# Semiparametric Ultra-High Dimensional Model Averaging of Nonlinear Dynamic Time Series

Jia Chen[*] Degui Li[†] Oliver Linton[‡] Zudi Lu[§]

February 3, 2016

## Abstract

We propose two semiparametric model averaging schemes for nonlinear dynamic time series regression models with a very large number of covariates including exogenous regressors and auto-regressive lags, aiming to obtain accurate forecasts of time series by using a large number of conditioning variables in a nonparametric way. In the first scheme, we introduce a Kernel Sure Independence Screening (KSIS) technique to screen out the regressors whose marginal regression (or auto-regression) functions do not make significant contribution to estimating the joint multivariate regression function; we then propose a semiparametric penalised method of Model Averaging MArginal Regression

[*] Department of Economics and Related Studies, University of York, Heslington, YO10 5DD, UK. E-mail: jia.chen@york.ac.uk

[†] Department of Mathematics, University of York, Heslington, YO10 5DD, UK. E-mail: degui.li@york.ac.uk.

[‡] Faculty of Economics, Cambridge University, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, UK. E-mail: obl20@cam.ac.uk.

[§] Statistical Sciences Research Institute and School of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. E-mail: Z.Lu@soton.ac.uk. Partially supported by the Marie Curie career integration grant of European Commission.

(MAMAR) for the regressors and auto-regressors that survive the screening procedure, to further select the regressors that have significant effects on estimating the multivariate regression function and predicting the future values of the response variable. In the second scheme, we impose an approximate factor modelling structure on the ultra-high dimensional exogenous regressors and use a popular principal component analysis to estimate the latent common factors; we then apply the penalised MAMAR method to select the estimated common factors and the lags of the response variable that are significant. In each of the two schemes, we ultimately determine the optimal combination of the significant marginal regression and auto-regression functions. Asymptotic properties for these two schemes are derived under some regularity conditions. Numerical studies including both simulation and an empirical application are illustrated on the proposed methodology.

*Keywords*: Kernel smoother, penalised MAMAR, principal component analysis, semiparametric approximation, sure independence screening, ultra-high dimensional time series.

# 1 Introduction

Nonlinear time series modelling taking account of both dynamic lags of response variable and exogenous regressors is of wide interest in applications. We suppose that $Y_t$, $t = 1, \ldots, n$, are $n$ observations collected from a stationary time series process, and often we are interested in the multivariate dynamic regression function

$$m(\mathbf{x}) = \mathsf{E}(Y_t \big| \mathbf{X}_t = \mathbf{x}), \tag{1.1}$$

where $Y_t$ is the response variable, and $\mathbf{X}_t = (\mathbf{Z}_t^\intercal, \mathbf{Y}_{t-1}^\intercal)^\intercal$ with $\mathbf{Z}_t = \big(Z_{t1}, Z_{t2}, \ldots, Z_{tp_n}\big)^\intercal$ and $\mathbf{Y}_{t-1} = (Y_{t-1}, Y_{t-2}, \ldots, Y_{t-d_n})^\intercal$ being a $p_n$-dimensional vector consisting of exogenous regressors and a vector of $d_n$ lags of $Y_t$, respectively. Here the superscript $^\intercal$ stands for the transpose of a vector (or a matrix). We allow that both $p_n$ and $d_n$

could increase with the sample size $n$, and $\mathbf{Z}_t$ could include lags of the exogenous regressors and has large dimension $p_n$, allowed to be even larger than the sample size $n$. Such an ultra-high dimensional time series setting poses a great challenge in estimating the regression function $m(\mathbf{x})$ and the subsequent forecast of the response.

When the dimension of $\mathbf{X}_t$ is low (say 1 or 2), it is well known that the conditional regression function $m(\mathbf{x})$ can be well estimated by using some commonly-used nonparametric methods such as the kernel method, the local polynomial method, and the spline method (c.f., Green and Silverman, 1994; Wand and Jones, 1995; Fan and Gijbels, 1996). However, if $\mathbf{X}_t$ is of large dimension, owing to the so-called "curse of dimensionality", a direct use of nonparametric methods leads to a very poor estimation and forecasting result. Hence, various semiparametric models, such as additive models, varying coefficient models and partially linear models, have been proposed to deal with the curse of dimensionality (c.f., Teräsvirta, Tjøstheim and Granger, 2010). Alternatively, Li, Linton and Lu (2015) recently developed a flexible semiparametric forecasting model, termed "*Model Averaging MArginal Regression*" (MAMAR). It seeks to optimally combine nonparametric low-dimensional marginal regressions, which helps to improve the accuracy of predicting future values of time series.

The idea of model averaging approach is to combine several candidate models by assigning higher weights to better candidate models. Under the linear regression setting with the dimension of covariates smaller than the sample size, there has been an extensive literature on various model averaging methods, see, for example, the AIC and BIC model averaging (Akaike, 1979; Raftery, Madigan and Hoeting,1997; Claeskens and Hjort, 2008), the Mallows $C_p$ model averaging (Hansen, 2007; Wan, Zhang and Zou, 2010), and the jackknife model averaging (Hansen and Racine, 2012). However, in the case of ultra-high dimensional time series, these methods may not perform well and the associated asymptotic theory fails. To address this issue, Ando and Li (2014) propose a two-step model averaging method for a high-dimensional linear regression with the dimension of the covariates larger than the sample size and show that such a method works well both theoretically and numerically. Cheng and Hansen (2015) study the model averaging of the factor-augmented linear regression by applying a principal component analysis on the high-dimensional covariates to estimate the unobservable factor regressors.

In this paper, our main objective is to propose semiparametric ultra-high dimensional model averaging schemes for studying the nonlinear dynamic regression structure for (1.1) which generalises the existing approaches. On one hand, we relax the restriction of linear modelling assumed in Ando and Li (2014) and Cheng and Hansen (2015), and on the other hand, we extend the recent work of Li, Linton and Lu (2015) to the ultra high dimensional case, thereby providing a much more flexible framework for nonlinear dynamic time series forecasting.

Throughout the paper, we assume that the dimension of the exogenous variables $\mathbf{Z}_t$, $p_n$, may diverge at an exponential rate of $n$, which implies that the potential explanatory variables $\mathbf{X}_t$ have the dimension of $p_n + d_n$ diverging at an exponential rate, i.e., $p_n + d_n = O(\exp\{n^{\delta_0}\})$ for some positive constant $\delta_0$. To ensure that our semiparametric model averaging scheme is feasible both theoretically and numerically, we need to reduce the dimension of the potential covariates $\mathbf{X}_t$ and select those variables that make a significant contribution to predicting the response. In this paper we propose two schemes to achieve the purpose of dimension reduction. The first scheme is called as the "*KSIS+PMAMAR*" method. It reduces the dimension of the potential covariates by first using the approach of *Kernel Sure Independence Screening* (KSIS), motivated by Fan and Lv (2008), to screen out the unimportant marginal regression (or auto-regression) functions, and then apply the so-called *Penalised Model Averaging MArginal Regression* (PMAMAR) to further select the most relevant regression functions. The second scheme is called as the "*PCA+PMAMAR*"method. In this scheme, we assume that the ultra-high dimensional exogenous regressors $\mathbf{Z}_t$ satisfy an approximate factor model which has been popular in many fields such as economic and financial data analysis (c.f., Chamberlain and Rothschild, 1983; Fama and French, 1992; Stock and Watson, 2002; Bai and Ng, 2002, 2006), and estimate the factor regressors using the *Principal Component Analysis* (PCA). Then, similarly to the second step in the first scheme, the PMAMAR method is applied to further select the significant estimated factor regressors and auto-regressors.

Under some regularity conditions, we develop the asymptotic properties of the proposed methods. For the KSIS procedure, we establish the sure screening property indicating that the covariates, whose marginal regression functions make truly significant contribution to estimating the multivariate regression function $m(\mathbf{x})$, would be selected with probability approaching to one to form a set of the regressors that would undergo a further selection in the PMAMAR procedure. For the PCA approach, we show that the estimated latent factors are uniformly consistent at a convergence rate that depends on both $n$ and $p_n$, and the kernel estimation of the marginal regression with the estimated factor regressors is asymptotically equivalent to that with rotated true factor regressors. For the PMAMAR procedure in either of the two semiparametric dimension-reduction schemes, we prove that the optimal weight estimation enjoys the well-known sparsity and oracle property that the estimated values of the true zero weights are forced to be zero. Furthermore, extensions of the proposed semiparametric dimension reduction approaches such as an iterative KSIS+PMAMAR procedure will be discussed. In the simulation studies, we find that our methods outperform some existing methods in terms of forecasting accuracy, and often have low prediction errors whose values are close to those using the oracle estimation. We finally apply our methods to forecasting quarterly inflation in the UK.

The rest of the paper is organised as follows. The two semiparametric model averaging schemes are proposed in Section 2. The asymptotic theory for them is then developed in Section 3. Section 4 discusses some extensions when the methods are implemented in practice. Numerical studies are reported in Section 5 including two simulated and one empirical data examples. Section 6 concludes. The proofs are given in a supplemental document.

# 2 Semiparametric model averaging

In this section, we propose two semiparametric model averaging approaches in dimension reduction of the ultra-high dimensional covariates, which are named as the KSIS+PMAMAR and the PCA+PMAMAR in Sections 2.1 and 2.2, respectively.

## 2.1 *KSIS+PMAMAR method*

As mentioned in Section 1, the KSIS+PMAMAR method is a two-step procedure. We first generalise the Sure Independence Screening (SIS) method introduced by Fan and Lv (2008) to the ultra-high dimensional dynamic time series and general semiparametric setting to screen out covariates whose nonparametric marginal regression functions have low correlations with the response. Then, for the covariates that have survived the screening, we propose a PMAMAR method with first-stage kernel smoothing to further select the exogenous regressors and the lags of the response variable which make significant contribution to estimating the multivariate regression function, and to determine an optimal linear combination of the significant marginal regression and auto-regression functions.

**Step one: KSIS**. For notational simplicity, we let

$$X_{tj} = \begin{cases} Z_{tj}, & j = 1, \dots, p_n, \\ Y_{t-(j-p_n)}, & j = p_n + 1, \dots, p_n + d_n. \end{cases}$$

To measure the contribution made by the univariate covariate $X_{tj}$ to estimating the multivariate regression function $m(\mathbf{x}) = \mathsf{E}(Y_t \big| \mathbf{X}_t = \mathbf{x})$, we consider the marginal regression function defined by

$$m_j(x_j) = \mathsf{E}\big(Y_t | X_{tj} = x_j\big), \quad j = 1, \ldots, p_n + d_n,$$

which is the projection of $Y_t$ onto the univariate component space spanned by $X_{tj}$. This function can also be seen as the solution to the following nonparametric optimisation problem(c.f., Fan, Feng and Song, 2011):

$$\min_{g_j \in \mathcal{L}_2(\mathsf{P})} \mathsf{E}\big[Y_t - g_j(X_{tj})\big]^2,$$

where $\mathcal{L}_2(\mathsf{P})$ is the class of square integrable functions under the probability measure $\mathsf{P}$. We estimate the functions $m_j(\cdot)$ by the commonly-used kernel smoothing method, although other nonparametric estimation methods such as the local polynomial smoothing and smoothing spline method are also applicable. The kernel smoother of $m_j(x_j)$ is

$$\hat{m}_j(x_j) = \frac{\sum_{t=1}^n Y_t K_{tj}(x_j)}{\sum_{t=1}^n K_{tj}(x_j)}, \quad K_{tj}(x_j) = K\Big(\frac{X_{tj} - x_j}{h_1}\Big), \quad j = 1, \ldots, p_n + d_n, \tag{2.1}$$

where $K(\cdot)$ is a kernel function and $h_1$ is a bandwidth. To make the above kernel estimation method feasible, we assume that the initial observations, $Y_{-1}, Y_{-2}, \ldots, Y_{-d_n}$, of the response are available.

When the observations are independent and the response variable has zero mean, the paper of Fan, Feng and Song (2011) ranks the importance of the covariates by calculating the $\mathcal{L}_2$-norm of $\hat{m}_j(\cdot)$, and chooses those covariates whose corresponding norms are larger than a pre-determined threshold value that usually tends to zero. However, in the time series setting for $j$ such that $j - p_n \to \infty$, we may show that under certain stationarity and weak dependence conditions

$$\hat{m}_j(x_j) \overset{P}{\to} m_j(x_j) \to \mathsf{E}[Y_t].$$

When $\mathsf{E}[Y_t]$ is non-zero, the norm of $\hat{m}_j(\cdot)$ would tend to a non-zero quantity. As a consequence, if covariates are chosen according to the $\mathcal{L}_2$-norm of their corresponding marginal regression functions, quite a few unimportant lags might be chosen. To address this issue, we consider ranking the importance of the covariates by calculating the correlation between the response variable and marginal regression

$$\mathsf{cor}(j) = \frac{\mathsf{cov}(j)}{\sqrt{\mathsf{v}(Y)\cdot\mathsf{v}(j)}} = \left[\frac{\mathsf{v}(j)}{\mathsf{v}(Y)}\right]^{1/2}, \tag{2.2}$$

where $\mathsf{v}(Y) = \mathsf{var}(Y_t)$, $\mathsf{v}(j) = \mathsf{var}(m_j(X_{tj}))$ and $\mathsf{cov}(j) = \mathsf{cov}(Y_t, m_j(X_{tj})) = \mathsf{var}(m_j(X_{tj})) = \mathsf{v}(j)$. Equation (2.2) indicates that the value of $\mathsf{cor}(j)$ is non-negative for all $j$ and the ranking of $\mathsf{cor}(j)$ is equivalent to the ranking of $\mathsf{v}(j)$ as $\mathsf{v}(Y)$ is positive and invariant across $j$. The sample version of $\mathsf{cor}(j)$ can be constructed as

$$\widehat{\mathsf{cor}}(j) = \frac{\widehat{\mathsf{cov}}(j)}{\sqrt{\hat{\mathsf{v}}(Y)\cdot\hat{\mathsf{v}}(j)}} = \left[\frac{\hat{\mathsf{v}}(j)}{\hat{\mathsf{v}}(Y)}\right]^{1/2}, \tag{2.3}$$

where

$$\hat{\mathsf{v}}(Y) = \frac{1}{n}\sum_{t=1}^{n} Y_t^2 - \Big(\frac{1}{n}\sum_{t=1}^{n} Y_t\Big)^2,$$

$$\widehat{\mathsf{cov}}(j) = \hat{\mathsf{v}}(j) = \frac{1}{n}\sum_{t=1}^{n} \hat{m}_j^2(X_{tj}) - \Big[\frac{1}{n}\sum_{t=1}^{n} \hat{m}_j(X_{tj})\Big]^2, \; j = 1, 2, \ldots, p_n + d_n.$$

The screened sub-model can be determined by,

$$\hat{\mathcal{S}} = \big\{ j = 1, 2, \ldots, p_n + d_n : \; \hat{\mathsf{v}}(j) \geq \rho_n \big\}, \tag{2.4}$$

where $\rho_n$ is a pre-determined positive number. By (2.3), the criterion in (2.4) is equivalent to

$$\hat{\mathcal{S}} = \big\{ j = 1, 2, \ldots, p_n + d_n : \; \widehat{\mathsf{cor}}(j) \geq \rho_n^{\diamond} \big\},$$

where $\rho_n^{\diamond} = \rho_n^{1/2}/\sqrt{\hat{\mathsf{v}}(Y)}$. As in Section 1, we let $\mathbf{X}_t^* = \big(X_{t1}^*, X_{t2}^*, \ldots, X_{tq_n}^*\big)^{\intercal}$ be the covariates chosen according to the criterion (2.4).

The above model selection procedure can be seen as the nonparametric kernel extension of the SIS method, which was first introduced by Fan and Lv (2008) in the context of linear regression models. Recent extensions to nonparametric additive models and varying coefficient models can be found in Fan, Feng and Song (2011), Fan, Ma and Dai (2014) and Liu, Li and Wu (2014). However, the existing literature only considers the case where the observations are independent, which rules out time series applications. In this paper, we relax such a restriction and show that the KSIS approach works well in the ultra-high dimensional time series and semiparametric setting. Also, differently from Fan, Feng and Song (2011) using the B-splines method, our paper applies the kernel smoothing method to estimate the marginal regression functions, with different mathematical tool needed to derive our asymptotic theory.

**Step two: PMAMAR.** In the second step, we propose using a semiparametric method of model averaging lower dimensional regression functions to estimate

$$m^*(\mathbf{x}) = \mathsf{E}(Y_t|\mathbf{X}_t^* = \mathbf{x}), \tag{2.5}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_{q_n})^\intercal$. Specifically, we approximate the conditional regression function $m^*(\mathbf{x})$ by an affine combination of one-dimensional conditional component regressions

$$m_j^*(x_j) = \mathsf{E}(Y_t|X_{tj}^* = x_j), \quad j = 1, \ldots, q_n.$$

Each marginal regression $m_j^*(\cdot)$ can be treated as a "nonlinear candidate model" and the number of such nonlinear candidate models is $q_n$. A weighted average of $m_j^*(x_j)$ is then used to approximate $m^*(\mathbf{x})$, i.e.,

$$m^*(\mathbf{x}) \approx w_0 + \sum_{j=1}^{q_n} w_j m_j^*(x_j), \tag{2.6}$$

where $w_j$, $j = 0, 1, \ldots, q_n$, are to be determined later and can be seen as the weights for different candidate models. The linear combination in (2.6) is called as Model Averaging MArginal Regressions or MAMAR (Li, Linton and Lu, 2015) and is applied by Chen *et al* (2016) in the dynamic portfolio choice with many conditioning variables. As the conditional component regressions $m_j^*(X_{tj}^*) = \mathsf{E}(Y_t|X_{tj}^*)$, $j = 1, \ldots, q_n$, are unknown but univariate, in practice, they can be well estimated by various nonparametric approaches that would not suffer from the curse of dimensionality problem.

Hence, the first stage in the semiparametric PMAMAR procedure is to estimate the marginal regression functions $m_j^*(\cdot)$ by the kernel smoothing method

$$\hat{m}_j^*(x_j) = \frac{\sum_{t=1}^n Y_t \overline{K}_{tj}(x_j)}{\sum_{t=1}^n \overline{K}_{tj}(x_j)}, \quad \overline{K}_{tj}(x_j) = K\Big(\frac{X_{tj}^* - x_j}{h_2}\Big), \quad j = 1, \ldots, q_n, \tag{2.7}$$

where $h_2$ is a bandwidth. Let

$$\hat{\mathcal{M}}(j) = \big[\hat{m}_j^*(X_{1j}^*), \ldots, \hat{m}_j^*(X_{nj}^*)\big]^\intercal$$

be the estimated values of

$$\mathcal{M}(j) = \big[m_j^*(X_{1j}^*), \ldots, m_j(X_{nj}^*)\big]^\intercal$$

for $j = 1, \ldots, q_n$. By using (2.7), we have

$$\hat{\mathcal{M}}(j) = \mathcal{S}_n(j)\mathcal{Y}_n, \quad j = 1, \ldots, q_n,$$

where $\mathcal{S}_n(j)$ is the $n \times n$ smoothing matrix whose $(k, l)$-component is $\overline{K}_{lj}(X_{kj}^*)/\big[\sum_{t=1}^n \overline{K}_{tj}(X_{kj}^*)\big]$, and $\mathcal{Y}_n = (Y_1, \ldots, Y_n)^\intercal$.

The second stage of PMAMAR is to replace $m_j^*(X_{tj}^*)$, $j = 1, \ldots, q_n$, by their corresponding nonparametric estimates $\hat{m}_j^*(X_{tj}^*)$, and use the penalised approach to select the significant marginal regression functions in the following "approximate linear model":

$$Y_t \approx w_0 + \sum_{j=1}^{q_n} w_j \hat{m}_j^*(X_{tj}^*). \tag{2.8}$$

Without loss of generality, we further assume that $\mathsf{E}(Y_t) = 0$, otherwise, we can simply replace $Y_t$ by $Y_t - \overline{Y} = Y_t - \frac{1}{n}\sum_{s=1}^n Y_s$. It is easy to show that the intercept term $w_0$ in (2.6) is zero under this assumption. In the sequel, we let $\mathbf{w}_o := \mathbf{w}_{on} = (w_{o1}, \ldots, w_{oq_n})$ be the optimal values of the weights in the model averaging. Based on the approximate linear modelling framework (2.8), for given $\mathbf{w}_n = (w_1, \ldots, w_{q_n})^\intercal$, we define the objective function by

$$\mathcal{Q}_n(\mathbf{w}_n) = \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_n)\big]^\intercal \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_n)\big] + n\sum_{j=1}^{q_n} p_\lambda(|w_j|), \tag{2.9}$$

where

$$\hat{\mathcal{M}}(\mathbf{w}_n) = \big[w_1\mathcal{S}_n(1) + \ldots + w_{q_n}\mathcal{S}_n(q_n)\big]\mathcal{Y}_n = \mathcal{S}_n(\mathcal{Y})\mathbf{w}_n,$$

$\mathcal{S}_n(\mathcal{Y}) = \big[\mathcal{S}_n(1)\mathcal{Y}_n, \ldots, \mathcal{S}_n(q_n)\mathcal{Y}_n\big]$, and $p_\lambda(\cdot)$ is a penalty function with a tuning parameter $\lambda$. The vector $\hat{\mathcal{M}}(\mathbf{w}_n)$ in (2.6) can be seen as the kernel estimate of

$$\mathcal{M}(\mathbf{w}_n) = \Big[\sum_{j=1}^{q_n} w_j m_j^*(X_{1j}^*), \ldots, \sum_{j=1}^{q_n} w_j m_j^*(X_{nj}^*)\Big]^{\intercal}$$

for given $\mathbf{w}_n$. Our semiparametric estimator of the optimal weights $\mathbf{w}_o$ can be obtained through minimising the objective function $\mathcal{Q}_n(\mathbf{w}_n)$:

$$\hat{\mathbf{w}}_n = \arg\min_{\mathbf{w}_n} \mathcal{Q}_n(\mathbf{w}_n). \tag{2.10}$$

There has been extensive discussion on the choice of the penalty function for parametric linear and nonlinear models. Many popular variable selection criteria, such as AIC and BIC, correspond to the penalised estimation method with $p_\lambda(|z|) = 0.5\lambda^2 I(|z| \neq 0)$ with different values of $\lambda$. However, as mentioned by Fan and Li (2001), such traditional penalised approaches are expensive in computational cost when $q_n$ is large. To avoid the expensive computational cost and the lack of stability, some other penalty functions have been introduced in recent years. For example, the LASSO penalty $p_\lambda(|z|) = \lambda|z|$ has been extensively studied by many authors (c.f., Tibshirani, 1996, 1997); Frank and Friedman (1993) consider the $L_q$-penalty $p_\lambda(|z|) = \lambda|z|^q$ for $0 < q < 1$; Fan and Li (2001) suggest using the SCAD penalty function which is defined by

$$p_\lambda'(z) = \lambda\left[I(z \le \lambda) + \frac{a_0\lambda - z}{(a_0 - 1)\lambda} I(z > \lambda)\right]$$

with $p_\lambda(0) = 0$, where $a_0 > 2$, $\lambda > 0$ and $I(\cdot)$ is the indicator function.

## 2.2 *PCA+PMAMAR method*

It is well known that we may also achieve dimension reduction through the use of factor models when analysing high-dimensional time series data. In this subsection, we assume that the high-dimensional exogenous variables $\mathbf{Z}_t$ follow the approximate factor model:

$$Z_{tk} = (\mathbf{b}_k^0)^\intercal \mathbf{f}_t^0 + u_{tk}, \quad k = 1, \ldots, p_n, \tag{2.11}$$

where $\mathbf{b}_k^0$ is an $r$-dimensional vector of factor loadings, $\mathbf{f}_t^0$ is an $r$-dimensional vector of common factors, and $u_{tk}$ is called an idiosyncratic error. The number of the common factors, $r$, is assumed to be fixed throughout the paper, but it is usually unknown in practice and its determination method will be discussed in Section 4 below.

From the approximate factor model (2.11), we can find that the main information in the exogenous regressors may be summarised in the common factors $\mathbf{f}_t^0$ that have a much lower dimension. The aim of dimension reduction can thus be achieved, and it may be reasonable to replace $\mathbf{Z}_t$ with an ultra-high dimension by the unobservable $\mathbf{f}_t$ with a fixed dimension in estimating the conditional multivariate regression function and predicting the future value of the response variable $Y_t$. In the framework of linear regression or autoregression, such an idea has been frequently used in the literature since Stock and Watson (2002) and Bernanke, Boivin and Eliasz (2005). However, so far as we know, there is virtually no work on combining the factor model (2.11) with the nonparametric nonlinear regression. The only exception is the paper by Härdle and Tsybakov (1995), which consider the additive regression model on principal components when the observations are independent and the dimension of the potential regressors is fixed. The latter restriction is relaxed in this paper.

Instead of directly studying the multivariate regression function $m(\mathbf{x})$ defined in (1.1), we next consider the multivariate regression function defined by

$$m_f(\mathbf{x}_1, \mathbf{x}_2) = \mathsf{E}\left(Y_t | \mathbf{f}_t^0 = \mathbf{x}_1, \mathbf{Y}_{t-1} = \mathbf{x}_2\right), \tag{2.12}$$

where $\mathbf{Y}_{t-1}$ is defined as in Section 1, $\mathbf{x}_1$ is $r$-dimensional and $\mathbf{x}_2$ is $d_n$-dimensional. In order to develop a feasible estimation approach for the factor augmented nonlinear regression function in (2.12), we need to estimate the unobservable factor regressors

$\mathbf{f}_t^0$ in the first step. This will be done through the PCA approach and we denote

$$\hat{\mathbf{X}}_{t,f}^* = \left(\hat{\mathbf{f}}_t^{\intercal}, \mathbf{Y}_{t-1}^{\intercal}\right)^{\intercal} = \left(\hat{f}_{t1}, \ldots, \hat{f}_{tr}, \ldots, \mathbf{Y}_{t-1}^{\intercal}\right)^{\intercal}$$

as a combination of the estimated factor regressors and lags of response variables, where $\hat{\mathbf{f}}_t$ is the estimated factor via PCA and $\hat{f}_{tk}$ is the $k$-th element of $\hat{\mathbf{f}}_t$, $k = 1, \ldots, r$. In the second step, we use the PMAMAR method to conduct a further selection among the $(r+d_n)$-dimensional regressors $\mathbf{X}_{t,f}^*$ and determine an optimal combination of the significant marginal regressions. This PCA+PMAMAR method substantially generalises the framework of factor-augmented linear regression or autoregression (c.f., Stock and Watson, 2002; Bernanke, Boivin and Eliasz, 2005; Bai and Ng, 2006; Pesaran, Pick and Timmermann, 2011; and Cheng and Hansen, 2015) to the general semiparametric framework.

**Step one: PCA on the exogenous regressors.** Letting

$$\mathbf{B}_n^0 = (\mathbf{b}_1^0, \ldots, \mathbf{b}_{p_n}^0)^{\intercal} \quad \text{and} \quad \mathbf{U}_t = (u_{t1}, \ldots, u_{tp_n})^{\intercal},$$

we may rewrite the approximate factor model (2.11) as

$$\mathbf{Z}_t = \mathbf{B}_n^0 \mathbf{f}_t^0 + \mathbf{U}_t. \tag{2.13}$$

We next apply the PCA approach to obtain the estimation of the common factors $\mathbf{f}_t^0$. Denote $\mathcal{Z}_n = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\intercal}$, the $n \times p_n$ matrix of the observations of the exogenous variables. We then construct $\hat{\mathcal{F}}_n = \left(\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_n\right)^{\intercal}$ as the $n \times r$ matrix consisting of the $r$ eigenvectors (multiplied by $\sqrt{n}$) associated with the $r$ largest eigenvalues of the $n \times n$ matrix $\mathcal{Z}_n \mathcal{Z}_n^{\intercal}/(np_n)$. Furthermore, the estimate of the factor loading matrix (with rotation) is defined as

$$\hat{\mathbf{B}}_n = \left(\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_{p_n}\right)^{\intercal} = \mathcal{Z}_n^{\intercal} \hat{\mathcal{F}}_n / n,$$

by noting that $\hat{\mathcal{F}}_n^{\intercal} \hat{\mathcal{F}}_n / n = I_r$.

As shown in the literature (see also Theorem 3 in Section 3.2 below), $\hat{\mathbf{f}}_t$ is a consistent estimator of the rotated common factor $\mathbf{H}\mathbf{f}_t$, where

$$\mathbf{H} = \hat{\mathbf{V}}^{-1}\left(\hat{\mathcal{F}}_n^{\intercal}\mathcal{F}_n^0/n\right)\left[(\mathbf{B}_n^0)^{\intercal}\mathbf{B}_n^0/p_n\right], \quad \mathcal{F}_n^0 = \left(\mathbf{f}_1^0, \ldots, \mathbf{f}_n^0\right)^{\intercal},$$

and $\hat{\mathbf{V}}$ is the $r \times r$ diagonal matrix of the first $r$ largest eigenvalues of $\mathcal{Z}_n\mathcal{Z}_n^{\intercal}/(np_n)$ arranged in descending order. Consequently, we may consider the following multivariate regression function with rotated latent factors:

$$m_f^*(\mathbf{x}_1, \mathbf{x}_2) = \mathsf{E}\left(Y_t|\mathbf{H}\mathbf{f}_t^0 = \mathbf{x}_1, \mathbf{Y}_{t-1} = \mathbf{x}_2\right). \tag{2.14}$$

In the subsequent PMAMAR step, we can use $\hat{\mathbf{f}}_t$ to replace $\mathbf{H}\mathbf{f}_t^0$ in the semiparametric procedure. The factor modelling and PCA estimation ensure that most of the useful information contained in the exogenous variables $\mathbf{Z}_t$ can be extracted before the second step of PMAMAR, which may lead to possible good performance in forecasting $Y_t$ through the use of the estimated common factors. In contrast, as discussed in some existing literature such as Fan and Lv (2008), when irrelevant exogenous variables are highly correlated with some relevant ones, they might be selected into a model by the SIS or KSIS procedure with higher priority than some other relevant exogenous variables, which results in high false positive rates and low true positive rates and leads to loss of useful information in the potential covariates, see, for example, the discussion in Section 4.1.

**Step two: PMAMAR using estimated factor regressors**. Recall that

$$\hat{\mathbf{X}}_{t,f}^* = \left(\hat{\mathbf{f}}_t^{\intercal}, \mathbf{Y}_{t-1}^{\intercal}\right)^{\intercal} = \left(\hat{f}_{t1}, \ldots, \hat{f}_{tr}, \mathbf{Y}_{t-1}^{\intercal}\right)^{\intercal},$$

where $\hat{f}_{tk}$ is the $k$-th element of $\hat{\mathbf{f}}_t$, $k = 1, \ldots, r$. We may apply the two-stage semiparametric PMAMAR procedure, which is exactly the same as that in Section 2.1 to the process $\left(Y_t, \hat{\mathbf{X}}_{t,f}^*\right)$, $t = 1, \ldots, n$, and then obtain the estimation of the optimal weights $\hat{\mathbf{w}}_{n,f}$. To save space, we next only sketch the kernel estimation of the marginal regression function with the estimated factor regressors obtained via PCA.

For $k = 1, \ldots, r$, define

$$m^*_{k,f}(z_k) = \mathsf{E}\left[Y_t | \tilde{f}^0_{tk} = z_k\right], \quad \tilde{f}^0_{tk} = e^\intercal_r(k)\mathbf{H}\mathbf{f}^0_t,$$

where $e_r(k)$ is an $r$-dimensional column vector with the $k$-th element being one and zeros elsewhere, $k = 1, \ldots, r$. As in Section 2.1, we estimate $m^*_{k,f}(z_k)$ by the kernel smoothing method:

$$\hat{m}^*_{k,f}(z_k) = \frac{\sum_{t=1}^n Y_t \widetilde{K}_{tk}(z_k)}{\sum_{t=1}^n \widetilde{K}_{tk}(z_k)}, \quad \widetilde{K}_{tk}(z_k) = K\Big(\frac{\hat{f}_{tk} - z_k}{h_3}\Big), \quad j = 1, \ldots r, \tag{2.15}$$

where $h_3$ is a bandwidth. In Section 3.2 below, we will show that $\hat{m}^*_{k,f}(z_k)$ is asymptotically equivalent to $\tilde{m}^*_{k,f}(z_k)$, which is defined as in (2.15) but with $\hat{f}_{tk}$ replaced by $\tilde{f}^0_{tk}$. The latter kernel estimation is infeasible in practice as the factor regressor involved is unobservable. As we may show that the asymptotic order of $\hat{m}^*_{k,f}(z_k) - \tilde{m}^*_{k,f}(z_k)$ is $o_P(n^{-1/2})$ under some mild conditions (c.f., Theorem 3), the influence of replacing $\tilde{f}^0_{tk}$ by the estimated factor regressors $\hat{f}_{tk}$ in the PMAMAR procedure is asymptotically negligible.

# 3 The main theoretical results

In this section, we establish the asymptotic properties for the methodologies developed in Section 2 above. The asymptotic theory for the KSIS+PMAMAR method is given in Section 3.1 and that for the PCA+PMAMAR method is given in Section 3.2.

## 3.1 *Asymptotic theory for KSIS+PMAMAR*

In this subsection, we first derive the sure screening property for the developed KSIS method, which implies that the covariates whose marginal regression functions make significant contribution to estimating the multivariate regression function $m(\mathbf{x})$ would be chosen in the screening with probability approaching one. The following regularity conditions are needed in the proof of this property.

**A1**. *The process $\{(Y_t, \mathbf{X}_t)\}$ is stationary and $\alpha$-mixing with the mixing coefficient decaying at a geometric rate: $\alpha(k) \sim c_\alpha \theta_0^k$, where $0 < c_\alpha < \infty$ and $0 < \theta_0 < 1$.*

**A2**. *Let $f_j(\cdot)$ be the marginal density function of $X_{tj}$, the $j$-th element of $\mathbf{X}_t$. Assume that $f_j(\cdot)$ has continuous derivatives up to the second order and $\inf_{x_j \in \mathcal{C}_j} f_j(x_j) > 0$, where $\mathcal{C}_j$ is the compact support of $X_{tj}$. For each $j$, the conditional density functions of $Y_t$ for given $X_{tj}$ exists and satisfies the Lipschitz continuous condition. Furthermore, the length of $\mathcal{C}_j$ is uniformly bounded by a positive constant.*

**A3**. *The kernel function $K(\cdot)$ is a Lipschitz continuous and bounded probability density function with a compact support. Let the bandwidth satisfy $h_1 \sim n^{-\theta_1}$ with $1/6 < \theta_1 < 1$.*

**A4**. *The marginal regression function $m_j(\cdot)$ has continuous derivatives up to the second order and there exists a positive constant $c_m$ such that $\sup_j \sup_{x_j \in \mathcal{C}_j} \big[|m_j(x_j)| + |m_j'(x_j)| + |m_j''(x_j)|\big] \leq c_m$.*

**A5**. *The response variable $Y_t$ satisfies $\mathsf{E}[\exp\{s|Y_t|\}] < \infty$ where $s$ is a positive constant.*

**Remark 1**. The condition **A1** imposes the stationary $\alpha$-mixing dependence structure on the observations, which is not uncommon in the time series literature (c.f., Bosq, 1998; Fan and Yao, 2003). It might be possible to consider a more general dependence structure such as the near epoch dependence studied in Lu and Linton (2007) and Li, Lu and Linton (2012), however, the technical proofs would be more involved. Hence, we impose the mixing dependence structure and focus on the ideas proposed. The restriction of geometric decaying rate on the mixing coefficient is due to the ultra-high dimensional setting and it may be relaxed if the dimension of the covariates diverges at a polynomial rate. The conditions **A2** and **A4** give some smoothness restrictions on the marginal density functions and marginal regression functions. To simplify the discussion, we assume that all of the marginal density functions have compact support. Such an assumption might be too restrictive for time series data, but it could be relaxed by slightly modifying our methodology. For example, if the marginal density function of $X_{tj}$ is the standard normal density which does not have a compact support, we can truncate the tail of $X_{tj}$ in the KSIS procedure by replacing $X_{tj}$ with $X_{tj}I\big(|X_{tj}| \leq \zeta_n\big)$ and $\zeta_n$ divergent to infinity at a slow rate. The condition **A3** is a commonly-used condition on the kernel function as well as the bandwidth. The strong moment condition on $Y_t$ in **A5** is also quite common in the SIS literature such as Fan, Feng and Song (2011) and Liu, Li and Wu (2014).

Define the index set of "true" candidate models as

$$\mathcal{S} = \big\{j = 1, 2, \ldots, p_n + d_n : \; \mathsf{v}(j) \neq 0\big\}.$$

The following theorem gives the sure screening property for the KSIS procedure.

**Theorem 1**. *Suppose that the conditions A1–A5 are satisfied.*

*(i) For any small $\delta_1 > 0$, there exists a positive constant $\delta_2$ such that*

$$\mathsf{P}\left(\max_{1\leq j\leq p_n+d_n}\left|\hat{\mathsf{v}}(j) - \mathsf{v}(j)\right| > \delta_1 n^{-2(1-\theta_1)/5}\right) = O\left(M(n)\exp\left\{-\delta_2 n^{(1-\theta_1)/5}\right\}\right), \quad (3.1)$$

*where $M(n) = (p_n + d_n)n^{(17+18\theta_1)/10}$ and $\theta_1$ is defined in the condition A3.*

*(ii) If we choose the pre-determined tuning parameter $\rho_n = \delta_1 n^{-2(1-\theta_1)/5}$ and assume*

$$\min_{j\in\mathcal{S}} \mathsf{v}(j) \geq 2\delta_1 n^{-2(1-\theta_1)/5}, \quad (3.2)$$

*then we have*

$$\mathsf{P}\big(\mathcal{S} \subset \hat{\mathcal{S}}\big) \geq 1 - O\left(M_{\mathcal{S}}(n)\exp\left\{-\delta_2 n^{(1-\theta_1)/5}\right\}\right), \quad (3.3)$$

*where $M_{\mathcal{S}}(n) = |\mathcal{S}|n^{(17+18\theta_1)/10}$ with $|\mathcal{S}|$ being the cardinality of $\mathcal{S}$.*

**Remark 2**. The above theorem shows that the covariates whose marginal regressions have not too small positive correlations with the response variable would be included in the screened model with probability approaching one at a possible exponential rate of $n$. The condition (3.2) guarantees that the correlations between the response and the marginal regression functions for covariates whose indices belong to $\mathcal{S}$ are bounded away from zero, but the lower bound may converge to zero. As $p_n+d_n = O(\exp\{n^{\delta_0}\})$, in order to ensure the validity of Theorem 1(i), we need to impose the restriction $\delta_0 < (1-\theta_1)/5$, which reduces to $\delta_0 < 4/25$ if the order of the optimal bandwidth in kernel smoothing (i.e., $\theta_1 = 1/5$) is used. Our theorem generalises the results in Fan, Feng and Song (2011) and Liu, Li, Wu (2014) to dynamic time series case and those in Ando and Li (2014) to the flexible nonparametric setting.

We next study the asymptotic properties for the PMAMAR method including the well-known the sparsity and oracle properties. As in Sections 1 and 2, we recall that $q_n = |\hat{\mathcal{S}}|$ and the dimension of the potential covariates is reduced from $p_n + d_n$ to $q_n$ after implementing the KSIS procedure. As above, we let $\mathbf{X}_t^*$ be the KSIS-chosen covariates, which may include both the exogenous regressors and lags of $Y_t$. Define

$$a_n = \max_{1\leq j\leq q_n} \left\{|p'_\lambda(|w_{oj}|)|,\ |w_{oj}| \neq 0\right\}$$

and

$$b_n = \max_{1\leq j\leq q_n} \left\{|p''_\lambda(|w_{oj}|)|,\ |w_{oj}| \neq 0\right\}.$$

We need to introduce some additional conditions to derive the asymptotic theory.

**A6**. *The matrix*

$$\mathbf{\Lambda}_n := \begin{pmatrix} \mathsf{E}\left[m_1^*(X_{t1}^*)m_1^*(X_{t1}^*)\right] & \dots & \mathsf{E}\left[m_1^*(X_{t1})m_{q_n}^*(X_{tq_n}^*)\right] \\ \vdots & \vdots & \vdots \\ \mathsf{E}\left[m_{q_n}^*(X_{tq_n}^*)m_1^*(X_{t1})\right] & \dots & \mathsf{E}\left[m_{q_n}(X_{tq_n}^*)m_{q_n}(X_{tq_n}^*)\right] \end{pmatrix}$$

*is positive definite with the largest eigenvalue bounded. The smallest eigenvalue of $\mathbf{\Lambda}_n$, $\chi_n$, is positive and satisfies $q_n = o(\sqrt{n}\chi_n)$.*

**A7**. *The bandwidth $h_2$ satisfies*

$$nh_2^4 \to 0, \quad n^{\frac{1}{2}-\xi}h_2 \to \infty, \quad q_n^2(\tau_n + h_2^2) = o(\chi_n) \tag{3.4}$$

*as $n \to \infty$, where $\xi$ is positive but arbitrarily small, and $\tau_n = \left(\frac{\log n}{nh_2}\right)^{1/2}$.*

**A8**. *Let $a_n = O(n^{-1/2}\chi_n^{-1})$, $b_n = o(\chi_n)$, $p_\lambda(0) = 0$, and there exit two positive constants $C_1$ and $C_2$ such that $\left|p''_\lambda(\vartheta_1) - p''_\lambda(\vartheta_2)\right| \le C_2|\vartheta_1 - \vartheta_2|$ when $\vartheta_1, \vartheta_2 > C_1\lambda$.*

**Remark 3**. The condition **A6** gives some regularity conditions on the eigenvalues of the $q_n \times q_n$ positive definite matrix $\mathbf{\Lambda}_n$. Note that we allow that some eigenvalues tend to zero at certain rates. In contrast, most of the existing literature dealing with independent observations assumes that the smallest eigenvalue of $\mathbf{\Lambda}_n$ is bounded away from zero, which may be violated for time series data. The restrictions in the condition **A7** imply that undersmoothing is needed in our semiparametric procedure and $q_n$ can only be divergent at a polynomial rate of $n$. The condition **A8** is a commonly-used condition on the penalty function $p_\lambda(\cdot)$, and would be similar to that in Fan and Peng (2004) if we let $\chi_n > \chi$ with $\chi$ being a positive constant.

Without loss of generality, define the vector of the optimal weights

$$\mathbf{w}_o = (w_{o1}, \ldots, w_{oq_n})^{\mathsf{T}} = \left[\mathbf{w}_o^{\mathsf{T}}(1),\ \mathbf{w}_o^{\mathsf{T}}(2)\right]^{\mathsf{T}},$$

where $\mathbf{w}_o(1)$ is composed of non-zero weights with dimension $s_n$ and $\mathbf{w}_o(2)$ is composed of zero weights with dimension $(q_n - s_n)$. In order to give the asymptotic normality for $\hat{\mathbf{w}}_n(1)$, the estimator of $\mathbf{w}_o(1)$, we need to introduce some further notation. Define

$$\eta_t^* = Y_t - \sum_{j=1}^{q_n} w_{oj} m_j^*(X_{tj}^*), \quad \eta_{tj}^* = Y_t - m_j^*(X_{tj}^*)$$

and $\boldsymbol{\xi}_t = \left(\xi_{t1}, \ldots, \xi_{ts_n}\right)^{\mathsf{T}}$ with $\xi_{tj} = \overline{\eta}_{tj}^* - \widetilde{\eta}_{tj}^*$, $\overline{\eta}_{tj}^* = m_j^*(X_{tj}^*)\eta_t^*$,

$$\widetilde{\eta}_{tj}^* = \sum_{k=1}^{q_n} w_{ok}\eta_{tk}^*\beta_{jk}(X_{tk}^*) = \sum_{k=1}^{s_n} w_{ok}\eta_{tk}^*\beta_{jk}(X_{tk}^*), \quad \beta_{jk}(x_k) = \mathsf{E}\left[m_j^*(X_{tj}^*)|X_{tk}^* = x_k\right].$$

Throughout the paper, we assume that the mean of $\boldsymbol{\xi}_t$ is zero, and define $\mathbf{\Sigma}_n = \sum_{t=-\infty}^{\infty} \mathsf{E}\left(\boldsymbol{\xi}_0\boldsymbol{\xi}_t^{\mathsf{T}}\right)$ and $\mathbf{\Lambda}_{n1}$ as the top-left $s_n \times s_n$ submatrix of $\mathbf{\Lambda}_n$. Let

$$\boldsymbol{\omega}_n = \left[p'_\lambda(|w_{o1}|)\mathsf{sgn}(w_{o1}), \ldots, p'_\lambda(|w_{os_n}|)\mathsf{sgn}(w_{os_n})\right]^{\mathsf{T}}$$

and

$$\mathbf{\Omega}_n = \mathsf{diag}\left\{p''_\lambda(|w_{o1}|), \ldots, p''_\lambda(|w_{os_n}|)\right\},$$

where $\mathsf{sgn}(\cdot)$ is the sign function. In the following theorem, we give the asymptotic theory of $\hat{\mathbf{w}}_n$ obtained by the PMAMAR method.

**Theorem 2**. *Suppose that the conditions A1–A8 are satisfied.*

*(i) There exists a local minimizer $\hat{\mathbf{w}}_n$ of the objective function $\mathcal{Q}_n(\cdot)$ defined in (2.9) such that*

$$\|\hat{\mathbf{w}}_n - \mathbf{w}_o\| = O_P\Big(\sqrt{q_n}(n^{-1/2}\chi_n^{-1} + a_n)\Big), \tag{3.5}$$

*where $\chi_n$ and $a_n$ are defined in the conditions A6 and A8, respectively, and $\|\cdot\|$ denotes the Euclidean norm.*

*(ii) Let $\hat{\mathbf{w}}_n(2)$ be the estimator of $\mathbf{w}_o(2)$ and further assume that*

$$\lambda \to 0, \quad \frac{\chi_n\sqrt{n}\lambda}{\sqrt{q_n}} \to \infty, \quad \liminf_{n\to\infty}\liminf_{\vartheta\to 0+}\frac{p'_\lambda(\vartheta)}{\lambda} > 0. \tag{3.6}$$

*Then, the local minimizer $\hat{\mathbf{w}}_n$ of the objective function $\mathcal{Q}_n(\cdot)$ satisfies $\hat{\mathbf{w}}_n(2) = \mathbf{0}$ with probability approaching one.*

*(iii) If we further assume that the eigenvalues of $\mathbf{\Lambda}_{n1}$ are bounded away from zero and infinity,*

$$\sqrt{n}\mathbf{A}_n\mathbf{\Sigma}_n^{-1/2}\big(\mathbf{\Lambda}_{n1} + \mathbf{\Omega}_n\big)\Big[\hat{\mathbf{w}}_n(1) - \mathbf{w}_o(1) - \big(\mathbf{\Lambda}_{n1} + \mathbf{\Omega}_n\big)^{-1}\boldsymbol{\omega}_n\Big] \xrightarrow{d} \mathsf{N}\big(\mathbf{0}, \mathbf{A}_0\big), \tag{3.7}$$

*where $\mathbf{0}$ is a null vector whose dimension may change from line to line, $\mathbf{A}_n$ is an $s \times s_n$ matrix such that $\mathbf{A}_n\mathbf{A}_n^{\intercal} \to \mathbf{A}_0$ and $\mathbf{A}_0$ is an $s \times s$ symmetric and non-negative definite matrix, $s$ is a fixed positive integer.*

**Remark 4**. Theorem 2(i) indicates that the convergence rate of the estimator $\hat{\mathbf{w}}_n$ is determined by the dimension of the covariates, by the matrix $\mathbf{\Lambda}_n$ and by the penalty function. The involvement of $\chi_n$ in the convergence rate makes Theorem 2(i) more general than the results obtained in the existing literature. If we assume that all the eigenvalues of the matrix $\mathbf{\Lambda}_n$ are bounded from zero and infinity with $\chi_n > \chi > 0$, the convergence rate would reduce to $O_P\big(\sqrt{q_n}(n^{-1/2}+a_n)\big)$, which is the same as that in Theorem 1 of Fan and Peng (2004). Furthermore, when $q_n$ is fixed and $a_n = O(n^{-1/2})$, we could derive the root-$n$ convergence rate for $\hat{\mathbf{w}}_n$ as in Theorem 3.1 of Li, Linton and Lu (2015). Theorem 2(ii) shows that the estimator of $\mathbf{w}_o(2)$ is equal to zero with probability approaching one, which indicates that the PMAMAR procedure possesses the well known sparsity property, and thus can be used as a model selector. Theorem 2(ii) and (iii) above shows that the proposed estimator of the optimal weights enjoy the oracle property, which takes $\mathbf{w}_o(2) = \mathbf{0}$ as a prerequisite. Furthermore, when $n$ is large enough and $\lambda$ tends to zero sufficiently fast for some penalty functions (such as the SCAD penalty), the asymptotic distribution in (3.7) would reduce to

$$\sqrt{n}\mathbf{A}_n\boldsymbol{\Sigma}_n^{-1/2}\boldsymbol{\Lambda}_{n1}\big[\hat{\mathbf{w}}_n(1)-\mathbf{w}_o(1)\big] \xrightarrow{d} \mathsf{N}\big(\mathbf{0},\mathbf{A}_0\big), \tag{3.8}$$

which is exactly the same as that in Theorem 3.3 of Li, Linton and Lu (2015).

## 3.2 *Asymptotic theory for PCA+PMAMAR*

In this subsection, we show that the estimated common factors consistently estimate the true common factors (with rotation), and the asymptotic order of the difference between $\hat{m}^*_{k,f}(z_k)$ defined in (2.15) and the infeasible kernel estimation $\tilde{m}^*_{k,f}(z_k)$ is $o_P(n^{-1/2})$ uniformly. The latter asymptotic result implies that the sparsity and oracle property for the PMAMAR approach developed in Theorem 2 still holds. We start with some regularity conditions that are used when proving the asymptotic results.

- **B1**. *The process $\{(Y_t, \mathbf{f}_t, \mathbf{U}_t)\}$ is stationary and $\alpha$-mixing with the mixing coefficient decaying at a geometric rate: $\alpha(k) \sim c_\alpha\theta_0^k$, where $c_\alpha$ and $0<\theta_0<1$ are defined as in the condition A1.*

- **B2**. *The random common factors satisfy the conditions that $\mathsf{E}\left[\mathbf{f}_t^0\right] = \mathbf{0}$, $\max_t \|\mathbf{f}_t^0\| = O_P(1)$, the $r\times r$ matrix $\boldsymbol{\Lambda}_F := \mathsf{E}\big[\mathbf{f}_t^0(\mathbf{f}_t^0)^\intercal\big]$ is positive definite and $\mathsf{E}\big[\|\mathbf{f}_t^0\|^{4+\tau}\big] < \infty$ for some $0<\tau<\infty$.*

**B3**. *The matrix* $(\mathbf{B}_n^0)^\intercal \mathbf{B}_n^0/p_n$ *is positive definite with the smallest eigenvalue bounded away from zero and* $\max_k \|\mathbf{b}_k^0\|$ *is bounded.*

**B4**. *The idiosyncratic error satisfies* $\mathsf{E}[u_{tk}] = 0$, $\mathsf{E}[u_{tk}\mathbf{f}_t^0] = \mathbf{0}$ *and* $\max_k \mathsf{E}\left[|u_{tk}|^8\right] < \infty$. *Furthermore, there exist two positive constants* $C_3$ *and* $C_4$ *such that*

$$\max_t \mathsf{E}\left[\left\|\sum_{k=1}^{p_n} u_{tk}\mathbf{b}_k^0\right\|^4\right] \le C_3 p_n^2 \tag{3.9}$$

*and*

$$\max_{t_1,t_2} \mathsf{E}\left[\left|\sum_{k=1}^{p_n} \{u_{t_1k}u_{t_2k} - \mathsf{E}[u_{t_1k}u_{t_2k}]\}\right|^4\right] \le C_4 p_n^2, \tag{3.10}$$

*and* $\max_k \mathsf{E}[\exp\{s\|u_{tk}\mathbf{f}_t^0\|\}] < \infty$ *where* $s$ *is a positive constant as in the condition A5.*

**B5**. **(i)** *The kernel function* $K(\cdot)$ *is positive and has continuous derivatives up to the second order with a compact support. In addition, the derivative functions of* $K(\cdot)$ *are bounded.*

**(ii)** *There exists* $0 < \gamma_0 < 1/6$ *such that* $n^{1-\gamma_0}h_3^3 \to \infty$. *In addition,* $n^3/(p_n^2 h_3^4) = o(1)$.

**(iii)** *The marginal regression functions (corresponding to the factor regressors)* $m_{k,f}^*(\cdot)$ *have continuous and bounded derivatives up to the second order.*

**Remark 5**. The above conditions have been commonly used in the literature. For example, the conditions **B2** and **B3** are similar to Assumptions A and B in Bai and Ng (2002), whereas the conditions **B1** and **B4** are similar to the corresponding conditions in Assumptions 3.2–3.4 in Fan, Liao and Mincheva (2013). In particular, the exponential bound $\max_k \mathsf{E}[\exp\{s\|u_{tk}\mathbf{f}_t^0\|\}] < \infty$ in the condition **B4** is crucial to ensure that $p_n$ can diverge at an exponential rate of $n$. The condition **B5** is mainly used for the proof of Theorem 3(ii) in Appendix B.

**Theorem 3**. *Suppose that the conditions B1–B4 are satisfied, and*

$$n = o(p_n^2), \quad p_n = O\left(\exp\{n^{\delta_*}\}\right), \quad 0 \leq \delta_* < 1/3. \tag{3.11}$$

*(i) For the PCA estimation $\hat{\mathbf{f}}_t$, we have*

$$\max_t \left\| \hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t^0 \right\| = O_P\left(n^{-1/2} + n^{1/4} p_n^{-1/2}\right), \tag{3.12}$$

*where $\mathbf{H}$ is defined in Section 2.2.*

*(ii) In addition, suppose that the conditions A5 and B5 are satisfied and the latent factor $\mathbf{f}_t^0$ has a compact support. Then we have*

$$\max_{1 \leq k \leq r} \sup_{z_k \in \mathcal{F}_k^*} \left| \hat{m}_{k,f}^*(z_k) - \tilde{m}_{k,f}^*(z_k) \right| = o_P\left(n^{-1/2}\right), \tag{3.13}$$

*where $\mathcal{F}_k^*$ is the compact support of $\tilde{f}_{tk}^0$.*

**Remark 6**. Theorem 3(i) gives the uniform consistency result for the estimation of the common factors, which is very similar to some existing results on PCA estimation of the high-dimensional factor models such as Theorem 3.3 in Fan, Liao and Mincheva (2013). If we further assume that $n^3 = o(p_n^2)$, which automatically holds when $p_n$ is divergent at an exponential rate of $n$, the uniform convergence rate in (3.12) would be $O_P\left(n^{-1/2}\right)$. Theorem 3(ii) shows that we may replace $\hat{m}_{k,f}^*(\cdot)$ by the infeasible kernel estimation $\tilde{m}_{k,f}^*(\cdot)$ when deriving the asymptotic theory for the PMAMAR method introduced in Section 3.2, and Theorem 2 in Section 3.1 still holds with some notational modifications (c.f., $q_n$ in (3.5) needs to be replaced by $d_n$). The restriction of compact support on $\mathbf{f}_t^0$ can be removed if we slightly modify the methodology as discussed in Remark 1.

# 4 Some extensions

This section will discuss some extensions by introducing an iterative KSIS+PMAMAR procedure when the covariates are highly correlated, and an extended PCA+PMAMAR approach with selection of the number of the latent factors in model (2.11).

## 4.1 *An iterative KSIS+PMAMAR procedure*

When the covariates are highly correlated with each other, difficulties in variable selection arise. As documented in Fan and Lv (2008), when the covariate dimension is large, even if the covariates are mutually independent, the data generated from them may exhibit significant spurious correlation. Fan and Lv (2008) noticed that when irrelevant covariates are highly correlated with some relevant ones, they might be selected into a model with higher priority than some other relevant covariates, which results in high false positive rates and low true positive rates. Such a problem may become even worse in the present situation of this paper due to the time series nature of the data, where both the response $Y_t$ and the covariates $\mathbf{X}_t$ are autocorrelated over time $t$. Since the covariates $X_{tj}$, $j = p_n+1, \ldots, p_n+d_n$, are generated from the lags of $Y_t$, both the temporal autocorrelation and the cross-sectional correlation among them arise. Hence, if we try to estimate or predict $Y_t$ with $p_n + d_n$ potential covariates by running firstly the KSIS and secondly the PMAMAR with those that have survived the screening process, the results could be very unsatisfactory. It is especially so when $p_n + d_n$ is much larger than the sample size $n$. Due to the autocorrelation in the response and the lagged covariates, the iterative procedure developed in Fan, Feng and Song (2011) can not apply in this context. This is because their iterative procedure involves a permutation step in which the observed data is randomly permuted to obtain a data-driven screening threshold for each iteration. When the data are autocorrelated, permutation would destroy the inherent serial dependence structure and hence may lead to erroneous thresholds obtained. To alleviate the problem, we propose an iterative version for the KSIS+PMAMAR procedure as follows.

**Step 1**: For each $j = 1, 2, \ldots, p_n + d_n$, estimate the marginal regression function $m_j(x_j)$ by the kernel method and denote the estimate as $\hat{m}_j(x_j)$. Then calculate the sample covariance between $Y_t$ and $\hat{m}_j(X_{tj})$:

$$\hat{\mathsf{v}}(j) = \frac{1}{n}\sum_{t=1}^{n} \hat{m}_j^2(X_{tj}) - \Big[\frac{1}{n}\sum_{t=1}^{n} \hat{m}_j(X_{tj})\Big]^2.$$

Select the variable with the largest $\hat{\mathsf{v}}(j)$ and let

$$\mathcal{S} = \Big\{ j : \ \hat{\mathsf{v}}(j) = \max_i(\hat{\mathsf{v}}(i)), 1 \leq i \leq p_n + d_n \Big\}.$$

**Step 2**: Run a linear regression of the response variable $Y$ on the estimated marginal regression functions of the selected variables in $\mathcal{S}$, and obtain the residuals $\widehat{e}^{\mathcal{S}}$.

**Step 3**: Run a linear regression the estimated marginal regression function of each variable in $\mathcal{S}^c$, which is defined as $\{1, 2, \ldots, p_n + d_n\} \backslash \mathcal{S}$, on the estimated marginal regression functions of the selected variables in $\mathcal{S}$, and obtain the residuals $\widehat{e}^{i\mathcal{S}}$ for each $i \in \mathcal{S}^c$.

**Step 4**: Compute the kernel estimate of the marginal regression function, $\widehat{m}_i^e$, of the residuals $\widehat{e}^{\mathcal{S}}$ from Step 2 on the residuals $\widehat{e}^{i\mathcal{S}}$ from Step 3 for each $i \in \mathcal{S}^c$, and calculate the sample covariance $\widehat{\mathsf{v}}^e(i)$ between $\widehat{e}^{\mathcal{S}}$ and $\widehat{m}_i^e$. Add the variable $j$ with the largest $\widehat{\mathsf{v}}^e(i)$ among all $i \in \mathcal{S}^c$ to the set $\mathcal{S}$.

**Step 5**: Run a PMAMAR regression with the SCAD penalty of $Y$ against $X_j$, $j \in \mathcal{S}$, as in (2.6), and discard any variables from $\mathcal{S}$ if their corresponding estimated weights are zero.

**Step 6**: Repeat Steps 2–5 until no new variable is recruited or until the number of variables selected in $\mathcal{S}$ hits $[n/\log(n)]$.

In Step 4, we treat the residuals, from the linear regression of the response variable on the marginal regression functions of the variables currently selected, as the new response variable, and the residuals, from linear regression of the marginal regression functions of the unselected variables on those of the selected variables, as the new covariates. We then carry out a nonparametric screening and select the variable with the largest resulting sample covariance $\widehat{\mathsf{v}}^e(i)$ as the candidate to be added to $\mathcal{S}$. The use of the residuals, instead of the original $Y$ and unselected $\widehat{m}_j$'s, reduces the priority of the remaining irrelevant variables, which are highly correlated with some selected relevant variables, being picked, and increases the priority of the remaining relevant variables, which are marginally insignificant but jointly significant, being picked. Hence, this iterative procedure may help reduce false positive rates and increase true positive rates. The variables in the selected set $\mathcal{S}$ then undergo the PMAMAR regression with the SCAD penalty. The set $\mathcal{S}$ is updated by discarding any variables having insignificant weights. Other penalty functions such as the LASSO and the MCP can equally apply in Step 5. The above iterative procedure can be seen as a greedy selection algorithm, since at most one variable is selected in each iteration. It starts with zero variable and keeps adding or deleting variables until none of the remaining variables are considered significant in the sense of significance of the weights in PMAMAR.

## 4.2 *The PCA+KSIS+PMAMAR procedure*

In reality, the number of common factors, $r$, in the approximate factor model (2.11) is usually unknown. We hence need to select it from an eigenanalysis of the matrix $\mathcal{Z}_n\mathcal{Z}_n^{\intercal}/(np_n)$. Two ways are possible to address this issue. The first is to set a maximum number, say $r_{\max}$ (not too large usually), for the factors. Since the factors extracted from the eigenanalysis are orthogonal to each other, the over-extracted insignificant factors will be discarded in the PMAMAR step. Another approach is to select the first few eigenvectors (corresponding to the first few largest eigenvalues) of $\mathcal{Z}_n\mathcal{Z}_n^{\intercal}/(np_n)$ so that a pre-determined amount, say 95%, of the total variation is accounted for. See Boneva, Linton and Vogt (2015) for more information on the selection of the number of common component functions. Other selection criteria such as BIC can be found in Bai and Ng (2002) and Fan, Liao and Mincheva (2013).

In the second step of the PCA+PMAMAR procedure proposed in Section 2.2, the estimated factors and the $d_n$ candidate lags of $Y$ undergo a PMAMAR regression. However, since the lags of $Y$ are often highly correlated, $d_n$ is hence large and the P-MAMAR regression usually cannot produce satisfactory results in selecting the truly significant lags. This may lead to poor performance of the PCA+PMAMAR procedure predicting the future values of $Y$. In order to alleviate this problem, a KSIS step can be added in between the PCA and PMAMAR steps so that the candidate lags of $Y$ first undergo a KSIS to preliminarily screen out some insignificant lags. The simulation results in Example 5.2 below confirm that this PCA+KSIS+PMAMAR procedure improves the prediction performance of the PCA+PMAMAR procedure.

# 5 Numerical studies

In this section, we report simulation studies (Examples 5.1 and 5.2) and an empirical application (Example 5.3). Throughout this section the rule of thumb bandwidth selection is used as our methods seem not sensitive to the choice of bandwidth.

## 5.1 Simulation studies

**Example 5.1**. In this example, the sample size is set to be $n = 100$, and the numbers of candidate exogenous covariates and lagged terms are $(p_n, d_n) = (30, 10)$ and $(p_n, d_n) = (150, 50)$. The data-generating model is defined by

$$Y_t = m_1(Z_{t1}) + m_2(Z_{t2}) + m_3(Z_{t3}) + m_4(Z_{t4}) + m_5(Y_{t-1}) + m_6(Y_{t-2}) + m_7(Y_{t-3}) + \varepsilon_t, \tag{5.1}$$

for $t \geq 1$, where, following Meier, van de Geer and Bühlmann (2009), we set

$$m_i(x) = \sin(0.5\pi x), \qquad i = 1, 2, \dots, 7, \tag{5.2}$$

the exogenous covariates $\mathbf{Z}_t = (Z_{t1}, Z_{t2}, \ldots, Z_{tp_n})^\intercal$ are independently drawn from $p_n$-dimensional Gaussian distribution with zero mean and covariance matrix $\mathrm{cov}(\mathbf{Z}) = I_{p_n}$ or $C_{\mathbf{Z}}$, whose main-diagonal entries are 1 and off-diagonal entries are 1/2. The error term $\varepsilon_t$ are independently generated from the $\mathsf{N}(0, 0.7^2)$ distribution. The real size of exogenous regressors is 4 and the real lag length is 3. We generate $100+n$ observations from the process (5.1) with initial states $Y_{-2} = Y_{-1} = Y_0 = 0$ and discard the first $100 - d_n$ observations.

The aim of this simulation is to compare the performance of the iterative KSIS+PMAMAR (IKSIS+PMAMAR) procedure in Section 4.1 with the (non-iterative) KSIS+PMAMAR procedure in Section 2.1. In order to further the comparison, we also employ the iterative sure independence screening (ISIS) method proposed in Fan and Lv (2008), the penalised least squares method for high-dimensional generalised additive models (penGAM) proposed in Meier, van de Geer and Bühlmann (2009), and the oracle semiparametric model averaging method (Oracle, in which the true relevant variables are known). For the KSIS+PMAMAR, we choose $[n/\log(n)]$ variables from the screening step, which then undergo a PMAMAR with the SCAD penalty. The measures of performance considered are the true positive (TP) and false positive (FP), defined, respectively, as the numbers of true and false relevant variables selected, the mean squared estimation error (MSEE) defined as MSEE $= \frac{1}{n}\sum_{t=1}^{n}(Y_t - \widehat{Y}_t)^2$, where $\widehat{Y}_t$ is the fitted value of $Y$ at $t$ obtained from a particular method. We also generate a prediction test set of size $n^* = n/10 = 10$ and calculate the $s$-step-ahead $(1 \leq s \leq n^*)$ forecasts for the response $Y$, from which the mean squared prediction error (MSPE) is obtained. The MSPE is defined as MSPE $= \frac{1}{n^*}\sum_{s=1}^{n^*}(Y_{n+s} - \widehat{Y}_{n+s})^2$, where $\widehat{Y}_{n+s}$ is the $s$-step-ahead forecast of $Y$. The smoothing parameters in the penalised regressions are chosen by the cross-validation. The SCAD penalised regression is implemented using the R package "ncvreg", the ISIS method implemented using the "SIS" R package and the penGAM method implemented using the "penGAM" package. The results in Table 5.1 are based on 200 simulation replications.

It can be seen from Table 5.1 that the iterative version of KSIS+PMAMAR generally increases the TP of the non-iterative version while at the same time decreases the FP. This results in a better performance of the IKSIS+PMAMAR in both estimation and prediction than the KSIS+PMAMAR. Among the 4 variable selection procedures (i.e., IKSIS+PMAMAR, KSIS+PMAMAR, penGAM, and ISIS), the penGAM has the smallest FP. In fact, it is the most conservative in variable selection and on average selects the least number of variables. This makes it the approach that has the highest MSEE, since within the same linear or nonlinear modelling framework it is generally the case that the more variables are selected the smaller the MSEE is. The ISIS, in contrast to the other approaches, assumes a linear modelling structure and hence is not able to correctly recognise the truly relevant and falsely relevant variables when the underlying data generating process is nonlinear, leading to low TP and high FP. This poor performance of the ISIS in variable selection also results in its poor predictive power. The predictive performance of an approach largely depends on its accuracy in variable selection, and a low TP and high FP will lead to a high MSPE. The results for the Oracle serve as a benchmark for those of the other approaches. The MSPEs from the IKSIS+PMAMAR and KSIS+PMAMAR are the closest among all the approaches to that of the Oracle. It can also be observed, by a comparison of the first two panels of Table 5.1 with the last two, that when the correlation among the exogenous variables increases, the performance of all approaches worsens.

**Example 5.2**. The exogenous variables $\mathbf{Z}_t$ in this example are generated through an approximate factor model:

$$\mathbf{Z}_t = \mathbf{B}\mathbf{f}_t + \mathbf{z}_t,$$

where the rows of the $p_n \times r$ loadings matrix $\mathbf{B}$ and the common factors $\mathbf{f}_t$, $t = 1, \cdots, n$, are independently generated from the multivariate $\mathsf{N}(\mathbf{0}, I_r)$ distribution, and the $p_n$-dimensional error terms $\mathbf{z}_t$, $t = 1, \cdots, n$, from the $0.1\mathsf{N}(\mathbf{0}, I_{p_n})$ distribution. We set

Table 5.1: Average results on variable selection and accuracy of estimation and prediction in Example 5.1 over 200 replications

| Model | Method | TP | FP | MSEE | MSPE |
|---|---|---|---|---|---|
| Example 5.1<br>$\mathrm{cov}(\mathbf{Z}) = I_{p_n}$<br>$(p_n, d_n) = (30, 10)$ | IKSIS+PMAMAR | 6.970(0.2437) | 6.815(5.3417) | 0.3487(0.0960) | 1.2760(0.7326) |
| | KSIS+PMAMAR | 6.940(0.2771) | 8.020(3.9735) | 0.3516(0.0659) | 1.3186(0.7777) |
| | penGAM | 6.040(1.0067) | 0.285(0.5340) | 1.7083(0.2783) | 2.2329(1.1247) |
| | ISIS | 5.380(0.8055) | 7.620(0.8055) | 1.7089(0.2729) | 2.7024(1.5347) |
| | Oracle | 7.000(0.0000) | 0.000(0.0000) | 0.4840(0.0789) | 0.9848(0.5942) |
| Example 5.1<br>$\mathrm{cov}(\mathbf{Z}) = I_{p_n}$<br>$(p_n, d_n) = (150, 50)$ | IKSIS+PMAMAR | 6.785(0.6170) | 9.510(5.5438) | 0.2419(0.1033) | 1.6758(0.9705) |
| | KSIS+PMAMAR | 6.290(0.8242) | 11.075(3.5768) | 0.3556(0.0811) | 1.7893(1.0595) |
| | penGAM | 5.995(1.0680) | 1.815(1.6414) | 1.6923(0.2855) | 2.3322(1.1407) |
| | ISIS | 4.435(1.0494) | 16.565 (1.0494) | 1.0371(0.2036) | 3.1249(1.6246) |
| | Oracle | 7.000(0.0000) | 0.000(0.0000) | 0.4780(0.0718) | 1.0300(0.5795) |
| Example 5.1<br>$\mathrm{cov}(\mathbf{Z}) = C_{\mathbf{Z}}$<br>$(p_n, d_n) = (30, 10)$ | IKSIS+PMAMAR | 5.845(1.3075) | 2.34(3.0382) | 0.7888(0.2352) | 1.8205(0.9793) |
| | KSIS+PMAMAR | 4.395(1.1293) | 2.715(3.3361) | 1.2163(0.3427) | 2.1788(1.1767) |
| | penGAM | 3.260(1.0186) | 0.085(0.2796) | 2.8712(0.3216) | 3.2532(1.3589) |
| | ISIS | 3.890(1.0788) | 8.790(1.6912) | 2.3324(0.5088) | 4.6481(3.1292) |
| | Oracle | 7.000(0.0000) | 0.000(0.0000) | 0.7867(0.0959) | 1.5681(0.9315) |
| Example 5.1<br>$\mathrm{cov}(\mathbf{Z}) = C_{\mathbf{Z}}$<br>$(p_n, d_n) = (150, 50)$ | IKSIS+PMAMAR | 4.615(1.5259) | 3.335(4.1773) | 0.8342(0.3272) | 2.3521(1.1848) |
| | KSIS+PMAMAR | 3.265(0.7600) | 2.980(2.7709) | 1.4383(0.2735) | 2.6098(1.7253) |
| | penGAM | 3.150(0.9655) | 0.585(0.8223) | 2.7857(0.3037) | 3.3010(1.5413) |
| | ISIS | 2.675(1.1515) | 18.3 (1.1342) | 1.3640(0.3241) | 8.6358(6.4155) |
| | Oracle | 7.000(0.0000) | 0.000(0.0000) | 0.7886(0.0976) | 1.6337(0.9636) |

$p_n = 30$ or $150$, $r = 3$, and generate the response variable via

$$Y_t = m_1(f_{t1}) + m_2(f_{t2}) + m_3(f_{t3}) + m_4(Y_{t-1}) + m_5(Y_{t-2}) + m_6(Y_{t-3}) + \varepsilon_t, \qquad (5.3)$$

where $f_{ti}$ is the $i$-th component of $\mathbf{f}_t$, $m_i(\cdot)$, $i = 1, \cdots, 6$, are the same as in (5.2), and $\varepsilon_t$, $t = 1, \cdots, n$, are independently drawn from the $\mathsf{N}(0, 0.7^2)$ distribution. In this example, we choose the number of candidate lags of $Y$ as $d_n = 10$. We compare the performance, in terms of estimation error and prediction error, of the following methods: PCA+PMAMAR, PCA+KSIS+PMAMAR, KSIS+PMAMAR, penGAM, ISIS, and Oracle. Since in reality both $r$ and the factors $\mathbf{f}_t$ are unobservable, in the first two methods, the factors are estimated by the first $\widehat{r}$ eigenvectors of $\mathcal{Z}_n\mathcal{Z}_n^\intercal/(np_n)$, where $\mathcal{Z}_n = (\mathbf{Z}_1, \cdots, \mathbf{Z}_n)^\top$, and $r$ is estimated by $\widehat{r}$, where $\widehat{r}$ is chosen so that 95% of the variation in $\mathcal{Z}_n$ is accounted for. In the PCA+PMAMAR method, the estimated factors and $d_n$ lags of $Y$ directly undergo a PMAMAR with the SCAD penalty, while in PCA+KSIS+PMAMAR the lags of $Y$ first undergo the KSIS and then the selected lags together with the estimated factors undergo a PMAMAR. The KSIS+PMAMAR, penGAM and ISIS deal directly with $p_n$ exogenous variables in $\mathbf{Z}_t$ and $d_n$ lags of $Y$ as in Example 5.1, and the Oracle uses the first 3 factors and the first 3 lags, as is the true case in the data generating process.

As in Example 5.1, the sample size is set as $n = 100$ and the experiment is repeated for 200 times. The results are summarised in Table 5.2. It can be seen from these results that when the number of exogenous variables $p_n$ is not so large compared with the sample size $n$ (i.e., 30 compared to 100), the KSIS+PMAMAR outperforms all the other approaches (except the Oracle), including the two PCA based approaches, in terms of estimation and prediction accuracy. However, when $p_n$ becomes larger than $n$, the PCA based approaches show their advantage in effective dimension reduction of the exogenous variables, which results in their lower MSEE and MSPE. The PCA+PMAMAR has a lower MSEE but higher MSPE than the PCA+KSIS+PMAMAR. This is due to the fact that without the screening step the

Table 5.2: Accuracy of estimation and prediction in Example 5.2 over 200 replications

| Model | Method | MSEE | MSPE |
|---|---|---|---|
| Example 5.2<br>$(p_n, d_n) = (30, 10)$ | PCA+PMAMAR | 0.7498(0.1313) | 2.2641(1.1040) |
| | PCA+KSIS+PMAMAR | 0.8846(0.1414) | 2.1239(1.0183) |
| | KSIS+PMAMAR | 0.5816(0.1116) | 2.1106(1.0122) |
| | penGAM | 1.9028(0.2561) | 2.6342(1.2488) |
| | ISIS | 2.1372(0.3876) | 11.6244(18.9164) |
| | Oracle | 0.9926(0.1551) | 1.9821(0.9775) |
| Example 5.2<br>$(p_n, d_n) = (150, 10)$ | PCA+PMAMAR | 0.7207(0.1240) | 2.1505(1.0793) |
| | PCA+KSIS+PMAMAR | 0.8469(0.1469) | 1.9355(0.9954) |
| | KSIS+PMAMAR | 0.9985(0.2731) | 2.8453(1.6823) |
| | penGAM | 1.8461(0.2526) | 2.6132(1.2584) |
| | ISIS | 1.8177(0.6077) | 43.4549(69.3956) |
| | Oracle | 0.9421(0.1626) | 1.7782(0.9229) |

PCA+PMAMAR selects more false lags of $Y$, and the higher FP leads to an higher MSPE and lower MSEE under the same PMAMAR framework. The above suggests that if one's main concern is to predict future values, there may be benefits in having the KSIS step to screen out insignificant lags between the PCA and PMAMAR steps.

## 5.2 An empirical application

**Example 5.3**. We next apply the proposed semiparametric model averaging methods to forecast inflation in the UK. The data were collected from the Office for National Statistics (ONS) and the Bank of England (BoE) websites and included quarterly observations on CPI and some other economics variables over the period Q1 1997 to Q4 2013. All the variables are seasonally adjusted. We use 53 series measuring aggregate real activity and other economic indicators to forecast CPI. Given the possible temporal persistence of CPI, we also add its 4 lags as predictors. Data from Q1 1997 to Q4 2012 are used as the training set and those between Q1 2013 and Q4 2013 are used for forecasting. As in Stock and Watson (1998, 1999), we make 4 types of transformations on different variables, depending on their nature: (i) logarithm, (ii) first difference of logarithms; (iii) first difference, and (iv) no transformation. Logarithms are usually taken on positive series that are not in rates or percentages, and first differences are taken of quantity series and of price indices. All series are standardised to have mean zero and unity variance after these transformations. Figure 5.1 plots both the original and transformed CPI series.

We use the training set to select or screen out the significant variables among the 53 exogenous economic variables and the 4 lags of CPI as well as to estimate the model averaging weights or model coefficients. These selected variables and estimated coefficients are then used to obtain the mean squared estimation error (MSEE) and form forecasts of CPI in the four quarters of 2013. We compare the forecasting capacity of the IKSIS+PMAMAR, KSIS+PMAMAR, PCA+PMAMAR, penGAM and ISIS methods via the mean squared prediction error (MSPE) and the mean absolute prediction error (MAPE), which are defined, respectively, as

$$\text{MSPE} = \frac{1}{4}\sum_{s=1}^{4}(Y_{n+s} - \widehat{Y}_{n+s})^2, \quad \text{MAPE} = \frac{1}{4}\sum_{s=1}^{4}\left|Y_{n+s} - \widehat{Y}_{n+s}\right|$$

where $\widehat{Y}_{n+s}$ is the $s$-step-ahead forecast of $Y$ from a particular method. In addition, we also compare the estimation accuracy of the methods via the mean squared estimation error (MSEE) and the mean absolute estimation error (MAEE), defined by

$$\text{MSEE} = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \widehat{Y}_t)^2, \quad \text{MAEE} = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \widehat{Y}_t\right|,$$

where $\widehat{Y}_t$ is the fitted value of $Y$ at time $t$.

Due to the small number of candidate lags of the response ($d = 4$), there is not much necessity to use the PCA+KSIS+PMAMAR approach in this example, and hence it is not included in the comparison. Similarly to Stock and Watson (2002), in the PCA+PMAMAR approach, common factors extracted from the exogenous variables together with lags of the response are used to forecast the response. The difference with Stock and Watson (2002)'s approach is that the PCA+PMAMAR allows these factors and lags to contribute to forecasting the response in a possibly nonlinear way. We also calculate forecasts based on the Phillips curve specification

$$I_{t+1} - I_t = \alpha + \beta(L)U_t + \gamma(L)\Delta I_t + \varepsilon_{t+1},$$

where $I_t$ is the CPI in the $t$-th quarter, $U_t$ is the unemployment rate, $\beta(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3$ and $\gamma(L) = \gamma_0 + \gamma_1 L + \gamma_2 L^2 + \gamma_3 L^3$ are lag polynomials with $L$ being the lag operator, and $\Delta$ is the first difference operator. We further employ some of the most commonly-used models from the BoE's suite of statistical forecasting models to model and forecast the CPI data. These include the autoregressive (AR) model, the vector autoregressive (VAR) model consisting of output, CPI, oil price, effective sterling exchange rate and BoE's base interest rate, and the smooth transition autoregressive (STAR) model. The order of autoregression in these models is selected by AIC, and the number of regimes in the STAR model is selected based on an LM test.

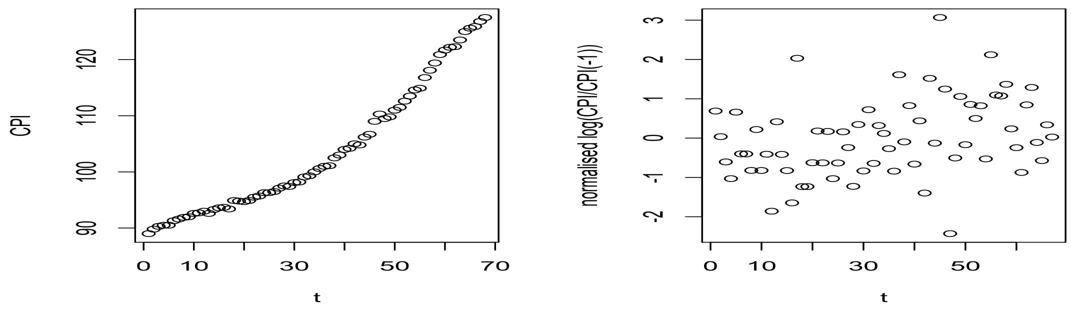

Figure 5.1: The plot for the UK CPI series. Left panel: the original UK CPI values from Q1 1997 to Q4 2013; and right panel: the normalised $\Delta \log(\text{CPI})$.

Table 5.3: Estimation and forecasting performance of various methods for the UK inflation data

| Method | MSEE | MSPE | MAEE | MAPE |
|---|---|---|---|---|
| IKSIS+PMAMAR | 0.1058 | 0.0360 | 0.2667 | 0.1815 |
| KSIS+PMAMAR | 0.3165 | 0.1130 | 0.4654 | 0.2802 |
| PCA+PMAMAR | 0.1303 | 0.0923 | 0.2977 | 0.2214 |
| penGAM | 0.6584 | 0.0865 | 0.6580 | 0.2855 |
| ISIS | 0.2714 | 0.1037 | 0.4317 | 0.3019 |
| Phillips Curve | 1.0225 | 1.1900 | 0.7655 | 1.0170 |
| AR | 1.0786 | 0.1138 | 0.8373 | 0.2621 |
| VAR | 1.0457 | 0.1027 | 0.8287 | 0.2456 |
| STAR | 1.0954 | 0.1558 | 0.8361 | 0.2962 |

The MSEEs, MSPEs, MAEEs and MAPEs of the above approaches are summarised in Table 5.3, which shows that the IKSIS+PMAMAR has the smallest MSPE followed by the penGAM and PCA+PMAMAR. The VAR, ISIS, KSIS+PMAMAR, and AR have comparablee MSPEs. However, the Phillips curve forecasts are much worse than those of the other methods. In terms of goodness of fit measured in either MSEE or MAEE, the AR, VAR, and STAR provide the worst fit. Among the variable selection/screening methods, the IKSIS+PMAMAR selects 12 exogenous variables and 3 lags of the response; the KSIS+PMAMAR selects 2 exogenous and 2 lags of response; the PCA+PMAMAR selects 17 common factors (which account for around 90% of the total variation) from the 53 exogenous variables and 2 lags of response; the penGAM selects 2 exogenous only; and the ISIS selects 10 exogenous and 2 lags. Figure 5.2 provides the fitted values of the CPI observations in the training set by using the methods described above, and Figure 5.3 provides the predicted values of the CPI from Q1 2013 to Q4 2013 using these methods. The findings from Figures

5.2 and 5.3 are consistent with those from Table 5.3.

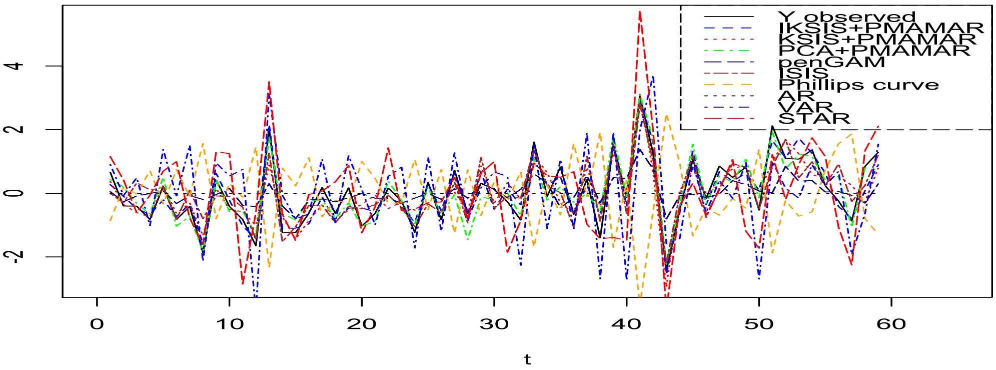

Figure 5.2: Observed $Y$ values and fitted values from the methods considered.

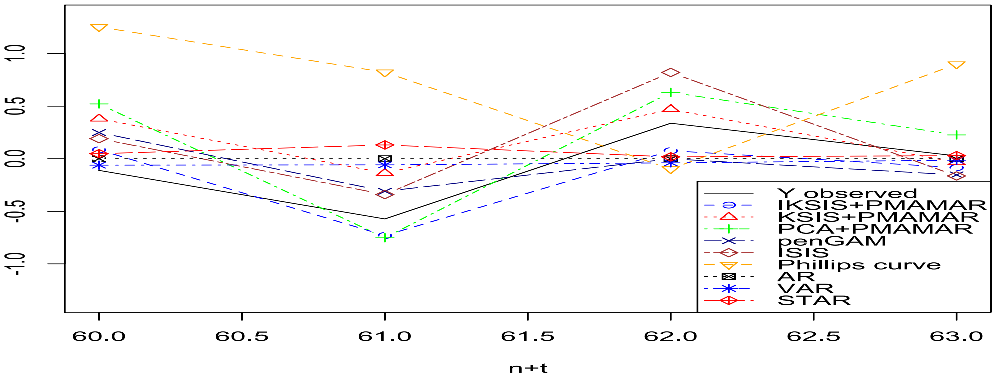

Figure 5.3: True values of $Y$ from Q1 2013 to Q4 2013 and their forecasts from the methods considered.

# 6 Conclusion

In this paper, we have developed two types of semiparametric methods to achieve dimension reduction on the candidate covariates and obtain good forecasting performance for the response variable. The KSIS technique, as the first step of the

KSIS+PMAMAR method and the generalisation of the SIS technique proposed by Fan and Lv (2008), screens out the regressors whose marginal regression functions do not make significant contribution to estimating the joint regression function and reduces the dimension of the regressors from an ultra large size to a moderately large size. The sure screening property developed in Theorem 1 shows that, through KSIS, the covariates whose marginal regression functions make truly significant contribution would be selected with probability approaching one. An iterative version of the KSIS is further developed in Section 4.1 and it can be seen as a possible solution to address the issue of false selection of some irrelevant covariates which are highly correlated to the significant covariates. The PMAMAR approach, as the second step of the two semiparametric dimension-reduction methods, is an extension of the MAMAR approximation introduced in Li, Linton and Lu (2015). Theorem 2 proves that the PMAMAR enjoys some well-known properties in high-dimensional variable selection such as the sparsity and oracle property. Both the simulated and empirical examples in Section 5 show that the KSIS+PMAMAR and its iterative version perform reasonably well in finite samples.

The second PCA+PMAMAR method is a generalisation of the well-known factor-augmented linear regression and auto-regression models (c.f., Stock and Watson, 2002; Bernanke, Boivin and Eliasz, 2005; Bai and Ng, 2006). By assuming an approximate factor structure on the ultra-high dimensional exogenous regressors and implementing the PCA, we estimate the unobservable factor regressors and achieve dimension reduction on the exogenous regressors. Our Theorem 3 shows that the estimated factor regressors are uniformly consistent and the asymptotic properties for the subsequent PMAMAR method (c.f., Theorem 2) remains valid for further selection of the estimated factor regressors and the time series lags. Example 5.2 shows that the P-CA+PMAMAR method performs well in predicting the future value of the time series when the sample size is small ($n = 100$). Furthermore, we may extend the methodology and theory developed in this paper to the more general case where some lags of the estimated factor regressors are included in the PMAMAR procedure.

# Supplemental document

The supplemental document contains the detailed proofs of the main asymptotic theorems given in Section 3 as well as some technical lemmas.

# References

[1] Akaike, H., 1979. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* 66, 237–242.

[2] Ando, T., Li, K., 2014. A model averaging approach for high-dimensional regression. *Journal of the American Statistical Association* 109, 254–265.

[3] Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. *Econometrica* 70, 191–221.

[4] Bai, J., Ng, S., 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74, 1135–1150.

[5] Bernanke, B., Boivin, J., Eliasz, P. S., 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics* 120, 387–422.

[6] Boneva, L., Linton, O., Vogt, M., 2015. A semiparametric model for heterogeneous panel data with fixed effects. *Journal of Econometrics* 188, 327–345.

[7] Bosq, D., 1998. *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer.

[8] Chen, J., Li, D., Linton, O., Lu, Z., 2016. Semiparametric dynamic portfolio choice with multiple conditioning variables. Forthcoming in *Journal of Econometrics*.

[9] Cheng, X., Hansen, B., 2015. Forecasting with factor-augmented regression: a frequentist model averaging approach. *Journal of Econometrics* 186, 280–293.

[10] Claeskens, G., Hjort, N., 2008. *Model Selection and Model Averaging*. Cambridge University Press.

[11] Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51, 1305–1324.

[12] Fama, E., French, K., 1992. The cross-section of expected stock returns. *Journal of Finance* 47, 427–465.

[13] Fan, J., Feng, Y., Song, R., 2011. Nonparametic independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association* 116, 544–557.

[14] Fan, J., Gijbels, I., 1996. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

[15] Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.

[16] Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements (with discussions). *Journal of the Royal Statistical Society: Series B* 75, 603–680.

[17] Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* 70, 849–911.

[18] Fan, J., Ma, Y., Dai, W., 2014. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association* 109, 1270–1284.

[19] Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* 32, 928–961.

[20] Frank, I. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35, 109–148.

[21] Green, P., Silverman, B., 1994. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC.

[22] Härdle, W., Tsybakov, A. B., 1995. Additive nonparametric regression on principal components. *Journal of Nonparametric Statistics* 5, 157–184.

[23] Hansen, B. E., 2007. Least squares model averaging. *Econometrica* 75, 1175–1189.

[24] Hansen, B. E., Racine, J., 2012. Jackknife model averaging. *Journal of Econometrics* 167, 38–46.

[25] Li, D., Linton, O., Lu, Z., 2015. A flexible semiparametric forecasting model for time series. *Journal of Econometrics* 187, 345–357.

[26] Li, D., Lu, Z., Linton, O., 2012. Local linear fitting under near epoch dependence: uniform consistency with convergence rates. *Econometric Theory* 28, 935–958.

[27] Liu, J., Li, R., Wu, R., 2014. Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Journal of the American Statistical Association* 109, 266–274.

[28] Lu, Z., Linton, O., 2007. Local linear fitting under near epoch dependence. *Econometric Theory* 23, 37–70.

[29] Meier, L., van de Geer, S., Bühlmann, P., 2009. High-dimensional additive modeling. *Annals of Statistics* 37, 3779–3821.

[30] Pesaran, M. H., Pick, A., Timmermann, A., 2011. Variable selection, estimation and inference for multi-period forecasting problems. *Journal of Econometrics* 164, 173–187.

[31] Raftery, A. E., Madigan, D., Hoeting, J. A., 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92, 179–191.

[32] Stock, J. H., Watson, M. W., 1998. Diffusion indexes. *NBER Working Paper 6702.*

[33] Stock, J. H., Watson, M. W., 1999. Forecasting inflation. *NBER Working Paper 7023.*

[34] Stock, J. H., Watson, M. W., 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–1179.

[35] Teräsvirta, T., Tjøstheim, D., Granger, C., 2010. *Modelling Nonlinear Economic Time Series.* Oxford University Press.

[36] Tibshirani, R. J., 1996. Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society Series B* 58, 267–288.

[37] Tibshirani, R. J., 1997. The LASSO Method for Variable Selection in the Cox Model. *Statistics in Medicine* 16, 385–395.

[38] Wan, A. T. K., Zhang, X., Zou, G., 2010. Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156, 277–283.

[39] Wand, M. P., Jones, M. C., 1995. *Kernel Smoothing.* Chapman and Hall.