Essays on Statistical and Machine Learning Methods for Dependent Data Analysis

学生番号 202081 **氏名** Ziyan Zhao

指導教員名 Hiroyuki Sano
Qingfeng Liu
Yongki Kim

２０２２年度提出

Essays on Statistical and Machine Learning Methods for Dependent Data Analysis

Ziyan Zhao

Otaru University of Commerce

January 19, 2023

Abstract

This dissertation proposes several novel statistical and machine learning methods, derives in-depth theoretical results, and conducts extensive empirical studies for dependent data analysis.

First, we propose a new factor model—time varying structural approximate dynamic factor model—by introducing time varying parameters into the classical approximate dynamic factor model, so that it can capture complex dynamic economic characteristics. Second, we propose a new estimation method—tying maximum likelihood estimation—using the parameter tying technique in Few-shot Learning to improve the performance of statistical and econometric models where most time series have long sample periods, while the other time series only have a few observations. Lastly, we provide new empirical insights into the impact of the COVID-19 pandemic on the consumer price index using a difference-in-difference approach; in addition, our static and dynamic empirical framework provides a valuable reference for other similar studies.

Preface

I thoroughly enjoyed my three years of studying for P.h.D in modern commerce at the Otaru University of Commerce, and there are many people whom I have to thank.

I would like to sincerely thank my three supervisors: professor Qingfeng Liu, professor Hiroyuki Sano, and professor Yongki Kim. Professor Qingfeng Liu taught and helped me a lot during my doctoral career, which actually went far beyond academic guidance. His kindness, patience, and conscientiousness not only shaped my academic ability but also encouraged me to keep moving forward in the field of econometrics, statistics, and machine learning. One day I hope to be a supervisor to others as he was to me. In addition, I learned a lot about industrial organization following professor Hiroyuki Sano, which both helps me understand economic markets in our reality better and benefits my study in the future. Professor Sano always gave me a detailed explanation for any question I asked, and his affinity, earnestness, and patience in teaching aroused my interest in game theory. Moreover, professor Yongki Kim always provided deep insight into the embedded corporation and interesting explanations for my questions in each class, which benefits me a lot. His humor, kindness, and attentiveness in teaching made it easy for me to understand human resource management and labor relations even if this is my first time to touch this field.

I am grateful to my coauthors and all teachers in the classes I attended. To professor Masamune Iwasawa, for his many useful comments, mathematical derivation, and detailed explanations about our study. To professor Tomohiko Kobayashi, for his helpful suggestions in studying business law and interesting explanations in our discussion. To professor Susumu Egashira, for his sincere help and interesting teaching about business and economic institution in the class.

I have to thank all members of the doctoral dissertation supervisory committee for their helpful comments and suggestions. In addition, I wish to appreciate all staff who helped me at the Otaru University of Commerce. Furthermore, I am also thankful for the financial support provided to me by the MEXT and Otaru University of Commerce.

Some chapters in this dissertation also benefited greatly from the comments and suggestions of many people and scientific research funds. Chapter 1: I would like to thank Florian Huber, Yasumasa Matsuda, Yang Feng, and Qihui Chen in the Asian meeting of the econometric society in China 2022 for their helpful comments, and acknowledge the financial support of the Japan Society for the Promotion of Science through KAKENHI Grant No. JP19K01582 (Liu) and the Nomura Foundation for Social Science Grant No. N21-3-E30-010 (Liu). Chapter 3: I have to thank professor Qingfeng Liu and two anonymous referees for their valuable comments and suggestions.

I am always deeply indebted to my grandmother, parents, and my girlfriend Siping Wang for their support and encouragement, and greatly miss my grandfather for what he taught me. In addition, I also thank my relatives and friends who helped me in my life.

To my grandmother, parents, Siping, and in memory of my grandfather.

Contents

Chapter 1

Introduction

This dissertation mainly consists of three independent chapters surrounding statistical and machine learning methods developed to solve some problems in dependent data analysis.

In the second chapter, which is co-authored with Qingfeng Liu (see SSRN working paper, Zhao and Liu, 2021)[1], we propose a time-varying structural approximate dynamic factor (TVS-ADF) model by extending the ADF model in state-space form. The TVS-ADF model considers time-varying coefficients and a time-varying variance–covariance matrix of its innovation terms, so that it can capture complex dynamic economic characteristics. We also propose an effective Markov chain Monte Carlo (MCMC) algorithm to estimate the TVS-ADF. To avoid the overparameterization caused by the time-varying characteristics of the TVS-ADF, we include the shrinkage and sparsification approaches in the MCMC algorithm. Extensive artificial simulations demonstrate that the TVS-ADF has better forecast performance than the ADF in almost all settings for different numbers of explained variables, numbers of explanatory variables, sparsity levels, and sample sizes. An empirical application to macroeconomic

[1] I undertake the main work of this research. This study was presented in the Asian meeting of the econometric society in China, 2022.

forecasting also indicates that our model can substantially improve predictive accuracy and capture the dynamic features of an economic system better than the ADF.

In the third chapter, which is co-authored with Qingfeng Liu and Masamune Iwasawa (see SSRN working paper, Iwasawa et al., 2022)[2], we propose a tying maximum likelihood estimation (TMLE) method to improve the performance of estimation of statistical and econometric models in which most time series have long sample periods, whereas the other time series are very short. The main idea of the TMLE is to tie the parameters of the long time series with those of the short time series together by introducing some restrictions on parameters so that some useful information can be transferred from the long series to the short series, which can help improve the estimation accuracy of parameters tied. We first provide asymptotic properties of the TMLE and show its finite-sample risk bound under a fixed tuning parameter which determines the strength of tying. In addition, we provide a bootstrap procedure to select the tuning parameter. Then a finite-sample theory about this bootstrap procedure is developed, which tells us how to conduct the bootstrap procedure effectively. Extensive artificial simulations and empirical applications show that the TMLE has an outstanding performance in point estimate and forecast.

In the fourth chapter, I provide empirical insight into the impact of the COVID-19 pandemic on the consumer price index (CPI) using a difference-in-difference approach (see Zhao, 2022). Using monthly panel data for eight CPI categories for China and considering two specifications (i.e., the average effect and month-by-month effect), we reveal that the pandemic had a persistent negative impact on housing and daily consumables, whereas no evidence was found for a strong effect on health care. Regarding education, culture, and recreation, the pandemic mainly had a persistent positive effect over the initial months of

[2]All authors contribute equally to this work.

the pandemic and then a negative effect for several months. In addition, the pandemic could have a positive effect on food, tobacco, and liquor, while it may have a persistent negative impact on clothing, transport, and communications. Furthermore, there could be a positive effect, which has increased slightly since the pandemic outbreak, on other articles and services.

Chapter 2

Time Varying Structural Approximate Dynamic Factor Model for Dependent Data

2.1 Introduction

Factor models have become increasingly popular in various economics and finance applications over the past two decades. For instance, latent factors can represent common shocks (e.g., technological shocks and financial crises) in macroeconometrics (e.g., Giannone and Lenza, 2010 and McCracken and Ng, 2016). Additionally, latent factors can represent the prices for unmeasured skills in microeconometrics (e.g., Cawley et al., 1997 and Carneiro et al., 2003), while in finance, they can represent unobservable factor returns (e.g., Chamberlain and Rothschild, 1982 and Zivot and Wang, 2006).

As a pioneer of the factor model, Spearman (1927) introduced an exact static factor model for analyzing independent and identically distributed (i.i.d.) data. Subsequent studies expanded the model to time series data analysis, with Geweke (1977) proposing an exact dynamic factor model and Chamberlain and Rothschild (1982) and Connor and Korajczyk (1986) an approximate static factor model. In a static factor model, the factors only exert a contemporaneous effect on the dependent variable, whereas in a dynamic factor model, the factors also affect the dependent variable through their lags. In the exact factor model, innovation (idiosyncratic component) has no cross-sectional dependence, unlike in the approximate factor model, where this is allowed. Combining the approximate and dynamic features of these models, Forni et al. (2009) proposed the approximate dynamic factor (ADF) model in state-space form, in which the dynamics of the factors capture comovements among the model variables. Moreover, the cross-sectional dependence of the innovations is permitted to reflect the impact of an innovation with respect to one variable on the other variables.

However, the ADF has a limitation, in that it does not consider the time-varying characteristics of coefficients (factor loadings) and the variance covariance matrix of the innovations, although this type of time-varying characteristics exists in many macroeconomic variables and financial time series. In an economic system, the relationships between economic variables can be time variant. Capturing the time-varying characteristics of an economic system is thus a crucial task in econometrics. Many studies have been devoted to this topic. For instance, Primiceri (2005) used time-varying parameters to measure policy changes and imply shifts in private sector behavior. Karakatsani and Bunn (2008) characterized the responses of prices to various market fundamentals using time-varying coefficients. Galí and Gambetti (2015) used time-varying parameters to analyze the response of stock prices to exogenous monetary policy shocks. Aharon and Demir (2022) used time-varying parameters to characterize the connectedness between returns for non-fungible tokens and other financial assets (i.e., equities, bonds, currencies, gold, oil, Ethereum) from January 2018 to June 2021.

Recently, vector autoregressive (VAR) models with time-varying parameters have enjoyed significant popularity in time series analysis. For example, based on the traditional VAR model proposed by Sims (1980), Cogley and Sargent (2005) and Cogley (2005) proposed a Bayesian VAR model with time-varying coefficients and variances of the innovations to capture the dynamics of economic data. However , in their models, the correlation between the elements of innovation are assumed to be time invariant. To allow for the time-varying covariance of the innovations, Primiceri (2005) proposed time-varying structural VAR (TVP-VAR). TVP-VAR can characterize the nonlinearities and time variation of both the relationships between variables and innovations. The TVP-VAR has been widely applied to time series analysis (see Koop et al. 2009; Nakajima et al. 2011; Korobilis 2013; Baumeister and Peersman 2013; Koop et al. 2019; Huber et al. 2020; Aharon and Demir 2022). The time-varying structure of the TVP-VAR, which can capture the dynamics of economic data, may be successfully applied to the ADF to address its time-invariant limitation.

In this study, we propose a new model—a time-varying structural approximate dynamic factor (TVS-ADF) model—by extending the ADF. The contributions of this study are threefold. First, we introduce a time-varying structure similar to that of the TVP-VAR into the ADF to form the TVS-ADF, which fully considers the time variations of the coefficients and the variance–covariance matrix of the innovations.

Second, we provide a Markov chain Monte Carlo (MCMC) algorithm for estimating the TVS-ADF. Although maximum likelihood estimation could be considered an alternative, the maximization of the likelihood function would be extremely difficult, if not impossible, when the dimensions of the TVS-ADF parameters are very high. The MCMC algorithm is thus a realizable choice for high-dimensional situations.

Third, we provide solutions for shrinkage and sparsification to avoid overparameterization. This is because the flexibility of the TVS-ADF arising from its time-varying characteristics comes at the cost of overparameterization, which can lead to perfect in-sample fit but poor out-of-sample forecast performance. To deal with this issue, we propose shrinkage and sparsification solutions. To shrink the TVS-ADF, we use the continuous shrinkage prior (Dirichlet–Laplace prior) proposed by Bhattacharya et al. (2015), which can be expressed as global–local scale mixtures of Gaussians and facilitate computation for high-dimensional situations. As Bhattacharya et al. (2015) pointed out, under the Bayesian paradigm, sparsity is routinely induced through two-component mixture priors with a probability mass of zero; however, such priors encounter daunting computational problems in high dimensions. Hence, we do not consider the sparsification-only approach for our TVS-ADF. As another solution, Huber et al. (2020) showed that carrying out sparsification after shrinkage can yield better predictive performance in some empirical applications. In their algorithm, the shrinkage procedure, which was conducted first, enabled them to adopt a sparsification procedure with low computation cost. We also adopt the approach of Huber et al. (2020) with both shrinkage and sparsification for our TVS-ADF.

It is worth mentioning that, although there are some studies that have developed dynamic factor models with time-varying parameters, they are different from our model. The models in these studies can be classified into two categories: parametric and semiparametric models. As for parametric models,[1] Del Negro and Otrok (2008) allowed the time-varying factor loadings and stochastic volatility of the innovations, but the variance–covariance matrix of the innovations is diagonal. Mumtaz and Surico (2012) allowed time-varying coefficients in the state equation of the factors, but the factor loadings and the variance of the innovations are time invariant. Combining the different characteristics of these two models, Bjørnland and Thorsrud (2019) proposed a new factor model in

[1] Note that we only focus on dynamic factor models in which the parameters are modeled as evolving stochastically. For positing a break in the parameters, see Stock and Watson (2016) for an overview.

which more time-varying parameters are allowed, but the variance–covariance matrix of the innovations is still diagonal. Marcellino et al. (2016) proposed a mixed frequency dynamic factor model in which the disturbances of both the factor and innovations have time-varying stochastic volatilities, but the factor loadings are time invariant. Mikkelsen et al. (2019) considered time-varying factor loadings, but the variance matrix of the innovations is both time invariant and diagonal. Based on Bjørnland and Thorsrud (2019) and Marcellino et al. (2016), Thorsrud (2020) introduced time-varying factor loadings with some restrictions, but the variance matrix of the innovations is still assumed to be both time invariant and diagonal. It is evident that all the aforementioned models assume the variance–covariance matrix of the innovations to be diagonal; in other words, innovations are not allowed to have cross-sectional dependence. As Barigozzi (2018) pointed out, the dynamic factor models in which innovations are allowed to have cross-sectional dependence are the most realistic. By contrast, our TVS-ADF not only considers the cross-sectional dependence of the innovations but also allows the variance–covariance matrix of the innovations to be time varying. Furthermore, in the models above, either the factor loadings or the variances of the innovations are time invariant. By contrast, our TVS-ADF allows both the factor loadings and the variance–covariance matrix of the innovations to be time-varying.

For semiparametric models, the main idea for capturing the dynamics of economic data is to model factor loadings as a function of time or observed variables (e.g., Motta et al., 2011; Eichler et al., 2011; Su and Wang, 2017; Ma et al., 2020; Cataño et al., 2021; Barigozzi et al., 2021; Pelger and Xiong, 2021). The main difference from our approach is that the variance–covariance matrix of the innovations in these models is specified as either diagonal or time invariant. Another drawback of these models is that semiparametric factor loadings are not easy to interpret in empirical applications.

The TVS-ADF is closely related to the class of factor augmented VAR (FAVAR) models with time-varying parameters. The models in this class have a different structure from the TVS-ADF and consist of two equations: the factor regression equation and the VAR equation. Most models in this class cannot capture all the time-varying characteristics that can be captured by our TVS-ADF. For example, the model proposed by Bianchi et al. (2009) includes the time-varying parameters in the VAR equation, but the factor loadings and the variances of innovations in the factor regression equation are time invariant. Liu et al. (2011) built a FAVAR model with time-varying parameters, in which the factor loadings are allowed to be time varying, but the variances of the innovations in the factor regression equation and the parameters in the VAR equation are time invariant. Korobilis (2013) proposed the time-varying parameter factor augmented VAR (TVP-FAVAR) model, in which the parameters in the VAR equation and the variances of the innovations in the factor regression equation are allowed to be time varying, but the factor loadings are constant. Subsequently, based on Korobilis (2013), Koop and Korobilis (2014) allowed the factor loadings to be time varying in the TVP-FAVAR. As mentioned below, this is different from our TVS-ADF, as are the other variants. All these models are different from our TVS-ADF as follows. Their variance–covariance matrices of the innovations in the factor regression equation are assumed to be diagonal. Hence, they cannot allow innovations to have cross-sectional dependence. By contrast, our TVS-ADF allows the cross-sectional dependence of the innovations and their time-varying variance–covariance matrix. Additionally, in the factor regression equations of some models above, either the factor loadings are time invariant or the innovations are time invariant. However, we allow them both to be time varying in the TVS-ADF.

The rest of the paper is organized as follows. Section 2 describes the proposed TVS-ADF model. Section 3 provides a detailed description of the estimation methodology with shrinkage and sparsification for our TVS-ADF. Section 4 conducts extensive artificial simulations. Section 5 carries out an empirical application of macroeconomic forecasting. Section 6 concludes. Further results about the artificial simulations and empirical simulation are provided in the Appendix.

2.2 Model

We construct the time-varying structural approximate dynamic factor model (TVS-ADF) as follows:

$$Y_t = B_t X_t + \Lambda_t F_t + \xi_t, \qquad \xi_t \sim N(0, \Gamma_t^{\xi}), \tag{2.1}$$

$$F_t = C_1 F_{t-1} + C_2 \eta_t, \qquad \eta_t \sim N(0, I), \tag{2.2}$$

where Y_t is an $n \times 1$ vector of explained variables, B_t is an $n \times m$ matrix of time-varying coefficients, X_t is an $m \times 1$ vector of observed explanatory variables, factor loading Λ_t is an $n \times r$ matrix, unobserved factor F_t is an $r \times 1$ vector, ξ_t is an $n \times 1$ vector of innovations allowed to have cross-sectional dependence, C_1 is an $r \times r$ matrix of coefficients, C_2 is an $r \times q$ matrix of coefficients, η_t is a $q \times 1$ vector of unobservable innovations, I is a $q \times q$ identity matrix, and Γ_t^{ξ} is an $n \times n$ positive definite matrix. Specifically,

$$Y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{pmatrix}, \qquad B_t = \begin{pmatrix} \beta'_{1t} & 0 & \cdots & 0 \\ 0 & \beta'_{2t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta'_{nt} \end{pmatrix}, \qquad X_t = \begin{pmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{nt} \end{pmatrix},$$

where β'_{it} is an $1\times m_i$ vector, x_{it} is an $m_i\times 1$ vector, $i=1,\cdots,n$, and $\sum_{i=1}^{n} m_i = m$. $\Lambda_t = (\lambda_{1t},\cdots \lambda_{nt})'$, where λ_{it} is an $r\times 1$ vector, $i=1,\cdots,n$. $F_t = (f'_t, f'_{t-1},\cdots,f'_{t-p})'$, where f_t is a $q\times 1$ vector consisting of q factors, and $q(p+1)=r$.

$$C_1 = \begin{pmatrix} c_1 & c_2 & \cdots & c_p & c_{p+1} \\ I_q & 0 & \cdots & 0 & 0 \\ 0 & I_q & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_q & 0 \end{pmatrix}, \qquad C_2 = \begin{pmatrix} I_q \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where c_i is a $q\times q$ matrix, $i=1,\cdots,p+1$, I_q is a $q\times q$ identity matrix. As Γ_t^{ξ} is positive definite, it can be factorized with Cholesky decomposition:

$$\Gamma_t^{\xi} = A_t^{-1} H_t A_t^{-1'},$$

$$A_t = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ a_{21,t} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1,t} & a_{n2,t} & \cdots & 1 \end{pmatrix}, \qquad H_t = \begin{pmatrix} h_{1t} & & & \\ & h_{2t} & & \\ & & \ddots & \\ & & & h_{nt} \end{pmatrix}.$$

This decomposition of the variance–covariance matrix above is a common technique used in time-varying models (see Cogley and Sargent, 2005, Cogley, 2005, Primiceri, 2005, Korobilis, 2013 and Koop and Korobilis, 2014). Similar to Primiceri (2005), we have the following transformed formula of (1), which is convenient for calculating the conditional distributions of A_t and H_t:

$$Y_t = B_t X_t + \Lambda_t F_t + A_t^{-1} H_t^{1/2} e_t, \qquad e_t \sim N(0, I). \tag{2.3}$$

Let $\beta_t = (\beta'_{1t},\cdots,\beta'_{nt})'$, $\lambda_t = (\lambda'_{1t},\cdots,\lambda'_{nt})'$, a_t be the vector of the non-zero and non-one elements of A_t (stacked by rows) and h_t be the vector of the diagonal elements of H_t. The dynamics of the time-varying parameters are specified as follows:

$$\beta_t = \beta_{t-1} + \bar{d}_t, \qquad \bar{d}_t \sim N(0, \bar{D}), \tag{2.4}$$

$$\lambda_t = \lambda_{t-1} + \tilde{d}_t, \qquad \tilde{d}_t \sim N(0, \tilde{D}), \tag{2.5}$$

$$a_t = a_{t-1} + u_t, \qquad u_t \sim N(0, S), \tag{2.6}$$

$$log(h_t) = log(h_{t-1}) + \gamma_t, \qquad \gamma_t \sim N(0, \Sigma), \tag{2.7}$$

where the diagonal matrices $\bar{D} = \text{diag}\left(\bar{D}_1, \cdots, \bar{D}_n\right)_{m \times m}$, $\tilde{D} = \text{diag}\left(\tilde{D}_1, \cdots, \tilde{D}_n\right)_{nr \times nr}$, for $i = 1, \cdots, n$, $\bar{D}_i$ is an $m_i \times m_i$ diagonal matrix that denotes the variance of β_{it}, $\tilde{D}_i$ is an $r \times r$ diagonal matrix that denotes the variance of λ_{it}, the diagonal matrix $S = \text{diag}\left(S_2, \cdots, S_n\right)_{\sum_{i=2}^{n}(i-1) \times \sum_{i=2}^{n}(i-1)}$, for $i = 2, \cdots, n$, S_i is a $(i-1) \times (i-1)$ diagonal matrix that denotes the variance of $(a_{i1,t}, \cdots, a_{ii-1,t})$. Note that e_t, η_t, $\bar{d}_t$, $\tilde{d}_t$, u_t, and γ_t are mutually independent, and $\bar{D}$, $\tilde{D}$, S, and Σ are all diagonal matrices. The random walk forms of equations (2.4) – (2.7) do not require any coefficient, which can reduce the number of parameters, especially for a large n compared to more general autoregressive specifications.

As previously discussed, in contrast to the ADF, our model considers additional issues related to the observed explanatory variables, time-varying coefficients, and time-varying variance–covariance matrix of the innovations. Incorporating the explanatory variables is intended to capture the impacts of known important economic variables on the explained variables other than the latent factors. Allowing for time variation in both the coefficients and variance–covariance matrix of the innovations can capture additional dynamic features of the economy. Specifically, the drifting coefficients, B_t and Λ_t, can capture time variation in the parameters or nonlinearities, which reflect the dynamic impact of the explanatory variables on the dependent variables. Furthermore, the time-varying variance–covariance matrix of multivariate stochastic volatility, Γ_t^{ξ}, can capture possible dynamic heteroscedasticity in the innovations and dynamic nonlinearities in the simultaneous relationships between model variables.

However, these time-varying parameters can lead to overparameterization. For instance, economic variables X_t may sometimes only have a very small effect on Y_t. Therefore, if we do not push β_t toward zero, it will cause poor results. Additionally, the impact of some variables on Y_t could be time invariant during a period. In such cases, it is critical to push the variance of β_t toward zero in that period; otherwise, it will cause excessive aggregate movements in β_t over time. As mentioned before, we take two approaches to shrink and sparsify the TVS-ADF to avoid overparameterization: (i) only shrink the model using the approach of Bhattacharya et al. (2015) and (ii) both shrink and sparsify the model using the approach of Huber et al. (2020). We use notation TVS-ADF(s) for the model that is only shrunk and TVS-ADF(ss) for the model that is both shrunk and sparsified. Shrinking or sparsifying our model involves the shrinkage or sparsification of β_t, λ_t, a_t, $log(h_t)$, $\bar{D}$, $\tilde{D}$, S, and Σ, that is, by shrinking the elements of these matrices toward zero or making some small elements become exactly zero. The methods for shrinkage and sparsification are incorporated into the MCMC algorithm and are described in detail in the next section.

Identification of factors Without restrictions, the loadings and factors cannot be identified, since, for an arbitrary $r \times r$ invertible matrix Q, it is evident that $\Lambda_t F_t = \Lambda_t Q Q^{-1} F_t$. Obviously, there are r^2 free elements in Q. Hence, we need r^2 restrictions to identify Λ_t and F_t. Note that the r^2 restrictions already exist in our model specification.

First, in the TVS-ADF, the variance–covariance matrix of η_t is set as an identity matrix, which is a standard normalization assumption for factor models. This provides $r(r+1)/2$ restrictions on the conditional variance–covariance matrix of F_t. Second, diagonal $\tilde{D}_i$ for all i (i.e., the loadings of different latent factors and their lags are independent) implies that the variance–covariance matrix of Λ_t (i.e., $E\Lambda_t'\Lambda_t - (E\Lambda_t)'E\Lambda_t$) is diagonal, which provides additional $r(r-1)/2$ restrictions. Finally, Λ_t and F_t are identified with the r^2 restrictions.

We have two remarks about identification. First, regarding time variation in the coefficients and factors, there is the concern that Q can be time varying instead of being constant. However, this is impossible under our model setting because rescaled factors $Q_t^{-1}F_t$ and rescaled loadings $\Lambda_t Q_t$ cannot satisfy (2.2) and (2.5). Second, if we multiply both Λ_t and F_t by -1, then $-\Lambda_t \times -F_t = \Lambda_t F_t$. Hence, the signs of loadings Λ_t and factors F_t are indeterminate. This is a common issue in all factor models. Please refer to Hamilton et al. (2007) for strategies of solving this issue. Of course, this is not a problem if we only focus on the forecasting of dependent variables or the scales of the latent factors and factor loadings.

The number of factors In this study, we assume that the number of factors is given. One can determine the number of factors according to some prior information or experience. Although there are some studies about the determination of the number of factors, these methods are mostly based on factor models with time-invariant parameters (e.g., Bai and Ng, 2002; Amengual and Watson, 2007; Hallin and Liška, 2007; Onatski, 2010; Alessi et al., 2010; Ahn and Horenstein, 2013; Trapani, 2018). By contrast, the parameters in our model are time varying, and the total number of factors depends on q and p. It is difficult to select an optimal number for q and p. However, as this issue is out of the scope of this study, we will attempt to solve it in our future research.

2.3 MCMC algorithm for estimation

We use the MCMC algorithm to estimate the model. As it is difficult and complex to obtain the joint posterior distribution of all parameters, we simulate the joint posterior distribution using Gibbs sampling, sequentially drawing the parameters of the TVS-ADF from the conditional posterior distributions. The detailed algorithm for the Gibbs sampling comprises seven steps, as follows:

2.3.1 *Step 1: drawing* B_t *and* Λ_t

B_t and Λ_t are drawn together, conditional on the remaining parameters. To simplify the drawing, shrinkage, and sparsification, we first undertake some transformations and introduce some additional notations. Note that in this step, we only carry out shrinkage on B_t and Λ_t, while their sparsification will be illustrated in step 5.

According to (2.1), for $i = 1, \cdots, n$, we have

$$y_{it} = \beta_{it}' x_{it} + \lambda_{it}' F_t + \xi_{it}, \xi_{it} \sim N(0, \Gamma_{t,ii}^{\xi}).$$

Then, by conflating β_{it} and λ_{it} and combining (2.4) and (2.5), we have

$$\begin{aligned} y_{it} &= b_{it}' p_{it} + \xi_{it}, \qquad \xi_{it} \sim N(0, \Gamma_{t,ii}^{\xi}), \\ b_{it} &= b_{it-1} + v_{it}, \qquad v_{it} \sim N(0, D_i), \end{aligned} \tag{2.8}$$

where $b_{it}' = (b_{i1t}, b_{i2t}, \cdots, b_{ik_it}) = (\beta_{it}', \lambda_{it}')$, $k_i = m_i + r$, $p_{it} = (x_{it}', F_t')'$, and $D_i = \mathrm{diag}(\bar{D}_i, \tilde{D}_i)$.

We introduce an $k_i \times 1$ vector $b_i = (b_{i1}, b_{i2}, \cdots, b_{ik_i})'$. Then, following

Frühwirth-Schnatter and Wagner (2010), we can transfer (2.8) to:

$$\begin{aligned} y_{it} &= \tilde{b}_{it}' D_i^{1/2} p_{it} + b_i' p_{it} + \xi_{it}, \qquad \xi_{it} \sim N(0, \Gamma_{t,ii}^{\xi}) \\ \tilde{b}_{it} &= \tilde{b}_{it-1} + \tilde{v}_{it}, \qquad \tilde{v}_{it} \sim N(0, I_{k_i}), \end{aligned} \tag{2.9}$$

where I_{k_i} is an $k_i \times k_i$ identity matrix, $\tilde{b}_{i0} = 0$ and

$$\tilde{b}_{it} = (\tilde{b}_{i1t}, \cdots, \tilde{b}_{ik_it})', \qquad \tilde{b}_{ijt} = \frac{b_{ijt} - b_{ij}}{\sqrt{D_{ij}}}, \qquad j = 1, \cdots, k_i, \tag{2.10}$$

where D_{ij} is the j-th diagonal element of D_i. Then, (2.9) also can be written as

$$y_{it} = \alpha_i' z_{it} + \xi_{it}, \tag{2.11}$$

with $\alpha_i = (\sqrt{D_{i1}}, \cdots, \sqrt{D_{ik_i}}, b_{i1}, \cdots, b_{ik_i})'$, $z_{it} = ((\tilde{b}_{it} \odot p_{it})', p_{it}')'$, and $\odot$ denotes element-wise multiplication.

Prior As we want to shrink parameters β_{it}, λ_{it}, and D_i, which have been collected and transformed to α_i, toward zero, we use a special prior, namely the Dirichlet–Laplace prior proposed by Bhattacharya et al. (2015). Specifically, α_{ij}, $j = 1, \cdots, 2k_i$ denotes the j-th element of α_i and follows a Gaussian distribution:

$$\alpha_{ij} \mid \omega_{ij}, \epsilon_{ij}, J_i \sim N(0, \omega_{ij} \epsilon_{ij}^2 \zeta_i^2), \tag{2.12}$$

with

$$\omega_{ij} \sim e(1/2) \quad \epsilon_{ij} \sim D(a, \cdots, a) \quad \zeta_i \sim G(2k_i a, 1/2), \tag{2.13}$$

where $e(\cdot)$ denotes the exponential distribution, a is specified as $(2k_i)^{-(1+\phi)}$ with ϕ being a positive number close to zero, $D(\cdot)$ is the Dirichlet distribution, and $G(\cdot)$ refers to the Gamma distribution.

This prior is adopted for the especially popular method of global–local shrinkage (e.g., Polson and Scott, 2010), which is both global (i.e., common to all parameters) and local (i.e., specific to each parameter). The shrinkages of the $2k_i$ parameters (i.e., global shrinkage) are controlled by ζ_i, while ω_{ij} and ϵ_{ij} handle the shrinkage of the j-th parameter (i.e., local shrinkage). Regarding equation (2.13), ζ_i and ω_{ij} both take very small positive values with a high probability, given the properties of the exponential and Gamma distributions; so does ϵ_{ij} because the marginal distributions of $D(a, \cdots, a)$ are beta distributions with $a < 1/2$. Therefore, in this type of setup, the value of α_{ij} will be close to zero with a high probability. Note that a plays an important role in determining the shrinkage behavior of the Dirichlet–Laplace prior. Following Huber et al. (2020), we draw a from its posterior distribution, which is obtained based on the prior of a uniform distribution bounded between $(2k_i)^{-1}$ and $1/2$.

Drawing process Now, we show how to simulate the full history of B_t and Λ_t, which have been transformed into $\tilde{b}_{it}$ in (2.9), using the Dirichlet–Laplace prior.

We need some additional notations, as follows. Let $b_i^T = (b_{i1}, \cdots, b_{iT})$, $b^T = (b_1^T, \cdots, b_n^T)$, $\tilde{b}_i^T = (\tilde{b}_{i1}, \cdots, \tilde{b}_{iT})$, and $\tilde{b}^T = (\tilde{b}_1^T, \cdots, \tilde{b}_n^T)$, similarly for F^T, A^T, H^T, Y^T, X^T. Furthermore, let $\omega_i = (\omega_{i1}, \cdots, \omega_{i2k_i})$, $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{i2k_i})$, and $M_1 = (\omega_i, \epsilon_i, \zeta_i, a)$.

According to Carter and Kohn (1994) and Frühwirth-Schnatter (1994), the conditional probability density function of $\tilde{b}_{it}$ can be factorized as

$$f(\tilde{b}_i^T|\Theta_1^T) = f(\tilde{b}_{iT}|\Theta_1^T) \prod_{t=1}^{T-1} f(\tilde{b}_{it}|\tilde{b}_{it+1}, \Theta_1^T), \tag{2.14}$$

where $\Theta_1^T = (F^T, A^T, H^T, \alpha_i, M_1, Y^T, X^T)$ and $f(\cdot|\cdot)$ stands for the conditional probability density function. Obviously, all conditional density functions in the equation are normal distributions. To conduct the drawing process, we first need to obtain the mean and variance of each conditional distribution.

For $f(\tilde{b}_{iT}|\Theta_1^T)$, we use the Kalman filter for (2.9) as follows:

$$\begin{aligned}
\tilde{b}_{it|t} =& \tilde{b}_{it|t-1} + \Gamma^{\tilde{b}_i}_{t|t-1} D_i^{1/2} p_{it} [p'_{it} D_i^{1/2} \Gamma^{\tilde{b}_i}_{t|t-1} D_i^{1/2} p_{it} + \Gamma^{\xi}_{t|t-1,ii}]^{-1} [y_{it} - \tilde{b}'_{it|t-1} D_i^{1/2} p_{it} - b'_i p_{it}],\\
\Gamma^{\tilde{b}_i}_{t|t} =& \Gamma^{\tilde{b}_i}_{t|t-1} - \Gamma^{\tilde{b}_i}_{t|t-1} D_i^{1/2} p_{it} [p'_{it} D_i^{1/2} \Gamma^{\tilde{b}_i}_{t|t-1} D_i^{1/2} p_{it} + \Gamma^{\xi}_{t|t-1,ii}]^{-1} p'_{it} D_i^{1/2} \Gamma^{\tilde{b}_i}_{t|t-1},\\
\tilde{b}_{it+1|t} =& \tilde{b}_{it|t},\\
\Gamma^{\tilde{b}_i}_{t+1|t} =& \Gamma^{\tilde{b}_i}_{t|t} + I_{k_i},
\end{aligned} \tag{2.15}$$

where $\tilde{b}_{i1|0} = 0$ and $\Gamma^{\tilde{b}_i}_{1|0}$ is a positive number close to zero. The final iteration of the Kalman filter provides the mean and variance of $f(\tilde{b}_{iT}|\Theta_1^T)$.

Following Carter and Kohn (1994), to obtain the mean and variance of $f(\tilde{b}_{it}|\tilde{b}_{it+1}, \Theta_1^T)$, we first conduct some equation transformations. We consider $\tilde{b}_{it+1} = \tilde{b}_{it} + \tilde{v}_{it+1}$ as additional observations on $\tilde{b}_{it}$. We then pre-multiply $\tilde{b}_{it+1} = \tilde{b}_{it} + \tilde{v}_{it+1}$ by L^{-1}, where L is from the Cholesky decomposition of $I_{k_i} = L'\Delta_i L$, and we have $L^{-1}\tilde{b}_{it+1} = L^{-1}\tilde{b}_{it} + L^{-1}\tilde{v}_{it+1}$.

We define $\overline{b}_{it+1} = L^{-1}\tilde{b}_{it+1}$ and $\overline{v}_{it+1} = L^{-1}\tilde{v}_{it+1}$. Then, for the j-th row of $\overline{b}_{it+1}$,

$$\overline{b}_{it+1,j} = L_j^{-1}\tilde{b}_{it} + \overline{v}_{it+1,j}, \qquad \overline{v}_{it+1,j} \sim N(0, \Delta_{i,jj}). \tag{2.16}$$

For $j = 1, \cdots, k_i$, let

$$\begin{aligned}
\tilde{b}_{it|t,j} &= E[\tilde{b}_{it}|\Theta_1^t, \tilde{b}_{it+1,1}, \cdots, \tilde{b}_{it+1,j-1}]\\
\Gamma^{\tilde{b}_i}_{t|t,j} &= Var[\tilde{b}_{it}|\Theta_1^t, \tilde{b}_{it+1,1}, \cdots, \tilde{b}_{it+1,j-1}],
\end{aligned}$$

where $\tilde{b}_{it+1,j}$ denotes the j-th row of $\tilde{b}_{it+1}$.

It is straightforward to obtain the following observation update equations using the Kalman filter for $\tilde{b}_{it|t,j-1}$:

$$\tilde{b}_{it|t,j} = \tilde{b}_{it|t,j-1} + \Gamma^{\tilde{b}_i}_{t|t,j-1} L_j^{-1'} [L_j^{-1} \Gamma^{\tilde{b}_i}_{t|t,j-1} L_j^{-1'} + \Delta_{i,jj}]^{-1} [\bar{b}_{it+1,j} - L_j^{-1} \tilde{b}_{it|t,j-1}],$$
$$\Gamma^{\tilde{b}_i}_{t|t,j} = \Gamma^{\tilde{b}_i}_{t|t,j-1} - \Gamma^{\tilde{b}_i}_{t|t,j-1} L_j^{-1'} [L_j^{-1} \Gamma^{\tilde{b}_i}_{t|t,j-1} L_j^{-1'} + \Delta_{i,jj}]^{-1} L_j^{-1} \Gamma^{\tilde{b}_i}_{t|t,j-1},$$

where $\tilde{b}_{it|t,0} = \tilde{b}_{it|t}$ and $\Gamma^{\tilde{b}_i}_{t|t,0} = \Gamma^{\tilde{b}_i}_{t|t}$, which are the outcomes of the Kalman filter in (2.15). To run the updated equations above, we need to obtain $\bar{b}_{it,j}$ for $t = 1, \cdots, T$. To do this, we first draw $\tilde{b}_{iT}$ from $f(\tilde{b}_{iT}|\Theta_1^T)$; then, $\bar{b}_{iT,j}$ can be obtained by $\bar{b}_{iT} = L^{-1}\tilde{b}_{iT}$. Based on $\tilde{b}_{iT}$, $\tilde{b}_{iT-1}$ can be drawn from $f(\tilde{b}_{iT-1}|\tilde{b}_{iT}, \Theta_1^T)$, and $\bar{b}_{iT-1,j}$ can be obtained as $\bar{b}_{iT-1} = L^{-1}\tilde{b}_{iT-1}$. The process is similar for $\bar{b}_{iT-2,j}, \cdots, \bar{b}_{i1,j}$. Now, we can run the above update equations k_i times. The final iteration gives the expectation and variance of $\tilde{b}_{it}|\tilde{b}_{it+1}, \Theta_1^T$. As (2.14) is a product of Gaussian densities, we can easily draw $\tilde{b}_{it}$ from it and then transform $\tilde{b}_{it}$ back to obtain b_{it} based on (2.10).

2.3.2 *Step 2: drawing* F_t

F_t is drawn from its conditional distribution:

$$f(F^T|\Theta_2^T) = f(F_T|\Theta_2^T) \prod_{t=1}^{T-1} f(F_t|F_{t+1}, \Theta_2^T), \tag{2.17}$$

where $\Theta_2^T = (b^T, A^T, H^T, C_1, Y^T, X^T)$. Combining (2.1) and (2.2), we can simulate the full history of F_t by following a similar drawing process as the one in step 1.

2.3.3 *Step 3: drawing* A_t

In this step, we implement shrinkage on A_t, while the sparsification for A_t is illustrated in step 5.

As preparation, we perform some equation transformations. From (2.3), we have $A_t \xi_t = H_t^{1/2} e_t$. This means:

$$
\begin{aligned}
\xi_{1t} &= \sqrt{h_{1t}} e_{1t} \\
a_{21,t}\xi_{1t} + \xi_{2t} &= \sqrt{h_{2t}} e_{2t} \\
a_{31,t}\xi_{1t} + a_{32,t}\xi_{2t} + \xi_{3t} &= \sqrt{h_{3t}} e_{3t} \\
&\vdots \\
a_{n1,t}\xi_{1t} + a_{n2,t}\xi_{2t} + \cdots + a_{nn-1,t}\xi_{n-1t} + \xi_{nt} &= \sqrt{h_{nt}} e_{nt}.
\end{aligned}
\tag{2.18}
$$

Then, for $i = 2, \cdots, n$, we have

$$
\frac{\xi_{it}}{\sqrt{h_{it}}} = -\frac{1}{\sqrt{h_{it}}} \xi^{i-1t'} a_{it} + e_{it}, \qquad e_{it} \sim N(0,1), \tag{2.19}
$$

where $\xi^{i-1t'} = (\xi_{1t}/\sqrt{h_{it}}, \cdots, \xi_{i-1t}/\sqrt{h_{it}})$ and $a_{it} = (a_{i1,t}, \cdots, a_{ii-1,t})'$. Moreover, (2.6) can be rewritten as

$$
a_{it} = a_{it-1} + u_{it}, \qquad u_{it} \sim N(0, S_i), \tag{2.20}
$$

where S_i is an $(i-1) \times (i-1)$ diagonal matrix.

Now, we introduce an $(i-1) \times 1$ vector $a_i = (a_{i1}, \cdots, a_{ii-1})'$ and let S_{ij} denote the j-th diagonal element of S_i. Then, we can rewrite (2.19) and (2.20) as

$$
\frac{\xi_{it}}{\sqrt{h_{it}}} = -\frac{1}{\sqrt{h_{it}}} \tilde{a}_{it} ? \sqrt{S_i} \xi^{i-1t} - \frac{1}{\sqrt{h_{it}}} a_{it} ? \xi^{i-1t} + e_{it}, \qquad e_{it} \sim N(0,1), \tag{2.21}
$$

$$
\tilde{a}_{it} = \tilde{a}_{it-1} + \tilde{u}_{it}, \qquad \tilde{u}_{it} \sim N(0, I_i), \tag{2.22}
$$

where I_i is an $(i-1) \times (i-1)$ identity matrix:

$$\tilde{a}_{it} = (\tilde{a}_{i1,t}, \cdots, \tilde{a}_{ii-1,t})', \qquad \tilde{a}_{ij,t} = \frac{a_{ij,t} - a_{ij}}{\sqrt{S_{ij}}}, \qquad j = 1, \cdots, i-1, \tag{2.23}$$

and $\tilde{a}_{i0} = 0$.

We then collect the parameters together and define the following new notations: $\psi_i = (\sqrt{S_{i1}}, \cdots, \sqrt{S_{ii-1}}, a_i')'$ and $c_{it} = ((-1/\sqrt{h_{it}}\tilde{a}_{it} \odot \xi^{i-1t})', -1/\sqrt{h_{it}}\xi^{i-1t'})'$. The following transformation of (2.21) is used subsequently: $\xi_{it}/\sqrt{h_{it}} = \psi_i' c_{it} + e_{it}$.

Prior For shrinking A_t, we use the Dirichlet–Laplace prior for ψ_i.

Drawing process We simulate the full history of $\tilde{a}_{it}$ using the Dirichlet–Laplace prior. Specifically, we define $\tilde{a}_i^T = (\tilde{a}_{i1}, \cdots, \tilde{a}_{iT})$, and the conditional probability density function of $\tilde{a}_{it}$ can be expressed as follows:

$$f(\tilde{a}_i^T | \Theta_3^T) = f(\tilde{a}_{iT} | \Theta_3^T) \prod_{t=1}^{T-1} f(\tilde{a}_{it} | \tilde{a}_{it+1}, \Theta_3^T), \tag{2.24}$$

where $\Theta_3^T = (b^T, F^T, H^T, \psi_i, M_2, Y^T, X^T)$ and M_2 denotes the hyperparameter in the prior of ψ_i. Based on (2.24), (2.21), and (2.22), we can obtain $\tilde{a}_i^T$ using a drawing process similar to that used in step 1. Finally, we transform $\tilde{a}_{it}$ back to get a_{it} based on (2.23).

2.3.4 *Step 4: drawing* H_t

We implement shrinkage in this step and sparsification in the next one on H_t. As a preparation, we make the following equation transformations.

We define $m_t = A_t \xi_t$; then, from $A_t \xi_t = H_t^{1/2} e_t$, for $i = 1, \cdots, n$, we have the i-th element of m_t, $m_{it} = \sqrt{h_{it}} e_{it}$. Consequently,

$$\begin{aligned} log(m_{it}^2) &= log(h_{it}) + log(e_{it}^2), \qquad e_{it}^2 \sim \chi^2(1), \\ &\approx -1.27 + log(h_{it}) + \phi_{it}, \qquad \phi_{it} \sim N(0, \frac{\pi^2}{2}). \end{aligned} \tag{2.25}$$

Next, we introduce element h_i and let σ_i^2 denote the i-th diagonal element of Σ. Combining (2.7) and (2.25), we have

$$log(m_{it}^2) = -1.27 + log(\tilde{h}_{it})\sigma_i + h_i + \phi_{it}, \qquad \phi_{it} \sim N(0, \frac{\pi^2}{2}), \tag{2.26}$$

$$log(\tilde{h}_{it}) = log(\tilde{h}_{it-1}) + \tilde{\gamma}_{it}, \qquad \tilde{\gamma}_{it} \sim N(0, 1), \tag{2.27}$$

where

$$log(\tilde{h}_{it}) = \frac{log(h_{it}) - h_i}{\sqrt{\sigma_i^2}}, \tag{2.28}$$

and $log(\tilde{h}_{i0}) = 0$. Further, we have $log(m_{it}^2) + 1.27 = \tau_i' l_{it} + \phi_{it}$, where $\tau_i' = (\sigma_i, h_i)$ and $l_{it} = (log(\tilde{h}_{it}), 1)'$.

Prior To shrink H_t, we use the Dirichlet–Laplace prior for τ_i.

Drawing process We simulate the full history of $log(\tilde{h}_{it})$ based on the conditional probability density function of $log(\tilde{h}_{it})$:

$$f(log(\tilde{h}_i)^T | \Theta_4^T) = f(log(\tilde{h}_{it}) | \Theta_4^T) \prod_{t=1}^{T-1} f(log(\tilde{h}_{it}) | log(\tilde{h}_{it+1}), \Theta_4^T), \tag{2.29}$$

where $log(\tilde{h}_i)^T = (log(\tilde{h}_{i1}), \cdots, log(\tilde{h}_{iT}))$, $\Theta_4^T = (b^T, F^T, A^T, \tau_i, M_3, Y^T, X^T)$ and M_3 refers to the hyperparameter in the prior of τ_i. Applying (2.26), (2.27), and (2.29), and following a similar drawing process to that in step 1, $log(\tilde{h}_{it})$ can be obtained. Finally, we transform $log(\tilde{h}_{it})$ back to obtain $log(h_{it})$ based on (2.28).

2.3.5 *Step 5: drawing* α_i, ψ_i, *and* τ_i

In this step, we first show how to draw α_i, ψ_i, and τ_i from their posteriors, and then illustrate how to sparsify them.

Posterior α_i, ψ_i, and τ_i are drawn from their posteriors, which can be obtained straightforwardly. Specifically,

$$\begin{aligned}
\alpha_i|\Theta_5^T &\sim N((\Omega_i^{-1} + z_i'\Gamma_i^{-1}z_i)^{-1}z_i'\Gamma_i^{-1}y_i,\ \ (\Omega_i^{-1} + z_i'\Gamma_i^{-1}z_i)^{-1}), \qquad (2.30)\\
\psi_i|\Theta_6^T &\sim N((\bar{\Omega}_i^{-1} + c_i'c_i)^{-1}c_i'\bar{y}_i,\ \ (\bar{\Omega}_i^{-1} + c_i'c_i)^{-1}),\\
\tau_i|\Theta_7^T &\sim N((\tilde{\Omega}_i^{-1} + (\frac{\pi^2}{2})^{-1}l_i'l_i)^{-1}(\frac{\pi^2}{2})^{-1}l_i'\tilde{y}_i,\ \ (\tilde{\Omega}_i^{-1} + (\frac{\pi^2}{2})^{-1}l_i'l_i)^{-1}),
\end{aligned}$$

where $\Theta_5^T = (b^T, F^T, H^T, A^T, M_1, Y^T, X^T)$, Ω_i refers to the variance of the prior of α_i, $\Gamma_i = \mathrm{diag}(\Gamma_{1,ii}^{\xi}, \cdots, \Gamma_{T,ii}^{\xi})$, $z_i = (z_{i1}, \cdots, z_{iT})'$, $\Theta_6^T = (b^T, F^T, H^T, A^T, M_2, Y^T, X^T)$, $\bar{\Omega}_i$ refers to the variance of the prior of c_i, $c_i = (c_{i1}, \cdots, c_{iT})'$, $\bar{y}_i = (\frac{\xi_{i1}}{\sqrt{h_{i1}}}, \cdots, \frac{\xi_{iT}}{\sqrt{h_{iT}}})'$, $\Theta_7^T = (b^T, F^T, H^T, A^T, M_3, Y^T, X^T)$, $\tilde{\Omega}_i$ refers to the variance of the prior of l_i, $l_i = (l_{i1}, \cdots, l_{iT})'$, and $\tilde{y}_i = (log(m_{i1}^2) + 1.27, \cdots, log(m_{iT}^2) + 1.27)'$.

Sparsification Following Huber et al. (2020) and given a draw $\alpha_i^{(nT)} = (\alpha_{i1}^{(nT)}, \cdots, \alpha_{i2k_i}^{(nT)})$ from (2.30), the sparsified α_i is obtained as

$$\bar{\alpha}_{ij} = sign(\alpha_{ij}^{(nT)}) \parallel z_{i,j} \parallel^{-2} (\mid \alpha_{ij}^{(nT)} \mid\parallel z_{i,j} \parallel^2 - \kappa_{ij})_+, \quad j = 1, \cdots, 2k_i, \qquad (2.31)$$

where $\kappa_{ij} =\mid \alpha_{ij}^{(nT)} \mid^{-2}$, $sign(x)$ returns the sign of x, $z_{i,j}$ denotes the j-th column of z_i, and $(x)_+ = \max(x, 0)$. Note that equation (2.31) is a soft-thresholding approach, in which the value of $\bar{\alpha}_{ij}$ below a certain value is set to zero.

Sparsification can be conducted similarly for ψ_i and τ_i.

2.3.6 *Step 6: drawing* C_1

From (2.2), we have

$$f_t = c' F_{t-1} + \eta_t, \;\; \eta_t \sim N(0, I),$$

where $c = (c_1, \cdots, c_{p+1})'$. For the i-th row of f_t,

$$f_{t,i} = c_i' F_{t-1} + \eta_{t,i}, \;\; \eta_{t,i} \sim N(0, 1).$$

Note that (2.2) is a VAR. There are several types of prior distributions available that can be used for VAR models, as summarized by Kadiyala and Karlsson (1997) and Gelman et al. (2013). Among these, we select the standard noninformative prior for c. Then, we draw c_i from its conditional posterior distribution,

$$c_i \mid F^T \sim N(\hat{c}_i, (F_{t-1}^{T'} F_{t-1}^{T})^{-1}),$$

where $\hat{c}_i$ denotes the OLS estimate of c_i.

2.3.7 *Step 7: drawing* M_1*,* M_2*,* M_3

In this step, we draw hyperparameters M_i for $i = 1, 2, 3$. For M_1, which includes ω_i, ϵ_i, ζ_i, and a in the prior of α_i, we have

$$\omega_{ij} \mid \alpha_{ij}, \epsilon_{ij}, \zeta_i \sim IG(\zeta_i \frac{\epsilon_{ij}}{\mid \alpha_{ij} \mid}, 1),$$

$$\zeta_i \mid \alpha_i, \epsilon_i \sim GIG(2k_i(a-1), 1, 2\sum_{j=1}^{2k_i} \frac{\mid \alpha_{ij} \mid}{\epsilon_{ij}}),$$

$$\epsilon_{ij} = \frac{T_{ij}}{\sum_{j=1}^{2k_i} T_{ij}}, \quad T_{ij} \mid \alpha_{ij} \sim GIG(a-1, 1, 2 \mid \alpha_{ij} \mid),$$

where IG denotes the inverse Gaussian distribution and GIG refers to the generalized inverted Gaussian distribution (see Bhattacharya et al., 2015 and Huber et al., 2020). Following Huber et al. 2020, we obtain the conditional posterior of a using a Metropolis–Hastings algorithm with a Gaussian proposal distribution truncated between $(2k_i)^{-1}$ and $1/2$. In the simulation and empirical application, the variance of the proposal distribution is tuned during the first 20% of the burn-in stage of the MCMC sampler, such that the acceptance rate is between 20% and 40%.

This is similar for M_2 and M_3.

2.4 Artificial simulation

Here, we present evidence on the performance of our model based on simulation experiments using artificial data generated from the TVS-ADF.

To assess how well the different models perform across different numbers of explained variables, numbers of explanatory variables, degrees of sparsity, and lengths of time series, we set $n = 10, 20, 30$. For each n, we consider three sparsity levels, labeled as dense (with 10% zeros in α_i, ϕ_i, and τ_i), moderate (with 50% zeros), and sparse (with 90% zeros). For each sparsity level, we consider $k_i = 4$ and 8 explanatory variables and sample sizes $T = 50$ and 200. We randomly generate $N = 1000$ simulated datasets for each variant. We set $x_{it} \sim N(0, 10I_x)$, $p = 1$ and $S = 0.01^2 I_S, \Sigma = 0.01^2 I_\Sigma, \bar{D} = 0.01^2 I_{\bar{D}}, \tilde{D} = 0.01^2 I_{\tilde{D}}$, where I_x, I_S, I_Σ, $I_{\bar{D}}$ and $I_{\tilde{D}}$ are identity matrices of dimensions $m_i \times m_i$, $[n \times (n-1)/2] \times [n \times (n-1)/2]$, $n \times n$, $nm_i \times nm_i$ and $nr \times nr$, respectively. Moreover, for $k_i = 4$, we set

$$m_i = 2, \qquad q = 1, \qquad C_1 = \begin{pmatrix} 0.2 & 0.1 \\ 1 & 0 \end{pmatrix}.$$

For $k_i = 8$, we set

$$m_i = 4, \qquad q = 2, \qquad C_1 = \begin{pmatrix} 0.2 & 0.2 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

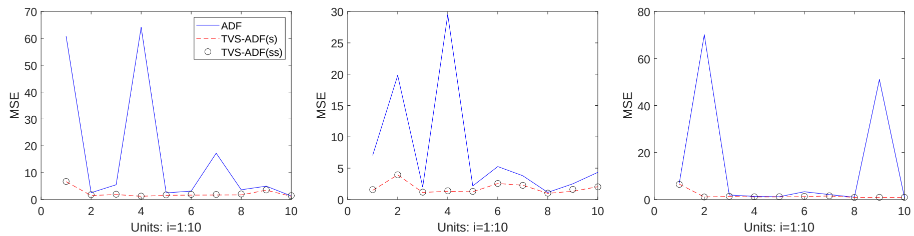
The initial value of F_t is set to zero. The initial value of Γ_t^ξ, Γ_1^ξ, is generated by $1/2(a + a')$, where a is an $n \times n$ matrix and each element of a is generated from a normal distribution, $N(0, 0.1)$. Each diagonal element of Γ_1^ξ is replaced by one to ensure Γ_1^ξ is a positive definite matrix. Then, we can easily obtain the initial values of A_t and H_t using the Cholesky decomposition for Γ_1^ξ, respectively.

We use the two-step method of Forni et al. (2009) to estimate the ADF and the MCMC algorithm described in Section 3 to estimate the TVS-ADF. We use the first T observations to estimate the models, and then the resulting estimates to predict the $T+1$-th observation. The precision of point forecasts is measured by the following two types of mean-squared errors:

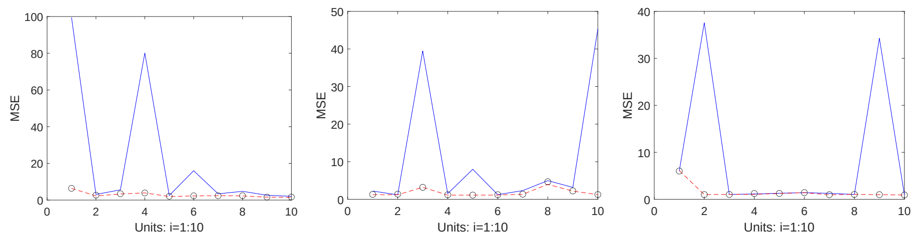
$$\begin{aligned} \mathrm{MSE}_i &= \frac{1}{N}\sum_{j=1}^{N}(y_{i,T+1}^j - \hat{y}_{i,T+1}^j)^2, \quad i = 1, \cdots, n, \\ \mathrm{MSE}^{(n)} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i, \end{aligned}$$

where $y_{i,T+1}^j$ refers to the $T+1$-th observation of the i-th unit (i.e., i-th explained variable) in the j-th simulated dataset and $\hat{y}_{i,T+1}^j$ denotes its fitted value. MSE_i measures the predictive precision of the i-th unit, while $\mathrm{MSE}^{(n)}$ measures the forecasting accuracy of all units.

Figures 2.1 and 2.2 plot the mean-squared errors of our model and the ADF with $n = 10$ for different numbers of explanatory variables, sparsity levels, and sample sizes. The results show that, for most units, the performance of the



(a) 10% sparsity level and $T = 50$ (b) 50% sparsity level and $T = 50$ (c) 90% sparsity level and $T = 50$

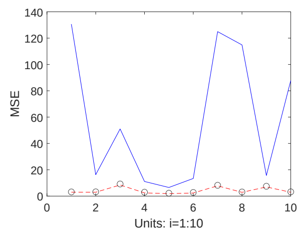


(d) 10% sparsity level and $T = 200$ (e) 50% sparsity level and $T = 200$ (f) 90% sparsity level and $T = 200$

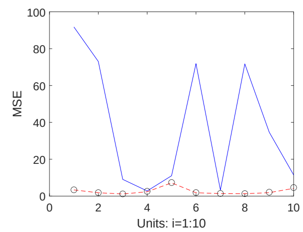
Figure 2.1: MSE_i of three models with 4 explanatory variables ($n = 10$)

TVS-ADF(s) and the TVS-ADF(ss) is always better than that of the ADF for all combinations with different numbers of explanatory variables, sparsity levels, and sample sizes. Additionally, the TVS-ADF(s) and TVS-ADF(ss) always have similar outcomes for all settings. Moreover, for different sample sizes and numbers of explanatory variables, the results of the ADF gradually approach those of our models as the sparsity level increases. The reason could be that, as the sparsity level rises, more parameters in the date-generating process will become constants, such that the time-varying feature of the data tends to become weaker, which is a favorable situation for the ADF. Similarly for $n = 20$ and $n = 30$ (see Figures A.1 – A.4 in Appendix A).

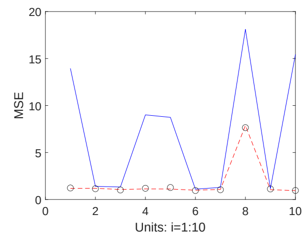
To further compare model performances, we tabulate MSE$^{(n)}$ for three models with different numbers of explained variables, numbers of explanatory variables, sample sizes, and sparsity levels in Table 1. We can make the following observations.



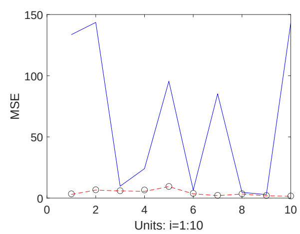
(a) 10% sparsity level and $T = 50$



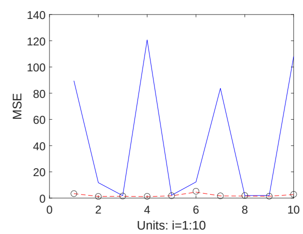
(b) 50% sparsity level and $T = 50$



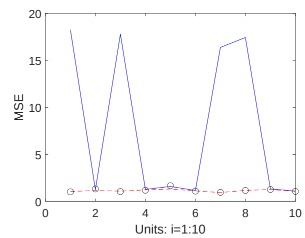
(c) 90% sparsity level and $T = 50$



(d) 10% sparsity level and $T = 200$



(e) 50% sparsity level and $T = 200$



(f) 90% sparsity level and $T = 200$

Figure 2.2: MSE_i of three models with 8 explanatory variables ($n = 10$)

First, regardless of the number of explained variables and explanatory variables, as the sparsity level gradually increases, the mean-squared errors of the three models become smaller across all sample sizes. As discussed in relation to Figures 1 and 2, it is obvious that the time-varying characteristic tends to become weaker as the sparsity level rises and more parameters become constants. This could cause MSE$^{(n)}$ for the ADF to decline gradually. Additionally, for both TVS-ADF(s) and TVS-ADF(ss), the increase in the sparsity level could be conducive to an improvement in their performance.

Second, MSE$^{(n)}$ for both TVS-ADF(s) and TVS-ADF(ss) decrease as sample size T increases in some cases, whereas in other cases, MSE$^{(n)}$ increase. This may be because although increasing the sample size benefits estimation accuracy, it also increases the number of unknown parameters; thus, an increase in the number of unknown parameters can offset the benefit of a larger sample. Moreover, the MSE$^{(n)}$ for the ADF has a similar tendency. A possible reason is that, in the random walk process, increasing time T can cause larger aggregate movements in the parameters, which can offset the benefit of a larger sample.

Third, the number of explanatory variables has little positive impact on the results. For $T = 200$ and 90% sparsity, $\text{MSE}^{(n)}$ for both TVS-ADF(s) and TVS-ADF(ss) decrease with the number of the variables. As a matter of fact, increasing the number of explanatory variables causes an increase in the number of parameters, which could decrease estimation accuracy. Fourth, TVS-ADF(s) and TVS-ADF(ss) show substantial improvements with respect to predictive accuracy relative to the ADF in all cases, which indicates that our models can better capture time-varying information. Moreover, the TVS-ADF(s) and TVS-ADF(ss) yield similar outcomes in all cases, but the TVS-ADF(s) performs slightly better than the TVS-ADF(ss) for most cases.

Overall, the TVS-ADF(s) and TVS-ADF(ss) almost always perform better than the ADF for different numbers of explained variables, numbers of explanatory variables, sparsity levels, and sample sizes. Furthermore, the TVS-ADF(s) and TVS-ADF(ss) display similar results.

Table 2.1: $\mathrm{MSE}^{(n)}$ of out-of-sample point forecasts

		ADF	TVS-ADF(s)	TVS-ADF(ss)	ADF	TVS-ADF(s)	TVS-ADF(ss)
$n = 10$			$k_i = 4, T = 50$			$k_i = 8, T = 50$	
	dense(10%)	16.56747	2.23583	2.42509	57.22385	4.16038	4.55003
	moderate(50%)	7.75462	1.82023	1.88252	38.07401	2.63993	2.78389
	sparse(90%)	13.98275	1.65553	1.67459	7.15185	1.75914	1.75670
			$k_i = 4, T = 200$			$k_i = 8, T = 200$	
	dense(10%)	21.96753	2.77822	2.88637	64.90106	4.24454	4.57604
	moderate(50%)	10.93620	1.77230	1.89306	43.38507	2.07157	2.25233
	sparse(90%)	8.59722	1.57637	1.58827	7.75183	1.13239	1.17176
$n = 20$			$k_i = 4, T = 50$			$k_i = 8, T = 50$	
	dense(10%)	9.98868	2.01296	2.18628	30.17183	3.75721	4.19641
	moderate(50%)	6.20357	1.53730	1.61561	25.03899	2.72899	2.96217
	sparse(90%)	7.62831	1.44730	1.48258	5.59381	1.76201	1.77271
			$k_i = 4, T = 200$			$k_i = 8, T = 200$	
	dense(10%)	12.78710	2.40539	2.52003	36.87788	4.31653	4.63485
	moderate(50%)	8.32359	1.54158	1.64028	22.86826	2.08616	2.25175
	sparse(90%)	8.36868	1.34532	1.36230	7.33960	1.21072	1.27215
$n = 30$			$k_i = 4, T = 50$			$k_i = 8, T = 50$	
	dense(10%)	9.95057	2.61056	2.87463	23.36062	4.58562	5.06217
	moderate(50%)	4.27196	1.50271	1.59289	19.20032	3.33551	3.63706
	sparse(90%)	5.99924	1.47907	1.54444	4.23643	1.76040	1.78358
			$k_i = 4, T = 200$			$k_i = 8, T = 200$	
	dense(10%)	9.53996	2.30794	2.41804	24.78437	4.26116	4.58184
	moderate(50%)	6.34530	1.55756	1.64761	16.83925	2.16652	2.33107
	sparse(90%)	6.21827	1.48551	1.50584	6.24828	1.15929	1.20680

2.5 Empirical application: Macroeconomic forecasting

We use the FRED-MD database from McCracken and Ng (2016), which consists of monthly US macroeconomic data. The sample period is from January 1995 to December 2020. The dataset includes economic variables in eight groups: output and income; labor market; housing; consumption, orders, and inventories; money and credit; interest and exchange rates; prices; and stock market.

Let us consider the ordering issue of variables before the formal empirical application. For simplicity, let us consider a two-variable example:

$$\Gamma_t^{\xi} = A_t^{-1} H_t A_t^{-1'} = \begin{pmatrix} 1 & \\ a_{21,t} & 1 \end{pmatrix} \begin{pmatrix} h_{1t} & \\ & h_{2t} \end{pmatrix} \begin{pmatrix} 1 & a_{21,t} \\ & 1 \end{pmatrix} = \begin{pmatrix} h_{1t} & a_{21,t}h_{1t} \\ a_{21,t}h_{1t} & a_{21,t}^2 h_{1t} + h_{2t} \end{pmatrix}$$
$$= \begin{pmatrix} e^{log(h_{1t-1})+\gamma_{1t}} & a_{21,t}h_{1t} \\ a_{21,t}h_{1t} & (a_{21,t-1} + u_{21,t})e^{log(h_{1t-1})+\gamma_{1t}} + e^{log(h_{2t-1})+\gamma_{2t}} \end{pmatrix}.$$

The expression above clarifies that, conditional on $t-1$, the distribution of the first diagonal element of Γ_t^{ξ} is a log-normal distribution, whereas the second diagonal element is not. Hence, different orderings will imply different distributions for the variables, which could affect the model's predictive results. In addition, the ordering issue can also be reflected by (2.18). In this equation, ξ_{1t} has a contemporaneous effect on ξ_{2t}, while ξ_{2t} does not have a contemporaneous effect on ξ_{1t}. In other words, the first variable reacts to the second one with at least one period of lag. Similarly, ξ_{1t} and ξ_{2t} have a contemporaneous effect on ξ_{3t}, but they react to ξ_{3t} with at least one period of lag. The situation is similar for the other variables. This characteristic gives us an important implication that one can determine the ordering of variables according to the contemporaneous relationships between different variables. For instance, in monetary policy analysis, inflation and unemployment react to the policy instrument (e.g., interest rate) with at least one period of lag. The contemporaneous relationships between variables can be speculated based on economic theories or experiences.

Now, we start our empirical application. Specifically, we choose 21 representative variables from different groups by selecting the highest-level indices in each group. Then, these variables are standardized and transformed to be stationary using the transformation codes provided by McCracken and Ng (2016). Subsequently, we use the block method of Belviso and Milani (2006) and Korobilis (2013) to determine the ordering of these variables. We divide these variables into six groups and the ordering is as follows: real activity; money, credit, and finance; exchange rate; price; expectations; and monetary policy (interest rate). Table A.1 (see Appendix A) details the variables. We conduct one-step-ahead point forecasts for these variables using the ADF and TVS-ADF, respectively.

We specify one latent factor and its first-order lags as latent variables, and take the first-order lags of the 21 economic variables as observed explanatory variables and two options, $T = 100$ and 200 as the sample sizes for the estimations. We adopt the following rolling window scheme. For $T = 100$ as the starting point of the rolling window, we use the first 100 observations from the sample period, January 1995 to April 2003, to estimate the models, which are then used to predict the outcomes for May 2003. Then, we move the rolling window one step ahead (i.e., the sample period is from February 1995 to May 2003) and use the resulting estimates to predict the outcomes for June 2003. We proceed recursively 100 times in this fashion and obtain a sequence of forecasts from May 2003 to August 2011. Similarly, for $T = 200$, we obtain a sequence of forecasts from October 2011 to February 2020.

We measure the precision of the one-step-ahead point forecasts for the i-th explained variable using the mean-squared error:

$$\mathrm{MSE}_i = \frac{1}{l}\sum_{t=T+1}^{T+1+l}(y_{it}-\hat{y}_{it})^2,\ \ i=1,\cdots,n,$$

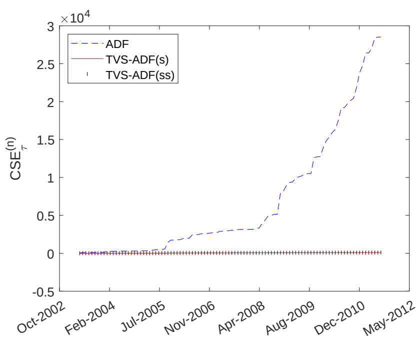
where $l = 100$ (i.e., 100 times) and $n = 21$ (i.e., 21 variables). Additionally, we measure the predictive accuracy of all explained variables using $\mathrm{MSE}^{(n)} = 1/n\sum_{i=1}^{n}\mathrm{MSE}_i$. To measure the predictive accuracy of all explained variables on the time dimension, we use the cumulative sum of forecasting errors:

$$\mathrm{CSE}_\tau^{(n)} = \sum_{t=T+1}^{\tau}\mathrm{SE}_t,\ \ \tau = T+1,\cdots,T+1+l,$$

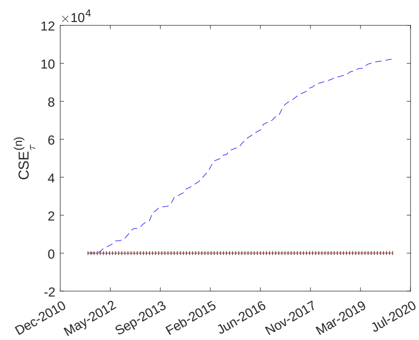
for all explained variables, where $\mathrm{SE}_t = 1/n\sum_{i=1}^{n}(y_{it}-\hat{y}_{it})^2$.

Table 2.2 presents the MSE_i and $\mathrm{MSE}^{(n)}$ for the one-step-ahead point forecasts of the three models for different sample sizes. The deep gray figures indicate the lowest MSE_i across the three models for a given sample size, while the light gray figures are the second lowest MSE_i. The last line of the table gives the $\mathrm{MSE}^{(n)}$. The values of MSE_i show that the predictive performances of the TVS-ADF(ss) and TVS-ADF(s) are better than that of the ADF for most economic variables, regardless of sample size, which could arise from the capacity of our models to capture economic dynamics. The results for the TVS-ADF(ss) and TVS-ADF(s) are similar. Moreover, the predictive accuracies of the TVS-ADF(ss) and TVS-ADF(s) are more stable for each economic variable relative to that of the ADF, for which there are several large values of MSE_i regardless of the sample size. The ADF without time-varying parameters cannot capture economic dynamics better, which could be responsible for its large mean-squared errors of some variables. Furthermore, the results of $\mathrm{MSE}^{(n)}$ indicate that the forecast errors of the TVS-ADF(ss) and TVS-ADF(s) decline with sample size, unlike those of the ADF.

Figure 2.3 presents $\text{CSE}_{\tau}^{(n)}$, the one-step-ahead point forecasts of the three models for different sample sizes, thus illustrating the increasing path of the predictive error. For $T = 100$, the TVS-ADF(s) and TVS-ADF(ss) have much smaller increases in the predictive error relative to the ADF for most time points, while the TVS-ADF(ss) and TVS-ADF(s) consistently beat the ADF during the whole forecasting period for $T = 200$. The cumulative predictive errors of the TVS-ADF(ss) and TVS-ADF(s) are also far smaller than those of their competitors. Moreover, the performance of the TVS-ADF(s) is similar to that of the TVS-ADF(ss).



(a) $T = 100$



(b) $T = 200$

Figure 2.3: $\text{CSE}_{\tau}^{(n)}$ of one-step-ahead point forecasts of three models

In sum, compared to the ADF, the TVS-ADF(s) and TVS-ADF(ss) can better capture economic dynamic features and thus substantially improve the predictive accuracy regardless of whether the sample size is $T = 100$ or 200. Additionally, the forecasting performance of our models is more stable than the ADF for different sample sizes. Moreover, the TVS-ADF(s) and TVS-ADF(ss) always have similar outcomes.

Table 2.2: MSE_i of one-step-ahead point forecasts of three models for different sample sizes

	ADF	TVS-ADF(s)	TVS-ADF(ss)	ADF	TVS-ADF(s)	TVS-ADF(ss)
		$T = 100$			$T = 200$	
Group 1: Real Activity						
CLF	0.00004	0.00000	0.00000	0.00002	0.00000	0.00000
CE	0.00030	0.00001	0.00001	0.00003	0.00001	0.00001
CUR	1.30923	0.03082	0.03332	0.42840	0.02121	0.02065
RPI	0.00065	0.00008	0.00007	0.00039	0.00005	0.00005
HSTNPO	18.39348	0.00686	0.00668	45.79897	0.00786	0.00761
NPHP	20.00081	0.00311	0.00315	47.66086	0.00257	0.00262
RPCE	0.00016	0.00002	0.00001	0.00019	0.00001	0.00001
Group 2: Money, Credit and Finance						
TRDI	0.09464	0.01754	0.01591	0.02762	0.00148	0.00145
M1MS	0.00170	0.00018	0.00020	0.00049	0.00013	0.00017
M2MS	0.00007	0.00002	0.00002	0.00039	0.00001	0.00001
CIL	0.00031	0.00006	0.00006	0.00056	0.00003	0.00003
Group 3: Exchange Rate						
EXJPUS	0.00230	0.00055	0.00054	0.01676	0.00045	0.00047
Group 4: Price						
PPI:CM	0.03470	0.00264	0.00287	0.01216	0.00081	0.00086
PPI:IM	0.00044	0.00016	0.00015	0.00005	0.00004	0.00004
PPI:FG	0.00035	0.00010	0.00010	0.00016	0.00005	0.00004
CPI:AI	0.00004	0.00002	0.00002	0.00003	0.00001	0.00001
PCE:CI	0.00002	0.00001	0.00001	0.00000	0.00001	0.00000
Group 5: Expectations						
CSI	5937.58482	21.75503	21.77418	21362.39305	11.56044	11.54999
NOCG	0.00696	0.00058	0.00056	0.00316	0.00021	0.00021
TBI	0.00360	0.00007	0.00006	0.00019	0.00002	0.00002
Group 6: Monetary policy (interest rate)						
EFFR	11.98383	0.02777	0.02894	1.25289	0.00421	0.00436
$\text{MSE}^{(n)}$	285.21040	1.04027	1.04128	1021.79030	0.55236	0.55184

2.6 Conclusions

This study proposed a new model, TVS-ADF, to help capture the time-varying characteristics of economic data. We also constructed an effective MCMC algorithm (seven-step Gibbs sampling) to estimate this model. Moreover, to avoid overparameterization, we offered shrinkage and sparsification methods for our model in two ways: (i) only shrink the model and (ii) both shrink and sparsify the model.

Using an artificial data experiment, we showed that the TVS-ADF(s) and TVS-ADF(ss) always yield more precise forecasts than the competing ADF for different numbers of explained variables, numbers of explanatory variables, sparsity levels, and sample sizes. Moreover, our proposed models have higher predictive accuracy as the sparsity level or sample size increases. An empirical application to macroeconomic forecasting indicated that our model also captures the dynamic features of a real economic system better than its competitor. We will attempt to address the issue of the determination of the number of factors in our future research.

Chapter 3

Tying Maximum Likelihood Estimation with Selection of Tuning Parameter for Dependent Data

3.1 Introduction

In empirical applications, we usually face a type of irregular dependent data problem that most of the time series in a data set have long sample periods, while the others only have very short sample periods due to some reasons (e.g., different listing time of stocks, emerging indices, and severe data missing). This problem makes it very hard to get a reliable estimation result. For example, if we endeavor to use a vector auto-regression model to analyze the data of two stocks with different lengths where the data length of one stock is 500, while the other one only has 10 observations, there is no doubt that the maximum likelihood estimation using only the data of the short time series (i.e., 10 observations) hardly gives us good point estimates.

There are some studies that focus on small sample data and unequal-length time series. For instance, Hoyle (1999) summarized some statistical strategies for analyzing data from small samples, but these methods are mainly appropriate for independent identically distributed small-sample data with equal lengths. Baltagi and Song (2006) provided a survey for the treatment of unbalanced panel data, but these treatments rely on an error component regression model, which means that the application range of these methods is greatly limited; in addition, these methods are mainly used for data with large sample sizes. Van de Schoot and Miočević (2020) provided guidelines and tools for implementing solutions to issues that arise in small-sample research, but these methods mainly focus on small-sample data with equal lengths. It is obvious that these studies can not directly provide an effective solution for the aforementioned problem. Although the two-stage quasi-maximum likelihood estimation (2SQMLE, see White, 1996) could be a solution, it has at least two significant limitations: (1) it can not ensure the estimation consistency of parameters of two stocks, for example, if one uses the 2SQMLE to estimate a VAR(1) model where the coefficients matrix of the lag one is non-diagonal; (2) the estimation for the short time series in the second-stage estimation may not be obtained because the degrees of freedom are not enough. As an alternative, the method provided by Lynch and Wachter (2013) may be able to deal with our irregular data problem. This method is based on the generalized method of moments (GMM), and is close to our newly proposed tying maximum likelihood estimation (TMLE, introduced later) without tying.

Recently, the parameter tying technique (see Yan et al., 2015; Goodfellow et al., 2016; Luo et al., 2017) has enjoyed popularity in Few-shot Learning (see Wang et al., 2020) that plays an important role in tackling small sample data in machine learning (see Zhou, 2021). The main idea of the parameter tying method is to transfer some useful information from other relevant but different data to the target data that only have a few observations. We apply the idea of the parameter tying to the maximum likelihood estimation to solve the aforementioned irregular data problem.

For this motivation, we propose the TMLE by tying some parameters of the long time series with the corresponding parameters of the short time series. The form of tying depends on the form of a penalty term added in the traditional likelihood function. The strength of tying depends on a tuning parameter λ. We provide a selection method of λ based on a bootstrap procedure to improve the estimation performance effectively.

The contributions of this study are fourfold. First, adopting the idea of the parameter tying, we propose the TMLE, which is a pioneering work in this direction. The idea of the parameter tying, to the best of our knowledge, has never appeared in econometrics literature. Moreover, the TMLE can be widely applied in various fields, such as economics and finance, as it can be used directly as long as the likelihood functions of econometric or statistical models exist.

Second, we derived the asymptotic properties of the TMLE. Under some regularity conditions, the asymptotic theories show the convergence rate of the estimator and also the asymptotic normality with $\lambda = o(1/\sqrt{T})$.

Third, we derived the finite-sample risk bound of the proposed estimator. The theory shows that the risk bound depends on the tuning parameter, the form of tying, and some other parameters, such as the sample sizes of the long and short time series. In addition, this theory also provides evidence on the advantage of the TMLE relative to the traditional MLE.

Fourth, to reduce the risk of estimation, we propose a bootstrap procedure to select the tuning parameter that determines the strength of tying. Furthermore, we also provided the finite-sample theory of this bootstrap procedure, which shows how one should carry out this procedure effectively in practice.

The rest of the paper is organized as follows. Section 2 describes the TMLE in detail. Section 3 provides the asymptotic properties of the TMLE. Section 4 shows the risk bound of the TMLE. Section 5 describes the bootstrap procedure and provides the finite-sample theory for it. In sections 6 and 7, we carry out extensive artificial simulations and empirical applications to investigate the finite sample performance of the TMLE. All technical proofs and additional results of the artificial simulations and empirical applications are provided in Appendix B.

3.2 Tying Maximum Likelihood Estimator

3.2.1 Irregular Data Sets

Let us denote the n-dimensional time series by $r_t = (r_{1t}, \cdots, r_{nt})'$. In this paper, we allow for the situation that the sample period of a part of the n-dimensional time series is different from that of other entries of r_t. This study considers two sets of time series, where the sample period of one set is different from that of the other. Extension of the results to multiple sets of time series with different sample periods may be possible without any further difficulties.

To be more specific, let us consider two bundles $I_1 \subset I \equiv \{1, 2, \ldots, n\}$ and $I_2 \subset I$. Data for time series that belong to I_1 are available for $t = 1, \ldots, \tau$, while that in I_2 are available for $t = \tau + 1, \ldots, T$. I_1 and I_2 are allowed to have non-empty intersection, $I_1 \cup I_2 = I$, and each bundle is allowed not to be a subset of the other bundle. The cardinality of I_1 and I_2 are denoted as n_1 and n_2, respectively. The time series that belong to the bundle I_j is denoted as $r_{I_j,t}$ for $j = 1, 2$. Recall that data available for $r_{I_1,t}$ and $r_{I_2,t}$ are $t = 1, \ldots, \tau$ and $t = \tau + 1. \ldots, T$, respectively.

We consider the scenario that the observation for $r_{I_2,t}$ is short in the sense that a constant a exists such that $(T - \tau) = T^a$ with $0 < a \leq 1$, while τ is assumed to grow in the same order with T. Figure 3.1 illustrates a simple example of data availability.

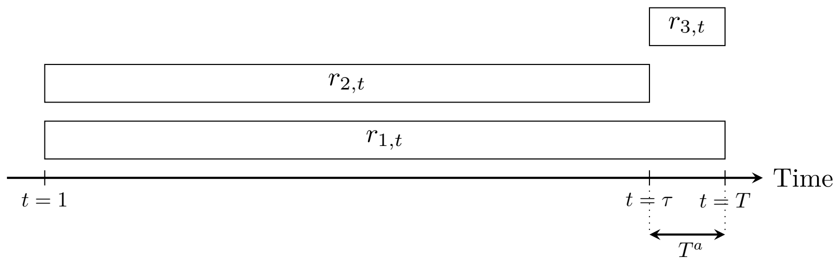


Figure 3.1: An example of data availability. In this example, r_t is a 3-dimensional time series ($n = 3$) with $I_1 = \{1, 2\}$, $I_2 = \{1, 3\}$, $r_{I_1,t} = (r_{1,t}, r_{2,t})'$, $r_{I_2,t} = (r_{1,t}, r_{3,t})'$, $n_1 = 2$, and $n_2 = 2$.

3.2.2 Quasi-Likelihood Function

Letting $\mathcal{F}_{t-1}$ be the sigma field generated by the past values of r_t, we denote the density functions of $r_{I_1,t}$ and $r_{I_2,t}$ conditional on $\mathcal{F}_{t-1}$ by $f_{1,t}(\theta_{I_1})$ and $f_{2,t}(\theta_{I_2})$, respectively, where θ_{I_1} and θ_{I_2} are K_1- and K_2-dimensional parameter vectors. Since I_1 and I_2 are allowed to have a non-empty intersection, θ_{I_1} and θ_{I_2} may have common parameters. Let $\check{\theta} \equiv \theta_{I_2} \setminus \theta_{I_1} \neq \emptyset$ and a parameter vector $\theta = (\theta'_{I_1}, \check{\theta}')'$ be K-dimensional.

The conditional log quasi-likelihood function is,

$$Q_T(\theta) = \frac{1}{T} \sum_{t=1}^{\tau} l_{I_1,t}(\theta_{I_1}) + \frac{1}{T} \sum_{t=\tau+1}^{T} l_{I_2,t}(\theta_{I_2}), \tag{3.1}$$

where $l_{I_1,t}(\theta_{I_1}) = -\log f_{1,t}(\theta_{I_1})$ and $l_{I_2,t}(\theta_{I_2}) = -\log f_{2,t}(\theta_{I_2})$.

When θ_{I_1} and θ_{I_2} have no common elements, the minimizer of $Q_T(\theta)$ is the standard maximum likelihood estimator (MLE) of parametric multivariate density models involving variables with histories of different lengths. When θ_{I_1} and θ_{I_2} have some common elements and the parameter of interest is θ_{I_2}

rather than θ_{I_1}, the minimizer of the second term is the familiar one-stage MLE (1SMLE) using only the overlapping data.

Another approach to estimate parameters is to employ two-stage quasi-maximum likelihood estimator (2SQMLE) (see White, 1996), when the multivariate model can be partitioned into elements relating only to the marginal distributions and elements only relating to the copula (see Patton, 2006 for 2SQMLE with time series of possibly different lengths). For all estimators mentioned above, precise estimation, especially for $\check{\theta}$, is unpromising, when the data available for the bundle I_2 is very short.

3.2.3 Tying Maximum Likelihood Estimator

Since the amount of data available for $r_{I_2,t}$ is short relative to that for $r_{I_1,t}$, less information are available for the estimation of $\check{\theta}$ compared to that of θ_{I_1}. To improve the finite sample performance for the estimation of $\check{\theta}$, we propose a novel estimation method that is inspired by the parameter tying technique (see Yan et al., 2015; Goodfellow et al., 2016; Luo et al., 2017) in Few-shot Learning (see Zhou, 2021). The proposed estimator transfers the information available for the estimation of θ_{I_1} to that of $\check{\theta}$ by imposing a penalty term on $Q_T\left(\theta\right)$. The penalized log-likelihood function is,

$$\begin{aligned} Q_\lambda(\theta) &= Q_T\left(\theta\right) + \lambda \left\|W'\theta\right\|^2 \\ &= \frac{1}{T}\sum_{t=1}^{\tau} l_{I_1,t}(\theta_{I_1}) + \frac{1}{T}\sum_{t=\tau+1}^{T} l_{I_2,t}(\theta_{I_2}) + \lambda \left\|W\theta\right\|^2, \end{aligned} \tag{3.2}$$

where $\lambda \geq 0$ is a tuning parameter that determines the scale of the penalty, W is a $m \times K$ restriction matrix that reflects prior information on the relationship among parameters, and $\|\cdot\|$ denotes the Euclidean norm.

The restriction matrix W consists of finite real numbers that are determined by researchers to introduce m restrictions on parameters. For example, to tie the first and the second element of θ, W will be the K-dimensional row vector such as $(1, -1, 0, \cdots, 0)$. Then the penalty term becomes $\lambda|\theta_1 - \theta_2|^2$, where θ_1 and θ_2 denotes, for now, the first and second element of θ. It may be possible to consider a tuning parameter, say λ_m. In this case, the penalty term becomes $\sum_{l=1}^{m} \lambda_l |W_l \theta|$, where the W_l is K-dimensional row restriction vector. For simplicity, however, this study focus on the single tuning parameter described in equation (3.2).

The tying maximum likelihood estimator (TMLE) is defined by

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} Q_\lambda(\theta), \tag{3.3}$$

where $\Theta = \Theta_{I_1} \times \check{\Theta} \subset \mathbb{R}^K$ is the parameter space of $\theta = (\theta'_{I_1}, \check{\theta}')'$. Similarly, the parameter space of θ_{I_2} is denoted as Θ_{I_2}. We denote the TMLEs of θ_{I_1}, θ_{I_2}, and $\check{\theta}$ by $\hat{\theta}_{I_1}$, $\hat{\theta}_{I_2}$, and $\hat{\check{\theta}}$, respectively.

The TMLE is an estimator of the pseudo-true parameter vector

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} Q_p(\theta), \tag{3.4}$$

where

$$Q_p(\theta) = \frac{1}{T} \sum_{t=1}^{\tau} E[l_{I_1,t}(\theta_{I_1})] + \frac{1}{T} \sum_{t=\tau+1}^{T} E[l_{I_2,t}(\theta_{I_2})] \tag{3.5}$$

is the population version of $Q_T(\theta)$. The pseudo-true values of $\hat{\theta}_{I_1}$, $\hat{\theta}_{I_2}$, and $\hat{\check{\theta}}$ are denoted as $\theta_{I_1,0}$, $\theta_{I_2,0}$, and $\check{\theta}_0$, respectively.

3.2.4 Notations

In order to show the theoretical results below, let us introduce some notations. For any function g, let $\nabla_{\cdot} g$ and $\nabla_{\cdot\cdot} g$ denote the partial derivative and the second derivative of g with respect to $\cdot$, respectively. We denote $\nabla_x g(\tilde{x}) = \nabla_x g(x)|_{x=\tilde{x}}$. For any integer p, 0_p denotes p-dimensional column zero vector. The L_p-norm and supremum norm are denoted as $\|\cdot\|_p$ and $\|\cdot\|_\infty$, respectively. For any matrix A, $\|A\|$ denotes the Frobenius norm. Let C denote a universal constant, which may vary at each occurrence. For any symmetric matrix A, $\iota_{min}(A)$ and $\iota_{max}(A)$ denote the smallest and largest eigenvalue of A, respectively. For any positive integer a, let I_a denote the $a \times a$ identity matrix. Similarly, let 0_{ab} and 0_a denote the $a \times b$ zero matrix and a-dimensional zero vector, respectively. We define the $K \times K$ matrix I_H such that

$$I_H = \begin{pmatrix} I_{K_1} & 0_{K_1(K-K_1)} \\ 0_{(K-K_1)K_1} & T^{1-a} I_{K-K_1} \end{pmatrix}. \tag{3.6}$$

Let $U_{1,t}(\theta)$ be the K_1-dimensional vector such that $U_{1,t}(\theta) = \nabla_{\theta_{I_1}} l_{I_1,t}(\theta_{I_1})$ for $t = 1, \cdots, \tau$ and $U_{1,t}(\theta) = \nabla_{\theta_{I_1}} l_{I_2,t}(\theta_{I_2})$ for $t = \tau+1, \cdots, T$. Let $U_{2,t}(\theta) = \nabla_{\breve{\theta}} l_{I_2,t}(\theta_{I_2})$ be a $(K-K_1)$-dimensional vector for $t = \tau+1, \cdots, T$.

Then, we define $\Sigma = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} E\left[U_t(\theta_0) U_s(\theta_0)'\right]$, where $U_t(\theta_0)$ is the K-dimensional vector such that $U_t(\theta_0) = (U_{1,t}(\theta_0)', 0_{K-K_1}')'$ for $t = 1, \cdots, \tau$ and $U_t(\theta_0) = (U_{1,t}(\theta_0)', \sqrt{T}/\sqrt{T_s} U_{2,t}(\theta_0)')'$, where $T_s \equiv T - \tau$, for $t = \tau+1, \cdots, T$ [1].

[1] Let

$$\begin{aligned}
\Sigma_1 &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} E\left[U_{1,t}(\theta_0) U_{1,s}(\theta_0)'\right], \\
\Sigma_2 &= \lim_{T\to\infty} \frac{1}{T^a} \sum_{t=\tau+1}^{T} \sum_{s=\tau+1}^{T} E\left[U_{2,t}(\theta_0) U_{2,s}(\theta_0)'\right], \\
\Sigma_{12} &= \lim_{T\to\infty} \frac{1}{\sqrt{T}\sqrt{T_s}} \sum_{t=1}^{T} \sum_{s=\tau+1}^{T} E\left[U_{1,t}(\theta_0) U_{2,s}(\theta_0)'\right], \\
\Sigma_{21} &= \lim_{T\to\infty} \frac{1}{\sqrt{T}\sqrt{T_s}} \sum_{t=1}^{T} \sum_{s=\tau+1}^{T} E\left[U_{2,t}(\theta_0) U_{1,s}(\theta_0)'\right].
\end{aligned} \tag{3.7}$$

Then, Σ is the $K \times K$ matrix

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix}. \tag{3.8}$$

Let $\nabla_\theta Q_p(\theta) \equiv \frac{1}{T}\sum_{t=1}^{\tau} E[\nabla_\theta l_{I_1,t}(\theta_{I_1})] + \frac{1}{T}\sum_{t=\tau+1}^{T} E[\nabla_\theta l_{I_2,t}(\theta_{I_2})]$ and $\nabla_{\theta\theta'} Q_P(\theta) \equiv \frac{1}{T}\sum_{t=1}^{\tau} E[\nabla_{\theta\theta'} l_{I_1,t}(\theta_{I_1})] + \frac{1}{T}\sum_{t=\tau+1}^{T} E[\nabla_{\theta\theta'} l_{I_2,t}(\theta_{I_2})]$.

3.3 Asymptotic Properties

We make the following assumptions.

Assumption 1 (Data). *For all $\theta \in \Theta$, $l_{I_1,t}(\theta_{I_1})$ and $l_{I_2,t}(\theta_{I_2})$ are $\mathcal{F}_t$-measurable. The process $\{r_s\}_{s=1}^{t}$ is a stationary strong mixing with mixing coefficients $\alpha(\cdot)$, where $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $0 < \rho < 1$.*

Assumption 2 (Parameter). *The parameter spaces Θ, Θ_{I_1}, and Θ_{I_2} are compact and convex subset of $\mathbb{R}^K$, $\mathbb{R}^{K_1}$, and $\mathbb{R}^{K_2}$, respectively. The true value θ_0 defined in equation (3.4) is unique and lies in the interior of Θ and satisfies $\nabla_\theta Q_p(\theta_0) = 0_K$.*

Assumption 3 (Model). *(1) $Q_T(\theta)$ is two-times continuously differentiable with respect to θ.*

(2) There exists a measurable function l_t such that, for all $j = 1, 2$ and $k, k' = 0, 1, \cdots, K$, $|\nabla_{\theta_k\theta_{k'}} l_{I_j,t}(\theta_{I_j}) - \nabla_{\theta_k\theta_{k'}} l_{I_j,t}(\bar{\theta}_{I_j})| < \|\theta_{I_j} - \bar{\theta}_{I_j}\| l_t$ for any $\theta_{I_j}, \bar{\theta}_{I_j} \in \Theta_{I_j}$, $\sup_{\theta_{I_j}\in\Theta_{I_j}} |\nabla_{\theta_k\theta_{k'}} l_{I_j,t}(\theta_{I_j})| \leq l_t$ and $E(|l_t|^q) < c_l$ for some constant $c_l < \infty$ and some $q > \max\{K_1 + 1, 4, K_2 + a\}$.

(3) $\nabla_{\theta\theta'} Q_P(\theta_0)$ exists and $\iota_{min}(I_H \nabla_{\theta\theta'} Q_P(\theta_0)) \geq c_H$ for a constant $c_H > 0$.

(4) There exists $H \equiv \lim_{T\to\infty}[I_H \nabla_{\theta\theta'} Q_P(\theta_0)]$ and $H > 0$.

(5) There exists Σ and $\Sigma > 0$.

Under Assumption 1, the density functions $l_{I_1,t}(\theta_{I_1})$ and $l_{I_2,t}(\theta_{I_2})$ are stationary strong mixing (e.g. Theorem 14.1 of Davidson (1994)). Note that, under

Assumptions 1, 2, and 3, the first and second derivatives of the log-likelihood functions are also stationary strong mixing processes because each of them is a measurable function of a stationary strong mixing process. As a matter of fact, many classical econometric models satisfy this assumption, such as the VAR model (see Yin, 2019), the GARCH BEKK model (see Comte and Lieberman, 2003), and the VEC model (see Hafner and Preminger, 2009).

Most of Assumptions 2 and 3 are standard assumptions for consistency and asymptotic normality of MLEs (e.g, Hayashi, 2000) adjusted for the models for dependent processes with different lengths. Assumption 3 (2) is about the smoothness and moment conditions on the objective function, which is commonly assumed for penalized estimator (see Su et al., 2016).

The population variant of the likelihood function defined in the equation (3.5) implies that the second term relating to the process $r_{I_2,t}$ is asymptotically negligible. To make the asymptotic properties of estimates relating only to the process $r_{I_2,t}$ non-negligible, the matrix I_H is introduced. Using the matrix, we are able to consider asymptotic properties of TMLE that reveal the consequences of having shorter sampling periods for $r_{I_2,t}$. This is reflected in Assumptions 3 (3) and (4), in which the Hessian is multiplied by I_H.

Lemma 1. ***(Consistency)*** *Suppose that Assumptions 1, 2, and 3 hold. For any δ, $P(\|\hat{\theta} - \theta_0\| > \delta) = o(T^{-1})$, when $\lambda \to 0$ as $T \to \infty$.*

Next, we consider the convergence rate of the estimator.

Theorem 2. *Suppose that Assumptions 1, 2, and 3 hold. Then,*

$$\hat{\theta} - \theta_0 = O_p(T^{1-\frac{3}{2}a}) + O_p(T^{1-a}\lambda).$$

The TMLE $\hat{\theta}$ has no asymptotic normality as long as the value of a that represents the sampling periods for $r_{I_2,t}$, that is, $(T - \tau) = T^a$ is not equal to 1.

This is because both the estimator and true value for $\check{\theta}$ vanishes with $T \to \infty$ for $a \neq 1$, which are implied by the likelihood functions (3.1) and (3.5).

For the asymptotic normality, let $a = 1$. This setting does not mean that the difference in data lengths is asymptotically negligible. Letting $T_s \equiv T - \tau$, we consider the case that $T_s/T \to \zeta$ for some $0 < \zeta \leq 1$. As shown below, the difference of data lengths matters as long as $\zeta \neq 1$ (see Patton, 2006 for asymptotic normality of 2SMLE with time series of different lengths).

Theorem 3. *Suppose that Assumptions 1, 2, and 3 hold. When* $\lambda = o(T^{-\frac{1}{2}})$*,*

$$\Sigma^{-1/2} H \mathbb{W}_T(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I_K),$$

where $\mathbb{W}_T$ *is the* $K \times K$ *diagonal matrix whose first* K_1 *diagonal elements are* $T^{1/2}$ *and the remaining* $K - K_1$ *diagonal elements are* $T_s^{1/2}$*.*

Following White (1996) (p.91), we say that C^* is the asymptotic covariance matrix of $\hat{\theta}$, when $(C^*)^{-1/2}\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I)$, which is denoted as $\mathrm{avar}(\hat{\theta}) = C^*$. For the TMLE, we have $\Sigma^{-1/2} H \mathbb{W}\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I)$, where $\mathbb{W}$ is the limit of $\mathbb{W}_T/\sqrt{T}$, that is,

$$\begin{aligned} \mathbb{W}_T/\sqrt{T} &= \begin{pmatrix} I_{K_1} & 0_{K_1(K-K_1)} \\ 0_{(K-K_1)K_1} & (T_s/T)^{1/2} I_{K-K_1} \end{pmatrix} \\ &\to \begin{pmatrix} I_{K_1} & 0_{K_1(K-K_1)} \\ 0_{(K-K_1)K_1} & \zeta^{1/2} I_{K-K_1} \end{pmatrix} = \mathbb{W}. \end{aligned} \tag{3.9}$$

Thus, asymptotic covariance matrix of the TMLE is,

$$\mathrm{avar}(\hat{\theta}) = (\mathbb{W}^{-1})'(H^{-1})'\Sigma H^{-1}\mathbb{W}^{-1}. \tag{3.10}$$

3.4 Risk Bound

This section shows the non-asymptotic property of the TMLE. The TMLE may be expressed as

$$\begin{aligned}\hat{\theta} &= \underset{\theta\in\Theta}{\operatorname{argmin}}\left\{Q_T(\theta) + \lambda\,\|W\theta\|^2\right\}\\ &= \underset{\theta\in\Theta}{\operatorname{argmin}}\left\{Q_T(\theta) - Q_p(\theta_0) + \lambda(\|W\theta\|^2 - \|W\theta_0\|^2)\right\},\end{aligned} \tag{3.11}$$

implying that the aim of the TMLE is to have the population version of the loss evaluated by itself, that is,

$$Q_p(\hat{\theta}) - Q_p(\theta_0) + \lambda\|W\hat{\theta}\|^2 - \lambda\|W\theta_0\|^2 \tag{3.12}$$

to be minimized. An upper bound of equation (3.12) that holds probability close to one is called risk bound, which reveals the non-asymptotic property of the estimator.

To do this, let us consider the parameter spaces $\Theta_\delta \equiv \{\theta \in \Theta : \|W'W(\theta - \theta_0)\| \leq \delta\}$ and $\tilde{\Theta}_\delta \equiv \{\theta_{I_1} \in \Theta_{I_1} : \theta = (\theta'_{I_1}, \check{\theta}')' \in \Theta_\delta\}$. Lemma 4 shows the parameter spaces to which the TMLE $\hat{\theta}$ belongs.

Lemma 4. *Let the estimator $\hat{\theta} = (\hat{\theta}'_{I_1}, \hat{\check{\theta}}')' \in \Theta$ of θ defined in equation (3.3) exists. Under Assumptions 1, 2, and 3, $\hat{\theta}_{I_1} \in \tilde{\Theta}_{\delta_\lambda}$ and $\hat{\theta} \in \Theta_{\delta_\lambda}$ with probability $1 - \epsilon_S$ for arbitrary small $\epsilon_S > 0$ and any $T \geq 2$, where $\delta_\lambda \equiv \frac{S}{2\lambda} + \|W'W\theta_0\|$ for some constant S.*

Remark 5. Lemma 4 shows that the penalized estimator $\hat{\theta}$ defined in equation (3.3) belongs to the restricted parameter space Θ_{δ_λ}, which is a subset of the entire parameter space Θ. Intuitively, this occurs because the parameters are tied in their estimation, which is reflected through some interesting features of the restricted parameter space Θ_{δ_λ}. First, the restrictions are set on the norm of $W'W(\theta - \theta_0)$, which depends on the weight W, rather than the norm of $\theta - \theta_0$ itself. This comes from the construction of the objective function in equation (3.3), in which parameters are tied with the weight. It indicates that the restriction is imposed only on parameters that are penalized with non-zero weights. Second, the complexity of the penalized parameter space depends on the tuning parameter λ. Furthermore, δ_λ is a decreasing function of λ, which can be close to zero when parameters are tied correctly, that is, $W\theta_0 = 0$. A larger value of λ indicates that the estimator belongs to a more restricted parameter space, which can be a strict subset of the entire parameter space Θ.

As shown in the proof of Lemma 4, S that appears in δ_λ is a positive constant such that $P(\|\nabla_\theta Q_T(\hat{\theta})\| > S) \leq \epsilon_S$ holds for any $T \geq 2$ and ϵ_S that can be arbitrarily small by taking S large. The size of S that makes ϵ_S small depends on the dependence of the observed processes, size of the parameter space Θ_{I_1} and Θ_{I_2}, existence of the moments of $\nabla_\theta Q_T(\theta)$, and smoothness of the likelihood function $Q_T(\theta)$ with respect to the parameter.

Adding the penalty to objective functions introduces a finite sample bias in the estimator. In contrast to this, the penalty restricts the parameter space to which the estimator belongs, as shown in Lemma 4. This trade-off makes room for an improvement of the finite-sample performance of the penalized estimator. The non-asymptotic property of the TMLE in terms of the risk bound is shown in the following theorem.

Theorem 6. *Suppose Assumptions 1, 2, and 3. In addition, we assume that conditional densities of $r_{I_1,t}$ and $r_{I_2,t}$ are bounded from above and away from 0 so that a constant l exists such that $|l_{I_j,t}(\cdot)| < l$ for $j = 1, 2$. For a fixed $\lambda > 0$, any $c > 0$ and arbitrary small $\epsilon_S > 0$, and all $\tau, T^a \geq$*

$\max\{-\log\rho/8, 2^7(-\log\rho)^{-1}, 2\}$, *the probability that*

$$\lambda\|W\hat{\theta}\|^2 + Q_p(\hat{\theta}) - Q_p(\theta_0) \leq \lambda\|W\theta_0\|^2 + \left(\frac{\tau^{3/4}}{T} + \frac{T^{3a/4}}{T}\right)\kappa\sqrt{c\tilde{\rho}\bar{K}c_l} + \left(\frac{\tau^{1/2}}{T} + \frac{T^{\frac{a}{2}}}{T}\right)\frac{2c\tilde{\rho}(l+c_l)}{3} \qquad (3.13)$$

is not less than $1 - 4(1+4e^{-2}c_\alpha)e^{-c} - 4\epsilon_S$, *where* $\kappa \equiv \sup_{\theta\in\Theta_{\delta_\lambda}}\|\theta-\theta_0\|$, $\tilde{\rho} \equiv (-8^3/\log\rho)^{1/2}$, *and* $\bar{K} \equiv \max\{K_1, K_2\}$.

Theorem 6 shows the non-asymptotic property of the penalized estimator $\hat{\theta}$. The results implies that, for a fixed $\lambda > 0$ and all $\tau, T^a \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$, the probability that

$$Q_p(\hat{\theta}) - Q_p(\theta_0) \leq B_T(\lambda) + V_T(\lambda), \qquad (3.14)$$

can be arbitrary close to one by taking c large and ϵ_S small, where

$$B_T(\lambda) \equiv \lambda\|W\theta_0\|^2, \qquad (3.15)$$

and

$$V_T(\lambda) \equiv \left(\frac{\tau^{3/4}}{T} + \frac{T^{3a/4}}{T}\right)\kappa\sqrt{c\tilde{\rho}\bar{K}c_l} + \left(\frac{\tau^{1/2}}{T} + \frac{T^{\frac{a}{2}}}{T}\right)\frac{2c\tilde{\rho}(l+c_l)}{3}. \qquad (3.16)$$

The size of $V_T(\lambda)$ relates to the number of parameters through $\bar{K} \equiv \max\{K_1, K_2\}$, the dependence of the sequence through $\tilde{\rho} \equiv (-8^3/\log\rho)^{1/2}$ in Assumption 1, and the existence of moments through c_l in Assumption 3 (2). Moreover, the size of $V_T(\lambda)$ relates to the size of the restricted parameter space Θ_{δ_λ} through $\kappa = \sup_{\theta\in\Theta_{\delta_\lambda}}\|\theta-\theta_0\|$. Thus, via S in δ_λ, the size of $V_T(\lambda)$ also relates to the smoothness as well as boundedness of the density functions of $r_{I_1,t}$ and $r_{I_2,t}$.

For $\lambda = 0$, that is, standard MLEs, Theorem B.1 shows that the right-hand side of the equation (3.14) turns out to be

$$V_T(0) \equiv \left(\frac{\tau^{3/4}}{T} + \frac{T^{3a/4}}{T}\right)\bar{\kappa}\sqrt{c\tilde{\rho}\bar{K}c_l} + \left(\frac{\tau^{1/2}}{T} + \frac{T^{\frac{a}{2}}}{T}\right)\frac{2c\tilde{\rho}(l+c_l)}{3}, \tag{3.17}$$

where $\bar{\kappa} \equiv \sup_{\theta\in\Theta}\|\theta - \theta_0\|$ measures the size of the entire parameter space Θ.

Comparing the upper bound for the TMLEs in (3.14) and that for standard MLEs in (3.17) reveals potential finite-sample advantages of TMLEs. When parameters are tied correctly, that is, $W\theta_0 = 0$, we have $B_T(\lambda) = 0$. Moreover, $V_T(\lambda) \leq V_T(0)$ for any λ because $\kappa \leq \bar{\kappa}$ by the restricted parameter spaces. When parameters are not tied correctly, that is, $W\theta_0 \neq 0$, $\Lambda \equiv \{\lambda : B_T(\lambda) + V_T(\lambda) < V_T(0)\}$ may be non-empty, especially when the densities of r_{I_1} and r_{I_2} are smooth so that κ is small relative to $\bar{\kappa}$.

It also reveals that for a fixed T, the finite-sample advantages of TMLEs over the standard MLEs can be large when $T - \tau$ is small. As discussed above, the advantages of TMLEs over the standard MLEs come from the restricted parameter space that makes the first term of $V_T(\lambda)$ smaller than that of $V_T(0)$ through $\kappa \leq \bar{\kappa}$. Note that the second term in $V_T(\lambda)$ and $V_T(0)$ are the same. Moreover, for $\tau > T/2$ with a fixed T, $(\tau^{3/4} + T^{3a/4})/(\tau^{1/2} + T^{a/2})$ increases with τ. Thus, the relative size between κ and $\bar{\kappa}$ become more important for larger τ, i.e., smaller $T - \tau$.

Let us consider the value of λ that is optimal in the sense that it minimizes $B_T(\lambda) + V_T(\lambda)$. The following remark considers the optimal value of λ, denoted as λ^*, to investigate the convergence rate of the optimal tuning parameter.

Remark 7. For simplicity, we consider the case that W is an identity matrix. In this case, we can obtain the upper bound of the inequality that is the same with those in (3.14) except that κ in $V_T(\lambda)$ is replaced with δ_λ.[2] Then, the optimal

[2]We obtain the same convergence rate of the optimal tuning parameter for any W by considering an oracle inequality that may be less sharp than that in (3.14) (see, Theorem B.2). The upper bound of the inequality is the same as those in (3.14) except that κ in $V_T(\lambda)$

tuning parameter that minimizes the upper bound of the oracle inequality is

$$\lambda^* = \arg\min_{\lambda>0}\{B_T(\lambda) + V_T(\lambda)\} = \frac{(S\sqrt{c\tilde{\rho}\bar{K}c_l})^{1/2}}{\|W\theta_0\|}\left(\frac{\tau^{3/4}}{T} + \frac{T^{3a/4}}{T}\right)^{1/2}, \quad (3.18)$$

implying that $\lambda^* = O(T^{-\frac{1}{8}})$. By replacing λ in the right hand side of equation (3.14) with λ^*, we can show that a constant $\bar{C}$ that is independent of T exists, such that,

$$P\left(Q_p(\hat{\theta}) - Q_p(\theta_0) \leq \bar{C}T^{-\frac{1}{8}}\right) \geq 1 - 4(1 + 4e^{-2}c_\alpha)e^{-c} - 4\epsilon_S. \quad (3.19)$$

Following the definition in Hearst et al. (1998), the rate of $T^{-\frac{1}{8}}$ in equation (3.19) is called the learning rate because it tells how well the method has learned (in terms of the Kullback-Leibler information criterion) from the given data of fixed length T. Equation (3.17) implies that the learning rate of the standard MLE is $T^{-\frac{1}{4}}$.

3.5 Selection of λ based on bootstrap

If the restriction matrix W set by users is absolutely correct, they can take λ large sufficiently to estimate parameters and this choice is reasonable according to Lemma 4 and Theorem 6. But it will become very hard for users to select an apt λ in practice if the restriction matrix is not completely right because, in this case, the performance of the TMLE for different λ is not always better than that of the MLE according to Theorem 6. Hence, it is necessary to provide a feasible and effective method to help users select a suitable λ to reduce the risk when they are not sure whether the restriction matrix W is completely correct.

In this study, we provide an effective bootstrap procedure to address this issue. Before we formally describe this bootstrap procedure, we need to introduce

is replaced by $\delta_\lambda + \omega\kappa$, where $\omega \equiv \|I_k - W'W\|$ is zero when W is an identity matrix.

some new notations and concepts. Specifically, we rewrite $Q_T(\theta)$ as $Q_T(\theta, X_T)$, where $X_T = (r_1, \cdots, r_T)$ belonging to $\mathscr{X}_T \subset \mathbb{R}^{n\times T}$ is a random variable that indicates the collection of sample data. Then we regrad $Q_T(\theta, X_T)$ as the loss function; in addition, we consider $Q_T(\theta, X_T)$ to be the true or a misspecified log-likelihood function. Let P_o denote the true distribution of data generating process. Theoretically (ideally), one seeks an optimal $\theta_0 \in \Theta$ that minimizes the expected loss function $Q_{P_o}(\theta)$, that is,

$$\theta_0 = \arg\min_{\theta\in\Theta} Q_{P_o}(\theta) = \arg\min_{\theta\in\Theta} \int Q_T(\theta, x_T) dP_o,$$

where the variable of integration is x_T.

As for the TMLE, since we consider the penalty term $\lambda \left\|W\theta\right\|^2$, the estimate of θ depends on λ for a given W. Note that here we assume that W is given (it can be correct or not) and our target is to select an apt λ. We rewrite λ as λ_T, which is intended to reflect that one can set different λ for different sample sizes T, and suppose that $\lambda_T \in [0, c]$ with $c \geq 0$. In addition, we also consider λ_T in a discrete set

$$\Lambda_T = \left\{0, \frac{c}{K(T)}, \frac{2c}{K(T)}, \cdots, \frac{[K(T)-1]c}{K(T)}, c\right\},$$

where $K(T)$ is positive real number. For a given λ_T and a sample data, one can obtain the point estimate of θ, which is denoted by $\hat{\theta}_{\lambda_T}$; this means that for $K(T)$ different λ_T, one can have $K(T)$ estimation values of θ. Then, to evaluate which λ_T is better, we can use new datasets to test. For example, there are two different λ_T: $\lambda_{T,1}$ and $\lambda_{T,2}$, then we can have two estimates of θ: $\hat{\theta}_{\lambda_{T,1}}$ and $\hat{\theta}_{\lambda_{T,2}}$. Next we calculate the values of $Q_T(\hat{\theta}_{\lambda_{T,1}}, x_T^*)$ and $Q_T(\hat{\theta}_{\lambda_{T,2}}, x_T^*)$ where x_T^* denotes additional new data. Then we select the λ_T corresponding to a smaller one of these two values. Formally, for any given λ_T (which means that

$\hat{\theta}_{\lambda_T} \in \Theta$ is given), one selects optimal $\check{\lambda}_T \in \Lambda_T$ and $\tilde{\lambda}_T \in [0, c]$ that minimize the expected loss function $Q_{P_o}(\lambda_T)$, that is,

$$Q_{P_o}(\lambda_T) = \int Q_T(\hat{\theta}_{\lambda_T}, x_T) dP_o,$$

$$\check{\lambda}_T = \underset{\lambda_T \in \Lambda_T}{\arg\min}\, Q_{P_o}(\lambda_T),$$

$$\tilde{\lambda}_T = \underset{\lambda_T \in [0,c]}{\arg\min}\, Q_{P_o}(\lambda_T),$$

where the variable of integration is only x_T, while $\hat{\theta}_{\lambda_T}$ is fixed (given). However, generally speaking, P_o is unknown, which means that one can not calculate the expected loss function above. Hence, we provide a bootstrap procedure as an alternative.

Bootstrap procedure Suppose that the model used by users for $r_1, \cdots, r_T$ can be written as

$$r_t = m\left(r_{t-1}, \cdots, r_{t-p}\right) + \varepsilon_t, \;\; \varepsilon_t \sim D,$$

where $m(\cdot)$ denotes a parametric function and ε_t means the error term following a distribution D. Giving a value of λ_T, we can obtain an estimate $\hat{\theta}_{\lambda_T}$. Then based on the estimate $\hat{\theta}_{\lambda_T}$, we can generate a new sequence, $X_T^b = (r_1^b, \cdots, r_T^b) \in \mathscr{X}_T^b \subset \mathbb{R}^{n \times T}$, using

$$r_t^b = \hat{m}\left(r_{t-1}, \cdots, r_{t-p}\right) + \varepsilon_t^b, \;\; \varepsilon_t^b \sim \hat{D},$$

where $\hat{m}\left(r_{t-1}, \cdots, r_{t-p}\right)$ denotes the fitted value of r_t and $\hat{D}$ means the distribution D with the estimates of its parameters. We call X_T^b as a *bootstrap sequence*. Then we generate B bootstrap sequences in this fashion. Note that one also can use other bootstrap schemes (e.g., the fixed-design wild bootstrap proposed by Gonçalves and Kilian (2004)) to generate B bootstrap sequences.

These bootstrap sequences are i.i.d. with a certain but unknown distribution, which is denoted by $P_T^b(\lambda_T)$. Then one selects an optimal $\hat{\lambda}_T \in \Lambda_T$ that minimizes the bootstrap expected loss function $Q_{P_T^b}(\lambda_T)$, that is,

$$Q_{P_T^b}(\lambda_T) = \int Q_T(\hat{\theta}_{\lambda_T}, x_T) dP_T^b(\lambda_T),$$
$$\hat{\lambda}_T = \underset{\lambda_T \in \Lambda_T}{\arg\min}\, Q_{P_T^b}(\lambda_T),$$

where the variable of integration is only x_T. Since $P_T^b(\lambda_T)$ is unknown, we use the empirical distribution of B bootstrap sequences, which is denoted by $P_T^B(\lambda_T)$. Then the bootstrap average loss function is defined as

$$\begin{aligned} Q_{P_T^B}(\lambda_T) &= \int Q_T(\hat{\theta}_{\lambda_T}, x_T) dP_T^B(\lambda_T) \\ &= \frac{1}{B} \sum_{i=1}^{B} Q_T(\hat{\theta}_{\lambda_T}, X_{T,i}^b), \end{aligned}$$

where $X_{T,i}^b$ refers to i-th bootstrap sequence, and its minimizer is denoted by

$$\bar{\lambda}_T = \underset{\lambda_T \in \Lambda_T}{\arg\min}\, Q_{P_T^B}(\lambda_T). \tag{3.20}$$

Now, we can use (3.20) to select an apt λ_T in practice.

This is the bootstrap procedure for the TMLE and we call this bootstrap procedure *TMLE bootstrap*. Next, we will present some interesting theorems about the TMLE bootstrap. It is worth mentioning that some of these theorems are also applicable to other bootstrap schemes that can generate B bootstrap sequences.

Theorems for bootstrap To show the finite-sample theorem for the TMLE bootstarp, we need the definition of Generalized entropy with bracketing. Consider a probability triple $(\Omega, \mathscr{F}, P)$, and sub-sigma algebras $\mathscr{F}_0 \subset \mathscr{F}_1 \subset \cdots \subset \mathscr{F}_T \subset \cdots \subset \mathscr{F}$. Let $Z_t = Z_t(\lambda)$ be $\mathscr{F}_t$-measurable random variables, $t = 1, \cdots, T, \cdots$, depending on a parameter $\lambda \in \Lambda$. Let M be a real positive constant. Define

$$\rho_M(Z_t) = 2M^2 E\left(e^{|Z_t|/M} - 1 - |Z_t|/M \big| \mathscr{F}_{t-1}\right), \;\; t = 1, \cdots, T,$$

and for $Z(\lambda) = (Z_1, \cdots, Z_T)$, write $\bar{\rho}_M^2(Z(\lambda)) = \frac{1}{T}\sum_{t=1}^{T} \rho_M(Z_t)$.

Definition 8. Generalized entropy with bracketing (refer to Definition 8.1 of Geer et al. (2000)): for $0 < \delta < \infty$, let $\left\{[Z_j^L, Z_j^U]\right\}_{j=1}^{\mathcal{N}}$ be a collection of pairs of random vectors $Z_j^L = (Z_{1,j}^L, \cdots, Z_{T,j}^L)$ and $Z_j^U = (Z_{1,j}^U, \cdots, Z_{T,j}^U)$, with $[Z_{t,j}^L, Z_{t,j}^U]$ $\mathscr{F}_t$-measurable, $t = 1, \cdots, T$, $j = 1, \cdots, \mathcal{N}$, such that for all $\lambda \in \Lambda$, there is a $j = j(\lambda) \in \{1, \cdots, \mathcal{N}\}$, with $\lambda \mapsto j(\lambda)$ non-random, such that

(i) $\bar{\rho}_M^2(Z_j^U - Z_j^L) \leq \delta^2$ on Ω,

(ii)$Z_{t,j}^L \leq Z_t \leq Z_{t,j}^U$, $t = 1, \cdots, T$, on Ω.

Let $\mathscr{N}_{Z(\lambda),M}(\delta, \Omega)$ be the smallest non-random value of $\mathcal{N}$ for which such a collection $\left\{[Z_j^L, Z_j^U]\right\}_{j=1}^{\mathcal{N}}$ exists. Then $\mathscr{H}_{Z(\lambda),M}(\delta, \Omega) = \log \mathscr{N}_{Z(\lambda),M}(\delta, \Omega)$ is called the generalized δ-entropy with bracketing.

Let $L_{I_1}(\lambda_T) = \left\{l_{I_1,1}(\hat{\theta}_{I_1,\lambda_T}), \cdots, l_{I_1,\tau}(\hat{\theta}_{I_1,\lambda_T})\right\}$ and $L_{I_2}(\lambda_T) = \left\{l_{I_2,\tau+1}(\hat{\theta}_{I_2,\lambda_T}), \cdots, l_{I_2,T}(\hat{\theta}_{I_2,\lambda_T})\right\}$. Applying Definition 8, we denote the generalized entropy with bracketing of $L_{I_i}(\lambda_T)$ by $\mathscr{H}_{L_{I_i}(\lambda_T),M}(\delta, \Omega_i)$, where Ω_i denotes the sample space satisfying $l_{I_i,t}(\hat{\theta}_{I_i,\lambda_T})$ w.r.t. $X_T : \Omega_i \to R_i \subset \mathbb{R}$, for $i = 1, 2$.

Theorem 9. *Suppose that for a given restriction matrix W and $\lambda_T \in [0, c]$, r_t has conditional log density function $l_{I_i,t}(\hat{\theta}_{I_i,\lambda_T})$ for $i = 1, 2$ such that*

$$\sup_{\hat{\theta}_{\lambda_T} \in \Theta,\, X_T^b \in \mathscr{X}_T^b} \left| Q_T(\hat{\theta}_{\lambda_T}, X_T^b) - Q_{P_T^b}(\lambda_T) \right| \leq C_2 < \infty \;\; a.s.,$$

$$\sup_{\hat{\theta}_{I_i,\lambda_T}\in\Theta_{I_i},\, X_T^{(\cdot)}} \left| l_{I_i,t}(\hat{\theta}_{I_i,\lambda_T}) \right| \leq M < \infty,$$

where $X_T^{(\cdot)} \in \{X_T, X_T^b\}$*,* $X_T \in \mathscr{X}_T$*, and* $X_T^b \in \mathscr{X}_T^b$*. Suppose that the real-valued function* $Q_{P_o}(\lambda_T)$ *is Lipschitz continuous. If* $\mathscr{H}_{L_{I_i}(\lambda_T),M}(\delta, \Omega_i^{(\cdot)})$ *exists, where* $\Omega_i^{(\cdot)} \in \{\Omega_i, \Omega_i^b\}$ *and* Ω_i^b *denotes the sample space satisfying* $l_{I_i,t}(\hat{\theta}_{I_i,\lambda_T})$ *w.r.t.* $X_T^b : \Omega_i^b \to R_i^b \subset \mathbb{R}$*, for* $i = 1, 2$*, then we have the following finite-sample result*

$$\begin{aligned}
0 \leq & E_{P_o} Q_{P_o}(\bar{\lambda}_T) - Q_{P_o}(\theta_0) \\
\leq & E_{P_o} Q_{P_o}(\tilde{\lambda}_T) - Q_{P_o}(\theta_0) \\
& + O\left(\frac{K(T)\sqrt{\tau}}{T}\right) + O\left(\frac{K(T)\sqrt{T-\tau}}{T}\right) + O\left(\frac{K(T)}{\sqrt{B}}\right) + O\left(\frac{c}{K(T)}\right) \\
& + A,
\end{aligned} \tag{3.21}$$

where

$$\begin{aligned}
A = & \frac{1}{T} E_{P_o} \int \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) \right] + \sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_2,t}(\hat{\theta}_{I_2,\bar{\lambda}_T}) \right] dP_o \\
& - \frac{1}{T} E_{P_o} \int \sum_{t=1}^{\tau} E_{P^b_{T|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) \right] + \sum_{t=\tau+1}^{T} E_{P^b_{T|\mathscr{F}_{t-1}}} \left[l_{I_2,t}(\hat{\theta}_{I_2,\bar{\lambda}_T}) \right] dP_T^b \\
& + \frac{1}{T} E_{P_o} \int \sum_{t=1}^{\tau} E_{P^b_{T|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\tilde{\lambda}_T}) \right] + \sum_{t=\tau+1}^{T} E_{P^b_{T|\mathscr{F}_{t-1}}} \left[l_{I_2,t}(\hat{\theta}_{I_2,\tilde{\lambda}_T}) \right] dP_T^b \\
& - \frac{1}{T} E_{P_o} \int \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\tilde{\lambda}_T}) \right] + \sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_2,t}(\hat{\theta}_{I_2,\tilde{\lambda}_T}) \right] dP_o.
\end{aligned}$$

Note that the third to sixth terms of the right-hand side of the inequality (3.21) all contain $K(T)$ (reflecting the number of λ_T users set over the interval $[0, c]$), which measures the bias of the point estimates of parameters caused by λ_T. Specifically, the third to fifth terms imply that the larger $K(T)$ is, the higher the risk is; however, the sixth term indicates that the larger $K(T)$ is, the smaller the bias is. Hence, users need to build a trade-off in practice when they set the number of λ_T. Now let us consider the other elements in these terms except for $K(T)$. First, the third and fourth terms measure the bias of point estimates of parameters caused by different lengths of the time series, i.e., T, τ, and $T-\tau$. Moreover, the fifth term measures the bias caused by the number of bootstrap sequences generated, which implies that one should take B sufficiently large. The last term measures the bias caused by c, the restriction matrix W, the bootstrap scheme, and whether the model is misspecified. There are several remarks about the last term. c and W can affect the estimates of parameters (thus affect P_T^b). If c is small, it probably causes A to be large. If W is not fully correct, then A may become large. But since $\lambda_T \to 0$ as $T \to \infty$, the penalty term $\lambda_T \|W\theta\|^2$ will disappear gradually whatever W is correct or not. In addition, if the bootstrap scheme one uses can not ensure $P_T^b \to P_o$ as $T \to \infty$ even if the model specification is correct, then the last term will not disappear. If the model is misspecified, then we believe that this term also is likely not to disappear as $T \to \infty$ no matter what the bootstrap scheme is.

Theorem 9 is general and it can directly apply to the bootstrap schemes that can generate B bootstrap sequences. According to (3.18) in section 4, we can take $c = O(1/\sqrt[8]{T})$. Then we have a general corollary, which is also applicable for the bootstrap schemes that can generate B bootstrap sequences, as follows.

Corollary 10. *Suppose that assumptions of Theorem 9 hold and take* $c = O(1/\sqrt[8]{T})$, *then we have the following finite-sample result*

$$
\begin{aligned}
0 \leq & E_{P_o} Q_{P_o}(\bar{\lambda}_T) - Q_{P_o}(\theta_0) \\
\leq & E_{P_o} Q_{P_o}(\tilde{\lambda}_T) - Q_{P_o}(\theta_0) \\
& + O\left(\frac{K(T)\sqrt{\tau}}{T}\right) + O\left(\frac{K(T)\sqrt{T-\tau}}{T}\right) + O\left(\frac{K(T)}{\sqrt{B}}\right) + O\left(\frac{1}{\sqrt[8]{T}K(T)}\right) \\
& + A.
\end{aligned}
$$

The explanation of each term in the inequality above is similar to (3.21). Now let us consider what conditions can make the last term A in (3.21) disappear. Theorem 11 provides an answer.

Theorem 11. *Suppose that the assumptions of Theorem 9 and Lemma 1 hold and $c \to 0$ as $T \to \infty$. If the model specification is correct and we use the TMLE bootstrap procedure to generate the bootstrap sequences, then we have the following result*

$$\begin{aligned}0 \leq & E_{P_o} Q_{P_o}(\bar{\lambda}_T) - Q_{P_o}(\theta_0) \\ \leq & E_{P_o} Q_{P_o}(\tilde{\lambda}_T) - Q_{P_o}(\theta_0) \\ & + O\left(\frac{K(T)\sqrt{\tau}}{T}\right) + O\left(\frac{K(T)\sqrt{T-\tau}}{T}\right) + O\left(\frac{K(T)}{\sqrt{B}}\right) + O\left(\frac{c}{K(T)}\right) \\ & + o_p(1).\end{aligned}$$

This theorem shows that A will disappear in probability as $T \to \infty$ if the model specification is correct and we use the TMLE bootstrap with $c = o(1)$. In addition, we can take $c = O(1/\sqrt[8]{T})$, then a corollary for the TMLE bootstrap holds as follows.

Corollary 12. *Suppose that assumptions of Theorem 11 hold and take $c = O(1/\sqrt[8]{T})$, then we have the following result*

$$\begin{aligned}0 \leq & E_{P_o} Q_{P_o}(\bar{\lambda}_T) - Q_{P_o}(\theta_0) \\ \leq & E_{P_o} Q_{P_o}(\tilde{\lambda}_T) - Q_{P_o}(\theta_0) \\ & + O\left(\frac{K(T)\sqrt{\tau}}{T}\right) + O\left(\frac{K(T)\sqrt{T-\tau}}{T}\right) + O\left(\frac{K(T)}{\sqrt{B}}\right) + O\left(\frac{1}{\sqrt[8]{T}K(T)}\right) \\ & + o_p(1).\end{aligned}$$

3.6 Artificial simulation

In this section, we present evidence on the performance of the TMLE based on simulation experiments using artificial data generated from a two-variable VAR model,

$$r_t = c + b * r_{t-1} + \xi_t, \ \ \xi_t \sim N(0, \Omega),$$

where

$$c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, b = \begin{pmatrix} b_{11} & 0 \\ 0 & b_{22} \end{pmatrix}, \Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix},$$

with a penalty term, $\lambda \times (b_{22} - b_{11})^2$. Considering its three competitors (i.e, the 1SMLE, 2SQMLE, and MLE), we also present their results. Note that setting b as a diagonal matrix is to ensure the estimation consistency of the 2SQMLE.

To assess how well the TMLE performs across different sample sizes of the long time series (i.e., T), sample sizes of the short time series (i.e., $T - \tau$), parameter values of the long time series and short time series (i.e., b_{22} and b_{11}), which can reflect whether the restriction matrix is fully correct, and degrees of tying parameters (i.e., λ), we consider three cases as follows:

$$\text{Case 1:} \quad c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \ b = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}, \ \Omega = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix},$$

$$\text{Case 2:} \quad c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \ b = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.2 \end{pmatrix}, \ \Omega = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix},$$

$$\text{Case 3:} \quad c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.3 \end{pmatrix}, \quad \Omega = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}.$$

Then for each case, we set $T - \tau = cT^a$, where $a = 1/2$, $c = 0.5, 1, 2$, $T = 100, 400, 900$; in addition, we also set $T = 900$ and $\tau = 450$ for each case. We randomly generate 1000 simulated datasets for each variant. Moreover, we set $\lambda \in \Lambda$, where

$$\Lambda = \{0, 0.2, 0.4, 0.6, 0.8, 1, 2, 4, 6, 8, 10, 20, 40, 60, 80, 100, 200, 400, 600, 800, 1000\},$$

and use the TMLE bootstrap and fixed-design wild bootstrap to select λ, respectively. We use notation TMLE_1 for the results of the fixed-design wild bootstrap and TMLE_2 for the results of the TMLE bootstrap. For each estimation method, we focus on the point estimate of b and measure the estimation accuracy of i-th diagonal element in b using the mean-squared error:

$$\text{MSE}_i = \frac{1}{1000} \sum_{j=1}^{1000} (\hat{b}_{ii}^j - b_{ii})^2, \quad i = 1, 2,$$

where $\hat{b}_{ii}^j$ refers to the fitted value of b_{ii} using j-th simulated dataset. In addition, we use $\overline{\text{MSE}} = \text{MSE}_1 + \text{MSE}_2$ to measure the estimation precision of b_{11} and b_{22} as a whole.

Table 3.1 presents the MSE of point estimates of b for different estimation methods. According to the results of Table 3.1, we can make the following observations. First, when the restriction matrix is absolutely correct (i.e., case 1), the performance of the TMLE_1 and TMLE_2 for the estimation results of b_{11} and b_{22} as a whole (i.e., $\overline{\text{MSE}}$) is better than that of the other estimation methods for all settings, which reflects the strength of transferring information between different series. In addition, the superiority of the TMLE_1 and TMLE_2

over the MLE is also consistent with Theorem 6. Moreover, although the results of the TMLE_1 are similar to the TMLE_1, the TMLE_2 outperforms TMLE_1 for some settings.

Second, TMLE_1 and TMLE_2 still have a higher estimation accuracy for most settings relative to the other methods even if the restriction matrix is not completely correct (i.e., case 2 and case 3), which implies that transferring information is still on working. In addition, as Theorem 6 shows, the risk bound of the TMLE still could be less than that of the MLE when the restriction matrix is not fully correct. Furthermore, TMLE_1 and TMLE_2 have similar results, but it is hard to say which is better.

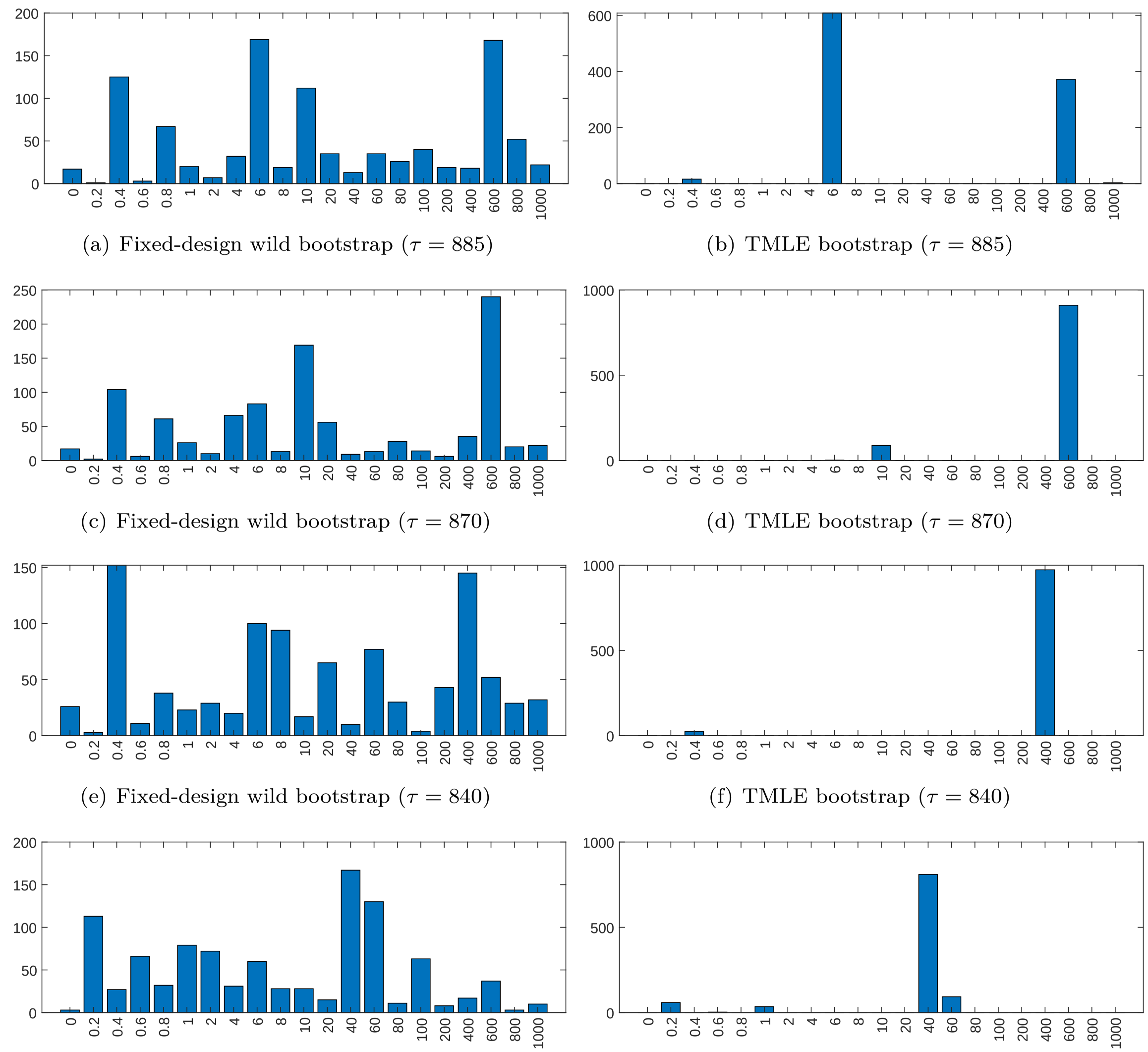
Third, as for the TMLE_1, no matter what T is, the MSE of b_{11} and b_{22} has a downward trend as the sample size of the short time series increases (i.e., τ decreases), which is reasonable because the larger the sample size is, the higher the estimation accuracy is. Similarly for the TMLE_2.

To show the effectiveness of the TMLE bootstrap and fixed-design wild bootstrap, we provide the bar charts of λ determined by two bootstrap procedures for 1000 simulated datasets of Cases 1 – 3 with $T = 900$ in Figures 3.2 – 3.4. As for Case 2 and Case 3, two bootstrap procedures are both inclined to select small λ when $\tau = 450$, which indicates that two bootstrap schemes are effective because, for a restriction matrix that is not absolutely correct, λ needs to become small when the sample size $T - \tau$ is large. However, for Case 1, two bootstrap procedures do not have an obvious trend for selecting λ, which is reasonable because the risk bound of the TMLE is always less than or equal to that of the MLE when the restriction matrix is completely correct.

For more details about the MSE of point estimates of b for different estimation methods, see Tables B.1 – B.30 and Figures B.1 – B.60 in Appendix B.

Table 3.1: MSE of point estimates of b for different estimation methods

		Case 1					Case 2					Case 3				
		1SMLE	2SQMLE	MLE	TMLE$_1$	TMLE$_2$	1SMLE	2SQMLE	MLE	TMLE$_1$	TMLE$_2$	1SMLE	2SQMLE	MLE	TMLE$_1$	TMLE$_2$
$T = 100$				$\tau = 95$					$\tau = 95$					$\tau = 95$		
	b_{11}	8.58727	0.00954	0.01204	0.01024	0.01029	8.89113	0.00954	0.01204	0.01034	0.01040	9.88110	0.00954	0.01217	0.01045	0.01059
	b_{22}	11.69063	5.62809	0.43382	0.01583	0.01484	9.98371	1.93667	0.44219	0.02894	0.03074	11.09923	2.16042	0.47232	0.06127	0.06653
	$\overline{\text{MSE}}$	20.27790	5.63763	0.44586	0.02607	0.02513	18.87483	1.94621	0.45423	0.03928	0.04113	20.98034	2.16996	0.48449	0.07172	0.07712
				$\tau = 90$					$\tau = 90$					$\tau = 90$		
	b_{11}	0.16719	0.00954	0.00989	0.00965	0.00953	0.17309	0.00954	0.00995	0.00949	0.00929	0.17185	0.00954	0.01000	0.00932	0.00906
	b_{22}	0.15697	0.14541	0.12073	0.01260	0.01234	0.16933	0.15746	0.13164	0.02462	0.02686	0.18383	0.17133	0.14769	0.05317	0.05793
	$\overline{\text{MSE}}$	0.32417	0.15495	0.13062	0.02225	0.02187	0.34242	0.16700	0.14159	0.03411	0.03615	0.35568	0.18087	0.15768	0.06249	0.06699
				$\tau = 80$					$\tau = 80$					$\tau = 80$		
	b_{11}	0.05995	0.00954	0.00988	0.00899	0.00892	0.05995	0.00954	0.00988	0.00850	0.00843	0.05993	0.00954	0.00989	0.00827	0.00823
	b_{22}	0.06553	0.06488	0.06108	0.01139	0.01152	0.06883	0.06819	0.06486	0.02212	0.02292	0.07183	0.07123	0.06992	0.04481	0.04646
	$\overline{\text{MSE}}$	0.12549	0.07442	0.07096	0.02038	0.02044	0.12878	0.07773	0.07474	0.03062	0.03135	0.13176	0.08077	0.07981	0.05308	0.05469
$T = 400$				$\tau = 390$					$\tau = 390$					$\tau = 390$		
	b_{11}	0.17447	0.00261	0.00263	0.00259	0.00258	0.17477	0.00261	0.00263	0.00258	0.00256	0.17293	0.00261	0.00263	0.00257	0.00254
	b_{22}	0.15926	0.14803	0.11628	0.00569	0.00314	0.17075	0.15851	0.12464	0.01575	0.01405	0.18143	0.16863	0.13645	0.04317	0.04395
	$\overline{\text{MSE}}$	0.33373	0.15064	0.11891	0.00828	0.00572	0.34552	0.16113	0.12727	0.01833	0.01661	0.35436	0.17125	0.13909	0.04574	0.04649
				$\tau = 380$					$\tau = 380$					$\tau = 380$		
	b_{11}	0.05915	0.00261	0.00262	0.00253	0.00254	0.05926	0.00261	0.00262	0.00250	0.00250	0.05937	0.00261	0.00262	0.00250	0.00249
	b_{22}	0.05323	0.05276	0.04970	0.00340	0.00261	0.05498	0.05439	0.05216	0.01343	0.01323	0.05647	0.05580	0.05584	0.04005	0.04163
	$\overline{\text{MSE}}$	0.11238	0.05537	0.05232	0.00593	0.00515	0.11425	0.05700	0.05478	0.01593	0.01573	0.11585	0.05841	0.05846	0.04255	0.04412
				$\tau = 360$					$\tau = 360$					$\tau = 360$		
	b_{11}	0.02674	0.00261	0.00262	0.00240	0.00241	0.02675	0.00261	0.00262	0.00237	0.00238	0.02678	0.00261	0.00262	0.00246	0.00247
	b_{22}	0.02688	0.02681	0.02602	0.00275	0.00250	0.02732	0.02723	0.02660	0.01156	0.01150	0.02751	0.02741	0.02739	0.03566	0.03526
	$\overline{\text{MSE}}$	0.05362	0.02942	0.02863	0.00515	0.00490	0.05407	0.02984	0.02921	0.01394	0.01387	0.05429	0.03002	0.03001	0.03812	0.03773
$T = 900$				$\tau = 885$					$\tau = 885$					$\tau = 885$		
	b_{11}	0.08477	0.00115	0.00115	0.00115	0.00116	0.08515	0.00115	0.00115	0.00115	0.00115	0.08552	0.00115	0.00115	0.00115	0.00116
	b_{22}	0.07754	0.07437	0.06909	0.00184	0.00116	0.08341	0.07968	0.07515	0.01195	0.01125	0.08937	0.08518	0.08354	0.04073	0.04086
	$\overline{\text{MSE}}$	0.16231	0.07552	0.07025	0.00300	0.00232	0.16856	0.08083	0.07630	0.01310	0.01240	0.17489	0.08633	0.08469	0.04188	0.04201
				$\tau = 870$					$\tau = 870$					$\tau = 870$		
	b_{11}	0.03477	0.00115	0.00115	0.00115	0.00115	0.03481	0.00115	0.00115	0.00115	0.00115	0.03486	0.00115	0.00115	0.00117	0.00117
	b_{22}	0.03519	0.03511	0.03346	0.00144	0.00115	0.03591	0.03580	0.03452	0.01103	0.01088	0.03641	0.03627	0.03614	0.03878	0.03952
	$\overline{\text{MSE}}$	0.06997	0.03626	0.03461	0.00259	0.00230	0.07072	0.03695	0.03567	0.01217	0.01203	0.07126	0.03742	0.03730	0.03994	0.04069
				$\tau = 840$					$\tau = 840$					$\tau = 840$		
	b_{11}	0.01576	0.00115	0.00115	0.00110	0.00110	0.01577	0.00115	0.00115	0.00113	0.00113	0.01578	0.00115	0.00115	0.00122	0.00123
	b_{22}	0.01829	0.01826	0.01815	0.00129	0.00113	0.01812	0.01808	0.01810	0.00993	0.00994	0.01775	0.01771	0.01809	0.03522	0.03551
	$\overline{\text{MSE}}$	0.03405	0.01941	0.01930	0.00240	0.00224	0.03389	0.01923	0.01926	0.01106	0.01107	0.03353	0.01886	0.01924	0.03644	0.03674
				$\tau = 450$					$\tau = 450$					$\tau = 450$		
	b_{11}	0.00225	0.00115	0.00114	0.00080	0.00081	0.00225	0.00115	0.00114	0.00158	0.00146	0.00225	0.00115	0.00114	0.00325	0.00241
	b_{22}	0.00218	0.00218	0.00218	0.00084	0.00086	0.00210	0.00210	0.00211	0.00399	0.00351	0.00199	0.00199	0.00200	0.01009	0.00666
	$\overline{\text{MSE}}$	0.00443	0.00333	0.00332	0.00164	0.00167	0.00435	0.00325	0.00325	0.00557	0.00498	0.00424	0.00314	0.00314	0.01334	0.00906



(a) Fixed-design wild bootstrap ($\tau = 885$)

(b) TMLE bootstrap ($\tau = 885$)

(c) Fixed-design wild bootstrap ($\tau = 870$)

(d) TMLE bootstrap ($\tau = 870$)

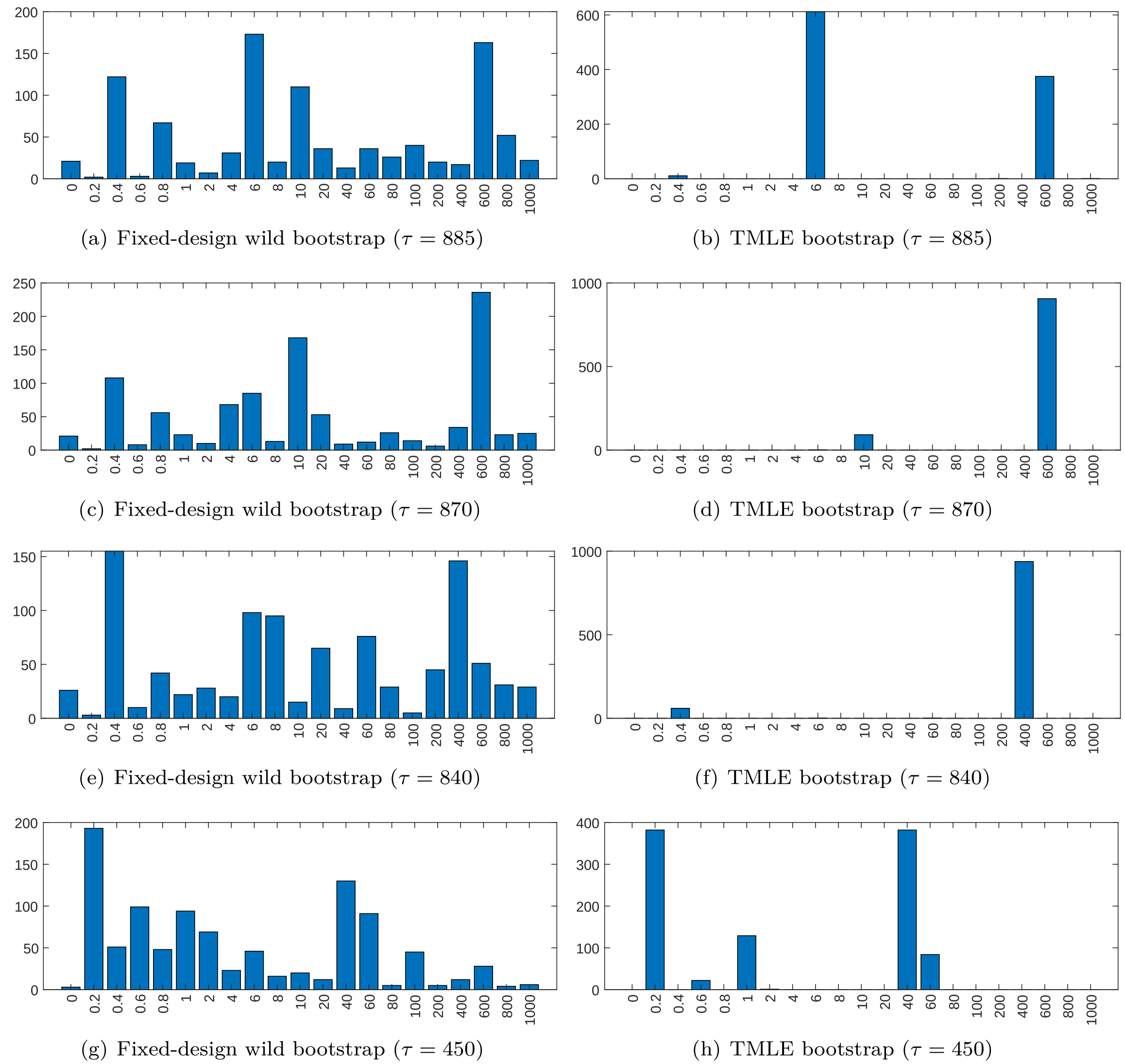
(e) Fixed-design wild bootstrap ($\tau = 840$)

(f) TMLE bootstrap ($\tau = 840$)

(g) Fixed-design wild bootstrap ($\tau = 450$)

(h) TMLE bootstrap ($\tau = 450$)

Figure 3.2: The bar charts of λ determined by different bootstrap procedures for 1000 simulated datasets of Case 1 with $T = 900$.



(a) Fixed-design wild bootstrap ($\tau = 885$)

(b) TMLE bootstrap ($\tau = 885$)

(c) Fixed-design wild bootstrap ($\tau = 870$)

(d) TMLE bootstrap ($\tau = 870$)

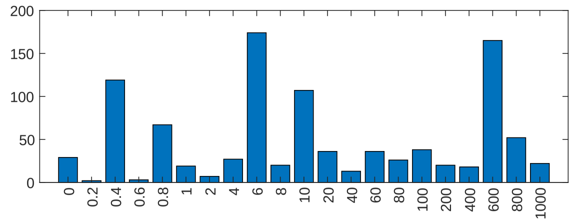
(e) Fixed-design wild bootstrap ($\tau = 840$)

(f) TMLE bootstrap ($\tau = 840$)

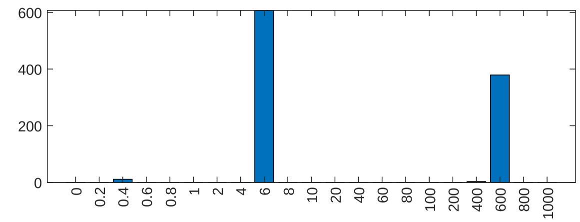
(g) Fixed-design wild bootstrap ($\tau = 450$)

(h) TMLE bootstrap ($\tau = 450$)

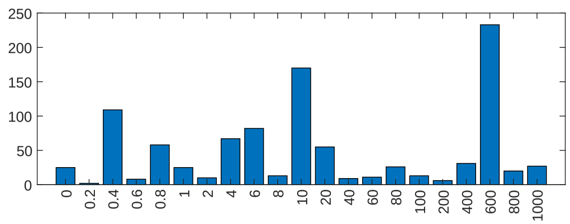
Figure 3.3: The bar charts of λ determined by different bootstrap procedures for 1000 simulated datasets of Case 2 with $T = 900$.



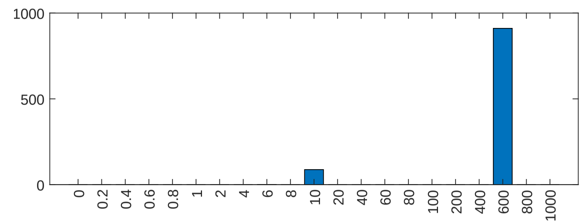
(a) Fixed-design wild bootstrap ($\tau = 885$)



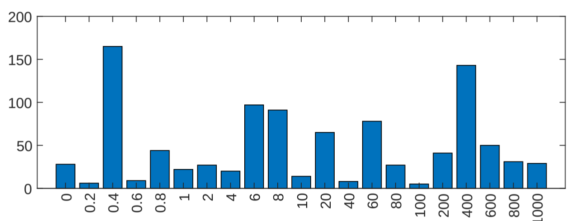
(b) TMLE bootstrap ($\tau = 885$)



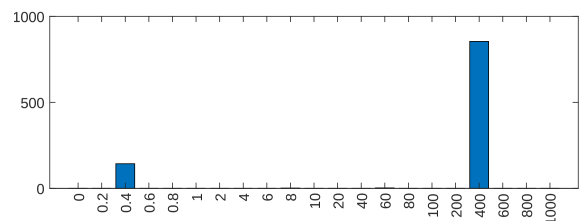
(c) Fixed-design wild bootstrap ($\tau = 870$)



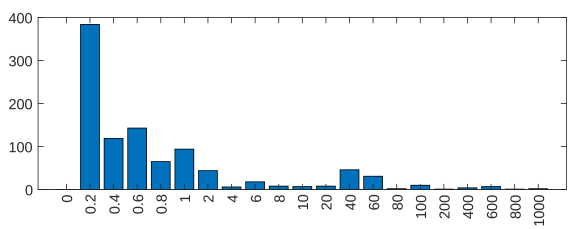
(d) TMLE bootstrap ($\tau = 870$)



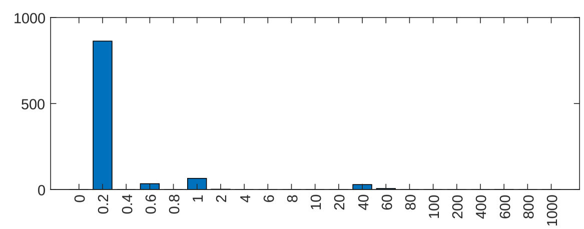
(e) Fixed-design wild bootstrap ($\tau = 840$)



(f) TMLE bootstrap ($\tau = 840$)



(g) Fixed-design wild bootstrap ($\tau = 450$)



(h) TMLE bootstrap ($\tau = 450$)

Figure 3.4: The bar charts of λ determined by different bootstrap procedures for 1000 simulated datasets of Case 3 with $T = 900$.

3.7 Empirical application

In this section, we present evidence on the performance of TMLE on point forecasts and point estimates using daily returns series of Facebook stock (starting on 18th May 2012) and Twitter stock (starting on 7th November 2013). Note that the initial time of Twitter stock is approximately one and a half years later than that of Facebook. Since Facebook and Twitter are both social networking sites that make it easy for people to get the latest news, communicate in short messages and share with family and friends online, it is reasonable that we can transfer some useful information between Facebook stock and Twitter stock to improve predictive accuracy.

Two time series are transformed to be stationary by the difference of the log return. We consider a simple VAR model for these two stocks:

$$r_t = c + b * r_{t-1} + \xi_t, \;\; \xi_t \sim N(0, \Omega),$$

where

$$r_t = \begin{pmatrix} \text{Facebook}_t \\ \text{Twitter}_t \end{pmatrix}, c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, b = \begin{pmatrix} b_{11} & 0 \\ 0 & b_{22} \end{pmatrix}, \Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}.$$

The penalty term is set as $\lambda \times (b_{22} - b_{11})^2$. As we mentioned before, to ensure the estimation consistency of 2SQMLE, we set b as a diagonal matrix. We set $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$.

Considering the real values of parameters are unknown (unlike the artificial simulation), we use the incremental window scheme and rolling window scheme (these two schemes will be introduced later) to evaluate the performance of the TMLE, MLE, 1SMLE, and 2SQMLE on point forecasts. For simplicity, we consider one-step-ahead point forecasts. Since the initial time of Twitter stock is 7th November 2013, which means that Facebook has 386 observations before this time point, we set $\tau = 386$.

Incremental window scheme (τ is fixed) First, for the TMLE, MLE, and 2SQMLE, we use the first $\tau + 10$ observations to predict $\tau + 11$, while for the 1SMLE, we only use 10 observations (i.e., from $\tau+1$ to $\tau+10$) to predict $\tau+11$. Then, for the TMLE, MLE, and 2SQMLE, we use the first $\tau + 11$ observations to predict $\tau + 12$, whereas for the 1SMLE, we use the sample from $\tau + 1$ to $\tau + 11$ to predict $\tau + 12$. We proceed recursively in this fashion until using the first $\tau + 30$ observations to predict $\tau + 31$ (i.e., there are 21 times repeats) and obtain a sequence of forecasts from $\tau+11$ to $\tau+31$ for each method. We call this kind of scheme *Incremental Window Scheme*. For each estimation method, we measure the precision of the one-step-ahead point forecast for the i-th variable in r_t using the mean-squared error:

$$\mathrm{MSE}_i = \frac{1}{21} \sum_{t=\tau+11}^{\tau+31} (r_{it} - \hat{r}_{it})^2, \ \ i = 1, 2,$$

where r_{it} and $\hat{r}_{it}$ means the real value and fitted value of the i-th variable, respectively. In addition, we use $\overline{\mathrm{MSE}} = \mathrm{MSE}_1 + \mathrm{MSE}_2$ to measure the predictive accuracy of both variables as a whole.

Table 3.2 presents the MSE of one-step-ahead point forecasts for different estimation methods. Note that in this table, to illustrate the performance of the TMLE, we also present the result of each fixed λ besides the results of the fixed-design wild bootstrap and TMLE bootstrap. Obviously, the performance of TMLE_1 and TMLE_2 for two stocks is better than the 1SMLE, 2SQMLE and MLE even though most fixed λ do not have lower MSE^n, which indicates the effectiveness of two bootstrap procedures and the importance of transferring information between Facebook and Twitter. Since we carry out the incremental window, which means that the samples for different windows are different (the sample sizes are also different), a fixed λ may not always have a good performance for different samples. In addition, if the penalty (i.e., $(b_{22} - b_{11})^2$) is not absolutely correct, as Theorem 6 shows, then the risk bound of the TMLE for a fixed λ is possible to be greater than that of the MLE. These factors could be responsible for the bad results of some fixed λ.

For more details about the MSE of one-step-ahead point forecasts for different estimation methods, please see Figures B.61 in Appendix B.

Table 3.2: MSE of one-step-ahead point forecasts ($\times 10^3$)

	1SMLE	2SQMLE	MLE	0.1	0.2
Facebook	0.82766	0.71878	0.72090	0.71737	0.72165
Twitter	1.78381	1.77383	2.93286	1094.68508	21.31037
$\overline{\text{MSE}}$	2.61147	2.49260	3.65376	1095.40246	22.03201
	0.3	0.4	0.5	0.6	0.7
Facebook	0.71612	0.72125	0.72990	0.71874	0.72034
Twitter	29.48607	2.55703	140.24896	1.35252	1.70274
$\overline{\text{MSE}}$	30.20219	3.27828	140.97886	2.07126	2.42308
	0.8	0.9	1	TMLE_1	TMLE_2
Facebook	0.71734	0.74328	0.72056	0.72019	0.71704
Twitter	8.31766	3.73829	2.63857	1.71385	1.56719
$\overline{\text{MSE}}$	9.03500	4.48157	3.35913	2.43404	2.28423

Rolling window scheme (τ is not fixed) In this scheme, we consider three cases: $T - \tau = 10$, 20, and 30. Then for each case, we conduct 100 times rolling windows. Firstly let us consider $T - \tau = 10$. For the TMLE, MLE, and 2SQMLE, we use the first T observations (i.e., from 1 to 396) to predict $T + 1$ (i.e., 397), while for the 1SMLE, we only use the sample from $\tau + 1$ to T (i.e., from 387 to 396) to predict 397. Then, for the TMLE, MLE, and 2SQMLE, we use the sample from 2 to 397 to predict 398, while for the 1SMLE, we use the sample from 388 to 397 to predict 398. We proceed recursively 100 times in this fashion and obtain a sequence of forecasts from $T+1$ to $T+100$ for each method. We call this kind of scheme *Rolling Window Scheme*. Similarly for $T-\tau=20$ and 30. For each estimation method, we measure the precision of the one-step-ahead point forecast for the i-th variable in r_t using the mean-squared error:

$$\text{MSE}_i = \frac{1}{100}\sum_{t=T+1}^{T+100}(r_{it}-\hat{r}_{it})^2, \ \ i=1,2.$$

In addition, we use $\overline{\text{MSE}} = \text{MSE}_1 + \text{MSE}_2$ to measure the predictive accuracy of both variables as a whole.

Table 3.3 presents the MSE of one-step-ahead point forecasts for different $T-\tau$. We can make the following observations. First, no matter what $T-\tau$ is, the TMLE_1 always has a smaller $\overline{\text{MSE}}$ relative to the MLE, 1SMLE, and 2SQMLE, which shows the strength of the TMLE. In addition, except for the case of $T-\tau=10$, the performance of the TMLE_2, which is similar to the TMLE_1, is also better than that of the MLE, 1SMLE, and 2SQMLE. A possible reason for the value of $\overline{\text{MSE}}$ of TMLE_2 in the case of $T-\tau=10$ is that the restriction matrix may not be completely correct. Second, the results of two bootstrap procedures always outperform those of most fixed λ regardless of the value of $T-\tau$, which illustrates the effectiveness of these two bootstrap schemes.

For more details about the MSE of one-step-ahead point forecasts for different $T-\tau$, please see Figures B.62 in Appendix B.

Table 3.3: MSE of one-step-ahead point forecasts ($\times 10^3$)

$T-\tau=10$		1SMLE	2SQMLE	MLE	0.1
	Facebook	1.20400	0.94183	0.94260	0.94196
	Twitter	3.69535	3.01583	485.97022	815.70323
	$\overline{\text{MSE}}$	4.89934	3.95766	486.91282	816.64518
		0.2	0.3	0.4	0.5
	Facebook	0.94247	0.94137	0.94066	0.94267
	Twitter	380.15132	17.27060	247.06155	542.10955
	$\overline{\text{MSE}}$	381.09380	18.21197	248.00221	543.05222
		0.6	0.7	0.8	0.9
	Facebook	0.94217	0.94127	0.94136	0.94289
	Twitter	717.20961	339.79032	221.06816	94.96909
	$\overline{\text{MSE}}$	718.15178	340.73159	222.00953	95.91198
		1	TMLE_1	TMLE_2	
	Facebook	0.93805	0.94112	0.94173	
	Twitter	30.71435	2.46913	28.31215	
	$\overline{\text{MSE}}$	31.65240	3.41024	29.25388	
$T-\tau=20$		1SMLE	2SQMLE	MLE	0.1
	Facebook	1.04551	0.93421	0.93723	0.93469
	Twitter	2.47891	2.51346	2.84234	2.71165
	$\overline{\text{MSE}}$	3.52442	3.44768	3.77957	3.64634
		0.2	0.3	0.4	0.5
	Facebook	0.93377	0.93356	0.93533	0.93587
	Twitter	2.48792	2.60636	3.30514	31.23080
	$\overline{\text{MSE}}$	3.42169	3.53992	4.24047	32.16667
		0.6	0.7	0.8	0.9
	Facebook	0.93754	0.93429	0.93403	0.93334
	Twitter	2.33397	2.51340	2.43808	2.86767
	$\overline{\text{MSE}}$	3.27151	3.44769	3.37211	3.80101
		1	TMLE_1	TMLE_2	
	Facebook	0.93880	0.93326	0.93601	
	Twitter	3.99274	2.45675	2.45774	
	$\overline{\text{MSE}}$	4.93154	3.39001	3.39375	
$T-\tau=30$		1SMLE	2SQMLE	MLE	0.1
	Facebook	0.92975	0.88209	0.88241	0.88282
	Twitter	2.71266	2.74225	2.80535	2.63770
	$\overline{\text{MSE}}$	3.64241	3.62435	3.68776	3.52052
		0.2	0.3	0.4	0.5
	Facebook	0.88365	0.88360	0.88357	0.88367
	Twitter	3.72562	2.63030	2.63407	2.67958
	$\overline{\text{MSE}}$	4.60926	3.51390	3.51764	3.56325
		0.6	0.7	0.8	0.9
	Facebook	0.88668	0.88429	0.88441	0.88399
	Twitter	2.62512	2.65238	2.62699	13.54607
	$\overline{\text{MSE}}$	3.51180	3.53667	3.51140	14.43006
		1	TMLE_1	TMLE_2	
	Facebook	0.88405	0.88505	0.88394	
	Twitter	2.79534	2.65173	2.63505	
	$\overline{\text{MSE}}$	3.67938	3.53678	3.51899	

3.8 Conclusion

In this study, we introduce the parameter tying technique in Few-shot Learning to econometrics and pioneer the TMLE to solve the problem that using the traditional MLE to estimate econometric or statistical models does not have a good performance for a type of irregular dependent data where the sample sizes of most series are very large, whereas the other series only have a few observations.

The proposed TMLE can be used directly as long as the likelihood functions of econometric or statistical models exist, which means that it has an enormous application range. We provide the asymptotic theory of the TMLE and detailedly describe its asymptotic properties. In addition, we provide the risk bound of the TMLE and present the strength of the TMLE relative to the traditional MLE. Moreover, we propose an effective bootstrap procedure to select an apt tuning parameter. Furthermore, we provide the finite-sample theory of this bootstrap, which presents some important implications for practical applications.

Extensive artificial simulations and empirical applications show that the performance of the TMLE is significantly better than the MLE, 1SMLE, and 2SQMLE.

Chapter 4

How Has the COVID-19 Pandemic Impacted the Consumer Price Index? Evidence from China

4.1 Introduction

The consumer price index (CPI) recently became one of the most important macroeconomic indicators to measure changes over time in the price level of consumer goods and services purchased by a country's residents. As we know, the unexpected disaster of the COVID-19 pandemic fundamentally impacts every facet of our existence by wreaking havoc in health-care systems, leading to a massive death toll and causing profound socioeconomic disruption. Unsurprisingly, the prices of a broad range of commodities are also affected.

There is no doubt that consumer prices affect people's livelihoods and that fluctuations in prices directly affect residential consumption and manufacturers' production. Hence, it is imperative to explore the impact of the pandemic on the prices of goods and services systematically, which will offer policymakers new insights into how to best combat the deleterious effects of the pandemic.

A large number of studies explore this topic. Specifically, quite a few studies focus on how COVID-19 affected the general price level of goods and services in different countries (e.g., Reinsdorf, 2020; Kouvavas et al., 2020; Cavallo, 2020; Yan and Qian, 2020; Mohsin et al., 2021; Mendez-Carbajo, 2021; CEPAL, 2021; Laskowski et al., 2022). Furthermore, some studies only focus on the impact of the pandemic on the prices of food (e.g.,Mead et al., 2020; Leone et al., 2020; Coluccia et al., 2021), alcohol (e.g., Castaldelli-Maia et al., 2021), and agriculture (e.g., Ramakumar, 2020; Pu and Zhong, 2020; Siche, 2020). However, these studies are primarily descriptive in nature. In addition, there are some studies that analyze the impact of the pandemic on prices by statistical modeling. For example, Ho et al. (2021) and Aliefendioğlu et al. (2021) analyze the impact of the pandemic on housing and transport prices using a multivariate linear regression model and nonlinear autoregressive distributed lag model, respectively. Liu and Rabinowitz (2021) applies a regression discontinuity design to characterize the immediate impacts of the pandemic on retail prices of dairy products in the United States. Lusk et al. (2021) uses a multivariate linear regression model to analyze beef and pork marketing margins and price spreads during the pandemic. Hillen (2021) and Bairagi et al. (2022) analyze the impact of the COVID-19 on food prices using a logit model and a reduced-form of inverse demand function, respectively. However, these studies do not separate out other factors that also affect the CPI, such as holidays or festivals. Although Amare et al. (2020), Akter (2020), Çakır et al. (2021), Clair (2021), among others, separate other factors that also affect the CPI of food, health-care, and housing prices, but they do not consider the dynamic features of the impact of the pandemic on prices.

Hence, we empirically analyze the impact of the COVID-19 pandemic on the different subindices of the CPI to address the limitations mentioned above.

Note that three studies are similar to ours (i.e., Zhang et al., 2020, Chen et al., 2021 and Uche et al., 2021), but they only focus on health-care services and food supplies, whereas we also consider other commodities and services.

Specifically, the innovations of our paper are threefold. First, our data are comprehensive. We collected a monthly CPI dataset of 31 provinces in China over a 24-month span between September 2018 and August 2020. This dataset comprises eight CPI categories: food, tobacco, and liquor; clothing; housing; daily consumables; transport and communications; education, culture, and recreation; health care; and other articles and services.

Second, the assessment of the consequences of the COVID-19 pandemic presents an empirical challenge because a simple pre- versus postpandemic comparison of CPI values, for example, will not adequately capture the effect of the pandemic when CPI changes are subject to inherent temporal trends. Therefore, we adopt the difference-in-difference (DID) method to capture the impact of the pandemic on the CPI. We regard the dataset from September 2019 to August 2020, which is a 12-month span and includes the onset of the COVID-19 pandemic, as the experiment group. To construct the missing counterfactuals depicting the CPI changes in the absence of the pandemic, we rely on the changes in the outcomes of the same set of the CPI categories observed during a 12-month span that closely resembles the experiment group from one year earlier. This feature renders the same set of the CPI that is observed from September 2018 to August 2019 as a suitable control group for the purposes of our analysis. For more details, see Section 2.

Third, to measure the impact of the pandemic on the eight CPI categories, we consider two specifications. We first consider the average effect of the pandemic on the CPI, that is, the impact on the CPI because of the outbreak of COVID-19. We then measure the dynamics of the effect on the CPI over a period of time. The pace of the spread of the virus has varied over time and, moreover, after the onset of the pandemic, some shops and restaurants introduced certain measures, such as socially distanced dining and measuring temperatures, to cope with the pandemic. Hence, the effect of the pandemic may vary from month to month.

The contribution of our paper is that we provide a more in-depth analysis of the impact of the COVID-19 pandemic on the CPI, obtain more definitive conclusions, and offer a deeper insight into policymaking using the monthly panel data of the eight CPI categories in China. In addition, our empirical framework provides a valuable reference for other similar studies. The empirical results indicate that from January to August 2020, the pandemic had a persistent negative impact on housing and daily consumables, whereas no evidence was found for a strong effect on health care prices. Regarding education, culture, and recreation, the pandemic mainly had a persistent positive effect on the price from January to June and then a negative effect for the next two months. In addition, the pandemic could have a persistent positive effect on the price of food, tobacco, and liquor from January to March and then a negative effect over the following several months, while it may have a persistent negative impact on clothing and transport and communications prices after January. Moreover, there could be a mild strengthening of the positive effect on the price of other articles and services following the outbreak of the pandemic.

The rest of the paper is organized as follows. Section 2 provides a detailed description of our data. Section 3 develops our empirical approach. Section 4 presents and discusses the results. The final section concludes.

4.2 Data

The source of our data is the China Economic Information Network Statistics Database [1]. We select the monthly CPI dataset for 31 provinces in China over a 24-month span between September 2018 and August 2020. This dataset comprises eight CPI categories: food, tobacco, and liquor; clothing; housing; daily consumables; transport and communications; education, culture, and recreation; health care; and other articles and services.

The first instance of pneumonia of unknown cause in China was officially registered on December 8, 2019. The first virus strain was successfully isolated on January 7, 2020 and medical professionals confirmed that the pathogen was a new type of coronavirus. On January 23, 2020, China's central government imposed a lockdown in Wuhan and other cities in Hubei Province in an effort to put the center of the COVID-19 outbreak into quarantine. This was an extremely critical point in time. In addition, although the first case was reported on December 8, 2019, most people did not realize the seriousness of this unknown virus during this month. Considering these points, we regard January 2020 as when the COVID-19 pandemic began in China.

We split this dataset into two contiguous, nonoverlapping 12-month subperiods. The first subperiod from September 2019 to August 2020 includes the onset of the COVID-19 pandemic in January 2020. We refer to this subperiod as the experiment group. The second subperiod from September 2018 and August 2019 covers the exact same number of months as the experiment group but begins one year earlier when the CPI of each province was not subject to any noteworthy shocks or legislative changes. We refer to this second subperiod as the control group. As we clarify in the next section, the two-group structure of our data allows the estimation of the effect of the COVID-19 pandemic on CPI.

[1] See https://db.cei.cn/

Table 4.1 presents the basic descriptive statistics for the CPI data that we use in the analysis for both the experiment group (part A) and control group (part B). In the experiment group, from September 2019 to December 2019 (part A1), the mean price and standard deviation of food, tobacco, and liquor are clearly much higher than those for the other CPI categories; therefore, food, tobacco, and liquor prices experienced large fluctuations over this time, which continued from January 2020 to August 2020 (part A2). Moreover, the standard deviations of some CPI categories, such as transport and communications, increased after the outbreak of the pandemic outbreak, which suggests that the pandemic could affect the prices of these items. Nevertheless, in the control group, the means and standard deviations of the eight CPI categories are all similar in the pre-January and post-January periods. In addition, the results for part B1 are similar to that of part B2.

Table 4.1: Descriptive statistics

	Part A: Experiment Group (Sep. 2019 – Aug. 2020)					
	Part A1: Sep. 2019 – Dec. 2019			Part A2: Jan. 2020 – Aug. 2020		
CPI	Obs.	Mean	SD	Obs.	Mean	SD
Food, tobacco, and liquor	124	112.84	3.83	248	116.35	4.24
Clothing	124	102.04	1.67	248	101.16	1.98
Housing	124	102.77	1.17	248	102.30	1.32
Daily consumables	124	101.82	1.32	248	101.78	1.46
Transport and communications	124	98.65	1.12	248	96.29	2.21
Education, culture, and recreation	124	104.22	1.79	248	105.11	1.91
Health care	124	104.36	2.57	248	105.58	2.86
Other articles and services	124	106.28	1.58	248	108.51	2.83
	Part B: Control Group (Sep. 2018 – Aug. 2019)					
	Part B1: Sep. 2018 – Dec. 2018			Part B2: Jan. 2019 – Aug. 2019		
	Obs.	Mean	SD	Obs.	Mean	SD
Food, tobacco, and liquor	124	101.97	1.25	248	105.26	1.78
Clothing	124	101.09	1.30	248	101.35	1.61
Housing	124	102.10	1.12	248	102.43	1.18
Daily consumables	124	101.24	0.72	248	101.65	1.03
Transport and communications	124	101.10	1.40	248	99.59	1.22
Education, culture, and recreation	124	102.60	1.46	248	103.51	1.61
Health care	124	102.20	0.93	248	103.36	1.78
Other articles and services	124	101.29	0.71	248	103.39	1.62

4.3 Empirical approach

Two simple approaches can be used to examine the consequences of the pandemic on the different CPI categories. One approach is to compare the value of the CPI after January with the value prior to January in the experiment group, that is, by contrasting the means in part A2 with part A1 of Table 4.1. However, this approach does not separate out other factors that also affect the CPI. For instance, the post-January period subsumes the holiday season in January and February, when the CPI naturally rises every year. Therefore, a post-January versus pre-January comparison alone would unlikely yield a compelling estimate for the effect of the pandemic. Alternatively, one might contrast the post-January outcomes in the experiment group with the post-January outcomes in the control group, that is, by comparing the mean for the outcomes in part A2 and part B2 of Table 4.1. However, the comparison of the post-January outcomes in the experiment group with the post-January outcomes in the control group does not address the concern that the experiment and control groups differ in unobserved ways, which confounds the estimate of the effect of the pandemic.

To address the deficiencies inherent in the two simple approaches described above, we use a DID approach and exploit the exogenous nature of the pandemic to analyze its impact on the CPI. First, we posit the following general model:

$$y_{group,it} = \beta_0 + \beta_1 post \times group + \beta_2 group + u_i + \lambda_t + \epsilon_{group,it}, \tag{4.1}$$

where *group* is equal to 1 if the observation is from the experiment group and 0 if it is from the control group, i refers to the i-th province, t means the month (from September to August in the following year), $y_{group,it}$ represents the eight CPI categories, which are listed in the first column in Table 4.1, *post* is a dummy variable equal to 1 if the observation is from January or later, u_i is the individual fixed effect, which absorbs the time-invariant impact on explained variables, λ_t is the month fixed effect, which absorbs the time-varying common trend of all units over time, $\epsilon_{group,it}$ denotes the error term.

The coefficient of interest in the regression equation (4.1) is β_1 and it denotes a DID estimate of the impact of the COVID-19 pandemic on the CPI. Table 4.2 presents the rationale of the DID estimation by (4.1). Specifically, the expected value of the CPI before January in the experiment group is $\beta_0 + \beta_2 + u_i + \lambda_{pre}$ in accordance with (4.1), whereas after January it becomes $\beta_0 + \beta_1 + \beta_2 + u_i + \lambda_{post}$. Hence, the difference, $\beta_1 + (\lambda_{post} - \lambda_{pre})$, captures the difference between the post-January and pre-January changes in the CPI in the experiment group. However, we can not observe the post-January CPI for the case where no pandemic began in 2020. To construct a pertinent counterfactual, we use the changes, $\lambda_{post} - \lambda_{pre}$, between the post- and pre-January CPI in the control group. By subtracting $\lambda_{post} - \lambda_{pre}$ from $\beta_1 + \lambda_{post} - \lambda_{pre}$, that is, β_1, provides a DID estimate of the effect of the pandemic on the CPI.

Table 4.2: DID estimate of the COVID-19 pandemic

	pre-Jan.	post-Jan.	difference
Experiment group (Sep.2019–Aug.2020)	$\beta_0 + \beta_2 + u_i + \lambda_{pre}$	$\beta_0 + \beta_1 + \beta_2 + u_i + \lambda_{post}$	$\beta_1 + \lambda_{post} - \lambda_{pre}$
Control group (Sep.2018–Aug.2019)	$\beta_0 + u_i + \lambda_{pre}$	$\beta_0 + u_i + \lambda_{post}$	$\lambda_{post} - \lambda_{pre}$
difference	β_2	$\beta_1 + \beta_2$	β_1

The estimate of β_1 based on (1) is informative of the average effect of the pandemic on the eight CPI categories. To gain further insight into whether, and if so how, the effect of the pandemic has varied over time, we estimate the following specification:

$$y_{group,it} = \theta_0 + \sum_t \theta_t month_t \times group + \theta_2 group + u_i + \lambda_t + e_{group,it}, \quad (4.2)$$

where $month_t$ is a dummy equal to 1 if the observation is from a specific month t from the 9 months from December to August in the following year. We omit December and use this month to compare all the month-by-month effects. $e_{group,it}$ is the error term. The remaining elements of the equation (4.2) are as defined in (4.1).

Note that in this study we estimate all models using OLS. Considering there could be a serial correlation in the error terms, then the cluster-robust standard errors (i.e., the robust standard errors clustered at the level of provinces, see Cameron and Miller, 2015) could be an alternative. But this standard errors, as Cameron and Miller (2015) and Greene (2018) mentioned, have a downward bias when the number of clusters (i.e., the number of provinces) is small. Since there are only 31 provinces in this study, which means that the number of clusters is small, it is hard to say that the cluster-robust standard errors are better than usual standrad errors. In this study we base inference on a larger one of two standard errors for a conservative inference and these two standard errors are presented in all tables.

If the CPI in the control group we chose can serve as a good control group for the CPI in the experiment group, then the change in CPI should be the same for both groups in the absence of the pandemic (i.e., two groups have parallel trends). We present the temporal evolution of the cross-sectional mean of the monthly CPI of 31 provinces from September 2019 to August 2020 (experiment group) and from September 2018 to August 2019 (control group) in Figure 4.1. In the figure, some CPI categories prior to January in the experiment and control groups exhibit comovement, which seems to indicate that the parallel trends assumption could be apt. To judge this better, we carry out two simple DID regressions as placebo tests. Specifically, for the first regression, we divide the dataset from September to December in the experiment and control groups into two subperiods. The first subperiod is from September to October and the second is from November to December. Then for these two subperiods, we use the equation (4.3) to estimate b_1.

$$y_{group,it} = b_0 + b_1 post^* \times group + b_2 group + u_i + \lambda_t + \xi_{group,it}, \tag{4.3}$$

where $post^*$ is equal to 1 if the observation is from the second subperiod (i.e., November and December) and 0 if it is from the first subperiod (i.e., September and October), $\xi_{group,it}$ is the error term, and the remaining elements are as defined in (4.1). The estimate of b_1 for each CPI is listed in Table 4.3. As for housing, daily consumables, education, culture, and recreation, and health care, the point estimates of b_1 are statistically non-significant, which means that the parallel trends assumption probably hold in our context for these four CPI categories. However, the results of the other CPI are statistically significant, which indicates that the control group we selected for these CPI categories could not serve as a good control group (i.e., the parallel trends assumption could not be apposite).

Now we consider the second DID regression as follows:

$$y_{group,it} = \phi_0 + \sum_t \phi_t month_t^* \times group + \phi_2 group + u_i + \lambda_t + \eta_{group,it}, \tag{4.4}$$

where $month_t^*$ is a dummy equal to 1 if the observation is from a specific month t from the 12 months from September to August in the following year. We omit December and use this month to compare the results of the other months. $\eta_{group,it}$ is the error term. The remaining explanatory variables of expression

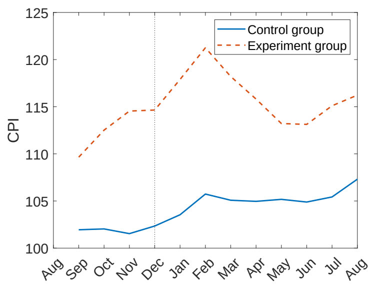
(4.4) are as defined in (4.1). Figure 4.2 shows the point estimates of ϕ_t and corresponding 95% confidence intervals based on a larger one of two standard errors for the eight CPI categories. For housing, daily consumables, education, culture, and recreation, and health care, all piont estimates prior to December are statistically non-significant, which means that the parallel trends assumption probably hold for these four CPI; in addition, there are some statistically significant results after December for housing, daily consumables, education, culture, and recreation, which indicates that the pandemic is likely to have an evident impact on these CPI. As for the remaining CPI, some estimation results prior to December are statistically significant, which means that the parallel trends assumption could not hold for these CPI.

Combining the results of two placebo tests, we believe that the parallel trends assumption seems an apposite one to make for housing, daily consumables, education, culture, and recreation, and health care. Hence, the results of the subsequent analysis for these four CPI categories are likely to be more convincing than for the remaining CPI. However, the reader should keep in mind that the parallel trends assumption is inherently untestable.

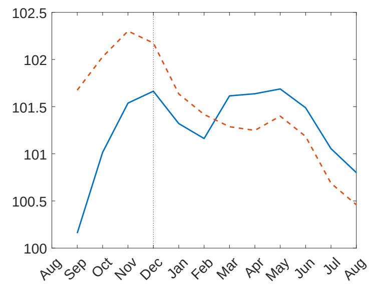
Table 4.3: The point estimates of b_1 for the eight CPI categories

CPI	Food, tobacco, and liquor	Clothing	Housing	Daily consumables
b_1	3.5623^{***} (0.4473) (0.2192)	-0.6274^{***} (0.1991) (0.1556)	-0.1226 (0.1244) (0.0839)	-0.1733 (0.1242) (0.0488)
CPI	Transport and communications	Education, culture, and recreation	Health care	Other articles and services
b_1	1.4585^{***} (0.1484) (0.1053)	0.0697 (0.1858) (0.1389)	-0.0972 (0.3187) (0.0746)	-1.1540^{***} (0.1816) (0.1203)

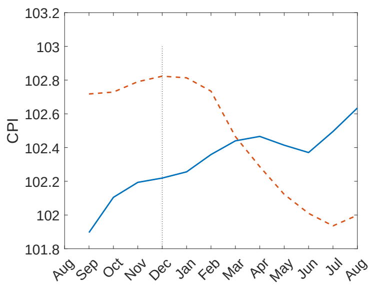
Notes: (a) we estimate b_1 in the regression equation, $y_{group,it} = b_0 + b_1 post^* \times group + b_2 group + u_i + \lambda_t + \xi_{group,it}$, using 248 observations for each CPI. (b) ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. The usual standard errors and the cluster-robust standard errors of b_1 are in the first and second parenthesis, respectively. We base inference on a larger one of two standard errors for a conservative inference.



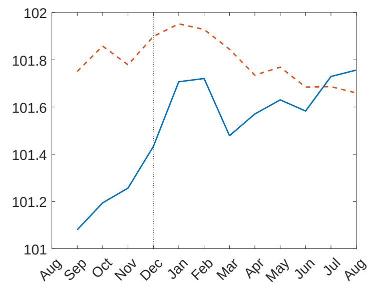
(a) Food, tobacco, and liquor



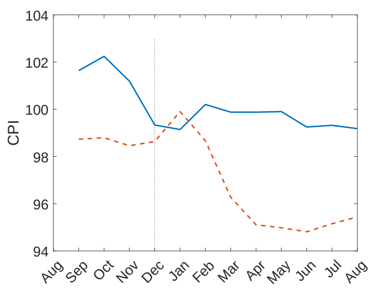
(b) Clothing



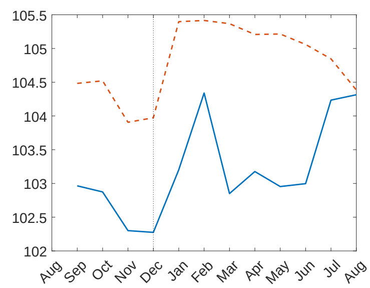
(c) Housing



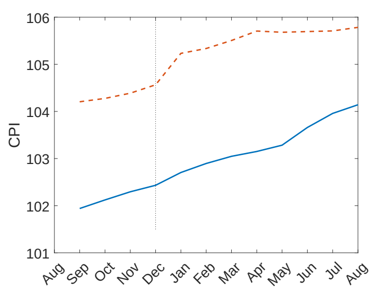
(d) Daily consumables



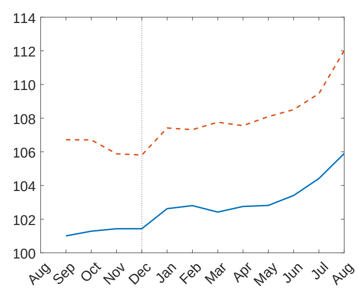
(e) Transport and communications



(f) Education, culture, and recreation

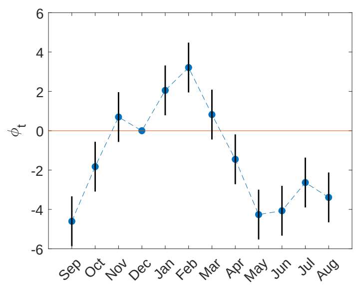


(g) Health care

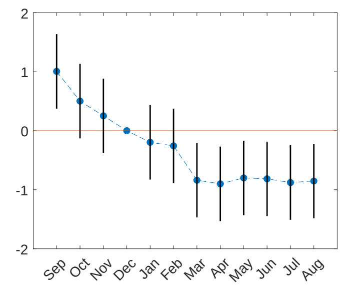


(h) Other articles and services

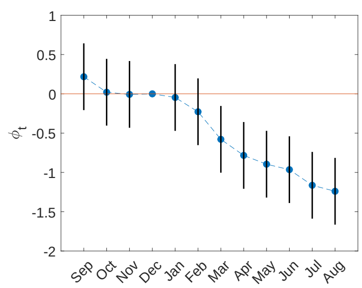
Figure 4.1: Temporal evolution of the cross-sectional mean of the monthly CPI of 31 provinces in the experiment group (Sep.2019-Aug.2020) and control group (Sep.2018-Aug.2019).



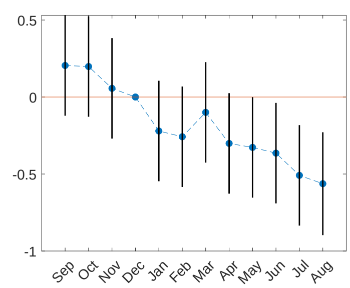
(a) Food, tobacco, and liquor



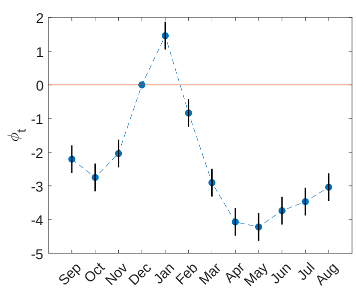
(b) Clothing



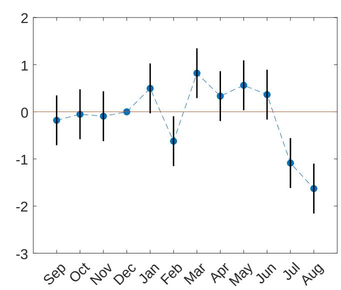
(c) Housing



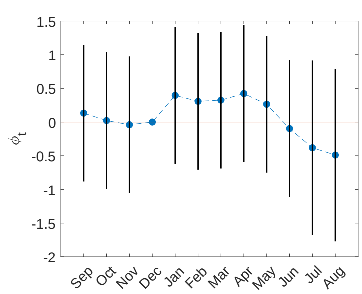
(d) Daily consumables



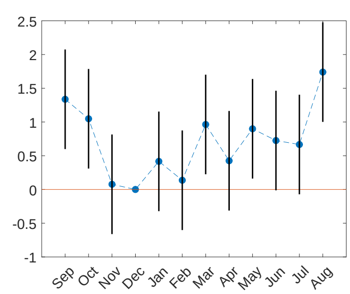
(e) Transport and communications



(f) Education, culture, and recreation



(g) Health care



(h) Other articles and services

Figure 4.2: (1) The point estimates and corresponding 95% confidence intervals (based on a larger one of two standard errors) of ϕ_t for the eight CPI categories. (2) We estimate ϕ_t in the regression, $y_{group,it} = \phi_0 + \sum_t \phi_t month_t^* \times group + \phi_2 group + u_i + \lambda_t + \eta_{group,it}$, using 744 observations for each CPI.

4.4 Results

In this section, we present the average and month-by-month effects of the COVID-19 pandemic on the eight CPI categories and provide some possible explanations for this.

The second column in Table 4.4 presents the estimation results of the average effect (β_1) of the pandemic on different CPI categories. The pandemic has a negative impact that is statistically significant on clothing, housing, daily consumables, transport and communications prices, while it does not have a significant effect on the remaining CPI categories.

The results from the third column to the last column in Table 4.4 show the estimates of the month-by-month effect (θ_t) from January 2020 to August 2020. For each CPI, apart from health care, most of the estimation results of β_t are statistically significant, which indicates that the pandemic has a significant effect on these CPI. To better show the dynamic trend of the month-by-month effect, each part of Figure 4.3 summarizes the results for a specific CPI category. Note that the omitted (comparison) month is December, which is the month immediately preceding the onset of the pandemic.

For food, tobacco, and liquor, part (a) shows that the pandemic has a persistent positive effect on the price in January, February, and March and then has a negative effect over the following several months. The offset positive and negative effects of the pandemic during different periods could explain the statistical non-significance of the estimation result of the average effect. These commodities, generally speaking, are necessities, hence the pandemic can not significantly affect people's demands. However, the supply of these goods can be affected by the pandemic because many stores and manufacturers are asked to be closed during the initial months of the pandemic. Hence, the demand can exceed the supply, which means that the price probably increases. But the supply recovers gradually as the pandemic eases, which means that the price could decrease. This provides a possible explanation for the dynamic effect of the pandemic.

Parts (b), (c), and (d) present that the pandemic has had a persistent negative effect on clothing, housing, and daily consumables prices since January and this negative effect is also reflected by the negative estimation results of the average effect. We believe that these commodities can not be purchased frequently for a short period of time except for some necessary expenses (e.g., water or electricity bills). In addition, people are likely to reduce the consumption of these goods because of falling incomes and rising unemployment risks. Hence, declining demand for these goods during the pandemic is a possible reason for falling prices.

Part (e) traces out the month-by-month effect of the pandemic on transport and communications prices. There is a positive effect in January and then a persistent negative effect over the following several months. The persistent negative effect from February to August exceeds the positive effect in January such that on the whole the pandemic has a negative effect on the prices, which is the same as the negative estimation result of the average effect. Many people are eager to return home because of the panic caused by the pandemic outbreak, which is likely to be responsible for a temporary rise in the price in January. Then the government asked people not to go out or travel unless necessary in order to prevent the spread of the pandemic, which could be a factor for a persistent drop in the price after February. As the pandemic eases gradually, people can go outside or travel freely, which provides a possible explanation for a persistent drop in the negative effect after May.

For education, culture, and recreation, part (f) presents that the pandemic mainly has a positive effect on the price of these commodities from January to June and then has a persistent negative effect over the following two months. The offset positive and negative effects of the pandemic during different periods could explain the statistical non-significance of the estimation result of the average effect. The decreased supply of these goods because many stores are asked to stop operations or shorten business hours could be responsible for the price increase during the initial months of the pandemic. The supply recovers gradually as the pandemic eases, which could lead to a drop in the price.

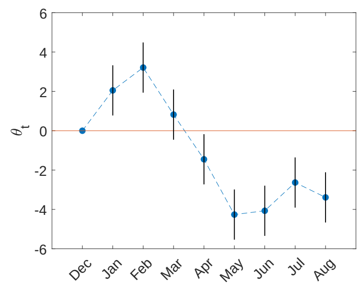
Part (g) shows that the month-by-month effect of the pandemic on health care is statistically non-significant, which is the same as that of the average effect. A possible reason is that, in China, the government controls the prices of most drugs and medical facilities, which means that the pandemic could not have a significant impact on their prices.

Part (h) reveals that the pandemic has had a persistent positive effect on other articles and services since January and this positive effect is also reflected by the positive estimation result of the average effect. Considering other articles and services mainly include insurance, beauty salons, jewelry, watches, and bags, we believe that people are likely to increase the demand for insurance out of concern for the uncertainty in the future caused by the pandemic, which is a possible factor for the price rise.

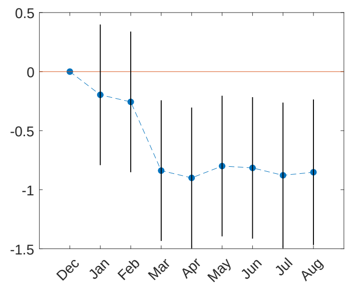
Table 4.4: Average effect (β_1) and month-by-month effect (θ_t) of the pandemic on the eight CPI categories

		θ_t							
CPI	β_1	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Food,	0.2180	2.0505***	3.2114***	0.8199	−1.4525**	−4.2624***	−4.0685***	−2.6332***	−3.3895***
tobacco,	(0.3404)	(0.6508)	(0.6508)	(0.6508)	(0.6508)	(0.6508)	(0.6508)	(0.6508)	(0.6508)
and liquor	(0.2259)	(0.2038)	(0.3462)	(0.2829)	(0.2200)	(0.2847)	(0.3391)	(0.4251)	(0.5505)
	−1.1318***	−0.1965	−0.2563	−0.8384***	−0.8998***	−0.7996***	−0.8154***	−0.8780***	−0.8520***
Clothing	(0.1405)	(0.3039)	(0.3039)	(0.3039)	(0.3039)	(0.3039)	(0.3039)	(0.3039)	(0.3039)
	(0.1992)	(0.0870)	(0.1437)	(0.2385)	(0.2764)	(0.2958)	(0.3056)	(0.3142)	(0.3144)
	−0.7953***	−0.0465	−0.2282	−0.5789***	−0.7836***	−0.8954***	−0.9648***	−1.1638***	−1.2399***
Housing	(0.0967)	(0.2195)	(0.2195)	(0.2195)	(0.2195)	(0.2195)	(0.2195)	(0.2195)	(0.2195)
	(0.1088)	(0.0463)	(0.0678)	(0.1044)	(0.1256)	(0.1448)	(0.1594)	(0.1470)	(0.1496)
Daily	−0.4452***	−0.2206	−0.2580*	−0.0997	−0.3013*	−0.3273**	−0.3644**	−0.5087***	−0.5629***
consumables	(0.0723)	(0.1537)	(0.1537)	(0.1537)	(0.1537)	(0.1537)	(0.1537)	(0.1537)	(0.1537)
	(0.0937)	(0.0622)	(0.1023)	(0.1208)	(0.1154)	(0.1256)	(0.1301)	(0.1449)	(0.1704)
Transport	−0.8533***	1.4609***	−0.8354***	−2.9054***	−4.0716***	−4.2210***	−3.7379***	−3.4682***	−3.0381***
and	(0.1552)	(0.2171)	(0.2171)	(0.2171)	(0.2171)	(0.2171)	(0.2171)	(0.2171)	(0.2171)
communications	(0.1076)	(0.1099)	(0.1080)	(0.1911)	(0.1882)	(0.1580)	(0.1593)	(0.1879)	(0.1845)
Education,	−0.0141	0.4972*	−0.6229**	0.8189***	0.3329	0.5621**	0.3642	−1.0869***	−1.6288***
culture,	(0.1280)	(0.2582)	(0.2582)	(0.2582)	(0.2582)	(0.2582)	(0.2582)	(0.2582)	(0.2582)
and recreation	(0.1955)	(0.1618)	(0.1628)	(0.2233)	(0.2139)	(0.2389)	(0.2270)	(0.2225)	(0.2352)
	0.0647	0.3960	0.3076	0.3253	0.4230	0.2638	−0.0969	−0.3816	−0.4900
Health care	(0.2238)	(0.5118)	(0.5118)	(0.5118)	(0.5118)	(0.5118)	(0.5118)	(0.5118)	(0.5118)
	(0.4308)	(0.3274)	(0.3503)	(0.3557)	(0.3812)	(0.4124)	(0.5031)	(0.6612)	(0.6534)
Other articles	0.1312	0.4169	0.1368	0.9634**	0.4257	0.8996**	0.7252*	0.6665*	1.7398***
and services	(0.1667)	(0.3874)	(0.3874)	(0.3874)	(0.3874)	(0.3874)	(0.3874)	(0.3874)	(0.3874)
	(0.1828)	(0.1338)	(0.2217)	(0.2081)	(0.2028)	(0.2548)	(0.2154)	(0.1968)	(0.2757)

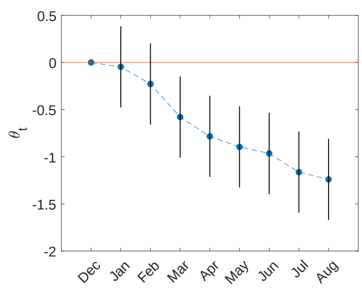
Notes: (a) β_1 in the regression equation, $y_{group,it} = \beta_0 + \beta_1 post \times group + \beta_2 group + u_i + \lambda_t + \epsilon_{group,it}$, means the average effect of the pandemic and the sample size is 744. (b) θ_t in the regression equation, $y_{group,it} = \theta_0 + \sum_t \theta_t month_t \times group + \theta_2 group + u_i + \lambda_t + e_{group,it}$, means the month-by-month effect of the pandemic and the sample size is 558. (c) ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. The usual standard errors and the cluster-robust standard errors of β_1 and θ_t are in the first and second parenthesis, respectively. We base inference on a larger one of two standard errors for a conservative inference.



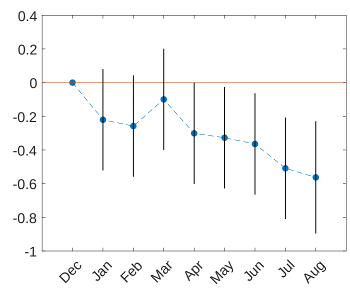
(a) Food, tobacco, and liquor



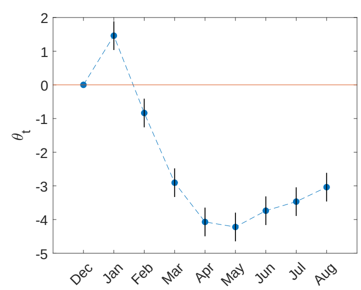
(b) Clothing



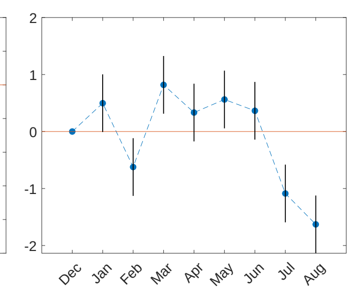
(c) Housing



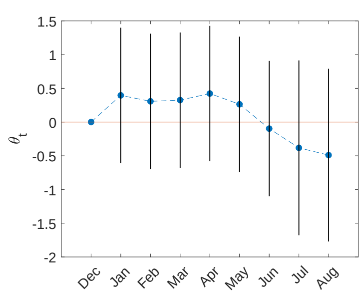
(d) Daily consumables



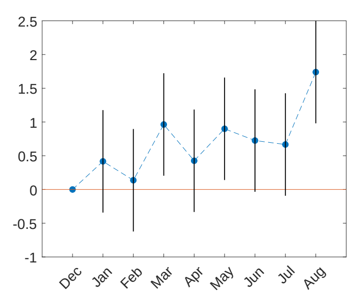
(e) Transport and communications



(f) Education, culture, and recreation



(g) Health care



(h) Other articles and services

Figure 4.3: (1) The point estimates and corresponding 95% confidence intervals (based on a larger one of two standard errors) of θ_t for the eight CPI categories. (2) We estimate θ_t in the regression, $y_{group,it} = \theta_0 + \sum_t \theta_t month_t \times group + \theta_2 group + u_i + \lambda_t + e_{group,it}$, using 558 observations for each CPI.

4.5 Conclusion

In this paper, we used a DID approach and a monthly panel for eight CPI categories in 31 provinces of China over a 24-month period between September 2018 and August 2020 to provide empirical insights into the consequences of the COVID-19 pandemic for the CPI.

The empirical results indicated that from January to August 2020, the pandemic had a persistent negative impact on housing and daily consumables, whereas no evidence was found for a strong effect on health care prices. Regarding education, culture, and recreation, the pandemic mainly had a persistent positive effect on the price from January to June and then a negative effect for the next two months. In addition, the pandemic could have a persistent positive effect on the price of food, tobacco, and liquor from January to March and then a negative effect over the following several months, while it may have a persistent negative impact on clothing and transport and communications prices after January. Moreover, there could be a mild strengthening of the positive effect on the price of other articles and services following the outbreak of the pandemic. Therefore, the government should implement certain measures, such as some type of fiscal stimulus, to increase the demand for housing and daily consumables so that their prices can recover to normal levels. Furthermore, it may be appropriate for the government to stimulate consumer demand for clothing and to allow more stores or manufacturers to open to increase the supply of other articles and services.

Chapter 5

Conclusion

In this dissertation, we develop several new econometric models and statistical methods to solve some problems in dependent data analysis.

In the second chapter, we propose the TVS-ADF model by extending the ADF model to capture the dynamic economic characteristics. In addition, we provide the effective MCMC algorithm including shrinkage and sparsification to estimate the TVS-ADF model. Since in this study, we suppose that the number of latent common factors is given, we will develop an effective method for the determination of the number of factors in future research.

In the third chapter, we pioneer the TMLE using the parameter tying technique to improve the performance of statistical and econometric models for a type of irregular dependent data that most of the time series have long sample periods, whereas the others are very short. We provide the asymptotic and finite-sample theories for the TMLE. In the future, we will consider different penalty forms and selection methods of the tuning parameter.

In the fourth chapter, we provide an in-depth empirical analysis of the consequences of the COVID-19 pandemic for the CPI using a DID method and the monthly panel data of the eight CPI categories in China, which gives a deeper insight into policymaking. Future research will consider the COVID-19 deaths and enlarge the samples to offer a more systematic analysis.

Bibliography

Aharon, D.Y., Demir, E., 2022. Nfts and asset class spillovers: Lessons from the period around the covid-19 pandemic. Finance Research Letters 47, 102515.

Ahn, S.C., Horenstein, A.R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227.

Akter, S., 2020. The impact of covid-19 related 'stay-at-home' restrictions on food prices in europe: findings from a preliminary analysis. Food Security 12, 719–725.

Alessi, L., Barigozzi, M., Capasso, M., 2010. Improved penalization for determining the number of factors in approximate factor models. Statistics & Probability Letters 80, 1806–1813.

Aliefendioğlu, Y., Tanrivermis, H., Salami, M.A., 2021. House price index (hpi) and covid-19 pandemic shocks: evidence from turkey and kazakhstan. International Journal of Housing Markets and Analysis .

Amare, M., Abay, K.A., Tiberti, L., Chamberlin, J., 2020. Impacts of COVID-19 on food security: Panel data evidence from Nigeria. volume 1956. Intl Food Policy Res Inst.

Amengual, D., Watson, M.W., 2007. Consistent estimation of the number of dynamic factors in a large n and t panel. Journal of Business & Economic Statistics 25, 91–96.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.

Bairagi, S., Mishra, A.K., Mottaleb, K.A., 2022. Impacts of the covid-19 pandemic on food prices: Evidence from storable and perishable commodities in india. PloS one 17, e0264355.

Baltagi, B.H., Song, S.H., 2006. Unbalanced panel data: A survey. Statistical Papers 47, 493–523.

Barigozzi, M., 2018. Dynamic factor models. Lecture notes. London School of Economics .

Barigozzi, M., Hallin, M., Soccorsi, S., von Sachs, R., 2021. Time-varying general dynamic factor models and the measurement of financial connectedness. Journal of Econometrics 222, 324–343.

Baumeister, C., Peersman, G., 2013. Time-varying effects of oil supply shocks on the us economy. American Economic Journal: Macroeconomics 5, 1–28.

Belviso, F., Milani, F., 2006. Structural factor-augmented vars (sfavars) and the effects of monetary policy. Topics in Macroeconomics 6.

Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B., 2015. Dirichlet–laplace priors for optimal shrinkage. Journal of the American Statistical Association 110, 1479–1490.

Bianchi, F., Mumtaz, H., Surico, P., 2009. Dynamics of the term structure of uk interest rates .

Bjørnland, H.C., Thorsrud, L.A., 2019. Commodity prices and fiscal policy design: Procyclical despite a rule. Journal of Applied Econometrics 34, 161–180.

Çakır, M., Li, Q., Yang, X., 2021. Covid-19 and fresh produce markets in the united states and china. Applied Economic Perspectives and Policy 43, 341–354.

Cameron, A.C., Miller, D.L., 2015. A practitioner's guide to cluster-robust inference. Journal of human resources 50, 317–372.

Carneiro, P., Hansen, K.T., Heckman, J.J., 2003. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college.

Carter, C.K., Kohn, R., 1994. On gibbs sampling for state space models. Biometrika 81, 541–553.

Castaldelli-Maia, J.M., Segura, L.E., Martins, S.S., 2021. The concerning increasing trend of alcohol beverage sales in the us during the covid-19 pandemic. Alcohol 96, 37–42.

Cataño, D.H., Rodríguez-Caballero, C.V., Chiann, C., Peña, D., 2021. Wavelet estimation for factor models with time-varying loadings. arXiv preprint arXiv:2110.04416 .

Cavallo, A., 2020. Inflation with Covid consumption baskets. Technical Report. National Bureau of Economic Research.

Cawley, J., Conneely, K., Heckman, J., Vytlacil, E., 1997. Cognitive ability, wages, and meritocracy, in: Intelligence, genes, and success. Springer, pp. 179–192.

CEPAL, N., 2021. The effects of the COVID-19 pandemic on the compilation of consumer price indices. ECLAC.

Chamberlain, G., Rothschild, M., 1982. Arbitrage, factor structure, and mean-variance analysis on large asset markets.

Chen, Y., Cai, M., Li, Z., Lin, X., Wang, L., 2021. Impacts of the covid-19 pandemic on public hospitals of different levels: Six-month evidence from shanghai, china. Risk Management and Healthcare Policy 14, 3635.

Clair, A., 2021. The effect of local housing allowance reductions on overcrowding in the private rented sector in england. International Journal of Housing Policy , 1–19.

Cogley, T., 2005. How fast can the new economy grow? a bayesian analysis of the evolution of trend growth. Journal of macroeconomics 27, 179–207.

Cogley, T., Sargent, T.J., 2005. Drifts and volatilities: monetary policies and outcomes in the post wwii us. Review of Economic dynamics 8, 262–302.

Coluccia, B., Agnusdei, G.P., Miglietta, P.P., De Leo, F., 2021. Effects of covid-19 on the italian agri-food supply and value chains. Food Control 123, 107839.

Comte, F., Lieberman, O., 2003. Asymptotic theory for multivariate garch processes. Journal of Multivariate Analysis 84, 61–84.

Connor, G., Korajczyk, R.A., 1986. Performance measurement with the arbitrage pricing theory: A new framework for analysis. Journal of financial economics 15, 373–394.

Davidson, J., 1994. Stochastic limit theory: An introduction for econometricians. OUP Oxford.

Del Negro, M., Otrok, C., 2008. Dynamic factor models with time-varying parameters: measuring changes in international business cycles. FRB of New York Staff Report .

Eichler, M., Motta, G., Von Sachs, R., 2011. Fitting dynamic factor models to non-stationary time series. Journal of Econometrics 163, 51–70.

Forni, M., Giannone, D., Lippi, M., Reichlin, L., 2009. Opening the black box: Structural factor models with large cross sections. Econometric Theory 25, 1319–1347.

Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. Journal of time series analysis 15, 183–202.

Frühwirth-Schnatter, S., Wagner, H., 2010. Stochastic model specification search for gaussian and partial non-gaussian state space models. Journal of Econometrics 154, 85–100.

Galí, J., Gambetti, L., 2015. The effects of monetary policy on stock market bubbles: Some evidence. American Economic Journal: Macroeconomics 7, 233–57.

Geer, S.A., van de Geer, S., Williams, D., 2000. Empirical Processes in M-estimation. volume 6. Cambridge university press.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data analysis. CRC press.

Geweke, J., 1977. The dynamic factor analysis of economic time series, in: Aigner, D.J., Goldberger, A.S. (Eds.), Latent variables in socio-economic models. North-Holland, Amsterdam, pp. 365–383.

Giannone, D., Lenza, M., 2010. The feldstein-horioka fact, in: NBER International Seminar on Macroeconomics, The University of Chicago Press Chicago, IL. pp. 103–117.

Gonçalves, S., Kilian, L., 2004. Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. Journal of econometrics 123, 89–120.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Greene, W.H., 2018. Econometric analysis (eighth edition). Pearson Education India.

Hafner, C.M., Preminger, A., 2009. On asymptotic theory for multivariate garch models. Journal of Multivariate Analysis 100, 2044–2054.

Hall, P., Heyde, C.C., 2014. Martingale limit theory and its application. Academic press.

Hallin, M., Liška, R., 2007. Determining the number of factors in the general dynamic factor model. Journal of the American Statistical Association 102, 603–617.

Hamilton, J.D., Waggoner, D.F., Zha, T., 2007. Normalization in econometrics. Econometric Reviews 26, 221–252.

Hang, H., Steinwart, I., 2014. Fast learning from α-mixing observations. Journal of Multivariate Analysis 127, 184–199.

Hayashi, F., 2000. Econometrics. princeton, nj: Princeton university press .

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intelligent Systems and their applications 13, 18–28.

Hillen, J., 2021. Online food prices during the covid-19 pandemic. Agribusiness 37, 91–107.

Ho, S.J., Xing, W., Wu, W., Lee, C.C., 2021. The impact of covid-19 on freight transport: Evidence from china. MethodsX 8, 101200.

Hoyle, R.H., 1999. Statistical strategies for small sample research. sage.

Huber, F., Koop, G., Onorante, L., 2020. Inducing sparsity and shrinkage in time-varying parameter models. Journal of Business & Economic Statistics , 1–48.

Iwasawa, M., Liu, Q., Zhao, Z., 2022. Tying maximum likelihood estimation for dependent data. Available at SSRN 4252842 .

Kadiyala, K.R., Karlsson, S., 1997. Numerical methods for estimation and inference in bayesian var-models. Journal of Applied Econometrics 12, 99–132.

Karakatsani, N.V., Bunn, D.W., 2008. Forecasting electricity prices: The impact of fundamentals and time-varying coefficients. International Journal of Forecasting 24, 764–785.

Koop, G., Korobilis, D., 2014. A new index of financial conditions. European Economic Review 71, 101–116.

Koop, G., Korobilis, D., Pettenuzzo, D., 2019. Bayesian compressed vector autoregressions. Journal of Econometrics 210, 135–154.

Koop, G., Leon-Gonzalez, R., Strachan, R.W., 2009. On the evolution of the monetary policy transmission mechanism. Journal of Economic Dynamics and Control 33, 997–1017.

Korobilis, D., 2013. Assessing the transmission of monetary policy using time-varying parameter dynamic factor models. Oxford Bulletin of Economics and Statistics 75, 157–179.

Kouvavas, O., Trezzi, R., Eiglsperger, M., Goldhammer, B., Gonçalves, E., et al., 2020. Consumption patterns and inflation measurement issues during the covid-19 pandemic. Economic Bulletin Boxes 7.

Laskowski, R., et al., 2022. Differences between online prices and the consumer prices index during covid-19 in germany. ACTA VSFS 16, 76–87.

Leone, L.A., Fleischhacker, S., Anderson-Steeves, B., Harper, K., Winkler, M., Racine, E., Baquero, B., Gittelsohn, J., 2020. Healthy food retail during the covid-19 pandemic: Challenges and future directions. International journal of environmental research and public health 17, 7397.

Liu, P., Mumtaz, H., Theophilopoulou, A., 2011. International transmission of shocks: A time-varying factor-augmented var approach to the open economy .

Liu, Y., Rabinowitz, A.N., 2021. The impact of the covid-19 pandemic on retail dairy prices. Agribusiness 37, 108–121.

Luo, Z., Zou, Y., Hoffman, J., Fei-Fei, L.F., 2017. Label efficient learning of transferable representations acrosss domains and tasks. Advances in neural information processing systems 30.

Lusk, J.L., Tonsor, G.T., Schulz, L.L., 2021. Beef and pork marketing margins and price spreads during covid-19. Applied Economic Perspectives and Policy 43, 4–23.

Lynch, A.W., Wachter, J.A., 2013. Using samples of unequal length in generalized method of moments estimation. Journal of Financial and Quantitative Analysis 48, 277–307.

Ma, S., Lan, W., Su, L., Tsai, C.L., 2020. Testing alphas in conditional time-varying factor models with high-dimensional assets. Journal of Business & Economic Statistics 38, 214–227.

Marcellino, M., Porqueddu, M., Venditti, F., 2016. Short-term gdp forecasting with a mixed-frequency dynamic factor model with stochastic volatility. Journal of Business & Economic Statistics 34, 118–127.

McCracken, M.W., Ng, S., 2016. Fred-md: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34, 574–589.

Mead, D., Ransom, K., Reed, S.B., Sager, S., 2020. The impact of the covid-19 pandemic on food price indexes and data collection. Monthly Lab. Rev. 143, 1.

Mendez-Carbajo, D., 2021. Consumer spending and the covid-19 pandemic. Page One Economics® .

Merlevède, F., Peligrad, M., Rio, E., et al., 2009. Bernstein inequality and moderate deviations under strong mixing conditions. High dimensional probability V: the Luminy volume 5, 273–292.

Mikkelsen, J.G., Hillebrand, E., Urga, G., 2019. Consistent estimation of time-varying loadings in high-dimensional factor models. Journal of Econometrics 208, 535–562.

Modha, D.S., Masry, E., 1996. Minimum complexity regression estimation with weakly dependent observations. IEEE Transactions on Information Theory 42, 2133–2145.

Mohsin, A., Hongzhen, L., Hossain, S.F.A., 2021. Impact of covid-19 pandemic on consumer economy: Countermeasures analysis. SAGE Open 11, 21582440211008875.

Motta, G., Hafner, C.M., Von Sachs, R., 2011. Locally stationary factor models: Identification and nonparametric estimation. Econometric Theory 27, 1279–1319.

Mumtaz, H., Surico, P., 2012. Evolving international inflation dynamics: world and country-specific factors. Journal of the European Economic Association 10, 716–734.

Nakajima, J., et al., 2011. Time-varying parameter var model with stochastic volatility: An overview of methodology and empirical applications .

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. The Review of Economics and Statistics 92, 1004–1016.

Patton, A.J., 2006. Estimation of multivariate models for time series of possibly different lengths. Journal of applied econometrics 21, 147–173.

Pelger, M., Xiong, R., 2021. State-varying factor models of large dimensions. Journal of Business & Economic Statistics , 1–19.

Polson, N.G., Scott, J.G., 2010. Shrink globally, act locally: Sparse bayesian regularization and prediction. Bayesian statistics 9, 105.

Primiceri, G.E., 2005. Time varying structural vector autoregressions and monetary policy. The Review of Economic Studies 72, 821–852.

Pu, M., Zhong, Y., 2020. Rising concerns over agricultural production as covid-19 spreads: Lessons from china. Global food security 26, 100409.

Ramakumar, R., 2020. Agriculture and the covid-19 pandemic: An analysis with special reference to india. Review of Agrarian Studies 10.

Reinsdorf, M., 2020. Covid-19 and the cpi: Is inflation underestimated? Available at SSRN 3758057 .

Van de Schoot, R., Miočević, M., 2020. Small sample size solutions: A guide for applied researchers and practitioners. Taylor & Francis.

Siche, R., 2020. What is the impact of covid-19 disease on agriculture? Scientia Agropecuaria 11, 3–6.

Sims, C.A., 1980. Macroeconomics and reality. Econometrica: journal of the Econometric Society , 1–48.

Spearman, C., 1927. The abilities of man: Their nature and measurement. Journal of Philosophical Studies 2.

Stock, J.H., Watson, M.W., 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics, in: Handbook of macroeconomics. Elsevier. volume 2, pp. 415–525.

Su, L., Shi, Z., Phillips, P.C., 2016. Identifying latent structures in panel data. Econometrica 84, 2215–2264.

Su, L., Wang, X., 2017. On time-varying factor models: Estimation and testing. Journal of Econometrics 198, 84–101.

Thorsrud, L.A., 2020. Words are the new numbers: A newsy coincident index of the business cycle. Journal of Business & Economic Statistics 38, 393–409.

Trapani, L., 2018. A randomized sequential procedure to determine the number of factors. Journal of the American Statistical Association 113, 1341–1349.

Uche, E., Nnamdi, M.S., Effiom, L., Okoronkwo, C., 2021. Food and healthcare accessibility during covid-19 pandemic. Heliyon , e08656.

Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M., 2020. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) 53, 1–34.

White, H., 1996. Estimation, inference and specification analysis. 22, Cambridge university press.

Yan, L., Qian, Y., 2020. The impact of covid-19 on the chinese stock market: An event study based on the consumer industry. Asian Economics Letters 1, 18068.

Yan, W., Yap, J., Mori, G., 2015. Multi-task transfer methods to improve one-shot learning for multimedia event detection., in: BMVC, pp. 37–1.

Yin, J., 2019. A note on mixing in high dimensional time series. arXiv preprint arXiv:1911.10648 .

Zhang, Y.N., Chen, Y., Wang, Y., Li, F., Pender, M., Wang, N., Yan, F., Ying, X.H., Tang, S.L., Fu, C.W., 2020. Reduction in healthcare services during the covid-19 pandemic in china. BMJ global health 5, e003421.

Zhao, Z., 2022. How has the covid-19 pandemic impacted the consumer price index? evidence from china. The Economic Review 73(4), Otaru University of Commerce (forthcoming) .

Zhao, Z., Liu, Q., 2021. The time varying structural approximate dynamic factor model. Available at SSRN 3996594 .

Zhou, Z.H., 2021. Machine learning. Springer Nature.

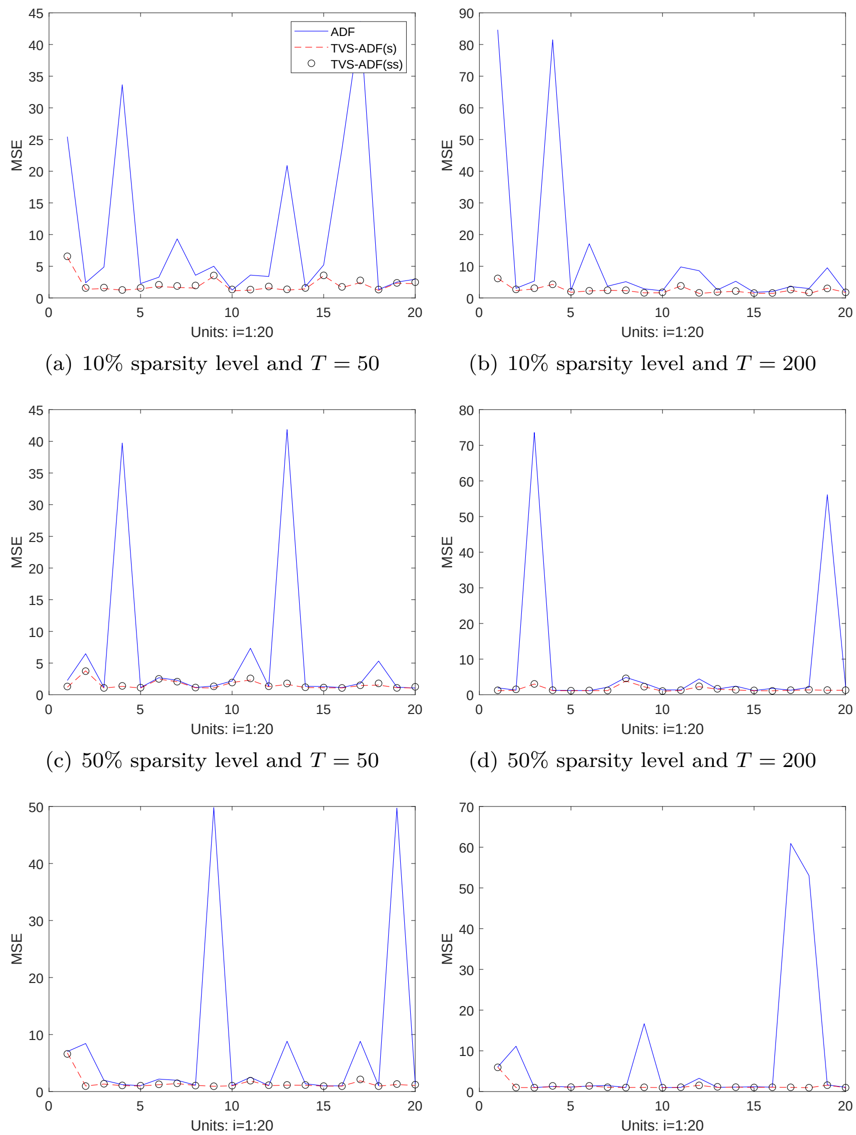
Zivot, E., Wang, J., 2006. Modeling financial time series with S-PLUS. volume 2. Springer.

Chapter A

Appendix A: Chapter 1

Table A.1: Data description

No.	Variable	Description
		Group 1: Real Activity
1	CLF	Civilian Labor Force
2	CE	Civilian Employment
3	CUR	Civilian Unemployment Rate
4	RPI	Real Personal Income
5	HSTNPO	Housing Starts: Total New Privately Owned
6	NPHP	New Private Housing Permits (SAAR)
7	RPCE	Real personal consumption expenditures
		Group 2: Money, Credit and Finance
8	TRDI	Total Reserves of Depository Institutions
9	M1MS	M1 Money Stock
10	M2MS	M2 Money Stock
11	CIL	Commercial and Industrial Loans
		Group 3: Exchange rate
12	EXJPUSx	Japan/U.S. Foreign Exchange Rate
		Group 4: Price
13	PPI:CM	PPI: Crude Materials
14	PPI:IM	PPI: Intermediate Materials
15	PPI:FG	PPI: Finished Goods
16	CPI:AI	PPI: All Items
17	PCE:CI	Personal Cons. Expend.: Chain Index
		Group 5: Expectations
18	CSI	Consumer Sentiment Index
19	NOCG	New Orders for Consumer Goods
20	TBI	Total Business Inventories
		Group 6: Monetary policy (interest rate)
21	EFFR	Effective Federal Funds Rate



(a) 10% sparsity level and $T = 50$

(b) 10% sparsity level and $T = 200$

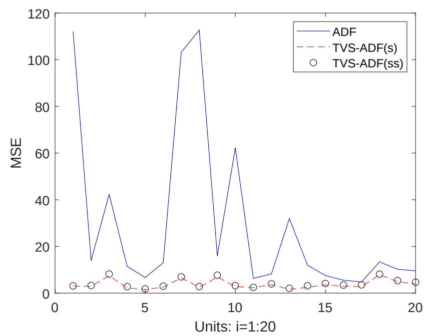
(c) 50% sparsity level and $T = 50$

(d) 50% sparsity level and $T = 200$

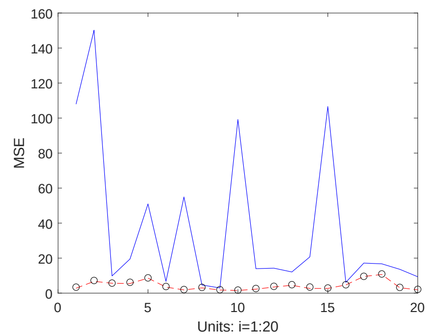
(e) 90% sparsity level and $T = 50$

(f) 90% sparsity level and $T = 200$

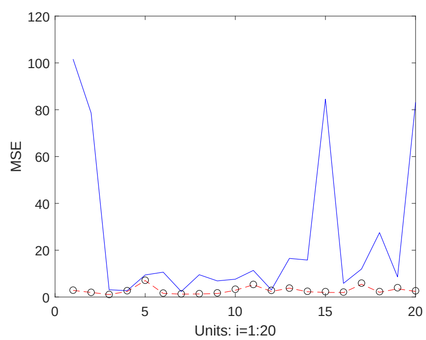
Figure A.1: MSE_i of three models with 4 explanatory variables ($n = 20$)



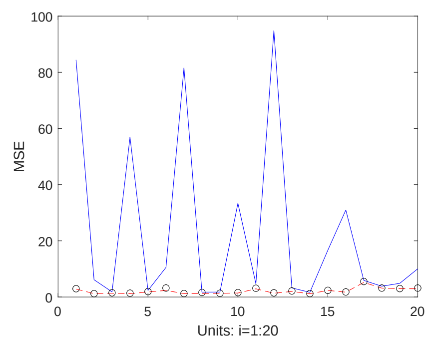
(a) 10% sparsity level and $T = 50$



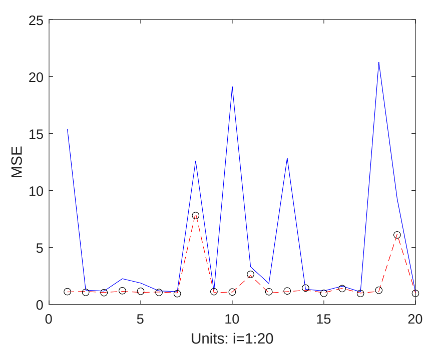
(b) 10% sparsity level and $T = 200$



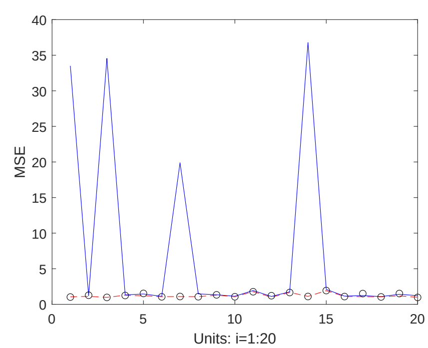
(c) 50% sparsity level and $T = 50$



(d) 50% sparsity level and $T = 200$

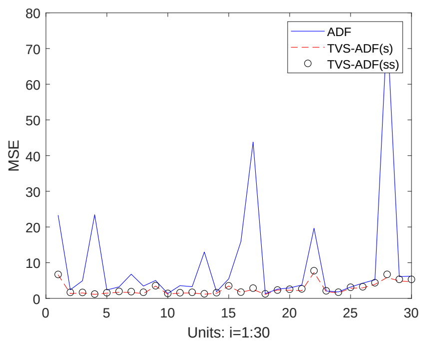


(e) 90% sparsity level and $T = 50$

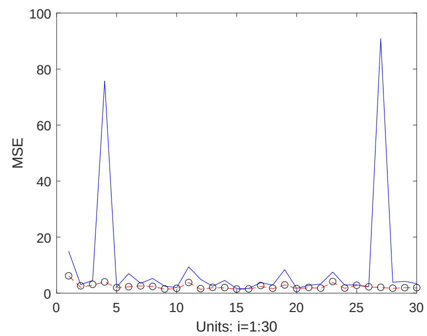


(f) 90% sparsity level and $T = 200$

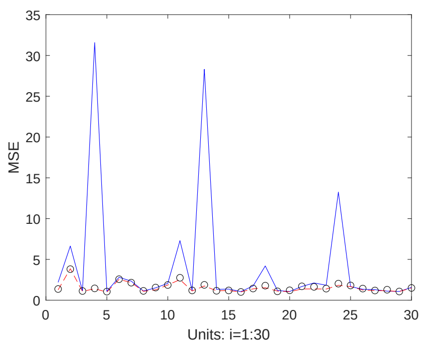
Figure A.2: MSE_i of three models with 8 explanatory variables ($n = 20$)



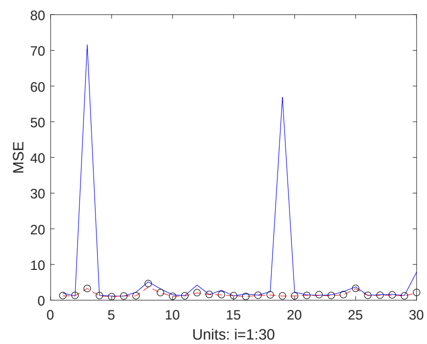
(a) 10% sparsity level and $T = 50$



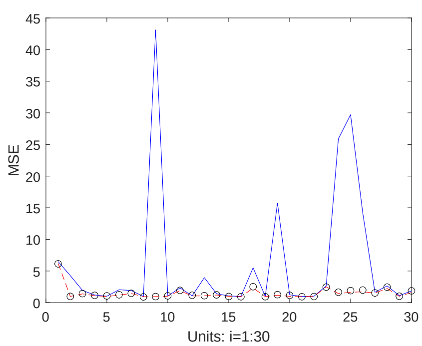
(b) 10% sparsity level and $T = 200$



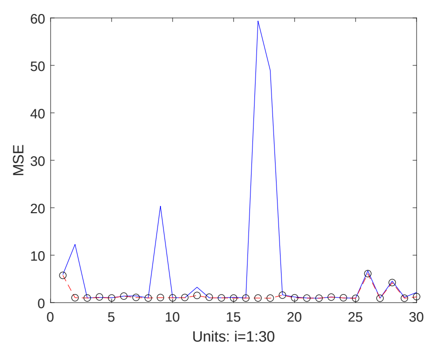
(c) 50% sparsity level and $T = 50$



(d) 50% sparsity level and $T = 200$

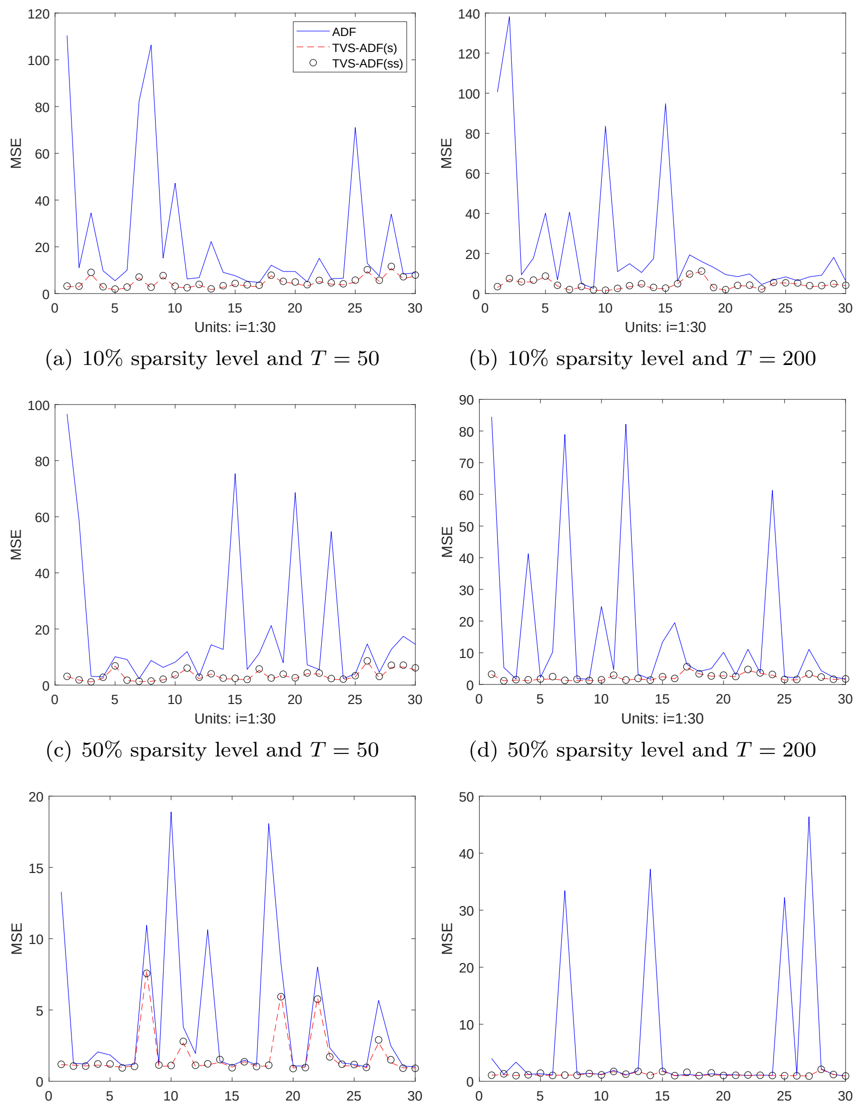


(e) 90% sparsity level and $T = 50$



(f) 90% sparsity level and $T = 200$

Figure A.3: MSE_i of three models with 4 explanatory variables ($n = 30$)



(a) 10% sparsity level and $T = 50$

(b) 10% sparsity level and $T = 200$

(c) 50% sparsity level and $T = 50$

(d) 50% sparsity level and $T = 200$

(e) 90% sparsity level and $T = 50$

(f) 90% sparsity level and $T = 200$

Figure A.4: MSE_i of three models with 8 explanatory variables ($n = 30$)

Chapter B

Appendix B: Chapter 2

B.1 Technical Lemmas

Lemma 13. *Let the parameter space* Θ *be a compact and convex subset of* $\mathbb{R}^K$*. For all* $\theta \in \Theta$*,* $l_t(\theta)$ *is a measurable function of a strong mixing process with mixing coefficients* $\alpha(\cdot)$*, where* $\alpha(\tau) \leq c_\alpha \rho^\tau$ *for some* $c_\alpha > 0$ *and* $0 < \rho < 1$*. A measurable function* $\tilde{l}_t$ *exists such that* $|l_t(\theta) - l_t(\bar{\theta})| \leq \|\theta - \bar{\theta}\|\tilde{l}_t$ *for any* $\theta, \bar{\theta} \in \Theta$*,* $\sup_{\theta\in\Theta} |l_t(\theta)| \leq \tilde{l}_t$*, and* $E(|\tilde{l}_t|^q) \leq C$ *for some integer* $q > \max\{K+1, 4\}$*, where* $C < \infty$ *is a constant. For any* $c > 0$*, we have*

$$P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{l_t(\theta) - E[l_t(\theta)]\}\right| \leq c\right) = 1 - o(T^{-1}). \tag{B.1}$$

Proof. Letting $L_t(\theta) = l_t(\theta) - E[l_t(\theta)]$, we denote $\mathbb{1}_t = \mathbb{1}\{|l_t(\theta) - E[l_t(\theta)]| \leq \sqrt{T}\} = \mathbb{1}\{|L_t(\theta)| \leq \sqrt{T}\}$ and $\bar{\mathbb{1}}_t = 1 - \mathbb{1}_t$. Since $E[L_t(\theta)] = 0$, we have

$$L_t(\theta) = L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t] + L_t(\theta)\bar{\mathbb{1}}_t - E[L_t(\theta)\bar{\mathbb{1}}_t]. \tag{B.2}$$

Then, it suffices to show that for any constants $c_1 > 0$ and $c_2 > 0$,

$$TP\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| > c_1\right) = o(1), \tag{B.3}$$

$$TP\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}L_t(\theta)\bar{\mathbb{1}}_t\right| > c_2\right) = o(1), \tag{B.4}$$

$$\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}E[L_t(\theta)\bar{\mathbb{1}}_t]\right| = o(1). \tag{B.5}$$

First, we show (B.5). By assumption, a positive constant C exists such that $E[L_t(\theta)^2] = E[l_t(\theta)^2] - E[l_t(\theta)]^2 \le E[\tilde{l}_t^2] < C$. Then, by the H lder inequality,

$$\begin{aligned}
\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}E[L_t(\theta)\bar{\mathbb{1}}_t]\right| &\le \frac{1}{T}\sum_{t=1}^{\tau}\sup_{\theta\in\Theta}|E[L_t(\theta)\bar{\mathbb{1}}_t]| \\
&\le \frac{1}{T}\sum_{t=1}^{\tau}\sup_{\theta\in\Theta}\left|E[L_t(\theta)^2]^{1/2}E(\bar{\mathbb{1}}_t)^{1/2}\right| \\
&\le \frac{C^{1/2}}{T}\sum_{t=1}^{\tau}\sup_{\theta\in\Theta}\left|E(\bar{\mathbb{1}}_t)^{1/2}\right| \\
&= \frac{C^{1/2}}{T}\sum_{t=1}^{\tau}\sup_{\theta\in\Theta}\left|P\left(|L_t(\theta)| > \sqrt{T}\right)^{1/2}\right| \\
&\le \frac{C^{1/2}}{T}\sum_{t=1}^{\tau}\sup_{\theta\in\Theta}\left|\left(\frac{E[L_t(\theta)^2]}{T}\right)^{1/2}\right| \\
&\le \frac{C^{1/2}}{T\sqrt{T}}\sum_{t=1}^{\tau}\sup_{\theta\in\Theta}\left|E[L_t(\theta)^2]^{1/2}\right| \\
&\le \frac{C\tau}{T\sqrt{T}} = O(T^{-1/2}).
\end{aligned} \tag{B.6}$$

Second, we show (B.4). By assumption, it holds that $\sup_{\theta\in\Theta}|L_t(\theta)| = \sup_{\theta\in\Theta}|l_t(\theta) - E[l_t(\theta)]| \le \sup_{\theta\in\Theta}|l_t(\theta)| + \sup_{\theta\in\Theta}E[|l_t(\theta)|] \le \tilde{l}_t + E(\tilde{l}_t)$. Note that, by the definition of $\bar{\mathbb{1}}_t = \mathbb{1}\{|L_t(\theta)| > \sqrt{T}\}$, $\left|\frac{1}{T}\sum_{t=1}^{\tau}L_t(\theta)\bar{\mathbb{1}}_t\right| > c_2$ implies that $\max_{1\le t\le\tau}\sup_{\theta\in\Theta}|L_t(\theta)| > \sqrt{T}$. Thus,

$$
\begin{aligned}
P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}L_t(\theta)\bar{\mathbb{1}}_t\right| > c_2\right) &\le P\left(\max_{1\le t\le\tau}\sup_{\theta\in\Theta}|L_t(\theta)| > \sqrt{T}\right)\\
&\le \tau\max_{1\le t\le\tau}P\left(\tilde{l}_t + E(\tilde{l}_t) > \sqrt{T}\right)\\
&\le \frac{\tau\max_{1\le t\le\tau}E\left[\left|\tilde{l}_t + E(\tilde{l}_t)\right|^q\right]}{T^{q/2}}\\
&= O(T^{1-q/2}) = o(T^{-1}). \qquad \text{(B.7)}
\end{aligned}
$$

Third, we show (B.3). Since Θ is assumed to be compact, subsets $\Theta_j \subset \Theta$ exist for $j = 1,\dots,n_\epsilon$ such that $\Theta \subset \cup_{j=1}^{n_\epsilon}\Theta_j$ and $\|\theta-\bar{\theta}\| \le \epsilon/T$ for any $\epsilon > 0$ and $\theta,\bar{\theta}\in\Theta_j$, where $n_\epsilon = O(T^K)$.

By the Boole inequality, we obtain

$$
\begin{aligned}
&P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| > c_1\right)\\
&\le \sum_{j=1}^{n_\epsilon}P\left(\sup_{\theta\in\Theta_j}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| > c_1\right). \qquad \text{(B.8)}
\end{aligned}
$$

When $\theta\in\Theta_j$, it holds by assumption that, for any $\bar{\theta}\in\Theta_j$, $|L_t(\theta) - L_t(\bar{\theta})| = |l_t(\theta) - l_t(\bar{\theta}) + E[l_t(\bar{\theta}) - l_t(\theta)]| \le \|\theta-\bar{\theta}\|[\tilde{l}_t + E(\tilde{l}_t)]$. Then,

$$
\begin{aligned}
\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| &\le \left|\frac{1}{T}\sum_{t=1}^{\tau}\left\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right|\\
&\quad + \left|\frac{1}{T}\sum_{t=1}^{\tau}\left\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t] - L_t(\bar{\theta})\mathbb{1}_t + E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right|\\
&\le \left|\frac{1}{T}\sum_{t=1}^{\tau}\left\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right|\\
&\quad + \frac{1}{T}\sum_{t=1}^{\tau}\left|L_t(\theta) - L_t(\bar{\theta})\right|\mathbb{1}_t + \frac{1}{T}\sum_{t=1}^{\tau}E\left[\left|L_t(\bar{\theta}) - L_t(\theta)\right|\mathbb{1}_t\right],
\end{aligned}
$$

(B.9)

where the second and third terms of the right hand side are

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{\tau}\left|L_t(\theta)-L_t(\bar{\theta})\right|\mathbb{1}_t+\frac{1}{T}\sum_{t=1}^{\tau}E\left[\left|L_t(\bar{\theta})-L_t(\theta)\right|\mathbb{1}_t\right]\\
&\le \frac{1}{T}\sum_{t=1}^{\tau}\left\|\theta-\bar{\theta}\right\|[\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t+\frac{1}{T}\sum_{t=1}^{\tau}\left\|\theta-\bar{\theta}\right\|E\{[\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t\}\\
&= \frac{1}{T}\sum_{t=1}^{\tau}\left\|\theta-\bar{\theta}\right\|\Big([\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t+E\{[\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t\}\Big)\\
&\le \frac{\epsilon}{T^2}\sum_{t=1}^{\tau}\tilde{L}_t, \qquad \text{(B.10)}
\end{aligned}
$$

where $\tilde{L}_t \equiv [\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t+E\{[\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t\}$. Equations (B.9) and (B.10) indicate that

$$
\begin{aligned}
\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t-E[L_t(\theta)\mathbb{1}_t]\}\right| &\le \left|\frac{1}{T}\sum_{t=1}^{\tau}\left\{L_t(\bar{\theta})\mathbb{1}_t-E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right|\\
&\quad+\frac{\epsilon}{T^2}\sum_{t=1}^{\tau}[\tilde{L}_t-E(\tilde{L}_t)]+\frac{\epsilon}{T^2}\sum_{t=1}^{\tau}E(\tilde{L}_t), \qquad \text{(B.11)}
\end{aligned}
$$

where $E(\tilde{L}_t)=2E\{[\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t\}<\infty$ by assumption. Then, equations (B.8) and (B.11) imply that

$$
\begin{aligned}
P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t-E[L_t(\theta)\mathbb{1}_t]\}\right|>c_1\right) &\le n_\epsilon P\left(\left|\frac{3}{T}\sum_{t=1}^{\tau}\left\{L_t(\bar{\theta})\mathbb{1}_t-E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right|>c_1\right)\\
&\quad+n_\epsilon P\left(\left|\frac{3\epsilon}{T^2}\sum_{t=1}^{\tau}[\tilde{L}_t-E(\tilde{L}_t)]\right|>c_1\right), \qquad \text{(B.12)}
\end{aligned}
$$

since $P(|\frac{\epsilon}{T^2}\sum_{t=1}^{\tau}E(\tilde{L}_t)|>c_1)=0$ by choosing ϵ small enough. Thus, it suffices to show that

$$n_\epsilon P\left(\left|\frac{1}{T}\sum_{t=1}^{\tau}\left\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right| > c_1\right) = o(T^{-1}), \tag{B.13}$$

$$n_\epsilon P\left(\left|\frac{\epsilon}{T^2}\sum_{t=1}^{\tau}[\tilde{L}_t - E(\tilde{L}_t)]\right| > c_1\right) = o(T^{-1}). \tag{B.14}$$

First, we show (B.13). abcNote that $L_t(\bar{\theta})$ is a measurable function of the strong mixing process with the mixing coefficient satisfying $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $0 < \rho < 1$. Moreover, we have $\sup_{1\leq t\leq\tau} |L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]| \leq 2\sqrt{T}$. Thus, applying Theorem 2 in Merlevède et al., 2009 (see also Lemma S1.1 in Su et al., 2016) yields

$$\begin{aligned}
&Tn_\epsilon P\left(\left|\sum_{t=1}^{\tau}\left\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right| > c_1 T\right)\\
&\leq Tn_\epsilon \exp\left(-\frac{C_0 c_1^2 T^2}{v_0^2 T + 4T + c_1 T\sqrt{T}[\log(T)]^2}\right)\\
&= \exp\left(-\frac{C_0 c_1^2 T}{v_0^2 + 4 + c_1\sqrt{T}[\log(T)]^2} + K\log(T) + \log(T)\right)
\end{aligned} \tag{B.15}$$

for some constant C_0 and v_o.[1] Since the right hand side of the above equation converges zero as $T \to \infty$, (B.13) holds.

Next, we show (B.14). By the Markov and H lder inequalities,

$$\begin{aligned}
n_\epsilon P\left(\left|\frac{\epsilon}{T^2}\sum_{t=1}^{\tau}[\tilde{L}_t - E(\tilde{L}_t)]\right| > c_1\right) &\leq n_\epsilon \frac{\epsilon^q E\left[\left|\frac{1}{T}\sum_{t=1}^{\tau}[\tilde{L}_t - E(\tilde{L}_t)]\right|^q\right]}{c_1^q T^q}\\
&\leq n_\epsilon \frac{\epsilon^q \frac{1}{T}\sum_{t=1}^{\tau} E\left[\left|\tilde{L}_t - E(\tilde{L}_t)\right|^q\right]}{c_1^q T^q}\\
&= O(T^{K-q}),
\end{aligned} \tag{B.16}$$

where the right hand side is $o(T^{-1})$ since $K + 1 < q$. □

[1] Let $L_t \equiv L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]$. Then, $v_o \equiv \sup_{t\geq 1}[\mathrm{Var}(L_t) + 2\sum_{s=t+1}^{\infty}\mathrm{Cov}(L_t, L_s)]$.

Lemma 14. *Let the parameter space* Θ *be a compact and convex subset of* $\mathbb{R}^K$*. For all* $\theta \in \Theta$*,* $l_t(\theta)$ *is a measurable function of a strong mixing process with mixing coefficients* $\alpha(\cdot)$*, where* $\alpha(\tau) \leq c_\alpha \rho^\tau$ *for some* $c_\alpha > 0$ *and* $0 < \rho < 1$*. A measurable function* $\tilde{l}_t$ *exists such that* $|l_t(\theta) - l_t(\bar{\theta})| \leq \|\theta - \bar{\theta}\| \tilde{l}_t$ *for any* $\theta, \bar{\theta} \in \Theta$*,* $\sup_{\theta \in \Theta} |l_t(\theta)| \leq \tilde{l}_t$*, and* $E(|\tilde{l}_t|^q) \leq C$ *for some integer* $q > \max\{K + a, 4\}$*, where* $C < \infty$ *is a constant. For any* $c > 0$*, we have*

$$P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\{l_t(\theta) - E[l_t(\theta)]\}\right| \leq c\right) = 1 - o(T^{-1}). \tag{B.17}$$

Proof. Letting $L_t(\theta) = l_t(\theta) - E[l_t(\theta)]$, we denote $\mathbb{1}_t = \mathbb{1}\{|l_t(\theta) - E[l_t(\theta)]| \leq \sqrt{T}\} = \mathbb{1}\{|L_t(\theta)| \leq \sqrt{T}\}$ and $\bar{\mathbb{1}}_t = 1 - \mathbb{1}_t$. Since $E[L_t(\theta)] = 0$, we have

$$L_t(\theta) = L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t] + L_t(\theta)\bar{\mathbb{1}}_t - E[L_t(\theta)\bar{\mathbb{1}}_t]. \tag{B.18}$$

Then, it suffices to show that for any constants $c_1 > 0$ and $c_2 > 0$,

$$TP\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| > c_1\right) = o(1), \tag{B.19}$$

$$TP\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}L_t(\theta)\bar{\mathbb{1}}_t\right| > c_2\right) = o(1), \tag{B.20}$$

$$\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}E[L_t(\theta)\bar{\mathbb{1}}_t]\right| = o(1). \tag{B.21}$$

First, we show (B.21). By assumption, a positive constant C exists such that

$E[L_t(\theta)^2] = E[l_t(\theta)^2] - E[l_t(\theta)]^2 \leq E[\tilde{l}_t^2] < C$. Then, by the H lder inequality,

$$
\begin{aligned}
\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=\tau+1}^{T} E[L_t(\theta)\bar{\mathbb{1}}_t]\right| &\leq \frac{1}{T}\sum_{t=\tau+1}^{T}\sup_{\theta\in\Theta}|E[L_t(\theta)\bar{\mathbb{1}}_t]| \\
&\leq \frac{1}{T}\sum_{t=\tau+1}^{T}\sup_{\theta\in\Theta}\left|E[L_t(\theta)^2]^{1/2}E(\bar{\mathbb{1}}_t)^{1/2}\right| \\
&\leq \frac{C^{1/2}}{T}\sum_{t=\tau+1}^{T}\sup_{\theta\in\Theta}\left|E(\bar{\mathbb{1}}_t)^{1/2}\right| \\
&= \frac{C^{1/2}}{T}\sum_{t=\tau+1}^{T}\sup_{\theta\in\Theta}\left|P\left(|L_t(\theta)| > \sqrt{T}\right)^{1/2}\right| \\
&\leq \frac{C^{1/2}}{T}\sum_{t=\tau+1}^{T}\sup_{\theta\in\Theta}\left|\left(\frac{E[L_t(\theta)^2]}{T}\right)^{1/2}\right| \\
&\leq \frac{C^{1/2}}{T\sqrt{T}}\sum_{t=\tau+1}^{T}\sup_{\theta\in\Theta}\left|E[L_t(\theta)^2]^{1/2}\right| \\
&\leq \frac{C(T-\tau)}{T\sqrt{T}} = O(T^{a-3/2}) = o(1). \qquad \text{(B.22)}
\end{aligned}
$$

Second, we show (B.20). By assumption, we obtain

$$
\sup_{\theta\in\Theta}|L_t(\theta)| = \sup_{\theta\in\Theta}|l_t(\theta) - E[l_t(\theta)]| \leq \sup_{\theta\in\Theta}|l_t(\theta)| + \sup_{\theta\in\Theta}E[|l_t(\theta)|] \leq \tilde{l}_t + E(\tilde{l}_t).
$$

Derivations similar with (B.7) yield

$$
\begin{aligned}
P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=\tau+1}^{T} L_t(\theta)\bar{\mathbb{1}}_t\right| > c_2\right) &\leq T^a \max_{\tau+1\leq t\leq T} P\left(\sup_{\theta\in\Theta}|L_t(\theta)| > \sqrt{T}\right) \\
&\leq \frac{T^a \max_{\tau+1\leq t\leq T} E\left[\left|\tilde{l}_t + E(\tilde{l}_t)\right|^q\right]}{T^{q/2}} \\
&= O(T^{a-q/2}) = o(T^{-1}). \qquad \text{(B.23)}
\end{aligned}
$$

Third, we show (B.19). Since Θ is assumed to be compact, subsets $\Theta_j \subset \Theta$ exist for $j = 1, \ldots, n_\epsilon$ such that $\Theta \subset \cup_j^{n_\epsilon}\Theta_j$ and $\|\theta - \bar{\theta}\| \leq \epsilon/T$ for any $\epsilon > 0$ and $\theta, \bar{\theta} \in \Theta_j$, where $n_\epsilon = O(T^K)$.

By the Boole inequality, we obtain

$$
\begin{aligned}
&P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| > c_1\right)\\
&\leq \sum_{j=1}^{n_\epsilon} P\left(\sup_{\theta\in\Theta_j}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| > c_1\right). \qquad \text{(B.24)}
\end{aligned}
$$

When $\theta \in \Theta_j$, it holds by assumption that, for any $\bar{\theta} \in \Theta_j$,

$$
|L_t(\theta) - L_t(\bar{\theta})| = |l_t(\theta) - l_t(\bar{\theta}) + E[l_t(\bar{\theta}) - l_t(\theta)]| \leq \|\theta - \bar{\theta}\|[\tilde{l}_t + E(\tilde{l}_t)].
$$

Then,

$$
\begin{aligned}
&\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right|\\
&\leq \left|\frac{1}{T}\sum_{t=\tau+1}^{T}\left\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right|\\
&+\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\left\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t] - L_t(\bar{\theta})\mathbb{1}_t + E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right|\\
&\leq \left|\frac{1}{T}\sum_{t=\tau+1}^{T}\left\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right|\\
&+\frac{1}{T}\sum_{t=\tau+1}^{T}\left|L_t(\theta) - L_t(\bar{\theta})\right|\mathbb{1}_t + \frac{1}{T}\sum_{t=\tau+1}^{T} E\left[\left|L_t(\bar{\theta}) - L_t(\theta)\right|\mathbb{1}_t\right], \qquad \text{(B.25)}
\end{aligned}
$$

where the second and third terms of the right hand side are

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=\tau+1}^{T}\left|L_t(\theta)-L_t(\bar\theta)\right|\mathbb{1}_t+\frac{1}{T}\sum_{t=\tau+1}^{T}E\left[\left|L_t(\bar\theta)-L_t(\theta)\right|\mathbb{1}_t\right]\\
&\leq\frac{1}{T}\sum_{t=\tau+1}^{T}\left\|\theta-\bar\theta\right\|[\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t+\frac{1}{T}\sum_{t=\tau+1}^{T}\left\|\theta-\bar\theta\right\|E\{[\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t\}\\
&=\frac{1}{T}\sum_{t=\tau+1}^{T}\left\|\theta-\bar\theta\right\|\Big([\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t+E\{[\tilde{l}_t+E(\tilde{l}_t)]\mathbb{1}_t\}\Big)\\
&\leq\frac{\epsilon}{T^2}\sum_{t=\tau+1}^{T}\tilde{L}_t, \qquad\qquad \text{(B.26)}
\end{aligned}
$$

where $\tilde{L}_t \equiv [\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t + E\{[\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t\}$. Equations (B.25) and (B.26) indicate that

$$
\begin{aligned}
\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\{L_t(\theta)\mathbb{1}_t-E[L_t(\theta)\mathbb{1}_t]\}\right| \leq \left|\frac{1}{T}\sum_{t=\tau+1}^{T}\left\{L_t(\bar\theta)\mathbb{1}_t-E[L_t(\bar\theta)\mathbb{1}_t]\right\}\right|\\
+\frac{\epsilon}{T^2}\sum_{t=\tau+1}^{T}[\tilde{L}_t-E(\tilde{L}_t)]+\frac{\epsilon}{T^2}\sum_{t=\tau+1}^{T}E(\tilde{L}_t),
\end{aligned}
\qquad \text{(B.27)}
$$

where $E(\tilde{L}_t) = 2E\{[\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t\} < \infty$ by assumption. Then, equations (B.24) and (B.27) imply that

$$
\begin{aligned}
&P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\{L_t(\theta)\mathbb{1}_t-E[L_t(\theta)\mathbb{1}_t]\}\right|>c_1\right)\\
&\leq n_\epsilon P\left(\left|\frac{3}{T}\sum_{t=\tau+1}^{T}\left\{L_t(\bar\theta)\mathbb{1}_t-E[L_t(\bar\theta)\mathbb{1}_t]\right\}\right|>c_1\right)\\
&+n_\epsilon P\left(\left|\frac{3\epsilon}{T^2}\sum_{t=\tau+1}^{T}[\tilde{L}_t-E(\tilde{L}_t)]\right|>c_1\right), \qquad\qquad \text{(B.28)}
\end{aligned}
$$

since $P(|\frac{\epsilon}{T^2}\sum_{t=\tau+1}^{T}E(\tilde{L}_t)| > c_1) = 0$ by choosing ϵ small enough. Thus, it suffices to show that

$$n_\epsilon P\left(\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\left\{L_t(\bar\theta)\mathbb{1}_t - E[L_t(\bar\theta)\mathbb{1}_t]\right\}\right| > c_1\right) = o(T^{-1}), \tag{B.29}$$

$$n_\epsilon P\left(\left|\frac{\epsilon}{T^2}\sum_{t=\tau+1}^{T}[\tilde L_t - E(\tilde L_t)]\right| > c_1\right) = o(T^{-1}). \tag{B.30}$$

First, we show (B.29). Note that $L_t(\bar\theta)$ is a measurable function of the strong mixing process with the mixing coefficient satisfying $\alpha(\tau) \le c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $0 < \rho < 1$. Moreover, we have $\sup_{\tau+1\le t\le T}|L_t(\bar\theta)\mathbb{1}_t - E[L_t(\bar\theta)\mathbb{1}_t]| \le 2\sqrt{T}$. Thus, applying Theorem 2 in Merlevède et al. (2009) (see also Lemma S1.1 in Su et al., 2016) yields

$$\begin{aligned}
&Tn_\epsilon P\left(\left|\sum_{t=\tau+1}^{T}\left\{L_t(\bar\theta)\mathbb{1}_t - E[L_t(\bar\theta)\mathbb{1}_t]\right\}\right| > c_1 T\right)\\
&\le Tn_\epsilon \exp\left(-\frac{C_0 c_1^2 T^2}{v_0^2(T-\tau) + 4T + c_1 T\sqrt{T}[\log(T-\tau)]^2}\right)\\
&= \exp\left(-\frac{C_0 c_1^2 T^2}{v_0^2 T^a + 4T + c_1 T\sqrt{T}[\log(T^a)]^2} + K\log(T) + \log(T)\right) \qquad \text{(B.31)}
\end{aligned}$$

for some constant C_0 and v_o.[2] Since the right hand side of the above equation converges zero as $T \to \infty$, (B.29) holds.

Next, we show (B.30). By the Markov and H lder inequalities,

$$\begin{aligned}
n_\epsilon P\left(\left|\frac{\epsilon}{T^2}\sum_{t=\tau+1}^{T}[\tilde L_t - E(\tilde L_t)]\right| > c_1\right) &\le n_\epsilon \frac{\epsilon^q E\left[\left|\frac{1}{T}\sum_{t=\tau+1}^{T}[\tilde L_t - E(\tilde L_t)]\right|^q\right]}{c_1^q T^q}\\
&\le n_\epsilon \frac{\epsilon^q \frac{1}{T}\sum_{t=\tau+1}^{T} E\left[\left|\tilde L_t - E(\tilde L_t)\right|^q\right]}{c_1^q T^q}\\
&= O(T^{a-1+K-q}), \qquad \text{(B.32)}
\end{aligned}$$

[2]Let $L_t \equiv L_t(\bar\theta)\mathbb{1}_t - E[L_t(\bar\theta)\mathbb{1}_t]$. Then, $v_o \equiv \sup_{t\ge1}[\mathrm{Var}(L_t) + 2\sum_{s=t+1}^{\infty}\mathrm{Cov}(L_t, L_s)]$.

where the right hand side is $o(T^{-1})$ since $K + a < q$. $\square$

Lemma 15. *Let f_t be a measurable function of a strong mixing process with mixing coefficients $\alpha(\cdot)$, where $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $0 < \rho < 1$, and $E|f_t|^q \leq C$ for some integer $q > 4$ and $C < \infty$. For any $c > 0$, we have*

$$P\left(\frac{1}{T}\sum_{t=1}^{\tau}[f_t - E(f_t)] \leq c\right) = 1 - o(T^{-1}). \tag{B.33}$$

Proof. Letting $L_t = f_t - E(f_t)$, we denote $\mathbb{1}_t = \mathbb{1}\{|f_t - E(f_t)| \leq \sqrt{T}\} = \mathbb{1}\{|L_t| \leq \sqrt{T}\}$ and $\bar{\mathbb{1}}_t = 1 - \mathbb{1}_t$. Since $E(L_t) = 0$, we have

$$L_t = L_t\mathbb{1}_t - E(L_t\mathbb{1}_t) + L_t\bar{\mathbb{1}}_t - E(L_t\bar{\mathbb{1}}_t). \tag{B.34}$$

Then, it suffices to show that for any constants $c_1 > 0$ and $c_2 > 0$,

$$TP\left(\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t\mathbb{1}_t - E(L_t\mathbb{1}_t)\}\right| > c_1\right) = o(1), \tag{B.35}$$

$$TP\left(\left|\frac{1}{T}\sum_{t=1}^{\tau}L_t\bar{\mathbb{1}}_t\right| > c_2\right) = o(1), \tag{B.36}$$

$$\left|\frac{1}{T}\sum_{t=1}^{\tau}E(L_t\bar{\mathbb{1}}_t)\right| = o(1). \tag{B.37}$$

First, we show (B.37). By assumption, a positive constant C exists such that

$E[L_t^2] = E[f_t^2] - E[f_t]^2 \leq E[f_t^2] < C$. Then, by the H lder inequality,

$$\begin{aligned}
\left| \frac{1}{T} \sum_{t=1}^{\tau} E(L_t \bar{\mathbb{1}}_t) \right| &\leq \frac{1}{T} \sum_{t=1}^{\tau} |E(L_t \bar{\mathbb{1}}_t)| \\
&\leq \frac{1}{T} \sum_{t=1}^{\tau} \left| E[L_t^2]^{1/2} E(\bar{\mathbb{1}}_t)^{1/2} \right| \\
&\leq \frac{C^{1/2}}{T} \sum_{t=1}^{\tau} \left| E(\bar{\mathbb{1}}_t)^{1/2} \right| \\
&= \frac{C^{1/2}}{T} \sum_{t=1}^{\tau} \left| P\left(|L_t| > \sqrt{T}\right)^{1/2} \right| \\
&\leq \frac{C^{1/2}}{T} \sum_{t=1}^{\tau} \left| \left(\frac{E[L_t^2]}{T} \right)^{1/2} \right| \\
&\leq \frac{C^{1/2}}{T\sqrt{T}} \sum_{t=1}^{\tau} \left| E[L_t^2]^{1/2} \right| \\
&\leq \frac{C\tau}{T\sqrt{T}} = O(T^{-1/2}).
\end{aligned} \tag{B.38}$$

Second, we show (B.36). Note that, by the assumption, a constant C exists such that $E(|L_t|^q) = E[|f_t - E(f_t)|^q] < C$. Derivations similar with (B.7) yield

$$\begin{aligned}
P\left(\left| \frac{1}{T} \sum_{t=1}^{\tau} L_t \bar{\mathbb{1}}_t \right| > c_2 \right) &\leq \tau \max_{1 \leq t \leq \tau} P\left(|L_t| > \sqrt{T}\right) \\
&\leq \frac{\tau \max_{1 \leq t \leq \tau} E\left[|f_t - E(f_t)|^q\right]}{T^{q/2}} \\
&= O(T^{1-q/2}) = o(T^{-1}).
\end{aligned} \tag{B.39}$$

Third, we show (B.35). Note that L_t is a measurable function of the strong mixing process with the mixing coefficient satisfying $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $0 < \rho < 1$. Moreover, we have $\sup_{1 \leq t \leq \tau} |L_t \mathbb{1}_t - E(L_t \mathbb{1}_t)| \leq 2\sqrt{T}$. Thus, applying Theorem 2 in Merlevède et al. (2009) (see also Lemma S1.1 in

Su et al., 2016) yields

$$TP\left(\left|\sum_{t=1}^{\tau}\{L_t\mathbb{1}_t - E(L_t\mathbb{1}_t)\}\right| > c_1 T\right) \le T\exp\left(-\frac{C_0c_1^2T^2}{v_0^2T + 4T + c_1T\sqrt{T}[\log(T)]^2}\right)$$
$$= \exp\left(-\frac{C_0c_1^2T}{v_0^2 + 4 + c_1\sqrt{T}[\log(T)]^2} + \log(T)\right) \tag{B.40}$$

for some constant C_0 and v_o.[3] Since the right hand side of the above equation converges zero as $T \to \infty$, (B.35) holds. ☐

Lemma 16. *Let f_t be a measurable function of a strong mixing process with mixing coefficients $\alpha(\cdot)$, where $\alpha(\tau) \le c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $0 < \rho < 1$, and $E|f_t|^q \le C$ for some integer $q > 4$ and $C < \infty$. For any $c > 0$, we have*

$$P\left(\frac{1}{T}\sum_{t=\tau+1}^{T}[f_t - E(f_t)] \le c\right) = 1 - o(T^{-1}). \tag{B.41}$$

Proof. Letting $L_t = f_t - E(f_t)$, we denote $\mathbb{1}_t = \mathbb{1}\{|f_t - E(f_t)| \le \sqrt{T}\} = \mathbb{1}\{|L_t| \le \sqrt{T}\}$ and $\bar{\mathbb{1}}_t = 1 - \mathbb{1}_t$. Since $E(L_t) = 0$, we have

$$L_t = L_t\mathbb{1}_t - E(L_t\mathbb{1}_t) + L_t\bar{\mathbb{1}}_t - E(L_t\bar{\mathbb{1}}_t). \tag{B.42}$$

Then, it suffices to show that for any constants $c_1 > 0$ and $c_2 > 0$,

$$TP\left(\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\{L_t\mathbb{1}_t - E(L_t\mathbb{1}_t)\}\right| > c_1\right) = o(1), \tag{B.43}$$

$$TP\left(\left|\frac{1}{T}\sum_{t=\tau+1}^{T}L_t\bar{\mathbb{1}}_t\right| > c_2\right) = o(1), \tag{B.44}$$

$$\left|\frac{1}{T}\sum_{t=\tau+1}^{T}E(L_t\bar{\mathbb{1}}_t)\right| = o(1). \tag{B.45}$$

[3]Let $\bar{L}_t \equiv L_t\mathbb{1}_t - E(L_t\mathbb{1}_t)$. Then, $v_o \equiv \sup_{t\ge1}[\mathrm{Var}(\bar{L}_t) + 2\sum_{s=t+1}^{\infty}\mathrm{Cov}(\bar{L}_t, \bar{L}_s)]$.

First, we show (B.45). By assumption, a positive constant C exists such that $E[L_t^2] = E[f_t^2] - E[f_t]^2 \leq E[f_t^2] < C$. Then, by the H lder inequality,

$$\begin{aligned}
\left| \frac{1}{T} \sum_{t=\tau+1}^{T} E(L_t \bar{\mathbb{1}}_t) \right| &\leq \frac{1}{T} \sum_{t=\tau+1}^{T} |E(L_t \bar{\mathbb{1}}_t)| \\
&\leq \frac{1}{T} \sum_{t=\tau+1}^{T} \left| E[L_t^2]^{1/2} E(\bar{\mathbb{1}}_t)^{1/2} \right| \\
&\leq \frac{C^{1/2}}{T} \sum_{t=\tau+1}^{T} \left| E(\bar{\mathbb{1}}_t)^{1/2} \right| \\
&= \frac{C^{1/2}}{T} \sum_{t=\tau+1}^{T} \left| P\left(|L_t| > \sqrt{T} \right)^{1/2} \right| \\
&\leq \frac{C^{1/2}}{T} \sum_{t=\tau+1}^{T} \left| \left(\frac{E[L_t^2]}{T} \right)^{1/2} \right| \\
&\leq \frac{C^{1/2}}{T\sqrt{T}} \sum_{t=\tau+1}^{T} \left| E[L_t^2]^{1/2} \right| \\
&\leq C T^{a-3/2} = O(T^{a-3/2}). \qquad \text{(B.46)}
\end{aligned}$$

Second, we show (B.44). Note that, by the assumption, a constant C exists such that $E(|L_t|^q) = E[|f_t - E(f_t)|^q] < C$. Derivations similar with (B.7) yield

$$\begin{aligned}
P\left(\left| \frac{1}{T} \sum_{t=\tau+1}^{T} L_t \bar{\mathbb{1}}_t \right| > c_2 \right) &\leq T^a \max_{\tau+1 \leq t \leq T} P\left(|L_t| > \sqrt{T} \right) \\
&\leq \frac{T^a \max_{\tau+1 \leq t \leq T} E\left[|f_t - E(f_t)|^q \right]}{T^{q/2}} \\
&= O(T^{a-q/2}) = o(T^{-1}). \qquad \text{(B.47)}
\end{aligned}$$

Third, we show (B.43). Note that L_t is a measurable function of the strong mixing process with the mixing coefficient satisfying $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $0 < \rho < 1$. Moreover, we have $\sup_{1 \leq t \leq \tau} |L_t \mathbb{1}_t - E(L_t \mathbb{1}_t)| \leq 2\sqrt{T}$. Thus, applying Theorem 2 in Merlevède et al. (2009) (see also Lemma S1.1 in

Su et al., 2016) yields

$$
\begin{aligned}
&TP\left(\left|\sum_{t=\tau+1}^{T}\{L_t\mathbb{1}_t - E(L_t\mathbb{1}_t)\}\right| > c_1T\right)\\
\leq &T\exp\left(-\frac{C_0c_1^2T^2}{v_0^2(T-\tau)+4T+c_1T\sqrt{T}[\log(T-\tau)]^2}\right) \qquad \text{(B.48)}\\
= &\exp\left(-\frac{C_0c_1^2T}{v_0^2T^{a-1}+4+c_1\sqrt{T}[a\log(T)]^2}+\log(T)\right) \qquad \text{(B.49)}
\end{aligned}
$$

for some constant C_0 and v_o.[4] Since the right hand side of the above equation converges zero as $T \to \infty$, (B.43) holds. □

Lemma 17. *Let the parameter space Θ be a compact and convex subset of $\mathbb{R}^K$. For all $\theta \in \Theta$, $l_t(\theta)$ is a measurable function of a strong mixing process with mixing coefficients $\alpha(\cdot)$, where $\alpha(\tau) \leq c_\alpha\rho^\tau$ for some $c_\alpha > 0$ and $0 < \rho < 1$. A measurable function $\tilde{l}_t$ exists such that $|l_t(\theta) - l_t(\bar{\theta})| \leq \|\theta - \bar{\theta}\|\tilde{l}_t$ for any $\theta, \bar{\theta} \in \Theta$, $\sup_{\theta\in\Theta}|l_t(\theta)| \leq \tilde{l}_t$, and $E(|\tilde{l}_t|^q) \leq C$ for some integer $q > \max\{K, 2\}$, where $C < \infty$ is a constant. For any $T \geq 2$ and arbitrary small $\epsilon_S > 0$, a constant S exists such that*

$$
P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{l_t(\theta) - E[l_t(\theta)]\}\right| \geq S\right) \leq \epsilon_S. \qquad \text{(B.50)}
$$

Proof. Letting $L_t(\theta) = l_t(\theta) - E[l_t(\theta)]$, we denote $\mathbb{1}_t = \mathbb{1}\{|l_t(\theta) - E[l_t(\theta)]| \leq \sqrt{ST}\} = \mathbb{1}\{|L_t(\theta)| \leq \sqrt{ST}\}$ and $\bar{\mathbb{1}}_t = 1 - \mathbb{1}_t$. Since $E[L_t(\theta)] = 0$, we have

$$
L_t(\theta) = L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t] + L_t(\theta)\bar{\mathbb{1}}_t - E[L_t(\theta)\bar{\mathbb{1}}_t]. \qquad \text{(B.51)}
$$

[4]Let $\bar{L}_t \equiv L_t\mathbb{1}_t - E(L_t\mathbb{1}_t)$. Then, $v_o \equiv \sup_{t\geq1}[\mathrm{Var}(\bar{L}_t) + 2\sum_{s=t+1}^{\infty}\mathrm{Cov}(\bar{L}_t, \bar{L}_s)]$.

Then, we obtain

$$\begin{aligned}P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}L_t(\theta)\right|\geq S\right) \leq\ & P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right|\geq S/3\right)\\ &+P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}L_t(\theta)\bar{\mathbb{1}}_t\right|\geq S/3\right)\\ &+P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}E[L_t(\theta)\bar{\mathbb{1}}_t]\right|\geq S/3\right) \qquad \text{(B.52)}\end{aligned}$$

By assumptions, it holds that

$$\sup_{\theta\in\Theta}|L_t(\theta)| = \sup_{\theta\in\Theta}|l_t(\theta)-E[l_t(\theta)]| \leq \sup_{\theta\in\Theta}|l_t(\theta)| + \sup_{\theta\in\Theta}E[|l_t(\theta)|] \leq \tilde{l}_t + E(\tilde{l}_t).$$

Thus, a constant C exists such that

$$c_t \equiv E[(\sup_{\theta\in\Theta}|L_t(\theta)|)^q] \leq E[|\tilde{l}_t + E(\tilde{l}_t)|^q] \equiv \tilde{c}_t \leq C. \qquad \text{(B.53)}$$

First, with respect to the third term of equation (B.52), it holds, by the H lder and Chebyshev inequalities, that

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}E[L_t(\theta)\bar{\mathbb{1}}_t]\right| &\leq \frac{1}{T}\sum_{t=1}^{\tau}\sup_{\theta\in\Theta}|E[L_t(\theta)\bar{\mathbb{1}}_t]|\\ &\leq \frac{\tau}{T}\max_{1\leq t\leq\tau}\sup_{\theta\in\Theta}\left|E[L_t(\theta)^q]^{1/q}E(\bar{\mathbb{1}}_t)^{1/q}\right|\\ &\leq \frac{\tau}{T}\max_{1\leq t\leq\tau}c_t^{1/q}\max_{1\leq t\leq\tau}\sup_{\theta\in\Theta}\left|E(\bar{\mathbb{1}}_t)^{1/q}\right|\\ &\leq \frac{\tau}{T}\max_{1\leq t\leq\tau}c_t^{1/q}\max_{1\leq t\leq\tau}\left|P\left(\sup_{\theta\in\Theta}|L_t(\theta)|>\sqrt{ST}\right)^{1/q}\right|\\ &\leq \frac{\tau}{T}\max_{1\leq t\leq\tau}c_t^{1/q}\max_{1\leq t\leq\tau}\left\{\frac{E\{[\sup_{\theta\in\Theta}|L_t(\theta)|]^q\}}{(ST)^{q/2}}\right\}^{1/q}\\ &= \frac{\tau}{T\sqrt{ST}}\max_{1\leq t\leq\tau}c_t^{2/q}. \qquad \text{(B.54)}\end{aligned}$$

Thus, when $S^{3/2} > \frac{3\tau}{T\sqrt{T}}\max_{1\leq t\leq\tau}c_t^{2/q}$, the probability in the third term of equation (B.52) is zero.

Second, we consider an upper bound of the second term of equation (B.52). Note that, by the definition of $\bar{\mathbb{1}}_t = \mathbb{1}\{|L_t(\theta)| > \sqrt{ST}\}$, $\left|\frac{1}{T}\sum_{t=1}^{\tau} L_t(\theta)\bar{\mathbb{1}}_t\right| > S/3$ implies that $\max_{1\leq t\leq \tau}\sup_{\theta\in\Theta}|L_t(\theta)| > \sqrt{ST}$. Thus, by Boole and Markov inequalities,

$$
\begin{aligned}
P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau} L_t(\theta)\bar{\mathbb{1}}_t\right| > S/3\right) &\leq P\left(\max_{1\leq t\leq\tau}\sup_{\theta\in\Theta}|L_t(\theta)| > \sqrt{ST}\right) \\
&\leq \tau \max_{1\leq t\leq\tau} P\left(\sup_{\theta\in\Theta}|L_t(\theta)| > \sqrt{ST}\right) \\
&\leq \frac{\tau\max_{1\leq t\leq\tau} E\left\{[\sup_{\theta\in\Theta}|L_t(\theta)|]^q\right\}}{(ST)^{q/2}} \\
&= \frac{\tau}{(ST)^{q/2}}\max_{1\leq t\leq\tau} c_t \qquad \text{(B.55)}
\end{aligned}
$$

Third, we consider an upper bound of the first term of equation (B.52). Since Θ is assumed to be compact, subsets $\Theta_j \subset \Theta$ exist for $j = 1, \ldots, n_\epsilon$ such that $\Theta \subset \cup_{j=1}^{n_\epsilon}\Theta_j$ and $\|\theta - \bar{\theta}\| \leq \epsilon/T$ for any $\epsilon > 0$ and $\theta, \bar{\theta} \in \Theta_j$, where $n_\epsilon = O(T^K)$.

By the Cauchy–Schwartz inequality, we have

$$
\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| \\
&\leq \left|\frac{1}{T}\sum_{t=1}^{\tau}\left\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right| \\
&+ \left|\frac{1}{T}\sum_{t=1}^{\tau}\left\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t] - L_t(\bar{\theta})\mathbb{1}_t + E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right| \\
&\leq \left|\frac{1}{T}\sum_{t=1}^{\tau}\left\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\right\}\right| \\
&+ \frac{1}{T}\sum_{t=1}^{\tau}\left|L_t(\theta) - L_t(\bar{\theta})\right|\mathbb{1}_t + \frac{1}{T}\sum_{t=1}^{\tau} E\left[\left|L_t(\bar{\theta}) - L_t(\theta)\right|\mathbb{1}_t\right]. \qquad \text{(B.56)}
\end{aligned}
$$

Since $\tilde{l}_t$ exists, by assumption, such that $|L_t(\theta) - L_t(\bar{\theta})| = |l_t(\theta) - l_t(\bar{\theta}) + E[l_t(\bar{\theta}) -$

$l_t(\theta)]| \leq \|\theta - \bar{\theta}\|[\tilde{l}_t + E(\tilde{l}_t)]$, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{\tau}\left|L_t(\theta) - L_t(\bar{\theta})\right|\mathbb{1}_t + \frac{1}{T}\sum_{t=1}^{\tau}E\left[\left|L_t(\bar{\theta}) - L_t(\theta)\right|\mathbb{1}_t\right] \\
&\leq \frac{1}{T}\sum_{t=1}^{\tau}\left\|\theta - \bar{\theta}\right\|[\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t + \frac{1}{T}\sum_{t=1}^{\tau}\left\|\theta - \bar{\theta}\right\|E\{[\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t\} \\
&= \frac{1}{T}\sum_{t=1}^{\tau}\left\|\theta - \bar{\theta}\right\|\Big([\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t + E\{[\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t\}\Big) \\
&\leq \frac{\epsilon}{T^2}\sum_{t=1}^{\tau}\tilde{L}_t,
\end{aligned} \tag{B.57}$$

where $\tilde{L}_t \equiv [\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t + E\{[\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t\}$. Then, equations (B.56) and (B.57) indicate that

$$\begin{aligned}
\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| &\leq \left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\}\right| \\
&\quad + \frac{\epsilon}{T^2}\sum_{t=1}^{\tau}[\tilde{L}_t - E(\tilde{L}_t)] + \frac{\epsilon}{T^2}\sum_{t=1}^{\tau}E(\tilde{L}_t),
\end{aligned} \tag{B.58}$$

Equation (B.58) and the Boole inequality yield

$$\begin{aligned}
&P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| > S/3\right) \\
&\leq \sum_{j=1}^{n_\epsilon}P\left(\sup_{\theta\in\Theta_j}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\}\right| > S/3\right). \\
&\leq n_\epsilon P\left(\left|\frac{3}{T}\sum_{t=1}^{\tau}\{L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]\}\right| > S/3\right) \\
&\quad + n_\epsilon P\left(\left|\frac{3\epsilon}{T^2}\sum_{t=1}^{\tau}[\tilde{L}_t - E(\tilde{L}_t)]\right| > S/3\right) \\
&\quad + n_\epsilon P\left(\left|\frac{3\epsilon}{T^2}\sum_{t=1}^{\tau}E(\tilde{L}_t)\right| > S/3\right)
\end{aligned} \tag{B.59}$$

With respect to the third term of the right hand side, it is able to select ϵ small enough such that $P(|\frac{3\epsilon}{T^2}\sum_{t=1}^{\tau}E(\tilde{L}_t)| > c_1) = 0$, because $E(\tilde{L}_t) =$

$2E\{[\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t\}$ is assumed to be bounded above by a constant. With respect to the first term, $L_t(\bar{\theta})$ is a measurable function of the strong mixing process with the mixing coefficient satisfying $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $0 < \rho < 1$. Moreover, since $\mathbb{1}_t = \mathbb{1}\{|L_t(\theta)| \leq \sqrt{ST}\}$, it holds that $\sup_{1\leq t\leq\tau} |L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]| \leq 2\sqrt{ST}$. Thus, applying Theorem 2 in Merlevède et al. (2009) (see also Lemma S1.1 in Su et al. 2016) yields

$$\begin{aligned} n_\epsilon P &\left(\left| \sum_{t=1}^{\tau} \left\{ L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t] \right\} \right| > TS/9 \right) \\ &\leq n_\epsilon \exp\left(-\frac{C_0 S^2 T}{v_0^2 + 4\sqrt{S} + (2/9)S\sqrt{ST}[\log(T)]^2} \right) \end{aligned} \tag{B.60}$$

for any $T \geq 2$ and any S and some constants C_0 and v_o.[5]

With respect to the second term, it holds that $\tilde{L}_t - E(\tilde{L}_t) = [\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t - E\{[\tilde{l}_t + E(\tilde{l}_t)]\mathbb{1}_t\} \leq \tilde{l}_t + E(\tilde{l}_t)$, since $\tilde{l}_t \geq 0$. Then, the Markov and H lder inequalities yield

$$\begin{aligned} n_\epsilon P\left(\left| \frac{\epsilon}{T^2} \sum_{t=1}^{\tau} [\tilde{L}_t - E(\tilde{L}_t)] \right| > S/9 \right) &\leq \frac{3^{2q} n_\epsilon \epsilon^q}{S^q T^q} E\left[\left| \frac{1}{T} \sum_{t=1}^{\tau} [\tilde{L}_t - E(\tilde{L}_t)] \right|^q \right] \\ &\leq \frac{3^{2q} n_\epsilon \epsilon^q}{S^q T^q} \frac{1}{T} \sum_{t=1}^{\tau} E\left[\left| \tilde{l}_t + E(\tilde{l}_t) \right|^q \right] \\ &\leq \frac{3^{2q} n_\epsilon \epsilon^q}{S^q T^q} \frac{\tau}{T} \max_{1\leq t\leq\tau} \tilde{c}_t. \end{aligned} \tag{B.61}$$

Therefore, we obtain

$$\begin{aligned} &P\left(\sup_{\theta\in\Theta} \left| \frac{1}{T} \sum_{t=1}^{\tau} \{L_t(\theta)\mathbb{1}_t - E[L_t(\theta)\mathbb{1}_t]\} \right| > S/3 \right) \\ &\leq n_\epsilon \exp\left(-\frac{C_0 S^2 T}{v_0^2 + 4\sqrt{S} + (2/9)S\sqrt{ST}[\log(T)]^2} \right) + \frac{3^{2q} n_\epsilon \epsilon^q}{S^q T^q} \frac{\tau}{T} \max_{1\leq t\leq\tau} \tilde{c}_t. \end{aligned} \tag{B.62}$$

[5] Let $L_t \equiv L_t(\bar{\theta})\mathbb{1}_t - E[L_t(\bar{\theta})\mathbb{1}_t]$. Then, $v_o \equiv \sup_{t\geq 1}[\mathrm{Var}(L_t) + 2\sum_{s=t+1}^{\infty} \mathrm{Cov}(L_t, L_s)]$.

Equations (B.52), (B.54), (B.55), and (B.62) indicate that for

$$S > \left(\frac{3\tau}{T\sqrt{T}} \max_{1 \le t \le \tau} c_t^{2/q} \right)^{3/2},$$

$$\begin{aligned}
&P\left(\sup_{\theta \in \Theta} \left| \frac{1}{T} \sum_{t=1}^{\tau} L_t(\theta) \right| \ge S \right) \\
&\le n_\epsilon \exp\left(-\frac{C_0 S^2 T}{v_0^2 + 4\sqrt{S} + (2/9)S\sqrt{ST}[\log(T)]^2} \right) + \frac{3^{2q} n_\epsilon \epsilon^q}{S^q T^q} \frac{\tau}{T} \max_{1 \le t \le \tau} \tilde{c}_t + \frac{\tau}{(ST)^{q/2}} \max_{1 \le t \le \tau} c_t \\
&\equiv \epsilon_S, \qquad \text{(B.63)}
\end{aligned}$$

where ϵ_S is bounded above uniformly in T when $q > \max\{K, 2\}$. Since ϵ_S is a decreasing function of S, ϵ_S can be as small as possible by taking S large. $\square$

Lemma 18. *Let Assumptions 1, 2, and 3 hold. For any $T \ge 2$ and arbitrary small $\epsilon_S > 0$, a positive constant S exists such that*

$$P(\|\nabla_\theta Q_T(\hat{\theta})\| > S) \le \epsilon_S. \qquad \text{(B.64)}$$

Proof. Since

$$
\begin{aligned}
\|\nabla_\theta Q_T(\hat{\theta})\|^2 =& \sum_{k=1}^{K} |\nabla_{\theta_k} Q_T(\hat{\theta})|^2 \\
\le& K \max_{1\le k\le K} |\nabla_{\theta_k} Q_T(\hat{\theta})|^2 \\
=& K \max_{1\le k\le K} \left| \frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_k} l_{I_1,t}(\hat{\theta}_{I_1}) + \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla_{\theta_k} l_{I_2,t}(\hat{\theta}_{I_2}) \right|^2 \\
\le& K \max_{1\le k\le K} \left| \frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_k} l_{I_1,t}(\hat{\theta}_{I_1}) \right|^2 \\
&+ K \max_{1\le k\le K} \left| \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla_{\theta_k} l_{I_2,t}(\hat{\theta}_{I_2}) \right|^2 \\
&+ 2K \max_{1\le k\le K} \left| \frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_k} l_{I_1,t}(\hat{\theta}_{I_1}) \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla_{\theta_k} l_{I_2,t}(\hat{\theta}_{I_2}) \right|, \quad \text{(B.65)}
\end{aligned}
$$

we obtain that

$$
\begin{aligned}
&P(\|\nabla_\theta Q_T(\hat{\theta})\| > S) = P(\|\nabla_\theta Q_T(\hat{\theta})\|^2 > S^2) \\
&\quad \le P\left(\max_{1\le k\le K} \left| \frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_k} l_{I_1,t}(\hat{\theta}_{I_1}) \right|^2 > \frac{S^2}{3K} \right) \\
&\quad + P\left(\max_{1\le k\le K} \left| \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla_{\theta_k} l_{I_2,t}(\hat{\theta}_{I_2}) \right|^2 > \frac{S^2}{3K} \right) \\
&\quad + P\left(2 \max_{1\le k\le K} \left| \frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_k} l_{I_1,t}(\hat{\theta}_{I_1}) \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla_{\theta_k} l_{I_2,t}(\hat{\theta}_{I_2}) \right| > \frac{S^2}{3K} \right). \quad \text{(B.66)}
\end{aligned}
$$

Then, it suffices to show that, for any $T \ge 2$ and arbitrary small $\epsilon_1, \epsilon_2, \epsilon_3 > 0$, a positive constants S_1, S_2, S_3 exist such that

$$P\left(\sup_{\theta_{I_1}\in\Theta_{I_1}}\left|\frac{1}{T}\sum_{t=1}^{\tau}\nabla_{\theta_k} l_{I_1,t}(\theta_{I_1})\right| > S_1\right) \le \epsilon_1 \tag{B.67}$$

$$P\left(\sup_{\theta_{I_2}\in\Theta_{I_2}}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\nabla_{\theta_k} l_{I_2,t}(\theta_{I_2})\right| > S_2\right) \le \epsilon_2 \tag{B.68}$$

$$P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{\tau}\nabla_{\theta_k} l_{I_1,t}(\theta_{I_1})\frac{1}{T}\sum_{t=\tau+1}^{T}\nabla_{\theta_k} l_{I_2,t}(\theta_{I_2})\right| > S_3\right) \le \epsilon_3, \tag{B.69}$$

for all $k = 1, \ldots, K$. We consider the cases in which $\theta_k \in \Theta_{I_1} \cap \Theta_{I_2}$ because otherwise the above probabilities can be zero. Furthermore, it suffices to show (B.67) and (B.68), since then the existence of S_3 in (B.69) is implied.

First, let us consider (B.67). It holds that $\sup_{\theta_{I_1}\in\Theta_{I_1}}\left|\frac{1}{T}\sum_{t=1}^{\tau} E[\nabla_{\theta_k} l_{I_1,t}(\theta_{I_1})]\right| \le \frac{\tau}{T}c_l$ by Assumption 3 (2). Then, we obtain

$$\begin{aligned}
&P\left(\sup_{\theta_{I_1}\in\Theta_{I_1}}\left|\frac{1}{T}\sum_{t=1}^{\tau}\nabla_{\theta_k} l_{I_1,t}(\theta_{I_1})\right| > S_1\right)\\
&\le P\left(\sup_{\theta_{I_1}\in\Theta_{I_1}}\left|\frac{1}{T}\sum_{t=1}^{\tau}\{\nabla_{\theta_k} l_{I_1,t}(\theta_{I_1}) - E[\nabla_{\theta_k} l_{I_1,t}(\theta_{I_1})]\}\right| > S_1 - \frac{\tau}{T}c_l\right).
\end{aligned} \tag{B.70}$$

Under Assumptions 1, 2, and 3, we can apply Lemma 17 to the right hand side of the above equation, where S in Lemma 17 is $S_1 - \frac{\tau}{T}c_l$. Let $c_{t,1} \equiv E[(\sup_{\theta_{I_1}\in\Theta_{I_1}}|\nabla_{\theta_k} l_{I_1,t}(\theta_{I_1}) - E[\nabla_{\theta_k} l_{I_1,t}(\theta_{I_1})]|)^q]$, which is, by Assumption 3 (2), bounded above by a constant. Then, for any $S_1 > \left(\frac{3\tau}{T\sqrt{T}}\max_{1\le t\le\tau} c_{t,1}^{2/q}\right)^{3/2} + \frac{\tau}{T}c_l$, we obtain

$$P\left(\sup_{\theta_{I_1}\in\Theta_{I_1}}\left|\frac{1}{T}\sum_{t=1}^{\tau}\nabla_{\theta_k} l_{I_1,t}(\theta_{I_1})\right| > S_1\right) \le \epsilon_{S_1}, \tag{B.71}$$

where ϵ_{S_1} is bounded above uniformly in T when $q > \max\{K_1, 2\}$, independent of k, and a decreasing function of S_1. Thus, ϵ_{S_1} can be as small as possible by taking S_1 large. This shows equation (B.67).

Second, equation (B.68) can be shown in the similar manner. Since

$$P\left(\sup_{\theta_{I_2}\in\Theta_{I_2}}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\nabla_{\theta_k}l_{I_2,t}(\theta_{I_2})\right|>S_2\right)$$
$$\leq P\left(\sup_{\theta_{I_2}\in\Theta_{I_2}}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\{\nabla_{\theta_k}l_{I_2,t}(\theta_{I_2})-E[\nabla_{\theta_k}l_{I_2,t}(\theta_{I_2})]\}\right|>S_2-\frac{T-\tau}{T}c_l\right), \tag{B.72}$$

we can apply Lemma 17 to the right hand side of the above equation, where S in Lemma 17 is $S_2-\frac{T-\tau}{T}c_l$. Letting $c_{t,2}\equiv E[(\sup_{\theta_{I_2}\in\Theta_{I_2}}|\nabla_{\theta_k}l_{I_2,t}(\theta_{I_2})-E[\nabla_{\theta_k}l_{I_2,t}(\theta_{I_2})]|)^q]$, which is, by Assumption 3 (2), bounded above by a constant, we obtain, for any $S_2>\left(\frac{3(T-\tau)}{T\sqrt{T}}\max_{\tau\leq t\leq T}c_{t,2}^{2/q}\right)^{3/2}+\frac{T-\tau}{T}c_l$, that

$$P\left(\sup_{\theta_{I_2}\in\Theta_{I_2}}\left|\frac{1}{T}\sum_{t=\tau+1}^{T}\nabla_{\theta_k}l_{I_2,t}(\theta_{I_2})\right|>S_2\right)\leq\epsilon_{S_2}, \tag{B.73}$$

where ϵ_{S_2} is bounded above uniformly in T when $q>\max\{K_2,2\}$, independent of k, and a decreasing function of S_2. Thus, ϵ_{S_2} can be as small as possible by taking S_2 large. This shows equation (B.68). □

B.2 Proofs of results

Proof of Lemma 1

Proof. Since $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} Q_\lambda(\theta)$, we have $Q_\lambda(\hat{\theta}) \leq Q_\lambda(\theta_0)$, that is, $Q_T(\hat{\theta}) + \lambda \|W'\hat{\theta}\|^2 \leq Q_T(\theta_0) + \lambda \|W'\theta_0\|^2$. Let $\epsilon = \inf_{\theta : \|\theta - \theta_0\| > \delta} Q_P(\theta) - Q_P(\theta_0)$, where $\epsilon > 0$ by the uniqueness of θ_0 in Assumption 2. Define an event $A = \{\sup_{\theta \in \Theta} |Q_T(\theta) - Q_P(\theta)| \leq \epsilon/3\}$, where $P(A) = 1 - P(A^c) = 1 - o(T^{-1})$ by Lemmas 13 and 14 under Assumptions 1, 2, and 3 (2). Conditional on A, we obtain

$$\begin{aligned}
\inf_{\theta : \|\theta - \theta_0\| > \delta} Q_T(\theta) + \lambda \|W'\theta\|^2 &\geq \inf_{\theta : \|\theta - \theta_0\| > \delta} [Q_P(\theta) + Q_T(\theta) - Q_P(\theta)] \\
&\geq \inf_{\theta : \|\theta - \theta_0\| > \delta} Q_P(\theta) - \frac{\epsilon}{3} \\
&= Q_P(\theta_0) + \epsilon - \frac{\epsilon}{3} \\
&\geq Q_T(\theta_0) + \epsilon - \frac{\epsilon}{3} - \frac{\epsilon}{3} \\
&= Q_T(\theta_0) + \lambda \|W'\theta_0\|^2 + \frac{\epsilon}{3} - \lambda \|W'\theta_0\|^2 \\
&\geq Q_T(\hat{\theta}) + \lambda \|W'\hat{\theta}\|^2 + \frac{\epsilon}{3} - \lambda \|W'\theta_0\|^2. \qquad \text{(B.74)}
\end{aligned}$$

When $\epsilon/3 \geq \lambda \|W'\theta_0\|^2$, the above inequality implies $\inf_{\theta : \|\theta - \theta_0\| > \delta} Q_T(\theta) + \lambda \|W'\theta\|^2 \geq Q_T(\hat{\theta}) + \lambda \|W'\hat{\theta}\|^2$, which implies $\|\hat{\theta} - \theta_0\| \leq \delta$.

Since $\epsilon/3 \geq \lambda \|W'\theta_0\|^2$ holds when $\|W'\theta_0\| = 0$, we consider the case of $\|W'\theta_0\| > 0$. Since the parameter space is assumed to be compact in Assumption 2, $\|W'\theta\|^2$ is bounded from above, implying that $\epsilon/3 \geq \lambda \|W'\theta\|^2$ holds for T large enough. Thus, we obtain

$$\begin{aligned}
P(\|\hat{\theta} - \theta_0\| \leq \delta) &= P(\|\hat{\theta} - \theta_0\| \leq \delta | A) P(A) + P(\|\hat{\theta} - \theta_0\| \leq \delta | A^c) P(A^c) \\
&\geq 1 - o(T^{-1}). \qquad \text{(B.75)}
\end{aligned}$$

$\square$

Proof of Theorem 2

Proof. By Assumption 3 (1), $\tilde{\theta} = (\tilde{\theta}_{I_1}', \tilde{\check{\theta}}')' \in \Theta$ exists such that it lies between $\hat{\theta}$ and θ_0 element-wise and satisfies

$$Q_T(\hat{\theta}) - Q_T(\theta_0) = (\hat{\theta} - \theta_0)' \nabla_\theta Q_T(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)' \nabla_{\theta\theta'} Q_T(\tilde{\theta})(\hat{\theta} - \theta_0). \quad \text{(B.76)}$$

We denote the K-dimensional vector $\nabla_\theta Q_T(\theta_0)$ by

$$\begin{aligned}\nabla_\theta Q_T(\theta_0) =& \frac{1}{T}\begin{pmatrix} \sum_{t=1}^{\tau} \nabla_{\theta_{I_1}} l_{I_1,t}(\theta_{I_1,0}) + \sum_{t=\tau+1}^{T} \nabla_{\theta_{I_1}} l_{I_2,t}(\theta_{I_2,0}) \\ \sum_{t=\tau+1}^{T} \nabla_{\check{\theta}} l_{I_2,t}(\theta_{I_2,0}) \end{pmatrix} \\ \equiv& \frac{1}{T}\begin{pmatrix} \sum_{t=1}^{T} U_{1,t}(\theta_0) \\ \sum_{t=\tau+1}^{T} U_{2,t}(\theta_0) \end{pmatrix}, \end{aligned} \quad \text{(B.77)}$$

where $U_{1,t}(\theta_0) = (u_{1,t,1}(\theta_0), \ldots, u_{1,t,K_1}(\theta_0))'$ is the K_1-dimensional vector such that $U_{1,t}(\theta_0) = \nabla_{\theta_{I_1}} l_{I_1,t}(\theta_{I_1,0})$ for $t = 1\ldots,\tau$ and $U_{1,t}(\theta_0) = \nabla_{\theta_{I_1}} l_{I_2,t}(\theta_{I_2,0})$ for $t = \tau+1\ldots,T$, and $U_{2,t}(\theta_0) = (u_{2,t,1}(\theta_0), \ldots, u_{2,t,K-K_1}(\theta_0))'$ is the $K - K_1$-dimensional vector such that $U_{2,t}(\theta_0) = \nabla_{\check{\theta}} l_{I_2,t}(\theta_{I_2,0})$ for $t = \tau+1\ldots,T$. By Assumption 2, we have $E[\nabla_\theta Q_T(\theta_0)] = 0_K$. Moreover, each element of $\sum_{t=1}^{T} U_{1,t}(\theta_0)$ is $O_p(T^{1/2})$, because, for all $j = 1, \ldots, K_1$, a constant C exists such that

$$\begin{aligned}
\operatorname{Var}\left(\sum_{t=1}^{T} u_{1,t,j}(\theta_0)\right) =& E\left\{\left[\sum_{t=1}^{T} u_{1,t,j}(\theta_0)\right]^2\right\} = \sum_{t=1}^{T}\sum_{s=1}^{T} E\left[u_{1,t,j}(\theta_0)u_{1,s,j}(\theta_0)\right] \\
=& \sum_{t=1}^{T} \operatorname{Var}\left[u_{1,t,j}(\theta_0)\right] + \sum_{t=1}^{T}\sum_{s\neq t}^{T} \operatorname{Cov}\left[u_{1,t,j}(\theta_0)u_{1,s,j}(\theta_0)\right] \\
\leq& T\|u_{1,t,j}(\theta_0)\|_2^2 + 8\sum_{t=1}^{T}\sum_{s\neq t}^{T} \||u_{1,t,j}(\theta_0)\|_q\||u_{1,s,j}(\theta_0)\|_q \alpha(|t-s|)^{1-2/q} \\
\leq& TC + TC\sum_{t=1}^{\infty} \alpha(t)^{1-2/q} = O(T), \qquad \text{(B.78)}
\end{aligned}$$

where the first inequality is derived by using the Davydov inequality (e.g.,Corollary A.2 in Hall and Heyde (2014)), the second inequality holds under Assumption 3 (2), $\sum_{t=1}^{T} \alpha(t)^{1-2/q}$ converges to a constant as $T \to \infty$ by Assumption 1, and $q > 3$ is an integer defined in Assumption 3 (2). Similarly, each element of $\sum_{t=\tau+1}^{T} U_{2,t}(\theta_0)$ is $O_p(T^{a/2})$, because, for all $j = 1, \ldots, K - K_1$, a constant C exists such that

$$\begin{aligned}
\operatorname{Var}\left(\sum_{t=\tau+1}^{T} u_{2,t,j}(\theta_0)\right) =& E\left\{\left[\sum_{t=\tau+1}^{T} u_{2,t,j}(\theta_0)\right]^2\right\} = \sum_{t=\tau+1}^{T}\sum_{s=\tau+1}^{T} E\left[u_{2,t,j}(\theta_0)u_{2,s,j}(\theta_0)\right] \\
=& \sum_{t=\tau+1}^{T} \operatorname{Var}\left[u_{2,t,j}(\theta_0)\right] + \sum_{t=\tau+1}^{T}\sum_{s\neq t}^{T} \operatorname{Cov}\left[u_{2,t,j}(\theta_0)u_{2,s,j}(\theta_0)\right] \\
\leq& T^a\|u_{2,t,j}(\theta_0)\|_2^2 + 8\sum_{t=\tau+1}^{T}\sum_{s\neq t}^{T} \||u_{2,t,j}(\theta_0)\|_q\||u_{2,s,j}(\theta_0)\|_q \alpha(|t-s|)^{1-2/q} \\
\leq& T^aC + T^aC\sum_{t=1}^{\infty} \alpha(t)^{1-2/q} = O(T^a). \qquad \text{(B.79)}
\end{aligned}$$

Thus, the first K_1 elements of

$$I_H \nabla_\theta Q_T(\theta_0) = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} U_{1,t}(\theta_0) \\ \frac{1}{T^a}\sum_{t=\tau+1}^{T} U_{2,t}(\theta_0) \end{pmatrix}, \qquad \text{(B.80)}$$

are $O_p(T^{-1/2})$ and the remaining $K - K_1$ elements are $O_p(T^{-a/2})$, where I_H is defined in equation (3.6).

We write the $K \times K$ matrix $I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})$ by

$$I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta}) = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}, \tag{B.81}$$

where

$$\begin{aligned}
H_{11} =& \frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_1,t}(\tilde{\theta}_{I_1}) + \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_2,t}(\tilde{\theta}_{I_2}), \\
H_{12} =& \frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_{I_1}\breve{\theta}'} l_{I_1,t}(\tilde{\theta}_{I_1}) + \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla_{\theta_{I_1}\breve{\theta}'} l_{I_2,t}(\tilde{\theta}_{I_2}), \\
H_{21} =& \frac{1}{T^a} \sum_{t=\tau+1}^{T} \nabla_{\breve{\theta}\theta'_{I_1}} l_{I_2,t}(\tilde{\theta}_{I_2}), \\
H_{22} =& \frac{1}{T^a} \sum_{t=\tau+1}^{T} \nabla_{\breve{\theta}\breve{\theta}'} l_{I_2,t}(\tilde{\theta}_{I_2}).
\end{aligned} \tag{B.82}$$

The first term of H_{11} can be decomposed as follows

$$\frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_1,t}(\tilde{\theta}_{I_1}) = \frac{1}{T} \sum_{t=1}^{\tau} E\left[\nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_1,t}(\theta_{I_1,0})\right] + H_{111} + H_{112}, \tag{B.83}$$

where

$$\begin{aligned}
H_{111} &= \frac{1}{T} \sum_{t=1}^{\tau} \left\{ \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_1,t}(\theta_{I_1,0}) - E\left[\nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_1,t}(\theta_{I_1,0})\right] \right\}, \\
H_{112} &= \frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_1,t}(\tilde{\theta}_{I_1}) - \frac{1}{T} \sum_{t=1}^{\tau} \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_1,t}(\theta_{I_1,0}).
\end{aligned} \tag{B.84}$$

Let $H_{111}^{(j,k)}$ be the j,k element of $K_1 \times K_1$ matrix H_{111}. Under Assumptions 1, 2, and 3 (2), we can apply Lemma 13 to the each element of H_{111}, which yields

$$P\left(\left|H_{111}^{(j,k)}\right| > \epsilon\right) = o(T^{-1}), \tag{B.85}$$

for any ϵ. For any $\delta > 0$, we define two events $A_1 = \{\|\hat{\theta}_{I_1} - \theta_{I_1,0}\| \leq \delta/2\}$ and $A_2 = \{\frac{1}{T}\sum_{t=1}^{\tau}[l_t - E(l_t)] \leq \delta/2\}$, where l_t is a function defined in Assumption 3 (2). By Lemmas 1 and 15, we have $P(A_1 \cap A_2) \geq 1 - P(A_1^c) - P(A_2^c) = 1 - o(T^{-1})$. Let $H_{112}^{(j,k)}$ be the j,k element of $K_1 \times K_1$ matrix H_{112}. Then, conditional on $A = A_1 \cap A_2$, for any $j,k = 1,2,\ldots,K_1$, we have

$$\begin{aligned}\left|H_{112}^{(j,k)}\right| &\leq \left|\frac{1}{T}\sum_{t=1}^{\tau}\|\tilde{\theta}_{I_1} - \theta_{I_1,0}\| l_t\right| \leq \frac{\delta}{2}\left|\frac{1}{T}\sum_{t=1}^{\tau} l_t\right| \\ &\leq \frac{\delta}{2}\left|E(l_t)\right| + \frac{\delta^2}{4} \leq \frac{\delta c_l}{2} + \frac{\delta^2}{4},\end{aligned} \tag{B.86}$$

where c_l is a constant defined in Assumption 3 (2). Thus, for any $\epsilon > 0$, we obtain

$$\begin{aligned}P\left(\left|H_{112}^{(j,k)}\right| > \epsilon\right) &= P\left(\left|H_{112}^{(j,k)}\right| > \epsilon | A\right) P(A) + P\left(\left|H_{112}^{(j,k)}\right| > \epsilon | A^c\right) P(A^c) \\ &\leq P(A^c) = o(T^{-1}).\end{aligned} \tag{B.87}$$

The second term of H_{11} can be decomposed as follows

$$\frac{1}{T}\sum_{t=\tau+1}^{T} \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_2,t}(\tilde{\theta}_{I_2}) = \frac{1}{T}\sum_{t=\tau+1}^{T} E\left[\nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_2,t}(\theta_{I_2,0})\right] + H_{113} + H_{114}, \tag{B.88}$$

where

$$H_{113} = \frac{1}{T} \sum_{t=\tau+1}^{T} \left\{ \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_2,t}(\theta_{I_2,0}) - E\left[\nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_2,t}(\theta_{I_2,0}) \right] \right\},$$

$$H_{114} = \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_2,t}(\tilde{\theta}_{I_2}) - \frac{1}{T} \sum_{t=\tau+1}^{T} \nabla_{\theta_{I_1}\theta'_{I_1}} l_{I_2,t}(\theta_{I_2,0}). \tag{B.89}$$

Let $H_{113}^{(j,k)}$ be the j,k element of $K_1 \times K_1$ matrix H_{113}. Under Assumptions 1, 2, and 3 (2), we can apply Lemma 14 to the each element of H_{113}, which yields

$$P\left(\left| H_{113}^{(j,k)} \right| > \epsilon \right) = o(T^{-1}), \tag{B.90}$$

for any $\epsilon > 0$.

For any $\delta > 0$, we define two events $A_3 = \{\|\hat{\theta}_{I_2} - \theta_{I_2,0}\| \le \delta/2\}$ and $A_4 = \{\frac{1}{T}\sum_{t=\tau+1}^{T}[l_t - E(l_t)] \le \delta/2\}$. By Lemmas 1 and 16, we have $P(A_3 \cap A_4) \ge 1 - P(A_3^c) - P(A_4^c) = 1 - o(T^{-1})$. Let $H_{114}^{(j,k)}$ be the j,k element of $K_1 \times K_1$ matrix H_{114}. Then, conditional on $\tilde{A} = A_3 \cap A_4$, for any $j,k = 1,2,\ldots,K_1$, we have

$$\begin{aligned} \left| H_{114}^{(j,k)} \right| &\le \left| \frac{1}{T} \sum_{t=\tau+1}^{T} \|\tilde{\theta}_{I_2} - \theta_{I_2,0}\| l_t \right| \le \frac{\delta}{2} \left| \frac{1}{T} \sum_{t=\tau+1}^{T} l_t \right| \\ &\le \frac{\delta}{2} \left| E(l_t) \right| + \frac{\delta^2}{4} \le \frac{\delta c_l}{2} + \frac{\delta^2}{4}, \end{aligned} \tag{B.91}$$

where c_l is a constant defined in Assumption 3 (2). Thus, for any $\epsilon > 0$, we obtain

$$P\left(\left| H_{114}^{(j,k)} \right| > \epsilon \right) = o(T^{-1}). \tag{B.92}$$

Equations (B.83), (B.85), (B.87), (B.88), (B.90), and (B.92) indicate that for any $\epsilon > 0$,

$$P\left(\left\|H_{11} - H_{11,T}^{*}\right\| > \epsilon\right) = o(T^{-1}), \tag{B.93}$$

where

$$H_{11,T}^{*} = \frac{1}{T}\sum_{t=1}^{\tau} E\left[\nabla_{\theta_{I_1}\theta_{I_1}'} l_{I_1,t}(\theta_{I_1,0})\right] + \frac{1}{T}\sum_{t=\tau+1}^{T} E\left[\nabla_{\theta_{I_1}\theta_{I_1}'} l_{I_2,t}(\theta_{I_2,0})\right]. \tag{B.94}$$

In the similar way to the derivation of (B.93), we can show that for any positive constant ϵ, we have

$$\begin{aligned} &P\left(\left\|H_{12} - H_{12,T}^{*}\right\| > \epsilon\right) = o(T^{-1}), \\ &P\left(\left\|H_{21} - H_{21,T}^{*}\right\| > \epsilon\right) = o(T^{-a}), \\ &P\left(\left\|H_{22} - H_{22,T}^{*}\right\| > \epsilon\right) = o(T^{-a}), \end{aligned} \tag{B.95}$$

where

$$\begin{aligned} H_{12,T}^{*} &= \frac{1}{T}\sum_{t=1}^{\tau} E\left[\nabla_{\theta_{I_1}\check{\theta}'} l_{I_1,t}(\theta_{I_1,0})\right] + \frac{1}{T}\sum_{t=\tau+1}^{T} E\left[\nabla_{\theta_{I_1}\check{\theta}'} l_{I_2,t}(\theta_{I_2,0})\right], \\ H_{21,T}^{*} &= \frac{1}{T^{a}}\sum_{t=\tau+1}^{T} E\left[\nabla_{\check{\theta}\theta_{I_1}'} l_{I_2,t}(\theta_{I_2,0})\right], \\ H_{22,T}^{*} &= \frac{1}{T^{a}}\sum_{t=\tau+1}^{T} E\left[\nabla_{\check{\theta}\check{\theta}'} l_{I_2,t}(\theta_{I_2,0})\right]. \end{aligned} \tag{B.96}$$

Thus, for any positive constant ϵ, we have

$$P\left(\left\|I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta}) - I_H \nabla_{\theta\theta'} Q_P(\theta_0)\right\| > \epsilon\right) = o(1), \tag{B.97}$$

where $\nabla_{\theta\theta'} Q_P(\theta_0)$ is the $K \times K$ matrix such that

$$\begin{aligned}
\nabla_{\theta\theta'} Q_P(\theta_0) =& \frac{1}{T}\sum_{t=1}^{\tau} E[\nabla_{\theta\theta'} l_{I_1,t}(\theta_{I_1,0})] + \frac{1}{T}\sum_{t=\tau+1}^{T} E[\nabla_{\theta\theta'} l_{I_2,t}(\theta_{I_2,0})] \\
=& \begin{pmatrix} H_{11,T}^* & H_{12,T}^* \\ \frac{1}{T^{1-a}} H_{21,T}^* & \frac{1}{T^{1-a}} H_{22,T}^* \end{pmatrix}. \qquad \text{(B.98)}
\end{aligned}$$

By the Weyl inequality, we have

$$\begin{aligned}
\iota_{min}(I_H \nabla_{\theta\theta'} Q_P(\theta_0)) &= \iota_{min}(I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta}) + I_H \nabla_{\theta\theta'} Q_P(\theta_0) - I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})) \\
&\leq \iota_{min}(I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})) + \iota_{max}(I_H \nabla_{\theta\theta'} Q_P(\theta_0) - I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})),
\end{aligned}$$

which implies

$$\iota_{min}(I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})) \geq \iota_{min}(I_H \nabla_{\theta\theta'} Q_P(\theta_0)) - \iota_{max}(I_H \nabla_{\theta\theta'} Q_P(\theta_0) - I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})).$$

Since $|\iota_{max}(A)| \leq \|A\|$ for any symmetric matrix A, we have

$$\begin{aligned}
\iota_{min}(I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})) &\geq \iota_{min}(I_H \nabla_{\theta\theta'} Q_P(\theta_0)) - \|I_H \nabla_{\theta\theta'} Q_P(\theta_0) - I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})\| \\
&\geq c_H - o_p(1), \qquad \text{(B.99)}
\end{aligned}$$

where the last inequality holds by Assumption 3 (3) and equation (B.97). This implies that a $K \times K$ matrix E exists such that satisfies $EE' = I$ and $I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta}) = E\Lambda E'$, where Λ is a $K \times K$ matrix whose diagonal elements are eigenvalue of $I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})$. Thus, we have $(\hat{\theta} - \theta_0)' I_H^{-1} I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})(\hat{\theta} - \theta_0) = (\hat{\theta} - \theta_0)' I_H^{-1} E\Lambda E'(\hat{\theta} - \theta_0) \geq \iota_{min}(I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta}))(\hat{\theta} - \theta_0)' I_H^{-1}(\hat{\theta} - \theta_0) \geq \iota_{min}(I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})) T^{a-1} \|\hat{\theta} - \theta_0\|^2 \geq 0$.

For now, let us denote the $m \times K$ restriction matrix by $W = (w_1, \ldots, w_m)'$, where w_i is a K-dimensional column vector. By a simple calculation, we obtain

$$\left\|W\hat{\theta}\right\|^2 = \sum_{i=1}^{m}(w_i'\hat{\theta})^2 = \sum_{i=1}^{m}(w_i'\theta_0)^2 + 2\sum_{i=1}^{m}\tilde{\theta}' w_i w_i'(\hat{\theta}-\theta_0)$$
$$= \|W\theta_0\|^2 + 2\tilde{\theta}'W'W(\hat{\theta}-\theta_0), \tag{B.100}$$

where $\tilde{\theta}$ lies between $\hat{\theta}$ and θ_0.

Since $Q_\lambda(\hat{\theta}) - Q_\lambda(\theta_0) \leq 0$ holds with probability 1 (w.p.1), it holds, along with equation (B.76), that

$$0 \geq Q_\lambda(\hat{\theta}) - Q_\lambda(\theta_0) = Q_T(\hat{\theta}) - Q_T(\theta_0) + \left[\left\|W\hat{\theta}\right\|^2 - \|W\theta_0\|^2\right]$$
$$= (\hat{\theta}-\theta_0)'\nabla_\theta Q_T(\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)'\nabla_{\theta\theta'}Q_T(\tilde{\theta})(\hat{\theta}-\theta_0) + 2\lambda\tilde{\theta}'W'W(\hat{\theta}-\theta_0)$$
$$= (\hat{\theta}-\theta_0)'I_H^{-1}I_H\nabla_\theta Q_T(\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)'I_H^{-1}I_H\nabla_{\theta\theta'}Q_T(\tilde{\theta})(\hat{\theta}-\theta_0) + 2\lambda\tilde{\theta}'W'W(\hat{\theta}-\theta_0), \tag{B.101}$$

which implies

$$-2(\hat{\theta}-\theta_0)'I_H^{-1}I_H\nabla_\theta Q_T(\theta_0) - 4\lambda\tilde{\theta}'W'W(\hat{\theta}-\theta_0)$$
$$\geq (\hat{\theta}-\theta_0)'I_H^{-1}I_H\nabla_{\theta\theta'}Q_T(\tilde{\theta})(\hat{\theta}-\theta_0) \tag{B.102}$$
$$\geq \iota_{min}(I_H\nabla_{\theta\theta'}Q_T(\tilde{\theta}))T^{a-1}\|\hat{\theta}-\theta_0\|^2$$
$$\geq 0. \tag{B.103}$$

Therefore, we obtain

$$\iota_{min}(I_H\nabla_{\theta\theta'}Q_T(\tilde{\theta}))T^{a-1}\|\hat{\theta}-\theta_0\|^2 \leq 2\left[\|I_H^{-1}\|\,\|I_H\nabla_\theta Q_T(\theta_0)\| + 2\lambda\|\tilde{\theta}'W'W\|\right]\|\hat{\theta}-\theta_0\|. \tag{B.104}$$

Since $\|I_H\nabla_\theta Q_T(\theta_0)\| = O_p(T^{-a/2})$ by equation (B.80) and $\|I_H^{-1}\| = O(1)$, we obtain

$$\|\hat{\theta}-\theta_0\| \leq 2\iota_{min}(I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta}))^{-1} T^{1-a}\left[\|I_H^{-1}\|\,\|I_H \nabla_\theta Q_T(\theta_0)\| + 2\lambda\|\tilde{\theta}' W' W\|\right]$$
$$= O_p(T^{1-\frac{3}{2}a}) + O_p(T^{1-a}\lambda). \quad \text{(B.105)}$$

□

Proof of Theorem 3

Proof. By the definition of the estimator and Assumption 3 (1), $\tilde{\theta} = (\tilde{\theta}'_{I_1}, \tilde{\tilde{\theta}}')' \in \Theta$ exists such that it lies between $\hat{\theta}$ and θ_0 element-wise and satisfies

$$\begin{aligned} 0_K = \nabla_\theta Q_\lambda(\hat{\theta}) &= \nabla_\theta Q_T(\hat{\theta}) + \nabla_\theta \lambda \left\|W\hat{\theta}\right\|^2 \\ &= \nabla_\theta Q_T(\theta_0) + \nabla_{\theta\theta'} Q_T(\tilde{\theta})(\hat{\theta}-\theta_0) + \nabla_\theta \lambda \left\|W\hat{\theta}\right\|^2 \\ &= I_H \nabla_\theta Q_T(\theta_0) + I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})(\hat{\theta}-\theta_0) + I_H \nabla_\theta \lambda \left\|W\hat{\theta}\right\|^2, \end{aligned} \quad \text{(B.106)}$$

where by equation (B.77),

$$\nabla_\theta Q_T(\theta_0) = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T U_{1,t}(\theta_0) \\ \frac{1}{T}\sum_{t=\tau+1}^T U_{2,t}(\theta_0) \end{pmatrix}, \quad \text{(B.107)}$$

and by equation (B.81),

$$I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta}) = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}. \quad \text{(B.108)}$$

Thus, we obtain

$$\hat{\theta} - \theta_0 = -[I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})]^{-1} I_H \nabla_\theta Q_T(\theta_0) - [I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})]^{-1} I_H \nabla_\theta \lambda \|W\hat{\theta}\|^2. \tag{B.109}$$

Let $\mathbb{W}_T$ be the $K \times K$ diagonal matrix whose first K_1 diagonal elements are $T^{1/2}$ and the remaining $K - K_1$ diagonal elements are $T_s^{1/2}$. Then,

$$\mathbb{W}_T(\hat{\theta} - \theta_0) = \begin{pmatrix} \sqrt{T}(\hat{\theta}_{I_1} - \theta_{I_1,0}) \\ \sqrt{T_s}(\hat{\check{\theta}} - \check{\theta}_0) \end{pmatrix} \equiv -\hat{H}\hat{S} - \hat{H}\hat{A}, \tag{B.110}$$

where

$$\begin{aligned} \hat{H} =& \mathbb{W}_T [I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta})]^{-1} \mathbb{W}_T^{-1} \\ \hat{S} =& \mathbb{W}_T I_H \nabla_\theta Q_T(\theta_0) \\ \hat{A} =& \mathbb{W}_T I_H \nabla_\theta \lambda \|W\hat{\theta}\|^2. \end{aligned} \tag{B.111}$$

Since $I_H \nabla_{\theta\theta'} Q_T(\tilde{\theta}) - I_H \nabla_{\theta\theta'} Q_P(\theta_0) = o_p(1)$ by equation (B.97) and $I_H \nabla_{\theta\theta'} Q_P(\theta_0) \to H > 0$ by Assumption 3 (4), it holds that

$$\hat{H} = H^{-1} + o_p(1) = \begin{pmatrix} H_{11}^* & H_{12}^* \\ H_{21}^* & H_{22}^* \end{pmatrix}^{-1} + o_p(1), \tag{B.112}$$

where $H_{11}^* \equiv \lim_{T\to\infty} H_{11,T}^*$, $H_{12}^* \equiv \lim_{T\to\infty} H_{12,T}^*$, $H_{21}^* \equiv \lim_{T\to\infty} H_{21,T}^*$, and $H_{22}^* \equiv \lim_{T\to\infty} H_{22,T}^*$.

Since $\hat{\theta}$ converges to θ_0 in probability and the parameter space is assumed to be compact in Assumption 2, we have $\|W\hat{\theta}\|^2 = O_p(1)$. Thus, the orders of K dimensional vector $\hat{A}$ are $\lambda\sqrt{T}$ for the first K_1 elements and $\lambda\sqrt{T_s}$ for the remaining $K - K_1$ elements, implying that $\hat{A} = o_p(1)$ when $\lambda = o(T^{-\frac{1}{2}})$.

By equation (B.77), $\hat{S}$ can be written as follows

$$\hat{S} = \mathbb{W}_T I_H \nabla_\theta Q_T(\theta_0) = \begin{pmatrix} \frac{1}{\sqrt{T}} \sum_{t=1}^T U_{1,t}(\theta_0) \\ \frac{1}{\sqrt{T_s}} \sum_{t=\tau+1}^T U_{2,t}(\theta_0) \end{pmatrix}. \tag{B.113}$$

Note that $E[\nabla_\theta Q_T(\theta_0)] = 0_K$ under Assumption 3 (2) and $\nabla_\theta Q_p(\theta_0) = 0_K$ by Assumption 2. Thus, we have

$$\begin{aligned} &\operatorname{Var}\left(\hat{S}\right) \\ =&E(\hat{S}\hat{S}') \qquad \text{(B.114)} \\ =&\begin{pmatrix} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T E\left[U_{1,t}(\theta_0) U_{1,s}(\theta_0)'\right] & \frac{1}{\sqrt{TT_s}} \sum_{t=1}^T \sum_{s=\tau+1}^T E\left[U_{1,t}(\theta_0) U_{2,s}(\theta_0)'\right] \\ \frac{1}{\sqrt{TT_s}} \sum_{t=1}^T \sum_{s=\tau+1}^T E\left[U_{2,t}(\theta_0) U_{1,s}(\theta_0)'\right] & \frac{1}{T_s} \sum_{t=\tau+1}^T \sum_{s=\tau+1}^T E\left[U_{2,t}(\theta_0) U_{2,s}(\theta_0)'\right] \end{pmatrix}. \end{aligned}$$

By Assumption 3 (5), the limit, in the sense of $T \to \infty$, of $\operatorname{Var}(\hat{S})$, denoted as Σ, exists and satisfies $\operatorname{Var}(\hat{S}) \to \Sigma > 0$.

To show the asymptotic normality of $\hat{S}$, we rewrite

$$\hat{S} = \frac{1}{\sqrt{T}} \begin{pmatrix} \sum_{t=1}^T U_{1,t}(\theta_0) \\ \sum_{t=\tau+1}^T \frac{\sqrt{T}}{\sqrt{T_s}} U_{2,t}(\theta_0) \end{pmatrix} = \frac{1}{\sqrt{T}} \sum_{t=1}^T \begin{pmatrix} U_{1,t}(\theta_0) \\ \tilde{U}_{2,t}(\theta_0) \end{pmatrix} \equiv \frac{1}{\sqrt{T}} \sum_{t=1}^T U_t(\theta_0), \tag{B.115}$$

where

$$\tilde{U}_{2,t}(\theta_0) = \begin{cases} 0 & t = 1, \ldots, \tau \\ \frac{\sqrt{T}}{\sqrt{T_s}} U_{2,t}(\theta_0) & t = \tau+1, \ldots, T. \end{cases}$$

We show the asymptotic normality of the sum of triangular stochastic arrays $Z_{T,t} \equiv \frac{1}{\sqrt{T}} \iota_K' U_t(\theta_0)$, where ι_K is arbitrary $K \times 1$ non-stochastic vector satisfying $\|\iota_K\| = 1$ (e.g., Theorem 25.6 in Davidson (1994)).

By Assumption 1, $Z_{T,t}$ is a measurable function of a zero mean strong mixing process, indicating that it is also near-epoch dependent in L_p-norm of any size on $\{r_t\}$. Let us define a positive constant array $c_{T,t} = \sqrt{\mathrm{Var}(Z_{T,t})}$, where $\mathrm{Var}(Z_{T,t}) = \mathrm{Var}(\iota_K' U_t(\theta_0))/T = \iota_{K_1}' E[U_t(\theta_0)U_t(\theta_0)']\iota_{K_1}/T$ exists for all t and T by assumptions. Then, we have

$$\sup_{t,T} E|Z_{T,t}/c_{T,t}|^q = \sup_{t,T} E|Z_{T,t}|^q/\mathrm{Var}(Z_{T,t})^{q/2} < 1, \tag{B.116}$$

by the H lder inequality, and the boundedness of

$$\sup_T \left\{ T \left(\max_{1\leq t\leq T} \{c_{T,t}\} \right)^2 \right\} = \sup_T \left\{ \left(\max_{1\leq t\leq T} \left\{ \sqrt{\iota_{K_1}' E[U_t(\theta_0)U_t(\theta_0)']\iota_{K_1}} \right\} \right)^2 \right\} \tag{B.117}$$

is implied by Assumption 3 (5) and $T_s/T \to \zeta$ for some $0 < \zeta \leq 1$. Then, by the central limit theorem for near-epoch dependent functions of strong mixing process (e.g., Corollary 24.7 in Davidson (1994)), we obtain

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} U_t(\theta_0) \xrightarrow{d} N(0, \Sigma). \tag{B.118}$$

Equations (B.110), (B.112), and (B.118) indicate that

$$\Sigma^{-1/2} H \mathbb{W}_T(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I_K). \tag{B.119}$$

□

Proof of Lemma 4

Proof. Since $\hat{\theta} = (\hat{\theta}_{I_1}', \hat{\hat{\theta}}')'$, it suffices to show that $\hat{\theta} \in \Theta_{\delta_\lambda}$. By the definition of Θ_{δ_λ}, it is proved when $\|W'W(\hat{\theta} - \theta_0)\| \leq \delta_\lambda$ holds with probability $1 - \epsilon_S$. By the definition of $\hat{\theta}$, we have

$$0 = \nabla_\theta Q_\lambda(\hat{\theta}) = \nabla_\theta Q_T(\hat{\theta}) + 2\lambda W'W(\theta_0 + \hat{\theta} - \theta_0), \tag{B.120}$$

which implies that

$$2\lambda W'W(\hat{\theta} - \theta_0) = -\nabla_\theta Q_T(\hat{\theta}) - 2\lambda W'W\theta_0. \tag{B.121}$$

Lemma 18 indicates that a positive constant S exists such that $P(\|\nabla_\theta Q_T(\hat{\theta})\| > S) \leq \epsilon_S$ for any $T \geq 2$ and arbitrary small $\epsilon_S > 0$. Thus, we obtain

$$\begin{aligned} 1 &= P\left(\|W'W(\hat{\theta} - \theta_0)\| \leq \frac{1}{2\lambda}\left[\|\nabla_\theta Q_T(\hat{\theta})\| + 2\lambda\|W'W\theta_0\|\right]\right) \\ &\leq P\left(\|W'W(\hat{\theta} - \theta_0)\| \leq \delta_\lambda\right) + \epsilon_S. \end{aligned} \tag{B.122}$$

□

Proof of Theorem 6

Proof. For simplicity, let ϵ denote ϵ_S in this proof. Let $Q_p(\theta_1, \theta_2) \equiv Q_p(\theta_1) - Q_p(\theta_2)$ and $Q_T(\theta_1, \theta_2) \equiv Q_T(\theta_1) - Q_T(\theta_2)$. By the definition of $\hat{\theta}$, it holds that $\lambda\|W'\hat{\theta}\|^2 + Q_T(\hat{\theta}, \theta_0) \leq \lambda\|W\theta_0\|^2$. By using this, we obtain

$$\begin{aligned} \lambda\|W\hat{\theta}\|^2 + Q_p(\hat{\theta}) - Q_p(\theta_0) &= \lambda\|W\hat{\theta}\|^2 + Q_T(\hat{\theta}, \theta_0) + Q_p(\hat{\theta}, \theta_0) - Q_T(\hat{\theta}, \theta_0) \\ &\leq \lambda\|W\theta_0\|^2 + Q_p(\hat{\theta}, \theta_0) - Q_T(\hat{\theta}, \theta_0) \\ &= \lambda\|W\theta_0\|^2 + V(\hat{\theta}, \theta_0), \end{aligned} \tag{B.123}$$

where $V(\hat{\theta}, \theta_0) \equiv Q_p(\hat{\theta}, \theta_0) - Q_T(\hat{\theta}, \theta_0)$. Note that $\|W\theta_0\|$ represents the correctness of the weight W, which can be zero with an appropriate weight.

We decompose $V(\hat{\theta}, \theta_0)$ as follows:

$$
\begin{aligned}
V(\hat{\theta}, \theta_0) &= Q_p(\hat{\theta}) - Q_p(\theta_0) - Q_T(\hat{\theta}) + Q_T(\theta_0) \\
&= \frac{1}{T}\sum_{t=1}^{\tau} E[l_{I_1,t}(\hat{\theta}_{I_1})] + \frac{1}{T}\sum_{t=\tau+1}^{T} E[l_{I_2,t}(\hat{\theta}_{I_2})] - \frac{1}{T}\sum_{t=1}^{\tau} E[l_{I_1,t}(\theta_{I_1,0})] - \frac{1}{T}\sum_{t=\tau+1}^{T} E[l_{I_2,t}(\theta_{I_2,0})] \\
&\quad - \frac{1}{T}\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1}) - \frac{1}{T}\sum_{t=\tau+1}^{T} l_{I_2,t}(\hat{\theta}_{I_2}) + \frac{1}{T}\sum_{t=1}^{\tau} l_{I_1,t}(\theta_{I_1,0}) + \frac{1}{T}\sum_{t=\tau+1}^{T} l_{I_2,t}(\theta_{I_2,0}) \\
&= \frac{1}{T}\sum_{t=1}^{\tau} \Big\{ E[l_{I_1,t}(\hat{\theta}_{I_1}) - l_{I_1,t}(\theta_{I_1,0})] - [l_{I_1,t}(\hat{\theta}_{I_1}) - l_{I_1,t}(\theta_{I_1,0})] \Big\} \\
&\quad + \frac{1}{T}\sum_{t=\tau+1}^{T} \Big\{ E[l_{I_2,t}(\hat{\theta}_{I_2}) - l_{I_2,t}(\theta_{I_2,0})] - [l_{I_2,t}(\hat{\theta}_{I_2}) - l_{I_2,t}(\theta_{I_2,0})] \Big\} \\
&\equiv \frac{1}{T}\sum_{t=1}^{\tau} \tilde{\xi}_t(\hat{\theta}_{I_1}) + \frac{1}{T}\sum_{t=\tau+1}^{T} \xi_t(\hat{\theta}_{I_2}). \qquad \text{(B.124)}
\end{aligned}
$$

For any $\theta_{I_1} \in \tilde{\Theta}_{\delta_\lambda}$, we obtain,

$$
\begin{aligned}
E[\tilde{\xi}_t(\theta_{I_1})^2] &= E\left\{[l_{I_1,t}(\theta_{I_1}) - l_{I_1,t}(\theta_{I_1,0})]^2\right\} - E[l_{I_1,t}(\theta_{I_1}) - l_{I_1,t}(\theta_{I_1,0})]^2 \\
&\le E\left\{[l_{I_1,t}(\theta_{I_1}) - l_{I_1,t}(\theta_{I_1,0})]^2\right\} \\
&\le E\left[\|\theta_{I_1} - \theta_{I_1,0}\|^2 \sup_{\theta_{I_1} \in \tilde{\Theta}_{\delta_\lambda}} \left\| \nabla_{\theta_{I_1}} l_{I_1,t}(\theta_{I_1}) \right\|^2 \right] \\
&\le \kappa_1^2 K_1 c_l, \qquad \text{(B.125)}
\end{aligned}
$$

where $\kappa_1 = \sup_{\theta_{I_1} \in \tilde{\Theta}_{\delta_\lambda}} \|\theta - \theta_0\|$ exist by the compactness of Θ_{δ_λ} in Assumption 2 and c_l is a constant defined in Assumption 3 (2). By Assumption 3 (2) and the boundedness of the conditional densities, we obtain $\|\tilde{\xi}_t(\cdot)\|_\infty \le 2(l + c_l)$. Then, applying Theorem 4.3 of Modha and Masry (1996) (see also Theorem 5.1 of Hang and Steinwart, 2014) yields

$$
P\bigg(\frac{1}{T}\sum_{t=1}^{\tau} \tilde{\xi}_t(\theta_{I_1}) \ge \frac{\tau^{3/4}}{T} \kappa_1 \sqrt{c\tilde{\rho} K_1 c_l} + \frac{\tau^{1/2}}{T} \frac{c\tilde{\rho}}{3} 2(l + c_l) \bigg) \le (1 + 4e^{-2} c_\alpha) e^{-c},
$$

for all $\theta_{I_1} \in \tilde{\Theta}_{\delta_\lambda}$, $\tau \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$, and $c > 0$, where $\tilde{\rho} \equiv (-8^3/\log\rho)^{1/2}$, and ρ and c_α are the upper bound of the mixing coefficient given in Assumption 1. Since $\hat{\theta}_{I_1} \in \tilde{\Theta}_{\delta_\lambda}$ with probability $1-\epsilon$ for arbitrary small ϵ by Lemma 4, we obtain

$$\begin{aligned}
&P\left(\frac{1}{T}\sum_{t=1}^{\tau}\tilde{\xi}_t(\hat{\theta}_{I_1}) \geq \frac{\tau^{3/4}}{T}\kappa_1\sqrt{c\tilde{\rho}K_1c_l} + \frac{\tau^{1/2}}{T}\frac{c\tilde{\rho}}{3}2(l+c_l)\right) \\
&\quad\leq P\left(\max_{\theta_{I_1}\in\tilde{\Theta}_{\delta_\lambda}}\frac{1}{T}\sum_{t=1}^{\tau}\tilde{\xi}_t(\theta_{I_1}) \geq \frac{\tau^{3/4}}{T}\kappa_1 c\tilde{\rho}K_1c_l + \frac{\tau^{1/2}}{T}\frac{c\tilde{\rho}}{3}2(l+c_l)\right) + \epsilon \\
&\quad\leq (1+4e^{-2}c_\alpha)e^{-c} + \epsilon, \qquad \text{(B.126)}
\end{aligned}$$

for all $\tau \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$ and $c > 0$.

Similar calculation yields, for any $\theta_{I_2} \subset \theta \in \Theta_{\delta_\lambda}$,

$$E[\xi_t(\theta_{I_2})^2] \leq \kappa_2^2 E\left(\sup_{\theta_{I_2}\subset\theta\in\Theta_{\delta_\lambda}}\left\|\nabla_{\theta_{I_2}} l_{I_2,t}(\theta_{I_2})\right\|^2\right) \leq \kappa_2^2 K_2 c_l, \qquad \text{(B.127)}$$

where $\kappa_2 = \sup_{\theta_{I_2}\in\{\theta_{I_2}:\Theta_{\delta_\lambda}\}}\|\theta-\theta_0\|$ and $\|\xi_t(\cdot)\|_\infty \leq 2(l+c_l)$. Since $\hat{\theta}_{I_2} \subset \hat{\theta} \in \Theta_{\delta_\lambda}$ with probability $1-\epsilon$ for arbitrary small ϵ by Lemma 4, applying Theorem 4.3 of Modha and Masry (1996) yields

$$P\left(\frac{1}{T}\sum_{t=\tau+1}^{T}\xi_t(\hat{\theta}_{I_2}) \geq \frac{T^{\frac{3a}{4}}}{T}\kappa_2\sqrt{c\tilde{\rho}K_2c_l} + \frac{T^{\frac{a}{2}}}{T}\frac{c\tilde{\rho}}{3}2(l+c_l)\right) \leq (1+4e^{-2}c_\alpha)e^{-c} + \epsilon, \qquad \text{(B.128)}$$

for all $T^a \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$ and $c > 0$.

From equations (B.124), (B.126), and (B.128), we obtain

$$\begin{aligned}
&P\left(V(\hat{\theta},\theta_0) \leq \left(\frac{\tau^{3/4}}{T} + \frac{T^{\frac{3a}{4}}}{T}\right)\kappa\sqrt{c\tilde{\rho}\bar{K}c_l} + \left(\frac{\tau^{1/2}}{T} + \frac{T^{\frac{a}{2}}}{T}\right)\frac{c\tilde{\rho}}{3}2(l+c_l)\right) \\
&\qquad > 1 - 4(1+4e^{-2}c_\alpha)e^{-c} - 4\epsilon \qquad \text{(B.129)}
\end{aligned}$$

for all $c > 0$ and $\tau, T^a \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$, where $\bar{K} \equiv \max\{K_1, K_2\}$ and $\kappa \equiv \sup_{\theta\in\Theta_{\delta_\lambda}} \|\theta - \theta_0\|$.

From equations (B.123) and (B.129), we obtain that for a fixed $\lambda > 0$, all $c > 0$, and $\tau, T^a \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$, the probability that

$$\begin{aligned}\lambda\|W\hat{\theta}\|^2 + Q_p(\hat{\theta}) - Q_p(\theta_0) \leq \lambda\|W\theta_0\|^2 &+ \left(\frac{\tau^{3/4}}{T} + \frac{T^{3a/4}}{T}\right)\kappa\sqrt{c\tilde{\rho}\bar{K}c_l} \\ &+ \left(\frac{\tau^{1/2}}{T} + \frac{T^{\frac{a}{2}}}{T}\right)\frac{c\tilde{\rho}}{3}2(l + c_l) \qquad \text{(B.130)}\end{aligned}$$

is not less than $1 - 4(1 + 4e^{-2}c_\alpha)e^{-c} - 4\epsilon$. $\square$

Risk Bound for Non-penalized Estimator (Theorem B.1)

Theorem B.1. *Let $\hat{\theta}$ be the ML estimator of θ, that is,*

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}} Q_T(\theta), \qquad \text{(B.131)}$$

where

$$Q_T(\theta) = \frac{1}{T}\sum_{t=1}^{\tau} l_{I_1,t}(\theta_{I_1}) + \frac{1}{T}\sum_{t=\tau+1}^{T} l_{I_2,t}(\theta_{I_2}). \qquad \text{(B.132)}$$

Suppose assumptions of Theorem 6. We assume that the ML estimator $\hat{\theta}$ is well-defined, measurable, and consistent. For any $c > 0$ and all $\tau, T^a \geq$ $\max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$, the probability that

$$Q_p(\hat{\theta}) - Q_p(\theta_0) \leq \left(\frac{\tau^{3/4}}{T} + \frac{T^{\frac{3a}{4}}}{T}\right)\bar{\kappa}\sqrt{c\tilde{\rho}\bar{K}c_l} + \left(\frac{\tau^{1/2}}{T} + \frac{T^{\frac{a}{2}}}{T}\right)\frac{2c\tilde{\rho}(l + c_l)}{3} \qquad \text{(B.133)}$$

is not less than $1 - 4(1 + 4e^{-2}c_\alpha)e^{-c}$, where $\bar{\kappa} \equiv \sup_{\theta\in\Theta} \|\theta - \theta_0\|$.

Proof. By the definition of $\hat{\theta}$, we have $Q_T(\hat{\theta}, \theta_0) \leq 0$. By using this, we obtain

$$\begin{aligned} Q_p(\hat{\theta}) - Q_p(\theta_0) &= Q_p(\hat{\theta}, \theta_0) - Q_T(\hat{\theta}, \theta_0) + Q_T(\hat{\theta}, \theta_0) \\ &\leq Q_p(\hat{\theta}, \theta_0) - Q_T(\hat{\theta}, \theta_0) \\ &= V(\hat{\theta}, \theta_0), \end{aligned} \tag{B.134}$$

where, by equation (B.124),

$$V(\hat{\theta}, \theta_0) = \frac{1}{T} \sum_{t=1}^{\tau} \tilde{\xi}_t(\hat{\theta}_{I_1}) + \frac{1}{T} \sum_{t=\tau+1}^{T} \xi_t(\hat{\theta}_{I_2}). \tag{B.135}$$

For any $\theta_{I_1} \in \Theta_{I_1}$,

$$\begin{aligned} E[\tilde{\xi}_t(\theta_{I_1})^2] &= E\left\{[l_{I_1,t}(\theta_{I_1}) - l_{I_1,t}(\theta_{I_1,0})]^2\right\} - E[l_{I_1,t}(\theta_{I_1}) - l_{I_1,t}(\theta_{I_1,0})]^2 \\ &\leq E\left\{[l_{I_1,t}(\theta_{I_1}) - l_{I_1,t}(\theta_{I_1,0})]^2\right\} \\ &\leq E\left[\|\theta_{I_1} - \theta_{I_1,0}\|^2 \sup_{\theta_{I_1} \in \Theta_{I_1}} \left\|\nabla_{\theta_{I_1}} l_{I_1,t}(\theta_{I_1})\right\|^2\right] \\ &\leq \bar{\kappa}_1^2 K_1 c_l, \end{aligned} \tag{B.136}$$

where $\bar{\kappa}_1 \equiv \sup_{\theta_{I_1} \in \Theta_{I_1}} \|\theta_{I_1} - \theta_{I_1,0}\|$ exists by the compactness of Θ_{I_1} in Assumption 2 and c_l is a constant defined in Assumption 3 (2). By Assumption 3 (2) and the boundedness of the conditional densities, we obtain $\|\tilde{\xi}_t(\cdot)\|_\infty \leq 2(l + c_l)$. Applying Theorem 4.3 of Modha and Masry (1996) yields

$$\begin{aligned} &P\left(\frac{1}{T} \sum_{t=1}^{\tau} \tilde{\xi}_t(\hat{\theta}_{I_1}) \geq \frac{\tau^{3/4}}{T} \bar{\kappa}_1 \sqrt{c\tilde{\rho} K_1 c_l} + \frac{\tau^{1/2}}{T} \frac{c\tilde{\rho}}{3} 2(l + c_l)\right) \\ &\quad P\left(\max_{\theta_{I_1} \in \Theta_{I_1}} \frac{1}{T} \sum_{t=1}^{\tau} \tilde{\xi}_t(\theta_{I_1}) \geq \frac{\tau^{3/4}}{T} \bar{\kappa}_1 \sqrt{c\tilde{\rho} K_1 c_l} + \frac{\tau^{1/2}}{T} \frac{c\tilde{\rho}}{3} 2(l + c_l)\right) \\ &\quad \leq (1 + 4e^{-2} c_\alpha) e^{-c} \end{aligned} \tag{B.137}$$

for all $\tau \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$ and $c > 0$, where $\tilde{\rho} \equiv (-8^3/\log\rho)^{1/2}$.

Applying similar arguments for $\xi_t(\theta_{I_2})$ yields, for any $\theta_{I_2} \in \Theta_{I_2}$,

$$E[\xi_t(\theta_{I_2})^2] \leq \bar{\kappa}_2^2 E\left[\sup_{\theta_{I_2}\in\Theta_{I_2}} \left\|\nabla_{\theta_{I_2}} l_{I_2,t}(\theta_{I_2})\right\|^2\right] \leq \bar{\kappa}_2^2 K_2 c_l, \tag{B.138}$$

where $\bar{\kappa}_2 \equiv \sup_{\theta_{I_2}\in\Theta_{I_2}} \|\theta_{I_2} - \theta_{I_2,0}\|$ and $\|\xi_t(\cdot)\|_\infty \leq 2(l + c_l)$. Applying Theorem 4.3 of Modha and Masry (1996) yields

$$P\left(\frac{1}{T}\sum_{t=\tau+1}^{T} \xi_t(\hat{\theta}_{I_2}) \geq \frac{T^{\frac{3a}{4}}}{T}\bar{\kappa}_2\sqrt{c\tilde{\rho}K_2 c_l} + \frac{T^{\frac{a}{2}}}{T}\frac{c\tilde{\rho}}{3}2(l + c_l)\right) \leq (1 + 4e^{-2}c_\alpha)e^{-c} \tag{B.139}$$

for all $T^a \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$ and $c > 0$.

Thus, we obtain

$$\begin{aligned} P\Bigg(Q_p(\hat{\theta}) - Q_p(\theta_0) \leq \left(\frac{\tau^{3/4}}{T} + \frac{T^{\frac{3a}{4}}}{T}\right)\bar{\kappa}\sqrt{c\tilde{\rho}\bar{K}c_l} + \left(\frac{\tau^{1/2}}{T} + \frac{T^{\frac{a}{2}}}{T}\right)\frac{c\tilde{\rho}}{3}2(l + c_l)\Bigg) \\ > 1 - 4(1 + 4e^{-2}c_\alpha)e^{-c} \end{aligned} \tag{B.140}$$

for all $c > 0$ and $\tau, T^a \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$, where $\bar{\kappa} \equiv \sup_{\theta\in\Theta}\|\theta - \theta_0\|$. □

Alternative Risk Bound for TMLE (Theorem B.2)

Theorem B.2. *Suppose assumptions in Theorem 6 holds. For a fixed $\lambda > 0$, any $c > 0$ and $\epsilon > 0$, and all $\tau, T^a \geq \max\{-\log\rho/8, 2^7(-\log\rho)^{-1}\}$, the probability that*

$$\begin{aligned} \lambda\|W\hat{\theta}\|^2 + Q_p(\hat{\theta}) - Q_p(\theta_0) \leq \lambda\|W\theta_0\|^2 + \left(\frac{\tau^{3/4}}{T} + \frac{T^{3a/4}}{T}\right)(\delta_\lambda + \omega\kappa)\sqrt{c\tilde{\rho}\bar{K}c_l} \\ + \left(\frac{\tau^{1/2}}{T} + \frac{T^{\frac{a}{2}}}{T}\right)\frac{2c\tilde{\rho}(l + c_l)}{3} \end{aligned} \tag{B.141}$$

is not less than $1 - 2(1 + 4e^{-2}c_\alpha)e^{-c} - 2\epsilon$*, where* $\omega \equiv \|I_k - W'W\|$*.*

Proof. With respect to equation (B.124), it holds for any $\theta_{I_1} \in \tilde{\Theta}_{\delta_\lambda}$ that

$$\begin{aligned}
E[\tilde{\xi}_t(\theta_{I_1})^2] &= E\left\{[l_{I_1,t}(\theta_{I_1}) - l_{I_1,t}(\theta_{I_1,0})]^2\right\} - E[l_{I_1,t}(\theta_{I_1}) - l_{I_1,t}(\theta_{I_1,0})]^2 \\
&\leq E\left\{[l_{I_1,t}(\theta_{I_1}) - l_{I_1,t}(\theta_{I_1,0})]^2\right\} \\
&\leq E\left[\|\theta_{I_1} - \theta_{I_1,0}\|^2 \sup_{\theta_{I_1}\in\tilde{\Theta}_{\delta_\lambda}} \left\|\nabla_{\theta_{I_1}} l_{I_1,t}(\theta_{I_1})\right\|^2\right] \\
&\leq E\left[\|W'W(\theta - \theta_0) + (I_k - W'W)(\theta - \theta_0)\|^2 \sup_{\theta_{I_1}\in\tilde{\Theta}_{\delta_\lambda}} \left\|\nabla_{\theta_{I_1}} l_{I_1,t}(\theta_{I_1})\right\|^2\right] \\
&\leq (\delta_\lambda + \omega\kappa)^2 E\left(\sup_{\theta_{I_1}\in\tilde{\Theta}_{\delta_\lambda}} \left\|\nabla_{\theta_{I_1}} l_{I_1,t}(\theta_{I_1})\right\|^2\right) \\
&\leq (\delta_\lambda + \omega\kappa)^2 K_1 c_l. \qquad \text{(B.142)}
\end{aligned}$$

Similarly, it holds for any $\theta_{I_2} \subset \theta \in \Theta_{\delta_\lambda}$ that

$$\begin{aligned}
E[\xi_t(\theta_{I_2})^2] &\leq (\delta_\lambda + \omega\kappa)^2 E\left(\sup_{\theta_{I_2}\subset\theta\in\Theta_{\delta_\lambda}} \left\|\nabla_{\theta_{I_2}} l_{I_2,t}(\theta_{I_2})\right\|^2\right) \\
&\leq (\delta_\lambda + \omega\kappa)^2 K_2 c_l. \qquad \text{(B.143)}
\end{aligned}$$

Then, by replacing equations (B.125) and (B.127) in the proof of Theorem 6 with (B.142) and (B.143), respectively, we obtain the assertion of this theorem. □

Proof of Theorem 9

Proof. Note that

$$
\begin{aligned}
0 \leq & Q_{P_o}(\bar{\lambda}_T) - Q_{P_o}(\theta_0) \\
= & Q_{P_o}(\bar{\lambda}_T) - Q_{P_T^b}(\bar{\lambda}_T) + Q_{P_T^b}(\bar{\lambda}_T) - Q_{P_T^B}(\bar{\lambda}_T) \\
& + Q_{P_T^B}(\bar{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T) + Q_{P_T^b}(\hat{\lambda}_T) - Q_{P_o}(\check{\lambda}_T) + Q_{P_o}(\check{\lambda}_T) - Q_{P_o}(\tilde{\lambda}_T) \\
& + Q_{P_o}(\tilde{\lambda}_T) - Q_{P_o}(\theta_0) \\
\leq & \left\{ Q_{P_o}(\bar{\lambda}_T) - Q_{P_T^b}(\bar{\lambda}_T) \right\} + \left\{ Q_{P_T^b}(\bar{\lambda}_T) - Q_{P_T^B}(\bar{\lambda}_T) \right\} \\
& + \left\{ Q_{P_T^B}(\hat{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T) \right\} + \left\{ Q_{P_T^b}(\check{\lambda}_T) - Q_{P_o}(\check{\lambda}_T) \right\} + \left\{ Q_{P_o}(\check{\lambda}_T) - Q_{P_o}(\tilde{\lambda}_T) \right\} \\
& + Q_{P_o}(\tilde{\lambda}_T) - Q_{P_o}(\theta_0) \,. \qquad \text{(B.144)}
\end{aligned}
$$

Since

$$
Q_T(\hat{\theta}_{\lambda_T}, X_T) = \frac{1}{T} \sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) + \frac{1}{T} \sum_{t=\tau+1}^{T} l_{I_2,t}(\hat{\theta}_{I_2,\lambda_T}),
$$

then

$$
\begin{aligned}
Q_{P_o}(\lambda_T) = & \int Q_T(\hat{\theta}_{\lambda_T}, x_T) dP_o \\
= & \int \frac{1}{T} \sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) + \frac{1}{T} \sum_{t=\tau+1}^{T} l_{I_2,t}(\hat{\theta}_{I_2,\lambda_T}) dP_o \\
= & \frac{\tau}{T} \int \frac{1}{\tau} \sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) dP_o + \frac{T-\tau}{T} \int \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} l_{I_2,t}(\hat{\theta}_{I_2,\lambda_T}) dP_o \\
= & \frac{\tau}{T} \int \frac{1}{\tau} \sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) - \frac{1}{\tau} \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) \right] dP_o \\
& + \frac{T-\tau}{T} \int \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} l_{I_2,t}(\hat{\theta}_{I_2,\lambda_T}) - \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_2,t}(\hat{\theta}_{I_2,\lambda_T}) \right] dP_o \\
& + \int \frac{1}{T} \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) \right] + \frac{1}{T} \sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_2,t}(\hat{\theta}_{I_2,\lambda_T}) \right] dP_o
\end{aligned}
$$

and

$$
\begin{aligned}
-Q_{P_o}(\lambda_T) \leq & \frac{\tau}{T}\int\left|\frac{1}{\tau}\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) - \frac{1}{\tau}\sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T})\right]\right| dP_o \\
& + \frac{T-\tau}{T}\int\left|\frac{1}{T-\tau}\sum_{t=\tau+1}^{T} l_{I_2,t}(\hat{\theta}_{I_2,\lambda_T}) - \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_2,t}(\hat{\theta}_{I_2,\lambda_T})\right]\right| dP_o \\
& - \int \frac{1}{T}\sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T})\right] + \frac{1}{T}\sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_2,t}(\hat{\theta}_{I_2,\lambda_T})\right] dP_o.
\end{aligned}
$$

We first deal with $Q_{P_o}(\bar{\lambda}_T)$. For $\frac{1}{\tau}\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) - \frac{1}{\tau}\sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T})\right]$, we can obtain that for all $\epsilon > 0$,

$$
\begin{aligned}
& P_o\left(\frac{1}{\tau}\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) - \frac{1}{\tau}\sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T})\right] > \epsilon\right) \\
& \leq P_o\left(\left|\frac{1}{\tau}\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) - \frac{1}{\tau}\sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T})\right]\right| > \epsilon\right) \\
& \leq K(T)\max_{\lambda_T\in\Lambda_T} P_o\left(\left|\frac{1}{\tau}\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) - \frac{1}{\tau}\sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T})\right]\right| > \epsilon\right).
\end{aligned}
\tag{B.145}
$$

Notice that $\hat{\theta}_{\lambda_T}$ is given (fixed), hence $\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) - \sum_{t=1}^{\tau} E_{P_o|\mathscr{F}_{t-1}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T})\right]$ is a martingale under the distribution of X_t conditional on $\hat{\theta}_{\lambda_T}$.

Since $\sup_{\hat{\theta}_{I_1,\lambda_T}\in\Theta_{I_1}, X_T\in\mathscr{X}_T}\left|l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T})\right| \leq M$ and $\mathscr{H}_{L_{I_1}(\lambda_T),M}(\delta,\Omega_1)$ exists by the assumption in Theorem 9, for

$$
\begin{aligned}
& \epsilon \leq C_1 C_3^2/M, \\
& \epsilon \leq 8C_3, \\
& \epsilon \geq \frac{1}{\sqrt{\tau}} C_0\left(\int_{\epsilon/2^6}^{C_3} \mathscr{H}^{1/2}_{L_{I_1}(\lambda_T),M}(u,\Omega_1)du \vee C_3\right), \\
& C_0^2 \geq C^2(C_1+1),
\end{aligned}
$$

where C, C_0, and C_1 are some positive constants and $C_3 = \sqrt{2M^2(e-2)}$, it follows that

$$
\begin{aligned}
&P_o\left(\left|\frac{1}{\tau}\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) - \frac{1}{\tau}\sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T})\right]\right| > \epsilon\right) \\
&= P_o\left(\sqrt{\tau}\left|\frac{1}{\tau}\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T}) - \frac{1}{\tau}\sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\lambda_T})\right]\right| > \sqrt{\tau}\epsilon\right) \\
&\leq C\exp\left[-\frac{(\sqrt{\tau}\epsilon)^2}{C^2(C_1+1)C_3^2}\right]
\end{aligned}
$$

by the uniform inequality for martingales (see Theorem 8.13 of Geer et al. (2000)).

Therefore, for (B.145), we have

$$
\begin{aligned}
&P_o\left(\frac{1}{\tau}\sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) - \frac{1}{\tau}\sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T})\right] > \epsilon\right) \\
&\leq K(T)C\exp\left[-\frac{(\sqrt{\tau}\epsilon)^2}{C^2(C_1+1)C_3^2}\right].
\end{aligned}
$$

Note that, for any random variable X,

$$
E(X) \leq E[I(X>0)X] = \int_0^\infty P(X>x)dx
$$

because, according to $P(X>x) = \int_x^\infty f(z)dz$ in which $f(\cdot)$ refers to the density function of X, it follows that

$$
\begin{aligned}
\int_0^\infty P(X>x)dx &= \int_0^\infty \int_x^\infty f(z)dzdx = \iint_{0<x<z<\infty} f(z)d(x,z) \\
&= \int_0^\infty \int_0^z f(z)dxdz = \int_0^\infty f(z)\int_0^z dxdz \\
&= \int_0^\infty zf(z)dz = E[I(X>0)X],
\end{aligned}
$$

where the third equation holds by Fubini's theorem.

Let $a = \frac{1}{\sqrt{\tau}} C_0 \left(\int_{\epsilon/2^6}^{C_3} \mathscr{H}^{1/2}_{L_{I_1}(\lambda_T),M}(u, \Omega_1) du \vee C_3 \right)$. Then we have

$$\begin{aligned}
&E_{P_o} \left[\frac{1}{\tau} \sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) - \frac{1}{\tau} \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) \right] \right] \\
\leq &\int_0^\infty P_o \left(\frac{1}{\tau} \sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) - \frac{1}{\tau} \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) \right] > x \right) dx \\
= &\int_0^a P_o \left(\frac{1}{\tau} \sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) - \frac{1}{\tau} \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) \right] > x \right) dx \\
&+ \int_a^\infty P_o \left(\frac{1}{\tau} \sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) - \frac{1}{\tau} \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) \right] > x \right) dx \\
\leq &a + \int_a^\infty K(T) \cdot C \exp \left[- \frac{(\sqrt{\tau} x)^2}{C^2 (C_1 + 1) C_3^2} \right] dx \\
= &a + \frac{K(T) \cdot C \sqrt{\pi C^2 (C_1 + 1) C_3^2} \mathrm{Erfc} \left[\sqrt{\tau / C^2 (C_1 + 1) C_3^2} a \right]}{2 \sqrt{\tau}} \\
\leq &a + \frac{2K(T) \cdot C \sqrt{\pi C^2 (C_1 + 1) C_3^2}}{2 \sqrt{\tau}} \\
= &O\left(\frac{1}{\sqrt{\tau}} \right) + O\left(\frac{K(T)}{\sqrt{\tau}} \right) \\
= &O\left(\frac{K(T)}{\sqrt{\tau}} \right),
\end{aligned}$$

which means that

$$E_{P_o} \left[\frac{1}{\tau} \sum_{t=1}^{\tau} l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) - \frac{1}{\tau} \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T}) \right] \right] \leq O\left(\frac{K(T)}{\sqrt{\tau}} \right). \tag{B.146}$$

Similarly for $\frac{1}{T-\tau} \sum_{t=\tau+1}^{T} l_{I_2,t}(\hat{\theta}_{I_2,\bar{\lambda}_T}) - \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_2,t}(\hat{\theta}_{I_2,\bar{\lambda}_T}) \right]$, we can show that

$$E_{P_o} \left[\frac{1}{T-\tau} \sum_{t=\tau+1}^{T} l_{I_2,t}(\hat{\theta}_{I_2,\bar{\lambda}_T}) - \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}} \left[l_{I_2,t}(\hat{\theta}_{I_2,\bar{\lambda}_T}) \right] \right] \leq O\left(\frac{K(T)}{\sqrt{T-\tau}} \right). \tag{B.147}$$

Combining (B.146), (B.147), and $Q_{P_o}(\bar{\lambda}_T)$, we have

$$Q_{P_o}(\bar{\lambda}_T) \leq O\left(\frac{K(T)\sqrt{\tau}}{T}\right) + O\left(\frac{K(T)\sqrt{T-\tau}}{T}\right) + \frac{1}{T}\int \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T})\right] + \sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_2,t}(\hat{\theta}_{I_2,\bar{\lambda}_T})\right] dP_o. \tag{B.148}$$

Second, we deal with $-Q_{P_T^b}(\bar{\lambda}_T)$, $Q_{P_T^b}(\check{\lambda}_T)$, and $-Q_{P_o}(\check{\lambda}_T)$ using the similar derivation of (B.148), then the following inequalities hold.

$$-Q_{P_T^b}(\bar{\lambda}_T) \leq O\left(\frac{K(T)\sqrt{\tau}}{T}\right) + O\left(\frac{K(T)\sqrt{T-\tau}}{T}\right) - \frac{1}{T}\int \sum_{t=1}^{\tau} E_{P^b_{T|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\bar{\lambda}_T})\right] + \sum_{t=\tau+1}^{T} E_{P^b_{T|\mathscr{F}_{t-1}}}\left[l_{I_2,t}(\hat{\theta}_{I_2,\bar{\lambda}_T})\right] dP_T^b, \tag{B.149}$$

$$Q_{P_T^b}(\check{\lambda}_T) \leq O\left(\frac{K(T)\sqrt{\tau}}{T}\right) + O\left(\frac{K(T)\sqrt{T-\tau}}{T}\right) + \frac{1}{T}\int \sum_{t=1}^{\tau} E_{P^b_{T|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\check{\lambda}_T})\right] + \sum_{t=\tau+1}^{T} E_{P^b_{T|\mathscr{F}_{t-1}}}\left[l_{I_2,t}(\hat{\theta}_{I_2,\check{\lambda}_T})\right] dP_T^b, \tag{B.150}$$

and

$$-Q_{P_o}(\check{\lambda}_T) \leq O\left(\frac{K(T)\sqrt{\tau}}{T}\right) + O\left(\frac{K(T)\sqrt{T-\tau}}{T}\right) - \frac{1}{T}\int \sum_{t=1}^{\tau} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_1,t}(\hat{\theta}_{I_1,\check{\lambda}_T})\right] + \sum_{t=\tau+1}^{T} E_{P_{o|\mathscr{F}_{t-1}}}\left[l_{I_2,t}(\hat{\theta}_{I_2,\check{\lambda}_T})\right] dP_o. \tag{B.151}$$

Third, let us consider $Q_{P_T^b}(\bar{\lambda}_T) - Q_{P_T^B}(\bar{\lambda}_T)$ and $Q_{P_T^B}(\hat{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T)$.

As for $Q_{P_T^B}(\hat{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T)$, we can obtain that for all $\epsilon > 0$,

$$\begin{aligned}
&P_{P_T^b}\left(Q_{P_T^B}(\hat{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T) > \epsilon\right) \\
&\leq P_{P_T^b}\left(\left|Q_{P_T^B}(\hat{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T)\right| > \epsilon\right) \\
&= P_{P_T^b}\left(\left|\frac{1}{B}\sum_{i=1}^{B} Q_T(\hat{\theta}_{\hat{\lambda}_T}, X_{T,i}^b) - Q_{P_T^b}(\hat{\lambda}_T)\right| > \epsilon\right) \\
&\leq K(T) \max_{\lambda_T \in \Lambda_T} P_{P_T^b}\left(\left|\frac{1}{B}\sum_{i=1}^{B} Q_T(\hat{\theta}_{\lambda_T}, X_{T,i}^b) - Q_{P_T^b}(\lambda_T)\right| > \epsilon\right).
\end{aligned}$$

Since $\sup_{\hat{\theta}_{\lambda_T}\in\Theta, X_T^b \in \mathscr{X}_T^b}\left|Q_T(\hat{\theta}_{\lambda_T}, X_T^b) - Q_{P_T^b}(\lambda_T)\right| \leq C_2 < \infty$ *a.s.* according to the assumption in Theorem 9 and B bootstrap sequences are i.i.d. (i.e., $X_{T,1}^b, \cdots, X_{T,B}^b$ are i.i.d.), then it follows that

$$P_{P_T^b}\left(\frac{1}{B}\sum_{i=1}^{B} Q_T(\hat{\theta}_{\lambda_T}, X_{T,i}^b) - Q_{P_T^b}(\lambda_T) > \epsilon\right) \leq \exp\left[-\frac{1}{2}\frac{\epsilon^2 B}{C_5 + \epsilon C_2/3}\right], \tag{B.152}$$

where $C_5 = E_{P_T^b}\left|Q_T(\hat{\theta}_{\lambda_T}, X_T^b) - Q_{P_T^b}(\lambda_T)\right|^2$, by the Bernstein inequality for i.i.d. random variables (see Theorem 4.1 in Modha and Masry (1996)). Note that (B.152) implies

$$P_{P_T^b}\left(\left|\frac{1}{B}\sum_{i=1}^{B} Q_T(\hat{\theta}_{\lambda_T}, X_{T,i}^b) - Q_{P_T^b}(\lambda_T)\right| > \epsilon\right) \leq 2\exp\left[-\frac{1}{2}\frac{\epsilon^2 B}{C_5 + \epsilon C_2/3}\right].$$

Hence,

$$P_{P_T^b}\left(Q_{P_T^B}(\hat{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T) > \epsilon\right) \leq 2K(T)\exp\left[-\frac{1}{2}\frac{\epsilon^2 B}{C_5 + 2\epsilon C_2/3}\right].$$

Since, for any random variable X,

$$E(X) \leq E[I(X>0)X] = \int_0^\infty P(X>x)dx,$$

we have

$$\begin{aligned}
& E_{P_T^b}\left(Q_{P_T^B}(\hat{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T)\right) \\
\leq & \int_0^\infty P_{P_T^b}\left(Q_{P_T^B}(\hat{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T) > x\right) dx \\
\leq & \int_0^\infty 2K(T)\exp\left[-\frac{1}{2}\frac{x^2B}{C_5+2xC_2/3}\right] dx \\
= & \int_0^1 2K(T)\exp\left[-\frac{1}{2}\frac{x^2B}{C_5+2xC_2/3}\right] dx + \int_1^\infty 2K(T)\exp\left[-\frac{1}{2}\frac{x^2B}{C_5+2xC_2/3}\right] dx \\
\leq & \int_0^1 2K(T)\exp\left[-\frac{1}{2}\frac{x^2B}{C_5+2C_2/3}\right] dx + \int_1^\infty 2K(T)\exp\left[-\frac{1}{2}\frac{xB}{C_5+2C_2/3}\right] dx \\
= & \frac{2K(T)\sqrt{\pi(C_5+2C_2/3)/2}\left[\mathrm{Erf}\left(\sqrt{\frac{B}{2(C_5+2C_2/3)}}\right)\right]}{\sqrt{B}} \\
& + \frac{4K(T)(C_5+2C_2/3)\exp\left[-\frac{B}{2(C_5+2C_2/3)}\right]}{B} \\
\leq & \frac{2K(T)\sqrt{\pi(C_5+2C_2/3)/2}}{\sqrt{B}} + \frac{4K(T)(C_5+2C_2/3)}{B} \\
= & O\left(\frac{K(T)}{\sqrt{B}}\right) + O\left(\frac{K(T)}{B}\right) \\
= & O\left(\frac{K(T)}{\sqrt{B}}\right).
\end{aligned}$$

Hence,

$$E_{P_T^b}\left[Q_{P_T^B}(\hat{\lambda}_T) - Q_{P_T^b}(\hat{\lambda}_T)\right] \leq O\left(\frac{K(T)}{\sqrt{B}}\right). \tag{B.153}$$

Similarly for $Q_{P_T^b}(\bar{\lambda}_T) - Q_{P_T^B}(\bar{\lambda}_T)$, we have

$$E_{P_T^b}\left[Q_{P_T^b}(\bar{\lambda}_T) - Q_{P_T^B}(\bar{\lambda}_T)\right] \leq O\left(\frac{K(T)}{\sqrt{B}}\right). \tag{B.154}$$

Lastly, we deal with $Q_{p_o}(\check{\lambda}_T) - Q_{p_o}(\tilde{\lambda}_T)$. Let $\lambda_T^* = \arg\min_{\lambda_T \in \Lambda_T} \left|\lambda_T - \tilde{\lambda}_T\right|$. Note that

$$Q_{p_o}(\check{\lambda}_T) - Q_{p_o}(\tilde{\lambda}_T) \leq Q_{p_o}(\lambda_T^*) - Q_{p_o}(\tilde{\lambda}_T).$$

Since $Q_{p_o}(\lambda_T)$ is Lipschitz continuous by the assumption in Theorem 9,

$$\left|Q_{p_o}(\lambda_T^*) - Q_{p_o}(\tilde{\lambda}_T)\right| \leq m \left|\lambda_T^* - \tilde{\lambda}_T\right| = O\left(\frac{c}{K(T)}\right), \tag{B.155}$$

where m is a positive real constant.

Combining (B.148), (B.149), (B.150), (B.151), (B.153), (B.154), and (B.155), we get the results by taking expectations on both sides of (B.144). □

Proof of Theorem 11

Proof. Since $\left\|\hat{\theta} - \theta_0\right\| = o_p(1)$ by Lemma 1, we have $\hat{D} = D_o + o_p(1)$ where D_o means the distribution with real values of parameters, which implies that $P_T^b = P_o + o_p(1)$. Therefore, $A = o_p(1)$. □

B.3 Artificial simulations and empirical applications

This section presents more details about 3 Cases in the artificial simulations and 2 schemes (i.e., the incremental window and rolling window) in the empirical applications. Note that for each table, we present the result of each fixed λ besides the results of the 1SMLE, 2SQMLE, MLE, TMLE_1, and TMLE_2.

Case 1

Case 1: $T = 100$, $\tau = 95$

Table B.1: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	8.58727	0.00954	0.01204	0.01083	0.01077	0.01068	0.01066
b_{22}	11.69063	5.62809	0.43382	0.02211	0.01453	0.01296	0.01218
$\overline{\text{MSE}}$	20.27790	5.63763	0.44586	0.03294	0.02530	0.02364	0.02284
	1	2	4	6	8	10	20
b_{11}	0.01069	0.01055	0.01054	0.01054	0.01054	0.01054	0.01054
b_{22}	0.01185	0.01096	0.01073	0.01066	0.01063	0.01061	0.01058
$\overline{\text{MSE}}$	0.02254	0.02151	0.02128	0.02120	0.02116	0.02116	0.02111
	40	60	80	100	200	400	600
b_{11}	0.01053	0.01053	0.01054	0.01056	0.01055	0.01048	0.01050
b_{22}	0.01055	0.01054	0.01055	0.01056	0.01055	0.01049	0.01050
$\overline{\text{MSE}}$	0.02108	0.02106	0.02109	0.02112	0.02110	0.02097	0.02100
	800	1000	TMLE$_1$	TMLE$_2$			
b_{11}	0.01033	0.01040	0.01024	0.01029			
b_{22}	0.01033	0.01040	0.01583	0.01484			
$\overline{\text{MSE}}$	0.02066	0.02080	0.02607	0.02513			

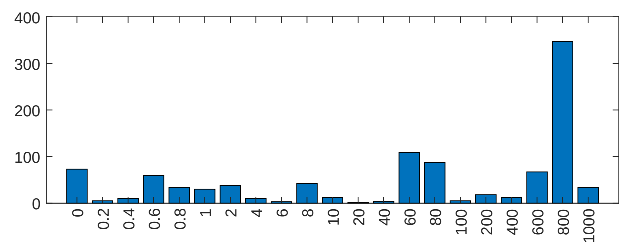


Figure B.1: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

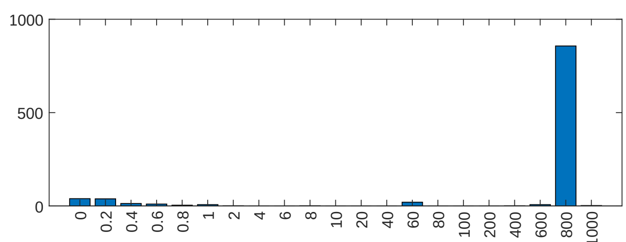


Figure B.2: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 1: $T = 100$**,** $\tau = 90$

Table B.2: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.16719	0.00954	0.00989	0.00950	0.00955	0.00957	0.00958
b_{22}	0.15697	0.14541	0.12073	0.01576	0.01179	0.01081	0.01041
$\overline{\text{MSE}}$	0.32417	0.15495	0.13062	0.02526	0.02134	0.02038	0.01999
	1	2	4	6	8	10	20
b_{11}	0.00959	0.00960	0.00961	0.00961	0.00961	0.00961	0.00961
b_{22}	0.01020	0.00985	0.00972	0.00968	0.00966	0.00965	0.00963
$\overline{\text{MSE}}$	0.01979	0.01945	0.01933	0.01929	0.01927	0.01926	0.01924
	40	60	80	100	200	400	600
b_{11}	0.00961	0.00961	0.00962	0.00961	0.00961	0.00962	0.00961
b_{22}	0.00962	0.00962	0.00963	0.00962	0.00962	0.00962	0.00961
$\overline{\text{MSE}}$	0.01923	0.01923	0.01925	0.01923	0.01923	0.01924	0.01923
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00964	0.00965	0.00965	0.00953			
b_{22}	0.00964	0.00965	0.01260	0.01234			
$\overline{\text{MSE}}$	0.01928	0.01931	0.02225	0.02187			

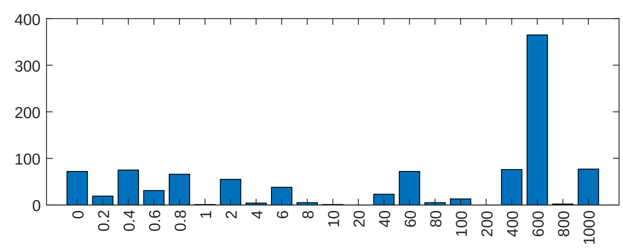


Figure B.3: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

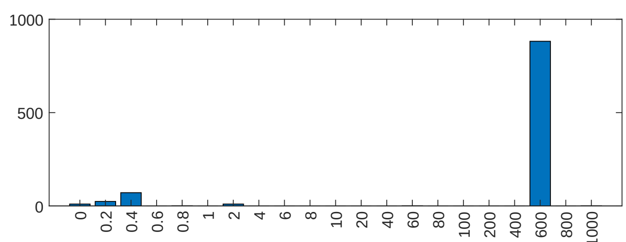


Figure B.4: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 1: $T = 100$**,** $\tau = 80$

Table B.3: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.05995	0.00954	0.00988	0.00889	0.00893	0.00896	0.00898
b_{22}	0.06553	0.06488	0.06108	0.01553	0.01165	0.01053	0.01004
$\overline{\text{MSE}}$	0.12549	0.07442	0.07096	0.02442	0.02058	0.01949	0.01902

	1	2	4	6	8	10	20
b_{11}	0.00900	0.00902	0.00904	0.00904	0.00905	0.00905	0.00905
b_{22}	0.00980	0.00935	0.00918	0.00914	0.00911	0.00910	0.00908
$\overline{\text{MSE}}$	0.01880	0.01837	0.01822	0.01818	0.01816	0.01815	0.01813

	40	60	80	100	200	400	600
b_{11}	0.00905	0.00905	0.00905	0.00906	0.00906	0.00904	0.00904
b_{22}	0.00907	0.00906	0.00906	0.00906	0.00906	0.00904	0.00904
$\overline{\text{MSE}}$	0.01812	0.01812	0.01812	0.01812	0.01811	0.01808	0.01807

	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00904	0.00900	0.00899	0.00892			
b_{22}	0.00904	0.00900	0.01139	0.01152			
$\overline{\text{MSE}}$	0.01807	0.01800	0.02038	0.02044			

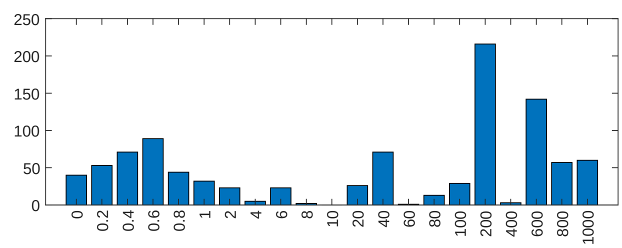


Figure B.5: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

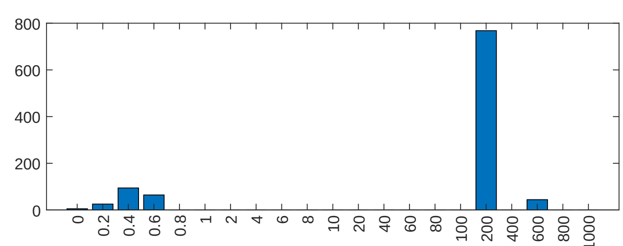


Figure B.6: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 1: $T = 400$**,** $\tau = 390$

Table B.4: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.17447	0.00261	0.00263	0.00259	0.00259	0.00259	0.00259
b_{22}	0.15926	0.14803	0.11628	0.00306	0.00273	0.00266	0.00263
$\overline{\text{MSE}}$	0.33373	0.15064	0.11891	0.00565	0.00531	0.00525	0.00522
	1	2	4	6	8	10	20
b_{11}	0.00259	0.00259	0.00259	0.00259	0.00259	0.00259	0.00259
b_{22}	0.00262	0.00260	0.00259	0.00259	0.00259	0.00259	0.00259
$\overline{\text{MSE}}$	0.00521	0.00519	0.00518	0.00518	0.00518	0.00518	0.00518
	40	60	80	100	200	400	600
b_{11}	0.00259	0.00259	0.00259	0.00259	0.00259	0.00259	0.00261
b_{22}	0.00259	0.00259	0.00259	0.00260	0.00259	0.00259	0.00261
$\overline{\text{MSE}}$	0.00518	0.00518	0.00518	0.00519	0.00518	0.00517	0.00522
	800	1000	TMLE$_1$	TMLE$_2$			
b_{11}	0.00263	0.00267	0.00259	0.00258			
b_{22}	0.00263	0.00267	0.00569	0.00314			
$\overline{\text{MSE}}$	0.00527	0.00534	0.00828	0.00572			

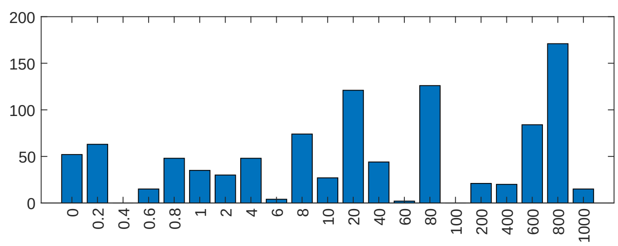


Figure B.7: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

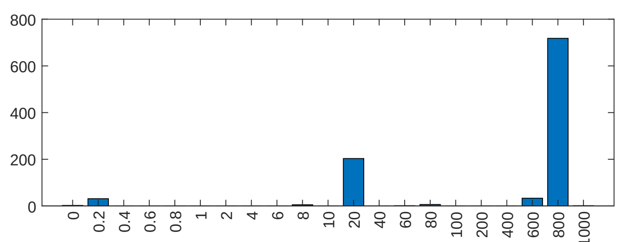


Figure B.8: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 1: $T = 400$**,** $\tau = 380$

Table B.5: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.05915	0.00261	0.00262	0.00254	0.00254	0.00254	0.00254
b_{22}	0.05323	0.05276	0.04970	0.00319	0.00273	0.00264	0.00260
$\overline{\text{MSE}}$	0.11238	0.05537	0.05232	0.00573	0.00527	0.00517	0.00514
	1	2	4	6	8	10	20
b_{11}	0.00253	0.00254	0.00254	0.00254	0.00254	0.00254	0.00254
b_{22}	0.00258	0.00255	0.00255	0.00254	0.00254	0.00254	0.00254
$\overline{\text{MSE}}$	0.00511	0.00509	0.00508	0.00508	0.00508	0.00508	0.00508
	40	60	80	100	200	400	600
b_{11}	0.00254	0.00254	0.00254	0.00254	0.00254	0.00254	0.00254
b_{22}	0.00254	0.00254	0.00254	0.00254	0.00254	0.00254	0.00254
$\overline{\text{MSE}}$	0.00509	0.00508	0.00509	0.00508	0.00508	0.00508	0.00509
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00255	0.00259	0.00253	0.00254			
b_{22}	0.00255	0.00259	0.00340	0.00261			
$\overline{\text{MSE}}$	0.00511	0.00519	0.00593	0.00515			

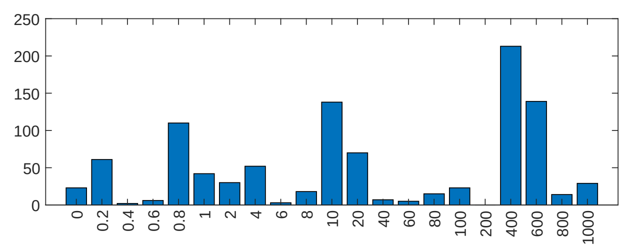


Figure B.9: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

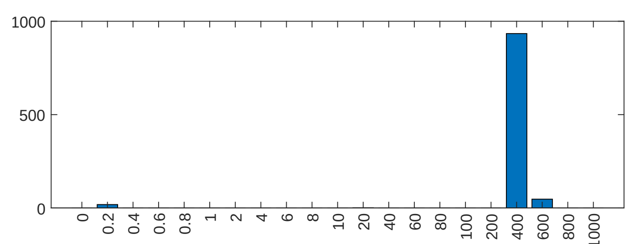


Figure B.10: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 1: $T = 400$, $\tau = 360$

Table B.6: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.02674	0.00261	0.00262	0.00241	0.00240	0.00240	0.00240
b_{22}	0.02688	0.02681	0.02602	0.00327	0.00266	0.00252	0.00247
$\overline{\text{MSE}}$	0.05362	0.02942	0.02863	0.00568	0.00506	0.00492	0.00487
	1	2	4	6	8	10	20
b_{11}	0.00240	0.00240	0.00240	0.00240	0.00240	0.00240	0.00240
b_{22}	0.00245	0.00241	0.00240	0.00240	0.00240	0.00240	0.00240
$\overline{\text{MSE}}$	0.00485	0.00481	0.00481	0.00480	0.00480	0.00480	0.00480
	40	60	80	100	200	400	600
b_{11}	0.00240	0.00240	0.00240	0.00240	0.00240	0.00241	0.00241
b_{22}	0.00240	0.00240	0.00240	0.00240	0.00240	0.00241	0.00241
$\overline{\text{MSE}}$	0.00480	0.00480	0.00480	0.00480	0.00480	0.00481	0.00482
	800	1000	TMLE$_1$	TMLE$_2$			
b_{11}	0.00240	0.00242	0.00240	0.00241			
b_{22}	0.00240	0.00242	0.00275	0.00250			
$\overline{\text{MSE}}$	0.00480	0.00483	0.00515	0.00490			

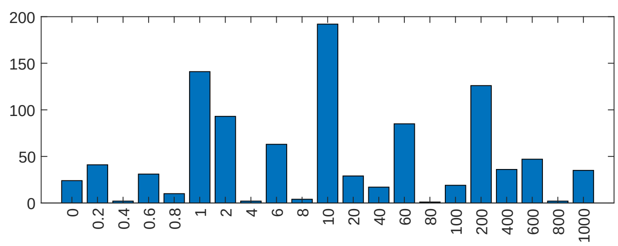


Figure B.11: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

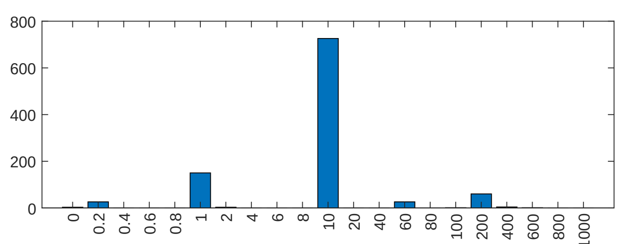


Figure B.12: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 1: $T = 900$**,** $\tau = 885$

Table B.7: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.08477	0.00115	0.00115	0.00115	0.00116	0.00115	0.00115
b_{22}	0.07754	0.07437	0.06909	0.00131	0.00121	0.00118	0.00117
$\overline{\text{MSE}}$	0.16231	0.07552	0.07025	0.00246	0.00237	0.00233	0.00233
	1	2	4	6	8	10	20
b_{11}	0.00115	0.00116	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.00117	0.00116	0.00116	0.00116	0.00115	0.00115	0.00115
$\overline{\text{MSE}}$	0.00232	0.00232	0.00231	0.00231	0.00231	0.00231	0.00231
	40	60	80	100	200	400	600
b_{11}	0.00115	0.00115	0.00116	0.00115	0.00115	0.00115	0.00115
b_{22}	0.00115	0.00115	0.00116	0.00115	0.00115	0.00115	0.00115
$\overline{\text{MSE}}$	0.00231	0.00231	0.00232	0.00231	0.00230	0.00231	0.00231
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00115	0.00117	0.00115	0.00116			
b_{22}	0.00115	0.00117	0.00184	0.00116			
$\overline{\text{MSE}}$	0.00230	0.00235	0.00300	0.00232			

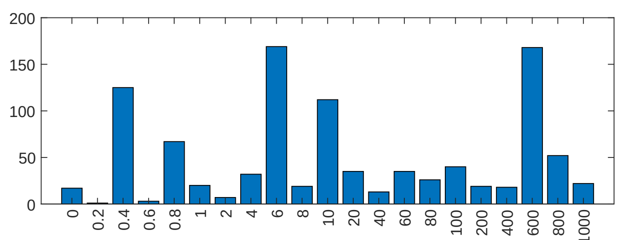


Figure B.13: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

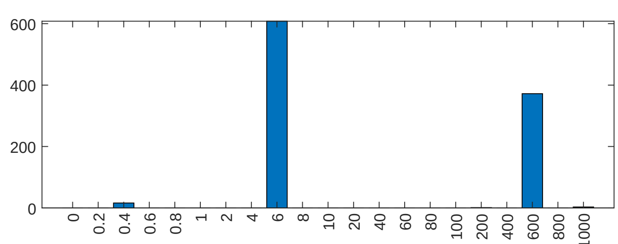


Figure B.14: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 1: $T = 900$**,** $\tau = 870$

Table B.8: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.03477	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.03519	0.03511	0.03346	0.00141	0.00124	0.00120	0.00118
$\overline{\text{MSE}}$	0.06997	0.03626	0.03461	0.00255	0.00239	0.00234	0.00233
	1	2	4	6	8	10	20
b_{11}	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.00118	0.00116	0.00115	0.00115	0.00115	0.00115	0.00115
$\overline{\text{MSE}}$	0.00233	0.00231	0.00230	0.00230	0.00230	0.00230	0.00230
	40	60	80	100	200	400	600
b_{11}	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
$\overline{\text{MSE}}$	0.00230	0.00230	0.00231	0.00231	0.00230	0.00230	0.00230
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00119	0.00121	0.00115	0.00115			
b_{22}	0.00119	0.00122	0.00144	0.00115			
$\overline{\text{MSE}}$	0.00238	0.00243	0.00259	0.00230			

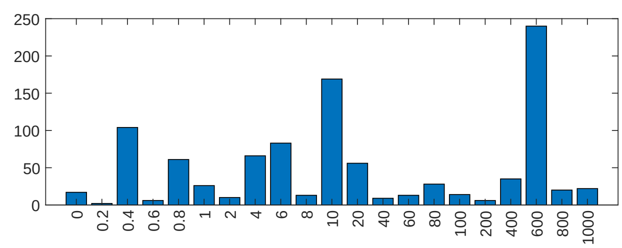


Figure B.15: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

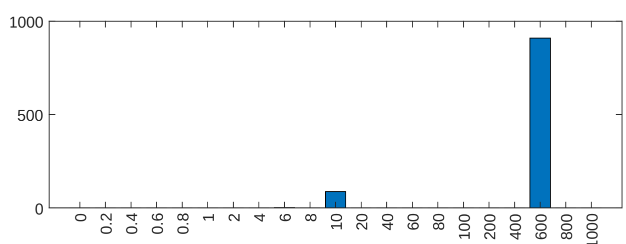


Figure B.16: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 1: $T = 900$**,** $\tau = 840$

Table B.9: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.01576	0.00115	0.00115	0.00110	0.00110	0.00110	0.00110
b_{22}	0.01829	0.01826	0.01815	0.00148	0.00123	0.00117	0.00114
$\overline{\text{MSE}}$	0.03405	0.01941	0.01930	0.00258	0.00233	0.00227	0.00225
	1	2	4	6	8	10	20
b_{11}	0.00110	0.00110	0.00110	0.00110	0.00110	0.00110	0.00110
b_{22}	0.00113	0.00111	0.00111	0.00111	0.00111	0.00111	0.00111
$\overline{\text{MSE}}$	0.00224	0.00222	0.00221	0.00221	0.00221	0.00221	0.00221
	40	60	80	100	200	400	600
b_{11}	0.00110	0.00110	0.00110	0.00110	0.00110	0.00110	0.00111
b_{22}	0.00110	0.00110	0.00110	0.00110	0.00110	0.00110	0.00111
$\overline{\text{MSE}}$	0.00221	0.00221	0.00221	0.00221	0.00221	0.00221	0.00222
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00110	0.00110	0.00110	0.00110			
b_{22}	0.00110	0.00110	0.00129	0.00113			
$\overline{\text{MSE}}$	0.00221	0.00221	0.00240	0.00224			

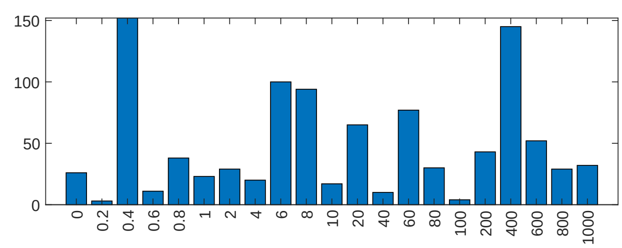


Figure B.17: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

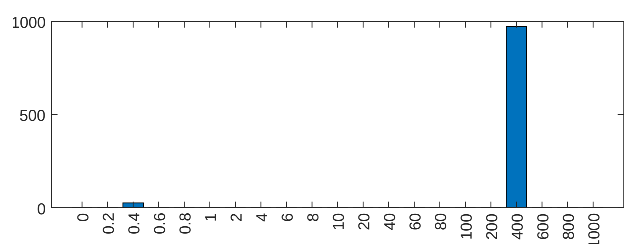


Figure B.18: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 1: $T = 900$, $\tau = 450$

Table B.10: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.00225	0.00115	0.00114	0.00086	0.00082	0.00080	0.00080
b_{22}	0.00218	0.00218	0.00218	0.00108	0.00091	0.00085	0.00083
$\overline{\text{MSE}}$	0.00443	0.00333	0.00332	0.00194	0.00172	0.00165	0.00162
	1	2	4	6	8	10	20
b_{11}	0.00079	0.00079	0.00078	0.00078	0.00078	0.00078	0.00078
b_{22}	0.00081	0.00079	0.00079	0.00078	0.00078	0.00078	0.00078
$\overline{\text{MSE}}$	0.00160	0.00158	0.00157	0.00157	0.00157	0.00157	0.00157
	40	60	80	100	200	400	600
b_{11}	0.00078	0.00078	0.00078	0.00078	0.00078	0.00078	0.00078
b_{22}	0.00078	0.00078	0.00078	0.00078	0.00078	0.00078	0.00078
$\overline{\text{MSE}}$	0.00157	0.00157	0.00157	0.00157	0.00157	0.00157	0.00157
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00078	0.00078	0.00080	0.00081			
b_{22}	0.00078	0.00078	0.00084	0.00086			
$\overline{\text{MSE}}$	0.00157	0.00157	0.00164	0.00167			

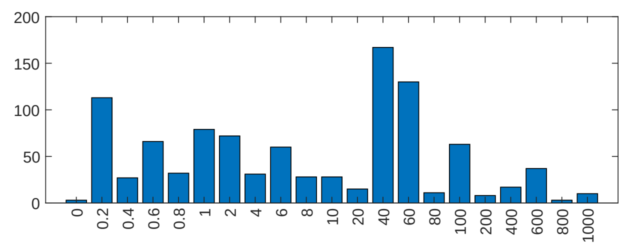


Figure B.19: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

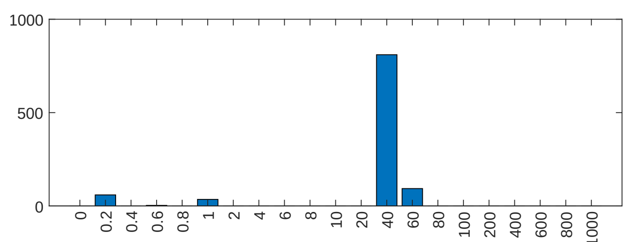


Figure B.20: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2

Case 2: $T = 100$, $\tau = 95$

Table B.11: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	8.89113	0.00954	0.01204	0.01080	0.01081	0.01080	0.01081
b_{22}	9.98371	1.93667	0.44219	0.04405	0.03239	0.02979	0.02859
$\overline{\text{MSE}}$	18.87483	1.94621	0.45423	0.05485	0.04320	0.04059	0.03940
	1	2	4	6	8	10	20
b_{11}	0.01081	0.01074	0.01089	0.01089	0.01066	0.01087	0.01064
b_{22}	0.02793	0.02652	0.02604	0.02581	0.02556	0.02565	0.02542
$\overline{\text{MSE}}$	0.03874	0.03726	0.03694	0.03671	0.03621	0.03653	0.03606
	40	60	80	100	200	400	600
b_{11}	0.01061	0.01061	0.01060	0.01086	0.01061	0.01056	0.01057
b_{22}	0.02532	0.02531	0.02526	0.02549	0.02529	0.02518	0.02520
$\overline{\text{MSE}}$	0.03593	0.03592	0.03586	0.03635	0.03591	0.03574	0.03577
	800	1000	TMLE$_1$	TMLE$_2$			
b_{11}	0.01066	0.01054	0.01034	0.01040			
b_{22}	0.02526	0.02509	0.02894	0.03074			
$\overline{\text{MSE}}$	0.03592	0.03563	0.03928	0.04113			

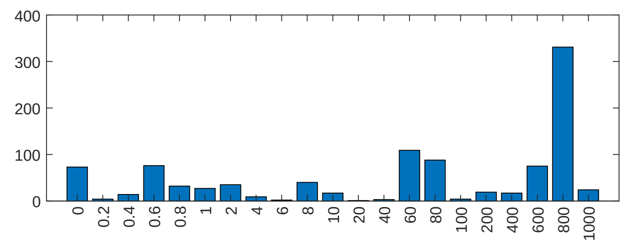


Figure B.21: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

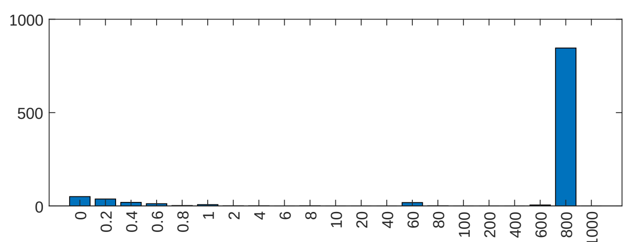


Figure B.22: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2: $T = 100$**,** $\tau = 90$

Table B.12: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.17309	0.00954	0.00995	0.00933	0.00936	0.00938	0.00939
b_{22}	0.16933	0.15746	0.13164	0.03089	0.02630	0.02506	0.02451
$\overline{\text{MSE}}$	0.34242	0.16700	0.14159	0.04022	0.03566	0.03443	0.03389
	1	2	4	6	8	10	20
b_{11}	0.00939	0.00940	0.00941	0.00941	0.00941	0.00941	0.00941
b_{22}	0.02421	0.02367	0.02344	0.02337	0.02334	0.02332	0.02328
$\overline{\text{MSE}}$	0.03360	0.03308	0.03285	0.03279	0.03275	0.03273	0.03269
	40	60	80	100	200	400	600
b_{11}	0.00941	0.00942	0.00942	0.00941	0.00941	0.00946	0.00941
b_{22}	0.02326	0.02325	0.02325	0.02323	0.02324	0.02330	0.02324
$\overline{\text{MSE}}$	0.03268	0.03267	0.03267	0.03264	0.03265	0.03276	0.03264
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00945	0.00948	0.00949	0.00929			
b_{22}	0.02328	0.02329	0.02462	0.02686			
$\overline{\text{MSE}}$	0.03273	0.03276	0.03411	0.03615			

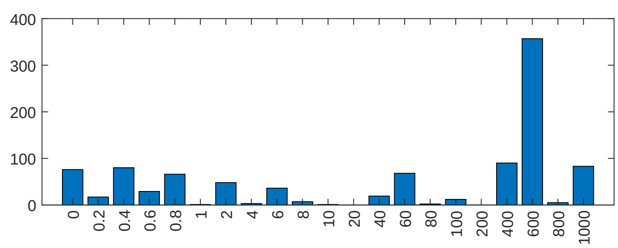


Figure B.23: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

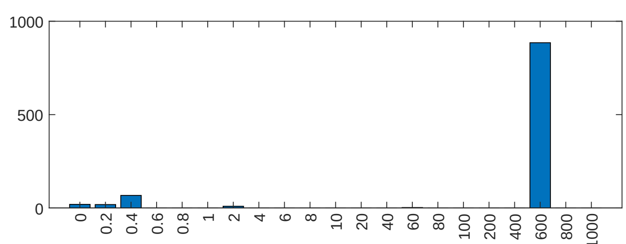


Figure B.24: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2: $T = 100$**,** $\tau = 80$

Table B.13: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.05995	0.00954	0.00988	0.00857	0.00856	0.00857	0.00858
b_{22}	0.06883	0.06819	0.06486	0.02547	0.02234	0.02151	0.02118
$\overline{\text{MSE}}$	0.12878	0.07773	0.07474	0.03404	0.03090	0.03008	0.02976
	1	2	4	6	8	10	20
b_{11}	0.00859	0.00862	0.00863	0.00863	0.00864	0.00864	0.00864
b_{22}	0.02102	0.02080	0.02073	0.02072	0.02072	0.02072	0.02072
$\overline{\text{MSE}}$	0.02961	0.02942	0.02936	0.02935	0.02935	0.02935	0.02936
	40	60	80	100	200	400	600
b_{11}	0.00864	0.00864	0.00864	0.00864	0.00866	0.00864	0.00865
b_{22}	0.02072	0.02072	0.02072	0.02072	0.02074	0.02073	0.02074
$\overline{\text{MSE}}$	0.02936	0.02936	0.02936	0.02936	0.02940	0.02937	0.02939
	800	1000	TMLE$_1$	TMLE$_2$			
b_{11}	0.00866	0.00865	0.00850	0.00843			
b_{22}	0.02078	0.02071	0.02212	0.02292			
$\overline{\text{MSE}}$	0.02944	0.02936	0.03062	0.03135			

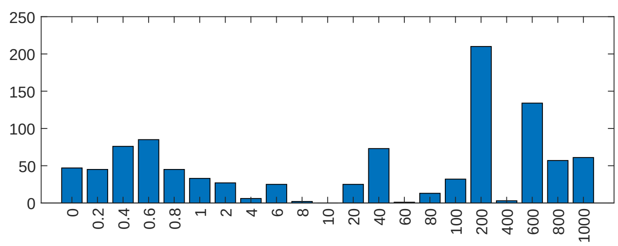


Figure B.25: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

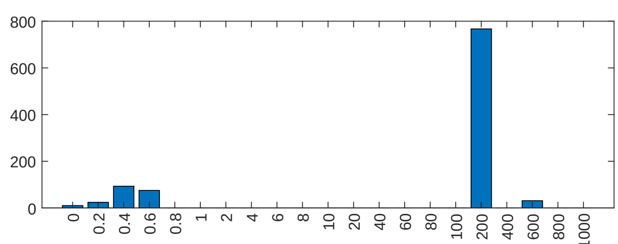


Figure B.26: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2: $T = 400$**,** $\tau = 390$

Table B.14: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.17477	0.00261	0.00263	0.00257	0.00257	0.00257	0.00257
b_{22}	0.17075	0.15851	0.12464	0.01422	0.01366	0.01352	0.01346
$\overline{\text{MSE}}$	0.34552	0.16113	0.12727	0.01679	0.01623	0.01609	0.01603
	1	2	4	6	8	10	20
b_{11}	0.00257	0.00257	0.00257	0.00257	0.00257	0.00257	0.00257
b_{22}	0.01342	0.01336	0.01332	0.01332	0.01331	0.01331	0.01331
$\overline{\text{MSE}}$	0.01599	0.01593	0.01589	0.01589	0.01588	0.01588	0.01588
	40	60	80	100	200	400	600
b_{11}	0.00257	0.00258	0.00257	0.00257	0.00258	0.00260	0.00261
b_{22}	0.01331	0.01330	0.01331	0.01331	0.01330	0.01334	0.01342
$\overline{\text{MSE}}$	0.01588	0.01588	0.01588	0.01588	0.01589	0.01594	0.01602
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00261	0.00259	0.00258	0.00256			
b_{22}	0.01337	0.01325	0.01575	0.01405			
$\overline{\text{MSE}}$	0.01598	0.01584	0.01833	0.01661			

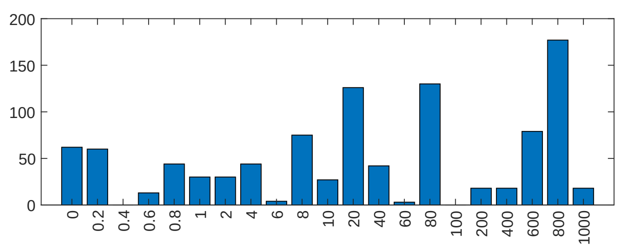


Figure B.27: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

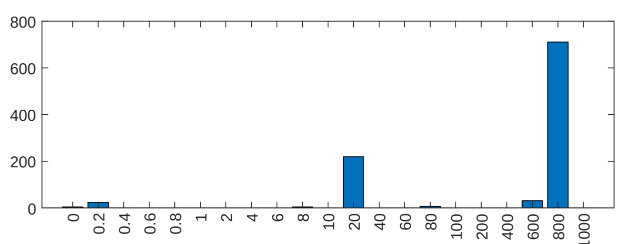


Figure B.28: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2: $T = 400$**,** $\tau = 380$

Table B.15: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.05926	0.00261	0.00262	0.00251	0.00251	0.00251	0.00251
b_{22}	0.05498	0.05439	0.05216	0.01283	0.01266	0.01267	0.01269
$\overline{\text{MSE}}$	0.11425	0.05700	0.05478	0.01534	0.01517	0.01518	0.01520
	1	2	4	6	8	10	20
b_{11}	0.00251	0.00251	0.00251	0.00251	0.00251	0.00251	0.00251
b_{22}	0.01268	0.01274	0.01277	0.01278	0.01278	0.01278	0.01279
$\overline{\text{MSE}}$	0.01519	0.01525	0.01528	0.01529	0.01529	0.01529	0.01530
	40	60	80	100	200	400	600
b_{11}	0.00251	0.00251	0.00251	0.00251	0.00251	0.00251	0.00252
b_{22}	0.01279	0.01280	0.01280	0.01280	0.01280	0.01281	0.01282
$\overline{\text{MSE}}$	0.01530	0.01531	0.01531	0.01531	0.01531	0.01532	0.01534
	800	1000	TMLE$_1$	TMLE$_2$			
b_{11}	0.00251	0.00252	0.00250	0.00250			
b_{22}	0.01280	0.01284	0.01343	0.01323			
$\overline{\text{MSE}}$	0.01532	0.01536	0.01593	0.01573			

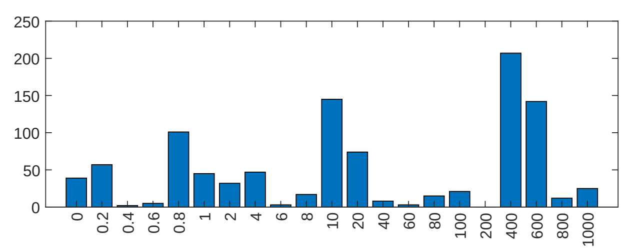


Figure B.29: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

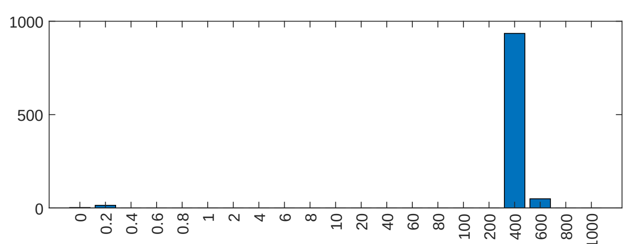


Figure B.30: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2: $T = 400$, $\tau = 360$

Table B.16: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.02675	0.00261	0.00262	0.00238	0.00238	0.00238	0.00238
b_{22}	0.02732	0.02723	0.02660	0.01077	0.01099	0.01119	0.01131
$\overline{\text{MSE}}$	0.05407	0.02984	0.02921	0.01315	0.01337	0.01356	0.01369
	1	2	4	6	8	10	20
b_{11}	0.00238	0.00238	0.00238	0.00238	0.00238	0.00238	0.00238
b_{22}	0.01140	0.01159	0.01170	0.01174	0.01176	0.01178	0.01180
$\overline{\text{MSE}}$	0.01377	0.01397	0.01408	0.01412	0.01414	0.01416	0.01418
	40	60	80	100	200	400	600
b_{11}	0.00238	0.00238	0.00238	0.00238	0.00238	0.00238	0.00238
b_{22}	0.01181	0.01182	0.01182	0.01182	0.01182	0.01183	0.01184
$\overline{\text{MSE}}$	0.01419	0.01420	0.01420	0.01420	0.01420	0.01421	0.01422
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00239	0.00239	0.00237	0.00238			
b_{22}	0.01193	0.01183	0.01156	0.01150			
$\overline{\text{MSE}}$	0.01432	0.01422	0.01394	0.01387			

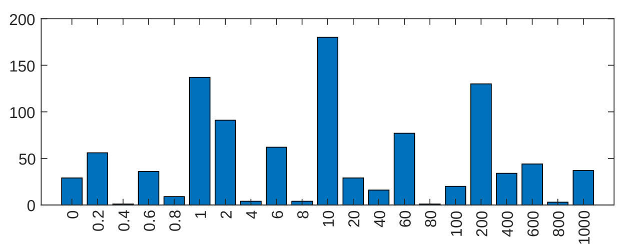


Figure B.31: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

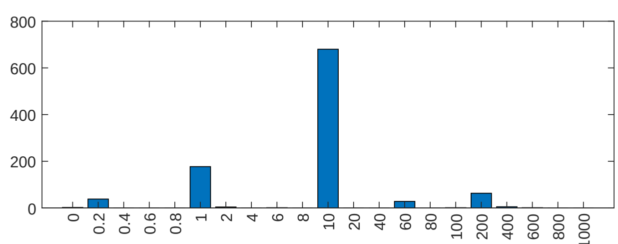


Figure B.32: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2: $T = 900$, $\tau = 885$

Table B.17: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.08515	0.00115	0.00115	0.00115	0.00115	0.00115	0.00116
b_{22}	0.08341	0.07968	0.07515	0.01150	0.01134	0.01127	0.01126
$\overline{\text{MSE}}$	0.16856	0.08083	0.07630	0.01265	0.01249	0.01243	0.01242
	1	2	4	6	8	10	20
b_{11}	0.00116	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.01127	0.01125	0.01125	0.01124	0.01124	0.01124	0.01124
$\overline{\text{MSE}}$	0.01243	0.01240	0.01240	0.01239	0.01239	0.01239	0.01239
	40	60	80	100	200	400	600
b_{11}	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115	0.00119
b_{22}	0.01124	0.01124	0.01124	0.01124	0.01124	0.01124	0.01124
$\overline{\text{MSE}}$	0.01239	0.01239	0.01239	0.01239	0.01239	0.01239	0.01243
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00115	0.00115	0.00115	0.00115			
b_{22}	0.01125	0.01124	0.01195	0.01125			
$\overline{\text{MSE}}$	0.01239	0.01239	0.01310	0.01240			

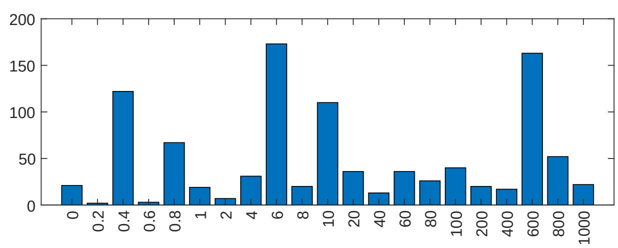


Figure B.33: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

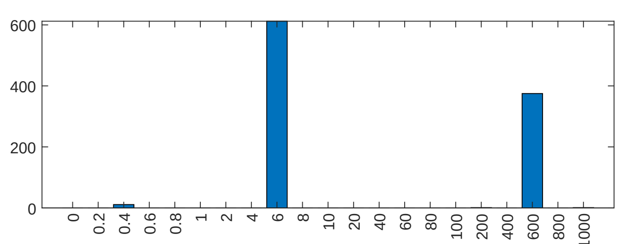


Figure B.34: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2: $T = 900$**,** $\tau = 870$

Table B.18: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.03481	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.03591	0.03580	0.03452	0.01043	0.01059	0.01068	0.01072
$\overline{\text{MSE}}$	0.07072	0.03695	0.03567	0.01158	0.01174	0.01183	0.01188
	1	2	4	6	8	10	20
b_{11}	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.01075	0.01081	0.01084	0.01085	0.01086	0.01086	0.01087
$\overline{\text{MSE}}$	0.01191	0.01196	0.01199	0.01200	0.01201	0.01201	0.01202
	40	60	80	100	200	400	600
b_{11}	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.01086	0.01087	0.01089	0.01088	0.01086	0.01090	0.01092
$\overline{\text{MSE}}$	0.01201	0.01203	0.01205	0.01204	0.01201	0.01205	0.01207
	800	1000	TMLE$_1$	TMLE$_2$			
b_{11}	0.00117	0.00118	0.00115	0.00115			
b_{22}	0.01089	0.01091	0.01103	0.01088			
$\overline{\text{MSE}}$	0.01206	0.01209	0.01217	0.01203			

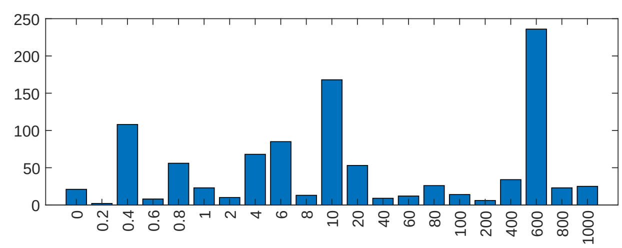


Figure B.35: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

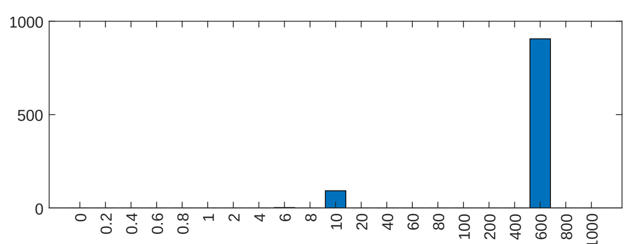


Figure B.36: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2: $T = 900$**,** $\tau = 840$

Table B.19: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.01577	0.00115	0.00115	0.00112	0.00112	0.00112	0.00113
b_{22}	0.01812	0.01808	0.01810	0.00869	0.00926	0.00952	0.00966
$\overline{\text{MSE}}$	0.03389	0.01923	0.01926	0.00981	0.01038	0.01064	0.01079
	1	2	4	6	8	10	20
b_{11}	0.00113	0.00113	0.00113	0.00113	0.00113	0.00113	0.00113
b_{22}	0.00975	0.00995	0.01005	0.01009	0.01011	0.01012	0.01014
$\overline{\text{MSE}}$	0.01088	0.01108	0.01118	0.01122	0.01124	0.01125	0.01127
	40	60	80	100	200	400	600
b_{11}	0.00113	0.00113	0.00113	0.00113	0.00113	0.00113	0.00117
b_{22}	0.01015	0.01016	0.01016	0.01016	0.01016	0.01016	0.01018
$\overline{\text{MSE}}$	0.01128	0.01128	0.01129	0.01129	0.01129	0.01129	0.01135
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00113	0.00113	0.00113	0.00113			
b_{22}	0.01017	0.01017	0.00993	0.00994			
$\overline{\text{MSE}}$	0.01129	0.01130	0.01106	0.01107			

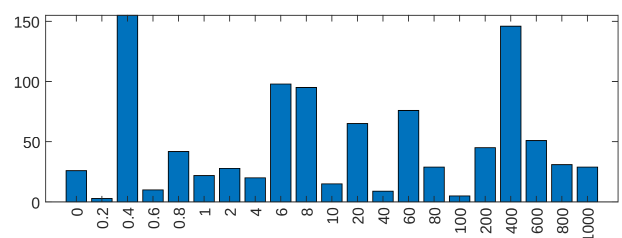


Figure B.37: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

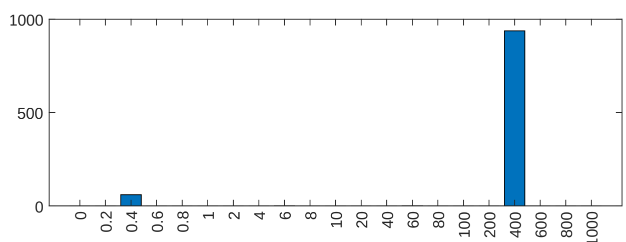


Figure B.38: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 2: $T = 900$, $\tau = 450$

Table B.20: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.00225	0.00115	0.00114	0.00117	0.00136	0.00148	0.00155
b_{22}	0.00210	0.00210	0.00211	0.00235	0.00306	0.00350	0.00380
$\overline{\text{MSE}}$	0.00435	0.00325	0.00325	0.00353	0.00441	0.00498	0.00535
	1	2	4	6	8	10	20
b_{11}	0.00161	0.00174	0.00183	0.00186	0.00187	0.00188	0.00190
b_{22}	0.00401	0.00451	0.00482	0.00493	0.00499	0.00502	0.00510
$\overline{\text{MSE}}$	0.00561	0.00625	0.00664	0.00679	0.00686	0.00690	0.00700
	40	60	80	100	200	400	600
b_{11}	0.00191	0.00191	0.00191	0.00192	0.00192	0.00192	0.00192
b_{22}	0.00513	0.00515	0.00515	0.00516	0.00516	0.00517	0.00517
$\overline{\text{MSE}}$	0.00704	0.00706	0.00707	0.00707	0.00708	0.00709	0.00709
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00192	0.00192	0.00158	0.00146			
b_{22}	0.00517	0.00517	0.00399	0.00351			
$\overline{\text{MSE}}$	0.00709	0.00709	0.00557	0.00498			

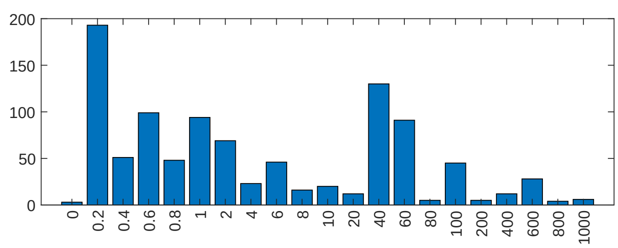


Figure B.39: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

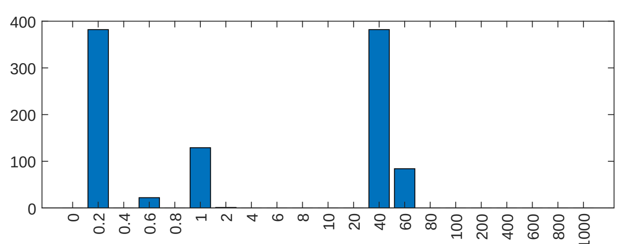


Figure B.40: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3

Case 3: $T = 100$, $\tau = 95$

Table B.21: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	9.88110	0.00954	0.01217	0.01100	0.01090	0.01093	0.01096
b_{22}	11.09923	2.16042	0.47232	0.08182	0.06912	0.06579	0.06439
$\overline{\text{MSE}}$	20.98034	2.16996	0.48449	0.09282	0.08002	0.07673	0.07535
	1	2	4	6	8	10	20
b_{11}	0.01102	0.01111	0.01151	0.01112	0.01113	0.01115	0.01143
b_{22}	0.06367	0.06184	0.06119	0.06078	0.06066	0.06057	0.06053
$\overline{\text{MSE}}$	0.07469	0.07296	0.07269	0.07190	0.07179	0.07172	0.07196
	40	60	80	100	200	400	600
b_{11}	0.01113	0.01110	0.01107	0.01112	0.01105	0.01135	0.01142
b_{22}	0.06032	0.06022	0.06036	0.06026	0.06007	0.06021	0.06018
$\overline{\text{MSE}}$	0.07145	0.07132	0.07143	0.07138	0.07112	0.07157	0.07160
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.01095	0.01103	0.01045	0.01059			
b_{22}	0.06001	0.06001	0.06127	0.06653			
$\overline{\text{MSE}}$	0.07096	0.07103	0.07172	0.07712			

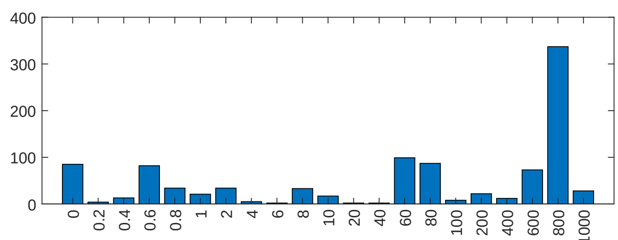


Figure B.41: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

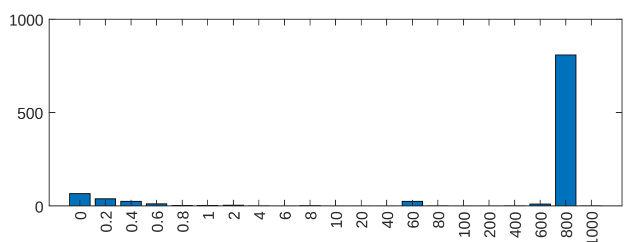


Figure B.42: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3: $T = 100$, $\tau = 90$

Table B.22: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.17185	0.00954	0.01000	0.00920	0.00923	0.00925	0.00926
b_{22}	0.18383	0.17133	0.14769	0.06053	0.05692	0.05604	0.05570
$\overline{\text{MSE}}$	0.35568	0.18087	0.15768	0.06973	0.06615	0.06530	0.06496
	1	2	4	6	8	10	20
b_{11}	0.00928	0.00928	0.00930	0.00929	0.00929	0.00929	0.00929
b_{22}	0.05555	0.05524	0.05518	0.05512	0.05511	0.05510	0.05509
$\overline{\text{MSE}}$	0.06482	0.06452	0.06448	0.06441	0.06440	0.06439	0.06438
	40	60	80	100	200	400	600
b_{11}	0.00929	0.00929	0.00929	0.00929	0.00929	0.00932	0.00950
b_{22}	0.05508	0.05508	0.05508	0.05508	0.05508	0.05513	0.05534
$\overline{\text{MSE}}$	0.06437	0.06437	0.06438	0.06438	0.06437	0.06444	0.06484
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00927	0.00930	0.00932	0.00906			
b_{22}	0.05499	0.05502	0.05317	0.05793			
$\overline{\text{MSE}}$	0.06426	0.06431	0.06249	0.06699			

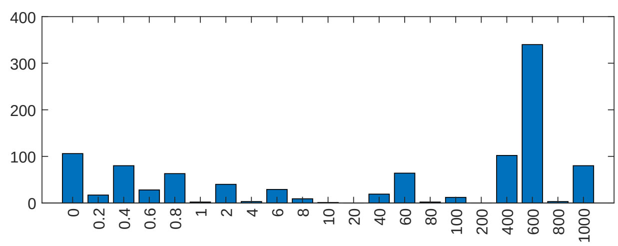


Figure B.43: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

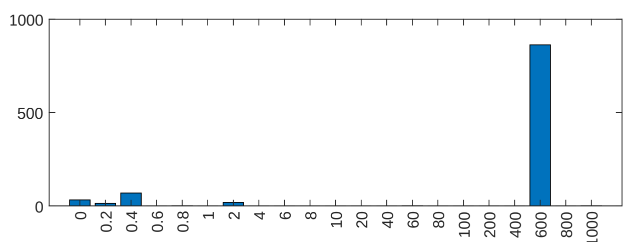


Figure B.44: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3: $T = 100$, $\tau = 80$

Table B.23: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.05993	0.00954	0.00989	0.00843	0.00845	0.00848	0.00851
b_{22}	0.07183	0.07123	0.06992	0.04440	0.04431	0.04484	0.04528
$\overline{\text{MSE}}$	0.13176	0.08077	0.07981	0.05283	0.05276	0.05333	0.05379
	1	2	4	6	8	10	20
b_{11}	0.00852	0.00856	0.00859	0.00859	0.00861	0.00860	0.00861
b_{22}	0.04561	0.04646	0.04699	0.04719	0.04730	0.04734	0.04746
$\overline{\text{MSE}}$	0.05414	0.05502	0.05557	0.05579	0.05591	0.05594	0.05607
	40	60	80	100	200	400	600
b_{11}	0.00861	0.00861	0.00861	0.00861	0.00861	0.00861	0.00861
b_{22}	0.04753	0.04755	0.04756	0.04756	0.04758	0.04759	0.04758
$\overline{\text{MSE}}$	0.05614	0.05616	0.05617	0.05618	0.05619	0.05620	0.05619
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00861	0.00864	0.00827	0.00823			
b_{22}	0.04765	0.04771	0.04481	0.04646			
$\overline{\text{MSE}}$	0.05626	0.05635	0.05308	0.05469			

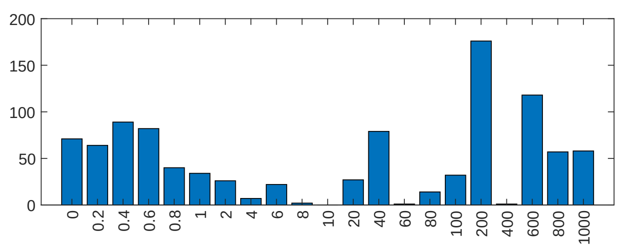


Figure B.45: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

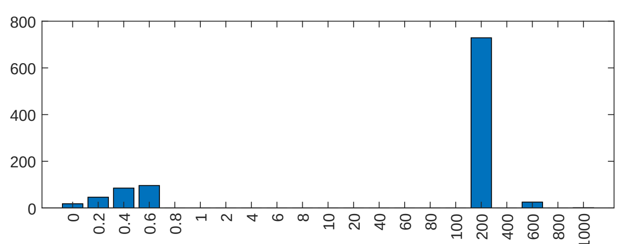


Figure B.46: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3: $T = 400$**,** $\tau = 390$

Table B.24: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.17293	0.00261	0.00263	0.00256	0.00256	0.00255	0.00255
b_{22}	0.18143	0.16863	0.13645	0.04345	0.04337	0.04337	0.04340
$\overline{\text{MSE}}$	0.35436	0.17125	0.13909	0.04601	0.04592	0.04593	0.04596
	1	2	4	6	8	10	20
b_{11}	0.00255	0.00255	0.00255	0.00255	0.00255	0.00255	0.00255
b_{22}	0.04342	0.04345	0.04347	0.04348	0.04348	0.04348	0.04349
$\overline{\text{MSE}}$	0.04598	0.04600	0.04602	0.04603	0.04603	0.04604	0.04604
	40	60	80	100	200	400	600
b_{11}	0.00255	0.00256	0.00255	0.00257	0.00258	0.00258	0.00257
b_{22}	0.04350	0.04348	0.04347	0.04346	0.04352	0.04355	0.04378
$\overline{\text{MSE}}$	0.04605	0.04605	0.04602	0.04603	0.04609	0.04614	0.04634
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00260	0.00261	0.00257	0.00254			
b_{22}	0.04354	0.04381	0.04317	0.04395			
$\overline{\text{MSE}}$	0.04613	0.04642	0.04574	0.04649			

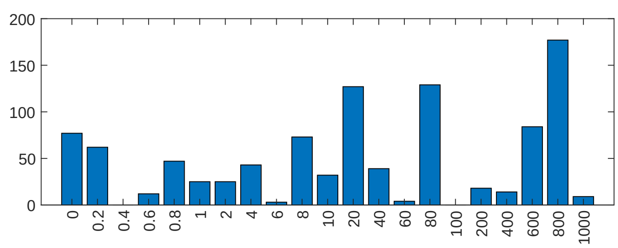


Figure B.47: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

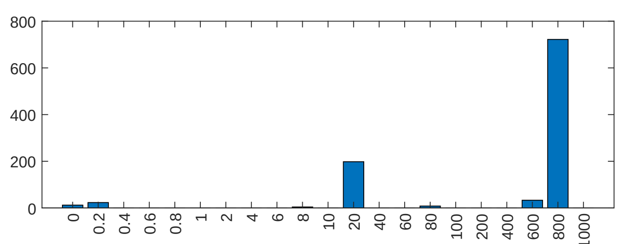


Figure B.48: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3: $T = 400$, $\tau = 380$

Table B.25: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.05937	0.00261	0.00262	0.00251	0.00251	0.00251	0.00251
b_{22}	0.05647	0.05580	0.05584	0.03783	0.03943	0.04008	0.04044
$\overline{\text{MSE}}$	0.11585	0.05841	0.05846	0.04034	0.04194	0.04259	0.04295
	1	2	4	6	8	10	20
b_{11}	0.00251	0.00251	0.00251	0.00251	0.00251	0.00251	0.00251
b_{22}	0.04066	0.04112	0.04136	0.04144	0.04149	0.04151	0.04156
$\overline{\text{MSE}}$	0.04317	0.04363	0.04387	0.04395	0.04400	0.04402	0.04407
	40	60	80	100	200	400	600
b_{11}	0.00251	0.00251	0.00251	0.00251	0.00251	0.00251	0.00252
b_{22}	0.04159	0.04159	0.04160	0.04160	0.04161	0.04163	0.04171
$\overline{\text{MSE}}$	0.04410	0.04410	0.04411	0.04411	0.04412	0.04414	0.04422
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00253	0.00250	0.00250	0.00249			
b_{22}	0.04170	0.04170	0.04005	0.04163			
$\overline{\text{MSE}}$	0.04423	0.04419	0.04255	0.04412			

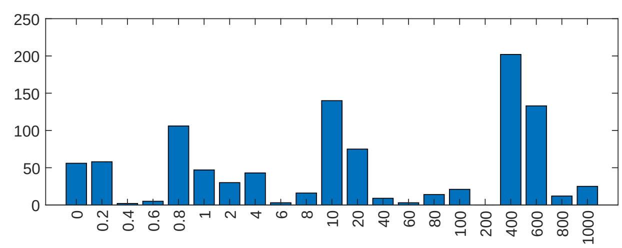


Figure B.49: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

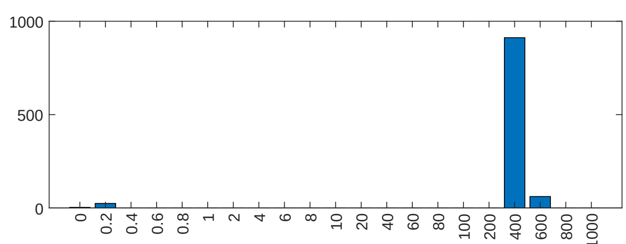


Figure B.50: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3: $T = 400$**,** $\tau = 360$

Table B.26: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.02678	0.00261	0.00262	0.00245	0.00247	0.00248	0.00248
b_{22}	0.02751	0.02741	0.02739	0.02968	0.03303	0.03449	0.03530
$\overline{\text{MSE}}$	0.05429	0.03002	0.03001	0.03213	0.03550	0.03697	0.03779
	1	2	4	6	8	10	20
b_{11}	0.00249	0.00249	0.00250	0.00250	0.00250	0.00250	0.00250
b_{22}	0.03582	0.03691	0.03749	0.03769	0.03779	0.03785	0.03797
$\overline{\text{MSE}}$	0.03830	0.03940	0.03999	0.04019	0.04029	0.04035	0.04047
	40	60	80	100	200	400	600
b_{11}	0.00250	0.00250	0.00250	0.00250	0.00250	0.00251	0.00258
b_{22}	0.03803	0.03805	0.03806	0.03807	0.03808	0.03808	0.03823
$\overline{\text{MSE}}$	0.04053	0.04055	0.04057	0.04057	0.04059	0.04059	0.04081
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00254	0.00252	0.00246	0.00247			
b_{22}	0.03809	0.03822	0.03566	0.03526			
$\overline{\text{MSE}}$	0.04062	0.04074	0.03812	0.03773			

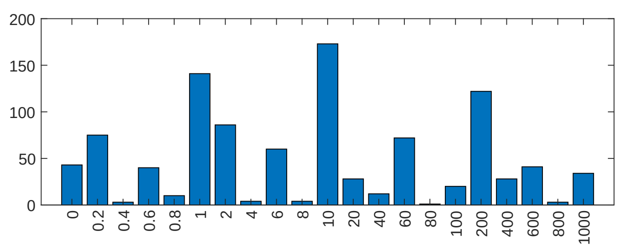


Figure B.51: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

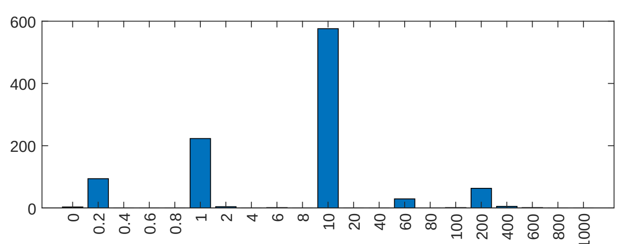


Figure B.52: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3: $T = 900$, $\tau = 885$

Table B.27: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.08552	0.00115	0.00115	0.00115	0.00116	0.00115	0.00116
b_{22}	0.08937	0.08518	0.08354	0.04015	0.04044	0.04058	0.04066
$\overline{\text{MSE}}$	0.17489	0.08633	0.08469	0.04130	0.04160	0.04173	0.04182
	1	2	4	6	8	10	20
b_{11}	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.04072	0.04080	0.04085	0.04086	0.04087	0.04088	0.04089
$\overline{\text{MSE}}$	0.04187	0.04195	0.04200	0.04202	0.04203	0.04203	0.04204
	40	60	80	100	200	400	600
b_{11}	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115	0.00115
b_{22}	0.04089	0.04089	0.04089	0.04089	0.04091	0.04088	0.04090
$\overline{\text{MSE}}$	0.04204	0.04204	0.04205	0.04205	0.04206	0.04203	0.04205
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00116	0.00115	0.00115	0.00116			
b_{22}	0.04089	0.04091	0.04073	0.04086			
$\overline{\text{MSE}}$	0.04205	0.04207	0.04188	0.04201			

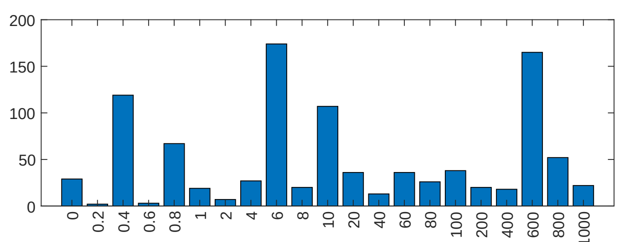


Figure B.53: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

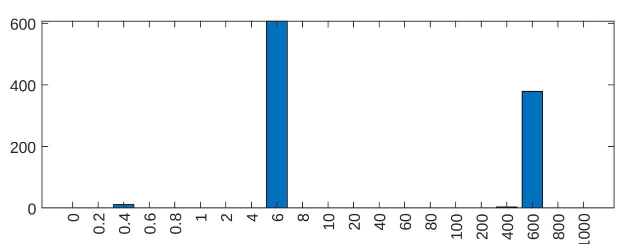


Figure B.54: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3: $T = 900$, $\tau = 870$

Table B.28: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.03486	0.00115	0.00115	0.00116	0.00117	0.00117	0.00117
b_{22}	0.03641	0.03627	0.03614	0.03590	0.03757	0.03818	0.03852
$\overline{\text{MSE}}$	0.07126	0.03742	0.03730	0.03706	0.03873	0.03935	0.03969
	1	2	4	6	8	10	20
b_{11}	0.00117	0.00117	0.00117	0.00117	0.00117	0.00117	0.00117
b_{22}	0.03869	0.03912	0.03931	0.03938	0.03942	0.03944	0.03951
$\overline{\text{MSE}}$	0.03986	0.04028	0.04048	0.04055	0.04059	0.04061	0.04068
	40	60	80	100	200	400	600
b_{11}	0.00117	0.00117	0.00117	0.00117	0.00117	0.00117	0.00117
b_{22}	0.03950	0.03953	0.03959	0.03953	0.03953	0.03962	0.03959
$\overline{\text{MSE}}$	0.04067	0.04070	0.04076	0.04070	0.04070	0.04079	0.04076
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00121	0.00118	0.00117	0.00117			
b_{22}	0.03965	0.03967	0.03878	0.03952			
$\overline{\text{MSE}}$	0.04086	0.04085	0.03994	0.04069			

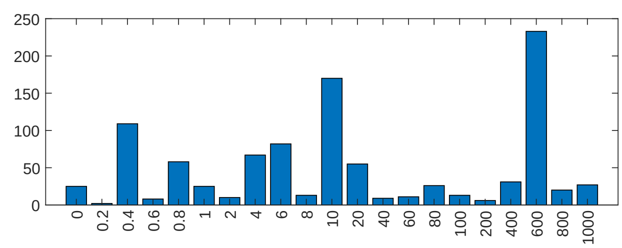


Figure B.55: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

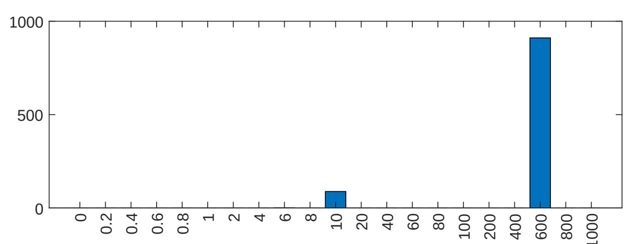


Figure B.56: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3: $T = 900$, $\tau = 840$

Table B.29: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.01578	0.00115	0.00115	0.00119	0.00121	0.00121	0.00122
b_{22}	0.01775	0.01771	0.01809	0.02913	0.03251	0.03385	0.03456
$\overline{\text{MSE}}$	0.03353	0.01886	0.01924	0.03032	0.03372	0.03506	0.03578
	1	2	4	6	8	10	20
b_{11}	0.00122	0.00122	0.00123	0.00123	0.00123	0.00123	0.00123
b_{22}	0.03501	0.03593	0.03641	0.03657	0.03666	0.03671	0.03681
$\overline{\text{MSE}}$	0.03622	0.03715	0.03764	0.03780	0.03788	0.03793	0.03803
	40	60	80	100	200	400	600
b_{11}	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123
b_{22}	0.03686	0.03687	0.03688	0.03689	0.03690	0.03691	0.03692
$\overline{\text{MSE}}$	0.03808	0.03810	0.03811	0.03811	0.03812	0.03813	0.03815
	800	1000	TMLE_1	TMLE_2			
b_{11}	0.00123	0.00123	0.00122	0.00123			
b_{22}	0.03691	0.03691	0.03522	0.03551			
$\overline{\text{MSE}}$	0.03814	0.03814	0.03644	0.03674			

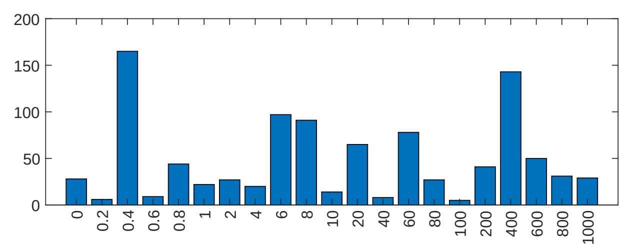


Figure B.57: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

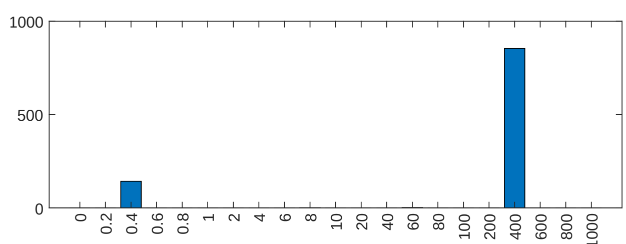


Figure B.58: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Case 3: $T = 900$**,** $\tau = 450$

Table B.30: MSE of point estimates of b_{11} and b_{22}

	1SMLE	2SQMLE	0	0.2	0.4	0.6	0.8
b_{11}	0.00225	0.00115	0.00114	0.00218	0.00310	0.00365	0.00401
b_{22}	0.00199	0.00199	0.00200	0.00591	0.00912	0.01103	0.01227
$\overline{\text{MSE}}$	0.00424	0.00314	0.00314	0.00809	0.01222	0.01468	0.01628
	1	2	4	6	8	10	20
b_{11}	0.00425	0.00484	0.00519	0.00532	0.00539	0.00543	0.00551
b_{22}	0.01314	0.01520	0.01645	0.01690	0.01714	0.01728	0.01757
$\overline{\text{MSE}}$	0.01739	0.02005	0.02164	0.02222	0.02252	0.02270	0.02308
	40	60	80	100	200	400	600
b_{11}	0.00555	0.00556	0.00557	0.00558	0.00558	0.00559	0.00559
b_{22}	0.01772	0.01777	0.01779	0.01781	0.01784	0.01785	0.01786
$\overline{\text{MSE}}$	0.02327	0.02333	0.02336	0.02338	0.02342	0.02344	0.02345
	800	1000	TMLE$_1$	TMLE$_2$			
b_{11}	0.00559	0.00559	0.00325	0.00241			
b_{22}	0.01786	0.01787	0.01009	0.00666			
$\overline{\text{MSE}}$	0.02345	0.02346	0.01334	0.00906			

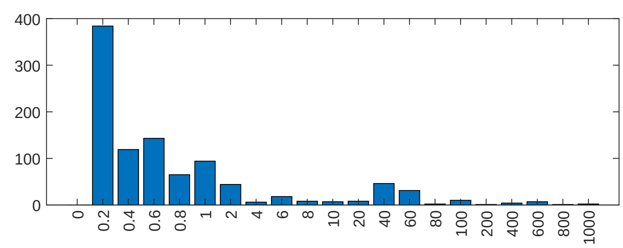


Figure B.59: The bar chart of λ determined by the fixed-design wild bootstrap for 1000 simulated datasets

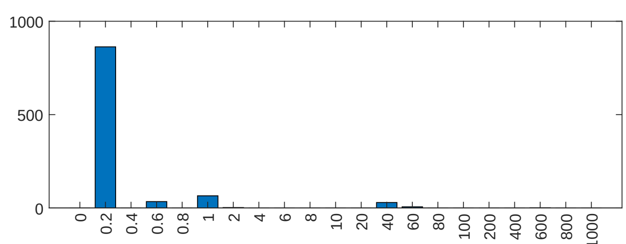
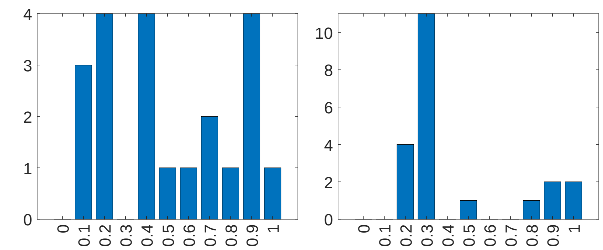


Figure B.60: The bar chart of λ determined by the TMLE bootstrap for 1000 simulated datasets

Empirical application

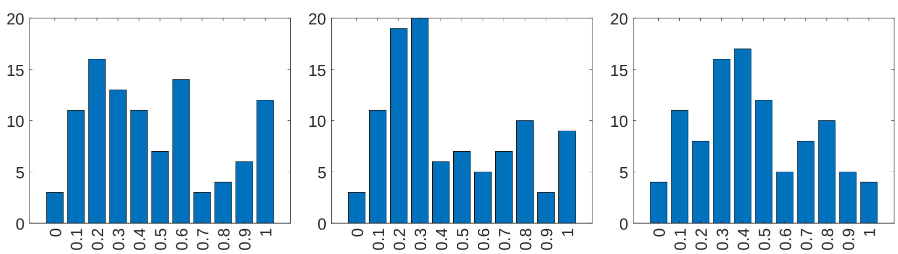
Incremental window



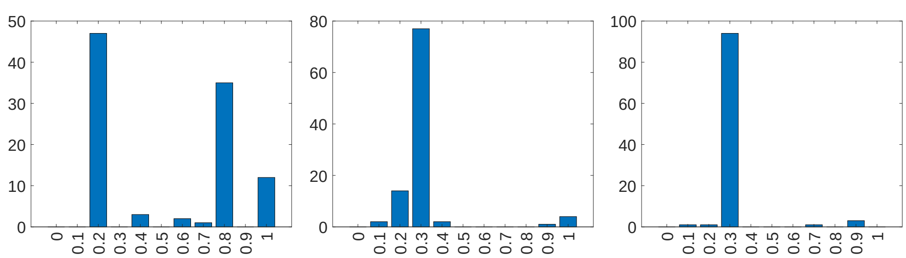
(a) Fix-design wild bootstrap (b) TMLE bootstrap

Figure B.61: The bar charts of λ determined by two bootstrap proceures for 21 incremental windows

Rolling window



(a) $T - \tau = 10$ (b) $T - \tau = 20$ (c) $T - \tau = 30$



(d) $T - \tau = 10$ (e) $T - \tau = 20$ (f) $T - \tau = 30$

Figure B.62: Bar charts of λ determined by the fixed-design bootstrap (a-c) and TMLE bootstrap (d-f) for 100 rolling windows