Taylor & Francis
Taylor & Francis Group

Check for updates

Targeting Predictors Via Partial Distance Correlation With Applications to Financial Forecasting

Kashif Yousuf[a] and Yang Feng[b]

[a]Department of Statistics, Columbia University, New York, NY; [b]Department of Biostatistics, School of Global Public Health, New York University, New York, NY

ABSTRACT

High-dimensional time series datasets are becoming increasingly common in various fields of economics and finance. Given the ubiquity of time series data, it is crucial to develop efficient variable screening methods that use the unique features of time series. This article introduces several model-free screening methods based on 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. We show the effectiveness of our methods via extensive simulation studies and an application on forecasting U.S. market returns.

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 observations (n), and one of the common ways statistical inference is made possible is by relying on the assumption of sparsity. The sparsity assumption, which states that only a small number of covariates contributes to the response, has led to a wealth of theoretical results and methods available for identifying important predictors in the high dimensional setting. These methods broadly fall into two classes: variable selection methods and screening methods. Variable selection methods aim to recover the true signal set $\mathcal{M}_*$, 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_n such that $P(\mathcal{M}_* \subset S_n) \rightarrow 1$ as $n \rightarrow \infty$. Ideally, we would also hope that $|S_n| \ll p_n$, thereby significantly reducing the dimension of the feature space for a second-stage method, such as penalized likelihood methods, or principal components regression on the set of targeted predictors selected during 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, Feng, and Song 2011), Cox proportional hazards model (Fan, Feng, and Wu 2010), linear quantile models (Ma, Li, and Tsai 2017), and varying coefficient models (Fan, Ma, and Dai 2014). In addition, model-free screening methods, which do not assume any particular model *a priori*, have also been developed. Some examples include a distance correlation based method in Li, Zhu, and Zhong (2012), the fused Kolmogorov filter in Mai et al. (2015), a conditional distance correlation method in Liu and Wang (2017), and a smoothing bandwidth based method in Feng, Wu, and Stefanski (2018). For a partial survey of screening methods, one can consult Liu, Zhong, and Li (2015). The main theoretical result of these methods is the so-called sure screening property, which states that under appropriate conditions we can reduce the dimension of the feature space from size $p_n = O\left(\exp\left(n^\alpha\right)\right)$ to a far smaller size d_n, while retaining all the relevant predictors with probability approaching 1. We note that variable screening methods are closely related to the targeted predictors framework more commonly used in econometrics. As introduced in Bai and Ng (2008), the targeted predictors framework was focused on selecting predictors using linear dependence measures for the specific setting of forecasting from a second-stage principal components regression. This can be thought of as a specific type of variable screening with linear dependence measures.

Although there has been a large amount of interest in developing screening methods, it is surprising to see that almost all of the works operate under the assumption of independent observations with a few exceptions (Chen et al. 2018; Yousuf 2018). This is even more surprising given the ubiquity of

time-dependent data in economics and finance. Some examples include forecasting low-frequency macroeconomic indicators, such as GDP or the inflation rate, or financial asset returns using a large number of macroeconomic and financial time series and their lags as possible covariates (Stock and Watson 2002a; Bai and Ng 2009; Gu, Kelly, and Xiu 2018; Yousuf and Ng 2020). These examples, among others, highlight the importance of developing screening methods specifically for time-dependent data.

In creating a screening method for stationary short-range dependent time series, we aim to account for some of the features of time series data including (a) a certain number of lags of the response variable are usually included in the model; (b) an ordered structure of the covariates, in which lower-order lags of covariates are thought to be more likely to be informative than higher-order lags; (c) the frequent occurrence of multivariate response models such as linear or nonlinear VAR models. Note that the idea of focusing on lower-order lags is popular in econometrics, for example, Doan, Litterman, and Sims (1984) and Litterman (1986) considered the Minnesota-type priors which shrink distant lags faster.

We also aim to have a model-free screening approach that can handle continuous, discrete, or grouped time series. Using a model-free approach allows us to avoid imposing assumptions on the structure of the model (i.e., linearity), thereby making our methods robust to model misspecification at the screening stage. This gives us full flexibility to select a nonlinear or nonparametric second-stage procedure. Our goal is thus to extend the targeted predictors framework to more general nonlinear or nonparametric models while accounting for the time dependence found in our data. This is especially useful given that recent work has shown the benefits of considering nonlinear and nonparametric models in forecasting macroeconomic and financial time series.[1] Finally, using a nonlinear dependence measure is helpful even when we aim to fit a second-stage linear model, as the marginal relationship between the predictors and the response can be highly nonlinear.

To achieve our goals, we will introduce several distance-correlation-based screening procedures. Distance correlation (DC) was introduced by Székely, Rizzo, and Bakirov (2007), for measuring dependence and testing independence between two random vectors. The consistency and weak convergence of sample distance correlation have been established for stationary time series in Zhou (2012) and Davis et al. (2016). DC has several useful properties including (a) the distance correlation of two random vectors equals zero if and only if these two random vectors are independent; (b) ability to handle discrete time series, as well as grouped predictors; (c) an easy to compute partial distance correlation (PDC) 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 features of time series data mentioned previously.

We will mainly be dealing with univariate response models, some examples of which include linear or nonlinear autoregressive models with exogenous predictors (NARX). Our methods can also be extended to multivariate response models such as linear or nonlinear VAR models. In both settings, we rely on PDC to build our screening procedures. PDC produces a wealthy family of screening methods by making 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. Using this prior knowledge often enhances the screening procedure, as shown in the case of generalized linear models in Barut, Fan, and Verhasselt (2016). Therefore our method 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 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 vector.

To the best of our knowledge, there have been only two works, Chen et al. (2018) and Yousuf (2018), studying screening methods in a stationary time series setting for a univariate response. Chen et al. (2018) extended the nonparametric independence screening (NIS) approach used for independent observations to the time series setting. However, the method does not use the serial dependence in the data or account for the unique properties of the time series data we outlined. In particular, to account for those properties, we would like to search for a conditional relationship, which would require a different dependency measure. Yousuf (2018) extended the theory of SIS to heavy-tailed and dependent data as well as proposing a generalized least square based screening method to correct for serial correlation. However, it is limited to linear models, and the other unique qualities of time series data outlined above are ignored.

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

The rest of the article is organized as follows. In Section 2, we first review the distance correlation-based methods, then introduce our screening procedures for univariate response and multivariate response models, respectively. Section 3 covers the asymptotic properties of our PDC screening based methods.

---

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

Section 4 covers extensive simulation studies comparing our methods with existing work. We present an application to forecasting monthly U.S. market returns in Section 5. We conclude the article with a short discussion in Section 6. The proofs for all theorems, along with additional simulations, are placed in the supplementary material.

2. Methods

Before introducing our method, we first review distance correlation and its related methods in Section 2.1, then introduce our new screening algorithms for univariate and multivariate time series models in Sections 2.2 and 2.3, respectively.

2.1. Review of Distance Correlation-Based Methods

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

Definition 2.1. For any random vectors $\boldsymbol{u} \in \mathbb{R}^q, \boldsymbol{v} \in \mathbb{R}^p$, let $\phi_{\boldsymbol{u}}(\boldsymbol{t}), \phi_{\boldsymbol{v}}(\boldsymbol{s})$ be the characteristic functions of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. The distance covariance between $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as in Székely, Rizzo, and Bakirov (2007):

$$\text{dcov}^2(\boldsymbol{u}, \boldsymbol{v}) = \int_{\mathbb{R}^{p+q}} |\phi_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{t}, \boldsymbol{s}) - \phi_{\boldsymbol{u}}(\boldsymbol{t})\phi_{\boldsymbol{v}}(\boldsymbol{s})|^2 w^{-1}(\boldsymbol{t}, \boldsymbol{s}) dt ds,$$

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

Given this choice of weight function, by Székely, Rizzo, and Bakirov (2007), we have a simpler formula for the distance covariance. Let $(\boldsymbol{u}', \boldsymbol{v}'), (\boldsymbol{u}'', \boldsymbol{v}'') \overset{\text{i.i.d.}}{\sim} (\boldsymbol{u}, \boldsymbol{v})$, and let:

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

Then, provided that second moments exist, we have by Remark 3 in Székely, Rizzo, and Bakirov (2007) and (1.2) in Székely and Rizzo (2014), $\text{dcov}^2(\boldsymbol{u}, \boldsymbol{v}) = S_1 + S_2 - 2S_3$. We can now estimate this quantity using moment-based methods. Suppose we observe $(\boldsymbol{u}_i, \boldsymbol{v}_i)_{i=1,\dots,n}$, the sample estimates for the distance covariance and distance correlation are

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

$$\widehat{\text{dcor}}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\widehat{\text{dcov}}(\boldsymbol{u}, \boldsymbol{v})}{\sqrt{\widehat{\text{dcov}}(\boldsymbol{u}, \boldsymbol{u})\widehat{\text{dcov}}(\boldsymbol{v}, \boldsymbol{v})}},$$

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

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

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

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

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

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

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

2.2. Screening Algorithms for Univariate Time Series Models

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

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

where $F(Y_t | \cdot)$ is the conditional cumulative distribution function of Y_t. The value h_n represents the maximum lag order we are considering for our response and predictor series. This value can be decided beforehand by the user, or can be selected using a data driven method. The value h_n could differ for different predictors, however, for simplicity of presentation we assume the same value h_n for all predictors.

2.2.1. Screening Algorithm I: PDC-SIS

In our first algorithm, PDC-SIS, we define the conditioning vector for the lth lag of predictor series k ($X_{t-l,k}$) at time t as: $\mathcal{S}_{k,l} = (Y_{t-1}, \dots, Y_{t-h_n}, X_{t-1,k}, \dots, X_{t-l+1,k})$, where $1 \leq l \leq h_n$. Since we are assuming a priori that a certain number of lags of Y_t are to be included in the model, $\{Y_{t-1}, \dots, Y_{t-h_n}\}$ is part of the conditioning vector for all possible covariates. Our conditioning vector also includes all lower-order lags for each lagged covariate we are considering. By doing so, our method tries to shrink toward sub-models with lower-order lags. To illustrate this, consider the case, where Y_t is strongly dependent on $X_{t-1,j}$

even while controlling for the effects of $Y_{t-1}, \ldots, Y_{t-h_n}$. Under this scenario, if $X_{t-1,j}$ has strong serial dependence, higher-order lags of $X_{t-1,j}$ can be mistakenly selected by our screening procedure even if they are not in our active set of covariates.

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

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

where the choice of γ_n will be discussed in Section 2.2.3. We use d_n to represent the size of $\widehat{\mathcal{M}}_{\gamma_n}$.

2.2.2. Screening Algorithm II: PDC-SIS+

As we have seen, the time ordering of the covariates allows us some additional flexibility in selecting the conditioning vector compared to the iid setting. PDC-SIS attempts to use the time series structure of our data by conditioning on lower lags of the covariate. However, rather than simply conditioning only on the lower lags of a covariate, we can condition on additional information available from lower lags of other covariates as well. One way to attempt this, and to potentially improve PDC-SIS, is to identify strong conditional signals at each lag level and add them to the conditioning vector for all higher-order lag levels. By using 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 auto-correlation, and cross-correlation.

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

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

where Γ_n is a threshold to be discussed in Section 2.2.3. We then use this information to form our next conditioning vector:

$$\widehat{\mathcal{S}}_{k,2} = \left(Y_{t-1}, \ldots, Y_{t-h_n}, X_{t-1,k}, \boldsymbol{z}_{t-1,\widehat{\mathcal{U}}_1^{\Gamma_n}} \right),$$

where $\boldsymbol{z}_{t-1,\widehat{\mathcal{U}}_1^{\Gamma_n}}$ is a subvector of $\boldsymbol{z}_{t-1}$ which is formed by extracting the indices contained in $\widehat{\mathcal{U}}_1^{\Gamma_n}$. We note that any duplicates which result from overlap between $X_{t-1,k}$ and $\boldsymbol{z}_{t-1,\widehat{\mathcal{U}}_1^{\Gamma_n}}$ are deleted. For convenience, we define $\widehat{C} = (\widehat{\mathcal{S}}_{1,1}, \ldots, \widehat{\mathcal{S}}_{m_n,1}, \widehat{\mathcal{S}}_{1,2}, \ldots, \widehat{\mathcal{S}}_{m_n,h_n})$ as our vector of estimated conditional sets. We then use $(\widehat{\mathcal{S}}_{k,2})_{k \leq m_n}$ to compute the strong conditional signal set for the 2nd lag level

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

Repeating this procedure we obtain

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

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

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

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

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

2.2.3. Threshold Selection

Threshold selection is a critical ingredient for the success of any screening based method. Common methods include resampling procedures (Barut, Fan, and Verhasselt 2016; Weng, Feng, and Qiao 2019), selecting a pre-specified number of predictors depending on the sample size (Fan and Lv 2008; Fan, Feng, and Song 2011), and using a data-driven method such as cross-validation. Callot et al. (2017) proposed an information criteria-based choice for the thresholding operation on a regularized estimator.

We first discuss how to select the parameter Γ_n for PDC-SIS+. For simplicity, we will only use a single threshold for all lag levels. The idea is to create pseudo data $\{(\boldsymbol{x}_t, Y_t^*)\}_{t=1,\ldots,n}$, where $\{Y_t^*\}_{t=1,\ldots,n}$ is formed using a stationary bootstrap on $\{Y_t\}_{t=1,\ldots,n}$. This resampling procedure creates a null model, where $\hat{\omega}_j^* = \widehat{pdcor}(Y_t^*, X_{t-1,j}; Y_{t-1}^*, Y_{t-2}^*, \ldots, Y_{t-h_n}^*)$. We can then choose the $\alpha = 0.99$ quantile of $\{\hat{\omega}_1^*, \ldots, \hat{\omega}_{p_n}^*\}$. Given that this threshold depends on a single resampling, we stabilize this threshold by constructing K (e.g., $K = 5$)[2] bootstrap samples. In order to avoid conditioning on too many variables, an upper bound of $\lceil n^{1/2} \rceil$ variables can be added to our conditioning vector at each lag level. This procedure is similar to the random decoupling approach used in Weng, Feng, and Qiao (2019) and Barut, Fan, and Verhasselt (2016) for the iid setting.

For both PDC-SIS and PDC-SIS+, we also need to select a threshold γ_n to form our targeted set of predictors. We give three possible methods to select this threshold. The first is to use the bootstrap resampling procedure detailed above, which is a data-driven method to select γ_n. Given we used $\alpha = 0.99$ to select Γ_n, we would want to use a quantile between 0.95 and 0.99 to select

---

[2]The performance is stable across different choices of K from our numerical experience.

γ_n. This is similar in spirit to thresholding by using a cutoff for the t-statistics of each marginal correlation measure used in Bai and Ng (2008). The second approach, which is most commonly used in the literature, is to select the top d_n predictors as ranked by our screening algorithm. When $p_n \gg n$, $d_n = \lceil n/\log(n) \rceil$ or $d_n = n - 1$ are common choices. Alternatively, d_n can be set by the researcher using prior knowledge of the data.[3] The third approach is to select γ_n via cross-validation if we have decided on the second-stage modeling procedure.

The choice of which threshold method to select depends on the goals of the user and prior information of the problem. As mentioned previously, the most common method in the literature, and the one we use in our empirical examples, is to select the top d_n predictors. This method is attractive for a number of reasons: it is agnostic as to the choice of second-stage model, and can be used as a quick way to reduce the dimensionality of the problem to make the resulting exploration and modeling more tractable. It also allows the user to control the size of the screened set directly. The bootstrap procedure is also a viable option, if the user does not have any prior opinion of how large they want their screened set, although it adds additional computational overhead. Whereas, the third option is limited to cases where one already has chosen their second-stage modeling procedure.

2.3. 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 useful 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 the 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, Shojaie, and Michailidis (2015), Kock and Callot (2015), and Nicholson, Bien, and Matteson (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 5×10^6. Additionally, these methods are restricted to linear VAR models, whereas there is considerable evidence of nonlinear effects such as the existence of thresholds, smooth transitions, regime switching, and varying coefficients in fields such as macroeconomics and finance (Kilian and Lütkepohl 2017).

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

$$x_t = g(x_{t-1}, \ldots, x_{t-k}) + \eta_t, \text{ where } x_t \in \mathcal{R}^{m_n}, \eta_t \text{ iid} \quad (1)$$

For simplicity, we let all groups be of size g_n, let $e_n = m_n/g_n$ denote the total number of groups for a given lag level, and denote our groups $(G_{t-1,1}, \ldots, G_{t-k,e_n})$. To get a sense of the computational benefits of screening on the group level, assume for example, $m_n = 500, k = 1$, and we have 25 groups all of size $g_n = 20$. For this linear VAR (1) model, when $n = 200$, we note it takes about 350 times longer to compute all $m_n^2 = 500^2$ pairwise distance correlations $\left\{ \widehat{\text{dcor}}(X_{t,j}, X_{t-1,k}) \right\}_{j \leq m_n, k \leq m_n}$ 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 PDCs, 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, Shojaie, and Michailidis 2015).

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

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

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

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

The sure screening properties of our group PDC-SIS procedure are 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, the 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

---

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

of PDC-SIS+ can also be developed similarly to the procedure in Section 2.2, however, we do not pursue this direction, as it usually leads to situations where we are conditioning on a large number of variables.

3. Asymptotic Properties

3.1. Dependence Measures

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

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

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

$$\boldsymbol{x}_i = \boldsymbol{h}\left(\ldots, \boldsymbol{\eta}_{i-1}, \boldsymbol{\eta}_i\right). \tag{3}$$

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

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

$$\delta_q(Y_i) = ||Y_i - g\left(\mathcal{F}_i^*\right)||_q = (E|Y_i - g\left(\mathcal{F}_i^*\right)|^q)^{1/q},$$
$$\delta_q(X_{ij}) = ||X_{ij} - h_j\left(\mathcal{H}_i^*\right)||_q = (E|X_{ij} - h_j\left(\mathcal{H}_i^*\right)|^q)^{1/q}, \tag{4}$$

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

$$\Delta_{0,q}(\boldsymbol{y}) = \sum_{i=0}^{\infty} \delta_q(Y_i) < \infty, \quad \text{and}$$

$$\Phi_{m,q}(\boldsymbol{x}) = \max_{j \le p_n} \sum_{i=m}^{\infty} \delta_q(X_{ij}) < \infty. \tag{5}$$

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

We note that compared to functional dependence measures, β-mixing coefficients can be defined for any stochastic processes, and are not limited to Bernoulli shift processes. On the other hand, functional dependence measures are easier to interpret and compute since they are related to the data-generating mechanism of the underlying process. In many cases using the functional dependence measure also requires less stringent assumptions (see Wu and Wu (2016), Yousuf (2018) for details). Although there is no direct relationship between these two dependence frameworks, fortunately, there are a large number of commonly used time series processes, which are β-mixing and satisfy Equation (5). For example, under appropriate conditions, linear processes, ARMA, GARCH, ARMA-ARCH, threshold autoregressive, Markov chain models, amongst others, can be shown to be β-mixing (see Pham and Tran (1985), Carrasco and Chen (2002), An and Huang (1996), and Lu (1998) for details).

3.2. Asymptotic Properties: PDC-SIS

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

Condition 3.1. $|\text{pdcor}(Y_t, Z_{t-1,k}; C_k)| \ge c_1 n^{-\kappa}$ for $k \in M_*$ and $\kappa \in (0, 1/2)$.

Condition 3.2. The response and the covariate processes have representations (2) and (3), respectively. Additionally, we assume the following decay rates $\Phi_{m,r}(\boldsymbol{x}) = O(m^{-\alpha_z}), \Delta_{m,q}(\boldsymbol{y}) = O(m^{-\alpha_y})$, for some $\alpha_z, \alpha_y > 0, q > 2, r > 4$ and $\tau = \frac{qr}{q+r} > 2$.

Condition 3.3. The response and the covariate processes have representations (2) and (3) respectively. Additionally assume $\upsilon_z = \sup_{q \ge 2} q^{-\tilde{\alpha}_z} \Phi_{0,q}(\boldsymbol{x}) < \infty$ and $\upsilon_y = \sup_{q \ge 2} q^{-\tilde{\alpha}_y} \Delta_{0,q}(\boldsymbol{y}) < \infty$, for some $\tilde{\alpha}_z, \tilde{\alpha}_y \ge 0$.

Condition 3.4. The process $\{(Y_t, \boldsymbol{x}_t)\}$ is β-mixing, with mixing rate $\beta_{xy}(a) = O(\exp(-a^{\lambda_1}))$, for some $\lambda_1 > 0$.

Condition 3.1 is a standard population-level assumption that allows covariates in the active set to be detected by our screening procedure. Condition 3.2 is similar to the one used in Yousuf (2018) and Wu and Wu (2016), and assumes both the response and covariate processes are causal Bernoulli shift processes, and have at least 2 and 4 finite moments respectively. Additionally it presents the dependence conditions on these processes, where higher values of $\alpha_x, \alpha_\epsilon$ indicate weaker temporal dependence. Examples of response processes which satisfy Condition 3.2 include stationary, causal, finite order ARMA, GARCH, ARMA-GARCH, bilinear, and threshold autoregressive processes, all of which have exponentially decaying functional dependence measures (see Wu (2011) for details). For the covariate process, assume $\boldsymbol{x}_i$ is a vector linear process: $\boldsymbol{x}_i = \sum_{l=0}^{\infty} A_l \boldsymbol{\eta}_{i-l}$. where $\{A_l\}$ are $m_n \times m_n$ coefficient matrices and $\{\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{im_n})^T\}$ are iid random vectors with $cov(\boldsymbol{\eta}_i) = \Sigma_\eta$. For simplicity, assume $\{\eta_{i,j}, j = 1, \ldots, m_n\}$ are identically distributed, then $\delta_q(X_{ij}) = ||A_{i,j} \boldsymbol{\eta}_0 - A_{i,j} \boldsymbol{\eta}_0^*||_q \le 2|A_{i,j}|||\eta_{0,1}||_q$, where $A_{i,j}$ is the jth column of A_i. Define $||A_i||_\infty$ as the maximum absolute row sum of A_i, then if $||A_i||_\infty = O(i^{-\beta})$ for

$\beta > 1$, we have $\Phi_{m,q}(x) = O(m^{-\beta+1})$. Other examples include stable VAR processes, and multivariate ARCH processes which have exponentially decaying cumulative functional dependence measures (Wu and Wu 2016; Yousuf 2018). We note that it is clear that if x_i satisfies Condition 3.2, then z_i trivially satisfies it as well. Condition 3.3 strengthens the moment requirements of Condition 3.2, and requires that all moments of the covariate and response processes are finite. To illustrate the role of the constants $\tilde{\alpha}_z$ and $\tilde{\alpha}_y$, consider the example where y_i is a linear process: $y_i = \sum_{j=0}^{\infty} f_j e_{i-j}$ with e_i iid and $\sum_{l=0}^{\infty} |f_l| < \infty$, then $\Delta_{0,q}(y) = ||e_0 - e_0^*||_q \sum_{l=0}^{\infty} |f_l|$. If we assume e_0 is sub-Gaussian, then $\tilde{\alpha}_y = 1/2$, since $||e_0||_q = O(\sqrt{q})$. Similarly, if e_i is sub-exponential, we have $\tilde{\alpha}_y = 1$.

To understand the inclusion of Condition 3.4, 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, Zhu, and Zhong 2012), Condition 3.4 is needed to control the rate at which the above bias vanishes as $n \to \infty$. Conditions 3.2 and 3.4 are frequently used when dealing with time series data (Wu and Pourahmadi 2009; Xiao and Wu 2012; Davis et al. 2016).

Throughout this article, let $\alpha = \min(\alpha_x, \alpha_y)$, and $\varrho = 1$, if $\alpha_z > 1/2 - 2/r$, otherwise $\varrho = r/4 - \alpha_z r/2$. Let $\iota = 1$ if $\alpha > 1/2 - 1/\tau$, otherwise $\iota = \tau/2 - \tau\alpha$, and let $\zeta = 1$, if $\alpha_y > 1/2 - 2/q$, otherwise $\zeta = q/4 - \alpha_y q/2$. Additionally, let $K_{y,q} = \sup_{m \geq 0}(m+1)^{\alpha_y} \Delta_{m,q}(y)$, and $K_{z,r} = \sup_{m \geq 0}(m+1)^{\alpha_z} \Phi_r(x)$. Given Condition 3.3, it follows that $K_{\epsilon,q}, K_{z,r} < \infty$. Let $t_n = \max_j \dim(C_j)$, be the maximum dimension of the conditional vectors. We define $\tilde{\psi} = \frac{2}{1 + 2\tilde{\alpha}_z + 2\tilde{\alpha}_y}, \tilde{\varphi} = \frac{2}{1 + 4\tilde{\alpha}_z}, \tilde{\alpha} = \frac{2}{1 + 4\tilde{\alpha}_y}$. Lastly, for ease of presentation, let $\hat{\omega} = (\hat{\omega}_1, \ldots, \hat{\omega}_{p_n})$, $\omega = (\omega_1, \ldots, \omega_{p_n})$, where $\omega_k = \mathrm{pdcor}(Y_t, Z_{t-1,k}; C_k)$, $\hat{\omega}_k = \widehat{\mathrm{pdcor}}(Y_t, Z_{t-1,k}; C_k)$. In addition, let

$$a_n = n^2 \left[\exp\left(-\frac{n^{1/2-\kappa}}{t_n \upsilon_y^2}\right)^{\tilde{\alpha}} + \exp\left(-\frac{n^{1/2-\kappa}}{t_n \upsilon_z \upsilon_y}\right)^{\tilde{\psi}} \right.$$
$$\left. + \exp\left(-\frac{n^{1/2-\kappa}}{t_n \upsilon_z^2}\right)^{\tilde{\varphi}} \right],$$

$$b_n = n^2 \left[\frac{t_n^{r/2} n^{\zeta} K_{y,r}^r}{n^{r/2 - r\kappa/2}} + \frac{t_n^{r/2} n^{\iota} K_{z,r}^{r/2} K_{y,r}^{r/2}}{n^{r/2 - r/2\kappa}} + \frac{t_n^{r/2} n^{\varrho} K_{z,r}^r}{n^{r/2 - r\kappa/2}} \right.$$
$$+ \exp\left(-\frac{n^{1-2\kappa}}{t_n^2 K_{z,r}^4}\right) + \exp\left(-\frac{n^{1-2\kappa}}{t_n^2 K_{z,r}^2 K_{y,r}^2}\right)$$
$$\left. + \exp\left(-\frac{n^{1-2\kappa}}{t_n^2 K_{y,r}^4}\right) \right],$$

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

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

Theorem 1. i. Suppose Conditions 3.1, 3.3, and 3.4 hold. For any $c_2 > 0$, we have

$$P(\max_{j \leq p_n} |\hat{\omega}_j - \omega_j| > c_2 n^{-\kappa}) \leq O(p_n a_n).$$

ii. Suppose Conditions 3.1, 3.3, and 3.4 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have

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

iii. Suppose Conditions 3.1, 3.2, and 3.4 hold. For any $c_2 > 0$, we have

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

if $r \geq 12$, $\quad P(\max_{j \leq p_n} |\hat{\omega}_j - \omega_j| > c_2 n^{-\kappa}) \leq O(p_n b_n).$

iv. Suppose Conditions 3.1, 3.2, and 3.4 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have

if $r < 12$, $\quad P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(d_n c_n);$

if $r \geq 12$, $\quad P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - O(d_n b_n).$

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

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

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

Li, Zhu, and Zhong (2012) obtained $p_n = o(\exp(n^{\frac{1-2\kappa}{3}}))$ for DC-SIS, which matches our rate. In the iid setting with finite polynomial moments, we can use the truncation method in their proof and combined with the Markov inequality to obtain $p_n = o(\exp(n^{r/4 - r\kappa/2 - 1}))$. Our results, which rely on a different proof strategy than the truncation method, provide a better bound even in this setting.

3.3. Asymptotic Properties: PDC-SIS+

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

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

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

$$\mathcal{U}_{l-1}^{\Gamma_n} = \left\{ (l-1)m_n + 1 \leq j \leq lm_n : |\text{pdcor}(Y_t, Z_{t-1,j}; C_j)| \quad (6) \right.$$
$$\left. \geq \lambda_n + \frac{c_1}{2} n^{-\kappa} \right\},$$

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

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

Condition 3.5 assumes the variables in the strong conditional signal set, $\mathcal{U}^{\Gamma_n}$, are easily identifiable from the rest of the covariates. This separation in the signal strength will allow us to ensure with high probability that our estimated conditional sets match their population-level counterparts. The assumption $\lambda_n n^{\kappa} \to \infty$, is introduced to ensure $d_n \gg$ (the size of the set $\mathcal{U}^{\lambda_n}$). Although the hope is that $\mathcal{U}^{\lambda_n} \subset \mathcal{M}_*$, this is not required to prove sure screening properties of our algorithm. The sure screening properties for PDC-SIS+ are similar to PDC-SIS, but for the sake of completeness, we state the theorem in full.

Theorem 2. i. Suppose Conditions 3.1, 3.3, 3.4, and 3.5 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have

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

ii. Suppose Conditions 3.1, 3.2, 3.4, and 3.5 hold. For $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_1/2$, we have

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

4. Simulations

We now evaluate the performance of PDC-SIS and PDC-SIS+ via extensive simulation studies. We also include the performance of four other screening methods whose properties have been investigated in the time series setting; these include marginal Pearson correlation screening (SIS) (Fan and Lv 2008), NIS (Fan, Feng, and Song 2011; Chen et al. 2018), generalized least-squares screening (GLSS) (Yousuf 2018), and distance correlation screening (DC-SIS) (Li, Zhu, and Zhong 2012). [4]

Unless noted otherwise, we fix the sample size $n = 200$, the maximum number of lags considered $h_n = 3$, and the conditioning vector always includes the first three lags of the 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. For each experiment, we report results related to out of sample forecasting, and the proportion of relevant variables selected in our screened set. The latter metric is defined as

$$\frac{\text{number of relevant variables in the screened set}}{\text{number of relevant variables in DGP}}.$$

Note that for all procedures considered, we will not be screening the lags of Y_t, therefore the previous metric does not include lags of Y_t in the numerator or denominator. For forecasting, we select our screening set using the first $n - 1$ observations and we select the top $d_n = \lceil n/\log(n) \rceil$ variables as our screened set. Using this screened set and the first $n - 1$ observations we fit a non-parametric sparse additive model (Ravikumar et al. 2009) and forecast the last observation.[5] We then calculate the square forecast error (MSFE). We repeat each experiment 500 times, and report the average MSFE, and the average proportion of relevant variables selected in our screened set.

We set $Y_0 = Y_{-1} = \ldots = Y_{-(h_n+1)} = 0$, and generate $n + 200$ samples of our model. We then discard the first $200 - h_n$ samples. To ensure stationarity when generating a nonlinear autoregressive model with exogenous predictors (NARX), we use the sufficient conditions provided in Masry and Tjøstheim (1997).

4.1. Data-Generation Processes

DGP 1:

$$Y_t = \sum_{j=1}^{6} \beta_j X_{t-1,j} + \epsilon_t, \text{ and } x_t = A_1 x_{t-1} + \eta_t,$$
$$\text{where } A_1 = 0.6 * I, \text{ and } \eta_t \overset{iid}{\sim} N(0, \Sigma_\eta). \quad (7)$$

For this model, we set $\Sigma_\eta = [0.3^{|i-j|}]_{ij}$, $\beta_j = 0.5$ for $j = 1, \ldots, 6$. The error follows an AR(1) process: $\epsilon_t = \alpha \epsilon_{t-1} + e_t$ where $\alpha = .6$, and we consider $e_t \overset{iid}{\sim} N(0, 1)$.

DGP 2:

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

---

[4]We use the R package energy to compute the partial DC and DC. The NIS estimator is computed using the R package mgcv. For computational efficiency, the GLSS estimator is computed using the nlme package using an AR(1) approximation for the residual covariance matrix. Simulations for our group PDC-SIS procedure are contained in the supplementary material.
[5]We use the R package SAM to fit the nonparametric sparse additive model.

where the functions are defined as follows:

$$g_1(x) = 0.25x, \quad g_2(x) = x\exp(-x^2/2),$$
$$g_3(x) = -0.6x + 0.3x(x > 0),$$
$$f_1(x) = 1.5x + 0.4x(x > 0), \quad f_2(x) = -x,$$
$$f_3(x) = 1.2x + 0.4x(x > 0), \quad f_4(x) = x^2\sin(2\pi x).$$

The covariate process is generated as in Equation (7), with $A_1 = [0.4^{|i-j|+1}]_{ij}$ and $\Sigma_\eta = I_{m_n}$. Additionally, we set $\epsilon_t \overset{iid}{\sim} N(0,1)$.
DGP 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 follows:

$$g_1(x) = 0.2x + 0.2x(x > 0), \quad g_2(x,y) = 0.2x + 0.1x(y > 0),$$
$$g_3(x,y) = x\exp(-y^2/2),$$
$$f_1(x,y) = f_2(x,y) = f_4(x,y) = x\left(1 + \frac{1}{1 + 0.5\exp(-y)}\right),$$
$$f_3(x,y) = x\left(2 + \frac{2}{1 + 0.5\exp(-y)}\right), \quad f_5(x) = f_6(x) = 2x,$$
$$f_7(x,y) = x\left(1 + \frac{1}{1 + \exp(-y)}\right).$$

The covariate process is a VAR(2) process: $\boldsymbol{x}_t = A_1\boldsymbol{x}_{t-1} + A_2\boldsymbol{x}_{t-2} + \boldsymbol{\eta}_t$, where $A_1 = [0.3^{|i-j|+1}]_{ij}$, $A_2 = [.2^{|i-j|+1}]_{ij}$, and $\boldsymbol{\eta}_t \overset{iid}{\sim} N(0, \Sigma_\eta)$ or $\boldsymbol{\eta}_t \overset{iid}{\sim} t_5(0, 3/5 * \Sigma_\eta)$, in which $\Sigma_\eta = [-.3^{|i-j|}]_{ij}$.
DGP 4:

$$Y_t = 0.25Y_{t-1} + 0.3Y_{t-2} + 0.3Y_{t-3} + f_1(X_{t-1,1}) + f_2(X_{t-2,1})$$
$$+ \beta_{1,t}f_3(X_{t-1,2}, X_{t-1,3}) + \beta_{2,t}f_4(X_{t-2,2}, X_{t-2,3})$$
$$+ \beta_{3,t}f_5(X_{t-1,3})$$
$$+ \beta_{4,t}f_6(X_{t-2,3}) + f_7(X_{t-1,2}) + f_8(X_{t-2,2}, X_{t-1,2}) + \epsilon_t,$$

where the functions are defined as follows:

$$f_1(x) = f_7(x) = 1.5x + 0.4x(x > 0), f_2(x) = 1.2x,$$
$$f_3(x,y) = f_4(x,y) = xy,$$
$$f_5(x) = f_6(x) = x, \quad f_8(x,y) = 1.2x + 0.4x(y > 0),$$
$$\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t} \overset{iid}{\sim} \text{Unif}(.5, 1) \quad \forall t.$$

The covariate process is generated as in Equation (7), with $A_1 = [0.4^{|i-j|+1}]_{ij}$ and $\Sigma_\eta = [-0.3^{|i-j|}]_{ij}$. We also note that the coefficients $\beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t}$, are random at each time t.
DGP 5: $Y_t = 0.25Y_{t-1} + 0.3Y_{t-2} + 0.3Y_{t-3} + X_{t-1,1} - X_{t-2,1} + 0.5X_{t-1,2} + 0.5X_{t-2,2} + \epsilon_t$.

The covariate process is generated as in Equation (7), with $A_1 = [0.4^{|i-j|+1}]_{ij}$ and $\Sigma_\eta = I_{m_n}$. Additionally we set $\epsilon_t \overset{iid}{\sim} N(0,1)$.

Table 1. Average MSFE over 500 repetitions.

	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5
		$n = 200, p_n = 1500$			
DC-SIS	1.91	1.98	14.25	9.40	2.79
SIS	1.90	1.99	14.08	9.01	2.78
NIS	1.92	2.06	14.17	9.23	2.81
PDC-SIS	1.89	1.97	9.54	6.11	2.05
GLSS	1.94	1.96	10.5	6.32	2.68
PDC-SIS+	**1.69**	1.96	**9.50**	**6.10**	**1.83**
		$n = 200, p_n = 4500$			
DC-SIS	2.48	2.03	18.2	12.26	3.19
SIS	2.40	2.02	17.85	11.47	3.10
NIS	2.53	2.11	18.98	11.66	3.22
PDC-SIS	2.43	2.03	14.61	7.78	2.57
GLSS	2.40	**2.00**	14.64	**7.63**	2.93
PDC-SIS+	**2.30**	2.02	**14.58**	7.77	**2.28**

NOTE: For each screening method we select the top $d_n = \lceil n/\log(n)\rceil$ variables as our screened set and estimate a one step ahead forecast using a non-parametric sparse additive model. In this and the following table, bold entries refer to the best performing model.

Table 2. Average proportion of relevant variables selected over 500 repetitions.

	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5
		$n = 200, p_n = 1500$			
DC-SIS	0.97	0.78	0.76	0.36	0.63
SIS	0.98	0.77	0.79	0.45	0.67
NIS	0.97	0.77	0.77	0.42	0.65
PDC-SIS	**0.99**	0.84	**0.89**	**0.85**	0.91
GLSS	0.98	0.73	0.76	0.64	0.58
PDC-SIS+	0.98	**0.87**	**0.89**	**0.85**	**0.94**
		$n = 200, p_n = 4500$			
DC-SIS	0.94	0.76	0.65	0.18	0.50
SIS	0.95	0.75	0.69	0.24	0.56
NIS	0.94	0.74	0.67	0.22	0.53
PDC-SIS	**0.97**	0.76	**0.83**	**0.76**	0.84
GLSS	**0.97**	0.72	0.69	0.52	0.54
PDC-SIS+	**0.97**	**0.80**	**0.83**	**0.76**	**0.86**

NOTE: For each screening method we select the top $d_n = \lceil n/\log(n)\rceil$ variables as our screened set and calculate the proportion of relevant variables selected (excluding the conditioning variables) in our screened set.

4.2. Results

The MSFE results are displayed in Table 1, and the average proportion of relevant variables results are reported in Table 2.

For DGP 1, we see that all methods perform well in this case with PDC-SIS+ performing slightly better than the rest. We see from Table 2 that all methods are able to capture the relevant variables. We note that for DGP 1, our PDC-based methods contains irrelevant variables in the conditioning set, namely the lags of Y_t.

For DGP 2, the nonlinear transformations used are mainly threshold functions which are popular nonlinear transformations for time series data (Teräsvirta, Tjostheim, and Granger 2010). We see that the proposed methods PDC-SIS and PDC-SIS+ outperform the other methods in terms of average proportion of relevant variables selected, and are comparable in terms of MSFE. As seen in Table 8 of the supplementary file, the covariate $X_{t-2,1}$ appears to be the most difficult to detect for

Table 3. Results for $n = 100, p_n = 1500$.

	DC-SIS	SIS	NIS	PDC-SIS	GLSS	PDC-SIS+
Average proportion of relevant variables selected.						
DGP 3	0.38	0.41	0.38	0.54	0.47	0.54
DGP 4	0.08	0.09	0.09	0.39	0.28	0.40
MSFE						
DGP 3	36.8	35.9	36.02	33.6	30.5	33.4
DGP 4	18.3	18.0	18.5	16.2	13.1	16.1

NOTE: See notes to Tables 1 and 2.

Table 4. Relative performance of the bootstrap approach to that of the top d_n predictor approach when $n = 200, p_n = 1500$.

	PDC-SIS $(\alpha = 0.975)$	PDC-SIS $(\alpha = 0.99)$	PDC-SIS+ $(\alpha = 0.975)$	PDC-SIS+ $(\alpha = 0.99)$
Ratio of the Average Proportions of Relevant Variables Selected				
DGP 3	0.97	0.92	0.97	0.93
DGP 4	0.98	0.91	0.98	0.91
Ratio of the MSFE				
DGP 3	1.0	0.99	0.99	1.01
DGP 4	1.0	1.0	1.0	1.0

the competing methods, and our conditioning scheme greatly improves the detection of this signal. For DGP 3, we apply a logistic smooth transition function to the covariates, and for the autoregressive terms, we mainly employ a hard threshold function. For this DGP, our PDC-based methods significantly outperform the other methods, with GLSS following next. The variable which appears to be the most difficult to detect is the transition variable, $X_{t-1,4}$. DGP 4 contains a mix of threshold functions, interactions, and random coefficients. The results for this are similar to DGP 3 with our PDC-based methods outperforming the remaining methods, especially in terms of the proportion of relevant variables selected. Looking at Table 10 of the supplementary file, we notice that the covariates $X_{t-1,3}$ and $X_{t-2,3}$, which only appear through random coefficient effects, are the most difficult to detect. Overall, we see that for DGPs 1-5, our PDC-based methods perform best, with PDC-SIS+ does as good, and in most cases, better than PDC-SIS.

4.3. Additional Results

We present some additional results which show the effect of the sample size, and the choice of threshold for our screening methods. Due to space considerations, we only present results for DGPs 3 and 4. We start with the effect of the sample size n, we set $n = 100, p_n = 1500$ and choose the top $d_n = \lceil n/\log n \rceil$ predictors as our screened set. From the results in Table 3, we observe that the MSFE has sharply increased and the proportion of relevant variables selected sharply decreased for both DGPs.[6]

Next, we examine the effect of setting alternate thresholds for our screening methods. Instead of selecting the top d_n predictors, we use the bootstrap resampling technique discussed in Section 2.2.3 to select the threshold γ_n. We provide results for

---

[6]This conclusion holds even if set the threshold to $d_n = \lceil 200/\log 200 \rceil$.

$\alpha = 0.99$ and $\alpha = 0.975$. The results can be seen in Table 4, and they are presented as ratios to the benchmark which selects the top $d_n = \lceil n/\log n \rceil$ predictors as our screened set. We see from the results that all three thresholds lead to essentially identical results for forecasting. For variable selection, setting $d_n = \lceil n/\log n \rceil$ does best with $\alpha = 0.975$ following closely behind.

5. Real Data Application: Forecasting Portfolio Returns

In this section, we present an application to forecasting US monthly equity portfolio returns, with the data originally analyzed in Kelly and Pruitt (2013). We first focus on forecasting market returns as measured by the CRSP value weighted index, and the SP500 index. Then, we forecast returns from 5 Fama-French (FF) portfolios sorted on market cap. For our predictor series, we use the book to market valuation ratios for FF size and value sorted portfolios, in which U.S. stocks are divided into 25 or 100 portfolios sorted by market cap and book to market ratios. Kelly and Pruitt (2013) built on the present value identity and argued both theoretically and empirically that this cross-section of dis-aggregated valuation ratios is predictive of future market returns.[7]

Let $\mathbf{x}_t$ denote the 100 (or 25) FF portfolios at time t, and Y_{t+1} denote the portfolio return at time $t + 1$. Since there exists autocorrelation in the returns, we set Y_t as a conditioning variable in PDC-SIS and PDC-SIS+, and our predictor set is $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \mathbf{x}_{t-3})$. Due to the strong autocorrelation present in $\mathbf{x}_t$, we have a high degree of cross sectional correlation in $\mathbf{z}_t$, and it is likely that many elements of $\mathbf{z}_t$ are unimportant.

The linear dynamic factor model (sometimes referred to as Diffusion Index (DI) model), in which the factors are estimated by principal components, is very commonly used in econometrics (Stock and Watson 2002a,b). One of the well-known weaknesses of the DI model is that the response is ignored when estimating the factors. Rather than estimating the principal components over the entire set of predictors $\mathbf{z}_t$, Bai and Ng (2008) and Bair et al. (2006), among others, have shown that estimating the principal components on a targeted set of predictors can often lead to greater predictive accuracy. This procedure is sometimes known as supervised principal components. Besides targeting predictors, another possible solution is to use a one-step supervised procedure, such as partial least squares (PLS), to form our factors (Kelly and Pruitt 2013, 2015).

Given the above discussion, we report the forecasting performance of 9 different models. The first is a linear AR(1) model: $\hat{Y}_{t+1} = \hat{\alpha}_0 + \hat{\alpha} Y_t$. The second model is the Diffusion Index model (DI)

$$\hat{Y}_{t+1} = \hat{\beta}_0 + \hat{\alpha}_1 Y_t + \hat{\boldsymbol{\gamma}} \hat{\boldsymbol{F}}_t, \tag{8}$$

where $\hat{\boldsymbol{F}}_t = (\hat{F}_{t,1}, \ldots, \hat{F}_{t,k})$ are k factors which are estimated as the first k principal components of $\mathbf{z}_t$. We then combine each of the six screening methods under consideration with a second-stage DI model. For each screening method, we select the top

---

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

Table 5. Percent improvement in MSFE over AR(1)

| | CRSP | | SP 500 | |
	25 FF Portfolios	100 FF Portfolios	25 FF Portfolios	100 FF Portfolios
SIS	−1.07	−3.38	−1.96	−3.99
PDC-SIS	**0.79**∗∗	**0.86**∗∗	**0.56**∗	**0.54**∗
DC-SIS	−0.29	0.00	−0.44	−0.70
PDC-SIS+	0.37	0.80∗∗	**0.56**∗	**0.54**∗
NIS	0.25	−0.32	−1.22	−1.60
GLSS	−1.11	−.3.41	−1.95	−3.88
DI	0.28	−1.42	−0.21	−0.11
PLS	−6.66	−11.08	−8.55	−12.00

NOTES: For all procedures besides "DI" we use a screened consisting of the top $d_n = \lceil p_n/10 \rceil$ predictors. Where our predictors consist of the book-to-market ratio of 25 or 100 Fama French size and value sorted portfolios along with three of their lags. Our response is the monthly returns of the CRSP and SP500 indices. In this and the following tables, ***, **, and * indicates significance at 0.01, 0.05, and 0.1 levels using the forecast encompassing test "ENC-NEW", respectively.

$d_n = \lceil p_n/10 \rceil$ predictors of z_t, and use these predictors form our k factors. We use BIC to select k among values between 2 and 5. Lastly, we consider the DI model (8) using PLS to estimate F_t.

Following Kelly and Pruitt (2013), we form expanding window out of sample forecasts, where the first out of sample forecast is for the time period 1980:1 (January 1980), and the last forecast is for time period 2010:11 (November 2010). To construct the forecast for 1980:1, we use the observations between 1930:1 and 1979:11 to estimate the factors and model parameters. Therefore, for the models described previously, $t = 1930:1$ to 1979:10. We then use the predictor values at $t = 1979:11$ to form our forecast for 1980:1. The next window uses observations from 1930:1 to 1980:1 to forecast 1980:2. This gives us a total of 372 out of sample forecasts. For each of our eight models, the predictive ability is reported through the percentage improvement in mean squared forecast error (MSFE) over the baseline AR model. Specifically, we have

$$100 * \left(1 - \frac{\sum_{t=T_0}^{2010:11}(\hat{Y}_t - Y_t)^2}{\sum_{t=T_0}^{2010:11}(\hat{Y}_{t,AR} - Y_t)^2}\right) \quad (9)$$

where $T_0 = 1980:1$ and $\hat{Y}_{t,AR}$ is the forecast for time t made with the AR(1) model. This measure ranges from $(-\infty, 100]$, where 100 indicates perfect out of sample prediction, and negative values indicating that the method is outperformed by a baseline AR(1) forecast. We conduct out of sample inference by using the forecast encompassing test "ENC-NEW" derived by Clark and McCracken (2001). We use this test for all methods which outperform the baseline AR(1).

We report the results when forecasting the SP500 and CRSP market index using either 25 FF portfolios and 100 FF portfolios in Table 5. We observe that in both cases that finding targeted predictors via PDC-SIS and PDC-SIS+ outperform the alternatives, with the next best model is formed using DC-SIS. On the other hand, linear screening procedures such as SIS and GLSS underperform a factor model estimated on all the predictors and underperform the mean forecast in most cases as well. Interestingly, PLS is our worst performing model. PLS takes into account the correlation between z_t and the response when estimating the factors, and it appears that the high degree of correlation between predictors as well as a large number of

Table 6. Percent improvement in MSFE over AR(1)

	Quintile 1 (Small)	Quintile 2	Quintile 3	Quintile 4	Quintile 5 (Large)
SIS	−1.80	−1.13	−1.76	−2.34	−1.56
PDC-SIS	0.31	1.18∗∗∗	0.81∗∗	**0.12**	−0.76
DC-SIS	0.65∗	0.97∗∗	**0.91**∗∗	−0.11	−0.79
PDC-SIS+	**0.71**∗∗	**1.25**∗∗∗	0.72∗∗	−0.33	−0.55
NIS	−2.36	−1.45	−1.50	−2.14	−0.58
GLSS	−1.19	−.44	−1.28	−2.02	−1.64
DI	−1.07	−0.42	−0.45	−0.89	**−0.14**

NOTES: See notes to Table 5. We provide results to forecasting the monthly returns to 5 FF size sorted portfolios.

irrelevant covariates is deteriorating its performance. From the results, we see that in general nonlinear screening methods, such as DC-SIS and PDC-SIS, outperform linear screening methods. This suggests that accounting for nonlinearities in marginal relationships is beneficial even when using linear second-stage procedures.

In Table 6 we report the results when forecasting the 5 FF size sorted portfolios.[8] The first quintile corresponds to small-cap stocks, and we see distance correlation methods strongly outperform other methods for this portfolio. Interestingly, in contrast to Kelly and Pruitt (2013), we obtain the highest predictability for this portfolio. Moreover, we generally find portfolios corresponding to smaller cap stocks easier to forecast than larger-cap stocks using distance correlation methods. Once again, in almost all cases, distance correlation-based methods outperform competing screening methods.

As stated previously, we used a sample split date of 1980:1 for our out of sample forecasts. In order to show the robustness of our results to this choice of split date, we plot the R^2_{OOS} for the range of sample split dates between $T_0 = 1960:1$ to $T_0 = 1995:1$ in Figure 1. We plot this for both the CRSP index and the SP 500 index, using 100 FF portfolios as predictors. For the convenience of the presentation, we omit the performance of GLSS and PDC-SIS+ models in our plot, given their very close performance with PDC-SIS and SIS, respectively. We see from the plot that PDC-SIS models outperform the alternatives over almost the entire range of sample split points. We also observe using a DI model estimated on all the predictors, along with linear screening methods underperform the baseline AR(1) forecast over the range of sample split points.

6. Discussion

In this work, we have introduced two classes of PDC-based screening procedures, which are applicable to univariate and multivariate time series models. These methods aim to use the unique features of time series data as an additional source of information, rather than treating temporal dependence as a nuisance. The methods introduced can be readily used by researchers, given that distance correlation methods are computable at low cost by existing statistical packages. Lastly, by using a model-free first stage procedure, we can expand the

---

[8]We used the 25 FF portfolios as possible predictors along with their lags. The results were qualitatively similar for the 100 FF portfolio setting; thus we omit its results due to space considerations. Due to the poor performance of PLS, we omit its results for the remainder of the section.
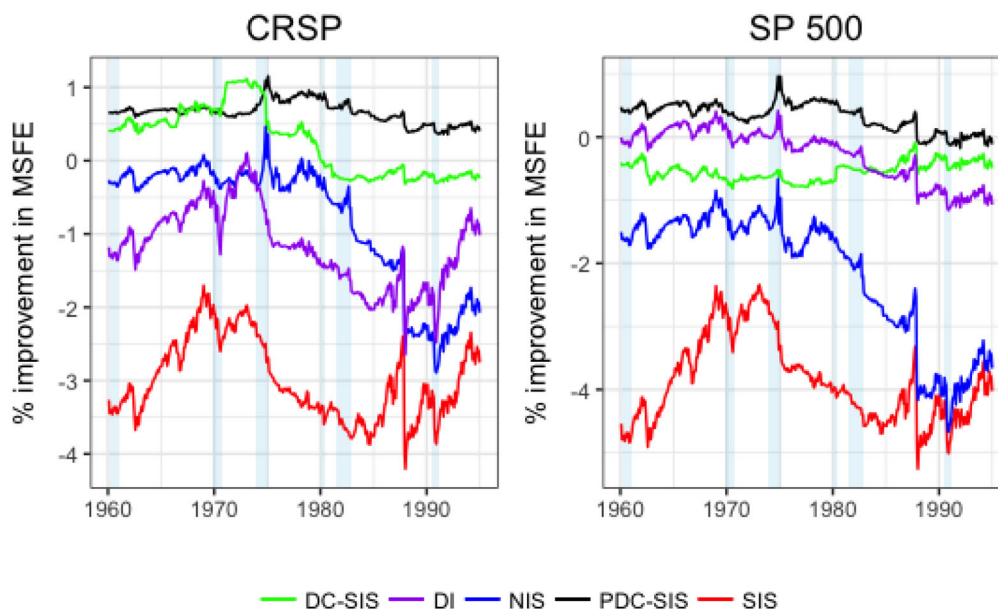
Figure 1. Percent improvement in MSFE vs. sample split date. We select each date between 1960:1–1995:1 as our sample split point and plot the corresponding percentage of improvement in MSFE over the baseline. We omit the values for GLSS and PDC+ due to having very close results to SIS and PDC, respectively. We used 100 FF portfolios and their lags as possible predictors.

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 build screening algorithms for time series data using measures that are more robust to heavy-tailed distributions. Also, our procedures were developed under the assumption that the underlying processes are weakly dependent and stationary. Although these assumptions are satisfied for an extensive range of applications, there are many instances where they are violated. For example, nonstationarity is commonly induced by time-varying parameters, structural breaks, and co-integrated processes, all of which are common in the fields of macroeconomics and finance. Therefore, developing new methodologies for certain classes of nonstationary processes, such as locally stationary processes, would be particularly welcome. Another important problem is to study the robustness of the PDC-SIS+ method for the multivariate case when the group structure is misspecified.

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 a comparison between PDC-SIS and CDC-SIS, the proofs for all the theorems, as well as more detailed tables of the results from Section 4.

Acknowledgments

The authors sincerely thank the editor, associate editor, and referees for insightful comments that significantly improved the article.

References

An, H., and Huang, F. (1996), "The Geometrical Ergodicity of Nonlinear Autoregressive Models," *Statistica Sinica*, 6, 943–956. [6]

Bai, J., and Ng, S. (2008), "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, 146, 304–317. [1,5,10]

——— (2009), "Boosting Diffusion Indices," *Journal of Applied Econometrics*, 24, 607–629. [2]

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by Supervised Principal Components," *Journal of the American Statistical Association*, 101, 119–137. [10]

Barut, E., Fan, J., and Verhasselt, A. (2016), "Conditional Sure Independence Screening," *Journal of the American Statistical Association*, 111, 1266–1277. [2,4]

Basu, S., and Michailidis, G. (2015), "Regularized Estimation in Sparse High-dimensional Time Series Models," *The Annals of Statistics*, 43, 1535–1567. [5]

Basu, S., Shojaie, A., and Michailidis, G. (2015), "Network Granger Causality With Inherent Grouping Structure," *Journal of Machine Learning Research*, 16, 417–453. [5]

Callot, L., Caner, M., Kock, A. B., and Riquelme, J. A. (2017), "Sharp Threshold Detection Based on Sup-norm Error Rates in High-dimensional Models," *Journal of Business & Economic Statistics*, 35, 250–264. [4]

Carrasco, M., and Chen, X. (2002), "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 18, 17–39. [6]

Chen, J., Li, D., Linton, O., and Lu, Z. (2018), "Semiparametric Ultra-high Dimensional Model Averaging of Nonlinear Dynamic Time Series," *Journal of the American Statistical Association*, 113, 919–932. [1,2,8]

Clark, T. E., and McCracken, M. W. (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85 – 110. Forecasting and empirical methods in finance and macroeconomics. [11]

Davidson, J. (1994), *Stochastic Limit Theory, An Introduction for Econometricians*. Oxford: Oxford University Press. [6]

Davis, R. A., Matsui, M., Mikosch, T., and Wan, P. (2016), "Applications of Distance Correlation to Time Series," *arXiv preprint:1606.05481*. [2,3,7]

Doan, T., Litterman, R., and Sims, C. (1984), "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews*, 3, 1–100. [2]

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. [1,4,8]

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. [1]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space w/ Discussion," *Journal of Royal Statistical Society*, Series B, 70, 849–911. [1,4,8]

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. [1]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-dimensionality," *Annals of Statistics*, 38, 3567–3604. [1]

Fan, J., Xue, L., and Yao, J. (2017), "Sufficient Forecasting Using Factor Models," *Journal of Econometrics*, 201, 292–306. [2]

Feng, Y., Wu, Y., and Stefanski, L. A. (2018), "Nonparametric Independence Screening Via Favored Smoothing Bandwidth," *Journal of Statistical Planning and Inference*, 197, 1–14. [1]

Fokianos, K., and Pitsillou, M. (2017), "Consistent Testing for Pairwise Dependence in Time Series," *Technometrics*, 59, 262–270. [3]

Gu, S., Kelly, B., and Xiu, D. (2018), "Empirical Asset Pricing Via Machine Learning," Working Paper 25398, National Bureau of Economic Research. [2]

——— (2019), "Autoencoder Asset Pricing Models," Available at SSRN. [2]

Kelly, B., and Pruitt, S. (2013), "Market Expectations in the Crosssection of Present Values," *The Journal of Finance*, 68, 1721–1756. [10,11]

——— (2015), "The Three-pass Regression Filter: A New Approach to Forecasting Using Many Predictors," *Journal of Econometrics*, 186, 294 – 316. High Dimensional Problems in Econometrics. [10]

Kilian, L., and Lütkepohl, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge: Cambridge University Press. [5]

Kock, A. and Callot, A. (2015), "Oracle Inequalities for High Dimensional Vector Autoregressions," *Journal of Econometrics*, 186, 325 – 344. [5]

Li, R., Zhu, L., and Zhong, W. (2012), "Feature Screening Via Distance Correlation," *The Journal of the American Statistical Association*, 107, 1129–1139. [1,7,8]

Litterman, R. B. (1986), "Forecasting With Bayesian Vector Autoregressionsfive Years of Experience," *Journal of Business & Economic Statistics*, 4, 25–38. [2]

Liu, J., Zhong, W., and Li, R. (2015), "A Selective Overview of Feature Screening for Ultrahigh-dimensional Data," *Science China Mathematics*, 58, 1–22. [1]

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*, 60, 1–19. [1,2]

Lu, Z. (1998), "On the Geometric Ergodicity of a Non-linear Autoregressive Model With an Autoregressive Conditional Heteroscedastic Term," *Statistica Sinica*, 8. 1205–1217. [6]

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*. Berlin: Springer. [5]

Ma, S., Li, R., and Tsai, C.-L. (2017), "Variable Screening Via Quantile Partial Correlation," *Journal of the American Statistical Association*, 112, 650–663. [1]

Mai, Q., Zou, H., et al. (2015), "The Fused Kolmogorov Filter: A Nonparametric Model-free Screening Method," *The Annals of Statistics*, 43, 1471–1497. [1]

Masry, E., and Tjøstheim, D. (1997), "Additive Nonlinear ARX Time Series and Projection Estimates," *Econometric Theory*, 13, 214–252. [8]

Nicholson, W. B., Bien, J., and Matteson, D. S. (2016), "Hierarchical Vector Autoregression," arXiv preprint:1412.5250. [5]

Pham, T. D. and Tran, L. T. (1985), "Some Mixing Properties of Time Series Models," *Stochastic Processes and their Applications*, 19(2):297–303. [6]

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society*, Series B, 71, 1009–1030. [8]

Stock, J. H., and Watson, M. W. (2002a), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [2,10]

——— (2002b), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–162. [10]

Székely, G. J., and Rizzo, M. L. (2014), "Partial Distance Correlation With Methods for Dissimilarities," *Annals Statistics*, 42, 2382–2412. [2,3]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *Annals Statistics*, 35, 2769–2794. [2,3]

Teräsvirta, T., Tjostheim, D., and Granger, C. (2010), *Modelling Nonlinear Economic Time Series*. Oxford: Oxford University Press. [9]

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. [5]

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015), "Conditional Distance Correlation," *Journal of the American Statistical Association*, 110, 1726–1734. PMID: 26877569. [2]

Wen, C., Pan, W., Huang, M., and Wang, X. (2018), "Sure Independence Screening Adjusted for Confounding Covariates With Ultrahigh-dimensional Data," *Statistica Sinica*, 28, 293–317. [2]

Weng, H., Feng, Y., and Qiao, X. (2019), "Regularization After Retention in Ultrahigh Dimensional Linear Regression Models," *Statistica Sinica*, 29, 387–407. [4]

Wu, W., and Pourahmadi, M. (2009), "Banding Sample Autocovaraince Matrices of Stationary Processes," *Statistica Sinica*, 19, 1755–1768. [6,7]

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. [6,7]

Wu, W. B. (2005), "Nonlinear System Theory: Another Look at Dependence," *Proceedings of the National Academy of Sciences*, 102, 14150–14154. [6]

——— (2009), "An Asymptotic Theory for Sample Covariances of Bernoulli Shifts," *Stochastic Processes and Their Applications*, 119, 453 – 467. [6]

——— (2011), "Asymptotic Theory for Stationary Processes," *Statistics and Its Interface*, 4, 207–226. [6]

Xiao, H., and Wu, W. B. (2012), "Covariance Matrix Estimation for Stationary Time Series," *Annals of Statistics*, 40, 466–493. [7]

Yousuf, K. (2018), "Variable Screening for High Dimensional Time Series," *Electronic Journal of Statistics*, 12, 667–702. [1,2,6,7,8]

Yousuf, K., and Ng, S. (2020), "Boosting High Dimensional Predictive Regressions With Time Varying Parameters," *Journal of Econometrics*. [2]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68,49–67. [5]

Zhou, Z. (2012), "Measuring Nonlinear Dependence in Time-series, a Distance Correlation Approach," *Journal of Time Series Analysis*, 33, 438–457. [2,3]

——— (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [5]