
Article:
Li, Jialiang, Zhang, Wenyang orcid.org/0000-0001-8391-1122 and Kong, Efang (2018) Factor Models for Asset Returns Based on Transformed Factors. Journal of Econometrics. pp. 432-448. ISSN 0304-4076

https://doi.org/10.1016/j.jeconom.2018.09.001

Factor Models for Asset Returns Based on Transformed Factors

Jialiang Li
Department of Statistics & Applied Probability
National University of Singapore, Singapore

Wenyang Zhang
Department of Mathematics
The University of York, UK

Efang Kong
School of Mathematical Science,
University of Electronic Science and Technology, China

September 14, 2018

Abstract

The Fama-French three factor models are commonly used in the description of asset returns in finance. Statistically speaking, the Fama-French three factor models imply that the return of an asset can be accounted for directly by the Fama-French three factors, i.e. market, size and value factor, through a linear function. A natural question is: would some kind of transformed Fama-French three factors work better? If so, what kind of transformation should be imposed on each factor in order to make the transformed three factors better account for asset returns? In this paper, we are going to address these questions through nonparametric modelling. We propose a data driven approach to construct the transformation for each factor concerned. A generalised maximum likelihood ratio based hypothesis test is also proposed to test whether transformations on the Fama-French three factors are needed for a given data set. Asymptotic properties are established to justify the proposed methods. Extensive simulation studies are conducted to show how the proposed methods perform with finite sample size. Finally, we apply the proposed methods to a real data set, which leads to some interesting findings.

KEY WORDS: Backfitting, factor models, generalised maximum likelihood ratio test, kernel smoothing, transformed factor.

SHORT TITLE: FM for Asset Returns.

1 Introduction

1.1 Preamble

During the past two decades, much literature is devoted to explore the common factors in asset returns, see Ang *et al.*(2006), Brennan *et al.*(1998), Davis *et al.*(2000), Fama (1998), Fama and French (1993, 1996, 2010, 2015), Petkova (2006), Vassalou and Xing (2004), and the references therein. Among the existing factor models, the Fama-French three factor models (FFTFM) are arguably the most common choices. They play a very important role in asset pricing and portfolio management. The application of the FFTFM may go beyond finance and economics. Fan *et al.*(2008), for example, apply the FFTFM to introduce a structure for high dimensional covariance matrices, which significantly improves the covariance estimation. Another example is related to measuring conditional dependence, which is a critical issue in statistics with broad applications, such as the graphical models. Based on the FFTFM, Fan *et al.*(2015) proposed a new conditional dependence measure to address this problem.

1.2 Motivating questions

Statistically speaking, the FFTFM implies that the return of an asset can be accounted for directly by the Fama-French three factors, commonly referred to as market (Rm-Rf), size (SMB) and value factor (HML), through a linear function. Precisely, Rm-Rf is a measure of market risk, computed as the difference between the return of the market portfolio and the risk-free return set; SMB stands for small market capitalization minus big market capitalization and HML for high book-to-market ratio minus low book-to-market ratio. These factors measure the historic excess returns of small caps over big caps and of value stocks over growth stocks. They are calculated with combinations of portfolios composed by ranked stocks and available historical market data. See Fama and French (1993) for more details. Since FFTFM has a very simple linear form, a natural question is: would some kind of transformed Fama-French three factors work better? If so, what kind of transformation should be imposed on each factor in order to make the transformed three factors better account for asset returns? We can go even further to question whether the linearity assumption in the FFTFM always holds.

To provide a motivation for the models we are going to propose and investigate in this paper, we first study a historical data set freely downloadable from Kenneth French's website:

`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

The original data set consists of the daily simple returns of $n = 49$ industry portfolios from 1927 to 2014. Let r_{tj} be the daily return of the jth portfolio at time t, $j = 1, \cdots, 49$, $t = 1, \cdots, T$,

x_{t1} (Rm-Rf), x_{t2} (SMB), x_{t3} (HML) be, respectively, the observations of the Fama-French three factors at time t. We note that the factors are all aggregated measures combining portfolios and thus do not depend on j. For each given j, $j = 1, \cdots, 49$, the FFTFM may be written as

$$r_{tj} = \alpha_j + \sum_{k=1}^{3} \beta_{jk} x_{tk} + \epsilon_{tj}, \quad t = 1, \cdots, T. \tag{1.1}$$

We fit this model to the downloaded data and compare with our proposed model in the real case studies of this paper. The transformed factors, denoted as $g_1(x_{t1})$, $g_2(x_{t2})$ and $g_3(x_{t3})$, might better account for asset returns than x_{t1}, x_{t2} and x_{t3} do. In particular, the estimated function of g_3 shows a clear nonlinear form (Figure 3 in section 6). Furthermore, using our proposed model yields more accurate numerical prediction for this data set than using the FFTFM, with more than 35% improvement in cross validation errors. It is thus worth considering such a transformation in a practical financial data analysis.

Now, the first question is to determine the transformations $g_1(\cdot)$, $g_2(\cdot)$ and $g_3(\cdot)$ from empirical data.

1.3 The proposed models

In order to find the transformations needed for the Fama-French three factors, we are going to propose a new factor model based on a mathematical transformation.

In general, suppose we have p factors, $x_1, \cdots, x_p$. In particular for FFTFM, we have $p = 3$. For $t = 1, \cdots, T$, let $x_{t1}, \cdots, x_{tp}$ be the observations of the factors at time t, and r_{tj}, the return of the jth asset at time t, $j = 1, \cdots, n$. We assume

$$r_{tj} = \alpha_j + \sum_{k=1}^{p} \beta_{jk} g_k(x_{tk}) + \epsilon_{tj}, \quad t = 1, \cdots, T; \;\; j = 1, \cdots, n, \tag{1.2}$$

where α_j, β_{jk}, and $g_k(\cdot)$, $j = 1, \cdots, n$; $k = 1, \cdots, p$, are unknown and need to be estimated, and

$$E(\epsilon_{tj}|x_{t1}, \cdots, x_{tp}) = 0, \quad \text{var}(\epsilon_{tj}|x_{t1}, \cdots, x_{tp}) = \sigma^2.$$

It is clear (1.2) is not identifiable. To make (1.2) identifiable, we assume

$$g_k(x_{1k}) = x_{1k} \text{ and } E\{g_k(x_k)\} = 0, \quad k = 1, \cdots, p. \tag{1.3}$$

Model (1.2) together with the identification condition (1.3) is the model we are going to address in the following. To connect the proposed model to the motivating questions, the $g_k(\cdot)$ in (1.2) is the transformation needed for the kth factor. In this paper we fix n and require $T \to \infty$. This sample size assumption is satisfied in financial studies where we usually investigate a fixed number

of interesting items with a large number of repeated observations. We assume that the distribution of the factors is not degenerate. The error distribution is left unspecified in this paper.

There is fundamental difference between the proposed model (1.2) and the additive models for panel data, which is the model (1.2) with $\beta_{jk}g_k(x_{tk})$ being replaced by a completely unknown function $G_{jk}(x_{tk})$. From statistical modelling point of view, the proposed model is more parsimonious, because there are only p unknown functions and $(p+1)n$ unknown parameters in the proposed model, whilst there are $(p+1)n$ unknown functions in the additive models for panel data. Most importantly, the proposed model (1.2) is more meaningful, because from finance point of view, $g_k(x_{tk})$, $k = 1, \cdots, p$, in (1.2) act as common risk factors, whilst $G_{jk}(x_{tk})$, $j = 1, \cdots, n$, $k = 1, \cdots, p$, in the additive models depend on individual asset, therefore cannot be viewed as common risk factors.

To estimate the unknown transformation, we will consider the familiar kernel smoothing approach. We notice that there exists a vast literature of using kernel methods for financial econometrics data analysis where support vector machine methods are perhaps the most popular (eg. Burges (1998), Ince and Trafalis (2006), Schebesch and stecking (2005) among others). Several other kernel learning algorithms are also available. For example, Lanckriet et al. (2004) and Ong et al. (2005) formulated kernel learning as semidefinite programming problems. Cristianini et al. (2006) examined the alignment between a kernel and the data to adapt the kernel matrix.

The rest of the paper is organized as follows. We begin in Section 2 with a description of the estimation procedure for the unknowns in (1.2). Hypothesis test about whether a transformation is needed for each factor is discussed in Section 3. Section 4 is devoted to the asymptotic properties of the proposed estimators and the hypothesis test. Simulation studies are conducted in Section 5 to show how accurate the proposed estimators are and how powerful the proposed hypothesis test is when sample size is finite. In Section 6, we apply the proposed modelling, estimation procedure and hypothesis test to the real data set mentioned in Section 1.2, and some interesting findings will be presented. All the detailed proofs are relegated to the appendix.

2 Estimation procedure

In this section, we are going to construct the estimation procedure for the unknowns in (1.2). We are going to address the estimation of $g_k(\cdot)$s first, then α_js and β_{jk}s. With a slight abuse of notation, from now on, for any random error appears in a synthetic model in this section, we shall denote it by e_{tj} to avoid repetition.

2.1 Estimation of $g_k(\cdot)$

Let $G_{jk}(x_{tk}) = \beta_{jk} g_k(x_{tk})$, and re-write (1.2) as

$$r_{tj} = \alpha_j + \sum_{k=1}^{p} G_{jk}(x_{tk}) + \epsilon_{tj}, \quad t = 1, \cdots, T; \;\; j = 1, \cdots, n.$$

For each given j, $j = 1, \cdots, n$, we apply the backfitting algorithm to estimate $G_{jk}(x_{tk})$, which is detailed as follows: Let

$$\hat{\alpha}_j = \frac{1}{T} \sum_{t=1}^{T} r_{tj} \tag{2.1}$$

and iterate the following two steps until convergence

1. Given the current $\tilde{G}_{jk}(x_{tk})$, $k = 1, \cdots, p$. For each l, $l = 1, \cdots, p$, we run the following synthetic univariate nonparametric regression

$$r_{tj} - \hat{\alpha}_j - \sum_{k=1}^{l-1} \hat{G}_{jk}(x_{tk}) - \sum_{k=l+1}^{p} \tilde{G}_{jk}(x_{tk}) = G_{jl}(x_{tl}) + e_{tj}, \quad t = 1, \cdots, T$$

by the local linear modelling, which is detailed as follows. For any given u, by the Taylor's expansion, we have

$$G_{jl}(x_{tl}) \approx G_{jl}(u) + \dot{G}_{jl}(u)(x_{tl} - u)$$

when x_{tl} is in a small neighbourhood of u. This leads to the following objective function for the local least squares estimation

$$\sum_{t=1}^{T} \left\{ r_{tj} - \hat{\alpha}_j - \sum_{k=1}^{l-1} \hat{G}_{jk}(x_{tk}) - \sum_{k=l+1}^{p} \tilde{G}_{jk}(x_{tk}) - c_{jl} - d_{jl}(x_{tl} - u) \right\}^2 K_h(x_{tl} - u), \tag{2.2}$$

where $K_h(\cdot) = K(\cdot/h)/h$, h is a bandwidth, $K(\cdot)$ is a kernel function, usually taken to be Epanechnikov kernel. Minimise (2.2) with respect to $(c_{jl},\ d_{jl})$, and denote the minimiser as $(\hat{c}_{jl},\ \hat{d}_{jl})$. The local linear estimator of $G_{jl}(u)$ is taken to be $\hat{c}_{jl}$, and denoted by $\check{G}_{jl}(u)$. By simple calculation, we have

$$\check{G}_{jl}(u) = (1,\ 0) \left(\boldsymbol{\Omega}_l(u)^{\mathrm{T}} \mathbf{W}_{l,h}(u) \boldsymbol{\Omega}_l(u) \right)^{-1} \boldsymbol{\Omega}_l(u)^{\mathrm{T}} \mathbf{W}_{l,h}(u) \boldsymbol{\eta}_{jl},$$

where $\mathbf{W}_{l,h}(u) = \mathrm{diag}\left(K_h(x_{1l} - u), \cdots, K_h(x_{Tl} - u) \right)$,

$$\boldsymbol{\Omega}_l(u) = \begin{pmatrix} 1 & x_{1l} - u \\ \vdots & \vdots \\ 1 & x_{Tl} - u \end{pmatrix}, \quad \boldsymbol{\eta}_{jl} = \begin{pmatrix} r_{1j} - \hat{\alpha}_j - \sum_{k=1}^{l-1} \hat{G}_{jk}(x_{1k}) - \sum_{k=l+1}^{p} \tilde{G}_{jk}(x_{1k}) \\ \vdots \\ r_{Tj} - \hat{\alpha}_j - \sum_{k=1}^{l-1} \hat{G}_{jk}(x_{Tk}) - \sum_{k=l+1}^{p} \tilde{G}_{jk}(x_{Tk}) \end{pmatrix}.$$

For each x_{tl}, the centralised $\check{G}_{jl}(x_{tl})$, denoted by $\hat{G}_{jl}(x_{tl})$, is

$$\hat{G}_{jl}(x_{tl}) = \check{G}_{jl}(x_{tl}) - \frac{1}{T}\sum_{t=1}^{T}\check{G}_{jl}(x_{tl}).$$

2. Let $\tilde{G}_{jk}(x_{tk})$ be $\hat{G}_{jl}(x_{tk})$, and go to step 1.

The iteration can be started by setting

$$\tilde{G}_{jk}(x_{tk}) = 0, \quad k = 1, \cdots, p.$$

With the final backfitting estimators $\hat{G}_{jl}(.)$s, we can construct the estimators of the functions $g_k(.)$s evaluated at the observation points as

$$\bar{g}_k(x_{tk}) = x_{1k}\frac{1}{n}\sum_{j=1}^{n}\hat{G}_{jk}(x_{tk})/\hat{G}_{jk}(x_{1k}), \quad k = 1, \cdots, p, \ t = 1, \cdots, T. \tag{2.3}$$

For each k, $k = 1, \cdots, p$, and any given u, viewing $\bar{g}_k(x_{tk})$ as a response variable, x_{tk} as a covariate, we have the following synthetic univariate nonparametric regression model

$$\bar{g}_k(x_{tk}) = g_k(x_{tk}) + e_{tk}, \quad t = 1, \cdots, T. \tag{2.4}$$

Applying the local linear modelling to (2.4), similar to what we have done in step 1 in the backfitting algorithm for estimating $G_{jk}(x_{tk})$, we get an estimator of $g_k(u)$

$$\hat{g}_k(u) = (1, \ 0)\left(\mathbf{\Omega}_k(u)^{\mathrm{T}}\mathbf{W}_{k,\tilde{h}}(u)\mathbf{\Omega}_k(u)\right)^{-1}\mathbf{\Omega}_k(u)^{\mathrm{T}}\mathbf{W}_{k,\tilde{h}}(u)\boldsymbol{\zeta}_k, \quad \boldsymbol{\zeta}_k = (\bar{g}_k(x_{1k}), \cdots, \bar{g}_k(x_{Tk}))\,,$$

where $\tilde{h}$ is a bandwidth. $\hat{g}_k(u)$ is our estimator of $g_k(u)$.

When implementing the proposed nonparametric estimation in the following numerical analysis, we carry out the cross-validation (CV) method to select the tuning bandwidth. Since at each step only univariate smoothing is needed the CV can be implemented efficiently for bandwidth selection. Our numerical results suggest that the performance of this method is quite stable.

2.2 Estimation of β_{jk}

Estimates $\hat{\alpha}_j$, $j = 1, \cdots, n$, from (2.1) and $\bar{g}_k(x_{tk})$, $t = 1, \cdots, T$, $k = 1, \cdots, p$, from (2.3) are plugged into (1.2) as substitutes for their corresponding true but unknown counterparts so that we have the following synthetic linear model

$$r_{tj} = \hat{\alpha}_j + \sum_{k=1}^{p}\beta_{jk}\bar{g}_k(x_{tk}) + e_{tj}, \quad t = 1, \cdots, T. \tag{2.5}$$

Let $\boldsymbol{\beta}_j = (\beta_{j1},\ \cdots,\ \beta_{jp})^\top$. We use the least squares estimator $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}_j$ to estimate $\boldsymbol{\beta}_j$, which is

$$\hat{\boldsymbol{\beta}}_j = (\bar{\mathbf{g}}^{\mathrm{T}}\bar{\mathbf{g}})^{-1}\,\bar{\mathbf{g}}^{\mathrm{T}}\mathbf{R}_j, \tag{2.6}$$

where

$$\bar{\mathbf{g}} = \begin{pmatrix} \bar{g}_1(x_{11}) & \cdots & \bar{g}_p(x_{1p}) \\ \vdots & \ddots & \vdots \\ \bar{g}_1(x_{T1}) & \cdots & \bar{g}_p(x_{Tp}) \end{pmatrix} \text{ and } \mathbf{R}_j = (r_{1j},\ \cdots,\ r_{Tj})^{\mathrm{T}}.$$

3 Hypothesis test

In this section, we are going to address whether or not a transformation on each factor is significantly needed for a given data set. We fomulate this question to a hypothesis test problem with null hypothesis

$$H_0 :\ g_1(x) = \cdots = g_p(x) = x. \tag{3.1}$$

and alternative hypothesis being that transformations on the factors are needed.

Our hypothesis test is based on the generalised maximum likelihood ratio test, see Fan *et al.*(2001). To construct the hypothesis test statistic, we first compute the residual sum of squares of the model (1.2) under null hypothesis (3.1). Under the null hypothesis (3.1), (1.2) becomes the following linear model

$$r_{tj} = \alpha_j + \sum_{k=1}^{p} \beta_{jk} x_{tk} + \epsilon_{tj}, \quad t = 1,\ \cdots,\ T;\ \ j = 1,\ \cdots,\ n. \tag{3.2}$$

Let

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T1} & \cdots & x_{Tp} \end{pmatrix}$$

By some simple calculations, we have the residual sum of squares of (3.2)

$$\mathrm{RSS}_0 = \sum_{j=1}^{n} \mathbf{R}_j^{\mathrm{T}} \left\{\mathbf{I}_T - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\right\} \mathbf{R}_j,$$

where $\mathbf{I}_T$ is an identity matrix of size T.

On the other hand, the residual sum of squares of (1.2) is

$$\mathrm{RSS}_1 = \sum_{j=1}^{n}\sum_{t=1}^{T} \left(r_{tj} - \hat{\alpha}_j - \sum_{k=1}^{p} \hat{\beta}_{jk}\bar{g}_k(x_{tk}) \right)^2.$$

Based on the idea in Fan *et al.*(2001), we propose the following test statistic for the null hypothesis (3.1)

$$\lambda = \frac{nT}{2}\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}.$$

We reject H_0 when $\lambda > c$, where c is determined by

$$P(\lambda > c|H_0) = \alpha,$$

α is the significant level.

In the implementation of the proposed hypothesis test, the distribution of λ under null hypothesis can be either estimated by bootstrap or approximated by its asymptotic distribution presented in Section 4.

4 Asymptotic properties

For each $k = 1, \cdots, p$, as far as the estimator of $g_k(u)$ is concerned, because the theoretical properties of $\hat{g}_k(u)$ easily follow from those of $\bar{g}_k(x_{tk})$ at the expense of further cumbersome notations, we only present the asymptotic properties of $\bar{g}_k(x_{tk})$.

For simplicity, we assume that observation points all lie in the interior of the support of $\mathbf{x}$ and focus on local polynomial fittings of odd degrees, as the expressions become considerably more complicated with boundary points or in the case of even degrees (Opsomer and Ruppert, 1997). Write $\boldsymbol{\epsilon}_j = (\epsilon_{1j}, \cdots, \epsilon_{\mathrm{T}j})^\top$, $j = 1, \cdots, n$. Then regarding the estimates discussed in Section 2, we have

Theorem 4.1 *Under the Assumptions given in the Appendix,*

(1) $\bar{g}_k(x_{tk}) = g_k(x_{tk}) + \gamma_{tk}^\top \mathbf{S}_k \frac{1}{n}\sum_{j=1}^{n} \beta_{jk}^{-1}\boldsymbol{\epsilon}_j + o_p(T^{-1/2})$ *uniformly with respect to* $t = 1, \cdots, T$ *and* $k = 1, \cdots, p$.

(2) $T^{1/2}(\hat{\alpha}_j - \alpha_j) \xrightarrow{D} N(0,\ \sigma^2)$

(3) $\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j = c_0(K)\frac{1}{Tn}\sum_{j'=1}^{n} \mathbf{A}_{j'|j}\boldsymbol{\epsilon}_{j'} + o_p(T^{-1/2})$ *uniformly over* $j = 1, \cdots, n$.

Definitions of $T \times 1$ vector γ_{tk}, $T \times T$ matrix $\mathbf{S}_k$, constant $c_0(K)$ and $p \times T$ matrix $\mathbf{A}_{j'|j}$ are given in the Appendix. It easily follows that $\hat{g}_k(.)$ converges at a nonparametric rate of $(Th_k)^{-1/2}$.

Let $R(K) = \int K^2(u)du$. For the testing statistic in Section 3, we have;

Theorem 4.2 *Suppose conditions in Theorem 4.1 hold, and for ease of exposition,* $h_1 = h_2 = \cdots = h_p = h$. *Then under the null hypothesis (3.1),*

$$P\{\sigma_T^{-1}[\lambda - npK(0)h^{-1}] < t\} \longrightarrow \Phi(t), \quad \textit{when } T \to \infty,$$

where $\Phi(\cdot)$ *is the standard normal distribution function,*

$$\sigma_T^2 = \sigma^4 R(K) h^{-1} \Big\{ \sum_{j,k} c_k \{4 + \sum_{j' \neq j} (\beta_{jk}/\beta_{j'k})^2\} + n(n-1) \sum_{k=1}^{p} c_k \Big\}.$$

Constant c_k is to be defined in the Appendix.

Theorem 4.2 provides us the asymptotic distribution of the proposed test statistic for the null hypothesis (3.1), which can be used to estimate the critical value of the proposed hypothesis test in Section 3.

5 Simulation studies

In this section, we are going to use a simulated example to demonstrate how accurate the proposed estimators are. We will also examine the power of the proposed hypothesis test for the null hypothesis (3.1). As the asymptotic distribution of the test statistic involves unknown parameters and some constants which are hard to calculate, we will use bootstrap approach to compute the critical value for the test.

We generate data according to model (1.2). Specifically, each element of $X_t = (x_{t1}, \cdots, x_{tp})^T$ is independently generated from a uniform distribution over $[-1, 1]$, and each random error ϵ_{tj} is generated from an auto-regressive time series with auto-regressive coefficient r_1 and a standard Guassian error. We set $p = 4$ and

$$\begin{gathered} g_1(x_1) = \sin(2.5\pi x_1), \quad g_2(x_2) = x_2^3, \quad g_3(x_3) = \sin(0.5\pi x_3), \\ g_4(x_4) = \left[1/\left\{1 + \exp(-x_4)\right\} - 0.5\right] / \left\{1/(1+e^{-1}) - 0.5\right\}. \end{gathered} \tag{5.1}$$

We will consider various n and T in our simulation study. For each n and T, the interecepts α_js in the model (1.2) are independently generated from $N(3,\ 0.5)$ and the slopes β_{jk}s are independently generated from $N(3.5,\ 0.5)$. Once these α_js and β_{jk}s are generated, we fix them across all simulations for the given n and T.

Let $\mathrm{MSE}(\hat{\alpha}_j)$ and $\mathrm{MSE}(\hat{\beta}_{jk})$ be the mean squared errors of $\hat{\alpha}_j$ and $\hat{\beta}_{jk}$, respectively. We consider ARMSE_α and ARMSE_β, which are defined as

$$\mathrm{ARMSE}_\alpha = \frac{1}{n} \sum_{j=1}^{n} \left\{ \alpha_j^{-2} \mathrm{MSE}(\hat{\alpha}_j) \right\}, \quad \mathrm{ARMSE}_\beta = \frac{1}{np} \sum_{j=1}^{n} \sum_{k=1}^{p} \left\{ \beta_{jk}^{-2} \mathrm{MSE}(\hat{\beta}_{jk}) \right\},$$

to assess the accuracy of our estimation for the interecepts α_js and for the slopes β_{jk}s, respectively. Let MISE_k be the mean integrated squared error of $\hat{g}_k(\cdot)$. We use ARMISE, which is defined as

$$\mathrm{ARMISE} = \frac{1}{p} \sum_{k=1}^{p} \mathrm{MISE}_k \left\{ \int g_k(u)^2 du \right\}^{-2}$$

to assess the accuracy of our estimation for the unknown functions $g_k(\cdot)$s.

We consider various n and T in the simulation. The ARMSE_α and ARMSE_β obtained from the 500 simulations are presented in Table 1, and the ARMISE obtained from the simulations are reported in 2. Additional simulation results for point estimation bias and standard deviations are available in the supplementary file of this paper. All the tabled simulation results show our estimation procedure works very well.

To compare with linear approach and the SVD approach, we also reported the average R^2 values which are the proportion of explained variation by the models. To implement the SVD approach, we carry out a factor analysis (Bai and Ng (2002)) of the data matrix and select the factors corresponding to the largest eigenvalues such that at least 80% variation is accounted. The results are summarized in Table 1.

Table 1: **Estimation Performance for Unknown Parameters and Comparison of R^2 for Four Approaches: R_1 is for linear model (FFTFM); R_2 is for our proposed nonparametric model; R_3 is the latent factor model with only top three factors; R_4 is the latent factor model where we choose the most important factors accounting for at least 80% variation.**

		$r_1 = 0$			$r_1 = 0.5$		
		$T = 500$	$T = 750$	$T = 1000$	$T = 500$	$T = 750$	$T = 1000$
$n = 50$	ARMSE_α	.1801	.0994	.0517	.0095	.0463	.0120
	ARMSE_β	.2326	.1096	.0567	.1790	.0662	.0415
	R_1	.7363	.7688	.6797	.8683	.8244	.7913
	R_2	.9647	.9547	.9731	.9731	.9441	.9830
	R_3	.9670	.9657	.9724	.9754	.9456	.9821
	R_4	.9864	.9875	.9979	.9947	.9856	.9934
$n = 100$	ARMSE_α	.963	.366	.228	.969	.368	.224
	ARMSE_β	2.64	1.61	1.05	2.14	1.24	1.01
	R_1	.8503	.8488	.8442	.8494	.8479	.8431
	R_2	.9901	.9941	.9942	.9889	.9902	.9935
	R_3	.9962	.0062	.9962	.9950	.9979	.9951
	R_4	.9982	.9982	.9981	.9980	.9979	.9978

We now examine how powerful the proposed hypothesis test is. To evaluate the performance

Table 2: **The ARMISEs of Our Estimation for Unknown Functions**

	$r_1 = 0$			$r_1 = 0.5$		
	$T = 500$	$T = 750$	$T = 1000$	$T = 500$	$T = 750$	$T = 1000$
$n = 50$	.0139	.0126	.0038	.0426	.0163	.0050
$n = 100$	.170	.152	.126	.156	.144	.095

of the proposed hypothesis test, we use the same data generating setting as described earlier and only modified the true functional forms of the factors to be

$$\mathbf{g} = \rho\,(g_1(x_1),\ g_2(x_2),\ g_3(x_3),\ g_4(x_4))^{\mathrm{T}} + (1-\rho)\mathbf{x}, \quad \mathbf{x} = (x_1,\ x_2,\ x_3,\ x_4)^{\mathrm{T}}$$

where each $g_k(\cdot)$ was given as in (5.1). When $\rho = 0$, the null hypothesis (3.1) is true. When ρ is away from zero, the true functional forms of the factors are not identity functions, and we should reject the null hypothesis (3.1).

We set the significance level to be 0.05, and consider the power function of the proposed test for various $n = 50$ and $T = 500$. We carry out 500 simulations for the serial dependence case $r_1 = 0.5$. In each simulation, we generate a data set and apply the proposed hypothesis test to the generated data to test the null hypothesis (3.1). The critical value is computed through a bootstrap sample, of size 1000, of the test statistic λ under null hypothesis. The value of the power function at ρ is defined as the rejection rate of the test among the 500 simulations, and actual size of the test is the value of the power function at $\rho = 0$. The obtained power function is reported in Figure 1, and the actual size is reported to be 0.056. Taking the Monte Carlo error, which is of size $(0.05 \times 0.95/500)^{1/2} \approx 0.01$, into account, we can safely conclude that the actual size of our test is very close to the nominal level. Figure 1 shows the rejection rates approach one as ρ becomes large, indicating that our test has high power to reject the null when it is false. Figure 2 displays the histograms for typical bootstrap samples of the test statistics under the null (left, $\rho = 0$) and the alternative (middle $\rho = .02$ and right $\rho = .05$), respectively.

6 Real data analysis

In this section, we apply the proposed methods to the data set mentioned in Section 1.2. We will show the transformations on the Fama-French three factors are quite necessary for this data set, and construct the transformation needed for each factor by the proposed estimation method. We will also show how much improvement the proposed transformation can result in, in terms of

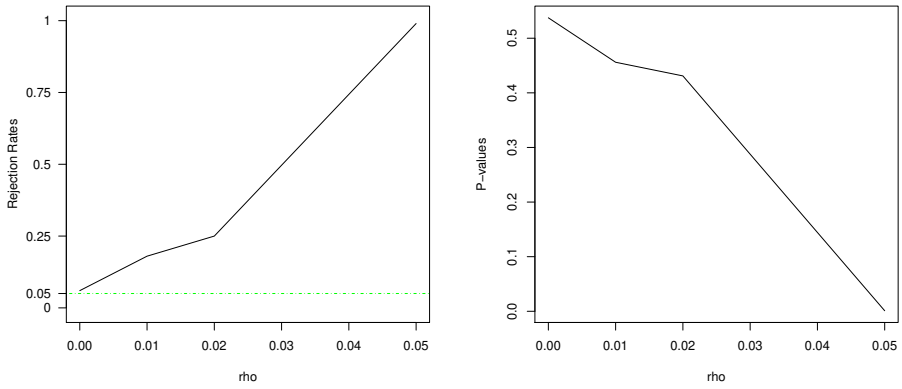

Figure 1: The power function (left) and the average P-value (right) of the proposed test when $n = 50$ and $T = 500$.

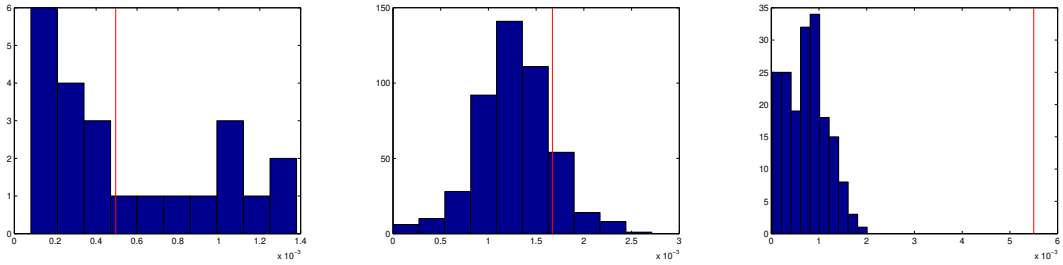

Figure 2: The bootstrap distribution of the test statistic under the null (left) and the alternative hypotheses (middle and right). The three panels correspond to $\rho = 0, 0.02$ and 0.05, respectively. The vertical lines indicate the observed test values for a particular simulation set.

accounting for the return of an asset. We mainly compare our approach with FFTFM. There are a few other *ad hoc* factor models but they are less commonly adopted in practice.

To investigate whether the FFTFM (1.1) is appropriate for this data set, we consider fitting the proposed model (1.2) to the data set described earlier. We remove the first 3000 records since these observations are from a long time ago and thus keep the most recent $T = 892$ observations for analysis. Using the proposed methods, we obtained the estimated functions $\hat{g}_k(\cdot)$ for the three factors and displayed them in Figure 3 along with 95% pointwise confidence intervals.

Figure 3 shows clearly $\hat{g}_k(\cdot)$, $k = 1,\ 2,\ 3$, differ from the identity functions (dotted lines), and $\hat{g}_3(\cdot)$ is not even a linear function. In particular, we observe that the estimated transformation for Rm-Rf factors is decreasing while the estimated transformation for SMB factors is increasing. However, this does not imply that Rm-Rf factors are negatively associated with the response and

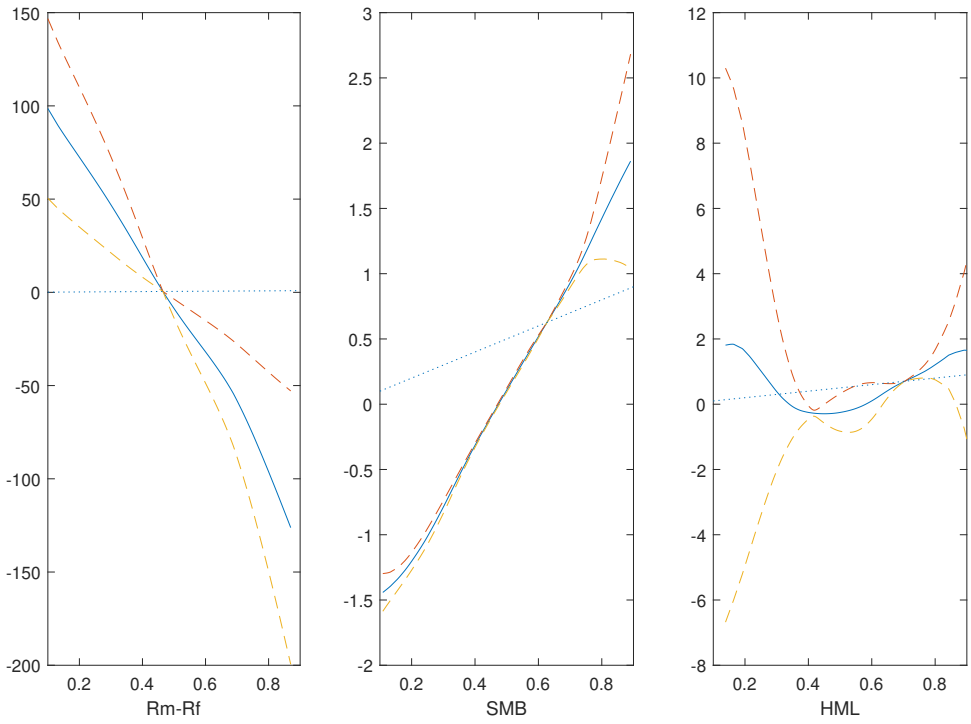

Figure 3: *The solid lines are estimated functions while the dashed lines are the 95% pointwise confidence intervales computed with the bootstrap resampling method. This dotted lines are identity functions.*

SMB factors are positively associated with the response, different from the interpretation for the traditional nonparametric additive model. In fact, the estimated functions serve as a common transformed factors for all the subjects. To obtain the effects on response we need to multiply the estimated coefficients (see Figure 4) for the subjects as well. The effects of HML show a nonlinear pattern, indicating a descending impact with low HML values and then an ascending impact with high HML values. The results in Figure 3 thus suggest some transformation to the original factors may be necessary. In practice, if the suggested transformation is close to a straight line, we may simply employ an identity function for the variable.

We next apply the proposed bootstrap test to this data set to test the null hypothesis (3.1). In this case we obtain a p-value of 0.0960, suggesting that there is weak evidence to reject the null hypothesis of linear model. The p-value is computed through a bootstrap sample, of size 1000, of the test statistic λ resampled under the null hypothesis. We therefore conclude that the FFTFM may fit the return data quite satisfactorily at the significance leve 0.05. Our analysis provides some

solid empirical support to the application of FFTFM in this case.

The estimated coefficients of the three transformed factors, $g_k(x_k)$, $k = 1,\ 2,\ 3$, for all $n = 49$ portforlios are shown in Figure 4. The coefficients for the transformed Rm-Rf, $g_1(x_1)$, were mostly negative and very close to -0.05. The coefficients for the transformed SMB, $g_2(x_2)$, are mostly positive around 0.50 and much greater than those for the transformed Rm-Rf. The coefficients for the transformed HML, $g_3(x_3)$, are not so homogeneous and may be quite different for the individual portforlios.

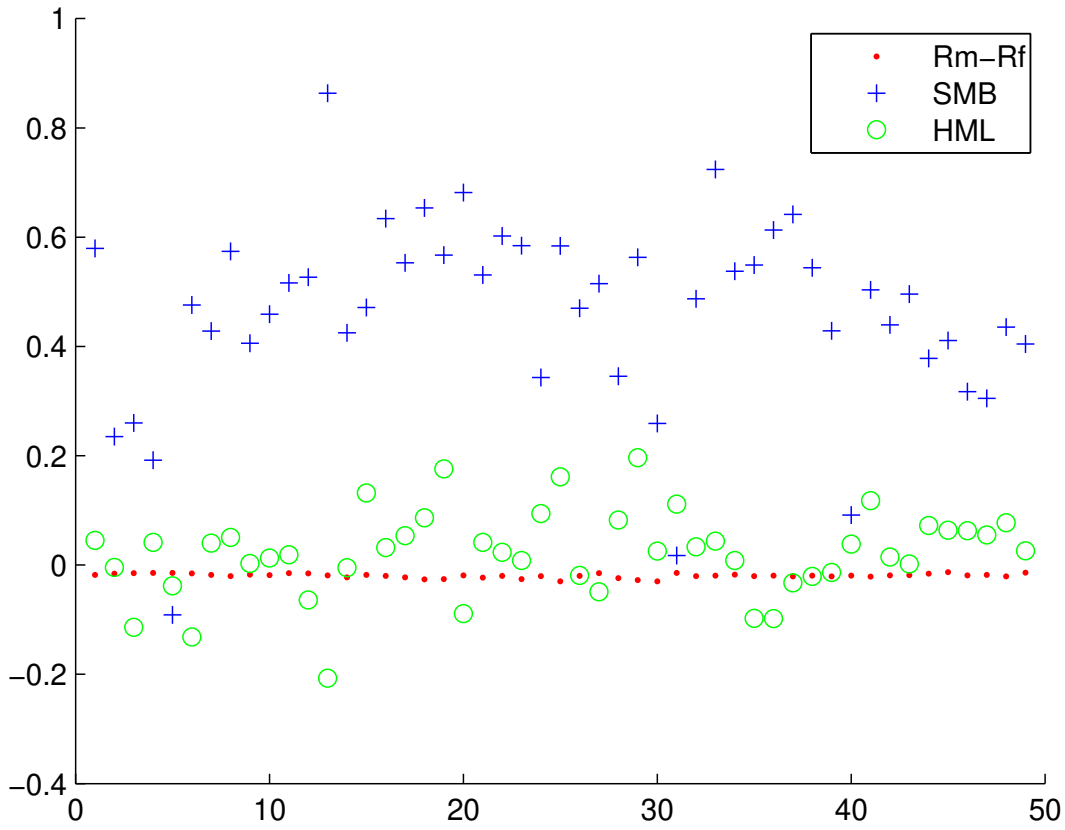

Figure 4: *Estimated coefficients for real data analysis.*

Though FFTFM enjoys good internal validity for this particular data analysis, it is equally important to examine the out-of-sample prediction performance. We now investigate how much improvement the transformed common factors can make in terms of accounting for the predicted return of an asset.

We consider FFTFM, our proposed model and latent factor model in the following. For a given fitted model, let $E_{ji} = (r_{T-i,j} - \hat{r}_{T-i,j})^2$ be the squared error of the prediction for the simple return

of the jth portfolio on the $(T-i)$th day in the sample, based on the fitted model using all the observations before the $(T-i)$th day. We construct such a cross-validation sum for the model for the last 30 days (i.e., $i = 1, \cdots, 30$), and define the prediction error as

$$\text{CV} = \frac{1}{30 \times 49} \sum_{i=1}^{30} \sum_{j=1}^{49} E_{ji}.$$

We compute, respectively, the CVs for the FFTFM and the proposed model (1.2), and find the ratio of CV of the FFTFM to the CV of the proposed model is 1.3694. Similarly we fit the latent factor model with factors selected to account for 80% sample variation. The ratio of CV of the latent factor model to the CV of the proposed model is 1.3505. These results indicate the proposed model can make more than 35% improvement in terms of accounting for the return of an asset. Our model is certainly better than the FFTFM and the latent factor model in terms of out-of-sample prediction. The FFTFM and the latent factor model, though performing very well for in-sample prediction, may not perform as accurately as our method for future forecasting.

In addition to the prediction on the exact numerical values, we also report in Table 3 the prediction accuracy on whether the next day return is increasing (up) or decreasing (down) from the current day. The high true positive and negative rates resulted from our proposed model confirm the superiority of our method. Since forecasting is a central issue for portfolio analysis, our analysis demonstrates the practical importance of the proposed model.

Table 3: **The Prediction Accuracy for Next Day Up-Down: Up means the price increases and Down means the price decreases; True means the predicted direction agrees with the actual direction and False means the two disagree.**

Method	True Up	Ture Down	False Up	False Down
Our model	0.8525	0.7449	0.2551	0.1475
FFTFM	0.8127	0.6410	0.3590	0.1873
Latent factor model	0.7965	0.6815	0.3185	0.2035

7 Discussion

There are all kinds of advantages of using Famma-French factors. However, the model has very restrictive parametric form and may not always hold in practice. When the data do not follow the model the estimation and prediction performance may be less satisfactory. In order to improve the

estimation and prediction performance we recommend to use our proposed flexible nonparametric model with nonlinear transformation. Even though the nonparametric functions may not be easily interpreted, they can be used to provide better numerical results, especially for forecasting.

We note that to introduce nonparametric structure to the FFTFM there could be many alternative specifications other than what we choose in this paper. For example, one may consider the self-modelling approach widely practiced for shape invariant models. Altman and Villarreal (2004) considered such a nonlinear semiparametric regression approach and proposed an efficient algorithm for parameter estimation. One computational advantage of this method is that one can exploit off-the-shelf software for fitting the nonlinear mixed effects models (Lindstrom and Bates (1990), Ke and Wang (2001)). The theoretical justification of this approach relies on the theory of profile likelihood (Murphy and van der Vaart (2000)).

The nonparametric transformations for the three factors are combined linearly in the model and thus resembles the basic structure of the FFTFM. The model interpretation may thus follow the FFTFM after the transformation is applied. It would be interesting to consider a more sophisticated nonlinear combination of all the factors. However, such a model might be even harder to interpret, not to mention the computational difficulty. The current semiparametric model may strike a balance between the parametric FFTFM and the fully nonparametric model.

We do not examine heterogeneity problem in this paper. In principle, we may also extend the works by Guo, Box and Zhang (2016) using the methodology proposed in this paper. Further, we do not discuss cluster effects in our modelling. A possible solution is to add a random-effects term (Palta (2003), Demidenko (2004)) and then apply the likelihood estimation methods. It is also quite interesting to extend our approaches to incorporate spatio-temporal modelling. In addition to local polynomial fitting, one may carry out basis approximation such as spline basis or Fourier basis (Xu et al. (2017)). We have implemented such estimation methods for a few numerical examples and found quite similar performance. However, the theoretical justification for basis approximation methods requires a non-trivial development. All these issues will be included in our future work.

References

Altman, N. S. and Villarreal, J. C. (2004). Self-modelling regression for longitudinal dta with time-invariant covariates. *Canadian Journal of Statistics* **32**, 251-268.

Ang, A., Hodrick, R. J., Xing, Y. and Zhang, X. (2006). The crosssection of volatility and expected returns. *The Journal of Finance*, **61**, 259-299.

Bai, J. and Ng, S. (2002). Determing the number of factors in approximate factor models. *Econometrica* **70**, 191-221.

Brennan, M. J., Chordia, T. and Subrahmanyam, A. (1998). Alternative factor specifications, security characteristics, and the cross-section of expected stock returns. *Journal of Financial Economics*, **49**, 345-373.

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models. *Ann. Statist* **17**, 543–555.

Burges, C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**, 121-167.

Cristianini, N., Kandola, J., Elisseeff, A., & Taylor, J. (2006). *On kernel target alignment. Innovations in machine learning: theory and applications*, Berlin: Springer.

Davis, J. L., Fama, E. F. and French, K. R. (2000). Characteristics, covariances, and average returns: 1929 to 1997. *The Journal of Finance*, **55**, 389-406.

Demidenko, E. (2004). *Mixed Models: Theory and Applications.* John Wiley.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics*, **49**, 283-306.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, **33**, 3-56.

Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, **51**, 55-84.

Fama, E. F. and French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance*, **65**, 1915-1947.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, **116**, 1-22.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics*, **147**, 186-197.

Fan, J., Feng, Y. and Xia, L. (2015). A conditional dependence measure with applications to undirected graphical models. arXiv:1501.01617.

Fan, J. and Jiang, J. (2005) Nonparametric inferences for additive models. *Journal of the American Statistical Association* **100**, 890-902.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153-93.

Guo, S., Box, J. and Zhang, W. (2016). A dynamic structure for high dimensional covariance matrices and its application in portfolio allocation. *Journal of the American Statistical Association*, to appear.

Ince H, Trafalis, TB (2006) Kernel methods for short-term portfolio management. *Expert Systems with Applications* **30**, 535-542.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association* **96**, 1272-1298.

Kong, E., Linton, O. and Xia, Y. (2010) Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory* **26**, 1529-1564.

Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., & Jordan, M. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, **5**, 27-72.

Le Cam, L. and Yang, G. (1990) *Asymptotic in Statistics: Some Basic Concepts.* New York: Springer-Verlag.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*. **46**, 673-687.

Masry, E. (1996a) Multivariate regression estimation: local polynomial fitting for time series. *Stochastic Processes and Their Applications* **65**, 81-101.

Masry, E. (1996b) Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis* **17**, 571-599.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95**, 449-482.

Ong, C., Smola, A., & Williamson, R. (2005). Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, **6**, 1043-1071.

Opsomer, J. (2000) Asymptotic Properties of Backfitting Estimators. *Journal of Multivariate Analysis* **73**, 166-179.

Opsomer, J. and Ruppert, D. (1997) Fitting a bivariate additive model by local polynomial regression. *Ann. Statist* **25**, 186–211.

Palta, M. (2003). *Quantitative Methods in Population Health.* John Wiley.

Petkova, R. (2006). Do the FamaFrench factors proxy for innovations in predictive variables? *The Journal of Finance*, **61**, 581-612.

Schebesch KB, Stecking R (2005) Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society* **56**, 1082-1088.

Vassalou, M. and Xing, Y. (2004). Default risk in equity returns. *The Journal of Finance*, **59**, 831-868.

Xu, M., Li, J., Chen, Y. (2017). Varying coefficient functional autoregressive model with application to the US Treasuries. *Journal of Multivariate Analysis.* **159**, 168-183.

Yu, B. (1994) Rates of convergence for empirical processes of stationary mixing sequences. *Annals of probability* **22**, 94-116.

Appendix

It is clear from the estimation procedure as described in Section 2.1 that the statistical properties of the estimated component functions $g_k(.)$ as well as those of $\hat{\alpha}_j$, $\hat{\beta}_{jk}$ could only be derived based on the asymptotics concerning the backfitting estimators $\hat{G}_{jk}(.)$. To present the relevant results on this aspect, we need to introduce more notations. Let $f(.)$ be the joint density function of $(x_{t1}, \cdots, x_{tp})$, and $f_k(.)$, $k = 1, \cdots, p$, the marginal density of the kth covariate x_{tk}. Denote by $f_{l;k}(.,.)$, the joint density of x_{tk} and $x_{(t+l)k}$; $f_{l;k,k'}(u,u,v,v)$, the joint pdf of $x_{tk}, x_{(t+l)k}, x_{tk'}, x_{(t+l)k'}$ evaluated at (u,u,v,v). For any $l \geq 1$, $k, k' = 1, \cdots, p, k \neq k'$, define

$$a_{l;k} = \int \frac{f_{l;k}(u,u)}{f_k^2(u)} du, \; b_{l;k,k'} = \int \frac{f_{l;k,k'}(u,u,v,v)}{f_k(u) f_{k'}(v)} du dv.$$

We assume that

$$c_k := \lim_{T\to\infty} \left| \frac{1}{T^2} \sum_{l=1}^{T-1} (T-l) a_{k,l} \right| < \infty, \quad \lim_{T\to\infty} \sup_{k\neq k'} \left| \frac{1}{T^2} \sum_{l=1}^{T-1} (T-l) b_{l;k,k'} \right| < \infty;$$

The following conditions are assumed throughout of the paper. First of all, we assume that for any given t, $\{\epsilon_{tj}, j = 1, \cdots, n\}$ is independent of each other. Such requirement is quite common in literature on panel data, since any correlation among the returns of different assets(stocks) are captured through the presence of common factors. Write $\boldsymbol{\epsilon}_t = (\epsilon_{t1}, \cdots, \epsilon_{tn})^\top$, $\mathbf{x}_t = (x_{t1}, \cdots, x_{tp})^\top$.

[A1] $\{\mathbf{x}_t, \boldsymbol{\epsilon}_t\}$ is strictly stationary and strongly mixing, i.e.

$$\gamma[\iota] := \sup_{\substack{A \in \mathbf{F}_{-\infty}^{0} \\ B \in \mathbf{F}_{\iota}^{\infty}}} |P[AB] - P[A]P[B]| \to 0, \text{ as } k \to \infty,$$

where $\mathbf{F}_{s_1}^{s_2}$ is the $\sigma-$ algebra of events generated by $\{\mathbf{x}_t : s_1 \leq t \leq s_2\}$ and $\gamma[\iota]$ is referred to as the strong mixing coefficient. Moreover, $\sum_{\iota=1}^{\infty} \iota^a \gamma[\iota]^{1-2/v} < \infty$ for some $v > 2$ and $a > 1 - 2/v$.

[A2] The kernel function $K(.)$ is bounded and continuous with a compact support; its first order derivative has a finite number of sign changes over its support.

[A3] Both the joint $f(.)$ and the marginal densities $f_k(.)$, $k = 1, \cdots, p$ are bounded and continuous with compact support; their first order derivatives also have a finite number of sign changes over their supports.

[A4] $\sup_{u,u'} |f_{l;k}(u, u') - f_k(u) f_k(u')| \leq A_1 < \infty$ for all $l \geq 1$.

[A5] As $T \to \infty$, $h_k \to 0$, $Th_k / \log T \to \infty$, $Th_k^{\iota_k+2} \to 0$ for all $k = 1, \cdots, p$.

[A6] There exists a sequence v_n of positive integers satisfying $v_{\mathrm{T}} \to \infty$ and $v_{\mathrm{T}} = o((nh)^{1/2})$ such that $(T/h)^{1/2} \gamma[v_{\mathrm{T}}] \to 0$ as $T \to \infty$.

Assumption [A1] is relevant since the backfitting estimator $\hat{G}_{jk}(.)$ is built on dependent observations, $\{r_{tj}, t = 1, \cdots, T\}$. Note that while Opsomer (2000) dealt with independent observations, the results he obtained are valid for time series sequence as well, as long as the dependence decreases quickly enough, such as described in [A1]. This has been made obvious by plenty of literature devoted to kernel smoothing for time series; e.g. Masry (1996a, 1996b), Kong et al. (2010). Strongly mixing could be replaced by a weaker condition, such as $\beta-$mixing or even $\phi-$mixing, but in that case additional requirement on these alternative mixing coefficients will then be necessary; see e.g. Masry (1996). [A2] could be relaxed to allows kernel functions of unbounded support provided that $u^{\iota_k+1} K(u) \to 0$ as $u \to \infty$.

For $l = 0, 1, \cdots$, write the lth moment of the kernel function $K(.)$ as $\mu_l(K) := \int u^l K(u) du$ and $R_l = \int u^l K^2(u) du$, and $R(K) = R_0$. For $k = 1, \cdots, p$, let $g_k^{(\iota)}(.)$ denote the ιth derivative of

component function $g_k(.)$, and write

$$\mathbf{g}_k^{(\iota)} = \begin{bmatrix} g_k^{(\iota)}(x_{1k}) \\ \vdots \\ g_k^{(\iota)}(x_{\mathrm{T}k}) \end{bmatrix}, \quad E(\mathbf{g}_k^{(\iota)}|\mathbf{X}^k) = \begin{bmatrix} E(g_k^{(\iota)}(x_{ik})|x_{1k}) \\ \vdots \\ E(g_k^{(\iota)}(x_{ik})|x_{\mathrm{T}k}) \end{bmatrix}, \ k \neq k'.$$

The backfitting algorithm described in Section 2.1 is based on local linear smoothing. Here we give a more general results on backfitting estimators based on local polynomial smoothing where functions $g_k(.)$ are locally approximated by a polynomial of degree ι_k, $k = 1, \cdots, p$. Define the following smoother matrix for the kth component function:

$$\mathbf{S}_k = (\mathbf{S}_{k,x_{1k}}, \cdots, \mathbf{S}_{k,x_{\mathrm{T}k}})^\top, \tag{A.1}$$

where $\mathbf{S}_{k,u}$ represents the transpose of the equivalent kernel for the kth covariate at the point u:

$$\mathbf{S}_{k,u} = \mathbf{K}_k(u)\mathbf{X}_k(u)\Big[\mathbf{X}_k(u)^\top \mathbf{K}_k(u)\mathbf{X}_k(u)\Big]^{-1}\mathbf{e}_{1k}^\top,$$

$\mathbf{e}_{1k}$ is the $(\iota_k + 1) \times 1$ vector with a one in the first position and zeros elsewhere,

$$\mathbf{X}_k(u) = \begin{bmatrix} 1 & x_{1k} - u & \cdots & (x_{1k} - u)^{\iota_k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{\mathrm{T}k} - u & \cdots & (x_{\mathrm{T}k} - u)^{\iota_k} \end{bmatrix}, \quad \mathbf{K}_k(u) = \mathrm{diag}\left(K_h(x_{1k} - u), \ \cdots, \ K_h(x_{Tk} - u)\right).$$

Further define the centered smoothing matrix $\mathbf{S}_k^* = (\mathbf{I} - \mathbf{1}_{\mathrm{T}}\mathbf{1}_{\mathrm{T}}^\top)\mathbf{S}_k$, $\mathbf{W}_{[-k]}$, the smoother matrix for the $(p-1)$-variate function $G_j^{(-k)}(.) = \sum_{l=1, l\neq k}^{p} G_{jl}(.)$, and $\mathbf{G}_{jk} = (G_{jk}(x_{1k}), \cdots, G_{jk}(x_{\mathrm{T}k}))^\top$, the vector of the kth component function evaluated at the observation points. Then regarding $\hat{\mathbf{G}}_{jk}$, the backfitting estimator of $\mathbf{G}_{jk}$, we have

Corollary 7.1 *Given* $\mathbf{X}$, *the conditional bias and variance of* $\hat{\mathbf{G}}_{jk}$, $j = 1, \cdots, n$, $k = 1, \cdots, p$, *are respectively*

$$\begin{aligned} E(\hat{\mathbf{G}}_{jk} - \mathbf{G}_{jk}|\mathbf{X}) &= (\mathbf{I} - \mathbf{S}_k^*\mathbf{W}_{[-k]})^{-1}\Big[\frac{1}{(\iota_k+1)!}h^{\iota_k+1}\mu_{\iota_k+1}(K)\beta_{jk}\Big(\mathbf{g}_k^{(\iota_k+1)} - E(\mathbf{g}_k^{(\iota_k+1)})\Big) - \mathbf{S}_k^*\mathbf{B}_{j[-k]}\Big] \\ &\quad + O_p(T^{-1/2}) + o_p(h^{\iota_k+1}), \\ Var(\hat{\mathbf{G}}_{jk}(x_{tk})|\mathbf{X}) &= \{nhf_k(x_{tk})\}^{-1}R_K\sigma^2 + o_p((nh)^{-1},) \end{aligned}$$

where

$$\mathbf{B}_{j[-k]} = E\Big(\mathbf{W}_{[-k]}(\mathbf{R}_j - \mathbf{G}_{jk})|\mathbf{X}\Big) - \sum_{l=1:l\neq k}^{p} \mathbf{G}_{jl}.$$

The bias expression in Corollary 7.1 is still a recursive formula, and as commented in Opsomer (2000), a non-recursive asymptotic bias expression can be derived, but the expressions become very complicated even for $p = 3$. Nevertheless, the order of the asymptotic bias could be easily decided for any p:

$$E(\hat{\mathbf{G}}_{jk} - \mathbf{G}_{jk}|\mathbf{X}) = O_p(\sum_{k=1}^{p} h_k^{\iota_k+1}).$$

Apparently, if $g_k(.)$, $k = 1, \cdots, p$ are all smooth enough, and with polynomial fitting of high enough ι_k degrees employed, this bias term could be made relatively negligible compared to asymptotic stochastic error. We will make use of this fact in later sections in the asymptotic study of $\hat{g}_k(.)$, and $\hat{\boldsymbol{\beta}}_j$.

We now move on to prove Theorem 4.1, starting with more notations. Let

$$c_0(K) = \sum_{\iota=0}^{\iota_k} [\mathbf{N}^{-1}]_{(\iota+1)1} \mu_\iota(K),$$

where $\mathbf{N}$ represents the $(\iota_k+1)\times(\iota_k+1)$ matrix, whose (i,j)th element is $\mu_{i+j-2}(K)$, and $[\mathbf{N}^{-1}]_{(\iota+1)1}$ stands for the $(\iota+1,1)$th element of its inverse matrix. Define the $T \times 1$ vectors

$$\gamma_{tk} = (-g_k(x_{tk}), 0, \cdots, 0, 1, 0, \cdots, 0)^\top, \quad t = 2, \cdots, T$$

with 1 as the tth entry. For any given $k, k' = 1, \cdots, p$, define

$$c_{k,k'}(u) = E[g_k(x_{tk})|x_{tk'} = u], \quad \mathbf{c}_{k,k'} = [c_{k,k'}(x_{1k'}), \cdots, c_{k,k'}(x_{\mathrm{T}k')}]^\top,$$

$$\mathbf{A}_{j'|j} = [\mathbf{a}_{1j'|j}, \cdots, \mathbf{a}_{pj'|j}]^\top, \quad \mathbf{a}_{kj'|j} = \sum_{k'=1}^{p} \frac{\beta_{jk'}}{\beta_{j'k'}} \mathbf{c}_{k,k'}, \quad j, j' = 1, \cdots, n; \ k = 1, \cdots, p.$$

Proof of Theorem 4.1 Similar computations as in the proof of the second assertion of Corollary 7.1 lead to

$$\hat{\mathbf{G}}_{jk} - E\hat{\mathbf{G}}_{jk} = \mathbf{S}_k \boldsymbol{\epsilon}_j + O_p(T^{-1/2}), \quad j = 1, \cdots, n, \ k = 1, \cdots, p,$$

uniformly in over all elements of the matrices; see, also Opsomer (2000, pp. 178). For ease of exposition, write the asymptotic bias and stochastic error of $\hat{\mathbf{G}}_{jk}$ as

$$\mathbf{b}_{jk} = E\hat{\mathbf{G}}_{jk} - \mathbf{G}_{jk} \equiv (b_{jk,1}, \cdots, b_{jk,\mathrm{T}})^\top, \quad \mathbf{v}_{jk} = \mathbf{S}_k \boldsymbol{\epsilon}_j \equiv (v_{jk,1}, \cdots, v_{jk,\mathrm{T}})^\top.$$

As a result, we have

$$\begin{aligned}
\frac{\hat{G}_{jk}(x_{tk})}{\hat{G}_{jk}(x_{1k})} &= \frac{\beta_{jk} g_k(x_{tk}) + b_{jk,t} + v_{jk,t}}{\beta_{jk} + b_{jk,1} + v_{jk,1}} \\
&= g_k(x_{tk}) + \frac{b_{jk,t}}{\beta_{jk}} + \frac{v_{jk,t}}{\beta_{jk}} - \frac{g_k(x_{tk}) b_{jk,1}}{\beta_{jk}} - \frac{g_k(x_{tk}) v_{jk,1}}{\beta_{jk}} + o_p(h_k^{\iota_k+1} + T^{-1/2}).
\end{aligned}$$

Since without loss of generality, we could always assume that $x_{1k} = 1$ and whence for each $t = 2, \cdots, T$,

$$\begin{aligned}\bar{g}_k(x_{tk}) &= \frac{1}{n}\sum_{j=1}^{n}\hat{G}_{jk}(x_{tk})/\hat{G}_{jk}(x_{1k}) \\ &= g_k(x_{tk}) + \frac{1}{n}\sum_{j=1}^{n}\Big(\frac{b_{jk,t}}{\beta_{jk}} - \frac{g_k(x_{tk})b_{jk,1}}{\beta_{jk}}\Big) \\ &\quad + \frac{1}{n}\sum_{j=1}^{n}\Big(\frac{v_{jk,t}}{\beta_{jk}} - \frac{g_k(x_{tk})v_{jk,1}}{\beta_{jk}}\Big) + o_p(h_k^{\iota_k+1} + T^{-1/2}), \end{aligned} \tag{A.2}$$

again uniformly in t and k.

Since the second (bias) term on the RHS of (A.2) is of order $o(T^{-1/2})$ if $g_k(.)$ is smooth enough and a large enough ι_k is used, we have

$$\bar{g}_k(x_{tk}) = g_k(x_{tk}) + \gamma_{tk}^{\top}\mathbf{S}_k\frac{1}{n}\sum_{j=1}^{n}\beta_{jk}^{-1}\boldsymbol{\epsilon}_j + o_p(T^{-1/2}).$$

Since $\boldsymbol{\epsilon}_j$, $j = 1, \cdots, n$ are all iid errors with zero mean and variance-covariance matrix V_T, the asymptotic variance of $\hat{g}_k(x_{tk})$ is such that

$$\Big(n^{-2}\sum_{j=1}^{n}\beta_{jk}^{-2}\Big)\gamma_{tk}^{\top}\mathbf{S}_k V_T \mathbf{S}_k^{\top}\gamma_{tk}. \tag{A.3}$$

Using standard results in polynomial smoothing (Masry, 1996) together with the fact that

$$[\mathbf{S}_k]_{ij} = \{f_k(x_{ik})\}^{-1}\frac{1}{Th_k}\sum_{\iota=0}^{\iota_k}[\mathbf{N}^{-1}]_{(\iota+1)1}\Big(\frac{x_{jk}-x_{ik}}{h_k}\Big)^{\iota}K\Big(\frac{x_{jk}-x_{ik}}{h_k}\Big), \tag{A.4}$$

we have

$$[\mathbf{S}_k V_T \mathbf{S}_k^{\top}]_{ii'} = \{f_k(x_{ik})f_k(x_{i'k})\}^{-1}\frac{\sigma^2}{Th_k}\sum_{\iota,\iota'=0}^{\iota_k}[\mathbf{N}^{-1}]_{(\iota+1)1}[\mathbf{N}^{-1}]_{(\iota'+1)1}R(i,i';\iota,\iota') + O_p((Th_k)^{-3/2})$$

where

$$R(i,i';\iota,\iota') = \int\Big(\frac{x_{ik}-x_{i'k}}{h_k} + t\Big)^{\iota'}t^{\iota}K(t)K(s+t)dt.$$

Therefore,

$$\begin{aligned}&\gamma_{tk}^{\top}\mathbf{S}_k = ([\mathbf{S}_k]_{tj} - g_k(x_{tk}) * [\mathbf{S}_k]_{1j}) = O((Th_k)^{-1}) \\ &\gamma_{tk}^{\top}\mathbf{S}_k V_T \mathbf{S}_k^{\top}\gamma_{tk} = \{g_k(x_{tk})\}^2[\mathbf{S}_k\mathbf{S}_k^{\top}]_{11} - 2[\mathbf{S}_k\mathbf{S}_k^{\top}]_{1t}g_k(x_{tk}) + [\mathbf{S}_k\mathbf{S}_k^{\top}]_{tt}\end{aligned}$$

This together with (A.3) implies that the asymptotic variance of $\hat{g}_k(x_{tk})$ is of order $O((Th_k)^{-1})$.

As for the estimates of the parameters, first note that the results on $\hat{\alpha}_j$ easily follow from (1.2), (1.3) and the strong mixing conditions [A1]. To examine the asymptotic properties of $\hat{\beta}_{jk}$, *least*

square estimate (2.6) derived from model (2.5), first note that according to Theorem 4.1, we have that

$$\bar{\mathbf{g}} = \mathbf{g} + O_p((Th_k)^{-1/2}), \quad (\frac{1}{T}\bar{\mathbf{g}}^\top\bar{\mathbf{g}})^{-1} = \Sigma_g^{-1} + O_p((Th_k)^{-1/2}), \tag{A.5}$$

uniformly in all elements of the matrix, where

$$\mathbf{g} = \begin{bmatrix} g_1(x_{11}) & \cdots & g_p(x_{1p}) \\ g_1(x_{21}) & \cdots & g_p(x_{2p}) \\ \vdots & \vdots & \vdots \\ g_1(x_{\mathrm{T}1}) & \cdots & g_p(x_{\mathrm{T}p}) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ g_1(x_{21}) & \cdots & g_p(x_{2p}) \\ \vdots & \vdots & \vdots \\ g_1(x_{\mathrm{T}1}) & \cdots & g_p(x_{\mathrm{T}p}) \end{bmatrix},$$

since without loss of generality, we have assumed that $x_{1k} = 1$ whence $g_k(x_{1k}) = x_{1k} = 1$. These, together with the decomposition $R_j = \hat{\alpha}_j\mathbf{1}_\mathrm{T} + (\alpha_j - \hat{\alpha}_j)\mathbf{1}_\mathrm{T} + \hat{\mathbf{g}}\boldsymbol{\beta}_j + (\mathbf{g} - \hat{\mathbf{g}})\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j$ and the root-T consistency of $\hat{\alpha}_j$, lead to

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_j &= (\bar{\mathbf{g}}^\top\bar{\mathbf{g}})^{-1}\bar{\mathbf{g}}^\top(R_j - \hat{\alpha}_j\mathbf{1}_\mathrm{T}) \\
&= \boldsymbol{\beta}_j + (\bar{\mathbf{g}}^\top\hat{\mathbf{g}})^{-1}\bar{\mathbf{g}}^\top(\alpha_j - \hat{\alpha}_j)\mathbf{1}_\mathrm{T} + (\bar{\mathbf{g}}^\top\hat{\mathbf{g}})^{-1}\bar{\mathbf{g}}^\top(\mathbf{g} - \bar{\mathbf{g}})\boldsymbol{\beta}_j + (\bar{\mathbf{g}}^\top\bar{\mathbf{g}})^{-1}\bar{\mathbf{g}}^\top(\mathbf{g} - \hat{\mathbf{g}})\boldsymbol{\epsilon}_j \\
&= \boldsymbol{\beta}_j + (\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top(\alpha_j - \hat{\alpha}_j)\mathbf{1}_\mathrm{T} + (\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top(\mathbf{g} - \bar{\mathbf{g}})\boldsymbol{\beta}_j + o_p(T^{-1/2}) \\
&= \boldsymbol{\beta}_j + \Sigma_g^{-1}T^{-1}\mathbf{g}^\top(\mathbf{g} - \bar{\mathbf{g}})\boldsymbol{\beta}_j + \Sigma_g^{-1}T^{-1}\mathbf{g}\boldsymbol{\epsilon}_j + o_p(T^{-1/2})
\end{aligned}$$

where we've used the following facts:

$$T^{-1}\mathbf{g}^\top\mathbf{1}_\mathrm{T} = O_p(T^{-1/2}), \quad T^{-1}(\mathbf{g} - \bar{\mathbf{g}})\boldsymbol{\epsilon}_j = O_p(T^{-1/2}).$$

This means the error arisen from the pre-estimation of α_j has been 'averaged out' and thus of no impact. To show that $\hat{\boldsymbol{\beta}}_j$ is asymptotically normal, first note that the kth element of $\mathbf{g}^\top(\mathbf{g} - \bar{\mathbf{g}})\boldsymbol{\beta}_j$ is given by

$$\begin{aligned}
&\frac{1}{n}\sum_{j'=1}^{n}\sum_{k'=1}^{p}\frac{\beta_{jk'}}{\beta_{j'k'}}\Big[\sum_{t=2}^{T}g_k(x_{tk})\gamma_{tk'}\Big]^\top\mathbf{S}_{k'}\boldsymbol{\epsilon}_{j'} \quad k = 1, \cdots, p; \quad \text{with} \\
&\sum_{t=2}^{T}g_k(x_{tk})\gamma_{tk'} = \Big[-\sum_{t=2}^{T}g_k(x_{tk})g_{k'}(x_{tk'}),\, g_k(x_{2k}), \cdots, g_k(x_{Tk})\Big]^\top.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Big[\sum_{t=2}^{T}g_k(x_{tk})\gamma_{tk'}\Big]^\top\mathbf{S}_{k'} &= c_0(K)\mathbf{c}_{k,k'}^\top + O_p((Th_k)^{-1/2}) \\
\tfrac{1}{n}\sum_{j'=1}^{n}\sum_{k'=1}^{p}\tfrac{\beta_{jk'}}{\beta_{j'k'}}\Big[\sum_{t=2}^{T}g_k(x_{tk})\gamma_{tk'}\Big]^\top\mathbf{S}_{k'}\boldsymbol{\epsilon}_{j'} &= c_0(K)\tfrac{1}{n}\sum_{j'=1}^{n}\Big[\sum_{k'=1}^{p}\tfrac{\beta_{jk'}}{\beta_{j'k'}}\mathbf{c}_{k,k'}\Big]^\top\boldsymbol{\epsilon}_{j'} + o_p(T^{1/2}).
\end{aligned}$$

Since $\boldsymbol{\epsilon}_{j'}$, $j' = 1, \cdots, n$ are independent with zero mean and variance V_T, the asymptotic normality of $T^{1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ thus follows with asymptotic variance given by

$$\sigma^2 c_0^2(K)\Sigma_g^{-1} n^{-2}\Big(\sum_{j'=1}^{n} T^{-1}\mathbf{A}_{j'|j}\mathbf{A}_{j'|j}^{\top}\Big)\Sigma_g^{-1},$$

which is finite. □

Proof of Theorem 4.2 Firstly, it is easy to see that $\mathrm{RSS}_1/(nT) \to \sigma^2$ in probability as $T \to \infty$. So we just need to concern us with the numerator for which we have

$$\mathrm{RSS}_0 - \mathrm{RSS}_1 = \sum_{j=1}^{n}\{\mathrm{RSS}_{0,j} - \mathrm{RSS}_{1,j}\},$$

where

$$\mathrm{RSS}_{1,j} = R_j^{\top}[\mathbf{I}_T - \tilde{\mathbf{g}}(\tilde{\mathbf{g}}^{\top}\tilde{\mathbf{g}})^{-1}\tilde{\mathbf{g}}^{\top}]R_j^{\top}; \quad \tilde{\mathbf{g}} = \begin{pmatrix} 1 & \bar{g}_1(x_{11}) & \cdots & \bar{g}_p(x_{1p}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \bar{g}_1(x_{T1}) & \cdots & \bar{g}_p(x_{Tp}) \end{pmatrix}$$
$$\mathrm{RSS}_{0,j} = R_j^{\top}[\mathbf{I}_T - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}]R_j^{\top} = \boldsymbol{\epsilon}_j^{\top}[\mathbf{I}_T - \bar{\mathbf{X}}(\bar{\mathbf{X}}^{\top}\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}^{\top}]\boldsymbol{\epsilon}_j^{\top}.$$

Note that the second identity follows from the fact that $\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}$ is invariant if $\mathbf{X}$ is replaced with $\mathbf{X}$ right-multiplied by a diagonal matrix and that

$$\bar{\mathbf{X}} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & g_1(x_{21}) & \cdots & g_p(x_{2p}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & g_1(x_{\mathrm{T}1}) & \cdots & g_p(x_{\mathrm{T}p}) \end{bmatrix} = \mathbf{X}\begin{bmatrix} 1 & x_{11}^{-1} & 0 & \cdots & 0 \\ 0 & 0 & x_{12}^{-1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & x_{1p}^{-1} \end{bmatrix}.$$

With a slight abuse of notation, we revert to the old notation of $\mathbf{g}$ in place of $\bar{\mathbf{X}}$. Write $\tilde{\mathbf{g}} = \mathbf{g} + \delta$, $\Delta = \mathbf{g}^{\top}\delta + \delta^{\top}\mathbf{g}$, $\Gamma = (\mathbf{g}^{\top}\mathbf{g})^{-1}\mathbf{g}^{\top}$ so that

$$\begin{aligned}
&\tilde{\mathbf{g}}^{\top}\tilde{\mathbf{g}} = \mathbf{g}^{\top}\mathbf{g} + \mathbf{g}^{\top}\delta + \delta^{\top}\mathbf{g} + \delta^{\top}\delta, \\
&(\tilde{\mathbf{g}}^{\top}\tilde{\mathbf{g}})^{-1} = (\mathbf{g}^{\top}\mathbf{g})^{-1} - (\mathbf{g}^{\top}\mathbf{g})^{-1}\Delta(\mathbf{g}^{\top}\mathbf{g})^{-1} + O_p((Th_k)^{-1}), \\
&\tilde{\mathbf{g}}(\tilde{\mathbf{g}}^{\top}\tilde{\mathbf{g}})^{-1}\tilde{\mathbf{g}}^{\top} = \mathbf{g}(\mathbf{g}^{\top}\mathbf{g})^{-1}\mathbf{g}^{\top} + \delta\Gamma + \Gamma^{\top}\delta^{\top} + \delta(\mathbf{g}^{\top}\mathbf{g})^{-1}\delta^{\top} - \Gamma^{\top}\Delta\Gamma \\
&\quad -\delta(\mathbf{g}^{\top}\mathbf{g})^{-1}\Gamma(\mathbf{g}^{\top}\mathbf{g})^{-1}\mathbf{g}^{\top} - \mathbf{g}(\mathbf{g}^{\top}\mathbf{g})^{-1}\Gamma(\mathbf{g}^{\top}\mathbf{g})^{-1}\delta^{\top} - \delta(\mathbf{g}^{\top}\mathbf{g})^{-1}\Gamma(\mathbf{g}^{\top}\mathbf{g})^{-1}\delta^{\top} + O_p((Th_k)^{-1}).
\end{aligned}$$

Since $R_j = \mathbf{g}\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j$, we have the following partition of the difference of the two Residual Sum of Squares:

$$\begin{aligned}
\mathrm{RSS}_{0,j} - \mathrm{RSS}_{1,j} = &-2R_j^{\top}\delta\Gamma R_j + R_j^{\top}\Gamma^{\top}\Delta\Gamma R_j - R_j^{\top}\delta(\mathbf{g}^{\top}\mathbf{g})^{-1}\delta^{\top}R_j \\
&+2R_j^{\top}\mathbf{g}(\mathbf{g}^{\top}\mathbf{g})^{-1}\Delta(\mathbf{g}^{\top}\mathbf{g})^{-1}\delta^{\top}R_j + R_j^{\top}\delta(\mathbf{g}^{\top}\mathbf{g})^{-1}\Delta(\mathbf{g}^{\top}\mathbf{g})^{-1}\delta^{\top}R_j. \quad \text{(A.6)}
\end{aligned}$$

We start with the third term on the RHS of (A.6), and will show that

$$R_j^\top \delta(\mathbf{g}^\top\mathbf{g})^{-1}\delta^\top R_j = o_p(h^{-1}), \tag{A.7}$$

uniformly in $j = 1, \cdots, n$. We will make use of the following results:

$$E[\boldsymbol{\epsilon}_j^\top \delta(\mathbf{g}^\top\mathbf{g})^{-1}\delta^\top\boldsymbol{\epsilon}_j] = \frac{1}{T}E[\boldsymbol{\epsilon}_j^\top\delta\Sigma_g^{-1}\delta^\top\boldsymbol{\epsilon}_j](1+O_p(1)) \le \frac{C}{T}E\|\delta^\top\boldsymbol{\epsilon}_j\|^2 = o(h_k^{-1}), \tag{A.8}$$

$$E\|\delta^\top\boldsymbol{\epsilon}_j\|^2 \le p\max_k E\Big(\sum_{t=2}^T[\gamma_{tk}^\top\mathbf{S}_k\sum_{j=1}^n\beta_{jk}^{-1}\boldsymbol{\epsilon}_j]\epsilon_{tj}\Big)^2,$$

$$\begin{aligned}E\Big(\sum_{t=2}^T[\gamma_{tk}^\top\mathbf{S}_k\sum_{j=1}^n\beta_{jk}^{-1}\boldsymbol{\epsilon}_j]\epsilon_{tj}\Big)^2 &= \sum_{t=2}^T E[\gamma_{tk}^\top\mathbf{S}_k\sum_{j=1}^n\beta_{jk}^{-1}\boldsymbol{\epsilon}_j]^2\epsilon_{tj}^2\\ &\quad+\sum_{t\neq t'}E\Big([\gamma_{tk}^\top\mathbf{S}_k\sum_{j=1}^n\beta_{jk}^{-1}\boldsymbol{\epsilon}_j][\gamma_{t'k}^\top\mathbf{S}_k\sum_{j=1}^n\beta_{jk}^{-1}\boldsymbol{\epsilon}_j]\epsilon_{tj}\epsilon_{t'j}\Big)\\ &= O(h^{-2}),\end{aligned}$$

where the last equality follows from the fact that $\gamma_{tk}^\top\mathbf{S}_k = ([\mathbf{S}_k]_{tj} - g_k(x_{tk}) * [\mathbf{S}_k]_{1j}) = O((Th)^{-1})$.

Apparently (A.7) follows from (A.8), if we could also show that $\boldsymbol{\beta}_j^\top\mathbf{g}^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\delta^\top\mathbf{g}\boldsymbol{\beta}_j = O(T^{-1}\mathbf{g}^\top\delta\Sigma_g^{-1}\delta^\top\mathbf{g}) = O_p(1)$: this could be done in a manner similar to (A.8). Specifically, for any $l, k = 1, \cdots, p$, the (k,l)th element of $\delta^\top\mathbf{g}$ is given by

$$\sum_{t=2}^T\frac{x_{tl}}{x_{1l}}\Big(\gamma_{tk}^\top\mathbf{S}_k\frac{1}{n}\sum_{j=1}^n\beta_{jk}^{-1}\boldsymbol{\epsilon}_j\Big) = \sum_{j=1}^n\beta_{jk}^{-1}\boldsymbol{\epsilon}_j^\top\mathbf{S}_k^\top\Big(\sum_{t=2}^T\frac{x_{tl}}{x_{1l}}\gamma_{tk}\Big) = O_p(1),$$

where for the last equality we make use of the following facts:

$$\begin{aligned}&\sum_{t=2}^T\frac{x_{tl}}{x_{1l}}\gamma_{tk} = \Big[-\sum_{t=2}^T\frac{x_{kl}x_{tl}}{x_{kl}x_{1l}}, \frac{x_{2l}}{x_{1l}}, \cdots, \frac{x_{Tl}}{x_{1l}}\Big]^\top,\\ &\sum_{t'=1}^T[\mathbf{S}_k]_{t'j}\Big(\sum_{t=1}^T\frac{x_{tl}}{x_{1l}}\gamma_{tk}\Big) = O(1) + O_p((Th)^{-1/2}).\end{aligned}$$

Next, we will show that for the last term on the RHS of (A.6) the following holds:

$$R_j^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\Delta(\mathbf{g}^\top\mathbf{g})^{-1}\delta^\top R_j = O_p((Th)^{-1}). \tag{A.9}$$

This is based on the following identities:

(A) $\boldsymbol{\epsilon}_j^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\Delta(\mathbf{g}^\top\mathbf{g})^{-1}\delta^\top\boldsymbol{\epsilon}_j = O_p((Th)^{-1})$;

(B) $\boldsymbol{\beta}_j^\top\mathbf{g}^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\delta^\top\mathbf{g}\boldsymbol{\beta}_j = O_p(T^{-2})$.

To prove (A), first note that $\boldsymbol{\epsilon}_j^\top \delta(\mathbf{g}^\top\mathbf{g})^{-1}\Delta(\mathbf{g}^\top\mathbf{g})^{-1}\delta^\top\boldsymbol{\epsilon}_j = 2\boldsymbol{\epsilon}_j^\top \delta(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\delta^\top\boldsymbol{\epsilon}_j$, and the kth ($k = 1, \cdots, p$) element of $\boldsymbol{\epsilon}_j^\top\delta$ is such that

$$\begin{aligned}\sum_{j'=1}^{n}\beta_{j'k}^{-1}\Big(\epsilon_{tj}\gamma_{tk}^\top\mathbf{S}_k\Big)\boldsymbol{\epsilon}_{j'} &= \sum_{j'=1}^{n}\beta_{j'k}^{-1}\Big([-\sum_{t=2}^{T}\frac{x_{tk}}{x_{1k}}\epsilon_{tj}, \epsilon_{2j}, \cdots, \epsilon_{Tj}]\mathbf{S}_k\Big)\boldsymbol{\epsilon}_{j'} \\ &= \sum_{j'=1}^{n}\beta_{j'k}^{-1}\Big[\sum_{t=2}^{T}\epsilon_{tj}[\mathbf{S}_k]_{t,t'} - \sum_{t=2}^{T}\frac{x_{tk}}{x_{1k}}\epsilon_{tj}[\mathbf{S}_k]_{1,t'}, t' = 1, \cdots, T\Big]\boldsymbol{\epsilon}_{j'}.\end{aligned}$$

Since $\sum_{t=2}^{T}\epsilon_{tj}[\mathbf{S}_k]_{t,t'} = O_p((Th)^{-1/2})$ and $\sum_{t=2}^{T}\frac{x_{tk}}{x_{1k}}\epsilon_{tj} = O_p(T^{-1/2})$, uniformly in $t' = 1, \cdots, T$, whence $\boldsymbol{\epsilon}_j^\top\delta = O_p((T/h)^{1/2})$. The proof of (B) is similar.

We now move on to the second term on the RHS of (A.6): $R_j^\top\Gamma^\top\Delta\Gamma R_j$, which again is bounded by two times the following term:

$$\boldsymbol{\epsilon}_j^\top\mathbf{g}(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\boldsymbol{\epsilon}_j + \boldsymbol{\beta}_j^\top\mathbf{g}^\top\mathbf{g}(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\mathbf{g}\boldsymbol{\beta}_j = O_p(1),$$

where for the last equality we used the fact that $\mathbf{g}^\top\boldsymbol{\epsilon}_j = O_p(T^{1/2})$.

Now the only term left to be dealt with is $R_j^\top\delta\Gamma R_j$, which equates to

$$\begin{aligned}R_j^\top\delta\boldsymbol{\beta}_j + R_j^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\boldsymbol{\epsilon}_j = \boldsymbol{\epsilon}_j^\top\delta\boldsymbol{\beta}_j + \boldsymbol{\beta}_j^\top\mathbf{g}^\top\delta\boldsymbol{\beta}_j + \boldsymbol{\beta}_j^\top\mathbf{g}^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\boldsymbol{\epsilon}_j \\ +\boldsymbol{\epsilon}_j^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\boldsymbol{\epsilon}_j;\end{aligned} \tag{A.10}$$

where $\boldsymbol{\beta}_j^\top\mathbf{g}^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\boldsymbol{\epsilon}_j = O_p(T^{-1/2})$ and $\boldsymbol{\beta}_j^\top\mathbf{g}^\top\delta\boldsymbol{\beta}_j = O_p(1)$. The kth element of $\boldsymbol{\epsilon}_j^\top\delta$:

$$\sum_{j'=1}^{n}\beta_{j'k}^{-1}\Big(\sum_{t=2}^{T}\epsilon_{tj}\gamma_{tk}^\top\Big)\mathbf{S}_k\boldsymbol{\epsilon}_{j'} = \sum_{j'\neq j}\beta_{j'k}^{-1}\Big(\sum_{t,t'=2}^{T}\epsilon_{tj}\epsilon_{t'j'}[\mathbf{S}_k]_{t,t'}\Big) + \beta_{jk}^{-1}\Big(\sum_{t,t'=2}^{T}\epsilon_{tj}\epsilon_{t'j}[\mathbf{S}_k]_{t,t'}\Big),$$

has a mean with the leading term given by

$$\beta_{jk}^{-1}\sigma^2\sum_{t=2}^{T}[\mathbf{S}_k]_{t,t} = K(0)\beta_{jk}^{-1}\sigma^2h^{-1}(1 + o_p(1)); \tag{A.11}$$

and a second moment with leading term given by

$$\begin{aligned}&\sigma^4\sum_{j'\neq j}\beta_{j'k}^{-2}\sum_{t,t'=2}^{T}[\mathbf{S}_k]_{t,t'}^2 + \beta_{jk}^{-2}\mu_4\sum_{t=2}^{T}[\mathbf{S}_k]_{t,t}^2 \\ &\qquad +\beta_{jk}^{-2}\sigma^4\sum_{t<t'}\{[\mathbf{S}_k]_{t,t'}^2 + [\mathbf{S}_k]_{t',t}^2 + 2[\mathbf{S}_k]_{t,t'}[\mathbf{S}_k]_{t',t} + 2[\mathbf{S}_k]_{t,t}[\mathbf{S}_k]_{t',t'}\} \\ &= \sigma^4\sum_{j'\neq j}\beta_{j'k}^{-2}\sum_{t,t'=2}^{T}[\mathbf{S}_k]_{t,t'}^2 + \beta_{jk}^{-2}(\mu_4 - \sigma^4)\sum_{t=2}^{T}[\mathbf{S}_k]_{t,t}^2 \\ &\qquad +\beta_{jk}^{-2}\sigma^4\sum_{t<t'}\{[\mathbf{S}_k]_{t,t'}^2 + [\mathbf{S}_k]_{t',t}^2 + 2[\mathbf{S}_k]_{t,t'}[\mathbf{S}_k]_{t',t}\} + \beta_{jk}^{-2}\sigma^4\Big(\sum_{t=2}^{T}[\mathbf{S}_k]_{t,t}\Big)^2.\end{aligned}$$

Thus its variance is such that

$$\Big(4\beta_{jk}^{-2}+\sum_{j'\neq j}\beta_{j'k}^{-2}\Big)\sigma^4R(K)h^{-1}T^{-2}\sum_{l=1}^{T-1}(T-l)a_{l;k}. \tag{A.12}$$

From (A.11) and (A.12), we could deduce that $\boldsymbol{\epsilon}_j^\top\delta\boldsymbol{\beta}_j$ has mean of $pK(0)\sigma^2h^{-1}$ and variance

$$\sigma^4R(K)h^{-1}T^{-2}\sum_{k=1}^{p}\{4+\sum_{j'\neq j}(\beta_{jk}/\beta_{j'k})^2\}\sum_{l=1}^{T-1}(T-l)a_{k,l}$$
$$+\sigma^4T^{-2}\sum_{k\neq k'}\{4+\sum_{j'\neq j}(\beta_{jk}/\beta_{j'k})^2\}\sum_{l=1}^{T-1}(T-l)b_{l;k,k'}.$$

Under assumption [A4], the variance of $\boldsymbol{\epsilon}_j^\top\delta\boldsymbol{\beta}_j$ could be further simplified as

$$\sigma^4R(K)h_k^{-1}\sum_{k=1}^{p}c_k\{4+\sum_{j'\neq j}(\beta_{jk}/\beta_{j'k})^2\}.$$

Now we deal with the fourth term in (A.10). As the kth element of $\boldsymbol{\epsilon}_j^\top\delta$ given by

$$\sum_{j'=1}^{n}\beta_{j'k}^{-1}\Big(\sum_{t=2}^{T}\epsilon_{tj}\gamma_{tk}^\top\Big)\mathbf{S}_k\boldsymbol{\epsilon}_{j'}=\sum_{j'\neq j}\beta_{j'k}^{-1}\Big(\sum_{t,t'=2}^{T}\epsilon_{tj}\epsilon_{t'j'}[\mathbf{S}_k]_{t,t'}\Big)+\beta_{jk}^{-1}\Big(\sum_{t,t'=2}^{T}\epsilon_{tj}\epsilon_{t'j}[\mathbf{S}_k]_{t,t'}\Big),$$

and the k'th element of $\mathbf{g}^\top\boldsymbol{\epsilon}_j$ given by $\sum_{t=1}^{T}\frac{x_{tk'}}{x_{1k'}}\epsilon_{tj}$, we have

$$\boldsymbol{\epsilon}_j^\top\delta(\mathbf{g}^\top\mathbf{g})^{-1}\mathbf{g}^\top\boldsymbol{\epsilon}_j=\frac{1}{T}\sum_{k,k'=1}^{p}\sigma_{k,k'}\beta_{jk}^{-1}\Big(\sum_{t=1}^{T}\frac{x_{tk'}}{x_{1k'}}\epsilon_{tj}\Big)\Big(\sum_{t,t'=2}^{T}\epsilon_{tj}\epsilon_{t'j}[\mathbf{S}_k]_{t,t'}\Big)$$
$$+\frac{1}{T}\sum_{k,k'=1}^{p}\sigma_{k,k'}\Big(\sum_{t=1}^{T}\frac{x_{tk'}}{x_{1k'}}\epsilon_{tj}\Big)\sum_{j'\neq j}\beta_{j'k}^{-1}\Big(\sum_{t,t'=2}^{T}\epsilon_{tj}\epsilon_{t'j'}[\mathbf{S}_k]_{t,t'}\Big),$$

which is of mean zero with its variance easily shown to be of order $O((Th)^{-1})$.

The fact that $\boldsymbol{\epsilon}_j^\top\delta$ is the dominating term in the partition (A.6) of $\mathrm{RSS}_{0,j}-\mathrm{RSS}_{1,j}$, applies to all $j=1,\ldots,n$. To derive the asymptotics of $\lambda(H_0)$, we also need to consider the covariance between $\mathrm{RSS}_{0,j}-\mathrm{RSS}_{1,j}$ and $\mathrm{RSS}_{0,j'}-\mathrm{RSS}_{1,j'}$ $(j,j'=1,\cdots,n,\ j\neq j')$. This in turn equals to that between $\boldsymbol{\epsilon}_j^\top\delta\boldsymbol{\beta}_j$ and $\boldsymbol{\epsilon}_{j'}^\top\delta\boldsymbol{\beta}_{j'}$, the leading term of which is easily seen to be given by

$$h^{-1}\sigma^4R(K)T^{-2}\sum_{k=1}^{p}\sum_{l=1}^{T-1}(T-l)a_{l;k}.$$

The proof is thus complete. □

Proof of Corollary 7.1 For backfitting estimation of additive models, Opsomer (2000) studied theoretical properties on general linear smoothers with independent observations. We now describe

the extension of his results to our case, i.e. for any given $j = 1, \cdots, n$, the estimation of $\{G_{jk}(.), k = 1, \cdots, p\}$ based on time series data $\{r_{tj},\, t = 1, \cdots, T\}$.

With linear smoother matrices such as the $T \times T$ matrices $\mathbf{S}_k$, $k = 1, \cdots, p$ of (A.1), the backfitting estimates of the additive component functions evaluated at the observation points are by definition the solution to the following system of equations for the unknown vectors of fits $\mathbf{G}_{j1}, \cdots, \mathbf{G}_{jp}$:

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \mathbf{S}_p & \cdots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{j1} \\ \mathbf{G}_{j2} \\ \vdots \\ \mathbf{G}_{jp} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_p \end{bmatrix} \mathbf{R}_j. \tag{A.13}$$

Conceptually the solution could be written as

$$\begin{bmatrix} \hat{\mathbf{G}}_{j1} \\ \hat{\mathbf{G}}_{j2} \\ \vdots \\ \hat{\mathbf{G}}_{jp} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \cdots & \mathbf{S}_2 \\ \vdots & \vdots\ddots & \vdots & \\ \mathbf{S}_p & \mathbf{S}_p & \cdots & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_p \end{bmatrix} \mathbf{R}_j \equiv \mathbf{M}^{-1}\mathbf{C}\mathbf{R}_j, \tag{A.14}$$

provided that $\mathbf{M}$ is invertible. Write

$$\mathbf{W}_k = \mathbf{E}_k \mathbf{M}^{-1} \mathbf{C},$$

where $\mathbf{E}_k$ is a partitioned matrix of dimension $T \times (T\,p)$ with an $T \times T$ identity matrix as the kth block and zero matrices else where, so that $\hat{\mathbf{G}}_{jk} = \mathbf{W}_k R_j$. According to Lemma 2.1 of Opsomer (2000), equation (A.13) solved through backfitting algorithm will converge to a unique solution if

$$\|\mathbf{S}_k \mathbf{W}_{[-k]}\| < 1 \tag{A.15}$$

for some $k \in \{1, \cdots, p\}$ and any matrix norm $\|.\|$, where recall that $\mathbf{W}_{[-k]}$ has been defined preceding Corollary 7.1. As pointed out in Buja et al. (1989) and Opsomer (2000), a necessary condition for (A.15) to hold for any of the major smoothing techniques unless the smoother matrices are centered, i.e. $\mathbf{S}_k$ replaced by its centered counterpart $\mathbf{S}_k^*$. In that case, the additive smoother with respect to the kth component function $G_{jk}(.)$ is written as

$$\mathbf{W}_k = \mathbf{I} - (\mathbf{I} - \mathbf{S}_k^* \mathbf{W}_{[-k]})^{-1}(\mathbf{I} - \mathbf{S}_k^*) = (\mathbf{I} - \mathbf{S}_k^* \mathbf{W}_{[-k]})^{-1} \mathbf{S}_k^* (\mathbf{I} - \mathbf{W}_{[-k]}). \tag{A.16}$$

The aymptotic bias and variance of $\hat{\mathbf{G}}_{jk}$, $j = 1, \cdots, T$, $k = 1, \cdots, P$ is then derived from (A.16) and that $\hat{\mathbf{G}}_{jk} = \mathbf{W}_k R_j$; see Theorem 3.1 in Opsomer (2000) in the case of iid observations. Here

we need to generalize these results to dependent sequences. The key intermediary step is, as in Opsomer and Ruppert (1997) and Opsomer (2000, pp. 178), to show that that

$$\mathbf{S}_k^* = \mathbf{S}_k - \mathbf{1}_{\mathrm{T}}\mathbf{1}_{\mathrm{T}}^\top/T + o_p(\mathbf{1}_{\mathrm{T}}\mathbf{1}_{\mathrm{T}}^\top/T),$$
$$(\mathbf{I} - \mathbf{S}_k^*\mathbf{W}_{[-k]})^{-1} = \mathbf{I} + O_p(\mathbf{1}_{\mathrm{T}}\mathbf{1}_{\mathrm{T}}^\top/T),$$

uniformly over all elements of the matrices. This follows from results given in Yu (1994) on rates of convergence for empirical processes of stationary mixing sequence. The rest of the proof are identical to that of Theorem 3.1 of Opsomer (2000). □