Spatial Lag Model with Time-lagged Effects and Spatial Weight Matrix Estimation

Clifford Lam*[1] and Cheng Qian†[1]

[1] *Department of Statistics, London School of Economics and Political Science, UK*

Abstract

This paper considers a spatial lag model with different spatial weight matrices for different time-lagged spatial effects, while allowing both the sample size T and the panel dimension N to grow to infinity together. To overcome potential misspecifications of these spatial weight matrices, we estimate each one by a linear combination of a set of M specified spatial weight matrices, with M being finite. Moveover, by penalizing on the coefficients of these linear combinations, oracle properties for these penalized coefficient estimators are proved, including their asymptotic normality and sign consistency. Other parameters of the model are estimated by profile-least square type of estimators after introducing covariates which serve similar functions as instrumental variables. Asymptotic normality for our estimators are developed under the framework of functional dependence used in Wu (2011), which is a measure of time series dependence. The proposed methods are illustrated using both simulated and real financial data.

Key words: Spatial econometrics; Spatial dynamic effect; Profile least square estimation; Adaptive lasso; Spatial sign consistency; Functional dependence measurement.

JEL classification: C31; C33.

*Clifford Lam is Associate Professor, Department of Statistics, London School of Economics. Email: C.Lam2@lse.ac.uk

†Cheng Qian is PhD student, Department of Statistics, London School of Economics. Email: C.Qian2@lse.ac.uk

1 Introduction

Spatial econometrics focus on models which allow for cross-sectional interactions among units under study. As it is easy to obtain panel data nowadays, the study of spatial econometrics is becoming more important. Since Cliff and Ord (1973), economists have investigated cross-sectional models, especially spatial autoregressive (SAR), or spatial lag models. Anselin et al. (2008) defined four types of spatial dynamic models. The first type is "pure space recursive" if only a spatial time lag is included. The second type is "time-space recursive" if both an individual time lag and a spatial time lag are included. The third type is "time-space simultaneous" if an individual time lag and a contemporaneous spatial lag are specified. And finally, "time-space dynamic" if all forms of lags are included. Elhorst (2005) considered a dynamic panel data model in spatial disturbance. Lee and Yu (2010) established asymptotic properties of quasi-maximum likelihood estimators for SAR panel data models with fixed effects and SAR disturbances. Based on these developments, spatial dynamic panel models can be used in regional markets in Keller and Shiue (2007), labour economics in Foote (2007) or public economics in Franzese and Hays (2007), to name but a few areas.

However, many estimation methods rely on the assumption that the spatial weight matrix, which measures the strength of interactions among units, is known. Applied researchers may specify a spatial weight matrix based on certain distance measures, for instance the contiguity of units. Overall, the choice of the spatial weight matrix very much depends on individual specifications. Even for a simple distance r between two units, we can specify an entry in the spatial weight matrix using r^{-1}, r^{-2} or r^{-3}. There are actually infinite possibilities, and it may be that certain such specifications are important while the others are not (see our real data analysis in Section 6, for instance). Because of this, spatial econometrics is often criticized, like that in Corrado and Fingleton (2012). Recently, Lam and Souza (2014) has provided an error upper bound for the spatial regression parameter estimators in a spatial lag model, showing that misspecification of the spatial weight matrix can indeed introduce large bias in the final estimates. To avoid such misspecification, non-parametric models are considered in past researches, see Tran and Yakowitz (1993) and Hallin et al. (2004) for instance. The Nadaraya-Watson kernel estimator is frequently used for nonparametric regression in econometrics. Robinson (2011) established consistency and asymptotic distribution theory for the Nadaraya-Watson estimator in a framework designed for various kinds of spatial data. Koroglu and Sun (2016) improved estimation accuracy by applying a nonparametric two-stage least squares estimation method. More specifically, the second-step estimator of the unknown functional coefficients are estimated by local linear regression. However, Kostov (2013) shows that it can lead to reduced efficiency of the estimators when the sample size is small.

With these drawbacks in mind, Bhattacharjee and Jensen-Butler (2013) proposes to estimate such a spatial weight matrix with a symmetric assumption, while Lam and Souza (2016) proposes to estimate the block pattern in such a matrix. With the development of high dimensional statistics, Ahrens and Bhattacharjee (2015) considers a two-step lasso estimation, which is based on the sparsity assumption of the spatial weight matrix. Meanwhile, adaptive lasso is used in Lam and Souza (2015) to estimate a sparse spatial weight matrix together with fixed effects in the spatial lag model. All these models do not consider time-lagged effects, however. The difficulty certainly lies in the fact that more than one spatial weight matrices have to be considered when time lags are added in the model, and hence such a time-space dynamic form has attracted little attentions so far.

Motivated by the evidence in our data example, our model includes pure dynamic effects and time lags simultaneously, while each spatial weight matrix involved in the model is estimated by a linear combination of user-specified spatial weight matrices. It helps avoid the risk of misspecification of the spatial weight matrices, while maintaining the overall parsimony of the model. As for the innate endogeneity in our dynamic spatial lag model, the direct least square estimation will result in inconsistent estimators. To overcome this difficulty, we introduce instrumental-like variables. In the particular case when the covariates are exogenous, they themselves can act as these instrumental-like variables. We estimate the "best" linear combination for each required spatial weight matrix, highlighting the relative contributions of each specified one. Asymptotic normality of all estimators are presented under the functional dependence measure of time series variables in Wu (2005) or Wu (2011), allowing both the sample size T and the panel size N grow to infinity together. With the input of different specified spatial weight matrices, the scope of applications of our model is expanded since there are many applications where there are numerous ways to specify a spatial weight matrix. See our theoretical results in Section 3.3.

The rest of the paper is organized as follows. Section 2 introduces our methodology, including the model and the estimation method. Properties of our estimators, including asymptotic normality are presented in Section 3. Simulation results and real data analysis are reported, respectively, in Section 4 and 5. All the technical proofs are relegated to the Appendix.

2 Methodology

2.1 The Model

Consider the following dynamic spatial lag model

$$y_t = \boldsymbol{\mu} + \boldsymbol{W}_0 y_t + \boldsymbol{W}_1 y_{t-1} + \cdots + \boldsymbol{W}_p y_{t-p} + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{2.1}$$

where $y_t = (y_{t1}, y_{t2}, ..., y_{tN})^T$ is an $N \times 1$ vector of observed time series variables. The data starts from y_{1-p}, and hence the true sample size is $T + p$. It does not affect our asymptotic analysis since p is finite in this paper. Hereafter when we talk about the sample size, we use T instead of $T + p$ for simplicity. For $j = 0, 1, \ldots, p$, $\boldsymbol{W}_j$ is an $N \times N$ spatial weight matrix with 0 on the main diagonal, and $\boldsymbol{\mu}$ is an $N \times 1$ constant vector. The $N \times K$ matrix of covariates $\boldsymbol{X}_t$ can contain y_{t-j} for $j = 1, \ldots, p$ in its columns on top of other covariates, while $\boldsymbol{\beta}$ is the $K \times 1$ vector of regression coefficients. The series $\{\boldsymbol{\epsilon}_t\}$ is an innovation process with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\epsilon$.

In many applied spatial econometrics applications, $\boldsymbol{W}_0$ is assumed known and there are no lagged terms $\boldsymbol{W}_j y_{t-j}$ for $j = 1, \ldots, p$. Instead of assuming all the spatial weight matrices are known, in this paper we assume that there are M specified spatial weight matrices $\boldsymbol{W}_{0i}$, $i = 1, \ldots, M$, such that each spatial weight matrix is a linear combination of the M specified ones. This is motivated by the fact that there are often more than one measures of spatial interactions. For instance, for the geographical distance r alone between two specific locations, we can specify three different entries r^{-1}, r^{-2} and r^{-3}, creating three specified spatial weight matrices. These are indeed our distance specifications included in our data application in Section 6. Spatial contiguity is also another popular choice in spatial econometrics. The linear combination

for each $\boldsymbol{W}_j$ is written as

$$\boldsymbol{W}_j = \sum_{i=1}^{M} \delta_{ji} \boldsymbol{W}_{0i},$$

where δ_{ji} for $i = 1, \ldots, M$, $j = 0, \ldots, p$ are unknown coefficients in the linear combinations.

On top of allowing for estimating the spatial weight matrices from pre-specified ones, our model also includes time-lagged spatial effects. In a differently specified spatial lag model, Dou et al. (2016) includes one lag to reflect such effects. We generalize this to p time-lagged effects, with p to be determined by data driven methods as described in Section 4. The pure dynamic effects are captured by the term $\boldsymbol{X}_t\boldsymbol{\beta}$, since we can allocate $\{y_{t-1}, \ldots, y_{t-p}\}$ to be the columns in $\boldsymbol{X}_t$, so that then $K \geq p$, and $K = p$ if no other covariates are present. Not counting the parameters in $\boldsymbol{\mu}$, there are $K + M(p+1)$ parameters to be estimated in total.

With $\boldsymbol{\mu}$, the spatial fixed effects of the model is then $(\boldsymbol{I}_N - \boldsymbol{W}_0)^{-1}\boldsymbol{\mu}$. For identifiability of such, we assume without loss of generality that $\mathbb{E}(X_t) = 0$. If not, we can write

$$\boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\mu} = (\boldsymbol{X}_t - \mathbb{E}(\boldsymbol{X}_t))\boldsymbol{\beta} + (\boldsymbol{\mu} + \mathbb{E}(\boldsymbol{X}_t)\boldsymbol{\beta})$$

so that the spatial fixed effects are now captured by $\boldsymbol{\mu} + \mathbb{E}(\boldsymbol{X}_t)\boldsymbol{\beta}$ rather than $\boldsymbol{\mu}$, and the covariates are of mean $\mathbf{0}$.

To present our model more neatly, we rewrite (2.1) as

$$y = \boldsymbol{\mu} \otimes 1_T + \boldsymbol{Z}_0\boldsymbol{V}_0\boldsymbol{\delta}_0 + \boldsymbol{Z}_1\boldsymbol{V}_0\boldsymbol{\delta}_1 + \cdots + \boldsymbol{Z}_p\boldsymbol{V}_0\boldsymbol{\delta}_p + \boldsymbol{X}_{\boldsymbol{\beta}}\mathrm{vec}(\boldsymbol{I}_N) + \boldsymbol{\epsilon},$$

where $y = \mathrm{vec}(y_1, \ldots, y_T)^T$, $\boldsymbol{\epsilon} = \mathrm{vec}(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)^T$, $\boldsymbol{Z}_j = \boldsymbol{I}_N \otimes (y_{1-j}, \ldots, y_{T-j})^T$ and $\boldsymbol{\delta}_j = (\delta_{j1}, \delta_{j2}, \ldots, \delta_{jM})^T$ for $j = 0, 1, \ldots, p$, $\boldsymbol{X}_{\boldsymbol{\beta}} = \boldsymbol{I}_N \otimes (\boldsymbol{I}_T \otimes \boldsymbol{\beta}^T)(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)^T$, and $\boldsymbol{V}_0 = (\mathrm{vec}(\boldsymbol{W}_{01}^T), \ldots, \mathrm{vec}(\boldsymbol{W}_{0M}^T))$. The notation $\otimes$ is the Kronecker product, and 1_T defines a vector of ones with size T. Simplifying, we have

$$y = \boldsymbol{\mu} \otimes 1_T + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{\delta} + \boldsymbol{X}_{\boldsymbol{\beta}}\mathrm{vec}(\boldsymbol{I}_N) + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{Z} = (\boldsymbol{Z}_0, \ldots, \boldsymbol{Z}_p)$, $\boldsymbol{\delta} = (\boldsymbol{\delta}_0^T, \ldots, \boldsymbol{\delta}_p^T)^T$, and $\boldsymbol{V} = \boldsymbol{I}_{p+1} \otimes \boldsymbol{V}_0$.

2.2 Profiled least square estimation

Note that $\boldsymbol{\mu}$ is of size N, while $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ together have $K + p(M+1)$ parameters which is potentially much smaller than N. To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ more efficiently while overcoming the problem of endogeneity contributed from $\boldsymbol{Z}_0$ in model (2.2), we assume that there are variables $\boldsymbol{B}_t$ of size $N \times K$ such that they are correlated with $\boldsymbol{X}_t$ but independent of $\boldsymbol{\epsilon}_t$ for each $t = 1, \ldots, T$. In particular, if $\boldsymbol{X}_t$ is exogenous, we can set $\boldsymbol{B}_t = \boldsymbol{X}_t$. Define

$$\boldsymbol{B} = T^{-1/2}N^{-a/2}(\boldsymbol{B}_{\boldsymbol{\gamma}} - \overline{\boldsymbol{B}}_{\boldsymbol{\gamma}}) = T^{-1/2}N^{-a/2}\boldsymbol{I}_N \otimes \{(\boldsymbol{I}_T \otimes \boldsymbol{\gamma}^T)(\boldsymbol{B}_1 - \bar{\boldsymbol{B}}, \ldots, \boldsymbol{B}_T - \bar{\boldsymbol{B}})^T\},$$

where $\bar{\boldsymbol{B}} = T^{-1}\sum_{t=1}^{T}\boldsymbol{B}_t$, and $\boldsymbol{\gamma} = K^{-1}1_K$. The value of a is not important in practice, and we set $a = 1$ in our algorithms. It is there only to adjust the order of eigenvalues of some constructs involving $\boldsymbol{B}$ in the proof of our theorems. See the technical assumptions in Section 7 for more details. The value of $\boldsymbol{\gamma}$ is

not the only choice, and we will introduce a way to choose a data driven one in Section 4.3. We naturally assumes $\boldsymbol{B}_t$ takes on different values so that $\boldsymbol{B}_t$ is different from $\bar{\boldsymbol{B}}$ in general.

To utilize $\boldsymbol{B}$, multiplying $\boldsymbol{B}^T$ on both sides of (2.2), we arrive at the augmented model

$$\boldsymbol{B}^T y = \boldsymbol{B}^T \boldsymbol{Z}\boldsymbol{V}\boldsymbol{\delta} + \boldsymbol{B}^T \boldsymbol{X}_{\boldsymbol{\beta}} \mathrm{vec}(\boldsymbol{I}_N) + \boldsymbol{B}^T \boldsymbol{\epsilon}. \tag{2.3}$$

The constant term disappears since $\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{\mu} \otimes 1_T) = \boldsymbol{0}$. Removing the N-dimensional constant term makes estimation much easier, while the error term $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\epsilon}$ is now weaker in correlations with the design matrix $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{V}$, so that least square estimation becomes viable again. This way, $\boldsymbol{B}$ serves a similar function as an instrumental variable.

In order to profile out $\boldsymbol{\beta}$ and estimate $\boldsymbol{\delta}$, we rewrite the augmented model (2.1) as

$$\boldsymbol{B}^{vT} y_0^v = \boldsymbol{B}^{vT}(\sum_{i=1}^{M} \delta_{0i} \boldsymbol{W}_{0i}^{\otimes}) y_0^v + \boldsymbol{B}^{vT} \sum_{j=1}^{p} (\sum_{i=1}^{M} \delta_{ji} \boldsymbol{W}_{0i}^{\otimes}) y_j^v + \boldsymbol{B}^{vT} \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{vT} \boldsymbol{\epsilon}^v,$$

where $y_j^v = (y_{1-j}^T, \ldots, y_{T-j}^T)^T$ for $j = 0, 1, \ldots, p$, $\boldsymbol{\epsilon}^v = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_T^T)^T$, $\boldsymbol{B}^v = ((\boldsymbol{B}_1 - \bar{\boldsymbol{B}})^T, \ldots, (\boldsymbol{B}_T - \bar{\boldsymbol{B}})^T)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_T^T)^T$ and $\boldsymbol{W}_{0i}^{\otimes} = \boldsymbol{I}_T \otimes \boldsymbol{W}_{0i}$ for $i = 1, \ldots, M$. Assuming $\boldsymbol{\delta}$ is known, we can estimate $\boldsymbol{\beta}$ by the least squared method, resulting in

$$\boldsymbol{\beta}(\boldsymbol{\delta}) = (\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \Big\{ (\boldsymbol{I}_{TN} - \sum_{i=1}^{M} \delta_{0i} \boldsymbol{W}_{0i}^{\otimes}) y_0^v - \sum_{j=1}^{p} (\sum_{i=1}^{M} \delta_{ji} \boldsymbol{W}_{0i}^{\otimes}) y_j^v \Big\}. \tag{2.4}$$

This formula provides a basis for a profile least square estimator for $\boldsymbol{\delta}$. We can show that by substituting the above into the augmented model (2.3) (proof omitted), the profile least square estimator for $\boldsymbol{\delta}$ is

$$\widehat{\boldsymbol{\delta}} = \{(\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z}\boldsymbol{V})^T (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z}\boldsymbol{V})\}^{-1} (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z}\boldsymbol{V})^T (\boldsymbol{K} y_0^v - \boldsymbol{B}^T y), \tag{2.5}$$

where

$$\begin{aligned}
\boldsymbol{K} &= T^{-1/2} N^{-a/2} (\sum_{t=1}^{T} \boldsymbol{X}_t \otimes (\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\gamma})(\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT}, \\
\boldsymbol{H} &= \boldsymbol{K} \left[\boldsymbol{W}_{01}^{\otimes}, \ldots, \boldsymbol{W}_{0M}^{\otimes} \right] (\boldsymbol{I}_M \otimes y_0^v, \boldsymbol{I}_M \otimes y_1^v, \ldots, \boldsymbol{I}_M \otimes y_p^v).
\end{aligned}$$

With this, the profile least square estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\delta}}) = (\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \Big\{ (\boldsymbol{I}_{TN} - \sum_{i=1}^{M} \widehat{\delta}_{0i} \boldsymbol{W}_i^{\otimes}) y_0^v - \sum_{j=1}^{p} (\sum_{i=1}^{M} \widehat{\delta}_{ji} \boldsymbol{W}_i^{\otimes}) y_j^v \Big\}. \tag{2.6}$$

Finally, to estimate $\boldsymbol{\mu}$, we can use

$$\widehat{\boldsymbol{\mu}} = \Big(\boldsymbol{I}_N - \sum_{j=0}^{p} \widehat{\boldsymbol{W}}_j \Big) \bar{y} - \bar{X} \widehat{\boldsymbol{\beta}}, \quad \text{where } \widehat{\boldsymbol{W}}_j = \sum_{i=1}^{M} \widehat{\delta}_{ji} \boldsymbol{W}_{0i}.$$

The corresponding spatial fixed effects estimator is then given by $(\boldsymbol{I}_N - \widehat{\boldsymbol{W}}_0)^{-1} \widehat{\boldsymbol{\mu}}$.

2.3 Selection of specified spatial weight matrices

Since $\widehat{\boldsymbol{\delta}}$ is a least square-type estimator, each element in it is not estimated to be exactly 0 in general. This hinders the selection of the specified spatial weight matrices, which is important for us to see which one contributes to the overall spatial weight matrices and which one does not. To ameliorate this, we can find a penalized profile least square estimator $\widetilde{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$, with

$$\widetilde{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \frac{1}{2T} \|\boldsymbol{B}^T y - \boldsymbol{B}^T \boldsymbol{Z}\boldsymbol{V}\boldsymbol{\delta} - \boldsymbol{B}^T \boldsymbol{X}_{\boldsymbol{\beta}(\boldsymbol{\delta})} \operatorname{vec}(\boldsymbol{I}_N)\|^2 + \gamma_T \boldsymbol{u}^T |\boldsymbol{\delta}|, \tag{2.7}$$

where $\boldsymbol{u} = (|\widehat{\delta}_{0,1}|^{-1}, \ldots, |\widehat{\delta}_{0,M}|, \ldots, |\widehat{\delta}_{p,1}|^{-1}, \ldots, |\widehat{\delta}_{p,M}|^{-1})^T$, and $|\boldsymbol{\delta}|$ represents the same vector $\boldsymbol{\delta}$ with all its entries taken absolute value. A more direct penalized least square formulation is given by

$$\begin{aligned}
&\widetilde{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \frac{1}{2T} \|\boldsymbol{B}^T y - (\boldsymbol{B}^T \boldsymbol{Z}\boldsymbol{V} - \boldsymbol{H})\boldsymbol{\delta} - \boldsymbol{g}\|^2 + \gamma_T \boldsymbol{u}^T |\boldsymbol{\delta}|, \text{ where} \\
&\boldsymbol{g} = T^{-1/2} N^{-a/2} (\sum_{t=1}^{T} \boldsymbol{X}_t \otimes (\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\gamma})(\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} y^v.
\end{aligned}$$

The tuning parameter γ_T can be found in Assumption R6 in the Appendix. For choosing an appropriate γ_T in practice, see Section 4.2.

3 Theoretical Properties

To present the theoretical properties of our estimators, we first present the notations used hereafter and introduce the measure of time dependence of all the time series variables involved.

Denote $\{\boldsymbol{b}_t\} = \{\operatorname{vec}(\boldsymbol{B}_t)\}$ and $\{\boldsymbol{x}_t\} = \{\operatorname{vec}(\boldsymbol{X}_t)\}$ the vectorized processes for $\{\boldsymbol{B}_t\}$ and $\{\boldsymbol{X}_t\}$ respectively, both with length NK. For $t = 1, \ldots, T$, we assume that

$$\boldsymbol{x}_t = \{f_j(\mathcal{F}_t)\}_{1 \le j \le NK},\ \boldsymbol{b}_t = \{g_j(\mathcal{G}_t)\}_{1 \le j \le NK},\ \boldsymbol{\epsilon}_t = \{h_l(\mathcal{H}_t)\}_{1 \le l \le N},$$

where the $f_j(\cdot)$, $g_j(\cdot)$ and $h_l(\cdot)$ are measurable functions defined on the real line, and $\mathcal{F}_t = (..., e_{x,t-1}, e_{x,t})$, $\mathcal{G}_t = (..., e_{b,t-1}, e_{b,t})$ and $\mathcal{H}_t = (..., e_{\epsilon,t-1}, e_{\epsilon,t})$ are definde by independent and identically distributed processes $\{e_{x,t}\}$, $\{e_{b,t}\}$ and $\{e_{\epsilon,t}\}$ respectively, with $\{e_{b,t}\}$ independent of $\{e_{\epsilon,t}\}$ but correlated with $\{e_{x,t}\}$.

We use the functional dependence measure introduced in Wu (2005) for gauging the serial dependence of a process. For $d > 0$, define

$$\begin{aligned}
\theta^x_{t,d,j} &= \|x_{tj} - x'_{tj}\|_d = (\mathbb{E}|x_{tj} - x'_{tj}|^d)^{1/d}, \\
\theta^b_{t,d,j} &= \|b_{tj} - b'_{tj}\|_d = (\mathbb{E}|b_{tj} - b'_{tj}|^d)^{1/d}, \\
\theta^\epsilon_{t,d,l} &= \|\epsilon_{tl} - \epsilon'_{tl}\|_d = (\mathbb{E}|\epsilon_{tl} - \epsilon'_{tl}|^d)^{1/d},
\end{aligned}$$

where $j = 1, ..., NK$, $l = 1, ..., N$ and $x'_{tj} = f_j(\mathcal{F}'_t)$, $\mathcal{F}'_t = (..., e_{x,-1}, e'_{x,0}, e_{x,1}, ..., e_{x,t})$, with $e'_{x,0}$ independent of all other $e_{x,j}$'s. Hence x'_{tj} is a coupled version of x_{tj} with $e_{x,0}$ replaced by an independent and identically distributed copy $e'_{x,0}$. Intuitively, a large $\theta^x_{t,d,j}$ means that serial correlation is strong at least for variables at most time t apart. Finally, we have similar definitions for b'_{tj} and ϵ'_{tl}.

3.1 Main assumptions

We present the main assumptions of the paper in this section.

M1. The elements in all $\boldsymbol{W}_i$'s can be negative and $\boldsymbol{W}_i$ itself can be asymmetric. Moreover, defining $S = \{s = 1, \ldots, K | \text{The } s\text{th column of } \boldsymbol{X}_t \text{ contains } y_{t-l},\ l = 1, \ldots, p\}$, we assume $\sum_{i=1}^{M} |\delta_{0i}| < 1$ and $\sum_{j=1}^{p} \sum_{i=1}^{M} |\delta_{ji}| + \sum_{s \in S} |\beta_s| < 1$.

M2. The processes $\{\boldsymbol{B}_t\}$, $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ are second-order stationary, with $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ having zero means, and $\{\boldsymbol{B}_t\}$ independent of $\{\boldsymbol{\epsilon}_t\}$. The tail condition $P(|Z| > v) \leq D_1 \exp(-D_2 v^q)$ is satisfied for the variables $B_{t,jk}$, $X_{t,jk}$ and $\epsilon_{t,j}$ by the same constants D_1, D_2 and q.

M3. Define

$$\Theta^x_{m,a} = \sum_{t=m}^{\infty} \max_{1 \leq j \leq NK} \theta^x_{t,a,j},\ \Theta^b_{m,a} = \sum_{t=m}^{\infty} \max_{1 \leq j \leq NK} \theta^b_{t,a,j},\ \Theta^\epsilon_{m,a} = \sum_{t=m}^{\infty} \max_{1 \leq j \leq N} \theta^\epsilon_{t,a,j}.$$

Then we assume that for some $w > 2$, $\Theta^x_{m,2w}, \Theta^b_{m,2w}, \Theta^\epsilon_{m,2w} \leq C m^{-\alpha}$ with $\alpha, C > 0$ being constants that can depend on w.

M4. (Identification condition) Assume that the two sets of parameters $(\boldsymbol{\delta}^*, \boldsymbol{\beta}^*)$ and $(\boldsymbol{\delta}^o, \boldsymbol{\beta}^o)$ both satisfy the proposed model (2.2). Write $\boldsymbol{\delta} = (\delta_\ell)_{1 \leq \ell \leq M(p+1)}$, and define the set H to be

$$H = \{\ell : \delta^*_\ell \neq 0 \text{ or } \delta^o_\ell \neq 0\}.$$

Then the identification condition is that the matrix $\boldsymbol{O}^T\boldsymbol{O}$ has all its eigenvalues uniformly bounded away from 0, where

$$\boldsymbol{O} = (T^{-1/2}\mathbb{E}(\boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}_H), T^{-1/2}\mathbb{E}(\boldsymbol{B}^T \widetilde{\boldsymbol{X}})), \text{ and}$$
$$\widetilde{\boldsymbol{X}} = (\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{T,1}, \ldots, \boldsymbol{x}_{1,N}, \ldots, \boldsymbol{x}_{T,N})^T.$$

The notation A_H means that the matrix A has columns restricted to the set H, while $\boldsymbol{x}^T_{t,j}$ is the jth row of $\boldsymbol{X}_t$.

Assumption M1 ensures that our model has a reduced form

$$y_t = \boldsymbol{\Pi}\boldsymbol{\mu} + \boldsymbol{\Pi}\boldsymbol{W}_1 y_{t-1} + \cdots + \boldsymbol{\Pi}\boldsymbol{W}_p y_{t-p} + \boldsymbol{\Pi}\boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\Pi}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\Pi} = (\boldsymbol{I}_N - \boldsymbol{W})^{-1}, \quad t = 1, \ldots, T.$$

The matrix $\boldsymbol{\Pi}$ exists with the assumption $\sum_{i=1}^{M} |\delta_{0i}| < 1$. The condition $\sum_{j=1}^{p} \sum_{i=1}^{M} |\delta_{ji}| + \sum_{s \in S} |\beta_s| < 1$ implies that each $\left\|\boldsymbol{W}_j\right\|_\infty < \sum_{i=1}^{M} |\delta_{ji}| \left\|\boldsymbol{W}_{0i}\right\|_\infty < 1$ since row-standardization means $\left\|\boldsymbol{W}_{0i}\right\|_\infty = 1$. At the same time, without loss of generality assuming $S = \phi$ and writing the model as

$$\boldsymbol{\Phi}(L) y_t = \boldsymbol{\Pi}\boldsymbol{\mu} + \boldsymbol{\Pi}\boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\Pi}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\Phi}(L) = (\boldsymbol{I}_N - \boldsymbol{\Pi}\boldsymbol{W}_1 L - \cdots - \boldsymbol{\Pi}\boldsymbol{W}_p L^p),$$

where L is the lag operator, then stationarity is ensured if $\det(\boldsymbol{\Phi}(z)) = 0$ has all roots lying outside the unit circle. This is ensured by the condition $\sum_{j=1}^{p} \sum_{i=1}^{M} |\delta_{ji}| + \sum_{s \in S} |\beta_s| < 1$, which is thus a sufficient condition for stationarity. In practice, we implement these restrictions when finding $\widetilde{\boldsymbol{\delta}}$ in Section 2.3.

The independence between $\{\boldsymbol{B}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ in M2 ensures that $\{\boldsymbol{B}_t\}$ serves a function similar to an instrument for model (2.2). The tail condition in M2 implies that all the random variables involved are with sub-exponential tails, which is a relaxation to strict normality.

The assumption $\Theta^x_{m,2w} \leq Cm^{-\alpha}$ essentially means that the strongest serial dependence for the x_{tj}'s with at least m time units apart is decaying polynomially as m increases. It allows for the application of a Nagaev-type inequality in Lemma 1 in the Appendix for our results to hold.

3.2 Identification of the model

To prove that the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ in model (2.2) are identified, we assume that we have two sets of parameters $(\boldsymbol{\beta}^*, \boldsymbol{\delta}^*)$ and $(\boldsymbol{\beta}^o, \boldsymbol{\delta}^o)$ that satisfy model (2.2), as stated in the identification condition M4. Then

$$\mathbf{0} = \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}_H (\boldsymbol{\delta}^*_H - \boldsymbol{\delta}^o_H) + \boldsymbol{B}^T \boldsymbol{X}_{\boldsymbol{\beta}^* - \boldsymbol{\beta}^o} \mathrm{vec}(\boldsymbol{I}_N).$$

But we can write

$$\begin{aligned} T^{-1/2} \boldsymbol{B}^T \boldsymbol{X}_{\boldsymbol{\beta}^* - \boldsymbol{\beta}^o} \mathrm{vec}(\boldsymbol{I}_N) &= N^{-a/2} \begin{pmatrix} T^{-1} \sum_{t=1}^T (\boldsymbol{B}_t - \bar{\boldsymbol{B}}) \boldsymbol{\gamma} \mathbf{x}_{t,1}^T (\boldsymbol{\beta}^* - \boldsymbol{\beta}^o) \\ \vdots \\ T^{-1} \sum_{t=1}^T (\boldsymbol{B}_t - \bar{\boldsymbol{B}}) \boldsymbol{\gamma} \mathbf{x}_{t,N}^T (\boldsymbol{\beta}^* - \boldsymbol{\beta}^o) \end{pmatrix} \\ &= T^{-1/2} \boldsymbol{B}^T \widetilde{\boldsymbol{X}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}^o), \end{aligned}$$

so that

$$[T^{-1/2} \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}_H \;\; T^{-1/2} \boldsymbol{B}^T \widetilde{\boldsymbol{X}}] \begin{pmatrix} \boldsymbol{\delta}^*_H - \boldsymbol{\delta}^o_H \\ \boldsymbol{\beta}^* - \boldsymbol{\beta}^o \end{pmatrix} = \mathbf{0}.$$

Hence taking expectation and multiplying $\boldsymbol{O}^T$ on both sides and then $(\boldsymbol{O}^T \boldsymbol{O})^{-1}$, we have shown that $\boldsymbol{\delta}^*_H = \boldsymbol{\delta}^o_H$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta}^o$.

Note that the matrix $\boldsymbol{O}$ has size $N^2 \times (|H| + K)$. Since $|H| \leq M(p+1)$ and K are finite and N^2 is much larger than $|H| + K$, assuming $\boldsymbol{O}$ has full rank is reasonable.

3.3 Main results

Define the rate $\lambda_T = cT^{-1/2} \log^{1/2}(T \vee N)$, where $c > 0$ is a constant. In all the theorems presented here, we assume that $\alpha \geq 1/2 - 1/w$ in Assumption M3, which is part of the further assumptions listed in Theorem 5 in the Appendix. See the Appendix for more details on the technical assumptions for this paper.

Theorem 1. *Let the assumptions in Section 3.1 and in Theorem 5 hold. Then defining the L_1 norm $\|a\|_1 = \sum_{i=1}^N |a_i|$ for a vector a, the estimators $\widehat{\boldsymbol{\delta}}$ in (2.5) and $\widehat{\boldsymbol{\beta}}$ in (2.6) satisfy*

$$\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1 = O_P(\lambda_T N^{-1/2+1/2w}) = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1.$$

Since $w > 2$ is assumed in M3, the above immediately implies $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1, \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1 \to 0$ in probability. This certainly makes sense as $T \to \infty$. It also makes perfect sense as $N \to \infty$ since we are accumulating more

information cross-sectionally for the finite-sized parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ as N goes to infinity. We present the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ in the following two theorems.

Theorem 2. *Let the assumptions in Section 3.1 and in Theorem 5 hold. Moreover, define the predictive dependence measures*

$$\boldsymbol{P}_0^b(B_{t,qk}) = \mathbb{E}(B_{t,qk}|\mathcal{G}_0) - \mathbb{E}(B_{t,qk}|\mathcal{G}_{-1}), \quad \boldsymbol{P}_0^\epsilon(\epsilon_{t,qk}) = \mathbb{E}(\epsilon_{t,qk}|\mathcal{H}_0) - \mathbb{E}(\epsilon_{t,qk}|\mathcal{H}_{-1}),$$

where $\mathcal{G}_t$ and $\mathcal{H}_t$ are defined in Section 3. Assume

$$\sum_{t\geq 0} \max_{1\leq q\leq N} \max_{1\leq k\leq K} \|\boldsymbol{P}_0^b(B_{t,qk})\| \leq \infty, \quad \sum_{t\geq 0} \max_{1\leq j\leq N} \|\boldsymbol{P}_0^\epsilon(\epsilon_{tj})\| \leq \infty.$$

Then we have

$$T^{1/2}\boldsymbol{\Sigma}_1^{-1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_K),$$

where $\boldsymbol{\Sigma}_1 = \boldsymbol{M}_1 \sum_{\tau\in\mathbb{Z}} \mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_{t+\tau}^T \boldsymbol{B}_{t+\tau}) \boldsymbol{M}_1^T$, with $\boldsymbol{M}_1 = (\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t))^{-1}\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)$.

Theorem 3. *Let the assumptions in Section 3.1 and in Theorem 5 hold. Assume that the predictive dependence measures $\boldsymbol{P}_0^b(B_{t,qk})$ and $\boldsymbol{P}_0^\epsilon(\epsilon_{t,qk})$ are as defined in Theorem 2 with the same assumptions applied. Then*

$$T^{1/2}\boldsymbol{\Sigma}_2^{-1/2}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_{M(p+1)}),$$

where $\boldsymbol{\Sigma}_2 = \boldsymbol{M}_2(\mathbf{S}_1 + \mathbf{S}_2 - \mathbf{S}_3 - \mathbf{S}_3^T)\boldsymbol{M}_2^T$, and

$$\begin{aligned}
\mathbf{S}_1 &= \sum_{\tau\in\mathbb{Z}} \mathbb{E}(\boldsymbol{M}\boldsymbol{B}_{t+\tau}^T \boldsymbol{\epsilon}_{t+\tau} \boldsymbol{\epsilon}_t^T \boldsymbol{B}_t^T \boldsymbol{M}^T), \\
\mathbf{S}_2 &= \sum_{\tau\in\mathbb{Z}} \left[\mathbb{E}(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_{t+\tau}^T) \otimes \mathbb{E}(\boldsymbol{B}_t \boldsymbol{\gamma}\boldsymbol{\gamma}^T \boldsymbol{B}_{t+\tau}^T)^T\right], \\
\mathbf{S}_3 &= \sum_{\tau\in\mathbb{Z}} \mathbb{E}(\boldsymbol{M}\boldsymbol{B}_{t+\tau}^T \boldsymbol{\epsilon}_{t+\tau} (\mathrm{vec}(\boldsymbol{B}_t \boldsymbol{\gamma} \boldsymbol{\epsilon}_t^T))^T), \\
\boldsymbol{M}_2 &= \{(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\}^{-1} (\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T, \ \textit{with} \\
\boldsymbol{H}_{10} &= [\boldsymbol{I}_N \otimes \mathbb{E}((\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\gamma} y_t^T), \ldots, \boldsymbol{I}_N \otimes \mathbb{E}((\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\gamma} y_{t-p}^T)]\boldsymbol{V}, \\
\boldsymbol{H}_{20} &= \boldsymbol{M}[\mathbb{E}((\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T \boldsymbol{W}_{01} y_t), \ldots, \mathbb{E}((\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T \boldsymbol{W}_{0M} y_t), \ldots, \\
&\qquad \mathbb{E}((\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T \boldsymbol{W}_{01} y_{t-p}), \ldots, \mathbb{E}((\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T \boldsymbol{W}_{0M} y_{t-p})],
\end{aligned}$$

where $\boldsymbol{M} = \mathbb{E}(\boldsymbol{X}_t \otimes (\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\gamma}) \left[\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t)\right]^{-1} \mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)$.

These two theorems are the main ones we use, since they provide the tools for practical data analysis such as hypothesis testing and confidence intervals construction. For calculating $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, we calculate all expectations by replacing them with the corresponding sample means. For the infinite summations in τ in $\mathbf{S}_1$ to $\mathbf{S}_3$, we check if the matrix at a particular τ has very small elements overall. If so, we discard the whole matrix and all the matrices beyond this particular τ. In the real data analysis in Section 6, we find that we always discard those with $\tau \geq 5$. See Section 4.1 for further treatments regarding the estimation of the matrices $\mathbf{S}_1$ to $\mathbf{S}_3$.

Theorem 4. *(Oracle property for $\widetilde{\boldsymbol{\delta}}$) Let the assumptions in Section 3.1 and in Theorem 5 hold. Then as*

$T, N \to \infty$, with probability approaching 1,

$$\text{sign}(\widetilde{\boldsymbol{\delta}}_H) = \text{sign}(\boldsymbol{\delta}_H), \quad \widetilde{\boldsymbol{\delta}}_{H_c} = 0,$$

where $H = \{\ell : \delta_\ell \neq 0\}$ and $\ell = 1, \ldots, M(p+1)$. Moreover, let the predictive dependence measures $\boldsymbol{P}_0^b(B_{t,qk})$ and $\boldsymbol{P}_0^\epsilon(\epsilon_{t,qk})$ be as defined in Theorem 2 with the same assumptions applied. Then

$$T^{1/2}\boldsymbol{\Sigma}_3^{-1/2}(\widetilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H) \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_{|H|}),$$

where $\boldsymbol{\Sigma}_3 = \boldsymbol{M}_3(\mathbf{S}_1 + \mathbf{S}_2 - \mathbf{S}_3 - \mathbf{S}_3^T)\boldsymbol{M}_3^T$, and $\boldsymbol{M}_3 = \{(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H\}^{-1}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T$

With Theorem 4, we can carry out the selection of the importance of the specified spatial weight matrices by the penalized estimator $\widetilde{\boldsymbol{\delta}}$, and the usual inferences on the non-zero elements in $\widetilde{\boldsymbol{\delta}}$. The practical performances of these estimators and the asymptotic normality results are presented in Section 5.

4 Practical Implementation

4.1 Regularized estimation of $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ in Theorem 3 and 4

In Theorem 3, the definitions of $\mathbf{S}_1$ to $\mathbf{S}_3$ involve some high dimensional matrices to be estimated. Since $\mathbf{S}_1$ to $\mathbf{S}_3$ are in fact all $N \times N$, in this paper we regularize $\mathbf{S}_1$ and $\mathbf{S}_3$ by banding them directly (see Bickel and Levina (2008) for more details). In simulations and real data analysis, we find that retaining only two off-diagonals (two upper and two lower, while setting 0 in all other off-diagonals) when $\tau = 0$, and retaining only one when $|\tau| \geq 1$ in the infinite summations in $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ achieves good results when N is moderate to large. Again similar to the discussion after Theorem 3, when $|\tau| \geq 5$, we set the matrices inside the summations in the definitions of $\mathbf{S}_1$ to $\mathbf{S}_3$ to exactly zero. For $\mathbf{S}_2$, there are two $N \times N$ matrices $\mathbb{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T)$ and $\mathbb{E}(\boldsymbol{B}_t\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{B}_{t+\tau}^T)$. We band them separately, again retaining only two off-diagonals each when $\tau = 0$, and only one when $|\tau| \geq 1$. We make these suggestions because in both simulations and real data analysis, using the 5-fold cross-validation procedure suggested in Bickel and Levina (2008), these are the banding numbers chosen for $|\tau| < 5$.

4.2 Choice of the number of time lags p, and γ_T

In our analysis, we assume that p in model (2.1) is fixed. For practical data analysis, we choose p by minimizing the following BIC criterion:

$$\text{BIC}(p) = \log\big(N^{-1}\|\boldsymbol{B}^T y - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}\widehat{\boldsymbol{\delta}} - \boldsymbol{B}^T\boldsymbol{X}_{\widehat{\boldsymbol{\beta}}}\text{vec}(\boldsymbol{I}_N)\|^2\big) + p\frac{\log T}{T}\log(\log T), \tag{4.8}$$

This BIC criterion follows the one in Wang et al. (2009), and proves to work well in practice. Note that in the definition of $\boldsymbol{B}$, there is a rate a which is unknown. However, because of the logarithmic operation in the first term in BIC(p), the value of a does not change where the minimum of BIC(p) is achieved.

For the choice of γ_T, we use the BIC criterion above, but with $\widehat{\boldsymbol{\delta}}$ replaced by $\widetilde{\boldsymbol{\delta}}$, so that we are effectively choosing p and γ_T together.

4.3 Choice of $\boldsymbol{\gamma}$ in $\boldsymbol{B}$

We have set $\boldsymbol{\gamma} = K^{-1}1_K$ as fixed in the definition of $\boldsymbol{B}$ in Section 2.2. In fact this can be estimated to provide maximal correlation between $\boldsymbol{B}$ and the response variable y_t through two-stage least squares. Consider the model

$$y_t = \boldsymbol{\alpha} + \boldsymbol{B}_t\boldsymbol{\gamma} + \boldsymbol{v}_t,$$

where $\boldsymbol{\alpha}$ is an $N \times 1$ vector of unknown coefficients, and $\boldsymbol{\gamma}$ is the $K \times 1$ vector of coefficients we want to estimate. To get $\widehat{\boldsymbol{\gamma}}$, we can consider the problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \sum_{t=1}^{T} \|y_t - \boldsymbol{\alpha} - \boldsymbol{B}_t\boldsymbol{\gamma}\|^2,$$

with solution

$$\widehat{\boldsymbol{\gamma}} = \Big(\sum_{t=1}^{T}(\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T(\boldsymbol{B}_t - \bar{\boldsymbol{B}})\Big)^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T(y_t - \bar{y}).$$

Implementing this does not change our proofs, since it is easy to show that $\left\|\widehat{\boldsymbol{\gamma}}\right\|_1 = O_P(1)$, which substitutes $\left\|\widehat{\boldsymbol{\gamma}}\right\|_1 = 1$ in all of our proofs. We have tried this in our simulations and real data analysis, and the practical differences between using this and $\boldsymbol{\gamma} = K^{-1}1_K$ is negligible.

5 Simulation Experiments

5.1 Setting and results

To generate y_t through model (2.1), we generate $\boldsymbol{X}_t$ by using $\text{vec}(\boldsymbol{X}_t) = 0.2 \cdot 1_K \otimes \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_t^{\boldsymbol{X}}$ with $K = 3$, where $\boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{I}_N)$ is the innovation series for model (2.1), with the $\boldsymbol{\epsilon}_t$'s being independent of each other. The $\boldsymbol{\epsilon}_t^{\boldsymbol{X}}$'s are independent of each other and of other variables, with $\boldsymbol{\epsilon}_t^{\boldsymbol{X}} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$, and

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \begin{bmatrix} 2\boldsymbol{I}_N & 0.5\boldsymbol{I}_N & 0.5\boldsymbol{I}_N \\ 0.5\boldsymbol{I}_N & 2\boldsymbol{I}_N & 0.5\boldsymbol{I}_N \\ 0.5\boldsymbol{I}_N & 0.5\boldsymbol{I}_N & 2\boldsymbol{I}_N \end{bmatrix}.$$

Since $\boldsymbol{X}_t$ depends on $\boldsymbol{\epsilon}_t$, we set $\boldsymbol{B}_t$ to be such that $\text{vec}(\boldsymbol{B}_t) = 0.7\boldsymbol{\epsilon}_t^{\boldsymbol{X}} + \boldsymbol{\epsilon}_t^{\boldsymbol{B}}$, where the $\boldsymbol{\epsilon}_t^{\boldsymbol{B}}$'s are drawn independently from the same distribution as $\boldsymbol{\epsilon}_T^{\boldsymbol{X}}$, and they are independent of all other variables.

We set $M = 3$ and $p = 2$ for the model. Each element of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ is generated independently from the uniform distribution $\boldsymbol{U}(0, 1)$. Elements in $\boldsymbol{\delta}$ are then randomly chosen to be 0 while maintaining $p = 2$. To make sure the stationarity of $\{y_t\}$, every element in $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ is then divided by 1.1 times the absolute sum of all elements in $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ respectively.

For the $M = 3$ specified spatial weight matrices, to facilitate stationarity of the model, we construct each $\boldsymbol{W}_i$ such that only the first three off-diagonals (upper and lower) have non-zero elements. This way, as N increases, we can still control the eigenvalues of $\boldsymbol{W}_i$ to be less than 1 in magnitude. In another setting, we generate an orthogonal matrix $\boldsymbol{V}_i$ and a diagonal matrix $\mathbf{D}_i$ with all values in $\mathbf{D}_i$ to be less than 1 in magnitude, such that $\boldsymbol{W}_i = \boldsymbol{V}_i\mathbf{D}_i\boldsymbol{V}_i^T$. Ultimately, both settings achieves very similar results, and hence we only show the results of the former setting.

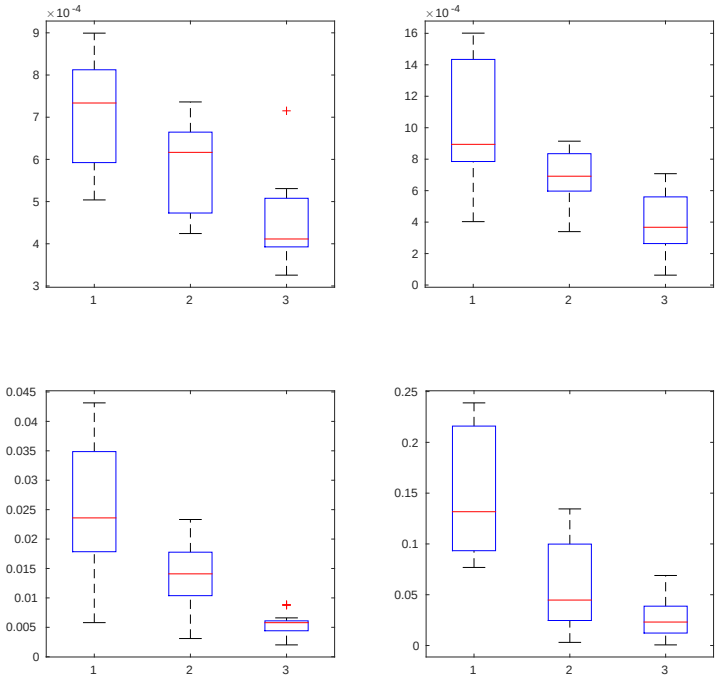

Figure 1: Boxplots of averaged L_1 errors. Upper row: $\sum_{i=1}^{3} |\widehat{\beta}_i - \beta_i|/3$. Bottom row: $\left\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right\|_1/9$. Left column (from left to right): $N = 40, 80, 120, T = 60$. Right column (from left to right): $T = 40, 80, 120, N = 60$.

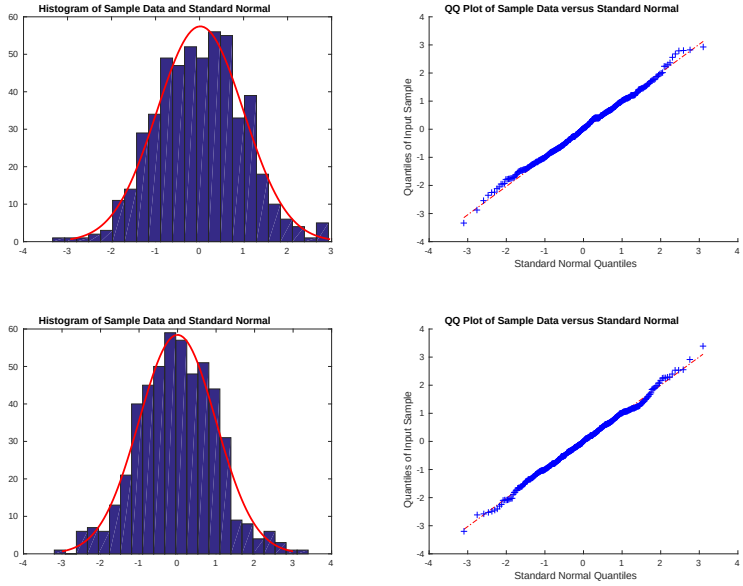

Figure 2: Histograms and normal probability plots for standardized $\widehat{\beta}_1$ (upper row) and $\widehat{\delta}_{1,3}$ (lower row) with $N = T = 80$. Standardization used respectively the asymptotic results from Theorem 2 and 3.

We repeat our simulations for 500 times, and report the averaged L_1-error for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ (i.e., respectively, $\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1/3=\sum_{i=1}^3|\widehat{\beta}_i-\beta_i|/3$ and $\left\|\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}\right\|_1/9$) in Figure 1, which illustrates the convergence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ respectively as N,T or both gets larger.

Next we consider the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$. We choose $\widehat{\beta}_1$ and $\widehat{\delta}_{1,3}$ with $(N,T)=(80,80)$ as examples for illustration. For each simulation, we construct $\widehat{\beta}_1$ and $\widehat{\delta}_{1,3}$, and standardize them according to the asymptotic results in Theorem 2 and Theorem 3 respectively. Figure 2 shows histograms and normal probability plots of the standardized estimators. They both show good fit for a standard normal distribution. It means that the asymptotic variance formulae in Theorem 2 and 3 are reliable for inference, and the way that we estimate any high dimensional covariance matrices mentioned in Section 4.1 helps in achieving an accurate estimation of the covariance matrices for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$. We actually get very similar good fits for the non-zero components of $\widetilde{\boldsymbol{\delta}}$, showing the asymptotic normality in Theorem 4 is reliable as well. The results are omitted here to save space.

On top of asymptotic normality, $\widetilde{\boldsymbol{\delta}}$ also enjoys sign consistency as shown in Theorem 4. We illustrate the selection consistency of $\widetilde{\boldsymbol{\delta}}$ in practice by calculating the specificity (i.e., proportion of correctly identified zeros) and the sensitivity (i.e., proportion of correctly identified non-zeros) of $\widetilde{\boldsymbol{\delta}}$. Table 1 shows that at various combinations of (N,T), the sensitivity and specificity are all 100%, showing perfect identifications of zeros and non-zeros. The table also shows the decreasing error for $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\delta}}$ as N or T increases.

		$T=40$	$T=80$	$T=120$
$N=60$	$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$	9.06(3.58)	6.26(1.04)	3.16(0.58)
	$\|\widetilde{\boldsymbol{\delta}}-\boldsymbol{\delta}\|_1$	0.13(0.05)	0.05(0.04)	0.02(0.02)
	$\widetilde{\boldsymbol{\delta}}$ Specificity	100%(0)	100%(0)	100%(0)
	$\widetilde{\boldsymbol{\delta}}$ Sensitivity	100%(0)	100%(0)	100%(0)
		$N=40$	$N=80$	$N=120$
$T=60$	$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$	7.60(0.89)	6.24(0.77)	3.51(0.65)
	$\|\widetilde{\boldsymbol{\delta}}-\boldsymbol{\delta}\|_1$	0.02(0.01)	0.01(0.00)	0.00(0.00)
	$\widetilde{\boldsymbol{\delta}}$ Specificity	100%(0)	100%(0)	100%(0)
	$\widetilde{\boldsymbol{\delta}}$ Sensitivity	100%(0)	100%(0)	100%(0)

Table 1: Mean L_1 error for $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\delta}}$. Standard deviations are shown in brackets. Sensitivity and specificity of $\widetilde{\boldsymbol{\delta}}$ are also shown for various combinations of T,N. The values of $\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_1$ are multiplied by 10^4.

5.2 Performance of BIC for choosing p

To examiner the performance of the BIC defined in (4.8), we run our simulations 100 times for each particular (N,T) combination using the same setting as before, except that each time p is randomly generated from 1 to 7. With each simulation, we construct the positive selection rate (PSR) and the false discovery rate (FDR), defined as

$$\text{PSR}=\frac{\sum_{j=1}^{100}|s_j^*\cap s_{0,j}|}{\sum_{j=1}^{100}|s_{0,j}|},\quad \text{FDR}=\frac{\sum_{j=1}^{100}|s_j^*\cap s_{0,j}^c|}{\sum_{j=1}^{100}|s_j^*|},$$

where $s_{0,j}$ represents the index set for all δ_{ir} that should be included in the model at the jth repetition. Since we do not set δ_{ir} to be exactly 0 in this experiment, we have $|s_{0,j}|=pM=3p$, where p is in fact different for different j. The set s_j^* is the index set for all $\widehat{\delta}_{ir}$ estimated when p is estimated as p^*. Clearly, if $p^*\le p$, then $|s_j^*\cap s_{0,j}|=|s_j^*|$ and $|s_j^*\cap s_{0,j}^c|=0$, meaning we may not be having the whole true set

$s_{0,j}$ but we do not falsely "discover" something that is not in $s_{0,j}$. On the other hand, if $p^* > p$, then $|s_j^* \cap s_{0,j}| = |s_{0,j}|$ and $|s_j^* \cap s_{0,j}^c| > 0$, meaning we have included all that are in $s_{0,j}$, but we have falsely "discovered" something outside of $s_{0,j}$. Hence in a sense, PSR measures an average number of times where we do not underestimate p, while FDR measures an average number of times we overestimate p. Ideally, we want PSR=100% while FDR = 0%. These two measures are also used in Chen and Chen (2008) and Chen and Chen (2012) in different contexts.

		$T = 40$	$T = 50$	$T = 60$
$N = 50$	PSR	100.00%	100.00%	98.00%
	FDR	2.00%	0.00%	0.00%
		$N = 40$	$N = 50$	$N = 60$
$T = 50$	PSR	98.00%	100.00%	100.00%
	FDR	0.00%	0.00%	2.00%

Table 2: Positive selection rate (PSR) and false discovery rate (FDR) for the choice of p using BIC defined in (4.8).

Table 2 shows the results. Our BIC definitely performs very well with PSR almost always equal 100% and FDR 0% in various (N, T) combinations.

6 Analysis of Stock Return Data

Spatial lag model has been widely applied to economic or geographic data, yet financial data is rarely analyzed using spatial econometrics tools. We illustrate the performance of our model using the daily log-returns of some important stocks in the Euro Stoxx 50 and S&P 500 in 2015. Our aim is to analyze the spatial interactions of these stocks and to see how different macroeconomic and financial indicators affect the dynamics of the returns.

The table below shows all the stocks we use:

France	Alstom, Total, BNP, Scociete, Sanofi, Carrefour, LVMH, Vivendi
Germany	Daimler, Allianz, Deutsche Bank
Italy	ENEL, ENI, Intesa, Unicredit, Tele Italy
Spain	Repsol, Banco, Telefonica
US	GM, PG, Nextera, American Express, Citi, Wells Frago, Amgen, Gilead, Johnson, Costco, Home, Centurylink, Verizon
Energy	Alstom, Total, ENEL, ENI, Repsol, PG, Nextera
Finance	BNP, Scociete, Allianz, Deutsche Bank, Intesa, Unicredit, Banco, American Express, Citi, Wells Fargo
Pharmacy	Sanofi, Amgen, Gilead, Johnson
Retails	Carrefour, LVMH, Costco, Home
Telecom	Vivendi, Tele Italy, Telefonica, Centurylink, Verizon
Auto	Daimler, GM

Arnold et al. (2013) illustrates, with the help of a spatial lag model, that the stocks belonging to the same country or the same industry are more related to each other, in the sense that spatial interactions of the log-returns are stronger. They analyze the Euro Stoxx 50 stock returns using a combination of three adjacency matrices as an estimator for the spatial weight matrix in their model. The first one being the weight of the stocks in Euro Stoxx 50, and the second and third ones being the adjacency matrices corresponding to the

same industry and to the same country, respectively. They found that all these matrices contribute to the final spatial weight matrix in their model, and improves risk estimation in a portfolio allocation exercise. However, no inferences on the estimated parameters are performed due to the lack of asymptotic results.

To fill in this gap and generalize on their model, we include four types of spatial weight matrix specifications instead of only three matrices as in Arnold et al. (2013). The first type is on the physical distance d_{ij} between city i and j where the headquarters of the stocks' associated companies are built. Three specified spatial weight matrices with elements $1/d_{ij}, 1/d_{ij}^2$ and $1/d_{ij}^3$ are included for selection. The second to fourth types coincide with the three adjacency matrices specified in Arnold et al. (2013). Namely, one contains the weights of stocks in Euro Stoxx 50 or S&P 500, and the remaining two having (i, j)th element equal to 1 if the corresponding stocks belong to the same industry or country respectively. This way, we have $M = 6$ specified spatial weight matrices for selection in our model. We have done row standardization on all of these six matrices.

As for the covariates $\boldsymbol{X}_t$, we use the Fama-French three factors (excess return = market return - risk free rate, SMB = Small (market capitalization) Minus Big, HML = High (book-to-market ratio) Minus Low), national stock index (S&P 500, CAC40, DAX, IBEX or MIB) and the corresponding European or US industry index for each stock. Hence $K = 5$, and we are treating these as exogenous covariates, so we set $\boldsymbol{B}_t = \boldsymbol{X}_t$, the same as the covariates. Minimizing the BIC defined in (4.8) results in $p = 1$.

	$1/d$	$1/d^2$	$1/d^3$	Stock weight	Country	Industry
$\widetilde{\boldsymbol{\delta}}_{0i}$	-0.0052	0.0811	-0.3880	0	0.0001	0.3122
	(0.0015)	(0.0036)	(0.0497)	(—)	(10^{-5})	(0.0346)
$\widetilde{\boldsymbol{\delta}}_{1i}$	0	0	-0.0612	0	2.22×10^{-5}	0.0610
	(—)	(—)	(0.0062)	(—)	(6.1×10^{-6})	(0.0051)

	Market excess return	SMB	HML	National Index	Industry Index
$\widehat{\boldsymbol{\beta}}$	1.516(0.616)	−4.884(2.662)	1.869(0.841)	14.788(5.563)	19.746(10.274)

Table 3: The values of $\widetilde{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}}$, where $p = 1$ and $\gamma_T = 1.6438$ are chosen by minimizing the BIC defined in (4.8). Estimated standard deviations are in brackets. All values associate with $\widehat{\boldsymbol{\beta}}$ are multiplied by 10^6.

Table 3 shows the values of $\widetilde{\boldsymbol{\delta}}$. Clearly, stock weights in their respective market indices do not contribute to the two spatial weight matrices $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$. However, the adjacency matrices for country and industry do contribute to both of the spatial weight matrices. For physical distance, clearly, a traditional approach where one chooses a distance $1/d$, $1/d^2$ or $1/d^3$ for the spatial weight matrix would fail, since it is clear that all three specified spatial weight matrices are significant and cannot be omitted for $\boldsymbol{W}_0$. Only the one for $1/d^3$ is significant to $\boldsymbol{W}_1$ though. In the same table, we can see that all factors in $\boldsymbol{X}_t$ are at least marginally significant, with national and industry indices play a more important role practically than the Fama-French three factors.

Figure 3 shows the heat map of the spatial weight matrices $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$. It is clear that there are some block patterns in these matrices, which mainly represent stocks in the same country or industry. Meanwhile, they are related strongly with each other in general if they are all from Europe or US, with France and Italy showing strong connections. It is interesting to note that the ninth stock Daimler, and the twentieth stock GM, are related to each other (two bright yellow dots on both $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$), although they belong to Germany and US auto-industry respectively. Since Daimler owns part of GM by spin-offs, the relation itself is not surprising. However, it means that our method of taking linear combination of

different specified spatial weight matrices can indeed reflect a general pattern of spatial interactions. In $\boldsymbol{W}_1$, we can also find some blocks for stocks in Germany and Spain.

References

Ahrens, A. and Bhattacharjee, A. (2015). Two-step lasso estimation of the spatial weights matrix. *Econometrics*, 3(1):128.

Anselin, L., Gallo, J. L., and Jayet, H. (2008). Spatial panel econometrics. In Mátyás, L. and Sevestre, P., editors, *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, pages 625–660. Springer Berlin Heidelberg, Berlin, Heidelberg.

Arnold, M., Stahlberg, S., and Wied, D. (2013). Modeling different kinds of spatial dependence in stock returns. *Empirical Economics*, 44(2):761–774.

Bhattacharjee, A. and Jensen-Butler, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, 43(4):617 – 634.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, pages 555–574.

Cliff, A. and Ord, J. (1973). *Spatial autocorrelation*. Monographs in spatial and environmental systems analysis. Pion.

Corrado, L. and Fingleton, B. (2012). Where is the economics in spatial econometrics? *Journal of Regional Science*, 52(2):210–239.

Dou, B., Parrella, M. L., and Yao, Q. (2016). Generalized yulewalker estimation for spatio-temporal models with unknown diagonal coefficients. *Journal of Econometrics*, 194(2):369 – 382.

Elhorst, J. P. (2005). Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels. *Geographical Analysis*, 37(1):85–106.

Foote, C. L. (2007). Space and time in macroeconomic panel data: young workers and state-level unemployment revisited. Working Papers 07-10, Federal Reserve Bank of Boston.

Franzese, R. J. and Hays, J. C. (2007). Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis*, 15(2):140–164.

Hallin, M., Lu, Z., Tran, L. T., et al. (2004). Local linear spatial regression. *The Annals of Statistics*, 32(6):2469–2500.

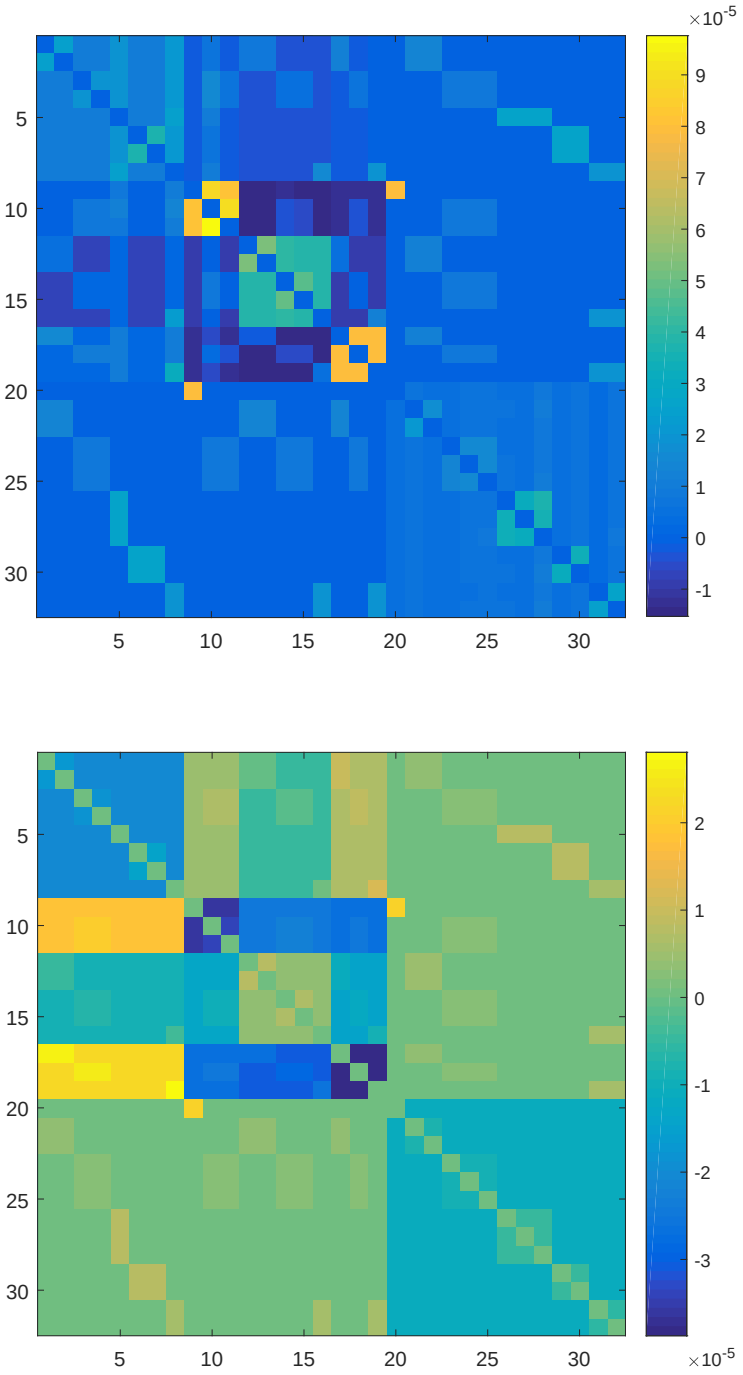

Figure 3: Upper: The estimate of $\boldsymbol{W}_0$. Lower: The estimate of $\boldsymbol{W}_1$. From 1 to 32, the stocks are Alstom, Total, BNP, Scociete, Sanofi, Carrefour, LVMH, Vivendi, Daimler, Allianz, Deutsche Bank, ENEL, ENI, Intesa, Unicredit, Tele Italy,Repsol, Banco, Telefonica, GM, PG, Nextera, American Express, Citi, Wells Frago, Amgen, Gilead, Johnson, Costco, Home, Centurylink and Verizon respectively.

Keller, W. and Shiue, C. H. (2007). The origin of spatial interaction. *Journal of Econometrics*, 140(1):304 – 332. Analysis of spatially dependent data.

Koroglu, M. and Sun, Y. (2016). Functional-coefficient spatial durbin models with nonparametric spatial weights: An application to economic growth. *Econometrics*, 4(1):6.

Kostov, P. (2013). Choosing the right spatial weighting matrix in a quantile regression model. *ISRN Economics*, 2013.

Lam, C. and Souza, P. (2014). Regularization for spatial panel time series using the adaptive lasso. Sticerd - econometrics paper series, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.

Lam, C. and Souza, P. C. (2015). One-step regularized spatial weight matrix and fixed effects estimation with instrumental variables. *Manuscript*.

Lam, C. and Souza, P. C. L. (2016). Detection and estimation of block structure in spatial weight matrix. *Econometric Reviews*, 35(8-10):1347–1376.

Lee, L.-f. and Yu, J. (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, 154(2):165 – 185.

Li, Y. (2003). A martingale inequality and large deviations. *Statistics & probability letters*, 62(3):317–321.

Robinson, P. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, 165(1):5 – 19. Moment Restriction-Based Econometric Methods.

Tran, L. and Yakowitz, S. (1993). Nearest neighbor estimators for random fields. *Journal of Multivariate Analysis*, 44(1):23 – 46.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14150–14154.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and Its Interface, 0*, pages 1–20.

7 Appendix

7.1 Technical assumptions

We present and explain the more technical assumptions of the paper in this section. Most of these assumptions are extended from Lam and Souza (2015).

R1. The column vectors $\text{vec}(\boldsymbol{W}_{0i}^T)$ in $\boldsymbol{V}_0$ are linearly independent to each other, such that there exists a constant $u > 0$ with $\sigma_M^2(\boldsymbol{V}_0) \geq u > 0$ uniformly as $N \to \infty$, where $\sigma_i(A)$ is the ith largest singular value of a matrix A. Moreover, $\max_{1 \leq i \leq M} \|\boldsymbol{W}_{0i}\|_1 \leq c < 1$ uniformly as $N \to \infty$ for some constant $c > 0$.

R2. Write $\boldsymbol{\epsilon}_t = \boldsymbol{\Sigma}_\epsilon^{1/2}\boldsymbol{\epsilon}_t^*$, where $\boldsymbol{\Sigma}_\epsilon$ is the covariance matrix for $\boldsymbol{\epsilon}_t$. Then the elements in $\boldsymbol{\Sigma}_\epsilon$ are all less than $\sigma^2_{\max}$ uniformly as $N \to \infty$. Same for the variance of the elements in $\boldsymbol{B}_t$. We also assume $\|\boldsymbol{\Sigma}_\epsilon^{1/2}\|_\infty \leq S_\epsilon < \infty$ uniformly as $N \to \infty$, with $\{\epsilon^*_{t,j}\}_{1\leq j\leq N}$ being a martingale difference with respect to the filtration generated by $\sigma(\epsilon^*_{t,1}, ..., \epsilon^*_{t,j})$. The tail condition $P(|Z| > v) \leq D_1 \exp(-D_2 v^q)$ is also satisfied by $\epsilon^*_{t,j}$.

R3. All singular values of $\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)$ are uniformly larger than Nu for some constant $u > 0$, while the maximum singular value is also of order N. Individual entries in the matrix $\mathbb{E}(\boldsymbol{x}_t\boldsymbol{b}_t^T)$ are uniformly bounded away from infinity.

R4. For the same constant a, we have for each N

$$\max_{1\leq i\leq N}\sum_{j=1}^{N}\Big\|\mathbb{E}(\sum_{q\geq 0}\boldsymbol{b}_{t,i}\boldsymbol{x}_{t-q,j}^T)\Big\|_{\max}, \quad \max_{1\leq j\leq N}\sum_{i=1}^{N}\Big\|\mathbb{E}(\sum_{q\geq 0}\boldsymbol{b}_{t,i}\boldsymbol{x}_{t-q,j}^T)\Big\|_{\max} \leq C_{bx}N^a$$

where $C_{bx} > 0$ is a constant and $\boldsymbol{b}_{t,i}$, $\boldsymbol{x}_{t,j}$ are the column vectors for the ith row of $\boldsymbol{B}_t$ and jth row of $\boldsymbol{X}_t$ respectively. At the same time, assume also that $\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\gamma})$ has all singular values of order N^{1+a}.

R5. Assume $0 < b < 1$. For fixed $= 1, \ldots, K$, the eigenvalues of $N^{-b}\text{var}(\boldsymbol{B}_{t,k})$ and $\text{var}(\boldsymbol{\epsilon}_T)$ are uniformly bounded away from 0 and infinity, and respectively dominates the singular values of the sum of $N^{-b}\text{cov}(\boldsymbol{B}_{t+\tau,k}, \boldsymbol{B}_{t,k})$ over $\tau \neq 0$ and the sum of $\mathbb{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T)$ over $\tau \neq 0$. Also, for each $i = 1, \ldots, N$, we assume that

$$\sum_\tau \sigma_i\big(N^{-b}\text{cov}(\boldsymbol{B}_{t+\tau,k}, \boldsymbol{B}_{t,k})\big) < \infty, \quad \sum_\tau \sigma_i\big(\mathbb{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau})\big) < \infty.$$

R6. Define $\lambda_T = cT^{-1/2}\log^{1/2}(T \vee N)$ for some constant $c > 0$. The tuning parameter γ_T is such that $\gamma_T = C\lambda_T$ for some constant $C > 0$.

R7. In all the assumptions above, we assume that as $N, T \to \infty$, $\lambda_T N^{1-a} = o(1)$, $N^{-a+b-1/w}\log^{-1}(T \vee N) = o(1)$, $\log(T \vee N)N^{1/w-b} = o(1)$ and $N^{b-a} = o(T\lambda_T)$.

Assumption R1 essentially requires that each specification $\boldsymbol{W}_{0i}$ is different from one another to a certain extent. This is intuitive, since if $\boldsymbol{W}_{0i}$ and $\boldsymbol{W}_{0l}$ are too similar to each other, the coefficients δ_{ji} and δ_{jl} are not well defined, and this will have anegative impact on the performance of our estimators.

The assumptions on $\boldsymbol{\Sigma}_\epsilon$ in R2 is mainly for the convenience of proofs, while the martingale difference assumption for $\boldsymbol{\epsilon}_t$ is a relaxation to independence.

Assumptions R3 and R4 are closely related. They paint a picture of how the exogenous variables in $\boldsymbol{B}_t$ are correlated with $\boldsymbol{X}_{t-q}$. Assumption R3 essentially says that the covariance between a variable in $\boldsymbol{B}_t$ and one in $\boldsymbol{X}_t$ is finite uniformly as $N \to \infty$. Then for $k = 1, \ldots, K$, considering the kth diagonal entry of $\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)$ is $\sum_{j=1}^N \mathbb{E}(X_{t,jk}B_{t,jk})$ with each $\mathbb{E}(X_{t,jk}B_{t,jk})$ being finite, it is indeed reasonable to assume that each diagonal entry in the matrix is of order N. This assumption is needed for the estimator $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\delta})$ to be well-defined.

Assumption R4 essentially describes how each row of variables in $\boldsymbol{B}_t$ are correlated with different rows of variables in $\boldsymbol{X}_t$. With this, we can actually derive easily that $\|\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\gamma})\|_1$ has order at most N^{1+a}. Hence the assumption of having all the singular values of $\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\gamma})$ of order N^{1+a} is reasonable.

Assumption R5 assumes a rate for the singular values of $\text{var}(\boldsymbol{B}_{t,k})$ essentially, which is important in certain asymptotic normality results. The rate N^b, possibly differing from N^a, is reasonable as well since the way that $\boldsymbol{B}_t$ and $\boldsymbol{X}_t$ are correlated do not directly indicate how the variables in $\boldsymbol{B}_t$ itself are correlated, unless of course when $\boldsymbol{B}_t = \boldsymbol{X}_t$ where $\boldsymbol{X}_t$ itself is exogenous, in which case $a = b$. The variance-covariance matrix being dominating the lag τ auto-covariances is for the ease of presentation of rates of convergence in the asymptotic normality results in this paper.

7.2 Proof of theorems

The followings are Lemma 1 and 2 of Lam and Souza (2015) respectively.

Lemma 1. *For a zero mean time series process $\mathbf{x}_t = \boldsymbol{f}(\mathcal{F})$ with dependence measure $\theta^x_{t,d,j}$ defined in Section 3, assume $\Theta^x_{m,a} \leq Cm^{-\alpha}$ as in Assumption M3. Then there exists constants C_1, C_2 and C_3 independent of v, T and the index j such that*

$$P(|1/T\sum_{t=1}^{T} x_{t,j}| > v) \leq \frac{C_1 T^{w(1/2-\widetilde{\alpha})}}{(Tv)^w} + C_2\exp(-C_3 T^{\widetilde{\beta}} v^2),$$

where $\widetilde{\alpha} = \alpha \wedge (1/2 - 1/w)$, and $\widetilde{\beta} = (3 + 2\widetilde{\alpha}w)/(1 + w)$.

Furthermore, assume another zero mean time series process z_t (can be the same process x_t) with both $\Theta^x_{m,2w}$, $\Theta^z_{m,2w} \leq Cm^{-\alpha}$, as in Assumption M3. Then provided $\max_j \|x_{tj}\|_{2w}$, $\max_j \|z_{tj}\|_{2w} \leq c_0 \leq \infty$ where c_0 is a constant, the above Nagaev-type inequality holds for the product process $\{x_{tj}z_{tl} - \mathbb{E}(x_{tj}z_{tl})\}$.

Lemma 2. *For any $N \times N$ matrix $\boldsymbol{H} = (h_1, \ldots, h_N)^T$ and any $N \times K$ matrix $\boldsymbol{M}$, define*

$$\boldsymbol{V}_{\boldsymbol{H}} = \begin{pmatrix} \boldsymbol{I}_K \otimes h_1 \\ \vdots \\ \boldsymbol{I}_K \otimes h_N \end{pmatrix}.$$

Then we have

$$\boldsymbol{H}\boldsymbol{M} = (\boldsymbol{I}_N \otimes \text{vec}^T(\boldsymbol{M}))\boldsymbol{V}_{\boldsymbol{H}}.$$

We first present an Theorem 5 which states that a set $\mathcal{M}$ is such that $P(\mathcal{M}) \to 1$ as $T, N \to \infty$, and our estimators enjoy nice properties on $\mathcal{M}$. This theorem is in fact exactly the same as Theorem S.1 of Lam and Souza (2015).

Denote $B_{t,ij}$ and $X_{t,ij}$ the (i,j) entry of $\boldsymbol{B}_t$ and $\boldsymbol{X}_t$ respectively, and define $\mathcal{M} = \cap_{i=1}^{7}\mathcal{A}_i$, where

$$\mathcal{A}_1 = \left\{ \max_{1\le i,k\le N} \max_{1\le j,l\le K} |T^{-1}\sum_{t=1}^{T}[B_{t,ij}X_{t,kl} - \mathbb{E}(B_{t,ij}X_{t,kl})]| < \lambda_T \right\},$$
$$\mathcal{A}_2 = \left\{ \max_{1\le i,k\le N} \max_{1\le j\le K} |T^{-1}\sum_{t=1}^{T} B_{t,ij}\epsilon_{t,k}| < \lambda_T \right\},$$
$$\mathcal{A}_3 = \left\{ \max_{1\le k\le K} |T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{N} B_{t,sk}\epsilon_{t,s}| < \lambda_T N^{1/2+1/2w} \right\},$$
$$\mathcal{A}_4 = \left\{ \max_{1\le i\le N} \max_{1\le j\le K} |\bar{B}_{.,ij} - \mathbb{E}(B_{t,ij})| < \lambda_T \right\},$$
$$\mathcal{A}_5 = \left\{ \max_{1\le j\le N} |\bar{\epsilon}_{.,j}| < \lambda_T \right\},$$
$$\mathcal{A}_6 = \left\{ \max_{1\le i\le N} \max_{1\le j\le K} |\bar{X}_{.,ij}| < \lambda_T \right\},$$
$$\mathcal{A}_7 = \left\{ \max_{1\le k\le K} |\sum_{s=1}^{N} \bar{B}_{.,sk}\bar{\epsilon}_{.,s}| < 2^{1/2}\lambda_T N^{1/2}\log^{1/2}(T\vee N) S_\epsilon(\max_{i.j}|\mathbb{E}(B_{t,ij})| + \lambda_T) \right\}.$$

Theorem 5. *Let Assumptions M1-M4 in Section 3.1 and R1-R7 in Section 7.1 hold. Suppose $\alpha \ge 1/2 - 1/w$ in Assumption M3, and for the application of the Nagaev-type inequality in Lemma 1 for the processes defined in $\mathcal{A}_1$ to $\mathcal{A}_7$, suppose the constants C_1, C_2 and C_3 are the same. Then with $c \ge \sqrt{3/C_3}$ where c is the constant defined in $\lambda_T = cT^{-1/2}\log^{1/2}(T\vee N)$, we have*

$$P(\mathcal{M}) \ge 1 - 8C_1K^2(C_3/3)^{w/2}\frac{N^2}{T^{w/2-1}\log^{w/2}(T\vee N)} - \frac{8C_2K^2N^2}{T^3\vee N^3} - \frac{2K}{T\vee N}.$$

It approaches 1 if we assume further that $N = o(T^{w/4-1/2}\log^{w/4}(T))$.

Proof of Theorem 1

From (2.4) and that

$$\begin{aligned} y_0^v &= \sum_{i=1}^{M}\delta_{0i}\boldsymbol{W}_{0i}^{\otimes}y_0^v + \sum_{j=1}^{p}\Big(\sum_{i=1}^{M}\delta_{ji}\boldsymbol{W}_{0i}^{\otimes}\Big)y_j^v + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^v + \mathbf{1}_T\otimes\boldsymbol{\mu} \\ &= \Big(\boldsymbol{I}_{TN} - \sum_{i=1}^{M}\delta_{0i}\boldsymbol{W}_{0i}^{\otimes}\Big)^{-1}\Big(\sum_{j=1}^{p}\Big(\sum_{i=1}^{M}\delta_{ji}\boldsymbol{W}_{0i}^{\otimes}\Big)y_j^v + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^v + \mathbf{1}_T\otimes\boldsymbol{\mu}\Big), \end{aligned}$$

it is easy to get, since $\boldsymbol{B}^{vT}(\mathbf{1}_T\otimes\boldsymbol{\mu}) = \mathbf{0}$, that

$$\boldsymbol{\beta}(\boldsymbol{\delta}) - \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v.$$

Moreover,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\delta}}) &= (\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\Big[\Big(\boldsymbol{I}_{TN} - \sum_{i=1}^{M}\widehat{\delta}_{0i}\boldsymbol{W}_{0i}^{\otimes}\Big)y_0^v - \sum_{j=1}^{p}\Big(\sum_{i=1}^{M}\widehat{\delta}_{ji}\boldsymbol{W}_{0i}^{\otimes}\Big)y_j^v\Big] \\
&= (\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\Big[\Big(\boldsymbol{I}_{TN} - \sum_{i=1}^{M}\delta_{0i}\boldsymbol{W}_{0i}^{\otimes}\Big)y_0^v - \sum_{j=1}^{p}\Big(\sum_{i=1}^{M}\delta_{ji}\boldsymbol{W}_{0i}^{\otimes}\Big)y_j^v \\
&+ \sum_{i=1}^{M}(\delta_{0i} - \widehat{\delta}_{0i})\boldsymbol{W}_{0i}^{\otimes}y_0^v + \sum_{j=1}^{p}\Big(\sum_{i=1}^{M}(\delta_{ji} - \widehat{\delta}_{ji})\boldsymbol{W}_{0i}^{\otimes}\Big)y_j^v\Big] \\
&= \boldsymbol{\beta}(\boldsymbol{\delta}) + (\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT} \\
&\qquad \cdot \Big[\sum_{i=1}^{M}(\delta_{0i} - \widehat{\delta}_{0i})\boldsymbol{W}_{0i}^{\otimes}y_0^v + \sum_{j=1}^{p}\Big(\sum_{i=1}^{M}(\delta_{ji} - \widehat{\delta}_{ji})\boldsymbol{W}_{0i}^{\otimes}\Big)y_j^v\Big].
\end{aligned}$$

Using the above, we can decompose

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= I_0 + I_1 + I_2 + I_3 + I_4 + I_5, \text{ where} \\
I_0 &= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}(\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t) - T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \\
I_1 &= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v, \\
I_2 &= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}(\sum_{i=1}^{M}(\delta_{0i} - \widehat{\delta}_{0i})\boldsymbol{W}_{0i}^{\otimes})\boldsymbol{\Pi}^{\otimes}\boldsymbol{X}\boldsymbol{\beta}, \\
I_3 &= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}(\sum_{i=1}^{M}(\delta_{0i} - \widehat{\delta}_{0i})\boldsymbol{W}_{0i}^{\otimes})\boldsymbol{\Pi}^{\otimes}\boldsymbol{\epsilon}^v, \\
I_4 &= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}(\sum_{i=1}^{M}(\delta_{0i} - \widehat{\delta}_{0i})\boldsymbol{W}_{0i}^{\otimes})\boldsymbol{\Pi}^{\otimes}(\sum_{j=1}^{p}(\sum_{i=1}^{M}\delta_{ji}\boldsymbol{W}_{0i}^{\otimes})y_j^v), \\
I_5 &= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}(\sum_{j=1}^{p}(\sum_{i=1}^{M}(\delta_{ji} - \widehat{\delta}_{ji})\boldsymbol{W}_{0i}^{\otimes}y_j^v)),
\end{aligned}$$

with $\boldsymbol{\Pi}^{\otimes} = (\boldsymbol{I}_{TN} - \sum_{i=1}^{M}\delta_{0i}\boldsymbol{W}_{0i}^{\otimes})^{-1}$. We need to find the rate of convergence of I_0 I_1, I_2, I_3, I_4 and I_5. To this end, using Assumption R3 in Section 7.1,

$$\|\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t)^{-1}\|_1 \leq \frac{K^{1/2}}{\lambda_{\min}(\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))} \leq \frac{K^{1/2}}{N^2u^2}.$$

Define $\boldsymbol{U} = \boldsymbol{I}_N \otimes T^{-1}\sum_{t=1}^{T}\text{vec}(\boldsymbol{B}_t - \bar{\boldsymbol{B}})\text{vec}^T(\boldsymbol{X}_t)$ and $\boldsymbol{U}_0 = \boldsymbol{I}_N \otimes \mathbb{E}(\boldsymbol{b}_t\boldsymbol{x}_t^T)$, then we can write $\mathrm{T}^{-1}\boldsymbol{X}^T\boldsymbol{B}^v = \boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}\boldsymbol{V}_{\boldsymbol{I}_N}$ and $\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t) = \boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0\boldsymbol{V}_{\boldsymbol{I}_N}$. Also, denote $\boldsymbol{W}_j^c$,$\boldsymbol{B}_{t,j}$ and $\boldsymbol{X}_{t,j}$ the jth column of $\boldsymbol{W}$, $\boldsymbol{B}_t$ and

$\boldsymbol{X}_t$ respectively, and let $\boldsymbol{\pi}_j^T$ be the jth row of $\boldsymbol{\Pi}$. Then on $\mathcal{M}$,

$$\begin{aligned}
\|I_0\|_1 &\le \|(\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}\|_1\|((\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t)) - T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_1 \\
&\le \frac{K^{1/2}}{N^2u^2}\left[\|\boldsymbol{V}_{\boldsymbol{I}_N}^T(\boldsymbol{U}_0-\boldsymbol{U})^T\boldsymbol{V}_{\boldsymbol{I}_N}\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0\|_1 + \|\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}^T\boldsymbol{V}_{\boldsymbol{I}_N}\boldsymbol{V}_{\boldsymbol{I}_N}^T(\boldsymbol{U}_0-\boldsymbol{U})\|_1\right]\|\boldsymbol{V}_{\boldsymbol{I}_N}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_1 \\
&\le \frac{K^{1/2}}{N^2u^2}\Bigg[K\|\boldsymbol{U}_0-\boldsymbol{U}\|_{\max}\cdot N\cdot K\|\boldsymbol{U}_0\|_{\max} \\
&+ (K\|\boldsymbol{V}_{\boldsymbol{I}_N}^T(\boldsymbol{U}-\boldsymbol{U}_0)^T\boldsymbol{V}_{\boldsymbol{I}_N}\|_{\max} + K\|\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0^T\boldsymbol{V}_{\boldsymbol{I}_N}\|_{\max})\cdot K\|\boldsymbol{U}_0-\boldsymbol{U}\|_{\max}\Bigg]\cdot N\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \\
&\le K^{1/2}(2\lambda_T\sigma_{bx}(1+\mu_{b,\max}+\lambda_T)+\lambda_T^2(1+\mu_{b,\max}+\lambda_T)^2)\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \\
&= O(\lambda_T\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1),
\end{aligned}$$

where $\mu_{b,\max} = \big\|\mathbb{E}(\boldsymbol{b}_t)\big\|_{\max}$. At the same time, on $\mathcal{M}$,

$$\begin{aligned}
\|I_1\|_1 &\le \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1\|T^{-1}\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v\|_1 \\
&\le \frac{K^{1/2}}{N^2u^2}\|\boldsymbol{V}_{\boldsymbol{I}_N}^T(\boldsymbol{U}-\boldsymbol{U}_0)\boldsymbol{V}_{\boldsymbol{I}_N} + \boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0^T\boldsymbol{V}_{\boldsymbol{I}_N}\|_1 \\
&\cdot (K\lambda_T N^{1/2+1/2w} + \sqrt{2}K\lambda_T N^{1/2}\log(T\vee N)S_\epsilon(\mu_{b,max}+\lambda_T)) \\
&\le \frac{K^{1/2}}{N^2u^2}N(\lambda_T(1+\mu_{b,\max}+\lambda_T)+\sigma_{bx}) \\
&\cdot (K\lambda_T N^{1/2+1/2w} + \sqrt{2}K\lambda_T N^{1/2}\log(T\vee N)S_\epsilon(\mu_{b,\max}+\lambda_T)) \\
&= O(\lambda_T N^{-1/2+1/2w}).
\end{aligned}$$

Recall that $\boldsymbol{W}_j = \sum_{i=1}^M \delta_{ji}\boldsymbol{W}_{0i}$, and denoting $\widehat{\boldsymbol{W}}_j = \sum_{i=1}^M \widehat{\delta}_{ji}\boldsymbol{W}_{0i}$ for $j=0,1,\ldots,p$, then on $\mathcal{M}$,

$$\begin{aligned}
\|I_2\|_1 &\le \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1\|T^{-1}\sum_{t=1}^T(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T(\boldsymbol{W}_0-\widehat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\boldsymbol{X}_t\|_1\|\boldsymbol{\beta}\|_1 \\
&\le \frac{K^{1/2}}{N^2u^2}O(N)\Big(K\cdot\max_{1\le r\le K}\Big|\sum_{j=1}^N(\boldsymbol{W}_{0,j}^c-\widehat{\boldsymbol{W}}_{0,j}^c)^TT^{-1}\sum_{t=1}^T(\boldsymbol{B}_{t,r}-\bar{\boldsymbol{B}}_{\cdot,r})\boldsymbol{X}_{t,r}^T\boldsymbol{\pi}_j\Big|\Big) \\
&\le O(N^{-1})(\sum_{j=1}^N(\lambda_T(1+\mu_{b,\max}+\lambda_T)+\sigma_{bx})\|\boldsymbol{W}_{0,j}^c-\widehat{\boldsymbol{W}}_{0,j}^c\|_1\|\boldsymbol{\pi}_j\|_1) \\
&\le O(N^{-1})(N\|\boldsymbol{\delta}_0-\widehat{\boldsymbol{\delta}}_0\|_1) = O(\|\boldsymbol{\delta}_0-\widehat{\boldsymbol{\delta}}_0\|_1).
\end{aligned}$$

Similarly, on $\mathcal{M}$,

$$\begin{aligned}
\|I_3\|_1 &\le \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1\|T^{-1}\sum_{t=1}^T(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T(\boldsymbol{W}_0-\widehat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\boldsymbol{\epsilon}_t\|_1 \\
&\le \frac{K^{1/2}}{N^2u^2}O(N)\Big(K\max_{1\le r\le K}\Big|\sum_{j=1}^N(\boldsymbol{W}_{0,j}^c-\widehat{\boldsymbol{W}}_{0,j}^c)^T\Big(T^{-1}\sum_{t=1}^T(\boldsymbol{B}_{t,r}-\bar{\boldsymbol{B}}_{\cdot,r})\boldsymbol{\epsilon}_t^T\Big)\boldsymbol{\pi}_j\Big|\Big) \\
&\le O(N^{-1})\cdot O(N\lambda_T\max_{1\le j\le N}\big\|\boldsymbol{\pi}_j\big\|_1\max_{1\le j\le N}\big\|\boldsymbol{W}_{0,j}^c-\widehat{\boldsymbol{W}}_{0,j}^c\big\|_1) = O(\lambda_T\big\|\boldsymbol{\delta}_0-\widehat{\boldsymbol{\delta}}_0\big\|_1).
\end{aligned}$$

For bounding $\|I_4\|_1$ and $\|I_5\|_1$, recall that from Section 3.1, we can express y_t as

$$y_t = \boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}(\boldsymbol{\mu} + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\epsilon}_t) = \sum_{q\geq 0} \boldsymbol{\Psi}_q \boldsymbol{\Pi}(\boldsymbol{\mu} + \boldsymbol{X}_{t-q}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{t-q}), \tag{7.9}$$

where $\boldsymbol{\Psi}_q$ is $N \times N$ such that $\sum_{q\geq 0} \|\boldsymbol{\Psi}_q\|_\infty < \infty$ because of stationarity. Then we can decompose

$$\begin{aligned}
\|I_4\|_1 &\leq \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1 \Big\|T^{-1}\sum_{t=1}^T (\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T(\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\sum_{j=1}^p \boldsymbol{W}_j\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}(\boldsymbol{X}_{t-j}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{t-j})\Big\|_1 \\
&= O(N^{-1}(\|I_{41}\|_1 + \|I_{42}\|_1)), \text{ where} \\
\|I_{41}\|_1 &= \Big\|T^{-1}\sum_{t=1}^T (\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T(\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\sum_{j=1}^p \boldsymbol{W}_j\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}\boldsymbol{X}_{t-j}\boldsymbol{\beta}\Big\|_1, \\
\|I_{42}\|_1 &= \Big\|T^{-1}\sum_{t=1}^T (\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T(\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\sum_{j=1}^p \boldsymbol{W}_j\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}\boldsymbol{\epsilon}_{t-j}\Big\|_1.
\end{aligned}$$

On $\mathcal{M}$, we have

$$\begin{aligned}
\|I_{41}\|_1 &\leq \max_{1\leq j\leq p}\max_{1\leq r,k\leq K} pK^2\|\boldsymbol{\beta}\|_1 \Big|\sum_{q\geq 0}\Big\{T^{-1}\sum_{t=1}^T (\boldsymbol{B}_{t,r} - \bar{\boldsymbol{B}}_{\cdot,r})^T(\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\boldsymbol{W}_j\boldsymbol{\Psi}_q\boldsymbol{\Pi}\boldsymbol{X}_{t-q-j,k}\Big\}\Big| \\
&= O(N\sigma_{bx}\|\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_0\|_\infty\|\boldsymbol{W}_j\|_\infty\|\boldsymbol{\Pi}\|_\infty^2\sum_{q\geq 0}\|\boldsymbol{\Psi}_q\|_\infty) \\
&= O(N\|\boldsymbol{\delta}_0 - \widehat{\boldsymbol{\delta}}_0\|_1),
\end{aligned}$$

where the second line is by Assumption R4. At the same time on $\mathcal{M}$,

$$\begin{aligned}
\|I_{42}\|_1 &\leq \max_{1\leq j\leq p}\max_{1\leq r\leq K} pK\Big|\sum_{q\geq 0}\Big\{T^{-1}\sum_{t=1}^T (\boldsymbol{B}_{t,r} - \bar{\boldsymbol{B}}_{\cdot,r})^T(\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\boldsymbol{W}_j\boldsymbol{\Psi}_q\boldsymbol{\Pi}\boldsymbol{\epsilon}_{t-q-j}\Big\}\Big| \\
&= O(N\lambda_T\|\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_0\|_\infty\|\boldsymbol{W}_j\|_\infty\|\boldsymbol{\Pi}\|_\infty^2\sum_{q\geq 0}\|\boldsymbol{\Psi}_q\|_\infty) \\
&= O(N\lambda_T\|\boldsymbol{\delta}_0 - \widehat{\boldsymbol{\delta}}_0\|_1),
\end{aligned}$$

where the second line follows from the rate on $\mathcal{A}_2$. These imply that on $\mathcal{M}$,

$$\|I_4\|_1 = O(\|\boldsymbol{\delta}_0 - \widehat{\boldsymbol{\delta}}_0\|_1).$$

To bound $\big\|I_5\big\|_1$, we can decompose

$$\begin{aligned}\|I_5\|_1 &\le \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1\Big\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\sum_{j=1}^{p}(\boldsymbol{W}_j-\widehat{\boldsymbol{W}}_j)\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}(\boldsymbol{X}_{t-j}\boldsymbol{\beta}+\boldsymbol{\epsilon}_{t-j})\Big\|_1\\ &= O(N^{-1}(\|I_{51}\|_1+\|I_{52}\|_1)),\quad \text{where}\\ \|I_{51}\|_1 &= \Big\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\sum_{j=1}^{p}(\boldsymbol{W}_j-\widehat{\boldsymbol{W}}_j)\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}\boldsymbol{X}_{t-j}\boldsymbol{\beta}\Big\|_1,\\ \|I_{52}\|_1 &= \Big\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\sum_{j=1}^{p}(\boldsymbol{W}_j-\widehat{\boldsymbol{W}}_j)\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}\boldsymbol{\epsilon}_{t-j}\Big\|_1.\end{aligned}$$

To bound $\big\|I_{51}\big\|_1$, similar to the treatment on $\big\|I_{41}\big\|_1$, on $\mathcal{M}$,

$$\begin{aligned}\big\|I_{51}\big\|_1 &\le \max_{1\le r,k\le K}K^2\big\|\boldsymbol{\beta}\big\|_1\Big|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_{t,r}-\bar{\boldsymbol{B}}_{\cdot,r})^T\sum_{j=1}^{p}(\boldsymbol{W}_j-\widehat{\boldsymbol{W}}_j)\sum_{q\ge0}\boldsymbol{\Psi}_q\boldsymbol{\Pi}\boldsymbol{X}_{t-q-j,k}\Big|\\ &= O(N\sigma_{bx}\sum_{j=1}^{p}\big\|\boldsymbol{W}_j-\widehat{\boldsymbol{W}}_j\big\|_\infty\big\|\boldsymbol{\Pi}\big\|_\infty\sum_{q\ge0}\big\|\boldsymbol{\Psi}_q\big\|_\infty)\\ &= O(N\big\|\boldsymbol{\delta}-\widehat{\boldsymbol{\delta}}\big\|_1).\end{aligned}$$

Finally, on $\mathcal{M}$,

$$\begin{aligned}\big\|I_{52}\big\|_1 &\le \max_{1\le r\le K}K\Big|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_{t,r}-\bar{\boldsymbol{B}}_{\cdot,r})^T\sum_{j=1}^{p}(\boldsymbol{W}_j-\widehat{\boldsymbol{W}}_j)\sum_{q\ge0}\boldsymbol{\Psi}_q\boldsymbol{\Pi}\boldsymbol{\epsilon}_{t-q-j}\Big|\\ &= O(N\lambda_T\sum_{j=1}^{p}\big\|\boldsymbol{W}_j-\widehat{\boldsymbol{W}}_j\big\|_\infty\big\|\boldsymbol{\Pi}\big\|_\infty\sum_{q\ge0}\big\|\boldsymbol{\Psi}_q\big\|_\infty)\\ &= O(N\lambda_T\big\|\boldsymbol{\delta}-\widehat{\boldsymbol{\delta}}\big\|_1).\end{aligned}$$

Hence on $\mathcal{M}$, we have

$$\big\|I_5\big\|_1 = O(\big\|\boldsymbol{\delta}-\widehat{\boldsymbol{\delta}}\big\|_1).$$

Combining the rates for $\big\|I_0\big\|_1$ to $\big\|I_5\big\|_1$, we can conclude that on $\mathcal{M}$,

$$\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big\|_1 = O(\lambda_T N^{-1/2+1/2w}+\big\|\boldsymbol{\delta}-\widehat{\boldsymbol{\delta}}\big\|_1). \tag{7.10}$$

We need to find the order of $\|\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}\|_1$. From (2.3) and (2.5), it is easy to show that

$$\boldsymbol{K}y_0^v-\boldsymbol{B}^Ty = \boldsymbol{K}y_0^v-(\boldsymbol{B}^T\boldsymbol{\epsilon}+\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}\boldsymbol{\delta}+\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}}\mathrm{vec}(\boldsymbol{I}_N)),$$

where

$$\begin{aligned}\boldsymbol{K}y_0^v &= \boldsymbol{H}\boldsymbol{\delta} + \boldsymbol{K}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}\boldsymbol{\epsilon}^v \\ &= \boldsymbol{H}\boldsymbol{\delta} + T^{-1/2}N^{-a/2}\sum_{t=1}^{T}\boldsymbol{X}_t \otimes (\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\gamma}\boldsymbol{\beta} + \boldsymbol{K}\boldsymbol{\epsilon}^v \\ &= \boldsymbol{H}\boldsymbol{\delta} + \boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}}\mathrm{vec}(\boldsymbol{I}_N) + \boldsymbol{K}\boldsymbol{\epsilon}^v.\end{aligned}$$

Hence,

$$\boldsymbol{K}y_0^v - \boldsymbol{B}^T y = -\boldsymbol{B}^T\boldsymbol{\epsilon} + (\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\boldsymbol{\delta} + \boldsymbol{K}\boldsymbol{\epsilon}^v.$$

Substituting the above back to (2.5), we can decompose

$$\begin{aligned}\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} &= \left[(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\right]^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T\left[\boldsymbol{K}\boldsymbol{\epsilon}^v - \boldsymbol{B}^T\boldsymbol{\epsilon}\right] = D_1 + D_2, \quad \text{where} \\ D_1 &= \left[(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\right]^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T\boldsymbol{K}\boldsymbol{\epsilon}^v, \\ D_2 &= -\left[(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\right]^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T\boldsymbol{B}^T\boldsymbol{\epsilon}.\end{aligned}$$

To bound $\|D_1\|_1$ and $\|D_2\|_1$, we introduce some notations and find their L_1 norm bounds first. For $i = 1, \ldots, M$, define

$$\boldsymbol{U}_q = \boldsymbol{I}_N \otimes T^{-1}\sum_{t=1}^{T}\mathrm{vec}(\boldsymbol{B}_t - \bar{\boldsymbol{B}})\mathrm{vec}^T(\boldsymbol{X}_{t-q}), \quad \boldsymbol{U}_{0q} = \boldsymbol{I}_N \otimes \mathbb{E}(\boldsymbol{b}_t\boldsymbol{x}_{t-q}^T).$$

Also, define for $i = 1, \ldots, M$ and $j = 1, \ldots, p$,

$$\begin{aligned}&\boldsymbol{A}_1 = T^{-1}\sum_{t=1}^{T}\boldsymbol{X}_t \otimes (\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\gamma}, && \boldsymbol{A}_1^0 = \mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\gamma}), \\ &\boldsymbol{A}_2 = (\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}^T\boldsymbol{V}_{\boldsymbol{I}_N}\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}\boldsymbol{V}_{\boldsymbol{I}_N})^{-1}, && \boldsymbol{A}_2^0 = (\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0^T\boldsymbol{V}_{\boldsymbol{I}_N}\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0\boldsymbol{V}_{\boldsymbol{I}_N})^{-1}, \\ &\boldsymbol{A}_3 = \boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}^T\boldsymbol{V}_{\boldsymbol{I}_N}, && \boldsymbol{A}_3^0 = \boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0^T\boldsymbol{V}_{\boldsymbol{I}_N}, \\ &\boldsymbol{A}_{4ij} = \sum_{q=0}^{\infty}\boldsymbol{V}_{\boldsymbol{W}_{0i}^T}^T\boldsymbol{U}_{q+j}\boldsymbol{V}_{\widetilde{\boldsymbol{\Pi}}_q}\boldsymbol{\beta}, && \boldsymbol{A}_{4ij}^0 = \sum_{q=0}^{\infty}\boldsymbol{V}_{\boldsymbol{W}_{0i}^T}^T\boldsymbol{U}_{0,q+j}\boldsymbol{V}_{\widetilde{\boldsymbol{\Pi}}_q}\boldsymbol{\beta}, \\ &\boldsymbol{A}_{5ij} = \sum_{q=0}^{\infty}\boldsymbol{V}_{\boldsymbol{W}_{0i}^T}^T\Big(\boldsymbol{I}_N \otimes T^{-1}\sum_{t=1}^{T}\mathrm{vec}(\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\epsilon}_{t-q-j}^T\Big)\mathrm{vec}(\widetilde{\boldsymbol{\Pi}}_q^T).\end{aligned} \tag{7.11}$$

where $\widetilde{\boldsymbol{\Pi}}_q = \boldsymbol{\Psi}_q\boldsymbol{\Pi}$. It is straightforward to see that, on $\mathcal{M}$,

$$\|\boldsymbol{A}_1 - \boldsymbol{A}_1^0\|_{\max} = O(\lambda_T) \tag{7.12}$$

Meanwhile, by Assumptions R4 and R7, on $\mathcal{M}$,

$$\|\boldsymbol{A}_1\|_1 \le \|\boldsymbol{A}_1^0\|_1 + \|\boldsymbol{A}_1 - \boldsymbol{A}_1^0\|_1 = O(N^{1+a} + \lambda_T N^2) = O(N^{1+a}). \tag{7.13}$$

Similarly, on $\mathcal{M}$,

$$\|\boldsymbol{A}_3^0\|_1 \le K\|\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0^T\boldsymbol{V}_{\boldsymbol{I}_N}\|_{\max} = O(N), \quad \|\boldsymbol{A}_3 - \boldsymbol{A}_3^0\|_1 = O(\lambda_T N), \quad \|\boldsymbol{A}_3\|_1 = O(N). \tag{7.14}$$

As $\boldsymbol{A}_2^0 = (\boldsymbol{A}_3^0\boldsymbol{A}_3^{0T})^{-1}$,

$$\|\boldsymbol{A}_2^0\|_1 \le \frac{K^{1/2}}{\lambda_{\min}(\boldsymbol{A}_3^0\boldsymbol{A}_3^{0T})} \le \frac{K^{1/2}}{N^2u^2} = O(N^{-2}). \tag{7.15}$$

Moreover, we know that $\boldsymbol{A}_2 - \boldsymbol{A}_2^0 = (\boldsymbol{A}_2 - \boldsymbol{A}_2^0)((\boldsymbol{A}_2^0)^{-1} - \boldsymbol{A}_2^{-1})\boldsymbol{A}_2^0 + \boldsymbol{A}_2^0((\boldsymbol{A}_2^0)^{-1} - \boldsymbol{A}_2^{-1})\boldsymbol{A}_2^0$, and on $\mathcal{M}$,

$$\|(\boldsymbol{A}_2^0)^{-1} - \boldsymbol{A}_2^{-1}\|_1 = \|\boldsymbol{A}_3^0\boldsymbol{A}_3^{0T} - \boldsymbol{A}_3\boldsymbol{A}_3^T\|_1 \le \|\boldsymbol{A}_3^0 - \boldsymbol{A}_3\|_1\|\boldsymbol{A}_3^{0T}\|_1 + \|\boldsymbol{A}_3\|_1\|\boldsymbol{A}_3^{0T} - \boldsymbol{A}_3^T\|_1 = O(\lambda_T N^2).$$

Therefore, on $\mathcal{M}$,

$$\|\boldsymbol{A}_2 - \boldsymbol{A}_2^0\|_1 \le \frac{\|(\boldsymbol{A}_2^0)^{-1} - \boldsymbol{A}_2^{-1}\|_1\|\boldsymbol{A}_2^0\|_1^2}{1 - O(\lambda_T N^2 N^{-2})} = O\Big(\frac{\lambda_T N^2 N^{-4}}{1 - \lambda_T N^2 N^{-2}}\Big) = O(\lambda_T N^{-2}). \tag{7.16}$$

As for $\|\boldsymbol{A}_{4ij}\|_1$, by Assumptions M1 and R4, defining $\widetilde{\boldsymbol{\pi}}_{q,r}^T$ to be the rth row of $\widetilde{\boldsymbol{\Pi}}_q$, we have on $\mathcal{M}$,

$$\begin{aligned}\|\boldsymbol{A}_{4ij}^0\|_1 &\le \sum_{q=0}^{\infty} K\|\boldsymbol{\beta}\|_1\|\boldsymbol{V}_{\boldsymbol{W}_{0i}^T}^T\boldsymbol{U}_{0,q+j}\boldsymbol{V}_{\widetilde{\boldsymbol{\Pi}}_q}\|_{\max} = \sum_{q=0}^{\infty} K\|\boldsymbol{\beta}\|_1 \max_{1\le k,m\le K}\Big|\sum_{r=1}^{N}\boldsymbol{W}_{0i,r}^{cT}\mathbb{E}(\boldsymbol{X}_{t-q-j,k}\boldsymbol{B}_{t,m}^T)\widetilde{\boldsymbol{\pi}}_{q,r}\Big| \\ &= O(\|\boldsymbol{W}_{0i}\|_1\sum_{q\ge 0}\|\widetilde{\boldsymbol{\Pi}}_q\|_\infty \cdot N) \le O(\|\boldsymbol{W}_{0i}\|_1\sum_{q\ge 0}(\|\boldsymbol{\Pi}\|_\infty\|\boldsymbol{\Psi}_q\|_\infty)\cdot N) = O(N).\end{aligned} \tag{7.17}$$

Similarly, we can easily show on $\mathcal{M}$ that

$$\|\boldsymbol{A}_{4ij} - \boldsymbol{A}_{4ij}^0\|_1 = O(\lambda_T N).$$

Hence we have

$$\|\boldsymbol{A}_{4ij}\|_1 = O(N). \tag{7.18}$$

To bound $\|\boldsymbol{A}_{5ij}\|_1$, an element in $\boldsymbol{A}_{5ij}$ is bounded on $\mathcal{M}$ by

$$\Big|\sum_{r=1}^{N}\sum_{q=0}^{\infty}\boldsymbol{W}_{0i,r}^{cT}T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_{t,k} - \bar{\boldsymbol{B}}_k)\boldsymbol{\epsilon}_{t-q-j}^T\widetilde{\boldsymbol{\pi}}_{q,r}\Big| = O(\lambda_T N), \quad \text{so} \quad \|\boldsymbol{A}_{5ij}\|_1 = O(\lambda_T N). \tag{7.19}$$

We now decompose $D_1 = F_1 + F_2 + F_3$, where

$$\begin{aligned}F_1 &= \big[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\big]^{-1} \\ &\quad \cdot \Big[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) - T^{-1}N^a(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\Big]D_1, \\ F_2 &= \big[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\big]^{-1} \\ &\quad \cdot (T^{-1/2}N^{a/2}\boldsymbol{H} - \boldsymbol{H}_{20} - T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V} + \boldsymbol{H}_{10})^T \cdot T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v, \\ F_3 &= \big[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\big]^{-1}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T \cdot T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v.\end{aligned}$$

Both $\boldsymbol{H}_{20}$ and $\boldsymbol{H}_{10}$ are $N^2 \times M(p+1)$ matrices defined in Theorem 3. By Assumptions R3 and R4, it is

easy to show that

$$\begin{aligned}\sigma_M(\boldsymbol{H}_{20}) &\geq \sigma_K(\boldsymbol{A}_1^0)\sigma_K(\boldsymbol{A}_2^0)\sigma_K(\boldsymbol{A}_3^0)\sigma_{\min}(\boldsymbol{A}_{410}^0,\ldots,\boldsymbol{A}_{4M0}^0,\ldots,\boldsymbol{A}_{41p}^0,\ldots,\boldsymbol{A}_{4Mp}^0)\\ &\geq \frac{CN^{1+a}\cdot N\cdot N}{\lambda_{\max}(\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))} \geq CN^{1+a},\\ \sigma_M^2(\boldsymbol{H}_{10}) &\geq \sigma_M^2(\boldsymbol{V}_0)\sigma_N^2\Big((\boldsymbol{I}_N\otimes\boldsymbol{\gamma}_T)\sum_{q=0}^{\infty}\mathbb{E}(\mathrm{vec}(\boldsymbol{B}_t^T)\mathrm{vec}(\boldsymbol{X}_t^T)^T)(\boldsymbol{I}_N\otimes\boldsymbol{\beta})\widetilde{\boldsymbol{\Pi}}_q\Big) \geq CN^{1+a}.\end{aligned}$$

Hence the smallest singular value of $\boldsymbol{H}_{20}$ dominates that of $\boldsymbol{H}_{10}$, and so for some constant $u>0$,

$$\sigma_{M(p+1)}^2(\boldsymbol{H}_{20}-\boldsymbol{H}_{10}) \geq uN^{1+a}. \tag{7.20}$$

With this, we have

$$\|[(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})]^{-1}\|_1 \leq \frac{M^{1/2}(p+1)^{1/2}}{\lambda_{\min}[(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})]} \leq \frac{M^{1/2}(p+1)^{1/2}}{uN^{1+a}} \tag{7.21}$$

To bound $\|D_1\|_1$, using (7.21), we have

$$\begin{aligned}\|F_1\|_1 \leq{}& \frac{M^{3/2}(p+1)^{1/2}}{N^{1+a}u}\Big[\|\boldsymbol{H}_{20}-\boldsymbol{H}_{10}\|_1\Big(\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_{\max}+\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_{\max}\Big)\\ &+\|T^{-1/2}N^{a/2}(\boldsymbol{H}-\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\|_{\max}\\ &\cdot\Big(\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_1+\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_1\Big)\Big]\|D_1\|_1, \qquad (7.22)\\ \|F_2\|_1 \leq{}& \frac{M^{3/2}(p+1)^{1/2}}{N^{1+a}\mu}\Big(\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_{\max}+\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_{\max}\Big)\\ &\cdot\|T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v\|_1, \qquad (7.23)\\ \|F_3\|_1 \leq{}& \frac{M^{3/2}(p+1)^{1/2}}{N^{1+a}\mu}\|\boldsymbol{H}_{20}-\boldsymbol{H}_{10}\|_{\max}\cdot\|T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v\|_1. \qquad (7.24)\end{aligned}$$

Now, to bound $\|F_1\|_1$, $\|F_2\|_1$ and $\|F_3\|_1$, we consider

$$\begin{aligned}&\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_{\max} = \max_{1\leq i\leq M}\max_{1\leq j\leq p}\|\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3(\boldsymbol{A}_{4ij}+\boldsymbol{A}_{5ij})-\boldsymbol{A}_1^0\boldsymbol{A}_2^0\boldsymbol{A}_3^0\boldsymbol{A}_{4ij}^0\|_{\max}\\ &\leq \max_{1\leq i\leq M}\max_{1\leq j\leq p}\|\boldsymbol{A}_1\|_{\max}\|\boldsymbol{A}_2\|_1\|\boldsymbol{A}_3\|_1\|\boldsymbol{A}_{5ij}\|_1+\\ &\quad\max_{1\leq i\leq M}\max_{1\leq j\leq p}\Big[\|\boldsymbol{A}_1\|_{\max}\|\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{A}_{4ij}-\boldsymbol{A}_2^0\boldsymbol{A}_3^0\boldsymbol{A}_{4ij}^0\|_1+\|\boldsymbol{A}_1-\boldsymbol{A}_1^0\|_{\max}\|\boldsymbol{A}_2^0\boldsymbol{A}_3^0\boldsymbol{A}_{4ij}^0\|_1\Big], \text{ with} \qquad (7.25)\\ &\|\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{A}_{4ij}-\boldsymbol{A}_2^0\boldsymbol{A}_3^0\boldsymbol{A}_{4ij}^0\|_1 \leq \max_{1\leq j\leq p}\Big[\|\boldsymbol{A}_2\|_1\|\boldsymbol{A}_3-\boldsymbol{A}_3^0\|_1\|\boldsymbol{A}_{4ij}\|_1+\\ &\qquad\qquad\|\boldsymbol{A}_2\|_1\|\boldsymbol{A}_3^0\|_1\|\boldsymbol{A}_{4ij}-\boldsymbol{A}_{4ij}^0\|_1+\|\boldsymbol{A}_2-\boldsymbol{A}_2^0\|_1\|\boldsymbol{A}_3^0\|_1\|\boldsymbol{A}_{4ij}^0\|_1\Big].\end{aligned}$$

Therefore, base on the rates found in (7.12) to (7.19), and (7.25), we have on $\mathcal{M}$ that

$$\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_{\max}=O(\lambda_T), \quad \text{and} \quad \|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_1=O(\lambda_T N^2). \tag{7.26}$$

Define $\mathbf{L}^q = T^{-1}\sum_{t=1}^T \mathrm{vec}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T)\mathrm{vec}^T(\boldsymbol{X}_{t-q})$ and $\mathbf{L}_0^q = \mathbb{E}(\mathrm{vec}(\boldsymbol{B}_t^T)\mathrm{vec}^T(\boldsymbol{X}_{t-q}^T)$. Then, by Assumption

R4 and on $\mathcal{M}$, we have

$$\begin{aligned}\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\|_1 &= \max_{0\le l\le p}\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\gamma}y_{t-l}\|_1 = \|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\gamma}y_t^T\|_1\\ &\le \|T^{-1}\sum_{t=1}^{T}\sum_{q=0}^{\infty}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\gamma}(\boldsymbol{\Psi}_q\boldsymbol{\Pi}(\boldsymbol{X}_{t-q}\boldsymbol{\beta}+\boldsymbol{\epsilon}_{t-q}))^T\|_1\\ &\le O(\lambda_T N+N^a+\lambda_T N)=O(N^a), \quad \text{and} \end{aligned} \tag{7.27}$$

$$\begin{aligned}\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_{\max} &= \max_{0\le l\le p}\max_{1\le i\le M,1\le j\le N}\|\Big(T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\gamma}y_{t-l}^T-\mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\gamma}y_{t-l}^T)\Big)\boldsymbol{W}_{0i,j}^c\|_{\max}\\ &= \max_{1\le i\le M,1\le j\le N}\|\Big(T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\gamma}y_t^T-\mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\gamma}y_t^T)\Big)\boldsymbol{W}_{0i,j}^c\|_{\max}\\ &\le \max_{1\le i\le M,1\le j\le N}\Big[\|\boldsymbol{I}_N\otimes\boldsymbol{\gamma}^T\|_\infty\sum_{q\ge0}\|\mathbf{L}^q-\mathbf{L}_0^q\|_{\max}\|\widetilde{\boldsymbol{\Pi}}_q^T\|_1\|\boldsymbol{I}_N\otimes\boldsymbol{\beta}\|_1\|\boldsymbol{W}_{0i,j}^c\|_1\\ &+\|\boldsymbol{I}_N\otimes\boldsymbol{\gamma}^T\|_\infty\sum_{q\ge0}\|T^{-1}\sum_{t=1}^{T}\mathrm{vec}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T)\boldsymbol{\epsilon}_{t-q}^T\|_{\max}\|\widetilde{\boldsymbol{\Pi}}_q^T\|_1\|\boldsymbol{W}_{0i,j}^c\|_1\Big]\\ &= O(\lambda_T). \end{aligned} \tag{7.28}$$

Hence on $\mathcal{M}$,

$$\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_1 = O(\lambda_T N^2). \tag{7.29}$$

Using the rates found in (7.12) to (7.19), on $\mathcal{M}$ (in particular using the rate on $\mathcal{A}_3$),

$$\|T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v\|_1 = \Big\|\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3(T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\epsilon_t)\Big\|_1 = O(\lambda_T N^{1/2+1/2w+a}). \tag{7.30}$$

Therefore, using results from (7.21) to (7.30), we know that

$$\begin{aligned}\|D_1\|_1 &\le \frac{M^{3/2}}{N^{1+a}u}\Big(o(\lambda_T N^2)+o(1)o(\lambda_T N^2+\lambda_T N^2)\Big)\|D_1\|_1\\ &+\frac{M^{3/2}}{N^{1+a}u}\Big(O(\lambda_T)O(\lambda_T N^{1/2+1/2w+a})\Big)+\frac{M^{3/2}}{N^{1+a}u}O(\lambda_T N^{1/2+1/2w+a})\\ &= O(\lambda_T N^{-1/2+1/2w}).\end{aligned}$$

For the rate of $\big\|D_2\big\|_1$, we refer to the proof of asymptotic normality of $\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}$ in Theorem 3 for the proof of the asymptotic normality of D_2 (along the exact same lines of proofs as in Theorem 3). Therefore, we state here the result that

$$T^{1/2}(\boldsymbol{M}_2\mathbf{S}_2\boldsymbol{M}_2^T)^{-1/2}D_2 \xrightarrow{\mathcal{D}} N(\mathbf{0},\boldsymbol{I}_m),$$

where $\mathbf{S}_2$ is defined in Theorem 3, and $\boldsymbol{M}_2 = \left[(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})\right]^{-1}(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})^T$. By As-

sumption R5, we conclude that all the eigenvalues of $\mathbf{S}_2$ are of order N^b. Hence by (7.20),

$$\begin{aligned}\lambda_{\max}(\boldsymbol{M}_2\mathbf{S}_2\boldsymbol{M}_2^T) &\leq \lambda_{\max}(\mathbf{S}_2)\lambda_{\max}(\big[(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})\big]^{-1}) \\ &\leq \frac{\lambda_{\max}(\mathbf{S}_2)}{\sigma^2_{M(p+1)}(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})} = O(N^{-1-a+b}),\end{aligned}$$

which can also be derived as the order for the lower bound of $\lambda_{\min}(\boldsymbol{M}_2\mathbf{S}_2\boldsymbol{M}_2^T)$. Hence we have $\|D_2\|_1 = O_p(T^{-1/2}N^{-(1+a-b)/2})$.

Finally, by Assumption R7 and the result of Theorem 5,

$$\begin{aligned}\|\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}\|_1 &= O_P(\big\|D_1\big\|_1 + \big\|D_2\big\|_1) = O_P(\lambda_T\cdot N^{1/2+a+1/2w}) + O_P(T^{-1/2}N^{-(1+a-b)/2}) \\ &= O_P(\lambda_T\cdot N^{-1/2+1/2w}).\end{aligned}$$

At the same time, using the result above,

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 = O_p(\lambda_T N^{-1/2+1/2w} + \|\boldsymbol{\delta}-\widehat{\boldsymbol{\delta}}\|_1) = O_p(\lambda_T N^{-1/2+1/2w}). \quad \square$$

Proof of Theorem 2.

It has been shown that $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta} = \sum_{i=0}^5 I_i$ in the proof of Theorem 1. From the rate of $\|\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}\|_1$ and Assumption R7, it is clear that

$$\|\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}\|_1 = O_P(\lambda_T N^{-1/2+1/2w}) = o_P(T^{-1/2}N^{-(1-b)/2}).$$

Therefore, if we can prove that I_1 is $T^{1/2}N^{(1-b)/2}$-convergent, then I_1 dominates I_2 to I_5, while $\big\|I_0\big\|_1 = O_P(\lambda_T\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big\|_1) = o_P(\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big\|_1)$.

We now prove that for $\boldsymbol{\alpha}\in\mathbb{R}^K$ such that $\big\|\boldsymbol{\alpha}\big\| = 1$, $\boldsymbol{\alpha}^T I_1$ is $T^{1/2}N^{(1-b)/2}$-convergent by proving its asymptotic normality. Recall that

$$\begin{aligned}I_1 &= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v \\ &= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}(T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v - \mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t))T^{-1}\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v \\ &\quad + (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)T^{-1}\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v.\end{aligned}$$

It is easy to show that the second term above dominates the first. Therefore, if we can prove that

$$\sum_{t\geq 0}\|\boldsymbol{P}_0(\boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t)\| < \infty, \tag{7.31}$$

where $\boldsymbol{M}_1 = (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)$, by Theorem 3(ii) of Wu (2011), we then have

$$T^{1/2}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{\alpha})^{-1/2}\boldsymbol{\alpha}^T I_1 \xrightarrow{\mathcal{D}} \boldsymbol{N}(0,1),$$

where $\boldsymbol{\Sigma}_1 = \boldsymbol{M}_1\sum_{\tau\in\mathbb{Z}}\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T\boldsymbol{B}_{t+\tau})\boldsymbol{M}_1^T$.

To determine the rate of the eigenvalues in $\boldsymbol{\Sigma}_1$, consider the (k,k) element of $\sum_\tau \mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_{t+\tau}^T \boldsymbol{B}_{t+\tau})$,

$$\begin{aligned}
&\sum_\tau \mathbb{E}(\boldsymbol{B}_{t,k}^T \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_{t+\tau}^T \boldsymbol{B}_{t+\tau,k}) = \sum_\tau \mathrm{tr}(\mathbb{E}(\boldsymbol{B}_{t+\tau,k}\boldsymbol{B}_{t,k}^T)\mathbb{E}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t^T)) \\
&= \sum_\tau \mathrm{tr}(\mathrm{cov}(\boldsymbol{B}_{t+\tau,k}\boldsymbol{B}_{t,k})\mathrm{cov}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t)) + \sum_\tau \mathrm{tr}(\boldsymbol{\mu}_{b,k}\boldsymbol{\mu}_{b,k}^T \mathrm{cov}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t)).
\end{aligned}$$

By Assumptions R5, the first term is N^{1+b}-convergent exactly and the second term's rate is

$$\sum_\tau \boldsymbol{\mu}_{b,k}^T \mathrm{cov}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t)\boldsymbol{\mu}_{b,k} \le \lambda_{\max}(\sum_\tau \mathrm{cov}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t))\|\boldsymbol{\mu}_{b,k}\|^2 = O(\|\boldsymbol{\mu}_{b,k}\|^2) = O(N).$$

Since K is finite, the order of the eigenvalues of $\sum_\tau \mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_{t+\tau}^T \boldsymbol{B}_{t+\tau})$ is exactly N^{1+b}. Also, for $i = 1,\ldots,K$,

$$\begin{aligned}
&\lambda_{\min}(\boldsymbol{M}_1\boldsymbol{M}_1^T)\lambda_{\min}(\sum_\tau \mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_{t+\tau}^T \boldsymbol{B}_{t+\tau})) \\
&\le \lambda_i(\boldsymbol{\Sigma}_1) \le \lambda_{\max}(\boldsymbol{M}_1\boldsymbol{M}_1^T)\lambda_{\max}(\sum_\tau \mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_{t+\tau}^T \boldsymbol{B}_{t+\tau})).
\end{aligned}$$

Since the order of the eigenvalues of $\boldsymbol{M}_1\boldsymbol{M}_1^T$ is N^{-2}, the order of all the eigenvalues of $\boldsymbol{\Sigma}_1$ is exactly N^{-1+b}. It means also that $\boldsymbol{\alpha}^T I_1$ is indeed $T^{1/2}N^{(1-b)/2}$-convergent, and so I_1 is $T^{1/2}N^{(1-b)/2}$-convergent in particular since K is finite. With the asymptotic normality for $\boldsymbol{\alpha}^T I_1$, we can then use the multivariate version of Theorem 3(ii) of Wu (2011) to conclude that

$$T^{1/2}\boldsymbol{\Sigma}_1^{-1/2} I_1 \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_K),$$

where we replaced $\boldsymbol{\alpha}$ by I_K.

It remains to prove (7.31). We decompose

$$\boldsymbol{P}_0(\boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t) = \boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{P}_0(\boldsymbol{B}_t^T)\mathbb{E}_0(\boldsymbol{\epsilon}_t) + \boldsymbol{\alpha}^T\boldsymbol{M}_1\mathbb{E}_{-1}(\boldsymbol{B}_t^T)\boldsymbol{P}_0(\boldsymbol{\epsilon}_t),$$

so that we have $\|\boldsymbol{P}_0(\boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t)\| \le C_{1,t} + C_{2,t}$, where

$$\begin{aligned}
C_{1,t}^2 &= \mathbb{E}(\boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{P}_0(\boldsymbol{B}_t^T)\mathbb{E}_0(\boldsymbol{\epsilon}_t)\mathbb{E}_0(\boldsymbol{\epsilon}_t^T)\boldsymbol{P}_0(\boldsymbol{B}_t)\boldsymbol{M}_1^T\boldsymbol{\alpha}) \\
&\le \boldsymbol{\alpha}^T\boldsymbol{M}_1\mathbb{E}(\boldsymbol{P}_0(\boldsymbol{B}_t^T)\boldsymbol{P}_0(\boldsymbol{B}_t))\boldsymbol{M}_1\boldsymbol{\alpha}\mathbb{E}(\lambda_{\max}(\mathbb{E}_0(\boldsymbol{\epsilon}_t)\mathbb{E}_0(\boldsymbol{\epsilon}_t^T))) \\
&\le \left\|\boldsymbol{\alpha}^T\boldsymbol{M}_1\right\|^2 \lambda_{\max}(\mathbb{E}(\boldsymbol{P}_0(\boldsymbol{B}_t^T)\boldsymbol{P}_0(\boldsymbol{B}_t)))\mathbb{E}(\mathbb{E}_0(\boldsymbol{\epsilon}_t^T)\mathbb{E}_0(\boldsymbol{\epsilon}_t)) \\
&= O(N^{-1}\max_{1\le k\le K}\mathbb{E}(\boldsymbol{P}_0(\boldsymbol{B}_{t,k}^T)\boldsymbol{P}_0(\boldsymbol{B}_{t,k}))\mathbb{E}(N^{-1}\mathbb{E}_0(\boldsymbol{\epsilon}_t^T)\mathbb{E}_0(\boldsymbol{\epsilon}_t))) \\
&= O(\max_{1\le k\le K}\max_{1\le s\le N}\|\boldsymbol{P}_0(B_{t,sk})\|^2 \max_{1\le j\le N}\mathbb{E}(\mathbb{E}_0^2(\epsilon_{t,j}))) \\
&= O(\max_{1\le k\le K}\max_{1\le s\le N}\|\boldsymbol{P}_0(B_{t,sk})\|^2\sigma_{\max}^2), \qquad (7.32)
\end{aligned}$$

so that $\sum_{t\ge 0} C_{1,t} < \infty$ by our assumption $\sum_{t\ge 0}\max_{1\le k\le K}\max_{1\le s\le N}\|\boldsymbol{P}_0^b(B_{t,sk})\| < \infty$.

Similarly, we have

$$\begin{aligned}
C_{2,t}^2 &= \mathbb{E}(\boldsymbol{\alpha}^T \boldsymbol{M}_1 \mathbb{E}_{-1}(\boldsymbol{B}_t^T) \boldsymbol{P}_0(\boldsymbol{\epsilon}_t) \boldsymbol{P}_0(\boldsymbol{\epsilon}_t^T) \mathbb{E}_{-1}(\boldsymbol{B}_t) \boldsymbol{M}_1^T \boldsymbol{\alpha}) \\
&\leq \boldsymbol{\alpha}^T \boldsymbol{M}_1 \mathbb{E}(\mathbb{E}_{-1}(\boldsymbol{B}_t^T) \mathbb{E}_{-1}(\boldsymbol{B}_t)) \boldsymbol{M}_1^T \boldsymbol{\alpha} \mathbb{E}(\lambda_{\max}(\boldsymbol{P}_0(\boldsymbol{\epsilon}_t) \boldsymbol{P}_0(\boldsymbol{\epsilon}_t^T))) \\
&\leq \left\| \boldsymbol{\alpha}^T \boldsymbol{M}_1 \right\|^2 \lambda_{\max}(\mathbb{E}(\mathbb{E}_{-1}(\boldsymbol{B}_t^T) \mathbb{E}_{-1}(\boldsymbol{B}_t))) \mathbb{E}(\boldsymbol{P}_0(\boldsymbol{\epsilon}_t^T) \boldsymbol{P}_0(\boldsymbol{\epsilon}_t)) \\
&= O(\max_{1 \leq k \leq K} \max_{1 \leq s \leq N} \mathbb{E}(\mathbb{E}_{-1}^2(B_{t,sk})) \max_{1 \leq j \leq N} \|\boldsymbol{P}_0(\epsilon_{t,j})\|^2) \\
&= O((\sigma_{\max}^2 + \max_{s,k} \mu_{b,sk}^2) \max_{1 \leq j \leq N} \|\boldsymbol{P}_0(\epsilon_{t,j})\|^2) \\
&= O(\max_{1 \leq j \leq N} \|\boldsymbol{P}_0^{\epsilon}(\epsilon_{t,j})\|^2), \qquad (7.33)
\end{aligned}$$

so that $\sum_{t \geq 0} C_{2,t} < \infty$ by our assumption of $\sum_{t \geq 0} \max_{1 \leq j \leq N} \|\boldsymbol{P}_0^{\epsilon}(\epsilon_{t,j})\| < \infty$. Hence (7.31) is established, and the proof of the theorem is completed. □

Proof of Theorem 3.

To prove the asymptotic normality of $\widehat{\boldsymbol{\delta}}$, we need to apply the same method we used for the proof of Theorem 2. Recall that from the proof of Theorem 1,

$$\begin{aligned}
\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} &= \left[(\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}) \right]^{-1} (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T \left[\boldsymbol{K} \boldsymbol{\epsilon}^v - \boldsymbol{B}^T \boldsymbol{\epsilon} \right], \\
\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} &= D_1 + D_2, \quad \text{where} \\
D_1 &= \left[(\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}) \right]^{-1} (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T \boldsymbol{K} \boldsymbol{\epsilon}^v, \\
D_2 &= - \left[(\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}) \right]^{-1} (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T \boldsymbol{B}^T \boldsymbol{\epsilon}.
\end{aligned}$$

Moreover, we further decompose D_1 as in the proof of Theorem 1 such that $D_1 = F_1 + F_2 + F_3$, where

$$\begin{aligned}
F_1 &= \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) \right]^{-1} \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) - T^{-1} N^a (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}) \right] D_1, \\
F_2 &= \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) \right]^{-1} (T^{-1/2} N^{a/2} \boldsymbol{H} - \boldsymbol{H}_{20} - T^{-1/2} N^{a/2} \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V} + \boldsymbol{H}_{10})^T \cdot T^{-1/2} N^{a/2} \boldsymbol{K} \boldsymbol{\epsilon}^v, \\
F_3 &= \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) \right]^{-1} (\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T \cdot T^{-1/2} N^{a/2} \boldsymbol{K} \boldsymbol{\epsilon}^v. \qquad (7.34)
\end{aligned}$$

From the proof of Theorem 1, it is clear that F_3 dominates all other terms in the decomposition of D_1. As for D_2, we can apply similar decomposition, and the term

$$F_4 = - \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) \right]^{-1} (\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T T^{-1/2} N^{a/2} \boldsymbol{B}^T \boldsymbol{\epsilon} \qquad (7.35)$$

dominates in the decomposition of D_2. Hence to show the asymptotic normality of $\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}$, we only consider

$$\begin{aligned}
F_3 + F_4 &= T^{-1/2} N^{a/2} \boldsymbol{M}_2 (\boldsymbol{K} \boldsymbol{\epsilon}^v - \boldsymbol{B}^T \boldsymbol{\epsilon}) \\
&= T^{-1} \sum_{t=1}^{T} \boldsymbol{M}_2 (\boldsymbol{M} \boldsymbol{B}_t^T \boldsymbol{\epsilon}_t - \mathrm{vec}(\boldsymbol{B}_t \boldsymbol{\gamma} \boldsymbol{\epsilon}_t^T)) (1 + o_P(1)),
\end{aligned}$$

where $\boldsymbol{M}_2 = \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) \right]^{-1} (\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T$. In view of the above and Theorem 3(ii) of

Wu (2011), to prove the asymptotic normality of $\boldsymbol{\alpha}^T(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})$ where $\boldsymbol{\alpha} \in \mathbb{R}^{M(p+1)}$, if we can show that

$$\sum_{t \geq 0} \left\| \boldsymbol{P}_0(\boldsymbol{\alpha}^T \boldsymbol{M}_2(\boldsymbol{M}\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t - \mathrm{vec}(\boldsymbol{B}_t \boldsymbol{\gamma} \boldsymbol{\epsilon}_t^T))) \right\| < \infty, \tag{7.36}$$

then we can conclude by Theorem 3(ii) of Wu (2011) that

$$T^{1/2}(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_2 \boldsymbol{\alpha})^{-1/2} \boldsymbol{\alpha}^T (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \xrightarrow{\mathcal{D}} \boldsymbol{N}(0, 1), \tag{7.37}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_2 &= \sum_{\tau \in \mathbb{Z}} \boldsymbol{M}_2 \mathrm{cov}(\boldsymbol{M}\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t - \mathrm{vec}(\boldsymbol{B}_t \boldsymbol{\gamma} \boldsymbol{\epsilon}_t^T), \boldsymbol{M}\boldsymbol{B}_{t+\tau}^T \boldsymbol{\epsilon}_{t+\tau} - \mathrm{vec}(\boldsymbol{B}_{t+\tau} \boldsymbol{\gamma} \boldsymbol{\epsilon}_{t+\tau}^T)) \boldsymbol{M}_2^T \\ &= \boldsymbol{M}_2(\mathbf{S}_1 + \mathbf{S}_2 - \mathbf{S}_3 - \mathbf{S}_3^T) \boldsymbol{M}_2^T, \end{aligned}$$

with $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ as defined in the statement of the theorem. A generalization to Theorem 3(ii) of Wu (2011) then gives us the asymptotic normality result after replacing $\boldsymbol{\alpha}$ by $\boldsymbol{I}_{M(p+1)}$.

It remains to show (7.36). Consider

$$\begin{aligned} \left\| \boldsymbol{\alpha}^T \boldsymbol{M}_2 \boldsymbol{M} \right\|^2 &\leq \lambda_{\max}(\boldsymbol{M}_2 \boldsymbol{M}_2^T) \lambda_{\max}(\boldsymbol{M}\boldsymbol{M}^T) = O(N^{-1-a}) \cdot O(N^{-2}) \cdot \lambda_{\max}(\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t \boldsymbol{\gamma}) \mathbb{E}(\boldsymbol{X}_t^T \otimes \boldsymbol{\gamma}^T \boldsymbol{B}_t^T)) \\ &= O(N^{-3-a}) \cdot \left\| \mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t \boldsymbol{\gamma}) \right\|_1 \left\| \mathbb{E}(\boldsymbol{X}_t^T \otimes \boldsymbol{\gamma}^T \boldsymbol{B}_t^T) \right\|_1 \\ &= O(N^{-3-a}) \cdot O(N^{1+a}) \cdot O(1) = O(N^{-2}), \end{aligned}$$

where the last line follows from Assumption R4. Then similar to showing (7.32), by the above, we have

$$\left\| \boldsymbol{P}_0(\boldsymbol{\alpha}^T \boldsymbol{M}_2 \boldsymbol{M}\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t) \right\| = O(\max_{1 \leq k \leq K} \max_{1 \leq s \leq N} \left\| \boldsymbol{P}_0(B_{t,sk}) \right\|), \tag{7.38}$$

so that $\sum_{t \geq 0} \left\| P_0(\boldsymbol{\alpha}^T \boldsymbol{M}_2 \boldsymbol{M}\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t) \right\| < \infty$ by the assumptions of the theorem. At the same time,

$$\begin{aligned} \boldsymbol{P}_0(\boldsymbol{\alpha}^T \boldsymbol{M}_2 \mathrm{vec}(\boldsymbol{B}_t \boldsymbol{\gamma} \boldsymbol{\epsilon}_t^T)) &= \boldsymbol{P}_0(\boldsymbol{\alpha}^T \boldsymbol{M}_2(\boldsymbol{\epsilon}_t \otimes \boldsymbol{B}_t \boldsymbol{\gamma})) \\ &= \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M}_2 \Big(\mathbb{E}_0(\boldsymbol{\epsilon}_t) \otimes \mathbb{E}_0(\boldsymbol{B}_t \boldsymbol{\gamma}) - \mathbb{E}_{-1}(\boldsymbol{\epsilon}_t) \otimes \mathbb{E}_{-1}(\boldsymbol{B}_t \boldsymbol{\gamma}) \Big) \\ &= \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M}_2 \boldsymbol{P}_0(\boldsymbol{\epsilon}_t) \otimes \mathbb{E}_0(\boldsymbol{B}_t \gamma) + \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M}_2 \mathbb{E}_{-1}(\boldsymbol{\epsilon}_t) \otimes \boldsymbol{P}_0(\boldsymbol{B}_t \gamma). \end{aligned}$$

Hence denote by $\boldsymbol{b}_{t,j}^T$ the jth row of $\boldsymbol{B}_t$,

$$
\begin{aligned}
&\big\|\boldsymbol{P}_0(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}_2\boldsymbol{\epsilon}_t \otimes \boldsymbol{B}_t\boldsymbol{\gamma})\big\| \\
&\le \Big\{2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}_2\mathbb{E}(\boldsymbol{P}_0(\boldsymbol{\epsilon}_t)\boldsymbol{P}_0(\boldsymbol{\epsilon}_t)^{\mathrm{T}}) \otimes \mathbb{E}\big(\mathbb{E}_0(\boldsymbol{B}_t\boldsymbol{\gamma})\mathbb{E}_0(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{B}_t^{\mathrm{T}})\big)\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{\alpha}\Big\}^{1/2} \\
&\quad + \Big\{2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}_2\mathbb{E}(\mathbb{E}_{-1}(\boldsymbol{\epsilon}_t)\mathbb{E}_{-1}(\boldsymbol{\epsilon}_t)^{\mathrm{T}}) \otimes \mathbb{E}\big(\boldsymbol{P}_0(\boldsymbol{B}_t\boldsymbol{\gamma})\boldsymbol{P}_0(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{B}_t^{\mathrm{T}})\big)\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{\alpha}\Big\}^{1/2} \\
&\le 2^{1/2}\big\|\boldsymbol{\alpha}\big\|_1\big\|\boldsymbol{M}_2\big\|_\infty \max_{1\le j\le N}\big\|\boldsymbol{P}_0(\epsilon_{tj})\big\| \cdot \max_{1\le j\le N}\mathrm{var}^{1/2}(\boldsymbol{b}_{t,j}^{\mathrm{T}}\boldsymbol{\gamma}) \\
&\quad + 2^{1/2}\big\|\boldsymbol{\alpha}\big\|_1\big\|\boldsymbol{M}_2\big\|_\infty \cdot \sigma_{\max} \cdot \max_{1\le j\le N}\big\|\boldsymbol{P}_0(\boldsymbol{b}_{t,j}^{\mathrm{T}}\boldsymbol{\gamma})\big\| \\
&\le 2^{1/2}\big\|\boldsymbol{\alpha}\big\|_1\big\|\boldsymbol{M}_2\big\|_\infty \max_{1\le j\le N}\big\|\boldsymbol{P}_0(\epsilon_{tj})\big\| \cdot \sigma_{\max}\big\|\boldsymbol{\gamma}\big\|_1 \\
&\quad + 2^{1/2}\big\|\boldsymbol{\alpha}\big\|_1\big\|\boldsymbol{M}_2\big\|_\infty \cdot \sigma_{\max} \cdot \max_{\substack{1\le j\le N\\ 1\le k\le K}}\big\|\boldsymbol{P}_0(B_{t,jk})\big\|\big\|\boldsymbol{\gamma}\big\|_1 \\
&= O(\max_{1\le j\le N}\big\|\boldsymbol{P}_0(\epsilon_{tj})\big\| + \max_{\substack{1\le j\le N\\ 1\le k\le K}}\big\|\boldsymbol{P}_0(B_{t,jk})\big\|),
\end{aligned}
$$

where the second inequality used the decomposition

$$\mathrm{var}(\cdot) = \mathrm{var}(\mathbb{E}_i(\cdot)) + \mathbb{E}(\mathrm{var}_i(\cdot)) \ge \mathrm{var}(\mathbb{E}_i(\cdot)),$$

and the third inequality used Assumption R2, while the last equality used $\big\|\boldsymbol{\gamma}\big\|_1 = 1$ and $\big\|\boldsymbol{M}_2\big\|_\infty = O(1)$. Hence $\sum_{t\ge 0}\big\|\boldsymbol{P}_0(\boldsymbol{\alpha}^T\boldsymbol{M}_2\mathrm{vec}(\boldsymbol{B}_t\boldsymbol{\gamma}\boldsymbol{\epsilon}_t^T)\big\| < \infty$, and together with (7.38), (7.36) is established. This completes the proof of the theorem. □

Proof of Theorem 4.

By the KKT condition, there exists a solution $\widetilde{\boldsymbol{\delta}}$ to (2.7) if and only if there exists a subgradient

$$\boldsymbol{h} = \partial(\boldsymbol{u}^{\mathrm{T}}|\widetilde{\boldsymbol{\delta}}|) = \Big\{\boldsymbol{h} \in \mathbb{R}^{M(p+1)} : \begin{cases} h_i = u_i\mathrm{sign}(\widetilde{\delta}_i), & \widetilde{\delta}_i \ne 0; \\ |h_i| \le u_i, & \text{otherwise.} \end{cases}\Big\},$$

such that differentiating the expression on the right hand side of (2.7) with respect to $\boldsymbol{\delta}$, we get

$$T^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\widetilde{\boldsymbol{\delta}} - T^{-1}(\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V} - \boldsymbol{H})^T(\boldsymbol{B}^T y - \boldsymbol{g}) = -\gamma_T\boldsymbol{h}.$$

We use a single index $i = 1, \ldots, M(p+1)$ to denote an element of $\boldsymbol{\delta}$ for easier notation in this proof. Since we have $\boldsymbol{B}^T y = \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}\boldsymbol{\delta} + \boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}}\mathrm{vec}(\boldsymbol{I}_N) + \boldsymbol{B}^T\boldsymbol{\epsilon}$, the above equation can be rewritten as

$$
\begin{aligned}
&T^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})(\widetilde{\boldsymbol{\delta}} - \boldsymbol{\delta}) + T^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}}\mathrm{vec}(\boldsymbol{I}_N) + \boldsymbol{H}\boldsymbol{\delta} - \boldsymbol{g}) \\
&+ T^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T\boldsymbol{B}^T\boldsymbol{\epsilon} = -\gamma_T\boldsymbol{h}.
\end{aligned}
$$

We can show easily that $-\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}(\boldsymbol{\delta})}\mathrm{vec}(\boldsymbol{I}_N) = \boldsymbol{H}\boldsymbol{\delta} - \boldsymbol{g}$, and hence there exists a sign consistent solution

$\widetilde{\boldsymbol{\delta}}$ if and only if

$$\begin{cases} T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)(\widetilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H) + T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T(\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}-\boldsymbol{\beta}(\boldsymbol{\delta})}\mathrm{vec}(\boldsymbol{I}_N)) \\ \qquad +T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\boldsymbol{B}^T\boldsymbol{\epsilon} = -\gamma_T\boldsymbol{h}_H, \\ |T^{-1}(\boldsymbol{H}_{H^c} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_{H^c})^T\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}-\boldsymbol{\beta}(\boldsymbol{\delta})}\mathrm{vec}(\boldsymbol{I}_N) + T^{-1}(\boldsymbol{H}_{H^c} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_{H^c})^T\boldsymbol{B}^T\boldsymbol{\epsilon}| \le -\gamma_T\boldsymbol{h}_{H^c}, \end{cases} \tag{7.39}$$

where $H = \{j : \delta_j \neq 0\}$.

From the first equation in (7.39), we decompose $\widetilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H = I_0 + I_1 + I_2 + I_3$, where

$$\begin{aligned} I_0 &= -(N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)^{-1}(T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H) \\ &\quad - N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)(\widetilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H), \\ I_1 &= (N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)^{-1}T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\boldsymbol{K}\boldsymbol{\epsilon}^v, \\ I_2 &= -(N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)^{-1}\gamma_T\boldsymbol{h}, \\ I_3 &= -(N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)^{-1}T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\boldsymbol{B}^T\boldsymbol{\epsilon}. \end{aligned}$$

The term I_1 has its form because of the identity $\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}(\boldsymbol{\delta})-\boldsymbol{\beta}}\mathrm{vec}(\boldsymbol{I}_N) = \boldsymbol{K}\boldsymbol{\epsilon}^v$. Similar to bounding $\big\|F_1\big\|_1$ to $\big\|F_3\big\|_1$ in (7.22) to (7.24) in the proof of Theorem 1, we can show that

$$\|I_0\|_{\max} = o_p(\lambda_T N^{1-a}\|\widetilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H\|_{\max}), \quad \|I_2\|_{\max} = O(\lambda_T N^{-1}).$$

We can show easily that

$$\begin{aligned} I_1 &= [(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H]^{-1}\Big(T^{-1/2}N^{a/2}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\Big)(T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v) \\ &= [(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H]^{-1}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v)(1 + o_P(1)), \\ I_3 &= -[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H]^{-1}\Big(T^{-1/2}N^{a/2}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\Big)(T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{\epsilon}) \\ &= -[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H]^{-1}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{\epsilon})(1 + o_P(1)). \end{aligned}$$

Hence I_1 is similar to F_3 in (7.34) and I_3 is similar to F_4 in (7.35) in the proof of Theorem 3, except that $\boldsymbol{H}_{20} - \boldsymbol{H}_{10}$ is now restricted to those columns with indices in H only. Using exactly the same lines of proof as in Theorem 3, we can conclude that

$$T^{1/2}\boldsymbol{\Sigma}_3^{-1/2}(I_1 + I_3) \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_{|H|}), \tag{7.40}$$

where $\boldsymbol{\Sigma}_3 = \boldsymbol{M}_3(\mathbf{S}_1 + \mathbf{S}_2 - \mathbf{S}_3 - \mathbf{S}_3^T)\boldsymbol{M}_3^T$, with $\boldsymbol{M}_3 = [(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H]^{-1}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T$. By Assumptions R4 and R5, we can show that then $I_1 + I_3$ is exactly $T^{1/2}N^{(1+a-b)/2}$-convergent. Since $0 < a, b < 1$, it is not difficult to see that I_2 is dominated by $I_1 + I_3$ then. Also, Assumption R7 ensures $\big\|I_0\big\|_{\max} = o_P(\big\|\widetilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\big\|_{\max})$. All these imply that

$$T^{1/2}\boldsymbol{\Sigma}_3^{-1/2}(\widetilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H) \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_{|H|}),$$

which is the asymptotic normality result we need, if we can also show that the second inequality in (7.39) is true.

From the above, since $\|I_0\|_{\max}, \|I_1\|_{\max}, \|I_2\|_{\max}$ and $\|I_3\|_{\max}$ are all $o_P(1)$, we have $\text{sign}(\widetilde{\boldsymbol{\delta}}_H) = \text{sign}(\boldsymbol{\delta}_H)$. It remains to show the second inequality in (7.39).

To this end, we can show from previous results that

$$\|T^{-1}(\boldsymbol{H}_{H_c} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_{H_c})^T\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}-\boldsymbol{\beta}(\boldsymbol{\delta})}\text{vec}(\boldsymbol{I}_N) + T^{-1}(\boldsymbol{H}_{H_c} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_{H_c})^T\boldsymbol{B}^T\boldsymbol{\epsilon}\|_{\max} = O_p(T^{-1/2}N^{(1+b-a)/2}),$$

while the right hand side of the second inequality has a minimum value of

$$\frac{\gamma_T}{\|\widetilde{\boldsymbol{\delta}}_{H_c}\|_{\max}} \geq \frac{\gamma_T}{\|\widetilde{\boldsymbol{\delta}}_{H_c} - \boldsymbol{\delta}_{H_c}\|_{\max}}.$$

Hence, it is sufficient to prove

$$(T^{-1/2}N^{(1+b-a)/2})(\|\widetilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_{\max}) = o_p(\lambda_T).$$

But the left hand side above has rate $T^{-1}N^{b-a} = o(\lambda_T)$ by Assumption R7. This completes the proof of the theorem. □