Journal of Multivariate Analysis 171 (2019) 270–283

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Penalized generalized empirical likelihood in high-dimensional weakly dependent data

Jia Zhang [a], Haoming Shi [b], Lemeng Tian [c,*], Fengjun Xiao [d]

[a] School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China
[b] School of Finance, Southwestern University of Finance and Economics, Chengdu 611130, China
[c] Guosheng Securities, Shanghai 200122, China
[d] School of Humanities and Social Sciences, Beihang University, Beijing 100083, China

ARTICLE INFO

Article history:
Received 27 December 2018
Available online 4 January 2019

AMS 2010 subject classifications:
primary 62G20
secondary 62H86

Keywords:
High-dimensional data analysis
Penalized likelihood
Smoothed generalized empirical likelihood
Smoothed moment functions
Weak dependence

ABSTRACT

In this paper, we propose a penalized generalized empirical likelihood (PGEL) approach based on the smoothed moment functions Anatolyev (2005), Smith (1997), Smith (2004) for parameters estimation and variable selection in the growing (high) dimensional weakly dependent time series setting. The dimensions of the parameters and moment restrictions are both allowed to grow with the sample size at some moderate rates. The asymptotic properties of the estimators of the smoothed generalized empirical likelihood (SGEL) and its penalized version (SPGEL) are then obtained by properly restricting the degree of data dependence. It is shown that the SPGEL estimator maintains the oracle property despite the existence of data dependence and growing (high) dimensionality. We finally present simulation results and a real data analysis to illustrate the finite-sample performance and applicability of our proposed method.

© 2019 Elsevier Inc. All rights reserved.

1. Introduction

In order to better capture large-scale economic dynamics or financial relations, models with a growing number of unknown parameters of interest are increasingly employed to analyze high-dimensional time series data. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be random vectors from an $\mathbb{R}^d$-valued stationary time series and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ be a vector of unknown parameters taking values in a parameter space Θ. Suppose that the data information is summarized by the moment restrictions

$$\mathrm{E}\{\mathbf{g}(\boldsymbol{X}_t, \boldsymbol{\theta}_0)\} = \mathbf{0},$$

where $\mathbf{g}(\boldsymbol{X}_t, \boldsymbol{\theta}) = (g_1(\boldsymbol{X}_t, \boldsymbol{\theta}), \dots, g_r(\boldsymbol{X}_t, \boldsymbol{\theta}))^\top$, and $\boldsymbol{\theta}_0 \in \Theta$ is the unknown true parameter vector. When p and r are fixed and finite, the generalized empirical likelihood (GEL) estimators [26,27], as theoretically seductive alternatives to the generalized method of moments (GMM) estimators of [12], enjoy the properties of Wilks' theorem [20,21,24] and Bartlett correction [6,7]. Besides, their higher order asymptotic properties are superior to the GMM estimators; see [1,19]. Refer to [22] for a general overview and [9] for a summary of the recent progress in a variety of fields.

When p and r are diverging with the sample size n, there are few studies of the asymptotic performance of the GEL estimators. The importance of inference under the growing (high) dimensional setting was first recognized in [8] and [13] simultaneously. Tang and Leng [28] and Leng and Tang [17] considered variable selection by adding a penalty term to the traditional likelihood ratio under the circumstances of mean vector and general moment restrictions, respectively. A recent

[*] Corresponding author.
E-mail address: tianlemeng@outlook.com (L. Tian).

https://doi.org/10.1016/j.jmva.2018.12.010
0047-259X/© 2019 Elsevier Inc. All rights reserved.

paper by Chang et al. [4] offered a new scope with the notion of penalized empirical likelihood, which allows both the dimensionalities of model parameters and estimating equations to grow exponentially with the sample size.

The results mentioned above were obtained in the context of independent data. Lahiri et al. [16] considered the dependent data case and introduced a penalized EL estimator for high-dimensional sparse mean parameters. Presumably, this modified EL does not maintain all of the excellent properties of the original EL estimator. Chang et al. [3] employed the blocking technique to handle the dependence in the original time series and the corresponding estimating functions. Hence, their GEL estimator preserves the self-studentized property of the traditional GEL estimator, which is not held in [16]. However, Chang et al. [3] did not give any criterion to choose the blocking number M, which seriously influences the GEL's estimation efficiency and application. Instead of the blocking technique, we can imagine that other methods, especially the popular local smoothing, may also work for growing (high) dimensional dependent time series data and may be able to circumvent this type of tuning parameter selection problem. This is because we already have several out-of-the-box solutions for the selection of the bandwidth parameter of the local smoothing method in the literature; see, e.g., [2]. This is exactly the starting point of this paper. In contrast, compared with the work of [3], our proposed method can use more data, which will lead to better finite-sample performance.

The idea of local smoothing has already appeared in the literature on empirical likelihood methods. Smith [26] incorporated the smoothed linear moment functions with the empirical likelihood to address the potential serial correlation in the moment functions. Subsequently, Smith [27] demonstrated that the smoothed empirical likelihood procedure offers alternative one-step estimators in the setting of weakly dependent data, which are asymptotically equivalent to their two-step GMM counterparts. Anatolyev [1] further derived the second order asymptotic bias of a smoothed generalized empirical likelihood estimator, which revealed that compared with GMM, this estimator avoids the bias term associated with the correlation between the moment function and its derivative, while the bias term associated with third moments depends on the kernel function. All the above work is restricted to the fixed p and r case, and extensions to diverging p or r cases are unclear.

Faced with growing (high) dimensionality, i.e., $p \to \infty$, a sparsity assumption is reasonable, and sparse models can improve the prediction accuracy to a certain extent. With the exception of the vast literature on the penalized likelihood approach, penalized EL or GEL has also been studied for general estimating equations with diverging dimensionality; see, e.g., [17] and [3] for reviews. Following this work, this paper proposes a smoothed penalized GEL method for diverging dimensional time series data. We use the kernel-based smoothed moment functions [1,26,27] to accommodate the temporal dependence among the data. Both r and p are allowed to diverge with the sample size n. When $r \geq p$, we first obtain the consistency, the rate of convergence and asymptotic normality of the smoothed GEL estimator [1] by properly restricting the growing rates of r, p and the truncating parameter h_n incorporated by the smoothed moment function. Then, when $p \geq r$ and the sparse assumption stands, we investigate the oracle property of the smoothed penalized GEL estimator, i.e., it identifies the true model with probability tending to 1. Furthermore, the estimated non-zero parameters remain asymptotically normal. It is worth noting that the oracle property mentioned above is tenable without imposing stringent distributional assumptions. Accordingly, our proposed estimator is robust against model misspecification.

The rest of the paper is organized as follows. In Section 2, we propose the smoothed generalized empirical likelihood (SGEL) and investigate its asymptotic properties. Then, when $p \geq r$ and the sparse assumption stands, the penalized version of SGEL (SPGEL) is given and studied in Section 3. Some implementation issues are discussed in Section 4. Several simulation results are presented in the next section to illustrate the finite-sample performance of the SGEL and SPGEL estimators. In Section 6, we employ Istanbul stock exchange data to demonstrate the applicability of our proposed method. Finally, Section 7 concludes this paper, and all the proofs are reported in the Appendix.

2. Smoothed generalized empirical likelihood

Suppose that the following unconditional moment restrictions sum up the available data information:

$$\mathrm{E}\{\boldsymbol{g}(\boldsymbol{X}_t, \boldsymbol{\theta}_0)\} = \mathbf{0}.$$

Here, we assume $r > p$. As mentioned in the Introduction, r and $p \to \infty$ as sample size $n \to \infty$. The dimension of $\boldsymbol{X}_t$, denoted by d, can be either diverging with n or fixed.

We now assume that the α-mixing condition [10] holds, i.e., as $k \to \infty$,

$$\alpha_X(k) = \sup_d \sup_{\boldsymbol{A}\in\mathcal{F}_{-\infty}^0, \boldsymbol{B}\in\mathcal{F}_k^{+\infty}} |\Pr(\boldsymbol{A}\cap\boldsymbol{B}) - \Pr(\boldsymbol{A})\Pr(\boldsymbol{B})| \to 0,$$

where $k \geq 1$ and $\mathcal{F}_u^v = \sigma(\boldsymbol{X}_t : u \leq t \leq v)$ is the σ-field generated by $\boldsymbol{X}_t$ from time u to v. This condition describes the degree of dependence among the data $\boldsymbol{X}_t$. In the context without dependence, it is easy to see that $\alpha_X(k) = 0$ for every integer $k \geq 1$.

To address the dependence, we introduce a kernel function $k(x)$ satisfying the following properties: (a) $k(x) : [-b, b] \to [-\bar{k}, \bar{k}]$, where b and $\bar{k}$ are finite; (b) for all $x \in [-b, b]$, $k(x) = k(-x)$; (c) $k(x)$ is continuous on $(-b, b)$; (d) $\int_{-b}^{b} k(x)\,dx = 1$. Various frequently-used kernels satisfy the properties given above, such as the truncated, Parzen and Bartlett kernel; see [2] for details. We then consider the smoothed moment function [1,26,27], viz.

$$\boldsymbol{m}_t(\boldsymbol{X}_t, \boldsymbol{\theta}) = \sum_{w=-h_n}^{h_n} \kappa(w)\boldsymbol{g}(\boldsymbol{X}_{t-w}, \boldsymbol{\theta}),$$

where $\kappa(w) = k(w/\delta_n)/\delta_n$, and δ_n is the bandwidth parameter growing to infinity much more slowly than n and chosen to ensure that $\sum_{w=-h_n}^{h_n} \kappa(w) = 1$, where $h_n = \lfloor b\delta_n \rfloor$, and $\lfloor a \rfloor$ means the integer part of a. The smoothed generalized empirical likelihood (SGEL) [1] is defined as

$$\ell(\boldsymbol{\theta}) = \sup\left\{\prod_{t=1}^{n} \pi_t : \pi_1, \ldots, \pi_n \in (0,1), \sum_{t=1}^{n} \pi_t = 1, \sum_{t=1}^{n} \pi_t \boldsymbol{m}_t(\boldsymbol{X}_t, \boldsymbol{\theta}) = 0\right\}.$$

As pointed in [26], the smoothed moment function with $\kappa(w)$ incorporated takes into account the potential serial correlation in the moment functions, which renders the implicit metric imposed by the generalized empirical likelihood appropriate for efficient estimation. The bandwidth parameter h_n may be viewed as reflecting the order of serial correlation in the moment functions. Hence, h_n will typically depend on n and will need to increase with the sample size n at a properly slow rate; see [2].

Following the conventional optimization procedure, the SGEL estimator $\hat{\boldsymbol{\theta}}_n$ together with the $r \times 1$ vector of Lagrange multipliers $\hat{\boldsymbol{\lambda}}$ can be obtained by solving the saddle point problem

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{\boldsymbol{\lambda} \in \hat{\Lambda}_n(\boldsymbol{\theta})} \sum_{t=1}^{n} \rho\{\boldsymbol{\lambda}^\top \boldsymbol{m}_t(\boldsymbol{X}_t, \boldsymbol{\theta})\}, \tag{1}$$

where $\hat{\Lambda}_n(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} \in \mathbb{R}^r : \boldsymbol{\lambda}^\top \boldsymbol{m}_t(\boldsymbol{X}_t, \boldsymbol{\theta}) \in \Upsilon, t \in \{1, \ldots, n\}, \boldsymbol{\theta} \in \Theta\}$, Υ is an open interval containing 0, and the link function $\rho(\upsilon)$ indexes the member of the GEL class. When $\rho(\upsilon) = \ln(1+\upsilon)$, it turns out to be the classical empirical likelihood (EL) estimator [20,24]; when $\rho(\upsilon) = -\exp(\upsilon)$, we obtain the exponential tilting (ET) estimator of [14]; when $\rho(\upsilon) = -\upsilon^2/2 - \upsilon$, it is the continuous updating (CU) estimator of [19]. In general, we assume that the link function $\rho(\upsilon)$ is a concave function such that (a) 0 is an interior point of the domain of ρ; (b) $\rho_\upsilon(0) \neq 0$ where $\rho_\upsilon(\upsilon) = \partial\rho(\upsilon)/\partial\upsilon$; (c) $\rho_{\upsilon\upsilon}(0) \leq 0$, where $\rho_{\upsilon\upsilon} = \partial^2\rho(\upsilon)/\partial^2\upsilon$; see [27].

Define

$$\hat{S}_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{t=1}^{n} \rho\{\boldsymbol{\lambda}^\top \boldsymbol{m}_t(\boldsymbol{\theta})\}, \tag{2}$$

where we denote $\boldsymbol{m}_t(\boldsymbol{X}_t, \boldsymbol{\theta})$ by $\boldsymbol{m}_t(\boldsymbol{\theta})$ for the sake of brevity, and we will maintain this notation hereinafter. Then we obtain the score function regarding $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\lambda}}$, viz.

$$\nabla_{\boldsymbol{\lambda}} \hat{S}_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\lambda}}) = \boldsymbol{0}.$$

By the implicit function theorem, e.g., Theorem 9.28 in [25], and in view of the concavity of $\hat{S}_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$ on $\boldsymbol{\lambda}$, $\hat{S}_n\{\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\} = \max_{\boldsymbol{\lambda} \in \hat{\Lambda}_n(\boldsymbol{\theta})} \hat{S}_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$. By the envelope theorem,

$$\boldsymbol{0} = \nabla_{\boldsymbol{\theta}} \hat{S}_n\{\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = \frac{1}{n}\sum_{t=1}^{n} \rho_\upsilon\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\{\nabla_{\boldsymbol{\theta}} \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}^\top \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n).$$

To establish the asymptotic properties of the SGEL estimator, we introduce the following notations and regularity conditions. In what follows, C_i denotes a positive finite constant which is different for each i. If A is a matrix, $\|A\|_F$ and $\|A\|_2$ denote its Frobenius-norm and operator-norm, respectively. For a vector a, we use $\|a\|_2$ to denote its L_2-norm. For convenience's sake, we abbreviate $\boldsymbol{g}(\boldsymbol{X}_t, \boldsymbol{\theta})$ by $\boldsymbol{g}_t(\boldsymbol{\theta})$ and denote the ith element of $\boldsymbol{g}(x, \boldsymbol{\theta})$ by $g_i(\boldsymbol{\theta})$. Then, the jth element of $\boldsymbol{g}_t(\boldsymbol{\theta})$ and $\boldsymbol{m}_t(\boldsymbol{\theta})$ are denoted by $g_{t,j}(\boldsymbol{\theta})$ and $m_{t,j}(\boldsymbol{\theta})$, respectively. Additionally, let $\bar{\boldsymbol{g}}(\boldsymbol{\theta}) = \{\boldsymbol{g}_1(\boldsymbol{\theta}) + \cdots + \boldsymbol{g}_n(\boldsymbol{\theta})\}/n$, $V_n = \mathrm{var}\{\sqrt{n}\,\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0)\}$, $\bar{m}(\boldsymbol{\theta}) = \{\boldsymbol{m}_1(\boldsymbol{\theta}) + \cdots + \boldsymbol{m}_n(\boldsymbol{\theta})\}/n$, and $U_h = \mathrm{var}\{\sqrt{h_n}\,\boldsymbol{m}_t(\boldsymbol{\theta}_0)\}$. Note that U_h can be seen as the covariance matrix of the smoothed moment function at time t. We need the following conditions throughout the paper.

(A.1) (i) $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is strictly stationary time series, and $\sum_{k=1}^{\infty} k\alpha_X^{1-2/\gamma} \leq \infty$ for some $\gamma > 2$; (ii) $\mathrm{E}\{\boldsymbol{g}_t(\boldsymbol{\theta}_0)\} = 0$, and

$$\inf_{\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \geq \varepsilon\}} \|\mathrm{E}\{\boldsymbol{g}_t(\boldsymbol{\theta})\}\|_2 \geq \triangle_1(r,p)\triangle_2(\varepsilon) > 0$$

for any ε and some positives functions $\triangle_1(r,p)$ and $\triangle_2(\varepsilon)$, where $\liminf_{r,p\to\infty} \triangle_1(r,p) > 0$; (iii) $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\bar{\boldsymbol{g}}(\boldsymbol{\theta}) - \mathrm{E}\{\boldsymbol{g}_t(\boldsymbol{\theta})\}\|_2 = O_p\{\triangle_1(r,p)\}$.

(A.2) (i) $\boldsymbol{\theta}_0 \in \mathrm{int}(\boldsymbol{\Theta})$, and a small $\|\cdot\|_2$-neighborhood of $\boldsymbol{\theta}_0$ is contained in $\boldsymbol{\Theta}$ in which $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})$ is continuously differentiable with respect to $\boldsymbol{\theta}$ for any $\boldsymbol{x} \in \boldsymbol{X}$, the domain of $\boldsymbol{X}_t$. Furthermore, for all $i \in \{1, \ldots, r\}$ and $i \in \{1, \ldots, p\}$,

$$\left|\frac{\partial}{\partial\theta_j} g_i(\boldsymbol{x}, \boldsymbol{\theta})\right| \leq T_{n,ij}(\boldsymbol{x})$$

for some functions $T_{n,ij}(\boldsymbol{x})$ satisfying $\mathrm{E}\{T_{n,ij}^2(\boldsymbol{X}_t)\} \leq C$ for any i, j; (ii) $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})\|_2 \leq \sqrt{r}\, B_n(\boldsymbol{x})$, where $\mathrm{E}\{B_n^\gamma(\boldsymbol{X}_t)\} \leq C$ for γ specified in (A.1)(i); (iii) $\mathrm{E}\{|g_{t,j}(\boldsymbol{\theta}_0)|^{2\gamma}\} \leq C$ for all $j \in \{1, \ldots, r\}$; (iv) the eigenvalues of $[\mathrm{E}\{\nabla_\theta \boldsymbol{g}_t(\boldsymbol{\theta})\}]^\top [\mathrm{E}\{\nabla_\theta \boldsymbol{g}_t(\boldsymbol{\theta})\}]$

in a $\|\cdot\|_2$- neighborhood of $\boldsymbol{\theta}_0$ are uniformly bounded away from zero and infinity, and

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\max}\left\{\frac{1}{n}\sum_{t=1}^{n}\mathbf{g}_t(\boldsymbol{\theta})\mathbf{g}_t(\boldsymbol{\theta})^\top\right\} \le C$$

with probability approaching 1.

(A.3) $\mathbf{g}(\boldsymbol{x},\boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ in a $\|\cdot\|_2$-neighborhood of $\boldsymbol{\theta}_0$ for any $\boldsymbol{x}\in\mathcal{X}$, and for all $i\in\{1,\dots,r\}$ and $k\in\{1,\dots,p\}$,

$$\left|\frac{\partial^2}{\partial\theta_j\partial\theta_k}g_i(\boldsymbol{x},\boldsymbol{\theta})\right| \le K_{n,ijk}(\boldsymbol{x})$$

for some functions $K_{n,ijk}(\boldsymbol{x})$ satisfying $\mathrm{E}\{K^2_{n,ijk}(\boldsymbol{X}_t)\}\le C$ for any i, j and k.

(A.4) $\liminf_{\tau\to 0}\liminf_{\theta\to 0+} p_\tau^\top(\boldsymbol{\theta})/\tau > 0$.

(A.5) $\max_{j\in\mathcal{A}} p_\tau(|\theta_{0j}|)\le C\tau$ for some positive constant C, where $\mathcal{A}$ will be defined later.

These regularities are frequently assumed in the literature; see, e.g., [3]. They are often used under the circumstances of weakly dependent time series data and diverging parameter space, and they are extensions of GEL for the fixed dimension setting. To construct the asymptotic properties of the SGEL estimator, we also need the following conditions:

$$r^2h_n^2n^{2/\gamma-1}=o(1), \tag{3}$$

$$\sup_n \mathrm{E}\{|\boldsymbol{\beta}_n^\top\mathbf{g}_t(\boldsymbol{\theta}_0)|^\gamma\}<\infty, \tag{4}$$

where $\boldsymbol{\beta}_n=-U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]\times\left[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]\right]^{-1/2}\boldsymbol{\alpha}_n$ for any vector $\boldsymbol{\alpha}_n$ with unit L_2-norm and $\gamma>2$ specified in (A.1)(i).

Next, we present the consistency, rate of convergence and asymptotic normality of the SGEL estimator.

Theorem 1. *Assume that the eigenvalues of U_h are uniformly bounded away from zero and infinity, and Conditions* (A.1), (A.2) *and* (3) *hold. Then $\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|_2\xrightarrow{p}0$. Furthermore, if $r^2ph_n^2/n=o(1)$, then $\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|_2=O_p(\sqrt{r/n})$, and $\|\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)\|_2=O_p(h_n\sqrt{r/n})$.*

This result echoes Theorem 1 in [3], where a blocking technique is used to handle the dependence among the data. It is noteworthy that our truncating parameter h_n incorporated by the smoothed moment function displays a similar effect with the block size M used in the blocking technique. However, it seems that we need fewer conditions for h_n than those needed for M in [3]. Moreover, if h_n is fixed, Condition (3) guarantees that $r=o(\sqrt{n})$ for large enough γ. Finally, Theorem 1 generalizes the existing results on the consistency of GEL estimator; see [24] and [19] for the case of fixed r and independent data, and [17] for the case of diverging r and independent data.

The following theorem states the asymptotic normality of the SGEL estimator.

Theorem 2. *Under Conditions* (A.1) *and* (A.3)*, assume that the eigenvalues of U_h and V_n are uniformly bounded away from zero and infinity. If*

$$r^3h_n^2n^{2/\gamma-1}=o(1),\quad r^3p^2/n=o(1)\quad\textit{and}\quad r^3ph_n^2/n=o(1), \tag{5}$$

then for any $\boldsymbol{\alpha}_n\in\mathbb{R}^p$ with unit L_2 norm such that (4) *holds, we have*

$$\sqrt{n}\,\boldsymbol{\alpha}_n^\top\left[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]^\top\right]^{-1/2}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}](\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$$

is asymptotically $\mathcal{N}(0,1)$ as $n\to\infty$.

Note that (4) is a necessary condition to use the Central Limit Theorem in the dependent data setting. From Theorem 2, if $\|V_n-U_h\|_2\to 0$, the SGEL estimator is asymptotically efficient. We can obtain the efficient estimator by properly choosing the diverging rate of h_n such that the conditions required for Theorem 2 hold.

3. Smoothed penalized generalized empirical likelihood

In high-dimensional data analysis, especially when $p>r$, the classical approach is to assume that only some of the covariates are active. Define $\mathcal{A}=\{j:\theta_{0j}\neq 0\}$, where θ_{0j} is the jth component of the true parameter vector $\boldsymbol{\theta}_0$ and let $s=|\mathcal{A}|$, i.e., the cardinality of $\mathcal{A}$ is s. Without loss of generality, denote $\boldsymbol{\theta}=(\boldsymbol{\theta}_1^\top,\boldsymbol{\theta}_2^\top)^\top$, where $\boldsymbol{\theta}_1\in\mathbb{R}^s$ and $\boldsymbol{\theta}_2\in\mathbb{R}^{p-s}$ correspond to nonzero and zero subsets of $\boldsymbol{\theta}$, respectively. Given such sparsity, we only need to assume $s\le r$ to ensure the identifiability of the relevant parameters. To estimate the relevant parameters efficiently, we add a penalty term to (1) and get the smoothed penalized generalized empirical likelihood (SPGEL) estimator

$$\hat{\boldsymbol{\theta}}_n^{(pe)}=\arg\min_{\boldsymbol{\theta}\in\Theta}\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})}\left\{\sum_{t=1}^{n}\rho\{\boldsymbol{\lambda}^\top\mathbf{m}_t(\boldsymbol{X}_t,\boldsymbol{\theta})\}+n\sum_{j=1}^{p}p_\tau(|\theta_j|)\right\},$$

where p_τ is some penalty function with a tuning parameter τ satisfying Conditions (A.4)–(A.5). Penalty functions such as the one defined in [11] and the minimax concave penalty function of [30] satisfy the aforementioned conditions. Define

$$\mathbf{S}(\boldsymbol{\theta}_0) = \left[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]\right]^{-1}\left[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]\right]$$
$$\times\left[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]\right]^{-1}.$$

We then decompose $\mathbf{S}(\boldsymbol{\theta}_0)$ as

$$\mathbf{S}(\boldsymbol{\theta}_0) = \begin{pmatrix} \mathbf{S}_{11}(\boldsymbol{\theta}_0) & \mathbf{S}_{12}(\boldsymbol{\theta}_0) \\ \mathbf{S}_{21}(\boldsymbol{\theta}_0) & \mathbf{S}_{22}(\boldsymbol{\theta}_0) \end{pmatrix}, \tag{6}$$

where $\mathbf{S}_{11}(\boldsymbol{\theta}_0)$ and $\mathbf{S}_{22}(\boldsymbol{\theta}_0)$ are $s \times s$ and $(p-s)\times(p-s)$ sub-matrices, respectively. To establish the oracle property of the SPGEL estimator, we need the following conditions regarding s and τ:

$$s\tau n/(rh_n) = O(1), \quad \tau\sqrt{n/r}\,/h_n \to \infty. \tag{7}$$

Denote the SPGEL estimator $\hat{\boldsymbol{\theta}}_n^{(pe)} = (\hat{\boldsymbol{\theta}}_n^{(1)^\top}, \hat{\boldsymbol{\theta}}_n^{(2)^\top})^\top$ and $\mathbf{S}_p(\boldsymbol{\theta}_0) = \mathbf{S}_{11}(\boldsymbol{\theta}_0) - \mathbf{S}_{12}(\boldsymbol{\theta}_0)\mathbf{S}_{22}^{-1}(\boldsymbol{\theta}_0)\mathbf{S}_{21}(\boldsymbol{\theta}_0)$. We have the following theorem.

Theorem 3. *Under Conditions* (A.1)–(A.5), *assume that the eigenvalues of U_h are uniformly bounded away from zero and infinity. If $\max_{j\in\mathcal{A}} p'_\tau(|\theta_{0j}|) = o(\sqrt{r/n})$, $\min_{j\in\mathcal{A}} |\theta_{0j}|/\tau \to \infty$, and (7) holds, we have the following results:*

(i) $\Pr\{\hat{\boldsymbol{\theta}}_n^{(2)} = \mathbf{0}\} \to 1$ as $n \to \infty$, provided that (3) holds and $r^2ph_n^2/n = o(1)$.

(ii) If the eigenvalues of V_n are uniformly bounded away from zero and infinity, then for any $\boldsymbol{\alpha}_n \in \mathbb{R}^s$ with unit L_2-norm, we have, as $n \to \infty$,

$$\sqrt{n}\,\boldsymbol{\alpha}_n^\top \boldsymbol{S}_p^{-1/2}(\boldsymbol{\theta}_0)\{\hat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}\} \rightsquigarrow \mathcal{N}(0, 1),$$

provided that

(a) for independent data, $r^3p^2/n = o(1)$ and $r^3n^{2/\gamma-1} = o(1)$;

(b) for dependent data, (5) holds and $\boldsymbol{\alpha}_n$ satisfies (4) with

$$\boldsymbol{\beta}_n = -\,U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]\left[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]\right]^{-1}$$
$$\times\,[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}^{(1)}}\mathbf{g}_t(\boldsymbol{\theta}_0)\}]\{\boldsymbol{S}_{11}(\boldsymbol{\theta}_0) - \boldsymbol{S}_{12}(\boldsymbol{\theta}_0)\boldsymbol{S}_{22}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{S}_{21}(\boldsymbol{\theta}_0)\}^{1/2}\boldsymbol{\alpha}_n.$$

Theorem 3 indicates that the zero components of the true parameter $\boldsymbol{\theta}_0$ can be estimated as zero with probability tending to 1. Comparing Theorem 3 with Theorem 2, it can be easily seen that the SPGEL estimator is more efficient than the SGEL estimator in estimating the nonzero components of $\boldsymbol{\theta}_0$. The efficiency is gained via penalization, by which we reduce the effective dimension of the parameter to be estimated. Our result for SPGEL estimator resembles that for the penalized GEL estimator based on the blocking technique when data dependence exists. There are no surprises since the smoothed estimating function is kind of a weighted average of the original estimating function $\mathbf{g}_t(\boldsymbol{\theta})$ while the estimating function constructed for a block is a simple average.

4. Implementation issues

There are nontrivial issues related to the computation of the SGEL and SPGEL estimators. For both of them, we need to choose a suitable kernel and the bandwidth or truncating parameter. Choices for the kernel function k and bandwidth parameter h_n should satisfy the properties given in Section 2 and assumptions required by Theorems 1–3.

For convenience of calculations, we choose the popular truncated kernel defined by $k(x) = 1$ for $|x| \le 1$ and $k(x) = 0$ otherwise, which implies a Bartlett kernel for the heteroscedasticity and autocorrelation consistent (HAC) matrix. However, how to practically choose a suitable bandwidth parameter is a challenging problem. Fortunately, Andrews [2] proposed an automated bandwidth estimator, which is asymptotically optimal under the asymptotic truncated mean squared error criterion for the covariance matrix estimation. While it is not clear whether this optimality holds for the smoothed moment function, we just follow this procedure as done in [5].

According to [2], the choice of the bandwidth parameter is closely related to the choice of the kernel function, and different kernels may imply different bandwidths. Conditions like (3) are quite mild, and the optimal bandwidth we choose can satisfy these conditions for some specific γ. For example, when r is fixed, the order of the optimal bandwidth for the Bartlett kernel is $O(n^{1/3})$; see [2]. Condition (3) requires that the order of the bandwidth should be $o(n^{1/2-1/\gamma})$. Clearly, the bandwidth we choose satisfies Condition (3) in this case. Numerical results based on different bandwidths were conducted to check whether the proposed method is sensitive to the bandwidth. The results are reported in Section 5.

For SPGEL itself, a non-differentiable penalty and a tuning parameter are introduced. For the selection of the tuning parameter, we use the BIC criterion [17,29]. Moreover, the Nelder–Mead algorithm [18] is used to solve the problem of non-differentiable penalty, since this method uses only function values and works reasonably well for non-differentiable functions.

Table 1
Empirical averages of the squared estimation errors ($\times 100$) of the smoothed GMM, EL and PEL with $c = 3$.

ψ	$n = 500$			$n = 1000$			$n = 2000$		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
GMM	17.87	19.38	22.23	15.26	15.76	17.08	12.54	12.77	13.36
EL	10.05	10.65	12.49	8.41	8.92	9.89	7.30	7.55	7.84
SEL	9.61	10.11	12.39	7.65	8.09	9.47	6.98	7.38	7.56
PEL	9.40	10.61	12.09	7.69	8.20	9.54	6.88	7.26	7.62
SPEL	8.78	10.01	11.43	6.77	7.29	8.38	5.94	6.34	6.16

5. Simulation results

We present several simulation results to investigate the finite-sample performance of the SGEL and SPGEL estimators in the high-dimensional weakly dependent time series setting. Their performances are then compared with that of traditional GEL, penalized GEL and smoothed GMM estimators with HAC positive definite weight matrices. We consider three members of the GEL family in the simulations: EL, ET and CU. The penalty function $p_\tau(\theta)$ we choose in the simulations satisfies

$$p'_\tau(\theta) = \tau \left\{ \mathbf{1}(\theta \le \tau) + \frac{(a\tau - \theta)_+}{(a-1)\tau} \mathbf{1}(\theta > \tau) \right\},$$

for $\theta > 0$, where $a = 3.7$ and $(x)_+ = x$ for $x > 0$ and $(x)_+ = 0$ otherwise; see [11].

We examine the generalized linear model with nonlinear moment restrictions. The covariates $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are generated from the vector autoregressive model (VAR) of order 1, viz. $\boldsymbol{Z}_t = \psi \boldsymbol{Z}_{t-1} + \boldsymbol{\varepsilon}_t$ where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$, $\boldsymbol{\Sigma}_\varepsilon = (\sigma_{i,j})_{p\times p}$, $\sigma_{i,i} = 1-\psi^2, \sigma_{i,i\pm1} = (1-\psi^2)/2$ and $\sigma_{i,j} = 0$ for $|i-j| > 1$. The stationary distribution of $\boldsymbol{Z}_t$ is $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_z)$ where $\boldsymbol{\Sigma}_z = (\tilde{\sigma}_{i,j})_{p\times p}$, $\tilde{\sigma}_{i,i} = 1, \tilde{\sigma}_{i,i\pm1} = 1/2$ and $\tilde{\sigma}_{i,j} = 0$ for $|i-j| > 1$. The response variable $Y_1, \ldots, Y_n$ take the values 0 or 1, and

$$\Pr(Y_t = 1|\boldsymbol{Z}_t) = \exp(1 + \boldsymbol{Z}_t^\top \boldsymbol{\theta}_0)/\{1 + \exp(1 + \boldsymbol{Z}_t^\top \boldsymbol{\theta}_0)\},$$

where the true parameter $\boldsymbol{\theta}_0 = (0.8, 0.2, 0, \ldots, 0)^\top \in \mathbb{R}^p$. Then we obtain

$$\mathrm{E}\{Y_t - \exp(1 + \boldsymbol{Z}_t^\top \boldsymbol{\theta}_0)/\{1 + \exp(1 + \boldsymbol{Z}_t^\top \boldsymbol{\theta}_0)\}|\boldsymbol{Z}_t\} = 0.$$

The corresponding moment restrictions can be constructed as

$$\boldsymbol{g}(\boldsymbol{X}_t, \boldsymbol{\theta}) = (\boldsymbol{Z}_t^\top, \boldsymbol{W}_t^\top)^\top \times \{Y_t - \exp(1 + \boldsymbol{Z}_t^\top \boldsymbol{\theta})/[1 + \exp(1 + \boldsymbol{Z}_t^\top \boldsymbol{\theta})]\},$$

where $\boldsymbol{Z}_t = (Z_{1,t}, \ldots, Z_{p,t})^\top$ and $\boldsymbol{W}_t = (Z_{1,t}^2, \ldots, Z_{p,t}^2)^\top$.

In this simulation model, we choose $n = 500$, 1000, and 2000, respectively, and we take p as the integer part of $cn^{2/15}$, where $c = 3$ and 4. The parameter ψ in the VAR process $\boldsymbol{Z}_t$ controlling the degree of serial dependence is set to be 0.1, 0.3 and 0.5, respectively. We then summarize the simulation results based on 100 repetitions. For each repetition, we get the SGEL, SPGEL and GMM estimators and compute the L_2 distance between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ as $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = \{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\}^{1/2}$. The numerical results confirm our theoretical findings concerning the SGEL and SPGEL estimators.

Tables 1 and 2 summarize the empirical averages of the squared estimation errors of the smoothed GMM, GEL, SGEL, PGEL and SPGEL estimators with $c = 3$ and $c = 4$, respectively. Notice that the results for the smoothed penalized exponential tilting method (SPET) and the smoothed penalized continuous updating method (SPCU) are quite similar to those for the smoothed penalized empirical likelihood (SPEL) method, so we only present the results pertaining to SPEL for simplicity. For the smoothed GMM, we choose the optimal Quadratic Spectral kernel with its corresponding optimal bandwidth as [2] suggested.

It can be seen easily that the performance of each estimator is improved when the sample size increases, which is expected given the convergence of our proposed SGEL and SPGEl estimators. We also observe that the SGEL and SPGEL estimators perform better than the GEL and PGEL estimators, which do not employ local smoothing for the moment restrictions. However, the efficiency gain is not so large in the tables since the local smoothing does not take effect in every replicate. If these failure samples are left out, we can obtain much more efficiency gain due to the local smoothing of the moment functions. The empirical averages of the squared estimation errors of the smoothed GMM estimators were much larger than those of the EL, SEL estimators and their penalized analogues, which can be deduced from the conclusions of [19] and [1] on GMM versus GEL for fixed finite-dimensional data settings. Finally, the penalized GEL and SGEL estimators have smaller empirical averages of the squared estimation errors, indicating the gain in efficiency by adding the penalty term for variable selection.

Of note, the proposed method does not work well when ψ, which characterizes the serial dependence of the data, is less than 0.1. This is easy to understand because the dependence under this setting is too weak for the local smoothing technique to take effect. Moreover, although our method does perform better in the settings given above, it cannot efface the bad influence of the data dependence completely. Generally speaking, the performance of the proposed estimators is closely related to the sample size, the diverging speed of the dimension p, the data dependence and the interaction between them.

Table 2
Empirical averages of the squared estimation errors ($\times 100$) of the smoothed GMM, EL and PEL with $c = 4$.

ψ	$n = 500$			$n = 1000$			$n = 2000$		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
GMM	15.49	16.66	19.90	12.50	13.11	14.51	10.52	10.25	11.14
EL	11.19	11.41	14.28	7.86	8.56	9.41	6.43	6.79	7.47
SEL	11.00	11.11	16.57	7.58	8.45	9.06	6.31	6.75	7.93
PEL	10.40	11.07	13.78	7.61	8.24	9.60	6.05	6.71	7.45
SPEL	8.98	9.77	13.02	6.41	6.98	8.41	5.30	5.72	6.42

Table 3
Empirical averages of the squared estimation errors ($\times 100$) of the smoothed GMM, EL and PEL with higher dimensions.

ψ	$c = 5$			$c = 6$		
	0.1	0.3	0.5	0.1	0.3	0.5
GMM	13.88	15.34	18.12	11.62	13.59	15.65
EL	11.39	12.41	14.65	10.98	12.82	14.84
SEL	11.74	12.82	18.54	11.78	13.83	18.44
PEL	7.45	11.11	12.86	9.97	11.42	14.11
SPEL	6.42	10.58	13.67	9.59	11.22	14.95

Table 4
Empirical averages of the squared estimation errors ($\times 100$) of the smoothed GMM, EL and SEL with different bandwidths.

GMM	EL	SEL				
	Bandwidth	$0.5h^*$	$0.75h^*$	h^*	$1.25h^*$	$1.5h^*$
				$c = 3$		
20.02	10.77	10.77	10.77	10.70	10.61	11.01
				$c = 4$		
17.01	12.17	12.17	12.17	12.15	12.29	13.68

Note that we did not compare our method with that of [3]. Because the performance of their method depends heavily on the choice of the blocking parameters, but they did not give any suggestion on the selection of these parameters. Nevertheless, our SPGEL method is a competitive alternative to theirs with better practicality.

We follow the Editor's suggestion to present more simulation results for higher dimensions in Table 3. In this setting, we set $n = 500$, $\psi = 0.1, 0.3, 0.5$, and $c = 5,6$. The results are quite similar to those given in Tables 1 and 2. Moreover, the results show that when c or ψ gets larger, both SEL and SPEL seem to lose efficacy, but SPEL seems to be more tolerant of dimensionality.

Bandwidth selection is of importance when kernel smoothing is employed. Numerical results are given in Table 4 to compare the SEL based on different bandwidths, where we use the optimal bandwidth h^* computed by the method given in [2] as the benchmark and compare the corresponding performance of SEL with those with bandwidth equaling $0.5h^*$, $0.75h^*$, $1.25h^*$ and $1.5h^*$. We choose $n = 500$, $\psi = 0.3$, and $c = 3$ and 4. Table 4 shows that our proposed method is robust to the bandwidth when the working bandwidth does not deviate too much from the benchmark.

6. Real data analysis

We use the Istanbul stock exchange data to illustrate the applicability of our proposed method. These data are available at http://archive.ics.uci.edu/ml/datasets/ISTANBUL+STOCK+EXCHANGE. The Istanbul stock exchange data include returns of the Istanbul stock exchange national 100 index (ISE) with seven other international indexes: Standard and Poor's 500 return index (SP), the stock market return index of Germany (DAX), the stock market return index of the United Kingdom (FTSE), the stock market return index of Japan (NIKKEI), the stock market return index of Brazil (BOVESPA), the MSCI European index (EU) and the MSCI emerging markets index (EM) from June 5, 2009 to February 22, 2011. There are thus 536 observations in total. There exist two measures of ISE: one is based on TL and the other one is USD based. The first one is employed in this study and is transformed to a dichotomous variable by $\mathbf{1}(\mathrm{ISE} > 0)$. Our aim is to investigate the influence of the seven indexes mentioned above on ISE. In order to explore the degree of dependence of these indexes, we compute their correlation matrix first and go further to construct a VAR model with order 1 for them. The results show that the indexes enjoy a weakly dependent structure.

It is appropriate to assume a logistic model, viz.

$$\Pr(\mathrm{ISE}_t = 1|\boldsymbol{X}_t) = \frac{\exp(\beta_0 + \boldsymbol{X}_t^\top \boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{X}_t^\top \boldsymbol{\beta})},$$

Table 5
The fitted coefficients and standard errors (in parentheses) for estimators from GMM, EL, SEL and SPEL.

Variable	GMM		EL		SEL		SPEL	
Intercept	0.27	(0.07)	−0.14	(0.91)	−0.10	(0.86)	0.00	(0.87)
SP	23.11	(11.60)	6.10	(14.28)	5.89	(13.93)	0.00	(16.45)
DAX	6.70	(20.75)	−5.33	(17.11)	−4.77	(14.91)	0.00	(14.98)
FTSE	38.67	(25.05)	39.64	(21.08)	39.66	(21.09)	41.66	(22.78)
NIKKEI	−4.57	(9.32)	0.18	(0.94)	0.18	(0.94)	0.00	(0.82)
BOVESPA	−29.56	(10.18)	−12.23	(16.69)	−12.24	(16.69)	0.00	(17.01)
EU	69.10	(16.76)	67.03	(21.57)	67.03	(21.11)	64.88	(25.52)
EM	93.22	(25.5)	97.69	(28.29)	97.08	(27.2)	93.67	(28.03)

where ISE_t is the transformed ISE, and $\boldsymbol{X}_t$ indicates the aforementioned seven indexes. Under this model, moment restrictions can be constructed as those in the simulation study. We then utilize GMM, EL, SEL and SPEL to estimate the coefficient vector $\boldsymbol{\beta}$ which is of our interest. The results are summarized in Table 5. To get the standard errors of the fitted coefficients, the block bootstrap method [15,23] is employed. The procedure is given below.

(i) Draw a sample X_k of size 1 from the original data set randomly with replacement. Here, $k \in \{1, \dots, 536\}$. Recall that we have $n = 536$ data points in total.
(ii) Choose a block of length $n/2 = 268$: $X_k, \dots, X_{k+267}$ when $k \le 268$, or $X_k, \dots, X_{k-267}$ when $k > 268$.
(iii) Use GMM, EL, SEL and SPEL to estimate the coefficients.
(iv) Repeat (i)–(iii) for $N = 100$ times, and derive the estimators $\hat{\beta}_1, \dots, \hat{\beta}_{100}$ for GMM, EL, SEL and SPEL.
(v) Compute the mean and standard error of the $\hat{\beta}_i$s for GMM, EL, SEL and SPEL, respectively.

From Table 5, the four kinds of estimators simultaneously identify EU and EM as influential variables for the change in ISE. This finding implies a close relationship between the three stock markets. Of note, the estimators of EL and SEL differ only slightly in this specific data set. In addition, the SPEL method performs satisfactorily in variable selection, highlighting the influence of EU and EM on ISE. Furthermore, these results are quite different from that of the L_1-regularized logistic regression method which identifies DAX, FTSE and EU as important variables.

7. Conclusion

In this paper, we have studied the asymptotic properties of SGEL and SPGEL estimators in the setting of growing (high) dimensional weakly dependent time series. The penalized version is implemented when $p > r$ but the true number of non-zero parameters is smaller than or equal to r. We show that the SPGEL estimator maintains the oracle property in spite of the existence of data dependence. To construct the estimators mentioned above, we introduce the smoothed moment functions. Although we use the method given by [2] to choose the truncating parameter h_n introduced by the smoothed moment function as done by [5], we are not sure whether the optimal bandwidth estimator for GMM is applicable to the GEL family. It is a challenging problem to select a uniformly optimal or a sub-optimal bandwidth estimator for the GEL family, and we leave it for further research.

Acknowledgments

We feel very grateful to the Editor-in-Chief, Prof. Christian Genest, an Associate Editor and two referees for very constructive comments and suggestions which have greatly improved our paper. This research benefited from support from the Fundamental Research Funds for the Central Universities, China (Grant No. JBK1707113) and the Joint Lab of Data Science and Business Intelligence at SWUFE, China (Grant No. JBK1805004).

Appendix

Throughout the Appendix, C denotes a generic positive finite constant that may be different in different uses. Let

$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{g}_t(\boldsymbol{\theta}), \quad \bar{\boldsymbol{m}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{m}_t(\boldsymbol{\theta}), \quad U_h = \operatorname{var}\{h_n^{1/2}\boldsymbol{m}_t(\boldsymbol{\theta}_0)\}, \quad \hat{\Omega}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{m}_t(\boldsymbol{\theta})\boldsymbol{m}_t(\boldsymbol{\theta})^\top$$

and $\Omega(\boldsymbol{\theta}) = \mathrm{E}\{\boldsymbol{m}_t(\boldsymbol{\theta})\boldsymbol{m}_t(\boldsymbol{\theta})^\top\}$.

A.1. Preliminary lemmas

The lemmas proposed in this subsection are used to prove Theorem 1.

Lemma A.1. *Under Conditions* (A.1)(ii) *and* (A.2)(ii), $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\bar{\boldsymbol{m}}(\boldsymbol{\theta})-\bar{\boldsymbol{g}}(\boldsymbol{\theta})\|_2 = O_p(\sqrt{r}\,h_n/n)$.

Proof. By Jensen's inequality,

$$\begin{aligned}\mathrm{E}\Big\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\bar{\boldsymbol{m}}(\boldsymbol{\theta})-\bar{\boldsymbol{g}}(\boldsymbol{\theta})\|_2\Big\} &\le \frac{1}{n}\left[\sum_{t=1}^{h_n}\left\{1-\sum_{s=-h_n}^{t-1}\kappa(s)\right\}+\sum_{t=n-h_n+1}^{n}\left\{1-\sum_{t-n}^{h_n}\kappa(s)\right\}\right]\times \mathrm{E}\Big\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{g}_t(\boldsymbol{\theta})\|_2\Big\}\\ &\le (h_n/n)\times \mathrm{E}\Big\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{g}_t(\boldsymbol{\theta})\|_2\Big\}.\end{aligned}$$

Hence, (A.2)(ii) leads to the conclusion. □

Lemma A.2. *Under Conditions* (A.1)(i) *and* (A.2)(iii), $\|\hat{\Omega}(\boldsymbol{\theta}_0)-\Omega(\boldsymbol{\theta}_0)\|_F = O_p(r/\sqrt{n})$.

Proof. Note that

$$\begin{aligned}\mathrm{E}\{\|\hat{\Omega}(\boldsymbol{\theta}_0)-\Omega(\boldsymbol{\theta}_0)\|_F^2\} &= \frac{1}{n}\,\mathrm{E}\big[\mathrm{tr}[\{\boldsymbol{m}_t(\boldsymbol{\theta}_0)\boldsymbol{m}_t(\boldsymbol{\theta}_0)^\top-\Omega(\boldsymbol{\theta}_0)\}^2]\big]\\ &+\frac{1}{n^2}\sum_{t_1\ne t_2}\mathrm{E}\big[\mathrm{tr}\{[\boldsymbol{m}_{t_1}(\boldsymbol{\theta}_0)\boldsymbol{m}_{t_1}(\boldsymbol{\theta}_0)^\top-\Omega(\boldsymbol{\theta}_0)]\times\{\boldsymbol{m}_{t_2}(\boldsymbol{\theta}_0)\boldsymbol{m}_{t_2}(\boldsymbol{\theta}_0)^\top-\Omega(\boldsymbol{\theta}_0)\}\}\big]\equiv A_1+A_2.\end{aligned}$$

As $A_1\le \mathrm{E}\{\|\boldsymbol{m}_t(\boldsymbol{\theta}_0)\|_2^4\}/n$, by Jensen's inequality and (A.2)(iii), $A_1=O(r^2/n)$. At the same time,

$$A_2=\frac{1}{n^2}\sum_{u,v=1}^{r}\sum_{t_1\ne t_2}\mathrm{E}\big[[m_{t_1,u}(\boldsymbol{\theta}_0)\{m_{t_1,v}(\boldsymbol{\theta}_0)-\Omega_{u,v}(\boldsymbol{\theta}_0)\}][m_{t_2,v}(\boldsymbol{\theta}_0)\{m_{t_2,u}(\boldsymbol{\theta}_0)-\Omega_{v,u}(\boldsymbol{\theta}_0)\}]\big],$$

where $\Omega_{u,v}(\boldsymbol{\theta}_0)$ denotes the (u,v)-element of $\Omega_{u,v}(\boldsymbol{\theta}_0)$. By Davydov's inequality and (A.2)(iii),

$$|A_2|\le cr^2n^{-2}\sum_{t_1\ne t_2}\alpha_m(|t_1-t_2|)^{1-2/\gamma}.$$

Hence, by (A.1)(i), $A_2=O(r^2/n)$. From Markov's inequality, $\|\hat{\Omega}(\boldsymbol{\theta}_0)-\Omega(\boldsymbol{\theta}_0)\|_F = O_p(r/\sqrt{n})$. □

Lemma A.3. *Under Conditions* (A.1)(ii), (A.2)(ii) *and* (A.2)(iv), $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\lambda_{\max}\{\hat{\Omega}(\boldsymbol{\theta})\}=O_p(1)$ *provided that* $rh_n/n=o(1)$.

Proof. Using the same approach as in the proof of Lemma A.2, we have

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\sup_{\|\boldsymbol{x}\|_2=1}\left\{\left|\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}^\top\boldsymbol{m}_t(\boldsymbol{\theta})\boldsymbol{m}_t(\boldsymbol{\theta})^\top\boldsymbol{x}-\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}^\top\boldsymbol{g}_t(\boldsymbol{\theta})\boldsymbol{g}_t(\boldsymbol{\theta})^\top\boldsymbol{x}\right|\right\}=O_p(rh_n/n).$$

Then $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\lambda_{\max}\{\hat{\Omega}(\boldsymbol{\theta})\}\le\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\lambda_{\max}\{\sum_{t=1}^{n}\boldsymbol{g}_t(\boldsymbol{\theta})\boldsymbol{g}_t(\boldsymbol{\theta})^\top/n\}+o_p(1)$. The result can then be deduced from (A.2)(iv). □

Lemma A.4. *Under Condition* (A.2)(ii), *define* $\delta_n=o(n^{-1/\gamma}/\sqrt{r})$ *and* $\Lambda_n=\{\boldsymbol{\lambda}\in\mathbb{R}^r:\|\boldsymbol{\lambda}\|_2\le\delta_n\}$, *we have*

$$\sup_{t\in\{1,\ldots,n\},\,\boldsymbol{\theta}\in\boldsymbol{\Theta},\,\boldsymbol{\lambda}\in\Lambda_n}|\boldsymbol{\lambda}^\top m_t(\boldsymbol{\theta})|\xrightarrow{p}0.$$

Also, with probability approaching 1, $\Lambda_n\subset\hat{\Lambda}_n(\boldsymbol{\theta})$ *for all* $\boldsymbol{\theta}\in\boldsymbol{\Theta}$.

Proof. From (A.2)(ii) and Markov's inequality, $\sup_{t\in\{1,\ldots,n\},\,\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{m}_t(\boldsymbol{\theta})\|_2=O_p(n^{1/\gamma}\sqrt{r})$. Then,

$$\sup_{t\in\{1,\ldots,n\},\,\boldsymbol{\theta}\in\boldsymbol{\Theta},\,\boldsymbol{\lambda}\in\Lambda_n}|\boldsymbol{\lambda}^\top\boldsymbol{m}_t(\boldsymbol{\theta})|\le\delta_n\sup_{t\in\{1,\ldots,n\},\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\boldsymbol{m}_t(\boldsymbol{\theta})\|_2\xrightarrow{p}0.$$

It also implies with probability approaching 1 that $|\boldsymbol{\lambda}^\top\boldsymbol{m}_t(\boldsymbol{\theta})|\in\nu$ for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and $\|\boldsymbol{\lambda}\|_2\le\delta_n$. □

Lemma A.5. *Under Conditions* (A.1)(i), (A.2)(i) *and* (A.2)(iii), *assume that* $\lambda_{\max}(U_h)$ *is uniformly bounded away from infinity. If* $r^2h_n^2/n=o(1)$, $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2=O_p(\tau_n)$ *and* $rph_n\tau_n^2=o(1)$, *then* $\|\hat{\Omega}(\boldsymbol{\theta})-\hat{\Omega}(\boldsymbol{\theta}_0)\|_2=O_p(\tau_n\sqrt{rp/h_n})$.

Proof. Choose $\boldsymbol{x}\in\mathbb{R}^r$ with unit L_2-norm such that $\lambda_{\max}\{\hat{\Omega}(\boldsymbol{\theta})-\hat{\Omega}(\boldsymbol{\theta}_0)\}=\boldsymbol{x}^\top\{\hat{\Omega}(\boldsymbol{\theta})-\hat{\Omega}(\boldsymbol{\theta}_0)\}\boldsymbol{x}$. Then,

$$|\lambda_{\max}\{\hat{\Omega}(\boldsymbol{\theta})-\hat{\Omega}(\boldsymbol{\theta}_0)\}|\le\frac{1}{n}\sum_{t=1}^{n}\|\boldsymbol{m}_t(\boldsymbol{\theta})-\boldsymbol{m}_t(\boldsymbol{\theta}_0)\|_2^2+2[\lambda_{\max}\{\hat{\Omega}(\boldsymbol{\theta}_0)\}]^{1/2}\left\{\frac{1}{n}\sum_{t=1}^{n}\|\boldsymbol{m}_t(\boldsymbol{\theta})-\boldsymbol{m}_t(\boldsymbol{\theta}_0)\|_2^2\right\}^{1/2}.$$

Note that $r^2h_n^2/n = o(1)$, by Lemma A.2 and $\lambda_{\max}(U_h)$ is uniformly bounded away from infinity, $\lambda_{\max}\{\hat{\Omega}(\boldsymbol{\theta}_0)\} = O_p(1/h_n)$. From (A.2)(i), $\sum_{t=1}^{n} \|\boldsymbol{m}_t(\boldsymbol{\theta}) - \boldsymbol{m}_t(\boldsymbol{\theta}_0)\|_2^2/n = rpO_p(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2)$. If $rph_n\tau_n^2 = o(1)$, then $|\lambda_{\max}\{\hat{\Omega}(\boldsymbol{\theta}) - \hat{\Omega}(\boldsymbol{\theta}_0)\}| = O_p(\tau_n\sqrt{rp/h_n})$. Using the same argument, $|\lambda_{\min}\{\hat{\Omega}(\boldsymbol{\theta}) - \hat{\Omega}(\boldsymbol{\theta}_0)\}| = O_p(\tau_n\sqrt{rp/h_n})$, This completes the argument. □

Lemma A.6. *Under Conditions* (A.1)(i), (A.1)(ii), (A.2)(i)–(A.2)(iii), *assume that the eigenvalues of* U_h *are uniformly bounded away from zero and infinity. If* $h_n/\sqrt{n} = o(1)$, $rh_npτ_n^2 = o(1)$, $r^2h_n^2n^{2/\gamma-1} = o(1)$, $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 = O_p(\tau_n)$, *and* $\|\bar{\boldsymbol{g}}(\tilde{\boldsymbol{\theta}})\|_2 = O_p(\sqrt{r/n})$, *then*

$$\hat{\boldsymbol{\lambda}}(\tilde{\boldsymbol{\theta}}) = \arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\tilde{\boldsymbol{\theta}})} \hat{S}_n(\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda})$$

exists, $\sup_{\boldsymbol{\lambda}\in\hat{\Lambda}_n} \hat{S}_n(\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}) = \rho(0) + O_p(rh_n/n)$ *and* $\|\hat{\boldsymbol{\lambda}}(\tilde{\boldsymbol{\theta}})\|_2 = O_p(h_n\sqrt{r/n})$, *where* $\hat{S}_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$ *is defined in* (2).

Proof. Pick $\delta_n = o(n^{-1/\gamma}/\sqrt{r})$ and $h_n/\sqrt{r/n} = o(\delta_n)$, which is guaranteed by $r^2h_n^2n^{2/\gamma-1=o(1)}$. From Lemma A.1 and the triangle inequality, $\|\bar{\boldsymbol{m}}(\tilde{\boldsymbol{\theta}})\|_2 \le \|\bar{\boldsymbol{g}}(\tilde{\boldsymbol{\theta}})\|_2 + O_p(\sqrt{r}\, h_n/n) = O_p(\sqrt{r/n})$. Let $\bar{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\Lambda_n} \hat{S}_n(\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda})$, where Λ_n is defined in Lemma A.4. By Lemmas A.2, A.4 and A.5, noting $\rho_{\nu\nu}(0) < 0$,

$$\begin{aligned}\rho(0) = \hat{S}_n(\tilde{\boldsymbol{\theta}}, \mathbf{0}) \le \hat{S}_n(\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{\lambda}}) &= \rho(0) + \rho_\nu(0)\bar{\boldsymbol{\lambda}}^\top\bar{\boldsymbol{m}}(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}\bar{\boldsymbol{\lambda}}^\top\left[\frac{1}{n}\sum_{t=1}^{n}\rho_{\nu\nu}\{\dot{\boldsymbol{\lambda}}^\top\boldsymbol{m}_t(\tilde{\boldsymbol{\theta}})\}\boldsymbol{m}_t(\tilde{\boldsymbol{\theta}})\boldsymbol{m}_t(\tilde{\boldsymbol{\theta}})^\top\right]\bar{\boldsymbol{\lambda}} \\ &\le \rho(0) + |\rho_\nu(0)|\|\bar{\boldsymbol{\lambda}}\|_2\|\bar{\boldsymbol{m}}(\tilde{\boldsymbol{\theta}})\|_2 - C\|\bar{\boldsymbol{\lambda}}\|_2^2\{1/h_n + o(1/h_n)\},\end{aligned}$$

where $\dot{\boldsymbol{\lambda}}$ lies on between $\mathbf{0}$ and $\bar{\boldsymbol{\lambda}}$. Hence, $\|\bar{\boldsymbol{\lambda}}\|^2 \le Ch_n\|\bar{\boldsymbol{m}}(\tilde{\boldsymbol{\theta}})\|_2\{1 + o_p(1)\} = O_p(h_n\sqrt{r/n}) = o(\delta_n)$. Thus $\bar{\boldsymbol{\lambda}} \in \operatorname{int}(\Lambda_n)$ with probability approaching 1. Since $\Lambda_n \subset \hat{\Lambda}_n(\tilde{\boldsymbol{\theta}})$ with probability approaching 1, $\hat{\boldsymbol{\lambda}}(\tilde{\boldsymbol{\theta}}) = \bar{\boldsymbol{\lambda}}$ with probability approaching 1 by the concavity of $\hat{S}_n(\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda})$ and $\hat{\Lambda}_n(\tilde{\boldsymbol{\theta}})$. Then,

$$\hat{S}_n\{\tilde{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}}(\tilde{\boldsymbol{\theta}})\} \le \rho(0) + |\rho_\nu(0)|\|\hat{\boldsymbol{\lambda}}(\tilde{\boldsymbol{\theta}})\|_2\|\bar{\boldsymbol{m}}(\tilde{\boldsymbol{\theta}})\|_2 - Ch_n^{-1}\|\hat{\boldsymbol{\lambda}}(\tilde{\boldsymbol{\theta}})\|_2^2\{1 + o_p(1)\}$$

leads to $\sup_{\boldsymbol{\lambda}\in\hat{\Lambda}_n} \hat{S}_n(\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}) = \rho(0) + O_p(rh_n/n)$. □

A.2. Proof of Theorem 1

Choose $\delta_n = o(n^{-1/\gamma}/\sqrt{r})$ and $h_n\sqrt{r/n} = o(\delta_n)$. Let $\bar{\boldsymbol{\lambda}} = \operatorname{sign}\{\rho_\nu(0)\}\delta_n\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)/\|\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)\|_2$, then $\bar{\boldsymbol{\lambda}} \in \Lambda_n$. By a Taylor expansion, Lemmas A.3 and A.4, noting $\rho_{\nu\nu}(0) < 0$,

$$\begin{aligned}\hat{S}_n(\hat{\boldsymbol{\theta}}_n, \bar{\boldsymbol{\lambda}}) &= \rho(0) + \rho_\nu(0)\bar{\boldsymbol{\lambda}}^\top\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n) + \frac{1}{2}\bar{\boldsymbol{\lambda}}^\top\left[\frac{1}{n}\sum_{t=1}^{n}\rho_{\nu\nu}\{\dot{\boldsymbol{\lambda}}^\top\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)^\top\right]\bar{\boldsymbol{\lambda}} \\ &\ge \rho(0) + |\rho_\nu(0)|\delta_n\|\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)\|_2 - CO_p(1)\|\bar{\boldsymbol{\lambda}}\|_2^2.\end{aligned}$$

Meanwhile, in the same way in the proof of Lemma A.2, $\|\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0) - \mathrm{E}\{\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}\|_2 = O_p(\sqrt{r/n})$. Since $\mathrm{E}\{\boldsymbol{g}_t(\boldsymbol{\theta}_0)\} = \mathbf{0}$, $\|\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0)\|_2 = O_p(\sqrt{r/n})$. Then, from Lemma A.6,

$$\hat{S}_n(\hat{\boldsymbol{\theta}}_n, \bar{\boldsymbol{\lambda}}) \le \sup_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\hat{\boldsymbol{\theta}}_n)} \hat{S}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\lambda}) \le \sup_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta}_0)} \hat{S}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}) = \rho(0) + O_p(rh_n/n).$$

Hence, $\|\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)\|_2 = O_p(\delta_n)$. Consider any $\varepsilon_n \to 0$, and let $\tilde{\boldsymbol{\lambda}} = \operatorname{sign}\{\rho_\nu(0)\}\varepsilon_n\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)$, then $\|\tilde{\boldsymbol{\lambda}}\|_2 = o_p(\delta_n)$. Using the same arguments given above, we can obtain

$$|\rho_\nu(0)|\varepsilon_n\|\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)\|_2^2 - CO_p(1)\varepsilon_n^2\|\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)\|_2^2 = O_p(rh_n/n).$$

Then, $\varepsilon_n\|\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)\|_2^2 = O_p(rh_n/n)$. Thus, $\|\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)\|_2^2 = O_p(rh_n/n)$.

From Lemma A.1, $\|\bar{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_n)\|_2 = O_p(\sqrt{rh_n/n})$. If $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2$ does not converge to zero in probability, then there exists a subsequence $\{(n_*, h_{n*}, r_*, p_*)\}$ such that $\|\hat{\boldsymbol{\theta}}_{n*} - \boldsymbol{\theta}_0\|_2 \ge \varepsilon$ almost surely for some positive constant ε. By (A.1)(iii), $\|\mathrm{E}\{\boldsymbol{g}_t(\hat{\boldsymbol{\theta}}_{n*})\}\|_2 = o_p\{\triangle_1(r_*p_*)\} + O_p(\sqrt{r_*h_{n*}/n_*})$. Furthermore, from (A.1)(ii), $\|\mathrm{E}\{\boldsymbol{g}_t(\hat{\boldsymbol{\theta}}_{n*})\}\|_2 \ge \triangle_1(r_*p_*)\triangle_2(\varepsilon)$. As $\liminf_{r,p\to\infty} \triangle_1(r, p) > 0$, it is a contradiction. Hence, $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 \xrightarrow{p} 0$.

By (A.2)(iv), $\|\bar{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_n) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_0)\|_2 \ge C\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2$ with probability approaching 1. Then, $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = O_p(\sqrt{h_n/n})$. In addition, if $r^2ph_n^2/n = o(1)$, from Lemmas A.2 and A.5, $\lambda_{\max}\{\hat{\Omega}(\hat{\boldsymbol{\theta}}_n)\} \le C/h_n$ with probability approaching 1. By repeating the above arguments, we can obtain $\|\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)\|_2 = O_p(\sqrt{r/n})$ and $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = O_p(\sqrt{r/n})$. From Lemma A.6, $\|\bar{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)\|_2 = O_p(h_n\sqrt{r/n})$. Therefore, we complete the proof of Theorem 1. □

A.3. Other subsidiary results

Lemma A.7. *Under Conditions* (A.1)–(A.2)*, assume that $\lambda_{\max}(U_h)$ is uniformly bounded away from infinity. If* (3) *holds and $r^2ph_n^2/n = o(1)$, then for any $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^r$,*

$$\left\| \frac{1}{n}\sum_{t=1}^{n} \rho_v\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{x} - \frac{1}{n}\sum_{t=1}^{n} \rho_v(0)\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{x} \right\|_2 = O_p(r\sqrt{p}\,h_n/\sqrt{n})\,\|\boldsymbol{x}\|_2,$$

and

$$\left| \frac{h_n}{n}\sum_{t=1}^{n} \boldsymbol{y}^\top \rho_{vv}\{\tilde{\boldsymbol{\lambda}}^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{z} - \frac{h_n}{n}\sum_{t=1}^{n} \rho_{vv}(0)\boldsymbol{y}^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{z} \right| = O_p(rh_n n^{1/\gamma-1/2})\,\|\boldsymbol{y}\|_2\|\boldsymbol{z}\|_2,$$

where $\tilde{\boldsymbol{\lambda}}$ is on the line joining $\boldsymbol{0}$ and $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)$.

Proof. From Theorem 1, both $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)$ and $\tilde{\boldsymbol{\lambda}}$ are of order $O_p(h_n\sqrt{r/n}) = o(\delta_n)$, where δ_n is defined in Lemma A.4. By a Taylor expansion and the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{t=1}^{n} \rho_v\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{x} - \frac{1}{n}\sum_{t=1}^{n} \rho_v(0)\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{x} \right\|_2^2 \\
&\quad \le \left[\frac{1}{n}\sum_{t=1}^{n} \rho_{vv}^2\{\dot{\boldsymbol{\lambda}}^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}^2\right] \times \left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{x}^\top\{\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}^\top\{\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\boldsymbol{x}\right],
\end{aligned}$$

where $\dot{\boldsymbol{\lambda}}$ lies on the line joining 0 and $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)$. From Lemma A.4 and $\lambda_{\max}\{\widehat{\Omega}(\hat{\boldsymbol{\theta}}_n)\} = O_p(1/h_n)$ which is implied by Lemmas A.2 and A.5, we obtain

$$\sum_{t=1}^{n} \rho_{vv}^2\{\dot{\boldsymbol{\lambda}}^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}^2 \le C\sum_{t=1}^{n}\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}^2\{1+o_p(1)\} = O_p(rh_n/n).$$

Furthermore,

$$\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{x}^\top\{\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}^\top\{\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\boldsymbol{x} \le \frac{2h_n+1}{n}\sum_{t=1}^{n}\sum_{s=-h_n}^{h_n} \kappa^2(s)\mathrm{E}\|\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_{t-s}(\hat{\boldsymbol{\theta}}_n)\boldsymbol{x}\|_2^2 = O_p(h_n rp)\|\boldsymbol{x}\|_2^2.$$

Hence, we obtain the first result. Using the same arguments, we can get the second result. □

Lemma A.8. *Under Conditions* (A.1)(i), (A.1)(ii) *and* (A.3)*, $\|\nabla_{\boldsymbol{\theta}}\bar{\boldsymbol{m}}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}\bar{\boldsymbol{m}}(\boldsymbol{\theta}^*)\|_F = O_p(\sqrt{r}\,p)\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2$ for any $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$ in a neighborhood of $\boldsymbol{\theta}_0$, and $\|\nabla_{\boldsymbol{\theta}}\bar{\boldsymbol{m}}(\boldsymbol{\theta}_0) - \mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}\|_F = O_p(\sqrt{rp/n})$ provided that $h_n = o(1/\sqrt{n})$.*

Proof. Using the Taylor expansion and noting (A.3), the first conclusion holds. Using the same method in the proof of Lemma A.2, $\|\nabla_{\boldsymbol{\theta}}\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0) - \mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}\|_F = O_p(\sqrt{rp/n})$. By the same way in the proof of Lemma A.1, $\|\nabla_{\boldsymbol{\theta}}\bar{\boldsymbol{m}}(\boldsymbol{\theta}_0) - \nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\|_F = O_p(\sqrt{rp}\,h_n/n)$. Hence, by the triangle inequality, we can obtain the second result. □

Proposition A.1. *Under Conditions* (A.1)–(A.3)*, assume that the eigenvalues of V_n and U_h are uniformly bounded away from zero and infinity, if $r^2ph_n^2/n = o(1)$ and* (3) *holds, then for any vector $\boldsymbol{\alpha}_n \in \mathbb{R}^p$ with unit L_2 norm,*

$$\begin{aligned}
&\sqrt{n}\,\boldsymbol{\alpha}_n^\top\Big[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\Big]^{-1/2}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}](\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) \\
&\quad = -\sqrt{n}\boldsymbol{\alpha}_n^\top\Big[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\Big]^{-1/2}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0) \\
&\qquad + O_p(r^{3/2}h_n n^{1/\gamma-1/2}) + O_p(r^{3/2}p/\sqrt{n}) + O_p(r^{3/2}p^{1/2}h_n/\sqrt{n}).
\end{aligned}$$

Proof. Define $\boldsymbol{\beta} = \Big[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\Big]^{-1/2}\boldsymbol{\alpha}_n$. Then, we have

$$\begin{aligned}
\|\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}\boldsymbol{\beta}\|_2^2 &= \boldsymbol{\alpha}_n^\top(U^\top U)^{-1/2}U^\top V_n^{-1/2}U_h^2V_n^{-1/2}U(U^\top U)^{-1/2}\boldsymbol{\alpha}_n \\
&\le \lambda_{\max}(V_n^{-1/2}U_h^2V_n^{-1/2})\|U(U^\top U)^{-1/2}\boldsymbol{\alpha}_n\|_2^2 = \lambda_{\max}^2(U_h)\lambda_{\min}^{-1}(V_n),
\end{aligned}$$

where $U = V_n^{1/2}U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]$. Therefore, $\|\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}\boldsymbol{\beta}\|_2 = o(1)$. Meanwhile,

$$\begin{aligned}
\|\boldsymbol{\beta}\|_2^2 &\le \lambda_{\min}^{-1}\Big[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\Big] \\
&\le \lambda_{\max}^2(U_h)\lambda_{\min}^{-1}([\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}])\lambda_{\min}^{-1}(V_n).
\end{aligned}$$

Hence, $\|\boldsymbol{\beta}\|_2^2 \le C$. From Lemma A.4, we obtain

$$\frac{h_n}{n}\sum_{t=1}^{n} \rho_{vv}\{\tilde{\boldsymbol{\lambda}}^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)^\top = \rho_{vv}(0)h_n\hat{\Omega}(\hat{\boldsymbol{\theta}}_n)\{1+o_p(1)\}.$$

From Lemmas A.3 and A.6, we know that the eigenvalues of $h_n\hat{\Omega}(\hat{\boldsymbol{\theta}}_n)$ are uniformly bounded away from zero and infinity with probability approaching 1. Hence, the eigenvalues of $h_n\sum_{t=1}^{n}\rho_{vv}\{\tilde{\boldsymbol{\lambda}}^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)^\top/n$ are uniformly bounded away from zero and infinity with probability approaching 1. By Lemma A.7,

$$\boldsymbol{\beta}^\top\nabla_{\boldsymbol{\theta}}\{\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)\}^\top\left[\frac{h_n}{n}\sum_{t=1}^{n}\rho_{vv}\{\tilde{\boldsymbol{\lambda}}^\top \boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\}\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)\boldsymbol{m}_t(\hat{\boldsymbol{\theta}}_n)^\top\right]^{-1}\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n) = O_p(r^{3/2}p^{1/2}h_n/n).$$

From Lemmas A.7 and A.8,

$$\boldsymbol{\beta}^\top[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top\{h_n\hat{\Omega}(\hat{\boldsymbol{\theta}}_n)\}^{-1}\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n) = O_p(r^{3/2}p^{1/2}h_n/n) + O_p(r^{3/2}h_n n^{1/\gamma-1}) + O_p(r^{3/2}p/n).$$

Note that by Lemmas A.2 and A.5,

$$\boldsymbol{\beta}^\top[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n) = O_p(r^{3/2}h_n n^{1/\gamma-1}) + O_p(r^{3/2}p/n) + O_p(r^{3/2}p^{1/2}h_n/n).$$

Expanding $\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n)$ around $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, by Lemmas A.7 and A.1,

$$\begin{aligned}&\boldsymbol{\beta}^\top[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}](\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)\\ &\quad= -\boldsymbol{\beta}^\top[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0) + O_p(r^{3/2}h_n n^{1/\gamma-1}) + O_p(r^{3/2}p/n) + O_p(r^{3/2}p^{1/2}h_n/n).\end{aligned}$$

Hence, we obtain Proposition A.1. □

A.4. Proof of Theorem 2

From Proposition A.1, we only need to show that, as $n\to\infty$,

$$S_n \equiv -\sqrt{n}\boldsymbol{\alpha}_n^\top\left[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\right]^{-1/2}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0) \rightsquigarrow \mathcal{N}(0,1).$$

Let

$$\boldsymbol{x}_{n,t} = -\boldsymbol{\alpha}_n^\top\left[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\right]^{-1/2}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}\boldsymbol{g}_t(\boldsymbol{\theta}_0) \equiv \boldsymbol{\beta}_n^\top\boldsymbol{g}_t(\boldsymbol{\theta}_0).$$

Then $S_n = (\boldsymbol{x}_{n,1}+\cdots+\boldsymbol{x}_{n,n})/\sqrt{n}$. As restriction (4) holds, $\sup_n\sup_{t\in\{1,\ldots,n\}}\mathrm{E}\{|\boldsymbol{x}_{n,t}|^\gamma\} < \infty$. Moreover, $\mathrm{var}(S_n) = 1$. From (A.1)(i) and the Central Limit Theorem proposed in [18], we have $S_n \rightsquigarrow \mathcal{N}(0,1)$ as $n\to\infty$, as claimed. □

Let

$$\hat{S}_n^{(pe)}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \frac{1}{n}\sum_{t=1}^{n}\rho\{\boldsymbol{\lambda}^\top\boldsymbol{m}_t(\boldsymbol{\theta})\} + \sum_{j=1}^{p}p_\tau(|\theta_j|)$$

for any $\boldsymbol{\theta}\in\Theta$ and $\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})$. Then, $\hat{\boldsymbol{\theta}}_n^{(pe)} = \arg\min_{\boldsymbol{\theta}\in\Theta}\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})}\hat{S}_n^{(pe)}(\boldsymbol{\theta},\boldsymbol{\lambda})$ and $\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta}\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})}\hat{S}_n(\boldsymbol{\theta},\boldsymbol{\lambda})$. The following lemma will be used to establish Theorem 3.

Lemma A.9. *Under Conditions* (A.1), (A.2) *and* (A.5), *assume that the eigenvalues of U_h are uniformly bounded away from zero and infinity. If* (3) *holds, $r^2ph_n^2/n = o(1)$ and $s\tau n/(rh_n) = O(1)$, then $\|\hat{\boldsymbol{\theta}}_n^{(pe)} - \boldsymbol{\theta}_0\|_2 = O_p(\sqrt{r/n})$.*

Proof. Choose $\delta_n = o(n^{-1/\gamma}/\sqrt{r})$ and $h_n\sqrt{r/n} = o(\delta_n)$. Let $\bar{\boldsymbol{\lambda}} = \mathrm{sign}\{\rho_v(0)\}\delta_n\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n^{(pe)})/\|\bar{\boldsymbol{m}}(\hat{\boldsymbol{\theta}}_n^{(pe)})\|_2$. Then $\bar{\boldsymbol{\lambda}}\in\Lambda_n$, where Λ_n is defined in Lemma A.5. By a Taylor expansion, Lemmas A.3 and A.4, noting that $\rho_{vv} < 0$, we have

$$\begin{aligned}\hat{S}_n\{\hat{\boldsymbol{\theta}}_n^{(pe)},\bar{\boldsymbol{\lambda}}\} &= \rho(0) + \rho_v(0)\bar{\boldsymbol{\lambda}}^\top\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\} + \frac{1}{2}\bar{\boldsymbol{\lambda}}^\top\left[\frac{1}{n}\sum_{t=1}^{n}\rho_{vv}[\dot{\boldsymbol{\lambda}}^\top\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}]\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}^\top\right]\bar{\boldsymbol{\lambda}}\\ &\ge \rho(0) + |\rho_v(0)|\delta_n\|\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}\|_2 - C\|\bar{\boldsymbol{\lambda}}\|_2^2O_p(1).\end{aligned}$$

Furthermore,

$$\hat{S}_n^{(pe)}\{\hat{\boldsymbol{\theta}}_n^{(pe)},\bar{\boldsymbol{\lambda}}\} \le \sup_{\boldsymbol{\lambda}\in\hat{\Lambda}_n\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}}\hat{S}_n^{(pe)}\{\hat{\boldsymbol{\theta}}_n^{(pe)},\boldsymbol{\lambda}\} \le \sup_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta}_0)}\hat{S}_n^{(pe)}(\boldsymbol{\theta}_0,\boldsymbol{\lambda}).$$

By Lemma A.6 and (A.5), as $s\tau r^{-1}h_n^{-1}n = O(1)$,

$$\sup_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta}_0)} \hat{S}_n^{(pe)}(\boldsymbol{\theta}_0, \boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta}_0)} \hat{S}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}) + \sum_{j=1}^{p} p_\tau(|\theta_{0j}|) = \rho(0) + O_p(rh_n/n + s\tau) = \rho(0) + O_p(rh_n/n).$$

Note that $\hat{S}_n^{(pe)}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \geq \hat{S}_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$ for any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\lambda} \in \hat{\Lambda}_n(\boldsymbol{\theta})$. This yields $\|\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}\|_2 = O_p(\delta_n)$. Consider any $\varepsilon_n \to 0$ and let $\tilde{\boldsymbol{\lambda}} = \mathrm{sign}\{\rho_v(0)\}\varepsilon_n \bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}$. Then $\|\tilde{\boldsymbol{\lambda}}\|_2 = o_p(\delta_n)$. Using the same procedure above, we can obtain

$$|\rho_v(0)|\varepsilon_n\|\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}\|_2^2 - O_p(1)\varepsilon_n^2\|\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}\|_2^2 = O_p(rh_n/n).$$

Then, $\varepsilon_n\|\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}\|_2^2 = O_p(rh_n/n)$. Thus, $\|\bar{\boldsymbol{m}}\{\hat{\boldsymbol{\theta}}_n^{(pe)}\}\|_2 = O_p(\sqrt{rh_n/n})$. Following the same arguments given in the proof of Theorem 1, we can obtain $\|\hat{\boldsymbol{\theta}}_n^{(pe)} - \boldsymbol{\theta}_0\|_2 = O_p(\sqrt{r/n})$. □

A.5. Proof of Theorem 3

Note that $\hat{\boldsymbol{\theta}}_n^{(pe)}$ and its Lagrange multiplier $\hat{\boldsymbol{\lambda}}^{(pe)}$ satisfy the score equation

$$\mathbf{0} = \nabla_{\boldsymbol{\lambda}}\hat{S}_n^{(pe)}\{\hat{\boldsymbol{\theta}}_n^{(pe)}, \hat{\boldsymbol{\lambda}}^{(pe)}\} = \nabla_{\boldsymbol{\lambda}}\hat{S}_n\{\hat{\boldsymbol{\theta}}_n^{(pe)}, \hat{\boldsymbol{\lambda}}^{(pe)}\}.$$

By the implicit function theorem as given, e.g., in Theorem 9.28 of [25], we have that for all $\boldsymbol{\theta}$ in a $\|\cdot\|_2$-neighborhood of $\hat{\boldsymbol{\theta}}_n^{(pe)}$, there exists a $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ such that $\nabla_{\boldsymbol{\lambda}}\hat{S}_n^{(pe)}\{\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\} = \mathbf{0}$, and $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$. By the concavity of $\hat{S}_n^{(pe)}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$, $\hat{S}_n^{(pe)}\{\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\} = \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})} \hat{S}_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$. From the envelop theorem,

$$\mathbf{0} = \nabla_{\boldsymbol{\theta}}\hat{\mathbf{S}}_{\mathbf{n}}^{(\mathbf{pe})}\{\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\} \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{pe})}} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{t}=\mathbf{1}}^{\mathbf{n}} \rho_v[\hat{\boldsymbol{\lambda}}\{\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{pe})}\}^\top \boldsymbol{m}_{\mathbf{t}}\{\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{pe})}\}][\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_{\mathbf{t}}\{\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{pe})}\}]^\top\hat{\boldsymbol{\lambda}}\{\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{pe})}\} + \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{p}} \mathbf{p}_\tau(|\theta_{\mathbf{j}}|) \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathbf{n}}^{(\mathbf{pe})}} .$$

For any $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 = O_p(\sqrt{r/n})$ and $\|\bar{\mathbf{g}}(\boldsymbol{\theta})\|_2 = O_p(\sqrt{r/n})$, define

$$\mathbf{h}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \rho_v\{\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})^\top \boldsymbol{m}_t(\boldsymbol{\theta})\}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{m}_t(\boldsymbol{\theta})\}^\top\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}) + \sum_{j=1}^{p} p_\tau(|\theta_j|).$$

Write $\mathbf{h}(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), \ldots, h_p(\boldsymbol{\theta}))^\top$. From Lemma A.6, we have that $\|\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\|_2 = O_p(h_n\sqrt{r/n})$, which implies that $\sup_{t\in\{1,\ldots,n\}}|\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})^\top \boldsymbol{m}_t(\boldsymbol{\theta})| = o_p(1)$. For each $j \in \{1, \ldots, p\}$,

$$h_j(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \rho_v(0)\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)^\top \frac{\partial}{\partial\theta_j}\boldsymbol{m}_t(\boldsymbol{\theta}_0) + \frac{1}{n}\sum_{t=1}^{n} \rho_v(0)\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)^\top \frac{\partial^2}{\partial\theta_j\partial\boldsymbol{\theta}}\boldsymbol{m}_t(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$
$$+p'_\tau(|\theta_j|)\mathrm{sign}(\theta_j) + \text{ higher order terms}.$$

From (A.4), there exists a positive constant C such that $p'_\tau(|\theta_j|) \geq C\tau$. Moreover, as $\tau\sqrt{n/r}/h_n \to \infty$,

$$\max_{j\notin\mathcal{A}}\left|\frac{1}{n}\sum_{t=1}^{n} \rho_v(0)\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)^\top \frac{\partial}{\partial\theta_j}\boldsymbol{m}_t(\boldsymbol{\theta}_0)\right| = O_p(h_n\sqrt{r/n}) = o_p(\tau).$$

Similarly, we can show

$$\max_{j\notin\mathcal{A}}\left|\frac{1}{n}\sum_{t=1}^{n} \rho_v(0)\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)^\top \frac{\partial^2}{\partial\theta_j\partial\boldsymbol{\theta}}\boldsymbol{m}_t(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right| = o_p(\tau).$$

Hence, $p'_\tau(|\theta_j|)\mathrm{sign}(\theta_j)$ dominates the sign of $h_j(\boldsymbol{\theta})$ uniformly for all $j \notin \mathcal{A}$. If $\hat{\boldsymbol{\theta}}_n^{(2)} \neq \mathbf{0}$, there exists some $j \notin \mathcal{A}$ such that $\hat{\theta}_{n,j} \neq 0$. From the above arguments, we find that $\Pr[h_j\{\hat{\boldsymbol{\theta}}_n^{(pe)}\} \neq 0] \to 1$. It is a contradiction. Hence, $\hat{\boldsymbol{\theta}}_n^{(2)} = \mathbf{0}$.

Next, we consider the second result. From (6), we deduce that

$$[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\big[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\big]^{-1}$$
$$\times[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}] = \begin{pmatrix} (\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1} & * \\ * & * \end{pmatrix}.$$

Let

$$[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\big[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1}V_nU_h^{-1}[\mathrm{E}\{\nabla_{\boldsymbol{\theta}}\boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]\big]^{-1/2} = \begin{pmatrix} \mathbf{U} & \mathbf{V} \\ \mathbf{V}^\top & * \end{pmatrix}.$$

where $\mathbf{U}$ is an $s \times s$ symmetric matrix. Then $\mathbf{U}\mathbf{U}^\top + \mathbf{V}\mathbf{V}^\top = (\boldsymbol{S}_{11} - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21})^{-1}$. For any $\boldsymbol{\alpha}_n \in \mathbb{R}^s$ such that $\|\boldsymbol{\alpha}_n\|_2 = 1$, define

$$\tilde{\boldsymbol{\alpha}}_n = \begin{pmatrix} \mathbf{U}^\top \\ \mathbf{V}^\top \end{pmatrix} (\boldsymbol{S}_{11} - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21})^{1/2}\boldsymbol{\alpha}_n.$$

Then, we have

$$\tilde{\boldsymbol{\alpha}}_n^\top \tilde{\boldsymbol{\alpha}}_n = \boldsymbol{\alpha}_n^\top (\boldsymbol{S}_{11} - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21})^{1/2}(\mathbf{U}\mathbf{U}^\top + \mathbf{V}\mathbf{V}^\top)(\boldsymbol{S}_{11} - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21})^{1/2}\boldsymbol{\alpha}_n = 1.$$

Following the same arguments of Proposition A.1, we know it still holds for $\hat{\boldsymbol{\theta}}_n^{(pe)}$. Note that

$$\begin{aligned} &\tilde{\boldsymbol{\alpha}}_n^\top \left[[\mathrm{E}\{\nabla_{\boldsymbol{\theta}} \boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1} V_n U_h^{-1} [\mathrm{E}\{\nabla_{\boldsymbol{\theta}} \boldsymbol{g}_t(\boldsymbol{\theta}_0)\}] \right]^{-1/2} [\mathrm{E}\{\nabla_{\boldsymbol{\theta}} \boldsymbol{g}_t(\boldsymbol{\theta}_0)\}]^\top U_h^{-1} [\mathrm{E}\{\nabla_{\boldsymbol{\theta}} \boldsymbol{g}_t(\boldsymbol{\theta}_0)\}] \{\hat{\boldsymbol{\theta}}_n^{(pe)} - \boldsymbol{\theta}_0^{(1)}\} \\ &\quad = \tilde{\boldsymbol{\alpha}}_n^\top (\boldsymbol{S}_{11} - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21})^{-1/2} \{\hat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}\}. \end{aligned}$$

We can then establish the second result following Proposition A.1. This concludes the argument. □

References

[1] S. Anatolyev, GMM, GEL, serial correlation, and asymptotic bias, Econometrica 73 (2005) 983–1002.

[2] D.W. Andrews, Heteroskedasticity and autocorrelation consistent covariance matrix estimation, Econometrica (1991) 817–858.

[3] J. Chang, S.X. Chen, X. Chen, High dimensional generalized empirical likelihood for moment restrictions with dependent data, J. Econometrics 185 (2015) 283–304.

[4] J. Chang, C.Y. Tang, T.T. Wu, A new scope of penalized empirical likelihood with high-dimensional estimating equations, Ann. Statist. 46 (2018) 3185–3216.

[5] P. Chaussé, Computing generalized method of moments and generalized empirical likelihood with R, J. Statist. Softw. 34 (2010) 1–35.

[6] S.X. Chen, H. Cui, On bartlett correction of empirical likelihood in the presence of nuisance parameters, Biometrika 93 (2006) 215–220.

[7] S.X. Chen, H. Cui, On the second-order properties of empirical likelihood with moment restrictions, J. Econometrics 141 (2007) 492–516.

[8] S.X. Chen, L. Peng, Y.L. Qin, Effects of data dimension on empirical likelihood, Biometrika 96 (2009) 711–722.

[9] S.X. Chen, I. Van Keilegom, A review on empirical likelihood methods for regression, Test 18 (2009) 415–447.

[10] P. Doukhan, Mixing: Properties and Examples, Springer, New York, 1994.

[11] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.

[12] L.P. Hansen, K.J. Singleton, Generalized instrumental variables estimation of nonlinear rational expectations models, Econometrica (1982) 1269–1286.

[13] N.L. Hjort, I.W. McKeague, I. Van Keilegom, Extending the scope of empirical likelihood, Ann. Statist. 37 (2009) 1079–1111.

[14] Y. Kitamura, M. Stutzer, An information-theoretic alternative to generalized method of moments estimation, Econometrica (1997) 861–874.

[15] H.R. Künsch, The jackknife and the bootstrap for general stationary observations, Ann. Statist. (1989) 1217–1241.

[16] S.N. Lahiri, S. Mukhopadhyay, A penalized empirical likelihood method in high dimensions, Ann. Statist. 40 (2012) 2511–2540.

[17] C. Leng, C.Y. Tang, Penalized empirical likelihood and growing dimensional general estimating equations, Biometrika 99 (2012) 703–716.

[18] J.A. Nelder, R. Mead, A simplex method for function minimization, Comput. J. 7 (1965) 308–313.

[19] W.K. Newey, R.J. Smith, Higher order properties of GMM and generalized empirical likelihood estimators, Econometrica 72 (2004) 219–255.

[20] A.B. Owen, Empirical likelihood ratio confidence intervals for a single functional, Biometrika 75 (1988) 237–249.

[21] A.B. Owen, Empirical likelihood ratio confidence regions, Ann. Statist. (1990) 90–120.

[22] A.B. Owen, Empirical Likelihood, Chapman and Hall/CRC, London, 2001.

[23] D.N. Politis, J.P. Romano, The stationary bootstrap, J. Amer. Statist. Assoc. 89 (1994) 1303–1313.

[24] J. Qin, J.F. Lawless, Empirical likelihood and general estimating equations, Ann. Statist. 22 (1994) 300–325.

[25] W.R. Rudin, Principles of Mathematical Analysis, Vol. 3, McGraw-Hill, New York, 1976.

[26] R.J. Smith, Alternative semi-parametric likelihood approaches to generalised method of moments estimation, Economic J. 107 (1997) 503–519.

[27] R.J. Smith, GEL criteria for moment condition models, Technical Report, CEMMAP Working Paper, Centre for Microdata Methods and Practice, 2004.

[28] C.Y. Tang, C. Leng, Penalized high-dimensional empirical likelihood, Biometrika 97 (2010) 905–920.

[29] H. Wang, B. Li, C. Leng, Shrinkage tuning parameter selection with a diverging number of parameters, J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2009) 671–683.

[30] C.H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2010) 894–942.