Partial Distance Correlation Screening for
High Dimensional Time Series *

Kashif Yousuf and Yang Feng

Abstract

High dimensional time series datasets are becoming increasingly common in various fields such as economics, finance, meteorology, and neuroscience. Given this ubiquity of time series data, it is surprising that very few works on variable screening discuss the time series setting, and even fewer works have developed methods which utilize the unique features of time series data. This paper introduces several model free screening methods based on the partial distance correlation and developed specifically to deal with time dependent data. Methods are developed both for univariate models, such as nonlinear autoregressive models with exogenous predictors (NARX), and multivariate models such as linear or nonlinear VAR models. Sure screening properties are proved for our methods, which depend on the moment conditions, and the strength of dependence in the response and covariate processes, amongst other factors. Dependence is quantified by functional dependence measures (Wu, 2005) and \( \beta \)-mixing coefficients, and the results rely on the use of Nagaev and Rosenthal type inequalities for dependent random variables. Finite sample performance of our methods is shown through extensive simulation studies, and we include an application to macroeconomic forecasting.

Keywords: Sure Independence Screening, Distance Correlation, Time Series, Variable Screening, Variable Selection, High Dimensionality.

*Kashif Yousuf, the corresponding author, is PhD Candidate (Email: ky2304@columbia.edu) and Yang Feng is Associate Professor (Email: yangfeng@stat.columbia.edu), Department of Statistics, Columbia University, New York, NY 10027. Yang Feng is partially supported by NSF Grant DMS-1554804.
1 Introduction

High dimensionality is an increasingly common characteristic of data being collected in fields as diverse as genetics, neuroscience, astronomy, finance, and macroeconomics. In these fields, we frequently encounter situations in which the number of candidate predictors \((p_n)\) is much larger than the number of samples \((n)\), and statistical inference is made possible by relying on the assumption of sparsity. The sparsity assumption, which states that only a small number of covariates contributes to the response, has led to a wealth of theoretical results and methods available for identifying important predictors in this high dimensional setting. These methods broadly fall into two classes: screening methods and penalized likelihood methods, and we focus on the screening approach in this work. For the case where \(p_n\) is much larger than \(n\), screening is more computationally feasible as a first stage method, which can be followed by a second stage method, such as penalized likelihood approaches, on the reduced subset of predictors selected at the screening stage.

Fan and Lv (2008) proposed Sure Independence Screening (SIS) for the linear model, and it is based on ranking the magnitudes of the marginal Pearson correlations of the covariates with the response. A large amount of work has been done since then to generalize the procedure to various other types of models including: generalized linear models (Fan and Song, 2010), nonparametric additive models (Fan et al., 2011), Cox proportional hazards model (Fan et al., 2010), linear quantile models (Ma et al., 2017), and varying coefficient models (Fan et al., 2014). Model-free screening methods, which do not assume any particular model a priori, have also been developed. Some examples include: a distance correlation based method in Li et al. (2012), the fused Kolmogorov filter in Mai et al. (2015), a conditional distance correlation method in Liu and Wang (2017), a method based on maximum correlation in Huang and Zhu (2016), a martingale difference based approach in Shao and Zhang (2014), and a smoothing bandwidth based method in Feng et al. (2017). For a partial survey of screening methods, one can consult Liu et al. (2015). The main theoretical result of these methods is the so called “sure screening property”, which states that under appropriate con-
ditions we can reduce the dimension of the feature space from size $p_n = O(\exp(n^\alpha))$ to a far smaller size $d_n$, while retaining all the relevant predictors with probability approaching 1.

Although there has been a large amount of interest in developing screening methods, it is surprising to see that almost all of the works operate under the assumption of independent observations. This is even more surprising given the ubiquity of time dependent data in many scientific disciplines. Data in fields such as climate science, neuroscience, political science, economics, and finance are frequently observed over time and/or space thereby exhibiting serial dependence. A specific example is in forecasting low frequency macroeconomic indicators such as GDP or inflation rate, where we can have a large number of macroeconomic and financial time series and their lags as possible covariates. Another example is identification of brain connectivity networks, where we have data from thousands of voxels collected over a relatively small number of time periods (Valdés-Sosa et al., 2005). These examples, amongst others, highlight the importance of developing screening methods for time dependent data.

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

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

Additionally, we aim to have a model free screening approach which can handle continuous, discrete or grouped time series. Using a model free approach makes our methods robust to model misspecification at the screening stage, and gives us full flexibility when considering a second stage procedure. The few works which have relaxed the assumption of independent observations include Cheng et al. (2014), and Xu et al. (2014) which dealt with longitudinal data. However, the dependence structure of longitudinal data is too restrictive to cover the
type of dependence present in most time series. To the best of our knowledge there have been only two works, Chen et al. (2017) and Yousuf (2018), dealing with the issue in a general stationary time series setting. The former work extended the nonparametric independence screening approach used for independent observations to the time series setting. However, the method does not utilize the serial dependence in the data, or account for the unique properties of time series data we outlined. The latter work (Yousuf, 2018) extended the theory of SIS to heavy tailed and/or dependent data as well as proposing a GLS based screening method to correct for serial correlation. However, this work is limited to the linear model and the other unique qualities of time series data outlined above are ignored. Additionally both of these works are only applicable to models with a univariate response.

In order to account for the unique characteristics of time series data mentioned above, and correct some of the limitations in previous works, we will introduce several distance correlation based screening procedures. Distance correlation (DC) was introduced by Székely et al. (2007), for measuring dependence and testing independence between two random vectors. The consistency, and weak convergence of sample distance correlation has been established for stationary time series in Zhou (2012) and Davis et al. (2016). DC has a number of useful properties such as:

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

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

Compared to the previous works on screening using distance correlation based methods
(Li et al., 2012; Liu and Wang, 2017), our work differs in a number of ways. First, our work deals with the time series setting, where both the covariates and response are stationary time series, and can be heavy tailed. Second, our screening procedures are developed specifically in order to account for certain unique features in time series data mentioned previously. Lastly, we choose to rely on partial DC, instead of conditional DC, when controlling for confounding variables. Partial DC is a DC based procedure which can be easily computed using pairwise distance correlations, whereas the computation of conditional DC is more involved and involves the choice of a bandwidth parameter, which can be difficult to choose.

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

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

The rest of the paper is organized as follows. Section 2 reviews the functional dependence measure and $\beta$ mixing coefficients, as well as comparisons between the two frameworks. We also discuss the assumptions placed on structure of the covariate and response processes. Section 3 introduces our screening procedures with their sure screening properties for models with a univariate response. Section 4 presents screening algorithms for multivariate response models. Section 5 covers simulation results, and a real data application is presented in Section 6. The concluding remarks are in Section 7. Lastly, the proofs for all theorems, along with additional simulations and data analysis results for sections 5 and 6 are placed in the supplementary material.

2 Dependence Measures

In order to establish asymptotic properties, we rely on two widely used dependence measures, the functional dependence measure and $\beta$-mixing coefficients. We first start with an overview of the functional dependence measure framework, before proceeding to $\beta$-mixing processes.

For univariate processes, $(Y_i \in \mathcal{R})_{i \in \mathbb{Z}}$, we assume $Y_i$ is a causal, strictly stationary, ergodic process with the following form:

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

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

\[ x_i = h(\ldots, \eta_{i-1}, \eta_i). \]  

(2)

Where \( \eta_i, i \in \mathbb{Z} \), are iid random vectors, \( h(\cdot) = (h_1(\cdot), \ldots, h_{pn}(\cdot)) \), \( x_i = (X_{i1}, \ldots, X_{ipn}) \), and \( X_{ij} = h_j(\ldots, \eta_{i-1}, \eta_i) \).

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

\[
\delta_q(Y_i) = ||Y_i - g(F_i^*)||_q = (E|Y_i - g(F_i^*)|^q)^{1/q},
\]

\[
\delta_q(X_{ij}) = ||X_{ij} - h_j(H_i^*)||_q = (E|X_{ij} - h_j(H_i^*)|^q)^{1/q},
\]

(3)

where \( F_i^* = (\ldots, e_{-1}, e_0^*, e_1, \ldots, e_i) \) with \( e_0^*, e_j, j \in \mathbb{Z} \) being iid. And for the multivariate case, \( H_i^* = (\ldots, \eta_{-1}, \eta_0^*, \eta_1, \ldots, \eta_i) \) with \( \eta_0^*, \eta_j, j \in \mathbb{Z} \) being iid. Since we are replacing \( e_0 \) by \( e_0^* \), we can think of this as measuring the dependency of \( y_i \) on \( e_0 \), since we are keeping all other inputs the same. We assume the cumulative functional dependence measures are finite:

\[
\Delta_{0,q}(y) = \sum_{i=0}^{\infty} \delta_q(Y_i) < \infty, \text{ and } \Phi_{m,q}(x) = \max_{j \leq pn} \sum_{i=m}^{\infty} \delta_q(X_{ij}) < \infty.
\]

(4)

This short range dependence condition implies, by the proof of theorem 1 in \( \text{Wu and Pourahmadi (2009)} \), the auto-covariances are absolutely summable.

There are many equivalent definitions given for \( \beta \)-mixing, and we use the one provided by \( \text{Doukhan (1994)} \):
Definition 2.1. Given a stationary multivariate process, \((x_i)_{i \in \mathbb{Z}}\), for each positive integer \(a\), the coefficient of absolute regularity or \(\beta\)-mixing coefficient, \(\beta(a)\), is:

\[
\beta(a) = \|P_{-\infty:0} \times P_{a:\infty} - P_{-\infty:0,a:\infty}\|_{TV}.
\]

Where \(\|\cdot\|_{TV}\) is the total variation norm, and \(P_{-\infty:0,a:\infty}\) is the joint distribution of the blocks \((x_{-\infty:0}, x_{a:\infty})\). A stochastic process is said to be \(\beta\)-mixing if \(\beta(a) \to 0\).

We note that compared to functional dependence measures, \(\beta\)-mixing coefficients can be defined for any stochastic processes, and are not limited to Bernoulli shift processes. On other hand, functional dependence measures are easier to interpret and compute since they are related to the data generating mechanism of the underlying process. In many cases using the functional dependence measure also requires less stringent assumptions (see Wu and Wu (2016), Yousuf (2018) for details). Although there is no direct relationship between these two dependence frameworks, fortunately there are a large number of commonly used time series processes which are \(\beta\)-mixing and satisfy (4). For example, under appropriate conditions, linear processes, ARMA, GARCH, ARMA-ARCH, threshold autoregressive, Markov chain models, amongst others, can be shown to be \(\beta\)-mixing (see Pham and Tran (1985), Carrasco and Chen (2002), An and Huang (1996), Lu (1998) for details).

3 Partial Distance Correlation Screening for Univariate Response Models

3.1 Preliminaries

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

Definition 3.1. For any random vectors \(u \in \mathcal{R}^q, v \in \mathcal{R}^p\), let \(\phi_u(t), \phi_v(s)\) be the character-
istic function of \( u \) and \( v \) respectively. The distance covariance between \( u \) and \( v \) is defined as Székely et al. (2007):

\[
dcov^2(u, v) = \int_{\mathbb{R}^{p+q}} |\phi_{u,v}(t, s) - \phi_u(t)\phi_v(s)|^2 \omega^{-1}(t,s) dt ds,
\]

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

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

\[
S_1 = E(|u - u'|_p|v - v'|_q), \quad S_2 = E(|u - u'|_p)E(|v - v'|_q), \quad S_3 = E(|u - u'|_p)E(|v - v''|_q).
\]

Then, by remark 3 in Székely et al. (2007), \( dcov^2(u, v) = S_1 + S_2 - 2S_3 \). We can now estimate this quantity using moment based methods. Suppose we observe \((u_i, v_i)_{i=1,\ldots,n}\), the sample estimates for the distance covariance and distance correlation are:

\[
\hat{dcov}^2(u, v) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3, \quad \hat{dcor}(u, v) = \frac{\hat{dcov}(u, v)}{\sqrt{\hat{dcov}(u, u)\hat{dcov}(v, v)}},
\]

where \( \hat{S}_1 = n^{-2} \sum_{i,j=1}^n |u_i - u_j|_p|v_i - v_j|_q \), \( \hat{S}_2 = n^{-2} \sum_{i,j=1}^n |u_i - u_j|_p n^{-2} \sum_{i,j=1}^n |v_i - v_j|_q \), and \( \hat{S}_3 = n^{-3} \sum_{i,j,l=1}^n |u_i - u_j|_p |v_i - v_l|_q \).

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

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

if $dcor(u, \mathcal{Z}), dcor(v, \mathcal{Z}) \neq 1$, otherwise $pdccor(u, v; \mathcal{Z}) = 0$. For more details and an interpretation of PDC, one can consult Székely and Rizzo (2014).

### 3.2 Screening Algorithm I: PDC-SIS

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

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

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

When developing screening algorithms for time series data, we would like to account for some of its unique properties as mentioned in the introduction. For models with a response, these would be:
• A prior belief that a certain number of lags of the response variable are to be in the model.

• An ordered structuring of the covariates, in which lower order lags of covariates are thought to be more informative than higher order lags.

The first property can be easily accounted for using partial distance correlation, while there are many different ways to account for the second property. In this section we present two partial distance correlation based screening algorithms, which attempt to account for the ordered structure of our covariates. In our first algorithm, PDC-SIS, we define the conditioning vector for the $l^{th}$ lag of predictor series $k$ as:

$$S_{k,l} = (Y_{t-1}, \ldots, Y_{t-h}, X_{t-1,k}, \ldots, X_{t-l+1,k}),$$

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

For convenience, let $C = \{S_{1,1}, \ldots, S_{m_n,1}, S_{1,2}, \ldots, S_{m_n,h}\}$ denote our set of conditioning vectors, where $C_{k+(l-1)*m_n} = S_{k,l}$ is the conditioning vector for covariate $Z_{t-1,(l-1)*m_n+k}$. Our screened sub-model is:

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

To establish sure screening properties, we introduce the following conditions.
Condition 3.1. Assume $|pdcor(Y_t, Z_{t-1,k}; C_k)| \geq c_1 n^{-\kappa}$ for $k \in M_*$ and $\kappa \in (0, 1/2)$.

Condition 3.2. Assume the response and the covariate processes have representations (1) and (A.1), respectively. Additionally, we assume the following decay rates $\Phi_{m,r}(x) = O(m^{-\alpha_x}), \Delta_{m,q}(y) = O(m^{-\alpha_y})$, for some $\alpha_x, \alpha_y > 0, q > 2, r > 4$ and $\tau = \frac{qr}{q+r} > 2$.

Condition 3.3. Assume the response and the covariate processes have representations (1) and (A.1) respectively. Additionally assume

$\upsilon_z = \sup_{q \geq 2} q^{-\tilde{\alpha}_z} \Phi_{0,q}(x) < \infty$ and $\upsilon_y = \sup_{q \geq 2} q^{-\tilde{\alpha}_y} \Delta_{0,q}(y) < \infty$, for some $\tilde{\alpha}_z, \tilde{\alpha}_y \geq 0$.

Condition 3.4. Assume the process $\{(Y_t, x_t)\}$ is $\beta$-mixing, with mixing rate $\beta_{xy}(a) = O(\exp(-a^{\lambda_1}))$, for some $\lambda_1 > 0$.

Condition 3.1 is a standard population level assumption which allows covariates in the active set to be detected by our screening procedure. Condition 1.2 is similar to the one used in Yousuf (2018) and Wu and Wu (2016), and assumes both the response and covariate processes are causal Bernoulli shift processes. Additionally it presents the dependence and moment conditions on these processes, where higher values of $\alpha_x, \alpha_t$ indicate weaker temporal dependence. Examples of response processes which satisfy condition 1.2 include stationary, causal, finite order ARMA, GARCH, ARMA-GARCH, bilinear, and threshold autoregressive processes, all of which have exponentially decaying functional dependence measures (see Wu (2011) for details). For the covariate process, assume $x_i$ is a vector linear process: $x_i = \sum_{l=0}^{\infty} A_l \eta_{i-l}$, where $\{A_l\}$ are $m_n \times m_n$ coefficient matrices and $\{\eta_i = (\eta_{i1}, \ldots, \eta_{im_n})\}$ are iid random vectors with $\text{cov}(\eta_i) = \Sigma_{\eta}$. For simplicity, assume $\{\eta_{i,j}, j = 1, \ldots, m_n\}$ are identically distributed, then $\delta_q(X_{ij}) = ||A_{i,j} \eta_0 - A_{i,j} \eta_0^\ast||_q \leq 2 |A_{i,j}|||\eta_0,1||_q$, where $A_{i,j}$ is the $j^{th}$ column of $A_i$. If $||A_i||_\infty = O(i^{-\beta})$ for $\beta > 1$, then $\Phi_{m,q}(x) = O(m^{-\beta+1})$. Other examples include stable VAR processes, and multivariate ARCH processes which have exponentially decaying cumulative functional dependence measures (Wu and Wu, 2016; Yousuf, 2018). Condition 1.3 strengthens the moment requirements of condition 1.2, and requires that all moments of the covariate and response processes are finite. To illustrate the role of the
constants \( \tilde{\alpha}_z \) and \( \tilde{\alpha}_y \), consider the example where \( y_t \) is a linear process: 
\[
y_t = \sum_{j=0}^{\infty} f_j e_{i-j}
\]
with \( e_i \) iid and \( \sum_{i=0}^{\infty} |f_i| < \infty \), then \( \Delta_{0,q}(\mathbf{y}) = \|e_0 - e_0^*\|_q \sum_{i=0}^{\infty} |f_i| \). If we assume \( e_0 \) is sub-Gaussian, then \( \tilde{\alpha}_y = 1/2 \), since \( \|e_0\|_q = O(\sqrt{q}) \). Similarly, if \( e_i \) is sub-exponential, we have \( \tilde{\alpha}_y = 1 \).

To understand the inclusion of condition 1.4, consider the \( U \)-statistic:
\[
U_r(S_{t_1}, \ldots, S_{t_r}) = \binom{n}{r} \sum_{t_1 \leq t_2 \leq \ldots \leq t_r \leq n} h(S_{t_1}, \ldots, S_{t_r}),
\]
which aims to estimate \( \theta(h) = \int h(S_{t_1}, \ldots, S_{t_r})d\mathcal{P}(S_1) \ldots d\mathcal{P}(S_r) \). When \( S_1, \ldots, S_n \) are iid, the \( U \)-statistic is an unbiased estimator of \( \theta(h) \), however for \( r > 1 \) the \( U \)-statistic is no longer unbiased if \( S_t \) is serially dependent. Since our sample distance correlation estimate can be written as a sum of \( U \)-statistics (Li et al., 2012), condition 1.4 is needed to control the rate at which the above bias vanishes as \( n \to \infty \). Conditions 1.2 and 1.4 are frequently used when dealing with time series data (Wu and Pourahmadi, 2009; Xiao and Wu, 2012; Davis et al., 2016).

Throughout this paper, let \( \alpha = \min(\alpha_x, \alpha_y) \), and \( q = 1 \), if \( \alpha_z > 1/2 - 2/r \), otherwise \( q = r/4 - \alpha_z r/2 \). Let \( \iota = 1 \) if \( \alpha > 1/2 - 1/\tau \), otherwise \( \iota = \tau/2 - \tau \alpha \), and let \( \zeta = 1 \), if \( \alpha_y > 1/2 - 2/q \), otherwise \( \zeta = q/4 - \alpha_y q/2 \). Additionally, let \( K_{y,q} = \sup_{m \geq 0}(m+1)^{\alpha_y} \Delta_{m,q}(\mathbf{y}) \), and \( K_{z,r} = \sup_{m \geq 0}(m+1)^{\alpha_z} \Phi_r(\mathbf{x}) \). Given condition 1.3, it follows that \( K_{x,q}, K_{z,r} < \infty \). Let \( t_n = \max_j \dim(C_j) \), be the maximum dimension of the conditional vectors. We define
\[
\hat{\psi} = \frac{2}{1+2\tilde{\alpha}_x+2\tilde{\alpha}_y}, \tilde{\phi} = \frac{2}{1+4\tilde{\alpha}_x}, \tilde{\alpha} = \frac{2}{1+4\tilde{\alpha}_y}.
\]
Lastly, for ease of presentation, let \( \hat{\omega} = (\hat{\omega}_1, \ldots, \hat{\omega}_p) \), \( \omega = (\omega_1, \ldots, \omega_p) \), where \( \omega_k = \text{pdcor}(Y_t, Z_{t-1,k}; C_k) \), \( \hat{\omega}_k = \widehat{\text{pdcor}}(Y_t, Z_{t-1,k}; C_k) \). In addition, let
\[
a_n = n^2 \left[ \exp \left( \frac{-n^{1/2-\kappa}}{t_n v^2_y} \right) \tilde{\alpha} + \exp \left( \frac{-n^{1/2-\kappa}}{t_n v^2_z} \right) \hat{\psi} + \exp \left( \frac{-n^{1/2-\kappa}}{t_n v^2_z} \right) \tilde{\phi} \right],
\]
\[
b_n = n^2 \left[ \frac{t_n^{r/2} n^r K_{y,r}^{r/2}}{n^{r/2-r\kappa/2}} + \frac{t_n^{r/2} n^r K_{z,r}^{r/2}}{n^{r/2-r\kappa/2}} + \frac{t_n^{r/2} n^{\kappa} K_{r,r}^{r/2}}{n^{r/2-r\kappa/2}} \right].
\]
\[ + \exp \left( -\frac{n^{1-2\kappa}}{t_n^2 K_{z,r}^4} \right) + \exp \left( -\frac{n^{1-2\kappa}}{t_n^2 K_{y,r}^2 K_{z,r}^2} \right) + \exp \left( -\frac{n^{1-2\kappa}}{t_n^2 K_{y,r}^4} \right), \]

\[ c_n = \frac{t_n^{r/2}}{n^{r/4-r\kappa/2}} K_{y,r}^{r/2} + \frac{t_n^{r/2}}{n^{r/4-r\kappa/2}} K_{z,r}^{r/2} + \frac{t_n^{r/2}}{n^{r/4-r\kappa/2}} K_{z,r}^{r/2}. \]

For simplicity and convenience of presentation, we assume \( q = r \), and one can consult the proof for the general case. The following theorem presents the sure screening properties of PDC-SIS for both the heavy tailed and light tailed settings.

**Theorem 1.**

1. Suppose conditions 3.1, 1.3, and 1.4 hold. For any \( c_2 > 0 \), we have:

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

2. Suppose conditions 3.1, 1.3, and 1.4 hold. For \( \gamma_n = c_3 n^{-\kappa} \) with \( c_3 \leq c_1/2 \), we have:

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

3. Suppose conditions 3.1, 1.2, and 1.4 hold. For any \( c_2 > 0 \), we have:

\[
\text{if } r < 12, \quad P(\max_{j \leq p_n} |\hat{\omega}_j - \omega_j| > c_2 n^{-\kappa}) \leq O(p_n c_n); \\
\text{if } r \geq 12, \quad P(\max_{j \leq p_n} |\hat{\omega}_k - \omega_k| > c_2 n^{-\kappa}) \leq O(p_n b_n).
\]

4. Suppose conditions 3.1, 1.2, and 1.4 hold. For \( \gamma_n = c_3 n^{-\kappa} \) with \( c_3 \leq c_1/2 \), we have:

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

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

For example, if we assume only finite polynomial moments with $r = q$ and $r < 12$, then $p_n = o(n^{r/4-rn/2})$. If we assume $\alpha \geq 1/2 - 2/r$ and $r > 12$, $p_n = o(n^{r/2-rn/2-3})$. The constants $K_{z,r}$ and $K_{y,q}$, which are related to the cumulative functional dependence measures, represent the effect of temporal dependence on our bounds when $\alpha \geq 1/2 - 2/r$. However, when using Nagaev type inequalities, there is an additional effect in the case of stronger dependence in the response or covariate process (i.e. $\alpha < 1/2 - 2/r$). For instance, if $\alpha_x = \alpha_\epsilon$ and $q = r$, the range for $p_n$ is reduced by a factor of $n^{r/4-ar/2}$ in the case of stronger dependence. For the case of exponentially decaying tails however, there is no level shift in the decay rate of our bounds due to the dependence of the response or covariate processes. We observe that if the response and covariates are sub-Gaussian, $p_n = o(n^{1-2\alpha/3})$, and if they are sub-exponential, $p_n = o(n^{1-2\alpha})$.

By choosing an empty conditional set for all the variables, our procedure reduces to the distance correlation screening (DC-SIS) introduced in Li et al. (2012) for the iid setting. Assuming sub-Gaussian response and covariates, Li et al. (2012) obtained $p_n = o(n^{1-2\alpha})$ for DC-SIS, which matches our rate. In the iid setting with finite polynomial moments, we can use the truncation method in their proof and combined with the Markov inequality to obtain $p_n = o(n^{r/4-rn/2-1})$. Our results, which rely on a different proof strategy than the truncation method, provide a better bound even in this setting.
3.3 Screening Algorithm II: PDC-SIS+

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

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

\[
\hat{U}_{1}^{\lambda_n} = \left\{ j \in \{1, \ldots, m_n\} : \left[pdcor(Y_t, Z_{t-1,j}; S_{j,1})\right] \geq \lambda_n + c_1 n^{-\kappa} \right\}.
\]

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

\[
\hat{S}_{k,2} = \left( Y_{t-1}, \ldots, Y_{t-h}, X_{t-1,k}, z_{t-1,\hat{U}_{1}^{\lambda_n}} \right),
\]

where \( z_{t-1,\hat{U}_{1}^{\lambda_n}} \) is a sub-vector of \( z_{t-1} \) which is formed by extracting the indices contained in \( \hat{U}_{1}^{\lambda_n} \). We note that any duplicates which result from overlap between \( X_{t-1,k} \) and \( z_{t-1,\hat{U}_{1}^{\lambda_n}} \) are deleted. For convenience, we define \( \hat{C} = (\hat{S}_{1,1}, \ldots, \hat{S}_{m_n,1}, \hat{S}_{1,2}, \ldots, \hat{S}_{m_n,h}) \) as our vector of estimated conditional sets. We then use \( (\hat{S}_{k,2})_{k \leq m_n} \) to compute the strong conditional signal
set for the 2nd lag level:

\[
\hat{U}_2^{\alpha_n} = \left\{ j \in \{m_n + 1, \ldots, 2m_n\} : |\widehat{pdcor}(Y_t, Z_{t-1,j}; \hat{C}_j)| \geq \lambda_n + c_1 n^{-\kappa} \right\}.
\]

Repeating this procedure we obtain:

\[
\hat{S}_{k,l} = \left( Y_{t-1}, \ldots, Y_{t-h}, X_{t-1,k}, \ldots, X_{t-l+1,k}, z_{t-1,\hat{U}_1^{\lambda_n}}, \ldots, z_{t-1,\hat{U}_{l-1}^{\lambda_n}} \right).
\]

We can also vary the threshold \( \lambda_n \) for each lag level; for simplicity we leave it the same for each of our levels here. Our sub-model obtained from this procedure is:

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

The asymptotic properties of this procedure are similar to PDC-SIS, and we present them in the supplementary material. To show the asymptotic properties associated with this algorithm, we denote

\[
\hat{S}_{k,l} = \left( Y_{t-1}, \ldots, Y_{t-h}, X_{t-1,k}, \ldots, X_{t-l+1,k}, z_{t-1,\hat{U}_1^{\lambda_n}}, \ldots, z_{t-1,\hat{U}_{l-1}^{\lambda_n}} \right),
\]

as the population level counterpart to \( \hat{S}_{k,l} \). In addition, let \( C = \{S_{1,1}, \ldots, S_{m_n,1}, S_{1,2}, \ldots, S_{m_n,h}\} \), and

\[
U_{l-1}^{\lambda_n} = \left\{ (l-1)m_n + 1 \leq j \leq lm_n : |\widehat{pdcor}(Y_t, Z_{t-1,j}; C_j)| \geq \lambda_n + \frac{c_2}{2} n^{-\kappa} \right\},
\]

represent the population level strong conditional signal set and the population level set of conditioning vectors, respectively. One of the difficulties in proving uniform convergence of our estimated partial distance correlations in this algorithm is the presence of an estimated conditioning set \( \hat{C} \). This issue becomes compounded as we estimate the conditioning vector for higher lag levels, since these rely on estimates of the conditioning vectors for lower ones.
To overcome this, we first denote the collection of strong signals from lag 1 to \( h - 1 \) as: 
\[
U^\lambda_n = \{U_1^\lambda_n, \ldots, U_{h-1}^\lambda_n\}.
\]
We will assume the following condition:

**Condition 3.5.** For any \( j \in \{1, \ldots, (h - 1) \ast m_n\} \setminus U^\lambda_n \), assume
\[
|pdcor(Y_t, Z_{t-1}; C_j)| \leq \lambda_n, \text{ where } \lambda_n n^\kappa \to \infty.
\]

Condition 3.5 assumes the variables in the strong conditional signal set, \( U^\lambda_n \), are easily identifiable from the rest of the covariates. This separation in the signal strength will allow us to ensure with high probability that our estimated conditional sets match their population level counterparts. The assumption \( \lambda_n n^\kappa \to \infty \), is introduced to ensure \( d_n = |\hat{M}_{\gamma_n}| \gg |U^\lambda_n| \).

Although the hope is that \( U^\lambda_n \subset M_* \), this is not required to prove sure screening properties of our algorithm. Additionally, as seen in Barut et al. (2016) for the case of generalized linear models, conditioning on irrelevant variables could also enhance the power of a screening procedure. We will discuss how to choose the threshold \( \Gamma_n = \lambda_n + c_1 n^{-\kappa} \) for \( \hat{U}^\lambda_n \) in section 3.4. In practice we would prefer not to condition on too many variables, therefore the threshold for adding a variable to \( U^\lambda_n \) would be high.

The sure screening properties for PDC-SIS+ are similar to PDC-SIS, but for the sake of completeness, we state the theorem in full.

**Theorem 2.** 1. Suppose conditions 3.1, 1.3, 1.4, and 3.5 hold. For \( \gamma_n = c_3 n^{-\kappa} \) with \( c_3 \leq c_1/2 \), we have
\[
P \left( M_* \subset \hat{M}_{\gamma_n} \right) \geq 1 - O(s_n a_n).
\]

2. Suppose conditions 3.1, 1.2, 1.4, and 3.5 hold. For \( \gamma_n = c_3 n^{-\kappa} \) with \( c_3 \leq c_1/2 \), we have
\[
\begin{align*}
\text{if } r < 12, & \quad P \left( M_* \subset \hat{M}_{\gamma_n} \right) \geq 1 - O(s_n c_n); \\
\text{if } r \geq 12, & \quad P \left( M_* \subset \hat{M}_{\gamma_n} \right) \geq 1 - O(s_n b_n).
\end{align*}
\]

Now, we have presented two classes of PDC screening methods. In the first class of methods, the conditional set of each covariate is known as *a priori*, while in the second class
the conditional set is estimated from the data. We can easily modify our algorithms for both procedures depending on the situation; for example we can screen groups of lags at a time for certain covariates in PDC-SIS. Additionally, for either procedure we can condition on a small number of lags of \(Y_t\), and leave the higher order lags of \(Y_t\) as possible covariates in our screening procedure.

### 3.4 Threshold Selection: PDC-SIS+

In order to implement PDC-SIS+, we need to select a threshold parameter \(\Gamma_n\). For simplicity we will only use a single threshold for all lag levels, and it is selected as follows: we first generate 1000 independent AR(1) variables, \(\xi_{t,1}, \ldots, \xi_{t,1000}\) where \(\xi_{t,i} = \beta_1 \xi_{t-1,i} + \theta_t \xi_t^{iid} \sim N(0,1)\), and \(\{\xi_{t,i}, i = 1, \ldots, 1000\}\) are independent of our response. We set \(\beta_1 = .4\), and estimate:

\[
\hat{\Upsilon} = (\hat{\Upsilon}_1, \ldots, \hat{\Upsilon}_{1000}), \text{ where } \hat{\Upsilon}_k = \hat{\text{pdcor}}(Y_t, \xi_{t,k}; Y_{t-1}, Y_{t-2}, Y_{t-3}).
\]  

We then set the 99th percentile of \(\hat{\Upsilon}\) as our estimated threshold parameter \(\hat{\Gamma}_n\). In order to avoid conditioning on too many variables, following Weng et al. (2017), we set an upper bound of \([n^{1/2}]\) variables which can be added to our conditioning vector at each lag level. Given that \(Y_t\) and \(\xi_{t,i}\) are independent, we have that \(\hat{\Upsilon}_i\) is a statistical estimate of zero. This procedure is similar to the random decoupling approach used in Weng et al. (2017) and Barut et al. (2016) for the iid setting. We note that we could also use a more computationally intensive alternative approach which involves forming a pseudo-sample \(\{Y_i, z_i^*\}_{i=1}^n\), where \((z_1^*, \ldots, z_n^*)\) is formed using a moving block bootstrap (Kunsch, 1989) or a stationary bootstrap (Politis and Romano, 1994). We then use the procedure outlined in Barut et al. (2016), using the pseudo-sample \(\{Y_i, z_i^*\}_{i=1}^n\), to select our threshold. We choose to go with the first approach due to computational efficiency.
4 Screening for Multivariate Time Series Models

Multivariate time series models, such as linear VAR models, are commonly used in fields such as macroeconomics (Lütkepohl, 2005), finance, and more recently neuroscience (Valdés-Sosa et al., 2005), and genomics. VAR models provide a convenient framework for forecasting, investigating Granger causality, and modeling the temporal and cross-sectional dependence for large numbers of series. Since the number of parameters grows quadratically with the number of component series, VAR models have traditionally been restricted to situations where the number of component series is small. One way to overcome this limitation is by assuming a sparse structure in our VAR process, and using penalized regression methods such as the Lasso and adaptive Lasso (Zou, 2006) to estimate the model. Examples of works which pursue this direction include Basu and Michailidis (2015), Basu et al. (2015), Kock and Callot (2015), and Nicholson et al. (2016). However, due to the quadratically increasing nature of the parameter space, penalized regression methods can quickly become computationally burdensome when we have a large panel of component series. For example, in a VAR($k$) process: $x_t = \sum_{i=1}^{k} B_i x_{t-i} + \eta_t$, where $x_t \in \mathcal{R}^{m_n}$, $m_n = 1000$, $k = 5$, the number of parameters to estimate is $5 \times 10^6$. Additionally, these methods are restricted to linear VAR models, whereas there is considerable evidence of non-linear effects such as the existence of thresholds, smooth transitions, regime switching, and varying coefficients in fields such as macroeconomics and finance (Kilian and Lütkepohl, 2017).

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

$$x_t = g(x_{t-1}, \ldots, x_{t-k}) + \eta_t, \text{ where } x_t \in \mathcal{R}^{m_n}, \eta_t iid.$$
For simplicity, we let all groups be of size $g_n$, let $e_n = m_n/g_n$ denote the total number of groups for a given lag level, and denote our groups $(G_{t-1,1}, \ldots, G_{t-k, e_n})$. To get a sense of the computational benefits of screening on the group level, assume for example, $m_n = 500$, $k = 1$, and we have 25 groups all of size $g_n = 20$. For this linear VAR (1) model, when $n = 200$, we note it takes about 350 times longer to compute all $m_n^2 = 500^2$ pairwise distance correlations \( \{ \hat{\text{dcor}}(X_{t,j}, X_{t-1,k}) \}_{j \leq m_n, k \leq m_n} \) vs. computing all $e_n^2 = 25^2$ group pairwise distance correlations. After the group screening, examples of second stage procedures include: screening at the individual series level using partial distance correlations, or using a group lasso type procedure (Yuan and Lin, 2006) which can handle sparsity between groups and within groups for a linear VAR model (Basu et al., 2015).

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

\[
\mathcal{M}^{(i)} = \{(i, k, j) \in A^{(i)} : F\left(G_{t,i} \mid G_{t-1,i}, \bigcup_{r=t-h}^{t-1} \{G_r\}_{l \leq e_n}\right) \text{ functionally depends on } G_{k,j}\}.
\]

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

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

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

5 Simulations

5.1 Univariate response models: PDC-SIS

In the first two subsections, we evaluate the performance of PDC-SIS and PDC-SIS+. We also include the performance of 4 other screening methods whose properties have been investigated in the time series setting, these include: marginal correlation screening (SIS), nonparametric independence screening (NIS), generalized least squares screening (GLSS), and distance correlation screening (DC-SIS). The NIS estimator is computed using the R package `mgcv`, and the distance and partial distance correlation estimators are computed using the R package `energy`. For computational efficiency, the GLSS estimator is computed using the R package `nlme` package using an AR(1) approximation for the residual covariance matrix. Simulations for our group PDC-SIS procedure are contained in the supplementary material.

Unless noted otherwise, we fix our sample size \(n = 200\), maximum number of lags considered \(h = 3\), and the conditioning vector always includes three lags of our response. We vary the number of candidate series, \(m_n\), from 500 to 1500, so the number of total covariates, \(p_n\), varies from 1500 to 4500. We repeat each experiment 200 times, and report the median minimum model size needed to include all the relevant covariates from \(z_{t-1} = (x_{t-1}, \ldots, x_{t-3})\). We note that for all procedures being considered, we will not be screening the lags of \(Y_t\). In the supplementary materials, we also report the median rank of our relevant covariates for each procedure. We set \(Y_0 = Y_{-1} = \ldots = Y_{-(h+1)} = 0\), and generate \(n + 200\) samples of our
model. We then discard the first 200 − h samples. To ensure stationarity when generating a nonlinear autoregressive model with exogenous predictors (NARX), we use the sufficient conditions provided in Masry and Tjøstheim (1997).

Model 1:

\[ Y_t = \sum_{j=1}^{6} \beta_j X_{t-1,j} + \epsilon_t, \text{ and } x_t = A_1 x_{t-1} + \eta_t, \]

where \( A_1 = .6 * I \), and \( \eta_t \overset{iid}{\sim} N(0, \Sigma_\eta) \), or \( \eta_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta) \). For this model, we set \( \Sigma_\eta = \{.3^{|i-j|}\}_{i,j \leq m_n} \). For the error process, we have an AR(1) process: \( \epsilon_t = \alpha \epsilon_{t-1} + \epsilon_t \) where \( \alpha = .6 \), and let \( \epsilon_t \overset{iid}{\sim} N(0, 1) \) or \( \epsilon_t \overset{iid}{\sim} t_5 \). This is a linear model with no autoregressive terms, therefore our conditional set contains irrelevant predictors which are the 3 lags of \( Y_t \).

The results are displayed in table 3, and the entries below “Gaussian” correspond to the setting where both \( \epsilon_t \) and \( \eta_t \) are drawn from a Gaussian distribution. Accordingly the entries under “t5” correspond to the case where \( \epsilon_t \) and \( \eta_t \) are drawn from a t5 distribution. We see that all methods perform well in this scenario, with GLSS performing best and PDC-SIS following closely even though the lags of \( Y_t \) are not significant variables in this example. The results also show that the effects of heavy tails deteriorates the performance of all methods for this model.

Model 2:

\[ Y_t = g_1(Y_{t-1}) + g_2(Y_{t-2}) + g_3(Y_{t-3}) + f_1(X_{t-1,1}) + f_2(X_{t-2,1}) + f_3(X_{t-1,2}) + f_4(X_{t-2,2}) + \epsilon_t, \]

where the functions are defined as:

\[
\begin{align*}
g_1(x) &= .25x, \quad g_2(x) = x \exp(-x^2/2), \quad g_3(x) = -.6x + .3x(x > 0), \\
f_1(x) &= 1.5x + .4x(x > 0), \quad f_2(x) = -x, \quad f_3(x) = 1.2x + .4x(x > 0), \\
f_4(x) &= x^2 \sin(2\pi x).\end{align*}
\]
The covariate process is generated as in (7), with $A_1 = \{0.4^{i-j+1}\}_{i,j \leq m}$ and we set $\Sigma_\eta = I_m$, with $\eta_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\eta_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$. Additionally, we set $\epsilon_t \overset{iid}{\sim} N(0, 1)$ or $t_5$. The nonlinear transformations used are mainly threshold or smooth threshold functions which are popular nonlinear transformations for time series data (Teräsvirta et al., 2010). Note that when $x$ is close to zero, $g_2(x) = x$, and when $|x|$ is far from zero the function is close to zero. The results are shown in table 3, and our method clearly outperforms the other methods across all scenarios. As seen in table 4 of the supplementary file, the covariate $X_{t-2.1}$ seems to be the most difficult to detect for the competing methods, and it appears our conditioning scheme greatly improves the detection of this signal.

Model 3:

$$Y_t = g_1(Y_{t-1}) + g_2(Y_{t-2}, Y_{t-1}) + g_3(Y_{t-3}, Y_{t-1}) + f_1(X_{t-1.1}, X_{t-1.4}) + f_2(X_{t-2.1}, X_{t-1.4}) + f_3(X_{t-1.2}, X_{t-1.4}) + f_4(X_{t-2.2}, X_{t-1.4}) + f_5(X_{t-1.3}, X_{t-1.4}) + f_6(X_{t-1.4}) + f_7(X_{t-1.3}, X_{t-1.4}) + \epsilon_t,$$

where the functions are defined as:

$$g_1(x) = .2x + .2x(x > 0), \; g_2(x, y) = .2x + .1x(y > 0),$$

$$g_3(x, y) = x \exp(-y^2/2), \; f_1(x, y) = f_3(x, y) = f_5(x, y) = x(1 + \exp(-y^2/2)),$$

$$f_2(x, y) = f_4(x, y) = (1 + .5 \exp(-y^2/2))x, \; f_6(x) = x \exp(-x^2/2), \; f_7(x, y) = xy.$$

The covariate process is a VAR(2) process: $x_t = A_1 x_{t-1} + A_2 x_{t-2} + \eta_t$, where $A_1 = \{0.3^{i-j+1}\}_{i,j \leq m}$, $A_2 = \{0.2^{i-j+1}\}_{i,j \leq m}$, and $\Sigma_\eta = \{-0.3^{i-j}\}_{i,j \leq m}$. As before, $\eta_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\eta_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$.

In this model the threshold variable for the autoregressive terms is $Y_{t-1}$, while the threshold variable for the covariates is $X_{t-1.4}$. For the covariates we apply a smooth threshold function, and for the autoregressive terms we mainly employ a hard threshold function.
Threshold transformations with a single threshold variable imply a change in that variable causes a shift in the effects of all other covariates, and many examples of this effect can be found in macroeconomics.

The results are displayed in table 3, and our method outperforms the rest of the methods, with distance correlation usually taking second place. The median ranks of each of our significant variables can be found in table 5 of the supplementary file. And the variable which appears to be the most difficult to detect seems to be the threshold variable, $X_{t-1.4}$.

As predicted by the theoretical results and inline with results for the previous models, the performance of all methods worsens as we encounter heavy tails.

Model 4:

$$Y_t = .25Y_{t-1} + .3Y_{t-2} + .3Y_{t-3} + f_1(X_{t-1.1}) + f_2(X_{t-2.1})$$
$$+ \beta_{1,t}f_3(X_{t-1.2}, X_{t-1.3}) + \beta_{2,t}f_4(X_{t-2.2}, X_{t-2.3}) + \beta_{3,t}f_5(X_{t-1.3})$$
$$+ \beta_{4,t}f_6(X_{t-2.3}) + f_7(X_{t-1.2}) + f_8(X_{t-2.2}, X_{t-1.2}) + \epsilon_t,$$

where the functions are defined as:

$$f_1(x), f_7(x) = 1.5x + .4x(x > 0), f_2(x) = 1.2x, f_3(x, y) = f_4(x, y) = xy,$$

$$f_5(x), f_6(x) = x, f_8(x, y) = 1.2x + .4x(y > 0), \beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t} \overset{iid}{\sim} Unif(.5, 1) \forall t.$$

The covariate process is generated as in (7), with $A_1 = \{4|\tilde{i}-\tilde{j}|+1\}_{i,j \leq m_n}$ and $\Sigma_\eta = \{-3|\tilde{i}-\tilde{j}|\}_{i,j \leq m_n}$. As in the previous examples, $\eta_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\eta_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$. We also note that the coefficients $\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t}$, are random at each time $t$. In this model the autoregressive portion is linear, and for the exogenous covariates we use a mix of threshold functions, interactions, and random coefficients. The results are displayed in table 3, and our method again does better than the rest of the methods considered, as the others seem to have difficulty dealing with the combination of high dependence and nonlinearities. Looking at table 6.
in the supplementary file, we notice that the covariates $X_{t-1,3}, X_{t-2,3}$, which only appear through random coefficient effects, are the most difficult to predict.

### 5.2 Univariate response models: PDC-SIS+

The previous subsection showed a range of nonlinear time series models in which the performance of PDC-SIS is superior to competing methods. In this subsection, we will compare the performance of PDC-SIS vs. PDC-SIS+, and investigate whether we can improve upon the performance of PDC-SIS by adding variables to our conditioning vector in a data driven way.

Looking at the performance of our previous simulations, we observe that in model 2 we have difficulty detecting $X_{t-2,1}$. Therefore, we start by comparing PDC-SIS vs. PDC-SIS+ on model 2. The results are displayed in table 2, and we report the median rank of $X_{t-1,2}$ as well as the median minimum model size (MMS) for both procedures. The results show that PDC-SIS+ clearly outperforms PDC-SIS in detecting $X_{t-2,1}$, and therefore has a much smaller MMS. This difference is seen clearest in the case where we have heavy tails and $p_n = 4500$. The median MMS for PDC-SIS+ is 125 vs. 275 for PDC-SIS.

Our next model is a linear ARDL model:

**Model 5:**

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

The covariate process is generated as in (7), with $A_1 = \{.4^{[i-j]+1}\}_{i,j \leq m}$ and we set $\Sigma_\eta = I_m$, with $\eta_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\eta_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$. Additionally we set $\epsilon_t \overset{iid}{\sim} N(0, 1)$ or $t_5$. By construction $X_{t-2,1}$ would be the most difficult to detect using a marginal approach. The results are in table 3, and show that PDC-SIS+ does better in all scenarios, with the difference being most pronounced in the case of heavy tails. The results from both model 2 and model 6 suggest that adding strong conditional signals to our conditioning vector improves upon the performance of PDC-SIS.
Table 1: Median Minimum Model Size

| Gaussian, $p_n = 1500$ | Model 1 | Model 2 | Model 3 | Model 4 |
|------------------------|---------|---------|---------|---------|
| PDC-SIS                | 7       | 61      | 29      | 42      |
| DC-SIS                 | 11      | 488     | 112     | 306.5   |
| NIS                    | 11      | 488     | 119.5   | 275     |
| SIS                    | 10      | 343.5   | 100.5   | 234.5   |
| GLSS                   | 6       | 179.5   | 813     | 800.5   |

| Gaussian, $p_n = 4500$ | Model 1 | Model 2 | Model 3 | Model 4 |
|------------------------|---------|---------|---------|---------|
| PDC-SIS                | 11      | 149     | 78.5    | 100.5   |
| DC-SIS                 | 19      | 1051    | 337     | 842.5   |
| NIS                    | 16      | 861     | 309     | 704     |
| SIS                    | 13      | 722     | 281     | 588     |
| GLSS                   | 6       | 592     | 2325.5  | 2214    |

| $t_5$, $p_n = 1500$ | Model 1 | Model 2 | Model 3 | Model 4 |
|---------------------|---------|---------|---------|---------|
| PDC-SIS             | 13      | 79.5    | 43      | 51      |
| DC-SIS              | 20      | 408.5   | 114     | 306     |
| NIS                 | 33      | 513.5   | 167     | 328     |
| SIS                 | 21.5    | 447     | 166.5   | 265     |
| GLSS                | 6       | 450.5   | 969.5   | 891.5   |

| $t_5$, $p_n = 4500$ | Model 1 | Model 2 | Model 3 | Model 4 |
|---------------------|---------|---------|---------|---------|
| PDC-SIS             | 36.5    | 275.5   | 78      | 104     |
| DC-SIS              | 68      | 951.5   | 301.5   | 814.5   |
| NIS                 | 114     | 1100.5  | 436.5   | 851.5   |
| SIS                 | 66.5    | 905     | 438     | 761     |
| GLSS                | 7       | 1386.5  | 3008    | 2843.5  |

Table 2: Model 2, PDC-SIS+

| Gaussian, $p_n = 1500$ | PDC-SIS+ | PDC-SIS+ | PDC-SIS | PDC-SIS |
|------------------------|----------|----------|---------|---------|
| MMS                    | 34       | 26       | 61      | 40.5    |
| $X_{t-2,1}$            | 79       | 43.5     | 149     | 141     |
| $t_5$, $p_n = 1500$    | 57.5     | 29       | 79.5    | 57.5    |
| $t_5$, $p_n = 4500$    | 121.5    | 88       | 275.5   | 239.5   |

| Gaussian, $p_n = 4500$ | PDC-SIS+ | PDC-SIS+ | PDC-SIS | PDC-SIS |
|------------------------|----------|----------|---------|---------|
| MMS                    | 22       | 22       | 24      | 23      |
| $X_{t-2,1}$            | 42       | 40.5     | 59      | 58      |
| $t_5$, $p_n = 1500$    | 42       | 39       | 52      | 52      |
| $t_5$, $p_n = 4500$    | 85       | 76.5     | 162.5   | 159.5   |

Table 3: Model 5, PDC-SIS+

| Gaussian, $p_n = 1500$ | PDC-SIS+ | PDC-SIS+ | PDC-SIS | PDC-SIS |
|------------------------|----------|----------|---------|---------|
| MMS                    | 22       | 22       | 24      | 23      |
| $X_{t-2,1}$            | 42       | 40.5     | 59      | 58      |
| $t_5$, $p_n = 1500$    | 42       | 39       | 52      | 52      |
| $t_5$, $p_n = 4500$    | 85       | 76.5     | 162.5   | 159.5   |
6 Real Data Application: Macroeconomic Forecasting

In this section, we present an application to forecasting univariate macroeconomic time series. An application to multivariate time series models can be found in the supplementary material. Our dataset consists of 132 monthly macroeconomic series which run from January 1984 to December 2011, and was obtained from the supplement to Jurado et al. (2015). The transformations needed to achieve approximate stationarity as well as descriptions of the series are given in Jurado et al. (2015). A start date of January 1984 is chosen since most macroeconomic series are widely thought to contain a structural break around the first quarter of 1984, resulting in significantly lower volatility in the U.S economy (Stock and Watson, 2002). This effect has been known as the great moderation in macroeconomics, and various explanations for this phenomenon are given in Boivin and Giannoni (2006); Stock and Watson (2002).

We focus on forecasting the 6 month ahead real personal income less transfer payments (RPI), and the 6 month ahead number of employees on nonfarm payrolls (EMP). Both of these series are major monthly economic series which are closely watched by the National Bureau of Economic Research (NBER) business cycle dating committee (Business Cycle Dating Committee, 2008). We utilize a rolling window scheme, where the first simulated out of sample forecast was for the time period 2000:1 (January 2000). To construct this forecast, we use the observations between 1984:4 to 1999:7 (the first three observations are used in forming lagged covariates) to estimate the factors, and the coefficients. Therefore for the models described above, \( t = 1984:4 \) to 1999:1. We then use the regressor values at \( t = 1999:7 \) to form our forecast for 2000:1. The next window uses observations from 1984:5 to 1999:8 to forecast 2000:2. Using this scheme we have a total of 144 out of sample forecasts, and for each window we use \( n = 178 \) observations.

We assume the following model for our forecasts:

\[
Y_{t+6}^6 = g(Y_t, \ldots, Y_{t-3}, z_t) + \epsilon_{t+6}^6. \tag{8}
\]
where $Y_{t+6}^{6} = \log(RPI_{t+6}/RPI_{t})$, and $Y_{t} = \log(RPI_{t}/RPI_{t-1})$, and we replace RPI with EMP in the previous definitions when forecasting EMP. Additionally, $z_{t} = (x_{t}, \ldots, x_{t-3})$, where $x_{t}$ are the 131 macroeconomic series apart from $y_{t}$, which gives us a total of 528 predictors. We report the forecasting performance of 15 different models. The first is a baseline linear AR(4) model: $\hat{Y}_{t+6}^{6} = \hat{\alpha}_{0} + \sum_{i=0}^{3} \hat{\alpha}_{i} Y_{t-i}$. We then combine each of the six screening methods under consideration with two classes of second stage procedures. The first class of models assumes $g(\cdot)$ is a linear model, which we will estimate with the adaptive Lasso. For the adaptive Lasso, we use the Lasso as our initial estimate, and we choose the penalty parameters for both the adaptive Lasso and the Lasso using the modified BIC (Wang et al., 2009). The second class of models are factor augmented autoregressions: $\hat{Y}_{t+6}^{6} = \hat{\beta}_{0} + \sum_{i=0}^{3} \hat{\alpha}_{i} Y_{t-i} + \hat{\gamma} \hat{F}_{t}$, where $\hat{F}_{t} = (\hat{F}_{t,1}, \ldots, \hat{F}_{t,4})$ are four factors which are computed as the first four principal components of the top $d_{n} = n - 1$ predictors of $z_{t}$, as ranked by the screening procedures. Lastly, we include the performance of both classes of models on the entire dataset of 528 predictors.

For PDC-SIS and PDC-SIS+, $(Y_{t}, \ldots, Y_{t-3})$ is part of the conditional vector for each variable, and we compute each of the screening methods as discussed in section 5. When using the adaptive Lasso as a second stage method, we select the top $d_{n} = \lceil n/\log(n) \rceil = 35$ predictors from $z_{t}$ as our screened set for all screening methods. We then add the vector $(Y_{t}, \ldots, Y_{t-3})$ to our screened set and estimate our model via the adaptive Lasso using this subset of predictors.

The results are reported in table 4, and the entries in bold refer to the best performing model(s). We report the mean squared error (MSE), and the mean absolute error (MAE) of the resulting forecasts relative to the MSE, MAE of the benchmark AR(4) forecasts respectively. For RPI, when using adaptive Lasso as our second stage method we generally see distance correlation and PDC based screening algorithms outperforming competing procedures across both error measures. For employment forecasting, using the adaptive Lasso as a stand alone or second stage procedure appears to do worse than the baseline AR(4) model.
Table 4: Employment (EMP) and Real Personal Income (RPI): 6 month ahead forecasts

|                | EMP MSE | EMP MAE | RPI MSE | RPI MAE |
|----------------|---------|---------|---------|---------|
| AR (4)         | 1.00    | 1.00    | 1.00    | 1.00    |
| SIS-Adaptive Lasso | 1.22    | 1.05    | .77     | .92     |
| GLSS-Adaptive Lasso | .89     | .98     | .78     | .91     |
| NIS-Adaptive Lasso  | 1.23    | 1.08    | .83     | .95     |
| DC-SIS Adaptive Lasso | 1.15    | 1.04    | .71     | .88     |
| PDC-SIS Adaptive Lasso | 1.19    | 1.06    | .72     | .89     |
| PDC-SIS+Adaptive Lasso | 1.24    | 1.07    | .73     | .91     |
| Adaptive Lasso    | 1.14    | 1.05    | .80     | .96     |
| SIS-Factor AR     | .83     | .84     | .56     | .77     |
| GLSS-Factor AR    | .82     | .87     | .66     | .83     |
| NIS-Factor AR     | .91     | .86     | .62     | .81     |
| DC-SIS Factor AR  | .78     | .81     | .57     | .78     |
| PDC-SIS Factor AR | .63     | .77     | .52     | .75     |
| PDC-SIS+ Factor AR| .68     | .79     | .52     | .75     |
| Factor AR         | .70     | .81     | .58     | .81     |

*MSE, MAE reported are relative to the benchmark AR(4) forecasts*

For both RPI and Employment, factor augmented autoregressions clearly outperform the baseline AR(4) and adaptive lasso forecasts for both error measures. For both series our PDC based factor forecasts outperform competing screening methods with the difference being larger when forecasting employment. It is also encouraging to see that our PDC based factor forecasts outperform the standalone factor forecasts. This is inline with results reported in Bai and Ng (2008), which showed that adding too many irrelevant variables can deteriorate its forecasting performance. We note that using a model free screening procedure gives us full flexibility in choosing a second stage procedure, so our results might be further improved by considering additional classes of second stage procedures.

7 Discussion

In this work, we have introduced two classes of partial distance correlation based screening procedures, which are applicable to univariate or multivariate time series models. These methods aim to utilize the unique features of time series data as an additional source of
information, rather than treating temporal dependence as a nuisance. By using a model free first stage procedure we are able to expand the choice of models which can be considered for a second stage procedure. This is especially helpful for the case of nonlinear or nonparametric models where estimation in high dimensions can be computationally challenging.

There are many opportunities for further research, such as developing a theoretical or data driven approach to selecting the number of lags considered in our algorithms. Additionally, we can develop screening algorithms for time series data using measures which are more robust to heavy tailed distributions. Lastly, our procedures were developed under the assumption that the underlying processes are weakly dependent and stationary. Although these assumptions are satisfied for a very wide range of applications, there are many instances where they are violated. For example, non-stationarity is commonly induced by time varying parameters, structural breaks, and cointegrated processes, all of which are common in the fields of macroeconomics and finance. In addition, long range dependence is a property which is prominent in economics, finance, climate studies, and the physical sciences (see Samorodnitsky (2006) for more details). Therefore, developing new methodologies for long range dependent processes, or certain classes of non-stationary processes, such as locally stationary processes, would be particularly welcome.

**Supplementary Material**

Due to space limitations, simulations and a real data application of our group PDC-SIS procedure are contained in the supplementary material. Additionally, the supplementary material also contains the proofs for all the theorems, as well as more detailed tables of the results from section 5.1.

**References**

An, H. and Huang, F. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statistica Sinica* 6, 943–956.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146, 304 – 317.
Barut, E., Fan, J., and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association* **111**, 1266–1277.

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

Basu, S., Shojaie, A., and Michailidis, G. (2015). Network granger causality with inherent grouping structure. *Journal of Machine Learning Research* **16**, 417–453.

Boivin, J. and Giannoni, M. P. (2006). Has monetary policy become more effective? *The Review of Economics and Statistics* **88**, 445–462.

Borisov, I. and Volodko, N. (2009). Exponential inequalities for the distributions of canonical u-and v-statistics of dependent observations. *Siberian Advances in Mathematics* **19**, 1–12.

Business Cycle Dating Committee, N. (2008). The nber’s recession dating procedure. Online.

Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory* **18**, 17–39.

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.

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.

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

Davis, R. A., Matsui, M., Mikosch, T., and Wan, P. (2016). Applications of distance correlation to time series. *arXiv preprint arXiv:1606.05481*.

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

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.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for cox’s proportional hazards model. *IMS Collections, Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown* **6**, 70–86.

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

Fan, J., 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.

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

Feng, Y., Wu, Y., and Stefanski, L. A. (2017). Nonparametric independence screening via favored smoothing bandwidth. *Journal of Statistical Planning and Inference* In press.

Han, F. (2016). An exponential inequality for u-statistics under mixing conditions. *Journal of Theoretical Probability*.
Huang, Q. and Zhu, Y. (2016). Model-free sure screening via maximum correlation. *Journal of Multivariate Analysis* 148, 89 – 106.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring uncertainty. *American Economic Review* 105, 1177–1216.

Kilian, L. and Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge University Press.

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

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* 17, 1217–1241.

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

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

Liu, W., Xiao, H., and Wu, W. B. (2013). Probability and moment inequalities under dependence. *Statistica Sinica* 23, 1257–1272.

Liu, Y. and Wang, Q. (2017). Model-free feature screening for ultrahigh-dimensional data conditional on some variables. *Annals of the Institute of Statistical Mathematics* pages 1–19.

Lu, Z. (1998). On the geometric ergodicity of a non-linear autoregressive model with an autoregressive conditional heteroscedastic term. *Statistica Sinica*.

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

Ma, S., Li, R., and Tsai, C.-L. (2017). Variable screening via quantile partial correlation. *Journal of the American Statistical Association* 112, 650–663.

Mai, Q., Zou, H., et al. (2015). The fused kolmogorov filter: a nonparametric model-free screening method. *The Annals of Statistics* 43, 1471–1497.

Masry, E. and Tjostheim, D. (1997). Additive nonlinear arx time series and projection estimates. *Econometric Theory* 13, 214–252.

Nicholson, W. B., Bien, J., and Matteson, D. S. (2016). Hierarchical vector autoregression. *arXiv preprint arXiv:1412.5250*.

Pham, T. D. and Tran, L. T. (1985). Some mixing properties of time series models. *Stochastic Processes and their Applications* 19, 297–303.

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

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

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.

Stock, J. H. and Watson, M. W. (2002). Has the business cycle changed and why? *NEBER Macroeconomics Annual* 17, 159–218.
Székely, G. J. and Rizzo, M. L. (2014). Partial distance correlation with methods for dissimilarities. *Ann. Statist.* **42**, 2382–2412.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794.

Teräsvirta, T., Tjostheim, D., and Granger, C. (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., and Canales-Rodríguez, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **360**, 969–981.

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.

Weng, H., Feng, Y., and Qiao, X. (2017). Regularization after retention in ultrahigh dimensional linear regression models. *Statistica Sinica* In press.

Wu, W. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica* **19**, 1755–1768.

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

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

Wu, W. B. (2009). An asymptotic theory for sample covariances of bernoulli shifts. *Stochastic Processes and their Applications* **119**, 453 – 467.

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

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

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

Yoshihara, K. i. (1976). Limiting behavior of u-statistics for stationary, absolutely regular processes. *Probability Theory and Related Fields* **35**, 237–252.

Yousuf, K. (2018). Variable screening for high dimensional time series. *Electronic Journal of Statistics* **12**, 667–702.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67.

Zhou, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis* **33**, 438–457.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**, 1418–1429.
Supplement to Partial Distance Correlation Screening for High Dimensional Time Series

This supplementary document is organized as follows: supplement A contains the sure screening properties, simulations, as well as a real data application of our group PDC-SIS procedure. Supplement B contains the proofs of theorems 1 and 2 found in our main paper. Lastly, supplement C provides more detailed results of the simulations in section 5.1 of our main paper.

Supplement A

1.1 Sure Screening Properties for Group PDC-SIS

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

\[ x_i = h(\ldots, \eta_{i-1}, \eta_i) . \]  

(A.1)

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

**Condition 1.1.** Assume \(|pdcor(G_{t,i}, G_{k,j}; G_{t-1,i})| \geq c_1 n^{-\kappa}\) for \((i, k, j) \in M^*, \kappa \in (0, 1/2)\).

**Condition 1.2.** Assume our multivariate response process has the representation (A.1).

Additionally, we assume the following decay rate \( \Phi_{m,r}(x) = O(m^{-\alpha_x}) \), for some \( \alpha_x > 0 \), \( r > 4 \).

**Condition 1.3.** Assume our multivariate response process \( x_t \) has the representation (A.1).

Additionally assume \( v_x = \sup_{q \geq 2} q^{-\tilde{\alpha}_x} \Phi_{0,q}(x) < \infty \), for some \( \tilde{\alpha}_x \geq 0 \).

**Condition 1.4.** Assume the process \( \{x_t\} \) is \( \beta \)-mixing, with mixing rate \( \beta_x(a) = O(\exp(-a^{\lambda_1})) \), for some \( \lambda_1 > 0 \).
Let $g = 1$, if $\alpha > 1/2 - 2/r$, otherwise $g = r/4 - \alpha x r/2$. And let $K_{x,r} = \sup_{m \geq 0}(m + 1)^{\alpha} \Phi_r(x)$. Recall that $t_n = \max_j \text{dim}(C_j)$ is the maximum dimension of the conditional vectors. Lastly let $\tilde{\varphi} = \frac{2}{1 + 4\alpha x}$. The results are similar to those in theorem 1, but for the sake of completeness we present them here as well:

**Corollary 3.**

1. Suppose conditions 1.1, 1.3, 1.4 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have:

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

2. Suppose conditions 1.1, 1.2, 1.4 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have

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

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

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

### 1.2 Simulations: Group PDC-SIS

We consider the following VAR(1) process,

**Model 6:**

$$x_t = A_1 x_{t-1} + \eta_t, \quad (A.2)$$
and assume we have 25 groups at each lag level \((e_n = 25)\) with equal size \(g_n = 20\). We assume a block upper triangular structure for \(A_1\), with two scenarios.

\[
A_1 = \begin{bmatrix}
B & 0 & C & 0 \\
& B & 0 & C \\
& & B & 0 \\
& & & B \\
\end{bmatrix}
\]

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

Since we are assuming the first lag for each group is in the model, we have 23 off-diagonal group connections we want to detect for each scenario. As in our main paper, the sample size is \(n = 200\), and we report the median MMS for group DC-SIS, and group PDC-SIS procedure for each scenario in table 1. The MMS in this case is defined as the minimum number of group connections which need to be selected for \(\mathcal{M}_s\) to be captured. In order to ensure a fair comparison, we do not evaluate \(dcor(G_{t,i}, G_{t-1,i})\) for each group \(i\) when using group DC-SIS. The results show that the procedures are robust to the level of sparsity within each group, and our group PDC-SIS procedure significantly outperforms the group DC-SIS for all scenarios.
Table 1: Model 6

|                | Scenario 1 | Scenario 1 | Scenario 2 | Scenario 2 |
|----------------|------------|------------|------------|------------|
|                | PDC-SIS    | DC-SIS     | PDC-SIS    | DC-SIS     |
| $N(0, \Sigma_\eta = \{.4^{|i-j|+1}\})$ | 33         | 53         | 32         | 52         |
| $N(0, \Sigma_\eta = \{-0.4^{|i-j|+1}\})$ | 68         | 139.5      | 66         | 140        |
| $t_3, \Sigma_\eta = \{.4^{|i-j|+1}\}$ | 38         | 46.5       | 37         | 45         |
| $t_3, \Sigma_\eta = \{-0.4^{|i-j|+1}\}$ | 89         | 159.5      | 83.5       | 145.5      |

1.3 Real data application: Group PDC-SIS

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

$$z_t = f(z_{t-1}, z_{t-2}) + \epsilon_t. \quad (A.4)$$

We utilize a rolling window scheme similar to the one described previously, except we are not computing out of sample forecasts. For the first window we use data from $t = 1984:3$ to $t = 1999:12$ to compute our correlations. We then move the window forward by one month, which gives us 144 windows in total and 191 observations for each window. As discussed in section 4 of our main paper, for each group $G_{t,i}$ we condition on the first lag $G_{t-1,i}$ for PDC-SIS. Let $\{G_{t,j}\}_{j \leq 8}$ represents the 8 economic groups at time $t$, and let $\mathcal{B} = \{(i,k,j) : i, j \leq 8, k \in \{t-1,t-2\}\} \setminus \{(i,t-1,i) : i \leq 8\}$ denotes the set of possible group connections between the 8 economic groups minus the connection between a group...
and its first lag. For each window, we select the top \( \lceil n / \log(n) \rceil = 37 \) group connections, and record the number of group connections which belong to \( B \). We note that all group connections which are to be screened and do not belong to \( B \) are spurious connections by construction.

The results are in table 2, and we report the median number of group connections which belong to \( B \) over the 144 windows. In order to ensure a fair comparison between group DC-SIS and group PDC-SIS, we do not evaluate \( dcor(G_{t,i}, G_{t-1,i}) \) for each group \( i \) when using group DC-SIS. We see that when \( \alpha = .4 \) and the noise is Gaussian, both group PDC-SIS and group DC-SIS are very effective at selecting connections between economic groups. When the dependence increases and heavy tailed variables are introduced, the performance of group DC-SIS greatly deteriorates with many spurious group connections selected, whereas group PDC-SIS remains effective.

**Supplement B: Proofs of Theorems 1 and 2**

*Proof of Theorem 1.*

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

\[
\text{pdcor}(Y_t, Z_{t-1,k}; C_k) = \frac{dcor^2(Y_t, Z_{t-1,k}) - dcor^2(Y_t, C_k)dcor^2(Z_{t-1,k}, C_k)}{\sqrt{1 - dcor^4(Y_t, C_k)}\sqrt{1 - dcor^4(Z_{t-1,k}, C_k)}}. \tag{A.5}
\]
To estimate this quantity, Székely and Rizzo (2014) proposed an unbiased estimator of the distance correlation to serve as the plug-in estimate. This estimate is different from the estimator proposed for the distance correlation in Székely et al. (2007), which is a biased but consistent estimate. In proving asymptotic properties we can use either estimate, and we will use the original estimator given in Székely et al. (2007).

To obtain a bound for $|\widehat{dcor}^2(Y_t, Z_{t-1,k}; C_k) - dcor^2(Y_t, Z_{t-1,k}; C_k)|$ in the numerator of (A.5). Recall that:

$$dcor^2(Y_t, Z_{t-1,k}) = \frac{d cov(Y_t, Y_t) d cov(Z_{t-1,k}, Z_{t-1,k})}{d cov(Y_t, Y_t) d cov(Z_{t-1,k}, Z_{t-1,k})}. \quad (A.6)$$

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

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

$$= |(\hat{T}_2^{-1} - T_2^{-1})(\hat{T}_1 - T_1) + (\hat{T}_1 - T_1)/T_2 + (\hat{T}_2^{-1} - T_2^{-1})T_1|. \quad (A.7)$$

Therefore

$$P\left(\left| \frac{\hat{T}_1}{T_2} - \frac{T_1}{T_2} \right| > cn^{-\kappa} \right) \leq P\left(\left| (\hat{T}_2^{-1} - T_2^{-1})(\hat{T}_1 - T_1) \right| > c_2 n^{-\kappa}/3 \right) \quad (A.8)$$

$$+ P\left(\left| (\hat{T}_1 - T_1)/T_2 \right| > c_2 n^{-\kappa}/3 \right) \quad (A.9)$$

$$+ P\left(\left| (\hat{T}_2^{-1} - T_2^{-1})T_1 \right| > c_2 n^{-\kappa}/3 \right). \quad (A.10)$$

For the RHS of (A.8), we obtain:

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

$$+ P\left(|\hat{T}_2^{-1} - E(T_2)^{-1}| > Cn^{-\kappa/2} \right).$$
So we focus on terms (A.9) and (A.10). For (A.9), recall that:

\[
dcov^2(Y_t, Z_{t-1,k}) = \hat{S}_{k1} + \hat{S}_{k2} - 2\hat{S}_{k3}, \tag{A.11}
\]

where

\[
\hat{S}_{k1} = n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} |Y_{i} - Y_{j}| \|Z_{i,k} - Z_{j,k}|,
\]
\[
\hat{S}_{k2} = n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} |Y_{i} - Y_{j}| n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} |Z_{i,k} - Z_{j,k}|,
\]
\[
\hat{S}_{k3} = n^{-3} \sum_{j=1}^{n} \sum_{i=1}^{n} \sum_{l=1}^{n} |Y_{i} - Y_{j}| \|Z_{i,k} - Z_{l,k}|. \tag{A.12}
\]

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

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

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

\[
P(|\hat{S}_{k1} - S_{k1}| > Cn^{-\kappa}) \leq P(|\hat{S}_{k1}^* - S_{k1}| > Cn^{-\kappa}). \tag{A.13}
\]

We also have the following decomposition:

\[
|\hat{S}_{k1}^* - S_{k1}| \leq |\hat{S}_{k1}^* - E(\hat{S}_{k1}^*)| + |E(\hat{S}_{k1}^*) - S_{k1}|. \tag{A.14}
\]

Observe that \(\hat{S}_{k1}^*\) is a \(U\)-statistic, and is a biased estimate of \(S_{k1}\) due to temporal dependence. By condition 3.4, we can control this bias, and we have \(|E(\hat{S}_{k1}^* - S_{k1})| = O(n^{-\frac{2}{3}})\) by Yoshihara (1976). Obtaining a bound on \(P(|\hat{S}_{k1}^* - S_{k1}| > Cn^{-\kappa})\) is difficult in a time series setting. Borisov and Volodko (2009) and Han (2016) introduced exponential inequalities for \(U\)-statistics in a time series setting under uniform mixing type conditions, in addition to
restrictions on the kernel function. These restrictions are often too strict and rule out most commonly used time series. For example, even AR(1) processes where the innovations have unbounded support are not uniform mixing (see example 14.8 in Davidson (1994)).

As a result, we will instead rely on Nagaev and Rosenthal type inequalities (Wu and Wu, 2016; Liu et al., 2013) to obtain our bounds. We first show the bounds obtained by using Nagaev inequalities, and then we show the results obtained using Rosenthal type inequalities.

Let \( \psi_i = (e_i, \eta_i) \) and \( H_{i,j} = |Y_i - Y_j||Z_{i,k} - Z_{j,k}| \). We have

\[
H_{i,j} = f(\ldots, \psi_0, \ldots, \psi_{\max(i,j)}) \quad \text{and} \quad \hat{S}_{k1}^* = 2[n(n-1)]^{-1} \sum_{l=1}^{n-1} \sum_{i=1}^{n-l} H_{i,i+l}. \tag{A.15}
\]

We can then write:

\[
P(\left| \sum_{l=1}^{n-1} \sum_{i=1}^{n-l} (H_{i,i+l} - E(H_{i,i+l})) \right| > Cn^{2-\kappa}) \tag{A.16}
\]

\[
\leq \sum_{l=1}^{n-1} P(\left| \sum_{i=1}^{n-l} (H_{i,i+l} - E(H_{i,i+l})) \right| > Cn^{1-\kappa}). \tag{A.17}
\]

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

\[
\sum_{i=m}^{\infty} \left| \left| Y_i - Y_{i+l} \right| \right|_r \left| \left| Z_{i,k} - Z_{i+l,k} \right| - \left| Z_{i,k}^* - Z_{i+l,k}^* \right| \right|_r \\
\leq \sum_{i=m}^{\infty} \left| \left| Y_i - Y_{i+l} \right| \right|_r \left| \left| Z_{i,k} - Z_{i+l,k} \right| - \left| Z_{i,k}^* - Z_{i+l,k}^* \right| \right|_q \\
+ \sum_{i=m}^{\infty} \left| \left| Z_{i,k}^* - Z_{i+l,k}^* \right|_q \right| \left| Y_i - Y_{i+l} \right| + \left| Y_i^* - Y_{i+l}^* \right|_r \\
\leq \sum_{i=m}^{\infty} \left| \left| Y_i - Y_{i+l} \right| \right|_r \left| \left| Z_{i,k} - Z_{i+l,k}^* \right| + \left| Z_{i+l,k} - Z_{i+l+k}^* \right| \right|_q \\
+ \sum_{i=m}^{\infty} \left| \left| Z_{i,k}^* - Z_{i+l,k}^* \right|_q \right| \left| Y_i - Y_i^* \right| + \left| Y_{i+l} - Y_{i+l}^* \right|_r \\
\leq 2\Delta_{0,q}(y)\Phi_{m,r}(x) + 2\Delta_{m,q}(y)\Phi_{0,r}(x) = O(m^{-\alpha}). \tag{A.18}
\]
The last inequality holds since \( \|Z_{ik}\|_r \leq \Phi_{0,r}(\mathbf{x}) \), by section 2 in Wu and Wu (2016). Therefore,

\[
\sup_{m}(m+1)^{\alpha} \sum_{i=m}^{\infty} \| Y_i - Y_{i+l} \| |Z_{i,k} - Z_{i+l,k}| - |Y^*_i - Y^*_i| |Z^*_i,k - Z^*_i+l,k| | \leq 4K_{z,r}K_{y,q} \tau. \quad (A.19)
\]

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

\[
P\left( \left| \sum_{i=1}^{n-l} (H_{i,i+l} - E(H_{i,i+l})) \right| > Cn^{1-\alpha} \right) \leq O\left( \frac{n^2K_{z,r}^\tau K_{y,q}^\tau}{n^{\tau-r\kappa}} \right) + O\left( \exp\left( -\frac{n^{1-2\kappa}}{K_{z,r}^2K_{y,q}^2} \right) \right). \quad (A.20)
\]

Using condition 3.4 along with (A.13),(A.14),(A.17), and (A.20), we obtain:

\[
P\left( |\hat{S}_{k1} - S_{k1}| > Cn^{-\kappa} \right) \leq O\left( \frac{n^2K_{z,r}^\tau K_{y,q}^\tau}{n^{\tau-r\kappa}} \right) + O\left( n \exp\left( -\frac{n^{1-2\kappa}}{K_{z,r}^2K_{y,q}^2} \right) \right). \quad (A.21)
\]

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

\[
P\left( |\hat{S}_{k2} - S_{k2}| > Cn^{-\kappa} \right) \leq P\left( |(\hat{S}_{k2,1} - S_{k2,1})S_{k2,2}| > Cn^{-\kappa} \right) + P\left( |(\hat{S}_{k2,1} - S_{k2,1})(\hat{S}_{k2,2} - S_{k2,2})| > Cn^{-\kappa} \right). \quad (A.22)
\]

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

\[
P\left( |\hat{S}_{k2} - S_{k2}| > Cn^{-\kappa} \right) \leq O\left( \frac{n^2K_{z,r}^\tau K_{y,q}^\tau}{n^{\tau-r\kappa}} \right) + O\left( n \exp\left( -\frac{n^{1-2\kappa}}{K_{z,r}^2K_{y,q}^2} \right) \right) + O\left( n^2K_{z,r}^\tau + n^2K_{y,q}^\tau \right) + O\left( \exp\left( -\frac{n^{1-2\kappa}}{K_{z,r}^2K_{y,q}^2} \right) \right). \quad (A.23)
\]

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

\[
\hat{S}_{k3}^* = [n(n-1)(n-2)]^{-1} \sum_{i<j<k} |Z_{ik} - Z_{jk}| |Y_j - Y_i| + |Z_{ik} - Z_{lk}| |Y_j - Y_l|
\]
\[ + |Z_{ik} - Z_{jk}| |Y_i - Y_l| + |Z_{lk} - Z_{jk}| |Y_i - Y_l| \]
\[ + |Z_{lk} - Z_{jk}| |Y_i - Y_j| + |Z_{lk} - Z_{ik}| |Y_i - Y_j|. \]  
(A.24)

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

\[ P(|\hat{S}_{k3} - S_{k3}| > C n^{-\kappa}) \leq P(|\hat{S}_{k3}^* - S_{k3}| > C n^{-\kappa}) \]
\[ + P(|\hat{S}_{k1}^* - S_{k1}| > C n^{-\kappa}). \]  
(A.25)

We have already dealt with (A.26), so we will proceed to (A.25). It suffices to deal with the first term in (A.24), since the rest can be bounded similarly. Let \( H_{i,j,l} = |Z_{ik} - Z_{jk}| |Y_j - Y_l| = f(\ldots, \psi_0, \ldots, \psi_{\max(i,j,l)}) \). We can then represent

\[ \sum_{i<j<l} |Z_{ik} - Z_{jk}| |Y_j - Y_l| = \sum_{l=1}^{n-2} \sum_{j=1}^{n-l-1} \sum_{i=1}^{n-j-l} H_{i,i+j,j+l}. \]  
(A.27)

Note that for fixed \( j, l \), \( \{H_{i,i+j,j+l}\}_{i \in \mathbb{Z}} \) is a Bernoulli shift process, whose cumulative functional dependence measure is the same as (A.18). We can then write:

\[ P(|\sum_{l=1}^{n-2} \sum_{j=1}^{n-l-1} \sum_{i=1}^{n-j-l} [H_{i,i+j,j+l} - E(H_{i,i+j,j+l})]| > C n^{3-\kappa}) \]
\[ \leq \sum_{l=1}^{n-2} \sum_{j=1}^{n-l-1} P(|\sum_{i=1}^{n-j-l} [H_{i,i+j,j+l} - E(H_{i,i+j,j+l})]| > C n^{1-\kappa}). \]  
(A.28)

Using condition 3.4, along with (A.21),(A.24),(A.25),(A.26),(A.28), and theorem 2 in Wu and Wu (2016), we obtain:

\[ P(|\hat{S}_{k3} - S_{k3}| > C n^{-\kappa}) \leq O(n^2 n^t K_{z,r}^\tau K_{y,q}^\tau) + O(n^2 \exp(-\frac{n^{1-2\kappa}}{K_{z,r}^2 K_{y,q}^2})). \]  
(A.29)

This gives us a bound for (A.9). For (A.10): \( |\hat{T}_2^{-1} - T_2^{-1}| = |\hat{T}_2 - T_2| / T_2 \) and \( T_2 \) is finite by
condition 3.4. Using this, we obtain:

\[
P(|\hat{T}_2^{-1} - T_2^{-1}| > Cn^{-\kappa}) \leq P(|\hat{T}_2 - T_2| > |\hat{T}_2|Cn^{-\kappa})
\leq P(|\hat{T}_2 - T_2| > CMn^{-\kappa}) + P(|\hat{T}_2| < M). \quad (A.30)
\]

We will deal with the first term in (A.30) and the second term can be handled similarly.

Using the definition of \(\hat{T}_2, T_2\) and the decomposition we used in (A.22), it suffices to analyze

\[
P(|\hat{dcov}(Y_t, Y_t) - dcov(Y_t, Y_t)| > Cn^{-\kappa}) \quad (A.31)
\]

and

\[
P(|\hat{dcov}(Z_{t-1,k}, Z_{t-1,k}) - dcov(Z_{t-1,k}, Z_{t-1,k})| > Cn^{-\kappa}). \quad (A.32)
\]

For (A.31) and (A.32), note that for \(a > 0, b > 0\) we have

\[
|\sqrt{a} - \sqrt{b}| = \frac{|a - b|}{\sqrt{a} + \sqrt{b}} < \frac{|a - b|}{\sqrt{b}}. \quad (A.33)
\]

Using this, along with (A.30) and the methods used to bound \(\hat{T}_1\), we obtain:

\[
P(|\hat{T}_2^{-1} - T_2^{-1}| > Cn^{-\kappa}) \leq O(n^2 \frac{n^\kappa K^{\tau}_{z,r}}{n^{\kappa/2}}) + O(n^2 \exp(-n^{1-2\kappa}K^{\tau}_{z,r}))
+ O(n^2 \frac{n^\kappa K^{\tau}_{y,q}}{n^{\kappa/2}}) + O(n^2 \exp(-n^{1-2\kappa}K^{\tau}_{y,q})). \quad (A.34)
\]

By (A.21),(A.23),(A.29),(A.33), we obtain:

\[
P(|\hat{T}_1 - T_1| > cn^{-\kappa}) \leq O(n^2 \frac{n^\kappa K^{\tau}_{z,r}}{n^{\kappa/2}}) + O(n^2 \exp(-n^{1-2\kappa}K^{\tau}_{z,r}))
+ O(n^2 \frac{n^\kappa K^{\tau}_{y,q}}{n^{\kappa/2}}) + O(n^2 \exp(-n^{1-2\kappa}K^{\tau}_{y,q})). \quad (A.34)
\]

The other terms in (A.5) deal with the conditioning vectors \(C_j\), and we need to account for the maximum dimension of the conditioning vectors \(\max_j[\text{dim}(C_j)] = t_n\). This comes into effect when computing the cumulative functional dependence measure. Recall that \(C_{k+(h-1)m_n} = S_{k,h}\), and for analyzing the cumulative functional dependence measure, we
define
\[ S_{k,h}(i) = \{ Y_{i-1}, \ldots, Y_{i-h}, X_{i-1}, \ldots, X_{i-h+1} \}, \]  
(A.35)
as the conditional vector of the \( h \)th lag of series \( k \) at time \( i \). Additionally recall that \( |a|_p \) stands for the Euclidean norm of \( a \in \mathbb{R}^p \). Assume \( \dim(S_{k,h}) = t_n \) and \( q = r \), we therefore have:

\[ \sum_{i=m}^{\infty} \left( ||S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} |S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} \right. \]

\[ - |S_{k,h}^*(i) - S_{k,h}^*(i+j)||_{t_n} |S_{k,h}^*(i) - S_{k,h}^*(i+j)||_{t_n} \right) \]

\[ \leq \sum_{i=m}^{\infty} \left( ||S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} ||S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} - |S_{k,h}^*(i) - S_{k,h}^*(i+j)||_{t_n} \right) \]

\[ + \sum_{i=m}^{\infty} \left( ||S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} ||S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} - |S_{k,h}^*(i) - S_{k,h}^*(i+j)||_{t_n} \right) \]

\[ \leq \sum_{i=m}^{\infty} \left( ||S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} ||S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} + |S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} \right) \]

\[ + \sum_{i=m}^{\infty} \left( ||S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} ||S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} + |S_{k,h}(i) - S_{k,h}(i+j)||_{t_n} \right) \]

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

To explain the last inequality, we analyze the term:

\[ \sum_{i=m}^{\infty} ||S_{k,h}(i) - S_{k,h}^*(i)||_{t_n} \]

\[ = \sum_{i=m}^{\infty} ||S_{k,h}(i) - S_{k,h}^*(i)||_{t_n}^{1/2} \]

\[ \leq (t_n/2)^{1/2} (\Delta_{0,q}(y) + \Phi_{m,q}(x)). \]  
(A.37)

Where the last inequality follows from Minkowski’s inequality and the definition of \( S_{k,h}(i) \).

Using this, the rest of the terms in (A.5) can be handled as done previously.

We now show the bounds obtained using a Rosenthal type inequality. We follow the same steps as previously, and it suffices to consider (A.25). As before we focus on the following
\[
\sum_{i<j<l} \left| Y_j - Y_l \right| = \sum_{l=1}^{n-2} \sum_{j=1}^{n-l-1} \sum_{i=1}^{n-j-l} H_{i,i+j,i+j+l}.
\] (A.38)

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

\[
P(|Q - E(Q)| > cn^{1-\kappa}) \leq \frac{||Q - E(Q)||^\tau}{n^{\tau-\tau\kappa}}.
\] (A.39)

Then using Minkowski’s inequality, we obtain:

\[
||Q - E(Q)||_\tau \leq \left\| \sum_{i=1}^{n-2} H_{i,i+1,i+2} - E(H_{i,i+1,i+2}) \right\|_\tau.
\] (A.40)

As we stated previously, for fixed \( j, l \), \( \{H_{i,i+j,i+j+l}\}_{i \in \mathbb{Z}} \) is a Bernoulli shift process whose cumulative functional dependence measure is the same as (A.18). By theorem 1 in Liu et al. (2013), we have:

\[
\left\| \sum_{i=1}^{n-2} H_{i,i+1,i+2} - E(H_{i,i+1,i+2}) \right\|_\tau \leq O(K_{z,r}^r K_{y,q}^q n^{\frac{3}{2}}).
\] (A.41)

Combining the above with (A.40), we obtain:

\[
P(|Q - E(Q)| > cn^{1-\kappa}) \leq O\left( \frac{K_{z,r}^r K_{y,q}^q n^{\frac{3}{2}}}{n^{\tau-\tau\kappa}} \right).
\] (A.42)

By repeating the same techniques we obtain:

\[
P(|\hat{\omega}_k - \omega_k| > c_2 n^{-\kappa}) \leq O\left( \frac{K_{y,q}^q n^{\frac{4}{2}}}{n^{q/2-q\kappa/2}} \right) + O\left( \frac{K_{z,r}^r K_{y,q}^q n^{\frac{3}{2}}}{n^{\tau-\tau\kappa}} \right) + O\left( \frac{K_{z,r}^r n^{\frac{r}{2}}}{n^{r/2-r\kappa/2}} \right).
\] (A.43)

For simplicity we assume \( r = q \), and we now compare the above result to (A.34), which was obtained using Nagaev type inequalities. Note that when \( q = r \) the above bound is of the order \( O(n^{r/4-r\kappa/2}) \). Using Nagaev type inequalities leads to the bound at most \( O(n^{r/2-r\kappa/2-3}) \).
Therefore, when $r < 12$, (A.43) provides a better bound. When $r > 12$, the comparison depends on the values of $\varrho, \iota, \zeta$ which are related to the dependence of the covariate and response processes. Applying the union bound gives us the desired result.

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

$$|\hat{\rho}_k| \geq |\rho_k| - |\hat{\rho}_k - \rho_k| \geq c_1 n^{-\kappa}/2, \; \forall k \in M_*.$$  \hfill (A.44)

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

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

$$\theta_q(y_i) = \|E(y_i|\mathcal{F}_0) - E(y_i|\mathcal{F}_{-1})\|_q,$$

$$\theta_q(Z_{ij}) = \|E(Z_{ij}|\mathcal{H}_0) - E(Z_{ij}|\mathcal{H}_{-1})\|_q.$$ \hfill (A.45)

With the cumulative predictive dependence measures defined as:

$$\Theta_{0,q}(y) = \max_{j \leq \rho_n} \sum_{i=0}^{\infty} \delta_q(Z_{ij}), \; \text{and} \; \Theta_{0,q}(\epsilon) = \sum_{i=0}^{\infty} \delta_q(\epsilon_i).$$ \hfill (A.46)

We follow the steps of the proof of part (iii). For $|\hat{d}_{\text{cor}}^2(Y_{t}, Z_{t-1,k}) - d_{\text{cor}}^2(Y_{t}, Z_{t-1,k})|$, it suffices to provide a bound for (A.28). Note that for fixed $j, l$, we have:

$$\sup_{q \geq 4} q^{-\tilde{\alpha}_z} \sum_{i=1}^{\infty} \theta_q(H_{t,i+j,i+j+l}) \leq \sup_{q \geq 4} q^{-\tilde{\alpha}_z} \sum_{i=1}^{\infty} \delta_q(H_{t,i+j,i+j+l})$$

$$\leq \sup_{q \geq 4} q^{-\tilde{\alpha}_z} \Delta_{0,q}(y) \Phi_{0,q}(x) < \infty,$$ \hfill (A.47)

14
where the first inequality follows from theorem 1 in Wu (2005), and the last inequality follows from condition 3.3. Using the above we have by theorem 3 in Wu and Wu (2016):

\[ (A.28) \leq O\left(n^2 \exp\left(-\frac{n^{1/2-\kappa}}{v_y}\right)^{\tilde{\jmath}}\right). \]  

We now provide a bound for (A.31) in a similar way. Let \( S_{i,j,l} = |Y_i - Y_j||Y_j - Y_i| = f_1(\ldots, e_{0,\ldots, e_{\max(i,j,l)}}). \) We then have:

\[
\sup_{q \geq 4} q^{-2\tilde{\alpha}_y} \sum_{i=1}^{\infty} \theta_q(S_{i,i,j,l}) \leq \sup_{q \geq 4} \sum_{i=1}^{\infty} \delta_q(S_{i,i,j,l}) \\
\leq \sup_{q \geq 4} q^{-2\tilde{\alpha}_y} \Delta^2_{0,q}(y) < \infty. \quad (A.49)
\]

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

\[
\sum_{l=1}^{n-2} \sum_{j=1}^{n-1-l} P(\sum_{i=1}^{n-j-l} (S_{i,i+j,i+j+l} - E(S_{i,i+j,i+j+l})) > Cn^{1-\kappa}) \\
\leq O\left(n^2 \exp\left(-\frac{n^{1/2-\kappa}}{v_y^{2}}\right)^{\tilde{\alpha}}\right). \quad (A.50)
\]

A similar result holds for (A.32). Following the steps in the proof of part (iii), and using the results above we obtain:

\[
P(\max_{j \leq p_n} |\hat{\omega}_k - \omega_k| > c_2 n^{-\kappa}) \leq p_n \left[ O(n^2 \exp\left(-\frac{n^{1/2-\kappa}}{v_y}\right)^{\tilde{\alpha}}\right) \\
+ O(n^2 \exp\left(-\frac{n^{1/2-\kappa}}{v_y}\right)^{\tilde{\psi}}\right) \\
+ O(n^2 \exp\left(-\frac{n^{1/2-\kappa}}{v_z}\right)^{\tilde{\phi}}\right].
\]

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

Proof of Theorem 2.
For simplicity we only prove part (i), and the proof for part (iii) follows similarly. Let 
\( \tilde{\omega} = (\tilde{\omega}_1, \ldots, \tilde{\omega}_{p_n}) \), where \( \tilde{\omega}_k = \hat{pdc}_{Y_t, Z_{t-1,k}}(\hat{\mathcal{C}}_k) \). We will work on the following set,

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

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

\[ \square \]

**Supplement C: Tables for Section 5.1**

Tables 3-6 provide more detailed results of the simulations in section 5.1. As stated in our main paper, tables 3-6 report the median minimum model size needed to include all the relevant predictors, as well as the median rank of the significant covariates for each procedure.
Table 3: Model 1

Gaussian, $p_n = 1500$

| MMS  | $X_{t-1,1}$ | $X_{t-1,2}$ | $X_{t-1,3}$ | $X_{t-1,4}$ | $X_{t-1,5}$ | $X_{t-1,6}$ |
|------|-------------|-------------|-------------|-------------|-------------|-------------|
| PDC-SIS | 7 | 6 | 3 | 2 | 2 | 3 | 5 |
| DC-SIS | 11 | 7 | 3.5 | 2 | 2 | 3 | 5.5 |
| NIS | 11 | 6 | 3 | 2 | 2 | 3 | 6 |
| SIS | 10 | 6 | 3 | 2 | 2 | 3 | 6 |
| GLSS | 6 | 5 | 3 | 2 | 2 | 3 | 5 |

Gaussian, $p_n = 4500$

| MMS  | $X_{t-1,1}$ | $X_{t-1,2}$ | $X_{t-1,3}$ | $X_{t-1,4}$ | $X_{t-1,5}$ | $X_{t-1,6}$ |
|------|-------------|-------------|-------------|-------------|-------------|-------------|
| PDC-SIS | 11 | 5 | 3 | 3 | 3 | 3 | 5 |
| DC-SIS | 19 | 6 | 3 | 3 | 3 | 3 | 6 |
| NIS | 16 | 6 | 3 | 3 | 3 | 3 | 6 |
| SIS | 13 | 5 | 3 | 2.5 | 3 | 3 | 6 |
| GLSS | 6 | 5 | 3 | 2 | 2 | 3 | 5 |

$t_5$, $p_n = 1500$

| MMS  | $X_{t-1,1}$ | $X_{t-1,2}$ | $X_{t-1,3}$ | $X_{t-1,4}$ | $X_{t-1,5}$ | $X_{t-1,6}$ |
|------|-------------|-------------|-------------|-------------|-------------|-------------|
| PDC-SIS | 13 | 5 | 3 | 3 | 3 | 3 | 5 |
| DC-SIS | 20 | 6 | 4 | 3 | 3 | 3 | 6 |
| NIS | 33 | 7 | 4 | 3 | 3 | 3 | 6 |
| SIS | 21.5 | 6 | 3 | 3 | 3 | 3 | 5 |
| GLSS | 6 | 5 | 3 | 2 | 2 | 3 | 5 |

$t_5$, $p_n = 4500$

| MMS  | $X_{t-1,1}$ | $X_{t-1,2}$ | $X_{t-1,3}$ | $X_{t-1,4}$ | $X_{t-1,5}$ | $X_{t-1,6}$ |
|------|-------------|-------------|-------------|-------------|-------------|-------------|
| PDC-SIS | 36.5 | 7 | 4 | 2 | 2 | 3 | 5 |
| DC-SIS | 68 | 10.5 | 4 | 2 | 3 | 3 | 7 |
| NIS | 114 | 16.5 | 4 | 2 | 3 | 4 | 9 |
| SIS | 66.5 | 10.5 | 4 | 3 | 3 | 4 | 7 |
| GLSS | 7 | 5 | 3 | 2 | 2 | 3 | 5 |
Table 4: Model 2

| Gaussian $p_n=1500$ | MMS  | $X_{t-1,1}$ | $X_{t-2,1}$ | $X_{t-1,2}$ | $X_{t-2,2}$ |
|---------------------|------|-------------|-------------|-------------|-------------|
| PDC-SIS             | 61   | 1           | 40.5        | 2           | 5           |
| DC-SIS              | 488  | 1           | 488         | 2           | 3           |
| NIS                 | 488  | 1           | 488         | 2           | 3           |
| SIS                 | 343.5| 1           | 341.5       | 2           | 3           |
| GLSS                | 179.5| 1           | 160.5       | 2           | 6.5         |

| Gaussian $p_n=4500$ | MMS  | $X_{t-1,1}$ | $X_{t-2,1}$ | $X_{t-1,2}$ | $X_{t-2,2}$ |
|---------------------|------|-------------|-------------|-------------|-------------|
| PDC-SIS             | 149  | 1           | 141         | 2           | 4           |
| DC-SIS              | 1051 | 1           | 1051        | 2           | 3           |
| NIS                 | 861  | 1           | 861         | 2           | 3           |
| SIS                 | 722  | 1           | 722         | 2           | 3           |
| GLSS                | 592  | 1           | 412.5       | 2           | 8           |

| $t_5 p_n=1500$      | MMS  | $X_{t-1,1}$ | $X_{t-2,1}$ | $X_{t-1,2}$ | $X_{t-2,2}$ |
|---------------------|------|-------------|-------------|-------------|-------------|
| PDC-SIS             | 79.5 | 1           | 57.5        | 2           | 5           |
| DC-SIS              | 408.5| 1           | 408.5       | 2           | 3           |
| NIS                 | 513.5| 1           | 492         | 2           | 4           |
| SIS                 | 447  | 1           | 440         | 2           | 4           |
| GLSS                | 450.5| 1           | 330.5       | 2           | 22          |

| $t_5 p_n=4500$      | MMS  | $X_{t-1,1}$ | $X_{t-2,1}$ | $X_{t-1,2}$ | $X_{t-2,2}$ |
|---------------------|------|-------------|-------------|-------------|-------------|
| PDC-SIS             | 275.5| 1           | 239.5       | 2           | 5           |
| DC-SIS              | 951.5| 1           | 951.5       | 2           | 3           |
| NIS                 | 1100.5| 1         | 984         | 2           | 4           |
| SIS                 | 905  | 1           | 859.5       | 2           | 3           |
| GLSS                | 1386.5| 1         | 995         | 2           | 18.5        |
Table 5: Model 3

Gaussian, $p_n = 1500$

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

Gaussian, $p_n = 4500$

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

$t_5, p_n = 1500$

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

$t_5, p_n = 4500$

|       | MMS  | $X_{t-1,1}$ | $X_{t-1,2}$ | $X_{t-1,3}$ | $X_{t-1,4}$ |
|-------|------|-------------|-------------|-------------|-------------|
| PDC-SIS | 78   | 2           | 5           | 4           | 3           | 11          | 24.5        |
| DC-SIS | 301.5| 14.5        | 8           | 11.5        | 4           | 33          | 113.5       |
| NIS    | 436.5| 14.5        | 8           | 14          | 4           | 33.5        | 124         |
| SIS    | 438  | 13          | 7           | 13          | 4           | 33          | 149.5       |
| GLSS   | 3008 | 85.5        | 690         | 1362        | 99.5        | 9           | 117.5       |
Table 6: Model 4

Gaussian, $p_n = 1500$

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

Gaussian, $p_n = 4500$

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

$t_5$, $p_n = 1500$

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

$t_5$, $p_n = 4500$

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