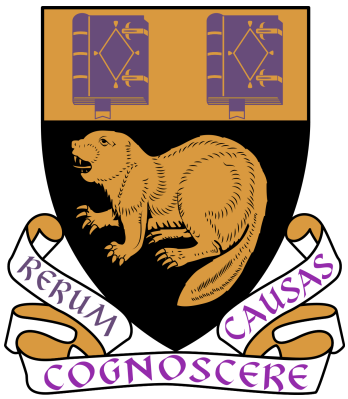
Autocovariance-based Statistical Inference for High-Dimensional Function/Scalar Time Series



Cheng Chen

The Department of Statistics

London School of Economics and Political Science

A thesis submitted for the degree of

Doctor of Philosophy

August 2021

This thesis is dedicated to
Rongxi Luo

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I confirm that Chapter 1 were jointly co-authored with my supervisor, Dr. Xinghao Qiao and Professor Shaojun Guo from Renmin University, China. Chapter 2 were jointly co-authored with Dr. Xinghao Qiao and Professor Jinyuan Chang from Southwestern University of Finance and Economics, China.

Chapter 1 has been published as Chen, C., Guo, S., and Qiao, X. (2020). Functional linear regression: dependence and error contamination. *Journal of Business & Economic Statistics*, 1-14. Chapter 2 has been submitted to a peer-reviewed statistical journal and we plan to submit Chapter 3 for publication soon.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Dr. Xinghao Qiao, for his constant help, patient guidance and professional advice for my PhD research and life. Without him, it is impossible for me to finish my PhD study. Therefore, I am deeply grateful for his time and effort spent on me.

I want to thank Professor Shaojun Guo and Professor Jinyuan Chang for the collaborative research projects consisting different chapters of my thesis. I am looking forward to collaborating with them on more exciting topics in the future.

I am also very grateful for the financial support from the China Scholarship Council. And I want to thank all faculty and staff in our department for providing me with a supportive learning environment. I would also like to thank Dr. Cheng Qian and Dr. Junyi Zhang for the stimulating discussions, the days we worked hard together, and all the fun we had in our life at LSE.

Last but not least, I would like to thank my parents for their continual support and unconditional love throughout my life. Additionally, I am also deeply indebted to my fiancée for her consistent encouragement and always love. Words cannot express how thankful to my family.

Abstract

High-dimensional time series analysis is an important area in modern statistics. Data arises in many fields, including finance, economics, environmental and medical studies, among others. Faced with corrupted data where measurement errors require particular attention, previous work devoted to clean data calls for further exploration. Motivated by the fact that the autocovariance of observed time series automatically filters out the noise term, this thesis investigates the autocovariance-based estimation and inferential studies based on high-dimensional functional and scalar time series models. The subsequent chapters are organised as follows. The first chapter studies the functional linear regression when observed functions are error contaminated. The second chapter extends the topic to high-dimensional functional linear models. The third chapter investigates not only the estimation but also the inference on scalar time series in high dimensions.

In the first chapter, we briefly introduce the motivation, main idea, contributions and limitations of this thesis.

In the second chapter, we consider functional linear regression with serially dependent observations of the functional predictor, where the contamination of the predictor by the white noise is genuinely functional with a fully nonparametric covariance structure. It is commonly assumed that samples of the functional predictor are independent realisations of an underlying stochastic process and are observed over a grid of points contaminated by i.i.d. measurement errors. In practice, however, the dynamical dependence across different curves may exist, and the parametric assumption on the error covariance structure could be unrealistic. Therefore, we propose a novel autocovariance-based generalised method-of-moments estimate of the slope function. We also develop a nonparametric smoothing approach to handle the scenario of partially observed functional predictors. The asymptotic properties of the resulting estimators under different scenarios are established. Finally, we

demonstrate that our proposed method significantly outperforms competing methods through an extensive set of simulations and an analysis of a public financial dataset.

In the third chapter, we model observed functional time series, which are subject to errors in the sense that each functional datum arises as the sum of two uncorrelated components, one dynamic and one white noise. We propose an autocovariance-based three-step procedure by first performing autocovariance-based dimension reduction and then formulating a novel autocovariance-based block regularised minimum distance (RMD) estimation framework to produce block sparse estimates, from which we can finally recover functional sparse estimates. We investigate non-asymptotic properties of relevant estimated terms under such an autocovariance-based dimension reduction framework. To provide theoretical guarantees for the second step, we present a convergence analysis of the block RMD estimator. Finally, we illustrate the proposed autocovariance-based learning framework using applications of three sparse high-dimensional functional time series models. With derived theoretical results, we study the convergence properties of the associated estimators. Using simulated and real datasets, we demonstrate that our proposed estimators significantly outperform the competitors.

In the fourth chapter, we study the high-dimensional linear regression with scalar serially dependent predictors that are error contaminated. To mitigate the influence of measurement errors, we propose an autocovariance-based de-bias regularised generalised method of moments (DRGMM) framework to obtain a high-quality estimator for regression coefficients. Moreover, we conduct an inferential study on the estimators within this framework. Theoretical results on estimation consistency and inference accuracy are provided. Finally, the finite sample performance of the proposed inference procedure is examined through simulation studies.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

To embrace the modern information age, time series analysts need to face the opportunities together with challenges brought by the availability of large time series datasets, which come from various sources including but not limited to financial markets, engineering, natural and social phenomena. Moreover, the data are recorded not only as data points but also in the form of functions, where observations are recorded continuously during a time interval or intermittently at several discrete time points. This thesis referred to these two forms of data as scalar time series and functional time series, respectively. In the meantime, the data are often recorded with errors, which are introduced into the observations through different manners. For both scalar and functional data, the errors may come from the missing value, inaccuracy and fault records. In addition, the generating of the functional data by interpolating and smoothing from the original discretely and incompletely observed trajectories could also induce errors. Therefore, this thesis encounters the difficulty consisting of high dimension and error contamination in terms of the function/scalar time series analysis.

High dimensional time series problems have drawn a lot of attention in recent years. On the other hand, it is always a challenge to model multiple time series even with moderately large dimensions. A simple extension of the approaches designed for small datasets such as the Autoregressive Integrated Moving Average (ARIMA) model and Vector Autoregression (VAR) to the high dimensional case is invalid for the reason of the well-known "curse of dimensionality". To reduce the number of parameters and to eliminate the non-identification issues, one popular approach is

to find a small number of factors to extract the information contains in the multiple time series; see Lam and Yao (2012), among others, for example. Another approach, which is of the main focus in Chapter 3 and Chapter 4 of this thesis, is the method based on the sparsity assumption that only a few variables among the whole data sets are effective, such as LASSO (Tibshirani, 1996) and Dantzig (Candes and Tao, 2007). These methods perform the variable selection and coefficients estimation simultaneously at the cost of introducing the bias into the model. Then, following a de-bias step, one can carry out the inferential studies.

For the purpose of characterising the temporal dependence in the time series data of interest, a strong mixing condition (Bosq, 2000), the functional stability measure (Guo and Qiao, 2020) and the physical dependence measure (Wu, 2005) are implemented in Chapter 2, Chapter 3 and Chapter 4, respectively. These conditions describe the dependence structures in different scenarios and facilitate the derivation of the theorems for estimation and inference in each chapter. Specifically, given the infinite-dimensional natural and serial dependence of the functional time series, the mixing condition in Chapter 2 and the sub-Gaussian assumption imposed in Chapter 3 simplify the derivatives of the theoretical results, and make the presentation of the main idea more concise. And in Chapter 4, some less strident conditions are imposed for scalar time series. See Yousuf (2018) for comparisons between physical dependence measures and strong mixing assumptions.

For the functional data that are inherently infinite-dimensional, dimension reduction is the key to functional data modelling and analysis. One of the most prevalent dimension reduction tools for multivariate data analysis is principal component analysis, which has been extended to functional data by Karhunen (1946) and Loève (1946). Via representing the infinite-dimensional functional data by a finite-dimensional random scores vector, functional principal component analysis (FPCA) facilities the modelling and simplifies the estimation for the functional data, and then, this approach becomes the most popular tool in functional data analysis. However, for the erroneous observed functional time series, the classical FPCA method is not directly available because of the existence of temporal dependence and error contamination. Alternatively, following the idea of Bathia et al. (2010), the autocovariance based method is implemented to tackle the problem. Chapter 2 implicitly uses this idea to estimates inverse operator and Chapter 3 takes advantage of dimensional reduction and error filtering of this approach to promote the solution of the high-dimensional problems.

Standing at the junction of the high dimensionality, temporal dependence and the various form of data observed with errors, the topic of this thesis is both interesting and challenging.

1.2 Contributions and limitations

This thesis studies the datasets consist of p-dimensional vector time series, $\mathbf{W}_t = \{W_{t1}, \ldots, W_{tp}\}^{\mathrm{T}}$ for $t = 1, \ldots, n$. To save the notation, W_{tj} could either be scalar in a real space or be functional defined on a compact interval $u \in \mathcal{U}$ as $W_{tj}(u)$, where $j = 1, \ldots, p$ and $p \geq 1$. We allows p to be finite in Chapter 2, or to be diverging with, or even larger than, n in a high-dimensional regime in Chapter 3 and Chapter 4.

Here we take the scalar time series for example as shown in Chapter 4. Define (auto)covariance matrices $\mathbf{\Sigma}_h^{\mathbf{W}} = \mathrm{Cov}\{\mathbf{W}_t, \mathbf{W}_{t+h}\}$ for any integer h, and similarly define the autocovariance functions $\mathbf{\Sigma}_h^{\mathbf{W}}(u, v),\ u, v \in \mathcal{U}$ as in Chapter 2 and Chapter 3. Then, suppose that the $\mathbf{W}_t$ are erroneous observed in the form of

$$\mathbf{W}_t = \mathbf{X}_t + \mathbf{e}_t,$$

where $\mathbf{X}_t$ is the p-dimensional signal series with (auto)covariance matrices denoted by $\mathbf{\Sigma}_h^{\mathbf{X}}$. And $\mathbf{e}_t$ are white noise sequence with zero mean and autocovariance $\mathbf{\Sigma}_h^{\mathbf{e}} = 0$ for any $h \neq 0$. This formation ensures that the signal series and the white noise sequence include all the dynamic elements and the errors of observations, respectively. Therefore, we have $\mathbf{\Sigma}_h^{\mathbf{W}} = \mathbf{\Sigma}_h^{\mathbf{X}}$ for $h \neq 0$, which implies that the autocovariance of $\mathbf{W}_t$ could filter the errors and be a consistent estimation of the autocovariance of unobservable $\mathbf{X}_t$. The idea of signal-error-decomposition and autocovariance-filtering has implemented by Lam et al. (2011) for high-dimensional scalar times and Bathia et al. (2010) for univariate functional time series.

In the previous studies, to handle the errors, some particular assumptions are applied to the error term. For example, in Hall and Vial (2006), the errors are assumed to vanish as the sample size increases. And the covariance matrices of the scalar errors are assumed known (Li et al., 2021) and some parametric structures of the covariance functions are imposed of the functional errors (Yao et al., 2005). And in our proposed method, we can relax these assumptions by taking advantages of the temporal dependence of the data. See more detailed discussions in the following chapters.

The proposed autocovariance based methods can handle linear regression models with response and covariates being functional/scalar time series, as well as functional vector autoregressive models in high dimensions. Throughout this thesis, the error contamination problem is taken into consideration. And the dynamical dependence across data facilitates the development of the proposed methods and makes the error contamination problem tractable.

We relax the assumptions that are usually placed on the covariance structure of the error terms, instead, we rely on the autocovariance of the observed time series to get rid of influence from errors. In the linear models considered in this thesis, the lag terms of $\mathbf{W}_t$ naturally play the role of the instrumental variables in the econometrics studies, which overtakes the difficult of finding the instrumental variables.

Given the infinite-dimensional nature of the functional data, the novel autocovariance based generalised method of moments approach and the block regularised minimum distance estimation following an autocovariance based dimensional reduction approach are proposed in Chapter 2 and Chapter 3, which contribute to the tool kit of functional time series analysis. And a three-step approach is implemented to perform the estimation and simultaneous inference study for the high-dimensional time series regression problem in Chapter 4.

We establish the convergence rate for our proposed estimators under varying model settings and provide the theoretical guarantees for the simultaneous influences. Some interesting phenomena are not only revealed by theoretical reasoning but also illustrated by empirical studies, which consists of both simulations and applications on financial datasets.

Even though our proposed autocovariance based approaches show many advantages mentioned above in handling the error contaminated linear time series problem, it relies on the major assumption that the temporal dependence of the variables is existent. Considering the case where the variables are independent indeed, there is no lag information we can obtain from the data. Therefore, our proposed method is no longer feasible because of the zero autocovariance. Even in the situation where the autocovariance is non-zero but very close to zero, that is, the temporal dependence is weak, our approach suffers from the inaccuracy of the corresponding estimators and the lack of efficiency. So, it is important to test the dependence beforehand and to apply our proposed method only when the relevant assumptions are fulfilled. To the best of our knowledge, there is no universal solution to this problem under our framework. And the dependence assumption could be regarded as the price to pay for removing the knowledge about the error terms required in previous work.

1.3 Summary of chapters

The rest of the thesis is organised as follows.

In Chapter 2, we study the linear regression model where predictors are functional time series with additive functional errors, whose covariance function is of fully non-

parametric. In Section 2.2, we present the model for regression with dependent functional errors-in-predictors and develop an Autocovariance-based generalised method-of moments (AGMM) fitting procedures for both scalar and functional responses. We also propose the regularised estimator by imposing some form of smoothness into the estimation procedure and discuss the selection of relevant tuning parameters. In Section 2.3, we present convergence results for our proposed estimators for the slope function under different functional scenarios. In Section 2.4, we develop a nonparametric smoothing approach for partially observed curve time series and investigate its asymptotic properties. Section 2.5 illustrates the finite sample performance of AGMM through a series of simulation studies and a public financial dataset.

In Chapter 3, we propose an autocovariance-based three-step procedure by first performing autocovariance-based dimension reduction and then formulating a novel autocovariance-based block regularised minimum distance (RMD) estimation framework to produce block sparse estimates, from which we can finally recover functional sparse estimates. In Section 3.2, we propose a general autocovariance-based three-step procedure with illustration using scalar-on-function linear additive regression (SFLR) as an example. In Section 3.3, we present the first step of autocovariance-based dimension reduction and establish essential deviation bounds in elementwise ℓ_∞-norm on relevant estimated terms used in subsequent analysis. In Section 3.4, we formulate the second step in a general block RMD estimation framework and investigate its theoretical properties. In Section 3.5, we illustrate the proposed autocovariance-based learning framework using applications of SFLR, function-on-function linear additive regression (FFLR) and vector functional autoregression (VFAR), and present convergence analysis of the associated estimators. In Section 3.6, we examine the finite-sample performance of the proposed estimators through both an extensive set of simulations and an analysis of a public financial dataset.

In Chapter 4, we consider the high-dimensional linear regression model where scalar serially dependent predictors that are error contaminated. we studied the coefficient estimation and the inference after a de-bias step. In Section 4.2, we present the high-dimensional time series linear regression model for which an autocovariance-based estimation and de-bias framework is proposed. In Section 4.3, we provided the theoretical guarantee for the estimation of sparse coefficients based on regularised minimum distant (RMD) estimation. In Section 4.4, we perform the inferential study on the de-biased regularised estimation, where the theoretical results on estimation consistency and inference accuracy are provided. Section 4.5 exams the finite sample performance of the proposed inference procedure through simulation studies.

All technical proofs are relegated to the appendices.

Chapter 2

Functional Linear Regression: Dependence and Error Contamination

2.1 Introduction

In functional data analysis, the linear regression problem depicting the linear relationship between a functional predictor and either a scalar or functional response, has recently received a great deal of attention. See Ramsay and Silverman (2005) for a thorough discussion of the issues involved with fitting such data. For examples of recent research on functional linear models, see Chakraborty and Panaretos (2017), Cho et al. (2013), Crambes et al. (2009), Hall and Horowitz (2007), Yao et al. (2005) and the references therein. We refer to Morris (2015) for an extensive review on recent developments for functional regression.

In functional regression literature, one typical assumption is to model observed functional predictors, denoted by $X_1(\cdot), \ldots, X_n(\cdot)$, as independent realisations of an underlying stochastic process. However, curves can also arise from segments of consecutive measurements over time. Examples include daily curves of financial transaction data (Horváth et al., 2014), intraday electricity load curves (Cho et al., 2013) and daily pollution curves (Aue et al., 2015). Such type of curves, also named as curve time series, violates the independence assumption, in the sense that the dynamical dependence across different curves exists. The other key assumption treats the functional predictor as being either fully observed (Hall and Horowitz,

2007) or incompletely observed, with measurement error, at a grid of time points (Crambes et al., 2009). In the latter case, errors associated with distinct observation points are assumed to be i.i.d., where the corresponding covariance function for the error process is diagonal with constant diagonal components. In the curve time series setting, $X_t(\cdot)$ are often recorded at discrete points and are subject to dependent and heteroskedastic errors (Bathia et al., 2010). Hence, the resulting error covariance matrix would be more nonparametric with varying diagonal entries and nonzero off-diagonal entries.

In this chapter, we consider the functional linear regression in a time series context, which involves serially dependent observations of the functional predictor contaminated by genuinely functional errors corresponding to a fully nonparametric covariance structure. We assume that the observed erroneous predictors, which we denote by $W_1(\cdot), \ldots, W_n(\cdot)$, are defined on a compact interval $\mathcal{U}$ and are subject to errors in the form of

$$W_t(u) = X_t(u) + e_t(u), \quad u \in \mathcal{U}, \tag{2.1}$$

where the error process $\{e_t(\cdot), t = 1, 2, \ldots\}$ is a sequence of white noise such that $E\{e_t(u)\} = 0$ for all t and $\mathrm{Cov}\{e_t(u), e_s(v)\} = 0$ for any $(u, v) \in \mathcal{U}^2$ provided $t \neq s$. We also assume that $X_t(\cdot)$ and $e_t(\cdot)$ are uncorrelated and correspond to unobservable signal and noise components, respectively. The error contamination model in (2.1) was also considered in Bathia et al. (2010). To fit the functional regression model, the conventional *least square* (LS) approach (Hall and Horowitz, 2007) relies on the sample covariance function of $W_t(\cdot)$, which is not a consistent estimator for the true covariance function of $X_t(\cdot)$, thus failing to account for the contamination that can result in substantial estimation bias. One can possibly implement the LS method in the resulting multiple linear regression after performing dimension reduction for $W_t(\cdot)$ to identify the dimensionality of $X_t(\cdot)$ (Bathia et al., 2010). However, this approach still suffers from unavoidable uncertainty due to $e_t(\cdot)$, while the inconsistency has been demonstrated by our simulations. Inspired from a simple fact that $\mathrm{Cov}\{W_t(u), W_{t+k}(v)\} = \mathrm{Cov}\{X_t(u), X_{t+k}(v)\}$ for any $k \neq 0$, which indicates that the impact from the unobservable noise term can be automatically eliminated, we develop an *autocovariance-based generalised method-of-moments* (AGMM) estimator for the slope function. This procedure makes the good use of the serial dependence information, which is the most relevant in the context of time series modelling.

To tackle the problem we consider, the conventional LS approach is not directly applicable in the sense that one cannot separate $X_t(\cdot)$ from $W_t(\cdot)$ in equation (2.1). This difficulty was resolved in Hall and Vial (2006) under the restrictive "low noise" setting, which assumes that the noise $e_t(\cdot)$ goes to zero as n grows to infinity. The

recent work by Chakraborty and Panaretos (2017) implements the regression calibration approach combined with the low rank matrix completion technique to separate $X_t(\cdot)$ from $W_t(\cdot)$. Their approach relies on the identifiability result that, provided real analytic and banded covariance functions for $X_t(\cdot)$ and $e_t(\cdot)$, respectively, the corresponding two covariance functions are identifiable (Descary and Panaretos, 2019). However, all the aforementioned methods are developed under the critical independence assumption, which would be inappropriate for the setting that $W_1(\cdot), \ldots, W_n(\cdot)$ are serially dependent.

The proposed AGMM method has four main advantages. First, it can handle regression with serially dependent observations of the functional predictor. The existence of dynamical dependence across different curves makes our problem tractable and facilitates the development of AGMM. Second, without placing any parametric assumption on the covariance structure of the error process, it relies on the autocovariance function to get rid of the effect from the genuinely functional error. Interestingly, it turns out that the operator in AGMM defined based on the autocovariance function of the curve process is identical to the nonnegative operator in Bathia et al. (2010), which is used to assess the dimensionality of $X_t(\cdot)$ in equation (2.1). Third, the proposed method can be applied to both scalar and functional responses with either finite or infinite dimensional functional predictors. To handle a practical scenario where functional predictors are partially observed, we also develop a local linear smoothing approach. Theoretically we establish relevant convergence rates for our proposed estimators under different model settings. In particular, our asymptotic results for partially observed functional predictors reveal interesting phase transition phenomena. Fourth, empirically we illustrate the superiority of AGMM relative to the potential competitors.

The rest of the chapter is organised as follows. In Section 2.2, we present the model for regression with dependent functional errors-in-predictors and develop AGMM fitting procedures for both scalar and functional responses. We also propose the regularised estimator by imposing some form of smoothness into the estimation procedure and discuss the selection of relevant tuning parameters. In Section 2.3, we present convergence results for our proposed estimators for the slope function under different functional scenarios. In Section 2.4, we develop a nonparametric smoothing approach for partially observed curve time series and investigate its asymptotic properties. Section 2.5 illustrates the finite sample performance of AGMM through a series of simulation studies and a public financial dataset. All technical proofs are relegated to the Appendix.

2.2 Methodology

2.2.1 Model setup

In this section, we describe the model setup for the functional linear regression with dependent errors-in-predictors we consider. Let $\mathcal{L}_2(\mathcal{U})$ denote a Hilbert space of square integrable functions defined on $\mathcal{U}$ equipped with the inner product $\langle f, g\rangle = \int_{\mathcal{U}} f(u)g(u)du$ for $f, g \in \mathcal{L}_2(\mathcal{U})$. Given a scalar response Y_t, a functional predictor $X_t(\cdot)$ in $\mathcal{L}_2(\mathcal{U})$, and, without loss of generality, assuming that $\{Y_t, X_t(\cdot)\}$ have been centred to have mean zero, the classical scalar-on-function linear regression model is of the form

$$Y_t = \int_{\mathcal{U}} X_t(u)\beta_0(u)du + \varepsilon_t, \quad t = 1, \ldots, n, \tag{2.2}$$

where the errors ε_t, independent of $X_{t+k}(\cdot)$ for any integer k, are generated according to a white noise process and $\beta_0(\cdot)$ is the unknown slope function. Generally, β_0 may not be uniquely determined. We will discuss how to identify β_0 we wish to estimate later.

We assume that the observed functional predictors $W_1(\cdot), \ldots, W_n(\cdot)$ satisfy the error contamination model in equation (2.1). The existence of the unobservable noise term $e_t(\cdot)$ indicates that the curves of interest, $X_t(\cdot)$, are not directly observed. Instead, they are recorded on a grid of points and are contaminated by the error process, $e_t(\cdot)$, without assuming any parametric structure on its covariance function, denoted by $C_e(u, v) = \mathrm{Cov}\{e_t(u), e_t(v)\}$. This model guarantees that all the dynamic elements of $W_t(\cdot)$ are included in the signal term $X_t(\cdot)$ and all the white noise elements are absorbed into the noise term $e_t(\cdot)$. Furthermore, we assume that predictor errors $e_t(\cdot)$ are uncorrelated with both $X_{t+k}(\cdot)$ and ε_{t+k}, for all integer k.

Here we turn to discuss the identification of β_0. Assume that $\big\{\big(Y_t, X_t(\cdot)\big)\big\}$ is strictly stationary and $C_0(u, v)$ is the covariance function of $X_t(\cdot)$, which admits the Karhunen-Loève expansion, $X_t(u) = \sum_{j=1}^{\infty} \xi_{tj}\phi_j(u)$, where $\xi_{tj} = \int_{\mathcal{U}} X_t(u)\phi_j(u)du$ and $\mathrm{Cov}(\xi_{tj}, \xi_{tj'}) = \lambda_j I(j = j')$ with $I(\cdot)$ denoting the indicator function. Then the eigenpairs $\{\lambda_j, \phi_j(\cdot)\}_{j\geq 1}$ satisfy the eigen-decomposition $\int_{\mathcal{U}} C_0(u, v)\phi_j(v)dv = \lambda_j\phi_j(u)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Define $S_0(u) = E\big\{Y_t X_t(u)\big\}$, $d = \sup_{i\geq 1}\{i : \lambda_i > 0\}$ and assume $\sum_{j=1}^{d} \lambda_j^{-2}\{\mathrm{Cov}(Y_1, \xi_{1j})\}^2 < \infty$. Obviously β_0 satisfies the following equation

$$S_0(u) = \int_{\mathcal{U}} C_0(u, v)\beta(v)dv, u \in \mathcal{U}. \tag{2.3}$$

If the span of eigenfunctions $\{\phi_1, \ldots, \phi_d\}$ is dense in the $\mathcal{L}_2$ space, which implies that all the elements of interest in the $\mathcal{L}_2$ space can be represented by the eigenfunctions.

It ensures that β_0 is the unique solution to (2.3) and hence can be identified. In a general scenario, β_0 can also be well defined. To make β_0 identifiable, we consider the following minimisation problem

$$\begin{aligned} &\min_{\beta\in\mathcal{L}^2(\mathcal{U})} \int_{\mathcal{U}} \beta^2(u)du, \\ &s.t.\ S_0(u) = \int_{\mathcal{U}} C_0(u,v)\beta(v)dv,\ u\in\mathcal{U}. \end{aligned} \tag{2.4}$$

Noting that the solution to (2.4) exists and is unique, we define the true slope function β_0 to be this unique minimizer in a closed form of $\beta_0 = \sum_{j=1}^{d} \lambda_j^{-1}\mathrm{Cov}(Y_1, \xi_{1j})\phi_j$, which holds for both $d < \infty$ and $d = \infty$. See also Cardot et al. (2003b) and He et al. (2010).

2.2.2 Main idea

In this section, we describe the main idea to facilitate the development of AGMM to estimate $\beta_0(\cdot)$ in (2.2). We choose $X_{t+k}(\cdot)$ for $k = 0, 1, \ldots,$ as functional instrumental variables, which are assumed to be uncorrelated with the error ε_t in (2.2). Let

$$g_k^X(\beta, u) = \mathrm{Cov}\big\{Y_t, X_{t+k}(u)\big\} - \int_{\mathcal{U}} \mathrm{Cov}\{X_t(v), X_{t+k}(u)\}\beta(v)dv. \tag{2.5}$$

The population moment conditions, $E\{\varepsilon_t X_{t+k}(u)\} = 0$ for any $u \in \mathcal{U}$, and equation (2.2) implies that

$$g_k^X(\beta_0, u) \equiv 0 \text{ for any } u \in \mathcal{U} \text{ and } k = 1, \ldots. \tag{2.6}$$

In particular, the conventional LS approach is based on (2.6) with $k = 0$. However, this approach is inappropriate when $X_t(\cdot)$ are replaced by the surrogates $W_t(\cdot)$ given the fact that $C_W(u,v) = \mathrm{Cov}\{W_t(u), W_t(v)\} = C_0(u,v) + C_e(u,v)$, and hence the sample version of $C_W(u,v)$ is not a consistent estimator for $C_0(u,v)$. See Hall and Vial (2006) for the identifiability of $C_0(u,v)$ and $C_e(u,v)$ under the assumption that the observed curves $W_1(\cdot), \ldots, W_n(\cdot)$ are independent and $e_t(\cdot)$ decays to zero as n goes to infinity.

To separate $X_t(\cdot)$ from $W_t(\cdot)$ under the serial dependence scenario, we develop a different approach without requiring the "low noise" condition. For an integer $k \geq 1$, denote the lag-k autocovariance function of $X_t(\cdot)$, by $C_k(u,v) = \mathrm{Cov}\{X_t(u), X_{t+k}(v)\}$,

which does not depend on t. Our method is based on the simple fact that

$$\mathrm{Cov}\{Y_t, W_{t+k}(u)\} = \mathrm{Cov}\{Y_t, X_{t+k}(u)\} \text{ and } \mathrm{Cov}\{W_t(u), W_{t+k}(v)\} = C_k(u,v)$$

for any $k \neq 0$. Then after substituting $X_t(\cdot)$ by $W_t(\cdot)$ in (2.5), we can also represent

$$g_k(\beta, u) = \mathrm{Cov}\big\{Y_t, W_{t+k}(u)\big\} - \int_{\mathcal{U}} \mathrm{Cov}\{W_t(v), W_{t+k}(u)\}\beta(v)dv = g_k^X(\beta, u),$$

and the moment conditions in (2.6) become

$$g_k(\beta_0, u) \equiv 0 \text{ for any } u \in \mathcal{U} \text{ and } k = 1 \ldots, L,$$

where L is some prescribed positive integer.

Under the over-identification setting, where the number of moment conditions exceeds the number of parameters, we borrow the idea of *generalized methods-of-moments* (GMM) based on minimizing the distance from $g_1(\beta, \cdot), \ldots, g_L(\beta, \cdot)$ to zero. This distance is defined by the quadratic form of

$$Q(\beta) = \sum_{k=1}^{L}\sum_{l=1}^{L}\int_{\mathcal{U}}\int_{\mathcal{U}} g_k(\beta, u)\Omega_{k,l}(u,v)g_l(\beta, v)dudv,$$

where $\mathbf{\Omega}(u,v) = \{\Omega_{k,l}(u,v)\}_{1\leq k,l\leq L}$ is an L by L weight matrix whose (k,l)-th element is $\Omega_{k,l}(u,v)$. A suitable choice of $\mathbf{\Omega}(u,v)$ must satisfy the properties of symmetry and positive-definiteness (Guhaniyogi et al., 2013), which are, to be specific, (i) $\Omega_{kl}(u,v) = \Omega_{lk}(v,u)$ for each $k, l = 1, \ldots, L$ and $(u,v) \in \mathcal{U}^2$; (ii) for any finite collection of time points $u_1, \ldots, u_T$, $\sum_{t=1}^{T}\sum_{t'=1}^{T} \mathbf{a}(u_t)^{\mathrm{T}}\mathbf{\Omega}(u_t, u_{t'})\mathbf{a}(u_{t'})$ must be positive for any $\mathbf{a}(\cdot) = \big(a_1(\cdot), \ldots, a_L(\cdot)\big)^{\mathrm{T}}$. In general, one can choose the optimal weight matrix $\mathbf{\Omega}$ and implement a two-step GMM. Functional linear regression (2.2) is equivalent to $Y_t = \int_{\mathcal{U}} W_t(u)\beta_0(u)du + \widetilde{\varepsilon}_t$, where $\widetilde{\varepsilon}_t = \varepsilon_t - \int_{\mathcal{U}} e_t(u)\beta_0(u)du$. Therefore, as suggested by Hansen (1982) and Arellano and Bond (1991), the optimal weighting matrix is $\mathbf{\Omega} = \widehat{\mathbf{S}}^{-1}$, where $\widehat{\mathbf{S}}$ should be the a consistent estimation of the L by L matrix $\mathbf{S}$ whose (k,l)-th element is $\mathbb{E}\{\widetilde{\varepsilon}^2 W_{t+k(u)}W_{t+l}(v)\}$. In general, one can implement a two-step GMM and estimate $\widehat{S}_{hl}(u,v) = (n-L)^{-1}\sum_{t=1}^{n-L}\{\widehat{\varepsilon}_t^2 W_{t+k}(u)W_{t+l}(v)\}$ with $\widehat{\varepsilon}_t = Y_t - \int_{\mathcal{U}} W_t(u)\widehat{\beta}^*(u)du$ for some preliminary consistent estimation $\widehat{\beta}^*$. However, the inverse problem needs further investigation and this would give a very slight improvement in our simulations. To simplify our derivation and accelerate the computation, we choose the identity weight matrix as $\Omega_{k,l}(u,v) = I(k=l)I(u=v)$ and

then minimise the resulting distance of

$$Q(\beta) = \sum_{k=1}^{L} \int_{\mathcal{U}} g_k(\beta, u)^2 du,$$

over $\beta(\cdot) \in \mathcal{L}_2(\mathcal{U})$. The minimiser of $Q(\beta)$, $\beta_0(\cdot)$, can be achieved by solving $\partial Q(\beta)/\partial\beta = 0$, i.e. for any $u \in \mathcal{U}$,

$$\sum_{k=1}^{L} \Big[\int_{\mathcal{U}} C_k(u,z) \mathrm{Cov}\big\{Y_t, W_{t+k}(z)\big\} dz - \int_{\mathcal{U}} \Big\{ \int_{\mathcal{U}} C_k(u,z) C_k(v,z) dz \Big\} \beta(v) dv \Big] = 0. \tag{2.7}$$

To ease our presentation, we define

$$R(u) = \sum_{k=1}^{L} \int_{\mathcal{U}} C_k(u,z) \mathrm{Cov}\{Y_t, W_{t+k}(z)\} dz \tag{2.8}$$

and

$$K(u,v) = \sum_{k=1}^{L} \int_{\mathcal{U}} C_k(u,z) C_k(v,z) dz. \tag{2.9}$$

Note that K can be viewed as the kernel of a linear operator acting on $\mathcal{L}_2(\mathcal{U})$, i.e. for any $f \in \mathcal{L}_2(\mathcal{U})$, K maps $f(u)$ to $\widetilde{f}(u) \equiv \int_{\mathcal{U}} K(u,v) f(v) dv$. For notational economy, we will use K to denote both the kernel and the operator. Indeed, the nonnegative definite operator K was proposed in Bathia et al. (2010) to identify the dimensionality of $X_t(\cdot)$ based on $W_t(\cdot)$ in (2.1). Substituting the relevant terms in (2.7), $\beta_0(\cdot)$ satisfies the following equation

$$R(u) = \int_{\mathcal{U}} K(u,v) \beta(v) dv \text{ for any } u \in \mathcal{U}. \tag{2.10}$$

See also functional extension of the least squares type of normal equation in (2.3).

Provided that $X_t(\cdot)$ is d-dimensional, it follows from Proposition 1 of Bathia et al. (2010) that, under regularity conditions, K has the spectral decomposition, $K(u,v) = \sum_{j=1}^{d} \theta_j \psi_j(u) \psi_j(v)$, with d nonzero eigenvalues $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_d$ and $\overline{\mathrm{span}}\{\psi_1, \ldots, \psi_d\}$ is the linear space spanned by the d eigenfunctions $\{\phi_1, \ldots, \phi_d\}$. This assertion still holds even for $d = \infty$.

Denote the null space of K and its orthogonal complement by $\ker(K) = \{x \in \mathcal{L}_2(\mathcal{U}) : Kx = 0\}$ and $\ker(K)^{\perp} = \{x \in \mathcal{L}_2(\mathcal{U}) : \langle x, y \rangle = 0, \forall y \in \ker(K)\}$, respectively. The inverse operator K^{-1} corresponds to the inverse of the restricted operator $\breve{K} =$

$K\big|\ker(K)^{\perp}$, which restricts the domain of K to $\ker(K)^{\perp}$. See Section 3.5 of Hsing and Eubank (2015) for details. When $d < \infty$, $\beta_0(\cdot)$ is indeed the unique solution to (2.10) in $\ker(K)^{\perp}$ in the form of

$$\beta_0(u) = \int_{\mathcal{U}} K^{-1}(u,v)R(v)dv = \sum_{j=1}^{d} \theta_j^{-1}\langle \psi_j, R\rangle \psi_j(u). \tag{2.11}$$

Provided K is a bounded operator when $d = \infty$, K^{-1} becomes an unbounded operator, which means it is discontinuous and cannot be estimated in a meaningful way. However, K^{-1} is usually associated with another function/operator, the composite function/operator can be reasonably assumed to be bounded, e.g. the regression operator (Li and Solea, 2018). If we further assume that the composite function $\int_{\mathcal{U}} K^{-1}(u,v)R(v)dv$ is bounded, or equivalently $\sum_{j=1}^{\infty} \theta_j^{-2}\langle \psi_j, R\rangle^2 < \infty$, $\beta_0(\cdot)$ is still the unique solution to (2.10) in $\ker(K)^{\perp}$ and is of the form

$$\beta_0(u) = \int_{\mathcal{U}} K^{-1}(u,v)R(v)dv = \sum_{j=1}^{\infty} \theta_j^{-1}\langle \psi_j, R\rangle \psi_j(u). \tag{2.12}$$

Both (2.11) and (2.12) motivate us to develop the estimation procedure for β_0 in Section 2.2.3.

2.2.3 Estimation procedure

In this section, we present the AGMM estimator for $\beta_0(\cdot)$ based on the main idea described in Section 2.2.2. We first provide the estimates of $C_k(u,v)$ and $\mathrm{Cov}\{Y_t, W_{t+k}(u)\}$ for $k = 1, \ldots, L$, i.e.

$$\begin{gathered} \widehat{C}_k(u,v) = \frac{1}{n-L}\sum_{t=1}^{n-L} W_t(u)W_{t+k}(v) \\ \text{and} \\ \widehat{\mathrm{Cov}}\{Y_t, W_{t+k}(u)\} = \frac{1}{n-L}\sum_{t=1}^{n-L} Y_t W_{t+k}(u). \end{gathered} \tag{2.13}$$

Combing (2.8), (2.9) and (2.13) gives the natural estimators for $K(u,v)$ and $R(u)$

as

$$\begin{aligned}\widehat{K}(u,v) &= \sum_{k=1}^{L}\int_{\mathcal{U}}\widehat{C}_k(u,z)\widehat{C}_k(v,z)dz \\ &= \frac{1}{(n-L)^2}\sum_{k=1}^{L}\sum_{t=1}^{n-L}\sum_{s=1}^{n-L}W_t(u)W_s(v)\langle W_{t+k},W_{s+k}\rangle\end{aligned} \tag{2.14}$$

and

$$\begin{aligned}\widehat{R}(u) &= \sum_{k=1}^{L}\int_{\mathcal{U}}\widehat{C}_k(u,z)\widehat{\mathrm{Cov}}\{Y_t,W_{t+k}(z)\}dz \\ &= \frac{1}{(n-L)^2}\sum_{k=1}^{L}\sum_{t=1}^{n-L}\sum_{s=1}^{n-L}W_t(u)Y_s\langle W_{t+k},W_{s+k}\rangle,\end{aligned} \tag{2.15}$$

respectively. Note we choose a fixed integer $L > 1$, as K pulls together the information at different lags, while $L = 1$ may lead to spurious estimation results. See Section 2.2.5 for the discussion on the selection of L.

We next perform an eigenanalysis on $\widehat{K}$ and thus obtain the estimated eigenpairs $\{\widehat{\theta}_j, \widehat{\psi}_j(\cdot)\}$ for $j = 1, 2, \ldots$. When the number of functional observations n is large, the accumulated errors in (2.14), (2.15) and the eigenanalysis on $\widehat{K}$ are relatively small, thus resulting in smooth estimates of $\psi_j(\cdot)$ and $\beta_0(\cdot)$. We refer to this implementation of our method as Base AGMM for the remainder of the chapter. However, in the setting without a sufficiently large n this version of AGMM suffers from a potential under-smoothing problem that the resulting estimate of $\beta_0(\cdot)$ wiggles quite a bit. To overcome this disadvantage, we can impose some level of smoothing in the eigenanalysis through the basis expansion approach, which converts the continuous functional eigenanalysis problem for $\widehat{K}$ to an approximately equivalent matrix eigenanalysis task. We explore this *basis expansion based AGMM*, simply referred to as AGMM from here on. To be specific, let $\mathbf{B}(u)$ be the J-dimensional orthonormal basis function, i.e. $\int_{\mathcal{U}}\mathbf{B}(u)\mathbf{B}^{\mathrm{T}}(u)du = \mathbf{I}_J$, such that for each $j = 1, \ldots, J$, $\psi_j(\cdot)$ can be well approximated by $\boldsymbol{\delta}_j^{\mathrm{T}}\mathbf{B}(\cdot)$, where $\boldsymbol{\delta}_j$ is the basis coefficients vector. Let

$$\widehat{\mathbf{K}} = \int_{\mathcal{U}}\int_{\mathcal{U}}\mathbf{B}(u)\mathbf{B}^{\mathrm{T}}(v)\widehat{K}(u,v)dudv. \tag{2.16}$$

We perform an eigen-decomposition on $\widehat{\mathbf{K}}$, which leads to the estimated eigenpairs $\{(\widehat{\theta}_j, \widehat{\boldsymbol{\delta}}_j)\}_{j=1}^{J}$. Then the j-th estimated principal component function is given by $\widehat{\psi}_j(\cdot) = \widehat{\boldsymbol{\delta}}_j^{\mathrm{T}}\mathbf{B}(\cdot)$. See Section 2.2.5 for the selection of J. A similar basis expansion technique can be applied to produce a smooth estimate $\widehat{R}(\cdot)$. Note that

$\widehat{\mathbf{K}}, \widehat{\theta}_j, \widehat{\psi}_j, j = 1, \ldots, d$, all depend on J, but for simplicity of notation, we will omit the corresponding superscripts where the context is clear.

Finally, we substitute the relevant terms in (2.11) and (2.12) by their estimated values. We discuss two situations corresponding to $d < \infty$ and $d = \infty$ as follows. (i) When $X_t(\cdot)$ is d-dimensional ($d < \infty$), we need to select the estimate $\widehat{d}$ of d in the sense that $\widehat{\theta}_1, \ldots, \widehat{\theta}_{\widehat{d}}$ are large eigenvalues of $\widehat{K}$ and $\widehat{\theta}_{\widehat{d}+1}$ drops dramatically. The estimate $\widehat{\beta}$ of β_0 is then given by

$$\widehat{\beta}(u) = \sum_{j=1}^{\widehat{d}} \widehat{\theta}_j^{-1} \langle \widehat{\psi}_j, \widehat{R} \rangle \widehat{\psi}_j(u). \tag{2.17}$$

(ii) When $X_t(\cdot)$ is an infinite dimensional functional object, we take the standard truncation approach by using the leading M eigenpairs of $\widehat{K}$ to approximate β_0 in (2.12). Specifically, we obtain the estimated slope function as

$$\widehat{\beta}(u) = \sum_{j=1}^{M} \widehat{\theta}_j^{-1} \langle \widehat{\psi}_j, \widehat{R} \rangle \widehat{\psi}_j(u). \tag{2.18}$$

Section 2.2.5 presents details to select $\widehat{d}$ and M. However, when $d = \infty$, the empirical performance of $\widehat{\beta}(\cdot)$ may be sensitive to the selected value of M. To improve the numerical stability, we suggest an alternative ridge-type method to estimate β_0. Specifically, we propose

$$\widehat{\beta}_{\text{ridge}}(u) = \sum_{j=1}^{\bar{M}} (\widehat{\theta}_j + \rho_n)^{-1} \langle \widehat{\psi}_j, \widehat{R} \rangle \widehat{\psi}_j(u), \tag{2.19}$$

where $\bar{M}$ is chosen to be reasonably larger than M and $\rho_n \geq 0$ is a ridge parameter. See also Hall and Horowitz (2007) for the ridge-type estimator in classical functional linear regression.

2.2.4 Generalisation to functional response

In this section, we consider the case when the response is also functional. Given a functional response $Y_t(\cdot)$ and a functional predictor $X_t(\cdot)$, both of which are in $\mathcal{L}_2(\mathcal{U})$ and have mean zero, the function-on-function linear regression takes the form of

$$Y_t(u) = \int_{\mathcal{U}} X_t(v) \gamma_0(u, v) dv + \varepsilon_t(u), \ u \in \mathcal{U}, \ t = 1, \ldots, n, \tag{2.20}$$

where $\gamma_0(u,v)$ is the slope function of interest and $\varepsilon_t(\cdot)$, independent of $X_{t+k}(\cdot)$ for any integer k, are random elements in the underlying separable Hilbert space. We still observe the erroneous version $W_t(\cdot)$ rather than the signal $X_t(\cdot)$ itself in equation (2.1).

To estimate the slope function in (2.20), we develop an AGMM approach analogous to that for the scalar case in Section 2.2 by solving the normal equation of

$$H(u,v) = \int_{\mathcal{U}} K(u,w)\gamma(w,v)dw \text{ for any } v \in \mathcal{U}, \tag{2.21}$$

where $H(u,v) = \sum_{k=1}^{L} \int_{\mathcal{U}} C_k(u,z)\text{Cov}\{Y_t(v), W_{t+k}(z)\}dz$ with its natural estimator

$$\widehat{H}(u,v) = \frac{1}{(n-L)^2} \sum_{k=1}^{L} \sum_{t=1}^{n-L} \sum_{s=1}^{n-L} W_t(u)Y_s(v)\langle W_{t+k}, W_{s+k}\rangle. \tag{2.22}$$

Accordingly, we can provide the estimate $\widehat{\gamma}$ of γ_0 under two functional scenarios including $d < \infty$ and $d = \infty$. (i) When $d < \infty$, $\gamma_0(u,v)$ is the unique solution of (2.21) in $\ker(K)^{\perp}$ and can be represented as

$$\gamma_0(u,v) = \int_{\mathcal{U}} K^{-1}(u,w)H(w,v)dw = \sum_{j=1}^{d} \theta_j^{-1}\langle \psi_j, H(\cdot,v)\rangle \psi_j(u). \tag{2.23}$$

The estimate of $\gamma_0(u,v)$ is then given by

$$\widehat{\gamma}(u,v) = \sum_{j=1}^{\widehat{d}} \widehat{\theta}_j^{-1}\widehat{\psi}_j(u)\langle \widehat{\psi}_j, \widehat{H}(\cdot,v)\rangle. \tag{2.24}$$

(ii) Under the infinite dimensional setting ($d = \infty$, if we assume the boundedness of the composite function $\int_{\mathcal{U}} K^{-1}(u,w)H(w,v)dw$ in the L_2 sense, the solution to (2.21) uniquely exists. Approximating the infinite dimensional $\gamma_0(u,v)$ in (2.23) by the first M components and substituting the relevant terms by their estimated values, we can obtain

$$\widehat{\gamma}(u,v) = \sum_{j=1}^{M} \widehat{\theta}_j^{-1}\widehat{\psi}_j(u)\langle \widehat{\psi}_j, \widehat{H}(\cdot,v)\rangle. \tag{2.25}$$

2.2.5 Selection of tuning parameters

Implementing AGMM requires choosing L (selected lag length in (2.7)), M (truncated dimension in (2.18) when $d = \infty$), $\widehat{d}$ (number of identified nonzero eigenvalues of $\widehat{K}$ when $d < \infty$) and J (dimension of the basis function $\mathbf{B}(u)$). First, the choice of L depends on the strength of serial dependence of the time series. A larger L may take advantages from the possible strong serial dependence and pull together the information at different lags, while using $L = 1$ may cause false choices of $\widehat{d}$. However, when L is too large, it will make $\widehat{K}$ less accurate because strongest autocorrelations usually appear at the small time lags and the "weak instrumental" problem may occur for long lags. Moreover, adding more terms will increase the model complexity and exacerbate estimation, especially when sample size n is small. Therefore, we tend to select a small value and set $L = 5$ in our empirical studies. See also Bathia et al. (2010) and Lam et al. (2011) for relevant discussions. One can also consider a test based method to choose L. Specifically, we may choose L to be the largest k, such that $H_0 : C_k(u, v) = 0$ is rejected while $H_0 : C_{k+1}(u, v) = 0$ is not rejected. Inference for the autocovariance operator of a functional time series has been extensively studied in the literature, see Kokoszka et al. (2017) for example.

Second, to select M when $d = \infty$, the typical approach is to find the largest M eigenvalues of $\widehat{K}$ such that the corresponding cumulative percentage of variation exceeds the pre-specified threshold value, e.g. 90% or 95%. Other available methods include the bootstrap test (Bathia et al., 2010) and the eigen-ratio-based estimator (Lam et al., 2011). Third, to determine $\widehat{d}$ when $d < \infty$, we take the bootstrap approach proposed in Bathia et al. (2010). Our task is to test the null hypothesis $H_0 : \theta_{d+1} = 0$. We reject H_0 if $\widehat{\theta}_{d+1} > c_\alpha$, where the critical value c_α is the $(1 - \alpha)$ quantile of $\widehat{\theta}_{d+1}$ corresponding to the significant level $\alpha \in (0, 1)$. And in practice, we use the bootstrap method to estimate it. We summarise the bootstrap procedure as follows.

1. Define $\widehat{W}_t(\cdot) = \sum_{j=1}^{\widehat{d}} \widehat{\eta}_{tj}\widehat{\psi}_j(\cdot)$, where $\widehat{\eta}_{tj} = \int_{\mathcal{U}} W_t(u)\widehat{\psi}_j(u)du$ for $j = 1, \ldots, \widehat{d}$. Let $\widehat{e}_t(\cdot) = W_t(\cdot) - \widehat{W}_t(\cdot)$.

2. Generate a bootstrap sample using $W_t^*(\cdot) = \widehat{W}_t(\cdot) + e_t^*(\cdot)$, where e_t^* are drawn with replacement from $\{\widehat{e}_1, \ldots, \widehat{e}_n\}$.

3. In an analogy to $\widehat{K}$ defined in (2.14), form an estimator $\widehat{K}^*$ by replacing $\{W_t\}$ with $\{W_t^*\}$. Then calculate the $(d+1)$-th largest eigenvalue θ_{d+1}^* of $\widehat{K}^*$.

We repeat Steps 2 and 3 above B-times and reject H_0 if the event of $\{\widehat{\theta}_{d+1} > \theta_{d+1}^*\}$ occurs more than $[(1-\alpha)B]$ times. Starting with $\widehat{d} = 1$, we sequentially test $\theta_{\widehat{d}+1} = 0$

and increase $\widehat{d}$ by one until the resulting null hypothesis fails to be rejected. Note that determine the dimensionality d is important in both theoretical study and practical application. The bootstrap test approach has shown to be working well (Bathia et al., 2010), but the joint difficulties of serial dependency, error contamination and the functional natural of the data make it hard to establish the theoretical guarantee, which is still an open question to the best of our knowledge. In practice, an "eyeball test" are usually adopted by checking the "elbow point" of a sequence of eigenvalues that decreasing substantially fast. Another choice is based on the percentage of the variance explained (PVE) by leading eigenvalues. Both methods are also feasible for the infinite dimensional case but effectively arbitrary. Alternatively, some ratio-based methods are proposed for better theoretical justification, see Lam and Yao (2012) and Xia et al. (2015) for details.

Fourth, to select J, we propose the following G-fold cross-validation (CV) approach.

1. Sequentially divide the set $\{1, \ldots, n\}$ into G blockwise groups, $\mathcal{D}_1, \ldots, \mathcal{D}_G$, of approximately equal size.

2. Treat the g-th group as a validation set. Implement the regularised eigenanalysis in Section 2.2.3 on the remaining $G-1$ groups, compute $\widehat{\mathbf{K}}^{(-g)}$ and let $\widehat{\boldsymbol{\delta}}_1^{(-g)}, \ldots, \widehat{\boldsymbol{\delta}}_d^{(-g)}$ be the top d eigenvectors of $\widehat{\mathbf{K}}^{(-g)}$.

3. Compute $\widehat{K}^{(g)}(u, v)$ and $\widehat{\mathbf{K}}^{(g)}$ based on the validation set. and Let $\widehat{\theta}_l^{(g)} = (\widehat{\boldsymbol{\delta}}_l^{(-g)})^{\mathrm{T}} \widehat{\mathbf{K}}^{(g)} \widehat{\boldsymbol{\delta}}_l^{(-g)}$ for $l = 1, \ldots, d$.

We repeat Steps 2 and 3 above G times and choose J as the value that minimize the following mean CV error

$$\mathrm{CV}(J) = \frac{1}{G} \sum_{g=1}^{G} \int_{\mathcal{U}} \int_{\mathcal{U}} \left\{ \widehat{K}^{(g)}(u, v) - \sum_{j=1}^{d} \widehat{\theta}_j^{(g)} (\widehat{\boldsymbol{\delta}}_j^{(-g)})^{\mathrm{T}} \mathbf{B}(u) \mathbf{B}(v)^{\mathrm{T}} \widehat{\boldsymbol{\delta}}_j^{(-g)} \right\}^2 du dv.$$

Given the time break on the training observations, the autocovariance assumption is jeopardised by $L = 5$ lagged terms. However, this effect on $\widehat{K}$ is negligible especially when n is sufficiently large, hence our proposed CV approach can still be practically applied. See also Bergmeir et al. (2018) for various CV methods for time dependent data.

2.3 Theoretical properties

In this section, we investigate the theoretical properties of our proposed estimators for both scalar-on-function and function-on-function linear regressions.

To present the asymptotic results, we need the following regularity conditions.

Condition 2.1. $\{W_t(\cdot), t = 1, 2, \dots\}$ is strictly stationary curve time series. Define the ψ-mixing with the mixing coefficients

$$\psi(l) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_l^\infty, P(A)P(B)>0} |1 - P(B|A)/P(B)|, \quad l = 1, 2, \dots,$$

where $\mathcal{F}_i^j$ denotes the σ-algebra generated by $\{W_t(\cdot), i \le t \le j\}$. Moreover, it holds that $\sum_{l=1}^\infty l\psi^{1/2}(l) < \infty$.

Condition 2.2. $E(\|W_t\|^4) < \infty$ and $E(\varepsilon_t^2) < \infty$.

The presentation of the ψ-mixing condition in Condition 2.1 is mainly for technical convenience. See Section 2.4 of Bosq (2000) on the mixing properties of curve time series. Condition 2.2 is the standard moment assumption in functional regression literature (Chakraborty and Panaretos, 2017, Hall and Horowitz, 2007).

Condition 2.3. (i) When d is fixed, $\theta_1 > \cdots > \theta_d > 0 = \theta_{d+1}$; (ii) When $d = \infty$, $\theta_1 > \theta_2 > \cdots > 0$, and there exist some positive constants c and $\alpha > 1$ such that $\theta_j - \theta_{j+1} \ge cj^{-\alpha-1}$ for $j \ge 1$; (iii) $\overline{\text{span}}\{\phi_1, \dots, \phi_d\} = \overline{\text{span}}\{\psi_1, \dots, \psi_d\}$.

Condition 2.4. When $d = \infty$, $\beta_0(u) = \sum_{j=1}^\infty b_j\psi_j(u)$ and there exist some positive constants $\tau \ge \alpha + 1/2$ and C such that $|b_j| \le Cj^{-\tau}$ for $j \ge 1$.

Condition 2.3 restricts the eigen-structure of K and assumes that all the nonzero eigenvalues of K are distinct from each other. Note that in the case where $W_t(\cdot)$ are independent, the autocovariance $C_k = 0$ for $k \ge 0$. Therefore, eigenvalues $\theta_j = 0$ for all $j \ge 0$, which means that there is no lag information available form the observation. So, the idea that using lag terms as the instrumental variables is no longer valid for violating Condition 2.3.

To this end, one may conciser to check the existence of the serial dependence beforehand. The serial correlation test can rely on the results from dimensional reduction via e.g. functional principal component (FPC) analysis, which transforms functional observations into a vector time series of FPC scores. Then, the multivariate time series technique could be used to investigate the dependence as summarised in Tsay (2013). An alternative way is to perform the test using functional objects. Since the linear correlation captured by the autocovariance operator is most relevant in our study, we can perform some test procedure to measure the (cumulative) significance of the first $L > 0$ empirical autocovariance. One can check the serial correlation by testing $H_0 : \forall_{h \in \{1,\dots,L\}} C_h(u, v) = 0$, see Kokoszka et al. (2017) for example.

Moreover, in our proposed autocovariance based frame work, we may also derive some procedure to perform the inference directly on the operator $K(u, v)$ by testing $H_0 : K(u, v) = 0$. And the techniques relying on a positive-definite operator, which has been investigated in Horváth et al. (2014) and Kokoszka et al. (2017), among others. And we will pursue the formal definition and theoretical properties of this test in the future study.

When $d = \infty$, Condition 2.3 (ii) prevents gaps between adjacent eigenvalues from being too small. The parameter α determines the tightness of eigen-gaps with larger values of α yielding tighter gaps. This condition also indicates that $\theta_j \geq c\alpha^{-1}j^{-\alpha}$ as $\theta_j = \sum_{k=j}^{\infty}(\theta_k - \theta_{k+1}) \geq c\sum_{k=j}^{\infty} k^{-\alpha-1}$, and can be used to derive the convergence rates of estimated eigenfunctions. See also Hall and Horowitz (2007) and Qiao et al. (2020). Condition 2.3 (iii) implies that no components in $\mathbf{X}_t$ are serial uncorrected. Therefore, the space spanned by the eigenfunctions of K is sufficient for recovering $\mathbf{X}_t$. Condition 2.4 restricts β_0 based on its expansion using eigenfunctions of K. The parameter τ determines the decay rate of slope basis coefficients, $\{b_j\}_{j=1}^{\infty}$. The assumption $\tau \geq \alpha + 1/2$ can be interpreted as requiring β_0 be sufficiently smooth relative to K, the smoothness of which can be implied by $\theta_j \geq c\alpha^{-1}j^{-\alpha}$ from Condition 2.3 (ii). See Hall and Horowitz (2007) for an analogous condition in functional linear regression.

Before presenting Theorem 2.1 for the asymptotic analysis of the scalar-on-function linear regression, we first solidify some notation. For any univariate function f, define $\|f\| = \sqrt{\langle f, f\rangle}$. We denote by $\|A\|_{\mathcal{S}}$ the Hilbert-Schmidt norm for any bivariate function A. The notation $a_n \asymp b_n$ for positive a_n and b_n means that the ratio a_n/b_n is bounded away from zero and infinity. To obtain $\widehat{\beta}$ in (2.17) when $d < \infty$, we use the consistent estimator for d defined as $\widehat{d} = \#\{j : \widehat{\theta}_j \geq \epsilon_n\}$, where ϵ_n satisfies the condition in Theorem 2.1 (i) below. Then by Theorem 3 of Bathia et al. (2010), $\widehat{d}$ converges in probability to d as $n \to \infty$.

Theorem 2.1. *Suppose that Conditions 2.1–2.4 hold. The following assertions hold as* $n \to \infty$:
(i) Let $\epsilon_n \to 0$ *and* $\epsilon_n^2 n \to \infty$ *as* $n \to \infty$. *When* d *is fixed, then*

$$\|\widehat{\beta} - \beta_0\| = O_P\big(n^{-1/2}\big).$$

(ii) When $d = \infty$, *if we further assume that* $M \asymp n^{1/(2\alpha+2\tau)}$, *then*

$$\|\widehat{\beta} - \beta_0\|^2 = O_P\big(M^{2\alpha+1}n^{-1} + M^{-2\tau+1}\big) = O_P\big(n^{-\frac{2\tau-1}{2\alpha+2\tau}}\big).$$

Remarks. (a) When d is fixed, the standard parametric root-n rate is achieved. The parametric rate is optimal, because the proof of Theorem 1.1(i) relies on the convergence of the eigenpairs $\{\widehat{\psi}_j, \widehat{\theta}_j\}$, and both $\widehat{\psi}_j$ and $\widehat{\theta}_j$, $j = 1, \ldots, d$ with fixed d enjoy the root-n consistency (and some faster rates for $j > d$). See discussion in Guo and Qiao (2020) and the references therein. (b) When $d = \infty$, the convergence rate is governed by two sets of parameters (1) dimensionality parameter, sample size (n); (2) internal parameters, truncated dimension of the curve time series (M), decay rate of the lower bounds for eigenvalues (α), decay rate of the upper bounds for slope basis coefficients (τ). It is easy to see that larger values of α (tighter eigengaps) yield a slower convergence rate, while increasing τ enhances the smoothness of $\beta_0(\cdot)$, thus resulting in a faster rate. The convergence rate consists of two terms, which reflects our familiar variance-bias tradeoff as commonly considered in nonparametric statistics. In particular, the bias is bounded by $O(M^{-\tau+1/2})$ and the variance is of the order $O_P(M^{2\alpha+1}n^{-1})$. To balance both terms, we choose the truncated dimension, $M \asymp n^{1/(2\alpha+2\tau)}$, while the optimal convergence rate then becomes $O_P\{n^{-(2\tau-1)/(2\alpha+2\tau)}\}$. It is also worth noting that this rate is slightly slower than the minimax rate $O_P\{n^{-(2\tau-1)/(\alpha+2\tau)}\}$ in Hall and Horowitz (2007), which considers independent observations of the functional predictor without any error contamination. In fact, we tackle a more difficult functional linear regression scenario, where extra complications come from the serial dependence and functional error contamination. From a theoretical perspective, whether the rate in part (ii) is optimal in the minimax sense is still of interest and requires further investigation.

Moreover, inference for the slope function in linear regression (2.2) has been widely studied, see Imaizumi and Kato (2019) and references therein for example. A typical concern is to check if the true slope coefficient function $\beta_0 = 0$ for the purpose of determining the explanatory power of the model. To address this problem, a hypothesis test $H_0 : \beta_0 = \beta_0^H$ v.s. $H_1 : \beta_0 \neq \beta_0^H$, is carried out by Cardot et al. (2003a), and specifically take $\beta_0^H = 0$ for instance (Kong et al., 2016a). Recently, Babii (2020) develops honest confidence bands of the functional instrumental variable (IV) regression for i.i.d. sample and Kutta et al. (2021) extended the test to functional time series. However, the test procedure for our proposed model is nontrivial because it should be carefully devised for functional times series regression with measurement error under GMM estimation. And we will pursue it in the future work.

Before presenting the asymptotic results for the function-on-function linear regression, we list Conditions 2.5 and 2.6 below, which are substitutes of Conditions 2.2 and 2.4, respectively, in the functional response case.

Condition 2.5. $E(\|W_t\|^4) < \infty$ and $E(\|\varepsilon_t\|^2) < \infty$.

Condition 2.6. When $d = \infty$, $\gamma_0(u,v) = \sum_{j=1}^{\infty}\sum_{\ell=1}^{\infty} b_{j\ell}\psi_j(u)\psi_\ell(v)$ and there exist some positive constants $\tau \geq \alpha + 1/2$ and C such that $|b_{j\ell}| \leq C(j+\ell)^{-\tau-1/2}$ for $j, \ell \geq 1$.

Theorem 2.2. *Suppose that Conditions 2.1, 2.3, 2.5 and 2.6 hold. The following assertions hold as* $n \to \infty$:
(i) Let $\epsilon_n \to 0$ *and* $\epsilon_n^2 n \to \infty$ *as* $n \to \infty$. *When* d *is fixed, then*

$$\|\widehat{\gamma} - \gamma_0\|_{\mathcal{S}} = O_P(n^{-1/2}).$$

(ii) When $d = \infty$, *if we further assume that* $M \asymp n^{1/(2\alpha+2\tau)}$, *then*

$$\|\widehat{\gamma} - \gamma_0\|_{\mathcal{S}}^2 = O_P\big(M^{2\alpha+1}n^{-1} + M^{-2\tau+1}\big) = O_P\big(n^{-\frac{2\tau-1}{2\alpha+2\tau}}\big).$$

2.4 Partially observed functional predictor

In this section, we consider a practical scenario where each $W_t(\cdot)$ is partially observed at random time points, $U_{t1}, \ldots, U_{tm_t} \in \mathcal{U} = [0,1]$, where for dense measurement designs all m_t's are larger than some order of n, and for sparse designs all m_t's are bounded (Qiao et al., 2020, Zhang and Wang, 2016). Let Z_{ti} represent the observed value of $W_t(U_{t_i})$ satisfying

$$Z_{ti} = W_t(U_{ti}) + \eta_{ti}, \quad i = 1, \ldots, m_t, \tag{2.26}$$

where η_{ti}'s are i.i.d. random errors with finite variance, independent of $W_t(\cdot)$.

Let $K(\cdot)$ be an univariate kernel function. We apply a local linear surface smoother to estimate the lag-k autocovariance function $C_k(u,v)$ for $k = 1, \ldots, L$ by minimizing

$$\sum_{t=1}^{n-L}\sum_{i=1}^{m_t}\sum_{j=1}^{m_{t+k}}\Big\{Z_{ti}Z_{(t+k)j} - a_0^{(k)} - a_1^{(k)}(U_{ti}-u) - a_2^{(k)}(U_{(t+k)j}-v)\Big\}^2 K_{k,i,j,t,h}(u,v) \tag{2.27}$$

with respect to $(a_0^{(k)}, a_1^{(k)}, a_2^{(k)})$, where $K_{k,i,j,t,h}(u,v) = K\left(\frac{U_{ti}-u}{h_C}\right)K\left(\frac{U_{(t+k)j}-v}{h_C}\right)$ with a bandwidth $h_C > 0$. Let the minimizer of (2.27) be $(\widehat{a}_0^{(k)}, \widehat{a}_1^{(k)}, \widehat{a}_2^{(k)})$ and the resulting lag-k autocovariance estimator is $\widetilde{C}_k(u,v) = \widehat{a}_0^{(k)}$. Similarly, we implement a local linear smoothing approach to estimate $S_k(u) = \mathrm{Cov}(Y_t, W_{t+k}(u))$ for $k = 1, \ldots, L$

by minimizing

$$\sum_{t=1}^{n-L}\sum_{i=1}^{m_t}\left\{Y_t Z_{(t+k)i} - b_0^{(k)} - b_1^{(k)}(U_{(t+k)i} - u)\right\}^2 K\left(\frac{U_{ti}-u}{h_S}\right) \tag{2.28}$$

with respect to $(b_0^{(k)}, b_1^{(k)})$ with a bandwidth $h_S > 0$. Then we obtain the estimate $\widetilde{S}_k(u) = \widehat{b}_0^{(k)}$. We also develop a basis expansion approach (Radchenko et al., 2015) to estimate C_k and S_k, where details can be found in Section 2.6.1 of the Appendix. Let $\widetilde{K}(u,v) = \sum_{k=1}^{L}\int_{\mathcal{U}}\widetilde{C}_k(u,z)\widetilde{C}_k(v,z)dz$ with estimated eigenpairs $(\widetilde{\theta}_j, \widetilde{\psi}_j)_{j\geq 1}$ and $\widetilde{R}(u) = \sum_{k=1}^{L}\int_{\mathcal{U}}\widetilde{C}_k(u,z)\widetilde{S}_k(z)dz$. In analogy to (2.17) and (2.18), we obtain the corresponding estimates $\widetilde{\beta}$ of β_0 by replacing $(\widehat{\theta}_j, \widehat{\psi}_j)_{j\geq 1}$ and $\widehat{R}$ with $(\widetilde{\theta}_j, \widetilde{\psi}_j)_{j\geq 1}$ and $\widetilde{R}$, respectively. Before presenting the main asymptotic results, we impose the following regularity conditions.

Condition 2.7. (i) The errors $\{\eta_{ti}\}$ are *i.i.d.* mean zero random variables with $E|\eta_{ti}|^{2s} < \infty$ for some $s > 2$; (ii) $\{W_t(\cdot), t = 1, 2, \cdots\}$ is strictly stationary with ψ-mixing coefficients $\psi(l)$ satisfying $\psi(l) \lesssim l^{-\lambda}$ with $\lambda > \frac{3s-2}{s-2}$ and $\sup_{u\in[0,1]} E|W_t(u)|^{2s} < \infty$.

Condition 2.8. $K(\cdot)$ is a symmetric probability density function on $[-1, 1]$ and is Lipschitz continuous.

Condition 2.9. $\{U_{ti}, i = 1, \ldots, m_t\}$ are i.i.d. copies of a random variable U defined on $[0, 1]$ and the density $f(\cdot)$ of U is twice continuously differentiable and is bounded from below and above over $[0, 1]$.

Condition 2.10. $\{W_t\}$ are independent of $\{U_{ti}\}$ and $\{\eta_{ti}\}$ are independent of $\{U_{ti}\}, \{W_t\}$.

Condition 2.11. (i) $\partial^2 C_k(u,v)/\partial u^2, \partial^2 C_k(u,v)/\partial u\partial v$ and $\partial^2 C_k(u,v)/\partial v^2$ for $k \geq 1$ are uniformly continuous and bounded on $[0, 1]^2$; (ii) $\partial^2 S_k(u)/\partial u^2$ for $k \geq 1$ are uniformly continuous and bounded on $[0, 1]$.

Condition 2.12. The number m_t of measurement locations in time t are independent random variables with distribution $m_t\rho_n^{-1} \sim \breve{m}$, where $\breve{m} \in \{1, \ldots, \bar{m}\}$ for some bounded $\bar{m}$ such that $P(\breve{m} > 1) > 0$.

Condition 2.13. The bandwidth parameters h_C and h_S satisfy

$$h_C \to 0,\ h_S \to 0,\ \frac{\log(n\rho_n^2)}{(n\rho_n^2)^{\theta_C}h_C^2} \to 0 \text{ and } \frac{\log(n\rho_n)}{(n\rho_n)^{\theta_S}h_S} \to 0,$$

with

$$\theta_C = \frac{\beta - 2 - (1+\beta)/(s-1)}{\beta + 2 - (1+\beta)/(s-1)},\quad \theta_S = \frac{\beta - 3 - (1+\beta)/(s-1)}{\beta + 1 - (1+\beta)/(s-1)}.$$

Conditions 2.7–2.13 are standard in local linear smoothing when the serial dependence exists (Hansen, 2008, Rubín and Panaretos, 2020). In Condition 2.12, we treat the number m_t of measurement locations as random variables, but possibly diverges with n at the order of ρ_n. When ρ_n is bounded, it corresponds to the sparse case in Rubín and Panaretos (2020).

We present the convergence rates of $\widetilde{C}_k, \widetilde{S}_k$ for $k \geq 1$ and $\widetilde{\beta}_0$ in the following Theorems 2.3 and 2.4, respectively.

Theorem 2.3. *Suppose that Conditions 2.7–2.13 hold. As $n \to \infty$, we have*

$$\|\widetilde{C}_k - C_k\|_{\mathcal{S}} = O_P\left(\delta_{n1}\right) \text{ and } \|\widetilde{S}_k - S_k\| = O_P\left(\delta_{n2}\right) \text{ for } k \geq 1,$$

where

$$\delta_{n1} = \frac{1}{\sqrt{n\rho_n^2 h_C^2}} + \frac{1}{\sqrt{n}} + h_C^2 \text{ and } \delta_{n2} = \frac{1}{\sqrt{n\rho_n h_S}} + \frac{1}{\sqrt{n}} + h_S^2.$$

Theorem 2.4. *Suppose that Conditions 2.3–2.4 and 2.7–2.13 hold. The following assertions hold as $n \to \infty$:*
(i) Let $\epsilon_n \to 0$ and $\epsilon_n^2 n \to \infty$ as $n \to \infty$. When d is fixed, then

$$\|\widetilde{\beta} - \beta_0\| = O_P\big(\delta_{n1} + \delta_{n2}\big).$$

(ii) When $d = \infty$, if we further assume that $M \asymp \delta_{n1}^{-2/(2\alpha+2\tau)} + \delta_{n2}^{-2/(2\alpha+2\tau)}$, then

$$\|\widetilde{\beta} - \beta_0\|^2 = O_P\Big\{M^{2\alpha+1}\big(\delta_{n1}^2 + \delta_{n2}^2\big) + M^{-2\tau+1}\Big\} = O_P\Big\{\delta_{n1}^{\frac{2(2\tau-1)}{2\alpha+2\tau}} + \delta_{n2}^{\frac{2(2\tau-1)}{2\alpha+2\tau}}\Big\}.$$

Remarks. (a) In the sparse case where ρ_n is bounded, the L_2 rates of convergence for $\widetilde{C}_k$ and $\widetilde{S}_k$ in Theorem 2.3 become $O_P(n^{-1/2}h_C^{-1} + h_C^2)$ and $O_P(n^{-1/2}h_S^{-1/2} + h_S^2)$, respectively, which are consistent to those yielded convergence rates of one-dimensional and surface local linear smoothers for independent and sparsely sampled functional data (Zhang and Wang, 2016). When ρ_n grows with n, the convergence result reveals interesting phase transition phenomena depending on the relative order of ρ_n to n. We use different rates of $\widetilde{C}_k$ ($k \geq 1$) to illustrate such phenomenon:

i. When $\rho_n/n^{1/4} \to 0$ with $n^{1/4}h \to \infty$, $\|\widetilde{C}_k - C_k\| = O_P(n^{-1/2}\rho_n^{-1}h_C^{-1} + h_C^2)$;

ii. When $\rho_n \asymp n^{1/4}$ with $h_C \asymp n^{-1/4}$ or $\rho_n/n^{1/4} \to \infty$ with $h_C = o(n^{-1/4})$ and $h_C\rho_n \to \infty$, $\|\widetilde{C}_k - C_k\| = O_P(n^{-1/2})$.

As ρ_n grows very fast, case (ii) results in the root-n rate, presenting the theory for very dense curve time series falls in the parametric paradigm. As ρ_n grows

moderately fast, case (i) corresponds to the rate faster than that for sparse data but slower than root-n. The rates under cases (i) and (ii) are respectively consistent to those of the estimated covariance function under categories of "dense" and "ultra-dense" functional data (Zhang and Wang, 2016). For $\widetilde{S}_k$ ($k \geq 1$), similar phase transition phenomenon occurs based on the ratio of ρ_n to $n^{1/4}$.

(b) The L_2 rates of $\widetilde{\beta}_0$ in Theorem 2.4 are governed by dimensionality parameters (n, ρ_n), bandwidth parameters (h_C, h_S) and those internal parameters in part (ii) of Theorem 2.1 when $d = \infty$. There also exists the phase transition based on the relative order of ρ_n to n. For example, when ρ_n is bounded and d is fixed, the rate of $\widetilde{\beta}_0$ is $O_P(n^{-1/2}h_C^{-1} + n^{-1/2}h_S^{-1/2} + h_C^2 + h_S^2)$. When ρ_n grows very fast with $\rho_n^{-1} = O(n^{-1/4})$ and suitable choices of h_C, h_S, the rates of $\widetilde{\beta}_0$ are identical to those for fully observed functional predictors in Theorem 2.1.

(c) Theorem 2.4 can be extended to functional response case. Let $N_k(u, v) = \mathrm{Cov}(Y_t(u), W_{t+k}(v))$ for $k = 1, \ldots, L$ and $\widetilde{N}_k$ be the local linear smoothing estimation of N_k. Then, in analogue to (2.24) and (2.25), we can get the estimates $\widetilde{\gamma}$ of γ_0 in the same manner as estimating β_0 by $\widetilde{\beta}$. Suppose we can show that $\|\widetilde{N}_k - N_k\|_{\mathcal{S}} = O_p(\delta_{n3})$, where δ_{n3} is likewise defined as δ_{n1} under some regularisation conditions imposed in Theorem 2.3, then we can extend Theorem 2.4 and get the same convergence rate for $\|\widetilde{\gamma} - \gamma_0\|_{\mathcal{S}}$.

2.5 Empirical studies

2.5.1 Simulation study

In this section, we evaluate the finite sample performance of the two versions of AGMM by a number of simulation studies. The basis expansion based AGMM is referred to as "AGMM", which relies on the regularised $\widehat{\mathbf{K}}$ defined in (2.16), while "Base AGMM" is based on $\widehat{K}$ in (2.14). The observed predictor curves, $W_t(u), u \in [0, 1]$, are generated from equation (2.1) with

$$X_t(u) = \sum_{j=1}^{d} \xi_{tj}\phi_j(u) \quad \text{and} \quad e_t(u) = \sum_{j=1}^{10} \nu_{tj}\zeta_j(u),$$

where $\{\xi_{tj}\}_{t=1}^n$ follows a linear AR(1) process with the coefficient $(-1)^j(0.9 - 0.5j/d)$. The slope functions are generated by $\beta_0(u) = \sum_{j=1}^d b_j\phi_j(u)$, where b_j's take values from the first d components in $(2, 1.6, -1.2, 0.8, -1, -0.6)$. We generate responses

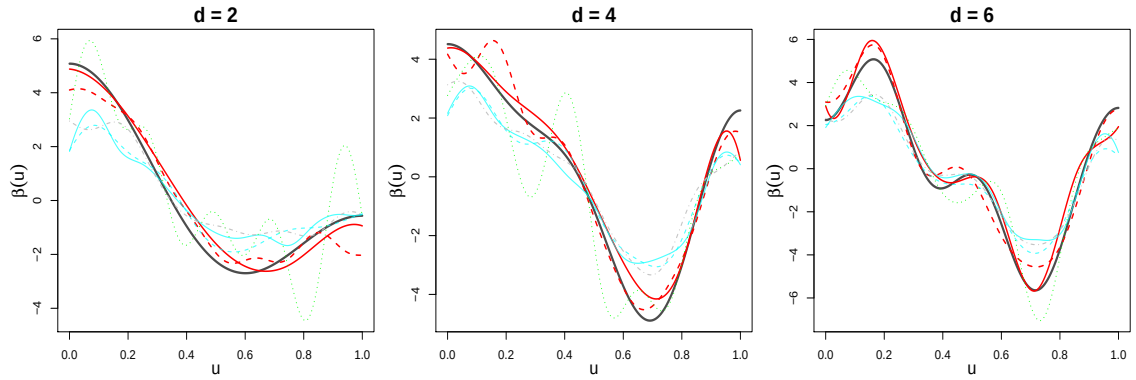


Figure 2.1: Example 1 with $n = 800$ and $d = 2, 4, 6$: Comparison of true $\beta(\cdot)$ functions (black solid) with median estimates over 100 simulation runs for AGMM (red solid), Base AGMM (red dashed), CLS (cyan solid), Base CLS (cyan dashed), Base CGMM (green dotted) and Base ALS (gray dash-dotted).

$Y_1, \ldots, Y_n$ from equation (2.2), where ε_t are independent $N(0, 1)$ variables. Finally, we consider two different scenarios to generate $\{\phi_j(\cdot)\}_{j=1}^d$, $\{\zeta_j(\cdot)\}_{j=1}^{10}$ and $\{\nu_{tj}\}_{n\times 10}$.

Example 1: This example is taken from Bathia et al. (2010) with

$$\phi_j(u) = \sqrt{2}\cos(\pi j u), \quad \zeta_j(u) = \sqrt{2}\sin(\pi j u),$$

and the innovations ν_{tj} being independent standard normal variables.

We compare two versions of AGMM with three competing methods: covariance-based LS (CLS), covariance-based GMM (CGMM), autocovariance-based LS (ALS). The three competing approaches are implemented as follows. In the first two methods, we perform eigenanalysis on the estimated covariance function $\widehat{C}_W$, which converts the functional linear regression to the multiple linear regression, and then implement either LS or GMM. The truncated dimension was chosen such that the selected principal components can explain more than 90% of the variation in the trajectory. We also tried the bootstrap method in Hall and Vial (2006) or to set a larger threshold level, e.g. 95%. However neither approach performed well, so we do not report the results here. The third ALS method relies on the eigenanalysis on the estimated autocovariance-based $\widehat{K}$ and the subsequent implementation of LS. In a similar fashion to the difference between Base AGMM and AGMM, we refer to each of the unregularised method as the "base" version.

The performance of four types of approaches are examined based on the mean integrated squared error for $\widehat{\beta}(u)$, i.e. $E[\int\{\widehat{\beta}(u) - \beta_0(u)\}^2 du]$. We consider different settings with $d = 2, 4, 6$ and $n = 200, 400, 800$, and ran each simulation 100 times. The regularised versions of CGMM and ALS did not give improvements in our simulation studies, so we do not report their results here. Figure 2.1 provides a graphical

illustration of the results for $n = 800$ and $d = 2, 4, 6$. The black solid lines correspond to the true $\beta(u)$ from which the data were generated. The median most accurate estimate is also plotted for each of the competing methods. It is easy to see that the AGMM methods apparently provide the highest level of accuracy. The top part of Table 2.1 reports numerical summaries for all simulation scenarios. We can observe that the advantage of AGMM over Base AGMM is prominent especially when either d or n is relatively small, while AGMM methods are superior to the competing methods when $n = 400$ or 800. However, under the setting with $n = 200$ and $d = 4$ or 6, the bootstrap test in Section 2.2.5 could not select $\widehat{d}$ very accurately, thus resulting in AGMM estimates inferior to some competitors.

Table 2.1: *Example 1*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs. The lowest values are in bold font.

$\widehat{d}$	n	d	Base CLS	CLS	Base CGMM	Base ALS	Base AGMM	AGMM
		2	1.320(0.026)	1.315(0.025)	2.215(0.099)	1.619(0.044)	1.187(0.052)	**0.720(0.033)**
	200	4	1.360(0.028)	**1.340(0.028)**	2.128(0.093)	2.451(0.102)	2.053(0.117)	1.704(0.107)
		6	1.337(0.030)	**1.320(0.029)**	1.912(0.102)	2.150(0.092)	1.847(0.098)	1.612(0.072)
		2	1.184(0.018)	1.181(0.019)	1.891(0.090)	1.338(0.026)	0.772(0.034)	**0.498(0.028)**
Est	400	4	1.198(0.021)	1.199(0.021)	1.939(0.090)	1.316(0.028)	0.701(0.034)	**0.584(0.034)**
		6	1.159(0.023)	1.154(0.022)	1.519(0.087)	1.323(0.034)	0.824(0.045)	**0.745(0.037)**
		2	1.159(0.012)	1.158(0.012)	1.792(0.080)	1.161(0.013)	0.346(0.013)	**0.211(0.012)**
	800	4	1.161(0.014)	1.160(0.014)	1.762(0.105)	1.122(0.014)	0.336(0.015)	**0.247(0.012)**
		6	1.123(0.014)	1.122(0.014)	1.297(0.091)	1.119(0.016)	**0.348(0.016)**	0.350(0.018)
		2	1.402(0.032)	1.238(0.030)	0.774(0.044)	1.637(0.044)	1.196(0.052)	**0.718(0.033)**
	200	4	1.365(0.030)	1.191(0.029)	0.924(0.056)	1.515(0.043)	1.214(0.071)	**0.797(0.046)**
		6	1.345(0.028)	1.272(0.027)	**1.150(0.065)**	1.465(0.036)	1.378(0.070)	1.196(0.057)
		2	1.226(0.019)	1.145(0.019)	0.503(0.027)	1.336(0.026)	0.772(0.034)	**0.498(0.028)**
True	400	4	1.199(0.021)	1.139(0.021)	0.529(0.024)	1.237(0.022)	0.653(0.032)	**0.488(0.029)**
		6	1.166(0.023)	1.139(0.022)	**0.656(0.038)**	1.170(0.023)	0.726(0.039)	0.704(0.042)
		2	1.174(0.012)	1.136(0.012)	0.269(0.011)	1.161(0.013)	0.346(0.013)	**0.211(0.012)**
	800	4	1.165(0.014)	1.131(0.014)	0.324(0.014)	1.130(0.014)	0.333(0.015)	0.**245(0.012)**
		6	1.121(0.014)	1.119(0.014)	**0.323(0.016)**	1.106(0.015)	0.336(0.015)	0.334(0.016)

To investigate the performance of AGMM after excluding the negative impact from the low accuracy of $\widehat{d}$ especially when $n = 200$, we also implement an "oracle" version, which uses the true d in the estimation. The numerical results are reported in the bottom part of Table 2.1. We can observe that GMM methods are superior to their LS versions, while CGMM slightly outperforms AGMM. These observations are due to the facts that, (i) top d eigenvalues for C_W and K correspond to the same signal components in Example 1, (ii) GMM methods are capable of removing the impact from the noise term, (iii) the estimate $\widehat{C}_W$ in CGMM does not consider the functional error, while $\widehat{K}$ in AGMM would suffer from error accumulations. To

better demonstrate the superiority of AGMM, we explore Example 2 below, where the covariance-based approach would fail to identify the signal components but its autocovariance-based version could.

Example 2: We generate $\{\zeta_j(\cdot)\}_{j=1}^{10}$ from a 10-dimensional orthonormal Fourier basis function, $\{\sqrt{2}\cos(2\pi ju), \sqrt{2}\sin(2\pi ju)\}_{j=1}^{5}$, and set $\phi_j(u) = \zeta_j(u)$ for $j = 1, \ldots, d$. The innovations ν_{tj} are independently sampled from $N(0, \sigma_j^2)$ with

$$\sigma_j^2 = \begin{cases} (1/2)^{j-1}, \text{ for } j = 1, \ldots, 6, \\ (2.6 - 0.1j) \times 1.1^{(d/2-3)}, \text{ for } j = 7, \ldots, 10. \end{cases}$$

In this example, provided the fact that $\{\phi_j(\cdot)\}_{j=1}^{d}$ shares the common basis functions with the first d elements in $\{\zeta_j(\cdot)\}_{j=1}^{10}$, we can calculate the variation in the trajectory explained by each of the 10 components under the population level. Table 2.2 reports the variance explained by each of the 10 components under the population level. For each of the three parts corresponding to $d = 2, 4$ and 6, the second and third rows provide the variance explained by each of the d signal components and 10 error components, respectively. The first row ranks the components based on the overall variance explained by each individual component, where the fourth row displays the corresponding values. Take $d = 4$ as an illustrative example, the autocovariance-based approach can correctly identify the first four signal components, while the covariance-based approach can only correctly identify "1" and "2", but incorrectly select "7" and "8" as signal components. Moreover, we consider another scenario for Example 2 by generating innovations $\{\nu_{tj}\}$ from a standard normal distribution, where the variance decomposition is illustrated via Table 2.3. Under this setting, we can observe that both approaches are capable of correctly identifying the d signal components.

Table 2.4 gives numerical summaries under the "oracle" scenario with true d in the estimation. As we would expect, two versions of AGMM provide substantially improved estimates, while Base AGMM is outperformed by AGMM in most of the cases. Under the scenario that $\widehat{d}$ is selected by the bootstrap approach, Figure 2.2 and Table 2.4 provide the graphical and numerical results, respectively. We observe similar trends as in Figure 2.1 and Table 2.1 with AGMM methods providing highly significant improvements over all the competitors.

Example 3: We use this example to demonstrate the sample performance of our proposed kernel smoothing approach to handle partially observed functional predictors. In each simulated scenario, we first generate $\{W_t(\cdot)\}$ and $\{e_t(\cdot)\}$ in the

Table 2.2: The variance explained by each of the components in Example 2. Top d components identified by covariance-based and autocovariance-based approaches are underlined and in bold font, respectively.

	Component	1	2	7	8	9	10	3	4	5	6
	Signal	**1.73**	**1.19**								
d=2	Error	1.00	0.50	1.57	1.49	1.40	1.32	0.25	0.13	0.06	0.03
	Sum	<u>2.73</u>	<u>1.69</u>	1.57	1.49	1.40	1.32	0.25	0.13	0.06	0.03
	Component	1	2	7	8	9	10	3	4	5	6
	Signal	**2.50**	**1.73**					**1.38**	**1.19**		
d=4	Error	1.00	0.50	1.73	1.64	1.55	1.45	0.25	0.13	0.06	0.03
	Sum	<u>3.50</u>	<u>2.23</u>	<u>1.73</u>	<u>1.64</u>	1.55	1.45	1.63	1.32	0.06	0.03
	Component	1	2	3	7	8	9	10	4	5	6
	Signal	**3.00**	**2.16**	**1.73**					**1.47**	**1.30**	**1.19**
d=6	Error	1.00	0.50	0.25	1.90	1.80	1.70	1.60	0.13	0.06	0.03
	Sum	<u>4.00</u>	<u>2.66</u>	<u>1.98</u>	<u>1.90</u>	<u>1.80</u>	<u>1.70</u>	1.60	1.60	1.37	1.22

Table 2.3: The variance explained by each of the components in Example 2 with $\{\nu_{tj}\}$ being $N(0,1)$ variables. Top d components identified by covaraicne-based and autocovariance-based approaches are underlined and in bold font, respectively.

	Component	1	2	3	4	5	6	7	8	9	10
	Signal	**1.73**	**1.19**								
d=2	Error	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Sum	<u>2.73</u>	<u>2.19</u>	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Component	1	2	3	4	5	6	7	8	9	10
	Signal	**2.50**	**1.73**	**1.38**	**1.19**						
d=4	Error	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Sum	<u>3.50</u>	<u>2.73</u>	<u>2.38</u>	<u>2.19</u>	1.00	1.00	1.00	1.00	1.00	1.00
	Component	1	2	3	4	5	6	7	8	9	10
	Signal	**3.00**	**2.16**	**1.73**	**1.47**	**1.30**	**1.19**				
d=6	Error	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Sum	<u>4.00</u>	<u>3.16</u>	<u>2.73</u>	<u>2.47</u>	<u>2.30</u>	<u>2.19</u>	1.00	1.00	1.00	1.00

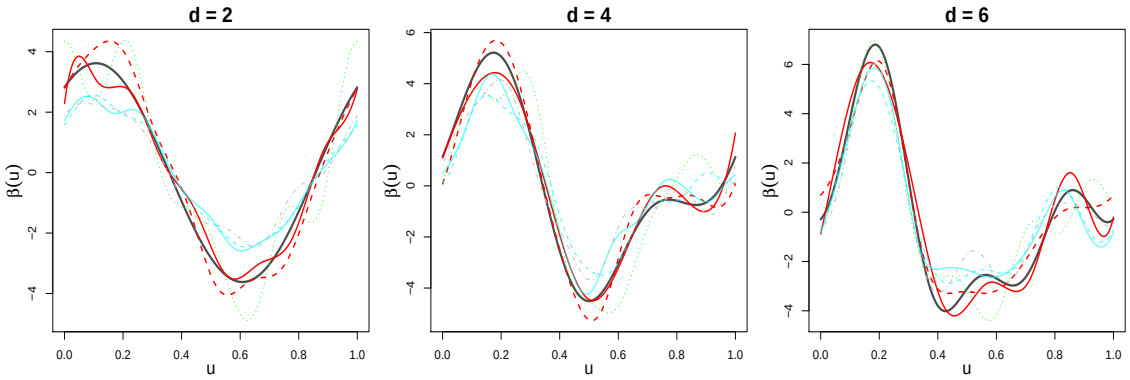


Figure 2.2: Example 2 with $n = 800$ and $d = 2, 4, 6$: Comparison of true $\beta(\cdot)$ functions (black solid) with median estimates over 100 simulation runs for AGMM (red solid), Base AGMM (red dashed), CLS (cyan solid), Base CLS (cyan dashed), Base CGMM (green dotted) and Base ALS (gray dash-dotted).

same way as Example 2 and then generate the observed values Z_{ti} from equation (2.26), where time points U_{ti} and errors η_{ti} are randomly sampled from Uniform$[0, 1]$ and $N(0, 0.5^2)$, respectively. We consider simulation settings $d = 2, 4, 6$, $n = 400, 800, 1200$ and $m_t = 10, 25, 50, 100$, changing from sparse to moderately dense to very dense measurement schedules. In each case, the optimal bandwidth parameters, h_C, h_S, are selected by the 10-fold cross-validation as explained in Rubín and Panaretos (2020) and $\widehat{d}$ is chosen so that the first $\widehat{d}$ eigenvalues explains over 95% of the total variation. Table 2.5 reports numerical summaries for all 36 cases. Several conclusions can be drawn. First, for each d, the estimation accuracy is improved as n and m_t increase. Second, as curves are very densely observed, e.g. $m_t = 100$, our proposed smoothing approach enjoys similar performance with AGMM in Table 2.4, providing empirical evidence to support our remark for Theorem 2.4 about the same convergence rate between very densely observed and fully observed functional scenarios.

Example 4: This example is used to demonstrate the superiority of AGMM methods under the setting where the dimension of the $\beta_0(\cdot)$ is less than the dimension of $X_t(\cdot)$. While the data are generated in the same fashion to Example 2, the slope functions are generated by $\beta_0(\cdot) = \sum_{j=1}^{d} \widetilde{b}_j \phi_j(\cdot)$ with $\widetilde{b}_j = b_j$ for $j = 1, \ldots, d-1$ and $\widetilde{b}_d = 0$ so that the dimension of $\beta_0(\cdot)$ is $d - 1 < d$. Table 2.6 provides numerical results under the oracle scenario with true d in the estimation. We obtain the same findings to those in Table 2.4, i.e. two versions of AGMM significantly outperform their competing methods, while AGMM is superior to Base AGMM in most of the cases.

Example 5: This example is used to illustrate the advantages of AGMM methods under the infinite dimensional setting. With a large enough d, e.g. $d = 25$, the data is generated as follows so that Conditions 2.3 and 2.4 are satisfied.

Table 2.4: *Example 2*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs. The lowest values are in bold font.

$\widehat{d}$	n	d	Base CLS	CLS	Base CGMM	Base ALS	Base AGMM	AGMM
		2	1.591(0.059)	0.990(0.046)	1.118(0.078)	1.165(0.030)	0.599(0.038)	**0.262(0.026)**
	400	4	2.026(0.066)	1.590(0.070)	2.310(0.112)	0.972(0.033)	0.686(0.041)	**0.448(0.034)**
		6	2.310(0.069)	1.932(0.077)	2.722(0.104)	0.938(0.035)	0.825(0.042)	**0.676(0.048)**
		2	1.377(0.051)	0.940(0.038)	0.884(0.085)	0.994(0.019)	0.337(0.020)	**0.138(0.010)**
True	800	4	1.934(0.051)	1.526(0.054)	2.268(0.105)	0.685(0.016)	0.318(0.016)	**0.208(0.013)**
		6	2.160(0.056)	1.872(0.055)	2.859(0.138)	0.575(0.015)	**0.339(0.017)**	0.364(0.020)
		2	1.294(0.053)	0.980(0.048)	0.750(0.081)	0.900(0.013)	0.203(0.011)	**0.080(0.005)**
	1200	4	1.959(0.053)	1.524(0.058)	2.426(0.121)	0.582(0.009)	0.167(0.008)	**0.124(0.006)**
		6	2.270(0.048)	2.002(0.050)	3.092(0.113)	0.494(0.011)	**0.217(0.010)**	0.248(0.010)
		2	0.817(0.012)	0.818(0.012)	0.980(0.059)	1.141(0.026)	0.575(0.030)	**0.248(0.018)**
	400	4	1.037(0.043)	0.725(0.036)	1.319(0.070)	1.097(0.038)	0.773(0.042)	**0.584(0.038)**
		6	0.913(0.041)	**0.811(0.038)**	1.305(0.068)	1.164(0.050)	0.999(0.051)	0.955(0.053)
		2	0.795(0.010)	0.795(0.010)	0.899(0.055)	0.989(0.019)	0.333(0.020)	**0.138(0.009)**
Est	800	4	1.093(0.033)	0.768(0.035)	1.471(0.065)	0.682(0.016)	0.319(0.016)	**0.212(0.013)**
		6	0.859(0.041)	0.809(0.039)	1.139(0.061)	0.571(0.016)	**0.335(0.017)**	0.369(0.020)
		2	0.779(0.007)	0.780(0.007)	0.747(0.044)	0.898(0.012)	0.205(0.012)	**0.079(0.005)**
	1200	4	1.055(0.026)	0.815(0.032)	1.344(0.052)	0.580(0.009)	0.166(0.008)	**0.130(0.007)**
		6	0.813(0.029)	0.808(0.029)	1.159(0.058)	0.492(0.011)	**0.216(0.011)**	0.243(0.009)

To be specific, we generate $X_t(u) = \sum_{j=1}^{d} \xi_{tj}\phi_j(u)$ based on $\xi_{tj} = 0.8\xi_{t-1,j} + \epsilon_{tj}$, where $\epsilon_{tj} \sim N(0, j^{-0.75})$. Some specific calculations yield lag-k autocovariance of ξ_{tj} as $\mathrm{Cov}(\xi_{tj}, \xi_{t+k,j}) = \frac{0.8^k \cdot j^{-0.75}}{0.36}$ and eigenvalues of K in equation (2.9) as $\theta_j = \sum_{k=1}^{L} \mathrm{Cov}(\xi_{tj}, \xi_{t+k,j})^2 = \frac{\sum_{k=1}^{L} 0.8^{2k}}{0.36^2} \cdot j^{-1.5} \asymp j^{-1.5}$ under the orthonormality of $\{\phi_j(\cdot)\}_{j\geq 1}$. Hence, Condition 2.3 is satisfied with $\alpha = 1.5$. Moreover, we set $\tau = 2$ in Condition 2.4 so that $\tau \geq \alpha + 1/2$ is satisfied and hence generate the slope function $\beta_0(\cdot) = \sum_{j=1}^{d} \widetilde{b}_j\phi_j(\cdot)$ with $\widetilde{b}_j = (-1)^{j-1} \cdot 2 \cdot j^{-2}$. The innovations $\{\nu_{tj}\}_{n\times 10}$ are independent $N(0,1)$ variables. The truncated dimension M is chosen so that the top M eigenvalues explains over 90% of the total variation. Table 2.7 reports numerical results for all comparison methods under two settings, where $\{\phi_j(\cdot)\}_{j=1}^{d}$ and $\{\zeta_j(\cdot)\}_{j=1}^{10}$ are generated from the corresponding basis functions used in Example 1 and 2, respectively. Again we observe the prominent superiority of two versions of AGMM methods over the competitors with AGMM significantly outperforming Base AGMM.

2.5.2 Real data analysis

In this section, we illustrate the proposed AGMM using a public financial dataset. The dataset was downloaded from Wharton Research Data Services and consists of

Table 2.5: *Example 3*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs.

n	d	$m_t = 10$	$m_t = 25$	$m_t = 50$	$m_t = 100$
	2	0.906(0.052)	0.374(0.019)	0.296(0.015)	0.227(0.011)
400	4	1.238(0.046)	0.637(0.027)	0.593(0.045)	0.395(0.020)
	6	1.168(0.051)	1.092(0.031)	0.906(0.028)	0.721(0.027)
	2	0.571(0.030)	0.194(0.009)	0.155(0.008)	0.142(0.007)
800	4	0.804(0.030)	0.375(0.015)	0.329(0.023)	0.231(0.010)
	6	1.130(0.039)	0.835(0.029)	0.481(0.019)	0.360(0.013)
	2	0.317(0.017)	0.145(0.007)	0.124(0.006)	0.107(0.005)
1200	4	0.632(0.025)	0.226(0.008)	0.214(0.013)	0.150(0.007)
	6	1.043(0.031)	0.505(0.016)	0.311(0.010)	0.269(0.009)

Table 2.6: *Example 4*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs. The lowest values are in bold font.

n	d	Base CLS	CLS	Base CGMM	Base ALS	Base AGMM	AGMM
	2	0.683(0.008)	0.577(0.007)	0.244(0.014)	0.646(0.008)	0.255(0.014)	**0.132(0.008)**
400	4	1.415(0.042)	0.993(0.043)	1.619(0.072)	0.756(0.014)	0.489(0.024)	**0.324(0.018)**
	6	1.990(0.051)	1.600(0.055)	2.312(0.066)	0.775(0.021)	0.647(0.025)	**0.500(0.024)**
	2	0.589(0.006)	0.560(0.006)	0.137(0.008)	0.593(0.006)	0.125(0.005)	**0.076(0.005)**
800	4	1.378(0.038)	0.855(0.037)	1.641(0.069)	0.620(0.008)	0.253(0.010)	**0.191(0.012)**
	6	1.817(0.036)	1.546(0.035)	2.351(0.077)	0.515(0.009)	**0.295(0.011)**	0.304(0.019)
	2	0.573(0.004)	0.552(0.004)	0.081(0.005)	0.576(0.004)	0.082(0.004)	**0.048(0.003)**
1200	4	1.383(0.035)	0.875(0.044)	1.732(0.072)	0.554(0.005)	0.142(0.006)	**0.108(0.006)**
	6	1.895(0.032)	1.623(0.036)	2.598(0.071)	0.462(0.007)	**0.196(0.007)**	0.197(0.013)

one-minute resolution prices of Standard & Poor's 500 index and inclusive stocks from $n = 251$ trading days in year 2017. The trading time (9:30-16:00) is then converted to minutes, $u \in [0, 390]$. Let $P_t(u_j)$ $(t = 1, \ldots, n, j = 1, \ldots, 390)$ be the price of a financial asset at the j-th minute after the opening time on the t-th trading day. Denote the *cumulative intraday return* (CIDR) trajectory, in percentage, by $r_t(u_j) = 100\big[\log\{P_t(u_j)\} - \log\{P_t(u_1)\}\big]$ (Horváth et al., 2014). Let $r_{m,t}(u)$ be the CIDR curves of the Standard & Poor's 500 index.

We extend the standard *capital asset pricing model* (CAPM) [Chapter 5 of Campbell et al. (1997)] to the functional domain by considering the functional linear regression

Table 2.7: *Example 5*: The mean and standard error (in parentheses) of the mean integrated squared error for $\widehat{\beta}(u)$ over 100 simulation runs. The lowest values are in bold font.

$\{\phi_j\}_{j=1}^{25}, \{\zeta_j\}_{j=1}^{10}$	n	Base CLS	CLS	Base CGMM	Base ALS	Base AGMM	AGMM
Example 1	400	0.972(0.017)	1.068(0.022)	0.913(0.030)	0.708(0.012)	0.582(0.013)	**0.390(0.017)**
	800	0.810(0.011)	0.849(0.012)	0.540(0.018)	0.535(0.008)	0.329(0.008)	**0.200(0.008)**
	1200	0.775(0.009)	0.800(0.009)	0.446(0.017)	0.463(0.006)	0.235(0.005)	**0.156(0.005)**
Example 2	400	0.677(0.012)	0.684(0.015)	0.838(0.026)	0.702(0.012)	0.590(0.014)	**0.376(0.017)**
	800	0.536(0.008)	0.541(0.007)	0.449(0.012)	0.546(0.007)	0.341(0.008)	**0.200(0.007)**
	1200	0.482(0.005)	0.486(0.005)	0.308(0.009)	0.478(0.004)	0.241(0.005)	**0.153(0.005)**

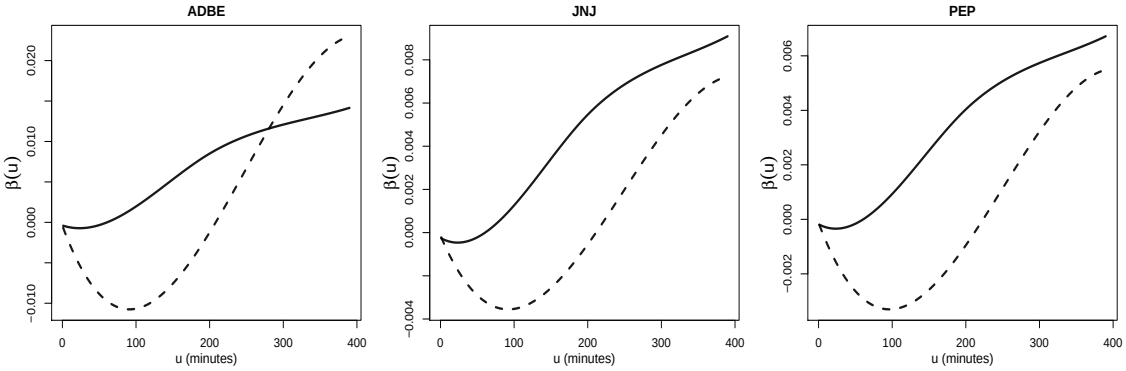


Figure 2.3: Estimated $\beta(\cdot)$ curves for AGMM (solid) and CLS (dashed).

with functional errors-in-predictors as follows

$$y_t = \alpha + \int x_t(u)\beta(u)du + \varepsilon_t, \quad r_{m,t}(u) = x_t(u) + e_t(u), \ t = 1, \ldots, n, \ u \in [0, 390], \tag{2.29}$$

where $x_t(\cdot)$ and $e_t(\cdot)$ represent the signal and error components in $r_{m,t}(\cdot)$, respectively, and y_t is the intraday return of a specific stock on the t-th trading day. Note that the slope parameter in the classical CAPM explains how strongly an asset return depends on the market portfolio. Analogously, $\beta(\cdot)$ in functional CAPM in (2.29) can be understood as the functional sensitivity measure of an asset return to the market CIDR trajectory.

Figure 2.3 plots the estimated $\beta(\cdot)$ functions using both AGMM and CLS for three large-cap-sector stocks, Adobe (ADBE), Johnson & Johnson (JNJ) and PepsiCo (PEP). A few trends are apparent. First, the AGMM estimates place more positive weights as u increases. This result seems reasonable given the fact that the daily most recent market price would contain the most information about the stock's closing price. Second, the CLS estimates first dip in the mid-morning and then start to increase until the end of the trading day. In general, the shapes of the estimated $\beta(\cdot)$ functions by either AGMM or CLS are quite similar across the three

stocks.

To formulate a prediction problem, we treat CIDR trajectories of the same stock as that in (2.29) up to current time $T < 390$ as $r_{y,t}(u), u \in [0, T]$, where, e.g., $T = 375$ corresponds to 15 minutes prior to the closing time of the trading day. Then we construct the same functional linear model as (2.29) by replacing $r_{m,t}(\cdot)$ with $r_{y,t}(\cdot)$. To judge which method produces superior predictions, we implement a rolling procedure to calculate the mean squared prediction error (MSPE) for $H = 30$ days. Specifically, for each $h = H, H-1, \ldots, 1$, we treat $\{y_{n-h+1}, r_{y,n-h+1}\}$ as a testing set, implementing each fitting method on the training set of $\{(y_t, r_{y,t}) : t = 1, \ldots, n-h\}$, calculate the squared error between y_{n-h+1} and its predicted value, and repeat this procedure H-times to compute the MSPE. We calculate the MSPEs over a grid of (d, J) values and choose the pair with the lowest error. We also include the prediction errors from the null model, using the mean of the training response to predict the test response. The resulting MSPEs, for various values of T and the same three stocks, are provided in Table 2.8. It is easy to observe that the prediction accuracy for AGMM and CLS improves as T approaches to 390 and AGMM significantly outperforms two competitors in almost all settings.

Table 2.8: Mean squared prediction errors up to different current times, $T = 330$, 345, 360, 375, 380 and 385 minutes, for AGMM and two competing methods. All entries have been multiplied by 10 for formatting reasons. The lowest MSPE for each value of T is bolded.

Stock	Method	$u \leq 330$	$u \leq 345$	$u \leq 360$	$u \leq 375$	$u \leq 380$	$u \leq 385$
ADBE	AGMM	1.276	**1.179**	**0.983**	**0.852**	**0.800**	**0.728**
	CLS	**1.272**	1.186	1.094	0.991	0.949	0.895
	Mean	12.224	12.224	12.224	12.224	12.224	12.224
JNJ	AGMM	**0.419**	**0.305**	**0.279**	**0.254**	**0.243**	**0.226**
	CLS	0.583	0.496	0.419	0.352	0.330	0.306
	Mean	3.077	3.077	3.077	3.077	3.077	3.077
PEP	AGMM	**0.749**	**0.659**	**0.557**	**0.466**	**0.429**	**0.384**
	CLS	0.781	0.687	0.596	0.502	0.468	0.429
	Mean	2.956	2.956	2.956	2.956	2.956	2.956

2.6 Appendix

Appendix 2.6.1 contains the basis expansion approach to address partially observed curve time series. The proofs of all theorems and technical lemmas are in the Appendix 2.6.2.

2.6.1 Basis expansion approach

We develop a standard basis expansion approach to estimate $K(u,v)$ and $R(u)$. Let $\mathbf{B}(u)$ be the J-dimensional orthonormal basis function, i.e. $\int_{\mathcal{U}} \mathbf{B}(u)\mathbf{B}^{\mathrm{T}}(u)du = \mathbf{I}_J$, such that each $C_k(u,v)$ can be well approximated by $\{\mathbf{B}(u)\}^T\mathbf{\Sigma}_k\mathbf{B}(v)$. In practice, J can be selected by a similar cross-validation procedure described in Section 2.2.5. Let $\mathbf{B}_{ti} = \mathbf{B}(U_{ti})$. We consider minimizing

$$\sum_{t=1}^{n-L}\sum_{i=1}^{m_t}\sum_{j=1}^{m_{t+k}}\left\{Z_{ti}Z_{(t+k)j} - \mathbf{B}_{ti}^T\mathbf{\Sigma}_k\mathbf{B}_{(t+k)j}\right\}^2 \tag{2.30}$$

with respect to $\mathbf{\Sigma}_k \in \mathbb{R}^{J\times J}$. Standard calculation shows that the estimate of $\mathbf{\Sigma}_k$ that minimizes (2.30) is

$$\mathrm{vec}(\widehat{\mathbf{\Sigma}}_k) = \left(\sum_{t,i,j}(\mathbf{B}_{(t+k)j}\otimes\mathbf{B}_i)(\mathbf{B}_{(t+k)j}\otimes\mathbf{B}_i)^T\right)^{-1}\sum_{t,i,j}(\mathbf{B}_{(t+k)j}\otimes\mathbf{B}_i)Z_{ti}Z_{(t+k)j},$$

where $\mathrm{vec}(\mathbf{B})$ denotes the vectorization of the matrix $\mathbf{B}$ formed by stacking its columns into a single column vector and $\otimes$ is the Kronecker product. Then the estimate of $K(u,v)$ is

$$\widetilde{K}(u,v) = \{\mathbf{B}(u)\}^{\mathrm{T}}\sum_{k=1}^{L}\widehat{\mathbf{\Sigma}}_k\widehat{\mathbf{\Sigma}}_k^{\mathrm{T}}\mathbf{B}(v).$$

Similarly, we can obtain a consistent estimator $\widehat{\mathrm{Cov}}\{Y_t, W_{t+k}(u)\} = \widehat{\boldsymbol{\delta}}_k^T\mathbf{B}(u)$, where $\widehat{\boldsymbol{\delta}}_k$ is obtained by minimizing

$$\sum_{t=1}^{n-L}\sum_{1\le i\le m_t}\left\{Y_tZ_{(t+k)i} - \boldsymbol{\delta}_k^T\mathbf{B}_{(t+k)i}\right\}^2$$

with repsect to $\boldsymbol{\delta}_k \in \mathbb{R}^J$. Then the estimate of $\boldsymbol{\delta}_k$ is

$$\widehat{\boldsymbol{\delta}}_k = \left(\sum_{t,i}\mathbf{B}_{(t+k)i}\mathbf{B}_{(t+k)i}^T\right)^{-1}\sum_{t,i}\mathbf{B}_{(t+k)i}Y_tZ_{(t+k)i}.$$

As a result, $R(u)$ can be estimated by

$$\widetilde{R}(u) = \sum_{k=1}^{L} \widehat{\boldsymbol{\delta}}_k^T \widehat{\boldsymbol{\Sigma}}_k^T \mathbf{B}(u).$$

2.6.2 Proofs

Proof of Theorem 2.1 (i)

Define $\breve{K}(u,v) = \sum_{j=1}^{d} \widehat{\theta}_j \widehat{\psi}_j(u)\widehat{\psi}_j(v)$ and $K^{-1}(u,v) = \sum_{j=1}^{d} \theta_j^{-1}\psi_j(u)\psi_j(v)$. Let $\breve{\beta}(u) = \int_{\mathcal{U}} \breve{K}^{-1}(u,v)\widehat{R}(v)dv$. For a large $\delta > 0$, by Lemma 2.4, we have

$$\begin{aligned} P\big(n^{1/2}\|\widehat{\beta} - \beta_0\| > \delta\big) &= P\big(n^{1/2}\|\widehat{\beta} - \beta_0\| > \delta, \widehat{d} = d\big) + P\big(n^{1/2}\|\widehat{\beta} - \beta_0\| > \delta, \widehat{d} \neq d\big) \\ &\leq P\big(n^{1/2}\|\breve{\beta} - \beta_0\| > \delta, \widehat{d} = d\big) + P\big(\widehat{d} \neq d\big) \\ &\leq P\big(n^{1/2}\|\breve{\beta} - \beta_0\| > \delta\big) + o(1), \end{aligned}$$

which means that, to prove $n^{1/2}\|\widehat{\beta} - \beta_0\| = O_P(1)$, it suffices to show that $\|\breve{\beta} - \beta_0\| = O_P(n^{-1/2})$. It is easy to show that

$$\|\breve{\beta} - \beta_0\| \leq \|\breve{K}^{-1} - K^{-1}\|_{\mathcal{S}}\|\widehat{R}\| + \|K^{-1}\|_{\mathcal{S}}\|\widehat{R} - R\|. \tag{2.31}$$

Then it follows from Lemmas 2.2, 2.3 and 2.5 that $\|\breve{\beta} - \beta_0\| = O_P(n^{-1/2})$.

Proof of Theorem 2.1 (ii)

Without any ambiguity, write $\langle q, K\rangle$, $\langle K, q\rangle$ and $\langle p, \langle K, q\rangle\rangle$ for

$$\int_{\mathcal{U}} K(u,v)q(u)du, \int_{\mathcal{U}} K(u,v)q(v)dv \quad \textit{and} \quad \int_{\mathcal{U}}\int_{\mathcal{U}} K(u,v)p(u)q(v)dudv,$$

respectively. In Lemma 2.6, we give expressions for $\widehat{\theta}_j - \theta_j$ and $\widehat{\psi}_j - \psi_j$ for $j \geq 1$. Let $\beta_M(u) = \sum_{j=1}^{M} \theta_j^{-1}\langle \psi_j, R\rangle \psi_j(u)$. By the triangle inequality, we have

$$\|\widehat{\beta} - \beta_0\|^2 \leq \|\widehat{\beta} - \beta_M\|^2 + \|\beta_M - \beta_0\|^2. \tag{2.32}$$

By (2.12) and orthonormality of $\{\psi_j(\cdot)\}$, we have $\|\beta_M - \beta_0\|^2 = \sum_{j=M+1}^{\infty} \theta_j^{-2}\langle \psi_j, R\rangle^2$.

It follows from Condition 2.4 and some specific calculations that

$$\|\beta_M - \beta_0\|^2 = \sum_{j=M+1}^{\infty} b_j^2 \le C \sum_{j=M+1}^{\infty} j^{-2\tau} = O(M^{-2\tau+1}). \tag{2.33}$$

Next we will show the convergence rate of $\|\widehat{\beta} - \beta_M\|^2$. Observe that

$$\begin{aligned}\widehat{\beta}(u) - \beta_M(u) &= \sum_{j=1}^{M} \big(\widehat{\theta}_j^{-1} - \theta_j^{-1}\big)\langle \psi_j, R\rangle \widehat{\psi}_j(u) + \sum_{j=1}^{M} \widehat{\theta}_j^{-1}\big(\langle \widehat{\psi}_j, \widehat{R}\rangle - \langle \psi_j, R\rangle\big)\widehat{\psi}_j(u) \\ &\quad + \sum_{j=1}^{M} \theta_j^{-1}\langle \psi_j, R\rangle\big\{\widehat{\psi}_j(u) - \psi_j(u)\big\}.\end{aligned}$$

Then we have

$$\begin{aligned}\|\widehat{\beta} - \beta_M\|^2 &\le 3\sum_{j=1}^{M} \big(\widehat{\theta}_j^{-1} - \theta_j^{-1}\big)^2\langle \psi_j, R\rangle^2 + 3\sum_{j=1}^{M} \widehat{\theta}_j^{-2}\big(\langle \widehat{\psi}_j, \widehat{R}\rangle - \langle \psi_j, R\rangle\big)^2 \\ &\quad + 3M\sum_{j=1}^{M} \theta_j^{-2}\langle \psi_j, R\rangle^2 \big\|\widehat{\psi}_j - \psi_j\big\|^2 \\ &= 3I_{n1} + 3I_{n2} + 3I_{n3}.\end{aligned} \tag{2.34}$$

Let $\widehat{\Delta} = \|\widehat{K} - K\|_{\mathcal{S}}$ and $\Omega_M = \{2\widehat{\Delta} \le \delta_M\}$. On the event Ω_M, we can see that $\sup_{j\le M} |\widehat{\theta}_j - \theta_j| \le \theta_M/2$, which implies that $2^{-1}\theta_j \le \widehat{\theta}_j \le 2\theta_j$. Moreover, we can show that $P(\Omega_M) \to 1$ since $n^{1/2}\delta_M \to \infty$ as $n \to \infty$. Hence it suffices to work with bounds that are established under the event Ω_M.

Provided that event Ω_M holds, it follows from $\sup_{j\ge 1} |\widehat{\theta}_j - \theta_j| = O_P(n^{-1/2})$ in Lemma 2.1(i) and some calculations that

$$I_{n1} \le 4\sum_{j=1}^{M} \big(\widehat{\theta}_j - \theta_j\big)^2\theta_j^{-4}\langle \psi_j, R\rangle^2 = 4\sum_{j=1}^{M} \theta_j^{-2} b_j^2\big(\widehat{\theta}_j - \theta_j\big)^2 = O_P\Big(n^{-1}\sum_{j=1}^{M} \theta_j^{-2} b_j^2\Big).$$

By Conditions 2.3–2.4, we have

$$\begin{aligned}I_{n1} &= O_P(n^{-1}) \cdot \Big(\sum_{j=1}^{M} j^{2\alpha-2\tau}\Big) \\ &= O_P(n^{-1}) \cdot \big(M + M^{2\alpha-2\tau+1}\big) \\ &= o_P\big(n^{-1}M^{2\alpha+1}\big).\end{aligned}$$

Consider the term I_{n3}. By $\|\widehat{\psi}_j - \psi_j\| = O_P\big(j^{1+\alpha} n^{-1/2}\big)$ in Lemma 2.1(iii) and Condition 2.4, we obtain that

$$I_{n3} \le M \sum_{j=1}^{M} b_j^2 \big\|\widehat{\psi}_j - \psi_j\big\|^2 = O_P\big(n^{-1} M^{2-2\tau+2\alpha+2}\big) = O_P\Big(n^{-1} M^{2\alpha+1}\Big),$$

where the last equality comes from $\alpha > 1$ and $2\alpha - 2\tau + 4 \le 2\alpha + 1$ implied by Condition 2.4.

Consider the term I_{n2}. On the event Ω_M, we have that

$$\begin{aligned} I_{n2} &\le 4 \sum_{j=1}^{M} \theta_j^{-2} \big(\langle \widehat{\psi}_j, \widehat{R}\rangle - \langle \psi_j, R\rangle\big)^2 \\ &\le 12 \sum_{j=1}^{M} \theta_j^{-2} \Big(\langle \widehat{\psi}_j - \psi_j, R\rangle^2 + \langle \psi_j, \widehat{R} - R\rangle^2 + \langle \widehat{\psi}_j - \psi_j, \widehat{R} - R\rangle^2\Big) \\ &\le 12 \sum_{j=1}^{M} \theta_j^{-2} \Big(\langle \widehat{\psi}_j - \psi_j, R\rangle^2 + \|\widehat{R} - R\|^2 + \|\widehat{\psi}_j - \psi_j\|^2 \|\widehat{R} - R\|^2\Big), \end{aligned} \tag{2.35}$$

where the last inequality comes from orthonormality of $\{\psi_j(\cdot)\}$ and Cauchy-Schwarz inequality. By Lemma 2.6 and some calculations, we can represent the term $\langle \widehat{\psi}_j - \psi_j, R\rangle$ as

$$\langle \widehat{\psi}_j - \psi_j, R\rangle = R_{j1} + R_{j2},$$

where $R_{j1} = \sum_{k:k\neq j} \theta_k b_k (\widehat{\theta}_j - \theta_k)^{-1} \langle \widehat{\psi}_j, \langle \widehat{K} - K, \psi_k\rangle\rangle$ and $R_{j2} = \theta_j b_j \langle \widehat{\psi}_j - \psi_j, \psi_j\rangle$. It follows from Condition 2.3–2.4, Lemma 2.1 and Cauchy-Schwarz inequality that

$$\sum_{j=1}^{M} \theta_j^{-2} R_{j2}^2 = O_P(n^{-1}) \cdot \Big(\sum_{j=1}^{M} j^{-2\tau+2\alpha+2}\Big) = o_P\big(n^{-1} M^{2\alpha+1}\big). \tag{2.36}$$

Note that on the event Ω_M, $|\widehat{\theta}_j - \theta_j| \le 2^{-1}|\theta_j - \theta_k|$ for $j = 1, \ldots, k-1, k+1, \ldots, M$ and hence $|\widehat{\theta}_j - \theta_k| \ge 2^{-1}|\theta_j - \theta_k|$. If we can show that

$$\sup_{j\ge 1} (\theta_j^2 j^{2\alpha})^{-1} \sum_{k:k\neq j} \theta_k^2 b_k^2 (\theta_j - \theta_k)^{-2} = O(1), \tag{2.37}$$

then, by Condition 2.4, Lemma 2.1 and on the event Ω_M, we have

$$\begin{aligned}\sum_{j=1}^{M}\theta_j^{-2}R_{j1}^2 &\leq 4\sum_{j=1}^{M}\theta_j^{-2}\sum_{k:k\neq j}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2}\|\widehat{K}-K\|_{\mathcal{S}}^2 \\ &= O_P(n^{-1})\cdot\sum_{j=1}^{M}\theta_j^{-2}\theta_j^2 j^{2\alpha} = O_P(n^{-1}M^{2\alpha+1}). \qquad (2.38)\end{aligned}$$

We turn to prove (2.37) as follows. Denote $[j/2]$ by the largest integer less than $j/2$. Then

$$\sum_{k:k\neq j}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2} = \left(\sum_{k=2(j+1)}^{\infty}+\sum_{k=[j/2]+1,k\neq j}^{k=2j+1}+\sum_{k=1}^{[j/2]}\right)\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2}.$$

Observe that for $k\geq 2(j+1)$,

$$\theta_j-\theta_k=\sum_{s=j}^{k-1}(\theta_s-\theta_{s+1})\geq c\int_{j+1}^{2(j+1)}s^{-\alpha-1}ds=-\frac{c}{\alpha}s^{-\alpha}\Big|_{j+1}^{2(j+1)}\geq\frac{c}{2\alpha}2^{-\alpha}j^{-\alpha},$$

and for $[j/2]+2\leq k\leq 2j+1$ but $k\neq j$,

$$|\theta_j-\theta_k|\geq\max(\theta_j-\theta_{j+1},\theta_{j-1}-\theta_j)\geq cj^{-\alpha-1}.$$

Therefore,

$$\begin{aligned}(\theta_j^2 j^{2\alpha})^{-1}\sum_{k=2(j+1)}^{\infty}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2} &= O(1)\cdot j^{2\alpha-2\tau}\sum_{k=2(j+1)}^{\infty}\theta_k^2 = O(1),\\ (\theta_j^2 j^{2\alpha})^{-1}\sum_{k=[j/2]+1}^{2j+1}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2} &\leq (\theta_j^2 j^{2\alpha})^{-1}\sum_{k=[j/2]+1}^{2j+1}2\{\theta_j^2+(\theta_j-\theta_k)^2\}b_k^2(\theta_j-\theta_k)^{-2}\\ &= O(1)\cdot\theta_j^{-2}j^{-2\alpha}(1+\theta_j^2 j^{2\alpha+3-2\tau})=O(1),\\ (\theta_j^2 j^{2\alpha})^{-1}\sum_{k=1}^{[j/2]}\theta_k^2 b_k^2(\theta_j-\theta_k)^{-2} &\leq O(1)\sum_{k=1}^{[j/2]}\theta_k^2 b_k^2(\theta_k-\theta_{2k})^{-2}=O(1)\cdot\theta_1^2 j^{2\alpha-2\tau+1}\\ &= O(1),\end{aligned}$$

uniformly in j. Then (2.37) follows.

Moreover, it follows from Condition 2.3, Lemmas 2.1–2.3 that

$$\sum_{j=1}^{M} \theta_j^{-2} \|\widehat{R} - R\|^2 = O_P(n^{-1}M^{2\alpha+1})$$

$$\text{and} \tag{2.39}$$

$$\sum_{j=1}^{M} \theta_j^{-2} \|\widehat{\psi}_j - \psi_j\|^2 \|\widehat{R} - R\|^2 = O_P(n^{-2}M^{4\alpha+3}).$$

Combing the results in (2.35)–(2.36) and (2.38)–(2.39), we have

$$I_{n2} = O_P\Big(n^{-2}M^{4\alpha+3} + n^{-1}M^{2\alpha+1}\Big). \tag{2.40}$$

Combining the results in (2.32),(2.33) and (2.40) and choosing $M \asymp n^{1/(2\alpha+2\tau)}$, we obtain that

$$\|\widehat{\beta} - \beta_0\|^2 = O_P\big(n^{-2}M^{4\alpha+3} + n^{-1}M^{2\alpha+1} + M^{-2\tau+1}\big) = O_P\big(n^{-\frac{2\tau-1}{2\alpha+2\tau}}\big).$$

Proof of Theorem 2.2

Following the similar arguments used in the proofs for Lemmas 2.2 and 2.3 under some regularity conditions, we can show that

$$\|\widehat{H} - H\|_{\mathcal{S}} = O_P(n^{-1/2}) \text{ and } \|H\|_{\mathcal{S}} = O(1). \tag{2.41}$$

Consider the case when d is fixed. Let $\breve{\gamma}(u,v) = \int_{\mathcal{U}} \breve{K}^{-1}(u,w)\widehat{H}(w,v)dw$. Then we have

$$\|\breve{\gamma} - \gamma_0\|_{\mathcal{S}} \le \|\breve{K}^{-1} - K^{-1}\|_{\mathcal{S}} \|\widehat{H}\|_{\mathcal{S}} + \|K^{-1}\|_{\mathcal{S}} \|\widehat{H} - H\|_{\mathcal{S}}. \tag{2.42}$$

It follows from Lemma 2.5 and (2.41) that $\|\breve{\gamma} - \gamma\|_{\mathcal{S}} = O_P(n^{-1/2} + n^{-1/2}) = O_P(n^{-1/2})$. Finally, applying the similar technique used in the proof for part (i) of Theorem 2.1, we can prove the result in part (i) of Theorem 2.2.

When $d = \infty$, let $\gamma_M(u,v) = \sum_{j=1}^{M} \theta_j^{-1}\psi_j(u)\langle \psi_j, H(\cdot,v)\rangle$. By the triangle inequality, we have

$$\|\widehat{\gamma} - \gamma_0\|_{\mathcal{S}}^2 \le \|\widehat{\gamma} - \gamma_M\|_{\mathcal{S}}^2 + \|\gamma_M - \gamma_0\|_{\mathcal{S}}^2. \tag{2.43}$$

It follows from Condition 2.6 and some specific calculations that

$$\begin{aligned}\|\gamma_M - \gamma_0\|_{\mathcal{S}}^2 &= O(1)\Big\| \sum_{j=M+1}^{\infty}\sum_{\ell=1}^{\infty} b_{j\ell}\psi_j(u)\psi_\ell(v)\Big\|_{\mathcal{S}}^2 \\ &= O(1)\sum_{j=M+1}^{\infty}\sum_{\ell=1}^{\infty} b_{j\ell}^2 = O(1)\sum_{j=M+1}^{\infty}\sum_{\ell=1}^{\infty}(j+\ell)^{-2\tau-1} \\ &= O(M^{-2\tau+1}).\end{aligned}$$

It remains to show that the convergence rate of $\|\widehat{\gamma} - \gamma_M\|_{\mathcal{S}}^2$. Observe that

$$\begin{aligned}\widehat{\gamma}(u,v) - \gamma_M(u,v) &= \sum_{j=1}^{M}\big(\widehat{\theta}_j^{-1} - \theta_j^{-1}\big)\langle\psi_j, H\rangle(v)\widehat{\psi}_j(u) \\ &\quad + \sum_{j=1}^{M}\widehat{\theta}_j^{-1}\big(\langle\widehat{\psi}_j, \widehat{H}\rangle(v) - \langle\psi_j, H\rangle\big)(v)\widehat{\psi}_j(u) \\ &\quad + \sum_{j=1}^{M}\theta_j^{-1}\langle\psi_j, H\rangle(v)\big\{\widehat{\psi}_j(u) - \psi_j(u)\big\}.\end{aligned}$$

Then we have,

$$\begin{aligned}\|\widehat{\gamma} - \gamma_M\|_{\mathcal{S}}^2 \;\le\; & 3\sum_{j=1}^{M}\big(\widehat{\theta}_j^{-1} - \theta_j^{-1}\big)^2\|\langle\psi_j, H\rangle\|^2 + 3\sum_{j=1}^{M}\widehat{\theta}_j^{-2}\big\|\langle\widehat{\psi}_j, \widehat{H}\rangle - \langle\psi_j, H\rangle\big\|^2 \\ & + 3M\sum_{j=1}^{M}\theta_j^{-2}\big\|\langle\psi_j, H\rangle\big\|^2\big\|\widehat{\psi}_j - \psi_j\big\|^2.\end{aligned}$$

Following the similar arguments used in the proof for Theorem 2.1 (ii), we can show that

$$\|\widehat{\gamma} - \gamma_M\|_{\mathcal{S}}^2 = O_P(M^{4\alpha+3}n^{-2} + M^{2\alpha+1}n^{-1}). \tag{2.44}$$

Combing the results in (2.43)–(2.44) and choosing $M \asymp n^{1/(2\alpha+2\tau)}$, we have

$$\|\widehat{\gamma} - \gamma_0\|_{\mathcal{S}}^2 = O_P\big(M^{2\alpha+1}n^{-1} + M^{-2\tau+1}\big) = O_P\big(n^{-\frac{2\tau-1}{2\alpha+2\tau}}\big).$$

which completes our proof for part (ii) of Theorem 2.2.

Proofs of Theorem 2.3

We begin with the L_2 rates of $\widetilde{C}_k$ for $k \geq 1$. We wish to prove them in the same fashion as the proof of Theorem 1 in Hansen (2008). For $p, q = 0, 1, 2$, define

$$\begin{aligned}
\widetilde{Z}^{(1)}_{p,q,i}(u,v) &= \sum_{i=1}^{m_t}\sum_{j=1}^{m_{t+k}} K_{k,i,j,h,t}(u,v)\left(\frac{U_{ti}-u}{h_C}\right)^p\left(\frac{U_{(t+k)j}-v}{h_C}\right)^q, \\
\widetilde{Z}^{(2)}_{p,q,i}(u,v) &= \sum_{i=1}^{m_t}\sum_{j=1}^{m_{t+k}} K_{k,i,j,h,t}(u,v)\left(\frac{U_{ti}-u}{h_C}\right)^p\left(\frac{U_{(t+k)j}-v}{h_C}\right)^q Z_{ti}Z_{(t+k)j}.
\end{aligned}$$

Let $S_{pq} = (n\rho_n^2 h_C^2)^{-1}\sum_{i=1}^n \widetilde{Z}^{(1)}_{p,q,i}$ and $G_{pq} = (n\rho_n^2 h_C^2)^{-1}\sum_{i=1}^n \widetilde{Z}^{(2)}_{p,q,i}$. Then we have

$$\widetilde{C}_k = \frac{(S_{20}S_{02}-S_{11}^2)G_{00} - (S_{10}S_{02}-S_{01}S_{11})G_{10} + (S_{10}S_{11}-S_{01}S_{20})G_{01}}{(S_{20}S_{02}-S_{11}^2)S_{00} - (S_{10}S_{02}-S_{01}S_{11})S_{10} + (S_{10}S_{11}-S_{01}S_{20})S_{01}}$$

so that $\widetilde{C}_k(u,v) - C_k(u,v)$ can be expressed as

$$\begin{aligned}
&= \frac{(S_{20}S_{02}-S_{11}^2)\{G_{00} - C_k(u,v)S_{00} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{10} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{01}\}}{(S_{20}S_{02}-S_{11}^2)S_{00} - (S_{10}S_{02}-S_{01}S_{11})S_{10} + (S_{10}S_{11}-S_{01}S_{20})S_{01}} \\
&- \frac{(S_{20}S_{02}-S_{11}^2)\{G_{10} - C_k(u,v)S_{10} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{20} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{11}\}}{(S_{20}S_{02}-S_{11}^2)S_{00} - (S_{10}S_{02}-S_{01}S_{11})S_{10} + (S_{10}S_{11}-S_{01}S_{20})S_{01}} \\
&+ \frac{(S_{20}S_{02}-S_{11}^2)\{G_{01} - C_k(u,v)S_{01} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{11} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{02}\}}{(S_{20}S_{02}-S_{11}^2)S_{00} - (S_{10}S_{02}-S_{01}S_{11})S_{10} + (S_{10}S_{11}-S_{01}S_{20})S_{01}}.
\end{aligned}$$

Let $\mathbb{U} = \{U_{ti}, i = 1, \ldots, m_t, t = 1, \ldots, n\}$. Suppose we have shown that for $p, q = 0, 1, 2$,

$$\left\|G_{pq} - E\{G_{pq}|\mathbb{U}\}\right\|_{\mathcal{S}} = O_P\left(\frac{1}{\sqrt{n\rho_n^2 h_C^2}} + \frac{1}{\sqrt{n}}\right), \tag{2.45}$$

and

$$\sup_{u,v\in[0,1]}\left|S_{pq}(u,v) - ES_{pq}(u,v)\right| = o_P(1). \tag{2.46}$$

By Taylor expansion, Condition 2.11 and (2.46),

$$\left\|E\{G_{00}|\mathbb{U}\} - C_k(u,v)S_{00} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{10} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{01}\right\|_{\mathcal{S}} = O_P(h_C^2). \tag{2.47}$$

Then combing (2.45) and (2.47) yields that

$$\left\| G_{00} - C_k(u,v)S_{00} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{10} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{01} \right\|_{\mathcal{S}} = O_P\left(\frac{1}{\sqrt{n\rho_n^2 h_C^2}} + \frac{1}{\sqrt{n}} + h_C^2\right). \quad (2.48)$$

Similarly, both $G_{10} - C_k(u,v)S_{10} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{20} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{11}$ and $G_{01} - C_k(u,v)S_{01} - h_C\frac{\partial C_k}{\partial u}(u,v)S_{11} - h_C\frac{\partial C_k}{\partial v}(u,v)S_{02}$ can be proved to have the same rate in (2.48). We can see from (2.46) that each denominator in $\widetilde{C}_k(u,v)$ is positive and bounded away from zero with probability approaching one, and as a consequence, part (i) of Theorem 2.3 follows.

Next, we turn to prove (2.45) and (2.46). For (2.45), it suffices to show that

$$\int\int E\left\{G_{00}(u,v) - E\big\{G_{00}(u,v)\big|\mathbb{U}\big\}\right\}^2 dudv \lesssim \frac{1}{n\rho_n^2 h_C^2} + \frac{1}{n},$$

where $a_n \lesssim b_n$ means $\limsup_{n\to\infty}|a_n/b_n| \le C$ for some positive constant $C > 0$. It is easy to see that

$$E\left\{G_{00} - E\big\{G_{00}\big|\mathbb{U}\big\}\right\}^2 \le \frac{1}{n^2\rho_n^4 h_C^4}\sum_{t=1}^{n} E\left|\widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\big\}\right|^2 + \frac{1}{n\rho_n^4 h_C^4}\sum_{t=1}^{n}\left|\mathrm{Cov}\Big\{\widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\big\}, \widetilde{Z}^{(2)}_{0,0,t+1} - E\big\{\widetilde{Z}^{(2)}_{0,0,t+1}|\mathbb{U}\big\}\Big\}\right|.$$

Let $\mathbb{W} = \{W_t(\cdot), U_{ti}, i = 1,\ldots,m_t, t = 1,\ldots,n\}$. Note that $\widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\big\} = \widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}\big|\mathbb{W}\big\} + E\big\{\widetilde{Z}^{(2)}_{0,0,1}\big|\mathbb{W}\big\} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\big\}$. Given $\mathbb{W}$, the first term $\widetilde{Z}^{(1)}_{0,0,1} - E\big\{\widetilde{Z}^{(1)}_{0,0,1}\big|\mathbb{W}\big\}$ is a U-type statistics and hence some specific calculations yield that $E\Big\{\widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}\big|\mathbb{W}\big\}\Big\}^2 \lesssim \rho_n^2 h_C^2 + \rho_n^3 h_C^3$ and $E\Big\{E\big\{\widetilde{Z}^{(2)}_{0,0,1}\big|\mathbb{W}\big\} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\big\}\Big\}^2 \lesssim \rho_n^4 h_C^4 + \rho_n^2 h_C^2$. As a result,

$$E|\widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\big\}|^2 \lesssim \rho_n^4 h_C^4 + \rho_n^2 h_C^2.$$

In a similar manner together with Marcinkiewicz-Zygmund inequality, we can show that

$$E|\widetilde{Z}^{(2)}_{0,0,1}(u,v) - E\big\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\big\}|^s \lesssim \rho_n^{2s} h_C^{2s} + \rho_n^s h_C^s + \rho_n^{s/2} h_C^2.$$

For each fixed (u,v) and h_C, under the Conditions 2.7–2.12, we see that $(\widetilde{Z}^{(2)}_{0,0,1},$

$\widetilde{Z}^{(2)}_{0,0,i},\ \ldots)$ is strictly stationary with ψ-mixing coefficients $\psi_Z(l)$ satisfying $\psi_Z(l) \lesssim (l-k)^{-\lambda}$ for $l \geq k+1$. For $j^* < j \leq \max(j^*+1, \rho_n^{-2}h_C^{-2})$ with fixed $j^* > k+1$, we have that

$$\Big|\mathrm{Cov}\big\{\widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\big\}, \widetilde{Z}^{(2)}_{0,0,t+1} - E\big\{\widetilde{Z}^{(2)}_{0,0,t+1}|\mathbb{U}\big\}\big\}\Big| \lesssim \rho_n^4 h_C^4.$$

For $j > \max(j^*+1, \rho_n^{-2}h_C^{-2}) + 1$, using Davydov's lemma, we show that

$$\begin{aligned}&\Big|\mathrm{Cov}\big\{\widetilde{Z}^{(2)}_{0,0,1} - E\big\{\widetilde{Z}^{(2)}_{0,0,1}|\mathbb{U}\big\}, \widetilde{Z}^{(2)}_{0,0,t+1} - E\big\{\widetilde{Z}^{(2)}_{0,0,t+1}|\mathbb{U}\big\}\big\}\Big| \\ &\qquad\qquad \lesssim j^{-2+2/s}\big(\rho_n^4 h_C^4 + \rho_n^2 h_C^2 + \rho_n h_C^{4/s}\big).\end{aligned}$$

Therefore, the rate in (2.45) follows from the steps to prove Theorem 1 in Hansen (2008). Similarly, together with Conditions 2.7–2.13, the rates in (2.46) and $\|\widehat{S}_k - S_k\|$ follows from the steps to prove Theorem 2 in Hansen (2008). The proof is complete.

Proofs of Theorem 2.4

By Theorem 2.3 for $k = 1, \ldots, L$, we can easily show that

$$\|\widetilde{K} - K\|_{\mathcal{S}} = O_P\big(\delta_{n1}\big) \text{ and } \|\widetilde{R} - R\| = O_P\big(\delta_{n1} + \delta_{n2}\big). \tag{2.49}$$

Following directly from the proof steps of Theorem 2.1 by replacing $\|\widehat{K} - K\| = O_P(n^{-1/2})$ and $\|\widehat{R} - R\| = O_P(n^{-1/2})$ with the corresponding rates in (2.49), we complete our proof.

Lemma 2.1 and its proof

Lemma 2.1. *Suppose that Conditions 2.1–2.3 hold and* $\langle\widehat{\psi}_j, \psi_j\rangle \geq 0$. *Then as* $n \to \infty$, *the following results hold:*
(i) $\|\widehat{K} - K\|_{\mathcal{S}} = O_P(n^{-1/2})$ *and* $\sup_{j\geq 1}|\widehat{\theta}_j - \theta_j| = O_P(n^{-1/2})$.
(ii) When d *is fixed,* $\|\widehat{\psi}_j - \psi_j\| = O_P(n^{-1/2})$ *for* $j = 1, \ldots, d$.
(iii) When $d = \infty$, $\|\widehat{\psi}_j - \psi_j\| = O_P(j^{1+\alpha}n^{-1/2})$ *for* $j = 1, 2, \ldots$.

Proof. The first result in part (i) can be found in Theorem 1 of Bathia et al. (2010) and hence the proof is omitted. By (4.43) of Bosq (2000), we have $\sup_{j\geq 1}|\widehat{\theta}_j - \theta_j| \leq \|\widehat{K} - K\|_{\mathcal{S}} = O_P(n^{-1/2})$, which completes the proof for the second result in

part (i). To prove parts (ii) and (iii), let $\delta_j = 2\sqrt{2}\max\{(\theta_{j-1}-\theta_j)^{-1}, (\theta_j-\theta_{j+1})^{-1}\}$ if $j \geq 2$ and $\delta_1 = 2\sqrt{2}(\theta_1-\theta_2)^{-1}$. It follows from Lemma 4.3 of Bosq (2000) that $\|\widehat{\psi}_j - \psi_j\| \leq \delta_j \|\widehat{K} - K\|_{\mathcal{S}} = O_P(\delta_j n^{-1/2})$. Under Condition 2.3(i) with a fixed d, root-n rate can be achieved. When $d = \infty$, Condition 2.3(ii) and (iii) imply that $\delta_j \leq Cj^{\alpha+1}$ with some positive constant C. This completes our proof for part (iii).

Lemma 2.2 and its proof

Lemma 2.2. *Suppose that Conditions 2.1-2.2 hold, then* $\|\widehat{R} - R\| = O_P(n^{-1/2})$.

Proof. Provided L is fixed, we may set $n \equiv n - L$. Let $\mathcal{S}$ denotes the space consisting of all the operators with a finite Hilbert-Schmidt norm and $\mathcal{H}$ denotes the space consisting of all the functions with a finite L_2 norm. Let $Z_{tk} = W_t \otimes W_{t+k} \in \mathcal{S}$ and $z_{tk} = Y_t W_{t+k} \in \mathcal{H}$. Now consider the kernel $\rho : \mathcal{S} \times \mathcal{H} \to \mathcal{H}$ given by $\rho(A, x) = Ax^*$ with $A \in \mathcal{S}$ and $x \in \mathcal{H}$. Let $c_k(\cdot) = \mathrm{Cov}\{Y_t, W_{t+k}(\cdot)\}$. We can represent $\widehat{C}_k\widehat{c}_k^* = n^{-2}\sum_{t=1}^n \sum_{t'=1}^n \rho(Z_{tk}, z_{t'k})$, which is simply a $\mathcal{H}$ valued Von Mises' functional (Borovskikh, 1996). For $d \geq 1$, neither of C_k and c_k is zero, it follows from Lemma 3 of Bathia et al. (2010) that $E\|\widehat{C}_k\widehat{c}_k^* - C_k c_k^*\|^2 = O(n^{-1})$. Then by the Chebyshev inequality, we have

$$\|\widehat{R} - R\| \leq \sum_{k=1}^{L} \|\widehat{C}_k\widehat{c}_k^* - C_k c_k^*\| = O_P(n^{-1/2}),$$

which completes the proof.

Lemma 2.3 and its proof

Lemma 2.3. *Suppose that Condition 2.2 holds, then* $\|R\| = O(1)$.

Proof. By the definitions of C_k and (2.8), we have

$$\|R\| \leq \sum_{k=1}^{L} \|C_k\|_{\mathcal{S}} \|\mathrm{Cov}(Y_t, W_{t+k})\| = \sum_{k=1}^{L} \|E\{W_t(u)W_{t+k}(v)\}\|_{\mathcal{S}} \|E(Y_t W_{t+k}(u))\|.$$

It follows from Cauchy-Schwartz inequality, Condition 2.2, Fubini Theorem and Jensen's inequality that $\|E\{W_t(u)W_{t+k}(v)\}\|_{\mathcal{S}}^2$

$$\begin{aligned} &= \int_{\mathcal{U}}\int_{\mathcal{U}}[E\{W_t(u)W_{t+k}(v)\}]^2 dudv \\ &\leq \int_{\mathcal{U}} E\{W_t(u)^2\}du \int_{\mathcal{U}} E\{W_{t+k}(v)^2\}dv = \Big[\int_{\mathcal{U}} E\{W_t(u)^2\}du\Big]^2 \\ &\leq E\Big\{\int_{\mathcal{U}} W_t(u)^2 du\Big\}^2 < \infty. \end{aligned}$$

Similarly, $\|E\{Y_t W_{t+k}(u)\}\|^2 \leq E(Y_t^2)\int_{\mathcal{U}} E\{W_{t+k}(u)^2\}du < \infty$. Combining the results leads to $\|R\| = O(1)$.

Lemma 2.4 and its proof

Lemma 2.4. *Suppose the Conditions 2.1, 2.2, 2.3 (i) and (iii) hold. Let* $\epsilon_n \to 0$, $\epsilon_n^2 n \to \infty$ *and as* $n \to \infty$. *Then when* $d < \infty$, $P\big(\widehat{d} \neq d\big) = O\{(\epsilon_n^2 n)^{-1}\} \to 0$.

Proof. This lemma, which holds for $d < \infty$, can be found in Theorem 3 of Bathia et al. (2010) and hence the proof is omitted.

Lemma 2.5 and its proof

Lemma 2.5. *Suppose that Conditions 2.1, 2.2, 2.3(i) and (iii) hold. Then the following results hold.*
(i) $\|\breve{K}^{-1} - K^{-1}\|_{\mathcal{S}} = O_P(n^{-1/2})$.
(ii) $\|K^{-1}\|_{\mathcal{S}} = O(1)$.

Proof. Observe that

$$\breve{K}^{-1} - K^{-1} = \sum_{j=1}^{d}(\widehat{\theta}_j^{-1} - \theta_j^{-1})\widehat{\psi}_j(u)\widehat{\psi}_j(v) + \sum_{j=1}^{d}\theta_j^{-1}\{\widehat{\psi}_j(u)\widehat{\psi}_j(v) - \psi_j(u)\psi_j(v)\}.$$

Then by the orthonormality of $\{\psi_j(\cdot)\}$ and $\{\widehat{\psi}_j(\cdot)\}$, we have

$$\|\breve{K}^{-1} - K^{-1}\|_{\mathcal{S}} \leq \sum_{j=1}^{d}\widehat{\theta}_j^{-1}\theta_j^{-1}|\widehat{\theta}_j - \theta_j| + 2\sum_{j=1}^{d}\theta_j^{-1}\|\widehat{\psi}_j - \psi_j\|. \tag{2.50}$$

When d is fixed, the smallest eigenvalue θ_d is bounded away from zero. It follows from Lemma 2.1 (i),(ii) and (2.50) that there exists some positive constant C such that $\|\breve{K}^{-1} - K^{-1}\|_{\mathcal{S}} \leq C(\theta_d^{-2} + \theta_d^{-1})n^{-1/2}$, which completes the proof for part (i).

Note that $\|K^{-1}\|_{\mathcal{S}} = \|\sum_{j=1}^{d} \theta_j^{-1}\psi_j(u)\psi_j(v)\|_{\mathcal{S}} = (\sum_{j=1}^{d} \theta_j^{-2})^{1/2} \leq d^{1/2}\theta_d^{-1}$. Then part (ii) follows as d is fixed and θ_d is bounded below from zero.

Lemma 2.6 and its proof

Lemma 2.6. *If* $\inf_{k\neq j} |\widehat{\theta}_j - \theta_k| > 0$, *then*

$$\widehat{\psi}_j - \psi_j = \sum_{k:k\neq j} (\widehat{\theta}_j - \theta_k)^{-1}\psi_k\langle\hat{\psi}_j, \langle\widehat{K} - K, \psi_k\rangle\rangle + \psi_j\langle\widehat{\psi}_j - \psi_j, \psi_j\rangle. \tag{2.51}$$

Proof. This lemma can be derived from Lemma 5.1 of Hall and Horowitz (2007) and hence the proof is omitted.

Chapter 3

An Autocovariance-based Learning Framework for High-Dimensional Functional Time Series

3.1 Introduction

Functional time series refers to functional data objects that are observed consecutively over time and constitutes an active research area. Existing research has mainly focused on extending standard univariate or low-dimensional multivariate time series methods to the functional domain with theoretical guarantees under an asymptotic framework, e.g., Aue et al. (2015), Bathia et al. (2010), Bosq (2000), Hörmann et al. (2015), Hörmann and Kokoszka (2010), Li et al. (2020), Panaretos and Tavakoli (2013), just to name a few. Rapid development of data collection technology has made high-dimensional functional time series datasets become increasingly common. Examples include hourly measured concentrations of various pollutants, e.g., PM10 trajectories (Hörmann et al., 2015) collected over a number of sites, daily electricity load curves (Cho et al., 2013) for a large number of households, cumulative intraday return trajectories (Horváth et al., 2014), daily return density curves (Bathia et al., 2010) and functional volatility processes (Müller et al., 2011) for a large collection of stocks, and annul temperature curves (Aue and van

Delft, 2020) at different measuring stations.

The datasets, in this context, consist of p-dimensional vector of functional time series, $\mathbf{W}_t(\cdot) = \{W_{t1}(\cdot), \ldots, W_{tp}(\cdot)\}^{\mathrm{T}}$ for $t = 1, \ldots, n$ with (auto)covariance functions $\mathbf{\Sigma}_h^W(u, v) = \mathrm{Cov}\{\mathbf{W}_t(u), \mathbf{W}_{t+h}(v)\}$ for any integer h and $u, v \in \mathcal{U}$ (a compact interval), where p can be diverging with, or even larger than, n in a high-dimensional regime. Suppose the observed curves $\mathbf{W}_t(\cdot)$ are subject to errors in the form of

$$\mathbf{W}_t(\cdot) = \mathbf{X}_t(\cdot) + \mathbf{e}_t(\cdot)\,, \tag{3.1}$$

where $\mathbf{X}_t(\cdot) = \{X_{t1}(\cdot), \ldots, X_{tp}(\cdot)\}^{\mathrm{T}}$ is p-dimensional functional time series of interest and $\mathbf{e}_t(\cdot) = \{e_{t1}(\cdot), \ldots, e_{tp}(\cdot)\}^{\mathrm{T}}$ is a white noise sequence. In the same manner as $\mathbf{\Sigma}_h^W(u, v)$, we define $\mathbf{\Sigma}_h^X(u, v)$ and $\mathbf{\Sigma}_h^e(u, v)$ by replacing $\mathbf{W}_t(\cdot)$ with $\mathbf{X}_t(\cdot)$ and $\mathbf{e}_t(\cdot)$, respectively. We call $\mathbf{e}_t(\cdot)$ is a white noise sequence if $\mathbb{E}\{\mathbf{e}_t(u)\} = \mathbf{0}$ and $\mathbf{\Sigma}_h^e(u, v) = \mathbf{0}$ for any $u, v \in \mathcal{U}$ and $h \neq 0$. This formulation guarantees that all dynamic elements of $\mathbf{W}_t(\cdot)$ are included in the signal term $\mathbf{X}_t(\cdot)$ and all white noise elements are absorbed into $\mathbf{e}_t(\cdot)$. The existence of $\mathbf{e}_t(\cdot)$ reflects that curves $\mathbf{X}_t(\cdot)$ are seldom completely observed. Instead, they are often only measured, with errors, at discrete locations. These noisy discrete data are smoothed to yield "observed" curves $\mathbf{W}_t(\cdot)$. Note that $\{\mathbf{X}_t(\cdot)\}_{t=1}^n$ and $\{\mathbf{e}_t(\cdot)\}_{t=1}^n$ are uncorrelated and unobservable. See Bathia et al. (2010) for the univariate case of model (3.1) with fully nonparametric structure on $\mathbf{\Sigma}_0^e$. When $\mathbf{W}_1(\cdot), \ldots, \mathbf{W}_n(\cdot)$ are univariate and independent, Hall and Vial (2006) addressed the same problem under a 'low noise' setting assuming that $\mathbf{e}_t(\cdot)$ goes to 0 as n goes to ∞. Imposing some parametrically specified structure in the univariate case, $\mathbf{\Sigma}_0^e$ is assumed to be diagonal in Yao et al. (2005) and banded under the assumption that $\mathbf{\Sigma}_0^X$ is finite rank in Descary and Panaretos (2019).

The standard estimation procedure for univariate or low-dimensional functional time series models consists of three steps (Aue et al., 2015). Due to the intrinsic infinite-dimensionality of functional data, the first step performs dimension reduction via e.g. functional principal components analysis (FPCA) to approximate each observed curve by the finite Karhunen-Loève representation, which transforms functional time series observations into a vector time series of FPC scores. The second step transforms the estimation of function-valued parameters involved in the models to the estimation of some vector- or matrix-valued parameters based on the estimated FPC scores. The third step utilises estimated eigenfunctions to obtain the function-valued estimate of interest from the vector- or matrix-valued estimate obtained in the second step. Estimation in the context of high-dimensional functional time series is often impossible without imposing some lower-dimensional structural assumption on the model parameters space. With imposed functional sparsity structure, the

second step needs to consider the estimation under a block (or group) sparsity constraint resulting from the first step, where variables belonging to the same group should be simultaneously included in or excluded from the model. In a regression setup, the group-lasso penalized least squares estimation (Yuan and Lin, 2006) can be implemented in the second step to obtain block sparse estimates, from which the third step can recover functional sparse estimates. Similar three-step procedures have been developed to estimate sparse high-dimensional functional models, see e.g., vector functional autoregression (VFAR) (Guo and Qiao, 2020), scalar-on-function linear additive regression (SFLR) (Fan et al., 2015, Kong et al., 2016b, Xue and Yao, 2020) and function-on-function linear additive regression (FFLR) (Fan et al., 2014, Luo and Qi, 2017) with serially dependent observations.

Under the error contamination model in (3.1), provided that both FPCA and penalized least squares estimation are based on the estimated covariance function of $\mathbf{W}_t(\cdot)$, i.e. $\widehat{\mathbf{\Sigma}}_0^W$, the standard covariance-based procedure is inappropriate given the fact that $\mathbf{\Sigma}_0^W = \mathbf{\Sigma}_0^X + \mathbf{\Sigma}_0^e$ and hence $\widehat{\mathbf{\Sigma}}_0^W$ is not a consistent estimator for $\mathbf{\Sigma}_0^X$. In this chapter, motivated from a simple fact that $\mathbf{\Sigma}_h^W = \mathbf{\Sigma}_h^X$ for any $h \neq 0$, which automatically removes the impact from the noise $\mathbf{e}_t(\cdot)$, we propose an autocovariance-based three-step learning framework. Differing from FPCA via Karhunen-Loève expansion of $W_{tj}(\cdot)$ for each j, our first step of dimension reduction is developed under an alternative data-driven basis expansion of $X_{tj}(\cdot)$ formed by performing eigenanlysis on a positive-definite operator defined based on autocovariance functions of $W_{tj}(\cdot)$. Different from the penalized least squares estimation applied in the second step, we make use of the autocovariance information of the basis coefficients to construct some moment equations and then apply our proposed block regularised method to estimate the associated block sparse vector- or matrix-valued parameters based on the estimated basis coefficients obtained in the first step. In the third step, the block sparse estimates obtained in the second step are re-transformed to sparse function-valued estimates via estimated basis functions obtained in the first step.

There exist several challenges in the theoretical analysis of the proposed autocovariance-based learning framework for high-dimensional functional time series gathering challenges of non-asymptotics (Wainwright, 2019) and infinite-dimensionality with serial dependence (Jirak, 2016). First, our proposed second step is applied to the estimated basis coefficients rather than the true coefficients to produce block sparse estimates whereas the conventional sparse estimation is applied directly to observed data. Accounting for such approximation is a major undertaking. Second, under a high-dimensional and serially dependent setting, it is essential to develop non-asymptotic theory that seeks to provide probabilistic bounds on relevant estimated

terms as a function of n, p and the truncated dimension under our autocovariance-based dimension reduction framework. Third, compared with non-functional data, the infinite-dimensional nature of functional data leads to the additional theoretical complexity that arises from specifying the block structure and controlling the bias terms formed by truncation errors in our dimension reduction step.

The main contribution of this chapter is fourfold.

1. Our autocovariance-based learning framework can address the error contamination model in (3.1) in the presence of infinite-dimensional signal curve dynamics with the addition of 'genuinely functional' white noise. It makes the good use of the serial correlation information in our estimation, which is most relevant in the context of time series modelling.

2. To provide theoretical guarantees for the first and third steps and to verify imposed conditions in the second step, we rely on functional stability measures (Fang et al., 2020, Guo and Qiao, 2020) to characterise the effect of serial dependence and investigate non-asymptotic properties of relevant estimated terms under the autocovariance-based dimension reduction framework we consider.

3. We utilise the autocovariance of basis coefficients to construct high-dimensional moment equations with partitioned group structure, based on which we formulate the second step in a general block regularised minimum distance (RMD) estimation framework so as to produce block sparse estimates. Within such framework, the group information can be explicitly encoded in a convex optimisation targeting at minimising the block ℓ_1 norm objective function subject to the block ℓ_∞ norm constraint. To theoretically support the second step, we also study convergence properties of the block RMD estimator.

4. Exemplarily, we illustrate the autocovariance-based three-step procedure using three sparse high-dimensional functional time series models, i.e. SFLR, FFLR and VFAR. Using our derived theoretical results, we establish convergence rates of the associated estimators in these models. Empirically, we demonstrate the superiority of these autocovariance-based estimators relative to their covariance-based counterparts.

This chapter is set out as follows. In Section 3.2, we propose a general autocovariance-based three-step procedure with illustration using SFLR as an example. In Section 3.3, we present the first step of autocovariance-based dimension reduction and establish essential deviation bounds in elementwise ℓ_∞-norm on relevant estimated

terms used in subsequent analysis. In Section 3.4, we formulate the second step in a general block RMD estimation framework and investigate its theoretical properties. In Section 3.5, we illustrate the proposed autocovariance-based learning framework using applications of SFLR, FFLR and VFAR, and present convergence analysis of the associated estimators. In Section 3.6, we examine the finite-sample performance of the proposed estimators through both an extensive set of simulations and an analysis of a public financial dataset. All technical proofs are relegated to the Appendix.

Notation. For a positive integer q, we denote $[q] = \{1, \ldots, q\}$. Let $L_2(\mathcal{U})$ be a Hilbert space of square integrable functions on a compact interval $\mathcal{U}$. The inner product of $f, g \in L_2(\mathcal{U})$ is $\langle f, g \rangle = \int_{\mathcal{U}} f(u)g(u)\,\mathrm{d}u$. For a Hilbert space $\mathbb{H} \subset L_2(\mathcal{U})$, we denote the p-fold Cartesian product by $\mathbb{H}^p = \mathbb{H} \times \cdots \times \mathbb{H}$ and the tensor product by $\mathbb{S} = \mathbb{H} \otimes \mathbb{H}$. For $\mathbf{f} = (f_1, \ldots, f_p)^{\mathrm{T}}$ and $\mathbf{g} = (g_1, \ldots, g_p)^{\mathrm{T}}$ in $\mathbb{H}^p$, we define $\langle \mathbf{f}, \mathbf{g} \rangle = \sum_{i=1}^p \langle f_i, g_i \rangle$. We use $\|\mathbf{f}\| = \langle \mathbf{f}, \mathbf{f} \rangle^{1/2}$ and $\|\mathbf{f}\|_0 = \sum_{i=1}^p I(\|f_i\| \neq 0)$ with $I(\cdot)$ being the indicator function to denote functional versions of induced norm and ℓ_0-norm, respectively. For an integral operator $\mathbf{K} : \mathbb{H}^p \to \mathbb{H}^q$ induced from the kernel function $\mathbf{K} = (K_{ij})_{q \times p}$ with each $K_{ij} \in \mathbb{S}$, $\mathbf{K}(\mathbf{f})(u) = \{\sum_{j=1}^p \langle K_{1j}(u, \cdot), f_j(\cdot) \rangle, \ldots, \sum_{j=1}^p \langle K_{qj}(u, \cdot), f_j(\cdot) \rangle\}^{\mathrm{T}} \in \mathbb{H}^q$ for any $\mathbf{f} \in \mathbb{H}^p$. For notational economy, we will also use $\mathbf{K}$ to denote both the kernel and the operator. We define functional versions of Frobenius and matrix ℓ_∞-norms by $\|\mathbf{K}\|_{\mathrm{F}} = (\sum_{i=1}^q \sum_{j=1}^p \|K_{ij}\|_{\mathcal{S}}^2)^{1/2}$ and $\|\mathbf{K}\|_\infty = \max_{i \in [q]} \sum_{j=1}^p \|K_{ij}\|_{\mathcal{S}}$, respectively, where $\|K_{ij}\|_{\mathcal{S}} = \{\int_{\mathcal{U}} \int_{\mathcal{U}} K_{ij}^2(u, v)\,\mathrm{d}u\mathrm{d}v\}^{1/2}$ denotes the Hilbert–Schmidt norm of K_{ij}. For any real matrix $\mathbf{B} = (b_{ij})_{q \times p}$, we write $\|\mathbf{B}\|_{\max} = \max_{i \in [q], j \in [p]} |b_{ij}|$ and use $\|\mathbf{B}\|_{\mathrm{F}} = (\sum_{i=1}^q \sum_{j=1}^p |b_{ij}|^2)^{1/2}$ and $\|\mathbf{B}\|_2 = \lambda_{\max}^{1/2}(\mathbf{B}^{\mathrm{T}}\mathbf{B})$ to denote its Frobenius norm and ℓ_2-norm, respectively. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exist a positive constant c such that $a_n/b_n \leq c$. We write $a_n \asymp b_n$ if and only if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously.

3.2 Autocovariance-based three-step procedure

Suppose we observe weakly stationary functional time series $\{\mathbf{W}_t(\cdot)\}_{t \in [n]}$ with mean zero and (auto)covariance functions $\mathbf{\Sigma}_h^W(u, v) = \{\Sigma_{h,jk}^W(u, v)\}_{j,k \in [p]}$ for integer $h \geq 0$ and $(u, v) \in \mathcal{U}^2$, whose sample estimators are given by

$$\widehat{\mathbf{\Sigma}}_h^W(u, v) = \frac{1}{n-h} \sum_{t=1}^{n-h} \mathbf{W}_t(u)\mathbf{W}_{t+h}(v)^{\mathrm{T}} = \{\widehat{\Sigma}_{h,jk}^W(u, v)\}_{j,k \in [p]}\,. \tag{3.2}$$

Our proposed autocovariance-based learning framework consists of the following three steps.

Step 1: Due to the infinite-dimensionality of functional data, for each j, we expand signal curves $X_{tj}(\cdot)$ through data-driven orthonormal basis functions, $\{\psi_{jl}(\cdot)\}_{l=1}^{\infty}$, and approximate them using d_j-dimensional truncation,

$$X_{tj}(\cdot) = \sum_{l=1}^{\infty} \eta_{tjl}\psi_{jl}(\cdot) \approx \boldsymbol{\eta}_{tj}^{\mathrm{T}}\boldsymbol{\psi}_j(\cdot)\,, \quad j \in [p]\,, \tag{3.3}$$

where $\eta_{tjl} = \langle X_{tj}, \psi_{jl}\rangle$, $\boldsymbol{\eta}_{tj} = (\eta_{tj1}, \ldots, \eta_{tjd_j})^{\mathrm{T}} \in \mathbb{R}^{d_j}$ and $\boldsymbol{\psi}_j = (\psi_{j1}, \ldots, \psi_{jd_j})^{\mathrm{T}} \in \mathbb{R}^{d_j}$. Given observed functional time series $\{W_{tj}(\cdot)\}_{t\in[n]}$, we adopt an autocovariance-based dimension reduction approach in Section 3.3, where we can obtain estimated basis functions $\widehat{\boldsymbol{\psi}}_j = (\hat{\psi}_{j1}, \ldots, \hat{\psi}_{jd_j})^{\mathrm{T}}$ and estimated basis coefficients $\widehat{\boldsymbol{\eta}}_{tj} = (\hat{\eta}_{tj1}, \ldots, \hat{\eta}_{tjd_j})^{\mathrm{T}}$ with $\hat{\eta}_{tjl} = \langle W_{tj}, \hat{\psi}_{jl}\rangle$ for $l \in [d_j]$.

Step 2: Based on the dimension reduction in Step 1, we can transform the estimation of function-valued parameters of interest under the sparsity constraint to the block sparse estimation of some vector- or matrix-valued parameters. Let $\mathbb{E}\{\boldsymbol{\eta}_{tj}\boldsymbol{\eta}_{(t+h)k}^{\mathrm{T}}\} = \{\sigma_{jklm}^{(h)}\}_{l\in[d_j],m\in[d_k]}$ with its estimator $(n-h)^{-1}\sum_{t=1}^{n-h}\widehat{\boldsymbol{\eta}}_{tj}\widehat{\boldsymbol{\eta}}_{(t+h)k}^{\mathrm{T}}$ $= \{\hat{\sigma}_{jklm}^{(h)}\}_{l\in[d_j],m\in[d_k]}$ for $j, k \in [p]$ and $h \geq 0$. To identify these vector- or matrix-valued parameters, we use the autocovariance information among the basis coefficients $\{\boldsymbol{\eta}_{tj}\}$ to construct high-dimensional moment equations with partitioned group structure and then rely on estimated autocovariance terms $\{\hat{\sigma}_{jklm}^{(h)} : j, k \in [p], l \in [d_j], m \in [d_k], h \geq 1\}$ to formulate the block RMD estimation as introduced in Section 3.4.

Step 3: We utilise $\{\widehat{\boldsymbol{\psi}}_j(\cdot)\}_{j\in[p]}$ to recover functional sparse estimates from those block sparse estimates obtained in Step 2.

We give some illustration on the rationality of our autocovariance-based procedure. Write $\boldsymbol{\Sigma}_h^X(u,v) = \{\Sigma_{h,jk}^X(u,v)\}_{j,k\in[p]}$ and $\boldsymbol{\Sigma}_h^e(u,v) = \{\Sigma_{h,jk}^e(u,v)\}_{j,k\in[p]}$. In the first step, the classical FPCA is implemented by the eigenanalysis of $\widehat{\Sigma}_{0,jj}^W$ for each j. However, such covariance-based estimation problem is insoluble in the sense that one cannot separate $X_{tj}(\cdot)$ from $W_{tj}(\cdot)$ due to $\Sigma_{0,jj}^W = \Sigma_{0,jj}^X + \Sigma_{0,jj}^e$ and hence $\widehat{\Sigma}_{0,jj}^W$ is no longer a consistent estimator for $\Sigma_{0,jj}^X$. Inspired from $\Sigma_{h,jj}^W = \Sigma_{h,jj}^X$ for any $h \neq 0$, which automatically filters out the impact from $e_{tj}(\cdot)$ and hence guarantees that $\widehat{\Sigma}_{h,jj}^W$ is a legitimate estimator for $\Sigma_{h,jj}^X$, our first step is developed under an alternative data-driven basis expansion of $X_{tj}(\cdot)$ formed by performing eigenanalysis on a positive-definite operator defined based on $\widehat{\Sigma}_{h,jj}^W$ for $h \geq 1$. In the second step,

the commonly adopted penalized least squares approach is based on the sample covariance among the estimated FPC scores $\{\hat{\sigma}^{(0)}_{jklm} : j, k \in [p], l \in [d_j], m \in [d_k]\}$, see e.g. Kong et al. (2016b). However, provided that $\sigma^{(h)}_{jklm} = \langle \psi_{jl}, \Sigma^X_{h,jk}(\psi_{km}) \rangle$ and $\hat{\sigma}^{(h)}_{jklm} = \langle \hat{\psi}_{jl}, \widehat{\Sigma}^W_{h,jk}(\hat{\psi}_{km}) \rangle$, such covariance-based penalized least squares approach is inappropriate due to the fact that $\widehat{\Sigma}^W_{0,jk}$ and $\hat{\sigma}^{(0)}_{jklm}$ are not consistent estimators for $\Sigma^X_{0,jk}$ and $\sigma^{(0)}_{jklm}$, respectively. Motivated from $\Sigma^W_{h,jk} = \Sigma^X_{h,jk}$ for any $h \neq 0$ ensuring that $\hat{\sigma}^{(h)}_{jklm}$ is a legitimate estimator for $\sigma^{(h)}_{jklm}$, the moment equations based on $\{\sigma^{(h)}_{jklm} : j, k \in [p], l \in [d_j], m \in [d_k], h \geq 1\}$ can be well approximated by its empirical version relied on $\{\hat{\sigma}^{(h)}_{jklm}\}$.

We next illustrate the proposed three-step procedure using SFLR as an example. Consider high-dimensional SFLR in the form of

$$Y_t = \sum_{j=1}^{p} \int_{\mathcal{U}} X_{tj}(u)\beta_{0j}(u)\,\mathrm{d}u + \varepsilon_t\,, \quad t \in [n]\,, \tag{3.4}$$

where p-dimensional functional covariates $\{\mathbf{X}_t(\cdot)\}_{t\in[n]}$ satisfying model (3.1) are independent of i.i.d. mean-zero random errors $\{\varepsilon_t\}_{t\in[n]}$, and $\{\beta_{0j}(\cdot)\}_{j\in[p]}$ are unknown functional coefficients. Given observations $\{(\mathbf{W}_t(\cdot), Y_t)\}_{t\in[n]}$, our goal is to estimate $\boldsymbol{\beta}_0(\cdot) = \{\beta_{01}(\cdot), \ldots, \beta_{0p}(\cdot)\}^{\mathrm{T}}$. To guarantee a feasible solution under high-dimensional scaling, we assume that $\boldsymbol{\beta}_0(\cdot)$ is functional s-sparse, i.e. s components in $\{\beta_{0j}(\cdot)\}_{j\in[p]}$ are nonzero with s being much smaller than p.

We expand each $X_{tj}(\cdot)$ according to (3.3) truncated at d_j and rewrite (3.4) as

$$Y_t = \sum_{j=1}^{p} \boldsymbol{\eta}_{tj}^{\mathrm{T}} \mathbf{b}_{0j} + r_t + \varepsilon_t\,,$$

where $\mathbf{b}_{0j} = \int_{\mathcal{U}} \boldsymbol{\psi}_j(u)\beta_{0j}(u)\,\mathrm{d}u \in \mathbb{R}^{d_j}$ and $r_t = \sum_{j=1}^{p}\sum_{l=d_j+1}^{\infty} \eta_{tjl}\langle \psi_{jl}, \beta_{0j}\rangle$ is the truncation error. Given some prescribed positive integer L, we choose $\{\boldsymbol{\eta}_{(t+h)k} : h \in [L], k \in [p]\}$ as vector-valued instrumental variables. Then $\mathbf{b}_0 = (\mathbf{b}_{01}^{\mathrm{T}}, \ldots, \mathbf{b}_{0p}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{\sum_{j=1}^{p} d_j}$ can be identified by the following moment equations:

$$\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\varepsilon_t\} = \mathbf{g}_{hk}(\mathbf{b}_0) + \mathbf{R}_{hk} = \mathbf{0}\,, \quad k \in [p]\,,\, h \in [L]\,, \tag{3.5}$$

where $\mathbf{g}_{hk}(\mathbf{b}_0) = \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k} Y_t\} - \sum_{j=1}^{p} \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{tj}^{\mathrm{T}}\mathbf{b}_{0j}\}$ and the bias term $\mathbf{R}_{hk} = -\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k} r_t\}$.

With $\{\widehat{\boldsymbol{\eta}}_{tj}\}_{t\in[n],j\in[p]}$ and $\{\widehat{\boldsymbol{\psi}}_j(\cdot)\}_{j\in[p]}$ obtained in the first step, for any $\mathbf{b} = (\mathbf{b}_1^{\mathrm{T}}, \ldots, \mathbf{b}_p^{\mathrm{T}})^{\mathrm{T}} \in$

$\mathbb{R}^{\sum_{j=1}^{p} d_j}$, we define

$$\widehat{\mathbf{g}}_{hk}(\mathbf{b}) = \frac{1}{n-h}\sum_{t=1}^{n-h}\widehat{\boldsymbol{\eta}}_{(t+h)k}Y_t - \frac{1}{n-h}\sum_{t=1}^{n-h}\sum_{j=1}^{p}\widehat{\boldsymbol{\eta}}_{(t+h)k}\widehat{\boldsymbol{\eta}}_{tj}^{\mathrm{T}}\mathbf{b}_j\,, \quad k\in[p]\,,\, h\in[L]\,, \tag{3.6}$$

which provides the empirical version of $\mathbf{g}_{hk}(\mathbf{b}) = \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}Y_t\} - \sum_{j=1}^{p}\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{tj}^{\mathrm{T}}\mathbf{b}_j\}$. It follows from (3.5) that

$$\widehat{\mathbf{g}}_{hk}(\mathbf{b}_0) \approx \mathbf{0}\,, \quad k\in[p]\,,\, h\in[L]\,. \tag{3.7}$$

Based on (3.7), applying the block RMD estimation introduced in Section 3.4 results in a block sparse estimator $\widehat{\mathbf{b}} = (\widehat{\mathbf{b}}_1^{\mathrm{T}},\ldots,\widehat{\mathbf{b}}_p^{\mathrm{T}})^{\mathrm{T}}$. Given that the recovery of functional sparsity in $\boldsymbol{\beta}_0(\cdot)$ is equivalent to estimating the block sparsity in $\mathbf{b}_0$, we can estimate functional sparse coefficients in the third step by

$$\hat{\beta}_j(\cdot) = \widehat{\boldsymbol{\psi}}_j(\cdot)^{\mathrm{T}}\widehat{\mathbf{b}}_j\,, \quad j\in[p]\,. \tag{3.8}$$

3.3 Autocovariance-based dimension reduction

3.3.1 Methodology

For each $j\in[p]$, we assume that signal curves $X_{tj}(\cdot)$ admit the Karhunen-Loève expansion $X_{tj}(\cdot) = \sum_{l=1}^{\infty}\xi_{tjl}\nu_{jl}(\cdot)$, where $\xi_{tjl} = \langle X_{tj},\nu_{jl}\rangle$ corresponds to a sequence of random variables with $\mathbb{E}(\xi_{tjl}) = 0$ and $\mathrm{Cov}(\xi_{tjl},\xi_{tjl'}) = \omega_{jl}I(l=l')$. Here $\omega_{j1}\geq\omega_{j2}\geq\cdots\geq 0$ are eigenvalues of $\Sigma_{0,jj}^{X}$ and $\nu_{j1}(\cdot),\nu_{j2}(\cdot),\ldots$ are the corresponding orthonormal eigenfunctions satisfying $\int_{\mathcal{U}}\Sigma_{0,jj}^{X}(u,v)\nu_{jl}(v)\,\mathrm{d}v = \omega_{jl}\nu_{jl}(u)$ for $l\geq 1$. The commonly adopted FPCA is based on applying Karhunen-Loève expansion to observed curves $\{W_{tj}(\cdot)\}_{t\in[n]}$. However, this covariance-based dimension reduction approach is inappropriate under the error contamination model in (3.1) as discussed in Section 3.2. Hall and Vial (2006) tackled such covariance-based problem under the assumption that $W_{1j}(\cdot),\ldots,W_{nj}(\cdot)$ are independent and the noise $e_{tj}(\cdot)$ goes to 0 as n grows to ∞.

Without requiring the restrictive 'low noise' and independence assumption, we follow Bathia et al. (2010) to implement an autocovariance-based dimension reduction approach for observed curves $\{W_{tj}(\cdot)\}_{t\in[n]}$ due to the fact $\Sigma_{h,jj}^{W} = \Sigma_{h,jj}^{X}$ for any $h\neq 0$, which ensures that $\widehat{\Sigma}_{h,jj}^{W}$ is a legitimate estimator for $\Sigma_{h,jj}^{X}$ when $h\neq 0$. Specifically,

we define a nonnegative operator K_{jj} to pull together the autocovariance information at different lags:

$$K_{jj}(u,v) = \sum_{h=1}^{L} \int_{\mathcal{U}} \Sigma_{h,jj}^{X}(u,z)\Sigma_{h,jj}^{X}(v,z)\,\mathrm{d}z = \sum_{h=1}^{L} \int_{\mathcal{U}} \Sigma_{h,jj}^{W}(u,z)\Sigma_{h,jj}^{W}(v,z)\,\mathrm{d}z\,, \quad (3.9)$$

where $L > 0$ is some prescribed fixed integer. See Lam and Yao (2012) for the selection of L in practice. It then follows from the infinite-dimensional analog of Proposition 1 in Bathia et al. (2010) that, under regularity conditions, K_{jj} has the spectral decomposition $K_{jj}(u,v) = \sum_{l=1}^{\infty} \lambda_{jl}\psi_{jl}(u)\psi_{jl}(v)$ with nonzero eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ and corresponding orthonormal eigenfunctions $\psi_{j1}(\cdot), \psi_{j2}(\cdot), \ldots$ such that the expansion in (3.3) holds. This expansion forms the foundation of autocovariance-based dimension reduction for error-contaminated functional time series and generalises the finite-dimensional formulation in Bathia et al. (2010) to the infinite-dimensional setting.

With legitimate estimators $\widehat{\Sigma}_{h,jj}^{W}$ for positive integer h in (3.2), a natural estimator for K_{jj} in (3.9) can be obtained by

$$\begin{aligned} \widehat{K}_{jj}(u,v) &= \sum_{h=1}^{L} \int_{\mathcal{U}} \widehat{\Sigma}_{h,jj}^{W}(u,z)\widehat{\Sigma}_{h,jj}^{W}(v,z)\,\mathrm{d}z \\ &= \frac{1}{(n-L)^2} \sum_{h=1}^{L} \sum_{t,s=1}^{n-L} W_{tj}(u)W_{sj}(v)\langle W_{(t+h)j}, W_{(s+h)j}\rangle\,. \end{aligned} \quad (3.10)$$

Performing eigenanalysis on $\widehat{K}_{jj}$ leads to the estimated eigenpairs $\{(\hat{\lambda}_{jl}, \hat{\psi}_{jl})\}_{l\geq 1}$. The infinite series in the expansion in (3.3) are then truncated at d_j, chosen data-adaptively. In practice, we only observe the erroneous versions $\{W_{tj}(\cdot)\}_{t\in[n]}$ instead of the signal components $\{X_{tj}(\cdot)\}_{t\in[n]}$ themselves, and the estimated basis coefficients are given by $\hat{\eta}_{tjl} = \langle W_{tj}, \hat{\psi}_{jl}\rangle$.

3.3.2 Rates in elementwise ℓ_∞-norm

To characterise the effect of dependence on relevant estimated terms, we will use the functional stability measure of $\{\mathbf{W}_t(\cdot)\}_{t\in\mathbb{Z}}$ proposed in Guo and Qiao (2020).

Condition 3.1. For $\{\mathbf{W}_t(\cdot)\}_{t\in\mathbb{Z}}$, the spectral density operator $\mathbf{f}_\theta^W = (2\pi)^{-1}\sum_{h\in\mathbb{Z}} \mathbf{\Sigma}_h^W e^{-ih\theta}$ for $\theta \in [-\pi,\pi]$ exists and the functional stability measure defined in (3.11) is finite, i.e.

$$\mathcal{M}^W = 2\pi \cdot \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi],\mathbf{\Phi}\in\mathbb{H}_0^p} \frac{\langle \mathbf{\Phi}, \mathbf{f}_\theta^W(\mathbf{\Phi})\rangle}{\langle \mathbf{\Phi}, \mathbf{\Sigma}_0^W(\mathbf{\Phi})\rangle} < \infty\,, \quad (3.11)$$

where $\mathbb{H}_0^p = \{\mathbf{\Phi} \in \mathbb{H}^p : \langle \mathbf{\Phi}, \mathbf{\Sigma}_0^W(\mathbf{\Phi})\rangle \in (0, \infty)\}$.

Here $\mathcal{M}^W$ in (3.11) is expressed proportional to functional Rayleigh quotients of $\mathbf{f}_\theta^W$ relative to $\mathbf{\Sigma}_0^W$ and hence it can more precisely capture the effect of eigenvalues of $\mathbf{f}_\theta^W$ relative to small decaying eigenvalues of $\mathbf{\Sigma}_0^W$, which is essential to handle truly infinite-dimensional functional objects $\{W_{tj}(\cdot)\}$. We next define the functional stability measure of all k-dimensional subsets of $\{\mathbf{W}_t(\cdot)\}_{t\in\mathbb{Z}}$, i.e. $\{(W_{tj}(\cdot) : j \in J)^{\mathrm{T}}\}_{t\in\mathbb{Z}}$ for $J \subset [p]$ with cardinality $|J| \le k$, by

$$\mathcal{M}_k^W = 2\pi \cdot \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi], \|\mathbf{\Phi}\|_0 \le k, \mathbf{\Phi}\in\mathcal{H}_0^p} \frac{\langle \mathbf{\Phi}, \mathbf{f}_\theta^W(\mathbf{\Phi})\rangle}{\langle \mathbf{\Phi}, \mathbf{\Sigma}_0^W(\mathbf{\Phi})\rangle}, \quad k \in [p]. \tag{3.12}$$

Under Condition 3.1, it is easy to verify that $\mathcal{M}_k^W \le \mathcal{M}^W < \infty$, which will be used in our non-asymptotic analysis. Provided that our non-asymptotic results are developed using the infinite-dimensional analog of Hanson–Wright inequality (Rudelson and Vershynin, 2013) in a general Hilbert space $\mathbb{H}$, we need to specify the sub-Gaussian random variables therein.

Definition 3.1. Let $Z_t(\cdot)$ be a mean zero random variable in $\mathbb{H}$ for any fixed t and $\Sigma_0 : \mathbb{H} \to \mathbb{H}$ be a covariance operator. Then $Z_t(\cdot)$ is a sub-Gaussian process if there exists a constant $c > 0$ such that $\mathbb{E}(e^{\langle x, Z\rangle}) \le e^{c^2\langle x, \Sigma_0(x)\rangle/2}$ for all $x \in \mathbb{H}$.

Condition 3.2. (i) $\{\mathbf{W}_t(\cdot)\}_{t\in\mathbb{Z}}$ is a sequence of multivariate functional linear processes with sub-Gaussian errors, namely sub-Gaussian functional linear processes, $\mathbf{W}_t(\cdot) = \sum_{l=0}^\infty \mathbf{B}_l(\boldsymbol{\varepsilon}_{t-l})$ for any $t \in \mathbb{Z}$, where $\mathbf{B}_l = (B_{l,jk})_{p\times p}$ with each $B_{l,jk} \in \mathbb{H}\otimes\mathbb{H}$, $\boldsymbol{\varepsilon}_t(\cdot) = \{\varepsilon_{t1}(\cdot), \ldots, \varepsilon_{tp}(\cdot)\}^{\mathrm{T}} \in \mathbb{H}^p$ and the components in $\{\boldsymbol{\varepsilon}_t(\cdot)\}_{t\in\mathbb{Z}}$ are independent sub-Gaussian processes satisfying Definition 3.1; (ii) The coefficient functions satisfy $\sum_{l=0}^\infty \|\mathbf{B}_l\|_\infty = O(1)$; (iii) $\omega_0^\varepsilon = \max_{j\in[p]} \int_{\mathcal{U}} \Sigma_{0,jj}^\varepsilon(u,u)\,\mathrm{d}u = O(1)$, where $\Sigma_{0,jj}^\varepsilon(u,u) = \mathrm{Cov}\{\varepsilon_{tj}(u), \varepsilon_{tj}(u)\}$.

The multivariate functional linear process can be seen as the generalisation of functional linear process (Bosq, 2000) to the multivariate setting as well as the extension of multivariate linear process (Hamilton, 1994) to the functional domain. According to Fang et al. (2020), Condition 3.2(ii) ensures the stationarity of $\{\mathbf{W}_t(\cdot)\}_{t\in\mathbb{Z}}$ and, together with Condition 3.2(iii), implies that $\omega_0^W = \max_{j\in[p]} \int_{\mathcal{U}} \Sigma_{0,jj}^W(u,u)\,\mathrm{d}u = O(1)$, which is essential in subsequent analysis.

Condition 3.3. (i) For each $j \in [p]$, it holds that $\lambda_{j1} > \lambda_{j2} > \cdots > 0$, and there exist some positive constants c_0 and $\alpha > 1$ such that $\lambda_{jl} - \lambda_{j(l+1)} \ge c_0 l^{-\alpha-1}$ for $l \ge 1$; (ii) For each $j \in [p]$, the linear space spanned by $\{\nu_{jl}(\cdot)\}_{l=1}^\infty$ is the same as that spanned by $\{\psi_{jl}(\cdot)\}_{l=1}^\infty$.

Condition 3.3(i) controls the lower bound of eigengaps with larger values of α yielding tighter gaps and also implies that $\lambda_{jl} \geq c_0 \alpha^{-1} l^{-\alpha}$. See similar conditions in Hall and Horowitz (2007) and Kong et al. (2016b). To simplify notation, we assume the same α across j, but this condition can be relaxed by allowing α to depend on j and our theoretical results can be generalised accordingly. We next establish the deviation bounds on estimated eigenpairs, $\{(\hat{\lambda}_{jl}, \hat{\psi}_{jl})\}$, and the sample autocovariance among estimated basis coefficients, $\{\hat{\sigma}^{(h)}_{jklm}\}$, in elementwise ℓ_∞-norm, which play a crucial role in further convergence analysis under high-dimensional scaling.

Theorem 3.1. *Let Conditions* 3.1–3.3 *hold, d be a positive integer possibly depending on (n, p). If $n \gtrsim \log(pd)$, then there exist some positive constants c_1 and c_2 independent of (n, p, d) such that*

$$\max_{j\in[p],l\in[d]} \left\{ |\hat{\lambda}_{jl} - \lambda_{jl}| + \left\| \frac{\hat{\psi}_{jl} - \psi_{jl}}{l^{\alpha+1}} \right\| \right\} \lesssim \mathcal{M}_1^W \sqrt{\frac{\log(pd)}{n}} \tag{3.13}$$

holds with probability greater than $1 - c_1(pd)^{-c_2}$, where $\mathcal{M}_1^W$ is defined in (3.12).

Theorem 3.2. *Let conditions in Theorem* 3.1 *hold and $h \geq 1$ be fixed. If $n \gtrsim d^{2\alpha+2}(\mathcal{M}_1^W)^2 \log(pd)$, then there exist some positive constants c_3 and c_4 independent of (n, p, d) such that*

$$\max_{j,k\in[p],l,m\in[d]} \frac{|\hat{\sigma}^{(h)}_{jklm} - \sigma^{(h)}_{jklm}|}{(l \vee m)^{\alpha+1}} \lesssim \mathcal{M}_1^W \sqrt{\frac{\log(pd)}{n}} \tag{3.14}$$

holds with probability greater than $1 - c_3(pd)^{-c_4}$, where $\mathcal{M}_1^W$ is defined in (3.12).

Remark 3.1. The parameter d in Theorems 3.1 and 3.2 can be understood as the truncated dimension of infinite-dimensional functional objects under the expansion in (3.3). In general, d can depend on j, say d_j, then the right-sides of (3.13) and (3.14) become $\mathcal{M}_1^W n^{-1/2} \log^{1/2}(\sum_{j=1}^p d_j)$. Compared with normalised deviation bounds on estimated eigenpairs, $\{(\hat{\omega}_{jl}, \hat{\nu}_{jl}\}$, and sample autocovariance among estimated FPC scores established in Guo and Qiao (2020), we obtain slower rates in (3.13) and (3.14) for decaying eigenvalues. Intuitively, as opposed to the expansion of X_{tj} through $\psi_{j1}, \psi_{j2}, \ldots$ with correlated coefficients, $\nu_{j1}, \nu_{j2}, \ldots$, provide the unique basis with respect to which X_{tj} can be expressed as Karhunen–Loève expansion with uncorrelated coefficients and gives the most rapidly convergent representation of X_{tj} in the L_2 sense. From a theoretical viewpoint, whether the rates in (3.13) and (3.14) are optimal in the minimax sense is of interest and requires further exploration.

3.4 Block RMD estimation framework

3.4.1 A general estimation procedure

In this section, we present the proposed second step in a general block RMD estimation framework. For high dimensional GMM problem, it is essential to select the moment conditions and the covariates simultaneously to prevent the accumulation of estimation errors (Fan and Liao, 2014). And the selection is implied by the RMD estimation in the sense that blocked ℓ_∞-norm are applied in the programming problem (3.16). An alternative method is LASSO-type regularisation proposed in Caner and Kock (2018), which, however, did not perform the moments selection and circumvent this problem by a more restricted assumption on the bounded minimum restricted eigenvalue instead of its blocked version shown in Condition 3.7.

Resulting from the dimension reduction step, the estimation of function-valued parameters involved in sparse high-dimensional functional models can be transformed to the block sparse estimation of some vector- or matrix-valued parameters, $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{0p}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{\sum_{j=1}^{p} d_j \times \tilde{d}}$ with each $\boldsymbol{\theta}_{0j} \in \mathbb{R}^{d_j \times \tilde{d}}$, under high-dimensional scaling. For SFLR with a scalar response, $\tilde{d} = 1$. For FFLR and VFAR, $\tilde{d} \geq 1$ is the truncated dimension of the functional response. Given some prescribed positive integer L and $q = pL$ target moment functions $\boldsymbol{\theta} \mapsto \mathbf{g}_i(\boldsymbol{\theta})$ mapping $\boldsymbol{\theta} \in \mathbb{R}^{\sum_{j=1}^{p} d_j \times \tilde{d}}$ to $\mathbf{g}_i(\boldsymbol{\theta}) \in \mathbb{R}^{d_k \times \tilde{d}}$ with $i = (h-1)p + k$ and $k \in [p]$ for $h \in [L]$, where both p and q are large, we assume that $\boldsymbol{\theta}_0$ can be identified by the following moment equations:

$$\mathbf{g}_i(\boldsymbol{\theta}_0) + \mathbf{R}_i = \mathbf{0}\,, \quad i \in [q]\,, \tag{3.15}$$

where $\mathbf{R}_i$'s are formed by autocovariance-based truncation errors due to the finite approximation in the first step. We are interested in estimating block sparse $\boldsymbol{\theta}_0$ based on empirical mappings $\boldsymbol{\theta} \mapsto \widehat{\mathbf{g}}_i(\boldsymbol{\theta})$ of $\boldsymbol{\theta} \mapsto \mathbf{g}_i(\boldsymbol{\theta})$ for $i \in [q]$. See Sections 3.2 and 3.5 for detailed expressions of $\mathbf{g}_i(\cdot)$ and $\widehat{\mathbf{g}}_i(\cdot)$ in some exemplified models.

We define the block RMD estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\theta}}_p^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{\sum_{j=1}^{p} d_j \times \tilde{d}}$ as a solution to the following convex optimisation problem:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{p} \|\boldsymbol{\theta}_j\|_{\mathrm{F}} \;\text{ subject to }\; \max_{i \in [q]} \|\widehat{\mathbf{g}}_i(\boldsymbol{\theta})\|_{\mathrm{F}} \leq \gamma_n\,, \tag{3.16}$$

where $\gamma_n \geq 0$ is a regularisation parameter. For SFLR or FFLR with $\tilde{d} = 1$, the matrix Frobenius norm in (3.16) degenerates to the vector ℓ_2-norm. For FFLR and VFAR with $\tilde{d} > 1$, the corresponding optimisation tasks are formulated under the

matrix Frobenius norm. The group information is encoded in the objective function, which forces the elements of $\widehat{\boldsymbol{\theta}}_j$ to either all be zero or nonzero, thus producing the block sparsity in $\widehat{\boldsymbol{\theta}}$. It is worth noting that, without the bias terms $\mathbf{R}_i$'s in (3.15), our proposed block RMD estimation framework can be seen as a blockwise generalisation of the RMD estimation (Belloni et al., 2018) by replacing $|\cdot|$ with the $\|\cdot\|_{\mathrm{F}}$. To solve the large-scale convex optimisation problem in (3.16), we use the R package `CVXR` (Fu et al., 2020), which is easy to implement and converges fast. In Sections 3.5.1, 3.5.2 and 3.5.3, we will illustrate our proposed autocovariance-based block RMD estimation framework using examples of SFLR, FFLR and VFAR, respectively, in the context of high-dimensional functional time series.

3.4.2 Theoretical properties

We begin with some notation that will be used in this section. For a block matrix $\mathbf{B} = (\mathbf{B}_{ij})_{i\in[N_1],j\in[N_2]} \in \mathbb{R}^{N_1 m_1 \times N_2 m_2}$ with the (i,j)-th block $\mathbf{B}_{ij} \in \mathbb{R}^{m_1\times m_2}$, we define $\|\mathbf{B}\|_{\max}^{(m_1,m_2)} = \max_{i\in[N_1],j\in[N_2]} \|\mathbf{B}_{ij}\|_{\mathrm{F}}$. When $N_2 = 1$, we also define $\|\mathbf{B}\|_1^{(m_1,m_2)} = \sum_{i=1}^{N_1} \|\mathbf{B}_i\|_{\mathrm{F}}$. To simplify notation in this section and theoretical analysis in Section 3.5, we assume the same truncated dimension $d_j = d$ across $j \in [p]$, but our theoretical results extend naturally to the more general setting where d_j's are different. Let $\mathbf{g}(\boldsymbol{\theta}) = \{\mathbf{g}_1(\boldsymbol{\theta})^{\mathrm{T}}, \ldots, \mathbf{g}_q(\boldsymbol{\theta})^{\mathrm{T}}\}^{\mathrm{T}}$ and $\mathbf{R} = (\mathbf{R}_1^{\mathrm{T}}, \ldots, \mathbf{R}_q^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{qd\times\tilde{d}}$. We focus on the case of which the moment function $\boldsymbol{\theta} \mapsto \mathbf{g}(\boldsymbol{\theta})$ mapping from $\mathbb{R}^{pd\times\tilde{d}}$ to $\mathbb{R}^{qd\times\tilde{d}}$ is linear with respect to $\boldsymbol{\theta}$ in the form of $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{G}\boldsymbol{\theta} + \mathbf{g}(\mathbf{0})$ for some $\mathbf{G} \in \mathbb{R}^{qd\times pd}$. This together with (3.15) implies that

$$\mathbf{G}\boldsymbol{\theta}_0 + \mathbf{g}(\mathbf{0}) + \mathbf{R} = \mathbf{0}\,, \tag{3.17}$$

the form of which can be easily verified for SFLR, FFLR and VFAR models we consider in this chapter. Now we reformulate the optimisation task in (3.16) as

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1^{(d,\tilde{d})} \;\text{ subject to }\; \|\widehat{\mathbf{g}}(\boldsymbol{\theta})\|_{\max}^{(d,\tilde{d})} \le \gamma_n\,, \tag{3.18}$$

where $\widehat{\mathbf{g}}(\boldsymbol{\theta}) = \widehat{\mathbf{G}}\boldsymbol{\theta} + \widehat{\mathbf{g}}(\mathbf{0})$ is the empirical version of $\mathbf{g}(\boldsymbol{\theta})$. It is worth noting that $\boldsymbol{\theta}_0$ is block s-sparse with support $S = \{j \in [p] : \|\boldsymbol{\theta}_{0j}\|_{\mathrm{F}} \neq 0\}$ and its cardinality $s = |S|$.

Before presenting properties of the block RMD estimator $\widehat{\boldsymbol{\theta}}$, we impose some high-level regularity conditions.

Condition 3.4. There exists $\epsilon_{n1}, \delta_{n1} > 0$ such that $\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\max}^{(d,d)} \vee \|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})\|_{\max}^{(d,\tilde{d})} \le \epsilon_{n1}$ with probability at least $1 - \delta_{n1}$.

Condition 3.5. There exists $\epsilon_2 > 0$ such that $\|\mathbf{R}\|_{\max}^{(d,\tilde{d})} \leq \epsilon_2$.

Condition 3.6. There exists $\delta_{n2} > 0$ such that $\|\widehat{\mathbf{g}}(\boldsymbol{\theta}_0)\|_{\max}^{(d,\tilde{d})} \leq \gamma_n$ with probability at least $1 - \delta_{n2}$.

Conditions 3.4 and 3.5 together ensure that the empirical moment functions are nicely concentrated around the target moment functions. Using our derived non-asymptotic results in Section 3.3.2, we can easily specify the concentration bounds in Condition 3.4 for SFLR, FFLR and VFAR. With further imposed smoothness conditions on coefficient functions, Condition 3.5 can also be verified. Condition 3.6 indicates that $\boldsymbol{\theta}_0$ is feasible in the optimisation problem (3.18) with high probability, in which case a solution $\widehat{\boldsymbol{\theta}}$ of (3.18) exists and satisfies $\|\widehat{\boldsymbol{\theta}}\|_1^{(d,\tilde{d})} \leq \|\boldsymbol{\theta}_0\|_1^{(d,\tilde{d})}$. The non-block version of such property typically plays a crucial role to tackle high-dimensional models in the literature.

Let $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$. We define a block ℓ_1-sensitivity coefficient

$$\kappa(\boldsymbol{\theta}_0) = \inf_{T:|T|\leq s} \inf_{\boldsymbol{\delta}\in C_T:\|\boldsymbol{\delta}\|_1^{(d,\tilde{d})}>0} \frac{\|\mathbf{G}\boldsymbol{\delta}\|_{\max}^{(d,\tilde{d})}}{\|\boldsymbol{\delta}\|_1^{(d,\tilde{d})}}, \tag{3.19}$$

where $C_T = \{\boldsymbol{\delta} \in \mathbb{R}^{pd\times\tilde{d}} : \|\boldsymbol{\delta}_{T^c}\|_1^{(d,\tilde{d})} \leq \|\boldsymbol{\delta}_T\|_1^{(d,\tilde{d})}\}$ for $T \subset [p]$. Provided that $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \in C_S$ under Condition 3.6 as justified in Lemma 3.1 of the Appendix, the lower bound of $\kappa(\boldsymbol{\theta}_0)$ is useful to establish the error bound for $\|\widehat{\boldsymbol{\delta}}\|_1^{(d,\tilde{d})}$. See also Gautier and Rose (2019) for non-block l_q-sensitivity quantities to handle high-dimensional instruments. We then need Condition 3.7 below to determine such lower bound. Before presenting this condition, we introduce some notation. Let $J \subset [q]$ and $M \subset [p]$, let $\mathbf{G}_{J,M} = (\mathbf{G}_{jk})_{j\in J,k\in M}$ with each $\mathbf{G}_{jk} \in \mathbb{R}^{d\times d}$ be the block submatrix of $\mathbf{G}$ consisting of all block rows $j \in J$ and all block columns $k \in M$ of $\mathbf{G}$. For an integer $m \geq s$, we define

$$\sigma_{\min}(m, \mathbf{G}) = \min_{|M|\leq m} \max_{|J|\leq m} \sigma_{\min}(\mathbf{G}_{J,M}) \quad \text{and} \quad \sigma_{\max}(m, \mathbf{G}) = \max_{|M|\leq m} \max_{|J|\leq m} \sigma_{\max}(\mathbf{G}_{J,M}),$$

where $\sigma_{\min}(\mathbf{G}_{J,M})$ and $\sigma_{\max}(\mathbf{G}_{J,M})$ are the smallest and largest singular values of $\mathbf{G}_{J,M}$.

Condition 3.7. There exists universal constants $c_5 > 0$ and $\mu > 0$ such that $\sigma_{\max}(m, \mathbf{G}) \geq c_5$ and $\sigma_{\min}(m, \mathbf{G})/\sigma_{\max}(m, \mathbf{G}) \geq \mu$ for $m = 16s/\mu^2$.

In Condition 3.7, the quantity μ serves as a key factor to determine the lower bound of $\kappa(\boldsymbol{\theta}_0)$, which is justified in Lemma 3.3 of the Appendix. When μ is bounded away

from zero, we have a strongly-identified model. When $\mu \to 0$, it corresponds to the scenario with weak instruments. See also Belloni et al. (2018) for similar conditions. We are now ready to present the theorem on the convergence rate of $\widehat{\boldsymbol{\theta}}$.

Theorem 3.3. *Suppose that Conditions* 3.4–3.7 *hold. If* $\|\boldsymbol{\theta}_0\|_1^{(d,\tilde{d})} \leq K$ *for some* $K > 0$ *and the regularisation parameter* $\gamma_n \lesssim (K+1)\epsilon_{n1} + \epsilon_2$, *then with probability at least* $1 - (\delta_{n1} + \delta_{n2})$, *the block RMD estimator* $\widehat{\boldsymbol{\theta}}$ *satisfies*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1^{(d,\tilde{d})} \lesssim s\mu^{-2}\{(K+1)\epsilon_{n1} + \epsilon_2\}. \tag{3.20}$$

Remark 3.2. (i) The error bound in (3.20) has the familiar variance-bias trade-off as commonly considered in nonparametrics statistics, suggesting us to carefully select the truncated dimension d so as to balance variance and bias terms for the optimal estimation. (ii) With commonly imposed smoothness conditions on functional coefficients, it is easy to verify that $K \vee \epsilon_2 = o(s)$ for SFLR, FFLR and VFAR in Section 3.5. (iii) For three examples we consider, $\mathbf{G}$ is formed by $\{\sigma_{jklm}^{(h)} : j,k \in [p], l,m \in [d], h \in [L]\}$ with the components $\sigma_{jklm}^{(h)}$ satisfying $|\sigma_{jklm}^{(h)}| \leq \{\mathbb{E}(\eta_{tjl}^2)\}^{1/2}[\mathbb{E}\{\eta_{(t+h)km}^2\}]^{1/2} = \lambda_{jl}^{1/2}\lambda_{km}^{1/2} \to 0$ as $l,m \to \infty$. Consider a general cross-covariance matrix $\mathbf{G} = \mathbb{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}}) \in \mathbb{R}^{qd \times pd}$ with entries decaying to zero as $d \to \infty$, where $\mathbf{x} = (x_1, \ldots, x_{qd})^{\mathrm{T}}$ with $\mathbb{E}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{y} = (y_1, \ldots, y_{pd})^{\mathrm{T}}$ with $\mathbb{E}(\mathbf{y}) = \mathbf{0}$, it is more sensible to impose Condition 3.7 on its normalised version $\widetilde{\mathbf{G}} = \mathbf{D}_x\mathbf{G}\mathbf{D}_y$ instead of $\mathbf{G}$ itself, where $\mathbf{D}_x = \mathrm{diag}\{\mathrm{Var}(x_1)^{-1/2}, \ldots, \mathrm{Var}(x_{qd})^{-1/2}\}$ and $\mathbf{D}_y = \mathrm{diag}\{\mathrm{Var}(y_1)^{-1/2}, \ldots, \mathrm{Var}(y_{pd})^{-1/2}\}$. For three exemplified models, $\mathbf{D}_x$ and $\mathbf{D}_y$ are formed by $\{\lambda_{jl}^{-1/2} : j \in [p], l \in [d]\}$.

Remark 3.2(iii) motivates us to present the following proposition that will be used in the theoretical analysis of associate estimators for SFLR, FFLR and VFAR in Section 3.5.

Proposition 3.1. *Suppose that all conditions in Theorem* 3.3 *hold except that Condition* 3.7 *holds for* $\widetilde{\mathbf{G}}$, *then with probability at least* $1 - (\delta_{n1} + \delta_{n2})$, *the block RMD estimator* $\widehat{\boldsymbol{\theta}}$ *satisfies*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1^{(d,\tilde{d})} \lesssim s\mu^{-2}\|\mathbf{D}_x\|_{\max}\|\mathbf{D}_y\|_{\max}\{(K+1)\epsilon_{n1} + \epsilon_2\}. \tag{3.21}$$

3.5 Applications

In this section, we present the proposed autocovariance-based estimation procedure with corresponding convergence analysis using applications of SFLR, FFLR and

VFAR models under high-dimensional scaling in Sections 3.5.1, 3.5.2 and 3.5.3, respectively.

3.5.1 High-dimensional SFLR

Within the learning framework in Section 3.2, we first perform autocovariance-based dimension reduction on $\{W_{tj}(\cdot)\}_{t\in[n]}$ for each $j \in [p]$. Following the optimisation framework in (3.16), we then develop the block RMD estimator $\widehat{\mathbf{b}}$ as a solution to the constrained optimisation problem below:

$$\widehat{\mathbf{b}} = \arg\min_{\mathbf{b}} \sum_{j=1}^{p} \|\mathbf{b}_j\|_2 \;\text{ subject to }\; \max_{k\in[p],h\in[L]} \|\widehat{\mathbf{g}}_{hk}(\mathbf{b})\|_2 \leq \gamma_n \,,$$

where $\gamma_n \geq 0$ is a regularisation parameter and $\widehat{\mathbf{g}}_{hk}(\mathbf{b})$ is defined in (3.6). Finally, we obtain estimated functional coefficients $\{\hat{\beta}_j(\cdot)\}_{j\in[p]}$ as in (3.8).

We next present the convergence analysis of $\{\hat{\beta}_j(\cdot)\}_{j\in[p]}$. To simplify notation, we assume the same truncated dimension $d_j = d$ across $j \in [p]$. We rewrite (3.5) in the form of (3.17), where $\mathbf{g} = (\mathbf{g}_{11}^{\mathrm{T}}, \ldots, \mathbf{g}_{1p}^{\mathrm{T}}, \ldots, \mathbf{g}_{L1}^{\mathrm{T}}, \ldots, \mathbf{g}_{Lp}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{R} = (\mathbf{R}_{11}^{\mathrm{T}}, \ldots, \mathbf{R}_{1p}^{\mathrm{T}}, \ldots, \mathbf{R}_{L1}^{\mathrm{T}}, \ldots, \mathbf{R}_{Lp}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{G} = (\mathbf{G}_{ij}) \in \mathbb{R}^{pLd\times pd}$ whose (i,j)-th block is $\mathbf{G}_{ij} = \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{tj}^{\mathrm{T}}\} \in \mathbb{R}^{d\times d}$ with $i = (h-1)p + k$ and $k \in [p]$ for $h \in [L]$. Applying Theorem 3.2 and Proposition 3.3 in the Appendix on $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{g}}(\mathbf{0})$, respectively, we can verify Condition 3.4 with the choice of $\epsilon_{n1} \asymp \mathcal{M}_{W,Y} d^{\alpha+2}\{\log(pd)/n\}^{1/2}$, where $\mathcal{M}_{W,Y}$ is specified in Proposition 3.3 in the Appendix. Before presenting the main theorem, we list two regularity conditions.

Condition 3.8. For each $j \in S = \{j \in [p] : \|\beta_{0j}\| \neq 0\}$, $\beta_{0j}(\cdot) = \sum_{l=1}^{\infty} a_{jl}\psi_{jl}(\cdot)$ and there exists some positive constant $\tau > \alpha + 1/2$ such that $|a_{jl}| \lesssim l^{-\tau}$ for $l \geq 1$.

Condition 3.9. Let $\widetilde{\mathbf{G}} = (\widetilde{\mathbf{G}}_{ij})$ be the normalised version of $\mathbf{G} = (\mathbf{G}_{ij})$ by replacing each $\mathbf{G}_{ij}$ with $\widetilde{\mathbf{G}}_{ij} = \mathbb{E}\{\mathbf{D}_k\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{tj}^{\mathrm{T}}\mathbf{D}_j\}$, $i = (h-1)p + k$, $k \in [p]$ for $h \in [L]$ and $j \in [p]$, where $\mathbf{D}_j = \mathrm{diag}(\lambda_{j1}^{-1/2}, \ldots, \lambda_{jd}^{-1/2})$. Then there exists an universal constant c_6 and $\mu > 0$ such that $\sigma_{\max}(m, \widetilde{\mathbf{G}}) \geq c_6$ and $\sigma_{\min}(m, \widetilde{\mathbf{G}})/\sigma_{\max}(m, \widetilde{\mathbf{G}}) \geq \mu$ for $m = 16s/\mu^2$.

Condition 3.8 restricts each component in $\{\beta_{0j}(\cdot) : j \in S\}$ based on its expansion through basis $\{\psi_{jl}(\cdot)\}_{l\geq 1}$. The parameter τ determines the decay rate of basis coefficients and hence control the level of smoothness with large values yielding smoother functions in $\{\beta_{0j}(\cdot) : j \in S\}$. See similar conditions in Hall and Horowitz (2007) and Kong et al. (2016b). Noting that components of $\mathbf{G}$ decay to zero as d grows

to infinity, we impose Condition 3.9 on $\widetilde{\mathbf{G}}$, which can be viewed as the normalised counterpart of Condition 3.7 for SFLR.

Applying Proposition 3.1 and Theorem 3.1 yields the convergence rate of the SFLR estimate $\widehat{\boldsymbol{\beta}}(\cdot) = \{\hat{\beta}_1(\cdot), \ldots, \hat{\beta}_p(\cdot)\}^{\mathrm{T}}$ under functional ℓ_1 norm in the following theorem.

Theorem 3.4. *Suppose that Conditions* 3.1–3.3 *and Condition* 3.13(ii) *in the Appendix hold for sub-Gaussian functional linear process* $\{\mathbf{W}_t(\cdot)\}$ *and sub-Gaussian linear process* $\{\mathbf{Y}_t\}$, *and also Conditions* 3.8–3.9 *hold. If the regularisation parameter* $\gamma_n \asymp s[d^{\alpha+2}\mathcal{M}_{W,Y}\{\log(pd)/n\}^{1/2} + d^{-\tau+1/2}]$, *then the estimate* $\widehat{\boldsymbol{\beta}}(\cdot)$ *satisfies*

$$\sum_{j=1}^{p} \|\hat{\beta}_j - \beta_{0j}\| = O_{\mathrm{p}}\left\{\mu^{-2}s^2\left(d^{2\alpha+2}\mathcal{M}_{W,Y}\sqrt{\frac{\log(pd)}{n}} + d^{\alpha-\tau+1/2}\right)\right\}. \tag{3.22}$$

Remark 3.3. The rate of convergence in (3.22) is governed by both dimensionality parameters (n, p, s) and internal parameters $(\mathcal{M}_{W,Y}, d, \alpha, \tau, \mu)$. Typically, the rate is better when τ, μ are large and $\mathcal{M}_{W,Y}$ and α are small. To balance variance and bias terms in (3.22) for the optimal estimation, we can choose the truncated dimension d satisfying $\mathcal{M}_{W,Y}^2 \log(pd) d^{2\tau+2\alpha+3} \asymp n$.

3.5.2 High-dimensional FFLR

Consider high-dimensional FFLR in the form of

$$Y_t(v) = \sum_{j=1}^{p} \int_{\mathcal{U}} X_{tj}(u)\beta_{0j}(u, v)\,\mathrm{d}u + \varepsilon_t(v)\,, \quad t \in [n]\,,\, v \in \mathcal{V}\,, \tag{3.23}$$

where $\{\mathbf{X}_t(\cdot)\}_{t\in[n]}$ satisfy model (3.1) and are independent of i.i.d. mean-zero functional errors $\{\varepsilon_t(\cdot)\}_{t\in[n]}$, and $\{\beta_{0j}(\cdot,\cdot)\}_{j\in[p]}$ are functional coefficients to be estimated. With observed data $\{(\mathbf{W}_t(u), Y_t(v)) : (u, v) \in \mathcal{U} \times \mathcal{V}, t \in [n]\}$, we target to estimate $\boldsymbol{\beta}_0 = \{\beta_{01}(\cdot,\cdot), \ldots, \beta_{0p}(\cdot,\cdot)\}^{\mathrm{T}}$ under a functional sparsity constraint when p is large. Specifically, we assume $\boldsymbol{\beta}_0$ is functional s-sparse with support $S = \{j \in [p] : \|\beta_{0j}\|_{\mathcal{S}} \neq 0\}$ and cardinality $s = |S| \ll p$.

Provided that each observed $Y_t(\cdot)$ is decomposed into the sum of dynamic and white noise components in (3.23), we approximate $Y_t(\cdot)$ under the Karhunen–Loève expansion truncated at $\tilde{d}$, i.e. $Y_t(\cdot) \approx \boldsymbol{\zeta}_t^{\mathrm{T}}\boldsymbol{\phi}(\cdot)$, where $\boldsymbol{\zeta}_t = (\zeta_{t1}, \ldots, \zeta_{t\tilde{d}})^{\mathrm{T}}$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{\tilde{d}})^{\mathrm{T}}$. Note that we can relax the independence assumption for $\{\varepsilon_t(\cdot)\}_{t\in[n]}$ and model observed response curves via $\widetilde{Y}_t(\cdot) = Y_t(\cdot) + e_t^Y(\cdot)$, where $Y_t(\cdot)$ and $e_t^Y(\cdot)$

correspond to the dynamic signal and white noise elements, respectively. Then $Y_t(\cdot)$ can be approximated under the autocovariance-based expansion in the sense of (3.3) and our subsequent analysis still follow. For each $j \in [p]$, we expand $X_{tj}(\cdot)$ according to (3.3) truncated at d_j. Some specific calculations lead to the representation of (3.23) as

$$\boldsymbol{\zeta}_t^{\mathrm{T}} = \sum_{j=1}^{p} \boldsymbol{\eta}_{tj}^{\mathrm{T}} \mathbf{B}_{0j} + \mathbf{r}_t^{\mathrm{T}} + \boldsymbol{\varepsilon}_t^{\mathrm{T}} , \tag{3.24}$$

where $\mathbf{B}_{0j} = \int_{\mathcal{U}\times\mathcal{V}} \boldsymbol{\psi}_j(u)\beta_{0j}(u,v)\boldsymbol{\phi}(v)^{\mathrm{T}}\,\mathrm{d}u\mathrm{d}v \in \mathbb{R}^{d_j\times\tilde{d}}$ and $\mathbf{r}_t = (r_{t1},\ldots,r_{t\tilde{d}})^{\mathrm{T}}$ is the truncation error with each $r_{tm} = \sum_{j=1}^{p}\sum_{l=d_j+1}^{\infty}\eta_{tjl}\langle\langle\psi_{jl},\beta_{0j}\rangle,\phi_m\rangle$ for $m \in [\tilde{d}]$. Let $\mathbf{B}_0 = (\mathbf{B}_{01}^{\mathrm{T}},\ldots,\mathbf{B}_{0p}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{\sum_{j=1}^{p} d_j\times\tilde{d}}$. We choose $\{\boldsymbol{\eta}_{(t+h)k} : h \in [L], k \in [p]\}$ as vector-valued instrumental variables, which are assumed to be uncorrelated with the random error $\boldsymbol{\varepsilon}_t$ in (3.24). Within the framework of (3.15), we assume that $\mathbf{B}_0$ is the unique solution to the following moment equations:

$$\mathbf{0} = \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\varepsilon}_t^{\mathrm{T}}\} = \mathbf{g}_{hk}(\mathbf{B}_0) + \mathbf{R}_{hk} , \quad h \in [L] , \ k \in [p] , \tag{3.25}$$

where $\mathbf{g}_{hk}(\mathbf{B}_0) = \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\zeta}_t^{\mathrm{T}}\} - \sum_{j=1}^{p}\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{tj}^{\mathrm{T}}\mathbf{B}_{0j}\}$ and $\mathbf{R}_{hk} = -\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\mathbf{r}_t^{\mathrm{T}}\}$. Given the recovery equivalence between functional sparsity in $\boldsymbol{\beta}_0$ and the block sparsity in $\mathbf{B}_0$, we aim to estimate the block sparse matrix $\mathbf{B}_0$ using the empirical versions $\mathbf{B} \mapsto \widehat{\mathbf{g}}_{hk}(\mathbf{B})$ for $h \in [L]$ and $k \in [p]$,

$$\widehat{\mathbf{g}}_{hk}(\mathbf{B}) = \frac{1}{n-h}\sum_{t=1}^{n-h}\widehat{\boldsymbol{\eta}}_{(t+h)k}\widehat{\boldsymbol{\zeta}}_t^{\mathrm{T}} - \frac{1}{n-h}\sum_{t=1}^{n-h}\sum_{j=1}^{p}\widehat{\boldsymbol{\eta}}_{(t+h)k}\widehat{\boldsymbol{\eta}}_{tj}^{\mathrm{T}}\mathbf{B}_j ,$$

where $\widehat{\boldsymbol{\zeta}}_t = (\hat{\zeta}_{t1},\ldots,\hat{\zeta}_{t\tilde{d}})^{\mathrm{T}}$ with $\hat{\zeta}_{tm} = \langle Y_t,\hat{\phi}_m\rangle$ for $m \in [\tilde{d}]$ and $\{\widehat{\boldsymbol{\eta}}_{tj}\}_{t\in[n],j\in[p]}$ are obtained in the first step. In the second step, according to (3.16), we formulate the block RMD estimator $\widehat{\mathbf{B}}$ by solving the convex optimisation problem below:

$$\widehat{\mathbf{B}} = \arg\min_{\mathbf{B}} \sum_{j=1}^{p}\|\mathbf{B}_j\|_{\mathrm{F}} \ \text{ subject to } \max_{k\in[p],h\in[L]}\|\widehat{\mathbf{g}}_{hk}(\mathbf{B})\|_{\mathrm{F}} \le \gamma_n ,$$

where $\gamma_n \ge 0$ is a regularisation parameter. In the third step, we estimate the coefficient functions by

$$\hat{\beta}_j(u,v) = \widehat{\boldsymbol{\psi}}_j(u)^{\mathrm{T}}\widehat{\mathbf{B}}_j\widehat{\boldsymbol{\phi}}(v) , \quad (u,v) \in \mathcal{U}\times\mathcal{V} , j \in [p] , \tag{3.26}$$

where $\{\widehat{\boldsymbol{\psi}}_j(u)\}_{j\in[p]}$ and $\widehat{\boldsymbol{\phi}}(v) = (\hat{\phi}_1(v),\ldots,\hat{\phi}_{\tilde{d}}(v))^{\mathrm{T}}$ are obtained in the first step.

In the following, we investigate the convergence property of $\{\hat{\beta}_j(\cdot,\cdot)\}_{j\in[p]}$ in (3.26). To simplify notation, we assume the same truncated dimension $d_j = d$ across $j \in [p]$. We first rewrite (3.25) in the form of (3.17) and apply Theorem 3.2 and Proposition 3.2 in the Appendix on $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{g}}(\mathbf{0})$ to verify Condition 3.4 with the choice of $\epsilon_{n1} \asymp \mathcal{M}_{W,Y} d^{\alpha\vee\tilde{\alpha}+2}\{\log(pd)/n\}^{1/2}$, where $\mathcal{M}_{W,Y}$ is specified in Proposition 3.2 in the Appendix. In a similar fashion to α, the parameter $\tilde{\alpha}$ as specified in Condition 3.14 in the Appendix determines the tightness of eigengaps of the covariance function of $\{Y_t(\cdot)\}$. We then impose the following smoothness condition on nonzero coefficient functions.

Condition 3.10. For each $j \in S$, $\beta_{0j}(u,v) = \sum_{l,m=1}^{\infty} a_{jlm}\psi_{jl}(u)\phi_m(v)$ and there exists some positive constant $\tau > \alpha \vee \tilde{\alpha} + 1/2$ such that $|a_{jlm}| \lesssim (l+m)^{-\tau-1/2}$ for $l, m \geq 1$.

Similar to Condition 3.8 in SFLR, Condition 3.10 ensures that smooth regression coefficients $\{\beta_{0j}(\cdot,\cdot) : j \in S\}$ in FFLR are expanded by its basis $\{\psi_{jl}(\cdot)\}_{l\geq 1}$ and $\{\phi_m(\cdot)\}_{m\geq 1}$, and the smoothness is determined by parameter τ. We are now ready to present the convergence rate of the FFLR estimate $\widehat{\boldsymbol{\beta}}(\cdot,\cdot) = \{\hat{\beta}_1(\cdot,\cdot),\ldots,\hat{\beta}_p(\cdot,\cdot)\}^{\mathrm{T}}$ under functional ℓ_1 norm in Theorem 3.5.

Theorem 3.5. *Suppose that Conditions* 3.1–3.3 *and Conditions* 3.13(i), 3.14 *in the Appendix hold for sub-Gaussian functional linear processes* $\{\mathbf{W}_t(\cdot)\}$ *and* $\{Y_t(\cdot)\}$, *and also Conditions* 3.9–3.10 *hold. Let* $d \asymp \tilde{d}$. *If the regularisation parameter* $\gamma_n \asymp s[d^{\alpha\vee\tilde{\alpha}+2}\mathcal{M}_{W,Y}\{\log(pd)/n\}^{1/2} + d^{-\tau+1/2}]$, *then the estimate* $\widehat{\boldsymbol{\beta}}(\cdot,\cdot)$ *satisfies*

$$\sum_{j=1}^{p} \|\hat{\beta}_j - \beta_{0j}\|_{\mathcal{S}} = O_{\mathrm{p}}\left\{\mu^{-2}s^2\left(d^{\alpha+\alpha\vee\tilde{\alpha}+2}\mathcal{M}_{W,Y}\sqrt{\frac{\log(pd)}{n}} + d^{\alpha-\tau+1/2}\right)\right\}. \quad (3.27)$$

Remark 3.4. With the same expression of $\mathbf{G}$ for both SFLR and FFLR, Condition 3.9 is required in both Theorems 3.4 and 3.5. Note we can further remove the assumption of $d \asymp \tilde{d}$, and establish the general convergence result in terms of $d, \tilde{d}$ and other parameters.

3.5.3 High-dimensional VFAR

The high-dimensional VFAR of a fixed lag order H, namely VFAR(H), takes the form of

$$\mathbf{X}_t(v) = \sum_{h'=1}^{H}\int_{\mathcal{U}} \mathbf{A}_0^{(h')}(u,v)\mathbf{X}_{t-h'}(u)\,\mathrm{d}u + \boldsymbol{\varepsilon}_t(v)\,, \quad t = H+1,\ldots,n\,, \quad (3.28)$$

where $\{\mathbf{X}_t(\cdot)\}$ satisfy model (3.1), the errors $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \ldots, \varepsilon_{tp})^{\mathrm{T}}$ are i.i.d. sampled from a p-dimensional vector of mean-zero functional processes, independent of $\mathbf{X}_{t-1}(\cdot), \mathbf{X}_{t-2}(\cdot), \ldots$, and $\mathbf{A}_0^{(h')} = \{A_{0,jj'}^{(h')}(\cdot,\cdot)\}_{j,j'\in[p]}$ is the unknown functional transition matrix at lag h'. In the special case $H = 1$ with $\mathbf{A}_0 = \mathbf{A}_0^{(1)}$, Theorem 3.1 of Bosq (2000) ensures the stationarity of $\{\mathbf{X}_t(\cdot)\}$ if there exists an integer l_0 such that $\sup_{\|\mathbf{f}\|\le 1} \|\mathbf{A}_0^{l_0}(\mathbf{f})\| < 1$ for $\mathbf{f} \in \mathbb{H}^p$. According to Guo and Qiao (2020), all VFAR(H) models can be reformulated as a VFAR(1) model and hence it is not hard to adjust the stationarity condition for the general case $H > 1$. To make a feasible fit to (3.28) under a high-dimensional regime based on observed curves $\{\mathbf{W}_t(\cdot)\}_{t\in[n]}$, we assume $\{\mathbf{A}_0^{(h')}\}_{h'\in[H]}$ is rowwise functional s-sparse with $s = \max_{j\in[p]} s_j \ll p$. To be specific, for the j-th row of components in $\{\mathbf{A}_0^{(h')}\}$, we denote the set of nonzero functions by $S_j = \{(j', h') \in [p] \times [H] : \|A_{0,jj'}^{(h')}\|_{\mathcal{S}} \neq 0\}$ and its cardinality by $s_j = |S_j|$ for $j \in [p]$. For each $j \in [p]$, we approximate $X_{tj}(\cdot)$ based on the expansion in (3.3) truncated at d_j. With some specific calculations, model (3.28) can be rowwisely rewritten as

$$\boldsymbol{\eta}_{tj}^{\mathrm{T}} = \sum_{h'=1}^{H}\sum_{j'=1}^{p} \boldsymbol{\eta}_{(t-h')j'}^{\mathrm{T}} \boldsymbol{\Omega}_{0,jj'}^{(h')} + \mathbf{r}_{tj}^{\mathrm{T}} + \boldsymbol{\varepsilon}_{tj}^{\mathrm{T}}, \quad j \in [p], \tag{3.29}$$

where $\boldsymbol{\Omega}_{0,jj'}^{(h')} = \int_{\mathcal{U}^2} \boldsymbol{\psi}_{j'}(u) A_{0,jj'}^{(h')}(u,v) \boldsymbol{\psi}_j(v)^{\mathrm{T}} \,\mathrm{d}u\mathrm{d}v \in \mathbb{R}^{d_{j'}\times d_j}$ and $\mathbf{r}_{tj} = (r_{tj1}, \ldots, r_{tjd_j})^{\mathrm{T}}$ is the truncation error with each $r_{tjm} = \sum_{h'=1}^{H}\sum_{j'=1}^{p}\sum_{l=d_{j'}+1}^{\infty} \eta_{(t-h')j'l} \langle\langle \psi_{j'l}, A_{0,jj'}^{(h')}\rangle, \psi_{jm}\rangle$ for $m \in [d_j]$. Let $\boldsymbol{\Omega}_{0j} = \{(\boldsymbol{\Omega}_{0,j1}^{(1)})^{\mathrm{T}}, \ldots, (\boldsymbol{\Omega}_{0,jp}^{(1)})^{\mathrm{T}}, \ldots, (\boldsymbol{\Omega}_{0,j1}^{(H)})^{\mathrm{T}}, \ldots, (\boldsymbol{\Omega}_{0,jp}^{(H)})^{\mathrm{T}}\}^{\mathrm{T}} \in \mathbb{R}^{H\sum_{j'=1}^{p} d_{j'}\times d_j}$. We choose $\{\boldsymbol{\eta}_{(t+h)k} : h \in [L], k \in [p]\}$ as vector-valued instrumental variables, which are assumed to be uncorrelated with the random error $\boldsymbol{\varepsilon}_{tj}$ in (3.29). Within the framework of (3.15), we assume that $\boldsymbol{\Omega}_{0j}$ is the unique solution to the following moment equations:

$$\mathbf{0} = \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\varepsilon}_{tj}^{\mathrm{T}}\} = \mathbf{g}_{j,hk}(\boldsymbol{\Omega}_{0j}) + \mathbf{R}_{j,hk}, \quad h \in [L], k \in [p], \tag{3.30}$$

where $\mathbf{g}_{j,hk}(\boldsymbol{\Omega}_{0j}) = \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{tj}^{\mathrm{T}}\} - \sum_{h'=1}^{H}\sum_{j'=1}^{p} \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{(t-h')j'}^{\mathrm{T}}\boldsymbol{\Omega}_{0,jj'}^{(h')}\}$ and $\mathbf{R}_{j,hk} = -\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\mathbf{r}_{tj}^{\mathrm{T}}\}$.

Given that estimating the functional sparsity in the j-th row of $\{\mathbf{A}_0^{(h')}\}_{h'\in[H]}$ is equivalent to estimating the block sparsity in $\boldsymbol{\Omega}_{0j}$ for each j, our goal is to estimate the block sparse matrix $\boldsymbol{\Omega}_{0j}$ using the empirical versions $\boldsymbol{\Omega}_j \mapsto \widehat{\mathbf{g}}_{j,hk}(\boldsymbol{\Omega}_j)$ for $h \in [L]$ and $k \in [p]$, where

$$\widehat{\mathbf{g}}_{j,hk}(\boldsymbol{\Omega}_j) = \frac{1}{n-h}\sum_{t=1}^{n-h} \widehat{\boldsymbol{\eta}}_{(t+h)k}\widehat{\boldsymbol{\eta}}_{tj}^{\mathrm{T}} - \frac{1}{n-h}\sum_{t=1}^{n-h}\sum_{h'=1}^{H}\sum_{j'=1}^{p} \widehat{\boldsymbol{\eta}}_{(t+h)k}\widehat{\boldsymbol{\eta}}_{(t-h')j'}^{\mathrm{T}}\boldsymbol{\Omega}_{jj'}^{(h')}$$

and $\{\widehat{\boldsymbol{\eta}}_{tj}\}_{t\in[n],j\in[p]}$ are obtained in the first step. The second step follows (3.16) to formulate the block RMD estimator $\widehat{\boldsymbol{\Omega}}_j$ by solving the following optimisation task:

$$\widehat{\boldsymbol{\Omega}}_j = \arg\min_{\boldsymbol{\Omega}_j} \sum_{h'=1}^{H}\sum_{j'=1}^{p} \|\boldsymbol{\Omega}_{jj'}^{(h')}\|_{\mathrm{F}} \ \text{ subject to } \max_{k\in[p],h\in[L]} \|\widehat{\mathbf{g}}_{j,hk}(\boldsymbol{\Omega}_j)\|_{\mathrm{F}} \le \gamma_{nj}\,,$$

where $\gamma_{nj} \ge 0$ is a regularisation parameter. The third step estimates functional transition matrices by

$$\hat{A}_{jj'}^{(h')}(u,v) = \widehat{\boldsymbol{\psi}}_{j'}(u)^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')}\widehat{\boldsymbol{\psi}}_{j}(v)\,, \quad (u,v)\in\mathcal{U}^2\,,\ j,j'\in[p]\,,\ h'\in[H]\,,$$

where $\{\widehat{\boldsymbol{\psi}}_j(\cdot)\}_{j\in[p]}$ are obtained in the first step.

We next present convergence analysis of $\{\hat{A}_{jj'}^{(h')}(\cdot,\cdot) : j,j'\in[p], h'\in[H]\}$. To simplify notation, we assume the same truncated dimension $d_j = d$ across $j\in[p]$. For each $j\in[p]$, we first express (3.30) in the form of

$$\mathbf{g}_j(\boldsymbol{\Omega}_{0j}) + \mathbf{R}_j = \mathbf{G}_j\boldsymbol{\Omega}_{0j} + \mathbf{g}_j(\mathbf{0}) + \mathbf{R}_j = \mathbf{0}\,,$$

where $\mathbf{g}_j = (\mathbf{g}_{j,11}^{\mathrm{T}},\ldots,\mathbf{g}_{j,1p}^{\mathrm{T}},\ldots,\mathbf{g}_{j,L1}^{\mathrm{T}},\ldots,\mathbf{g}_{j,Lp}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{R}_j = (\mathbf{R}_{j,11}^{\mathrm{T}},\ldots,\mathbf{R}_{j,1p}^{\mathrm{T}},\ldots,\mathbf{R}_{j,L1}^{\mathrm{T}},\ldots,\mathbf{R}_{j,Lp}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{G}_j = (\mathbf{G}_{j,ii'}) \in \mathbb{R}^{pLd\times pHd}$ whose (i,i')-th block is $\mathbf{G}_{j,ii'} = \mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{(t-h')j'}^{\mathrm{T}}\} \in \mathbb{R}^{d\times d}$ with $i = (h-1)p+k$, $k\in[p]$ for $h\in[L]$ and $i' = (h'-1)p+j'$, $j'\in[p]$ for $h'\in[H]$. Applying Theorem 3.2 on $\widehat{\mathbf{G}}_j$ and $\widehat{\mathbf{g}}_j(\mathbf{0})$, we can verify Condition 3.4 with the choice of $\epsilon_{n1} \asymp \mathcal{M}_1^W d^{\alpha+2}\{\log(pd)/n\}^{1/2}$. Similarly, we then give two regularity conditions.

Condition 3.11. For each $j\in[p]$ and $(j',h')\in S_j$, $A_{0,jj'}^{(h')}(u,v) = \sum_{l,m=1}^{\infty} a_{jj'lm}^{(h')}\psi_{j'm}(u)\psi_{jl}(v)$ and there exists some constant $\tau > \alpha + 1/2$ such that $|a_{jj'lm}^{(h')}| \lesssim (l+m)^{-\tau-1/2}$ for $l,m\ge 1$.

Condition 3.12. For each $j\in[p]$, let $\widetilde{\mathbf{G}}_j = (\widetilde{\mathbf{G}}_{j,ii'})$ be the normalised version of $\mathbf{G}_j = (\mathbf{G}_{j,ii'})$ by replacing each $\mathbf{G}_{j,ii'}$ with $\widetilde{\mathbf{G}}_{j,ii'} = \mathbb{E}\{\mathbf{D}_k\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{(t-h')j'}^{\mathrm{T}}\mathbf{D}_{j'}\}$ for $i = (h-1)p+k$ and $i' = (h'-1)p+j'$ with $k,j'\in[p]$, $h\in[L]$ and $h'\in[H]$, where $\mathbf{D}_j = \mathrm{diag}(\lambda_{j1}^{-1/2},\ldots,\lambda_{jd}^{-1/2})$. Then there exists an universal constant $\tilde{c}_j$ and $\mu_j > 0$ such that $\sigma_{\max}(m,\widetilde{\mathbf{G}}_j) \ge \tilde{c}_j$ and $\sigma_{\min}(m,\widetilde{\mathbf{G}}_j)/\sigma_{\max}(m,\widetilde{\mathbf{G}}_j) \ge \mu_j$ for $m = 16s_j/\mu_j^2$.

Following the spirit of Condition 3.8 and 3.10, Condition 3.11 determines the basis $\{\psi_{jl}(\cdot)\}_{j\in[p],l\ge1}$ on which the functional transition matrices are expanded and also controls the smoothness of the functional transition matrices by parameter τ, where a smaller (larger) τ implies smoother (rougher) coefficients. And Condition 3.12

extends Condition 3.9 to VFAR, where $\widetilde{\mathbf{G}}_j$ is the normalised version of $\mathbf{G}_j$, whose components vanish as d increase to infinity. Therefore, working with $\widetilde{\mathbf{G}}_j$ is preferred as argued in Remark 3.2.

We finally establish convergence rate of the VFAR estimate $\{\hat{A}_{jj'}^{(h')}\}_{j,j'\in[p],h'\in[H]}$ in the sense of functional matrix ℓ_∞ norm as follows.

Theorem 3.6. *Suppose that Conditions* 3.1–3.3 *hold for sub-Gaussian functional linear process* $\{\mathbf{W}_t(\cdot)\}$, *and Conditions* 3.11–3.12 *also hold. If regularisation parameters satisfy* $\gamma_{nj} \asymp s_j[d^{\alpha+2}\mathcal{M}_1^W\{\log(pd)/n\}^{1/2} + d^{-\tau+1/2}]$ *for* $j \in [p]$ *and* $\mu = \min_{j\in[p]}\mu_j$, *the estimate* $\{\hat{A}_{jj'}^{(h')}\}$ *satisfies*

$$\max_{j\in[p]}\sum_{j'=1}^{p}\sum_{h'=1}^{H}\|\hat{A}_{jj'}^{(h')} - A_{0,jj'}^{(h')}\|_{\mathcal{S}} = O_{\mathrm{p}}\left\{\mu^{-2}s^2\left(d^{2\alpha+2}\mathcal{M}_1^W\sqrt{\frac{\log(pd)}{n}} + d^{\alpha-\tau+1/2}\right)\right\}. \tag{3.31}$$

3.6 Empirical studies

3.6.1 Simulation study

In this section, we conduct a number of simulations to evaluate the finite-sample performance of the proposed autocovariance-based estimators for SFLR, FFLR and VFAR models.

In each simulated scenario, to mimic the infinite-dimensional nature of signal curves, we generate $X_{tj}(u) = \sum_{l=1}^{25}\eta_{tjl}\psi_l(u) = \boldsymbol{\eta}_{tj}^{\mathrm{T}}\boldsymbol{\psi}(u)$ with $\boldsymbol{\eta}_{tj} = (\eta_{tj1},\ldots,\eta_{tj25})^{\mathrm{T}}$ and $\boldsymbol{\psi}(\cdot) = \{\psi_1(\cdot),\ldots,\psi_{25}(\cdot)\}^{\mathrm{T}}$ for $t \in [n], j \in [p]$ and $u \in \mathcal{U} = [0,1]$, where $\{\psi_l(u)\}_{1\le l\le 25}$ is formed by 25-dimensional Fourier basis functions, 1, $\sqrt{2}\cos(2\pi lu)$, $\sqrt{2}\sin(2\pi lu)$ for $l = 1,\ldots,12$ and each $\boldsymbol{\eta}_t = (\boldsymbol{\eta}_{t1}^{\mathrm{T}},\ldots,\boldsymbol{\eta}_{tp}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{25p}$ is generated from a stationary vector autoregressive (VAR) model, $\boldsymbol{\eta}_t = \boldsymbol{\Omega}\boldsymbol{\eta}_{t-1} + \boldsymbol{\epsilon}_t$, with block transition matrix $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_{jk})_{j,k\in[p]} \in \mathbb{R}^{25p\times 25p}$ and $\boldsymbol{\epsilon}_t = (\boldsymbol{\epsilon}_{t1},\ldots,\boldsymbol{\epsilon}_{tp})^{\mathrm{T}}$, with $\boldsymbol{\epsilon}_{tj} \in \mathbb{R}^{25}$ whose components are independently sampled according to $\epsilon_{tjl} \sim \mathcal{N}(0, 0.7 - 0.1l)$ for $l = 1,\ldots,5$ and $\mathcal{N}(0, l^{-2})$ for $l = 6,\ldots,25$. Therefore, $\mathbf{X}_t(\cdot)$ follows a VFAR(1) model satisfying $\mathbf{X}_t(v) = \int_{\mathcal{U}}\mathbf{A}(u,v)\mathbf{X}_{t-1}(u)\,\mathrm{d}u + \boldsymbol{\varepsilon}_t(v)$, where $\boldsymbol{\varepsilon}_{tj}(v) = \boldsymbol{\psi}(v)^{\mathrm{T}}\boldsymbol{\epsilon}_{tj}$ and autocoefficient functions satisfy $A_{jk}(u,v) = \boldsymbol{\psi}(v)^{\mathrm{T}}\boldsymbol{\Omega}_{jk}\boldsymbol{\psi}(u)$ for $j,k \in [p]$ and $u,v \in \mathcal{U}$. In our simulations, we generate $n = 100, 200, 400$ serially dependent observations of $p = 40, 80$ functional variables. The observed curves are generated from $W_{tj}(u) = X_{tj}(u) + e_{tj}(u)$, where white noise curves $e_{tj}(u) = \sum_{l=1}^{5} z_{tjl}\psi_l(u)$ and

$\{(z_{tj1}, \ldots, z_{tj5})^{\mathrm{T}}\}_{t\in[n]}$ are independently sampled from multivariate normal distribution with mean zero and covariance $\mathrm{diag}(1, 0.8, 0.3, 1.5, 1.6)$. For each of the three models, the data is generated as follows.

VFAR: We generate block sparse $\mathbf{\Omega}$ with 5% or 10% nonzero blocks for $p = 80$ or $p = 40$, respectively. Specifically, for the j-th block row, we set the diagonal block $\mathbf{\Omega}_{jj} = \mathrm{diag}(0.60, 0.59, 0.58, 0.3, 0.2, 6^{-2}, \ldots, 25^{-2})$ and randomly choose one off-diagonal block being $0.4\mathbf{\Omega}_{jj}$ and two off-diagonal blocks being $0.1\mathbf{\Omega}_{jj}$. Such block sparse design on $\mathbf{\Omega}$ can guarantee the stationarity of the VFAR(1) process. It is worth noting that estimating VFAR(1) results in a very high-dimensional task, since, e.g. even under the most 'low-dimensional' setting with $p = 40, n = 400$ and truncated dimension $d = 3$, one needs to estimate $40^2 \times 3^2 = 14,400$ parameters based on only 400 observations. The p-dimensional functional covariates $\{\mathbf{X}_t(\cdot)\}_{t\in[n]}$ for SFLR and FFLR below are generated in the same way as those for VFAR.

SFLR: We generate the scalar responses $\{Y_t\}_{t\in[n]}$ from model (3.4), where ε_t's are independent $\mathcal{N}(0, 1)$ variables. For each $j \in S = \{1, \ldots, 5\}$, we generate $\beta_j(u) = \sum_{l=1}^{25} b_{jl}\psi_l(u)$ for $u \in \mathcal{U}$, where b_{j1}, b_{j2}, b_{j3} are sampled from the uniform distribution with support $[-1, -0.5] \cup [0.5, 1]$ and $b_{jl} = (-1)^l\, l^{-2}$ for $l = 4, \ldots, 25$. For $j \in [p] \setminus S$, we let $\beta_j(u) = 0$.

FFLR: We generate the functional responses $\{Y_t(v) : v \in \mathcal{V}\}_{t\in[n]}$ with $\mathcal{V} = [0, 1]$ from model (3.23), where $\varepsilon_t(v) = \sum_{m=1}^{5} g_{tm}\psi_m(v)$ with g_{tm}'s being independent $\mathcal{N}(0, 1)$ variables. For $j \in S$, we generate $\beta_j(u, v) = \sum_{l,m=1}^{25} b_{jml}\psi_l(u)\psi_m(v)$ for $(u, v) \in \mathcal{U} \times \mathcal{V}$, where components in $\{b_{jlm}\}_{1\le l,m\le 3}$ are sampled from the uniform distribution with support $[-1, -0.5] \cup [0.5, 1]$ and $b_{jlm} = (-1)^{l+m}(l + m)^{-2}$ for l or $m = 4, \ldots, 25$. For $j \in [p] \setminus S$, we let $\beta_j(u, v) = 0$.

Implementing our proposed autocovariance-based learning framework (AUTO) requires choosing L and d_j's. As our simulated results suggest that the estimators are not sensitive to the choice of L, we set $L = 3$ in simulations. To select d_j, we take the standard approach by selecting the largest d_j eigenvalues of $\widehat{K}_{jj}$ in (3.10) such that the cumulative percentage of selected eigenvalues exceeds 90%. To choose the regularisation parameter(s) for each model and comparison method, there are several possible methods one could adopt such as AIC, BIC and cross-validation. The BIC and AIC methods require the calculation of the effective degrees of freedom, which leads to a very challenging task given the high-dimensional, functional and dependent nature of the model structure and hence is left for future research. In our simulations, we generate a training sample of size n and a separate validation sample of the same size. Using the training data, we compute a series of estimators with 30 different values of the regularisation parameters, i.e. $\{\widehat{\mathbf{b}}_j^{(\gamma_n)}\}_{j\in[p]}$ (or $\{\widehat{\mathbf{B}}_j^{(\gamma_n)}\}_{j\in[p]}$) as a function of γ_n for SFLR (or FFLR) and $\{\widehat{\mathbf{\Omega}}_{jk}^{(\gamma_{nj})}\}_{k\in[p]}$ as a function of γ_{nj} for

VFAR, calculate the squared error between observed and fitted values on the validation set, i.e. $\sum_{t=1}^{n}[Y_t - \sum_{j=1}^{p}\{\widehat{\mathbf{b}}_j^{(\gamma_n)}\}^{\mathrm{T}}\widehat{\boldsymbol{\eta}}_{tj}]^2$ for SFLR, $\sum_{t=1}^{n}\|\widehat{\boldsymbol{\zeta}}_t - \sum_{j=1}^{p}(\widehat{\mathbf{B}}_j^{(\gamma_n)})^{\mathrm{T}}\widehat{\boldsymbol{\eta}}_{tj}\|^2$ for FFLR and $\sum_{t=1}^{n}\|\widehat{\boldsymbol{\eta}}_{tj} - \sum_{k=1}^{p}(\widehat{\boldsymbol{\Omega}}_{jk}^{(\gamma_{nj})})^{\mathrm{T}}\widehat{\boldsymbol{\eta}}_{(t-1)k}\|^2$ for VFAR, and choose the one with the smallest error.

We compare AUTO with the standard covariance-based estimation framework (COV), which proceeds in the following three steps. The first step performs FPCA on $\{W_{tj}(\cdot)\}_{t\in[n]}$ for each j, where the truncated dimension was selected in a similar way as d_j. Therefore, estimating SFLR and FFLR models are transformed into fitting multiple linear regressions with univariate response (Kong et al., 2016b) and multivariate response (Fang et al., 2020), respectively and the VFAR estimation is converted to the VAR estimation (Guo and Qiao, 2020). The second step considers minimising the covariance-based criterion, essentially the least squares with the addition of a group lasso type penalty. Such criterion can be optimised using an efficient block version of fast iterative shrinkage-thresholding algorithm developed in Guo and Qiao (2020), which converges faster than the commonly adopted block coordinate descent algorithm (Fan et al., 2015). The third step recovers functional sparse estimates using estimated eigenfunctions.

We examine the performance of COV and AUTO for three models in terms of the relative estimation accuracy, i.e. $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathrm{F}}/\|\mathbf{A}\|_{\mathrm{F}}$ for VFAR, $(\sum_{j=1}^{p}\|\hat{\beta}_j - \beta_{0j}\|^2)^{1/2}/(\sum_{j=1}^{p}\|\beta_{0j}\|^2)^{1/2}$ for SFLR and $(\sum_{j=1}^{p}\|\hat{\beta}_j - \beta_{0j}\|_{\mathcal{S}}^2)^{1/2}/(\sum_{j=1}^{p}\|\beta_{0j}\|_{\mathcal{S}}^2)^{1/2}$ for FFLR. We ran each simulation 100 times. Figure 3.1 displays boxplots of relative estimation errors for three models, while Table 3.2 in the Appendix gives numerical summaries. Several conclusions can be drawn from Figure 3.1. First, AUTO significantly outperforms COV for three models under all scenarios we consider. Second, as discussed in Section 3.2, AUTO provides consistent estimates, while the consistency of COV estimates is jeopardized by the white noise contamination. This can be demonstrated by our empirical results that AUTO provides more substantially improved estimates over COV as n increases from 100 to 400 especially for SFLR and FFLR. Third, the performance of AUTO slightly deteriorates as p increases from 40 to 80, providing empirical evidence to support that the rates in (3.22), (3.27) and (3.31) for SFLR, FFLR and VFAR models, respectively, all depend on the $(\log p)^{1/2}$ term.

3.6.2 Real data analysis

We further illustrate our developed methodology using a public financial dataset, which was obtained from the Wharton Research Data Services and consists of high-frequency observations of prices for S&P 100 index and component stocks (list

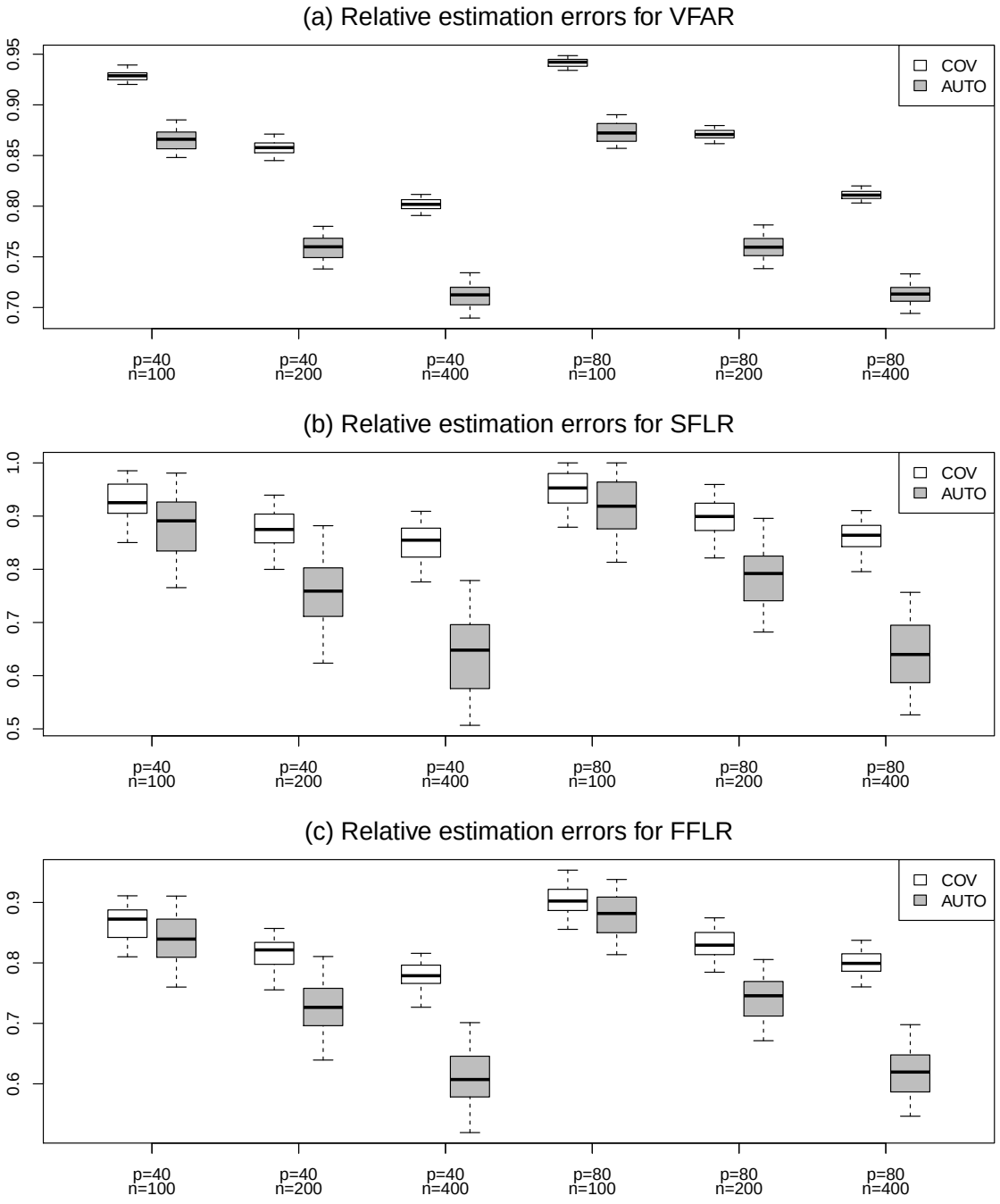


Figure 3.1: The boxplots of relative estimation errors for (a) VFAR, (b) SFLR and (c) FFLR.

available in Table 3.3 of the Appendix, we removed several stocks for which the data were not available so that $p = 98$ in our analysis) in year 2017 comprising 251 trading days. We obtain one-minute resolution prices by using the last transaction price in each one-minute interval after removing the outliers, and hence convert the trading period (9:30–16:00) to minutes $[0, 390]$. We construct cumulative intraday return (CIDR) trajectories (Horváth et al., 2014), in percentage, by $W_{tj}(u_k) = 100[\log\{P_{tj}(u_k)\} - \log\{P_{tj}(u_1)\}]$, where $P_{tj}(u_k)$ $(t \in [n], j \in [p], k \in [N])$ denotes the price of the j-th stock at the k-th minute after the opening time on the t-th trading day. We work with mildly smoothed CIDRs obtained by expanding the data with respect to a 45-dimensional B-spline basis. Such CIDR curves always start from zero and have nearly the same shape as the original price curves, but make the stationarity assumption more plausible. We performed the functional KPSS test (Horváth et al., 2014) on CIDR curves for each stock using the R package `"fsta"` (Shang, 2013). The p-values are greater than 1% for all the companies, indicating that these CIRDs are stationary.

Our interest is in predicting the intraday return of the S&P 100 index based on observed CIDR trajectories of component stocks, $W_{tj}(u), u \in \mathcal{U} = [0, N]$ up to time N, where, e.g. $N = 360$ corresponds to 30 minutes prior to the closing time of the trading day. With this in mind, we construct a sparse SFLR model with erroneous functional predictors as follows

$$Y_t = \sum_{j=1}^{p} \int_{\mathcal{U}} X_{tj}(u)\beta_{0j}(u)\,\mathrm{d}u + \varepsilon_t, \quad W_{tj}(u) = X_{tj}(u) + e_{tj}(u), \quad t \in [n],\, j \in [p]\,, \tag{3.32}$$

where Y_t is the intraday return of the S&P 100 index on the t-th trading day, $X_{tj}(\cdot)$ and $e_{tj}(\cdot)$ represent the signal and noise components in $W_{tj}(\cdot)$, respectively. We split the whole dataset into three subsets: training, validation and test sets consisting of the first 171, subsequent 40 and last 40 observations, respectively. We apply the validation set approach to select the regularisation parameters for AUTO and COV, based on which we estimate sparse functional coefficients in (3.32) and calculate the mean squared prediction errors (MSPEs) on the test set. For comparison, we also implement autocovariance-based generalised method-of-moments (AGMM) (Chen et al., 2020) and covariance-based least squares method (CLS) (Hall and Horowitz, 2007) to fit the unvariate version of (3.32) for each component stock, among which we choose the best models leading to the lowest test MSPEs. Finally, we include the null model, using the mean of the training response to predict the test response.

The resulting test MSPEs for different values of N and all comparison approaches are presented in Table 3.1. We observe a few apparent patterns. First, in all

Table 3.1: MSPEs up to different current times, $N = 300$, 315, 330, 345, 360, 370 and 380 minutes, for AUTO and four competing methods. All entries have been multiplied by 100 for formatting reasons. The lowest MSPE for each value of N is in bold font.

Method	$u \leq 300$	$u \leq 315$	$u \leq 330$	$u \leq 345$	$u \leq 360$	$u \leq 370$	$u \leq 380$
AUTO	**5.068**	**4.936**	**4.814**	**4.161**	**3.892**	**3.798**	**3.726**
COV	5.487	5.360	5.222	5.090	4.976	4.927	4.882
AGMM	6.506	6.470	6.454	6.441	6.408	6.385	6.364
CLS	6.859	6.798	6.730	6.655	6.583	6.546	6.507
Mean	8.832	8.832	8.832	8.832	8.832	8.832	8.832

scenarios we consider, AUTO provides the best predictive performance, while the autocovariance-based methods are superior to the covariance-based counterparts. Second, the predictive accuracy for functional regression type of methods improves as N approaches to 390 providing more recent information into the predictors. Third, AUTO and COV significantly outperform AGMM and CLS, while Mean gives the worst results. This indicates that using multiple selected functional predictors from the trading histories indeed improves the prediction results.

3.7 Appendix

Appendix 3.7.1 contains further non-asymptotic results. Additional empirical results are presented in Appendix 3.7.2. Technical proofs of main theoretical results, additional technical lemmas and their proofs are in Appendix 3.7.3.

3.7.1 Further non-asymptotic results

To provide the theoretical support for proposed estimators in Sections 3.5.1 and 3.5.2, we present essential non-asymptotic results for relevant estimated cross-(auto) covariance terms based on the functional cross-spectral stability measure (Fang et al., 2020) between $\{\mathbf{W}_t(\cdot)\}_{t\in\mathbb{Z}}$ and $\tilde{p}$-dimensional mean-zero functional time series (or scalar time series) $\{\mathbf{Y}_t(\cdot)\}_{t\in\mathbb{Z}}$ (or $\{\mathbf{Z}_t\}_{t\in\mathbb{Z}}$). Define $\mathbf{\Sigma}_h^{W,Y}(u,v) = \mathrm{Cov}\{\mathbf{W}_t(u), \mathbf{Y}_{t+h}(v)\}$ and $\mathbf{\Sigma}_h^{W,Z}(u) = \mathrm{Cov}\{\mathbf{W}_t(u), \mathbf{Z}_{t+h}\}$ for $h \in \mathbb{Z}$ and $(u,v) \in \mathcal{U} \times \mathcal{V}$.

Condition 3.13. (i) For $\{\mathbf{W}_t(\cdot)\}_{t\in\mathbb{Z}}$ and $\{\mathbf{Y}_t(\cdot)\}_{t\in\mathbb{Z}}$, the cross-spectral density function $\mathbf{f}_\theta^{W,Y} = (2\pi)^{-1}\sum_{h\in\mathbb{Z}}\mathbf{\Sigma}_h^{W,Y}e^{-ih\theta}$ for $\theta\in[-\pi,\pi]$ exists and the functional cross-spectral stability measure defined in (3.33) is finite, i.e.

$$\mathcal{M}^{W,Y} = 2\pi\cdot \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi],\mathbf{\Phi}_1\in\mathbb{H}_0^p,\mathbf{\Phi}_2\in\mathbb{H}_0^{\tilde{p}}} \frac{|\langle\mathbf{\Phi}_1,\mathbf{f}_\theta^{W,Y}(\mathbf{\Phi}_2)\rangle|}{\sqrt{\langle\mathbf{\Phi}_1,\mathbf{\Sigma}_0^W(\mathbf{\Phi}_1)\rangle}\sqrt{\langle\mathbf{\Phi}_2,\mathbf{\Sigma}_0^Y(\mathbf{\Phi}_2)\rangle}} < \infty\,, \quad (3.33)$$

where $\mathbb{H}_0^p = \{\mathbf{\Phi}\in\mathbb{H}^p : \langle\mathbf{\Phi},\mathbf{\Sigma}_0^W(\mathbf{\Phi})\rangle\in(0,\infty)\}$ and $\mathbb{H}_0^{\tilde{p}} = \{\mathbf{\Phi}\in\mathbb{H}^{\tilde{p}} : \langle\mathbf{\Phi},\mathbf{\Sigma}_0^Y(\mathbf{\Phi})\rangle\in(0,\infty)\}$.
(ii) For $\{\mathbf{W}_t(\cdot)\}_{t\in\mathbb{Z}}$ and $\{\mathbf{Z}_t\}_{t\in\mathbb{Z}}$, the cross-spectral density function $\mathbf{f}_\theta^{W,Z} = (2\pi)^{-1}\sum_{h\in\mathbb{Z}}\mathbf{\Sigma}_h^{W,Z}e^{-ih\theta}$ for $\theta\in[-\pi,\pi]$ exists and the functional cross-spectral stability measure defined in (3.34) is finite, i.e.

$$\mathcal{M}^{W,Z} = 2\pi\cdot \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi],\mathbf{\Phi}\in\mathbb{H}_0^p,\mathbf{v}\in\mathbb{R}_0^{\tilde{p}}} \frac{|\langle\mathbf{\Phi},\mathbf{f}_\theta^{W,Z}\mathbf{v}\rangle|}{\sqrt{\langle\mathbf{\Phi},\mathbf{\Sigma}_0^X(\mathbf{\Phi})\rangle}\sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}_0^Z\mathbf{v}}} < \infty\,, \quad (3.34)$$

where $\mathbb{R}_0^{\tilde{p}} = \{\boldsymbol{\nu}\in\mathbb{R}^{\tilde{p}} : \mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}_0^Z\mathbf{v}\in(0,\infty)\}$.

In analogy to (3.12), we can define the functional cross-spectral stability measure of all k_1-dimensional subsets of $\{\mathbf{W}_t(\cdot)\}$ and k_2-dimensional subsets of $\{\mathbf{Y}_t(\cdot)\}$ (or $\{\mathbf{Z}_t\}$) as $\mathcal{M}_{k_1,k_2}^{W,Y}$ (or $\mathcal{M}_{k_1,k_2}^{W,Z}$). It is easy to verify that $\mathcal{M}_{k_1,k_2}^{W,Y}\le\mathcal{M}^{W,Y}<\infty$ (or $\mathcal{M}_{k_1,k_2}^{W,Z}\le\mathcal{M}^{W,Z}<\infty$) for $k_1\in[p]$ and $k_2\in[\tilde{p}]$. For scalar time series $\{\mathbf{Z}_t\}$, the non-functional stability measure degenerates to

$$\mathcal{M}^Z = 2\pi\cdot \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi],\mathbf{v}\in\mathbb{R}_0^{\tilde{p}}} \frac{\mathbf{v}^{\mathrm{T}}\mathbf{f}_\theta^Z\mathbf{v}}{\mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}_0^Z\mathbf{v}}.$$

and the stability measure of all k-dimensional subsets of $\{\mathbf{Z}_t\}$, i.e. $\mathcal{M}_k^Z$ for $k\in[\tilde{p}]$, can be similarly defined according to (3.12).

For each $k\in[\tilde{p}]$, we represent $Y_{tk}(\cdot) = \sum_{m=1}^\infty\zeta_{tkm}\phi_{km}(\cdot)$ under the Karhunen-Loève expansion, where $\zeta_{tkm} = \langle Y_{tk},\phi_{km}\rangle$ and $\{(\theta_{km},\phi_{km})\}_{m\ge1}$ are pairs of eigenvalues and eigenfunctions of $\Sigma_{0,kk}^Y$. Let $\{(\hat{\theta}_{km},\hat{\phi}_{km})\}_{m\ge1}$ be estimated eigenpairs of $\widehat{\Sigma}_{0,kk}^Y$ and $\hat{\zeta}_{tkm} = \langle Y_{tk},\hat{\phi}_{km}\rangle$. We next impose a condition on eigenvalues $\{\theta_{km}\}_{m\ge1}$ and then develop the deviation bound in elementwise ℓ_∞-norm on how $\hat{\sigma}_{h,jklm}^{W,Y} = (n-h)^{-1}\sum_{t=1}^{n-h}\hat{\eta}_{tjl}\hat{\zeta}_{(t+h)km}$ concentrates around $\sigma_{h,jklm}^{W,Y} = \mathrm{Cov}\{\eta_{tjl},\zeta_{(t+h)km}\}$, which plays a crucial role in investigating the convergence property of the FFLR estimate in Section 3.5.2.

Condition 3.14. (i) For each $k \in [\tilde{p}]$, $\theta_{k1} > \theta_{k2} > \cdots > 0$, and there exist some positive constants $\tilde{c}$ and $\tilde{\alpha} > 1$ such that $\theta_{km} - \theta_{k(m+1)} \geq \tilde{c}m^{-\tilde{\alpha}-1}$ for $m \geq 1$; (ii) $\max_{k\in[\tilde{p}]} \sum_{m=1}^{\infty} \theta_{km} = O(1)$.

Proposition 3.2. *Suppose that Conditions* 3.1–3.3, 3.13(i) *and* 3.14 *hold for sub-Gaussian functional linear processes,* $\{\mathbf{W}_t(\cdot)\}$, $\{\mathbf{Y}_t(\cdot)\}$, *and* h *is fixed. Let* d *and* $\tilde{d}$ *be positive integers possibly depending on* $(n, p, \tilde{p})$ *and* $\mathcal{M}_{W,Y} = \mathcal{M}_1^W + \mathcal{M}_1^Y + \mathcal{M}_{1,1}^{W,Y}$. *If* $n \gtrsim (d^{2\alpha+2} \vee \tilde{d}^{2\tilde{\alpha}+2})(\mathcal{M}_{W,Y})^2 \log(p\tilde{p}d\tilde{d})$, *then there exist some positive constants* c_7 *and* c_8 *independent of* $(n, p, \tilde{p}, d, \tilde{d})$ *such that*

$$\max_{j\in[p],k\in[\tilde{p}],l\in[d],m\in[\tilde{d}]} \frac{|\hat{\sigma}_{h,jklm}^{W,Y} - \sigma_{h,jklm}^{W,Y}|}{l^{\alpha+1} \vee m^{\tilde{\alpha}+1}} \lesssim \mathcal{M}_{W,Y} \sqrt{\frac{\log(p\tilde{p}d\tilde{d})}{n}} \tag{3.35}$$

holds with probability greater than $1 - c_7(p\tilde{p}d\tilde{d})^{-c_8}$.

We next consider a mixed process scenario consisting of $\{\mathbf{W}_t(\cdot)\}$ and $\{\mathbf{Z}_t\}$ and establish the deviation bound in elementwise ℓ_∞-norm on how $\hat{\varrho}_{h,jkl}^{X,Z} = (n-h)^{-1} \sum_{t=1}^{n-h} \hat{\eta}_{tjl} Z_{(t+h)k}$ concentrates around $\varrho_{h,jkl}^{X,Z} = \mathrm{Cov}\{\eta_{tjl}, Z_{(t+h)k}\}$, which is essential in the convergence analysis of the SFLR estimate in Section 3.5.1.

Proposition 3.3. *Suppose that Conditions* 3.1–3.3 *and* 3.13(ii) *hold for sub-Gaussian functional linear process* $\{\mathbf{W}_t(\cdot)\}$, *sub-Gaussian linear process* $\{\mathbf{Z}_t\}$ *and* h *is fixed. Let* d *be a positive integer possibly depending on* $(n, p, \tilde{p})$ *and* $\mathcal{M}_{W,Z} = \mathcal{M}_1^W + \mathcal{M}_1^Z + \mathcal{M}_{1,1}^{W,Z}$. *If* $n \gtrsim (\mathcal{M}_{W,Z})^2 \log(p\tilde{p}d)$, *then there exist some positive constants* c_9 *and* c_{10} *independent of* $(n, p, \tilde{p}, d)$ *such that*

$$\max_{j\in[p],k\in[\tilde{p}],l\in[d]} \frac{|\hat{\varrho}_{h,jkl}^{W,Z} - \varrho_{h,jkl}^{W,Z}|}{l^{\alpha+1}} \lesssim \mathcal{M}_{W,Z} \sqrt{\frac{\log(p\tilde{p}d)}{n}}, \tag{3.36}$$

holds with probability greater than $1 - c_9(p\tilde{p}d)^{-c_{10}}$.

3.7.2 Additional simulation results

Table 3.2 reports numerical summaries of relative errors for VFAR, SFLR and FFLR. Table 3.3 presents the list of S&P 100 component stocks used in Section 3.6.2.

3.7.3 Proofs

Throughout, we use $c, c_1, c_2, \ldots$ to denote positive constants.

Table 3.2: The mean and standard error (in parentheses) of relative estimation errors over 100 simulation runs.

Model	p	40			80		
	n	100	200	400	100	200	400
VFAR	COV	0.928(0.005)	0.858(0.006)	0.802(0.005)	0.942(0.004)	0.871(0.005)	0.811(0.004)
	AUTO	0.865(0.010)	0.759(0.012)	0.712(0.011)	0.873(0.010)	0.759(0.011)	0.713(0.010)
SFLR	COV	0.927(0.035)	0.874(0.035)	0.852(0.034)	0.950(0.033)	0.897(0.035)	0.863(0.027)
	AUTO	0.883(0.058)	0.757(0.061)	0.639(0.073)	0.917(0.050)	0.785(0.056)	0.642(0.065)
FFLR	COV	0.866(0.029)	0.816(0.024)	0.777(0.022)	0.904(0.024)	0.831(0.022)	0.801(0.020)
	AUTO	0.840(0.040)	0.728(0.044)	0.611(0.047)	0.879(0.034)	0.742(0.036)	0.617(0.039)

Proof of Theorem 3.1

Applying similar techniques to prove Theorem 1 of Fang et al. (2020) and Proposition 1 of Guo and Qiao (2020), we obtain that for $h \geq 1$

$$\mathbb{P}\bigg\{\bigg|\frac{\langle \boldsymbol{\Phi}_1, (\widehat{\boldsymbol{\Sigma}}_h^W - \boldsymbol{\Sigma}_h^W)(\boldsymbol{\Phi}_2)\rangle}{\langle \boldsymbol{\Phi}_1, \boldsymbol{\Sigma}_0^W(\boldsymbol{\Phi}_1)\rangle + \langle \boldsymbol{\Phi}_2, \boldsymbol{\Sigma}_0^W(\boldsymbol{\Phi}_2)\rangle}\bigg| > 2\mathcal{M}_k^W\delta\bigg\} \leq 8\exp\big\{-c_1 n\min(\delta^2,\delta)\big\}. \tag{3.37}$$

For each $j \in [p]$, consider the spectral decomposition $\Sigma_{0,jj}^W(u,v) = \sum_{l=1}^{\infty}\omega_{jl}^W\nu_{jl}^W(u)\nu_{jl}^W(v)$ and $\omega_0 = \max_j\sum_{l=1}^{\infty}\omega_{jl}^W = O(1)$, implied from Lemma 3.4 in Appendix. For each (j,k,l,m), choosing $\boldsymbol{\Phi}_1 = \{0,\ldots,0,(\omega_{jl}^W)^{-1/2}\nu_{jl}^W,0,\ldots,0\}^{\mathrm{T}}$ and $\boldsymbol{\Phi}_2 = \{0,\ldots,0,(\omega_{km}^W)^{-1/2}\nu_{km}^W,0,\ldots,0\}^{\mathrm{T}}$ on (3.37) and following the same developments to prove Theorem 2 of Guo and Qiao (2020) with the choice of suitable constant c_2, we can obtain that

$$\mathbb{P}\big\{\|\widehat{\Sigma}_{h,jk}^W - \Sigma_{h,jk}^W\|_{\mathcal{S}} > \mathcal{M}_1^W\delta\big\} \leq 8\exp\big\{-c_2 n\min(\delta^2,\delta)\big\}. \tag{3.38}$$

It follows from (3.9), (3.10) and Cauchy–Schwartz inequality that

$$\|\widehat{K}_j - K_j\|_{\mathcal{S}}^2 \leq 2L\sum_{h=1}^{L}\|\widehat{\Sigma}_{h,jj}^W - \Sigma_{h,jj}^W\|_{\mathcal{S}}^2\|\Sigma_{h,jj}^W\|_{\mathcal{S}}^2 + L\sum_{h=1}^{L}\|\widehat{\Sigma}_{h,jj}^W - \Sigma_{h,jj}^W\|_{\mathcal{S}}^4.$$

Let $\Omega_{\omega,jk}^{(h)} = \{\|\widehat{\Sigma}_{h,jk}^W - \Sigma_{h,jk}^W\|_{\mathcal{S}} \leq \omega_0\}$ and $\Omega_{jk}^{(h)} = \{\|\widehat{\Sigma}_{h,jk}^W - \Sigma_{h,jk}^W\|_{\mathcal{S}} \leq \mathcal{M}_1^W\delta\}$. On the event $\Lambda_j = \Omega_{\omega,jj}^{(1)} \cap \cdots \cap \Omega_{\omega,jj}^{(L)} \cap \Omega_{jj}^{(1)} \cap \cdots \cap \Omega_{jj}^{(L)}$, it follows from the above results and Lemma 3.5 that

$$\|\widehat{K}_j - K_j\|_{\mathcal{S}} \leq \sqrt{3}L\omega_0\mathcal{M}_1^W\delta. \tag{3.39}$$

Table 3.3: List of S&P 100 stocks.

Ticker	Company name	Ticker	Company name
AAPL	APPLE INC	JPM	JPMORGAN CHASE & CO
ABBV	ABBVIE INC	KHC	KRAFT HEINZ
ABT	ABBOTT LABORATORIES	KMI	KINDER MORGAN INC
ACN	ACCENTURE PLC CLASS A	KO	COCA-COLA
AGN	ALLERGAN	LLY	ELI LILLY
AIG	AMERICAN INTERNATIONAL GROUP INC	LMT	LOCKHEED MARTIN CORP
ALL	ALLSTATE CORP	LOW	LOWES COMPANIES INC
AMGN	AMGEN INC	MA	MASTERCARD INC CLASS A
AMZN	AMAZON COM INC	MCD	MCDONALDS CORP
AXP	AMERICAN EXPRESS	MDLZ	MONDELEZ INTERNATIONAL INC CLASS A
BA	BOEING	MDT	MEDTRONIC PLC
BAC	BANK OF AMERICA CORP	MET	METLIFE INC
BIIB	BIOGEN INC	MMM	3M
BK	BANK OF NEW YORK MELLON CORP	MO	ALTRIA GROUP INC
BLK	BLACKROCK INC	MON	MONSANTO
BMY	BRISTOL MYERS SQUIBB	MRK	MERCK & CO INC
C	CITIGROUP INC	MS	MORGAN STANLEY
CAT	CATERPILLAR INC	MSFT	MICROSOFT CORP
CELG	CELGENE CORP	NEE	NEXTERA ENERGY INC
CHTR	CHARTER COMMUNICATIONS INC CLASS A	NKE	NIKE INC CLASS B
CL	COLGATE-PALMOLIVE	ORCL	ORACLE CORP
COF	CAPITAL ONE FINANCIAL CORP	OXY	OCCIDENTAL PETROLEUM CORP
COP	CONOCOPHILLIPS	PCLN	THE PRICELINE GROUP INC
COST	COSTCO WHOLESALE CORP	PEP	PEPSICO INC
CSCO	CISCO SYSTEMS INC	PFE	PFIZER INC
CVS	CVS HEALTH CORP	PG	PROCTER & GAMBLE
CVX	CHEVRON CORP	PM	PHILIP MORRIS INTERNATIONAL INC
DHR	DANAHER CORP	PYPL	PAYPAL HOLDINGS INC
DIS	WALT DISNEY	QCOM	QUALCOMM INC
DUK	DUKE ENERGY CORP	RTN	RAYTHEON
EMR	EMERSON ELECTRIC	SBUX	STARBUCKS CORP
EXC	EXELON CORP	SLB	SCHLUMBERGER NV
F	F MOTOR	SO	SOUTHERN
FB	FACEBOOK CLASS A INC	SPG	SIMON PROPERTY GROUP REIT INC
FDX	FEDEX CORP	T	AT&T INC
FOX	TWENTY-FIRST CENTURY FOX INC CLASS B	TGT	TARGET CORP
FOXA	TWENTY-FIRST CENTURY FOX INC CLASS A	TWX	TIME WARNER INC
GD	GENERAL DYNAMICS CORP	TXN	TEXAS INSTRUMENT INC
GE	GENERAL ELECTRIC	UNH	UNITEDHEALTH GROUP INC
GILD	GILEAD SCIENCES INC	UNP	UNION PACIFIC CORP
GM	GENERAL MOTORS	UPS	UNITED PARCEL SERVICE INC CLASS B
GOOG	ALPHABET INC CLASS C	USB	US BANCORP
GS	GOLDMAN SACHS GROUP INC	UTX	UNITED TECHNOLOGIES CORP
HAL	HALLIBURTON	V	VISA INC CLASS A
HD	HOME DEPOT INC	VZ	VERIZON COMMUNICATIONS INC
HON	HONEYWELL INTERNATIONAL INC	WBA	WALGREEN BOOTS ALLIANCE INC
IBM	INTERNATIONAL BUSINESS MACHINES CO	WFC	WELLS FARGO
INTC	INTEL CORPORATION CORP	WMT	WALMART STORES INC
JNJ	JOHNSON & JOHNSON	XOM	EXXON MOBIL CORP

Applying (3.38) and choosing $\delta = (\mathcal{M}_1^W)^{-1}\omega_0$ for $\Omega_{\omega,jj}^{(1)}, \ldots, \Omega_{\omega,jj}^{(L)}$ yields

$$\mathbb{P}(\Lambda_j^C) \le 8L \exp\big\{ - c_2 n \min(\delta^2, \delta)\big\} + 8L \exp\big[- c_2 n \min\{(\mathcal{M}_1^W)^{-2}\omega_0^2, (\mathcal{M}_1^W)^{-1}\omega_0\}\big] .$$

Combing the above results and choosing suitable constants c_3, c_4, we obtain

$$\mathbb{P}\big(\|\widehat{K}_j - K_j\|_{\mathcal{S}} > \mathcal{M}_1^W \delta\big) \le c_4 \exp\big\{ - c_3 n \min(\delta^2, \delta)\big\} + c_4 \exp(-c_3 n) . \tag{3.40}$$

For each $j \in [p]$, it follows from Lemma 4.3 of Bosq (2000) and Condition 3.3 that

$$|\hat{\lambda}_{jl} - \lambda_{jl}| \le \|\widehat{K}_j - K_j\|_{\mathcal{S}} \ \text{ and } \ \|\hat{\psi}_{jl} - \psi_{jl}\| \le 2\sqrt{2} c^{-1} l^{\alpha+1} \|\widehat{K}_j - K_j\|_{\mathcal{S}} . \tag{3.41}$$

Combining (3.40), (3.41) and the union bound of probability yields that

$$\begin{aligned}
&\mathbb{P}\Big(\max_{j\in[p], l\in[d]} |\hat{\lambda}_{jl} - \lambda_{jl}| > \mathcal{M}_1^W \delta\Big) \vee \mathbb{P}\Big\{ \max_{j\in[p], l\in[d]} (\|\hat{\psi}_{jl} - \psi_{jl}\|/l^{\alpha+1}) > 2\sqrt{2} c^{-1} \mathcal{M}_1^W \delta\Big\} \\
&\qquad \le c_4 pd \exp\big\{ - c_3 n \min(\delta^2, \delta)\big\} + c_4 pd \exp(-c_3 n) .
\end{aligned}$$

Let $\delta = \rho\sqrt{\log(pd)/n} \le 1$. Choosing suitable positive constants c_5 and $c_6 = 1 - c_3\rho^2$, we obtain that (3.13) holds with probability greater than $1 - c_5(pd)^{-c_6}$, which completes the proof of Theorem 3.1. □

Proof of Theorem 3.2

For each (j, k, l, m) and $h \ge 1$, we write

$$\begin{aligned}
\hat{\sigma}_{jklm}^{(h)} - \sigma_{jklm}^{(h)} &= \langle \hat{\psi}_{jl}, \widehat{\Sigma}_{h,jk}^W(\widehat{\psi}_{km})\rangle - \langle \psi_{jl}, \Sigma_{h,jk}^W(\psi_{km})\rangle \\
&= \langle (\hat{\psi}_{jl} - \psi_{jl}), \widehat{\Sigma}_{h,jk}^W(\hat{\psi}_{km} - \psi_{km})\rangle + \langle \psi_{jl}, (\widehat{\Sigma}_{h,jk}^W - \Sigma_{h,jk}^W)(\psi_{km})\rangle \\
&\quad + \big\{ \langle (\hat{\psi}_{jl} - \psi_{jl}), (\widehat{\Sigma}_{h,jk}^W - \Sigma_{h,jk}^W)(\psi_{km})\rangle + \\
&\qquad \langle \psi_{jl}, (\widehat{\Sigma}_{h,jk}^W - \Sigma_{h,jk}^W)(\hat{\psi}_{km} - \psi_{km})\rangle \big\} \\
&\quad + \big\{ \langle (\hat{\psi}_{jl} - \psi_{jl}), \Sigma_{h,jk}^W(\psi_{km})\rangle + \langle \psi_{jl}, \Sigma_{h,jk}^W(\hat{\psi}_{km} - \psi_{km})\rangle \big\} \\
&= J_1 + J_2 + J_3 + J_4 .
\end{aligned}$$

On the event $\widetilde{\Omega}_{jk} = \Omega_{\omega,jk}^{(h)} \cap \Omega_{jk}^{(h)} \cap \Lambda_j \cap \Lambda_k$, it follows from Lemma 3.5, (3.39), (3.41), the orthonormality of $\{\psi_{jl}\}, \{\psi_{km}\}$ that $|J_1| \lesssim (l \vee m)^{2(\alpha+1)} (\mathcal{M}_1^W)^2 \delta^2$, $|J_2| \le \mathcal{M}_1^W \delta$,

$|J_3| \lesssim (l \vee m)^{\alpha+1} \mathcal{M}_1^W \delta$ and $|J_4| \lesssim (l \vee m)^{\alpha+1} \mathcal{M}_1^W \delta$. Combing the above results implies that

$$\sum_{i=1}^{4} |J_i| \leq c_7 (l \vee m)^{\alpha+1} \mathcal{M}_1^W \delta + c_8 (l \vee m)^{2(\alpha+1)} (\mathcal{M}_1^W)^2 \delta^2 \,.$$

Applying (3.38) and choosing $\delta = (\mathcal{M}_1^W)^{-1} \omega_0$ for $\Omega_{\omega,jk}^{(h)}$, $\Omega_{\omega,jj}^{(1)}$, $\ldots$, $\Omega_{\omega,jj}^{(L)}$, $\Omega_{\omega,kk}^{(1)}$, $\ldots$, $\Omega_{\omega,kk}^{(L)}$ yields

$$\begin{aligned} \mathbb{P}(\widetilde{\Omega}_{jk}^c) \leq\; & (16L+8) \exp\big\{ - c_2 n \min(\delta^2, \delta) \big\} \\ & + (16L+8) \exp\big[- c_2 n \min\{ (\mathcal{M}_1^W)^{-2} \omega_0^2, (\mathcal{M}_1^W)^{-1} \omega_0 \} \big] \,. \end{aligned}$$

Combing the above results, choosing suitable positive constants c_9, c_{10}, c_{11}, and applying the union bound of probability yields

$$\begin{aligned} & \mathbb{P}\bigg\{ \max_{j,k \in [p], l,m \in [d]} \bigg| \frac{\hat{\sigma}_{jklm}^{(h)} - \sigma_{jklm}^{(h)}}{(l \vee m)^{\alpha+1}} \bigg| > \mathcal{M}_1^W \delta + c_{11} (l \vee m)^{\alpha+1} (\mathcal{M}_1^W)^2 \delta^2 \bigg\} \\ & \qquad \leq c_{10} p^2 d^2 \exp\big\{ - c_9 n \min(\delta^2, \delta) \big\} + c_{10} p^2 d^2 \exp(-c_9 n) \,. \end{aligned} \tag{3.42}$$

Choosing $\delta = \rho_1 \sqrt{\log(pd)/n} \leq 1$ and $1 + c_{11} d^{\alpha+1} \mathcal{M}_1^W \delta \leq \rho_2$ for some positive constants ρ_1, ρ_2, which can be achieved for sufficiently large $n \gtrsim d^{2\alpha+2} (\mathcal{M}_1^W)^2 \log(pd)$, it follows from (3.42) that there exists positive constants c_{12}, c_{13} such that, with probability greater than $1 - c_{12}(pd)^{-c_{13}}$,

$$\max_{j,k \in [p], l,m \in [d]} \bigg| \frac{\hat{\sigma}_{jklm}^{(h)} - \sigma_{jklm}^{(h)}}{(l \vee m)^{\alpha+1}} \bigg| \leq \rho_1 \rho_2 \mathcal{M}_1^W \sqrt{\frac{\log(pd)}{n}} \,,$$

which completes the proof of Theorem 3.2. □

Proof of Proposition 3.2

For each (h, j, k, l, m), we write

$$\begin{aligned}
&\hat{\sigma}^{W,Y}_{h,jklm} - \sigma^{W,Y}_{h,jklm} \\
= \;& \big\langle (\hat{\psi}_{jl} - \psi_{jl}), \widehat{\Sigma}^{W,Y}_{h,jk}(\hat{\phi}_{km} - \phi_{km}) \big\rangle + \big\langle \psi_{jl}, (\widehat{\Sigma}^{W,Y}_{h,jk} - \Sigma^{W,Y}_{h,jk})(\phi_{km}) \big\rangle \\
& + \Big\{ \big\langle (\hat{\psi}_{jl} - \psi_{jl}), (\widehat{\Sigma}^{W,Y}_{h,jk} - \Sigma^{W,Y}_{h,jk})(\phi_{km}) \big\rangle + \big\langle \psi_{jl}, (\widehat{\Sigma}^{W,Y}_{h,jk} - \Sigma^{W,Y}_{h,jk})(\hat{\phi}_{km} - \phi_{km}) \big\rangle \Big\} \\
& + \Big\{ \big\langle (\hat{\psi}_{jl} - \psi_{jl}), \Sigma^{W,Y}_{h,jk}(\phi_{km}) \big\rangle + \big\langle \psi_{jl}, \Sigma^{W,Y}_{h,jk}(\hat{\phi}_{km} - \phi_{km}) \big\rangle \Big\} \\
= \;& I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

Let $\Omega^Y_{0kk} = \{\|\widehat{\Sigma}^Y_{0,kk} - \Sigma^Y_{0,kk}\|_{\mathcal{S}} \le \mathcal{M}^Y_1 \delta\}$, $\Omega^{W,Y}_{hjk} = \{\|\widehat{\Sigma}^{W,Y}_{h,jk} - \Sigma^{W,Y}_{h,jk}\|_{\mathcal{S}} \le \mathcal{M}_{W,Y}\delta\}$. On the event $\Lambda_j \cap \Omega^Y_{0,kk} \cap \Omega^{W,Y}_{h,jk}$, it follows from $\|\langle \Sigma^{W,Y}_{h,jk}, \phi_{km} \rangle\| \le \omega_0^{1/2}\theta_{km}^{1/2}$ and $\|\langle \psi_{jl}, \Sigma^{W,Y}_{h,jk} \rangle\| \le \omega_0^{1/2}\theta_0^{1/2}$, derived by the similar techniques to prove Lemma 3.5, together with Lemma 3.4, Lemma 4.3 of Bosq (2000), (3.39), (3.41), the orthonormality of $\{\psi_{jl}\}, \{\phi_{km}\}$ and Condition 3.14 that

$$\begin{aligned}
|I_1| &\lesssim l^{\alpha+1}\mathcal{M}^W_1\delta m^{\tilde{\alpha}+1}\mathcal{M}^Y_1\delta \lesssim l^{2(\alpha+1)}(\mathcal{M}^W_1)^2\delta^2 + m^{2(\tilde{\alpha}+1)}(\mathcal{M}^Y_1)^2\delta^2, \\
|I_2| &\le \mathcal{M}_{W,Y}\delta, \\
|I_3| &\lesssim l^{\alpha+1}\mathcal{M}^W_1\mathcal{M}_{W,Y}\delta^2 + m^{\tilde{\alpha}+1}\mathcal{M}^Y_1\mathcal{M}_{W,Y}\delta^2, \\
|I_4| &\lesssim l^{\alpha+1}\mathcal{M}^W_1\delta + m^{\tilde{\alpha}+1}\mathcal{M}^Y_1\delta
\end{aligned}$$

Combing the above results and $\mathcal{M}_{W,Y} = \mathcal{M}^W_1 + \mathcal{M}^Y_1 + \mathcal{M}^{W,Y}_{1,1}$ yields that

$$\sum_{i=1}^4 |I_i| \le c_{14}(l^{\alpha+1} \vee m^{\tilde{\alpha}+1})\mathcal{M}_{W,Y}\delta + c_{15}(l^{2(\alpha+1)} \vee m^{2(\tilde{\alpha}+1)})(\mathcal{M}_{W,Y})^2\delta^2.$$

Following the same developments to prove (3.42), we apply (3.40), Theorem 2, Lemma 24 of Fang et al. (2020) and the union bound of probability, choose suitable positive constants c_{16}, c_{17}, c_{18} and hence obtain that

$$\begin{aligned}
&\mathbb{P}\left\{ \max_{j\in[p],k\in[\tilde{p}],l\in[d],m\in[\tilde{d}]} \frac{|\hat{\sigma}^{W,Y}_{h,jklm} - \sigma^{W,Y}_{h,jklm}|}{l^{\alpha+1} \vee m^{\tilde{\alpha}+1}} > \mathcal{M}_{W,Y}\delta + c_{18}(l^{\alpha+1} \vee m^{\tilde{\alpha}+1})(\mathcal{M}_{W,Y})^2\delta^2 \right\} \\
&\le c_{17}p\tilde{p}d\tilde{d}\exp\Big\{ - c_{16}n\min(\delta^2, \delta)\Big\} + c_{17}p\tilde{p}d\tilde{d}\exp\big(- c_{16}n\big).
\end{aligned} \tag{3.43}$$

Choosing $\delta = \rho_3\sqrt{\log(p\tilde{p}d\tilde{d})/n} \le 1$ and $1 + c_{19}(d^{\alpha+1} \vee \tilde{d}^{\tilde{\alpha}+1})\mathcal{M}_{W,Y}\delta \le \rho_4$ for some positive constants ρ_3, ρ_4, which can be achieved for sufficiently large $n \gtrsim (d^{2\alpha+2} \vee \tilde{d}^{2\tilde{\alpha}+2})(\mathcal{M}_{W,Y})^2\log(p\tilde{p}d\tilde{d})$, it follows from (3.43) that there exists positive constants

c_{20}, c_{21} such that, with probability greater than $1 - c_{20}(p\tilde{p}d\tilde{d})^{-c_{21}}$,

$$\max_{j\in[p],k\in[\tilde{p}],l\in[d],m\in[\tilde{d}]} \frac{|\hat{\sigma}^{W,Y}_{h,jklm} - \sigma^{W,Y}_{h,jklm}|}{l^{\alpha+1} \vee m^{\tilde{\alpha}+1}} \leq \rho_3\rho_4 \mathcal{M}_{W,Y} \sqrt{\frac{\log(p\tilde{p}d\tilde{d})}{n}},$$

which completes the proof of Proposition 3.2. □

Proof of Proposition 3.3

For each (h, j, k, l), we write

$$\begin{aligned} & \hat{\varrho}^{W,Z}_{h,jkl} - \varrho^{W,Z}_{h,jkl} \\ = \; & \big\langle (\hat{\psi}_{jl} - \psi_{jl}), (\hat{\Sigma}^{W,Z}_{h,jk} - \Sigma^{W,Z}_{h,jk}) \big\rangle + \big\langle \psi_{jl}, (\widehat{\Sigma}^{W,Z}_{h,jk} - \Sigma^{W,Z}_{h,jk}) \big\rangle + \big\langle (\hat{\psi}_{jl} - \psi_{jl}), \Sigma^{W,Z}_{h,jk} \big\rangle \\ = \; & T_1 + T_2 + T_3. \end{aligned}$$

Let $\Omega^{W,Z}_{hjk} = \big\{ \|\widehat{\Sigma}^{W,Z}_{h,jk} - \Sigma^{W,Z}_{h,jk}\|_{\mathcal{S}} \leq \mathcal{M}_{W,Z}\delta \big\}$. On the event $\Lambda_j \cap \Omega^{W,Z}_{hjk}$, it follows from (3.39), (3.41), the orthonormality of $\{\psi_{jl}\}$ and $\|\Sigma^{WZ}_{h,jk}\| \leq \omega_0^{1/2}\sigma^Z_{0,kk}$ that

$$\begin{aligned} |T_1| &\lesssim l^{\alpha+1}\mathcal{M}^W_1\delta\mathcal{M}_{W,Z}\delta, \\ |T_2| &\leq \mathcal{M}_{W,Z}\delta, \\ |T_3| &\lesssim l^{\alpha+1}\mathcal{M}^W_1\delta. \end{aligned}$$

Combing the above results and $\mathcal{M}_{W,Z} = \mathcal{M}^W_1 + \mathcal{M}^Z_1 + \mathcal{M}^{W,Z}_{1,1}$ implies that

$$\sum_{i=1}^{3} |T_i| \leq c_{22} l^{\alpha+1}\mathcal{M}_{W,Z}\delta + c_{23} l^{\alpha+1}(\mathcal{M}_{W,Z})^2\delta^2.$$

Following the same developments to prove (3.42), we apply (3.40), Remark 3 and Lemma 28 of Fang et al. (2020) and the union bound of probability, choose suitable positive constants c_{24}, c_{25}, c_{26} and hence obtain that

$$\begin{aligned} & \mathbb{P}\left\{ \max_{j\in[p],k\in[\tilde{p}],l\in[d]} \frac{|\hat{\varrho}^{W,Z}_{h,jkl} - \varrho^{W,Z}_{h,jkl}|}{l^{\alpha+1}} > \mathcal{M}_{W,Z}\delta + c_{26}(\mathcal{M}_{W,Z})^2\delta^2 \right\} \\ & \leq c_{25}p\tilde{p}d \exp\Big\{ -c_{24}n\min(\delta^2, \delta) \Big\} + c_{25}p\tilde{p}d \exp\big(-c_{24}n \big). \end{aligned} \tag{3.44}$$

Choosing $\delta = \rho_5\sqrt{\log(p\tilde{p}d)/n} \leq 1$ and $1 + c_{26}\mathcal{M}_{W,Z}\delta \leq \rho_6$ for some positive constants ρ_5, ρ_6, which can be achieved for sufficiently large $n \gtrsim (\mathcal{M}_{W,Z})^2 \log(p\tilde{p}d)$, it follows

from (3.44) that there exists positive constants c_{27}, c_{28} such that, with probability greater than $1 - c_{27}(p\tilde{p}d)^{-c_{28}}$,

$$\max_{j\in[p],k\in[\tilde{p}],l\in[d]} \frac{|\hat{\varrho}_{h,jkl}^{W,Z} - \varrho_{h,jkl}^{W,Z}|}{l^{\alpha+1}} \leq \rho_5\rho_6\mathcal{M}_{W,Z}\sqrt{\frac{\log(p\tilde{p}d)}{n}},$$

which completes the proof of Proposition 3.3. □

Proof of Theorem 3.3

By $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{G}\boldsymbol{\theta} + \mathbf{g}(\mathbf{0})$ and (3.17), we have $\mathbf{g}(\widehat{\boldsymbol{\theta}}) = \mathbf{G}\widehat{\boldsymbol{\theta}} + \mathbf{g}(\mathbf{0})$, $\mathbf{G}\boldsymbol{\theta}_0 + \mathbf{g}(\mathbf{0}) + \mathbf{R} = \mathbf{0}$ and $\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) = \widehat{\mathbf{G}}\widehat{\boldsymbol{\theta}} + \widehat{\mathbf{g}}(\mathbf{0})$. Consider event $A = \big\{\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\max}^{(d,d)} \vee \|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})\|_{\max}^{(d,\tilde{d})} \leq \epsilon_{n1}\big\} \cap \big\{\|\widehat{\mathbf{g}}(\boldsymbol{\theta}_0)\|_{\max}^{(d,\tilde{d})} \leq \gamma_n\big\}$. By the union bound of probability and Conditions 3.4 and 3.6, this event occurs with probability at least $1 - \delta_{n1} - \delta_{n2}$. On event A, we have

$$\begin{aligned}
\|\mathbf{G}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|_{\max}^{(d,\tilde{d})} &\leq \|\mathbf{g}(\widehat{\boldsymbol{\theta}})\|_{\max}^{(d,\tilde{d})} + \|\mathbf{R}\|_{\max}^{(d,\tilde{d})} \\
&\leq \|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) - \mathbf{g}(\widehat{\boldsymbol{\theta}})\|_{\max}^{(d,\tilde{d})} + \|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}})\|_{\max}^{(d,\tilde{d})} + \|\mathbf{R}\|_{\max}^{(d,\tilde{d})} \\
&\leq \|(\widehat{\mathbf{G}} - \mathbf{G})\widehat{\boldsymbol{\theta}}\|_{\max}^{(d,\tilde{d})} + \|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})\|_{\max}^{(d,\tilde{d})} + \|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}})\|_{\max}^{(d,\tilde{d})} + \|\mathbf{R}\|_{\max}^{(d,\tilde{d})} \\
&\leq \|\widehat{\mathbf{G}} - \mathbf{G}\|_{\max}^{(d,d)}\|\boldsymbol{\theta}_0\|_1^{(d,\tilde{d})} + \|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})\|_{\max}^{(d,\tilde{d})} + \\
&\qquad \|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}})\|_{\max}^{(d,\tilde{d})} + \|\mathbf{R}\|_{\max}^{(d,\tilde{d})} \\
&\leq K\epsilon_{n1} + \epsilon_{n1} + \gamma_n + \epsilon_2,
\end{aligned} \tag{3.45}$$

where, in the last two inequalities, we have used the facts that $\|(\widehat{\mathbf{G}} - \mathbf{G})\widehat{\boldsymbol{\theta}}\|_{\max}^{(d,\tilde{d})} = \max_{i\in[q]} \sum_{j=1}^{p} \|(\widehat{\mathbf{G}} - \mathbf{G})_{ij}\widehat{\boldsymbol{\theta}}_j\|_{\mathrm{F}} \leq \max_{i,j} \|(\widehat{\mathbf{G}} - \mathbf{G})_{ij}\|_{\mathrm{F}} \sum_j \|\widehat{\boldsymbol{\theta}}_j\|_{\mathrm{F}} = \|\widehat{\mathbf{G}} - \mathbf{G}\|_{\max}^{(d,d)} \|\widehat{\boldsymbol{\theta}}\|_1^{(d,\tilde{d})}$, $\|\widehat{\boldsymbol{\theta}}\|_1^{(d,\tilde{d})} \leq \|\boldsymbol{\theta}_0\|_1^{(d,\tilde{d})} \leq K$ and $\|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}})\|_{\max}^{(d,\tilde{d})} \leq \gamma_n$ by the definition of the block RMD estimator in (3.18) and $\|\mathbf{R}\|_{\max}^{(d,\tilde{d})} \leq \epsilon_2$ by Condition 3.5. On event A, choosing the set $T = S$ in (3.19) and applying Lemma 3.1 under Condition 3.6 yields $\|\widehat{\boldsymbol{\delta}}_{S^C}\|_1^{(d,\tilde{d})} \leq \|\widehat{\boldsymbol{\delta}}_S\|_1^{(d,\tilde{d})}$ and hence $\widehat{\boldsymbol{\delta}} \in C_S$. Then by (3.19), (3.45) and Lemma 3.3 under Condition 3.7, we have

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1^{(d,\tilde{d})} \leq \kappa(\boldsymbol{\theta}_0)^{-1} \cdot \|\mathbf{G}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|_{\max}^{(d,\tilde{d})} \lesssim \frac{s\{(K+1)\epsilon_{n1} + \gamma_n + \epsilon_2\}}{\mu^2},$$

which completes the proof. □

Proof of Proposition 3.1

Define $\tilde{\kappa}(\boldsymbol{\theta}_0)$ by substituting $\mathbf{G}$ in (3.19) by $\widetilde{\mathbf{G}}$. By $\widetilde{\mathbf{G}} = \mathbf{D}_x\mathbf{G}\mathbf{D}_y$ with $\mathbf{D}_x$ and $\mathbf{D}_y$ being diagonal matrices, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1^{(d,\tilde{d})} &\leq \tilde{\kappa}(\boldsymbol{\theta}_0)^{-1} \cdot \|\widetilde{\mathbf{G}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|_{\max}^{(d,\tilde{d})} \\ &\leq \tilde{\kappa}(\boldsymbol{\theta}_0)^{-1} \cdot \|\mathbf{D}_x\|_{\max}\|\mathbf{D}_y\|_{\max}\|\mathbf{G}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|_{\max}^{(d,\tilde{d})}.\end{aligned}$$

Following the same procedure to prove Theorem 3.3, we can obtain (3.21). □

Lemma 3.1 and its proof

Lemma 3.1. *Suppose that Condition* 3.6 *holds. Then* $\|\widehat{\boldsymbol{\delta}}_{S^c}\|_1^{(d,\tilde{d})} \leq \|\widehat{\boldsymbol{\delta}}_S\|_1^{(d,\tilde{d})}$ *with probability at least* $1 - \delta_{n2}$.

Proof. It follows from Condition 3.6 and $\boldsymbol{\theta}_{0,S^c} = \mathbf{0}$ by definition that with probability at least $1 - \delta_{n2}$, $\|\widehat{\boldsymbol{\theta}}\|_1^{(d,\tilde{d})} \leq \|\boldsymbol{\theta}_0\|_1^{(d,\tilde{d})} = \|\boldsymbol{\theta}_{0,S}\|_1^{(d,\tilde{d})}$, which implies that

$$\begin{aligned}\|\boldsymbol{\theta}_{0,S}\|_1^{(d,\tilde{d})} &\geq \|\widehat{\boldsymbol{\theta}}_S\|_1^{(d,\tilde{d})} + \|\widehat{\boldsymbol{\theta}}_{S^c}\|_1^{(d,\tilde{d})} \\ &\geq \|\boldsymbol{\theta}_{0,S}\|_1^{(d,\tilde{d})} - \|\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}\|_1^{(d,\tilde{d})} + \|\widehat{\boldsymbol{\theta}}_{S^c}\|_1^{(d,\tilde{d})}.\end{aligned}$$

By cancelling $\|\boldsymbol{\theta}_{0,S}\|_1^{(d,\tilde{d})}$ on both sides above, we obtain $\|\widehat{\boldsymbol{\theta}}_{S^c} - \boldsymbol{\theta}_{0,S^c}\|_1^{(d,\tilde{d})} \leq \|\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0,S}\|_1^{(d,\tilde{d})}$. □

Lemma 3.2 and its proof

To simplify notation in this section, we will use $\sigma_{\min}(m)$ and $\sigma_{\max}(m)$ to represent $\sigma_{\min}(m, \mathbf{G})$ and $\sigma_{\max}(m, \mathbf{G})$, respectively.

Lemma 3.2. *It holds that*

$$\kappa(\boldsymbol{\theta}_0) \geq \max_{m \geq s} \left\{ \frac{\sigma_{\min}(m)}{\sqrt{m}} - \frac{2\sigma_{\max}(m)}{\sqrt{m}} \sqrt{\frac{s}{m}} \right\} \frac{s^{-1/2}}{2(1 + 2\sqrt{s/m})}.$$

Proof. Let $T \subset [p]$ and $\|\boldsymbol{\delta}_{T^c}\|_1^{(d,\tilde{d})} \leq \|\boldsymbol{\delta}_T\|_1^{(d,\tilde{d})}$ by (3.19). Let T_1 denote the largest m components of $\{\|\boldsymbol{\delta}_i\|_{\mathrm{F}}\}_{i\in[p]}$, and T_2 be the subsequent m-largest, etc. Let $\mathbf{V}_{\boldsymbol{\mu}} = \mathrm{diag}(\boldsymbol{\mu} \otimes \mathbf{1}_d)$ where $\boldsymbol{\mu} \in \mathbb{R}^q$ with $\sum_{i=1}^q I(|\mu_i| \neq 0) \leq m$ and $\mathbf{1}_d = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^d$.

We let $\|\boldsymbol{\mu}\| = (\sum_{i=1}^{q} \mu_i^2)^{1/2}$ and $\|\boldsymbol{\mu}\|_\infty = \max_{i\in[q]} |\mu_i|$. Then, we have

$$\begin{aligned}\|\mathbf{G}\boldsymbol{\delta}\|_{\max}^{(d,\tilde{d})} &= \max_{\boldsymbol{\mu}} \left\| \frac{1}{\|\boldsymbol{\mu}\|} \mathbf{V}_{\boldsymbol{\mu}} \mathbf{G}\boldsymbol{\delta} \right\|_{\mathrm{F}} \geq \left\| \frac{1}{\sqrt{m}\|\boldsymbol{\mu}\|_\infty} \mathbf{V}_{\boldsymbol{\mu}} \mathbf{G}\boldsymbol{\delta} \right\|_{\mathrm{F}} \\ &\geq \left\| \frac{1}{\sqrt{m}\|\boldsymbol{\mu}\|_\infty} \mathbf{V}_{\boldsymbol{\mu}} \mathbf{G}_{\cdot,T_1} \boldsymbol{\delta}_{T_1} \right\|_{\mathrm{F}} - \sum_{j\geq 2} \left\| \frac{1}{\sqrt{m}\|\boldsymbol{\mu}\|_\infty} \mathbf{V}_{\boldsymbol{\mu}} \mathbf{G}_{\cdot,T_j} \boldsymbol{\delta}_{T_j} \right\|_{\mathrm{F}}, \end{aligned} \tag{3.46}$$

where $\mathbf{G}_{\cdot,T_j}$ is the block submatrix of $\mathbf{G}$ consisting of all rows and all block columns in T_j of $\mathbf{G}$ for $j \geq 1$. Define $\tilde{J}_1 = \arg\max_{|J|\leq m} \sigma_{\min}(\mathbf{G}_{J,T_1})$. We can let $\boldsymbol{\mu} = (\mu_i)$ with $\mu_i = 1$ if $i \in \tilde{J}_1$ and 0 otherwise, so that $\|\boldsymbol{\mu}\|_\infty = 1$. Then the first term in (3.46) becomes

$$\begin{aligned}\left\| \frac{1}{\sqrt{m}\|\boldsymbol{\mu}\|_\infty} \mathbf{V}_{\boldsymbol{\mu}} \mathbf{G}_{\cdot T_1} \boldsymbol{\delta}_{T_1} \right\|_{\mathrm{F}} &= \left\| \frac{1}{\sqrt{m}} \mathbf{G}_{\tilde{J}_1,T_1} \boldsymbol{\delta}_{T_1} \right\|_{\mathrm{F}} \\ &\geq \frac{\sigma_{\min}(\mathbf{G}_{\tilde{J}_1,T_1})}{\sqrt{m}} \|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}} = \frac{1}{\sqrt{m}} \max_{|J|\leq m} \sigma_{\min}(\mathbf{G}_{J,T_1}) \|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}} \\ &\geq \frac{1}{\sqrt{m}} \min_{|M|\leq m} \max_{|J|\leq m} \sigma_{\min}\left(\mathbf{G}_{J,M}\right) \|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}} = \frac{\sigma_{\min}(m)}{\sqrt{m}} \|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}}, \end{aligned} \tag{3.47}$$

where the first inequality comes from Lemma 3.6. Define $\tilde{J}_j = \arg\max_{|J|\leq m}$ $\sigma_{\max}(\mathbf{G}_{J,T_j})$ for each $j \geq 2$. By the similar arguments as above, the second term in (3.46) becomes

$$\begin{aligned}\sum_{j\geq 2} \left\| \frac{1}{\sqrt{m}\|\boldsymbol{\mu}\|_\infty} \mathbf{V}_{\boldsymbol{\mu}} \mathbf{G}_{\cdot,T_j} \boldsymbol{\delta}_{T_j} \right\|_{\mathrm{F}} &= \frac{1}{\sqrt{m}} \sum_{j\geq 2} \left\| \mathbf{G}_{\tilde{J}_j,T_j} \boldsymbol{\delta}_{T_j} \right\|_{\mathrm{F}} \\ &\leq \frac{1}{\sqrt{m}} \sum_{j\geq 2} \sigma_{\max}(\mathbf{G}_{\tilde{J}_j,T_j}) \|\boldsymbol{\delta}_{T_j}\|_{\mathrm{F}} \\ &= \frac{1}{\sqrt{m}} \sum_{j\geq 2} \max_{|J|\leq m} \sigma_{\max}(\mathbf{G}_{J,T_j}) \|\boldsymbol{\delta}_{T_j}\|_{\mathrm{F}} \\ &\leq \frac{1}{\sqrt{m}} \max_{|M|\leq m} \max_{|J|\leq m} \sigma_{\max}\left(\mathbf{G}_{J,M}\right) \sum_{j\geq 2} \|\boldsymbol{\delta}_{T_j}\|_{\mathrm{F}} \\ &= \frac{\sigma_{\max}(m)}{\sqrt{m}} \sum_{j\geq 2} \|\boldsymbol{\delta}_{T_j}\|_{\mathrm{F}}. \end{aligned} \tag{3.48}$$

By the construction of sets $\{T_j\}_{j\geq 1}$, we have $\|\boldsymbol{\delta}_{T_j}\|_1^{(d,\tilde{d})} = \sum_{l\in T_j} \|\boldsymbol{\delta}_l\|_{\mathrm{F}} \geq m\|\boldsymbol{\delta}_{T_{j+1}}\|_{\max}^{(d,\tilde{d})}$

$\geq \sqrt{m}\|\boldsymbol{\delta}_{T_{j+1}}\|_{\mathrm{F}}$, which implies that

$$\sum_{j\geq 2}\|\boldsymbol{\delta}_{T_j}\|_{\mathrm{F}} \leq \sum_{j\geq 1}\|\boldsymbol{\delta}_{T_j}\|_1^{(d,\tilde{d})}/\sqrt{m} \leq \|\boldsymbol{\delta}\|_1^{(d,\tilde{d})}/\sqrt{m}. \tag{3.49}$$

Combining (3.47), (3.48) and (3.49) yields

$$\begin{aligned}\|\mathbf{G}\boldsymbol{\delta}\|_{\max}^{(d,\tilde{d})} &\geq \frac{\sigma_{\min}(m)}{\sqrt{m}}\|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}} - \frac{\sigma_{\max}(m)}{\sqrt{m}}\|\boldsymbol{\delta}\|_1^{(d,\tilde{d})}/\sqrt{m} \\ &\geq \frac{\sigma_{\min}(m)}{\sqrt{m}}\|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}} - \frac{\sigma_{\max}(m)}{\sqrt{m}}2\sqrt{\frac{s}{m}}\|\boldsymbol{\delta}_T\|_{\mathrm{F}} \\ &= \left\{\frac{\sigma_{\min}(m)}{\sqrt{m}} - 2\frac{\sigma_{\max}(m)}{\sqrt{m}}\sqrt{\frac{s}{m}}\frac{\|\boldsymbol{\delta}_T\|_{\mathrm{F}}}{\|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}}}\right\}\|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}}\end{aligned} \tag{3.50}$$

where the second inequality comes from $\|\boldsymbol{\delta}\|_1^{(d,\tilde{d})} \leq 2\|\boldsymbol{\delta}_T\|_1^{(d,\tilde{d})} \leq 2\sqrt{s}\|\boldsymbol{\delta}_T\|_{\mathrm{F}}$ with $|T| \leq s$. This fact together with (3.49) implies that

$$\|\boldsymbol{\delta}\|_{\mathrm{F}} \leq \|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}} + \sum_{j\geq 2}\|\boldsymbol{\delta}_{T_j}\|_{\mathrm{F}} \leq \|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}} + 2\sqrt{s/m}\|\boldsymbol{\delta}_T\|_{\mathrm{F}} \leq (1+2\sqrt{s/m})\|\boldsymbol{\delta}_{T_1}\|_{\mathrm{F}}. \tag{3.51}$$

Combing (3.50) and (3.51) yields that

$$\begin{aligned}\|\mathbf{G}\boldsymbol{\delta}\|_{\max}^{(d,\tilde{d})} &\geq \left\{\frac{\sigma_{\min}(m)}{\sqrt{m}} - 2\frac{\sigma_{\max}(m)}{\sqrt{m}}\sqrt{\frac{s}{m}}\right\}\frac{\|\boldsymbol{\delta}\|_{\mathrm{F}}}{(1+2\sqrt{s/m})} \\ &\geq \left\{\frac{\sigma_{\min}(m)}{\sqrt{m}} - 2\frac{\sigma_{\max}(m)}{\sqrt{m}}\sqrt{\frac{s}{m}}\right\}\frac{\|\boldsymbol{\delta}\|_1^{(d,\tilde{d})}/\sqrt{s}}{2(1+2\sqrt{s/m})},\end{aligned} \tag{3.52}$$

where the second inequality comes from $\|\boldsymbol{\delta}\|_{\mathrm{F}} \geq \|\boldsymbol{\delta}_T\|_{\mathrm{F}} \geq \|\boldsymbol{\delta}_T\|_1^{(d,\tilde{d})}/\sqrt{s} \geq \|\boldsymbol{\delta}\|_1^{(d,\tilde{d})}/\sqrt{4s}$. We complete our proof by (3.19) and dividing $\|\boldsymbol{\delta}\|_1^{(d,\tilde{d})}$ on both sides of (3.52). □

Lemma 3.3 and its proof

Lemma 3.3. *Suppose that Condition* 3.7 *holds. Then there exists some constant c such that*

$$\kappa(\boldsymbol{\theta}_0) \geq \frac{c\mu^2}{24s}.$$

Proof. Applying Lemma 3.2 and choosing $m = 16s/\mu^2$ yields that

$$\begin{aligned}\kappa(\boldsymbol{\theta}_0) &\geq \max_{m \geq s} \frac{\sigma_{\max}(m, \mathbf{G})}{\sqrt{m}} \left\{ \frac{\sigma_{\min}(m, \mathbf{G})}{\sigma_{\max}(m, \mathbf{G})} - \frac{\mu}{2} \right\} \frac{s^{-1/2}}{2(1 + \mu/2)} \\ &\geq \frac{c\mu}{4\sqrt{s}} (\mu - \frac{\mu}{2}) [2(1 + \frac{\mu}{2})]^{-1} s^{-1/2} \geq \frac{c\mu^2}{24s} . \square\end{aligned}$$

Proof of Theorem 3.4

We first verify Condition 3.4 for SFLR. Define events

$$I_1 = \left\{ \max_{j,k \in [p], h \in [L], l, m \in [d]} \left| \hat{\sigma}^{(h)}_{jklm} - \sigma^{(h)}_{jklm} \right| \leq c_1 d^{\alpha+1} \mathcal{M}_1^W \sqrt{\frac{\log(pd)}{n}} \right\}, \tag{3.53}$$

$$I_2 = \left\{ \max_{k \in [p], h \in [L], m \in [d]} \left| \frac{1}{(n-h)} \sum_{t=h}^{n} \hat{\eta}_{(t+h)km} Y_t - \mathbb{E}\{\eta_{(t+h)km} Y_t\} \right| \leq c_2 d^{\alpha+1} \mathcal{M}_{W,Y} \sqrt{\frac{\log(pd)}{n}} \right\}.$$

On event $I_1 \cap I_2$, we have

$$\begin{aligned}\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\max}^{(d,d)} &= \max_{j,k \in [p], h \in [L]} \left\| \frac{1}{(n-h)} \sum_{t=h}^{n} \widehat{\boldsymbol{\eta}}_{(t+h)k} \widehat{\boldsymbol{\eta}}_{tj}^{\mathrm{T}} - E\{\boldsymbol{\eta}_{(t+h)k} \boldsymbol{\eta}_{tj}^{\mathrm{T}}\} \right\|_{\mathrm{F}} \\ &\leq c_1 d^{\alpha+2} \mathcal{M}_1^W \sqrt{\frac{\log(pd)}{n}},\end{aligned} \tag{3.54}$$

$$\begin{aligned}\|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})\|_{\max}^{(d,1)} &= \max_{k \in [p], h \in [L]} \left\| \frac{1}{(n-h)} \sum_{t=h}^{n} \widehat{\boldsymbol{\eta}}_{(t+h)k} Y_t - \mathbb{E}(\boldsymbol{\eta}_{tj} Y_t) \right\| \\ &\leq c_2 d^{\alpha+3/2} \mathcal{M}_{W,Y} \sqrt{\frac{\log(pd)}{n}}.\end{aligned} \tag{3.55}$$

By Theorem 3.2, Proposition 3.3 and the union bound of probability, $P(I_1 \cap I_2) \geq 1 - c_3(pd)^{-c_4}$. By (3.54) and (3.55), Condition 3.4 can be verified by choosing $\delta_{n1} = c_3(pd)^{-c_4}$ (pd depends on n) and

$$\epsilon_{n1} = c_5 d^{\alpha+2} \mathcal{M}_{W,Y} \sqrt{\frac{\log(pd)}{n}}. \tag{3.56}$$

We next verify Condition 3.5 for SFLR. If follows from $r_t = \sum_{j=1}^{p}\sum_{l=d+1}^{\infty}\eta_{tjl}\langle\psi_{jl}, \beta_{0j}\rangle$, orthonormality of $\{\psi_{jl}\}$, Cauchy–Schwartz inequality and Condition 3.8 that

$$\begin{aligned}
\big\{\|\mathbf{R}\|_{\max}^{(d,1)}\big\}^2 &= \max_{k\in[p],h\in[L]}\|\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}r_t\}\|^2 = \max_{k,h}\sum_{m=1}^{d}\left\{\mathbb{E}\Big(\eta_{(t+h)km}\sum_{j=1}^{p}\sum_{l=d+1}^{\infty}\eta_{tjl}a_{jl}\Big)\right\}^2 \\
&\le \max_{k,h}\sum_{m=1}^{d}\Big\{\sum_{j\in S}\sum_{l=d+1}^{\infty}\sqrt{\mathbb{E}(\eta_{(t+h)km}^2)\mathbb{E}(\eta_{tjl}^2)}a_{jl}\Big\}^2 \\
&\le s^2\max_{k,j}\sum_{m=1}^{d}\Big(\sum_{l=d+1}^{\infty}\lambda_{km}^{1/2}\lambda_{jl}^{1/2}a_{jl}\Big)^2 \\
&\le s^2\max_{k}\sum_{m=1}^{d}\lambda_{km}\max_{j}\Big\{\sum_{l=d+1}^{\infty}\lambda_{jl}\sum_{l=d+1}^{\infty}a_{jl}^2\Big\} \\
&\lesssim \lambda_0^2 s^2\sum_{l=d+1}^{\infty}l^{-2\tau} = O(s^2d^{-2\tau+1}).
\end{aligned}$$

where the asymptotic inequality comes from Condition 3.8 and $\lambda_0 = \max_j\sum_{l=1}^{\infty}\lambda_{jl}$ $= O(1)$ implied by some calculations based on (3.9) and Lemma 3.4. Therefore

$$\|\mathbf{R}\|_{\max}^{(d,1)} \le c_6 s d^{-\tau+1/2} = \epsilon_2. \tag{3.57}$$

By the similar technique above and Condition 3.8,

$$\|\mathbf{b}_0\|_1^{(d,1)} = \sum_{j\in S}(\sum_{l=1}^{d}a_{jl}^2)^{1/2} \lesssim s\max_{j\in S}(\sum_{l=1}^{d}l^{-2\tau})^{1/2} = O(s). \tag{3.58}$$

Finally, we verify Condition 3.6 for SFLR. On event $I_1\cap I_2$, combing (3.56) (3.57) and (3.58) yields that

$$\begin{aligned}
\|\widehat{\mathbf{g}}(\mathbf{b}_0)\|_{\max}^{(d,1)} &\le \|\widehat{\mathbf{g}}(\mathbf{b}_0) - \mathbf{g}(\mathbf{b}_0)\|_{\max}^{(d,1)} + \|\mathbf{R}\|_{\max}^{(d,1)} \\
&\le \|(\widehat{\mathbf{G}} - \mathbf{G})\mathbf{b}_0\|^{(d,1)} + \|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})\|_{\max}^{(d,1)} + \|\mathbf{R}\|_{\max}^{(d,1)} \\
&\le \|\widehat{\mathbf{G}} - \mathbf{G}\|_{\max}^{(d,d)}\|\mathbf{b}_0\|_1^{(d,1)} + \|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})\|_{\max}^{(d,1)} + \|\mathbf{R}\|_{\max}^{(d,1)} \\
&\le c_7 s\Big(d^{\alpha+2}\mathcal{M}_{W,Y}\sqrt{\frac{\log(pd)}{n}} + d^{-\tau+1/2}\Big) = \gamma_n.
\end{aligned} \tag{3.59}$$

By Condition 3.3 with $\max_j\|\mathbf{D}_j\|_{\max} \le \max_j\lambda_{jd}^{-1/2} = O(d^{\alpha/2})$ and Proposition 3.1

under Condition 3.9, we have

$$\|\widehat{\mathbf{b}} - \mathbf{b}_0\|_1^{(d,1)} = O_{\mathrm{p}}\left\{\mu^{-2}s^2 d^{\alpha}\Big(d^{\alpha+2}\mathcal{M}_{W,Y}\sqrt{\frac{\log(pd)}{n}} + d^{-\tau+1/2}\Big)\right\}. \tag{3.60}$$

For each $j \in [p]$, let $R_j(u) = \sum_{l=d+1}^{\infty} a_{jl}\psi_{jl}(u)$. By the orthonormality of $\{\psi_{jl}\}$ and $\|R_j\|^2 = \|\sum_{l=d+1}^{\infty} a_{jl}\psi_{jl}\|^2 = \sum_{l=d+1}^{\infty} a_{jl}^2 \lesssim d^{-2\tau+1}$ for $j \in S$ under Condition 3.8, we have

$$\begin{aligned}
\|\hat{\beta}_j - \beta_{0j}\| &= \|\widehat{\boldsymbol{\psi}}_j^{\mathrm{T}}\widehat{\mathbf{b}}_j - \boldsymbol{\psi}_j^{\mathrm{T}}\mathbf{b}_{0j} - R_j\| \\
&\le \|(\widehat{\boldsymbol{\psi}}_j - \boldsymbol{\psi}_j)^{\mathrm{T}}\widehat{\mathbf{b}}_j\| + \|\boldsymbol{\psi}_j^{\mathrm{T}}\{\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}\}\| + \|R_j\| \\
&\le d^{1/2}\max_{l\in[d]}\|\hat{\psi}_{jl} - \psi_{jl}\|\|\widehat{\mathbf{b}}_j\| + \|\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}\| + O(d^{-\tau+1/2}),
\end{aligned}$$

which implies that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \le d^{1/2}\max_{j\in[p],l\in[d]}\|\hat{\psi}_{jl} - \psi_{jl}\|\|\widehat{\mathbf{b}}\|_1^{(d,1)} + \|\widehat{\mathbf{b}} - \mathbf{b}_0\|_1^{(d,1)} + O(sd^{-\tau+1/2}),$$

where the third term above is of a smaller order of the second term due to (3.60). By $\|\widehat{\mathbf{b}}\|_1^{(d,1)} \le \|\widehat{\mathbf{b}} - \mathbf{b}_0\|_1^{(d,1)} + \|\mathbf{b}_0\|_1^{(d,1)}$, (3.58) and Theorem 3.1, the first term above is of a smaller order of the second term. Hence, we obtain (3.22) from (3.60), which completes the proof. □

Proof of Theorem 3.5

We first verify Condition 3.4 for FFLR. In addition to event I_1 in (3.53), we define event

$$I_3 = \left\{\max_{k\in[p],h\in[L],m\in[d],l\in[\tilde{d}]}\left|\frac{1}{(n-h)}\sum_{t=h}^{n}\hat{\eta}_{(t+h)km}\hat{\zeta}_{tl} - \mathbb{E}\{\eta_{(t+h)km}\zeta_{tl}\}\right| \le c_2 d^{\alpha\vee\tilde{\alpha}+1}\mathcal{M}_{W,Y}\sqrt{\frac{\log(pd\tilde{d})}{n}}\right\}.$$

On event $I_1 \cap I_3$, we have

$$\begin{aligned}\|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})\|_{\max}^{(d,\tilde{d})} &= \max_{k\in[p],h\in L}\Big\|\frac{1}{(n-h)}\sum_{t=h}^{n}\widehat{\boldsymbol{\eta}}_{(t+h)k}\widehat{\boldsymbol{\zeta}}_t^{\mathrm{T}} - \mathbb{E}\{\boldsymbol{\eta}_{tj}\boldsymbol{\zeta}_t^{\mathrm{T}}\}\Big\|_{\mathrm{F}} \\ &\le c_2 d^{\alpha\vee\tilde{\alpha}+2}\mathcal{M}_{W,Y}\sqrt{\frac{\log(pd\tilde{d})}{n}}.\end{aligned} \tag{3.61}$$

By Theorem 3.2, Proposition 3.2 and the union bound probability, $P(I_1 \cap I_3) \ge 1 - c_3(pd)^{-c_4}$. By (3.54) and (3.61), Condition 3.4 can be verified with the choice of

$$\epsilon_{n1} = c_5 d^{\alpha\vee\tilde{\alpha}+2}\mathcal{M}_{W,Y}\sqrt{\frac{\log(pd)}{n}}. \tag{3.62}$$

We next verify Condition 3.5 for FFLR. If follows from $\mathbf{r}_t = (r_{t1},\ldots,r_{t\tilde{d}})^{\mathrm{T}}$ with each $r_{tm'} = \sum_{j=1}^{p}\sum_{l=d+1}^{\infty}\eta_{tjl}\langle\langle\psi_{jl},\beta_{0j}\rangle,\phi_{m'}\rangle$, orthonormality of $\{\psi_{jl}\}$, $\{\phi_{m'}\}$, Cauchy–Schwartz inequality and Condition 3.10 that

$$\begin{aligned}\Big\{\|\mathbf{R}\|_{\max}^{(d,\tilde{d})}\Big\}^2 &= \max_{k\in[p],h\in[L]}\|\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\mathbf{r}_t^{\mathrm{T}}\}\|_{\mathrm{F}}^2 \\ &= \max_{k,h}\sum_{m=1}^{d}\sum_{m'=1}^{\tilde{d}}\left\{\mathbb{E}\Big(\eta_{(t+h)km}\sum_{j=1}^{p}\sum_{l=d+1}^{\infty}\eta_{tjl}a_{jlm'}\Big)\right\}^2 \\ &\le \max_{k,h}\sum_{m=1}^{d}\sum_{m'=1}^{\tilde{d}}\Big\{\sum_{j\in S}\sum_{l=d+1}^{\infty}\sqrt{\mathbb{E}(\eta_{(t+h)km}^2)\mathbb{E}(\eta_{tjl}^2)}a_{jlm'}\Big\}^2 \\ &\le s^2\max_{k,j}\sum_{m=1}^{d}\sum_{m'=1}^{\tilde{d}}\Big(\sum_{l=d+1}^{\infty}\lambda_{km}^{1/2}\lambda_{jl}^{1/2}a_{jlm'}\Big)^2 \\ &\le s^2\max_{k}\sum_{m=1}^{d}\lambda_{km}\max_{j}\Big\{\sum_{l=d+1}^{\infty}\lambda_{jl}\sum_{m'=1}^{\tilde{d}}\sum_{l=d+1}^{\infty}a_{jlm'}^2\Big\} \\ &\lesssim \lambda_0^2 s^2\sum_{m'=1}^{\tilde{d}}\sum_{l=d+1}^{\infty}(l+m')^{-2\tau-1} = O(s^2 d^{-2\tau+1}),\end{aligned}$$

which implies that

$$\|\mathbf{R}\|_{\max}^{(d,\tilde{d})} \le c_6 s d^{-\tau+1/2} = \epsilon_2. \tag{3.63}$$

By the similar technique above and Condition 3.10,

$$\|\mathbf{B}_0\|_1^{(d,\widetilde{d})} = \sum_{j\in S}(\sum_{l=1}^{d}\sum_{m=1}^{\widetilde{d}}a_{jlm}^2)^{1/2} \lesssim s\max_{j\in S}\Big\{\sum_{l=1}^{d}\sum_{m=1}^{\widetilde{d}}(l+m)^{-2\tau-1}\Big\}^{1/2} = O(s). \tag{3.64}$$

Finally, we verify Condition 3.6 for FFLR. On event $I_1 \cap I_3$, combing (3.62) (3.63) and (3.64) and applying the similar techniques for SFLR, we have

$$\begin{aligned}\|\widehat{\mathbf{g}}(\mathbf{B}_0)\|_{\max}^{(d,\tilde{d})} &\le \|\widehat{\mathbf{G}} - \mathbf{G}\|_{\max}^{(d,d)} \|\mathbf{B}_0\|_1^{(d,\tilde{d})} + \|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})\|_{\max}^{(d,\tilde{d})} + \|\mathbf{R}\|_{\max}^{(d,\tilde{d})} \\ &\le c_7 s\Big(d^{\alpha\vee\tilde{\alpha}+2}\mathcal{M}_{W,Y}\sqrt{\frac{\log(pd)}{n}} + d^{-\tau+1/2}\Big) = \gamma_n.\end{aligned} \tag{3.65}$$

By Condition 3.3 and Proposition 3.1 under Condition 3.9, we have

$$\|\widehat{\mathbf{B}} - \mathbf{B}_0\|_1^{(d,\tilde{d})} = O_{\mathrm{p}}\left\{\mu^{-2}s^2 d^{\alpha}\Big(d^{\alpha\vee\tilde{\alpha}+2}\mathcal{M}_{W,Y}\sqrt{\frac{\log(pd)}{n}} + d^{-\tau+1/2}\Big)\right\}. \tag{3.66}$$

For each $j \in [p]$, let $R_j(u,v) = (\sum_{l=1}^{d}\sum_{m=1}^{\tilde{d}} - \sum_{l,m=1}^{\infty})a_{jlm}\psi_{jl}(u)\phi_m(v)$ and write

$$\begin{aligned}\hat{\beta}_j(u,v) - \beta_{0j}(u,v) &= \widehat{\boldsymbol{\psi}}_j(u)^{\mathrm{T}}\widehat{\mathbf{B}}_j\widehat{\boldsymbol{\phi}}(v) - \boldsymbol{\psi}_j(u)^{\mathrm{T}}\mathbf{B}_{0j}\boldsymbol{\phi}(v) + R_j(u,v) \\ &= \widehat{\boldsymbol{\psi}}_j(u)^{\mathrm{T}}\widehat{\mathbf{B}}_j\{\widehat{\boldsymbol{\phi}}(v) - \boldsymbol{\phi}(v)\} + \{\widehat{\boldsymbol{\psi}}_j(u) - \boldsymbol{\psi}_j(u)\}^{\mathrm{T}}\widehat{\mathbf{B}}_j\boldsymbol{\phi}(v) \\ &\quad + \boldsymbol{\psi}_j(u)^{\mathrm{T}}\{\widehat{\mathbf{B}}_j - \mathbf{B}_{0j}\}\boldsymbol{\phi}(v) + R_j(u,v).\end{aligned}$$

By Lemma 9 of Guo and Qiao (2020), we bound the first three terms by

$$\begin{aligned}&\big\|\widehat{\boldsymbol{\psi}}_j^{\mathrm{T}}\widehat{\mathbf{B}}_j(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi})\big\|_{\mathcal{S}} \le \tilde{d}^{1/2}\max_{m\in[\tilde{d}]}\|\widehat{\phi}_m - \phi_m\|\|\widehat{\mathbf{B}}_j\|_{\mathrm{F}}, \\ &\big\|(\widehat{\boldsymbol{\psi}}_j - \boldsymbol{\psi}_j)^{\mathrm{T}}\widehat{\mathbf{B}}_j\boldsymbol{\phi}\big\|_{\mathcal{S}} \le d^{1/2}\max_{l\in[d]}\|\widehat{\psi}_{jl} - \psi_{jl}\|\|\widehat{\mathbf{B}}_j\|_{\mathrm{F}}, \\ &\big\|\boldsymbol{\psi}_j^{\mathrm{T}}(\widehat{\mathbf{B}}_j - \mathbf{B}_{0j})\boldsymbol{\phi}\big\|_{\mathcal{S}} = \|\widehat{\mathbf{B}}_j - \mathbf{B}_{0j}\|_{\mathrm{F}}.\end{aligned} \tag{3.67}$$

We next bound the fourth term. For $j \in S$, by the orthonormality of $\{\psi_{jl}\}$ and $\{\phi_m\}$,

$$\begin{aligned}\|R_j\|_{\mathcal{S}}^2 &= \|(\sum_{l=1}^{d}\sum_{m=1}^{\tilde{d}} - \sum_{l,m=1}^{\infty})a_{jlm}\psi_{jl}\phi_m\|_{\mathcal{S}}^2 \\ &= O(1)(\sum_{l=1}^{d}\sum_{m=\tilde{d}+1}^{\infty} a_{jlm}^2 + \sum_{l=1}^{\infty}\sum_{m=1}^{\tilde{d}} a_{jlm}^2) \\ &= O(1)\Big\{\sum_{l=1}^{d}\sum_{m=\tilde{d}+1}^{\infty}(l+m)^{-2\tau-1} + \sum_{l=1}^{\infty}\sum_{m=1}^{\tilde{d}}(l+m)^{-2\tau-1}\Big\} = O(d^{-2\tau+1}).\end{aligned} \tag{3.68}$$

Combing (3.67) and (3.68), we obtain

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \;\le\; & \|\widehat{\mathbf{B}}\|_1^{(d,\tilde{d})}\Big\{\tilde{d}^{1/2}\max_{m\in[\tilde{d}]}\|\widehat{\phi}_m-\phi_m\|+d^{1/2}\max_{j\in[p],l\in[d]}\|\widehat{\psi}_{jl}-\psi_{jl}\|\Big\}\\ & +\|\widehat{\mathbf{B}}-\mathbf{B}_0\|_1^{(d,\tilde{d})}+O(sd^{-\tau+1/2}),\end{aligned}$$

where the third term above is of a smaller order of the second term due to (3.66). By $\|\widehat{\mathbf{B}}\|_1^{(d,\tilde{d})}\le\|\widehat{\mathbf{B}}-\mathbf{B}_0\|_1^{(d,\tilde{d})}+\|\mathbf{B}_0\|_1^{(d,\tilde{d})}$, (3.64) and Theorem 3.1, the first term is of a smaller order of the second term. According to (3.66), we complete the proof. □

Proof of Theorem 3.6

For each $j\in[p]$, we first verify Condition 3.4 for VFAR. On event I_1 in (3.53),

$$\begin{aligned}\|\widehat{\mathbf{G}}_j-\mathbf{G}_j\|_{\max}^{(d,d)} &= \max_{j',k\in[p],h\in[L],h'\in[H]}\Big\|\frac{1}{(n-h)}\sum_{t=h}^{n}\widehat{\boldsymbol{\eta}}_{(t+h)k}\widehat{\boldsymbol{\eta}}_{(t-h')j'}^{\mathrm{T}}-\\ &\qquad\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{(t-h')j'}^{\mathrm{T}}\}\Big\|_{\mathrm{F}}\\ &\le c_1d^{\alpha+2}\mathcal{M}_1^W\sqrt{\frac{\log(pd)}{n}},\end{aligned}\tag{3.69}$$

$$\begin{aligned}\|\widehat{\mathbf{g}}_j(\mathbf{0})-\mathbf{g}_j(\mathbf{0})\|_{\max}^{(d,d)} &= \max_{k\in[p],h\in[L]}\Big\|\frac{1}{(n-h)}\sum_{t=h}^{n}\widehat{\boldsymbol{\eta}}_{(t+h)k}\widehat{\boldsymbol{\eta}}_{tj}^{\mathrm{T}}-\mathbb{E}(\boldsymbol{\eta}_{(t+h)k}\boldsymbol{\eta}_{tj}^{\mathrm{T}})\Big\|_{\mathrm{F}}\\ &\le c_2d^{\alpha+2}\mathcal{M}_1^W\sqrt{\frac{\log(pd)}{n}}.\end{aligned}\tag{3.70}$$

It follows from Theorem 3.2 that $P(I_1)\ge 1-c_3(pd)^{-c_4}$. By (3.69) and (3.70), Condition 3.4 can be verified by choosing

$$\epsilon_{n1}=c_5d^{\alpha+2}\mathcal{M}_1^W\sqrt{\frac{\log(pd)}{n}}.\tag{3.71}$$

We next verify Condition 3.5 for VFAR. It follows from $\mathbf{r}_{tj}=(r_{tj1},\ldots,r_{tjd})^{\mathrm{T}}$ with each $r_{tjm'}=\sum_{h'=1}^{H}\sum_{j'=1}^{p}\sum_{l=d+1}^{\infty}\eta_{(t-h')j'l}\big\langle\langle\psi_{j'l},A_{0,jj'}^{(h')}\rangle,\psi_{jm'}\big\rangle$, orthonormality

of $\{\psi_{jl}\}$, Cauchy–Schwartz inequality and Condition 3.11 that

$$\begin{aligned}
\big\{\|\mathbf{R}_j\|_{\max}^{(d,d)}\big\}^2 &= \max_{k\in[p],h\in[L]}\|\mathbb{E}\{\boldsymbol{\eta}_{(t+h)k}\mathbf{r}_{tj}^{\mathrm{T}}\}\|_{\mathrm{F}}^2 \\
&= \max_{k,h}\sum_{m=1}^{d}\sum_{m'=1}^{d}\left\{\mathbb{E}\Big(\eta_{(t+h)km}\sum_{h'=1}^{H}\sum_{j'=1}^{p}\sum_{l=d+1}^{\infty}\eta_{(t-h')j'l}a_{jj'lm'}^{(h')}\Big)\right\}^2 \\
&\le \max_{k,h}\sum_{m=1}^{d}\sum_{m'=1}^{d}\Big\{\sum_{(j',h')\in S_j}\sum_{l=d+1}^{\infty}\sqrt{\mathbb{E}(\eta_{(t+h)km}^2)\mathbb{E}(\eta_{(t-h')j'l}^2)}a_{jj'lm'}^{(h')}\Big\}^2 \\
&\le s_j^2\max_{k,j',h'}\sum_{m=1}^{d}\sum_{m'=1}^{d}\Big(\sum_{l=d+1}^{\infty}\lambda_{km}^{1/2}\lambda_{j'l}^{1/2}a_{jj'lm'}^{(h')}\Big)^2 \\
&\le s_j^2\max_{k}\sum_{m=1}^{d}\lambda_{km}\max_{j',h'}\Big\{\sum_{l=d+1}^{\infty}\lambda_{j'l}\sum_{m'=1}^{d}\sum_{l=d+1}^{\infty}(a_{jj'lm'}^{(h')})^2\Big\} \\
&\lesssim \lambda_0^2 s_j^2\sum_{m'=1}^{d}\sum_{l=d+1}^{\infty}(l+m')^{-2\tau-1} = O(s_j^2 d^{-2\tau+1}),
\end{aligned}$$

which implies that

$$\|\mathbf{R}_j\|_{\max}^{(d,d)} \le c_6 s_j d^{-\tau+1/2} = \epsilon_2. \tag{3.72}$$

By the similar technique above and Condition 3.11, we have

$$\begin{aligned}
\|\boldsymbol{\Omega}_{0j}\|_1^{(d,d)} &= \sum_{(j',h')\in S_j}\Big\{\sum_{l=1}^{d}\sum_{m=1}^{d}(a_{jj'lm}^{(h')})^2\Big\}^{1/2} \\
&\lesssim s_j\max_{(j',h')\in S_j}\Big(\sum_{l=1}^{d}\sum_{m=1}^{d}\big\{l+m\big)^{-2\tau-1}\Big\}^{1/2} = O(s_j).
\end{aligned} \tag{3.73}$$

Finally, we verify Condition 3.6 for VFAR. On event I_1, combing (3.71) (3.72),(3.73) and applying the similar techniques, we have

$$\begin{aligned}
\|\widehat{\mathbf{g}}_j(\boldsymbol{\Omega}_{0j})\|_{\max}^{(d,d)} &\le \|\widehat{\mathbf{G}}_j - \mathbf{G}_j\|_{\max}^{(d,d)}\|\boldsymbol{\Omega}_{0j}\|_1^{(d,d)} + \|\widehat{\mathbf{g}}_j(\mathbf{0}) - \mathbf{g}_j(\mathbf{0})\|_{\max}^{(d,d)} + \|\mathbf{R}_j\|_{\max}^{(d,d)} \\
&\le c_7 s_j\Big(d^{\alpha+2}\mathcal{M}_1^W\sqrt{\frac{\log(pd)}{n}} + d^{-\tau+1/2}\Big) = \gamma_{nj}.
\end{aligned} \tag{3.74}$$

By Condition 3.3 and Proposition 3.1 under Condition 3.12, we have

$$\|\widehat{\boldsymbol{\Omega}}_j - \boldsymbol{\Omega}_{0j}\|_1^{(d,d)} = O_{\mathrm{p}}\left\{\mu_j^{-2}s_j^2 d^{\alpha}\left(d^{\alpha+2}\mathcal{M}_1^W\sqrt{\frac{\log(pd)}{n}} + d^{-\tau+1/2}\right)\right\}. \tag{3.75}$$

For each $j' \in [p]$, let $R_{jj'}^{(h')}(u,v) = (\sum_{l=1}^{d}\sum_{m=1}^{d} - \sum_{l,m=1}^{\infty})a_{jj'lm}^{(h')}\psi_{j'm}(u)\psi_{jl}(v)$ and write

$$\begin{aligned}
\hat{A}_{jj'}^{(h')}(u,v) - A_{0,jj'}^{(h')}(u,v) &= \widehat{\boldsymbol{\psi}}_{j'}(u)^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')}\widehat{\boldsymbol{\psi}}_{j}(v) - \boldsymbol{\psi}_{j'}(u)^{\mathrm{T}}\boldsymbol{\Omega}_{0,jj'}^{(h')}\boldsymbol{\psi}_{j}(v) + R_{jj'}^{(h')}(u,v) \\
&= \widehat{\boldsymbol{\psi}}_{j'}(u)^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')}\{\widehat{\boldsymbol{\psi}}_{j}(v) - \boldsymbol{\psi}_{j}(v)\} + \\
&\quad \{\widehat{\boldsymbol{\psi}}_{j'}(u) - \boldsymbol{\psi}_{j'}(u)\}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')}\boldsymbol{\psi}_{j}(v) + \\
&\quad \boldsymbol{\psi}_{j'}(u)^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')} - \boldsymbol{\Omega}_{0,jj'}^{(h')}\}\boldsymbol{\psi}_{j}(v) + R_{jj'}^{(h')}(u,v)\,.
\end{aligned}$$

By the same techniques to prove (3.67), we bound the first three terms

$$\begin{aligned}
&\big\|\widehat{\boldsymbol{\psi}}_{j'}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')}(\hat{\boldsymbol{\psi}}_{j} - \boldsymbol{\psi}_{j})\big\|_{\mathcal{S}} \le d^{1/2}\max_{l\in[d]}\|\widehat{\psi}_{jl} - \psi_{jl}\|\|\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')}\|_{\mathrm{F}}\,, \\
&\big\|(\widehat{\boldsymbol{\psi}}_{j'} - \boldsymbol{\psi}_{j'})^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')}\boldsymbol{\psi}_{j}\big\|_{\mathcal{S}} \le d^{1/2}\max_{m\in[d]}\|\hat{\psi}_{j'm} - \psi_{j'm}\|\|\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')}\|_{\mathrm{F}}\,, \qquad (3.76) \\
&\big\|\boldsymbol{\psi}_{j'}^{\mathrm{T}}\{\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')} - \boldsymbol{\Omega}_{0,jj'}^{(h')}\}\boldsymbol{\psi}_{j}\big\|_{\mathcal{S}} = \|\widehat{\boldsymbol{\Omega}}_{jj'}^{(h')} - \boldsymbol{\Omega}_{0,jj'}^{(h')}\|_{\mathrm{F}}\,.
\end{aligned}$$

We next bound the fourth term. For $(j',h') \in S_j$, by the orthonormality of $\{\psi_{jl}\}$,

$$\begin{aligned}
\|R_{jj'}^{(h')}\|_{\mathcal{S}}^2 &= \|(\sum_{l=1}^{d}\sum_{m=1}^{d} - \sum_{l,m=1}^{\infty})a_{jj'lm}^{(h')}\psi_{jl}\psi_{j'm}\|_{\mathcal{S}}^2 \\
&= O(1)\Big\{\sum_{l=1}^{d}\sum_{m=d+1}^{\infty}(a_{jj'lm}^{(h')})^2\Big\} \qquad (3.77) \\
&= O(1)\Big\{\sum_{l=1}^{d}\sum_{m=d+1}^{\infty}(l+m)^{-2\tau-1}\Big\} = O(d^{-2\tau+1}).
\end{aligned}$$

Combing (3.76) and (3.77), we obtain

$$\begin{aligned}
&\max_{j\in[p]}\sum_{j'=1}^{p}\sum_{h'=1}^{H}\|\hat{A}_{jj'}^{(h')} - A_{0,jj'}^{(h')}\|_{\mathcal{S}} \\
\le\ &\max_{j}\|\widehat{\boldsymbol{\Omega}}_{j}\|_1^{(d,d)}\Big\{d^{1/2}\max_{j\in[p],l\in[d]}\|\hat{\psi}_{jl} - \psi_{jl}\| + d^{1/2}\max_{j'\in[p],m\in[d]}\|\hat{\psi}_{j'm} - \psi_{j'm}\|\Big\} \\
&+ \max_{j}\|\widehat{\boldsymbol{\Omega}}_{j} - \boldsymbol{\Omega}_{0j}\|_1^{(d,d)} + O(s_j d^{-\tau+1/2})\,,
\end{aligned}$$

where the third term above is of a smaller order of the second term due to (3.75). By $\max_j \|\widehat{\boldsymbol{\Omega}}_j\|_1^{(d,d)} \le \max_j \|\widehat{\boldsymbol{\Omega}}_j - \boldsymbol{\Omega}_{0j}\|_1^{(d,d)} + \max_j \|\boldsymbol{\Omega}_{0j}\|_1^{(d,d)}$, (3.73) and Theorem 3.1, the first term is of a smaller order of the second term. Applying (3.75) with $\mu = \min_j \mu_j$

and $s = \max_j s_j$ completes our proof. □

Lemma 3.4 and its proof

Lemma 3.4. *Suppose that Condition 3.2 holds. Then we have* $\omega_0 = \max_j \sum_{l=1}^{\infty} \omega_{jl}^W = O(1)$.

Proof. This lemma follows directly from Lemma 2 of Fang et al. (2020) and hence the proof is omitted here. □

Lemma 3.5 and its proof

Lemma 3.5. *For $p \times p$ lag-h autocovariance of $\{W_t(\cdot)\}$, $\{\Sigma_{h,jk}^W\}_{1\le j,k\le p}$, we have*

$$\|\Sigma_{h,jk}^W\|_{\mathcal{S}} \le \omega_0, \quad \|\Sigma_{h,jk}^W(\psi_{km})\|_{\mathcal{S}} \le \omega_{km}^{1/2}\omega_0^{1/2} \text{ for } m \ge 1.$$

Proof. This lemma follows directly from Lemma 8 of Guo and Qiao (2020) and hence the proof is omitted here. □

Lemma 3.6 and its proof

Lemma 3.6. *For $\mathbf{A} \in \mathbb{R}^{q\times p}$ with rank$(\mathbf{A}) \le \min(p, q)$ and $\mathbf{x} \in \mathbb{R}^{p\times d}$, let $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$ be the singular value decomposition of $\mathbf{A}$ with $\mathbf{\Lambda} = \mathrm{diag}\{\sigma_1, \ldots, \sigma_{\mathrm{r}}\}$ and $\sigma_1 \ge \cdots \ge \sigma_r > 0$. Then we have*

$$\sigma_r\|\mathbf{x}\|_F \le \|\mathbf{A}\mathbf{x}\|_F \le \sigma_1\|\mathbf{x}\|_F.$$

Proof. Let $\mathbf{v}_j$ denotes the j-th row of $\mathbf{V}^{\mathrm{T}}\mathbf{x}$ for $j \in [r]$. Write

$$\sigma_r^2\|\mathbf{x}\|_{\mathrm{F}}^2 \le \|\mathbf{A}\mathbf{x}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x}) = \mathrm{tr}(\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^{\mathrm{T}}\mathbf{x}) = \Big(\sum_{j=1}^{r} \sigma_j^2\mathbf{v}_j^{\mathrm{T}}\mathbf{v}_j\Big)^{1/2} \le \sigma_1^2\|\mathbf{x}\|_{\mathrm{F}}^2,$$

where, in the inequalities above, we have used $\|\mathbf{V}^{\mathrm{T}}\mathbf{x}\|_{\mathrm{F}} = \|\mathbf{x}\|_{\mathrm{F}}$ due to the orthonormality of $\mathbf{V}$. Taking the squared root completes the proof of this lemma. □

Chapter 4

De-Biased Learning for High Dimensional Time Series Linear Regression

4.1 Introduction

In the recent years, significant progress has been made on high-dimensional linear regression models. Consider the model

$$Y_t = \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_t,\ 1 \le t \le n, \tag{4.1}$$

where $\mathbf{X}_t = (X_{t1}, \ldots, X_{tp}) \in \mathbb{R}^p$ is a stationary vector time series with autocovariance $\boldsymbol{\Sigma}_h^X = \mathrm{Cov}(\mathbf{X}_{t+h}, \mathbf{X}_t)$ for any integer h and ε_t is the random disturbance independent of $\mathbf{X}_t$. Without loss of generality, we assume that $\mathbb{E}\{\mathbf{X}_t\} = \mathbf{0}$. And we consider the error-in-variables case where signal $\mathbf{X}_t$ are not observable, instead, $\mathbf{W}_t = \mathbf{X}_t + \mathbf{e}_t$ are observed, where the white noise sequence $\mathbf{e}_t$ are independent of $\mathbf{X}_t$ and ε_t, with zero mean and $\boldsymbol{\Sigma}_h^e = \mathbf{0}$ for $h \neq 0$. For identification in high-dimension, we assume that β_0 is sparse with only s components are non-zero and s is much smaller than p.

Regression model with measurement errors has been substantially developed in the literature. Wang et al. (2019) considered the case where the covariates are missing completely at random, which could be regraded as an error-in-variables problem with

multiplicative errors that are Bernoulli distributed. Li et al. (2021) introduced the additive error into the covariates of interest. Both additive and multiplicative error were modelled in Datta and Zou (2017), but they all assumed the key information of the distribution for errors are known. In our proposed problem, all the covariates are measured with error and we do not impose the assumption that variance of the error is known. Instead, we can use autocovariance $\mathbf{\Sigma}_h^W$ to filter out the impact of $\mathbf{e}_t$ in the sense that $\mathbf{\Sigma}_h^W = \mathbf{\Sigma}_h^X$ for $h \neq 0$. This idea was used in Bathia et al. (2010), where the autocovariance information are extracted to perform the dimensional reduction and dimensional identification for functional data. Then the "low noise" condition (Hall and Vial, 2006) which assume that the error goes to zero as sample size increase is not required.

Let the lag terms $\mathbf{Z}_t = (\mathbf{W}_{t+1}^{\mathrm{T}}, \ldots, \mathbf{W}_{t+L}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{pL}$ be the instrumental variables (IVs), where $L > 0$ is a prescribed integer. Given the stationarity of the process $\mathbf{W}_t$, where the dependence quickly decaying to zero, a small L is suggested. Then denote the covariance of $\mathbf{Z}_t$ by $\mathbf{\Gamma} = \mathbb{E}\{\mathbf{Z}_t\mathbf{Z}_t^{\mathrm{T}}\} = (\mathbf{\Sigma}_{i-j}^{\mathbf{W}})_{i,j\in[L]} \in \mathbb{R}^{pL\times pL}$. Consider estimating regression coefficient $\boldsymbol{\beta}_0$ in (4.1) by Generalised Method of Moments (GMM) approach, and for $\boldsymbol{\beta} \in \mathbb{R}^p$, let

$$g_t(\boldsymbol{\beta}) = \mathbf{Z}_t(Y_t - \mathbf{W}_t^T\boldsymbol{\beta}), \quad t \in [n-L], \tag{4.2}$$

and denote expecation of $g_t(\boldsymbol{\beta})$ by $\mathbf{g}(\boldsymbol{\beta}) = \mathbb{E}\{g_t(\boldsymbol{\beta})\}$. Then the moment conditions are

$$\mathbf{g}(\boldsymbol{\beta}_0) = \mathbf{G}\boldsymbol{\beta}_0 + \mathbf{g}(\mathbf{0}) = \mathbf{0}, \tag{4.3}$$

where $\mathbf{G} = -\mathbb{E}\{\mathbf{Z}_t\mathbf{W}_t^{\mathrm{T}}\} = -(\mathbf{\Sigma}_h^{\mathbf{W}})_{h\in[L]} \in \mathbb{R}^{pL\times p}$ and $\mathbf{g}(\mathbf{0}) = \mathbb{E}\{\mathbf{Z}_tY_t\} \in \mathbb{R}^{pL}$. And $\widehat{\boldsymbol{\beta}}$, the estimation of the coefficients $\boldsymbol{\beta}_0$, could be obtained by some regularised estimates like Lasso or Dantzig. Belloni et al. (2018) studied the Dantzig type method for independent data. And Caner and Kock (2018) estimate the high dimension linear GMM with inference using LASSO. However, the estimation $\widehat{\boldsymbol{\beta}}$ derived from the regularised methods suffers from the regularisation bias, which motivates us to invoke some de-bias approach. For example, Javanmard and Montanari (2014) and Wang et al. (2019) proposed the de-biased estimator based on the Lasso estimation and remove the bias by subtracting the term proportional to the sub-gradient of ℓ_1 norm in the Lasso solution. Ning and Liu (2017) and Li et al. (2021) provided the de-biased estimator by solving the estimation equation consists of the first order approximation (decorrelated) score function. However, they assume a fixed number for covariates of interest, even the dimension of nuisance parameter are allowed to be larger than n. Belloni et al. (2018) proposed de-biased regularised GMM (DRGMM)

approach, which can mitigate the impact of the bias. The DRGMM estimator is derived from the regularised GMM (RGMM) estimator with an approximately linear form which fits into the Many Approximate Mean (MAM) framework. Moreover, MAM provides a tool kits for further inferential study on $\boldsymbol{\beta}_0$.

To extend the DRGMM approach to dependent data case, it is important to measure the dependence or stability. Basu and Michailidis (2015) introduce a measure of stability for stationary processes using their spectral properties that provides insight into the effect of dependence on the accuracy of the regularised estimates. Guo and Qiao (2020) and Fang et al. (2020) propose a stability measure for Gaussian and sub-Gaussian functional data, which serve as a fundamental tool for further consistency analysis. Wu (2005) developed a functional dependence measure, which cover a large class of time series model and facilitates the inferential study. With the help of tools from Zhang and Wu (2017, 2021), the deviation of autocovariance based estimation are bounded.

Once we have a de-biased estimation, on which the hypothesis test and confidence intervals can be constructed. Adamek et al. (2020) and Caner and Kock (2018) bring Lasso to the time series setting and establish the uniform asymptotic normality for desparsified Lasso method, allowing for inference in high-dimensional time series.

In this chapter, we derive a DRGMM estimation and perform inference on $\boldsymbol{\beta}_0$ for high-dimensional linear time series model. Our proposed work consists of the following three steps.

Step 1: Apply RMD approach to obtain initial RGMM estimation by;

Step 2: Update the initial estimation by DRGMM;

Step 3: Perform simultaneous inference on $\boldsymbol{\beta}_0$.

This chapter is organised as follows. In Section 4.2, we present the high-dimensional time series linear regression model for which an autocovariance-based estimation and de-bias framework is proposed. In Section 4.3, we provided the theoretical guarantee for the estimation of sparse coefficients based on regularised minimum distant estimation. In Section 4.4, we perform the inferential study on the de-biased regularised estimation, where the theoretical results on estimation consistency and inference accuracy are provided. Section 4.5 exams the finite sample performance of the proposed inference procedure through simulation studies.

Notation. For a vector $\mathbf{v} = (v_1, \ldots, v_p)^{\mathrm{T}}$ we define $|\mathbf{v}|_q = (\sum_{j=1}^p |v_j|^q)^{1/q}, q > 0$, $|\mathbf{v}|_\infty = \max_{j\in[p]} |v_j|$ and $|\mathbf{v}|_0 = \sum_{j=1}^p I\{v_j \neq 0\}$. For any real matrix $\mathbf{A} = (a_{ij})_{q\times p}$, we write the ℓ_1, ℓ_∞ and the entrywise max norm as $|\mathbf{A}|_1 = \max_{j\in[p]} \sum_{i=1}^q |a_{ij}|$,

$|\mathbf{A}|_\infty = \max_{i\in[q]} \sum_{j=1}^{p} |a_{ij}|$ and $|\mathbf{A}|_{\max} = \max_{i\in[q],j\in[p]} |a_{ij}|$, respectively. For a random variable X, let $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$, $q > 0$. For two real numbers, set $x \vee y = \max(x, y)$ and $x \wedge y = \min(x, y)$. Let $\{a_n\}$ and $\{b_n\}$ be two sequences of positive numbers and denote $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a positive constant c such that $a_n/b_n \leq c$. If $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously, we can write $a_n \asymp b_n$. For a positive integer p, let $[p] = \{1, \ldots, p\}$.

4.2 Autocovariance-based DRGMM estimation

4.2.1 Autocovariance-based estimations

To characterise the effect of dependence and develop asymptotic results for estimators, we impose dependence measures as follow. Let $\mathcal{F}_t = (\ldots, \nu_{t-1}, \nu_t)$, where $(\nu_t)_{t\in\mathbb{Z}}$ are i.i.d. random elements. Suppose that $\mathbf{W}_t$ is a zero mean stationary process satisfying

$$\mathbf{W}_t = (W_{t1}, \ldots, W_{tp})^{\mathrm{T}} = \boldsymbol{\zeta}(\mathcal{F}_t), \tag{4.4}$$

where $\boldsymbol{\zeta}(\cdot) = (\zeta_1(\cdot), \ldots, \zeta_p(\cdot))^{\mathrm{T}}$ is an $\mathbb{R}^p$-valued measurable function. (4.4) defines a large class of linear and nonlinear time series model (Wu, 2005). Moreover, for the simplicity we assume that ν_t is a martingale difference sequence with $\mathbb{E}\{\nu_t|\mathcal{F}_{t-1}\} = 0$, $\mathbb{E}\{\nu_t^2|\mathcal{F}_{t-1}\} = \sigma_\nu^2$ and $\mathbb{E}\{W_t\nu_t\} = \mathbf{0}$, which restrict the dependency structure of the random disturbance term.

Let ν_0 be replaced by its i.i.d. copy ν_0' and $\mathbf{W}_t' = \boldsymbol{\zeta}(\mathcal{F}_t')$, where $\mathcal{F}_t' = (\ldots, \nu_0', \ldots, \nu_{t-1}, \nu_t)$. Then define the dependence adjusted norms of $\mathbf{W}_t$ by

$$\|\mathbf{W}_{\cdot,j}\|_{q,\theta} = \sup_{h\geq 0}(h+1)^\theta \sum_{t=h}^{\infty} \|W_{tj} - W_{tj}'\|_q, \tag{4.5}$$

where $q \geq 1$ and $\theta > 0$ depicts the decay rate of the cumulative tail dependence. Thus, a larger θ implies weaker temporal dependence. For the p-dimensional stationary process $\mathbf{W}_t$, to address the high-dimensionality, we further define the ℓ_∞ dependence adjusted norm

$$\||\mathbf{W}_\cdot|_\infty\|_{q,\theta} = \sup_{h\geq 0}(h+1)^\theta \sum_{t=h}^{\infty} \||\mathbf{W}_t - \mathbf{W}_t'|_\infty\|_q.$$

Then define the overall and the uniform dependence adjusted norm as $\Theta_{q,\theta}^{\mathbf{W}} = (\sum_{j\in[p]} \|\mathbf{W}_{\cdot,j}\|_{q,\theta}^{q/2})^{2/q}$ and $\Psi_{q,\theta}^{\mathbf{W}} = \max_{j\in[p]} \|\mathbf{W}_{\cdot,j}\|_{q,\theta}$, respectively. Let $\boldsymbol{\Sigma}_h^{\mathbf{W}}$ be the

lag-h autocovariance matrices of $\mathbf{W}_t$ estimated by

$$\widehat{\mathbf{\Sigma}}_h^{\mathbf{W}} = \frac{1}{n-L}\sum_{t=1}^{n-L}\mathbf{W}_{t+h}\mathbf{W}_t^{\mathrm{T}}$$

for $h \leq L$. When $h = 0$, $\mathbf{\Sigma}_0^{\mathbf{W}}$ is covariance matrix. Let $\widehat{\mathbf{g}}(\mathbf{0}) = (n-L)^{-1}\sum_{t=1}^{n-L} g_t(\mathbf{0})$ be the estimation of $\mathbf{g}(\mathbf{0})$. By (4.1) and (4.2), we have

$$g_t(\mathbf{0}) = \mathbf{Z}_t\mathbf{W}_t^{\mathrm{T}}\boldsymbol{\beta}_0 + \mathbf{R}_t,$$

where $\mathbf{R}_t = \mathbf{Z}_t(\varepsilon_t - \mathbf{e}_t^{\mathrm{T}}\boldsymbol{\beta}_0)$. Note that $\mathbf{e}_t$ and ε_t are zero mean white noise sequences independent of zero mean stationary sequence $\mathbf{Z}_t$. Therefore, $\mathbf{R}_t$ is also stationary with zero mean (Wecker, 1978). For the (pL)-dimensional process $\mathbf{R}_t$ we can similarly define the uniform dependence adjusted norm $\Psi_{q,\theta}^{\mathbf{R}} = \max_{j\in[pL]}\|\mathbf{R}_{\cdot,j}\|_{q,\theta}$. Denote that

$$\epsilon_n^{\mathbf{W}} = \max_{0\leq h\leq L}|\widehat{\mathbf{\Sigma}}_h^{\mathbf{W}} - \mathbf{\Sigma}_h^{\mathbf{W}}|_{\max} \quad \text{and} \quad \epsilon_n^{\mathbf{R}} = (n-L)^{-1}\left|\sum_{t=1}^{n-L}\mathbf{R}_t\right|_{\infty},$$

and Theorem 4.1 and 4.2 below provide, respectively, the non-asymptotic bounds on $\epsilon_n^{\mathbf{W}}$ and $\epsilon_n^{\mathbf{R}}$, which consists the estimation for $\mathbf{G}$, $\mathbf{\Gamma}$ and $\mathbf{g}(\mathbf{0})$.

Theorem 4.1. *Suppose that $\mathbf{W}_t$ is a zero mean stationary process of the form* (4.4) *and $q_1 > 4$. Let*

$$\mathcal{H}_1 = \frac{H_{1,n}^{2/q_1}}{n}(\log p\||\mathbf{W}_\cdot|_\infty\|_{q_1,\theta_1} \wedge \Theta_{q_1,\theta_1}^{\mathbf{W}})^2,$$
$$\mathcal{H}_2 = \sqrt{\frac{\log\{(L+1)p\}}{n}}(\Psi_{4,\theta_1}^{\mathbf{W}})^2 \quad \textit{and} \quad \mathcal{H}_3 = \frac{(1+(L+1)^{-\theta_1+1})\Psi_{2,0}^{\mathbf{W}}\Psi_{2,\theta_1}^{\mathbf{W}}}{n},$$

where $H_{1,n} = (L+1)^{q_1/4}n$ for $\theta_1 > 1/2 - 2/q_1$ and $H_{1,n} = (L+1)^{q_1/4}n + (L+1)n^{q_1/4-\theta_1 q_1/2}$ for $\theta_1 < 1/2 - 2/q_1$. Then we have

$$\epsilon_n^{\mathbf{W}} = O_p\left(\mathcal{H}_1 + \mathcal{H}_2 + \mathcal{H}_3\right).$$

One can prove Theorem 4.1 by applying Lemma 4.9 and 4.10 in the appendix, therefore, the proof is omitted. Given the preset small integer L, if $\Psi_{2,0}^{\mathbf{W}}$ and $\Psi_{2,\theta_1}^{\mathbf{W}}$ are bounded by constants, $\mathcal{H}_3$ is always smaller in order than $\mathcal{H}_1$ and $\mathcal{H}_2$, which respectively imply the polynomial tail and sub-Gaussian type tail. For large deviation, the

polynomial tail $\mathcal{H}_1$ dominates while the sub-Gaussian tail $\mathcal{H}_2$ dominates for small deviation. Both $\mathcal{H}_1$ and $\mathcal{H}_2$ involve dimension p, and we next investigate the effect of p.

Consider the case where $\mathcal{H}_1$ dominates when dependence adjusted norms $\Psi^{\mathbf{W}}_{4,\theta_1} \vee \|\mathbf{W}_{\cdot,j}\|_{q_1,\theta_1} \asymp 1$ and $\||\mathbf{W}_\cdot|_\infty\|_{q_1,\theta_1} \asymp p^{\tau_1}$ for some $\tau_1 \geq 0$. For example, τ_1 could be $1/q_1$ if there exists constants $C_1,\ C_2 > 0$ such that $C_1 \leq \|\mathbf{W}_{\cdot,j}\|_{q_1,\theta_1} \leq C_2$ for $j \in [p]$, where each component process of $\mathbf{W}_t$ satisfies a balanced order of dependence adjusted norm. Then we have $\||\mathbf{W}_\cdot|_\infty\|_{q_1,\theta_1} \leq (\sum_{j=1}^p \|\mathbf{W}_{\cdot,j}\|^{q_1}_{q_1,\theta_1})^{1/q_1} \asymp p^{1/q_1}$ and $\mathbf{\Theta}_{q_1,\theta_1} \asymp p^{2/q_1}$. Therefore, in $\mathcal{H}_1$, $\log p\||\mathbf{W}_\cdot|_\infty\|_{q_1,\theta_1}$ is smaller in order than $\Theta^{\mathbf{W}}_{q_1,\theta_1}$ and $\log p\||\mathbf{W}_\cdot|_\infty\|_{q_1,\theta_1} \wedge \Theta^{\mathbf{W}}_{q_1,\theta_1} \lesssim p^{\tau_1}\log p$. Note that for fixed L, there exists b_1 such that $H^{2/q_1}_{1,n}/n \asymp n^{-b_1}$ for $1/2 < b_1 \leq 1 - 2/q_1$. Then it allows p to be polynomial increase with n as $p^{2\tau_1}(\log p)^{3/4} \gtrsim n^{b_1+1/2}$. On the contrary, if we assume that dependence adjusted norms $\||\mathbf{W}_\cdot|_\infty\|_{q_1,\theta_1}$, $\Theta^{\mathbf{W}}_{q_1,\theta_1}$ and $\Psi^{\mathbf{W}}_{4,\theta_1}$ are bounded by constants, it can be shown that $\mathcal{H}_2$ dominates and we have $\epsilon^{\mathbf{W}}_n = O_p\left(n^{-1/2}(\log p)^{1/2}\right)$, which is the same rate as shown in Section 3.3.2, where sub-Gaussian assumption is applied.

Theorem 4.2. *Suppose that $\mathbf{R}_t$ is a zero mean stationary process of the form* (4.4) *and $q_2 > 2$. Let*

$$\mathcal{H}_4 = \frac{H_{2,n}\{\log(Lp)\}^{3/2}}{n}\||\mathbf{R}_\cdot|_\infty\|_{q_2,\theta_2}, \qquad \mathcal{H}_5 = \sqrt{\frac{\log(Lp)}{n}}\Psi^{\mathbf{R}}_{2,\theta_2},$$

where $H_{2,n} = n^{1/q_2}$ if $\theta_2 > 1/2 - 1/q_2$ and $H_{2,n} = n^{1/2-\theta_2}$ if $\theta_2 < 1/2 - 1/q_2$, then we have

$$\epsilon^{\mathbf{R}}_n = O_p\left(\mathcal{H}_4 + \mathcal{H}_5\right).$$

One can prove Theorem 4.2 by applying Lemma 4.11 in the appendix, therefore, the proof is omitted. Consider the case where dependence adjusted norms $\Psi^{\mathbf{R}}_{2,\theta_2} \asymp 1$ and $\||\mathbf{R}_\cdot|_\infty\|_{q_2,\theta_2} \asymp p^{\tau_2}$ for some $\tau_2 \geq 0$, and L is preset. Theorem 4.2 implies that $H_{2,n}/n \asymp n^{-b_2}$ for $1/2 < b_2 \leq 1-1/q_1$. Then, $\mathcal{H}_4$ dominates when $n^{b_2-1/2} \lesssim p^{\tau_2}\log p$. In the special case where $\Psi^{\mathbf{R}}_{2,\theta_2}$ and $\||\mathbf{R}_\cdot|_\infty\|_{q_2,\theta_2}$ are bounded by constants, it allows $\log p = o(n^{1/2-b_2})$, which leads to concentration rate $\epsilon^{\mathbf{R}}_n = O_p\left(n^{-1/2}(\log p)^{1/2}\right)$.

Theorem 4.1 and 4.2 facilitate us to establish the deviation bounds on estimators $\widehat{\mathbf{G}}$, $\widehat{\mathbf{\Gamma}}$ and $\widehat{\mathbf{g}}(\mathbf{0})$, which are crucial for convergence analysis of the RMD estimations in Section 4.3. Given the construction of $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{\Gamma}}$, it is clear that $|\widehat{\mathbf{G}} - \mathbf{G}|_{\max} \leq |\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}|_{\max} \leq \epsilon^{\mathbf{W}}_n$. Moreover, by Hölder's inequality, we have

$$|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})|_\infty \leq \max_{h\in[L]}|\widehat{\Sigma}^{\mathbf{W}}_h - \Sigma^{\mathbf{W}}_h|_{\max}|\boldsymbol{\beta}_0|_1 + \epsilon^{\mathbf{R}}_n.$$

Suppose that $|\boldsymbol{\beta}_0|_1 \leq K$ for some constant $K < \infty$. Then for $\epsilon_n^{\mathbf{g}} = |\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})|_\infty$, we have $\epsilon_n^{\mathbf{g}} \leq K\epsilon_n^{\mathbf{W}} + \epsilon_n^{\mathbf{R}}$.

To summarise the above discussion of Theorem 4.1 and 4.2 and to conclude the concentration analysis, we present the following condition.

Condition 4.1. Suppose that $\mathbf{W}_t$ and $\mathbf{R}_t$ are zero mean stationary process of the form (4.4). For $q_0 > 2$, there exist $\theta_0 > 0$ such that dependence adjusted norms satisfy $\|\mathbf{W}_{\cdot,j}\|_{2q_0,\theta_0} \vee \Psi_{4,\theta_0}^{\mathbf{W}} \vee \Psi_{2,\theta_0}^{\mathbf{R}} \asymp 1$, $\||\mathbf{W}_\cdot|_\infty\|_{2q_0,\theta_0} \lesssim p^{\tau_0}$ and $\||\mathbf{R}_\cdot|_\infty\|_{q_0,\theta_0} \lesssim p^{\tau_0}(\log p)^{1/2}$ for $\tau_0 > 0$.

For the two processes of interest $\mathbf{W}_t$ and $\mathbf{R}_t$, Condition 4.1 controls the the moments condition and the strength of dependence by q and θ. It balances the dimensionality of the ℓ_∞ dependence adjusted norms. Let $\epsilon_n = \epsilon_n^{\mathbf{W}} \vee \epsilon_n^{\mathbf{R}}$, the following proposition establishes a unified convergence rates.

Proposition 4.1. *Under Condition* 4.1*, there exists* b *with* $1/2 < b \leq 1 - 1/q_0$ *such that*

$$\epsilon_n = O_p\left(\frac{p^{\tau_0}(\log p)^2}{n^b} + \sqrt{\frac{\log p}{n}}\right).$$

Proposition 4.1 restricts the dimensionality and dependency structure of the time series. As a natural requirement of consistency, we need $p^{\tau_0}(\log p)^2 = o(n^b)$ and $\log p = o(n)$. If $p^{\tau_0}(\log p)^{3/2} \lesssim n^{b-1/2}$, then sub-Gaussian tail dominates and $\epsilon_n = O_p\left((\log p)^{1/2} n^{-1/2}\right)$. Otherwise, $p^{\tau_0}(\log p)^{3/2} \gtrsim n^{b-1/2}$ and $\epsilon_n = O_p\left(p^{\tau_0}(\log p)^2 n^{-b}\right)$.

4.2.2 DRGMM estimation

For the high-dimensional regression (4.1), we assume a sparse regression coefficient β_0 satisfying $|\boldsymbol{\beta}_0|_0 = s \ll p$. And let the coefficient has moderate magnitude with $|\boldsymbol{\beta}_0|_1 \leq K$ for some $K \geq 0$. Then we can acquire a regularised GMM (RGMM) estimation from the regularised minimum distance (RMD) estimator defined by

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} |\boldsymbol{\beta}|_1 \quad \text{s.t.} \quad |\widehat{\mathbf{g}}(\boldsymbol{\beta})|_\infty \leq \lambda_n^{\boldsymbol{\beta}}, \tag{4.6}$$

where $\lambda_n^{\boldsymbol{\beta}} \geq 0$ is a regularisation parameter.

We further consider a de-biased estimator by updating the biased estimation (4.6). Let $\boldsymbol{\Omega} = \mathbb{E}\{g_t(\boldsymbol{\beta}_0) g_t(\boldsymbol{\beta}_0)^\top\} \in \mathbb{R}^{pL \times pL}$ be the variance of the scores $g_t(\boldsymbol{\beta}_0)$. If the homoskedastic assumption is applied, then $\boldsymbol{\Omega} = \sigma^2 \mathbb{E}\{\mathbf{Z}_t \mathbf{Z}_t^\top\} = \sigma^2 \boldsymbol{\Gamma}$ with a finite σ^2. Suppose that the initial estimation is the nuisance parameters denoted by $\boldsymbol{\eta} \in \mathbb{R}^p$,

and the updating parameters $\boldsymbol{\alpha} \in \mathbb{R}^p$ to be the target. Considering the moment condition defined in (4.3), we have $\mathbf{g}(\boldsymbol{\beta}_0) = \mathbf{G}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}) + \mathbf{g}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, this motivate us to construct the moment equations for parameters $(\boldsymbol{\alpha}, \boldsymbol{\eta})$ as

$$\mathbf{M}(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \mathbf{G}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\left[\mathbf{G}(\boldsymbol{\alpha} - \boldsymbol{\eta}) + \mathbf{g}(\boldsymbol{\eta})\right], \tag{4.7}$$

Given the true values $\boldsymbol{\alpha} = \boldsymbol{\eta} = \boldsymbol{\beta}_0$, it follows that $\mathbf{M}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0) = 0$ and the Neyman orthogonality property $\partial_{\boldsymbol{\eta}^{\mathrm{T}}}\mathbf{M}(\boldsymbol{\beta}_0, \boldsymbol{\eta})|_{\boldsymbol{\eta}=\boldsymbol{\beta}_0} = \mathbf{0}$ is satisfied. It ensures that the moment equations are first order insensitive to the nuisance parameter $\boldsymbol{\eta}$ around the true value. Then we could define an "oracle" linear estimator $\bar{\mathbf{b}}$ of $\boldsymbol{\alpha}_0$ as the root of $\bar{\mathbf{M}}(\bar{\mathbf{b}}, \boldsymbol{\beta}_0) = \mathbf{G}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\left[\mathbf{G}(\bar{\mathbf{b}} - \boldsymbol{\beta}_0) + \widehat{\mathbf{g}}(\boldsymbol{\beta}_0)\right] = \mathbf{0}$, that is

$$\sqrt{n}(\bar{\mathbf{b}} - \boldsymbol{\beta}_0) = -(\mathbf{G}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\sqrt{n}\widehat{\mathbf{g}}(\boldsymbol{\beta}_0), \tag{4.8}$$

where $\sqrt{n}(\bar{\mathbf{b}} - \boldsymbol{\beta}_0)$ approximates $N(\mathbf{0}, (\mathbf{G}^{\mathrm{T}}\boldsymbol{\Gamma}^{-\mathbf{1}}\mathbf{G})^{-\mathbf{1}})$, under some regular conditions in Sections 4.4. This motivate us to update the estimator of the target parameter and obtain the DRGMM estimation

$$\widehat{\mathbf{b}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}}\widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}}), \tag{4.9}$$

where $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\mu}}$ are the plug-in estimator for $\boldsymbol{\gamma}_0 = \mathbf{G}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}$ and $\boldsymbol{\mu}_0 = (\boldsymbol{\gamma}_0\mathbf{G})^{-1}$, respectively, and $\widehat{\boldsymbol{\Gamma}} = (n-L)^{-1}\sum_{t=1}^{n-L}\mathbf{Z}_t\mathbf{Z}_t^{\mathrm{T}}$. Note that in the high-dimensional setting, $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\mu}}$ are singular due to the rank deficiency. Therefore, we consider to RMD estimation again to get $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\mu}}$. In particular, for each $j \in [p]$, let $\boldsymbol{\gamma}_j$ denote the j-th row of $\boldsymbol{\gamma}$ and $\widehat{\boldsymbol{\gamma}}_j$ be the solution of

$$\underset{\boldsymbol{\gamma}_j \in \mathbb{R}^{pL}}{\arg\min}|\boldsymbol{\gamma}_j|_1 \text{ s.t. } |\boldsymbol{\gamma}_j\widehat{\boldsymbol{\Gamma}} - (\widehat{\mathbf{G}}^{\mathrm{T}})_j|_\infty \leq \lambda_{nj}^{\gamma}, \tag{4.10}$$

where $(\widehat{\mathbf{G}}^{\mathrm{T}})_j$ denotes the j-th row of $\widehat{\mathbf{G}}^{\mathrm{T}}$. Likewise, let $\boldsymbol{\mu}_j$ be j-th row of $\boldsymbol{\mu}$, then, for each $j \in [p]$, $\widehat{\boldsymbol{\mu}}_j$ is the solution of

$$\underset{\boldsymbol{\mu}_j \in \mathbb{R}^{p}}{\arg\min}|\boldsymbol{\mu}_j|_1 \text{ s.t. } |\boldsymbol{\mu}_j\widehat{\boldsymbol{\gamma}}\widehat{\mathbf{G}} - \mathbf{I}_j|_\infty \leq \lambda_{nj}^{\mu}, \tag{4.11}$$

where $\mathbf{I}_j$ is the j-th row of the identity matrix $\mathbf{I} \in \mathbb{R}^{p\times p}$.

4.3 Theoretical results

In this section, we will show the consistency of the RMD estimator in (4.6), and the results could be applied to (4.10) and (4.11) in the same way. Before presenting properties of the RMD estimator $\widehat{\boldsymbol{\beta}}$, we impose the following high-level regularity conditions.

Condition 4.2. There exists $\epsilon_n^{\mathbf{G}}$, $\epsilon_n^{\mathbf{g}}$ and $\delta_n = o(1)$ such that $\{|\widehat{\mathbf{G}} - \mathbf{G}|_{\max} \leq \epsilon_n^{\mathbf{G}}$ and $|\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})|_\infty \leq \epsilon_n^{\mathbf{g}}\}$ hold with probability $1 - \delta_n$.

Let $\epsilon_n^{\mathbf{G}} \leq \epsilon_n$ and $\epsilon_n^{\mathbf{g}} \leq (K+1)\epsilon_n$, Condition 4.2 can be verified directly by Proposition 4.1. The concentration of the empirical moment equations is indicated by Condition 4.2, under which we can check that the $\boldsymbol{\beta}_0$ is a feasible solution in the optimisation problem (4.6) with high probability. It can be guaranteed by the following lemma.

Lemma 4.1. *Suppose that Condition* 4.1 *and* 4.2 *hold. If regularisation parameter in* (4.6) *satisfies* $\lambda_n^{\boldsymbol{\beta}} > K\epsilon_n^{\mathbf{G}} + \epsilon_n^{\mathbf{g}}$, *then* $|\widehat{\mathbf{g}}(\boldsymbol{\beta}_0)|_\infty \leq \lambda_n^{\boldsymbol{\beta}}$ *holds with probability* $1 - \delta_n$.

Lemma 4.1 ensures that, with high probability, the solution $\widehat{\boldsymbol{\beta}}$ of (4.6) exists and satisfies $|\widehat{\boldsymbol{\beta}}|_1 \leq |\boldsymbol{\beta}_0|_1$, which implies $|\widehat{\boldsymbol{\delta}}_{S^c}|_1 \leq |\widehat{\boldsymbol{\delta}}_S|_1$ for $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ as justified in Lemma 4.5 of the Appendix. It is a important property to handle high-dimensional models because for $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$, we can define the ℓ_1-sensitivity coefficient

$$\kappa(\boldsymbol{\beta}_0, \mathbf{G}) = \min_{T:|T|\leq s} \left\{ \min_{\boldsymbol{\delta}\in C_T:|\boldsymbol{\delta}|_1=1} |\mathbf{G}\boldsymbol{\delta}|_\infty \right\}, \tag{4.12}$$

where $C_T = \{\boldsymbol{\delta} \in \mathbb{R}^p : |\boldsymbol{\delta}_{T^c}|_1 \leq |\boldsymbol{\delta}_T|_1\}$ for $T \subset [p]$. Therefore, it follows that $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \in C_S$, by which we can establish the error bound for $|\widehat{\boldsymbol{\delta}}|_1$ in consequence of determined lower bond of $\kappa(\boldsymbol{\beta}_0, \mathbf{G})$ under Condition 4.3 below. Let $J \subset [q]$ and $M \subset [p]$, let $\mathbf{G}_{J,M} = (\mathbf{G}_{jk})_{j\in J, k\in M}$ with each $\mathbf{G}_{jk} \in \mathbb{R}^{|J|\times|M|}$ be the submatrix of $\mathbf{G}$ consisting of all rows $j \in J$ and all columns $k \in M$ of $\mathbf{G}$. For an integer $m \geq s$, we define

$$\sigma_{\min}(m, \mathbf{G}) = \min_{|M|\leq m} \max_{|J|\leq m} \sigma_{\min}(\mathbf{G}_{J,M}) \quad \text{and} \quad \sigma_{\max}(m, \mathbf{G}) = \max_{|M|\leq m} \max_{|J|\leq m} \sigma_{\max}(\mathbf{G}_{J,M}),$$

where $\sigma_{\min}(\mathbf{G}_{J,M})$ and $\sigma_{\max}(\mathbf{G}_{J,M})$ are the smallest and largest singular values of $\mathbf{G}_{J,M}$, respectively.

Condition 4.3. For a pair $(\boldsymbol{\beta}_0, \mathbf{G})$ with $S = \{j : \boldsymbol{\beta}_{0j} \neq 0\}$ and sparsity $s = |S|$, there exists universal constants $\sigma_0 > 0$ and $\mu > 0$ such that $\sigma_{\max}(m, \mathbf{G}) \geq \sigma_0$ and $\sigma_{\min}(m, \mathbf{G})/\sigma_{\max}(m, \mathbf{G}) \geq \mu$ for $m \leq 16s/\mu^2$.

We denote that Condition 4.3 holds for $(\boldsymbol{\beta}_0, \mathbf{G})$ since it is indexed by the parameter of interest $\boldsymbol{\beta}_0$ and the matrix $\mathbf{G}$. Provided that the closeness between population autocovariance $\mathbf{G}$ and its estimation $\widehat{\mathbf{G}}$ are bounded in Proposition 4.1, we choose to formulate Condition 4.3 on the population autocovariance matrix, rather than on the sample autocovariance matrix. Note that the quantity μ plays a crucial role in determining the lower bound of $\kappa(\beta_0, \mathbf{G})$. It is a strongly identified model if μ is bounded away from zero. Otherwise, we have a weak identified model when $\mu \to 0$. The identification of the problem is determined by the strength of the instrumental variables, see Belloni et al. (2018) for more details.

Lemma 4.2. *Suppose that Condition* 4.1, 4.2 *and* 4.3 *hold. If there exists* $\epsilon_n^{\boldsymbol{\beta}} > 0$ *such that regularisation parameter* $\lambda_n^{\boldsymbol{\beta}}$ *in* (4.6) *satisfies* $K\epsilon_n^{\mathbf{G}} + \epsilon_n^{\mathbf{g}} \leq \lambda_n^{\boldsymbol{\beta}} \leq \epsilon_n^{\boldsymbol{\beta}}$, *then with probability* $1 - \delta_n$ *we have*

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \leq 2C_{\sigma_0,\mu} s\epsilon_n^{\boldsymbol{\beta}},$$

where constant $C_{\sigma_0,\mu}$ *depends only on* σ_0 *and* μ.

It suffices to choose $\epsilon_n^{\boldsymbol{\beta}} = (2K+1)\epsilon_n$ and we have that $|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1$ is bounded by ϵ_n shown in Proposition 4.1. Next we proceed to analyse the convergence properties of the estimators $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\mu}}$. An empirical moment condition is imposed.

Condition 4.4. There exists $\epsilon_n^{\mathbf{G}}$, $\epsilon_n^{\boldsymbol{\Gamma}}$ and $\delta_n = o(1)$ such that $\{|\widehat{\mathbf{G}} - \mathbf{G}|_{\max} \leq \epsilon_n^{\mathbf{G}}$ and $|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}|_{\max} \leq \epsilon_n^{\boldsymbol{\Gamma}}\}$ hold with probability $1 - \delta_n$.

Condition 4.4 plays a similar role as Condition 4.2 of Lemma 4.2 and can be verified under Proposition 4.1 by setting $\epsilon_n^{\mathbf{G}} \leq \epsilon_n^{\boldsymbol{\Gamma}} \leq \epsilon_n$. Under the similar conditions as shown in Lemma 4.2 and by carefully choosing the regularisation parameters, we can derive the ℓ_1-rate of convergence for the rows of estimators.

Lemma 4.3. *Suppose that Condition* 4.1 *and* 4.4 *hold and* $\max_{j\in[p]} |\boldsymbol{\gamma}_{0j}|_1 \leq K$. *If there exists* $\epsilon_n^{\boldsymbol{\gamma}} > 0$ *such that the regularisation parameters* $\lambda_{nj}^{\boldsymbol{\gamma}}$ *in* (4.10) *satisfy* $K\epsilon_n^{\boldsymbol{\Gamma}} + \epsilon_n^{\mathbf{G}} \leq \lambda_{nj}^{\boldsymbol{\gamma}} \leq \epsilon_n^{\boldsymbol{\gamma}}$ *for* $j \in [p]$. *Suppose that Condition 4.3 holds for* $(\boldsymbol{\gamma}_{0j}, \boldsymbol{\Gamma})$, $j \in [p]$. *Then with probability* $1 - \delta_n$ *we have*

$$\max_{j\in[p]} |\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j}|_1 \leq 2C_{\sigma_0,\mu} s\epsilon_n^{\boldsymbol{\gamma}}.$$

Lemma 4.4. *Suppose that Condition* 4.1 *and* 4.4 *holds and* $\max_{j\in[p]} |\boldsymbol{\mu}_{0j}|_1 \leq K$. *If there exists* $\epsilon_n^{\boldsymbol{\mu}} > 0$ *such that the regularisation parameters* $\lambda_{nj}^{\boldsymbol{\gamma}}$ *in* (4.11) *satisfy*

$2K\epsilon_n^{\mathbf{G}} + K^2\epsilon_n^{\mathbf{\Gamma}} + K\max_{j\in[p]}\lambda_{nj}^{\boldsymbol{\gamma}} \le \lambda_{nj}^{\boldsymbol{\mu}} \le \epsilon_n^{\boldsymbol{\mu}}$ *for* $j \in [p]$. *Suppose that Condition 4.3 holds for* $(\boldsymbol{\mu}_{0j}, \mathbf{G}^{\mathrm{T}}\mathbf{\Gamma}^{-1}\mathbf{G}),\ j \in [p]$. *Then with probability* $1-\delta_n$ *we have*

$$\max_{j\in[p]} |\widehat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_{0j}|_1 \le 2C_{\sigma_0,\mu} s\epsilon_n^{\boldsymbol{\mu}}.$$

It is satisfied to choose $\epsilon_n^{\boldsymbol{\gamma}} = (K+2)\epsilon_n$ and $\epsilon_n^{\boldsymbol{\mu}} = (1\vee K^2)(K+2)\epsilon_n$. Then we have $\max_{j\in[p]} |\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j}|_1$ and $\max_{j\in[p]} |\widehat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_{0j}|_1$ are bounded by ϵ_n shown in Proposition 4.1.

4.4 Inferential study on the de-biased estimation

In this section we perform the inference based on de-biased estimator in Section 4.2.2. We invoke the high-dimensional central limit theorem proposed in Zhang and Wu (2017) to construct the simultaneous confidence intervals of the regression coefficients.

4.4.1 Influence decomposition

We first analysis the accuracy of the estimator $\widehat{\mathbf{b}}$ in (4.9). Let $\mathcal{U}_t = (u_{tj})_{j\in[p]}$, where $u_{tj} = -(\boldsymbol{\mu}_0\boldsymbol{\gamma}_0)_j g_t(\boldsymbol{\beta}_0)$, then, as justified in Lemma 4.8, the deviance $\widehat{\mathbf{b}} - \boldsymbol{\beta}_0$ is decomposed into two parts as

$$\sqrt{n}(\widehat{\mathbf{b}} - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\mathcal{U}_t + \mathbf{r}_n,$$

where $\mathbf{r}_n = \mathbf{r}_{1n} + \mathbf{r}_{2n}$ with $\mathbf{r}_{1n} = \sqrt{n}(\mathbf{I} - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}}\widehat{\mathbf{G}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\mathbf{r}_{2n} = \sqrt{n}(\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}} - \boldsymbol{\mu}_0\boldsymbol{\gamma}_0)\widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}})$. $\mathcal{U}_t$ are the influence factors and $\mathbf{r}_n$ is error. We can show in Theorem 4.3 that $|\mathbf{r}_n|_\infty$ is asymptotically negligible under mild conditions. Then we apply a central limit theorem to the influence factors.

Theorem 4.3. *Suppose that conditions in Lemma* 4.2, 4.3 *and* 4.4 *hold with* $|\boldsymbol{\beta}_0|_1 \le K$, $\max_{j\in[p]}|\boldsymbol{\gamma}_{0j}|_1 \le K$ *and* $\max_{j\in[p]}|\boldsymbol{\mu}_{0j}|_1 \le K$. *Let regularisation parameters satisfy* $\lambda_n^{\boldsymbol{\beta}} \le (2K+1)\epsilon_n$, $\lambda_{nj}^{\boldsymbol{\gamma}} \le (K+2)\epsilon_n$ *and* $\lambda_{nj}^{\boldsymbol{\mu}} \le 2(1\vee K^2)(K+2)\epsilon_n$ *for* $j \in [p]$. *Then there exists constant* C_K *only depending on* K *and* $C_{\sigma_0,\mu}$ *only depending on* σ_0 *and* μ *such that*

$$|\mathbf{r}_n|_\infty \le C_K C_{\sigma_0,\mu}\sqrt{n}s\epsilon_n^2.$$

The above result implies that $|\mathbf{r}_n|_\infty$ approximate zero if $\sqrt{n}s\epsilon_n^2 = o(1)$. By the results in Proposition 4.1, it allows dimension p polynomial increase with n as $p^a \lesssim n$ for some constant $a > 0$ such that $a\tau_0 < b - 1/4$, then $|\mathbf{r}_n|_\infty = o\left(sn^{1/2-2b}p^{2\tau_0}(\log p)^4\right)$ with $s \ll p$. Note that as discussed in Proposition 4.1, the rate can be improved by employing the stronger tail assumptions.

4.4.2 Simultaneous Inference

Consider the inference on $H_0 : \beta_{0j} = 0$ for $j \in J \subseteq [p]$. Suppose that $\mathbf{u}_t = \{u_{tj}\}_{j\in J}$ is a stationary process of form (4.4). Likewise, for $\mathbf{u}_t$ we define the dependence adjusted norm $\|\mathbf{u}_\cdot\|_{q,\theta}$ as (4.5). Moreover, to address the high-dimensionality, we define the following quantities:

$$\Psi_{q,\theta}^{\mathbf{u}} = \max_{j\in J} \|\mathbf{u}_{\cdot j}\|_{q,\theta}, \qquad \Upsilon_{q,\theta} = \left(\sum_{j\in J} \|\mathbf{u}_{\cdot j}\|_{q,\theta}^q\right)^{1/q},$$

$$\Phi_{q,\theta} = \Upsilon_{q,\theta} \wedge \{\|\mathbf{u}_\cdot\|_{q,\theta}(\log|J|)^{3/2}\}, \qquad L_1 = \{\Psi_{2,\theta}\Psi_{2,0}(\log|J|)^2\}^{1/\theta},$$

$$W_1 = (\Psi_{3,0}^6 + \Psi_{4,0}^4)\{\log(|J|n)\}^7, \qquad W_2 = \Psi_{2,\theta}^2\{\log(|J|n)\}^4,$$

$$W_3 = \left[n^{-\theta}\{\log(|J|n)\}^{3/2}\Phi_{q,\theta}\right]^{1/(1/2-\theta-1/q)}, \qquad N_1 = (n/\log|J|)^{q/2}\Phi_{q,\theta}^{-q},$$

$$N_2 = n(\log|J|)^{-2}\Psi_{2,\theta}^{-2}, \qquad N_3 = \{n^{1/2}(\log|J|)^{-1/2}\Phi_{q,\theta}^{-1}\}^{1/(1/2-\theta)}.$$

Condition 4.5. (i) (Weak dependency case) Suppose that $\Phi_{q,\theta} < \infty$ holds with $q \geq 2$ and $\theta > 1/2-1/q$, then $\Phi_{q,\theta}\{\log(|J|n)\}^{3/2}n^{1/q-1/2} \to 0$ and $L_1\max(W_1, W_2) = o(1)\min(N_1, N_2)$. (ii) (Strong dependency case) Suppose that $0 < \theta < 1/2 - 1/q$, then $\Phi_{q,\theta}(\log|J|)^{1/2} = o(n^\theta)$ and $L_1\max(W_1, W_2, W_3) = o(1)\min(N_2, N_3)$.

Condition 4.5 restricts the dependency structure of the influence factors $\mathbf{u}_t$. For example, consider the weak dependency case where $\theta > 1/2-1/q$ for $\Phi_{q,\theta} = O(|J|^{1/q})$ and $\Psi_{q,\theta}^{\mathbf{u}} = O(1)$. Then $\Phi_{q,\theta}\{\log(|J|n)\}^{3/2}n^{1/q-1/2} \to 0$ implies $|J|\log(|J|n)^{3q/2} = o(n^{q/2-1})$, which further ensures $L_1\max(W_1, W_2) = o(1)\min(N_1, N_2)$. Therefore, Condition 4.5 holds under the requirement $|J|\log(|J|)^{3q/2} = o(n^{q/2-1})$, which is a admissible rate compared with Proposition 4.1.

Let c_α be the $(1-\alpha)$ quantile of $\max_{j\in J}|\rho_{1j}|$, where $(\rho_{1j})_{j\in J}$ are drawn from standard normal distribution. When $\boldsymbol{\mu}_0 = (\mu_{0,ij})_{i,j\in[p]} = (\mathbf{G}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{G})^{-1}$ is known with $\min_{j\in J}\mu_{0,jj} > c$ for some constant c, under Condition 4.5 and the conditions in

Theorem 4.3, we have

$$\lim_{n\to\infty}\left|\Pr\left(\max_{j\in J}\frac{\sqrt{n}|\widehat{b}_j-\beta_{0j}|}{\mu_{0,jj}}\geq c_\alpha\right)-\alpha\right|=0$$

hold with probability $1-o(1)$ (Chernozhukov et al., 2021). When $\boldsymbol{\mu}_0$ is unknown, we shall use a consistent estimation $\widehat{\boldsymbol{\mu}}$ to replace $\boldsymbol{\mu}_0$, and employ a multiplier bootstrap method to estimate the quantifies necessary for the inference. Let $\widehat{u}_{tj}=-(\widehat{\boldsymbol{\mu}}_0\widehat{\boldsymbol{\gamma}}_0)_j g_t(\widehat{\boldsymbol{\beta}})$ and define multiplier bootstrap estimator τ_j by

$$\tau_j=-\frac{1}{\sqrt{n}}\sum_{i=1}^{l_n}\rho_{2i}\sum_{t=(i-1)b_n+1}^{id_n}\widehat{u}_{tj},\ j\in J, \tag{4.13}$$

where $(\rho_{2i})_{i\in l_n}$ are independent standard normal variables, and l_n, d_n are the number of blocks and block size, respectively, with $l_n=\lceil n/d_n\rceil$ and $d_n\to\infty$. In particular, to guarantee the availability of the multiplier bootstrap estimation (4.13), the following conditions on d_n are required.

Condition 4.6. Define $F_\theta=n,$ for $\theta>1-2/q$, $F_\theta=l_n d_n^{q/2-q\theta/2},$ for $1/2-2/q<\theta<1-2/q$, and $F_\theta=l_n^{q/4-q\theta/2}d_n^{q/2-q\theta/2},$ for $\theta<1/2-2/q$, it satisfies that

$$d_n=o\left(n(log|J|)^{-4}(\Psi^{\mathbf{u}}_{q,\theta})^{-4}\wedge n(log|J|)^{-5}(\Psi^{\mathbf{u}}_{4,\theta})^{-4}\right),\ F_\theta=o\left(n^{q/2}(\log|J|)^{-q}|J|^{-1}\Upsilon^{-q}_{q,\theta}\right).$$

$$\Psi^{\mathbf{u}}_{2,0}\Psi^{\mathbf{u}}_{2,\theta}\{d_n^{-1}+\log(n/d_n)n^{-1}+(n-d_n)\log d_n(nd_n)^{-1}\}(\log|J|)^2=o(1),\ \text{if } \theta=1;$$

$$\Psi^{\mathbf{u}}_{2,0}\Psi^{\mathbf{u}}_{2,\theta}\{d_n^{-1}+n^{-\theta}+(n-d_n)d_n^{-\theta+1}(nd_n)^{-1}\}(\log|J|)^2=o(1),\ \text{if } \theta<1;$$

$$\Psi^{\mathbf{u}}_{2,0}\Psi^{\mathbf{u}}_{2,\theta}\{d_n^{-1}+n^{-1}d_n^{-\theta+1}+(n-d_n)(nd_n)^{-1}\}(\log|J|)^2=o(1),\ \text{if } \theta>1.$$

Condition 4.6 controls the rate on which the bootstrap block size d_n diverges with n and $|J|$. Consider the weak dependency case where $1>\theta>1-2/q$ for $\Psi^{\mathbf{u}}_{q,\theta}=O(1)$ and $\Upsilon_{q,\theta}=O(|J|^{1/q})$. Then Condition 4.6 holds for $d_n=o\left(n(\log p)^{-5}\right)$ and $p^2(\log p)^q=o(n^{q/2-1})$. We find that Condition 4.5 and Condition 4.6 are mild and it is valid to assume these two conditions hold at the same time. Then a theorem is provided for the simultaneous confidence intervals of the de-biased estimator.

Theorem 4.4. *Let c^*_α be the $(1-\alpha)$ quantile of $\max_{j\in J}|\tau_j|$. Under Condition* 4.5 *and* 4.6 *and the same conditions of Theorem* 4.3*, and assume that $\Psi^{\mathbf{u}}_{q,\theta}<\infty$ with*

$q > 4$, we have

$$\lim_{n\to\infty}\left|\Pr\left(\max_{j\in J}\frac{\sqrt{n}|\widehat{b}_j-\beta_{0j}|}{\widehat{\mu}_{jj}}\geq c^*_\alpha\right)-\alpha\right|=0.$$

The theorem 4.4 is relying on Theorem 5.1 of Zhang and Wu (2017) and the proof is similar to theorem 5.8 of Chernozhukov et al. (2021).

4.5 Simulation study

In this section, we illustrate the finite-sample properties of our proposed method by simulation studies. Let $\mathbf{X}_t,\ t\in[n]$ generate from a stationary vector autoregressive (VAR) model $\mathbf{X}_t=\mathbf{A}\mathbf{X}_{t-1}+\boldsymbol{\epsilon}_t$ with $\mathbf{A}\in\mathbb{R}^{p\times p}$ and $\boldsymbol{\epsilon}_t$ is a vector independently sampled from multivariate normal distribution $N(\mathbf{0},\boldsymbol{\Sigma}^{\boldsymbol{\epsilon}})$, where $\boldsymbol{\Sigma}^{\boldsymbol{\epsilon}}_{jk}=\rho^{|j-k|}$ for $j,k\in[p]$. Furthermore, we fix the coefficients of interest $\boldsymbol{\beta}_0=(1,1,\mathbf{0}^{\mathrm{T}}_{p-8},0.5,\mathbf{0}^{\mathrm{T}}_5)$ with $\mathbf{0}_k$ being a zero vector of length k. Then the response variables are generated by $Y_t=\mathbf{X}_t^{\mathrm{T}}\boldsymbol{\beta}_0+\varepsilon_t$, where ε_t are independent normally distribution with mean 0 and the variance is chosen to ensure the SNR $=4$. The error contaminated regressors are generated from $W_{tj}=X_{tj}+e_{tj}$, where the observation errors e_{tj} are independently sampling from $N(0,\delta^e\cdot\mathrm{Var}(X_{tj}))$ for $j\in[p]$.

The inference performance is evaluated by computing the average rejection rate on the null hypothesis $H_0^j:\beta_{0j}=0$ for either $j\in\{j:\beta_{0j}=0\}$ and $j\in\{j:\beta_{0j}\neq 0\}$, respectively, as the empirical size and power.

We use five fold cross validation to select $\lambda^{\boldsymbol{\beta}}_n$ for the initial estimator $\widehat{\boldsymbol{\beta}}$ in the problem (4.6). And in the problem (4.10) and (4.11), as in Gold et al. (2020), we choose $\lambda^{\boldsymbol{\gamma}}_{nj}=1.2\cdot\inf_{\boldsymbol{\gamma}_j\in\mathbb{R}^{pL}}|\boldsymbol{\gamma}_j\widehat{\boldsymbol{\Gamma}}-(\widehat{\mathbf{G}}^{\mathrm{T}})_j|_\infty$ and $\lambda^{\boldsymbol{\mu}}_{nj}=1.2\cdot\inf_{\boldsymbol{\gamma}_j\in\mathbb{R}^{pL}}|\boldsymbol{\mu}_j\widehat{\boldsymbol{\gamma}}\widehat{\mathbf{G}}-\mathbf{I}^{\mathrm{T}}_j|_\infty$ for each $j\in[p]$. The corresponding large-scale minimisation problems are solved by the MOSEK optimiser for R (ApS, 2021).

We set the diagonal elements of $\mathbf{A}$ equal to 0.8 and 0 otherwise. Then $\delta^e=0.4,\ 0.8$ are chosen for the cases of moderate and large measurement errors, respectively. And we also check influence from the correlation by setting $\rho=0.5,\ 0.8$. We take $n=200,\ 300$ for $p=100,\ 150$, and repeat each design for 500 times for significance levels at $\alpha=1\%,5\%$ and 10%.

The averaged type I error rates at different significance levels are presented in Table 4.1. We can see that the size are stable and close to the nominal significance levels in all the designs, indicating a good performance of our proposed method.

Table 4.1: Type I error of the hypothesis test at different significance levels.

		$p = 100$				$p = 150$			
		$\rho = 0.5$		$\rho = 0.8$		$\rho = 0.5$		$\rho = 0.8$	
δ^e	α	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$
	1%	0.9%	1.1%	1.1%	1.2%	0.7%	0.9%	0.8%	1.0%
0.4	5%	4.9%	5.3%	4.9%	5.2%	4.2%	4.5%	4.3%	4.8%
	10%	10.0%	10.5%	10.1%	10.1%	9.0%	9.2%	8.9%	9.4%
	1%	0.9%	0.9%	1.1%	1.3%	0.7%	0.7%	0.9%	1.1%
0.8	5%	4.6%	4.8%	5.0%	5.0%	4.1%	4.2%	4.5%	4.7%
	10%	9.6%	9.8%	10.0%	10.0%	8.6%	8.6%	9.1%	9.4%

Table 4.2: Power of the hypothesis test at different significance levels.

		$p = 100$				$p = 150$			
		$\rho = 0.5$		$\rho = 0.8$		$\rho = 0.5$		$\rho = 0.8$	
δ^e	α	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$	$n = 200$	$n = 300$
	1%	75.8%	84.8%	59.5%	66.3%	50.3%	86.7%	37.0%	69.4%
0.4	5%	87.7%	92.8%	72.6%	78.6%	64.4%	95.5%	52.4%	81.8%
	10%	92.7%	95.7%	79.0%	83.4%	72.2%	97.7%	60.5%	88.1%
	1%	66.2%	74.9%	53.9%	59.9%	38.6%	77.4%	31.3%	64.2%
0.8	5%	80.7%	87.0%	67.0%	71.5%	55.0%	87.3%	46.1%	77.6%
	10%	85.6%	91.2%	75.4%	78.2%	64.9%	93.7%	54.5%	82.9%

The power are summarised in Table 4.2 at different significance levels. We observe that power increases when the sample size increases, the variance of measurement error decreases and ρ decreases. And as significance level increases, the higher tolerance of the type I error leads to more powerful test. We can also observe that the empirical size and power slightly deteriorates as p increases.

4.6 Appendix

4.6.1 Technical Proofs

Proof of Theorem 4.3

By applying Proposition 4.1, we have that for $j \in [p]$, $|\widehat{\boldsymbol{\gamma}}_j|_1 \leq |\boldsymbol{\gamma}_{0j}|_1$ and $|\widehat{\boldsymbol{\mu}}_j|_1 \leq |\boldsymbol{\mu}_{0j}|_1$ hold with probability $1 - \delta_n$. Lemma 4.2, 4.3 and 4.4 yields with probability $1 - \delta_n$ that $|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \leq 2C_{\sigma_0,\mu}(2K+1)s\epsilon_n$, $\max_{j\in[p]} |\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j}|_1 \leq 2C_{\sigma_0,\mu}(K+2)s\epsilon_n$ and $\max_{j\in[p]} |\widehat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_{0j}|_1 \leq 2C_{\sigma_0,\mu}(1 \vee K^2)(K+2)s\epsilon_n$. Then, we can bound $\mathbf{r}_{1n}$ and

$\mathbf{r}_{2n}$. It follows from Höder's inequality that $\mathbf{r}_{1n}$ as

$$\begin{aligned}
|\mathbf{r}_{1n}|_\infty &\leq \sqrt{n}|\mathbf{I} - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}}\widehat{\mathbf{G}}|_{\max}|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \\
&\leq \sqrt{n}\max_{j\in[p]} \lambda_{nj}^{\boldsymbol{\mu}} \cdot |\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \\
&\leq 4(1 \vee K^2)(K+2)(2K+1)C_{\sigma_0,\mu}\sqrt{n}s\epsilon_n^2.
\end{aligned}$$

Next, by triangle inequality and Höder's inequality, we have

$$\begin{aligned}
|\mathbf{r}_{2n}|_\infty &\leq \sqrt{n}\max_{j\in[p]} \left(|(\boldsymbol{\mu}_{0j} - \widehat{\boldsymbol{\mu}}_j)\boldsymbol{\gamma}_0\widehat{\mathbf{g}}(\boldsymbol{\beta}_0)| + |\widehat{\boldsymbol{\mu}}_j(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}})\widehat{\mathbf{g}}(\boldsymbol{\beta}_0)|\right) \\
&\leq \sqrt{n}\max_{j\in[p]} |\boldsymbol{\mu}_{0j} - \widehat{\boldsymbol{\mu}}_j|_1|\boldsymbol{\gamma}_0\widehat{\mathbf{g}}(\boldsymbol{\beta}_0)|_\infty \\
&\qquad + \sqrt{n}\max_{j\in[p]} |\widehat{\boldsymbol{\mu}}_j|_1|(\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}})\widehat{\mathbf{g}}(\boldsymbol{\beta}_0)|_\infty \\
&\leq \sqrt{n}\max_{j\in[p]} |\boldsymbol{\gamma}_{0j}|_1 \max_{j\in[p]} |\boldsymbol{\mu}_{0j} - \widehat{\boldsymbol{\mu}}_j|_1|\widehat{\mathbf{g}}(\boldsymbol{\beta}_0)|_\infty \\
&\qquad + \sqrt{n}\max_{j\in[p]} |\widehat{\boldsymbol{\mu}}_{0j}|_1 \max_{j\in[p]} |\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0j}|_1|\widehat{\mathbf{g}}(\boldsymbol{\beta}_0)|_\infty \\
&\leq 6K(1 \vee K^2)(K+2)(2K+1)C_{\sigma_0,\mu}\sqrt{n}s\epsilon_n^2.
\end{aligned}$$

Therefore, by applying Proposition 4.1, we have

$$|\mathbf{r}_n|_\infty \leq |\mathbf{r}_{1n}|_\infty + |\mathbf{r}_{2n}|_\infty \leq C_K C_{\sigma_0,\mu}\sqrt{n}s\epsilon_n^2,$$

where $C_K = 2(1 \vee K^2)(K+2)(2K+1)(3K+2)$. □

Proof of Lemma 4.1

It hold from Condition 4.2 that with probability $1 - \delta_n$ we have

$$\begin{aligned}
|\widehat{\mathbf{g}}(\boldsymbol{\beta}_0)|_\infty &= |\widehat{\mathbf{g}}(\boldsymbol{\beta}_0) - \mathbf{g}(\boldsymbol{\beta}_0)|_\infty \\
&\leq |(\widehat{\mathbf{G}} - \mathbf{G})\boldsymbol{\beta}_0|_\infty + |\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})|_\infty \\
&\leq |\widehat{\mathbf{G}} - \mathbf{G}|_{\max}|\boldsymbol{\beta}_0|_1 + |\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})|_\infty, \\
&\leq K\epsilon_n^{\mathbf{G}} + \epsilon_n^{\mathbf{g}} \leq \lambda_n^{\boldsymbol{\beta}}. \square
\end{aligned}$$

Proof of Lemma 4.2

Consider event $A = \big\{|\widehat{\mathbf{G}} - \mathbf{G}|_{\max} \leq \epsilon_n^{\mathbf{G}} \text{ and } |\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})|_\infty \leq \epsilon_n^{\mathbf{g}}\big\} \cap \big\{|\widehat{\mathbf{g}}(\boldsymbol{\beta}_0)|_{\max} \leq \lambda_n^{\boldsymbol{\beta}}\big\}$. Under Condition 4.2 and by applying Lemma 4.1, this event occurs with probability $1 - \delta_n$. On event A, we have

$$\begin{aligned} |\mathbf{G}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)|_\infty &\leq |\mathbf{g}(\widehat{\boldsymbol{\beta}})|_\infty \\ &\leq |\widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}}) - \mathbf{g}(\widehat{\boldsymbol{\beta}})|_\infty + |\widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}})|_\infty \\ &\leq |(\widehat{\mathbf{G}} - \mathbf{G})\widehat{\boldsymbol{\beta}}|_\infty + |\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})|_\infty + |\widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}})|_\infty \\ &\leq |\boldsymbol{\beta}_0|_1 |\widehat{\mathbf{G}} - \mathbf{G}|_{\max} + |\widehat{\mathbf{g}}(\mathbf{0}) - \mathbf{g}(\mathbf{0})|_\infty + |\widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}})|_\infty \\ &\leq K\epsilon_n^{\mathbf{G}} + \epsilon_n^{\mathbf{g}} + \lambda_n^{\boldsymbol{\beta}} \leq 2\epsilon_n^{\boldsymbol{\beta}}, \end{aligned} \tag{4.14}$$

where, in the last two inequalities, we have used the Hölder's inequality and the facts that $|\widehat{\boldsymbol{\beta}}|_1 \leq |\boldsymbol{\beta}_0|_1 \leq K$ and $|\widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}})|_{\max} \leq \lambda_n^{\boldsymbol{\beta}}$ by the definition of the RMD estimator in (4.6).

On event A, choosing the set $T = S$ in (4.12) and applying Lemma 4.1 and 4.5 yields $|\widehat{\boldsymbol{\delta}}_{S^C}|_1 \leq |\widehat{\boldsymbol{\delta}}_S|_1$ and hence $\widehat{\boldsymbol{\delta}} \in C_S$. Then by (4.12), (4.14) and Lemma 4.7 under Condition 4.3, we have

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \leq \kappa(\boldsymbol{\beta}_0)^{-1} \cdot |\mathbf{G}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)|_\infty \leq 2C_{\sigma_0,\mu} s\epsilon^{\boldsymbol{\beta}},$$

which completes the proof. □

Proof of Lemma 4.3

The proof follows the same idea as Theorem 4.2. We observe that problem (4.10) fit into (4.6) by redefining $\widehat{\mathbf{G}} = \widehat{\boldsymbol{\Gamma}}$ and $\widehat{\mathbf{g}}(\mathbf{0}) = (\widehat{\mathbf{G}}^{\mathrm{T}})_j$. Note that Condition 4.3 and Condition 4.4, as a counterpart of Condition 4.2, are assumed for (4.10). Next we need to verify the feasibility of the regularisation parameter $\lambda_{nj}^{\boldsymbol{\gamma}}$ as shown in Lemma 4.1. Under Condition 4.4, with probability $1 - \delta_n$ we have

$$\begin{aligned} |\boldsymbol{\gamma}_{0j}\widehat{\boldsymbol{\Gamma}} - (\widehat{\mathbf{G}}^{\mathrm{T}})_j|_\infty &\leq |\boldsymbol{\gamma}_{0j}\boldsymbol{\Gamma} - (\mathbf{G}^{\mathrm{T}})_j|_\infty + |\boldsymbol{\gamma}_{0j}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})|_\infty + |(\mathbf{G}^{\mathrm{T}})_j - (\widehat{\mathbf{G}}^{\mathrm{T}})_j|_\infty \\ &\leq 0 + |\boldsymbol{\gamma}_{0j}|_1 |\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}|_{\max} + \epsilon_n^{\mathbf{G}} \\ &\leq K\epsilon_n^{\boldsymbol{\Gamma}} + \epsilon_n^{\mathbf{G}} \end{aligned}$$

Therefore, if $\lambda_j^{\boldsymbol{\gamma}} \geq K\epsilon_n^{\boldsymbol{\Gamma}} + \epsilon_n^{\mathbf{G}}$, $\boldsymbol{\gamma}_{0j}$ is feasible for all $j \in [p]$. Thus, the rest of steps to

prove Lemma 4.3 follows directly from that in Lemma 4.2. □

Proof of Lemma 4.4

We observe that problem (4.11) fit into (4.6) by redefining $\widehat{\mathbf{G}} = \widehat{\boldsymbol{\gamma}}\widehat{\mathbf{G}}$ and $\widehat{\mathbf{g}}(\mathbf{0}) = \mathbf{I}_j$. Note that Condition 4.3 and Condition 4.4, as a counterpart of Condition 4.2, are assumed for (4.11). Next we need to verify the feasibility of the regularisation parameter $\lambda_{nj}^{\boldsymbol{\mu}}$ as shown in Lemma 4.1. Under Condition 4.4, we have with probability greater than $1 - \delta_n$ that

$$
\begin{aligned}
|\widehat{\boldsymbol{\gamma}}\widehat{\mathbf{G}} - \boldsymbol{\gamma}\mathbf{G}|_{\max} &\leq |\widehat{\boldsymbol{\gamma}}(\widehat{\mathbf{G}} - \mathbf{G})|_{\max} + |(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\mathbf{G}|_{\max} \\
&\leq \max_{j\in[p]} |\boldsymbol{\gamma}_{0j}|_1 |\widehat{\mathbf{G}} - \mathbf{G}|_{\max} + |\widehat{\boldsymbol{\gamma}}\boldsymbol{\Gamma}\boldsymbol{\gamma}_0^{\mathrm{T}} - \boldsymbol{\gamma}_0\boldsymbol{\Gamma}\boldsymbol{\gamma}_0^{\mathrm{T}}|_{\max} \\
&\leq K\epsilon_n^{\mathbf{G}} + |\widehat{\boldsymbol{\gamma}}\boldsymbol{\Gamma}\boldsymbol{\gamma}_0^{\mathrm{T}} - \widehat{\boldsymbol{\gamma}}\widehat{\boldsymbol{\Gamma}}\boldsymbol{\gamma}_0^{\mathrm{T}} + \widehat{\boldsymbol{\gamma}}\widehat{\boldsymbol{\Gamma}}\boldsymbol{\gamma}_0^{\mathrm{T}} - \mathbf{G}^{\mathrm{T}}\boldsymbol{\gamma}_0^{\mathrm{T}}|_{\max} \\
&\leq K\epsilon_n^{\mathbf{G}} + |\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})\boldsymbol{\gamma}_0^{\mathrm{T}}|_{\max} + |(\widehat{\boldsymbol{\gamma}}\widehat{\boldsymbol{\Gamma}} - \widehat{\mathbf{G}}^{\mathrm{T}} + \widehat{\mathbf{G}}^{\mathrm{T}} - \mathbf{G}^{\mathrm{T}})\boldsymbol{\gamma}_0^{\mathrm{T}}|_{\max} \\
&\leq K\epsilon_n^{\mathbf{G}} + K\left\{|\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})|_{\max} + |\widehat{\boldsymbol{\gamma}}\widehat{\boldsymbol{\Gamma}} - \widehat{\mathbf{G}}^{\mathrm{T}}|_{\max} + |\widehat{\mathbf{G}} - \mathbf{G}|_{\max}\right\} \\
&\leq K\epsilon_n^{\mathbf{G}} + K\max_{j\in[p]} |\boldsymbol{\gamma}_{0j}|_1 |\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}|_{\max} + K\max_{j\in[p]} |\boldsymbol{\gamma}_{0j}\widehat{\boldsymbol{\Gamma}} - (\widehat{\mathbf{G}}^{\mathrm{T}})_j|_\infty \\
&\quad + K|\widehat{\mathbf{G}} - \mathbf{G}|_{\max} \\
&\leq 2K\epsilon_n^{\mathbf{G}} + K^2\epsilon_n^{\boldsymbol{\Gamma}} + K\max_{j\in[p]} \lambda_{nj}^{\boldsymbol{\gamma}},
\end{aligned}
$$

where we used $\mathbf{G} = \boldsymbol{\Gamma}\boldsymbol{\gamma}_0^{\mathrm{T}}$. Therefore, $\boldsymbol{\mu}_{0j}$ is feasible for all $j \in [p]$, if $\lambda_j^{\boldsymbol{\mu}} \geq 2K^2\epsilon_n^{\mathbf{G}} + K^3\epsilon_n^{\boldsymbol{\Gamma}} + K^2 \max_{j\in[p]} \lambda_{nj}^{\boldsymbol{\gamma}}$. Thus, the rest of steps to prove Lemma 4.4 follows directly from that in Lemma 4.2. □

Lemma 4.5 and its proof

Lemma 4.5. *Suppose that Condition* 4.2 *holds. Then* $|\widehat{\boldsymbol{\delta}}_{S^c}|_1 \leq |\widehat{\boldsymbol{\delta}}_S|_1$ *with probability* $1 - \delta_n$.

Proof. It follows from Lemma 4.1 under Condition 4.2 and $\boldsymbol{\beta}_{0,S^c} = \mathbf{0}$ by definition that with probability $1 - \delta_n$, $|\widehat{\boldsymbol{\beta}}|_1 \leq |\boldsymbol{\beta}_0|_1 = |\boldsymbol{\beta}_{0,S}|_1$, which implies that

$$
\begin{aligned}
|\boldsymbol{\beta}_{0,S}|_1 &\geq |\widehat{\boldsymbol{\beta}}_S|_1 + |\widehat{\boldsymbol{\beta}}_{S^c}|_1 \\
&\geq |\boldsymbol{\beta}_{0,S}|_1 - |\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}|_1 + |\widehat{\boldsymbol{\beta}}_{S^c}|_1 .
\end{aligned}
$$

By cancelling $|\boldsymbol{\beta}_{0,S}|_1$ on both sides above, we obtain $|\widehat{\boldsymbol{\beta}}_{S^c} - \boldsymbol{\beta}_{0,S^c}|_1 \leq |\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0,S}|_1$. □

Lemma 4.6 and its proof

Lemma 4.6. *It holds that*

$$\kappa(\boldsymbol{\beta}_0, \mathbf{G}) \geq \max_{m \geq s} \left\{ \frac{\sigma_{\min}(m, \mathbf{G})}{\sqrt{m}} - \frac{2\sigma_{\max}(m, \mathbf{G})}{\sqrt{m}} \sqrt{\frac{s}{m}} \right\} \frac{s^{-1/2}}{2(1 + 2\sqrt{s/m})}.$$

Proof. This lemma follows directly from Theorem 1 of Belloni et al. (2019) and hence the proof is omitted here. □

Lemma 4.7 and its proof

Lemma 4.7. *Suppose that Condition* 4.3 *holds. Then there exists some constant* $C_{\sigma_0,\mu}$ *denpending on* σ_0 *and* μ *such that*

$$\kappa(\boldsymbol{\beta}_0, \mathbf{G}) \geq C_{\sigma_0,\mu}^{-1} s^{-1}.$$

Proof. Applying Lemma 4.6 under Condition 4.3 yields that

$$\begin{aligned} \kappa(\boldsymbol{\beta}_0, \mathbf{G}) \geq & \max_{m \geq s} \frac{\sigma_{\max}(m, \mathbf{G})}{\sqrt{m}} \left\{ \frac{\sigma_{\min}(m, \mathbf{G})}{\sigma_{\max}(m, \mathbf{G})} - \frac{\mu}{2} \right\} \frac{s^{-1/2}}{2(1 + \mu/2)} \\ & \geq \frac{\sigma_0 \mu}{4\sqrt{s}} (\mu - \frac{\mu}{2}) [2(1 + \frac{\mu}{2})]^{-1} s^{-1/2} \geq \frac{\sigma_0 \mu^2}{24s} \,. \, \square \end{aligned}$$

Lemma 4.8 and its proof

Lemma 4.8. *According to* (4.9)*, we have*

$$\widehat{\mathbf{b}} = \boldsymbol{\beta}_0 - \boldsymbol{\mu}_0 \gamma_0 \widehat{\mathbf{g}}(\boldsymbol{\beta}_0) + (\mathbf{I} - \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\gamma}} \widehat{\mathbf{G}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\boldsymbol{\mu}_0 \gamma_0 - \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\gamma}}) \widehat{\mathbf{g}}(\boldsymbol{\beta}_0).$$

Proof. By elementary expansion and some algebra, we have

$$\begin{aligned}
\widehat{\mathbf{b}} &= \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}}\widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}}) \\
&= \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 + \boldsymbol{\beta}_0 - \boldsymbol{\mu}_0\boldsymbol{\gamma}_0\widehat{\mathbf{g}}(\boldsymbol{\beta}_0) + \boldsymbol{\mu}_0\boldsymbol{\gamma}_0\widehat{\mathbf{g}}(\boldsymbol{\beta}_0) - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}}\widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}}) \\
&= \boldsymbol{\beta}_0 - \boldsymbol{\mu}_0\boldsymbol{\gamma}_0\widehat{\mathbf{g}}(\boldsymbol{\beta}_0) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\boldsymbol{\mu}_0\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}})\widehat{\mathbf{g}}(\boldsymbol{\beta}_0) + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}}\left(\widehat{\mathbf{g}}(\boldsymbol{\beta}_0) - \widehat{\mathbf{g}}(\widehat{\boldsymbol{\beta}})\right) \\
&= \boldsymbol{\beta}_0 - \boldsymbol{\mu}_0\boldsymbol{\gamma}_0\widehat{\mathbf{g}}(\boldsymbol{\beta}_0) + (\boldsymbol{\mu}_0\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}})\widehat{\mathbf{g}}(\boldsymbol{\beta}_0) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}}(\widehat{\mathbf{G}}\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{G}}\boldsymbol{\beta}_0) \\
&= \boldsymbol{\beta}_0 - \boldsymbol{\mu}_0\boldsymbol{\gamma}_0\widehat{\mathbf{g}}(\boldsymbol{\beta}_0) + (\mathbf{I} - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}}\widehat{\mathbf{G}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\boldsymbol{\mu}_0\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\gamma}})\widehat{\mathbf{g}}(\boldsymbol{\beta}_0).\square
\end{aligned}$$

4.6.2 Some useful lemmas

Lemma 4.9. *(Proposition 3.3 of Zhang and Wu (2021)). Let $\mathbf{X}_t$ be a zero mean stationary process of the form (4.4), which satisfies $\Psi^{\mathbf{X}}_{4,\theta} < \infty$ and $\Theta^{\mathbf{X}}_{q,\theta} < \infty$ for some $q > 4$ and $\theta > 0$. Then there exists absolute constant C, constant C_θ only depending on θ and constant $C_{q,\theta}$ only depending on q and θ such that for any $\eta > 0$,*

$$\begin{aligned}
\Pr\left(\max_{0\le h\le L} |\widehat{\boldsymbol{\Sigma}}^{\mathbf{X}}_h - \mathbb{E}\widehat{\boldsymbol{\Sigma}}^{\mathbf{X}}_h|_{\max} \ge \eta\right) \le & \frac{C_{q,\theta} H_{n,L}(\log p \||\mathbf{X}_\cdot|_\infty\|_{q,\theta} \wedge \Theta_{q,\theta})^q}{(n\eta)^{q/2}} \\
& + C(L+1)p^2 \exp\left\{-\frac{n\eta^2}{C_\theta(\Psi^{\mathbf{X}}_{4,\theta})^4}\right\},
\end{aligned}$$

where $H_{n,L} = (L+1)^{q/4}n$ for $\theta > 1/2 - 2/q$ and $H_{n,L} = (L+1)^{q/4}n + (L+1)n^{q/4-\theta q/2}$ for $\theta < 1/2 - 2/q$.

Lemma 4.10. *(Corollary 3.4 of Zhang and Wu (2021)). Under the conditions in Lemma 4.9, we have*

$$\max_{0\le h\le L} |\mathbb{E}\widehat{\boldsymbol{\Sigma}}^{\mathbf{X}}_h - \boldsymbol{\Sigma}^{\mathbf{X}}_h|_{\max} = O_p\left\{\frac{(1 + (L+1)^{-\theta+1})\Psi^{\mathbf{X}}_{2,0}\Psi^{\mathbf{X}}_{2,\theta}}{n}\right\}.$$

Lemma 4.11. *(Theorem 6.2 of Zhang and Wu (2017)). Let $\mathbf{S}_n = \sum_{t=1}^n \mathbf{X}_t$, where $\mathbf{X}_t \in \mathbb{R}^p$ is a zero mean process of the form (4.4), satisfying $\||\mathbf{X}_\cdot|_\infty\|_{q,\theta} \le \infty$ and*

$\Psi^{\mathbf{X}}_{2,\theta} < \infty$ *with* $q > 2$ *and* $\theta > 0$. *(i) If* $\theta > 1/2 - 1/q$, *then for* $\eta \gtrsim (n \log p)^{1/2} \Psi^{\mathbf{X}}_{2,\theta} + n^{1/q} (\log p)^{3/2} \| |\mathbf{X}_\cdot|_\infty \|_{q,\theta}$,

$$\Pr(|\mathbf{S}_n|_\infty \geq \eta) \leq \frac{C_{q,\theta} n (\log p)^{q/2} \| |\mathbf{X}_\cdot|_\infty \|^q_{q,\theta}}{\eta^q} + C_{q,\theta} \exp\left\{ -\frac{C_{q,\theta} \eta^2}{n (\Psi^{\mathbf{X}}_{2,\theta})^2} \right\}.$$

(ii) If $\theta < 1/2 - 1/q$, *then for* $\eta \gtrsim (n \log p)^{1/2} \Psi^{\mathbf{X}}_{2,\theta} + n^{1/2-\theta} (\log p)^{3/2} \| |\mathbf{X}_\cdot|_\infty \|_{q,\theta}$,

$$\Pr(|\mathbf{S}_n|_\infty \geq \eta) \leq \frac{C_{q,\theta} n^{q/2-\theta q} (\log p)^{q/2} \| |\mathbf{X}_\cdot|_\infty \|^q_{q,\theta}}{\eta^q} + C_{q,\theta} \exp\left\{ -\frac{C_{q,\theta} \eta^2}{n (\Psi^{\mathbf{X}}_{2,\theta})^2} \right\}.$$

Bibliography

Adamek, R., Smeekes, S., and Wilms, I. (2020). Lasso Inference for High-Dimensional Time Series. *arXiv:2007.10952*, pages 1–49.

ApS, M. (2021). *MOSEK Rmosek package 9.2.49*.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.

Aue, A., Norinho, D., and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110:378–392.

Aue, A. and van Delft, A. (2020). Testing for stationarity of functional time series in the frequency domain. *The Annals of Statistics*, To appear.

Babii, A. (2020). Honest confidence sets in nonparametric iv regression and other ill-posed models. *Econometric Theory*, 36(4):658–706.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43:1535–1567.

Bathia, N., Yao, Q., and Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, 38:3352–3386.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018). High-dimensional econometrics and regularized GMM. *arXiv:1806.01888v2*.

Belloni, A., Hansen, C., and Newey, W. (2019). Simultaneous confidence intervals for high-dimensional linear models with many endogenous variables. *arXiv:1712.08102v4*.

Bergmeir, C., Hyndman, R., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis*, 120:70–83.

Borovskikh, Y. V. (1996). *U Statistics in Banach Spaces*. VSP, Netherlands.

Bosq, D. (2000). *Linear Processes in Function Spaces*. Springer, New York.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press, New Jersey.

Candes, E. and Tao, T. (2007). The dantzig selection: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351.

Caner, M. and Kock, A. B. (2018). High dimensional linear gmm. *arXiv:1811.08779*.

Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003a). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics*, 30(1):241–255.

Cardot, H., Ferraty, F., and Sarda, P. (2003b). Splines estimators for the functional linear model. *Statistica Sinica*, 13:571–591.

Chakraborty, A. and Panaretos, V. M. (2017). Regression with genuinely functional errors-in-covariates. *arXiv:1712.04290.*

Chen, C., Guo, S., and Qiao, X. (2020). Functional linear regression: dependence and error contamination. *Journal of Business & Economic Statistics*, To appear.

Chernozhukov, V., Härdle, W. K., Huang, C., and Wang, W. (2021). Lasso-driven inference in time and space. *The Annals of Statistics*, 49(3):1702–1735.

Cho, H., Goude, Y., Brossat, X., and Yao, Q. (2013). Modelling and forecasting daily electricity load curves: a hybrid approach. *Journal of the American Statistical Association*, 108:7–21.

Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37:35–72.

Datta, A. and Zou, H. (2017). Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):2400–2426.

Descary, M.-H. and Panaretos, V. M. (2019). Functional data analysis by matrix completion. *Annals of Statistics*, 47:1–38.

Fan, J. and Liao, Y. (2014). Endogeneity in high dimensions. *The Annals of Statistics*, 42(3):872–917.

Fan, Y., Foutz, N., James, G., and Jank, W. (2014). Functional forecasting of demand decay rates using online virtual stock markets. *The Annals of Applied Statistics*, 8:2435–2460.

Fan, Y., James, G., and Radchenko, P. (2015). Functional additive regression. *The Annals of Statistics*, 43:2296–2325.

Fang, Q., Guo, S., and Qiao, X. (2020). A new perspective on dependence in high-dimensional functional/scalar time series: finite sample theory and applications. *arXiv:2004.07781*.

Fu, A., Narasimhan, B., and Boyd, S. (2020). CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, To appear.

Gautier, E. and Rose, C. (2019). High-dimensional instrumental variables regression and confidence sets. *arXiv:1105.2454v6*.

Gold, D., Lederer, J., and Tao, J. (2020). Inference for high-dimensional instrumental variables regression. *Journal of Econometrics*, 217(1):79–111.

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Kobe, R. (2013). Modeling complex spatial dependencies: low rank spatially varying cross-covariances with application to soil nutrient data. *Journal of Agricultural, Biological and Environmental Statistics*, 18:274–298.

Guo, S. and Qiao, X. (2020). On consistency and sparsity for high-dimensional functional time series with application to autoregressions. *arXiv:2003.11462*.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35:70–91.

Hall, P. and Vial, C. (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B*, 68:689–705.

Hamilton, J. D. (1994). *Time series analysis*, volume 2. Princeton New Jersey.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24:726–748.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054.

He, G., Mueller, H. G., Wang, J. L., and Yang, W. (2010). Functional linear regression via canonical analysis. *Bernoulli*, 16:705–729.

Hörmann, S., Kidzinski, L., and Hallin, M. (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B*, 77:319–348.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, 38:1845–1884.

Horváth, L., Kokoszka, P., and Rice, G. (2014). Testing stationary of functional time series. *Journal of Econometrics*, 179:66–82.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, Chichester.

Imaizumi, M. and Kato, K. (2019). A simple method to construct confidence bands in functional linear regression. *Statistica Sinica*, 29(4):2055–2081.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

Jirak, M. (2016). Optimal eigen expansions and uniform bounds. *Probability Theory and Related Fields*, 166:753–799.

Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI*, 34.

Kokoszka, P., Rice, G., and Shang, H. L. (2017). Inference for the autocovariance of a functional time series under conditional heteroscedasticity. *Journal of Multivariate Analysis*, 162:32–50.

Kong, D., Staicu, A.-M., and Maity, A. (2016a). Classical testing in functional linear models. *Journal of nonparametric statistics*, 28(4):813–838.

Kong, D., Xue, K., Yao, F., and Zhang, H. (2016b). Partially functional linear regression in high dimensions. *Biometrika*, 103:147–159.

Kutta, T., Dierickx, G., and Dette, H. (2021). Statistical inference for the slope parameter in functional linear regression. *arXiv:2108.07098*.

Lam, C. and Yao, Q. (2012). Factor modelling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40:694–726.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98:901–918.

Li, B. and Solea, E. (2018). A nonparametric graphical model for functional data with application to brain networks based on fmri. *Journal of the American Statistical Association*, 113(524):1637–1655.

Li, D., Robinson, P. M., and Shang, H. L. (2020). Long-range dependent curve time series. *Journal of the American Statistical Association*, 115(530):957–971.

Li, M., Li, R., and Ma, Y. (2021). Inference in high dimensional linear measurement error models. *Journal of Multivariate Analysis*, 184:104759.

Loève, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique*, 84:159–162.

Luo, R. and Qi, X. (2017). Function-on-function linear regression by signal compression. *Journal of the American Statistical Association*, 112:690–705.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359.

Müller, H., Sen, R., and Stadtmüller, U. (2011). Functional data analysis for volatility. *Journal of Econometrics*, 165:233–245.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45:158–195.

Panaretos, V. and Tavakoli, S. (2013). Fourier analysis of stationary time series in function space. *The Annals of Statistics*, 41:568–603.

Qiao, X., Qian, C., James, G., and Guo, S. (2020). Doubly functional graphical models in high dimensions. *Biometrika*, 107:415–431.

Radchenko, P., Qiao, X., and James, G. (2015). Index models for sparsely sampled functional data. *Journal of the American Statistical Association*, 110:824–836.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis (2nd ed.)*. Springer, New York.

Rubín, T. and Panaretos, V. M. (2020). Sparsely observed functional time series: Estimation and prediction. *Electronic Journal of Statistics*, 14:1137–1210.

Rudelson, M. and Vershynin, R. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9.

Shang, L. H. (2013). ftsa : An r package for analyzing functional time series. *The R Journal*, 5(1):64–72.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tsay, R. (2013). *Multivariate Time Series Analysis: With R and Financial Applications*. Wiley Series in Probability and Statistics. Wiley.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge.

Wang, Y., Wang, J., Balakrishnan, S., and Singh, A. (2019). Rate optimal estimation and confidence intervals for high-dimensional regression with missing covariates. *Journal of Multivariate Analysis*, 174:104526.

Wecker, W. E. (1978). A note on the time series which is the product of two stationary time series. *Stochastic Processes and their Applications*, 8(2):153–157.

Wharton Research Data Services (n.d.). "TAQ". `http://wrds.wharton.upenn.edu`.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154.

Xia, Q., Xu, W., and Zhu, L. (2015). Consistently determining the number of factors in multivariate volatility modelling. *Statistica Sinica*, pages 1025–1044.

Xue, K. and Yao, F. (2020). Hypothesis testing in large-scale functional linear regression. *Statistica Sinica*, To appear.

Yao, F., Müller, H. G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590.

Yousuf, K. (2018). Variable screening for high dimensional time series. *Electronic Journal of Statistics*, 12(1):667–702.

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

Zhang, D. and Wu, W. B. (2017). Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45:1895–1919.

Zhang, D. and Wu, W. B. (2021). Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *The Annals of Statistics*, 49(1):233–254.

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics*, 44:2281–2321.