{T}_2 - T_2| > CMn^{-\kappa}) + P(|\hat{T}_2| < M). \end{aligned} \tag{2.37}$$

We will deal with the first term in (2.37) and the second term can be handled similarly. Using the definition of $\hat{T}_2, T_2$ and the decomposition we used in (2.29), it suffices to analyze

$$P(|\widehat{dcov}(Y_t, Y_t) - dcov(Y_t, Y_t)| > Cn^{-\kappa}) \tag{2.38}$$

$$\text{and } P(|\widehat{dcov}(Z_{t-1,k}, Z_{t-1,k}) - dcov(Z_{t-1,k}, Z_{t-1,k})| > Cn^{-\kappa}). \tag{2.39}$$

For (2.38) and (2.39), note that for $a > 0, b > 0$ we have $|\sqrt{a} - \sqrt{b}| = \frac{|a-b|}{\sqrt{a}+\sqrt{b}} < \frac{|a-b|}{\sqrt{b}}$. Using this, along with (2.37) and the methods used to bound $\hat{T}_1$, we obtain:

$$\begin{aligned} P(|\hat{T}_2^{-1} - T_2^{-1}| > Cn^{-\kappa}) \leq\ & O(n^2 \frac{n^{\zeta} K_{z,r}^{r}}{n^{r/2 - r\kappa/2}}) + O(n^2 \exp(-\frac{n^{1-2\kappa}}{(K_{z,r})^2})) \\ & + O(n^2 \frac{n^{\varrho} K_{y,q}^{q}}{n^{q/2 - q\kappa/2}}) + O(n^2 \exp(-\frac{n^{1-2\kappa}}{(K_{y,q})^2})). \end{aligned} \tag{2.40}$$

By (2.28),(2.30),(2.36),(2.40), we obtain:

$$
\begin{aligned}
P(|\frac{\hat{T}_1}{\hat{T}_2} - \frac{T_1}{T_2}| > cn^{-\kappa}) \leq\ & O(n^2 \frac{n^{\zeta} K_{z,r}^{r}}{n^{r/2 - r\kappa/2}}) + O(n^2 \exp(-\frac{n^{1-2\kappa}}{K_{z,r}^2})) \\
& + O(n^2 \frac{n^{\varrho} K_{y,q}^{q}}{n^{q/2 - q\kappa/2}}) + O(n^2 \exp(-\frac{n^{1-2\kappa}}{K_{y,q}^2})) \\
& + O(n^2 \frac{n^{\iota} K_{z,r}^{\tau} K_{y,q}^{\tau}}{n^{\tau - \tau\kappa}}) + O(n^2 \exp(-\frac{n^{1-2\kappa}}{K_{z,r}^2 K_{y,q}^2})).
\end{aligned} \tag{2.41}
$$

The other terms in (2.12) deal with the conditioning vectors C_j, and we need to account for the maximum dimension of the conditioning vectors $\max_j[dim(C_j)] = t_n$. This comes into effect when computing the cumulative functional dependence measure. Recall that $C_{k+(h-1)*m_n} = \mathcal{S}_{k,h}$, and for analyzing the cumulative functional dependence measure, we define

$$
\mathcal{S}_{k,h}(i) = \{Y_{i-1}, \ldots, Y_{i-h}, X_{i-1,k}, \ldots, X_{i-h+1,k}\}, \tag{2.42}
$$

as the conditional vector of the h^{th} lag of series k at time i. Additionally recall that $|\boldsymbol{a}|_p$ stands for the Euclidean norm of $\boldsymbol{a} \in \mathcal{R}^p$. Assume $dim(S_{k,h}) = t_n$ and $q = r$, we

therefore have:

$$\sum_{i=m}^{\infty}\Big(|||\mathcal{S}_{k,h}(i)-\mathcal{S}_{k,h}(i+j)|_{t_n}|\mathcal{S}_{k,h}(i)-\mathcal{S}_{k,h}(i+j)|_{t_n} \tag{2.43}$$

$$- |\mathcal{S}^*_{k,h}(i)-\mathcal{S}^*_{k,h}(i+j)|_{t_n}|\mathcal{S}^*_{k,h}(i)-\mathcal{S}^*_{k,h}(i+j)|_{t_n}||_{q/2}\Big)$$

$$\leq \sum_{i=m}^{\infty} |||\mathcal{S}_{k,h}(i)-\mathcal{S}_{k,h}(i+j)|_{t_n}||_q|||\mathcal{S}_{k,h}(i)-\mathcal{S}_{k,h}(i+j)|_{t_n}-|\mathcal{S}^*_{k,h}(i)-\mathcal{S}^*_{k,h}(i+j)|_{t_n}||_q$$

$$+ \sum_{i=m}^{\infty} |||\mathcal{S}^*_{k,h}(i)-\mathcal{S}^*_{k,h}(i+j)|_{t_n}||_q|||\mathcal{S}_{k,h}(i)-\mathcal{S}_{k,h}(i+j)|_{t_n}-|\mathcal{S}^*_{k,h}(i)-\mathcal{S}^*_{k,h}(i+j)|_{t_n}||_q$$

$$\leq \sum_{i=m}^{\infty} |||\mathcal{S}_{k,h}(i)-\mathcal{S}_{k,h}(i+j)|_{t_n}||_q|||\mathcal{S}_{k,h}(i)-\mathcal{S}^*_{k,h}(i)|_{t_n}+|\mathcal{S}_{k,h}(i+j)-\mathcal{S}^*_{k,h}(i+j)|_{t_n}||_q$$

$$+ \sum_{i=m}^{\infty} |||\mathcal{S}^*_{k,h}(i)-\mathcal{S}^*_{k,h}(i+j)|_{t_n}||_q|||\mathcal{S}_{k,h}(i)-\mathcal{S}^*_{k,h}(i)|_{t_n}+|\mathcal{S}_{k,h}(i+j)-\mathcal{S}^*_{k,h}(i+j)|_{t_n}||_q$$

$$\leq t_n(\Delta_{0,q}(\boldsymbol{y})+\Phi_{m,q}(\boldsymbol{x}))^2.$$

To explain the last inequality, we analyze the term:

$$\sum_{i=m}^{\infty} |||\mathcal{S}_{k,h}(i)-\mathcal{S}^*_{k,h}(i)|_{t_n}||_q = \sum_{i=m}^{\infty} |||\mathcal{S}_{k,h}(i)-\mathcal{S}^*_{k,h}(i)|^2_{t_n}||^{1/2}_{q/2}$$

$$\leq (t_n/2)^{1/2}(\Delta_{0,q}(\boldsymbol{y})+\Phi_{0,q}(\boldsymbol{x})). \tag{2.44}$$

Where the last inequality follows from Minkowski's inequality and the definition of $S_{k,h}(i)$. Using this, the rest of the terms in (2.12) can be handled as done previously.

We now show the bounds obtained using a Rosenthal type inequality. We follow the same steps as previously, and it suffices to consider (2.32). As before we focus on

the following term

$$\sum_{i<j<l}[|Z_{ik}-Z_{jk}||Y_j-Y_l|]=\sum_{l=1}^{n-2}\sum_{j=1}^{n-l-1}\sum_{i=1}^{n-j-l}H_{i,i+j,i+j+l}. \tag{2.45}$$

Let $Q=[(n-1)(n-2)]^{-1}\sum_{l=1}^{n-2}\sum_{j=1}^{n-l-1}\sum_{i=1}^{n-j-l}H_{i,i+j,i+j+l}$. Then by Markov's inequality we obtain:

$$P(|Q-E(Q)|>cn^{1-\kappa})\leq\frac{||Q-E(Q)||_\tau^\tau}{n^{\tau-\tau\kappa}}. \tag{2.46}$$

Then using Minkowski's inequality, we obtain:

$$||Q-E(Q)||_\tau\leq||\sum_{i=1}^{n-2}H_{i,i+1,i+2}-E(H_{i,i+1,i+2})||_\tau. \tag{2.47}$$

As we stated previously, for fixed j,l, $\{H_{i,i+j,i+j+l}\}_{i\in\mathcal{Z}}$ is a Bernoulli shift process whose cumulative functional dependence measure is the same as (2.25). By theorem 1 in Liu et al. (2013), we have:

$$||\sum_{i=1}^{n-2}H_{i,i+1,i+2}-E(H_{i,i+1,i+2})||_\tau\leq O(K_{z,r}K_{y,q}n^{\frac{1}{2}}). \tag{2.48}$$

Combining the above with (2.47), we obtain:

$$P(|Q-E(Q)|>cn^{1-\kappa})\leq O(\frac{K_{z,r}^\tau K_{y,q}^\tau n^{\frac{\tau}{2}}}{n^{\tau-\tau\kappa}}). \tag{2.49}$$

By repeating the same techniques we obtain:

$$P(|\hat{\omega}_k-\omega_k|>c_2n^{-\kappa})\leq O(\frac{K_{y,q}^q n^{\frac{q}{4}}}{n^{q/2-q\kappa/2}})+O(\frac{K_{z,r}^\tau K_{y,q}^\tau n^{\frac{\tau}{2}}}{n^{\tau-\tau\kappa}})+O(\frac{K_{z,r}^r n^{\frac{r}{4}}}{n^{r/2-r\kappa/2}}). \tag{2.50}$$

For simplicity we assume $r = q$, and we now compare the above result to (2.41), which was obtained using Nagaev type inequalities. Note that when $q = r$ the above bound is of the order $O(n^{r/4-r\kappa/2})$. Using Nagaev type inequalities leads to the bound at most $O(n^{r/2-r\kappa/2-3})$. Therefore, when $r < 12$, (2.50) provides a better bound. When $r > 12$, the comparison depends on the values of ϱ, ι, ζ which are related to the dependence of the covariate and response processes. Applying the union bound gives us the desired result.

For part (iv), let $\mathcal{A}_n = \{\max_{k\in M_*} |\hat{\rho}_k - \rho_k| \leq \frac{c_1 n^{-\kappa}}{2}\}$. On the set $\mathcal{A}_n$, by condition 3.1, we have:

$$|\hat{\rho}_k| \geq |\rho_k| - |\hat{\rho}_k - \rho_k| \geq c_1 n^{-\kappa}/2, \ \forall k \in M_*. \tag{2.51}$$

Hence by our choice of γ_n, we obtain $P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) > P(\mathcal{A}_n)$. By applying part (i), the result follows.

For part(i), we first define the *predictive dependence measure* introduced by Wu (2005). The predictive dependence measure for a univariate process and multivariate processes is defined respectively as:

$$\begin{aligned} \theta_q(y_i) &= ||\mathrm{E}\left(y_i|\mathcal{F}_0\right) - \mathrm{E}\left(y_i|\mathcal{F}_{-1}\right)||_q, \\ \theta_q(Z_{ij}) &= ||\mathrm{E}\left(Z_{ij}|\mathcal{H}_0\right) - \mathrm{E}\left(Z_{ij}|\mathcal{H}_{-1}\right)||_q. \end{aligned} \tag{2.52}$$

With the cumulative predictive dependence measures defined as:

$$\Theta_{0,q}(\boldsymbol{x}) = \max_{j\leq p_n} \sum_{i=0}^{\infty} \delta_q(Z_{ij}), \text{ and } \Theta_{0,q}(\boldsymbol{\epsilon}) = \sum_{i=0}^{\infty} \delta_q(\epsilon_i). \tag{2.53}$$

We follow the steps of the proof of part (iii). For $|\widehat{dcor}^2(Y_t, Z_{t-1,k}) - dcor^2(Y_t, Z_{t-1,k})|$, it suffices to provide a bound for (2.35). Note that for fixed j, l, we have:

$$\begin{aligned}\sup_{q\geq 4} q^{-(\tilde{\alpha}_z+\tilde{\alpha}_y)} \sum_{i=1}^{\infty} \theta_q(H_{i,i+j,i+j+l}) &\leq \sup_{q\geq 4} q^{-(\tilde{\alpha}_z+\tilde{\alpha}_y)} \sum_{i=1}^{\infty} \delta_q(H_{i,i+j,i+j+l}) \\ &\leq \sup_{q\geq 4} q^{-(\tilde{\alpha}_z+\tilde{\alpha}_y)} \Delta_{0,q}(\boldsymbol{y})\Phi_{0,q}(\boldsymbol{x}) < \infty, \end{aligned} \tag{2.54}$$

where the first inequality follows from theorem 1 in Wu (2005), and the last inequality follows from condition 3.3. Using the above we have by theorem 3 in Wu and Wu (2016):

$$(2.35) \leq O\left(n^2 \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_z \upsilon_y}\right)^{\tilde{\psi}}\right). \tag{2.55}$$

We now provide a bound for (2.38) in a similar way. Let $S_{i,j,l} = |Y_i - Y_j||Y_j - Y_l| = f_1(\ldots, e_0, \ldots, e_{\max(i,j,l)})$. We then have:

$$\begin{aligned}\sup_{q\geq 4} q^{-2\tilde{\alpha}_y} \sum_{i=1}^{\infty} \theta_q(S_{i,i+j,i+j+l}) &\leq \sup_{q\geq 4} q^{-2\tilde{\alpha}_y} \sum_{i=1}^{\infty} \delta_q(S_{i,i+j,i+j+l}) \\ &\leq \sup_{q\geq 4} q^{-2\tilde{\alpha}_y} \Delta_{0,q}^2(\boldsymbol{y}) < \infty. \end{aligned} \tag{2.56}$$

Then by theorem 3 in Wu and Wu (2016):

$$\begin{aligned}\sum_{l=1}^{n-2} \sum_{j=1}^{n-l-1} P(|\sum_{i=1}^{n-j-l} (S_{i,i+j,i+j+l} - E(S_{i,i+j,i+j+l}))| > Cn^{1-\kappa}) \\ \leq O\left(n^2 \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_y^2}\right)^{\tilde{\alpha}}\right). \end{aligned} \tag{2.57}$$

A similar result holds for (2.39). Following the steps in the proof of part (iii), and

using the results above we obtain:

$$P(\max_{j\leq p_n}|\hat{\omega}_k-\omega_k|>c_2n^{-\kappa})\leq p_n\Bigg[O(n^2\exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_y^2}\right)^{\tilde{\alpha}})+O(n^2\exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_z\upsilon_y}\right)^{\tilde{\psi}})+O(n^2\exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_z^2}\right)^{\tilde{\varphi}})\Bigg].$$

The proof for part (ii) is similar to the proof for part (iv) and we omit its details. □

Proof of Theorem 4.

For simplicity we only prove part (i), and the proof for part (iii) follows similarly. Let $\tilde{\boldsymbol{\omega}}=(\tilde{\omega}_1,\ldots,\tilde{\omega}_{p_n})$, where $\tilde{\omega}_k=\widehat{pdcor}(Y_t,Z_{t-1,k};\hat{C}_k)$. We will work on the following set,

$$\mathcal{A}_n=\{\max_{k\leq p_n}|\tilde{\omega}_k-\omega_k|\leq\frac{c_1}{2}n^{-\kappa}\}.$$

The main difference in the proof for this procedure vs. PDC-SIS lies in the randomness which results from estimating the conditional sets at each lag level. We claim that on the set $\mathcal{A}_n$, $\hat{\boldsymbol{C}}=\boldsymbol{C}$. To see this, note that on the first lag level: $\max_{k\leq m_n}|\tilde{\omega}_k-\omega_k|\leq\frac{c_1}{2}n^{-\kappa}$, which implies $\hat{\mathcal{U}}_1^{\lambda_n}=\mathcal{U}_1^{\lambda_n}$. Now due to $\hat{\mathcal{U}}_1^{\lambda_n}=\mathcal{U}_1^{\lambda_n}$, we have $\hat{C}_j=C_j$ for $k\in m_n+1,\ldots,2m_n$, which implies $\tilde{\omega}_k=\hat{\omega}_k$ for $k\in m_n+1,\ldots,2m_n$. Continuing this argument we see that on the set $\mathcal{A}_n$ we have $\hat{\boldsymbol{C}}=\boldsymbol{C}$, and therefore $\tilde{\boldsymbol{\omega}}=\hat{\boldsymbol{\omega}}$. The result then follows from the results in theorem 1. □

2.12 Appendix D: Tables for Section 2.5

Tables 2.10-2.13 provide more detailed results of the simulations in section 5.1. As stated in our main paper, tables 2.10-2.13 report the median minimum model size needed to include all the relevant predictors, as well as the median rank of the significant covariates for each procedure.

Table 2.10: Model 1

Gaussian, $p_n = 1500$							
	MMS	$X_{t-1,1}$	$X_{t-1,2}$	$X_{t-1,3}$	$X_{t-1,4}$	$X_{t-1,5}$	$X_{t-1,6}$
PDC-SIS	7	6	3	2	2	3	5
DC-SIS	11	7	3.5	2	2	3	5.5
NIS	11	6	3	2	2	3	6
SIS	10	6	3	2	2	3	6
GLSS	6	5	3	2	2	3	5
Gaussian, $p_n = 4500$							
	MMS	$X_{t-1,1}$	$X_{t-1,2}$	$X_{t-1,3}$	$X_{t-1,4}$	$X_{t-1,5}$	$X_{t-1,6}$
PDC-SIS	11	5	3	3	3	3	5
DC-SIS	19	6	3	3	3	3	6
NIS	16	6	3	3	3	3	6
SIS	13	5	3	2.5	3	3	6
GLSS	6	5	3	2	2	3	5
t_5, $p_n = 1500$							
	MMS	$X_{t-1,1}$	$X_{t-1,2}$	$X_{t-1,3}$	$X_{t-1,4}$	$X_{t-1,5}$	$X_{t-1,6}$
PDC-SIS	13	5	3	3	3	3	5
DC-SIS	20	6	4	3	3	3	6
NIS	33	7	4	3	3	3	6
SIS	21.5	6	3	3	3	3	5
GLSS	6	5	3	2	2	3	5
t_5, $p_n = 4500$							
	MMS	$X_{t-1,1}$	$X_{t-1,2}$	$X_{t-1,3}$	$X_{t-1,4}$	$X_{t-1,5}$	$X_{t-1,6}$
PDC-SIS	36.5	7	4	2	2	3	5
DC-SIS	68	10.5	4	2	3	3	7
NIS	114	16.5	4	2	3	4	9
SIS	66.5	10.5	4	3	3	4	7
GLSS	7	5	3	2	2	3	5

Table 2.11: Model 2

Gaussian p_n=1500					
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$
PDC-SIS	61	1	40.5	2	5
DC-SIS	488	1	488	2	3
NIS	488	1	488	2	3
SIS	343.5	1	341.5	2	3
GLSS	179.5	1	160.5	2	6.5
Gaussian p_n=4500					
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$
PDC-SIS	149	1	141	2	4
DC-SIS	1051	1	1051	2	3
NIS	861	1	861	2	3
SIS	722	1	722	2	3
GLSS	592	1	412.5	2	8
t_5 p_n=1500					
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$
PDC-SIS	79.5	1	57.5	2	5
DC-SIS	408.5	1	408.5	2	3
NIS	513.5	1	492	2	4
SIS	447	1	440	2	4
GLSS	450.5	1	330.5	2	22
t_5 p_n=4500					
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$
PDC-SIS	275.5	1	239.5	2	5
DC-SIS	951.5	1	951.5	2	3
NIS	1100.5	1	984	2	4
SIS	905	1	859.5	2	3
GLSS	1386.5	1	995	2	18.5

Table 2.12: Model 3

Gaussian, $p_n = 1500$							
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$	$X_{t-1,3}$	$X_{t-1,4}$
PDC-SIS	29	2	4	4	3	7	11
DC-SIS	112	8	4.5	8	4	19	34.5
NIS	119.5	8	4	8	3	18	48.5
SIS	100.5	7	4	7	3	16	42
GLSS	813	14.5	164	535.5	13	2	18

Gaussian, $p_n = 4500$							
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$	$X_{t-1,3}$	$X_{t-1,4}$
PDC-SIS	78.5	3	4	3.5	2	10	20
DC-SIS	337	15	6.5	10	3	19	34.5
NIS	309	14	6	9	3	39	137
SIS	281	11.5	5	8	2	31	130
GLSS	2325.5	30.5	364	1709.5	36.5	2	73.5

t_5, $p_n = 1500$							
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$	$X_{t-1,3}$	$X_{t-1,4}$
PDC-SIS	43	3	3.5	4	3	6.5	16
DC-SIS	114	8	5	9	4	16	64.5
NIS	167	9	4	11	4	15	51
SIS	166.5	8	4	10	4	18.5	71
GLSS	969.5	42	202	453	60.5	3	44

t_5, $p_n = 4500$							
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$	$X_{t-1,3}$	$X_{t-1,4}$
PDC-SIS	78	2	5	4	3	11	24.5
DC-SIS	301.5	14.5	8	11.5	4	33	113.5
NIS	436.5	14.5	8	14	4	33.5	124
SIS	438	13	7	13	4	33	149.5
GLSS	3008	85.5	690	1362	99.5	9	117.5

Table 2.13: Model 4

Gaussian, $p_n = 1500$							
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$	$X_{t-1,3}$	$X_{t-2,3}$
PDC-SIS	42	5	5	3	2	20	10
DC-SIS	306.5	114.5	53	64	22.5	162.5	73
NIS	275	105.5	47	46	16	149	80
SIS	234.5	95	42	41	15	129.5	72.5
GLSS	800.5	1	12	5.5	10	552.5	103

Gaussian, $p_n = 4500$							
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$	$X_{t-1,3}$	$X_{t-2,3}$
PDC-SIS	100.5	8	6	4	2	33	16
DC-SIS	842.5	338	144	148	53	350	181
NIS	704	255.5	104.5	119	38	322	158
SIS	588	224	95.5	103.5	35	307	142
GLSS	2214	1	29	13.5	22	1490.5	291.5

t_5, $p_n = 1500$							
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$	$X_{t-1,3}$	$X_{t-2,3}$
PDC-SIS	51	4	5	5	4	19	9
DC-SIS	306	108.5	54.5	75	34.5	132	59
NIS	328	90.5	39	70	27	136	61
SIS	265	79.5	33	62.5	24.5	133	57
GLSS	891.5	3	48.5	47.5	43	476	162

t_5, $p_n = 4500$							
	MMS	$X_{t-1,1}$	$X_{t-2,1}$	$X_{t-1,2}$	$X_{t-2,2}$	$X_{t-1,3}$	$X_{t-2,3}$
PDC-SIS	104	8	8	4	3	33	18.5
DC-SIS	814.5	322	157	155.5	61.5	395	196
NIS	851.5	283	139.5	144	54.5	418.5	181
SIS	761	249	120	120.5	46	372	181
GLSS	2843.5	5	137	81	80	1760	554.5

Chapter 3

Variable Selection for Linear High Dimensional Time Series Models

This chapter is based on the article Yousuf (2018) with the title "Variable Screening for High Dimensional Time Series", authored by Kashif Yousuf. It is published in the Electronic Journal of Statistics.

3.1 Introduction

With the advancement of data acquisition technology, high dimensionality is a characteristic of data being collected in fields as diverse as health sciences, genomics, neuroscience, astronomy, finance, and macroeconomics. Applications where we have a large number of predictors for a relatively small number of observations are becoming increasingly common. For example, in disease classification we usually have thousands of variables, such as expression of genes, which are collected, while the sample size is usually in the tens. Other examples include fMRI data, where the number of voxels can number in the thousands and far outnumber the observations. For an overview of high dimensionality in economics and finance, see Fan et al. (2011b). For the biological sciences, see Fan and Ren (2006); Bickel et al. (2009a) and references

therein. The main goals in these situations according to Bickel (2008) are:

- To construct as effective a method as possible to predict future observations.
- To gain insight into the relationship between features and response for scientific purposes, as well as hopefully, to construct an improved prediction method.

More formally we are dealing with the case where

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

with $\boldsymbol{y} = (Y_1, \ldots, Y_n)^T$ being an n-vector of responses, $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})^T$ being an $n \times p_n$ random design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is a random vector of errors. In addition, when the dimensionality of the predictors (p_n) is large we usually make the assumption that the underlying coefficient vector ($\boldsymbol{\beta}$) is sparse. Sparsity is a characteristic that is frequently found in many scientific applications Fan and Lv (2008),Johnstone (2009). For example, in disease classification it is usually the case that only a small amount of genes are relevant to predicting the outcome.

Indeed, there are a wealth of theoretical results and methods that are devoted to this issue. Our primary focus is on screening procedures. Sure Independence Screening (SIS) as originally introduced in Fan and Lv (2008), was applicable to the linear model, and is based on a ranking of the absolute values of the marginal correlations of the predictors with the response. This method allows one to deal with situations in which the number of predictors is of an exponential order of the number of observations, which they termed as ultrahigh dimensionality. Further work on the topic has expanded the procedure to cover the case of generalized linear models Fan and Song (2010), non-parametric additive models Fan et al. (2011a), Cox proportional hazards model Fan et al. (2010), single index hazard rate models Gorst-

Rasmussen and Scheike (2013), and varying coefficient models Fan et al. (2014). Model-free screening methods have also been developed. For example; screening using distance correlation was analyzed in Li et al. (2012b), a martingale difference correlation approach was introduced in Shao and Zhang (2014), additional works include Zhu et al. (2011), Huang and Zhu (2016) among others. For an overview of works related to screening procedures, one can consult Liu et al. (2015). The main result introduced with these methods is that, under appropriate conditions, we can reduce the predictor dimension from size $p_n = O\left(\exp\left(n^{\alpha}\right)\right)$, for some $\alpha < 1$, to a size d_n, while retaining all the relevant predictors with probability approaching 1.

Another widely used class of methods is based on the penalized least squares approach. An overview of these methods is provided in Fan and Lv (2010) and Bickel et al. (2009a). Examples of methods in this class are the Lasso Tibshirani (1996), and the adaptive Lasso Zou (2006a). Various theoretical results have been discovered for these class of methods. They broadly fall into analyzing the prediction error $|X(\hat{\beta} - \beta)|_2^2$, parameter estimation error $|\hat{\beta} - \beta|_1$, model selection consistency, as well as limiting distributions of the estimated parameters (see Buhlmann and Van de Geer (2011) for a comprehensive summary). Using screening procedures in conjunction with penalized least squares methods, such as the adaptive Lasso, presents a powerful tool for variable selection. Variable screening can allow us to quickly reduce the parameter dimension p_n significantly, which weakens the assumptions needed for model selection consistency of the adaptive Lasso Huang et al. (2008); Medeiros and Mendes (2016).

A key limitation of the results obtained for screening methods, is the assumption of independent observations. In addition, it is usually assumed that the covariates and the errors are sub-Gaussian (or sub-exponential). However, there are many examples of real world data where these assumptions are violated. Data which is observed

over time and/or space such as meteorological data, longitudinal data, economic and financial time series frequently exhibit covariates and/or errors which are serially correlated. One specific example is the case of fMRI time series, where there can exist a complicated spatial-temporal dependence structure in the errors and the covariates (see Worsley et al. (2002)). Another example is in forecasting macroeconomic indicators such as GDP or inflation rate, where we have large number of macroeconomic and financial time series, along with their lags, as possible covariates. Examples of heavy tailed and dependent errors and covariates can be found most prominently in financial, insurance and macroeconomic data.

These examples stress why it is extremely important for variable selection methods to be capable of handling scenarios where the assumption of independent sub-Gaussian (or sub-exponential) observations is violated. Some works related to this goal for the Lasso include Wang et al. (2007), which extended the Lasso to jointly model the autoregressive structure of the errors as well as the covariates. However, their method is applicable only to the case where $p_n < n$, and they assume an autoregressive structure where the order of the process is known. Whereas Wu and Wu (2016) studied the theoretical properties of the Lasso assuming a fixed design in the case of heavy tailed and dependent errors. Additionally **?**, and Kock and Callot (2015) investigated theoretical properties of the Lasso for high-dimensional Gaussian processes. Most recently Medeiros and Mendes (2016) analyzed the adaptive Lasso for high dimensional time series while allowing for both heavy tailed covariate and errors processes, with the additional assumption that the error process is a martingale difference sequence.

Some works related to this goal for screening methods include Li et al. (2012a), which allows for heavy tailed errors and covariates. Additionally Chang et al. (2013), Wu et al. (2014), and Zhu et al. (2011) also relax the Gaussian assumption, with the

first two requiring the tails of the covariates and the response to be exponentially light, while the latter allows for heavy tailed errors provided the covariates are sub-exponential. Although these works relax the moment and distributional assumptions on the covariates and the response, they still remain in the framework of independent observations. A few works have dealt with correlated observations in the context of longitudinal data (see Cheng et al. (2014),Xu et al. (2014)). However, the dependence structure of longitudinal data is too restrictive to cover the type of dependence present in most time series. Most recently Chen et al. (2017) proposed a non-parametric kernel smoothing screening method applicable to time series data. In their work they assume a sub-exponential response, covariates that are bounded and have a density, as well as assuming the sequence $\{(Y_i, \boldsymbol{x}_i)\}$ is strong mixing, with the additional assumption that the strong mixing coefficients decay geometrically. These assumptions can be quite restrictive; they exclude, for example, heavy tailed time series, and discrete valued time series which are common in fields such as macroeconomics, finance, neuroscience, amongst others Davis et al. (2016a).

In this work, we study the theoretical properties of SIS for the linear model with dependent and/or heavy tailed covariates and errors. This allows us to substantially increase the number of situations in which SIS can be applied. However, one of the drawbacks to using SIS in a time series setting is that the temporal dependence structure between observations is ignored. In an attempt to correct this, we introduce a generalized least squares screening (GLSS) procedure, which utilizes this additional information when estimating the marginal effect of each covariate. By using GLS to estimate the marginal regression coefficient for each covariate, as opposed to OLS used in SIS, we correct for the effects of serial correlation. Our simulation results show the effectiveness of GLSS over SIS, is most pronounced when we have strong levels of serial correlation and weak signals. Using the adaptive Lasso as a second stage

estimator after applying the above screening procedures is also analyzed. Probability bounds for our combined two stage estimator being sign consistent are provided, along with comparisons between our two stage estimator and the adaptive Lasso as a stand alone procedure.

Compared to previous work, we place no restrictions on the distribution of the covariate and error processes besides existence of a certain number of finite moments. In order to quantify dependence, we rely on the functional dependence measure framework introduced by Wu (2005), rather than the usual strong mixing coefficients. Comparisons between functional dependence measures and strong mixing assumptions are discussed in section 3.2. For both GLSS and SIS, we present the sure screening properties and show the range of p_n can vary from the high dimensional case, where p_n is a power of n, to the ultrahigh dimensional case discussed in Fan and Lv (2008). We detail how the range of p_n and the sure screening properties are affected by the strength of dependence and the moment conditions of the errors and covariates, the strength of the underlying signal, and the sparsity level, amongst other factors.

The rest of the paper is organized as follows: Section 3.2 reviews the functional and predictive dependence measures which will allow us to characterize the dependence in the covariate $(\boldsymbol{x_i}, i = 1, ..., n)$ and error processes. We also discuss the assumptions placed on structure of the covariate and error processes; these assumptions are very mild, allowing us to represent a wide variety of stochastic processes which arise in practice. Section 3.3 presents the sure screening properties of SIS under a range of settings. Section 3.4 introduces the GLSS procedure and presents its sure screening properties. Combining these screening procedures with the adaptive Lasso will discussed in Section 3.5. Section 3.6 covers simulation results, while section 3.7 discusses an application to forecasting the US inflation rate. Lastly, concluding remarks are in Section 3.8, and the proofs for all the results follow in the appendix.

3.2 Preliminaries

We shall assume the error sequence is a strictly stationary, ergodic process with the following form:

$$\epsilon_i = g\left(\ldots, e_{i-1}, e_i\right) \tag{3.2}$$

Where $g(\cdot)$ is a real valued measurable function, and e_i are iid random variables. This representation includes a very wide range of stochastic processes such as linear processes, their non-linear transforms, Volterra processes, Markov chain models, non-linear autoregressive models such as threshold auto-regressive (TAR), bilinear, GARCH models, among others (for more details see Wu (2011),Wu (2005)). This representation allows us to use the functional and predictive dependence measures introduced in Wu (2005). The functional dependence measure for the error process is defined as the following:

$$\delta_q(\epsilon_i) = ||\epsilon_i - g\left(\mathcal{F}_i^*\right)||_q = \left(E|\epsilon_i - g\left(\mathcal{F}_i^*\right)|^q\right)^{1/q} \tag{3.3}$$

where $\mathcal{F}_i^* = (\ldots, e_{-1}, e_0^*, e_1, \ldots, e_i)$ with $e_0^*, e_j, j \in \mathbb{Z}$ being iid. Since we are replacing e_0 by e_0^*, we can think of this as measuring the dependency of ϵ_i on e_0 as we are keeping all other inputs the same. The cumulative functional dependence measure is defined as $\Delta_{m,q}(\boldsymbol{\epsilon}) = \sum_{i=m}^{\infty} \delta_q(\epsilon_i)$. We assume weak dependence of the form:

$$\Delta_{0,q}(\boldsymbol{\epsilon}) = \sum_{i=0}^{\infty} \delta_q(\epsilon_i) < \infty \tag{3.4}$$

The predictive dependence measure is related to the functional dependence measure,

and is defined as the following:

$$\theta_q(\epsilon_l) = ||\mathrm{E}\left(\epsilon_l|\mathcal{F}_0\right) - \mathrm{E}\left(\epsilon_l|\mathcal{F}_{-1}\right)||_q = ||\mathcal{P}_0\epsilon_l||_q \tag{3.5}$$

where $\mathcal{F}_i = (\ldots, e_{-1}, e_0, e_1, \ldots, e_i)$ with $e_i, i \in \mathbb{Z}$ being iid. The cumulative predictive dependence measure is defined as $\Theta_q(\boldsymbol{\epsilon}) = \sum_{l=0}^{\infty}\theta_q(\epsilon_l)$, and by Theorem 1 in Wu (2005) we obtain $\Theta_q(\boldsymbol{\epsilon}) \leq \Delta_{0,q}(\boldsymbol{\epsilon})$.

Similarly the covariate process is of the form:

$$\boldsymbol{x}_{\boldsymbol{i}}^{(n)} = \boldsymbol{h}\left(\ldots, \boldsymbol{\eta}_{i-1}^{(n)}, \boldsymbol{\eta}_i^{(n)}\right) \tag{3.6}$$

Where $\boldsymbol{\eta}_i^{(n)} \in \mathcal{R}^{p_n}, i \in \mathbb{Z}$, are iid random vectors, $\boldsymbol{h(\cdot)} = (h_1(\cdot)\ldots, h_{p_n}(\cdot))$, $\boldsymbol{x}_{\boldsymbol{i}}^{(n)} = (X_{i1}, ..., X_{ip_n})$ and $X_{ij} = h_j(..., \boldsymbol{\eta}_{i-1}^{(n)}, \boldsymbol{\eta}_i^{(n)})$. The superscript (n) denotes that the dimension of vectors is a function of n, however for presentational clarity we suppress the superscript (n) from here on and use $\boldsymbol{x}_i$ and $\boldsymbol{\eta}_i$ instead. Let $\mathcal{H}_i^* = (\ldots, \boldsymbol{\eta}_{-1}, \boldsymbol{\eta}_0^*, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_i)$. As before the functional dependence measure is $\delta_q(X_{ij}) = ||X_{ij} - h_j\left(\mathcal{H}_i^*\right)||_q$ and the cumulative dependence measure for the covariate process is defined as:

$$\Phi_{m,q}(\boldsymbol{x}) = \sum_{i=m}^{\infty}\max_{j\leq p_n}\delta_q(X_{ij}) < \infty \tag{3.7}$$

The representations (3.2), and (3.6), along with the functional and predictive dependence measures have been used in various works including Wu and Pourahmadi (2009),Xiao and Wu (2012), and Wu and Wu (2016) amongst others. Compared to strong mixing conditions, which are often difficult to verify, the above dependence measures are easier to interpret and compute since they are related to the data generating mechanism of the underlying process Wu (2011). In many cases using the functional dependence measure also requires less stringent assumptions. For exam-

ple, consider the case of a linear process, $\epsilon_i = \sum_{j=0}^{\infty} f_j e_{i-j}$, with e_i iid. Sufficient conditions for a linear process to be strong mixing involve: the density function of the innovations (e_i) being of bounded variation, restrictive assumptions on the decay rate of the coefficients (f_j), and invertibility of the process (see Theorem 14.9 in Davidson (1994) for details). Additional conditions are needed to ensure strong mixing if the innovations for the linear process are dependent Doukhan (1994).

As a result many simple processes can be shown to be non-strong mixing. A prominent example involves an AR(1) model with iid Bernoulli (1/2) innovations: $\epsilon_i = \rho\epsilon_{i-1} + e_i$ is non-strong mixing if $\rho \in (0, 1/2]$ Andrews (1984). These cases can be handled quite easily in our framework, since we are not placing distributional assumptions on the innovations, e_i, such as the existence of a density. For linear processes with iid innovations, representation (3.2) clearly holds and (3.4) is satisfied if $\sum_{j=0}^{\infty} |f_j| < \infty$. For dependent innovations, suppose we have: $e_i = h(\ldots, a_{i-1}, a_i)$, where $h(\cdot)$ is a real valued measurable function and $a_i, i \in \mathbb{Z}$, are iid. Then $\epsilon_i = \sum_{j=0}^{\infty} f_j e_{i-j}$, has a causal representation, and satisfies (3.4) if: $\sum_{i=0}^{\infty} \delta_q(e_i) < \infty$, and $\sum_{j=0}^{\infty} |f_j| < \infty$ (see Wu and Min (2005)).

3.3 SIS with Dependent Observations

Sure Independence Screening, as introduced by Fan and Lv Fan and Lv (2008), is a method of variable screening based on ranking the magnitudes of the p_n marginal regression estimates. Under appropriate conditions, this simple procedure is shown to possess the sure screening property. The method is as follows, let:

$$\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_{p_n}), \text{ where } \hat{\rho}_j = (\sum_{t=1}^{n} X_{tj}^2)^{-1} (\sum_{t=1}^{n} X_{tj} Y_t) \tag{3.8}$$

Therefore, $\hat{\rho}_j$ is the OLS estimate of the linear projection of Y_t onto X_{tj}. Now let

$$\mathcal{M}_* = \{1 \leq i \leq p_n : \beta_i \neq 0\} \tag{3.9}$$

and let $|\mathcal{M}_*| = s_n << n$ be the size of the true sparse model. We then sort the elements of $\hat{\boldsymbol{\rho}}$ by their magnitudes. For any given γ_n, define a sub-model

$$\hat{\mathcal{M}}_{\gamma_n} = \{1 \leq i \leq p_n : |\hat{\rho}_i| \geq \gamma_n\} \tag{3.10}$$

and let $|\hat{\mathcal{M}}_{\gamma_n}| = d_n$ be the size of the selected model. The sure screening property states that for an appropriate choice of γ_n, we have $P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \to 1$.

Throughout this paper let: $Y_t = \sum_{i=1}^{p_n} X_{ti}\beta_i + \epsilon_t$, $\boldsymbol{x_t} = (X_{t1}, ..., X_{tp_n})$, $\Sigma = cov(\boldsymbol{x_t})$, and $\boldsymbol{X}_k$ be k^{th} column of $\boldsymbol{X}$. In addition, we assume $Var(Y_t), Var(X_{tj}) = O(1)$, $\forall j \leq p_n$. Note that $\boldsymbol{x}_t$ can contain lagged values of Y_t. Additionally, let $\rho_j = (E(X_{tj}^2))^{-1}E(X_{tj}Y_t)$, and $\mathcal{M}_{\gamma_n} = \{1 \leq i \leq p : |\rho_i| \geq \gamma_n\}$. For a vector $\boldsymbol{a} = (a_1, ..., a_n)$, $sgn(\boldsymbol{a})$ denotes its sign vector, with the convention that $sgn(0) = 0$, and $|\boldsymbol{a}|_p^p = \sum_{i=1}^n |a_i|^p$. For a square matrix $\boldsymbol{A}$, let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$, denote the minimum eigenvalue, and maximum eigenvalue respectively. For any matrix $\boldsymbol{A}$, let $||\boldsymbol{A}||_\infty$, and $||\boldsymbol{A}||_2$ denote the maximum absolute row sum of $\boldsymbol{A}$, and the spectral norm of $\boldsymbol{A}$ respectively. Lastly we will use C, c to denote generic positive constants which can change between instances.

3.3.1 SIS with dependent, heavy tailed covariates and errors

To establish sure screening properties, we need the following conditions:

Condition A: $|\rho_k| \geq c_1 n^{-\kappa}$ for $k \in M_*$, $\kappa < 1/2$

Condition B: $E(\epsilon_t), E(X_{tj}), E(X_{tj}\epsilon_t) = 0 \; \forall j, t.$

Condition C: Assume the error and the covariate processes have representations (3.2), and (3.6) respectively. Additionally, we assume the following decay rates $\Phi_{m,r}(\boldsymbol{x}) = O(m^{-\alpha_x}), \Delta_{m,q}(\boldsymbol{\epsilon}) = O(m^{-\alpha_\epsilon})$, for some $\alpha_x, \alpha_\epsilon > 0$, $q > 2, r > 4$ and $\tau = \frac{qr}{q+r} > 2$.

Condition **A** is standard in screening procedures, and it assumes the marginal signals of the active predictors cannot be too small. Condition **B** assumes the covariates and the errors are contemporaneously uncorrelated. This is significantly weaker than independence between the error sequence and the covariates usually assumed. Condition **C** presents the structure, dependence and moment conditions on the covariate and error processes. Notice that higher values of $\alpha_x, \alpha_\epsilon$ indicate weaker temporal dependence.

Examples of error and covariate processes which satisfy Condition C are: If ϵ_i is a linear process, $\epsilon_i = \sum_{j=0}^{\infty} f_j e_{i-j}$ with e_i iid and $\sum_{j=0}^{\infty} |f_j| < \infty$ then $\delta_q(\epsilon_i) = |f_i|||e_0 - e_0^*||_q$. If $f_i = O(i^{-\beta})$ for $\beta > 1$ we have $\Delta_{m,q} = O(m^{-\beta+1})$ and $\alpha_\epsilon = \beta - 1$. We have a geometric decay rate in the cumulative functional dependence measure, if ϵ_i satisfies the geometric moment contraction (GMC) condition, see Shao and Wu (2007). Conditions needed for a process to satisfy the GMC condition are given in Theorem 5.1 of Shao and Wu (2007). Examples of processes satisfying the GMC condition include stationary, causal finite order ARMA, GARCH, ARMA-GARCH, bilinear, and threshold autoregressive processes, amongst others (see Wu (2011) for details).

For the covariate process, if we assume $\boldsymbol{x}_i$ is a vector linear process: $\boldsymbol{x}_i = \sum_{l=0}^{\infty} A_l \boldsymbol{\eta}_{i-l}$. Where A_l are $p_n \times p_n$ coefficient matrices and $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ip_n})$ are iid random vectors with $cov(\boldsymbol{\eta_i}) = \Sigma_\eta$. For simplicity, assume $\eta_{i,j} (j = 1, \ldots, p_n)$ are identically distributed, then

$$\delta_q(X_{ij}) = ||A_{i,j}\boldsymbol{\eta}_0 - A_{i,j}\boldsymbol{\eta}_0^*||_q \leq 2|A_{i,j}|||\eta_{0,1}||_q \tag{3.11}$$

where $A_{i,j}$ is the j^{th} column of A_i. If $||A_i||_\infty = O(i^{-\beta})$ for $\beta > 1$, then $\Phi_{m,q} = O(m^{-\beta+1})$.

In particular for stable VAR(1) processes,
$\boldsymbol{x}_t = B_1\boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t$, $\Phi_{m,q}(\boldsymbol{x}) = O(||\lambda_{\max}(B)_1||^m)$ Chen et al. (2013). For stable VAR(k) processes, $\boldsymbol{x}_t = \sum_{i=1}^{k} B_i\boldsymbol{x}_{t-i} + \boldsymbol{\eta}_t$, we can rewrite this as a VAR(1) process, $\tilde{\boldsymbol{x}}_t = \tilde{B}_1\tilde{\boldsymbol{x}}_{t-1} + \tilde{\boldsymbol{\eta}}_t$, with:

$$\tilde{\boldsymbol{x}}_t = \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \\ \vdots \\ \boldsymbol{x}_{t-k+1} \end{bmatrix}_{kp\times 1} \quad \tilde{B}_1 = \begin{pmatrix} B_1 & \cdots & B_{k-1} & B_k \\ I_{p_n} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & I_{p_n} & \mathbf{0} \end{pmatrix}_{kp\times kp} \quad \tilde{\boldsymbol{\eta}}_t = \begin{bmatrix} \boldsymbol{\eta}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}_{kp\times 1} \tag{3.12}$$

And by section 11.3.2 in Lütkepohl (2005), the process $\tilde{\boldsymbol{x}}_t$ is stable if and only if $\boldsymbol{x}_t$ is stable. Therefore if $\tilde{B}_1$ is diagonalizable, we have $\Phi_{m,q}(\boldsymbol{x}) = O(a^m)$, where a represents the largest eigenvalue in magnitude of $\tilde{B}_1$. And by the stability of $\boldsymbol{x}_t$, $a \in (0, 1)$. Additional examples of error and covariate processes which satisfy Condition C are given in Wu and Pourahmadi (2009) and Wu and Wu (2016) respectively.

Define $\alpha = min(\alpha_x, \alpha_\epsilon)$ and let $\omega = 1$ if $\alpha_x > 1/2 - 2/r$, otherwise $\omega = r/4 - \alpha_x r/2$. Let $\iota = 1$ if $\alpha > 1/2 - 1/\tau$, otherwise $\iota = \tau/2 - \tau\alpha$. Additionally, let $K_{\epsilon,q} =$

$\sup_{m\geq 0}(m+1)^{\alpha_\epsilon}\Delta_{m,q}(\boldsymbol{\epsilon})$ and $K_{x,r}=\max_{j\leq p_n}\sup_{m\geq 0}(m+1)^{\alpha_x}\sum_{i=m}^{\infty}\delta_r(X_{ij})$. Given Condition C, it follows that $K_{\epsilon,q},K_{x,r}<\infty$. For ease of presentation we let:

$$\vartheta_n=\frac{s_n n^{\omega}K_{x,r}^r}{(n/s_n)^{r/2-r\kappa/2}}+\frac{n^{\iota}K_{x,r}^{\tau}K_{\epsilon,q}^{\tau}}{n^{\tau-\tau\kappa}}+\exp\left(-\frac{n^{1-2\kappa}}{s_n^2K_{x,r}^4}\right)+\exp\left(-\frac{n^{1-2\kappa}}{K_{x,r}^2K_{\epsilon,q}^2}\right) \tag{3.13}$$

The following theorem gives the sure screening properties, and provides a bound on the size of the selected model:

Theorem 6. *Suppose Conditions A,B,C hold.*

(i) For any $c_2>0$, we have:

$$P\left(\max_{j\leq p_n}|\hat{\rho}_j-\rho_j|>c_2n^{-\kappa}\right)\leq O(p_n\vartheta_n)$$

(ii) For $\gamma_n=c_3n^{-\kappa}$ with $c_3\leq c_1/2$, we have:

$$P\left(\mathcal{M}_*\subset\hat{\mathcal{M}}_{\gamma_n}\right)\geq 1-O(s_n\vartheta_n)$$

(iii) For $\gamma_n=c_3n^{-\kappa}$ with $c_3\leq c_1/2$, we have:

$$P\left(|\hat{\mathcal{M}}_{\gamma_n}|\leq O(n^{2\kappa}\lambda_{max}(\Sigma))\right)\geq 1-O(p_n\vartheta_n)$$

In Theorem 6 we have two types of bounds, for large n the polynomial terms dominate, whereas for small values of n the exponential terms dominate. The covariate dimension (p_n) can be as large as $o(min(\frac{s_n(n/s_n)^{r/2-r\kappa/2}}{n^{\omega}},\frac{n^{\tau-\tau\kappa}}{n^{\iota}}))$. The range of p_n depends on the dependence in both the covariate and the error processes, the strength of the signal (κ), the moment condition, and the sparsity level (s_n). If we assume $s_n=O(1), r=q$, and $\alpha\geq 1/2-2/r$ then $p_n=o(n^{r/2-r\kappa/2-1})$. For the case of iid

errors and covariates, we would replace $K_{x,r}, K_{\epsilon,q}$ in Theorem 6 with $\max_{j \leq p_n} ||X_{ij}||_{r/2}$ and $||\epsilon_i||_q$ respectively. Therefore for the case of weaker dependence in the covariate and error processes (i.e. $\alpha_x > 1/2 - 2/r$ and $\alpha > 1/2 - 1/\epsilon$), our range for p_n is reduced only by a constant factor. However, our range for p_n is significantly reduced in the case of stronger dependence in the error or covariate processes (i.e. either $\alpha_x < 1/2 - 2/r$ or $\alpha_\epsilon < 1/2 - 2/q$). For instance if $\alpha_x = \alpha_\epsilon$ and $q = r$, our range for p_n is reduced by a factor of $n^{r/4 - \alpha r/2}$ in the case of stronger dependence.

In the iid setting, to achieve sure screening in the ultrahigh dimensional case, Fan and Lv (2008) assumed the covariates and errors are jointly normally distributed. Future works applicable to the linear model, such as Fan and Song (2010),Fan et al. (2011a) among others, relaxed this Gaussian assumption, but generally assumed the tails of the covariates and errors are exponentially light. Compared to the existing results for iid observations, our moment conditions preclude us from dealing with the ultrahigh dimensional case. However, our setting is far more general in that it allows for dependent and heavy tailed covariates and errors. In addition, we allow for the covariates and error processes to be dependent on each other, with the mild restriction that $E(X_{tj}\epsilon_t) = 0$, $\forall j \leq p_n$.

3.3.2 Ultrahigh Dimensionality under dependence

It is possible to achieve the sure screening property in the ultrahigh dimensional setting with dependent errors and covariates. However, we need to make stronger assumptions on the moments of both the error and covariate processes. Until now we have assumed the existence of a finite qth moment, which restricted the range of p to a power of n. If the error and covariate processes are assumed to follow a stronger moment condition, such as $\Delta_{0,q}(\boldsymbol{\epsilon}) < \infty$ and $\Phi_{0,q}(\boldsymbol{x}) < \infty$ for arbitrary $q > 0$, we

can achieve a much larger range of p_n which will cover the ultrahigh dimensional case discussed in Fan and Lv (2008). More formally, we have:

Condition D: Assume the error and the covariate processes have representations (3.2), and (3.6) respectively. Additionally assume $\upsilon_x = \sup_{q\geq 2} q^{-\tilde{\alpha}_x}\Phi_{0,q}(\boldsymbol{x}) < \infty$ and $\upsilon_\epsilon = \sup_{q\geq 2} q^{-\tilde{\alpha}_\epsilon}\Delta_{0,q}(\boldsymbol{\epsilon}) < \infty$, for some $\tilde{\alpha}_x, \tilde{\alpha}_\epsilon \geq 0$.

By Theorem 3 in Wu and Wu (2016), Condition D implies the tails of the covariate and error processes are exponentially light. There are a wide range of processes which satisfy the above condition. For example, if ϵ_i is a linear process: $\epsilon_i = \sum_{j=0}^{\infty} f_j e_{i-j}$ with e_i iid and $\sum_{l=0}^{\infty} |f_l| < \infty$ then $\Delta_{0,q}(\epsilon_l) = ||e_0 - e_0^*||_q \sum_{l=0}^{\infty} |f_l|$. If we assume e_0 is sub-Gaussian, then $\tilde{\alpha}_\epsilon = 1/2$, since $||e_0||_q = O(\sqrt{q})$. Similarly if e_i is sub-exponential we have $\tilde{\alpha}_\epsilon = 1$. More generally, for $\epsilon_i = \sum_{j=0}^{\infty} f_j e_{i-j}^p$, if e_i is sub-exponential, we have $\tilde{\alpha}_\epsilon = p$. Similar results hold for vector linear processes discussed previously.

Condition D is primarily a restriction on the rate at which $||\epsilon_i||_q, \max_{j\leq p_n} ||X_{ij}||_q$ increase as $q \to \infty$. We remark that, for any fixed q, we are not placing additional assumptions on the temporal decay rate of the covariate and error processes besides requiring $\Delta_{0,q}(\boldsymbol{\epsilon})$,$\Phi_{0,q}(\boldsymbol{x}) < \infty$. In comparison, in the ultrahigh dimensional setting, Chen et al. (2017) requires geometrically decaying strong mixing coefficients, in addition to requiring sub-exponential tails for the response. As an example, if we assume $\epsilon_i = \sum_{j=0}^{\infty} f_j e_{i-j}$, geometrically decaying strong mixing coefficients would require the coefficients, f_j, to decay geometrically. Whereas in Condition D, the only restrictions we place on the coefficients, f_j, is absolute summability.

Theorem 7. *Suppose Conditions A,B,D hold. Define* $\tilde{\alpha}' = \frac{2}{1+2\tilde{\alpha}_x+2\tilde{\alpha}_\epsilon}$, *and* $\tilde{\alpha} = \frac{2}{1+4\tilde{\alpha}_x}$.

(i) For any $c_2 > 0$ we have:

$$P\left(\max_{j\leq p_n}|\hat{\rho}_j - \rho_j| > c_2 n^{-\kappa}\right) \leq O\left(s_n p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x^2 s_n}\right)^{\tilde{\alpha}}\right) + O\left(p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x \upsilon_\epsilon}\right)^{\tilde{\alpha}'}\right)$$

(ii) For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have:

$$P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O\left(s_n^2 \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x^2 s_n}\right)^{\tilde{\alpha}}\right) - O\left(s_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x \upsilon_\epsilon}\right)^{\tilde{\alpha}'}\right)$$

(iii) For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have:

$$P\left(|\hat{\mathcal{M}}_{\gamma_n}| \leq O(n^{2\kappa}\lambda_{max}(\Sigma))\right) \geq 1 - O\left(s_n p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x^2 s_n}\right)^{\tilde{\alpha}}\right) - O\left(p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x \upsilon_\epsilon}\right)^{\tilde{\alpha}'}\right)$$

From Theorem 7, we infer the covariate dimension (p_n) can be as large as $o(\min[\exp\left(\frac{Cn^{1/2-\kappa}}{s_n}\right)^{\tilde{\alpha}}/s_n, \exp(Cn^{1/2-\kappa})^{\tilde{\alpha}'}])$. As in Theorem 6, the range of p_n depends on the dependence in both the covariate and the error processes, the strength of the signal (κ), the moment condition, and the sparsity level (s_n). For the case of iid covariates and errors, we would replace υ_x and υ_ϵ with $\mu_{r/2} = \max_{j\leq p_n}||X_{ij}||_{r/2}$ and $||\epsilon_i||_q$ respectively. In contrast to Theorem 6, temporal dependence affects our range of p_n only by a constant factor.

If we assume $s_n = O(1)$, and both the covariate and error processes are sub-Gaussian we obtain $p_n = o(\exp(n^{\frac{1-2\kappa}{3}}))$, while for sub-exponential distributions we

obtain $p_n = o(\exp(n^{\frac{1-2\kappa}{5}}))$. In contrast, Fan and Lv Fan and Lv (2008), assuming independent observations, allow for a larger range $p_n = o(\exp(n^{1-2\kappa})$. However, their work relied critically on the Gaussian assumption. Fan and Song Fan and Song (2010), relax the Gaussian assumption by allowing for sub-exponential covariates and errors, and our rates are similar to theirs up to a constant factor. Additionally, in our work we relax the sub-exponential assumption, provided the tails of the covariates and errors are exponentially light.

3.4 Generalized Least Squares Screening (GLSS)

Consider the marginal model:

$$Y_t = X_{tk}\rho_k + \epsilon_{t,k} \tag{3.14}$$

where ρ_k is the linear projection of y_t onto X_{tk}. In SIS, we rank the magnitudes of the OLS estimates of this projection. In a time series setting, if we are considering the marginal model (3.14) it is likely the case that the marginal errors ($\epsilon_{t,k}$) will be serially correlated. This holds even if we assume that the errors (ϵ_t) in the full model (3.1) are serially uncorrelated. A procedure which accounts for this serial correlation, such as Generalized Least Squares (GLS), will provide a more efficient estimate of ρ_k.

We first motivate our method by considering a simple univariate model. Assume $Y_t = \beta X_t + \epsilon_t$ and the errors follow an AR(1) process, $\epsilon_t = \rho\epsilon_{t-1} + \theta_t$, where θ_t, and X_t are iid standard Gaussian. We set $\beta = .5$, $n = 200$, and estimate the model using both OLS and GLS for values of ρ ranging from .5 to .95. The mean absolute errors for both procedures is plotted in figure 3.1. We observe that the performance

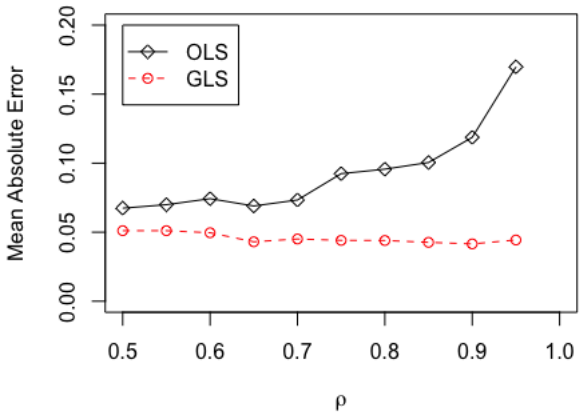

Figure 3.1: GLS vs OLS error comparison for values of ρ between .5 and .95 incrementing by .05. Absolute error averaged over 200 replications.

of OLS steadily deteriorates for increasing values of ρ, while the performance of GLS stays constant. This suggests that a screening procedure based on GLS estimates will be most useful in situations where we have weak signals and high levels of serial correlation.

The infeasible GLS estimate for ρ_k is:

$$\tilde{\beta}_k^M = (\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{y} \tag{3.15}$$

Where $\boldsymbol{X}_k$ is the k^{th} column of $\boldsymbol{X}$, and $\Sigma_k = (\gamma_{i-j,k})_{1 \leq i,j \leq n}$ is the auto-covariance matrix of $\boldsymbol{\epsilon}^k = (\epsilon_{t,k}, t = 1, ..., n)$. Given that Σ_k needs to be estimated to form our GLS estimates, we use the banded autocovariance matrix estimator introduced in Wu and Pourahmadi (2009), which is defined as:

$$\hat{\Sigma}_{k,l_n} = \left(\hat{\gamma}_{i-j,k} \mathbb{1}_{|i-j| \leq l_n}\right)_{1 \leq i,j \leq n} \tag{3.16}$$

Where l_n is our band length, $\hat{\gamma}_{r,k} = \frac{1}{n}\sum_{t=1}^{n-|r|} \hat{\epsilon}_{t,k}\hat{\epsilon}_{t+|r|,k}$, with $\hat{\epsilon}_{t,k} = y_t - X_{tk}\hat{\rho}_k$, and $\hat{\rho}_k$ is the OLS estimate of ρ_k. Our GLS estimator is now:

$$\hat{\beta}_k^M = (\boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{y} \tag{3.17}$$

When $E(\boldsymbol{\epsilon}^k | \boldsymbol{X}_k) = 0$, by the Gauss-Markov theorem it is clear that $\tilde{\beta}_k^M$ is efficient relative to the OLS estimator. Amemiya (1973) showed that under non-stochastic regressors and appropriate conditions on the error process, a two stage sieve type GLS estimator has the same limiting distribution as the infeasible GLS estimator $\tilde{\beta}_k^M$. In the appendix, we provide the appropriate conditions under which our GLS estimator, $\hat{\beta}_k^M$, and the infeasible GLS estimate, $\tilde{\beta}_k^M$, have the same asymptotic distribution.

For positive definite Σ_k, the banded estimate for Σ_k is not guaranteed to be positive definite, however it is asymptotically positive definite (see Lemma 9). For small samples, we can preserve positive definiteness by using the tapered estimate: $\hat{\Sigma}_k * R_{l_n}$, where R_{l_n} is a positive definite kernel matrix, and $*$ denotes coordinate-wise multiplication. For example, we can choose $R_{l_n} = (\max(1 - \frac{|i-j|}{l_n}, 0))_{1 \le i,j \le n}$. We need the following conditions for the sure screening property to hold:

Condition E: Assume the marginal error process, $\epsilon_{t,k}$, is a stationary AR(L_k) process, $\epsilon_{t,k} = \sum_{i=1}^{L_k} \alpha_i \epsilon_{t-i,k} + e_t$. Where $L_k < K < \infty$, $\forall k \le p_n$.

Condition F: For $k \in M_*$, $\kappa < 1/2$:
$\beta_k^M = E(y_t - \sum_{i=1}^{L_k} \alpha_i y_{t-i})(X_{t,k} - \sum_{i=1}^{L_k} \alpha_i X_{t-i,k})/(E(X_{t,k} - \sum_{i=1}^{L_k} \alpha_i X_{t-i,k})^2) \ge c_6 n^{-\kappa}$.

Condition G: Assume $E(X_{tk}), E(\epsilon_t), E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{\epsilon}) = 0$

Condition H: Assume $\epsilon_{t,k}$, ϵ_t are of the form (3.2), and the covariate process is of the form (3.6). Additionally we assume the following decay rates $\Delta_{m,q}(\boldsymbol{\epsilon}) = O(m^{-\alpha_\epsilon})$, $\Phi_{m,r}(\boldsymbol{x}) = O(m^{-\alpha_x}), \chi_{m,q'} = \sum_{i=m}^{\infty} \max_{k \leq p_n} \delta_q(\epsilon_{i,k}) = O(m^{-\alpha})$, for some $\alpha_x, \alpha_\epsilon > 0$, $\alpha = min(\alpha_x, \alpha_\epsilon)$, and $q' = \min(q, r) \geq 4$.

Condition I: Assume $\epsilon_{t,k}$, ϵ_t are of the form (3.2), and the covariate process is of the form (3.6). Additionally assume $\upsilon_x = \sup_{q \geq 4} q^{-\tilde{\alpha}_x} \Phi_{0,q}(\boldsymbol{x}) < \infty$,$\upsilon_\epsilon = \sup_{q \geq 4} q^{-\tilde{\alpha}_\epsilon} \Delta_{0,q}(\boldsymbol{\epsilon}) < \infty, \phi = \sup_{q \geq 4} q^{-\varphi} \chi_{0,q} < \infty$ for some $\tilde{\alpha}_x, \tilde{\alpha}_\epsilon \geq 0$, and $\varphi = \max(\tilde{\alpha}_\epsilon, \tilde{\alpha}_x)$.

In Condition E, we can let the band length K diverge to infinity at a slow rate, e.g $O(\log(n))$, for simplicity we set K to be a constant. Assuming a finite order AR model for the marginal error process is reasonable in most practical situations, since any stationary process with a continuous spectral density function can be approximated arbitrarily closely by a finite order linear AR process (see corollary 4.4.2 in Brockwell and Davis (1991)). For further details on linear AR approximations to stationary processes, see Amemiya (1973) and Bühlmann (1995). We remark that compared to previous works Amemiya (1973); Koreisha and Fang (2001), knowledge about the structure of the marginal errors is not necessary in estimating β_k^M, since we use a non-parametric estimate of Σ_k. Therefore Condition E is assumed strictly for technical reasons.

For Condition F, from (3.14), we have $\beta_k^M = \rho_k$, iff $E(\epsilon_{t,k} - \sum_{i=1}^{L_k} \alpha_i \epsilon_{t-i,k})(X_{t,k} - \sum_{i=1}^{L_k} \alpha_i X_{t-i,k}) = 0$. When $\beta_k^M \neq \rho_k$, recall that:

$$\beta_k^M = E(y_t - \sum_{i=1}^{L_k} \alpha_i y_{t-i})(X_{t,k} - \sum_{i=1}^{L_k} \alpha_i X_{t-i,k}) / (E(X_{t,k} - \sum_{i=1}^{L_k} \alpha_i X_{t-i,k})^2) \tag{3.18}$$

If we assume the cross covariance, $\gamma_{\boldsymbol{X}_k,Y}(h)$, is proportional to $E(X_{tk}Y_t)$, i.e. $\gamma_{\boldsymbol{X}_k,Y}(h) \propto E(X_{tk}Y_t)$, for $h \in \{-L_k,\ldots,-1,1,\ldots,L_k\}$, then $\beta_k^M \propto \rho_k$ whenever $|\beta_k^M| > 0$. And for $|\rho_k| > 0$, it is likely the case that $\beta_k^M \propto \rho_k$ if we assume $\gamma_{\boldsymbol{X}_k,Y}(h) \propto E(X_{tk}Y_t)$, for $h \in \{-L_k,\ldots,-1,1,\ldots,L_k\}$. When $\beta_k^M \neq \rho_k$, we believe the advantage in using GLSS is due to the GLS estimator being robust to serial correlation in the marginal error process (see the appendix for details).

For Condition H, since $\epsilon_{t,k} = Y_t - X_{tk}\rho_k$, we have $\epsilon_{t,k} = r_k(\ldots,\boldsymbol{\theta}_{t-1},\boldsymbol{\theta}_t)$, where $r_k(\cdot)$ is a measurable function and $\boldsymbol{\theta}_t = (\boldsymbol{\eta}_t, e_t)$. If we assume e_t, and $\boldsymbol{\eta}_i$ are independent for $i \neq t$, then $\boldsymbol{\theta}_i$ are iid. We then have:

$$\begin{aligned}\delta_{q'}(\epsilon_{t,k}) =& || \sum_{i\in M_*} X_{ti}\beta_i + \epsilon_t - X_{tk}\rho_k - (\sum_{i\in M_*} X^*_{ti}\beta_i + \epsilon^*_t - X^*_{tk}\rho_k)||_{q'} \\ &\leq \sum_{i\in M_*} |\beta_i|\delta_{q'}(X_{ti}) + \delta_{q'}(\epsilon_t) + |\rho_k|\delta_{q'}(X_{tk})\end{aligned}$$

Therefore, $\chi_{m,q'} = O(m^{-\alpha})$, if we assume $\sum_{i\in M_*}|\beta_i| = O(1)$, $\Delta_{m,q}(\boldsymbol{\epsilon}) = O(m^{-\alpha_\epsilon})$, and $\Phi_{m,r}(\boldsymbol{x}) = O(m^{-\alpha_x})$.

For GLSS; define $\hat{\mathcal{M}}_{\gamma_n} = \left\{1 \leq i \leq p_n : |\hat{\beta}_k^M| \geq \gamma_n\right\}$, $\alpha = min(\alpha_x, \alpha_\epsilon)$, $\tau = \frac{qr}{q+r}, \tau' = \frac{qq'}{q+q'} = \min(q/2, \tau)$. Let $\iota = 1$ if $\alpha > 1/2 - 1/\tau'$, otherwise $\iota = \tau'/2 - \tau'\alpha$. Let $\zeta = 1$, if $\alpha > 1/2 - 2/q'$, otherwise $\zeta = q'/4 - \alpha q'/2$ and let $\omega = 1$, if $\alpha_x > 1/2 - 2/r$, otherwise $\omega = r/4 - \alpha_x r/2$. Additionally, let $K_{x,r} = \max_{j\leq p_n}\sup_{m\geq 0}(m+1)^{\alpha_x}\sum_{i=m}^{\infty}\delta_r(X_{ij})$, $\tilde{K}_{\epsilon,q'} = \max_{k\leq p_n}\sup_{m\geq 0}(m+1)^{\alpha}\sum_{i=m}^{\infty}\delta_{q'}(\epsilon_{i,k})$. Given Condition H, it follows that $K_{x,r}, \tilde{K}_{\epsilon,q'} < \infty$. For the case of exponentially light tails, we define

$\tilde{\varphi}' = \frac{2}{1+2\tilde{\alpha}_x+2\varphi}, \tilde{\varphi} = \frac{2}{1+4\varphi}$, and $\tilde{\alpha} = \frac{2}{1+4\tilde{\alpha}_x}$. Lastly, for ease of presentation let:

$$a_n = l_n \left[\frac{n^{\iota} l_n^{\tau'} K_{x,r}^{\tau'} \tilde{K}_{\epsilon,q'}^{\tau'}}{n^{\tau'-\tau'\kappa}} + \frac{n^{\zeta} l_n^{q'/2} \tilde{K}_{\epsilon,q'}^{q'}}{n^{q'/2-q'\kappa/2}} + \frac{n^{\omega} l_n^{r/2} K_{x,r}^{r}}{n^{r/2}} \right] \tag{3.19}$$

$$b_n = l_n \left[\exp\left(-\frac{n^{1/2}}{l_n \upsilon_x^2} \right)^{\tilde{\alpha}} + \exp\left(-\frac{n^{1/2-\kappa}}{l_n \upsilon_x \phi} \right)^{\tilde{\varphi}'} + \exp\left(-\frac{n^{1/2-\kappa}}{l_n \phi^2} \right)^{\tilde{\varphi}} \right] \tag{3.20}$$

We first present the following lemma, which provides deviation bounds on $||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2$. This lemma, which is of independent interest, will allow us to obtain deviation bounds on our GLSS estimates.

Lemma 9. *Assume the band length, $l_n = c\log(n)$ for sufficiently large $c > 0$.*

(i) Assume Condition H holds. For $\kappa \in [0, 1/2)$ we have the following:

$$P(||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 > cn^{-\kappa}) \leq O(a_n)$$

(ii) Assume Condition I holds. For $\kappa \in [0, 1/2)$ we have the following:

$$P(||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 > cn^{-\kappa}) \leq O(b_n)$$

The following theorem gives the sure screening properties of GLSS:

Theorem 8. *Assume the band length, $l_n = c\log(n)$ for sufficiently large $c > 0$.*

(i) Assume Conditions E,F,G,H hold, for any $c_2 > 0$ we have:

$$P\left(\max_{j \leq p_n} |\hat{\beta}_k^M - \beta_k^M| > c_2 n^{-\kappa} \right) \leq O(p_n a_n)$$

(ii) Assume Conditions E,F,G,H hold, then for $\gamma_n = c_5 n^{-\kappa}$ *with* $c_5 \leq c_6/2$*:*

$$P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(s_n a_n)$$

(iii) Assume Conditions E,F,G,I hold, for any $c_2 > 0$ *we have:*

$$P\left(\max_{j \leq p_n} |\hat{\beta}_k^M - \beta_k^M| > c_2 n^{-\kappa}\right) \leq O(p_n b_n)$$

(iv) Assume Conditions E,F,G,I hold, then for $\gamma_n = c_5 n^{-\kappa}$ *with* $c_5 \leq c_6/2$*:*

$$P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(s_n b_n)$$

In Lemma 9, the rate of decay also depends on the band length (l_n). The band length primarily depends on the decay rate of the autocovariances of the process $\epsilon_{t,k}$. Since we are assuming an exponential decay rate, we can set $l_n = O(\log(n))$. If $\gamma_{i,k} = O(i^{-\beta})$ for $\beta > 1$, then we require $l_n^{-\beta+1} = o(n^{-\kappa})$. We omit the exponential terms in the bounds for part part (i) of Lemma 9, and parts (i), and (ii) of Theorem 8 to conserve space and provide a cleaner result. For GLSS, the range for p_n also depends on the band length (l_n), in addition to the moment conditions and the strength of dependence in the covariate and error processes. For example, if we assume $r = q$, and $\alpha \geq 1/2 - 2/r$ then $p_n = o(n^{r/2-r\kappa/2-1}/l_n^{r/2+1})$. Compared to SIS, we have a lower range of p_n by a factor of $l_n^{r/2+1}$. We conjecture that this is due to our proof strategy, which relies on using a deviation bound on $||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2$, and uses the functional dependence measure, rather than autocorrelation, to quantify dependence. In practice, we believe using GLSS, which corrects for serial correlation, and uses an estimator with lower asymptotic variance will achieve better performance. We

illustrate this in more detail in our simulations section, and in the appendix (section 3.9.2).

Similar to SIS, we can control the size of the model selected by GLSS. For the case when $\beta_k^M = \rho_k \; \forall k$, the bound on the selected model size is the same as in SIS. However, we need to place an additional assumption when $\beta_k^M \neq \rho_k$: If the cross covariance, $\gamma_{\boldsymbol{X}_k,Y}(h) \propto E(X_{tk}Y_t)$, for $h \in \{-L_k, \ldots, -1, 1, \ldots, L_k\}$, we can bound the selected model size by the model size selected by SIS. More formally we have:

Corollary 9. *Assume the cross covariance,* $\gamma_{\boldsymbol{X}_k,Y}(h) \propto E(X_{k,t}Y_t)$, *for* $h \in \{-L_k, \ldots, -1, 1, \ldots, L_k\}$

(i) Assume Conditions E,F,G,H hold, then for $\gamma_n = c_5 n^{-\kappa}$ *with* $c_5 \leq c_6/2$*:*

$$P\left(|\hat{\mathcal{M}}_{\gamma_n}| \leq O(n^{2\kappa}\lambda_{max}(\Sigma))\right) \geq 1 - O(p_n a_n)$$

(ii) Assume Conditions E,F,G,I hold, then for $\gamma_n = c_5 n^{-\kappa}$ *with* $c_5 \leq c_6/2$*:*

$$P\left(|\hat{\mathcal{M}}_{\gamma_n}| \leq O(n^{2\kappa}\lambda_{max}(\Sigma))\right) \geq 1 - O(p_n b_n)$$

3.5 Second Stage Selection with Adaptive Lasso

The adaptive Lasso, as introduced by Zou (2006a), is the solution to the following:

$$argmin_\beta \; ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_n \sum_{j=1}^{p_n} w_j|\beta_j|, \text{ where } w_j = |\hat{\beta}_{I,j}|^{-1}, \tag{3.21}$$

and $\hat{\beta}_{I,j}$ is our initial estimate. For sign consistency; when $p_n >> n$, the initial estimates can be the marginal regression coefficients provided the design matrix satisfies

the partial orthogonality condition as stated in Huang et al. (2008), or we can use the Lasso as our initial estimator provided the restricted eigenvalue condition holds (see Medeiros and Mendes (2016)). Both of these conditions can be stringent when $p_n >> n$. This makes the adaptive Lasso a very attractive option as a second stage variable selection method, after using screening to significantly reduce the dimension of the feature space. We have the following estimator:

$$\tilde{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}} = argmin_{\boldsymbol{\beta}_{\hat{\mathcal{M}}_{\gamma_n}}} \; ||\boldsymbol{y} - \boldsymbol{X}_{\hat{\mathcal{M}}_{\gamma_n}} \boldsymbol{\beta}_{\hat{\mathcal{M}}_{\gamma_n}}||^2 + \lambda_n \sum_{j=1}^{d_n} w_j |\beta_j|, w_j = |\hat{\beta}_{I,j}|^{-1} \tag{3.22}$$

Where $\boldsymbol{X}_{\hat{\mathcal{M}}_{\gamma_n}}$ denotes the $n \times d_n$ submatrix of $\boldsymbol{X}$ that is obtained by extracting its columns corresponding to the indices in $\hat{\mathcal{M}}_{\gamma_n}$. We additionally define $\boldsymbol{X}_{\mathcal{M}_{\gamma_n}}$ accordingly. Our initial estimator $\hat{\boldsymbol{\beta}}_I = (\hat{\beta}_{I,1}, \ldots, \hat{\beta}_{I,d_n})$ is obtained using the Lasso. Let $\hat{\Sigma}_{\mathcal{M}_{\gamma_n}} = \boldsymbol{X}^T_{\mathcal{M}_{\gamma_n}} \boldsymbol{X}_{\mathcal{M}_{\gamma_n}}/n$, and let $\Sigma_{\mathcal{M}_{\gamma_n}}$ be its population counterpart. Our two stage estimator, $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}}$, is then formed by inserting zeroes corresponding to the covariates which were excluded in the screening step, and inserting the adaptive Lasso estimates, $\tilde{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}}$, for covariates which were selected by the screening step. We need the following conditions for the combined two stage estimator to achieve sign consistency:

Condition J: The matrix $\Sigma_{\mathcal{M}_{\frac{\gamma_n}{2}}}$ satisfies the restricted eigenvalue condition, RE(s_n,3)(see Bickel et al. (2009b) for details):

$$\phi_0 = \min_{S \subseteq \{1,\ldots,d'_n\}, |S| \leq s_n} \; \min_{\boldsymbol{v} \neq 0, |\boldsymbol{v}_{S^c}| \leq 3|\boldsymbol{v}_S|} \frac{\boldsymbol{v}^T \Sigma_{\mathcal{M}_{\frac{\gamma_n}{2}}} \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}} \geq c > 0, \tag{3.23}$$

where $\boldsymbol{v} = (v_1, \ldots, v_{d'_n})$ and $\boldsymbol{v}_S = (v_i, i \in S), \boldsymbol{v}_{S^c} = (v_i, i \in S^c)$.

Condition K: Let λ_n and $\lambda_{I,n}$ be the regularization parameters of the adaptive

lasso and the initial lasso estimator respectively. For some $\psi \in (0,1)$, we assume:

$$cn^{1-\frac{\psi}{2}}(\frac{\phi_0}{s_n})^{3/2} \geq \lambda_{I,n} \geq \lambda_n n^{\psi/2} \tag{3.24}$$

Condition L: Let $\beta_{min} = \min_{i \leq s_n} |\beta_i|$, and $w_{\max} = \max_{i \leq s_n} w_i > 0$. Assume $\beta_{min} > \frac{2}{w_{\max}}$ and $\beta_{min} > 2c\frac{\lambda_{I,n} s_n}{\phi_0 n}$.

Condition J allows us to use the Lasso as our initial estimator. Notice that we placed the RE(s_n,3) assumption on the matrix $\Sigma_{\mathcal{M}_{\frac{\gamma_n}{2}}}$, rather than the matrix $\hat{\Sigma}_{\hat{\mathcal{M}}_{\gamma_n}}$, given the indices in $\hat{M}_{\gamma_n}$ are random as a result of our screening procedure. Recall that for SIS, $\mathcal{M}_{\frac{\gamma_n}{2}} = \{1 \leq i \leq p : |\rho_i| \geq \gamma_n/2\}$, and $|\mathcal{M}_{\frac{\gamma_n}{2}}| = d'_n = O(d_n)$, and for GLSS we have a similar definition. Therefore, we are placing the RE(s_n,3) assumption on the population covariance matrix of a fixed set of d'_n predictors. Conditions K and L are standard assumptions, and are similar to the ones used in Medeiros and Mendes (2016). Condition K primarily places restrictions on the rate of increase of λ_n, and $\lambda_{I,n}$. Condition L places a lower bound on the magnitude of the non-zero parameters which decays with the sample size.

The next theorem deals with the two stage SIS-Adaptive Lasso estimator. A very similar result applies to the two stage GLSS-Adaptive Lasso estimator, if we replace Conditions A,B,C (resp. D) with Conditions E,F,G,H (resp. I), to avoid repetition we omit the result. For the following theorem, the terms $\iota, \omega, K_{x,r}$, and $K_{\epsilon,q}$ have been defined in the paragraph preceding Theorem 6, and $\tilde{\alpha}', \tilde{\alpha}$ have been defined in Theorem 7.

Theorem 10. *(i) Assume Conditions A,B,C,J,K,L hold, then for $\gamma_n = c_3 n^{-\kappa}$ with*

$c_3 \leq c_1/2$ *we have:*

$$P(sgn(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}}) = sgn(\boldsymbol{\beta})) \geq 1 - O\left(s_n p_n \left[\frac{n^{\omega} K_{x,r}^{r}}{(n/s_n)^{r/2-r\kappa/2}} - \exp(-\frac{n^{1-2\kappa}}{s_n^2 K_{x,r}^4})\right]\right)$$
$$- O\left(p_n \left[\frac{n^{\iota} K_{x,r}^{\tau} K_{\epsilon,q}^{\tau}}{n^{\tau-\tau\kappa}} - \exp(-\frac{n^{1-2\kappa}}{K_{x,r}^2 K_{\epsilon,q}^2})\right]\right)$$
$$- O\left(d_n^{'2} \left[\frac{n^{\omega} K_{x,r}^{r}}{(n/s_n)^{r/2}} - \exp(-\frac{n}{s_n^2 K_{x,r}^4})\right]\right)$$
$$- O\left(d_n' \left[\frac{n^{\iota} K_{x,r}^{\tau} K_{\epsilon,q}^{\tau}}{\lambda_n^{\tau} n^{\tau\psi/2}} + \exp(-\frac{\lambda_n^2 n^{\psi-1}}{K_{x,r}^2 K_{\epsilon,q}^2})\right]\right)$$

(ii) Assume Conditions A,B,C,J,K,L hold, then for $\gamma_n = c_3 n^{-\kappa}$ *with* $c_3 \leq c_1/2$ *we have:*

$$P(sgn(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}}) = sgn(\boldsymbol{\beta})) \geq 1 - O(s_n p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x^2 s_n}\right)^{\tilde{\alpha}})$$
$$-O(p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x \upsilon_\epsilon}\right)^{\tilde{\alpha}'}) - O(d_n^{'2} \exp\left(-\frac{n^{1/2}}{\upsilon_x^2 s_n}\right)^{\tilde{\alpha}})$$
$$-O(d_n' \exp\left(-\frac{\lambda_n n^{\psi/2-1/2}}{\upsilon_x \upsilon_\epsilon}\right)^{\tilde{\alpha}'})$$

To achieve sign consistency for the case of finite polynomial moments we require:

Condition M: Assume $\lambda_n n^{\psi/2-1/2} \to \infty$ and $p_n = o(\min(\frac{s_n (n/s_n)^{r/2-r\kappa/2}}{n^{\omega}}, \frac{n^{\tau-\tau\kappa}}{n^{\iota}}))$, $d_n' = o(min((n/s_n)^{r/4-\omega/2}, \lambda_n^{\tau} n^{\tau\psi/2-\iota}))$

For the case of exponential moments, we require:

Condition N: Assume $\lambda_n n^{\psi/2-1/2} \to \infty$,
$p_n = o(\min(\exp\left(\frac{C n^{1/2-\kappa}}{s_n}\right)^{\tilde{\alpha}}/s_n, \exp(C n^{1/2-\kappa})^{\tilde{\alpha}'}))$,
and $d_n' = o(\min(\exp\left(\frac{n^{1/2}}{s_n}\right)^{\tilde{\alpha}/2}, \exp\left(\lambda_n n^{\psi/2-1/2}\right)^{\tilde{\alpha}'}))$

From Conditions M, N, and Theorem 8, we see an additional benefit of using the two stage selection procedure as opposed to using the adaptive Lasso as a stand alone procedure. For example, if we assume $d_n \leq n^{2\kappa}\lambda_{\max}(\Sigma) = O(n)$, and that both the error and covariate processes are sub-Gaussian, we obtain $p_n = o(\exp(n^{\frac{1-2\kappa}{3}}))$ for the two stage estimator. By setting $d_n' = p_n$, we obtain the result when using the adaptive Lasso as a stand alone procedure, with the Lasso as its initial estimator. Under the scenario detailed above, the dimension of the feature space, which depends on λ_n and ψ, for the stand alone adaptive Lasso can be at most $p_n = o(\exp(n^{\frac{1}{6}}))$. Therefore for $\kappa < 1/4$, we obtain a larger range for p_n and a faster rate of decay using the two stage estimator. For $\kappa \geq 1/4$ it is not clear whether the two stage estimator has a larger range for p_n, compared to using the adaptive Lasso alone.

The sign consistency of the stand alone adaptive Lasso estimator in the time series setting was established in Medeiros and Mendes (2016). Their result was obtained under strong mixing assumptions on the covariate and error processes, with the additional assumption that the error process is a martingale difference sequence. Additionally, in the ultrahigh dimensional setting they require a geometric decay rate on the strong mixing coefficients. In contrast, we obtain results for both the two stage and stand alone adaptive lasso estimator, and our results are obtained using the functional dependence measure framework. Besides assuming moment conditions, we are not placing any additional assumptions on the temporal decay of the covariate and error processes other than $\Delta_{0,q}(\boldsymbol{\epsilon}),\Phi_{0,q}(\boldsymbol{x}) < \infty$. Furthermore, we weaken the martingale difference assumption they place on the error process, thereby allowing for serial correlation in the error process. Finally, by using Nagaev type inequalities introduced in Wu and Wu (2016), our results are easier to interpret and also allow us obtain a higher range for p_n.

3.6 Simulations

In this section, we evaluate the performance of SIS, GLSS, and the two stage selection procedure using the adaptive Lasso. For GLSS instead of using the banded estimate for Σ_k we use a tapered estimate: $\hat{\Sigma}_k * R_{l_n}$, where $\hat{\Sigma}_k = (\hat{\gamma}_{i-j,k})_{1\leq i,j\leq n}$ and $R_{l_n} = (\max(1 - \frac{|i-j|}{l_n}, 0))_{1\leq i,j\leq n}$ is the triangular kernel. We fix $l_n = 15$, and we observed the results were fairly robust to the choice of l_n. In our simulated examples, we fix $n = 200, s_n = 6$ and $d_n = n - 1$, while we vary p_n from 1000 to 5000. We repeat each experiment 200 times. For screening procedures, we report the proportion of times the true model is contained in our selected model. For the two stage procedure using the adaptive Lasso, we report the proportion of times there was a λ_n on the solution path which *selected the true model*.

Case 1: Uncorrelated Features

Consider the model (3.1), for the covariate process we have:

$$\boldsymbol{x}_t = A_1 \boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t \tag{3.25}$$

Where $A_1 = diag(\gamma)$, and we vary γ from .4 to .6. We set $\boldsymbol{\eta}_t \sim N(0, \Sigma_\eta)$, or $\boldsymbol{\eta}_t \sim t_5(0, V)$ in which case the covariance matrix is $\Sigma_\eta = (5/3) * V$. For this scenario we will be dealing with uncorrelated predictors, we set $\Sigma_\eta = I_{p_n}$. For the error process, we have an AR(1) process: $\epsilon_i = \alpha\epsilon_{i-1} + e_i$. We let α vary from .6 to .9, and let $e_i \sim t_5$ or $e_i \sim N(0, 1)$. We set $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, where $\boldsymbol{\beta}_1 = (.5, .5, .5, .5, .5, .5)$ and $\boldsymbol{\beta}_2 = \mathbf{0}$. Even though the features are uncorrelated, this is still a challenging setting, given the low signal to noise ratio along with heavy tails and serial dependence being present.

Table 3.1: Case 1

	SIS			GLSS		
(γ, α)	(.4,.6)	(.5,.8)	(.6,.9)	(.4,.6)	(.5,.8)	(.6,.9)
Gaussian						
$p_n = 1000$	.95	.63	.15	.99	.99	.98
$p_n = 5000$	.62	.11	.01	.95	.95	.97
t_5						
$p_n = 1000$	.58	.26	.06	.83	.84	.83
$p_n = 5000$	.21	.01	0	.55	.49	.50

The results are displayed in table 3.1. The entries below "Gaussian" correspond to the setting where both e_i and $\boldsymbol{\eta}_i$ are drawn from a Gaussian distribution. Accordingly the entries under "t_5" correspond to the case where e_i and $\boldsymbol{\eta}_i$ are drawn from a t_5 distribution. We see from the results that the performance of SIS, and GLSS are comparable when $p_n = 1000$, with moderate levels of temporal dependence, along with Gaussian covariates and errors. Interestingly, in this same setting, switching to heavy tails seems to have a much larger effect on the performance of SIS vs GLSS. In all cases, the performance of GLSS appears to be robust to the effects of serial correlation in the covariate and the error processes. Whereas, for SIS the performance severely deteriorates as we increase the level of serial correlation. For example, for our highest levels of serial correlation, SIS nearly always fails to contain the true model.

Case 2: Correlated Features

We now compare the performance of SIS and GLSS for the case of correlated predictors. We have two scenarios:

Scenario A: The covariate process is generated from (3.25), with $A_1 = diag(.4)$. $\boldsymbol{\eta}_t \sim N(0, \Sigma_\eta)$, or $\boldsymbol{\eta}_t \sim t_5(0, V)$, with $\Sigma_\eta = \{.3^{|i-j|}\}_{i,j \leq p_n}$ for both cases. Therefore $\Sigma = \sum_{i=0}^{\infty} .4^{2i} \Sigma_\eta$. We set $\boldsymbol{\beta}_1 = (1, -1, 1, -1, 1, -1)$ and $\boldsymbol{\beta}_2 = \mathbf{0}$. We have an AR(1)

Table 3.2: Case 2: Scenario A

	SIS			GLSS		
α	.4	.6	.8	.4	.6	.8
Gaussian						
$p_n = 1000$	.83	.73	.55	.95	.90	.90
$p_n = 5000$	.38	.30	.07	.63	.63	.57
t_5						
$p_n = 1000$	.44	.42	.21	.56	.56	.53
$p_n = 5000$	.01	.04	0	.16	.14	.16

process for the errors: $\epsilon_i = \alpha\epsilon_{i-1} + e_i$, we vary α from .4 to .8, and set $e_i \sim t_5$ or $e_i \sim N(0,1)$

Scenario B: The covariate process is generated from (3.25), with $A_1 = \{.4^{|i-j|+1}\}_{i,j\leq p_n}$. And $\boldsymbol{\eta}_t \sim N(0,\Sigma_\eta)$, or $\boldsymbol{\eta}_t \sim t_5(0,V)$, with $\Sigma_\eta = I_{p_n}$ for both cases. Therefore $\Sigma = \sum_{i=0}^{\infty}(A_1^T)^i A_1^i$. We set $\boldsymbol{\beta}_1 = (1,-1,1,-1,1,-1)$ and $\boldsymbol{\beta}_2 = \mathbf{0}$. We have an AR(1) process for the errors: $\epsilon_i = \alpha\epsilon_{i-1} + e_i$, and we vary α from .4 to .8. The errors are generated in the same manner as in scenario A above.

The results are displayed in tables 3.2, and 3.3 respectively. In scenario A, we have a Toeplitz covariance matrix for the predictors, and moderate levels of serial dependence in the predictors. The trends are similar to the ones we observed in case 1. The performance of SIS is sensitive to the effects of increasing the serial correlation in the errors, with the effect of serial dependence being more pronounced as we encounter heavy tail distributions. In contrast, increasing the level of serial dependence has a negligible impact on the performance of GLSS. For scenario B, we observe similar trends as in scenario A.

Case 3: Two Stage Selection

Table 3.3: Case 2: Scenario B

	SIS			GLSS		
α	.4	.6	.8	.4	.6	.8
Gaussian						
$p_n = 1000$	.90	.82	.68	.99	1.00	1.00
$p_n = 5000$	.71	.64	.26	.95	.97	.98
t_5						
$p_n = 1000$	.76	.63	.40	.92	.90	.92
$p_n = 5000$	.37	.26	.06	.76	.74	.75

We test the performance of the two stage GLSS-AdaLasso procedure. We also compare its performance with using the adaptive Lasso on its own. We use the Lasso as our initial estimator and select $\lambda_{I,n}$ using the modified BIC introduced in Wang et al. (2009). Fan and Tang (2013) extended the theory of the modified BIC to the case where $p > n$, $p = o(n^a)$, $a > 1$, and independent observations. We conjecture that the same properties hold in a time series setting. We have two scenarios:

Scenario A: The covariate process is generated from (3.25), with $A_1 = diag(.4)$. And $\boldsymbol{\eta}_t \sim N(0, \Sigma_\eta)$, or $\boldsymbol{\eta}_t \sim t_5(0, V)$, with $(\Sigma_\eta)_{i,j} = \{.8^{|i-j|}\}_{i,j \leq p_n}$. We set $\boldsymbol{\beta}_1 = (.5, .5, .5, .5, .5, .5)$ and $\boldsymbol{\beta}_2 = \mathbf{0}$. We have an AR(1) process for the errors: $\epsilon_i = \alpha\epsilon_{i-1} + e_i$, we vary α from .4 to .6, and set $e_i \sim t_5$ or $e_i \sim N(0, 1)$

Scenario B: The covariate process is generated from (3.25), with $A_1 = \{.4^{|i-j|+1}\}_{i,j \leq p_n}$. And $\boldsymbol{\eta}_t \sim N(0, \Sigma_\eta)$, or $\boldsymbol{\eta}_t \sim t_5(0, V)$, with $(\Sigma_\eta)_{i,j} = .8$ for $i \neq j$ and 1 otherwise. We set $\boldsymbol{\beta}_1 = (.75, .75, .75, .75, .75, .75)$ and $\boldsymbol{\beta}_2 = \mathbf{0}$. The errors are generated the same as in scenario A above.

In both scenarios we have a high degree of correlation between the predictors, low signal to noise ratio, along with mild to moderate levels of serial correlation in the covariate and error processes. The results are displayed in tables 3.4 and 3.5 for

Table 3.4: Case 3: Scenario A

	GLSS-AdaLasso			AdaLasso		
α	.4	.5	.6	.4	.5	.6
Gaussian						
$p_n = 1000$	.79	.65	.49	.60	.49	.35
$p_n = 5000$	.84	.65	.46	.66	.43	.29
t_5						
$p_n = 1000$	.45	.37	.23	.32	.22	.14
$p_n = 5000$	.36	.32	.18	.24	.18	.10

Table 3.5: Case 3: Scenario B

	GLSS-AdaLasso			AdaLasso		
α	.4	.5	.6	.4	.5	.6
Gaussian						
$p_n = 1000$	.86	.72	.59	.57	.49	.34
$p_n = 5000$	.69	.59	.43	.60	.44	.25
t_5						
$p_n = 1000$	.48	.41	.22	.30	.19	.10
$p_n = 5000$	.35	.25	.19	.25	.16	.11

scenarios A and B respectively. We observe that the two stage estimator outperforms the standalone adaptive Lasso for both scenarios, with the difference being more pronounced in scenario B. For both scenarios, going from mild to moderate levels of serial correlation in the errors appears to significantly deteriorate the performance of the adaptive Lasso. This affects our results for the two stage estimator primarily at the second stage of selection. This sensitivity to serial correlation appears to increase as we encounter heavy tailed distributions.

3.7 Real Data Example: Forecasting Inflation Rate

In this section we focus on forecasting the 12 month ahead inflation rate. We use two major monthly price indexes as measures of inflation: the consumer price index (CPI), and the producer price index less finished goods (PPI). Specifically we are

forecasting:

$$y_{t+12}^{12} = 100 \times \log\left(\frac{CPI_{t+12}}{CPI_t}\right), \text{ or } y_{t+12}^{12} = 100 \times \log\left(\frac{PPI_{t+12}}{PPI_t}\right) \tag{3.26}$$

Therefore the above quantities are approximately the percentage change in CPI or PPI over 12 months. Our data was obtained from the supplement to Jurado et al. (2015), and it consists of 132 monthly macroeconomic variables from January 1960 to December 2011, for a total of 624 observations. Apart from $\log(CPI)$ and $\log(PPI)$ which we are treating as $I(1)$, the remaining 130 macroeconomic time series have been transformed to achieve stationarity according to Jurado et al. (2015). Treating $\log(CPI)$, and $\log(PPI)$ as $I(1)$, has been found to provide an adequate description of the data according to Stock and Watson (2002c),Stock and Watson (1999),Medeiros and Mendes (2016).

We consider forecasts from 8 different models. Similar to Medeiros and Mendes (2016); Stock and Watson (2002c) our benchmark model is an AR(4) model: $\hat{y}_{t+12}^{12} = \hat{\alpha}_0 + \sum_{i=0}^{3} \hat{\alpha}_i y_{t-i}$, where $y_t = 1200 \times log(CPI_t/CPI_{t-1})$ when forecasting CPI, and $y_t = 1200 \times log(PPI_t/PPI_{t-1})$ when forecasting PPI. For comparison, we also consider an AR(4) model augmented with 4 factors. Specifically we have:

$$\hat{y}_{t+12}^{12} = \hat{\beta}_0 + \sum_{i=0}^{3} \hat{\alpha}_i y_{t-i} + \hat{\boldsymbol{\gamma}} \hat{\boldsymbol{F}}_t \tag{3.27}$$

Where $\hat{\boldsymbol{F}}_t$ are four factors which are estimated by taking the first four principal components of the 131 predictors along with three of their lags. We also consider forecasts estimated by the Lasso and the adaptive Lasso. And lastly we include forecasts estimated by the following two stage procedures: GLSS-Lasso, GLSS-adaptive Lasso, SIS-Lasso, and SIS-Adaptive Lasso. Our forecasting equation for the penalized re-

gression and two stage forecasts is:

$$y_{t+12}^{12} = \beta_0 + \boldsymbol{x}_t\boldsymbol{\beta} + \epsilon_{t+12}^{12} \tag{3.28}$$

Where $\boldsymbol{x}_t$ consists of y_t and three of its lags along with the other 131 predictors and three of their lags, additionally we also include the first four estimated factors $\hat{F}_t$. Therefore $\boldsymbol{x}_t$ consists of 532 covariates in total. For each of the two stage methods, we set $d_n = \lceil n/\log(n) \rceil = 73$ for the first stage screening procedure. For the second stage selection, and the standalone lasso/adaptive lasso models, we select the tuning parameters and initial estimators using the approach described in section 3.6.

We utilize a rolling window scheme, where the first simulated out of sample forecast was for January 2000 (2000:1). To construct this forecast, we use the observations between 1960:6 to 1999:1 (the first five observations are used in forming lagged covariates and differencing) to estimate the factors, and the coefficients. Therefore for the models described above, t=1960:6 to 1998:1. We then use the regressor values at t=1999:1 to form our forecast for 2000:1. Then the next window uses observations from 1960:7 to 1999:2 to forecast 2000:2. Using this scheme, in total we have 144 out of sample forecasts, and for each window we use $n = 451$ observations for each regression model. The set-up described above allows us to simulate real-time forecasting.

Table 3.6 shows the mean squared error (MSE), and the mean absolute error (MAE) of the resulting forecasts relative to the MSE and MAE of the baseline AR(4) forecasts. We observe that the two stage GLSS methods clearly outperform the benchmark AR(4) model, and appear to have the best forecasting performance overall for both CPI and PPI, with the difference being more substantial when comparing by MSE. Furthermore GLSS-lasso and GLSS-adaptive Lasso do noticeably better than

Table 3.6: Inflation Forecasts: 12 month horizon

	CPI-MSE	CPI-MAE	PPI-MSE	PPI-MAE
AR(4)	1.00	1.00	1.00	1.00
Lasso	.94	.99	.69	.89
Adaptive Lasso	1.08	1.05	.80	.99
SIS-Lasso	.96	.97	.76	.95
SIS-Adaptive Lasso	1.03	1.00	.82	1.00
GLSS-Lasso	.84	.98	.65	.87
GLSS-Adaptive Lasso	.94	1.00	.70	.92
AR(4) + 4 Factors	1.18	.99	1.08	1.09

their SIS based counterparts with the differences being greater when forecasting PPI. We also note that the widely used factor augmented autoregressions do worse than the benchmark model AR(4) model.

3.8 Discussion

In this paper we have analyzed the sure screening properties of SIS in the presence of dependence and heavy tails in the covariate and error processes. In addition, we have proposed a generalized least squares screening (GLSS) procedure, which utilizes the serial correlation present in the data when estimating our marginal effects. Lastly, we analyzed the theoretical properties of the two stage screening and adaptive Lasso estimator using the Lasso as our initial estimator. These results will allow practitioners to apply these techniques to many real world applications where the assumption of light tails and independent observations fails.

There are plenty of avenues for further research, for example extending the theory of model-free screening methods such as distance correlation, or robust measures of dependence such as rank correlation to the setting where we have heavy tails and dependent observations. Other possibilities include extending the theory in this work, or to develop new methodology for long range dependent processes, or certain

classes of non-stationary processes. Long range dependence, is a property which is prominent in a number of fields such as physics, telecommunications, econometrics, and finance (see Samorodnitsky (2006) and references therein). If we assume the error process (ϵ_i) is long range dependent, then by the proof of Theorem 1 in Wu and Pourahmadi (2009) we have $\Delta_{0,q}(\boldsymbol{\epsilon}) = \infty$. A similar result holds for the covariate process, therefore we may need to use a new dependence framework when dealing with long range dependent processes. Lastly, developing new methodology which aims to utilize the unique qualities of time series data such as serial dependence, and the presence of lagged covariates, would be a particularly fruitful area of future research.

3.9 Appendix

3.9.1 Proofs of Results

Proof of Theorem 7.

We first prove part (i), we start by obtaining a bound on:

$$P(|\hat{\rho}_j - \rho_j| > c_2 n^{-\kappa}) \tag{3.29}$$

Let $T_1 = \sum_{t=1}^{n} X_{tj}^2/n$, $T_2 = \sum_{t=1}^{n} X_{tj}Y_t/n$. Then $|\hat{\rho}_j - \rho_j| = |T_2/T_1 - E(T_2)/E(T_1)| =$
$|(T_1^{-1} - E(T_1)^{-1})(T_2 - E(T_2)) + (T_2 - E(T_2))/E(T_1) + (T_1^{-1} - E(T_1)^{-1})E(T_2)|$

Therefore:

$$P(|\hat{\rho}_j - \rho_j| > c_2 n^{-\kappa}) \leq P(|(T_1^{-1} - E(T_1)^{-1})(T_2 - E(T_2))| > c_2 n^{-\kappa}/3) \tag{3.30}$$

$$+ P(|(T_2 - E(T_2))/E(T_1) > c_2 n^{-\kappa}/3|) \tag{3.31}$$

$$+ P(|(T_1^{-1} - E(T_1)^{-1})E(T_2)| > c_2 n^{-\kappa}/3) \tag{3.32}$$

For the RHS of (3.30), we obtain:

$$(3.30) \leq P(|(T_2 - E(T_2))| > C n^{-\kappa/2}) + P(|(T_1^{-1} - E(T_1)^{-1})| > C n^{-\kappa/2}) \tag{3.33}$$

Therefore it suffices to focus on terms (3.31), (3.32). For (3.31), recall that

Recall that $T_2 = \sum_{t=1}^n X_{tj}(\boldsymbol{x}_t\boldsymbol{\beta} + \epsilon_t)/n = \sum_{t=1}^n X_{tj}(\sum_{k=1}^{p_n} X_{tk}\beta_k + \epsilon_t)/n$. Now we let:

$$S_1 = \sum_{t=1}^n X_{tj}(\sum_{k=1}^{p_n} X_{tk}\beta_k)/n \text{ and } S_2 = \sum_{t=1}^n X_{tj}\epsilon_t/n \tag{3.34}$$

By Condition B, $E(X_{tj}\epsilon_t) = 0$, therefore

$$P(|T_2 - E(T_2)| > C n^{-\kappa}) \leq P(|S_1 - E(S_1)| > C n^{-\kappa}/2) + P(|S_2| > C n^{-\kappa}) \tag{3.35}$$

Recall that $\sum_{k=1}^{p_n} \mathbb{1}_{|\beta_k|>0} = s_n$, thus:

$$P\left(|S_1 - E(S_1)| > \frac{c_2 n^{-\kappa}}{2}\right) \leq \sum_{k \in M_*} P\left(|\sum_{t=1}^n \frac{X_{tj}(X_{tk}\beta_k)}{n} - \beta_k E(X_{tj}X_{tk})| > \frac{c_2 n^{-\kappa}}{2 s_n}\right) \tag{3.36}$$

From section 2 in Wu and Wu (2016): $||X_{ij}||_r \leq \Delta_{0,r}(\boldsymbol{X}_j) \leq \Phi_{0,r}(\boldsymbol{x})$. Using this we

compute the cumulative functional dependence measure of $X_{tk}X_{tj}$ as:

$$\sum_{t=m}^{\infty}||X_{tj}X_{tk}-X^*_{tj}X^*_{tk}||_{r/2}\leq\sum_{t=m}^{\infty}(||X_{tj}||_r||X_{tk}-X^*_{tk}||_r+||X_{tk}||_r||X_{tj}-X^*_{tj}||_r)$$
$$\leq 2\Phi_{0,r}(\boldsymbol{x})\Phi_{m,r}(\boldsymbol{x})=O(m^{-\alpha_x}) \tag{3.37}$$

Therefore we obtain: $\sup_m(m+1)^{\alpha_x}\sum_{t=m}^{\infty}||X_{tj}X_{tk}-X^*_{tj}X^*_{tk}||_{r/2}\leq 2K^2_{x,r}$. Combining this with (3.36), and Theorem 2 in Wu and Wu (2016), yields:

$$P\left(|S_1-E(S_1)|>\frac{c_2n^{-\kappa}}{2}\right)\leq Cs_n\left(\frac{n^{\omega}K^r_{x,r}}{(n/s_n)^{r/2-r\kappa/2}}+\exp\left(-\frac{n^{1-2\kappa}}{s^2_nK^4_{x,r}}\right)\right) \tag{3.38}$$

Similarly for $X_{tj}\epsilon_t$, by using Holder's inequality we obtain:

$$\sum_{t=m}^{\infty}||X_{tj}\epsilon_t-X^*_{tj}\epsilon^*_t||_{\tau}\leq\sum_{t=m}^{\infty}(||X_{tj}||_r||\epsilon_t-\epsilon^*_t||_q+||\epsilon_t||_q||X_{tj}-X^*_{tj}||_r)$$
$$\leq\Delta_{0,q}(\boldsymbol{\epsilon})\Phi_{m,r}(\boldsymbol{x})+\Delta_{m,q}(\boldsymbol{\epsilon})\Phi_{0,r}(\boldsymbol{x})=O(m^{-\alpha}) \tag{3.39}$$

Therefore $\sup_m(m+1)^{\alpha}\sum_{t=m}^{\infty}||X_{tj}\epsilon_t-X^*_{tj}\epsilon^*_t||_{\tau}\leq 2K_{x,r}K_{\epsilon,q}$. Using Theorem 2 in Wu and Wu (2016), we obtain:

$$P\left(|S_2|>\frac{c_2n^{-\kappa}}{2}\right)\leq O\left(\frac{n^{\iota}K^{\tau}_{x,r}K^{\tau}_{\epsilon,q}}{n^{\tau-\tau\kappa}}+\exp\left(-\frac{n^{1-2\kappa}}{K^2_{x,r}K^2_{\epsilon,q}}\right)\right) \tag{3.40}$$

For (3.32), assuming $E(X^2_{ij})=O(1)\ \forall j\leq p_n$, and $\max_{j\leq p_n}E(X_{tj}Y_t)<L<\infty$ we obtain:

$$(3.32)\leq P(|T_1-E(T_1)|>T_1Cn^{-\kappa})\leq P(|T_1-E(T_1)|>MCn^{-\kappa})+P(T_1<M) \tag{3.41}$$

We set $M < \min_{j \leq p_n} E(X_{ij}^2) - \epsilon$, for $\epsilon > 0$. We then have:

$$P(T_1 < M) \leq P(|T_1 - E(T_1)| > E(T_1) - M) \tag{3.42}$$

We can then bound the above two equations similar to (3.38). By combining (3.33)(3.35),(3.38),(3.40),(3.41), along with union bound we obtain:

$$P\left(\max_{j \leq p_n} |\hat{\rho}_j - \rho_j| > c_2 n^{-\kappa}\right) \leq O\left(s_n p_n \left[\frac{n^{\omega} K_{x,r}^r}{(n/s_n)^{r/2 - r\kappa/2}} + \exp\left(-\frac{n^{1-2\kappa}}{s_n^2 K_{x,r}^4}\right)\right]\right) + O\left(p_n \left[\frac{n^{\iota} K_{x,r}^{\tau} K_{\epsilon,q}^{\tau}}{n^{\tau - \tau\kappa}} + \exp(-n^{1-2\kappa}/K_{x,r}^2 K_{\epsilon,q}^2)\right]\right)$$

To prove part (ii), we follow the steps in the proof of Theorem 2 in Li et al. (2012a). Let $\mathcal{A}_n = \{\max_{k \in M_*} |\hat{\rho}_k - \rho_k| \leq \frac{c_1 n^{-\kappa}}{2}\}$. On the set $\mathcal{A}_n$, by Condition **A**, we have:

$$|\hat{\rho}_k| \geq |\rho_k| - |\hat{\rho}_k - \rho_k| \geq c_1 n^{-\kappa}/2, \ \forall k \in M_* \tag{3.43}$$

Hence by our choice of γ_n, we obtain $P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) > P(\mathcal{A}_n)$. By applying part (i), the result follows.

For part (iii) we follow the steps in the proof of Theorem 3 in Li et al. (2012a). Using $Var(Y_t), Var(X_{tj}) = O(1)$ for $j \leq p_n$, along with Condition B, we obtain $\sum_{k=1}^{p_n} \rho_k^2 = O(\lambda_{\max}(\Sigma))$. Then on the set $\mathcal{B}_n = \{\max_{k \leq p_n} |\hat{\rho}_k - \rho_k| \leq c_4 n^{-\kappa}\}$, the number of $\{k : |\hat{\rho}_k| > 2c_4 n^{-\kappa}\}$ cannot exceed the number of $\{k : |\rho_k| > c_4 n^{-\kappa}\}$ which is bounded by $O(n^{2\kappa}\lambda_{\max}(\Sigma))$. Therefore, by setting $c_4 = c_3/2$ we obtain:

$$P\left(|\hat{\mathcal{M}}_{\gamma_n}| < O(n^{2\kappa}\lambda_{\max}(\Sigma))\right) > P(\mathcal{B}_n) \tag{3.44}$$

The result then follows from part (i).

□

Proof of Theorem 7.

We follow the steps from the proof of Theorem 6. Let $\boldsymbol{T} = (T_1, \ldots, T_n)$ where $T_i = X_{ij}X_{ik}$, and let $\boldsymbol{R} = (R_1, \ldots, R_n)$ where $R_i = X_{ij}\epsilon_i$. We need to bound the sums: $\sum_{i=1}^{n}(T_i - E(T_i))/n$ and $\sum_{i=1}^{n} R_i/n$.

By Theorem 1 in Wu (2005), $\Theta_q(\boldsymbol{T}) \leq \Delta_{0,q}(\boldsymbol{T})$, and from Section 2 in Wu and Wu (2016): $||X_{ij}||_q \leq \Delta_{0,q}(\boldsymbol{X}_j) \leq \Phi_{0,q}(\boldsymbol{x})$. Additionally, by Holders inequality we have

$$\Delta_{0,q}(\boldsymbol{T}) \leq \sum_{t=0}^{\infty}(||X_{tj}||_{2q}||X_{tk} - X^*_{tk}||_{2q} + ||X_{tk}||_{2q}||X_{tj} - X^*_{tj}||_{2q}) \leq 2\Phi^2_{0,2q}(\boldsymbol{x}) \quad (3.45)$$

Using these, along with Condition D we obtain:

$$\sup_{q\geq 4} q^{-2\tilde{\alpha}_x}\Theta_q(\boldsymbol{T}) \leq \sup_{q\geq 4} q^{-2\tilde{\alpha}_x}\Delta_{0,q}(\boldsymbol{T}) \leq \sup_{q\geq 4} 2q^{-2\tilde{\alpha}_x}\Phi^2_{0,2q}(\boldsymbol{x}) < \infty \quad (3.46)$$

Combining the above and using Theorem 3 in Wu and Wu (2016), we obtain:

$$P\left(\left|\sum_{i=1}^{n} T_i - E(T_i)\right| > \frac{c_2 n^{1-\kappa}}{2}\right) \leq C\exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x^2}\right)^{\tilde{\alpha}} \quad (3.47)$$

Similarly, using the same procedure we obtain:

$$P\left(\left|\sum_{i=1}^{n} R_i\right| > \frac{c_2 n^{1-\kappa}}{2}\right) \leq C\exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x\upsilon_\epsilon}\right)^{\tilde{\alpha}'} \quad (3.48)$$

Now using the above bounds and following the steps in the proof of Theorem 6 we obtain the results.

□

Proof of Lemma 9.

By the proof of Theorem 2 in Wu and Pourahmadi (2009), we have:

$$||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 \leq 2\sum_{i=1}^{l_n} |\hat{\gamma}_{i,k} - \gamma_{i,k}| + 2\sum_{i=l_n+1}^{\infty} |\gamma_{i,k}| \tag{3.49}$$

Recall that $\hat{\rho}_k$ is the OLS estimate of the marginal projection, by (3.14) we have $\hat{\epsilon}_{t,k} = \epsilon_{t,k} - X_{tk}(\hat{\rho}_k - \rho_k) = \epsilon_{t,k} - X_{tk}(\frac{\sum_{j=1}^n X_{jk}\epsilon_{j,k}/n}{\sum_{j=1}^n X_{jk}^2/n})$. Which gives us:

$$\hat{\gamma}_{i,k} = \frac{1}{n}\sum_{t=1}^{n-|i|}\Bigg[\epsilon_{t,k}\epsilon_{t+|i|,k} - \epsilon_{t,k}X_{t+|i|,k}\left(\sum_{j=1}^n X_{jk}\epsilon_{j,k}/n\right) \tag{3.50}$$

$$- \epsilon_{t+|i|,k}X_{tk}\left(\sum_{j=1}^n X_{jk}\epsilon_{j,k}/n\right) + X_{tk}X_{t+|i|,k}\left(\sum_{j=1}^n X_{jk}\epsilon_{j,k}/n\right)^2\Bigg] \tag{3.51}$$

By Condition E and $l_n = c\log(n)$, for sufficiently large c, we have: $\sum_{i=l_n+1}^{\infty} |\gamma_{i,k}| = o(n^{-\kappa})$, so we focus on the term $\sum_{i=1}^{l_n} |\hat{\gamma}_{i,k} - \gamma_{i,k}|$ in (3.49). We then have:

$$P(||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 > cn^{-\kappa}) \leq \sum_{i=1}^{l_n} P\left(|\hat{\gamma}_{i,k} - \gamma_{i,k}| > cn^{-\kappa}/l_n\right) \tag{3.52}$$

And

$$P(|\hat{\gamma}_{i,k}-\gamma_{i,k}|>\frac{cn^{-\kappa}}{l_n}) \leq P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}\epsilon_{t+|i|,k}-E(\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}\epsilon_{t+|i|,k})\right| \right. \tag{3.53}$$

$$\left. + \left|E(\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}\epsilon_{t+|i|,k})-\gamma_{i,k}\right| > cn^{-\kappa}/4l_n\right) \tag{3.54}$$

$$+P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k}\left(\frac{\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n}{\sum_{j=1}^{n}X_{jk}^2/n}\right)\right| > cn^{-\kappa}/4l_n\right) \tag{3.55}$$

$$+P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t+|i|,k}X_{tk}\left(\frac{\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n}{\sum_{j=1}^{n}X_{jk}^2/n}\right)\right| > cn^{-\kappa}/4l_n\right) \tag{3.56}$$

$$+P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}X_{tk}X_{t+|i|,k}\left(\frac{\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n}{\sum_{j=1}^{n}X_{jk}^2/n}\right)^2\right| > cn^{-\kappa}/4l_n\right) \tag{3.57}$$

For (3.54), the bias $|E(\sum_{t=1}^{n-|i|}\frac{\epsilon_{t,k}\epsilon_{t+|i|,k}}{n}-\gamma_{i,k}| \leq \frac{i\gamma_{i,k}}{n}$. Using the techniques in the proof of Theorem 6 we can then bound (3.53). For (3.55) we have:

$$(3.55) \leq P\left(\frac{|\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n|}{\sum_{j=1}^{n}X_{jk}^2/n} > cn^{-\kappa}/Ml_n\right)+P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k}\right| > M\right) \tag{3.58}$$

And $P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k}\right| > M\right)$

$$\leq P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k}-E(\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k})\right| > M-\left|E(\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k})\right|\right) \tag{3.59}$$

And we set $M > \max_{k\leq p_n}\max_{i\leq l_n}2|E(\epsilon_{t,k}X_{t+|i|,k})|+\epsilon$, for some $\epsilon>0$. Similarly we have $P(\frac{|\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n|}{\sum_{j=1}^{n}X_{jk}^2/n} > cn^{-\kappa}/Ml_n)$

$$\leq P\left(\left|\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n\right| > M_1Cn^{-\kappa}/l_n\right)+P\left(\sum_{j=1}^{n}X_{jk}^2/n < M_1\right) \tag{3.60}$$

Where we set $M_1 < \min_{j \leq p_n} E(X_{ij}^2) - \epsilon$, for $\epsilon > 0$. The same method we used for (3.53) can be applied to (3.56), (3.57). Using the techniques in the proof of Theorem 6, and (3.52), we obtain the result. For (ii), we follow the same procedure as in (i), and apply the methods seen in the proof of Theorem 7.

□

Proof of Theorem 8.

For (i), as before we start with a bound on: $P(|\hat{\beta}_k^M - \beta_k^M| > c_2 n^{-\kappa})$. Using Condition E, we can write:

$$\beta_k^M = (E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{X}_k)/n)^{-1} E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{y}^k/n) + O(1/n)$$

After combining this with (3.14), it suffices to obtain a bound for:

$$P(|(\boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{X}_k/n)^{-1} \boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{\epsilon}^k/n - (E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{X}_k))^{-1} E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{\epsilon}^k)| > c n^{-\kappa}) \tag{3.61}$$

Similar to the proof of Theorem 6 we let $T_1 = \boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{X}_k/n$,
$T_2 = \boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{\epsilon}^k/n$, $T_3 = E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{X}_k)$, and $T_4 = E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{\epsilon}^k)$. Then:

$$\begin{aligned} |\hat{\beta}_k^M - \beta_k^M| = |T_2/T_1 - T_4/T_3| = |(T_1^{-1} - T_3^{-1})(T_2 - T_4) \\ + (T_2 - T_4)/T_3 + (T_1^{-1} - T_3^{-1})T_4| \end{aligned} \tag{3.62}$$

Following the steps in the proof of Theorem 6, it suffices to focus on the terms:

$$P(|T_1 - T_3| > c n^{-\kappa}) \text{ and } P(|T_2 - T_4| > c n^{-\kappa}) \tag{3.63}$$

We then have:

$$P(|T_2 - T_4| > Cn^{-\kappa}) \leq P(|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/n| > Cn^{-\kappa}/2) + P(|\boldsymbol{X}_k^T\Sigma_k^{-1}\boldsymbol{\epsilon}^k/n - E(\boldsymbol{X}_k^T\Sigma_k^{-1}\boldsymbol{\epsilon}^k)| > Cn^{-\kappa}/2) \tag{3.64}$$

We first deal with the term $\boldsymbol{X}_k^T\Sigma_k^{-1}\boldsymbol{\epsilon}^k/n$. We can rewrite this term as $\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{\epsilon}}^k/n$, where $\tilde{\boldsymbol{X}}_k = V_k\boldsymbol{X}_k, \tilde{\boldsymbol{\epsilon}}^k = V_k\boldsymbol{\epsilon}^k/n$, V_k is a lower triangle matrix and the square root of Σ_k^{-1}. Ignoring the first L_k observations, we can express:

$$\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{\epsilon}}^k/n = \sum_{t=L_k+1}^{n}\left(\epsilon_{t,k} - \sum_{i=1}^{L_k}\alpha_{i,k}\epsilon_{t-i,k}\right)\left(X_{t,k} - \sum_{i=1}^{L_k}\alpha_{i,k}X_{t-i,k}\right) \tag{3.65}$$

, where $(\alpha_{1,k}, \ldots, \alpha_{L_k,k})$ are the autoregressive coefficients of the process $\epsilon_{t,k}$.

We compute the cumulative functional dependence measure of $\tilde{X}_{t,k}\tilde{\epsilon}_{t,k}$ as:

$$\sum_{l=m}^{\infty}||\tilde{X}_{l,k}\tilde{\epsilon}_{l,k} - \tilde{X}_{l,k}^*\tilde{\epsilon}_{l,k}^*||_{\tau'} \leq \sum_{l=m}^{\infty}(||\tilde{X}_{l,k}||_r||\tilde{\epsilon}_{l,k} - \tilde{\epsilon}_{l,k}^*||_{q'} + ||\tilde{\epsilon}_{l,k}||_{q'}||\tilde{X}_{l,k} - \tilde{X}_{l,k}^*||_r) \tag{3.66}$$

We have: $||\tilde{X}_{l,k} - \tilde{X}_{l,k}^*||_r \leq ||X_{l,k} - X_{l,k}^*||_r + \sum_{i=1}^{L_k}|\alpha_i|||X_{k,l-i} - X_{k,l-i}^*||_r$. And by our assumptions $||\tilde{\epsilon}_{l,k} - \tilde{\epsilon}_{l,k}^*||_{q'} = 0$, for $l > 0$. From which we obtain:

$$\sum_{l=m}^{\infty}||\tilde{X}_{l,k}\tilde{\epsilon}_{l,k} - \tilde{X}_{l,k}^*\tilde{\epsilon}_{l,k}^*||_{\tau'} \leq C\Phi_{m,r} = O(m^{-\alpha_x}) \tag{3.67}$$

Using Theorem 2 in Wu and Wu (2016):

$$P(|\boldsymbol{X}_k^T\Sigma_k^{-1}\boldsymbol{\epsilon}^k/n - E(\boldsymbol{X}_k^T\Sigma_k^{-1}\boldsymbol{\epsilon}^k)| > Cn^{-\kappa}) \leq O\left(\frac{n^{\iota}K_{x,r}^{\tau'}}{n^{\tau'-\tau'\kappa}}\right) \tag{3.68}$$

For the term $|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/n|$, using Cauchy-Schwarz inequality:

$$\frac{|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/n|}{||\boldsymbol{X}_k||_2||\boldsymbol{\epsilon}^k||_2} \leq \frac{||(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k||_2}{n||\boldsymbol{\epsilon}^k||_2} \leq \frac{||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2}{n} \tag{3.69}$$

Using (3.69) we obtain:

$$P(|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/n| > Cn^{-\kappa}) \leq P(||\boldsymbol{X}_k||_2||\boldsymbol{\epsilon}^k||_2||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2/n > Cn^{-\kappa}) \tag{3.70}$$

Where the right hand side of (3.70) is:

$$\leq P(||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2 > Cn^{-\kappa}/\sqrt{M}) + P\left(\left(\sum_{i=1}^{n} X_{ik}^2/n\right)\left(\sum_{i=1}^{n} \epsilon_{i,k}^2/n\right) > M\right) \tag{3.71}$$

Let $M = M_1M_2$, where $M_1 \geq \max_{k\leq p_n} E(X_{i,k}^2) + \epsilon$, and $M_2 = \max_{k\leq p_n} E(\epsilon_{i,k}^2) + \epsilon$, for some $\epsilon > 0$. The second term of (3.71) is:

$$\leq P\left(\sum_{i=1}^{n} X_{ik}^2/n > M_1\right) + P\left(\sum_{i=1}^{n} \epsilon_{i,k}^2/n > M_2\right) \tag{3.72}$$

We can bound the above using the same techniques as in the previous proofs.

By Condition E, the spectral density of the process $\epsilon_{t,k}, \forall k \leq p_n$ is bounded away from zero and infinity. Therefore, $0 < C_1 \leq \lambda_{min}(\Sigma_k) \leq \lambda_{max}(\Sigma_k) \leq C_2 < \infty, \forall k \leq p_n$ Wu and Pourahmadi (2009). We then use:

$$\begin{aligned} \lambda_{\min}(\Sigma_k)||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2 &\leq ||\Sigma_k^{\frac{1}{2}}(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\Sigma_k^{\frac{1}{2}}||_2 \\ &= ||\Sigma_k^{\frac{1}{2}}\hat{\Sigma}_{k,l_n}^{-1}\Sigma_k^{\frac{1}{2}} - I_n||_2 \end{aligned} \tag{3.73}$$

Let $a_1 \geq a_2 \geq \ldots \geq a_n$ be the ordered eigenvalues of $\Sigma_k^{-\frac{1}{2}}\hat{\Sigma}_{k,l_n}\Sigma_k^{-\frac{1}{2}}$, therefore

$||\Sigma_k^{\frac{1}{2}}\hat{\Sigma}_{k,l_n}^{-1}\Sigma_k^{\frac{1}{2}} - I_n||_2 = \max_i|\frac{1}{a_i} - 1| = \max_i|\frac{a_i - 1}{a_i}|$. We then have

$$\max_i |a_i - 1| = ||\Sigma_k^{-\frac{1}{2}}\hat{\Sigma}_{k,l_n}\Sigma_k^{-\frac{1}{2}} - I_n||_2 \leq \lambda_{\max}(\Sigma_k^{-1})||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 \tag{3.74}$$

Let $a_j = argmin_{a_i}|a_i^{-1}|$, using this and (3.73),(3.74) we obtain:

$$\begin{aligned} P(||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2 > Cn^{-\kappa}) &\leq P(||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 > Ca_jn^{-\kappa}) \\ &\leq P(||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 > CM_3n^{-\kappa}) + P(|a_j| < M_3) \end{aligned} \tag{3.75}$$

Where $M_3 \in (0, 1-\epsilon)$ for $\epsilon > 0$. We then have

$$P(|a_j| < M_3) \leq P(|a_j - 1| > 1 - M_3) \leq P(||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 > 1 - M_3)$$

Combining the above with (3.75) and Lemma 9, we obtain:

$$P(||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2 > Cn^{-\kappa}) \leq l_nO\left(\frac{n^\iota l_n^{\tau'}K_{x,r}^{\tau'}\tilde{K}_{\epsilon,q'}^{\tau'}}{n^{\tau'-\tau'\kappa}} + \frac{n^\zeta l_n^{q'/2}\tilde{K}_{\epsilon,q'}^{q'}}{n^{q'/2-q'\kappa/2}} + \frac{n^\omega l_n^{r/2}K_{x,r}^r}{n^{r/2}}\right) \tag{3.76}$$

By (3.64),(3.68),(3.70),(3.72),(3.76) we obtain a bound for $P(|T_2 - E(T_2)| > Cn^{-\kappa})$. For the term $P(|T_1 - E(T_1)| > Cn^{-\kappa})$, we proceed in a similar fashion:

$$\begin{aligned} P(|T_1 - E(T_1)| > Cn^{-\kappa}) &\leq P(|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{X}^k/n| > Cn^{-\kappa}/2) \\ &\quad + P(|\boldsymbol{X}_k^T\Sigma_k^{-1}\boldsymbol{X}^k/n - E(\boldsymbol{X}_k^T\Sigma_k^{-1}\boldsymbol{X}^k)| > Cn^{-\kappa}/2) \end{aligned}$$

We can then obtain a bound on the above terms by following a similar procedure as before. Combining these gives us the result for (i). For (ii), using the result from (i) we follow a similar procedure to the proof of Theorem 6. For (iii) and (iv) we follow the same procedure as (i) and (ii), and apply the methods seen in the proof of

Theorem 7; we omit the details.

□

Proof of Corollary 9.

Recall that:

$$\beta_k^M = E(y_t - \sum_{i=1}^{L_k} \alpha_i y_{t-i})(X_{t,k} - \sum_{i=1}^{L_k} \alpha_i X_{t-i,k})/(E(X_{t,k} - \sum_{i=1}^{L_k} \alpha_i X_{t-i,k})^2)$$

Therefore by our assumption, we have that $\beta_k^M \propto \rho_k$ whenever $\beta_k^M > 0$. Using this we obtain $\sum_{k=1}^{p_n} (\beta_k^M)^2 = O(\sum_{k=1}^{p_n} \rho_k^2) = O(\lambda_{\max}(\Sigma))$. We obtain the result, by following the procedure in the proof of Theorem 6 and using the results from Theorem 8.

□

Proof of Theorem 10.

For simplicity we only prove part (i), the proof for part (ii) follows similarly. We will work on the following set $\mathcal{D}_n = \mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n$, where

$$\begin{aligned}
\mathcal{A}_n &= \{max_{k \le p_n} |\hat{\rho}_k - \rho_k| \le c_3 n^{-\kappa}/2\} \\
\mathcal{B}_n &= \{max_{i,j \le d_n'} |[\Sigma_{\mathcal{M}_{\frac{\gamma_n}{2}}} - \hat{\Sigma}_{\mathcal{M}_{\frac{\gamma_n}{2}}}]_{i,j}| \le \frac{\phi_0}{16 s_n}\} \\
\mathcal{C}_n &= \{max_{k \le d_n'} |\sum_{i=1}^{n} X_{ik}\epsilon_i| \le \lambda_n n^{\psi/2}\}
\end{aligned}$$

On the set $\mathcal{A}_n$, if we apply screening as a first stage procedure, by our choice of γ_n, we obtain:

$$\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n} \subset \mathcal{M}_{\gamma_n/2} \tag{3.77}$$

Next we need to use Lemma 7 and 8 in Medeiros and Mendes (2016), specifically we

need to show our reduced model satisfies conditions DGP 3,DESIGN, and WEIGHTS in Medeiros and Mendes (2016). On the set $\mathcal{B}_n$, by Lemma 1 in Medeiros and Mendes (2016), we have $\phi_{\Sigma_{\hat{\mathcal{M}}_{\gamma_n/2}}} = \phi_{\Sigma_{\mathcal{M}_{\gamma_n/2}}} = \phi_0$. Therefore, we have:

$$\phi_{\Sigma_{\hat{\mathcal{M}}_{\gamma_n}}} = \min_{S\subseteq\{1,\ldots,d_n\},|S|\leq s_n} \min_{\boldsymbol{v}\neq 0,|\boldsymbol{v}_{S^c}|\leq 3|\boldsymbol{v}_S|} \frac{\boldsymbol{v}^T\Sigma_{\hat{\mathcal{M}}_{\gamma_n}}\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}} \geq \phi_0 \tag{3.78}$$

Using this along with Lemma 1 in Medeiros and Mendes (2016) and Condition J, we have that DESIGN 3a is satisfied with $\phi_{\min} = \phi_0/16$, where $\inf_{v^Tv=1} v^T\Sigma_{11}v > 2\phi_{\min} > 0$, and Σ_{11} is the covariance matrix of the relevant predictors. On the set $\mathcal{D}_n$, by Conditions K and L in our work, and Lemma 2 and proposition 1 in Medeiros and Mendes (2016), assumption WEIGHTS is satisfied. On the set $\mathcal{A}_n \cap \mathcal{B}_n$, DGP 3 and DESIGN 3b are satisfied, while DESIGN 2 is satisfied by Condition L.

Now by proposition 2, Lemmas 7 and 8 in Medeiros and Mendes (2016) we obtain:

$$P(sgn(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}}) = sgn(\boldsymbol{\beta})) \geq P(\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n) \geq 1 - P(\mathcal{A}_n^{\complement}) - P(\mathcal{B}_n^{\complement}) - P(\mathcal{C}_n^{\complement}) \tag{3.79}$$

$P(\mathcal{A}_n^{\complement})$ is given in Theorem 6 part i. For $P(\mathcal{B}_n^{\complement})$ using the method in the proof for Theorem 6, we obtain:

$$P(\mathcal{B}_n^{\complement}) \leq d_n'^2 O\left(\frac{n^{\omega}K_{x,r}^r}{n^{r/2}} + \exp\left(-n/K_{x,r}^4\right)\right) \tag{3.80}$$

And for $P(\mathcal{C}_n^{\complement})$:

$$P(\mathcal{C}_n^{\complement}) \leq d_n' O\left(\frac{n^{\iota}K_{x,r}^{\tau}K_{\epsilon,q}^{\tau}}{\lambda_n^{\tau}n^{\tau\psi/2}} + \exp\left(-\lambda_n^2 n^{\psi-1}/K_{x,r}^2K_{\epsilon,q}^2\right)\right) \tag{3.81}$$

To prove part ii) we follow the same steps from part i). We obtain $P(\mathcal{A}_n^{\complement})$, $P(\mathcal{B}_n^{\complement}), P(\mathcal{C}_n^{\complement})$ by following the method in the proof of Theorem 7, and using The-

orem 3 in Wu and Wu (2016). □

3.9.2 Asymptotic Distribution of GLS estimator

Lemma 10. *Assume conditions E,F,G,H hold, then* $\sqrt{n}(\hat{\beta}_k^M - \beta_k^M)$ *and* $\sqrt{n}(\tilde{\beta}_k^M - \beta_k^M)$ *have the same asymptotic distribution.*

Proof of Lemma 10. It is clear that sufficient conditions for the feasible GLS estimator $\hat{\beta}_k^M$, and $\tilde{\beta}_k^M$ to have the same asymptotic distribution are Davidson and MacKinnon (2004):

$$\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/\sqrt{n} \to 0$$
$$\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{X}_k/n \to 0$$

By the proof of theorem 8, both these conditions are satisfied, therefore $\hat{\beta}_k^M$, and $\tilde{\beta}_k^M$ have the same asymptotic distribution. □

We use the above lemma, and rely on the asymptotic distribution of $\tilde{\beta}_k^M$ to provide an explanation for the superior performance of GLSS, and its robustness to increasing levels of serial correlation in $\epsilon_{t,k}$. We deal with three cases, and we assume an AR(1) process for the errors for simplicity and ease of presentation. The results can be generalized to AR(p) processes, by using the moving average representation of $\epsilon_{t,k}$:

Case 1:

We start with the setting used in figure 3.1, assume $x_{t,k}$ is iid and $\epsilon_{t,k} = \alpha\epsilon_{t-1,k} + e_t$, with $x_{t,k}$, and $\epsilon_{t,k}$ being independent $\forall t$. Using Gordin's central limit theorem Hayashi (2000), we calculate the asymptotic distribution of $\sqrt{n}(\tilde{\beta}_k^M - \beta_k^M) \to N(0, J)$, where

$J = \frac{\sigma_e^2}{\sigma_{x_k}^2(1+\alpha^2)}$, $\sigma_e^2 = var(e_t)$, and $\sigma_{x_k}^2 = var(x_{t,k})$. Using the same methods we calculate the asymptotic distribution of the marginal OLS estimator as $\sqrt{n}(\hat{\rho}_k - \rho_k) \to N(0, V)$, where $V = \frac{\sigma_e^2}{\sigma_{x_k}^2(1-\alpha^2)}$. Therefore the variance of the OLS estimator increases without bound as α increases towards 1. Whereas the variance of the GLS estimator actually decreases as α increases.

Case 2:

We expand this to the case when $x_{t,k}$ is temporally dependent, for simplicity we let $x_{t,k} = \phi x_{t-1,k} + \eta_t$. We still assume $x_{j,k}$ and ϵ_t are independent $\forall j, t$, and $\epsilon_{t,k} = \alpha\epsilon_{t-1,k} + e_t$. This is the setting for the first model in the simulations section. Using Gordin's central limit theorem, and elementary calculations: $\sqrt{n}(\tilde{\beta}_k^M - \beta_k^M) \to N(0, J)$, where $J = \frac{(1-\phi^2)\sigma_e^2}{(1+\alpha^2-2\phi\alpha)\sigma_\eta^2}$. And for the marginal OLS estimator $\sqrt{n}(\hat{\rho}_k - \rho_k) \to N(0, V)$, where $V = \frac{(1+\phi^2)\sigma_e^2}{(1-\alpha^2)\sigma_\eta^2}$. We clearly see that for fixed ϕ, the GLS estimate is robust to increasing α, whereas the variance of the OLS estimator increases without bound as α increases towards 1. This sensitivity to α provides an explanation for the results seen in case 1 of the simulations, which show the performance of SIS severely deteriorates for high levels of serial correlation in $\epsilon_{t,k}$

Case 3:

In both the previous cases, it is easy to see the GLS estimator is asymptotically efficient to the OLS estimator. For the case where $\boldsymbol{X}_k = (x_{t,k}, t = 1, \ldots, n)$ and $\boldsymbol{\epsilon}^k = (\epsilon_{t,k}, t = 1, \ldots, n)$ are dependent on each other, it is more complicated. In this setting, it is likely the case that $\rho_k \neq \beta_k^M$. Assume $\epsilon_{t,k} = \alpha\epsilon_{t-1,k} + e_t$, and let $x_{t,k} - \alpha x_{t-1,k} = \tilde{x}_{t,k}$, and $W_1 = \sum_{i=-\infty}^{\infty} \gamma(i)$, where $\gamma(i) = cov(\tilde{x}_{t,k}e_t, \tilde{x}_{t-i,k}e_{t-i})$. We

start by examining the asymptotic distribution of $\sqrt{n}(\tilde{\beta}_k^M - \beta_k^M) \to N(0, J)$, where $J = W_1/(var(\tilde{x}_{t,k}))^2$. By the proof of theorem 1 in Wu and Pourahmadi (2009), $W_1 \leq (\sum_{t=0}^{\infty} \delta_2(\tilde{x}_{t,k} e_t))^2$, which gives us:

$$J \leq \frac{(\sum_{t=0}^{\infty} \delta_2(\tilde{x}_{t,k}\epsilon_t))^2}{var(\tilde{x}_{t,k})^2} \leq \left(\frac{2||e_t||_4 \Delta_{0,4}(\tilde{\boldsymbol{X}}_k)}{var(\tilde{x}_{t,k})} \right)^2$$

Where the last inequality follows from: $\delta_2(\tilde{x}_{t,k}e_t) = ||e_0||_4||\tilde{x}_{t,k} - \tilde{x}^*_{t,k}||_4 + ||\tilde{x}_{0,k}||_4||e_t - e_t*||_4$. Since e_t is iid $||e_t - e_t*||_4 = 0, \forall t > 0$. If we assume, $x_{t,k} = \phi x_{t-1,k} + \eta_t$, by writing $\tilde{x}_{t,k} = \eta_t + (\phi - \alpha)x_{t-1,k}$, we have:

$$J \leq \left(\frac{2||e_t||_4||\eta_t||_4|\phi - \alpha|}{(1 - |\phi|)var(\tilde{x}_{t,k})} + \frac{2||e_t||_4||\eta_t||_4}{var(\tilde{x}_{t,k})} \right)^2$$

From these results we see that the asymptotic variance of the GLS estimator is bounded when α increases towards 1, and is largely robust to increasing levels of serial correlation in $\epsilon_{t,k}$. This result seems to provide an explanation for GLSS being robust to increasing levels of serial correlation in our simulations.

For the OLS estimator we obtain, $(\hat{\rho}_k - \rho_k) \to N(0, V)$, where $V = W_2/(var(x_{t,k}))^2$ and $W_2 = \sum_{i=-\infty}^{\infty} cov(x_{t,k}\epsilon_t, x_{t-i}\epsilon_{t-i})$. As before, we can bound:

$$V \leq \frac{(\sum_{t=0}^{\infty} \delta_2(x_{t,k}e_t))^2}{var(x_{t,k}))^2} \leq \left(\frac{||\epsilon_{t,k}||_4 \Delta_{0,4}(\boldsymbol{X}_k)}{var(x_{t,k})} + \frac{2||X_{k,t}||_4||e_t||_4}{(1 - |\alpha|)var(x_{t,k})} \right)^2$$

We see the above bound is very sensitive to increasing serial correlation in $\epsilon_{t,k}$. Although this is an upper bound to the asymptotic variance, it seems to explain the deterioration in performance of SIS when increasing the serial correlation of $\epsilon_{t,k}$ in our simulations.

Bibliography

Amemiya, T. (1973). Generalized least squares with an estimated autocovariance matrix. *Econometrica*, 41(4):723–732.

An, H. and Huang, F. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statistica Sinica*, 6:943–956.

Andrews, D. (1984). Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21(4):930–934.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

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

Bai, J. and Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, 24(4):607–629.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Barut, E., Fan, J., and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43:1535–1567.

Basu, S., Shojaie, A., and Michailidis, G. (2015). Network granger causality with inherent grouping structure. *Journal of Machine Learning Research*, 16:417–453.

Bates, B. J., Plagborg-Møller, M., Stock, J. H., and Watson, M. W. (2013). Consistent factor estimation in dynamic factor models with structural instability. *Journal of Econometrics*, 177(2):289–304.

Bickel, P.J.and Brown, B., Huang, H., and Li, Q. (2009a). An overview of recent developments in genomics and associated statistical methods . *Phil. Transactions of the Roy. Soc. A*, 367:4313–4337.

Bickel, P. (2008). Discussion of Sure independence screening for ultrahigh dimensional feature space . *J.Roy. Statist. Soc. B.*, 70:883–884.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009b). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37:1705–1732.

Borisov, I. and Volodko, N. (2009). Exponential inequalities for the distributions of canonical u-and v-statistics of dependent observations. *Siberian Advances in Mathematics*, 19(1):1–12.

Breitung, J. and Eickmeier, S. (2011). Testing for structural breaks in dynamic factor models. *Journal of Econometrics*, 163(1):71–84.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer.

Bühlmann, P. (1995). Moving-average representation of autoregressive approximations. *Stochastic Processes and their Applications*, 60(2):331 – 342.

Buhlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.*, 2(34):559–583.

Buhlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505.

Buhlmann, P. and Van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136(1):163–188.

Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, 18(1):17–39.

Carrasco, M. and Rossi, B. (2016). In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics*, 34(3):313–338.

Casini, A. and Perron, P. (2018). Structural breaks in time series. *arXiv preprint arXiv:1805.03807*.

Chan, J. C., Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2012). Time varying dimension models. *Journal of Business & Economic Statistics*, 30(3):358–367.

Chang, J., Tang, C. Y., and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.*, 41(3):2123–2148.

Chen, B. (2015). Modeling and testing smooth structural changes with endogenous regressors. *Journal of Econometrics*, 185(1):196–215.

Chen, B. and Hong, Y. (2012). Testing for smooth structural changes in time series models via nonparametric regression. *Econometrica*, 80(3):1157–1183.

Chen, J., Li, D., Linton, O., and Lu, Z. (2017). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, In Press.

Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.*, 41:2994–3021.

Cheng, M.-Y., Honda, T., Li, J., and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *Ann. Statist.*, 42(5):1819–1849.

Clements, M. P. and Hendry, D. F. (1996). Intercept corrections and structural change. *Journal of Applied Econometrics*, 11(5):475–494.

Cogley, T. and Sargent, T. J. (2001). Evolving post-world war ii US inflation dynamics. *NBER macroeconomics annual*, 16:331–373.

Dahlhaus, R. (1996). On the kullback-leibler information divergence of locallystationary processes. *Stochastic processes and their applications*, 62(1):139–168.

Dahlhaus, R. (2012). Locally stationary processes. In *Handbook of statistics*, volume 30, pages 351–413. Elsevier.

Dahlhaus, R. et al. (1997). Fitting time series models to nonstationary processes. *The annals of Statistics*, 25(1):1–37.

Dahlhaus, R., Richter, S., and Wu, W. B. (2018). Towards a general theory for non-linear locally stationary processes. *Bernoulli*. In press.

Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106(1):157–181.

Davidson, J. (1994). *Stochastic Limit Theory, An Introduction for Econometricians*. Oxford University Press.

Davidson, R. and MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford University Press.

Davis, R. A., Holan, S. H., Lund, R., and Ravishanker, N., editors (2016a). *Handbook of Discrete-Valued Time Series*. CRC Press.

Davis, R. A., Matsui, M., Mikosch, T., and Wan, P. (2016b). Applications of distance correlation to time series. *arXiv preprint arXiv:1606.05481*.

Ding, X., Qiu, Z., Chen, X., et al. (2017). Sparse transition matrix estimation for high-dimensional and locally stationary vector autoregressive models. *Electronic Journal of Statistics*, 11(2):3871–3902.

Doukhan, P. (1994). *Mixing: Properties and Examples*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag New York.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Eickmeier, S., Lemke, W., and Marcellino, M. (2015). Classical time varying factor-augmented vector auto-regressive modelsâestimation, forecasting and structural analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):493–533.

Fan, J., Feng, Y., and Song, R. (2011a). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, 106:544–557.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for cox's proportional hazards model. *IMS Collections, Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, 6:70–86.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space w/ discussion. *J.Roy. Statist. Soc. B.*, 70:849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space . *Statistica Sinica*, 20:101–148.

Fan, J., Lv, J., and Qi, L. (2011b). Sparse High dimensional models in economics. *Annual Review of Economics*, 3:291–317.

Fan, J., Lv, J., and Qi, L. (2011c). Sparse high-dimensional models in economics.

Fan, J., Ma, J., and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284.

Fan, J. and Ren, Y. (2006). Statistical Analysis of DNA Microarray Data in Cancer Research . *Clinical Cancer Research*, 12:4469–4473.

Fan, J. and Song, R. (2010). Sure Independence Screening in generalized linear models with NP-dimensionality . *Annals of Statistics*, 38:3567–3604.

Fan, J., Xue, L., and Yao, J. (2017). Sufficient forecasting using factor models. *Journal of econometrics*, 201(2):292–306.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high-dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, 75(3):531–552.

Feng, Y., Wu, Y., and Stefanski, L. A. (2017). Nonparametric independence screening via favored smoothing bandwidth. *Journal of Statistical Planning and Inference*. In press.

Fokianos, K. and Pitsillou, M. (2017). Consistent testing for pairwise dependence in time series. *Technometrics*, 59(2):262–270.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). Discussion of boosting papers. *Ann. Statist*, 32:102–107.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Giraitis, L., Kapetanios, G., and Price, S. (2013). Adaptive forecasting in the presence of recent and ongoing structural change. *Journal of Econometrics*, 177(2):153–170.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):217–245.

Goyal, A. and Welch, I. (2003). Predicting the equity premium with dividend ratios. *Management Science*, 49(5):639–654.

Groen, J. J., Paap, R., and Ravazzolo, F. (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics*, 31(1):29–44.

Gu, S., Kelly, B., and Xiu, D. (2018). Empirical asset pricing via machine learning. Working Paper 25398, National Bureau of Economic Research.

Gu, S., Kelly, B. T., and Xiu, D. (2019). Autoencoder asset pricing models. *Available at SSRN*.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384.

Han, F. (2016). An exponential inequality for u-statistics under mixing conditions. *Journal of Theoretical Probability*.

Han, Y. and Tsay, R. S. (2017). High-dimensional linear regression for dependent observations with application to nowcasting. *arXiv preprint arXiv:1706.07899*.

Hastie, T., Taylor, J., Tibshirani, R., and Walther, G. (2007). Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29.

Hayashi, F. (2000). *Econometrics*. Princeton Univ Press.

Hepp, T., Schmid, M., Gefeller, O., Waldmann, E., and Mayr, A. (2016). Approaches to regularized regression–a comparison between gradient boosting and the lasso. *Methods of information in medicine*, 55(05):422–430.

Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M. (2014). Model-based boosting in r: a hands-on tutorial using the r package mboost. *Computational statistics*, 29(1-2):3–35.

Hu, L., Huang, T., and You, J. (2018). Estimation and identification of a varying-coefficient additive model for locally stationary processes. *Journal of the American Statistical Association*. In Press.

Huang, J., Ma, S., and Zhang, C. (2008). Adaptive Lasso for sparse high-dimensional regression models . *Statistica Sinica*, 18:1603–1618.

Huang, Q. and Zhu, Y. (2016). Model-free sure screening via maximum correlation. *Journal of Multivariate Analysis*, 148:89 – 106.

Inoue, A., Jin, L., and Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics*, 196(1):55–67.

Johnstone, I.M.and Tetterington, M. (2009). Statistical challenges of high dimensional data . *Phil. Transactions of the Roy. Soc. A*, 367:4237–4253.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring uncertainty. *American Economic Review*, 105(3):1177–1216.

Kelly, B. and Pruitt, S. (2013). Market expectations in the crossâsection of present values. *The Journal of Finance*, 68(5):1721–1756.

Kelly, B. and Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294 – 316. High Dimensional Problems in Econometrics.

Kilian, L. and Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge University Press.

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

Kock, A. and Callot, A. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186:325 – 344.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter vars. *Journal of Econometrics*, 177(2):185–198.

Koop, G. M. (2013). Forecasting with medium and large bayesian vars. *Journal of Applied Econometrics*, 28(2):177–203.

Koreisha, S. G. and Fang, Y. (2001). Generalized least squares with misspecified serial correlation structures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, 63(3).

Lee, E. R., Mammen, E., et al. (2016). Local linear smoothing for sparse high dimensional varying coefficient models. *Electronic Journal of Statistics*, 10(1):855–894.

Li, G., Peng, H., Jun, Z., and Zhu, L. (2012a). Robust rank correlation based screening . *Annals of Statistics*, 40:1846–1877.

Li, R., Zhu, L., and Zhong, W. (2012b). Feature Screening via distance correlation. . *J. Amer. Statist. Assoc.*, 107:1129–1139.

Liu, J., Zhong, W., and Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Sci. China Math.*, 58:1–22.

Liu, W., Xiao, H., and Wu, W. B. (2013). Probability and moment inequalities under dependence. *Statistica Sinica*, 23:1257–1272.

Liu, Y. and Wang, Q. (2017). Model-free feature screening for ultrahigh-dimensional data conditional on some variables. *Annals of the Institute of Statistical Mathematics*, pages 1–19.

Lu, Z. (1998). On the geometric ergodicity of a non-linear autoregressive model with an autoregressive conditional heteroscedastic term. *Statistica Sinica*, (1205-1217).

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Lutz, R. W. and Bühlmann, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statistica Sinica*, 16(2):471–494.

Ma, S., Li, R., and Tsai, C.-L. (2017). Variable screening via quantile partial correlation. *Journal of the American Statistical Association*, 112(518):650–663.

Mai, Q., Zou, H., et al. (2015). The fused kolmogorov filter: a nonparametric model-free screening method. *The Annals of Statistics*, 43(4):1471–1497.

Masry, E. and Tjøstheim, D. (1997). Additive nonlinear arx time series and projection estimates. *Econometric Theory*, 13(2):214–252.

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

Medeiros, M. and Mendes, E. (2016). ℓ1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, 191:255–271.

Ng, S. (2013). Variable selection in predictive regressions. In Elliott, G. and Timmermann, A., editors, *Handbook of Forecasting*, number 754-786 in 2B. North Holland.

Ng, S. (2014). Viewpoint: Boosting recessions. *Canadian Journal of Economics/Revue canadienne d'Ã©conomique*, 47(1):1–34.

Nicholson, W. B., Bien, J., and Matteson, D. S. (2016). Hierarchical vector autoregression. *arXiv preprint arXiv:1412.5250*.

Orbe, S., Ferreira, E., and Rodriguez-Poo, J. (2005). Nonparametric estimation of time varying parameters under shape restrictions. *Journal of Econometrics*, 126(1):53–77.

Orbe, S., Ferreira, E., and Rodriguez-Poo, J. (2006). On the estimation and testing of time varying constraints in econometric models. *Statistica Sinica*, pages 1313–1333.

Paye, B. S. and Timmermann, A. (2006). Instability of return prediction models. *Journal of Empirical Finance*, 13(3):274–315.

Perron, P. et al. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics*, 1(2):278–352.

Pesaran, M. H. and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137(1):134–161.

Pettenuzzo, D. and Timmermann, A. (2017). Forecasting macroeconomic variables under model instability. *Journal of Business & Economic Statistics*, 35(2):183–201.

Pham, T. D. and Tran, L. T. (1985). Some mixing properties of time series models. *Stochastic Processes and their Applications*, 19(2):297–303.

Phillips, P. C., Li, D., and Gao, J. (2017). Estimating smooth structural change in cointegration models. *Journal of Econometrics*, 196(1):180–195.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2):821–862.

Richter, S. and Dahlhaus, R. (2018). Cross validation for locally stationary processes. *Ann. Statist.* In Press.

Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. In *Statistical Analysis and Forecasting of Economic Structural Change*, pages 253–264. Springer.

Robinson, P. M. (1991). Time-varying nonlinear regression. In *Economic Structural Change*, pages 179–190. Springer.

Rosset, S., Zhu, J., and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5(Aug):941–973.

Rossi, B. (2013). Advances in forecasting under instability. In *Handbook of economic forecasting*, volume 2, pages 1203–1324. Elsevier.

Rossi, B. and Sekhposyan, T. (2010). Have economic modelsâ forecasting performance for US output growth and inflation changed over time, and when? *International Journal of Forecasting*, 26(4):808–835.

Samorodnitsky, G. (2006). Long Range Dependence. *Foundations and Trends in Stochastic systems*, 1:163–257.

Schinasi, G. J. and Swamy, P. A. V. B. (1989). The out-of-sample forecasting performance of exchange rate models when coefficients are allowed to change. *Journal of International Money and Finance*, 8(3):375–390.

Shao, X. and Wu, W. B. (2007). Asymptotic spectral theory for nonlinear time series. *Ann. Statist.*, 35(4):1773–1801.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318.

Shapiro, J. M. (2017). *Is Big Data a Big Deal for Applied Microeconomics?*, volume 2 of *Econometric Society Monographs*, page 35â52. Cambridge University Press.

Stock, J. H. and Watson, M. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, 41(3):788–829.

Stock, J. H. and Watson, M. (2009). Forecasting in dynamic factor models subject to structural instability. *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, 173:205.

Stock, J. H. and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1):11–30.

Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2):293 – 335.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Has the business cycle changed and why? *NBER Macroeconomics Annual*, 17:159–218.

Stock, J. H. and Watson, M. W. (2002c). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Székely, G. J. and Rizzo, M. L. (2014). Partial distance correlation with methods for dissimilarities. *Ann. Statist.*, 42(6):2382–2412.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35(6):2769–2794.

Temlyakov, V. N. (2000). Weak greedy algorithms. *Advances in Computational Mathematics*, 12(2-3):213–227.

Terasvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Assocation*, 89(425):208–218.

Teräsvirta, T., Tjostheim, D., and Granger, C. (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J.Roy. Statist. Soc. B.*, 58:267–288.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., and Canales-Rodríguez, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1457):969–981.

Vogt, M. et al. (2012). Nonparametric regression for locally stationary time series. *The Annals of Statistics*, 40(5):2601–2633.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. . *J.Roy. Statist. Soc. B.*, 71:671–683.

Wang, H., Li, G., and Tsai, C. (2007). Regression coefficient and autoregressive order shrinkage and selection via lasso. . *J.Roy. Statist. Soc. B.*, 69:63–78.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734. PMID: 26877569.

Wen, C., Pan, W., Huang, M., and Wang, X. (2018). Sure independence screening adjusted for confounding covariates with ultrahigh-dimensional data. *Statistica Sinica*, 28:293–317.

Weng, H., Feng, Y., and Qiao, X. (2017). Regularization after retention in ultrahigh dimensional linear regression models. *Statistica Sinica*. In press.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fmri data. *NeuroImage*, 15(1):1 – 15.

Wu, S., Xue, H., Wu, Y., and Wu, H. (2014). Variable selection for sparse high-dimensional nonlinear regression models by combining nonnegative garrote and sure independence screening. *Statistica Sinica*, 24(3):1365–1387.

Wu, W. and Pourahmadi, M. (2009). Banding sample autocovaraince matrices of stationary processes. *Statistica Sinica*, 19:1755–1768.

Wu, W. and Wu, Y. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10:352–379.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154.

Wu, W. B. (2009). An asymptotic theory for sample covariances of bernoulli shifts. *Stochastic Processes and their Applications*, 119(2):453 – 467.

Wu, W. B. (2011). Asymptotic theory for stationary processes . *Statistics and its Interface*, 4(2):207–226.

Wu, W. B. and Min, W. (2005). On linear processes with dependent innovations. *Stochastic Processes and their Applications*, 115(6):939 – 958.

Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *Ann. Statist.*, 40(1):466–493.

Xu, P., Zhu, L., and Li, Y. (2014). Ultrahigh dimensional time course feature selection. *Biometrics*, 70(2):356–365.

Yoshihara, K. i. (1976). Limiting behavior of u-statistics for stationary, absolutely regular processes. *Probability Theory and Related Fields*, 35(3):237–252.

Yousuf, K. (2018). Variable screening for high dimensional time series. *Electronic Journal of Statistics*, 12(1):667–702.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, T., Wu, W. B., et al. (2015). Time-varying nonlinear regression models: Nonparametric estimation and model selection. *The Annals of Statistics*, 43(2):741–768.

Zhou, Z. (2010). Nonparametric inference of quantile curves for nonstationary time series. *The Annals of Statistics*, 38(4):2187–2217.

Zhou, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, 33(3):438–457.

Zhou, Z. and Wu, W. B. (2009). Local linear quantile estimation for nonstationary time series. *The Annals of Statistics*, 37(5B):2696–2729.

Zhu, L., Li, L., Li, R., and Zhu, L. (2011). Model-Free Feature Selection for Ultrahigh Dimensional Data. . *J. Amer. Statist. Assoc.*, 106:1464–1475.

Zou, H. (2006a). The adaptive Lasso and its oracle properties. . *J. Amer. Statist. Assoc.*, 101:1418–1429.

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