Electronic Journal of Statistics
Vol. 12 (2018) 667–702
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1402

Variable screening for high dimensional time series

Kashif Yousuf

Columbia University
Department of Statistics
1255 Amsterdam Ave
New York, NY 10027
e-mail: ky2304@columbia.edu

Abstract: Variable selection is a widely studied problem in high dimensional statistics, primarily since estimating the precise relationship between the covariates and the response is of great importance in many scientific disciplines. However, most of theory and methods developed towards this goal for the linear model invoke the assumption of iid sub-Gaussian covariates and errors. This paper analyzes the theoretical properties of Sure Independence Screening (SIS) (Fan and Lv [20]) 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 sure screening properties, which depend on the moment conditions, and the strength of dependence in the error and covariate processes, amongst other factors. Additionally, combining these screening procedures with the adaptive Lasso is analyzed. Dependence is quantified by functional dependence measures (Wu [49]), and the results rely on the use of Nagaev-type and exponential inequalities for dependent random variables. We also conduct simulations to demonstrate the finite sample performance of these procedures, and include a real data application of forecasting the US inflation rate.

MSC 2010 subject classifications: Primary 62F07; secondary 62J07.
Keywords and phrases: High-dimensional statistics, sparsity, lasso, time series, functional dependence measure, variable selection, Nagaev inequality, sure independence screening.

Received May 2017.

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 [22]. For the biological sciences, see [24, 5] and references therein. The main goals in these situations according to [4] 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

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{1}$$

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 [20],[31]. 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 [20], 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 [25], non-parametric additive models [19], Cox proportional hazards model [18], single index hazard rate models [27], and varying coefficient models [23]. Model-free screening methods have also been developed. For example; screening using distance correlation was analyzed in [35], a martingale difference correlation approach was introduced in [42], additional works include [57], [30] among others. For an overview of works related to screening procedures, one can consult [37]. 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 [21] and [5]. Examples of methods in this class are the Lasso [45], and the adaptive Lasso [58]. 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 [9] 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 [29, 39].

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 [48]). 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 [46], 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 [53] studied the theoretical properties of the Lasso assuming a fixed design in the case of heavy tailed and dependent errors. Additionally [3], and [33] investigated theoretical properties of the Lasso for high-dimensional Gaussian processes. Most recently [39] 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 [36], which allows for heavy tailed errors and covariates. Additionally [10], [54], and [57] 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 [13],[56]). However, the dependence structure of longitudinal data is too restrictive to cover the type of dependence present in most time series. Most recently [12] 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 [16].

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 [49], rather than the usual strong mixing coefficients. Comparisons between functional dependence measures and strong mixing assumptions are discussed in section 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 [20]. 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 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 presents the sure screening properties of SIS under a range of settings. Section 4 introduces the GLSS procedure and presents its sure screening properties. Combining these screening procedures with the adaptive Lasso will discussed in Section 5. Section 6 covers simulation results, while section 7 discusses an application to forecasting the US inflation rate. Lastly, concluding remarks are in Section 8, and the proofs for all the results follow in the appendix.

2. Preliminaries

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

$$\epsilon_i = g\left(\ldots, e_{i-1}, e_i\right) \tag{2}$$

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 [50],[49]). This representation allows us to use the functional and predictive dependence measures introduced in [49]. The functional dependence measure for the error process is defined as the following:

$$\delta_q(\epsilon_i) = ||\epsilon_i - g\left(\mathcal{F}_i^*\right)||_q = \left(E|\epsilon_i - g\left(\mathcal{F}_i^*\right)|^q\right)^{1/q} \tag{3}$$

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:

$$\Delta_{0,q}(\boldsymbol{\epsilon}) = \sum_{i=0}^{\infty} \delta_q(\epsilon_i) < \infty \tag{4}$$

The predictive dependence measure is related to the functional dependence measure, and is defined as the following:

$$\theta_q(\epsilon_l) = ||\mathrm{E}\left(\epsilon_l|\mathcal{F}_0\right) - \mathrm{E}\left(\epsilon_l|\mathcal{F}_{-1}\right)||_q = ||\mathcal{P}_0\epsilon_l||_q \tag{5}$$

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 [49] we obtain $\Theta_q(\boldsymbol{\epsilon}) \leq \Delta_{0,q}(\boldsymbol{\epsilon})$.

Similarly the covariate process is of the form:

$$\boldsymbol{x}_{\boldsymbol{i}}^{(n)} = \boldsymbol{h}\left(\ldots, \boldsymbol{\eta}_{i-1}^{(n)}, \boldsymbol{\eta}_i^{(n)}\right) \tag{6}$$

Where $\boldsymbol{\eta}_i^{(n)} \in \mathcal{R}^{p_n}, i \in \mathbb{Z}$, are iid random vectors, $\boldsymbol{h}(\boldsymbol{\cdot}) = (h_1(\cdot)\ldots, h_{p_n}(\cdot))$, $\boldsymbol{x}_{\boldsymbol{i}}^{(n)} = (X_{i1}, ..., X_{ip_n})$ 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:

$$\Phi_{m,q}(\boldsymbol{x}) = \sum_{i=m}^{\infty} \max_{j \leq p_n} \delta_q(X_{ij}) < \infty \tag{7}$$

The representations (2), and (6), along with the functional and predictive dependence measures have been used in various works including [52],[55], and [53] 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 [50]. In many cases using the functional dependence measure also requires less stringent assumptions. For example, 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 [14] for details). Additional conditions are needed to ensure strong mixing if the innovations for the linear process are dependent [17].

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 $\rho \in (0, 1/2]$ [2]. 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 (2) clearly holds and (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 (4) if: $\sum_{i=0}^{\infty} \delta_q(e_i) < \infty$, and $\sum_{j=0}^{\infty} |f_j| < \infty$ (see [51]).

3. SIS with dependent observations

Sure Independence Screening, as introduced by Fan and Lv [20], 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:

$$\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_{p_n}), \text{ where } \hat{\rho}_j = (\sum_{t=1}^{n} X_{tj}^2)^{-1} (\sum_{t=1}^{n} X_{tj} Y_t) \tag{8}$$

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

$$\mathcal{M}_* = \{1 \leq i \leq p_n : \beta_i \neq 0\} \tag{9}$$

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

$$\hat{\mathcal{M}}_{\gamma_n} = \{1 \leq i \leq p_n : |\hat{\rho}_i| \geq \gamma_n\} \tag{10}$$

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.

3.1. 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 (2), and (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 [41]. Conditions needed for a process to satisfy the GMC condition are given in Theorem 5.1 of [41]. Examples of processes satisfying the GMC condition include stationary, causal finite order ARMA, GARCH, ARMA-GARCH, bilinear, and threshold autoregressive processes, amongst others (see [50] 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

$$\delta_q(X_{ij}) = ||A_{i,j}\boldsymbol{\eta}_0 - A_{i,j}\boldsymbol{\eta}_0^*||_q \leq 2|A_{i,j}|||\eta_{0,1}||_q \tag{11}$$

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, $\boldsymbol{x}_t = B_1 \boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t$, $\Phi_{m,q}(\boldsymbol{x}) = O(||\tilde{B}_1||_2^m)$ [11]. 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:

$$\tilde{\boldsymbol{x}}_t = \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \\ \vdots \\ \boldsymbol{x}_{t-k+1} \end{bmatrix}_{kp\times 1} \quad \tilde{B}_1 = \begin{pmatrix} B_1 & \cdots & B_{k-1} & B_k \\ I_{p_n} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & I_{p_n} & \mathbf{0} \end{pmatrix}_{kp\times kp} \quad \tilde{\boldsymbol{\eta}}_t = \begin{bmatrix} \boldsymbol{\eta}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}_{kp\times 1} \tag{12}$$

And by section 11.3.2 in [38], 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 $O(a^m)$, 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 [52] and [53] 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:

$$\vartheta_n = \frac{s_n n^\omega K_{x,r}^r}{(n/s_n)^{r/2 - r\kappa/2}} + \frac{n^\iota K_{x,r}^\tau K_{\epsilon,q}^\tau}{n^{\tau-\tau\kappa}} + \exp\left(-\frac{n^{1-2\kappa}}{s_n^2 K_{x,r}^4}\right) + \exp\left(-\frac{n^{1-2\kappa}}{K_{x,r}^2 K_{\epsilon,q}^2}\right) \tag{13}$$

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

Theorem 1. *Suppose Conditions A,B,C hold.*

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

$$P\left(\max_{j\leq p_n} |\hat{\rho}_j - \rho_j| > c_2 n^{-\kappa}\right) \leq O(p_n \vartheta_n)$$

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

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

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

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

In Theorem 1 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 1 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, [20] assumed the covariates and errors are jointly normally distributed. Future works applicable to the linear model, such as [25],[19] 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 $E(X_{tj}\epsilon_t) = 0$, $\forall j \leq p_n$.

3.2. 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 [20]. More formally, we have:

Condition D: Assume the error and the covariate processes have representations (2), and (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 [53], 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 $||\epsilon_i||_q$, $\max_{j \le p_n} ||X_{ij}||_q$ 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, [12] 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 2. *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:*

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

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

$$P\left(\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n}\right) \ge 1 - O\left(s_n^2 \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x^2 s_n}\right)^{\tilde{\alpha}}\right) - O\left(s_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x \upsilon_\epsilon}\right)^{\tilde{\alpha}'}\right)$$

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

$$P\left(|\hat{\mathcal{M}}_{\gamma_n}| \le O(n^{2\kappa}\lambda_{max}(\Sigma))\right) \ge 1 - O\left(s_n p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x^2 s_n}\right)^{\tilde{\alpha}}\right) - O\left(p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x \upsilon_\epsilon}\right)^{\tilde{\alpha}'}\right)$$

From Theorem 2, 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 1, 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 1, 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 [20], 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 [25], 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.

4. Generalized least squares screening (GLSS)

Consider the marginal model:

$$Y_t = X_{tk}\rho_k + \epsilon_{t,k} \tag{14}$$

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 (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 (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 1. We observe that the performance 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:

$$\tilde{\beta}_k^M = (\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{y} \tag{15}$$

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 autocovariance 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 [52], which is defined as:

$$\hat{\Sigma}_{k,l_n} = \left(\hat{\gamma}_{i-j,k} \mathbb{1}_{|i-j| \leq l_n}\right)_{1 \leq i,j \leq n} \tag{16}$$

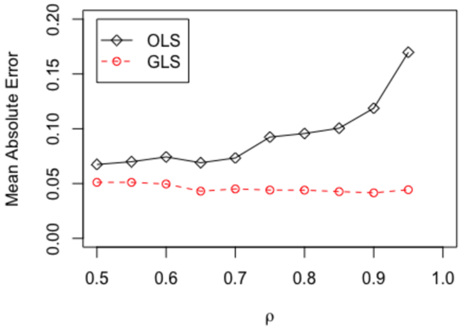

FIG 1. *GLS vs OLS error comparison for values of ρ between .5 and .95 incrementing by .05. Absolute error averaged over 200 replications.*

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:

$$\hat{\beta}_k^M = (\boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{y} \tag{17}$$

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. [1] 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 1). 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 (2), and the covariate process is of the form (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 (2), and the covariate process is of the form (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 [7]). For further details on linear AR approximations to stationary processes, see [1] and [8]. We remark that compared to previous works [1, 34], 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 (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:

$$\beta_k^M = E(y_t - \sum_{i=1}^{L_k} \alpha_i y_{t-i})(X_{t,k} - \sum_{i=1}^{L_k} \alpha_i X_{t-i,k})/(E(X_{t,k} - \sum_{i=1}^{L_k} \alpha_i X_{t-i,k})^2) \quad (18)$$

If we assume the cross covariance, $\gamma_{\mathbf{X}_k,Y}(h)$, is proportional to $E(X_{tk}Y_t)$, i.e. $\gamma_{\mathbf{X}_k,Y}(h) \propto E(X_{tk}Y_t)$, 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 $\gamma_{\mathbf{X}_k,Y}(h) \propto E(X_{tk}Y_t)$, 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 $\hat{\mathcal{M}}_{\gamma_n} = \Big\{1 \leq i \leq p_n : |\hat{\beta}_k^M| \geq \gamma_n\Big\}$, $\alpha = min(\alpha_x, \alpha_\epsilon)$, $\tau = \frac{qr}{q+r}, \tau' = \frac{qq'}{q+q'} = \min(q/2, \tau)$. 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:

$$a_n = l_n\left[\frac{n^{\iota}l_n^{\tau'}K_{x,r}^{\tau'}\tilde{K}_{\epsilon,q'}^{\tau'}}{n^{\tau'-\tau'\kappa}} + \frac{n^{\zeta}l_n^{q'/2}\tilde{K}_{\epsilon,q'}^{q'}}{n^{q'/2-q'\kappa/2}} + \frac{n^{\omega}l_n^{r/2}K_{x,r}^{r}}{n^{r/2}}\right] \tag{19}$$

$$b_n = l_n\left[\exp\left(-\frac{n^{1/2}}{l_n\upsilon_x^2}\right)^{\tilde{\alpha}} + \exp\left(-\frac{n^{1/2-\kappa}}{l_n\upsilon_x\phi}\right)^{\tilde{\varphi}'} + \exp\left(-\frac{n^{1/2-\kappa}}{l_n\phi^2}\right)^{\tilde{\varphi}}\right] \tag{20}$$

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 1. *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 3. *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:

$$P\left(\max_{j\leq p_n}|\hat{\beta}_k^M - \beta_k^M| > c_2n^{-\kappa}\right) \leq O(p_na_n)$$

(ii) Assume Conditions E,F,G,H hold, then for $\gamma_n = c_5n^{-\kappa}$ with $c_5 \leq c_6/2$:

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

(iii) Assume Conditions E,F,G,I hold, for any $c_2 > 0$ we have:

$$P\left(\max_{j\leq p_n}|\hat{\beta}_k^M - \beta_k^M| > c_2n^{-\kappa}\right) \leq O(p_nb_n)$$

(iv) Assume Conditions E,F,G,I hold, then for $\gamma_n = c_5n^{-\kappa}$ with $c_5 \leq c_6/2$:

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

In Lemma 1, 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 1, and parts (i), and (ii) of Theorem 3 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 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 $h \in \{-L_k, \dots, -1, 1, \dots, L_k\}$, we can bound the selected model size by the model size selected by SIS. More formally we have:

Corollary 4. *Assume the cross covariance, $\gamma_{\boldsymbol{X}_k,Y}(h) \propto E(X_{k,t}Y_t)$, for $h \in \{-L_k, \dots, -1, 1, \dots, L_k\}$*

(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 $c_5 \leq c_6/2$:

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

5. Second stage selection with adaptive lasso

The adaptive Lasso, as introduced by [58], is the solution to the following:

$$argmin_\beta \ ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda_n \sum_{j=1}^{p_n} w_j|\beta_j|, \ \text{where } w_j = |\hat{\beta}_{I,j}|^{-1}, \tag{21}$$

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 [29], or we can use the

Lasso as our initial estimator provided the restricted eigenvalue condition holds (see [39]). 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:

$$\tilde{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}} = argmin_{\boldsymbol{\beta}_{\hat{\mathcal{M}}_{\gamma_n}}} \ ||\boldsymbol{y} - \boldsymbol{X}_{\hat{\mathcal{M}}_{\gamma_n}} \boldsymbol{\beta}_{\hat{\mathcal{M}}_{\gamma_n}}||^2 + \lambda_n \sum_{j=1}^{d_n} w_j |\beta_j|, w_j = |\hat{\beta}_{I,j}|^{-1} \quad (22)$$

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 $\hat{\boldsymbol{\beta}}_I = (\hat{\beta}_{I,1}, \dots, \hat{\beta}_{I,d_n})$ 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 [6] for details):

$$\phi_0 = \min_{S \subseteq \{1,\dots,d'_n\}, |S| \le s_n} \ \min_{\boldsymbol{v} \neq 0, |\boldsymbol{v}_{S^c}| \le 3 |\boldsymbol{v}_S|} \frac{\boldsymbol{v}^T \Sigma_{\mathcal{M}_{\frac{\gamma_n}{2}}} \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}} \ge c > 0, \quad (23)$$

where $\boldsymbol{v} = (v_1, \dots, v_{d'_n})$ 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:

$$cn^{1-\frac{\psi}{2}} (\frac{\phi_0}{s_n})^{3/2} \ge \lambda_{I,n} \ge \lambda_n n^{\psi/2} \quad (24)$$

Condition L: Let $\beta_{min} = \min_{i \le s_n} |\beta_i|$, and $w_{\max} = \max_{i \le s_n} w_i > 0$. 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, $\mathcal{M}_{\frac{\gamma_n}{2}} = \{1 \le i \le p : |\rho_i| \ge \gamma_n/2\}$, 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 [39]. 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 1, and $\tilde{\alpha}', \tilde{\alpha}$ have been defined in Theorem 2.

Theorem 5. *(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:*

$$\begin{aligned}
&P(sgn(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}}) \\
&\quad = sgn(\boldsymbol{\beta})) \\
&\quad \geq 1 - O\left(s_n p_n \left[\frac{n^{\omega} K_{x,r}^{r}}{(n/s_n)^{r/2 - r\kappa/2}} - \exp(-\frac{n^{1-2\kappa}}{s_n^2 K_{x,r}^4})\right]\right) \\
&\quad - O\left(p_n \left[\frac{n^{\iota} K_{x,r}^{\tau} K_{\epsilon,q}^{\tau}}{n^{\tau - \tau\kappa}} - \exp(-\frac{n^{1-2\kappa}}{K_{x,r}^2 K_{\epsilon,q}^2})\right]\right) \\
&\quad - O\left(d_n'^2 \left[\frac{n^{\omega} K_{x,r}^{r}}{(n/s_n)^{r/2}} - \exp(-\frac{n}{s_n^2 K_{x,r}^4})\right]\right) \\
&\quad - O\left(d_n' \left[\frac{n^{\iota} K_{x,r}^{\tau} K_{\epsilon,q}^{\tau}}{\lambda_n^{\tau} n^{\tau\psi/2}} + \exp(-\frac{\lambda_n^2 n^{\psi-1}}{K_{x,r}^2 K_{\epsilon,q}^2})\right]\right)
\end{aligned}$$

(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:

$$\begin{aligned}
P(sgn(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}}) = sgn(\boldsymbol{\beta})) &\geq 1 - O(s_n p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x^2 s_n}\right)^{\tilde{\alpha}}) \\
&-O(p_n \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x \upsilon_\epsilon}\right)^{\tilde{\alpha}'}) - O(d_n'^2 \exp\left(-\frac{n^{1/2}}{\upsilon_x^2 s_n}\right)^{\tilde{\alpha}}) \\
&-O(d_n' \exp\left(-\frac{\lambda_n n^{\psi/2-1/2}}{\upsilon_x \upsilon_\epsilon}\right)^{\tilde{\alpha}'})
\end{aligned}$$

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 $\lambda_n n^{\psi/2-1/2} \to \infty$,
$p_n = o(\min(\exp\left(\frac{Cn^{1/2-\kappa}}{s_n}\right)^{\tilde{\alpha}} / s_n, \exp(Cn^{1/2-\kappa})^{\tilde{\alpha}'}))$,
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 3, 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 [39]. 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 [53], our results are easier to interpret and also allow us obtain a higher range for p_n.

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 (1), for the covariate process we have:

$$\boldsymbol{x}_t = A_1\boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t \tag{25}$$

TABLE 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

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.

The results are displayed in table 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 (25), with $A_1 = diag(.4)$. $\boldsymbol{\eta}_t \sim N(0, \Sigma_\eta)$, or $\boldsymbol{\eta}_t \sim t_5(0, V)$, with $\Sigma_\eta = \{.3^{|i-j|}\}_{i,j \le p_n}$ 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) 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 (25), with $A_1 = \{.4^{|i-j|+1}\}_{i,j \le p_n}$. 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.

TABLE 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

TABLE 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

The results are displayed in tables 2, and 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

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 [47]. [26] 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 (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 (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.

Table 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 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

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 4 and 5 for 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.

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:

$$y^{12}_{t+12} = 100 \times \log\left(\frac{CPI_{t+12}}{CPI_t}\right), \text{ or } y^{12}_{t+12} = 100 \times \log\left(\frac{PPI_{t+12}}{PPI_t}\right) \tag{26}$$

Therefore the above quantities are approximately the percentage change in CPI or PPI over 12 months. Our data was obtained from the supplement to [32], 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 [32]. Treating $\log(CPI)$, and $\log(PPI)$ as $I(1)$, has been found to provide an adequate description of the data according to [44],[43],[39].

We consider forecasts from 8 different models. Similar to [39, 44] our benchmark model is an AR(4) model: $\hat{y}^{12}_{t+12} = \hat{\alpha}_0 + \sum_{i=0}^{3} \hat{\alpha}_i y_{t-i}$, 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:

$$\hat{y}^{12}_{t+12} = \hat{\beta}_0 + \sum_{i=0}^{3} \hat{\alpha}_i y_{t-i} + \hat{\boldsymbol{\gamma}} \hat{\boldsymbol{F}}_t \tag{27}$$

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 regression and two stage forecasts is:

$$y^{12}_{t+12} = \beta_0 + \boldsymbol{x}_t \boldsymbol{\beta} + \epsilon^{12}_{t+12} \tag{28}$$

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 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 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 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.

Table 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

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 [40] and references therein). If we assume the error process (ϵ_i) is long range dependent, then by the proof of Theorem 1 in [52] 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.

9. Appendix

9.1. Proofs of results

Proof of Theorem 1. We first prove part (i), we start by obtaining a bound on:

$$P(|\hat{\rho}_j - \rho_j| > c_2 n^{-\kappa}) \tag{29}$$

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:

$$\begin{aligned} P(|\hat{\rho}_j - \rho_j| > c_2 n^{-\kappa}) &\leq P(|(T_1^{-1} - E(T_1)^{-1})(T_2 - E(T_2))| > c_2 n^{-\kappa}/3) \quad (30) \\ &+ P(|(T_2 - E(T_2))/E(T_1) > c_2 n^{-\kappa}/3|) \quad (31) \\ &+ P(|(T_1^{-1} - E(T_1)^{-1})E(T_2)| > c_2 n^{-\kappa}/3) \quad (32) \end{aligned}$$

For the RHS of (30), we obtain:

$$(30) \leq P(|(T_2 - E(T_2))| > Cn^{-\kappa/2}) + P(|(T_1^{-1} - E(T_1)^{-1})| > Cn^{-\kappa/2}) \quad (33)$$

Therefore it suffices to focus on terms (31), (32). For (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:

$$S_1 = \sum_{t=1}^{n} X_{tj}(\sum_{k=1}^{p_n} X_{tk}\beta_k)/n \text{ and } S_2 = \sum_{t=1}^{n} X_{tj}\epsilon_t/n \quad (34)$$

By Condition B, $E(X_{tj}\epsilon_t) = 0$, therefore

$$P(|T_2 - E(T_2)| > Cn^{-\kappa}) \leq P(|S_1 - E(S_1)| > Cn^{-\kappa}/2) + P(|S_2| > Cn^{-\kappa}) \quad (35)$$

Recall that $\sum_{k=1}^{p_n} \mathbb{1}_{|\beta_k|>0} = s_n$, thus:

$$\begin{aligned} &P\left(|S_1 - E(S_1)| > \frac{c_2 n^{-\kappa}}{2}\right) \\ &\leq \sum_{k \in M_*} P\left(|\sum_{t=1}^{n} \frac{X_{tj}(X_{tk}\beta_k)}{n} - \beta_k E(X_{tj}X_{tk})| > \frac{c_2 n^{-\kappa}}{2s_n}\right) \quad (36) \end{aligned}$$

From section 2 in [53]: $||X_{ij}||_r \leq \Delta_{0,r}(\boldsymbol{X}_j) \leq \Phi_{0,r}(\boldsymbol{x})$. Using this we compute the cumulative functional dependence measure of $X_{tk}X_{tj}$ as:

$$\begin{aligned} \sum_{t=m}^{\infty} ||X_{tj}X_{tk} - X_{tj}^* X_{tk}^*||_{r/2} &\leq \sum_{t=m}^{\infty} (||X_{tj}||_r ||X_{tk} - X_{tk}^*||_r + ||X_{tk}||_r ||X_{tj} - X_{tj}^*||_r) \\ &\leq 2\Phi_{0,r}(\boldsymbol{x})\Phi_{m,r}(\boldsymbol{x}) = O(m^{-\alpha_x}) \quad (37) \end{aligned}$$

Therefore we obtain: $\sup_m (m+1)^{\alpha_x} \sum_{t=m}^{\infty} ||X_{tj}X_{tk} - X_{tj}^* X_{tk}^*||_{r/2} \leq 2K_{x,r}^2$. Combining this with (36), and Theorem 2 in [53], yields:

$$P\left(|S_1 - E(S_1)| > \frac{c_2 n^{-\kappa}}{2}\right) \leq C s_n \left(\frac{n^{\omega} K_{x,r}^{r}}{(n/s_n)^{r/2 - r\kappa/2}} + \exp\left(-\frac{n^{1-2\kappa}}{s_n^2 K_{x,r}^4}\right)\right) \quad (38)$$

Similarly for $X_{tj}\epsilon_t$, by using Holder's inequality we obtain:

$$\sum_{t=m}^{\infty} ||X_{tj}\epsilon_t - X^*_{tj}\epsilon^*_t||_\tau \leq \sum_{t=m}^{\infty}(||X_{tj}||_r||\epsilon_t - \epsilon^*_t||_q + ||\epsilon_t||_q||X_{tj} - X^*_{tj}||_r)$$
$$\leq \Delta_{0,q}(\boldsymbol{\epsilon})\Phi_{m,r}(\boldsymbol{x}) + \Delta_{m,q}(\boldsymbol{\epsilon})\Phi_{0,r}(\boldsymbol{x}) = O(m^{-\alpha}) \quad (39)$$

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 [53], we obtain:

$$P\left(|S_2| > \frac{c_2 n^{-\kappa}}{2}\right) \leq O\left(\frac{n^\iota K^\tau_{x,r}K^\tau_{\epsilon,q}}{n^{\tau-\tau\kappa}} + \exp\left(-\frac{n^{1-2\kappa}}{K^2_{x,r}K^2_{\epsilon,q}}\right)\right) \quad (40)$$

For (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:

$$(32) \leq P(|T_1 - E(T_1)| > T_1 C n^{-\kappa}) \leq P(|T_1 - E(T_1)| > MCn^{-\kappa}) + P(T_1 < M) \quad (41)$$

We set $M < \min_{j\leq p_n} E(X^2_{ij}) - \epsilon$, for $\epsilon > 0$. We then have:

$$P(T_1 < M) \leq P(|T_1 - E(T_1)| > E(T_1) - M) \quad (42)$$

We can then bound the above two equations similar to (38). By combining (33)(35),(38),(40),(41), along with union bound we obtain:

$$P\left(\max_{j\leq p_n}|\hat{\rho}_j - \rho_j| > c_2 n^{-\kappa}\right) \leq O\left(s_n p_n\left[\frac{n^\omega K^r_{x,r}}{(n/s_n)^{r/2-r\kappa/2}} + \exp\left(-\frac{n^{1-2\kappa}}{s^2_n K^4_{x,r}}\right)\right]\right)$$
$$+ O\left(p_n\left[\frac{n^\iota K^\tau_{x,r}K^\tau_{\epsilon,q}}{n^{\tau-\tau\kappa}} + \exp(-n^{1-2\kappa}/K^2_{x,r}K^2_{\epsilon,q})\right]\right)$$

To prove part (ii), we follow the steps in the proof of Theorem 2 in [36]. 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:

$$|\hat{\rho}_k| \geq |\rho_k| - |\hat{\rho}_k - \rho_k| \geq c_1 n^{-\kappa}/2,\ \forall k \in M_* \quad (43)$$

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 [36]. Using $Var(Y_t), Var(X_{tj}) = O(1)$ for $j \leq p_n$, along with Condition B, we obtain $\sum_{k=1}^{p_n}\rho^2_k = O(\lambda_{\max}(\Sigma))$. 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:

$$P\left(|\hat{\mathcal{M}}_{\gamma_n}| < O(n^{2\kappa}\lambda_{\max}(\Sigma))\right) > P(\mathcal{B}_n) \quad (44)$$

The result then follows from part (i). □

Proof of Theorem 2. We follow the steps from the proof of Theorem 1. Let $\boldsymbol{T} = (T_1, \dots, T_n)$ where $T_i = X_{ij}X_{ik}$, and let $\boldsymbol{R} = (R_1, \dots, R_n)$ 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 [49], $\Theta_q(\boldsymbol{T}) \leq \Delta_{0,q}(\boldsymbol{T})$, and from Section 2 in [53]: $||X_{ij}||_q \leq \Delta_{0,q}(\boldsymbol{X}_j) \leq \Phi_{0,q}(\boldsymbol{x})$. Additionally, by Holders inequality we have

$$\Delta_{0,q}(\boldsymbol{T}) \leq \sum_{t=0}^{\infty} (||X_{tj}||_{2q} ||X_{tk} - X_{tk}^*||_{2q} + ||X_{tk}||_{2q} ||X_{tj} - X_{tj}^*||_{2q}) \leq 2\Phi_{0,2q}^2(\boldsymbol{x}) \tag{45}$$

Using these, along with Condition D we obtain:

$$\sup_{q \geq 4} q^{-2\tilde{\alpha}_x} \Theta_q(\boldsymbol{T}) \leq \sup_{q \geq 4} q^{-2\tilde{\alpha}_x} \Delta_{0,q}(\boldsymbol{T}) \leq \sup_{q \geq 4} 2q^{-2\tilde{\alpha}_x} \Phi_{0,2q}^2(\boldsymbol{x}) < \infty \tag{46}$$

Combining the above and using Theorem 3 in [53], we obtain:

$$P\left(\left| \sum_{i=1}^n T_i - E(T_i) \right| > \frac{c_2 n^{1-\kappa}}{2} \right) \leq C \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x^2} \right)^{\tilde{\alpha}} \tag{47}$$

Similarly, using the same procedure we obtain:

$$P\left(\left| \sum_{i=1}^n R_i \right| > \frac{c_2 n^{1-\kappa}}{2} \right) \leq C \exp\left(-\frac{n^{1/2-\kappa}}{\upsilon_x \upsilon_\epsilon} \right)^{\tilde{\alpha}'} \tag{48}$$

Now using the above bounds and following the steps in the proof of Theorem 1 we obtain the results. □

Proof of Lemma 1. By the proof of Theorem 2 in [52], we have:

$$||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 \leq 2 \sum_{i=1}^{l_n} |\hat{\gamma}_{i,k} - \gamma_{i,k}| + 2 \sum_{i=l_n+1}^{\infty} |\gamma_{i,k}| \tag{49}$$

Recall that $\hat{\rho}_k$ is the OLS estimate of the marginal projection, by (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:

$$\hat{\gamma}_{i,k} = \frac{1}{n} \sum_{t=1}^{n-|i|} \Bigg[\epsilon_{t,k}\epsilon_{t+|i|,k} - \epsilon_{t,k} X_{t+|i|,k} \left(\sum_{j=1}^n X_{jk}\epsilon_{j,k}/n \right) \tag{50}$$

$$- \epsilon_{t+|i|,k} X_{tk} \left(\sum_{j=1}^n X_{jk}\epsilon_{j,k}/n \right) + X_{tk} X_{t+|i|,k} \left(\sum_{j=1}^n X_{jk}\epsilon_{j,k}/n \right)^2 \Bigg] \tag{51}$$

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 (49). We then have:

$$P(||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 > cn^{-\kappa}) \leq \sum_{i=1}^{l_n} P\left(|\hat{\gamma}_{i,k} - \gamma_{i,k}| > cn^{-\kappa}/l_n \right) \tag{52}$$

And

$$P(|\hat{\gamma}_{i,k} - \gamma_{i,k}| > \frac{cn^{-\kappa}}{l_n}) \le P\Bigg(\Bigg|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}\epsilon_{t+|i|,k} - E(\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}\epsilon_{t+|i|,k})\Bigg| \tag{53}$$

$$+ \Bigg|E(\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}\epsilon_{t+|i|,k}) - \gamma_{i,k}\Bigg| > cn^{-\kappa}/4l_n\Bigg) \tag{54}$$

$$+ P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k}\left(\frac{\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n}{\sum_{j=1}^{n}X_{jk}^2/n}\right)\right| > cn^{-\kappa}/4l_n\right) \tag{55}$$

$$+ P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t+|i|,k}X_{tk}\left(\frac{\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n}{\sum_{j=1}^{n}X_{jk}^2/n}\right)\right| > cn^{-\kappa}/4l_n\right) \tag{56}$$

$$+ P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}X_{tk}X_{t+|i|,k}\left(\frac{\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n}{\sum_{j=1}^{n}X_{jk}^2/n}\right)^2\right| > cn^{-\kappa}/4l_n\right) \tag{57}$$

For (54), the bias $|E(\sum_{t=1}^{n-|i|}\frac{\epsilon_{t,k}\epsilon_{t+|i|,k}}{n} - \gamma_{i,k}| \le \frac{i\gamma_{i,k}}{n}$. Using the techniques in the proof of Theorem 1 we can then bound (53). For (55) we have:

$$(55) \le P\left(\frac{|\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n|}{\sum_{j=1}^{n}X_{jk}^2/n} > cn^{-\kappa}/Ml_n\right) + P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k}\right| > M\right) \tag{58}$$

And $P\left(\left|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k}\right| > M\right)$

$$\le P\Bigg(\Bigg|\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k} - E(\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k})\Bigg|$$

$$> M - \Bigg|E(\frac{1}{n}\sum_{t=1}^{n-|i|}\epsilon_{t,k}X_{t+|i|,k})\Bigg|\Bigg) \tag{59}$$

And we set $M > \max_{k\le p_n}\max_{i\le l_n} 2|E(\epsilon_{t,k}X_{t+|i|,k})| + \epsilon$, for some $\epsilon > 0$. Similarly we have $P(\frac{|\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n|}{\sum_{j=1}^{n}X_{jk}^2/n} > cn^{-\kappa}/Ml_n)$

$$\le P\left(\left|\sum_{j=1}^{n}X_{jk}\epsilon_{j,k}/n\right| > M_1Cn^{-\kappa}/l_n\right) + P\left(\sum_{j=1}^{n}X_{jk}^2/n < M_1\right) \tag{60}$$

Where we set $M_1 < \min_{j\le p_n} E(X_{ij}^2) - \epsilon$, for $\epsilon > 0$. The same method we used for (53) can be applied to (56), (57). Using the techniques in the proof of Theorem 1, and (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 2. □

Proof of Theorem 3. 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 (14), it suffices to obtain a bound for:

$$P(|(\boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{X}_k/n)^{-1} \boldsymbol{X}_k^T \hat{\Sigma}_{k,l_n}^{-1} \boldsymbol{\epsilon}^k/n - (E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{X}_k))^{-1} E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{\epsilon}^k)| > cn^{-\kappa}) \tag{61}$$

Similar to the proof of Theorem 1 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:

$$\begin{aligned} |\hat{\beta}_k^M - \beta_k^M| = |T_2/T_1 - T_4/T_3| = |(T_1^{-1} - T_3^{-1})(T_2 - T_4) \\ + (T_2 - T_4)/T_3 + (T_1^{-1} - T_3^{-1})T_4| \end{aligned} \tag{62}$$

Following the steps in the proof of Theorem 1, it suffices to focus on the terms:

$$P(|T_1 - T_3| > cn^{-\kappa}) \text{ and } P(|T_2 - T_4| > cn^{-\kappa}) \tag{63}$$

We then have:

$$\begin{aligned} P(|T_2 - T_4| > Cn^{-\kappa}) \leq P(|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/n| > Cn^{-\kappa}/2) \qquad (64) \\ + P(|\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{\epsilon}^k/n - E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{\epsilon}^k)| > Cn^{-\kappa}/2) \end{aligned}$$

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:

$$\tilde{\boldsymbol{X}}_k^T \tilde{\boldsymbol{\epsilon}}^k/n = \sum_{t=L_k+1}^{n} \left(\epsilon_{t,k} - \sum_{i=1}^{L_k} \alpha_{i,k} \epsilon_{t-i,k} \right) \left(X_{t,k} - \sum_{i=1}^{L_k} \alpha_{i,k} X_{t-i,k} \right) \tag{65}$$

, where $(\alpha_{1,k}, \dots, \alpha_{L_k,k})$ 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:

$$\begin{aligned} &\sum_{l=m}^{\infty} ||\tilde{X}_{l,k}\tilde{\epsilon}_{l,k} - \tilde{X}_{l,k}^* \tilde{\epsilon}_{l,k}^*||_{\tau'} \\ &\quad \leq \sum_{l=m}^{\infty} (||\tilde{X}_{l,k}||_r ||\tilde{\epsilon}_{l,k} - \tilde{\epsilon}_{l,k}^*||_{q'} + ||\tilde{\epsilon}_{l,k}||_{q'} ||\tilde{X}_{l,k} - \tilde{X}_{l,k}^*||_r) \end{aligned} \tag{66}$$

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:

$$\sum_{l=m}^{\infty} ||\tilde{X}_{l,k}\tilde{\epsilon}_{l,k} - \tilde{X}_{l,k}^* \tilde{\epsilon}_{l,k}^*||_{\tau'} \leq C\Phi_{m,r} = O(m^{-\alpha_x}) \tag{67}$$

Using Theorem 2 in [53]:

$$P(|\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{\epsilon}^k/n - E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{\epsilon}^k)| > Cn^{-\kappa}) \leq O\left(\frac{n^{\iota} K_{x,r}^{\tau'}}{n^{\tau' - \tau' \kappa}} \right) \tag{68}$$

For the term $|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/n|$, using Cauchy-Schwarz inequality:

$$\frac{|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/n|}{||\boldsymbol{X}_k||_2||\boldsymbol{\epsilon}^k||_2} \leq \frac{||(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k||_2}{n||\boldsymbol{\epsilon}^k||_2} \leq \frac{||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2}{n} \tag{69}$$

Using (69) we obtain:

$$\begin{aligned} &P(|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/n| > Cn^{-\kappa}) \\ &\leq P(||\boldsymbol{X}_k||_2||\boldsymbol{\epsilon}^k||_2||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2/n > Cn^{-\kappa}) \end{aligned} \tag{70}$$

Where the right hand side of (70) is:

$$\leq P(||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2 > Cn^{-\kappa}/\sqrt{M}) + P\left(\left(\sum_{i=1}^n X_{ik}^2/n\right)\left(\sum_{i=1}^n \epsilon_{i,k}^2/n\right) > M\right) \tag{71}$$

Let $M = M_1 M_2$, 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 (71) is:

$$\leq P\left(\sum_{i=1}^n X_{ik}^2/n > M_1\right) + P\left(\sum_{i=1}^n \epsilon_{i,k}^2/n > M_2\right) \tag{72}$$

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, $0 < C_1 \leq \lambda_{min}(\Sigma_k) \leq \lambda_{max}(\Sigma_k) \leq C_2 < \infty, \forall k \leq p_n$ [52]. We then use:

$$\begin{aligned} \lambda_{\min}(\Sigma_k)||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2 &\leq ||\Sigma_k^{\frac{1}{2}}(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\Sigma_k^{\frac{1}{2}}||_2 \\ &= ||\Sigma_k^{\frac{1}{2}}\hat{\Sigma}_{k,l_n}^{-1}\Sigma_k^{\frac{1}{2}} - I_n||_2 \end{aligned} \tag{73}$$

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

$$\max_i |a_i - 1| = ||\Sigma_k^{-\frac{1}{2}}\hat{\Sigma}_{k,l_n}\Sigma_k^{-\frac{1}{2}} - I_n||_2 \leq \lambda_{\max}(\Sigma_k^{-1})||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 \tag{74}$$

Let $a_j = argmin_{a_i}|a_i^{-1}|$, using this and (73),(74) we obtain:

$$\begin{aligned} P(||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2 > Cn^{-\kappa}) &\leq P(||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 > Ca_j n^{-\kappa}) \\ &\leq P(||\hat{\Sigma}_{k,l_n} - \Sigma_k||_2 > CM_3 n^{-\kappa}) + P(|a_j| < M_3) \end{aligned} \tag{75}$$

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 (75) and Lemma 1, we obtain:

$$P(||\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1}||_2 > Cn^{-\kappa}) \\ \leq l_n O\left(\frac{n^{\iota} l_n^{\tau'} K_{x,r}^{\tau'} \tilde{K}_{\epsilon,q'}^{\tau'}}{n^{\tau' - \tau'\kappa}} + \frac{n^{\zeta} l_n^{q'/2} \tilde{K}_{\epsilon,q'}^{q'}}{n^{q'/2 - q'\kappa/2}} + \frac{n^{\omega} l_n^{r/2} K_{x,r}^{r}}{n^{r/2}} \right) \tag{76}$$

By (64),(68),(70),(72),(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:

$$P(|T_1 - E(T_1)| > Cn^{-\kappa}) \leq P(|\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{X}^k/n| > Cn^{-\kappa}/2) \\ + P(|\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{X}^k/n - E(\boldsymbol{X}_k^T \Sigma_k^{-1} \boldsymbol{X}^k)| > Cn^{-\kappa}/2)$$

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 1. For (iii) and (iv) we follow the same procedure as (i) and (ii), and apply the methods seen in the proof of Theorem 2; we omit the details. □

Proof of Corollary 4. 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 1 and using the results from Theorem 3. □

Proof of Theorem 5. 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

$$\mathcal{A}_n = \{max_{k \leq p_n} |\hat{\rho}_k - \rho_k| \leq c_3 n^{-\kappa}/2\}$$
$$\mathcal{B}_n = \{max_{i,j \leq d_n'} |[\Sigma_{\mathcal{M}_{\frac{\gamma_n}{2}}} - \hat{\Sigma}_{\mathcal{M}_{\frac{\gamma_n}{2}}}]_{i,j}| \leq \frac{\phi_0}{16 s_n}\}$$
$$\mathcal{C}_n = \{max_{k \leq d_n'} |\sum_{i=1}^{n} X_{ik}\epsilon_i| \leq \lambda_n n^{\psi/2}\}$$

On the set $\mathcal{A}_n$, if we apply screening as a first stage procedure, by our choice of γ_n, we obtain:

$$\mathcal{M}_* \subset \hat{\mathcal{M}}_{\gamma_n} \subset \mathcal{M}_{\gamma_n/2} \tag{77}$$

Next we need to use Lemma 7 and 8 in [39], specifically we need to show our reduced model satisfies conditions DGP 3,DESIGN, and WEIGHTS in [39]. On the set $\mathcal{B}_n$, by Lemma 1 in [39], we have $\phi_{\Sigma_{\hat{\mathcal{M}}_{\gamma_n/2}}} = \phi_{\Sigma_{\mathcal{M}_{\gamma_n/2}}} = \phi_0$. Therefore, we have:

$$\phi_{\Sigma_{\hat{\mathcal{M}}_{\gamma_n}}} = \min_{S \subseteq \{1,\ldots,d_n\}, |S| \leq s_n} \min_{\boldsymbol{v} \neq 0, |\boldsymbol{v}_{S^c}| \leq 3|\boldsymbol{v}_S|} \frac{\boldsymbol{v}^T \Sigma_{\hat{\mathcal{M}}_{\gamma_n}} \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}} \geq \phi_0 \tag{78}$$

Using this along with Lemma 1 in [39] 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 [39], 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 [39] we obtain:

$$P(sgn(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}_{\gamma_n}}) = sgn(\boldsymbol{\beta})) \geq P(\mathcal{A}_n \cap \mathcal{B}_n \cap \mathcal{C}_n) \geq 1 - P(\mathcal{A}_n^{\complement}) - P(\mathcal{B}_n^{\complement}) - P(\mathcal{C}_n^{\complement}) \quad (79)$$

$P(\mathcal{A}_n^{\complement})$ is given in Theorem 1 part i. For $P(\mathcal{B}_n^{\complement})$ using the method in the proof for Theorem 1, we obtain:

$$P(\mathcal{B}_n^{\complement}) \leq {d'_n}^2 O\left(\frac{n^{\omega}K_{x,r}^r}{n^{r/2}} + \exp\left(-n/K_{x,r}^4\right)\right) \quad (80)$$

And for $P(\mathcal{C}_n^{\complement})$:

$$P(\mathcal{C}_n^{\complement}) \leq d'_n O\left(\frac{n^{\iota}K_{x,r}^{\tau}K_{\epsilon,q}^{\tau}}{\lambda_n^{\tau}n^{\tau\psi/2}} + \exp\left(-\lambda_n^2 n^{\psi-1}/K_{x,r}^2 K_{\epsilon,q}^2\right)\right) \quad (81)$$

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 2, and using Theorem 3 in [53]. □

9.2. Asymptotic distribution of GLS estimator

Lemma 2. *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 2. 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 [15]:

$$\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{\epsilon}^k/\sqrt{n} \to 0$$
$$\boldsymbol{X}_k^T(\hat{\Sigma}_{k,l_n}^{-1} - \Sigma_k^{-1})\boldsymbol{X}_k/n \to 0$$

By the proof of theorem 3, 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 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 [28], 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 [52], $W_1 \leq (\sum_{t=0}^{\infty} \delta_2(\tilde{x}_{t,k}e_t))^2$, 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.

Acknowledgments

The author would like to thank the two anonymous referees, the Associate Editor, and the Editor for their helpful comments that have improved the paper.

References

[1] Amemiya, T. (1973). Generalized Least Squares with an Estimated Autocovariance Matrix. *Econometrica* **41** 723-732.

[2] Andrews, D. W. K. (1984). Non-Strong Mixing Autoregressive Processes. *Journal of Applied Probability* **21** 930-934.

[3] Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535-1567.

[4] Bickel, P. J. (2008). Discussion of Sure independence screening for ultra-high dimensional feature space. *J.Roy. Statist. Soc. B.* **70** 883-884.

[5] Bickel, P. J., Brown, B. B., Huang, H. and Li, Q. (2009). An overview of recent developments in genomics and associated statistical methods. *Phil. Transactions of the Roy. Soc. A* **367** 4313-4337.

[6] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705-1732.

[7] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.* Springer.

[8] Bühlmann, P. (1995). Moving-average representation of autoregressive approximations. *Stochastic Processes and their Applications* **60** 331-342.

[9] Buhlmann, P. and Van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer.

[10] Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.* **41** 2123-2148.

[11] Chen, X., Xu, M. and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41** 2994-3021.

[12] 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**.

[13] 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** 1819-1849.

[14] Davidson, J. (1994). *Stochastic Limit Theory, An Introduction for Econometricians*. Oxford University Press.

[15] Davidson, R. and MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford University Press.

[16] Davis, R. A., Holan, S. H., Lund, R. and Ravishanker, N., eds. (2016). *Handbook of Discrete-Valued Time Series*. CRC Press.

[17] Doukhan, P. (1994). *Mixing: Properties and Examples. Lecture Notes in Statistics* **85**. Springer-Verlag New York.

[18] 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.

[19] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models. *Journal of the American Statistical Association* **106** 544-557.

[20] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space w/ discussion. *J.Roy. Statist. Soc. B.* **70** 849-911.

[21] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20** 101-148.

[22] Fan, J., Lv, J. and Qi, L. (2011). Sparse High dimensional models in economics. *Annual Review of Economics* **3** 291-317.

[23] 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** 1270-1284.

[24] Fan, J. and Ren, Y. (2006). Statistical Analysis of DNA Microarray Data in Cancer Research. *Clinical Cancer Research* **12** 4469-4473.

[25] Fan, J. and Song, R. (2010). Sure Independence Screening in generalized linear models with NP-dimensionality. *Annals of Statistics* **38** 3567-3604.

[26] 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** 531-552.

[27] 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** 217-245.

[28] Hayashi, F. (2000). *Econometrics*. Princeton Univ Press.

[29] Huang, J., Ma, S. and Zhang, C. H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18** 1603-1618.

[30] Huang, Q. and Zhu, Y. (2016). Model-free sure screening via maximum correlation. *Journal of Multivariate Analysis* **148** 89-106.

[31] Johnstone, I. M. and Tetterington, M. (2009). Statistical challenges of high dimensional data. *Phil. Transactions of the Roy. Soc. A* **367** 4237-4253.

[32] Jurado, K., Ludvigson, S. C. and Ng, S. (2015). Measuring Uncertainty. *American Economic Review* **105** 1177-1216.

[33] Kock, A. and Callot, A. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics* **186** 325-344.

[34] 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**.

[35] Li, R., Zhu, L. P. and Zhong, W. (2012). Feature Screening via distance correlation. *J. Amer. Statist. Assoc.* **107** 1129-1139.

[36] Li, G., Peng, H., Jun, Z. and Zhu, L. (2012). Robust rank correlation based screening. *Annals of Statistics* **40** 1846-1877.

[37] Liu, J., Zhong, W. and Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Sci. China Math.* **58** 1-22.

[38] Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis.* Springer.

[39] 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.

[40] Samorodnitsky, G. (2006). Long Range Dependence. *Foundations and Trends in Stochastic systems* **1** 163-257.

[41] Shao, X. and Wu, W. B. (2007). Asymptotic spectral theory for nonlinear time series. *Ann. Statist.* **35** 1773-1801.

[42] Shao, X. and Zhang, J. (2014). Martingale Difference Correlation and Its Use in High-Dimensional Variable Screening. *Journal of the American Statistical Association* **109** 1302-1318.

[43] Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics* **44** 293-335.

[44] Stock, J. H. and Watson, M. W. (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics* **20** 147-162.

[45] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J.Roy. Statist. Soc. B.* **58** 267-288.

[46] Wang, H., Li, G. and Tsai, C. L. (2007). Regression coefficient and autoregressive order shrinkage and selection via lasso. *J.Roy. Statist. Soc. B.* **69** 63-78.

[47] 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.

[48] Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F. and Evans, A. C. (2002). A General Statistical Analysis for fMRI Data. *NeuroImage* **15** 1-15.

[49] Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences* **102** 14150-14154.

[50] Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and its Interface* **4** 207-226.

[51] Wu, W. B. and Min, W. (2005). On linear processes with dependent innovations. *Stochastic Processes and their Applications* **115** 939-958.

[52] Wu, W. B. and Pourahmadi, M. (2009). Banding sample autocovaraince matrices of stationary processes. *Statistica Sinica* **19** 1755-1768.

[53] Wu, W. B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics* **10** 352-379.

[54] 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** 1365-1387.

[55] Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *Ann. Statist.* **40** 466-493.

[56] Xu, P., Zhu, L. and Li, Y. (2014). Ultrahigh dimensional time course feature selection. *Biometrics* **70** 356-365.

[57] Zhu, L., Li, L., Li, R. and Zhu, L. X. (2011). Model-Free Feature Selection for Ultrahigh Dimensional Data. *J. Amer. Statist. Assoc.* **106** 1464-1475.

[58] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418-1429.