cemmap

centre for microdata methods and practice

Semiparametric model averaging of ultra-high dimensional time series

Jia Chen
Degui Li
Oliver Linton
Zudi Lu

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP62/15

Semiparametric Model Averaging of Ultra-High Dimensional Time Series

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

October 5, 2015

Abstract

In this paper, we consider semiparametric model averaging of the nonlinear dynamic time series system where the number of exogenous regressors is ultra large and the number of auto-regressors is moderately large. In order to accurately forecast the response variable, we propose two semiparametric approaches of dimension reduction among the exogenous regressors and auto-regressors (lags of the response variable). In the first approach, we introduce a Kernel Sure Independence Screening (KSIS) technique for the nonlinear time series setting which screens out the regressors whose marginal regression (or auto-regression) functions do not make significant contribution to estimating the joint multivariate regression function and thus reduces the dimension of the regressors from a possible exponential rate to a certain polynomial rate, typically smaller than the sample size; then we consider a semiparametric method of Model Averaging MArginal Regression (MAMAR) for the regressors and auto-regressors that survive the screening procedure, and propose a penalised MAMAR method to further select the regressors which have significant effects on estimating the multivariate regression function and predicting the future values of the response variable. In the second approach, we impose an approximate factor modelling structure on the ultra-high dimensional exogenous regressors and use a well-known principal component analysis to estimate the latent common factors, and then

[*]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.

[§]School of Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. E-mail: Z.Lu@soton.ac.uk.

apply the penalised MAMAR method to select the estimated common factors and lags of the response variable which are significant. Through either of the two approaches, we can finally determine the optimal combination of the significant marginal regression and auto-regression functions. Under some regularity conditions, we derive the asymptotic properties for the two semiparametric dimension-reduction approaches. Some numerical studies including simulation and an empirical application are provided to illustrate the proposed methodology.

JEL subject classifications: C14, C22, C52.

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

1 Introduction

Suppose that Y_t, $t = 1, \ldots, n$, are n observations collected from a stationary time series process. In practical applications, it is often interesting to study the multivariate regression function:

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

where $\mathbf{X}_t = (\mathbf{Z}_t^{\intercal}, \mathbf{Y}_{t-1}^{\intercal})^{\intercal}$ with $\mathbf{Z}_t = \left(Z_{t1}, Z_{t2}, \ldots, Z_{tp_n}\right)^{\intercal}$ being a p_n-dimensional vector of exogenous regressors and $\mathbf{Y}_{t-1} = (Y_{t-1}, Y_{t-2}, \ldots, Y_{t-d_n})^{\intercal}$ being a vector of d_n lags of the response variable Y_t, the superscript $^{\intercal}$ stands for the transpose of a vector (or a matrix). In the nonlinear dynamic time series analysis, it is reasonable to assume that both p_n and d_n increase with the sample size n, and the dimension of the exogenous regressors can be even larger than the sample size. Such an ultra-high dimensional time series setting poses challenges in estimating the regression function $m(\mathbf{x})$ and the subsequent forecasting of the response. It is well known that when the dimension of $\mathbf{X}_t$ is very small (say 1 or 2), 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 the dimension is large, owing to the so-called "curse of dimensionality", the direct use of nonparametric methods might lead to a very poor estimation result and forecasting performance. Hence, various nonparametric and semiparametric models, such as additive models, varying coefficient models and partially linear models, have been proposed to deal with the curse of dimensionality in the literature for the dynamic time series data (c.f., Teräsvirta, Tjøstheim and Granger, 2010).

It is well-known that the approach of model averaging is useful for improving the accuracy of predicting future values of the response variable in time series analysis. The model averaging approach advocates combining 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 may fail. 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 shows that such a method works well both theoretically and numerically; while 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, we relax the restriction of linear modelling framework assumed in Ando and Li (2014) and Cheng and Hansen (2015) by studying the nonlinear dynamic regression structure for (1.1) which would provide a much more flexible framework.

Throughout the paper, we assume that the dimension of the exogenous variables $\mathbf{Z}_t$, p_n, may be diverging at certain exponential rate of n, which indicates that the dimension of the potential explanatory variables $\mathbf{X}_t$, p_n+d_n, can be diverging at an exponential rate, i.e., $p_n+d_n = O(\exp\{n^{\delta_0}\})$ for some positive constant δ_0. To ensure that our semiparametric model averaging technique 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. To achieve the aim of dimension reduction, in this paper we propose two methods both of which include two steps in the respective algorithm.

The first dimension reduction method is called as the "KSIS+PMAMAR" method which reduces the dimension of the potential covariates via two steps introduced as follows. In the first step, we use the approach of Kernal Sure Independence Screening (KSIS) which is motivated by Fan and Lv (2008)'s Sure Independence Screening (SIS) method in the context of linear regression to screen out the unimportant marginal regression (or auto-regression) functions, and reduce the dimension of the potential covariates from the exponential rate to a certain polynomial rate of n which is typically smaller than the sample size. This is done by first calculating the correlations between the response variable Y_t and the marginal regression or auto-regression functions $\mathsf{E}[Y_t|X_{tj}]$ with X_{tj} being a univariate element chosen from $\mathbf{Z}_t$ or $\mathbf{Y}_{t-1}$, and then removing those covariates whose corresponding correlation coefficients are smaller than a pre-determined threshold value. Then we denote the chosen covariates by $\mathbf{X}_t^* = \left(X_{t1}^*, X_{t2}^*, \ldots, X_{tq_n}^*\right)^\intercal$ which may include both exogenous variables and lags of the response variable, where q_n might be diverging but is smaller than the sample size n. In the second step, we propose using a semiparametric method of model averaging lower dimensional regression functions to estimate

$$m^*(\mathbf{x}) = \mathsf{E}(Y_t|\mathbf{X}_t^* = \mathbf{x}), \tag{1.2}$$

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.,

$$m^*(\mathbf{x}) \approx w_0 + \sum_{j=1}^{q_n} w_j m_j^*(x_j), \tag{1.3}$$

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. 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 which would not suffer from the curse of dimensionality problem. By replacing $m_j^*(X_{tj}^*)$, $j = 1, \ldots, q_n$, by their corresponding nonparametric estimates $\hat{m}_j^*(X_{tj}^*)$, we have the following "approximate linear model":

$$Y_t \approx w_0 + \sum_{j=1}^{q_n} w_j \hat{m}_j^*(X_{tj}^*). \tag{1.4}$$

To further select the significant components $m_j^*(X_{tj}^*)$ in (1.4), we use the penalisation device to force some weights to zero as penalisation with linear regression models. For example, Tibshirani (1996, 1997) proposes the penalised least squares estimation with the L_1 penalty, which is well known as the least absolute shrinkage and selection operator (LASSO). Frank and Friedman (1993) and Fu (1998) study the penalised regression with general L_q penalty, which leads to the bridge regression. Fan and Li (2001) and Fan and Peng (2004) use the smoothly clipped absolute deviation (SCAD) penalty in the penalised likelihood method to carry out the estimation and variable selection simultaneously. Bühlmann and van de Geer (2011) review the recent developments on this popular topic. As in Fan and Li (2001), we will select the significant covariates and estimate the optimal weights simultaneously in the second step and use the optimal combination of the significant marginal regressions for predicting the response. This method is called as Penalised Model Averaging MArginal Regression (PMAMAR).

The second dimension-reduction method is called as "PCA+PMAMAR" which is also a two-step procedure. In the first step, we assume that the ultra-high dimensional exogenous regressors $\mathbf{Z}_t$ satisfy the approximate factor model which has been commonly used in economic and financial data analysis (c.f., Chamberlain and Rothschild, 1983; Fama and French, 1992; Stock and Watson, 2002; Bai and Ng, 2002, 2006):

$$Z_{tk} = (\mathbf{b}_k^0)^\intercal \mathbf{f}_t^0 + u_{tk}, \quad k = 1, \ldots, p_n, \tag{1.5}$$

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. We then apply the technique of Principal Component Analysis (PCA) to estimate the latent factors which capture a large proportion of the information contained in the exogenous regressors $\mathbf{Z}_t$, and thus achieve dimension reduction on $\mathbf{Z}_t$. We denote

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

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 sketched above 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. The proposed 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.

Under some regularity conditions, we establish the asymptotic properties of the developed semiparametric approaches. For the KSIS procedure, we establish the sure screening property, which indicates 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 one to form $\mathbf{X}_t^*$ which would undergo a further selection in the PMAMAR procedure. For the PCA approach, we show that the estimated latent factors are uniformly consistent with convergence rate dependent on both n and p_n, and the kernel estimation of the marginal regression with 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 approaches, we prove that the optimal weight estimation enjoys the well-known sparsity and oracle property with the estimated values of the true zero weights forced to be zero.

In addition, we further discuss some extensions of the proposed semiparametric dimension reduction approaches such as an iterative KSIS+PMAMAR procedure when implementing them in practice. Through the simulation studies, we show 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. Finally, we apply the developed semiparametric model averaging methods to forecast the quarterly inflation in the UK and compare the results with those using other commonly-used methods.

The rest of the paper is organised as follows. The semiparametric model averaging methods are introduced in Section 2. The asymptotic theory of the developed methodology is established in Section 3. Section 4 discusses some extensions when the methods are implemented in practice. Section 5 gives some numerical studies to investigate the finite sample behaviour of the proposed methodology and Section 6 concludes this paper. The proofs of the asymptotic results are provided in an appendix.

2 Semiparametric model averaging

In this section, we introduce two types of semiparametric model averaging approaches which result in dimension reduction of the possibly ultra-high dimensional covariates. One is the KSIS+PMAMAR method proposed in Section 2.1 and the other is the PCA+PMAMAR method proposed in Section 2.2.

2.1 *KSIS+PMAMAR method*

As mentioned in Section 1, the KSIS+PMAMAR method is a two-step procedure. We first generalise Fan and Lv (2008)'s SIS method 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 $m^*(\cdot)$ defined in (1.2), and use an optimal combination of the significant marginal regression and auto-regression functions to approximate $m^*(\cdot)$.

Step one: KSIS. For notational simplicity, we let

$$X_{tj} = \begin{cases} Z_{tj}, & j = 1, 2, \ldots, p_n, \\ Y_{t-(j-p_n)}, & j = p_n + 1, p_n + 2, \ldots, p_n + d_n. \end{cases}$$

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 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 splines 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 simply 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 which 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,

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

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

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

where

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

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

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 ρ_n is a pre-determined positive number. By (2.3), the criterion in (2.4) is equivalent to

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

where $\rho_n^{\diamond} = \rho_n^{1/2}/\sqrt{\hat{\mathsf{v}}(Y)}$. As in Section 1, we let $\mathbf{X}_t^* = \big(X_{t1}^*, X_{t2}^*, \ldots, X_{tq_n}^*\big)^{\mathsf{T}}$ 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 is 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 might be too restrictive for data arising from economics and finance. In this paper, we relax such a restriction and show that the developed KSIS approach works well for the ultra-high dimensional time series and semiparametric setting. Another difference between our paper and the paper by Fan, Feng and Song (2011) is that the kernel smoothing method is used in this paper to estimate the marginal regression functions whereas the B-splines method is used in Fan, Feng and Song (2011). Hence, a different mathematical tool is needed to derive our asymptotic theory.

Step two: PMAMAR. We next consider the multivariate regression function defined in (1.2) which can be seen as an approximation to the multivariate regression function defined in (1.1) after screening out the unimportant covariates. In order to avoid the curse of dimensionality, we approximate $m^*(\cdot)$ by an affine combination of one-dimensional marginal regression functions $m_j^*(\cdot)$ with $j = 1, \ldots, q_n$. Such a method is called as model averaging marginal regressions or MAMAR (Li, Linton and Lu, 2015) and is applied by Chen *et al* (2015) in the dynamic portfolio choice with many conditioning variables. Since some of the marginal regression functions may not have a significant effect on estimating $m^*(\cdot)$, they should be excluded in order to further enhance the predictability of the semiparametrically approximated model. Hence, we next introduce a penalised version of the MAMAR technique to simultaneously determine which marginal regression functions should be included in the model averaging and obtain the affine weights.

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

$$\hat{m}_j^*(x_j) = \frac{\sum_{t=1}^n Y_t \overline{K}_{tj}(x_j)}{\sum_{t=1}^n \overline{K}_{tj}(x_j)}, \quad \overline{K}_{tj}(x_j) = K\Big(\frac{X_{tj}^* - x_j}{h_2}\Big), \quad j = 1, \ldots, q_n, \tag{2.5}$$

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 (2.5), 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 $\mathcal{Y}_n = (Y_1, \ldots, Y_n)^\intercal$.

As introduced in (1.4) of Section 1, the second stage of PMAMAR is to replace the marginal regression functions by their corresponding kernel estimates, and then use the penalised approach to select the significant marginal regression functions. 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 (1.3) 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 (1.4), for given $\mathbf{w}_n = (w_1, \ldots, w_{q_n})^\intercal$, we define the objective function by

$$\mathcal{Q}_n(\mathbf{w}_n) = \left[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_n)\right]^\intercal \left[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_n)\right] + n\sum_{j=1}^{q_n} p_\lambda(|w_j|), \tag{2.6}$$

where

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

$\mathcal{S}_n(\mathcal{Y}) = \left[\mathcal{S}_n(1)\mathcal{Y}_n, \ldots, \mathcal{S}_n(q_n)\mathcal{Y}_n\right]$, and $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ. 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 $\mathcal{Q}_n(\mathbf{w}_n)$:

$$\hat{\mathbf{w}}_n = \arg\min_{\mathbf{w}_n} \mathcal{Q}_n(\mathbf{w}_n). \tag{2.7}$$

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 λ. 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, LASSO which is the L_1-penalty $p_\lambda(|z|) = \lambda|z|$ has been extensively studied by many authors (see, for example, Bühlmann and van de Geer, 2011); 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

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

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 defined in (1.5). 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, we can find that the main information in the exogenous regressors may be summarised in the common factors $\mathbf{f}_t^0$ which 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 (1.5) 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

$$m_f(\mathbf{x}_1, \mathbf{x}_2) = \mathsf{E}\left(Y_t | \mathbf{f}_t^0 = \mathbf{x}_1, \mathbf{Y}_{t-1} = \mathbf{x}_2\right), \tag{2.8}$$

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 (2.8), we need to estimate the unobservable factor regressors $\mathbf{f}_t^0$. This will be done through the PCA approach.

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 (1.5) as

$$\mathbf{Z}_t = \mathbf{B}_n^0 \mathbf{f}_t^0 + \mathbf{U}_t. \tag{2.9}$$

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 $\hat{\mathcal{F}}_n = \left(\hat{\mathbf{f}}_1, \dots, \hat{\mathbf{f}}_n\right)^\intercal$ 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

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

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

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

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:

$$m_f^*(\mathbf{x}_1, \mathbf{x}_2) = \mathsf{E}\left(Y_t | \mathbf{H}\mathbf{f}_t^0 = \mathbf{x}_1, \mathbf{Y}_{t-1} = \mathbf{x}_2\right). \tag{2.10}$$

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 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. As in Section 1, we define

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

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

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

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

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

$$\hat{m}^*_{k,f}(z_k) = \frac{\sum_{t=1}^n Y_t \widetilde{K}_{tk}(z_k)}{\sum_{t=1}^n \widetilde{K}_{tk}(z_k)}, \quad \widetilde{K}_{tk}(z_k) = K\Big(\frac{\hat{f}_{tk} - z_k}{h_3}\Big), \quad j = 1, \dots r, \tag{2.11}$$

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 (2.11) 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 α-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 $\sup_j \sup_{x_j \in \mathcal{C}_j} \left[|m_j(x_j)|+|m_j'(x_j)|+|m_j''(x_j)|\right] \leq c_m$.*

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 α-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 ζ_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 δ_2 such that

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

where $M(n) = (p_n + d_n)n^{(17+18\theta_1)/10}$ and θ_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

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

then we have

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

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 marginal regression functions and the response 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

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

and

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

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

A6. *The matrix*

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

is positive definite with the largest eigenvalue bounded. The smallest eigenvalue of $\mathbf{\Lambda}_n$, χ_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* ξ *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* $\big|p_\lambda''(\vartheta_1) - p_\lambda''(\vartheta_2)\big| \le C_2|\vartheta_1 - \vartheta_2|$ *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 χ being a positive constant.

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

$$\mathbf{w}_o = (w_{o1}, \dots, w_{oq_n})^\intercal = \big[\mathbf{w}_o^\intercal(1),\ \mathbf{w}_o^\intercal(2)\big]^\intercal,$$

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 $\boldsymbol{\xi}_t = \big(\xi_{t1}, \ldots, \xi_{ts_n}\big)^{\intercal}$ 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 $\boldsymbol{\Sigma}_n = \sum_{t=-\infty}^{\infty} \mathsf{E}\big(\boldsymbol{\xi}_0\boldsymbol{\xi}_t^{\intercal}\big)$ and $\boldsymbol{\Lambda}_{n1}$ as the top-left $s_n \times s_n$ submatrix of $\boldsymbol{\Lambda}_n$. Let

$$\boldsymbol{\omega}_n = \left[p'_\lambda(|w_{o1}|)\mathsf{sgn}(w_{o1}), \ldots, p'_\lambda(|w_{os_n}|)\mathsf{sgn}(w_{os_n})\right]^{\intercal}$$

and

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

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 (2.6) 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 χ_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 $\boldsymbol{\Lambda}_{n1}$ are bounded away from zero and infinity,

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

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, the matrix $\boldsymbol{\Lambda}_n$ and the penalty function. The involvement of χ_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 λ tends to zero sufficiently fast for some penalty functions (such as the SCAD penalty), the asymptotic distribution in (3.7) would reduce to

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

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 are the consistent estimation of the true common factors (with rotation), and the asymptotic order of the difference between $\hat{m}^*_{k,f}(z_k)$ defined in (2.11) 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 which are used when proving the asymptotic results.

- **B1**. *The process $\{(Y_t, \mathbf{f}_t, \mathbf{U}_t)\}$ is stationary and α-mixing with the mixing coefficient decaying at a geometric rate: $\alpha(k) \sim c_\alpha\theta_0^k$, where c_α and $0 < \theta_0 < 1$ are defined as in the condition* **A1**.
- **B2**. *The random common factors satisfy the conditions that $\mathsf{E}\,[\mathbf{f}_t^0] = \mathbf{0}$, $\max_t \|\mathbf{f}_t^0\| = O_P(1)$, the $r \times r$ matrix $\mathbf{\Lambda}_F := \mathsf{E}\big[\mathbf{f}_t^0(\mathbf{f}_t^0)^\intercal\big]$ 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*

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

and

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

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,* $n^3/(p_n^2h_3^4)=o(1)$.

(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 of the methodology

In this section, we first introduce an iterative KSIS+PMAMAR procedure which is expected to work well when the covariates are highly correlated, and then discuss how to select the number of the latent factors in the approximate factor model (1.5) and an extension of the PCA+PMAMAR approach.

4.1 An iterative KSIS+PMAMAR procedure

Difficulties in variable selection arise when the covariates are highly correlated with each other. It is documented in Fan and Lv (2008) that even if the covariates are mutually independent, the data generated from them may exhibit significant spurious correlation when the covariate dimension is large. As discussed in Fan and Lv (2008), 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 is even more severe in the present paper. Due to the time series nature of the data, both the response Y_t and the covariates $\mathbf{X}_t$ are likely to be autocorrelated over time t. This results in both the autocorrelation between X_{tj} and X_{sj}, $t \neq s$, and the inter-correlation between covariate X_{tj} and X_{tk}, $j \neq k$, as X_{tj}, $j = p_n + 1, \ldots, p_n + d_n$, are generated from the lags of Y_t. 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 the components that have survived the screening process, then the results could be very unsatisfactory. This is especially so when $p_n + d_n$ is much larger than the sample size n. To alleviate this problem, we propose below an iterative version for the KSIS+PMAMAR procedure.

Due to the autocorrelation in the response Y_t and the lagged covariates X_{tj}, $j = p_n + 1, \ldots, p_n + d_n$, the iterative procedure developed in Fan, Feng and Song (2011) is not applicable in this context. This is because their iterative procedure includes 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, as is the case in our context, permutation would destroy the inherent serial dependence structure and hence may lead to erroneous thresholds being obtained. Our iterative KSIS+PMAMAR procedure is 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 $\widehat{e}^S$.

Step 3: Run a linear regression the estimated marginal regression function of each variable in $\mathcal{S}^c$ which is defined as $\{1, 2, \ldots, p_n + d_n\} \setminus \mathcal{S}$ on the estimated marginal regression functions of the selected variables in $\mathcal{S}$, and obtain the residuals $\widehat{e}^{iS}$ 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 $\widehat{e}^S$ from Step 2 on the residuals $\widehat{e}^{iS}$ from Step 3 for each $i \in \mathcal{S}^c$, and calculate the sample covariance $\widehat{\mathsf{v}}^e(i)$ between $\widehat{e}^S$ 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 of the above procedure, we treat the residuals from linearly regressing the response variable on the marginal regression functions of currently selected variables as the new response variable and the residuals from linearly regressing the marginal regression functions of the unselected variables on those of the selected variables as the new covariates, and 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 those remaining irrelevant variables that are highly correlated with some selected relevant variables being picked and increases the priority of the remaining relevant variables that 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 then updated with any variables having insignificant weights being discarded. Other penalty functions such as the LASSO and the MCP are equally applicable in Step 5. The above iterative procedure can be considered as a greedy selection algorithm, since at most one variable is selected in each iteration. This algorithm starts with zero variables 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 selection of number of factors and the PCA+KSIS+PMAMAR procedure

In reality, the number of common factors, r, in the approximate factor model (1.5) is usually unknown. Hence, we need to select the number of factors to be extracted from an eigenanalysis of the matrix $\mathcal{Z}_n\mathcal{Z}_n^{\intercal}/(np_n)$. There could be two ways to address this issue. The first is to set a maximum number, say $r_{\max}$ (which is usually not too large), 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 selected 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. The reader is referred to Boneva, Linton and Vogt (2015) for more information on the selection of the number of common component functions. Other commonly-used 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, when d_n is large, the PMAMAR regression usually cannot produce satisfactory results in selecting the truly significant lags. This could lead to poor performance of the PCA+PMAMAR procedure in the prediction of 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 show that this PCA+KSIS+PMAMAR procedure improves the prediction performance of the PCA+PMAMAR procedure.

5 Numerical studies

In this section, we give simulation studies (Examples 5.1 and 5.2) and an empirical application (Example 5.3) of the methodology developed in Section 2 and the extensions discussed in Section 4.

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 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

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

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 $\text{cov}(\mathbf{Z}) = I_{p_n}$ or $C_{\mathbf{Z}}$, whose the main-diagonal entries are 1 and off-diagonal entries are 1/2. The error term ε_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 simulated example is to compare the performance of the iterative KSIS+PMAMAR (IKSIS+PMAMAR) procedure proposed in Section 4.1 with the (non-iterative) KSIS+PMAMAR procedure proposed 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 SCAD-penalised MAMAR regression. 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 estimation error (EE) defined as the mean squared error, the prediction error (PE) defined as the mean squared prediction error. We generate a prediction test set of size $n/10 = 10$ and calculated the one-step-ahead forecasts for the response Y, from which the PE is obtained. 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[1]. 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 EE, since within the same linear or nonlinear modelling framework it is generally the case that the more variables are selected the smaller the EE 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 PE. The results for the Oracle serve as a benchmark for those of the other approaches. The PEs 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 via 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

[1] The authors thank Dr Lukas Meier for kindly providing the "penGAM" package.

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	EE	PE
Example 5.1 $\mathrm{cov}(\mathbf{Z}) = I_{p_n}$ $(p_n, d_n) = (30, 10)$	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 $\mathrm{cov}(\mathbf{Z}) = I_{p_n}$ $(p_n, d_n) = (150, 50)$	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 $\mathrm{cov}(\mathbf{Z}) = C_{\mathbf{Z}}$ $(p_n, d_n) = (30, 10)$	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 $\mathrm{cov}(\mathbf{Z}) = C_{\mathbf{Z}}$ $(p_n, d_n) = (150, 50)$	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)

terms $\mathbf{z}_t$, $t = 1, \cdots, n$, are independently drawn from the $0.1\mathsf{N}(\mathbf{0}, I_{p_n})$ distribution. We set $p_n = 30$ or 150, $r = 3$, and generate the response variable via

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

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 ε_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 EE and PE. The PCA+PMAMAR has a lower EE but higher PE than the PCA+KSIS+PMAMAR. This is due to the fact that without the screening step the PCA+PMAMAR selects more false lags of Y, and the higher FP leads to an higher PE and lower EE under the same PMAMAR framework. The above suggests that if one's main concern is to predict future values of the response, there may be benefits in having the KSIS step to screen out some insignificant lags between the PCA and PMAMAR step.

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

Table 5.2: Average results on accuracy of estimation and prediction in Example 5.2 over 200 replications

Model	Method	EE	PE
Example 5.2 $(p_n, d_n) = (30, 10)$	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 $(p_n, d_n) = (150, 10)$	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)

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 time 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.

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 form forecasts of CPI in the four quarters of 2013. As in the simulation, we compare the forecasting capacity of the IKSIS+PMAMAR, KSIS+PMAMAR, PCA+PMAMAR, penGAM and ISIS methods. Note that due to the small number of candidate lags of the response ($d = 4$), there is not much necessity

Table 5.3: Mean squared prediction errors of various approaches in forecasting inflation in the UK

Method	IKSIS+PMAMAR	KSIS+PMAMAR	PCA+PMAMAR	penGAM	ISIS	Phillips curve
PE	0.0360	0.1130	0.0787	0.0865	0.3275	1.1900

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:

$$I_{t+1} - I_t = \alpha + \beta(L)U_t + \gamma(L)\Delta I_t + \varepsilon_{t+1}, \tag{5.4}$$

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 Δ is the first difference operator.

The prediction errors (PE) of the above approaches are summarised in Table 5.3, which shows the IKSIS+PMAMAR has the smallest PE followed by the PCA+PMAMAR and penGAM and then the KSIS+PMAMAR and ISIS. The Phillips curve forecasts are much worse than those of the other 5 methods. 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 4 common factors from the 53 exogenous variables and 3 lags of response; the penGAM selects 2 exogenous only; and the ISIS selects 17 exogenous and 2 lags.

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 some 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). Through assuming an approximate factor modelling 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. Theorem 3 in Section 3.2 indicates that the estimated factor regressors are uniformly consistent and the asymptotic properties for the subsequent PMAMAR method (c.f., Theorem 2) remains valid for the further selection of the estimated factor regressors and the lags of the response variable. Example 5.2 shows that the PCA+PMAMAR method performs well in predicting the future value of the response variable when the sample size is small ($n = 100$). Furthermore, we may extend the methodology and theory to the more general case that some lags of the estimated factor regressors are included in the PMAMAR procedure.

Appendix A: Some technical lemmas

In this appendix, we give some technical lemmas which will be used in the proof the main results. The first result in a well-known exponential inequality for the α-mixing sequence which can be found in some existing literature such as Bosq (1998).

Lemma 1. *Let $\{Z_t\}$ be a zero-mean α-mixing process satisfying $\mathsf{P}(|Z_t| \leq B) = 1$ for all $t \geq 1$. Then for each integer $q \in [1, n/2]$ and each $\epsilon > 0$, we have*

$$\mathsf{P}\Big(\Big|\sum_{t=1}^{n} Z_t\Big| > n\epsilon\Big) \leq 4\exp\Big(-\frac{\epsilon^2 q}{8v^2(q)}\Big) + 22\Big(1+\frac{4B}{\epsilon}\Big)^{1/2} q\alpha([p]), \tag{A.1}$$

where $v^2(q) = 2\sigma^2(q)/p^2 + B\epsilon/2$, $p = n/(2q)$,

$$\sigma^2(q) = \max_{1\leq j\leq 2q-1} \mathsf{E}\ \Big\{([jp]+1-jp)Z_{[jp]+1} + Z_{[jp]+2} + \ldots + Z_{[(j+1)p]} + ((j+1)p - [(j+1)p])Z_{[(j+1)p]+1}\Big\}^2$$

and $[\cdot]$ denotes the integer part.

Define

$$\mathcal{G}_{ji} = \Big\{\sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^{n}\Big\{\big(\frac{X_{tj}-x_j}{h_1}\big)^i K\big(\frac{X_{tj}-x_j}{h_1}\big) - \mathsf{E}\big[\big(\frac{X_{tj}-x_j}{h_1}\big)^i K\big(\frac{X_{tj}-x_j}{h_1}\big)\big]\Big\}\Big| \geq (nh_1)^{\frac{1}{2}+\kappa_2}\Big\}, \tag{A.2}$$

where $0 < \kappa_2 < 1/2$ and $i = 0, 1, \ldots$. The following Lemma gives the upper bound of the probability of the event $\mathcal{G}_{j0} \cup \mathcal{G}_{j1} \cup \mathcal{G}_{j2}$.

Lemma 2. *Suppose that the conditions A1–A3 in Section 3.1 are satisfied. Then we have*

$$\mathsf{P}\big(\mathcal{G}_{j0}\cup\mathcal{G}_{j1}\cup\mathcal{G}_{j2}\big) \leq M_1 n^{(5+9\theta_1-2\kappa_2+2\kappa_2\theta_1)/4}\Big[\exp\big\{-c_1 n^{2(1-\theta_1)\kappa_2}\big\} + \exp\Big\{-c_2 n^{(1-\theta_1)(\frac{1}{2}-\kappa_2)}\Big\}\Big] \tag{A.3}$$

for $j = 1, 2, \ldots, p_n + d_n$, where c_1, c_2 and M_1 are positive constants which are independent of j, and θ_1 is defined in the condition A3.

Proof. We next only prove that

$$\mathsf{P}\big(\mathcal{G}_{j0}\big) \leq \frac{M_1}{3}\Big(n^{(5+9\theta_1-2\kappa_2+2\kappa_2\theta_1)/4}\Big[\exp\big\{-c_1 n^{2(1-\theta_1)\kappa_2}\big\} + \exp\Big\{-c_2 n^{(1-\theta_1)(\frac{1}{2}-\kappa_2)}\Big\}\Big]\Big), \tag{A.4}$$

as the same conclusion also holds for $\mathcal{G}_{j1}$ and $\mathcal{G}_{j2}$ with a similar proof. Then the proof of (A.3) can be completed. We cover the uniformly bounded set $\mathcal{C}_j$ by a finite number of intervals $\mathcal{C}_j(k)$ with centre $s_j(k)$ and radius $h_1(nh_1)^{\frac{1}{2}+\kappa_2}/(3c_K n)$, where c_K is a positive constant such that $|K(u) - K(v)| \leq c_K|u-v|$. Let $N_n(j)$ be the total number of $\mathcal{C}_j(k)$ and it is easy to see that the order of $N_n(j)$ is $O\big(nh_1^{-1}(nh_1)^{-\frac{1}{2}-\kappa_2}\big)$.

Note that

$$\begin{aligned}
& \sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^n\Big\{K\big(\frac{X_{tj}-x_j}{h_1}\big)-\mathsf{E}\big[K\big(\frac{X_{tj}-x_j}{h_1}\big)\big]\Big\}\Big| \\
\leq\ & \max_{1\leq k\leq N_n(j)}\Big|\sum_{t=1}^n\Big\{K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)-\mathsf{E}\big[K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)\big]\Big\}\Big|+ \\
& \max_{1\leq k\leq N_n(j)}\sup_{x_j\in\mathcal{C}_j(k)}\Big|\sum_{t=1}^n\Big[K\big(\frac{X_{tj}-x_j}{h_1}\big)-K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)\Big]\Big|+ \\
& \max_{1\leq k\leq N_n(j)}\sup_{x_j\in\mathcal{C}_j(k)}\Big|\sum_{t=1}^n\mathsf{E}\Big[K\big(\frac{X_{tj}-x_j}{h_1}\big)-K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)\Big]\Big| \\
\leq\ & \max_{1\leq k\leq N_n(j)}\Big|\sum_{t=1}^n\Big\{K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)-\mathsf{E}\big[K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)\big]\Big\}\Big|+\frac{2}{3}(nh_1)^{\frac{1}{2}+\kappa_2},
\end{aligned}$$

which indicates that

$$\begin{aligned}
\mathsf{P}\big(\mathcal{G}_{j0}\big)\ \leq\ & \mathsf{P}\Big(\max_{1\leq k\leq N_n(j)}\Big|\sum_{t=1}^n\Big\{K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)-\mathsf{E}\big[K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)\big]\Big\}\Big|>\frac{1}{3}(nh_1)^{\frac{1}{2}+\kappa_2}\Big) \\
\leq\ & \sum_{k=1}^{N_n(j)}\mathsf{P}\Big(\Big|\sum_{t=1}^n\Big\{K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)-\mathsf{E}\big[K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)\big]\Big\}\Big|>\frac{1}{3}(nh_1)^{\frac{1}{2}+\kappa_2}\Big). \qquad \text{(A.5)}
\end{aligned}$$

Then, by taking $Z_t=K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)-\mathsf{E}\big[K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)\big]$, $B=2\sup_u K(u)$, $p=(nh_1)^{\frac{1}{2}-\kappa_2}$ and $\epsilon=(nh_1)^{\frac{1}{2}+\kappa_2}/(3n)$ in Lemma 1 and noting that $h_1\sim n^{-\theta_1}$, we may show that

$$\begin{aligned}
& \mathsf{P}\Big(\Big|\sum_{t=1}^n\Big\{K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)-\mathsf{E}\big[K\big(\frac{X_{tj}-s_j(k)}{h_1}\big)\big]\Big\}\Big|>\frac{1}{3}(nh_1)^{\frac{1}{2}+\kappa_2}\Big) \\
\leq\ & 4\exp\left\{-c_1(nh_1)^{2\kappa_2}\right\}+c_3n^{3/2}(nh_1)^{(2\kappa_2-3)/4}\exp\left\{-c_2(nh_1)^{\frac{1}{2}-\kappa_2}\right\} \\
=\ & 4\exp\left\{-c_1n^{2(1-\theta_1)\kappa_2}\right\}+c_3n^{[2\kappa_2+3-(2\kappa_2-3)\theta_1]/4}\exp\left\{-c_2n^{(1-\theta_1)(\frac{1}{2}-\kappa_2)}\right\}, \qquad \text{(A.6)}
\end{aligned}$$

where c_1, c_2 and c_3 are positive constants which are independent of j. Combining (A.5) and (A.6) and using the definition of $N_n(j)$, we can prove (A.4), completing the proof of Lemma 2. ■

Lemma 3. *Let $\eta_{tj}=Y_t-m_j(X_{tj})$. Suppose that the conditions A1–A5 are satisfied. Then we have for any $\xi>0$ and $j=1,2,\dots,p_n+d_n$,*

$$\begin{aligned}
& \mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^n\eta_{tj}K\big(\frac{X_{tj}-x_j}{h_1}\big)\Big|>\xi(nh_1)n^{-\kappa_1}\Big) \\
\leq\ & M_2n^{1+\frac{7\kappa_1}{4}+\frac{5\theta_1}{2}}\left[\exp\left\{-c_4n^{1-2\kappa_1-\theta_1}\right\}+\exp\left\{-c_5n^{\kappa_1/2}\right\}\right], \qquad \text{(A.7)}
\end{aligned}$$

where $0 < \kappa_1 < (1-\theta_1)/2$, c_4, c_5 and M_2 are positive constants which are independent of j.

Proof. As $\mathsf{E}[\exp\{s|Y_t|\}] < \infty$ assumed in the condition A5, we may show that

$$\mathsf{E}[\exp\{s|\eta_{tj}|\}] \leq \mathsf{E}[\exp\{s|Y_t| + s|m_j(X_{tj})|\}] \leq e^{sc_m}\mathsf{E}[\exp\{s|Y_t|\}] < \infty. \tag{A.8}$$

Let

$$\nu_n = n^{\kappa_1/2}, \quad \overline{\eta}_{tj} = \eta_{tj}I\big(|\eta_{tj}| \leq \nu_n\big), \quad \widetilde{\eta}_{tj} = \eta_{tj}I\big(|\eta_{tj}| > \nu_n\big).$$

As $\mathsf{E}[\eta_{tj}] = 0$, it is easy to show that

$$\eta_{tj} = \eta_{tj} - \mathsf{E}[\eta_{tj}] = \overline{\eta}_{tj} - \mathsf{E}[\overline{\eta}_{tj}] + \widetilde{\eta}_{tj} - \mathsf{E}[\widetilde{\eta}_{tj}].$$

Hence, we have

$$\sum_{t=1}^{n}\eta_{tj}K\big(\frac{X_{tj}-x_j}{h_1}\big) = \sum_{t=1}^{n}\{\overline{\eta}_{tj} - \mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-x_j}{h_1}\big) + \sum_{t=1}^{n}\{\widetilde{\eta}_{tj} - \mathsf{E}[\widetilde{\eta}_{tj}]\}K\big(\frac{X_{tj}-x_j}{h_1}\big). \tag{A.9}$$

For sufficiently large k, by (A.8) and the choice of ν_n, we can prove that

$$\mathsf{E}[|\widetilde{\eta}_{tj}|] = \mathsf{E}[|\eta_{tj}|\mathsf{I}(|\eta_{tj}| > \nu_n)] \leq \mathsf{E}[|\eta_{tj}|^{k+1}\nu_n^{-k}] = O(\nu_n^{-k}) = o(h_1 n^{-\kappa_1}). \tag{A.10}$$

Then, we can show that

$$\begin{aligned}
& \mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^{n}\{\widetilde{\eta}_{tj} - \mathsf{E}[\widetilde{\eta}_{tj}]\}K\big(\frac{X_{tj}-x_j}{h_1}\big)\Big| > \frac{1}{2}\xi(nh_1)n^{-\kappa_1}\Big) \\
\leq\ & \mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^{n}\widetilde{\eta}_{tj}K\big(\frac{X_{tj}-x_j}{h_1}\big)\Big| > \frac{1}{4}\xi(nh_1)n^{-\kappa_1}\Big) \\
\leq\ & \mathsf{P}\Big(\max_{1\leq t\leq n}|\eta_{tj}| > \nu_n\Big) \leq \sum_{t=1}^{n}\mathsf{P}\Big(|\eta_{tj}| > \nu_n\Big) \\
\leq\ & n\frac{\mathsf{E}[\exp\{s|\eta_{tj}|\}]}{\exp\{s\nu_n\}} = M_2(n\exp\{-sn^{\kappa_1/2}\})/2,
\end{aligned} \tag{A.11}$$

where M_2 is a sufficiently large positive constant which is independent of j.

We next consider the upper bound for the probability of the event:

$$\Big\{\sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^{n}\{\overline{\eta}_{tj} - \mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-x_j}{h_1}\big)\Big| > \frac{1}{2}\xi(nh_1)n^{-\kappa_1}\Big\}.$$

The argument is similar to the proof of (A.4) above. We cover $\mathcal{C}_j$ by a finite number of intervals $\mathcal{C}_j^*(k)$ with centre $s_j^*(k)$ and radius $\xi h_1^2 n^{-\kappa_1}/(6c_K\nu_n)$, where c_K is defined as in the proof of Lemma 2.

Letting $N_n^*(j)$ be the total number of $\mathcal{C}_j^*(k)$, the order of $N_n^*(j)$ is $O\big(n^{\kappa_1}h_1^{-2}\nu_n\big)$. By some standard arguments, we have

$$
\begin{aligned}
& \sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^n\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-x_j}{h_1}\big)\Big| \\
\leq\ & \max_{1\leq k\leq N_n^*(j)}\Big|\sum_{t=1}^n\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-s_j^*(k)}{h_1}\big)\Big| + \\
& \max_{1\leq k\leq N_n^*(j)}\sup_{x_j\in\mathcal{C}_j^*(k)}\Big|\sum_{t=1}^n\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\}\Big[K\big(\frac{X_{tj}-x_j}{h_1}\big)-K\big(\frac{X_{tj}-s_j^*(k)}{h_1}\big)\Big]\Big| \\
\leq\ & \max_{1\leq k\leq N_n^*(j)}\Big|\sum_{t=1}^n\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-s_j^*(k)}{h_1}\big)\Big| + \frac{1}{3}\xi(nh_1)n^{-\kappa_1}.
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
& \mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^n\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-x_j}{h_1}\big)\Big| > \frac{1}{2}\xi(nh_1)n^{-\kappa_1}\Big) \\
\leq\ & \mathsf{P}\Big(\max_{1\leq k\leq N_n^*(j)}\Big|\sum_{t=1}^n\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-s_j^*(k)}{h_1}\big)\Big| > \frac{1}{6}\xi(nh_1)n^{-\kappa_1}\Big) \\
\leq\ & \sum_{k=1}^{N_n^*(j)}\mathsf{P}\Big(\Big|\sum_{t=1}^n\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-s_j^*(k)}{h_1}\big)\Big| > \frac{1}{6}\xi(nh_1)n^{-\kappa_1}\Big). \qquad \text{(A.12)}
\end{aligned}
$$

Then, by taking $Z_t=\big\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\big\}K\big(\frac{X_{tj}-s_j^*(k)}{h_1}\big)$, $B=2\nu_n\sup_u K(u)$, $p=n^{\kappa_1}/\nu_n=n^{\kappa_1/2}$ and $\epsilon=\xi h_1 n^{-\kappa_1}/6$ in Lemma 1 and noting that $h_1\sim n^{-\theta_1}$, we can show that

$$
\begin{aligned}
& \mathsf{P}\Big(\Big|\sum_{t=1}^n\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-s_j^*(k)}{h_1}\big)\Big| > \frac{1}{6}\xi(nh_1)n^{-\kappa_1}\Big) \\
\leq\ & 4\exp\big\{-c_4n^{1-2\kappa_1}h_1\big\}+c_6n^{1+\frac{\kappa_1}{4}}h_1^{-1/2}\exp\big\{-c_5n^{\kappa_1/2}\big\} \\
=\ & 4\exp\big\{-c_4n^{1-2\kappa_1-\theta_1}\big\}+c_6n^{1+\frac{\kappa_1}{4}+\frac{\theta_1}{2}}\exp\big\{-c_5n^{\kappa_1/2}\big\}, \qquad \text{(A.13)}
\end{aligned}
$$

where c_4, c_5 and c_6 are positive constants which are independent of j. By (A.12), (A.13) and the definition of $N_n^*(j)$, we can prove that

$$
\begin{aligned}
& \mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^n\{\overline{\eta}_{tj}-\mathsf{E}[\overline{\eta}_{tj}]\}K\big(\frac{X_{tj}-x_j}{h_1}\big)\Big| > \frac{1}{2}\xi(nh_1)n^{-\kappa_1}\Big) \\
\leq\ & \frac{M_2}{2}n^{1+\frac{7\kappa_1}{4}+\frac{5\theta_1}{2}}\left[\exp\big\{-c_4n^{1-2\kappa_1-\theta_1}\big\}+\exp\big\{-c_5n^{\kappa_1/2}\big\}\right]. \qquad \text{(A.14)}
\end{aligned}
$$

We can complete the proof of (A.7) by using (A.9), (A.11) and (A.14). ■

We next derive the upper bound for the probability of the event

$$\Big\{\sup_{x_j\in\mathcal{C}_j}|\hat{m}_j(x_j)-m_j(x_j)|>\xi n^{-2(1-\theta_1)/5}\Big\}$$

for any $\xi>0$, where Lemmas 2 and 3 will play a crucial role.

Lemma 4. *Suppose that the conditions A1–A5 in Section 3.1 are satisfied. Then we have for any* $\xi>0$ *and* $j=1,2,\ldots,p_n+d_n$,

$$\mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}|\hat{m}_j(x_j)-m_j(x_j)|>\xi n^{-2(1-\theta_1)/5}\Big)\leq M_1^*(n)+M_2^*(n), \tag{A.15}$$

where

$$M_1^*(n)=2M_1n^{(7+14\theta_1)/6}\exp\left\{-c_7n^{(1-\theta_1)/3}\right\},\quad M_2^*(n)=2M_2n^{(17+18\theta_1)/10}\exp\left\{-c_8n^{(1-\theta_1)/5}\right\},$$

$c_7=\min(c_1,c_2)$, $c_8=\min(c_4,c_5)$, M_1, c_1 *and* c_2 *are defined in Lemma 2, and* M_2, c_4 *and* c_5 *are defined in Lemma 3.*

Proof. Let $\mathcal{G}_j=\mathcal{G}_{j0}\cup\mathcal{G}_{j1}\cup\mathcal{G}_{j2}$ and the complement $\mathcal{G}_j^c=\mathcal{G}_{j0}^c\cap\mathcal{G}_{j1}^c\cap\mathcal{G}_{j2}^c$. Notice that

$$\begin{aligned}&\mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}|\hat{m}_j(x_j)-m_j(x_j)|>\xi n^{-\kappa_1}\Big)\\ \leq\ &\mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}|\hat{m}_j(x_j)-m_j(x_j)|>\xi n^{-\kappa_1},\mathcal{G}_j^c\Big)+\mathsf{P}\big(\mathcal{G}_j\big).\end{aligned}\tag{A.16}$$

By Lemma 2 with $\kappa_2=1/6$, we may show that

$$\mathsf{P}\big(\mathcal{G}_j\big)\leq 2M_1n^{(7+14\theta_1)/6}\exp\left\{-c_7n^{(1-\theta_1)/3}\right\}=:M_1^*(n). \tag{A.17}$$

Consider the decomposition:

$$\begin{aligned}\hat{m}_j(x_j)-m_j(x_j)\ &=\ \frac{\sum_{t=1}^n\left[Y_t-m_j(x_j)\right]K\left(\frac{X_{tj}-x_j}{h_1}\right)}{\sum_{t=1}^nK\left(\frac{X_{tj}-x_j}{h_1}\right)}\\ &=\ \frac{\sum_{t=1}^n\left[Y_t-m_j(X_{tj})\right]K\left(\frac{X_{tj}-x_j}{h_1}\right)}{\sum_{t=1}^nK\left(\frac{X_{tj}-x_j}{h_1}\right)}+\frac{\sum_{t=1}^n\left[m_j(X_{tj})-m_j(x_j)\right]K\left(\frac{X_{tj}-x_j}{h_1}\right)}{\sum_{t=1}^nK\left(\frac{X_{tj}-x_j}{h_1}\right)}\\ &=:\ I_{n1}(x_j)+I_{n2}(x_j).\end{aligned}\tag{A.18}$$

By the condition A4 and Taylor's expansion for $m_j(\cdot)$, we have

$$m_j(X_{tj})-m_j(x_j)=m'(x_j)(X_{tj}-x_j)+\frac{1}{2}m_j''(x_{tj})(X_{tj}-x_j)^2,$$

where x_{tj} lies between X_{tj} and x_j. Hence, for $I_{n2}(x_j)$, we have

$$I_{n2}(x_j) \;=\; m'(x_j)\frac{\sum_{t=1}^n (X_{tj}-x_j)K\big(\frac{X_{tj}-x_j}{h_1}\big)}{\sum_{t=1}^n K\big(\frac{X_{tj}-x_j}{h_1}\big)} + \frac{1}{2}\cdot\frac{\sum_{t=1}^n m_j''(x_{tj})(X_{tj}-x_j)^2 K\big(\frac{X_{tj}-x_j}{h_1}\big)}{\sum_{t=1}^n K\big(\frac{X_{tj}-x_j}{h_1}\big)}.$$

On the event $\mathcal{G}_j^c$ with $\kappa_2 = 1/6$, as $\theta_1 > 1/6$ and choosing κ_1 as $2(1-\theta_1)/5$,

$$I_{n2}(x_j) = O\big(h_1^2 + (nh_1)^{-1/3}h_1\big) = o(n^{-\kappa_1}). \tag{A.19}$$

Hence, we have

$$\mathsf{P}\big(\sup_{x_j\in\mathcal{C}_j} |\hat{m}_j(x_j) - m_j(x_j)| > \xi n^{-\kappa_1}, \mathcal{G}_j^c\big) \le \mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^n \eta_{tj}K\big(\frac{X_{tj}-x_j}{h_1}\big)\Big| > \xi_1(nh_1)n^{-\kappa_1}\Big), \tag{A.20}$$

where $\xi_1 = \frac{1}{2}\xi \inf_{x_j\in\mathcal{C}_j} f_j(x_j)$. By Lemma 3 with $\kappa_1 = 2(1-\theta_1)/5$, we have

$$\mathsf{P}\Big(\sup_{x_j\in\mathcal{C}_j}\Big|\sum_{t=1}^n \eta_{tj}K\big(\frac{X_{tj}-x_j}{h_1}\big)\Big| > \xi_1(nh_1)n^{-\kappa_1}\Big) \le 2M_2 n^{(17+18\theta_1)/10}\exp\left\{-c_8 n^{(1-\theta_1)/5}\right\},$$

with $c_8 = \min(c_4, c_5)$, which indicates that

$$\mathsf{P}\big(\sup_{x_j\in\mathcal{C}_j} |\hat{m}_j(x_j) - m_j(x_j)| > \xi n^{-\kappa_1}, \mathcal{G}_j^c\big) \le 2M_2 n^{(17+18\theta_1)/10}\exp\left\{-c_8 n^{(1-\theta_1)/5}\right\} =: M_2^*(n). \tag{A.21}$$

We can complete the proof of (A.15) by combining (A.16), (A.17) and (A.21). ■

Lemma 5. *Suppose that the conditions B1–B4 and (3.11) in Section 3.2 are satisfied. Then, we have*

$$\left\|\boldsymbol{d}_n(\hat{\mathcal{F}}_n)\right\| := \left\|\mathsf{vec}\left(\mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0\right)/\sqrt{n}\right\| = o_P(1), \tag{A.22}$$

where $\mathsf{vec}(\cdot)$ *denotes the vectorization of a matrix.*

Proof. It is easy to see that the PCA method is equivalent to the following constrained least squares method:

$$\left(\hat{\mathcal{F}}_n, \hat{\mathbf{B}}_n\right) = \arg\min_{\mathbf{b}_k,\mathbf{f}_t}\sum_{k=1}^{p_n}\sum_{t=1}^n \left(Z_{tk} - \mathbf{b}_k^\intercal \mathbf{f}_t\right)^2 = \arg\min_{\mathcal{F}_n,\mathbf{B}_n}\left\|\mathcal{Z}_n - \mathcal{F}_n\mathbf{B}_n^\intercal\right\|_F^2, \tag{A.23}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix, and the $n\times r$ matrix $\mathcal{F}_n$ and the $p_n\times r$ matrix $\mathbf{B}_n$ need to satisfy

$$\frac{1}{n}\mathcal{F}_n^\intercal\mathcal{F}_n = I_r,\quad \frac{1}{p_n}\mathbf{B}_n^\intercal\mathbf{B}_n \text{ is diagonal.} \tag{A.24}$$

Denote $\mathcal{M}_{\mathcal{F}_n} = I_n - \mathcal{F}_n\left(\mathcal{F}_n^\intercal\mathcal{F}_n\right)^{-1}\mathcal{F}_n^\intercal =: I_n - \mathcal{P}_{\mathcal{F}_n}$ and $\mathcal{L}_n(\mathcal{F}_n) = \mathsf{Tr}\left(\mathcal{Z}_n^\intercal\mathcal{M}_{\mathcal{F}_n}\mathcal{Z}_n\right)$, where $\mathsf{Tr}(\cdot)$ is the trace of a square matrix.

From (A.23), we may show that

$$\mathcal{L}_n(\hat{\mathcal{F}}_n) - \mathcal{L}_n(\mathcal{F}_n^0) = \mathsf{Tr}\left(\mathcal{Z}_n^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n} \mathcal{Z}_n\right) - \mathsf{Tr}\left(\mathcal{Z}_n^\intercal \mathcal{M}_{\mathcal{F}_n^0} \mathcal{Z}_n\right) \leq 0. \tag{A.25}$$

Using the fact that $\mathcal{M}_{\mathcal{F}_n^0}\mathcal{F}_n^0 = \mathbf{0}$ and $(\mathcal{F}_n^0)^\intercal \mathcal{M}_{\mathcal{F}_n^0} = \mathbf{0}$, and then by (2.9), we have

$$\begin{aligned}\mathcal{L}_n(\hat{\mathcal{F}}_n) - \mathcal{L}_n(\mathcal{F}_n^0) &= \mathsf{Tr}\left(\mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal \mathbf{B}_n^0(\mathcal{F}_n^0)^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\right) + \mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{U}_n\right) - \\ &\quad \mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{M}_{\mathcal{F}_n^0}\mathcal{U}_n\right) + \mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\right) + \\ &\quad \mathsf{Tr}\left(\mathbf{B}_n^0(\mathcal{F}_n^0)^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{U}_n\right) \\ &= \mathsf{Tr}\left(\mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal \mathbf{B}_n^0(\mathcal{F}_n^0)^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\right) + \mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{P}_{\hat{\mathcal{F}}_n}\mathcal{U}_n\right) - \\ &\quad \mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{P}_{\mathcal{F}_n^0}\mathcal{U}_n\right) + \mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\right) + \\ &\quad \mathsf{Tr}\left(\mathbf{B}_n^0(\mathcal{F}_n^0)^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{U}_n\right),\end{aligned} \tag{A.26}$$

where $\mathcal{U}_n = (\mathbf{U}_1, \ldots, \mathbf{U}_n)^\intercal$.

We next prove that the last four terms on the right hand side of (A.26) are $o_P(np_n)$. Start with $\mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\right)$. Note that

$$\mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\right) = \mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\right) - \frac{1}{n}\mathsf{Tr}\left(\mathcal{U}_n^\intercal \hat{\mathcal{F}}_n\hat{\mathcal{F}}_n^\intercal \mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\right) \tag{A.27}$$

using the restriction of $\hat{\mathcal{F}}_n^\intercal \hat{\mathcal{F}}_n/n = I_r$. By the conditions B2 and B4, the Cauchy-Schwarz inequality and some standard arguments, we may show that

$$\begin{aligned}\mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\right) &= \sum_{t=1}^{n}\sum_{k=1}^{p_n} u_{tk}(\mathbf{f}_t^0)^\intercal \mathbf{b}_k^0 = \left(\sum_{t=1}^{n}\left\|\mathbf{f}_t^0\right\|^2\right)^{1/2}\left(\sum_{t=1}^{n}\left\|\sum_{k=1}^{p_n} u_{tk}\mathbf{b}_k^0\right\|^2\right)^{1/2} \\ &= O_P(n^{1/2})\cdot O_P(n^{3/4}p_n^{1/2}) = O_P(n^{5/4}p_n^{1/2}).\end{aligned} \tag{A.28}$$

On the other hand, by some similar calculations and using the fact that $\|\hat{\mathcal{F}}_n^\intercal \mathcal{F}_n^0\|_F = O_P(n)$, we can also prove that

$$\begin{aligned}\frac{1}{n}\mathsf{Tr}\left(\mathcal{U}_n^\intercal \hat{\mathcal{F}}_n\hat{\mathcal{F}}_n^\intercal \mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\right) &\leq C\left(\sum_{t=1}^{n}\left\|\mathbf{f}_t^0\right\|^2\right)^{1/2}\left(\sum_{t=1}^{n}\left\|\sum_{k=1}^{p_n} u_{tk}\mathbf{b}_k^0\right\|^2\right)^{1/2} \\ &= O_P(n^{5/4}p_n^{1/2}),\end{aligned} \tag{A.29}$$

where C is a positive constant whose value may change from line to line. By (A.27)–(A.29) and using the condition of $n = o(p_n^2)$, we have

$$\mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\right) = O_P(n^{5/4}p_n^{1/2}) = o_P(np_n). \tag{A.30}$$

Analogously, we may also show that

$$\mathsf{Tr}\left(\mathbf{B}_n^0(\mathcal{F}_n^0)^{\intercal}\mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{U}_n\right) = o_P(np_n). \tag{A.31}$$

We next consider $\mathsf{Tr}\left(\mathcal{U}_n^{\intercal}\mathcal{P}_{\hat{\mathcal{F}}_n}\mathcal{U}_n\right)$. Note that

$$\begin{aligned}\mathsf{Tr}\left(\mathcal{U}_n^{\intercal}\mathcal{P}_{\hat{\mathcal{F}}_n}\mathcal{U}_n\right) &= \frac{1}{n}\sum_{k=1}^{p_n}\sum_{t_1=1}^{n}\sum_{t_2=1}^{n}\mathsf{Tr}(\hat{\mathbf{f}}_{t_1}\hat{\mathbf{f}}_{t_2}^{\intercal})u_{t_1k}u_{t_2k} \\ &= \frac{1}{n}\sum_{k=1}^{p_n}\sum_{t_1=1}^{n}\sum_{t_2=1}^{n}\mathsf{Tr}(\hat{\mathbf{f}}_{t_1}\hat{\mathbf{f}}_{t_2}^{\intercal})\mathsf{E}\left[u_{t_1k}u_{t_2k}\right] + \\ &\quad \frac{1}{n}\sum_{k=1}^{p_n}\sum_{t_1=1}^{n}\sum_{t_2=1}^{n}\mathsf{Tr}(\hat{\mathbf{f}}_{t_1}\hat{\mathbf{f}}_{t_2}^{\intercal})\left(u_{t_1k}u_{t_2k} - \mathsf{E}\left[u_{t_1k}u_{t_2k}\right]\right), \end{aligned} \tag{A.32}$$

where we again have used the fact of $\hat{\mathcal{F}}_n^{\intercal}\hat{\mathcal{F}}_n/n = I_r$. For the first term on the right hand side of (A.32), by the conditions B1 and B4, and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\left|\sum_{k=1}^{p_n}\sum_{t_1=1}^{n}\sum_{t_2=1}^{n}\mathsf{Tr}(\hat{\mathbf{f}}_{t_1}\hat{\mathbf{f}}_{t_2}^{\intercal})\mathsf{E}\left[u_{t_1k}u_{t_2k}\right]\right| \\ \leq\ & Cp_n\left(\sum_{t_1=1}^{n}\left\|\hat{\mathbf{f}}_{t_1}\right\|^2\sum_{t_2=1}^{n}\left\|\hat{\mathbf{f}}_{t_2}\right\|^2\right)^{1/2}\left(\sum_{t_1=1}^{n}\sum_{t_1=1}^{n}\alpha^{3/2}(|t_1-t_2|)\right)^{1/2} \\ =\ & p_n\cdot O_P(n)\cdot O_P(n^{1/2}).\end{aligned} \tag{A.33}$$

For the second term on the right hand side of (A.32), letting $u(t_1,t_2) = \sum_{k=1}^{p_n}\left(u_{t_1k}u_{t_2k} - \mathsf{E}\left[u_{t_1k}u_{t_2k}\right]\right)$, by (3.10) in the condition B4 and the Cauchy-Schwarz inequality again, we have

$$\begin{aligned}&\left|\sum_{t_1=1}^{n}\sum_{t_2=1}^{n}\mathsf{Tr}(\hat{\mathbf{f}}_{t_1}\hat{\mathbf{f}}_{t_2}^{\intercal})\sum_{k=1}^{p_n}\left(u_{t_1k}u_{t_2k} - \mathsf{E}\left[u_{t_1k}u_{t_2k}\right]\right)\right| \\ \leq\ & \left(\sum_{t_1=1}^{n}\left\|\hat{\mathbf{f}}_{t_1}\right\|^2\sum_{t_2=1}^{n}\left\|\hat{\mathbf{f}}_{t_2}\right\|^2\right)^{1/2}\left(\sum_{t_1=1}^{n}\sum_{t_1=1}^{n}u^2(|t_1-t_2|)\right)^{1/2} \\ =\ & O_P(n)\cdot O_P(np_n^{1/2}n^{1/4}) = O_P(n^{9/4}p_n^{1/2}).\end{aligned} \tag{A.34}$$

In view of (A.32)–(A.34), we have

$$\mathsf{Tr}\left(\mathcal{U}_n^{\intercal}\mathcal{P}_{\hat{\mathcal{F}}_n}\mathcal{U}_n\right) = O_P(n^{5/4}p_n^{1/2} + n^{1/2}p_n) = o_P(np_n). \tag{A.35}$$

We finally consider $\mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{P}_{\mathcal{F}_n^0}\mathcal{U}_n\right)$. By the conditions B1 and B2 as well as the central limit theorem, we have

$$\frac{1}{n}\sum_{t=1}^n \mathbf{f}_t^0(\mathbf{f}_t^0)^\intercal - \mathbf{\Lambda}_F = O_P(n^{-1/2}),$$

which indicates that

$$\begin{aligned}\mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{P}_{\mathcal{F}_n^0}\mathcal{U}_n\right) &= \frac{1}{n}\sum_{k=1}^{p_n}\sum_{t_1=1}^n\sum_{t_2=1}^n \mathsf{Tr}(\mathbf{f}_{t_1}^0\mathbf{\Lambda}_F^{-1}(\mathbf{f}_{t_2}^0)^\intercal)u_{t_1k}u_{t_2k} + O_P(n^{-3/2})\sum_{k=1}^{p_n}\left\|\sum_{t_1=1}^n \mathbf{f}_{t_1}^0 u_{t_1k}\right\|\left\|\sum_{t_2=1}^n \mathbf{f}_{t_2}^0 u_{t_2k}\right\| \\ &= \frac{1}{n}\sum_{k=1}^{p_n}\sum_{t_1=1}^n\sum_{t_2=1}^n \mathsf{Tr}(\mathbf{f}_{t_1}^0\mathbf{\Lambda}_F^{-1}(\mathbf{f}_{t_2}^0)^\intercal)u_{t_1k}u_{t_2k} + O_P(p_n n^{1/2}) \\ &= \frac{1}{n}\sum_{k=1}^{p_n}\sum_{t_1=1}^n\sum_{t_2=1}^n \mathsf{Tr}(\mathbf{f}_{t_1}^0\mathbf{\Lambda}_F^{-1}(\mathbf{f}_{t_2}^0)^\intercal)u_{t_1k}u_{t_2k} + o_P(np_n), \qquad (A.36)\end{aligned}$$

where we have used the result that

$$\max_{1\le k\le p_n}\left\|\sum_{t=1}^n \mathbf{f}_t^0 u_{tk}\right\| = O_P(n), \qquad (A.37)$$

which can be proved by using the exponential inequality in Lemma 1 and the arguments in Lemmas 2 and 3. Following the arguments in the proof of (A.35), we can similarly show that

$$\frac{1}{n}\sum_{k=1}^{p_n}\sum_{t_1=1}^n\sum_{t_2=1}^n \mathsf{Tr}(\mathbf{f}_{t_1}^0\mathbf{\Lambda}_F^{-1}(\mathbf{f}_{t_2}^0)^\intercal)u_{t_1k}u_{t_2k} = O_P(n^{5/4}p_n^{1/2} + n^{1/2}p_n), \qquad (A.38)$$

which together with (A.36), implies that

$$\mathsf{Tr}\left(\mathcal{U}_n^\intercal \mathcal{P}_{\mathcal{F}_n^0}\mathcal{U}_n\right) = o_P(np_n). \qquad (A.39)$$

Hence, by (A.30), (A.31), (A.35) and (A.39), we have

$$\frac{1}{np_n}\left[\mathcal{L}_n(\hat{\mathcal{F}}_n) - \mathcal{L}_n(\mathcal{F}_n^0)\right] = \frac{1}{np_n}\mathsf{Tr}\left(\mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\mathbf{B}_n^0(\mathcal{F}_n^0)^\intercal\mathcal{M}_{\hat{\mathcal{F}}_n}\right) + o_P(1). \qquad (A.40)$$

Define

$$\mathbf{\Sigma}_n(\mathbf{B}_n^0) = \frac{1}{p_n}(\mathbf{B}_n^0)^\intercal\mathbf{B}_n^0 \otimes I_n, \quad \boldsymbol{d}_n(\hat{\mathcal{F}}_n) = \mathsf{vec}\left(\mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0\right)/\sqrt{n},$$

where $\otimes$ denotes the Kronecker product. It is easy to verify that

$$\frac{1}{np_n}\mathsf{Tr}\left(\mathcal{M}_{\hat{\mathcal{F}}_n}\mathcal{F}_n^0(\mathbf{B}_n^0)^\intercal\mathbf{B}_n^0(\mathcal{F}_n^0)^\intercal\mathcal{M}_{\hat{\mathcal{F}}_n}\right) = \mathbf{d}_n^\intercal(\hat{\mathcal{F}}_n)\mathbf{\Sigma}_n(\mathbf{B}_n^0)\mathbf{d}_n(\hat{\mathcal{F}}_n) + o_P(1). \qquad (A.41)$$

By the condition B3, the smallest eigenvalue of $\mathbf{\Sigma}_n(\mathbf{B}_n^0)$ is positive and bounded away from zero. Therefore we can prove that

$$0 \leq \mathbf{d}_n^{\intercal}(\hat{\mathcal{F}}_n)\mathbf{\Sigma}_n(\mathbf{B}_n^0)\mathbf{d}_n(\hat{\mathcal{F}}_n) = o_P(1), \tag{A.42}$$

which leads to (A.22), completing the proof of Lemma 5. ■

Appendix B: Proofs of the main results

In this appendix, we provide the detailed proofs of the asymptotic results given in Section 3.

Proof of Theorem 1 (i) By the definition of $\hat{\mathsf{v}}(j)$, we have for $j = 1, 2, \ldots, p_n + d_n$,

$$\begin{aligned}\hat{\mathsf{v}}(j) - \mathsf{v}(j) &= \frac{1}{n}\sum_{t=1}^n \hat{m}_j^2(X_{tj}) - \Big[\frac{1}{n}\sum_{t=1}^n \hat{m}_j(X_{tj})\Big]^2 - \mathsf{var}\big(m_j(X_{tj})\big) \\ &= \Big\{\frac{1}{n}\sum_{t=1}^n \big[\hat{m}_j^2(X_{tj}) - m_j^2(X_{tj})\big]\Big\} - \Big\{\Big[\frac{1}{n}\sum_{t=1}^n \hat{m}_j(X_{tj})\Big]^2 - \Big[\frac{1}{n}\sum_{t=1}^n m_j(X_{tj})\Big]^2\Big\} + \\ &\quad \Big\{\frac{1}{n}\sum_{t=1}^n \big(m_j^2(X_{tj}) - \mathsf{E}[m_j^2(X_{tj})]\big)\Big\} - \Big\{\Big[\frac{1}{n}\sum_{t=1}^n m_j(X_{tj})\Big]^2 - \mathsf{E}^2\big[m_j(X_{tj})\big]\Big\} \\ &=: \Pi_{nj}(1) + \Pi_{nj}(2) + \Pi_{nj}(3) + \Pi_{nj}(4).\end{aligned} \tag{B.1}$$

For $\Pi_{nj}(1)$, note that

$$\begin{aligned}\big|\Pi_{nj}(1)\big| &= \frac{1}{n}\sum_{t=1}^n \big|\hat{m}_j^2(X_{tj}) - m_j^2(X_{tj})\big| \\ &= \frac{1}{n}\sum_{t=1}^n \big[\hat{m}_j(X_{tj}) - m_j(X_{tj})\big]^2 + \frac{2}{n}\sum_{t=1}^n \big|m_j(X_{tj})\big|\big|\hat{m}_j(X_{tj}) - m_j(X_{tj})\big| \\ &\leq \sup_{x_j\in\mathcal{C}_j}\big|\hat{m}_j(x_j) - m_j(x_j)\big|^2 + 2c_m \cdot \sup_{x_j\in\mathcal{C}_j}\big|\hat{m}_j(x_j) - m_j(x_j)\big|.\end{aligned} \tag{B.2}$$

By (B.2) and Lemma 4, we readily obtain

$$\mathsf{P}\Big(\big|\Pi_{nj}(1)\big| > \frac{\delta_1}{4}n^{-2(1-\theta_1)/5}\Big) \leq 2M_2(n)\exp\big\{-c_8 n^{(1-\theta_1)/5}\big\}, \tag{B.3}$$

where $M_2(n) = M_2 n^{(17+18\theta_1)/10}$ and c_8 is defined in Lemma 4. Analogously, we can also show that

$$\mathsf{P}\Big(\big|\Pi_{nj}(2)\big| > \frac{\delta_1}{4}n^{-2(1-\theta_1)/5}\Big) \leq 2M_2(n)\exp\big\{-c_8 n^{(1-\theta_1)/5}\big\}. \tag{B.4}$$

Using Lemma 1 with $Z_t = m_j(X_{tj})$ or $m_j^2(X_{tj})$, we may show that

$$\mathsf{P}\Big(\big|\Pi_{nj}(3)\big| > \frac{\delta_1}{4}n^{-2(1-\theta_1)/5}\Big) + \mathsf{P}\Big(\big|\Pi_{nj}(4)\big| > \frac{\delta_1}{4}n^{-2(1-\theta_1)/5}\Big) = o\big(M_2(n)\exp\left\{-c_8 n^{(1-\theta_1)/5}\right\}\big). \quad \text{(B.5)}$$

Then, by (B.1) and (B.3)–(B.5), we can prove that

$$\mathsf{P}\Big(\big|\hat{\mathsf{v}}(j) - \mathsf{v}(j)\big| > \delta_1 n^{-2(1-\theta_1)/5}\Big) \leq 5M_2(n)\exp\left\{-c_8 n^{(1-\theta_1)/5}\right\}, \quad \text{(B.6)}$$

which indicates that

$$\begin{aligned} & \mathsf{P}\Big(\max_{1\leq j\leq p_n+d_n}\big|\hat{\mathsf{v}}(j) - \mathsf{v}(j)\big| > \delta_1 n^{-2(1-\theta_1)/5}\Big) \\ \leq & \sum_{j=1}^{p_n+d_n}\mathsf{P}\Big(\big|\hat{\mathsf{v}}(j) - \mathsf{v}(j)\big| > \delta_1 n^{-2(1-\theta_1)/5}\Big) \\ \leq & \ O\left((p_n+d_n)M_2(n)\exp\left\{-c_8 n^{(1-\theta_1)/5}\right\}\right). \quad \text{(B.7)} \end{aligned}$$

Choosing $M(n) = (p_n + d_n)n^{(17+18\theta_1)/10}$ and $\delta_2 = c_8$, we can complete the proof of Theorem 1(i).

(ii) By the definition of $\hat{\mathcal{S}}$, using the condition that $\min_{j\in\mathcal{S}}\mathsf{v}(j) \geq 2\delta_1 n^{-2(1-\delta_1)/5}$ and following the proof of Theorem 1(i), we have

$$\begin{aligned} \mathsf{P}\big(\mathcal{S}\subset\hat{\mathcal{S}}\big) &= \mathsf{P}\big(\min_{j\in\mathcal{S}}\hat{\mathsf{v}}(j) \geq \rho_n\big) = \mathsf{P}\big(\min_{j\in\mathcal{S}}\hat{\mathsf{v}}(j) \geq \delta_1 n^{-2(1-\theta_1)/5}\big) \\ &= \mathsf{P}\big(\min_{j\in\mathcal{S}}\mathsf{v}(j) - \min_{j\in\mathcal{S}}\hat{\mathsf{v}}(j) \leq \min_{j\in\mathcal{S}}\mathsf{v}(j) - \delta_1 n^{-2(1-\theta_1)/5}\big) \\ &\geq \mathsf{P}\big(\min_{j\in\mathcal{S}}\mathsf{v}(j) - \min_{j\in\mathcal{S}}\hat{\mathsf{v}}(j) \leq 2\delta_1 n^{-2(1-\theta_1)/5} - \delta_1 n^{-2(1-\theta_1)/5}\big) \\ &\geq \mathsf{P}\big(\max_{j\in\mathcal{S}}\big|\hat{\mathsf{v}}(j) - \mathsf{v}(j)\big| \leq \delta_1 n^{-2(1-\theta_1)/5}\big) \\ &= 1 - \mathsf{P}\big(\max_{j\in\mathcal{S}}\big|\hat{\mathsf{v}}(j) - \mathsf{v}(j)\big| > \delta_1 n^{-2(1-\theta_1)/5}\big) \\ &\geq 1 - O\big(M(n)\exp\big\{-\delta_2 n^{(1-\theta_1)/5}\big\}\big). \quad \text{(B.8)} \end{aligned}$$

Then, we complete the proof of (3.3).

The proof of Theorem 1 has been completed. ■

Proof of Theorem 2 (i) Recall that $\mathbf{w}_n = (w_1, \dots, w_{q_n})^\intercal$ and $\mathbf{w}_o = (w_{o1}, \dots, w_{oq_n})^\intercal = \big[\mathbf{w}_o^\intercal(1), \mathbf{w}_o^\intercal(2)\big]^\intercal$, 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 $(q_n - s_n)$. Let $\mathcal{Q}_n(\cdot)$ be defined as in (2.6) and $\epsilon_n = \sqrt{q_n}\big[(\sqrt{n}\chi_n)^{-1} + a_n\big]$. In order to prove the convergence rate in Theorem 2, as in Fan and Peng (2004), it suffices to show that there exists a sufficiently large constant $C > 0$ such that

$$\lim_{n\to\infty}\mathsf{P}\left(\inf_{\|\mathbf{u}\|=C}\mathcal{Q}_n(\mathbf{w}_o + \epsilon_n\mathbf{u}) > \mathcal{Q}_n(\mathbf{w}_o)\right) = 1, \quad \text{(B.9)}$$

where $\mathbf{u} = (u_1, \ldots, u_{q_n})^{\intercal}$. In fact, (B.9) implies that there exists a minimum $\hat{\mathbf{w}}_n$ in the ball $\{\mathbf{w}_o + \epsilon_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$, such that $\|\hat{\mathbf{w}}_n - \mathbf{w}_o\| = O_P(\epsilon_n)$.

Observe that

$$\begin{aligned}
& \mathcal{Q}_n(\mathbf{w}_o + \epsilon_n \mathbf{u}) - \mathcal{Q}_n(\mathbf{w}_o) \\
= \ & \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o + \epsilon_n \mathbf{u})\big]^{\intercal} \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o + \epsilon_n \mathbf{u})\big] + n \sum_{j=1}^{q_n} p_\lambda(|w_{oj} + \epsilon_n u_j|) \\
& - \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o)\big]^{\intercal} \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o)\big] - n \sum_{j=1}^{q_n} p_\lambda(|w_{oj}|) \\
\geq \ & \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o + \epsilon_n \mathbf{u})\big]^{\intercal} \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o + \epsilon_n \mathbf{u})\big] - \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o)\big]^{\intercal} \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o)\big] \\
& + n \sum_{j=1}^{s_n} p_\lambda(|w_{oj} + \epsilon_n u_j|) - n \sum_{j=1}^{s_n} p_\lambda(|w_{oj}|) \\
= \ & \Xi_{n1} + \Xi_{n2}, \qquad \text{(B.10)}
\end{aligned}$$

where

$$\begin{aligned}
\Xi_{n1} &= \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o + \epsilon_n \mathbf{u})\big]^{\intercal} \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o + \epsilon_n \mathbf{u})\big] - \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o)\big]^{\intercal} \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o)\big], \\
\Xi_{n2} &= n \sum_{j=1}^{s_n} \big[p_\lambda(|w_{oj} + \epsilon_n u_j|) - p_\lambda(|w_{oj}|)\big].
\end{aligned}$$

By the definition of $\hat{\mathcal{M}}(\cdot)$ in Section 2.1 and some elementary calculations, we have

$$\begin{aligned}
\Xi_{n1} &= -2\epsilon_n \mathbf{u}^{\intercal} \mathcal{S}_n^{\intercal}(\mathcal{Y}) \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o)\big] + \epsilon_n^2 \mathbf{u}^{\intercal} \mathcal{S}_n^{\intercal}(\mathcal{Y}) \mathcal{S}_n(\mathcal{Y}) \mathbf{u} \\
&=: \Xi_{n1}(1) + \Xi_{n1}(2).
\end{aligned}$$

Following the proof of Theorem 3.3 in Li, Linton and Lu (2015), we can show that

$$\big\| \mathcal{S}_n^{\intercal}(\mathcal{Y}) \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_o)\big] \big\| = O_P(\sqrt{n q_n}),$$

which indicates that

$$|\Xi_{n1}(1)| = O_P(\epsilon_n \sqrt{n q_n}) \cdot \|\mathbf{u}\|. \qquad \text{(B.11)}$$

We next consider $\Xi_{n1}(2)$. By the definition of $m_j^*(\cdot)$ in Section 1 and the uniform consistency result in Theorem 3.1 of Li, Lu and Linton (2012), we have, uniformly for x_j and $j = 1, 2, \ldots, q_n$,

$$\hat{m}_j^*(x_j) - m_j^*(x_j) = O_P(\tau_n + h_2^2), \qquad \text{(B.12)}$$

where τ_n is defined in Section 3.1. Observe that

$$\mathcal{S}_n^{\intercal}(\mathcal{Y})\mathcal{S}_n(\mathcal{Y}) = \mathcal{M}_n^{\intercal}\mathcal{M}_n + \big(\mathcal{S}_n(\mathcal{Y}) - \mathcal{M}_n\big)^{\intercal}\mathcal{M}_n + \mathcal{M}_n^{\intercal}\big(\mathcal{S}_n(\mathcal{Y}) - \mathcal{M}_n\big) + \big(\mathcal{S}_n(\mathcal{Y}) - \mathcal{M}_n\big)^{\intercal}\big(\mathcal{S}_n(\mathcal{Y}) - \mathcal{M}_n\big), \tag{B.13}$$

where $\mathcal{M}_n = \big[\mathcal{M}(1), \dots, \mathcal{M}(q_n)\big]$ and $\mathcal{M}(j)$, $j = 1, 2, \dots, q_n$, are defined in Section 2.1. By (3.4) in the condition A7, there exists a sequence $\{\chi_n^*\}$ such that $\chi_n^* = o(\chi_n)$ and $q_n = o(\sqrt{n}\chi_n^*)$. Then, for any $\xi > 0$, by Chebyshev's inequality and following the proof of Lemma 8 in Fan and Peng (2004), we have

$$\mathsf{P}\left(\left\|\frac{1}{n}\mathcal{M}_n^{\intercal}\mathcal{M}_n - \mathbf{\Lambda}_n\right\|_F > \xi\chi_n^*\right) \leq \frac{\mathsf{E}\big[\|\mathcal{G}_n^{\intercal}\mathcal{G}_n - \mathbf{\Lambda}_n\|_F^2\big]}{\xi^2 n^2 (\chi_n^*)^2} = O\left(\frac{q_n^2}{n(\chi_n^*)^2}\right) = o(1).$$

Hence, we have

$$\left\|\frac{1}{n}\mathcal{M}_n^{\intercal}\mathcal{M}_n - \mathbf{\Lambda}_n\right\|_F = o_P(\chi_n^*). \tag{B.14}$$

Equation (B.14) and the fact of $\chi_n^* = o(\chi_n)$ imply that the smallest eigenvalue of $\mathcal{M}_n^{\intercal}\mathcal{M}_n/n$ is larger than $\chi_n/2$ with probability approaching one. As $q_n^2(\tau_n + h_2^2) = o(\chi_n)$ in the condition A7, we can easily prove that the Frobenius norm for the last three matrices on the right hand side of (B.13) tend to zero with convergence rates faster than χ_n. Hence, we have

$$\Xi_{n1}(2) \geq \frac{n\epsilon_n^2\chi_n}{2} \cdot \|\mathbf{u}\|^2 \tag{B.15}$$

in probability. By (B.11), (B.15) and taking the constant C sufficiently large, $\Xi_{n1}(1)$ would be dominated by $\Xi_{n1}(2)$ asymptotically.

For Ξ_{n2}, by the condition A8 and Taylor's expansion for the penalty function, we have

$$\begin{aligned} \Xi_{n2} &= n\sum_{j=1}^{s_n}\big[p_\lambda(|w_{oj} + \epsilon_n u_j|) - p_\lambda(|w_{oj}|)\big] \\ &= O_P(n\epsilon_n a_n\sqrt{q_n}) \cdot \|\mathbf{u}\| + O_P(n\epsilon_n^2 b_n) \cdot \|\mathbf{u}\|^2, \end{aligned} \tag{B.16}$$

where w_{oj}^* lies between w_{oj} and $w_{oj} + \epsilon_n u_j$. By the condition A8 , Ξ_{n2} would be also dominated by $\Xi_{n1}(2)$ by taking the constant C sufficiently large. We thus complete the proof of (B.9) in view of (B.10), (B.11), (B.15) and (B.16).

(ii) Let $\hat{\mathbf{w}}_n(1)$ and $\hat{\mathbf{w}}_n(2)$ be the estimators of $\mathbf{w}_o(1)$ and $\mathbf{w}_o(2)$, respectively. To prove Theorem 2(ii), it suffices to show that for any constant C and any given $\mathbf{w}_n(1)$ satisfying $\|\mathbf{w}_n(1) - \mathbf{w}_o(1)\| = O_P(\epsilon_{n*})$, we have

$$\mathcal{Q}_n\big(\big[\mathbf{w}_n^{\intercal}(1), \mathbf{0}^{\intercal}\big]^{\intercal}\big) = \min_{\|\mathbf{w}_n(2)\| \leq C\epsilon_{n*}} \mathcal{Q}_n\big(\big[\mathbf{w}_n^{\intercal}(1), \mathbf{w}_n^{\intercal}(2)\big]^{\intercal}\big), \tag{B.17}$$

where $\epsilon_{n*} = \frac{\sqrt{q_n}}{\sqrt{n}\chi_n}$, $\mathbf{w}_n(2)$ is a $(q_n - s_n)$-dimensional vector. By (B.17), Theorem 2(i) and noting that $a_n = O\big(\frac{1}{\sqrt{n}\chi_n}\big)$ in the condition A8, it is easy to prove that $\hat{\mathbf{w}}_n(2) = \mathbf{0}$.

As in Fan and Li (2001), to prove (B.17), it is sufficient to show that, with probability approaching one, for any q_n-dimensional vector $\mathbf{w}_n^{\intercal} = \big[\mathbf{w}_n^{\intercal}(1), \mathbf{w}_n^{\intercal}(2)\big]$ with $\mathbf{w}_n(1)$ satisfying $\|\mathbf{w}_n(1) - \mathbf{w}_o(1)\| = O_P(\epsilon_{n*})$ and for $j = s_n + 1, \ldots, q_n$,

$$\frac{\partial \mathcal{Q}_n(\mathbf{w}_n)}{\partial w_j} > 0, \quad 0 < w_j < \epsilon_{n*}, \tag{B.18}$$

and

$$\frac{\partial \mathcal{Q}_n(\mathbf{w}_n)}{\partial w_j} < 0, \quad -\epsilon_{n*} < w_j < 0, \tag{B.19}$$

where $\mathbf{w}_n^{\intercal}(2) = (w_{s_n+1}, \ldots, w_{d_n})$.

Note that

$$\frac{\partial \mathcal{Q}_n(\mathbf{w}_n)}{\partial w_j} = \frac{\partial \mathcal{L}_n(\mathbf{w}_n)}{\partial w_j} + np'_\lambda(|w_j|)\mathsf{sgn}(w_j)$$

for $j = s_n + 1, \ldots, d_n$, where

$$\mathcal{L}_n(\mathbf{w}_n) = \big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_n)\big]^{\intercal}\big[\mathcal{Y}_n - \hat{\mathcal{M}}(\mathbf{w}_n)\big]$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}_n(\mathbf{w}_n)}{\partial w_j} &= \mathcal{Y}_n^{\intercal}\mathcal{S}_n(j)\big[\mathcal{Y}_n - \mathcal{S}_n(\mathcal{Y})\mathbf{w}_n\big] \\ &= \mathcal{Y}_n^{\intercal}\mathcal{S}_n(j)\big[\mathcal{Y}_n - \mathcal{S}_n(\mathcal{Y})\mathbf{w}_o\big] - \mathcal{Y}_n^{\intercal}\mathcal{S}_n(j)\mathcal{S}_n(\mathcal{Y})\big(\mathbf{w}_n - \mathbf{w}_o\big) \\ &=: \Xi_{n3} + \Xi_{n4}.\end{aligned}$$

As in the proof of Theorem 2(i), it is easy to prove that

$$|\Xi_{n3}| = O_P\big(\sqrt{q_n n}\chi_n^{-1}\big) \ \text{ and } \ |\Xi_{n4}| = O_P\big(\sqrt{q_n n}\chi_n^{-1}\big), \tag{B.20}$$

which indicate that

$$\frac{\partial \mathcal{L}_n(\mathbf{w}_n)}{\partial w_j} = O_P\big(\sqrt{q_n n}\chi_n^{-1}\big). \tag{B.21}$$

Hence, by (B.21), we have

$$\begin{aligned}\frac{\partial \mathcal{Q}_n(\mathbf{w}_n)}{\partial w_j} &= O_P\big(\sqrt{q_n n}\chi_n^{-1}\big) + np'_\lambda(|w_j|)\mathsf{sgn}(w_j) \\ &= O_P\big(\sqrt{q_n n}\chi_n^{-1}\big) + n\lambda\big[\lambda^{-1}p'_\lambda(|w_j|)\mathsf{sgn}(w_j)\big] \\ &= O_P\Big(\sqrt{q_n n}\chi_n^{-1}\big\{1 + \frac{\chi_n\sqrt{n}\lambda}{\sqrt{q_n}}\big[\lambda^{-1}p'_\lambda(|w_j|)\mathsf{sgn}(w_j)\big]\big\}\Big).\end{aligned} \tag{B.22}$$

Since $\frac{\chi_n\sqrt{n}\lambda}{\sqrt{q_n}} \to \infty$, we can show that (B.18) and (B.19) hold by using (B.22). We thus complete the proof of Theorem 2(ii).

(iii) Let $\hat{\mathbf{w}}_0^\intercal(n) = (\hat{\mathbf{w}}_n^\intercal(1), \mathbf{0}^\intercal)$ and $\hat{w}_j$ be the estimator of w_{oj} for $j = 1, 2, \ldots, q_n$. By Theorem 2(ii), we have

$$\frac{\partial \mathcal{Q}_n\big(\hat{\mathbf{w}}_n\big)}{\partial w_j} = \frac{\partial \mathcal{Q}_n\big(\hat{\mathbf{w}}_0(n)\big)}{\partial w_j} = 0 \tag{B.23}$$

for $j = 1, 2, \ldots, s_n$. By Taylor's expansion and Theorem 2(ii), we have for $j = 1, 2, \ldots, s_n$,

$$\begin{aligned} \frac{\partial \mathcal{Q}_n\big(\hat{\mathbf{w}}_0(n)\big)}{\partial w_j} &= \frac{\partial \mathcal{Q}_n\big(\mathbf{w}_o\big)}{\partial w_j} + \sum_{l=1}^{q_n} \frac{\partial^2 \mathcal{Q}_n\big(\mathbf{w}_n^*\big)}{\partial w_j \partial w_l}(\hat{w}_l - w_{ol}) \\ &= \frac{\partial \mathcal{Q}_n\big(\mathbf{w}_o\big)}{\partial w_j} + \sum_{l=1}^{s_n} \frac{\partial^2 \mathcal{Q}_n\big(\mathbf{w}_n^*\big)}{\partial w_j \partial w_l}(\hat{w}_l - w_{ol}), \end{aligned} \tag{B.24}$$

where $\mathbf{w}_n^*$ lies between $\hat{\mathbf{w}}_0(n)$ and $\mathbf{w}_o$.

Define

$$\boldsymbol{\Theta}_n^\intercal\big(\mathbf{w}_o\big) = \left[\frac{\partial \mathcal{Q}_n(\mathbf{w}_o)}{\partial w_1}, \ldots, \frac{\partial \mathcal{Q}_n(\mathbf{w}_o)}{\partial w_{s_n}}\right]$$

and $\boldsymbol{\Phi}_n\big(\mathbf{w}^*(n)\big)$ be the $s_n \times s_n$ matrix whose (j,k)-th component is $\frac{\partial^2 \mathcal{Q}_n(\mathbf{w}^*(n))}{\partial w_j \partial w_k}$. Then, by (B.24), we have

$$\hat{\mathbf{w}}_n(1) - \mathbf{w}_o(1) = \boldsymbol{\Phi}_n^{-1}\big(\mathbf{w}^*(n)\big)\boldsymbol{\Theta}_n\big(\mathbf{w}_o\big). \tag{B.25}$$

Following the proof of Theorem 3.3 in Li, Linton and Lu (2015), we can show that

$$\frac{1}{n}\boldsymbol{\Theta}_n\big(\mathbf{w}_o\big) \overset{P}{\sim} \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\xi}_t + \boldsymbol{\omega}_n, \tag{B.26}$$

where $\boldsymbol{\omega}_n$ is defined in Section 3.1. On the other hand, we can also show that

$$\frac{1}{n}\boldsymbol{\Phi}_n\big(\mathbf{w}^*(n)\big) \overset{P}{\sim} \boldsymbol{\Lambda}_{n1} + \boldsymbol{\Omega}_n, \tag{B.27}$$

where $\boldsymbol{\Lambda}_{n1}$ and $\boldsymbol{\Omega}_n$ are defined in Section 3.1. Letting $\boldsymbol{u}_{nt} = \mathcal{A}_n\boldsymbol{\Sigma}_n^{-1/2}\boldsymbol{\xi}_t$, by (B.25)–(B.27), it suffices to show that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \boldsymbol{u}_{nt} \xrightarrow{d} \mathsf{N}\big(\mathbf{0}, \mathcal{A}_0\big), \tag{B.28}$$

which can be proved by using the central limit theorem for the stationary α-mixing sequence. The proof of Theorem 2(iii) has thus been completed. ■

Proof of Theorem 3 (i) The proof is similar to the proof of Theorem 1 in Bai and Ng (2002) and the proof of Theorem 3.3 in Fan, Liao and Mincheva (2013). By the definition of $\hat{\mathbf{f}}_t$, we readily have

$$\begin{aligned}\hat{\mathbf{V}}\left(\hat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t^0\right) &= \frac{1}{np_n}\left(\sum_{s=1}^n\sum_{k=1}^{p_n}\hat{\mathbf{f}}_s(\mathbf{f}_s^0)^\intercal\mathbf{b}_k^0u_{tk} + \sum_{s=1}^n\sum_{k=1}^{p_n}\hat{\mathbf{f}}_s(\mathbf{f}_t^0)^\intercal\mathbf{b}_k^0u_{sk} + \right.\\ &\qquad \left.\sum_{s=1}^n\sum_{k=1}^{p_n}\hat{\mathbf{f}}_s\mathsf{E}\left[u_{sk}u_{tk}\right] + \sum_{s=1}^n\sum_{k=1}^{p_n}\hat{\mathbf{f}}_s\left\{u_{sk}u_{tk} - \mathsf{E}\left[u_{sk}u_{tk}\right]\right\}\right) \qquad \text{(B.29)}\end{aligned}$$

for any $1 \leq t \leq n$.

By the conditions B1 and B4, and following the proof of (A.35) in Appendix A, we may show that uniformly for $1 \leq t \leq n$,

$$\frac{1}{np_n}\left(\sum_{s=1}^n\sum_{k=1}^{p_n}\hat{\mathbf{f}}_s\mathsf{E}\left[u_{sk}u_{tk}\right]\right) = O_P(n^{-1/2}) \qquad \text{(B.30)}$$

and

$$\frac{1}{np_n}\left(\sum_{s=1}^n\sum_{k=1}^{p_n}\hat{\mathbf{f}}_s\left\{u_{sk}u_{tk} - \mathsf{E}\left[u_{sk}u_{tk}\right]\right\}\right) = O_P(n^{1/4}p_n^{-1/2}). \qquad \text{(B.31)}$$

Noting that $\left\|\sum_{s=1}^n\hat{\mathbf{f}}_s(\mathbf{f}_s^0)^\intercal\right\| = O_P(n)$ by the condition B2, and $\max_t\|\sum_{k=1}^{p_n}\mathbf{b}_k^0u_{tk}\| = O_P(n^{1/4}p_n^{1/2})$ by (3.9) in the condition B4, we have

$$\frac{1}{np_n}\left(\sum_{s=1}^n\sum_{k=1}^{p_n}\hat{\mathbf{f}}_s(\mathbf{f}_s^0)^\intercal\mathbf{b}_k^0u_{tk}\right) = O_P(n^{1/4}p_n^{-1/2}) \qquad \text{(B.32)}$$

uniformly for $1 \leq t \leq n$.

Notice that

$$\begin{aligned}\max_t\left\|\sum_{s=1}^n\sum_{k=1}^{p_n}\hat{\mathbf{f}}_s(\mathbf{f}_t^0)^\intercal\mathbf{b}_k^0u_{sk}\right\| &\leq \max_t\|\mathbf{f}_t^0\|\left(\sum_{s=1}^n\left\|\hat{\mathbf{f}}_s\right\|^2\right)^{1/2}\left(\sum_{s=1}^n\left\|\sum_{k=1}^{p_n}\mathbf{b}_k^0u_{sk}\right\|^2\right)^{1/2}\\ &= O_P(1)\cdot O_P(n^{1/2})\cdot O_P(n^{3/4}p_n),\end{aligned}$$

by using the conditions B2 and B4. Hence, we have

$$\frac{1}{np_n}\left(\sum_{s=1}^n\sum_{k=1}^{p_n}\hat{\mathbf{f}}_s(\mathbf{f}_t^0)^\intercal\mathbf{b}_k^0u_{sk}\right) = O_P(n^{1/4}p_n^{-1/2}) \qquad \text{(B.33)}$$

uniformly for $1 \leq t \leq n$.

By the definition of $\hat{\mathcal{F}}_n$ and following the arguments in the proof of Lemma 5 in Appendix A, we may show that

$$\hat{\mathbf{V}} - \left(\frac{1}{n}\hat{\mathcal{F}}_n^{\intercal}\mathcal{F}_n^0\right)\left(\frac{1}{p_n}(\mathbf{B}_n^0)^{\intercal}\mathbf{B}_n^0\right)\left(\frac{1}{n}(\mathcal{F}_n^0)^{\intercal}\hat{\mathcal{F}}_n\right) = o_P(1). \tag{B.34}$$

Furthermore, by Lemma 5 again,

$$\frac{1}{n}(\mathcal{F}_n^0)^{\intercal}(\mathcal{F}_n^0) - \left(\frac{1}{n}(\mathcal{F}_n^0)^{\intercal}\hat{\mathcal{F}}_n\right)\left(\frac{1}{n}\hat{\mathcal{F}}_n^{\intercal}\mathcal{F}_n^0\right) = o_P(1),$$

which together with the condition B2, implies that $\hat{\mathcal{F}}_n^{\intercal}\mathcal{F}_n^0/n$ is asymptotically invertible. By (B.34) and noting that $(\mathbf{B}_n^0)^{\intercal}\mathbf{B}_n^0/p_n$ is positive definite, we may show that $\hat{\mathbf{V}}$ is also asymptotically invertible. We can then complete the proof of (3.12) in Theorem 3(i) by using this fact and (B.29)–(B.33).

(ii) Let

$$\eta_{tk,f}^* = Y_t - m_{k,f}^*(\tilde{f}_{tk}^0) = Y_t - \mathsf{E}\big[Y_t|\tilde{f}_{tk}^0\big],$$

where $\tilde{f}_{tk}^0 = e_r(k)\mathbf{H}\mathbf{f}_t^0$ is defined as in Section 2.2. Given the $r\times r$ matrix $\mathbf{H}$, $\big\{\big(\eta_{tk,f}^*, \tilde{f}_{tk}^0\big), k = 1, \cdots, r\big\}$ is a stationary α-mixing process over t which satisfies the condition B1. Note that

$$\begin{aligned}
\hat{m}_{k,f}^*(z_k) - \tilde{m}_{k,f}^*(z_k) &= \big[\hat{m}_{k,f}^*(z_k) - m_{k,f}^*(z_k)\big] - \big[\tilde{m}_{k,f}^*(z_k) - m_{k,f}^*(z_k)\big] \\
&= \frac{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)\big[Y_t - m_{k,f}^*(z_k)\big]}{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)} - \frac{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)\big[Y_t - m_{k,f}^*(z_k)\big]}{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)} \\
&= \left[\frac{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)\eta_{tk,f}^*}{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)} - \frac{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)\eta_{tk,f}^*}{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)}\right] + \\
&\quad \left[\frac{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)\Delta_{tk,m}}{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)} - \frac{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)\Delta_{tk,m}}{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)}\right] \\
&=: \Gamma_{n1}(z_k) + \Gamma_{n2}(z_k),
\end{aligned} \tag{B.35}$$

where $\Delta_{tk,m} = m_{k,f}^*(\tilde{f}_{tk}^0) - m_{k,f}^*(z_k)$.

We first consider the uniform convergence for $\Gamma_{n1}(z_k)$. It is easy to show that

$$\begin{aligned}
\Gamma_{n1}(z_k) &= \left[\frac{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)\eta_{tk,f}^*}{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)} - \frac{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)\eta_{tk,f}^*}{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)}\right] + \\
&\quad \left[\frac{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)\eta_{tk,f}^*}{\sum_{t=1}^n K\big(\frac{\hat{f}_{tk}-z_k}{h_3}\big)} - \frac{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)\eta_{tk,f}^*}{\sum_{t=1}^n K\big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\big)}\right] \\
&=: \Gamma_{n1,1}(z_k) + \Gamma_{n1,2}(z_k).
\end{aligned} \tag{B.36}$$

Following the arguments in the proof of Lemma 3, we may prove

$$\max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\frac{1}{nh_3}\sum_{t=1}^n K\Big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\Big)\eta_{tk,f}^*\right| = O_P\left(\sqrt{\log n/(nh_3)}\right). \tag{B.37}$$

By (3.11) and (3.12) in Theorem 3(i), we have

$$\max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\frac{1}{nh_3}\sum_{t=1}^n K\Big(\frac{\hat{f}_{tk}-z_k}{h_3}\Big)-\frac{1}{nh_3}\sum_{t=1}^n K\Big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\Big)\right| = O_P\left(n^{-1/2}h_3^{-1}+n^{1/4}p_n^{-1/2}h_3^{-1}\right). \tag{B.38}$$

By (B.37), (B.38) and the condition B5(ii), we readily have

$$\max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}|\Gamma_{n1,2}(z_k)| = O_P\left(n^{-1}h_3^{-3/2}(\log n)^{1/2}+n^{-1/4}h_3^{-3/2}p_n^{-1/2}(\log n)^{1/2}\right)=o_P\left(n^{-1/2}\right). \tag{B.39}$$

By the condition B5(i), we apply Taylor's expansion to the kernel function:

$$K\Big(\frac{\hat{f}_{tk}-z_k}{h_3}\Big)-K\Big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\Big)=\frac{\hat{f}_{tk}-\tilde{f}_{tk}^0}{h_3}K'\Big(\frac{\tilde{f}_{tk}^*-z_k}{h_3}\Big),$$

where $\tilde{f}_{tk}^*$ lies between $\tilde{f}_{tk}^0$ and $\hat{f}_{tk}$. Using the above expansion, (B.29) and (B.31)–(B.33), we have

$$\begin{aligned}
&\max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\frac{1}{nh_3}\sum_{t=1}^n \eta_{tk,f}^*K\Big(\frac{\hat{f}_{tk}-z_k}{h_3}\Big)-\frac{1}{nh_3}\sum_{t=1}^n \eta_{tk,f}^*K\Big(\frac{\tilde{f}_{tk}^0-z_k}{h_3}\Big)\right| \\
=\;& \max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\frac{1}{n^2h_3^2p_n}\sum_{s=1}^n\hat{f}_{sk}\sum_{t=1}^n\sum_{k=1}^{p_n}\mathsf{E}\left[u_{sk}u_{tk}\right]\eta_{tk,f}^*K'\Big(\frac{\tilde{f}_{tk}^*-z_k}{h_3}\Big)\right|+O_P\left(n^{1/4}p_n^{-1/2}h_3^{-1}\right) \\
=\;& \max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\frac{1}{n^2h_3^2p_n}\sum_{s=1}^n\hat{f}_{sk}\sum_{t=1}^n\sum_{k=1}^{p_n}\mathsf{E}\left[u_{sk}u_{tk}\right]\eta_{tk,f}^*K'\Big(\frac{\tilde{f}_{tk}^*-z_k}{h_3}\Big)\right|+o_P\left(n^{-1/2}\right).
\end{aligned} \tag{B.40}$$

By the conditions B1 and B4, for $\kappa_n=[n^{\gamma_0}]$, we have

$$\sum_{|s-t|>\kappa_n}\mathsf{E}\left[u_{sk}u_{tk}\right]=O\left(\theta_0^{3\kappa_n/4}\right),\quad 0<\theta_0<1,$$

which implies that

$$
\begin{aligned}
& \max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\frac{1}{n^2h_3^2p_n}\sum_{s=1}^{n}\hat f_{sk}\sum_{t=1}^{n}\sum_{k=1}^{p_n}\mathsf{E}\left[u_{sk}u_{tk}\right]\eta^*_{tk,f}K'\Big(\frac{\tilde f^*_{tk}-z_k}{h_3}\Big)\right| \\
= & \max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\frac{1}{n^2h_3^2p_n}\sum_{s=1}^{n}\hat f_{sk}\sum_{|t-s|\le\kappa_n}\sum_{k=1}^{p_n}\mathsf{E}\left[u_{sk}u_{tk}\right]\eta^*_{tk,f}K'\Big(\frac{\tilde f^*_{tk}-z_k}{h_3}\Big)\right| + o_P\left(n^{-1/2}\right) \\
\le & \max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\frac{1}{n^2h_3^2p_n}\sum_{s=1}^{n}\hat f_{sk}\sum_{|t-s|\le\kappa_n}\sum_{k=1}^{p_n}\mathsf{E}\left[u_{sk}u_{tk}\right]\eta^*_{tk,f}K'\Big(\frac{\tilde f^0_{tk}-z_k}{h_3}\Big)\right| + o_P\left(n^{-1/2}\right) + \\
& O_P\left(\frac{\kappa_n}{n^{3/2}h_3^3}+\frac{\kappa_n}{n^{3/4}p_n^{1/2}h_3^3}\right) \\
= & O_P\left(\frac{\kappa_n}{nh_3}\right) + O_P\left(\frac{\kappa_n}{n^{3/2}h_3^3}+\frac{\kappa_n}{n^{3/4}p_n^{1/2}h_3^3}\right) + o_P\left(n^{-1/2}\right) = o_P\left(n^{-1/2}\right),
\end{aligned}
\tag{B.41}
$$

where we have used (3.12) in Theorem 3(i) and the condition B5(ii). By (B.38), (B.40) and (B.41), we may prove that

$$\max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\Gamma_{n1,1}(z_k)\right| = o_P\left(n^{-1/2}\right). \tag{B.42}$$

By (B.36), (B.39) and (B.42), we readily have

$$\max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\Gamma_{n1}(z_k)\right| = o_P\left(n^{-1/2}\right). \tag{B.43}$$

On the other hand, using Taylor's expansion for $m^*_{k,f}(\cdot)$ and following the arguments in the proofs of (B.39) and (B.42), we may also show that

$$\max_{1\le k\le r}\sup_{z_k\in\mathcal{F}_k^*}\left|\Gamma_{n2}(z_k)\right| = o_P\left(n^{-1/2}\right), \tag{B.44}$$

which, together with (B.43), completes the proof of Theorem 3(ii). ■

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] Bühlmann, P., van de Geer, S., 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer Series in Statistics, Springer.

[9] Chen, J., Li, D., Linton, O., Lu, Z., 2015. Semiparametric dynamic portfolio choice with multiple conditioning variables. Forthcoming in *Journal of Econometrics.*

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

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

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

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

[14] 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.

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

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

[17] 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.

[18] 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.

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

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

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

[22] Fu, W., 1998. Penalized regression: the bridge versus LASSO. *Journal of Computational and Graphical Statistics* 7, 397–416.

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

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

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

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

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

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

[29] 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.

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

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

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

[33] 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.

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

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

[36] 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.

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

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

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

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

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