Non-likelihood based methods for the estimation and inference in econometric models

Hyeonseok Park

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Yanqin Fan, Chair

Chang-Jin Kim

Fang Han

Jing Tao

Program Authorized to Offer Degree:

Economics

University of Washington

Abstract

Non-likelihood based methods for the estimation and inference in econometric models

Hyeonseok Park

Chair of the Supervisory Committee:
Professor Yanqin Fan
Department of Economics

This dissertation aims to address estimation and inference in econometric models when the likelihood-based estimations may not be applicable.

Chapter 1 proposes simple, robust estimation and inference methods for the transition matrix of a high-dimensional semiparametric Gaussian copula vector autoregressive (VAR) process with unknown, possibly fat-tailed marginal distributions. In this model, the observable variable is a monotonic transformation of the latent variable, and the latent variable follows the Gaussian VAR process. Since the marginal distribution is unknown, conventional approaches that use the sample variance and auto-covariances such as OLS are not applicable. This chapter circumvents the problem by constructing the rank estimators of the variance and auto-covariance matrices of the latent process. This chapter derives rates of convergence of the estimator based on which we develop de-biased inference for Granger causality.

Chapter 2 develops a simple, robust method for the estimation and inference in structural models using sliced distances between empirical and model-induced quantile functions (distribution functions). In state-space models, observable variables could be driven by fewer latent variables. This causes stochastic singularity, and the likelihood function does not exist. For the models with parameter-dependent support such as in the one-sided and two-sided models, the likelihood function may not be smooth depending on the parameter. Therefore,

the asymptotic theory for MLE may not be robust to the parameter. We handle these issues using sliced distances since they are well-defined for stochastic singular models and models with parameter-dependent support. In contrast to MLE and likelihood-based inference, we show that under mild regularity conditions, our estimator is asymptotically normally distributed, leading to simple inference regardless of the possible presence of "stochastic singularity" and parameter-dependent supports. Furthermore, our estimator applies to generative models with intractable likelihood functions but from which one can easily draw synthetic samples. We provide simulation results based on a stochastic singular state-space model, a term structure model, and an auction model.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

The completion of my dissertation would not have been possible without the support of many people, whom I would like to express my gratitude.

I would like to express my deepest gratitude to my supervisor, Professor Yanqin Fan for her continuous support, invaluable advice, and patience during my Ph.D. life. She was the light of my life navigating through my research journey. I would like to extend my sincere thank to Professor Fang Han for his invaluable advice, guidance, and patience. I would also like to express my deepest appreciation to Professor Chang-Jin Kim for the continuous support and advice throughout my life. I would like to extend my sincere gratitude to Professor Jing Tao and Professor Yen-Chi Chen for their critical advice and support. I am also grateful to Simon Reeve-Parker, Heidi Hannah, and Kim H Lee for the administrative support.

Last but not least, I would like to express my deepest gratitude to my parent for their unconditional love and support.

DEDICATION

to my parents.

Chapter 1

ESTIMATION AND INFERENCE IN HIGH-DIMENSIONAL SEMIPARAMETRIC GAUSSIAN COPULA VECTOR AUTOREGRESSION [1]

1.1 Introduction

Estimation and inference in high-dimensional stationary Gaussian vector autoregressive (VAR) models have been considered in recent works such as Negahban and Wainwright (2011), Han et al. (2015), and Basu and Michailidis (2015) under the assumption that the underlying process is observable. In many applications in economics and finance, however, the observable process may be fat-tailed and thus does not follow a Gaussian VAR process. This paper proposes a high-dimensional stationary Gaussian copula VAR process with unknown, possibly fat-tailed marginal distributions. We develop simple methods for the estimation and inference of the transition matrices characterizing the dependence of the Gaussian copula VAR process.

To simplify notation and exposition, we focus on Gaussian copula VAR(1) and refer to it as Gaussian copula VAR in the paper. Extensions to Gaussian copula VAR(p) process for any finite and fixed p are provided in Appendix A.2. We construct the Gaussian copula VAR process by first transforming the observed process by an increasing but unknown transformation and then modeling the transformed latent process by a high-dimensional stationary Gaussian VAR model. In detail, we let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$, with $\boldsymbol{X}_t \in \mathbb{R}^d$, denote an *observable process* such that

$$\boldsymbol{X}_t := \boldsymbol{f}(\boldsymbol{Z}_t) = (f_1(Z_{t,1}), \dots, f_d(Z_{t,d}))^\top,$$

[1]THIS CHAPTER IS A JOINT WORK WITH YANQIN FAN AND FANG HAN FROM THE UNIVERSITY OF WASHINGTON.

where $\boldsymbol{f} := \{f_1, \ldots, f_d\}$ consists of univariate strictly increasing *unknown* functions and $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$, with $\mathbb{Z}$ standing for the set of integers and $\boldsymbol{Z}_t$ in the d-dimensional real space $\mathbb{R}^d$, denote a high-dimensional *latent process* following a stationary Gaussian VAR model with transition matrix $\mathbf{A}$:

$$\boldsymbol{Z}_t = \mathbf{A}\boldsymbol{Z}_{t-1} + \boldsymbol{E}_t,\ \ \boldsymbol{E}_t \overset{i.i.d.}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_E). \tag{1.1}$$

Model (1.1) and the relation $\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{Z}_t)$ imply that the observable process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ follows a stationary semiparametric Gaussian copula VAR model with unknown marginal distributions. Unlike the Gaussian VAR model, the Gaussian copula VAR model does not impose any finite moment restrictions on the process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ and hence is more suitable for modeling potentially fat-tailed time series. The dimension $d \equiv d_n$, the parameters $\mathbf{A}$ and $\boldsymbol{\Sigma}_0 := \mathrm{Var}(\boldsymbol{Z}_t)$ in (1.1) are all allowed to change with the sample size n and d_n may even be larger than n. The scenario thus falls into the application regime of a high-dimensional triangular array framework, which has been explicitly discussed recently in Han and Wu (2019). For notational compactness, we omit the dependence of all model parameters on the sample size n in the rest of this paper.

Following Negahban and Wainwright (2011), Han et al. (2015), and Basu and Michailidis (2015), we impose sparsity constraint on the transition matrix $\mathbf{A}$ and aim at estimating, and recovering the sparsity pattern of $\mathbf{A}$ in model (1.1) based on observations $\{\boldsymbol{X}_t\}_{t=1}^n$. Existing methods in Negahban and Wainwright (2011), Han et al. (2015), and Basu and Michailidis (2015) are inapplicable because we don't have access to the latent process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$. We propose a new rank-based estimator of $\mathbf{A}$ based on $\{\boldsymbol{X}_t\}_{t=1}^n$, bypassing the unknown transformations $\{f_1, \ldots, f_d\}$ without estimating them. Our estimator is formulated in two steps. In the first step, we construct rank-based estimators of the variance and first autocovariance matrices of $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ from $\{\boldsymbol{X}_t\}_{t=1}^n$ avoiding the estimation of unknown transformations $\{f_1, \ldots, f_d\}$. This is made possible by the nature of rank estimators, which are invariant to increasing transformations, and the relation between the (Pearson) covariance of a bivariate Gaussian distribution and its Kendall's tau. In the second step, we construct a regularized estimator of $\mathbf{A}$ from rank estimators of the variance and first autocovariance matrices of $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ using the relation between $\mathbf{A}$ and population variance-autocovariance matrices, known as the Yule-Walker equation.

To justify the proposed estimator, we first derive its rate of convergence and show its consistency to the truth in a high-dimensional asymptotic regime. Under the adopted assumptions, the latent process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ and the observable process $\{\boldsymbol{X}_t\}_{t=1}^n$ are both α-mixing with geometric decaying rates. This paves the way for theoretical studies. However, compared with regression-based estimators of the variance and first autocovariance matrices in Negahban and Wainwright (2011), Han et al. (2015), and Basu and Michailidis (2015), establishing the asymptotic properties of our rank estimators appears to be much more challenging due to their dependence on large random matrices with elements given by U-statistics of α-mixing processes. The analysis is now possible thanks to a recent progress made in Shen et al. (2019), who laid out the path to establishing exponential inequalities for (degenerate) U-statistics of α-mixing processes.

To handle the regularization bias inherent in the estimator of $\mathbf{A}$, we adopt a recently established de-biasing inference framework (Neykov et al., 2018a) to develop tests for non-Granger causality in model (1.1), which is equivalent to testing non-zeroness of entries of $\mathbf{A}$ (Lütkepohl, 2005). This is via combining the de-biasing strategy with a circular-block-bootstrap (CBB) based variance estimator. The latter has been studied in, e.g, Dehling and Wendler (2010) and Fan et al. (2016), although the analysis here is more demanding due to the complicated form of the target of interest. In addition, when the presence of Granger causality is detected, a sign-consistent estimator of Granger causality is immediate following the strategy in Qiu et al. (2015). Numerical results are presented to illustrate the finite sample behavior of the estimator of $\mathbf{A}$ and also the proposed tests.

This paper contributes to the growing literature on modelling the possibly nonlinear temporal dependence (Hsing and Wu, 2004; Beare, 2010; Patton, 2012, 2013; Fan and Patton, 2014; Wang and Xia, 2015), and in particular, the literature on copula-based nonlinear time series models including both the univariate Gaussian copula Markov chains introduced in Chen and Fan (2006b) and studied further in Chen et al. (2009a) and Chen et al. (2009b), and copula-based multivariate time series models studied in Chen and Fan (2006a), Lee and Long (2009), Min and Czado (2014), Rémillard et al. (2012), and Chen et al. (2018). In contrast to the current paper, the model of $\boldsymbol{X}_t$ (and hence also the dimension d) in all these works is assumed to be fixed, while the high-dimensional triangular array framework in this paper induces more challenges.

The rest of this paper is organized as follows. Section 1.2 specifies the latent model and the observable process, introduces our estimator, and establishes its rates of convergence. Section 1.3 develops two de-biasing tests for non-Granger causality and introduces a sign consistent estimator of Granger causality. Section 1.4 presents numerical results. The last section concludes and discusses extensions. Technical proofs are relegated to Appendix A.

We now introduce notations used in this paper. For any positive integer d, let $[d] := \{1, 2, \ldots, d\}$. Let $\boldsymbol{v} = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$ and $\mathbf{M} = [M_{jk}] \in \mathbb{R}^{d \times d}$ be a vector and a matrix of interest. We denote $\boldsymbol{v}_I$ and $\mathbf{M}_{I,J}$ to be the subvector of $\boldsymbol{v}$ and submatrix of $\mathbf{M}$ whose entries are indexed by a set $I \subset [d]$ and sets $I, J \subset [d]$. We denote $\mathbf{M}_{I*}$ and $\mathbf{M}_{*J}$ to be submatrices of $\mathbf{M}$ whose rows are indexed by I and whose columns are indexed by J. For any $p \in (0, \infty]$, we define $\|\boldsymbol{v}\|_p$ to be the vector L_p norm of $\boldsymbol{v}$. Let $\|\mathbf{M}\|_\infty := \max_j \sum_{k=1}^d |M_{jk}|$ be the matrix operator ∞-norm, $\|\mathbf{M}\|_2$ be the matrix spectral norm, and $\|\mathbf{M}\|_{\max}$ be the elementwise maximum norm. For a symmetric matrix $\mathbf{M}$, let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote the largest and smallest eigenvalues of $\mathbf{M}$. For any univariate function $f(\cdot) : \mathbb{R} \to \mathbb{R}$, define $f(\boldsymbol{v}) = (f(v_1), \ldots, f(v_d))^\top$ and $f(\mathbf{M}) := [f(M_{jk})]$. Let $\mathrm{diag}(\mathbf{M})$ be the diagonal matrix with diagonals $M_{11}, \ldots, M_{dd}$. Let $\mathbf{I}_d$ be the d-dimensional identity matrix. For a random vector $\boldsymbol{X}_t \in \mathbb{R}^d$, let $X_{t,j}$ denote the j-th element of $\boldsymbol{X}_t$. For any $x \in \mathbb{R}$, denote $\mathrm{sign}(x) = \mathbb{1}(x > 0) - \mathbb{1}(x < 0)$, with $\mathbb{1}(\cdot)$ standing for the indicator function. For any two real sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists an absolute positive constant C such that $|a_n| \le C|b_n|$ for any large enough n. We write $a_n \asymp b_n$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$ hold. We write $a_n = o(b_n)$ if for any absolute positive constant C, we have $|a_n| \le C|b_n|$ for any large enough n. We write $a_n = O_{\mathbb{P}}(b_n)$ and $a_n = o_{\mathbb{P}}(b_n)$ if $a_n = O(b_n)$ and $a_n = o(b_n)$ hold stochastically.

1.2 Estimation Method and Asymptotic Properties

1.2.1 The Model, Granger Causality, and α-mixing Property

The observable process and the sample information are characterized by Assumption S below.

Assumption S (Observable Process and Sample Information). (i) Let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ denote the observable process satisfying $\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{Z}_t)$, where $\boldsymbol{f} := \{f_1, \ldots, f_d\}$ consists of univariate strictly increasing *unknown* functions; (ii) Let $\{\boldsymbol{X}_t : t = 1, 2, \ldots, n\}$ denote a length-n segment of $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ and constitute the data accessible.

The latent process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is characterized by (1.1) and Assumption M below.

Assumption M (Latent Process). (i) $\operatorname{diag}(\boldsymbol{\Sigma}_0) = \mathbf{I}_d$, i.e., $\boldsymbol{\Sigma}_0$ is a correlation matrix; (ii) $\{\boldsymbol{E}_t\}_{t\in\mathbb{Z}} \overset{\text{i.i.d}}{\sim} N(0, \boldsymbol{\Sigma}_{\boldsymbol{E}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{E}}$ satisfies: $0 < c_{\boldsymbol{E}} < \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{E}}) \le \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{E}}) < C_{\boldsymbol{E}} < \infty$ for some absolute constants $c_{\boldsymbol{E}}$ and $C_{\boldsymbol{E}}$; (iii) The transition matrix $\mathbf{A}$ satisfies: $\|\mathbf{A}\|_2 \le C_{\mathbf{A}} < 1$ for some absolute constant $C_{\mathbf{A}}$ and the following sparsity constraint:

$$\mathbf{A} \in \mathcal{M}(s, M) \;\; \text{with } \mathcal{M}(s, M) := \left\{ \mathbf{M} \in \mathbb{R}^{d\times d} : \max_{1\le j\le d} \sum_{k=1}^{d} \mathbb{1}(M_{jk} \neq 0) \le s, \|\mathbf{M}\|_\infty \le M \right\}. \quad (1.2)$$

Here s is a positive integer and M is a positive constant, both of which may depend on d and hence also on n.

Assumption M(i) is a normalization condition ensuring identifiability of $\mathbf{A}$ and $\boldsymbol{f}$. It is not essential and is removable if only the sparsity pattern of $\mathbf{A}$ is of interest. Although it is assumed that $C_{\mathbf{A}} < 1$, there is no such restriction on M, which is allowed to increase to infinity along with d and n.

Under Assumption M, it holds that

$$\operatorname{Var}(\boldsymbol{Z}_t) = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{\boldsymbol{E}} + \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{E}}\mathbf{A}^\top + \mathbf{A}^2\boldsymbol{\Sigma}_{\boldsymbol{E}}(\mathbf{A}^\top)^2 + \cdots,$$

which is also positive definite and the eigenvalues of $\boldsymbol{\Sigma}_0$ are bounded away from zero and infinity by absolute constants. We will show in Proposition 1.2.2 below that Assumption M yields that the latent process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is geometrically α-mixing, which implies weak temporal dependence even in high dimensions and is crucial in justifying our proposed methods.

Assumptions M and S(i) define a stationary semiparametric Gaussian copula VAR model for the observable process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ with marginal distribution functions $\Phi(f_1^{-1}),\ldots,\Phi(f_d^{-1})$, where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Although the Gaussian process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ has finite moments of all orders, the observable process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ may not have any finite moments and is thus more suitable for modeling financial time series than a Gaussian VAR model.

For the Gaussian VAR process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$, it is a well-known fact that the sequence $\{Z_{t,k}\}_{t\in\mathbb{Z}}$ *Granger causes* $\{Z_{t,j}\}_{t\in\mathbb{Z}}$ if and only if the (j,k)-th entry of the transition matrix is nonzero (cf. Corollary 2.2.1 in Lütkepohl (2005)). The following proposition, which is a direct consequence of Theorem 6.2 in Qiu et al. (2015), shows that this fact also applies to the observable process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$.

Proposition 1.2.1. *Under Model (1.1) with Assumption M(ii) and (iii), and Assumption S(i), we have that the sequence $\{X_{t,k}\}_{t\in\mathbb{Z}}$ Granger causes $\{X_{t,j}\}_{t\in\mathbb{Z}}$ if and only if $A_{jk}\neq 0$.*

We then proceed to characterize an important property of the observable process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$, its weak (temporal) dependence. To this end, let's first introduce the α-mixing measure of dependence. Given a probability space $(\Omega,\mathcal{F},\mathbb{P})$, for any two σ-algebras $\mathcal{A}$ and $\mathcal{B}$ in $\mathcal{F}$, define the α-mixing measure of dependence between $\mathcal{A}$ and $\mathcal{B}$ as

$$\alpha(\mathcal{A},\mathcal{B}) := \sup_{A\in\mathcal{A},B\in\mathcal{B}} |\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)|.$$

For any measurable process denoted as $\{\boldsymbol{W}_t\}_{t\in\mathbb{Z}}$, we define its α-mixing coefficient as

$$\alpha(n;\{\boldsymbol{W}_t\}_{t\in\mathbb{Z}}) := \sup_{t\in\mathbb{Z}} \alpha(\mathcal{G}_{-\infty}^{t},\mathcal{G}_{t+n}^{\infty}), \quad \text{for each } n=1,2,\ldots,$$

where for arbitrary $j\in\mathbb{Z}$, $\mathcal{G}_{-\infty}^{j} := \sigma(\boldsymbol{W}_t, t\le j)$ and $\mathcal{G}_{j}^{\infty} := \sigma(\boldsymbol{W}_t, t\ge j)$ stand for the sigma fields generated by $\{\boldsymbol{W}_t\}_{t\le j}$ and $\{\boldsymbol{W}_t\}_{t\ge j}$ respectively.

With these concepts introduced, we are now ready to present the weak dependence property of $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$, i.e., its α-mixing coefficient is uniformly geometrically decaying to 0, which holds even if the dimension d explodes as $n\to\infty$. This proposition hinges on the Gaussian assumption and will be heavily used in building up our theory in the follow-up sections.

Proposition 1.2.2 (Kolmogorov and Rozanov (1960), explicitly stated as Theorem 3.1 in Han and Wu (2019))**.** *Suppose that the latent process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ follows the latent Gaussian VAR model (1.1) with Assumption M and the observable process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ satisfies Assumption S(i). Then both $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ and $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ are geometrically α-mixing such that*

$$\alpha(n;\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}) = \alpha(n;\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}) \le \kappa_1 \exp(-\kappa_2 n), \quad \text{for all } n \ge 1.$$

Here $\kappa_1 = \sqrt{\frac{C_{\boldsymbol{E}}}{c_{\boldsymbol{E}}(1-C_{\mathbf{A}}^2)}}$ and $\kappa_2 = -\ln(C_{\mathbf{A}})$ are two absolute positive constants, where $C_{\boldsymbol{E}}$, $c_{\boldsymbol{E}}$, and $C_{\mathbf{A}}$ are defined in Assumption M (ii) and (iii).

1.2.2 Estimators

The celebrated Yule-Walker formula yields that $\mathbf{A} = \boldsymbol{\Sigma}_1^\top \boldsymbol{\Sigma}_0^{-1}$, where $\boldsymbol{\Sigma}_1 := \mathbb{E}[\boldsymbol{Z}_t \boldsymbol{Z}_{t+1}^\top]$. Notice that $(\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top)^\top$ has the following covariance matrix,

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_1^\top & \boldsymbol{\Sigma}_0 \end{pmatrix}.$$

For estimating $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, it is then sufficient to estimate $\boldsymbol{\Omega}$. It is well-known – cf. Kruskal (1958, p. 823) – that

$$\boldsymbol{\Omega} = \sin\left(\frac{\pi}{2} \cdot \underbrace{\mathbb{E}\left[\operatorname{sign}\begin{pmatrix} \boldsymbol{Z}_1 - \widetilde{\boldsymbol{Z}}_1 \\ \boldsymbol{Z}_2 - \widetilde{\boldsymbol{Z}}_2 \end{pmatrix} \operatorname{sign}\begin{pmatrix} \boldsymbol{Z}_1 - \widetilde{\boldsymbol{Z}}_1 \\ \boldsymbol{Z}_2 - \widetilde{\boldsymbol{Z}}_2 \end{pmatrix}^\top \right]}_{\mathbf{T}} \right), \tag{1.3}$$

where $(\widetilde{\boldsymbol{Z}}_1^\top, \widetilde{\boldsymbol{Z}}_2^\top)^\top$ is an independent copy of $(\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top)^\top$. Since $\boldsymbol{f}$ is a monotonic transformation, we have

$$\begin{aligned}
\mathbf{T} &= \mathbb{E}\left[\operatorname{sign}\begin{pmatrix}\boldsymbol{Z}_1 - \widetilde{\boldsymbol{Z}}_1\\ \boldsymbol{Z}_2 - \widetilde{\boldsymbol{Z}}_2\end{pmatrix}\operatorname{sign}\begin{pmatrix}\boldsymbol{Z}_1 - \widetilde{\boldsymbol{Z}}_1\\ \boldsymbol{Z}_2 - \widetilde{\boldsymbol{Z}}_2\end{pmatrix}^\top\right]\\
&= \mathbb{E}\left[\operatorname{sign}\begin{pmatrix}\boldsymbol{f}(\boldsymbol{Z}_1) - \boldsymbol{f}(\widetilde{\boldsymbol{Z}}_1)\\ \boldsymbol{f}(\boldsymbol{Z}_2) - \boldsymbol{f}(\widetilde{\boldsymbol{Z}}_2)\end{pmatrix}\operatorname{sign}\begin{pmatrix}\boldsymbol{f}(\boldsymbol{Z}_1) - \boldsymbol{f}(\widetilde{\boldsymbol{Z}}_1)\\ \boldsymbol{f}(\boldsymbol{Z}_2) - \boldsymbol{f}(\widetilde{\boldsymbol{Z}}_2)\end{pmatrix}^\top\right]\\
&= \mathbb{E}\left[\operatorname{sign}\begin{pmatrix}\boldsymbol{X}_1 - \widetilde{\boldsymbol{X}}_1\\ \boldsymbol{X}_2 - \widetilde{\boldsymbol{X}}_2\end{pmatrix}\operatorname{sign}\begin{pmatrix}\boldsymbol{X}_1 - \widetilde{\boldsymbol{X}}_1\\ \boldsymbol{X}_2 - \widetilde{\boldsymbol{X}}_2\end{pmatrix}^\top\right] \qquad (1.4)
\end{aligned}$$

where $(\widetilde{\boldsymbol{X}}_1^\top, \widetilde{\boldsymbol{X}}_2^\top)^\top = (\boldsymbol{f}(\widetilde{\boldsymbol{Z}}_1)^\top, \boldsymbol{f}(\widetilde{\boldsymbol{Z}}_2^\top))^\top$.

The Yule-Walker formula and (1.4) suggest our two-step approach for estimating $\mathbf{A}$, as presented below.

Step 1. We construct "consistent estimators" of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. We propose the following estimator of $\boldsymbol{\Omega}$ based on (1.3) and (1.4):

$$\widehat{\boldsymbol{\Omega}} = \sin\left(\frac{\pi}{2}\widehat{\mathbf{T}}\right),$$

where

$$\widehat{\mathbf{T}} = \frac{2}{(n-1)(n-2)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n-1}\left[\operatorname{sign}\begin{pmatrix}\boldsymbol{X}_i - \boldsymbol{X}_j\\ \boldsymbol{X}_{i+1} - \boldsymbol{X}_{j+1}\end{pmatrix}\operatorname{sign}\begin{pmatrix}\boldsymbol{X}_i - \boldsymbol{X}_j\\ \boldsymbol{X}_{i+1} - \boldsymbol{X}_{j+1}\end{pmatrix}^\top\right] \qquad (1.5)$$

is a high-dimensional matrix of entries formulated as second-order U statistics. We then denote

$$\widehat{\boldsymbol{\Sigma}}_0 = \widehat{\boldsymbol{\Omega}}_{[d],[d]} \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}}_1 = \widehat{\boldsymbol{\Omega}}_{[d],d+[d]}$$

to be the estimators of variance and auto-covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$.

Step 2. Using the formula $\mathbf{A} = \boldsymbol{\Sigma}_1^\top\boldsymbol{\Sigma}_0^{-1}$, we propose to estimate $\mathbf{A}$ by the following Dantzig-type linear programming estimator,

$$\widehat{\mathbf{A}} := \underset{\mathbf{M}\in\mathbb{R}^{d\times d}}{\operatorname{argmin}}\sum_{j=1}^{d}\sum_{k=1}^{d}|M_{jk}| \quad \text{s.t. } \|\widehat{\boldsymbol{\Sigma}}_0\mathbf{M}^\top - \widehat{\boldsymbol{\Sigma}}_1\|_{\max} \le \lambda. \qquad (1.6)$$

Here λ is a tuning parameter to encourage sparsity of the output. We refer to Section 3 in Han et al. (2015) for the numerical implementation of this linear programming estimator.

We note that an alternative approach to constructing estimators of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ is to apply the likelihood or normal-score based estimator of a high dimensional variance matrix proposed in Liu et al. (2009). However, for random samples, Liu et al. (2012) show that in high-dimensional setting, this estimator does not achieve optimal convergence rate, but rank based estimators do, see Section 2.3 in Liu et al. (2012)). From a more practical point of view, the rank based estimators are also known to be more robust to outliers/noises than likelihood based estimators (cf. the simulations conducted in Section 5.2 in Liu et al. (2012)).

1.2.3 Rates of Convergence

Estimating matrices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_0$ via the Kendall's tau matrix $\widehat{\mathbf{T}}$ in (1.5) allows us to estimate $\mathbf{A}$ without estimating the unknown functions $\{f_1, \ldots, f_d\}$. However, in high dimensions, the asymptotic properties of $\widehat{\mathbf{T}}$ are much harder to establish than the sample covariance matrix used in Negahban and Wainwright (2011), Han et al. (2015), Basu and Michailidis (2015), and more recently in Hecq et al. (2019). Although the asymptotic properties of $\widehat{\mathbf{T}}$ in a finite-dimensional time series model have been well understood (see, for example, Dehling et al. (2017) and the references therein), to our knowledge, Fan et al. (2016) is the only paper in the literature exploring $\widehat{\mathbf{T}}$ in a high-dimensional time series triangular array set-up. Their analysis is, however, based on the assumption that the underlying process is ϕ-mixing. Assuming ϕ-mixing in our case is unrealistic, as a ϕ-mixing Gaussian VAR process is m-dependent (cf. Proposition 1 in Section 2.1 of Doukhan (1994)). Instead, our analysis is built upon Proposition 1.2.2 that a Gaussian VAR process is α-mixing under conditions therein, and uses heavily the newly developed probability tools in Shen et al. (2019) for α-mixing processes.

In addition to Assumptions M and S, we impose the following scaling condition on (d, n).

Assumption E $\log^5(d)/n = O(1)$ so that there exists an absolute positive constant K_0

satisfying $\log^5(d) \leq K_0 n$ for all large enough n.

Assumption E is very mild, allowing the dimension d to be even exponentially larger than the sample size n. It is added mainly for the purpose of presentational simplicity. With Assumptions M, S, and E, we are now ready to introduce our first theorem, which is nonasymptotic and characterizes the deviation probability of $\hat{\mathbf{T}}$ from $\mathbf{T}$.

Theorem 1.2.1 (Stochastic bound for $\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\max}$)**.** *Suppose Assumptions M, S, and E hold. Then, for all sufficiently large n, we have*

$$\mathbb{P}\left(\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\max} \geq 2z + \frac{1}{n-2}\right) \leq 16e^{-4}d^{-2} + K_1\{(n-1)d\}^{-2},$$

for any z such that

$$\begin{aligned} z \geq\, & Q\sqrt{\frac{\log(ed)}{n-1}} + \frac{C\log^2(n-1)}{(n-1)} + \frac{K_2}{2\sqrt{(n-1)\log(ed)}} \\ & + \frac{C}{2}\sqrt{\frac{\log(ed)}{n-1}} + \frac{2C\sqrt{K_1}}{(n-1)^2d^2} + \frac{2C^2\sqrt{K_1}\log^2(n-1)}{(n-1)^2d^2} + \frac{C\sqrt{K_1}K_2}{d^2(n-1)\sqrt{(n-1)\log(ed)}}, \end{aligned} \tag{1.7}$$

in which Q only depends on κ_1, κ_1 and K_0 introduced in Proposition 1.2.2 and Assumption E, respectively; K_1 only depends on $C_{\mathbf{A}}$, $c_{\boldsymbol{E}}$ and $C_{\boldsymbol{E}}$; K_2 only depends on K_0; C is an absolute positive constant. In particular, we have

$$\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{\log(ed)}{n}}\right).$$

Remark 1.2.1. Because $\sin(x)$ is a Lipschitz continuous function, Theorem 1.2.1 implies that, with probability no smaller than $1 - \epsilon_0$, where

$$\epsilon_0 := 16e^{-4}d^{-2} + K_1\{(n-1)d\}^{-2}, \tag{1.8}$$

we have

$$\|\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\max} \leq \frac{\pi}{2}\left[2z + \frac{1}{n-1}\right] \quad \text{and} \quad \|\widehat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\|_{\max} \leq \frac{\pi}{2}\left[2z + \frac{1}{n-1}\right]$$

in which z is chosen to satisfy (1.7). Also, we have

$$\|\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{\log(ed)}{n}}\right) \quad \text{and} \quad \|\widehat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{\log(ed)}{n}}\right).$$

Exploiting the proof of Theorem 1 in Han et al. (2015) with Lemmas 1 and 2 therein replaced by the corresponding bounds in Remark 1.2.1, we immediately have the following theorem that quantifies the deviation of $\hat{\mathbf{A}}$ from $\mathbf{A}$.

Theorem 1.2.2 (Stochastic bounds for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\max}$ and $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\infty}$)**.** *Suppose Assumptions M, S, and E hold. Let the tuning parameter λ in equation (1.6) be chosen such that*

$$\lambda \geq \frac{\pi}{2}(M+1)\left[2z + \frac{1}{n-2}\right],$$

where z is defined in equation (1.7). Then, for all sufficiently large n, it holds that

$$\mathbb{P}(\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\max} \geq 2\|\mathbf{\Sigma}_0^{-1}\|_{\infty}\lambda) \leq 16e^{-4}d^{-2} + K_1\{(n-1)d\}^{-2} \quad \textit{and}$$
$$\mathbb{P}(\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\infty} \geq 4s\|\mathbf{\Sigma}_0^{-1}\|_{\infty}\lambda) \leq 16e^{-4}d^{-2} + K_1\{(n-1)d\}^{-2}.$$

In particular, when $\lambda = CM\sqrt{\frac{\log(ed)}{n-1}}$ for some large enough absolute constant C, we obtain that

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\max} = O_{\mathbb{P}}\left(\|\mathbf{\Sigma}_0^{-1}\|_{\infty}M\sqrt{\frac{\log(ed)}{n}}\right) \quad \textit{and} \quad \|\widehat{\mathbf{A}} - \mathbf{A}\|_{\infty} = O_{\mathbb{P}}\left(s\|\mathbf{\Sigma}_0^{-1}\|_{\infty}M\sqrt{\frac{\log(ed)}{n}}\right).$$

To further illustrate consequences of Theorem 1.2.2, one can now show that $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\max} = o_{\mathbb{P}}(1)$ provided that $\|\mathbf{\Sigma}_0^{-1}\|_{\infty}M\sqrt{\frac{\log(ed)}{n}} = o(1)$ and $\lambda = CM\sqrt{\log(ed)/n}$ for some pre-specified large enough constant C. Similarly, one has $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\infty} = o_{\mathbb{P}}(1)$ when $s\|\Sigma_0^{-1}\|_{\infty}M\sqrt{\frac{\log(ed)}{n}} = o(1)$.

Remark 1.2.2. Consider a particular case, Example 1 in Han et al. (2015). There, the authors proved that $\|\mathbf{\Sigma}_0^{-1}\|_{\infty} = O(1)$ as long as $\mathbf{\Sigma}_0$ is strictly diagonal dominant (Horn and Johnson (2012)), satisfying that, for all $j \in [d]$,

$$\delta_j := |\mathbf{\Sigma}_{0,jj}| - \sum_{j \neq k}|\mathbf{\Sigma}_{0,jk}| \geq \underline{\delta} > 0,$$

where $\underline{\delta}$ is some positive absolute constant. Then, the assumption that $\|\boldsymbol{\Sigma}_0^{-1}\|_\infty M\sqrt{\log(ed)/n} = o(1)$ is equivalent to $M = o\left(\sqrt{n/\log(ed)}\right)$. Similarly, the assumption that $s\|\boldsymbol{\Sigma}_0^{-1}\|_\infty M\sqrt{\log(ed)/n} = o(1)$ implies that $sM = o\left(\sqrt{n/\log(ed)}\right)$. Both allow the parameters s and M to diverge with n.

1.3 Inference on Granger Causality

In this section, we first construct a test for non-Granger causality of one individual series on another individual series in $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ using the fact that for any $k \in [d]$, $j \in [d]$, and $j \neq k$, $\{X_{t,k}\}_{t\in\mathbb{Z}}$ *Granger causes* $\{X_{t,j}\}_{t\in\mathbb{Z}}$ if and only if the (j,k)-th entry of the transition matrix, namely A_{jk}, is nonzero. In the presence of Granger causality, we then construct a sign-consistent estimator of the causality relation.

Without loss of generality and also in line with Neykov et al. (2018a), we focus on inference for A_{m1} for some $m \in [d]$. Inference for A_{mj} for $j = 2, ..., d$ is obtained by changing $\mathbf{e}_1 := (1, 0, \dots, 0)^\top$ to $\mathbf{e}_j := (\underbrace{0, \dots, 0}_{j-1}, 1, 0, \dots, 0)^\top$ below. First, we develop a test for non-Granger causality, i.e., testing

$$H_0 : A_{m1} = 0 \text{ against } H_1 : A_{m1} \neq 0. \tag{1.9}$$

The test is constructed in two steps. In the first step, we follow Neykov et al. (2018a) to construct a de-biased estimator of A_{m1} and establish its asymptotic normality. In the second step, we construct a consistent estimator of the asymptotic variance of the de-biased estimator under the null hypothesis. This is done via CBB, and requires more delicate analysis given the strategies paved by Dehling and Wendler (2010) and Fan et al. (2016).

1.3.1 A De-biased Estimator and its Asymptotic Normality

For simplicity of notation, we write $\boldsymbol{\beta} := \mathbf{A}_{m*} \in \mathbb{R}^d$, the m-th row of $\mathbf{A}$. Decompose $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^\top)^\top$, where θ is the first element of $\boldsymbol{\beta}$, i.e., $\theta = A_{m1}$. We construct a de-biased estimator of $\theta = A_{m1}$ via the application of Algorithm 1 in Neykov et al. (2018a). It consists of the following three steps.

Step 1 Let $\hat{\boldsymbol{\beta}} = (\widehat{\theta}, \widehat{\boldsymbol{\gamma}}^\top)^\top$, where $\widehat{\mathbf{A}}$ is defined in (1.6).

Step 2 Estimate $\boldsymbol{w} := [\boldsymbol{\Sigma}_0^{-1}]_{*1}$ by

$$\widehat{\boldsymbol{w}} := \underset{\boldsymbol{v}\in\mathbb{R}^d}{\operatorname{argmin}} \|\boldsymbol{v}\|_1 \text{ such that } \|\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{v} - \mathbf{e}_1\|_\infty \le \lambda',$$

where λ' is another tuning parameter.

Step 3 Estimate θ by solving the estimating equation $\widehat{S}((\theta, \widehat{\boldsymbol{\gamma}}^\top)^\top) = 0$ with

$$\widehat{S}(\boldsymbol{v}) = \widehat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{v} - \widehat{\boldsymbol{\Sigma}}_{1,*m}).$$

It is easy to see that the solution θ to the above equation has the following closed form:

$$\widetilde{\theta} = -\frac{\widehat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\beta}}(0) - \widehat{\boldsymbol{\Sigma}}_{1,*m})}{\widehat{\boldsymbol{w}}^\top[\widehat{\boldsymbol{\Sigma}}_0]_{*1}} = \widehat{\theta} - \frac{\widehat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Sigma}}_{1,*m})}{\widehat{\boldsymbol{w}}^\top[\widehat{\boldsymbol{\Sigma}}_0]_{*1}}, \tag{1.10}$$

where $\widehat{\boldsymbol{\beta}}(0) := (0, \widehat{\boldsymbol{\gamma}}^\top)^\top$.

Consistency and asymptotic normality of the above de-biased estimator $\widetilde{\theta}$ are now stated in the following two theorems.

Theorem 1.3.1. *Suppose Assumptions M, S, and E hold. Further, suppose that*

$$s\|\boldsymbol{\Sigma}_0^{-1}\|_\infty M\sqrt{\frac{\log(ed)}{n}} = o(1) \text{ and } s_{\boldsymbol{w}}\|\boldsymbol{\Sigma}_0^{-1}\|_\infty^2\sqrt{\frac{\log(ed)}{n}} = o(1),$$

where $s_{\boldsymbol{w}} := \sum_{k=1}^d \mathbb{1}(w_k \neq 0)$ is a positive integer. Let $\lambda = CM\sqrt{\log(ed)/n}$ and $\lambda' = C'\|\boldsymbol{\Sigma}_0^{-1}\|_\infty\sqrt{\log(ed)/n}$ for some sufficiently large absolute constants C, C'. Then, we have $\widetilde{\theta} - \theta = o_{\mathbb{P}}(1)$.

We note that the assumptions for consistency of $\widetilde{\theta}$ are similar to those for consistency of $\widehat{\mathbf{A}}$ in Section 2. As discussed there, M is allowed to diverge. However for asymptotic normality, we impose the assumption $M = O(1)$ following Neykov et al. (2018a) to ensure that the asymptotic variance of our estimator $\widetilde{\theta}$ is finite and the bootstrap variance proposed in the next section converges to the true one.

Theorem 1.3.2. *Suppose all conditions in Theorem 1.3.1 hold. In addition,* $\|\boldsymbol{\Sigma}_0^{-1}\|_\infty = O(1)$*,* $M = O(1)$*, and*

$$\max\{s_{\boldsymbol{w}}, s\}\frac{\log(ed)}{\sqrt{n}} = o(1).$$

Let

$$\sigma_n^2 := (n-1)\mathrm{Var}\Big\{\boldsymbol{w}^\top\Big(\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}\Big)\Big\}. \tag{1.11}$$

Assume that $\sigma_n^2 \geq C > 0$ *is bounded away from zero by some absolute constant* $C > 0$ *and for all sufficiently large* n*. Let*

$$U_n = \frac{\sqrt{n-1}}{\sigma_n}(\widetilde{\theta} - \theta).$$

Then $\lim_{n\to\infty}|\mathbb{P}(U_n \leq t) - \Phi(t)| = 0$ *for any fixed* $t \in \mathbb{R}$*.*

1.3.2 Bootstrap Estimation of the Asymptotic Variance

In this section, we use CBB to estimate the asymptotic variance σ_n^2 in (1.11) under H_0. Because our estimator contains $\widehat{\boldsymbol{\Sigma}}_1$, our bootstrap method is based on $\boldsymbol{Y}_t := (\boldsymbol{X}_t^\top, \boldsymbol{X}_{t+1}^\top)^\top$. Note that under H_0, we have $\theta = 0$.

Step 1 Construct $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n-1}\}$, where $\boldsymbol{Y}_t = (\boldsymbol{X}_t^\top, \boldsymbol{X}_{t+1}^\top)^\top$. Then, draw CBB samples from $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n-1}\}$. In detail, following Fan et al. (2016), setting ℓ to be some positive integer, we first extend $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n-1}\}$ by defining $\boldsymbol{Y}_{i+(n-1)} = \boldsymbol{Y}_i$ for $i \geq 1$, randomly draw a block of ℓ consecutive observations, and then obtain a bootstrap sample $\{\boldsymbol{Y}_t^*\}_{t=1}^{b\ell}$ by repeating this $b = \lfloor (n-1)/\ell \rceil$ times so that for each $k = 0, \ldots, b-1$,

$$\mathbb{P}^*(\boldsymbol{Y}_{k\ell+1}^* = \boldsymbol{Y}_j, \ldots, \boldsymbol{Y}_{(k+1)\ell}^* = \boldsymbol{Y}_{j+\ell-1}) = \frac{1}{n-1}$$

for any $j \in [n-1]$. Here $\mathbb{P}^*$ is the bootstrap distribution conditional on $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n-1}\}$ and $\lfloor x \rceil$ stands for the nearest integer value of $x \in \mathbb{R}$.

Step 2 Construct the bootstrap version of $\widehat{\boldsymbol{\Omega}}$,

$$\widehat{\boldsymbol{\Omega}}^* = \sin\left(\frac{\pi}{2}\cdot \underbrace{\frac{2}{b\ell(b\ell-1)}\sum_{i<j}\operatorname{sign}(\boldsymbol{Y}_i^*-\boldsymbol{Y}_j^*)\operatorname{sign}(\boldsymbol{Y}_i^*-\boldsymbol{Y}_j^*)^\top}_{\widehat{\mathbf{T}}^*}\right),$$

and let $\widehat{\boldsymbol{\Sigma}}_0^* = \widehat{\boldsymbol{\Omega}}^*_{[d],[d]}$ and $\widehat{\boldsymbol{\Sigma}}_1^* = \widehat{\boldsymbol{\Omega}}^*_{[d],d+[d]}$.

Step 3 Use the bootstrap variance defined as

$$\widehat{\sigma}_n^{2*} = (b\ell)\mathrm{Var}^*\left\{\widehat{\boldsymbol{w}}^\top\left(\widehat{\boldsymbol{\Sigma}}_0^*\widehat{\boldsymbol{\beta}}(0) - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*\right)\right\}$$

for the consistent estimator of σ_n^2. Here $\mathrm{Var}^*(\cdot)$ stands for the variance operator under $\mathbb{P}^*$.

Theorem 1.3.3. *In addition to assumptions in Theorem 1.3.2, assume further that* $\ell\to\infty$, $\ell^2/n = o(1)$, *and* $\max\{s_{\boldsymbol{w}}^2, s^2\} = o(\sqrt{n}/\log d)$. *Then under* H_0, $|\widehat{\sigma}_n^{2*} - \sigma_n^2| = o_{\mathbb{P}}(1)$.

1.3.3 Tests for Granger non-Causality

This section develops two tests for Granger non-causality, i.e., for testing H_0 in (1.9) based on the following two test statistics, where the latter is a modification to the original t type statistic,

$$T_n = \frac{\sqrt{n-1}}{\widehat{\sigma}_n^*}\widetilde{\theta} \quad\text{and}\quad T_n^{\mathrm{adj}} := (\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Sigma}}_{0,*1})T_n. \tag{1.12}$$

Theorems 1.3.2 and 1.3.3 combined together yield that T_n is asymptotically standard normal under H_0. Similar conclusion also applies to the adjusted statistic T_n^{adj} because $\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Sigma}}_{0,*1} = 1 + o_{\mathbb{P}}(1)$. However, in finite samples, $\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Sigma}}_{0,*1}$ is usually not strictly equal to 1. This motivates the adjusted test statistic that is shown to enjoy better finite sample performance in our simulation setups in Section 1.4.

Based on T_n and T_n^{adj}, we define the following two tests,

$$\mathbb{T}_{n,\alpha} = \mathbb{1}\left\{|T_n| > z_{1-\alpha/2}\right\} \text{ and } \mathbb{T}_{n,\alpha}^{\mathrm{adj}} = \mathbb{1}\left\{\left|T_n^{\mathrm{adj}}\right| > z_{1-\alpha/2}\right\},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. We then have the following corollary, which justifies the validity of the tests, and is also a direct consequence of Theorems 1.3.2 and 1.3.3.

Corollary 1.3.4 (by Theorems 1.3.2 and 1.3.3)**.** *Suppose all conditions in Theorems 1.3.2 and 1.3.3 hold. We then have*

$$\mathbb{P}(\mathbb{T}_{n,\alpha} = 1 | H_0) = \alpha + o(1) \text{ and } \mathbb{P}(\mathbb{T}^{\mathrm{adj}}_{n,\alpha} = 1 | H_0) = \alpha + o(1).$$

Remark 1.3.1. The tests developed here can be extended to testing Granger causality from a finite number of series to a single series by applying results in VAR(p) model in Appendix A.2. We can also construct tests for Granger causality from a finite number of series to any other finite number of series in a similar way. To illustrate, consider testing Granger causality from $\{X_{t,1}, X_{t,2}\}_{t\in\mathbb{Z}}$ to $\{X_{t,m}, X_{t,k}\}_{t\in\mathbb{Z}}$, i.e., testing $H_0 : A_{m1} = A_{m2} = A_{k1} = A_{k2} = 0$. With a slight abuse of notation, we define

$$\boldsymbol{\beta}_1 = \mathbf{A}_{m*},\ \boldsymbol{\beta}_2 = \mathbf{A}_{k*},\ \boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top,\ \text{where } \boldsymbol{\theta}_1 = (A_{m1}, A_{m2})^\top,\ \boldsymbol{\theta}_2 = (A_{k1}, A_{k2})^\top.$$

For any generic vector $\boldsymbol{v} = (v_1, \dots, v_d)^\top \in \mathbb{R}^d$ and a vector $\boldsymbol{x} = (x_1, x_2)^\top \in \mathbb{R}^2$, let

$$\boldsymbol{v}(\boldsymbol{x}) = (x_1, x_2, v_3, \dots, v_d)^\top.$$

That is, $\boldsymbol{v}(\boldsymbol{x})$ is a vector such that the value of $(v_1, v_2)^\top$ is replaced by $\boldsymbol{x}$. We can implement the tests as follows.

Step 1 Estimate $\boldsymbol{\theta}_1 = (A_{m1}, A_{m2})^\top$, $\boldsymbol{\theta}_2 = (A_{k1}, A_{k2})^\top$, $\boldsymbol{w}_1 = [\boldsymbol{\Sigma}_0^{-1}]_{*1}$, and $\boldsymbol{w}_2 = [\boldsymbol{\Sigma}_0^{-1}]_{*2}$ by following the steps below.

Step 1-1 Let $\widehat{\boldsymbol{\beta}}_1 = \widehat{\mathbf{A}}_{m*}$ and $\widehat{\boldsymbol{\beta}}_2 = \widehat{\mathbf{A}}_{k*}$.

Step 1-2 Estimate $\boldsymbol{w}_1 = [\boldsymbol{\Sigma}_0^{-1}]_{*1}$ and $\boldsymbol{w}_2 = [\boldsymbol{\Sigma}_0^{-1}]_{*2}$ by

$$\widehat{\boldsymbol{w}}_1 := \underset{\boldsymbol{v}\in\mathbb{R}^d}{\operatorname{argmin}} \|\boldsymbol{v}\|_1 \text{ such that } \|\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{v} - \mathbf{e}_1\|_\infty \le \lambda',$$

$$\widehat{\boldsymbol{w}}_2 := \underset{\boldsymbol{v}\in\mathbb{R}^d}{\operatorname{argmin}} \|\boldsymbol{v}\|_1 \text{ such that } \|\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{v} - \mathbf{e}_2\|_\infty \le \lambda'.$$

We denote $\widehat{\mathbf{W}} = (\widehat{\boldsymbol{w}}_1^\top, \widehat{\boldsymbol{w}}_2^\top)^\top$.

Step 1-3 Estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ by solving equations $\widehat{S}_1(\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\theta}_1)) = \mathbf{0}_{2\times 1}$ and $\widehat{S}_2(\widehat{\boldsymbol{\beta}}_2(\boldsymbol{\theta}_2)) = \mathbf{0}_{2\times 1}$ with

$$\widehat{S}_1(\boldsymbol{v}) = \widehat{\mathbf{W}}^\top(\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{v} - \widehat{\boldsymbol{\Sigma}}_{1,*m}) \text{ and } \widehat{S}_2(\boldsymbol{v}) = \widehat{\mathbf{W}}^\top(\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{v} - \widehat{\boldsymbol{\Sigma}}_{1,*k}).$$

Step 2 Construct the bootstrap variance defined as $\boldsymbol{\Sigma}_n^* := (b\ell)\mathrm{Var}^* \begin{pmatrix} \widehat{\mathbf{W}}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\widehat{\boldsymbol{\beta}}_m(\mathbf{0}_{2\times 1}) - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*) \\ \widehat{\mathbf{W}}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\widehat{\boldsymbol{\beta}}_k(\mathbf{0}_{2\times 1}) - \widehat{\boldsymbol{\Sigma}}_{1,*k}^*) \end{pmatrix}$, where $\widehat{\boldsymbol{\Sigma}}_0^*$ and $\widehat{\boldsymbol{\Sigma}}_1^*$ are the bootstrap versions of $\widehat{\boldsymbol{\Sigma}}_0$ and $\widehat{\boldsymbol{\Sigma}}_1$ defined in Section 1.3.2.

Step 3 Define test statistics:

$$\mathcal{W}_n = (n-1)\begin{pmatrix}\hat{\boldsymbol{\theta}}_1, & \hat{\boldsymbol{\theta}}_2\end{pmatrix}[\boldsymbol{\Sigma}_n^*]^{-1}\begin{pmatrix}\hat{\boldsymbol{\theta}}_1 \\ \hat{\boldsymbol{\theta}}_2\end{pmatrix} \text{ and}$$

$$\mathcal{W}_n^{adj} = (n-1)\begin{pmatrix}\widehat{\mathbf{W}}^\top\widehat{\boldsymbol{\Sigma}}_{0,*G}\hat{\boldsymbol{\theta}}_1 \\ \widehat{\mathbf{W}}^\top\widehat{\boldsymbol{\Sigma}}_{0,*G}\hat{\boldsymbol{\theta}}_2\end{pmatrix}^\top [\boldsymbol{\Sigma}_n^*]^{-1}\begin{pmatrix}\widehat{\mathbf{W}}^\top\widehat{\boldsymbol{\Sigma}}_{0,*G}\hat{\boldsymbol{\theta}}_1 \\ \widehat{\mathbf{W}}^\top\widehat{\boldsymbol{\Sigma}}_{0,*G}\hat{\boldsymbol{\theta}}_2\end{pmatrix},$$

where $G = \{1, 2\}$.

Based on $\mathcal{W}_n$ and $\mathcal{W}_n^{adj}$, we can implement the tests using χ_4^2 distribution.

1.3.4 Estimation of the Sign of Granger Causality

When a test detects Granger causality, it may be of interest to further infer the sign of A_{m1}. Following Qiu et al. (2015), we define a new estimator of $\mathbf{A}$ as $\widetilde{\mathbf{A}} = [\widetilde{\mathbf{A}}_{jk}]$, where $\widetilde{\mathbf{A}}_{jk} := \widehat{\mathbf{A}}_{jk}\mathbb{1}(|\widehat{\mathbf{A}}_{jk}| \geq \gamma)$ for some threshold level γ. Then, similar to Theorem 6.4 in Qiu et al. (2015), we have the following sign-consistency theorem.

Theorem 1.3.5 (Theorem 6.4 in Qiu et al. (2015))**.** *Assume that the conditions in Theorem 1.2.1 hold. Let* $\gamma = 2\|\boldsymbol{\Sigma}_0^{-1}\|_\infty(M+1)z$*. Then, with probability no smaller than* $1-\epsilon_0$*, where* z *is defined in equation (1.7) and* ϵ_0 *is defined in Remark 1.2.1, we have* $\mathrm{sign}(\widetilde{\mathbf{A}}) = \mathrm{sign}(\mathbf{A})$*, provided that* $\min_{\{(j,k):\mathbf{A}_{jk}>0\}}|\mathbf{A}_{jk}| \geq 2\gamma$.

1.4 Monte-Carlo Simulation

In this section, we investigate the finite sample performance of the proposed estimator $\widehat{\mathbf{A}}$ and tests for non-Granger causality based on T_n and T_n^{adj}. The data generating process (DGP) used in the simulation is characterized by model (1.1) for the latent process and the transformation $\boldsymbol{f}$ generating the observed process.

1.4.1 Data Generating Processes

We consider three DGPs in the simulation. Noting that our proposed estimator and tests are invariant to any increasing $\boldsymbol{f}$, we use the following $\boldsymbol{f}$ from Xue and Zou (2012) in all three DGPs,

$$\boldsymbol{f} = \{f_1, f_2, f_3, f_4, f_5, f_1, f_2, f_3, f_4, f_5, \dots\},$$

where

$$\begin{aligned}
&f_1(x) = x, \quad f_2(x) = \exp(x), \quad f_3(x) = x^3, \quad f_4(x) = \frac{1}{1+\exp(-x)}, \quad \text{and} \\
&f_5(x) = (f_2(x) - 3)\mathbb{I}(x < -1) + f_1(x)\mathbb{1}(-1 \le x \le 1) + (f_4(x-1)+1)\mathbb{1}(x > 1). \qquad (1.13)
\end{aligned}$$

The matrix $\boldsymbol{\Sigma}_0$ for all three DGPs is generated as follows: $\boldsymbol{\Sigma}_0$ is the normalized $\boldsymbol{\Xi}^{-1}$ with diagonal elements taking the value 1, where $\boldsymbol{\Xi}$ is a tri-diagonal matrix with diagonal elements 1 and sub-diagonal and superdiagonal elements 1/3.

The three DGPs are distinguished by the transition matrix $\mathbf{A}$.

DGP A The matrix $\mathbf{A}$ is a tri-diagonal matrix such that its diagonal elements are $|\rho_A|$ and sub-diagonal and superdiagonal elements are $\rho_A/3$, where ρ_A is chosen so that $\boldsymbol{\Sigma}_E = \boldsymbol{\Sigma}_0 - \mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^\top$ is a positive-definite matrix. In this simulation, we choose $\rho_A = 0.54$.

DGP B and DGP C are based on Experiments B and D in Kock and Callot (2015).

DGP B The matrix $\mathbf{A}$ is a block diagonal matrix, where the blocks are 5×5 matrices with all entries of the blocks of $\mathbf{A}$ equaling to 0.15.

DGP C An element $\mathbf{A}_{ij}$ of $\mathbf{A}$ satisfies the following condition:

$$\mathbf{A}_{ij} = \begin{cases} (-1)^{|i-j|}\rho^{|i-j|+1} & \text{if } |\rho|^{|i-j|+1} > c, \\ 0 & \text{otherwise,} \end{cases}$$

where $|\rho| < 1$, and c is some absolute positive constant. In the simulation, we choose $\rho = 0.4$, and $c = 0.05$.

Given sample size n, dimension d, matrix $\mathbf{A}$, and matrix $\mathbf{\Sigma}_0$, we generate data from each DGP in two steps:

DGP Step 1 Generate $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\}$ recursively. That is, generate $\boldsymbol{Z}_1$ from $N(\mathbf{0}, \mathbf{\Sigma}_0)$ and given $\boldsymbol{Z}_{t-1}$, generate $\boldsymbol{Z}_t$ from $\boldsymbol{Z}_t = \mathbf{A}\boldsymbol{Z}_{t-1} + \boldsymbol{E}_t$, where $\boldsymbol{E}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{E}})$;

DGP Step 2 Transform $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ to $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ using $\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{Z}_t)$, for $t = 1, \ldots, n$.

Once the data $\{\boldsymbol{X}_t\}_{t=1}^n$ is generated, we construct $\{\mathbf{Y}_t\}_{t=1}^{n-1}$, where $\mathbf{Y}_t = (\boldsymbol{X}_t^\top, \boldsymbol{X}_{t+1}^\top)^\top$ and compute $\widehat{\mathbf{A}}$ using the tuning parameter $\lambda = C\sqrt{\log d/(n-1)}$. We choose $d = 80, 120$. The sample sizes are $n = 251, 501, 751$. All the results are obtained from 5000 simulation repetitions. To increase computational speed, we used the fast algorithms for the calculation of Kendall's tau in codes implemented in pcaPP package by Filzmoser et al. (2021). We report two sets of results below. The first set of results measures the performance of the estimator $\widehat{\mathbf{A}}$ in terms of its ability to detect the true sparsity pattern in $\mathbf{A}$. The second set of results reports the size and power performance of the two proposed tests for non-Granger causality.

1.4.2 The Performance of $\widehat{\mathbf{A}}$

Following Liu et al. (2012), we use the receiver operating characteristic (ROC) curves to illustrate the ability of the estimator $\widehat{\mathbf{A}}$ in detecting the sparsity pattern of A. For this, we define the sparsity sets $S_{\mathbf{A}}$ and $S_{\widehat{\mathbf{A}}}$ as follows:

- the pair (i, j) is not an element of the set $S_{\mathbf{A}}$ if and only if $A_{ij} = 0$;
- the pair (i, j) is not an element of the set $S_{\widehat{\mathbf{A}}}$ if and only if $\widehat{A}_{ij} = 0$.

Then, following equations (5.2) and (5.3) in Liu et al. (2012), we define the false positive rate (FPR) and the false negative rate (FNR) given the tuning parameter λ as follows. We first calculate the false positive and false negative numbers.

- $\mathrm{FP}(\lambda)$: the number of pairs in $S_{\widehat{\mathbf{A}}}$ not in $S_{\mathbf{A}}$;
- $\mathrm{FN}(\lambda)$: the number of pairs in $S_{\mathbf{A}}$ not in $S_{\widehat{\mathbf{A}}}$.

Similar to equation (5.4) in Liu et al. (2012), we calculate the $\mathrm{FNR}(\lambda)$, the $\mathrm{FPR}(\lambda)$, and true positive rate ($\mathrm{TPR}(\lambda)$) as follows,

$$\mathrm{FNR}(\lambda) = \frac{\mathrm{FN}(\lambda)}{\mathrm{card}\,(S_{\mathbf{A}})}, \quad \mathrm{FPR}(\lambda) = \frac{\mathrm{FP}(\lambda)}{d^2 - \mathrm{card}\,(S_{\mathbf{A}})}, \quad \mathrm{TPR}(\lambda) = 1 - \mathrm{FNR}(\lambda),$$

where $\mathrm{card}(\cdot)$ outputs the number of elements in the input. Then, we plot the averages of $\mathrm{FPR}(\lambda)$ and $\mathrm{TPR}(\lambda)$ over 5,000 repetitions for given n and d. Here λ is set to be $C\sqrt{\log d/(n-1)}$, where $C \in \{0.10, 0.11, 0.12, ..., 1.00\}$.

Since the qualitative results are the same for all three DGPs. We present ROC plots for DGP A only in Figure 1.1 below.

Several observations are in-line. First, for each pair of (n, d), the true positive rate increases along with the false positive rate. This is as expected since, as more non-zero $\widehat{A}_{ij}$'s are selected, it is also likely that we will pick more non-zero A_{ij}'s. Secondly, as the sample size increases, the true positive rate increases while the false positive rate decreases. For example, for DGP A, when the sample size is 751, the true positive rate is above 0.9 while the false positive rate is lower than 0.2. This demonstrates the powerfulness of the proposed method in variable selection. Lastly, comparing the three different pairs of (n, d), it could be observed that, as the sample size alone increases, the ROC curves grow higher. This indicates that the proposed method is more capable of correctly selecting the non-zero coefficients as more information is obtained, which is as expected.

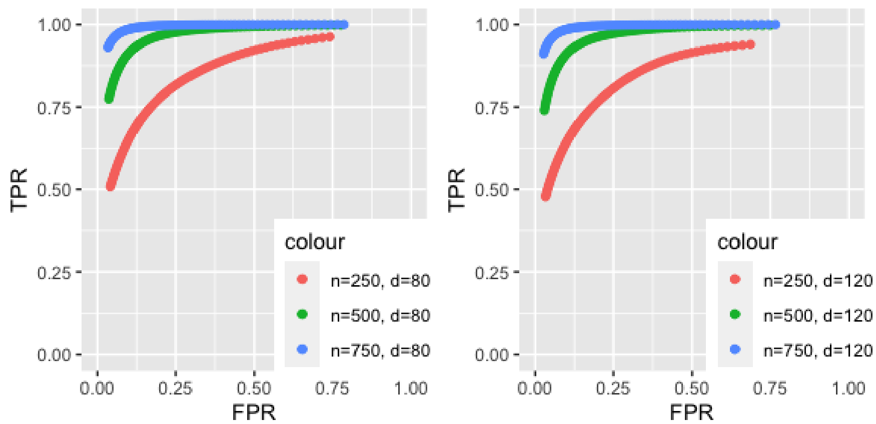

Figure 1.1: The ROC plots for DGP A

1.4.3 The Size and Power Performance

For each DGP, we examine the size performance of T_n and T_n^{adj} by testing four null hypotheses: $A_{1,j} = 0$ for $j = 10, 20, 30, 40$. For power performance, we test the bull hypothesis: $A_{1,2} = 0$ for DGP A; $A_{1,5} = 0$ for DGP B; and $A_{1,j} = 0$ for $j = 2, 3$ for DGP C. Following Section 5.2 of Neykov et al. (2018a), we compute the de-biased estimator using the tuning parameters $\lambda = \lambda' = 0.5\sqrt{\frac{\log d}{n-1}}$. For CBB, we set $\ell = [(n-1)^{1/3}]$ following the discussions in Dehling and Wendler (2010) and compute the bootstrap variance based on $2,000$ bootstrap samples.

Table 1.1 reports the results for DGPs A, B, C for nominal sizes $5\%, 10\%$. Several qualitative conclusions emerge. First, for all settings, the adjusted test has better size than the unadjusted test; second, for almost all settings, the sizes of both tests improve and approach the nominal size as n gets larger; third, the power of both tests increases quickly with the sample size; and lastly the results are qualitatively similar across different values of d and across the three DGPs.

1.5 Concluding Remarks

In this paper, we have constructed a simple rank-based estimator of the transition matrix in a high dimensional stationary Gaussian copula VAR(p) model for any finite order p and established its rates of convergence. We have also developed tests for Granger non-causality of one or a finite number of individual series on another or a finite number of individual series in the high-dimensional model and provided numerical evidence on their finite sample performance.

It is worth mentioning that all the methods developed in this paper apply directly to the corresponding high dimensional stationary Gaussian copula VAR(p) model with non-zero mean, although we are unable to estimate the mean. Several extensions are possible. First, if it is desirable to test Granger non-causality of an increasing number of individual series on another series in the high-dimensional latent model, the idea underlying multiple testing and the method of multiplier bootstrap in Chernozhukov et al. (2013) may be adopted. Second, allowing for both increasing d and increasing p is interesting but technically challenging. Third, extensions of our rank-based methodology to certain structural models in economics and finance such as structural asset pricing models, where the observed equity price is an increasing transformation (unknown) of the latent firm's market value, and structural auction models, where the observed bid is an increasing transformation (unknown) of the bidder's private value, are important and will be studied in future work.

Acknowledgements

We are grateful to the Editor, Serena Ng, an Associate Editor, and three referees for their insightful comments on the previous version of this paper. We thank Yandi Shen, Kevin Song, participants of the Sixth Seattle-Vancouver Econometrics Conference, and seminar participants at Penn State University for helpful discussions. Numerical computation in this paper is done via the Statistics cluster in the Department of Statistics at the University of Washington, CSDE Simulation Cluster, which has received partial support from an Eunice

Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2C HD042828, and the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system and funded by the STF at the University of Washington.

Table 1.1: The power and size for DGPs A-C

	n=251				n=501				n=751			
	5%		10%		5%		10%		5%		10%	
Index	T_n	T_n^{adj}	T_n	T_n^{adj}	T_n	T_n^{adj}	T_n	T_n^{adj}	T_n	T_n^{adj}	T_n	T_n^{adj}
DGP A												
d=80												
(1, 2)	0.6664	0.6270	0.7534	0.7242	0.9376	0.9276	0.9654	0.9602	0.9940	0.9926	0.9970	0.9966
(1, 10)	0.0920	0.0712	0.1514	0.1250	0.0754	0.0608	0.1314	0.1146	0.0700	0.0616	0.1240	0.1092
(1, 20)	0.0908	0.0708	0.1486	0.1254	0.0786	0.0682	0.1414	0.1206	0.0722	0.0606	0.1334	0.1194
(1, 30)	0.0902	0.0720	0.1468	0.1230	0.0782	0.0662	0.1410	0.1232	0.0712	0.0604	0.1214	0.1082
(1, 40)	0.0954	0.0748	0.1592	0.1300	0.0766	0.0666	0.1362	0.1204	0.0710	0.0594	0.1300	0.1174
d=120												
(1, 2)	0.6500	0.6060	0.7428	0.7102	0.9274	0.9134	0.9558	0.9494	0.9870	0.9834	0.9926	0.9922
(1, 10)	0.0918	0.0730	0.1584	0.1306	0.0746	0.0640	0.1326	0.1190	0.0722	0.0606	0.1324	0.1172
(1, 20)	0.0924	0.0704	0.1574	0.1306	0.0772	0.0636	0.1314	0.1144	0.0686	0.0572	0.1254	0.1124
(1, 30)	0.0976	0.0730	0.1658	0.1362	0.0804	0.0680	0.1398	0.1212	0.0728	0.0618	0.1362	0.1198
(1, 40)	0.0896	0.0674	0.1554	0.1274	0.0824	0.0708	0.1418	0.1204	0.0688	0.0604	0.1226	0.1080
DGP B												
d=80												
(1, 5)	0.4698	0.4272	0.5788	0.5384	0.7648	0.7384	0.8474	0.8278	0.9138	0.9030	0.9498	0.9434
(1, 10)	0.0914	0.0702	0.1558	0.1276	0.0726	0.0612	0.1280	0.1142	0.0678	0.0588	0.1246	0.1108
(1, 20)	0.0798	0.0630	0.1420	0.1154	0.0762	0.0634	0.1356	0.1196	0.0732	0.0636	0.1262	0.1158
(1, 30)	0.0848	0.0652	0.1458	0.1192	0.0732	0.0584	0.1302	0.1138	0.0676	0.0576	0.1228	0.1096
(1, 40)	0.0844	0.0682	0.1462	0.1204	0.0742	0.0622	0.1322	0.1156	0.0648	0.0548	0.1220	0.1084
d=120												
(1, 5)	0.4600	0.4158	0.5630	0.5226	0.7578	0.7304	0.8390	0.8178	0.9120	0.8994	0.9500	0.9436
(1, 10)	0.0900	0.0712	0.1514	0.1246	0.0804	0.0662	0.1382	0.1192	0.0714	0.0608	0.1308	0.1186
(1, 20)	0.0936	0.0728	0.1542	0.1266	0.0746	0.0640	0.1288	0.1122	0.0604	0.0534	0.1172	0.1028
(1, 30)	0.0940	0.0744	0.1546	0.1272	0.0748	0.0612	0.1344	0.1142	0.0694	0.0584	0.1274	0.1108
(1, 40)	0.0854	0.0626	0.1484	0.1194	0.0746	0.0602	0.1326	0.1132	0.0666	0.0578	0.1202	0.1062
DGP C												
d=80												
(1, 2)	0.6702	0.6304	0.7620	0.7248	0.9272	0.9158	0.9578	0.9522	0.9858	0.9822	0.9938	0.9930
(1, 3)	0.2240	0.1914	0.3164	0.2794	0.3288	0.3002	0.4372	0.4072	0.4288	0.4022	0.5402	0.5172
(1, 10)	0.0830	0.0610	0.1418	0.1148	0.0700	0.0552	0.1254	0.1050	0.0678	0.0568	0.1248	0.1104
(1, 20)	0.0726	0.0550	0.1376	0.1110	0.0714	0.0564	0.1326	0.1136	0.0644	0.0558	0.1228	0.1094
(1, 30)	0.0774	0.0596	0.1366	0.1114	0.0706	0.0578	0.1306	0.1090	0.0630	0.0564	0.1160	0.1040
(1, 40)	0.0850	0.0670	0.1442	0.1192	0.0744	0.0608	0.1318	0.1134	0.0730	0.0622	0.1238	0.1126
d=120												
(1, 2)	0.6282	0.5892	0.7170	0.6858	0.9134	0.8986	0.9524	0.9434	0.9852	0.9818	0.9942	0.9934
(1, 3)	0.2262	0.1898	0.3178	0.2778	0.3384	0.3060	0.4430	0.4116	0.4346	0.4058	0.5514	0.5218
(1, 10)	0.0794	0.0568	0.1388	0.1124	0.0752	0.0608	0.1320	0.1106	0.0708	0.0588	0.1266	0.1140
(1, 20)	0.0788	0.0604	0.1446	0.1164	0.0674	0.0580	0.1268	0.1070	0.0630	0.0530	0.1202	0.1050
(1, 30)	0.0844	0.0644	0.1396	0.1134	0.0704	0.0590	0.1318	0.1116	0.0714	0.0632	0.1294	0.1164
(1, 40)	0.0796	0.0600	0.1400	0.1140	0.0760	0.0624	0.1370	0.1162	0.0658	0.0558	0.1200	0.1072

Chapter 2

MINIMUM SLICED DISTANCE ESTIMATION IN STRUCTURAL MODELS [1]

2.1 Introduction

Estimation and inference in structural models in economics and finance often pose challenges to the classical likelihood-based method. Examples of such models include Dynamic Stochastic General Equilibrium (DSGE) models in macroeconomics and asset pricing models in finance; BLP aggregate demand models in empirical industrial organization, and econometric models with parameter-dependent supports such as auction models and equilibrium job-search models. In DSGE and asset pricing models, the observable economic/financial variables are often driven by a few unobserved state variables and as a result, the mapping from the unobserved state variables to the observed variables is non-invertible. Such a system is known as *a stochastic singular system* and MLE is not well-defined for stochastic singular systems. To handle this issue, a common practice in the current literature is to add enough measurement errors to the model such that the mapping from unobserved variables including the original unobserved state variables and added measurement errors to the observable variables is invertible, see An and Schorfheide (2007) and Komunjer and Ng (2011) for DSGE models and Pastorello et al. (2003) and Chernov (2003) for asset pricing models. However, there is no theoretical guidance on how the measurement errors should be introduced and numerical evidence reveals sensitivity of conclusions to the types of measurement errors and how the measurement errors are added to the system. In BLP models, the dimension of the observables is the same as that of the unobservables, but invertibility is not always guaran-

[1] THIS CHAPTER IS A JOINT WORK WITH YANQIN FAN FROM THE UNIVERSITY OF WASHINGTON

teed, see Berry (1994); Berry et al. (1995) for sufficient conditions ensuring the invertibility of the aggregate demand function in the mean utility. In addition to the invertibility of the aggregate demand function, computing MLE may be challenging: it requires computing the Jacobian matrix of the mapping and the integrals defining the mapping as well as a tractable likelihood function.

In econometric models with parameter-dependent supports such as the one-sided and two-sided models studied in Chernozhukov and Hong (2004),[2] MLE is well defined but the classical likelihood-based inference may be invalid, see Chernozhukov and Hong (2004) for the non-normal asymptotic theory for MLE and Bayes estimation (BE) and Hirano and Porter (2003) for efficiency considerations according to the local asymptotic minimax criterion for conventional loss functions. In the one-sided model in Chernozhukov and Hong (2004) and Hirano and Porter (2003), the conditional density function of the dependent variable is assumed to be strictly bounded away from zero at the boundary; In the two-sided models in Chernozhukov and Hong (2004), the conditional density function has a jump at the boundary, see Example 2.2.3 for a rigorous statement of this assumption for both one-sided and two-sided models. Under this assumption, likelihood-based inference relies on the potentially complex asymptotic distribution which may be difficult to implement.[3] Another impediment to likelihood-based inference is the dichotomy of the asymptotic theory. For example, in the two-sided models, MLE is asymptotically normal when the conditional density function of the dependent variable has no jump at the boundary and non-normal otherwise.

The sensitivity of likelihood-based inference to model assumptions and parameter values in the afore-mentioned structural models is partly due to the distinct properties of the Kullback-Leibler (KL) divergence for these models. Using a singular example (Example 1 in

---

[2]Chernozhukov and Hong (2004) reviews many early works in economics that belong to this class of models such as auction models in Paarsch (1992) and Donald and Paarsch (2002); parametric equilibrium search in Bowlus et al. (2001) among others.

[3]Chernozhukov and Hong (2004) shows that Bayesian credible intervals based on posterior quantiles, which are computationally attractive, are valid in large samples and perform well in small samples.

Arjovsky et al. (2017)), a one-sided uniform model, and a two-sided uniform model, we show that for the singular model, the KL divergence is unbounded rendering MLE undefined; for the one-sided model, the KL divergence is bounded but not differentiable at the true parameter value and as a result, the classical normal asymptotic theory for MLE is invalid; for the two-sided model, the KL divergence is bounded but may or may not be first order differentiable at the true parameter value resulting in a dichotomy of asymptotic theory for MLE.

The aim of this paper is to develop a simple, robust method for the estimation and inference in a broad class of structural models including asset pricing models, BLP, and models with parameter-dependent supports. This is accomplished by introducing a class of minimum distance estimators based on *sliced* L_2-distances between some "empirical measure" of the observed data and the corresponding parametric or semiparametric measure induced by the parametric structural model. We call this class of estimators *minimum sliced distance* (MSD) estimators. By choosing different measures of the distribution, we obtain different MSD estimators. Three important estimators we focus on in this paper are the minimum sliced Wasserstein Distance (MSWD) estimator, the minimum sliced Cramer Distance (MSCD) estimator, and *Simulated* MSD estimator. The SMSD estimator is proposed for generative models from which one can easily draw synthetic samples for any given parameter value in its parameter space. In contrast to MLE, the MSD estimator is well defined for structural models ranging from stochastic singular models to models with parameter-dependent supports as well as generative models.

As our first theoretical contribution, we establish consistency and asymptotic normality of the general MSD estimator under a set of high-level assumptions. In contrast to likelihood-based inference, inference using our MSD estimator is standard. We then verify the high-level assumptions for the MSWD, MSCD estimators and their simulated versions under primitive conditions for three broad classes of structural models: (i) unconditional models which induce parametric distributions such as asset pricing models; (ii) conditional models which induce semiparametric distributions such as BLP and parameter-dependent support models; and

(iii) unconditional and conditional generative models which have intractable distributions but can be simulated easily. For parameter dependent support models in Chernozhukov and Hong (2004) and Hirano and Porter (2003), our primitive conditions imply that our estimator is asymptotically normally distributed regardless of the presence of a jump in the conditional density function at the boundary or not. We confirm our theoretical findings via two simulation experiments based on the stochastic singular state-space model and the auction models in Paarsch (1992), Donald and Paarsch (2002), and Li (2010).

This paper makes contributions to several literatures. First, Wasserstein distances have recently been used in the statistics and machine learning literatures for testing and estimation of generative models. For example, Bernton et al. (2019) proposes minimum WD estimator based on the 1-Wasserstein distance and establishes non-normal asymptotic distribution for *univariate* unconditional parametric distributions; Nadjahi et al. (2020b) proposes a sliced version of the estimator in Bernton et al. (2019) and extends the non-normal asymptotic distribution in Bernton et al. (2019). The technical analysis of Bernton et al. (2019) (and Nadjahi et al. (2020b)) relies critically on the equivalence between the 1-Wasserstein distance and the 1-Cramer distance for univariate distributions. Such an equivalence no longer holds for the 2-Wasserstein distance adopted in the current paper and the analysis in Bernton et al. (2019) breaks down. The current paper is the first to adopt the sliced 2-Wasserstein distance in the estimation of a broad range of structural models including generative models and establish its asymptotic normality under general conditions. For univariate models, our MSWD estimator reduces to the L_2-quantile distance estimator of parametric distributions in LaRiccia (1982, 1984) and the matching quantile estimator of linear models in Sgouropoulos et al. (2015) and Qin and Wu (2020). Second, the sliced Cramer distance is used in Zhu et al. (1997) for goodness-of-fit testing in the i.i.d. case. Our paper is the first to adopt the sliced CD for estimation. In the univariate case, the MSCD estimator reduces to the estimator in Hettmansperger et al. (1994) and Öztürk and Hettmansperger (1997). Beutner and Bordes (2011) extends Öztürk and Hettmansperger (1997) to censored data. Third, for generative models, a popular machine learning tool is the Generative Adversarial Network

(GAN) originally proposed in Goodfellow et al. (2014). A GAN estimator is formulated as the solution to a minimax problem between a generator and a discriminator. The discriminator maximizes the accuracy the its classification while the generator minimizes it. Goodfellow et al. (2014) shows that at discriminator optimality, the GAN generator minimizes the Jensen-Shannon (JS) divergence between the true/target distribution and the distribution induced by the generative model. Like the KL divergence, the JS divergence is not continuous for stochastic singular systems rendering instable behavior of the original GAN. This motivates Arjovsky et al. (2017) to propose Wasserstein GAN based on the Earth Mover or 1-Wasserstein distance and Lei et al. (2019) to propose Wasserstein GAN based on the 2-Wasserstein distance. Both are implemented using the dual forms of the Wasserstein distances. Neither paper establishes asymptotic theory for their estimators. Although our MSWD estimator is a minimum distance estimator, we demonstrate via Brenier theorem that it is in fact a sliced 2-Wasserstein GAN estimator. By using the primal form directly, our MSWD estimator overcomes the computational challenge and the difficulty of finding the encoding map needed to implement the estimator in Lei et al. (2019), see Section 2.3.3 for a detailed discussion.[4]

Fourth, the current paper offers complementary results to Kaji et al. (2020) which extends the original GAN in Goodfellow et al. (2014) to estimate structural models. It is the first paper to establish a rigorous asymptotic theory including asymptotic normality of the original GAN estimator for structural models. It remains to see if their assumptions can be verified for stochastic singular systems and models with parameter-dependent supports. Like Kaji et al. (2020), our SMSD estimator is related to important works in the econometrics literature on indirect inference and it is free from the choice of an auxillary model. We refer interested readers to Kaji et al. (2020) for a nice discussion of the indirect inference literature.

Technically we make use of Andrews (1999) to establish asymptotic theory for the general MSD estimator under high level assumptions and rely heavily on modern empirical processes

[4] Kolouri et al. (2018) use the sliced Wasserstein distance for the auto-encoder problem for the generative models.

and U-processes results in Sherman (1994), Newey (1991), and Brown and Wegkamp (2002) to verify high level assumptions for the MSWD, MSCD, and their simulated versions under primitive conditions.

The rest of the paper is organized as follows. Section 2 introduces our general framework and the three motivating examples. Section 3 introduces our MSD estimator and applies it to the motivating examples. Using simple examples, Section 4 illustrates the different properties of the KL divergence, JS divergence, the 2-Wasserstein distance, and Cramer distance, as well as the different distributions of the estimators based on these divergences/distances. Section 5 establishes the consistency and asymptotic normality of the MSD estimator under high level assumptions. Sections 6-8 verify the high level assumptions for the unconditional model, conditional model, and generative model under primitive conditions. Section 9 presents numerical results using synthetic data generated from the stochastic singular state-space model, the term structure model in Backus et al. (1998), and the auction model in Paarsch (1992), Donald and Paarsch (2002), and Li (2010). Section 10 concludes. A series of appendices contains technical proofs of the main results in the text and additional materials for univariate unconditional models and the singular example in Arjovsky et al. (2017).

2.2 The General Framework and Three Motivating Examples

Let $\{Z_t : t = 1, 2, ...\}$ denote a strictly stationary process. We consider three general models for $\{Z_t : t = 1, 2, ...\}$. The first model is an *unconditional model* in which $\{Z_t : t = 1, 2, ...\}$ has a stationary parametric distribution denoted as $F(\cdot; \psi_0)$ for some unknown $\psi_0 \in \Psi \subset \mathbb{R}^{d_\psi}$; the second model is a *conditional model*, where $Z_t^\top = \left(Y_t^\top, X_t^\top\right)$, the conditional distribution of Y_t given $X_t = x$ is of parametric form denoted by $F(\cdot|x, \psi_0)$ but *the distribution of X_t is unspecified*; and the third model is a generative model from which one can easily draw synthetic samples at any given $\psi \in \Psi$. In all three models, the support of the distribution of Z_t is allowed to depend on the unknown parameter ψ_0, and we are interested in the estimation and inference for ψ_0.

Assumption 2.2.1. *The sample information denoted as $\{Z_t : t = 1, 2, ..., T\}$ is either a random sample or a strictly stationary time series from either the unconditional model characterized by the distribution function $F(\cdot;\psi_0)$ of Z_t; the conditional model characterized by the conditional distribution function of Y_t given $X_t = x$ denoted as $F(\cdot|x,\psi_0)$; or the generative model with parameter ψ_0.*

When the density function of either $F(\cdot;\psi)$ or $F(\cdot|x,\psi)$ exists, MLE or the conditional MLE is a popular approach to estimating the unknown parameter ψ_0. The asymptotic distribution of MLE or conditional MLE and the associated inference depend critically on the model assumptions. The classical Wald, QLR, and score tests rely on smoothness assumptions on the density function and the assumption that the true parameter ψ_0 is in the interior of Ψ. Many important models in economics and finance violate one or more assumptions underlying the classical likelihood theory motivating the development of alternative methods of estimation and inference such as those in the references discussed in Section 1.

Below, we present three classes of structural models in economics and finance for which likelihood-based estimation and inference are either not well defined or difficult to implement. We will use them to illustrate our new method in subsequent sections. Since the three models are from different areas, we adopt the conventional notations in the corresponding literatures so that notations in each example are specific to that example.

Example 2.2.1 (Asset Pricing/State Space Models)**.** *In important financial and macroeconomic models such as asset pricing and DSGE models, the observables are typically driven by a few unobservable or latent factors. To apply likelihood-based inference, a common practice is to add enough measurement errors to the model to make the system stochastically non-singular. However, the resulting conclusions depend critically on how and what types of measurement errors are added, see Section 6 of Pastorello et al. (2003) and Chernov (2003) for detailed discussions on the implications in asset pricing models. We use a simple generic state space model as a running example. Let $\{Y_t\}_{t=1}^{\infty}$ be the observable process and $\{Y_t^*\}_{t=1}^{\infty}$*

be the latent process such that

$$Y_t = h(Y_t^*, \theta_0), \tag{2.1}$$

where Y_t *is a vector of observable variables of dimension* d_y*, and* Y_t^* *is a vector of latent state variables of dimension* d^* *such that* $\{Y_t^*\}_{t=1}^{\infty}$ *is a strictly stationary Markov process of order 1 with transition density* $f^*\left(\cdot|y_{t-1}^*;\gamma_0\right)$*. Let* $\psi_0 = (\theta_0, \gamma_0)$*. We are interested in estimating* ψ_0 *from time series* $\{Y_t\}_{t=1}^{T}$*. Chernov (2003) summarizes nicely the different scenarios in asset pricing models:*

(i) *When* $d^* < d_y$*, (e.g., Chen and Scott (1993)), the function* h *in (2.1) is non-invertible. A common practice is to assume that* d^* *asset prices are observed without any error and add* $(d_y - d^*)$ *measurement errors to the system of equations in (2.1) assuming that* $(d_y - d^*)$ *asset prices are observed with an error (either to address microstructure effects or transactions costs, or to resolve the statistical singularityproblem);*

(ii) *When* $d^* = d_y$*, (e.g., Pan (2002)), the function* h *in (2.1) is typically invertible;*

(iii) *When* $d^* > d_y$*, the inverse of* h *does not exist. The examples are as follows:*

  (a) $d_y = 1$*, that is, one estimates a multifactor model for one financial asset, be it equity returns, interest, or exchange rate (e.g., Andersen et al. (2002); Chernov et al. (2003); Eraker et al. (2003)).*

  (b) *All observables have pricing errors; for example, some term structure studies assume that all bond yields are observed with an error (e.g., Jegadeesh and Pennacchi (1996)).*

The mapping in (2.1) is non-invertible in both Case (i) and Case (iii). However, in Case (i), the distribution of Y_t *is singular with support lying in a low dimensional manifold of* R^{d_y} *rendering MLE invalid. On the other hand, MLE is valid in Case (iii), but may be difficult to compute due to the intractibility of the likelihood function.*

Let $Z_t = Y_t$ *or* $\left(Y_t^\top, Y_{t-1}^\top\right)^\top$*. This is an example of the unconditional model.*

Example 2.2.2 (A BLP Model)**.** *Notations and discussion in this example follow closely Gandhi and Houde (2019). They consider a special case of the random-utility model considered by Berry et al. (1995), in which product characteristics are exogenous. Consider a market* t *with* $J_t + 1$ *differentiated products. Each product* j *is characterized by a vector of observed (to the econometrician) product characteristics* $x_{jt} \in \mathbb{R}^K$ *and an unobserved characteristic* ξ_{jt}*. Let* $x_t = (x_{1t}, ..., x_{J_t,t})$ *denote a summary of the observed market structure-the entire menu of observed product characteristics available to consumers in market* t *(i.e.,* $J_t \times K$ *matrix). Similarly,* $s_t = (s_{1t}, ..., s_{J_t,t})$ *denotes the vector of observed market shares, which is defined such that* $1 - \sum_{j=1}^{J_t} s_{jt} = s_{0t}$ *is the market share of the "outside" good available to all consumers in market* t*. We normalize the characteristics of the outside good such that* $x_{0t} = 0$ *and* $\xi_{0t} = 0$*.*

The preference of consumers can be summarized by a linear-in-characteristics random-utility model with a single-index unobserved quality. Thus each characteristic can be interpreted in terms of differences relative to the outside good. Specifically, each consumer i *has linear preferences for products* $j = 0, 1, ..., J_t$*:*

$$u_{ijt} = \delta_{jt} + \sum_{k=1}^{K_2} v_{ik} x_{jt,k}^{(2)} + \epsilon_{ijt},$$

where $\delta_{jt} = x'_{jt}\beta + \xi_{jt}$ *is the "mean utility" of product* j *such that* $E[\xi_{jt}|x_t] = 0$*,* $x_{jt}^{(2)}$ *is a sub-vector of* x_{jt} *(i.e. non-linear attributes),* $\epsilon_{ijt} \sim T1EV(0,1)$ *is an idiosyncratic utility shock for product* j*, and* $v_i = (v_{i1}, ..., v_{iK_2})$ *denotes the random coefficient vector with cdf* $F(v_i; \lambda_v)$*. Then the aggregate demand function for product* j *can be written as follows:*

$$\sigma_j(\delta_t, x_t^{(2)}; \lambda_v) = \int \frac{\exp(\sum_{k=1}^{K_2} v_{ik} x_{jt,k}^{(2)} + \delta_{jt})}{1 + \sum_{j'=1}^{J_t} \exp(\sum_{k=1}^{K_2} v_{ik} x_{j't,k}^{(2)} + \delta_{j't})} dF(v_i; \lambda_v), \tag{2.2}$$

where $x_t^{(2)} = \left(x_{1t}^{(2)}, ..., x_{J_t,t}^{(2)}\right)$ *and* $\delta_t = (\delta_{1t}, ..., \delta_{J_t,t})$*.*

Let $\sigma(\delta_t, x_t^{(2)}; \lambda_v) = \left(\sigma_1(\delta_t, x_t^{(2)}; \lambda_v), ..., \sigma_J(\delta_t, x_t^{(2)}; \lambda_v)\right)$*. Then*

$$s_t = \sigma(\delta_t, x_t^{(2)}; \lambda_v),$$

where $\theta_v = (\beta, \lambda_v)$ *is the full parameter vector of dimension* m. *Suppose the conditional distribution function of* $\xi_t = (\xi_{1t}, ..., \xi_{J_t,t})^\top$ *given* x_t *is non-degenerate of parametric form* $G(\cdot|x_t; \lambda_\xi)$. *Then likelihood-based estimation and inference may be applied provided that* σ *is invertible in* δ_t. *Unlike (2.1) in Example 1, the mapping the dimension of* s_t *is the same as the dimension of* δ_t. *However, the invertibility of* σ *is not automatic. Berry (1994); Berry et al. (1995) present sufficient conditions for the mapping* σ *to be invertible in* δ_t. *Examples of* $G(\cdot|x_t; \lambda_\xi)$ *include* $N(0, \tau^2 I_{J_t})$ *in Jiang et al. (2009) and the distribution of* $\eta_{jt} = \lambda_j' f_t + e_{jt}$ *in Moon et al. (2018), where* λ_j, f_t, e_t *are three vectors of latent variables following* $N(0, \sigma_\lambda^2 I_J)$, $N\left(0, \sigma_f^2 I_J\right)$, $N(0, \sigma_e^2 I_J)$ *respectively.*

To summarize, computing MLE based on a random sample $\{s_t, x_t\}_{t=1}^T$ *requires*

1. *the mapping* σ *to be invertible in* $\boldsymbol{\delta}_t$;
2. *the computation of the Jacobian matrix;*
3. *numerical computation of the integrals defining the mapping* σ;
4. *the likelihood to be tractable.*

   This is an example of the conditional model, where $Z_t = \left(s_t^\top, x_t^\top\right)^\top$. *It is also an example of the generative model. We note in passing that if the conditional distribution function of* $\xi_t = (\xi_{1t}, ..., \xi_{J_t,t})^\top$ *given* x_t *is degenerate, then we run into the same problem as Case (i) in Example 1.*

Example 2.2.3 (Parameter-Dependent Support Models)**.** *Consider the one-sided models in Chernozhukov and Hong (2004) and Hirano and Porter (2003) and two-sided models in Chernozhukov and Hong (2004). A scalar random variable* Y *given a vector of covariates* X *follows*

$$Y = g(X, \theta) + \epsilon,$$

where $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $\gamma \in \Gamma \subset \mathbb{R}^{d_\gamma}$ are finite-dimensional parameters, and the error term ϵ has conditional density function $f_\epsilon(\epsilon|X,\theta,\gamma)$. In one-sided models, $f_\epsilon(\epsilon|X,\theta,\gamma) = 0$ for all $\epsilon \leq 0$. Let $\mathcal{X} \subset R^{d_x}$ denote the support of X. Hirano and Porter (2003) assume that for X in some subset of $\mathcal{X}$ with positive probability, the conditional density of ϵ at its support boundary $\epsilon = 0$ is strictly positive. In two-sided models, one can express the conditional density of the error term ϵ given a vector of covariates X as

$$f_\epsilon(\epsilon|X,\theta,\gamma) := \begin{cases} f_{L,\epsilon}(\epsilon|X,\theta,\gamma) & \text{if } \epsilon < 0, \\ f_{U,\epsilon}(\epsilon|X,\theta,\gamma) & \text{if } \epsilon \geq 0. \end{cases}$$

Chernozhukov and Hong (2004) assume that for any $x \in \mathcal{X}$,

$$\begin{aligned} &\lim_{\epsilon \uparrow 0} f_\epsilon(\epsilon|x,\theta,\gamma) = f_{L,\epsilon}(0|x,\theta,\gamma),\ \lim_{y \downarrow 0} f(\epsilon|x,\theta,\gamma) = f_{U,\epsilon}(0|x,\theta,\gamma), \\ &f_{U,\epsilon}(0|x,\theta,\gamma) > f_{L,\epsilon}(0|x,\theta,\gamma) + \eta \text{ for some } \eta > 0, \\ &\text{for all } (\theta,\gamma) \in \Theta \times \Gamma. \end{aligned}$$

Many structural econometrics models lead to one-sided or two-sided models satisfying the assumptions in Hirano and Porter (2003) and Chernozhukov and Hong (2004) so that the non-normal asymptotic theory they develop applies albeit their implementation may be challenging. In other models, however, it may be difficult to check whether these assumptions hold. When there is no jump in the conditional density function, the non-normal asymptotic theory in Chernozhukov and Hong (2004) may not apply.

Let $F(y|x,\theta,\gamma)$ denote the conditional cdf of Y given $X = x$. For one-sided models,

$$F(y|x,\theta,\gamma) = \int_{g(x,\theta)}^{y} f(u|x,\theta,\gamma)\, du,\ y \geq g(x,\theta); \tag{2.3}$$

for two-sided models,

$$F(y|x,\theta,\gamma) = \begin{cases} \int_{-\infty}^{y} f_L(u|x,\theta,\gamma)\mathrm{d}u, & \text{if } y \leq g(x,\theta), \\ \int_{-\infty}^{g(x,\theta)} f_L(u|x,\theta,\gamma)\mathrm{d}u + \int_{g(x,\theta)}^{y} f_U(u|x,\theta,\gamma)\mathrm{d}u, & \text{if } y > g(x,\theta). \end{cases} \tag{2.4}$$

Let $\psi = (\theta,\gamma)' \in \Psi = \Theta \times \Gamma$. This is an example of the conditional model, where $Z_t = \left(Y_t^\top, X_t^\top\right)^\top$.

2.3 Minimum Sliced Distance Estimation and Examples

Minimum sliced distance estimation falls within the framework of minimum distance estimation based on a sliced distance between an "empirical measure" and a model induced measure. In this section, we propose a general MSD estimator which includes the MSWD esitmator, the MSCD estimator, and their simulated versions.

2.3.1 Sliced Wasserstein and Sliced Cramer Distances

Let $\mathcal{P}_2(\mathcal{Z})$ denote the space of probability measures with support $\mathcal{Z} \subset \mathbb{R}^d$ and finite second moments. For two probability measures μ and ν from $\mathcal{P}_2(\mathcal{Z})$, we denote by $\mathcal{W}_2(\mu, \nu)$ their 2-Wasserstein distance or simply Wassserstein distance. It is a finite metric on $\mathcal{P}_2(\mathcal{Z})$ defined by the optimal transport problem:

$$\mathcal{W}_2(\mu, \nu) = \left[\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|x - y\|^2 \mathrm{d}\gamma(x, y) \right]^{1/2},$$

where $\Gamma(\mu, \nu)$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν.

When $d = 1$, Proposition 2.17 in Santambrogio (2015) or Theorem 6.0.2 in Ambrosio et al. (2008) implies that

$$\mathcal{W}_2^2(\mu, \nu) = \int_0^1 \left(F_\mu^{-1}(s) - F_\nu^{-1}(s) \right)^2 \mathrm{d}s, \tag{2.5}$$

where $F_\mu(\cdot)$ and $F_\nu(\cdot)$ are the distribution functions associated with the measure μ and ν, respectively, and F_μ^{-1} and F_ν^{-1} are quantile functions. Unlike the 1-Wasserstein distance also known as the Earth Mover distance which equals to $\int_{-\infty}^{\infty} |F_\mu(s) - F_\nu(s)| \, \mathrm{d}s$, the 2-Wasserstein distance differs from the Cramer distance:

$$\mathcal{C}_2^2(\mu, \nu) = \int_{-\infty}^{\infty} (F_\mu(s) - F_\nu(s))^2 \mathrm{d}s. \tag{2.6}$$

In fact, the following result holds.

Lemma 2.3.1 (Theorem 2.11 in Bobkov and Ledoux (2019))**.** *Suppose μ and ν are probability measures in $\mathcal{P}_2(\mathbb{R})$. Then,*

$$\begin{aligned}\mathcal{W}_2^2(\mu,\nu) &= \int_{-\infty}^{\infty}(F_\mu(u\wedge v)-F_\nu(u\vee v))^+\mathrm{d}u\mathrm{d}v+\int_{-\infty}^{\infty}(F_\nu(u\wedge v)-F_\mu(u\vee v))^+\mathrm{d}u\mathrm{d}v\\ &= 2\iint_{u\le v}\left[(F_\mu(u)-F_\nu(v))^+ + (F_\nu(u)-F_\mu(v))^+\right]\mathrm{d}u\mathrm{d}v. \end{aligned} \tag{2.7}$$

When $d>1$, the Wasserstein distance $\mathcal{W}_2$ is computationally costly, and the sliced Wasserstein distance is introduced in the literature to ease computational burden associated with the Wasserstein distance $\mathcal{W}_2$ (c.f. Bonneel et al. (2015)). Let $\mathbb{S}^{d-1}=\{u\in\mathbb{R}^d:\|u\|_2=1\}$ be the unit-sphere in $\mathbb{R}^d$. For $u\in\mathbb{S}^{d-1}$ and $z\in\mathbb{R}^d$, let $u^*(z)=u^\top z$ be the 1D (or scalar) projection of z to u. For a probability measure μ, we denote by $u^*_\sharp\mu$ the push-forward measure of μ by u^*. The sliced Wasserstein distance $\mathcal{SW}(\mu,\nu)$ is defined as follows:

$$\mathcal{SW}(\mu,\nu)=\left[\int_{\mathbb{S}^{d-1}}\mathcal{W}^2(u^*_\#\mu,u^*_\#\nu)d\varsigma(u)\right]^{1/2}, \tag{2.8}$$

where $\varsigma(u)$ is the uniform distribution on $\mathbb{S}^{d-1}$. It is well known that SW is a well-defined metric, see Nadjahi et al. (2020a). Also, the sample complexity of the sliced distance does not depend on the dimension d (Nadjahi et al. (2020a)). We can see later that the MSWD estimators converges to the true parameter with $n^{-1/2}$, which is independent of the dimension. This was already discussed in Nadjahi et al. (2020b) with the sliced 1-Wasserstein distance.

For each $u\in\mathbb{S}^{d-1}$, let

$$G_\mu(s;u)=\int I(u^\top z\le s)\mathrm{d}F_\mu(z).$$

Define $G_\nu(s;u)$ similarly. Since

$$\mathcal{W}^2(u^*_\#\mu,u^*_\#\nu)=\int_0^1\left(G_\mu^{-1}(s;u)-G_\nu^{-1}(s;u)\right)^2\mathrm{d}s,$$

we obtain that

$$\mathcal{SW}(\mu,\nu)=\left[\int_{\mathbb{S}^{d-1}}\int_0^1\left(G_\mu^{-1}(s;u)-G_\nu^{-1}(s;u)\right)^2\mathrm{d}s\mathrm{d}\varsigma(u)\right]^{1/2}.$$

We introduce a weighted Sliced Wasserstein distance:

$$\mathcal{SW}_w(\mu,\nu) = \left[\int_{\mathbb{S}^{d-1}}\int_0^1 \left(G_\mu^{-1}(s;u) - G_\nu^{-1}(s;u)\right)^2 w\left(s\right) \mathrm{d}s d\varsigma(u)\right]^{1/2},$$

where $w(s)$ is a nonnegative function such that $\int_0^1 w(s)\, ds = 1$. Similarly, we define the weighted Sliced Cramer distance:

$$S\mathcal{C}_w(\mu,\nu) = \left(\int_{\mathbb{S}^{d-1}}\int_{-\infty}^{\infty} (G_\mu(s;u) - G_\nu(s;u))^p w(t) \mathrm{d}s d\varsigma(u)\right)^{1/p}.$$

2.3.2 The General Estimator and Examples

Given the sample information $\{Z_t\}_{t=1}^T$, let $Q_T(\cdot;u)$ denote an empirical measure such as the empirical quantile or distribution function of $\left\{u^\top Z_t\right\}_{t=1}^T$ and $\widehat{Q}_T(\cdot;u,\psi)$ denote a possibly random function depending on the model, $\psi \in \Psi \subset \mathbb{R}^{d_\psi}$. *A general minimum sliced distance* (MSD hereafter) estimator denoted by $\hat{\psi}_T$ is defined as

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u) - \widehat{Q}_T(s;u,\hat{\psi}_T))^2 w(s) \mathrm{d}s d\varsigma(u) \\
= \ & \inf_{\psi\in\Psi}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u) - \widehat{Q}_T(s;u,\psi))^2 w(s) \mathrm{d}s d\varsigma(u) + o_p(T^{-1}),
\end{aligned} \tag{2.9}$$

where $\mathcal{S}$ is the domain of $Q_T(s;u)$ with respect to s.

When $Q_T(s;u)$ is the empirical quantile function of $\left\{u^\top Z_t\right\}_{t=1}^T$ and $\widehat{Q}_T(\cdot;u,\psi)$ is the model induced quantile function, $\hat{\psi}_T$ is the minimum weighted SW distance (MSWD hereafter) estimator; when $Q_T(s;u)$ is the empirical distribution function of $\left\{u^\top Z_t\right\}_{t=1}^T$ and $\widehat{Q}_T(\cdot;u,\psi)$ is the model induced distribution function, $\hat{\psi}_T$ is the minimum weighted SC distance (MSCD hereafter) estimator.

The form of $\widehat{Q}_T(\cdot;u,\psi)$ also differs for unconditional, conditional, and generative models. For *unconditional models* such as Example 2.2.1, $\widehat{Q}_T(\cdot;u,\psi)$ is deterministic denoted as $Q(\cdot;u,\psi)$ such as parametric quantile function of $u^\top Z_t$ induced by the parametric distribution of Z_t for MSWD or the parametric distribution function of $u^\top Z_t$ for MSCD. Regardless of the invertibility of the mapping h in (2.1), the distribution function and quantile function of Z_t always exist and our estimators are always well defined. This is in sharp contrast to MLE.

Example 2.2.1 (Cont'd). We will consider one-factor discrete-time Vasicek model in Section 6 of Backus et al. (1998). In the one-factor discrete-time Vasicek model, the short-term interest rate Y_t^* follows

$$(Y_t^* - c) = \rho(Y_{t-1}^* - c) + \sigma\epsilon_t,$$

where $0 < \rho < 1$, c and σ are positive constants, and ϵ_t follows the standard normal distribution.

The yield $Y_t(\tau)$ of zero-coupon bond at t of maturity τ is determined by

$$Y_t(\tau) = a(\tau) + b(\tau)Y_t^*,$$

where $a(\tau) = A(\tau)/\tau$, and $b(\tau) = B(\tau)/\tau$ in which $A(\tau)$ and $B(\tau)$ are calculated by recursion: $A(1) = 0$, $B(1) = 1$, and

$$\begin{aligned} A(\tau) &= A(\tau-1) + \frac{1}{2}\lambda^2 + B(\tau-1)(1-\rho)c - \frac{1}{2}(\lambda + B(\tau-1)\sigma)^2, \\ B(\tau) &= 1 + B(\tau-1)\rho. \end{aligned}$$

When we consider $\tau = \tau_1, \ldots, \tau_K$, we have

$$Y_t = a + bY_t^*, \tag{2.10}$$

where

$$Y_t = \begin{pmatrix} Y_t(\tau_1) & \ldots & Y_t(\tau_K) \end{pmatrix}^\top, \quad a = \begin{pmatrix} a(\tau_1) & \ldots & a(\tau_K) \end{pmatrix}^\top, \quad b = \begin{pmatrix} b(\tau_1) & \ldots & b(\tau_K) \end{pmatrix}^\top.$$

Let $Z_t = (Y_t^\top, Y_{t-1}^\top)^\top$. For each $u \in \mathbb{S}^{2d_Y-1}$, the projected variable $u^\top Z_t$ follows a normal distribution:

$$u^\top Z_t \sim N\left(u^\top \begin{pmatrix} a+bc \\ a+bc \end{pmatrix}, \frac{\sigma^2}{1-\rho^2} u^\top \left[\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \otimes bb^\top \right] u \right),$$

where $\otimes$ is the Kronecker product. Both the distribution function and quantile function have closed form expressions. We will use this in the numerical section to estimate the model. Compared to conventional approaches (such as Pastorello et al. (2003)), we will estimate the parameters with measurement error to the equation (2.10).

For *conditional models* such as Examples 2.2.2 and 2.2.3, $\widehat{Q}_T(\cdot; u, \psi)$ is random.

Example 2.2.3 (Cont'd). Let $Z = \left(Y, X^\top\right)^\top$. Then the cdf of Z and $u^\top Z$ are given by

$$F\left(z; \psi\right) = \mathbb{E}\left[F\left(y|X, \psi\right) I\left(X \leq x\right)\right] \text{ and}$$

$$\begin{aligned}
G(s; u, \psi) &= \Pr(u_1 Y + u_2^\top X \leq s) \\
&= \mathbb{E}\left[\int_{-\infty}^{\infty} I(u_1 y + u_2^\top X \leq s) f(y|X, \psi) \mathrm{d}y\right] \\
&= \begin{cases} \mathbb{E}\left[F(u_1^{-1}(s - u_2^\top X)|X, \psi)\right] & \text{if } u_1 > 0 \\ \mathbb{E}\left[I(u_2^\top X \leq s)\right] & \text{if } u_1 = 0 \\ 1 - \mathbb{E}\left[F(u_1^{-1}(s - u_2^\top X)|X, \psi)\right] & \text{if } u_1 < 0, \end{cases}
\end{aligned}$$

where $F(y|X, \psi)$ is given in equations 2.3 and 2.4. For the MSCD estimator, $\widehat{Q}_T(\cdot; u, \psi) = \widehat{G}_T(\cdot; u, \psi)$; for the MSWD estimator, $\widehat{Q}_T(\cdot; u, \psi) = \widehat{G}_T^{-1}(\cdot; u, \psi)$, where

$$\begin{aligned}
\widehat{G}_T(s; u, \psi) &= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} I(u_1 y + u_2^\top X_i \leq s) f(y|X_i, \psi) \mathrm{d}y \\
&= \begin{cases} \frac{1}{T} \sum_{t=1}^{T} F(u_1^{-1}(s - u_2^\top X_t)|X_t, \psi) & \text{if } u_1 > 0 \\ \frac{1}{T} \sum_{t=1}^{T} I(u_2^\top X_t \leq s) & \text{if } u_1 = 0 \\ 1 - \frac{1}{T} \sum_{t=1}^{T} F(u_1^{-1}(s - u_2^\top X_t)|X_t, \psi) & \text{if } u_1 < 0. \end{cases}
\end{aligned}$$

For generative models, $\widehat{Q}_T(\cdot; u, \psi)$ is random as well.

Example 2.3.1. ***Generative Model***

The distribution function or quantile function of the observed variables in complex structural models may not be available in closed form, but the model may be easily simulated for *any parameter value $\psi \in \Psi$. We follow the machine learning literature and call such models generative models. For generative models, let*

$$\widehat{Q}_T(\cdot; u, \psi) := \frac{1}{K} \sum_{k=1}^{K} \widehat{Q}_m^{(k)}(\cdot; u, \psi),$$

where for each $k = 1, \ldots, K$, we generate a simulated sample with size m from the parametric model with parameter ψ and then construct $\widehat{Q}_m^{(k)}(\cdot; u, \psi)$ from the simulated sample. Here, m is a function of T, and $\widehat{Q}_m^{(k)}(\cdot; u, \psi)$ is independent of $Q_T(\cdot; u)$. Let $\hat{\psi}_{T,m}$ satisfy

$$\begin{aligned} &\int_{\mathbb{S}^{d-1}} \int (Q_T(s; u) - \widehat{Q}_T(s; u, \hat{\psi}_{T,m}))^2 w(s) \mathrm{d}s d\varsigma(u) \\ &= \inf_{\psi \in \mathcal{H}} \int_{\mathbb{S}^{d-1}} \int (Q_T(s; u) - \widehat{Q}_T(s; u, \psi))^2 w(s) \mathrm{d}s d\varsigma(u) + \nu_T, \end{aligned} \tag{2.11}$$

where $\nu_T = o_p(T^{-1})$. We call $\hat{\psi}_{T,m}$ a simulated minimum sliced distance (SMSD) estimator.

2.3.3 Adversarial Perspective

Both the MSWD and MSCD estimators are variants of the GAN estimator originally proposed in Goodfellow et al. (2014) and extended to structural models in Kaji et al. (2020). The population criterion function for the original GAN estimator is the Jensen-Shannon divergence between the true distribution and the distribution induced by the generative model. Arjovsky et al. (2017) propose Wasserstein GAN based on the Earth Mover or 1-Wasserstein distance and implement it using the dual form of the 1-Wasserstein distance. Lei et al. (2019) develop Wasserstein GAN based on the 2-Wasserstein distance. The main contribution of Lei et al. (2019) is to show via Brenier theorem that the optimal discriminator is in fact uniquely determined by the generator so the optimization can be carried out in one step. However, since the dimension of the observed variables is typically larger than the dimension of the latent variables in generative models, in order to apply Brenier theorem, Lei et al. (2019) propose to first transform the empirical measure of the data to a corresponding measure for the latent variables via *an encoding map* and then use the 2-Wasserstein distance between the empirical measure of the latent variables obtained from applying the encoding map and that induced by the generative model. For structural models, it is unclear how to find the encoding map. Besides, the 2-Wasserstein distance is costly to compute when the dimension is high. We demonstrate below that our MSWD estimator overcomes the potential drawbacks of the Wasserstein GAN in Lei et al. (2019): no encoding map is needed; since the two distributions being compared are of the same dimension 1 and the sliced WD is easy to compute.

Let μ be probability measure, and $\mu(\psi)$ be the model-induced probability measure which depends on parameter ψ. From the dual formulation for $\mathcal{W}^2(u^*_{\#}\mu, u^*_{\#}\mu\,(\psi))$, we obtain that

$$\mathcal{W}^2(u^*_{\#}\mu, u^*_{\#}\mu\,(\psi)) = \max_{\varphi}\left[\int \varphi\,(x)\,du^*_{\#}\mu\,(x) + \int \varphi^*\,(y)\,du^*_{\#}\mu\,(y;\psi)\right], \tag{2.12}$$

where φ^* is the Legendre-Fenchel transform of φ defined as

$$\varphi^*\,(y) := \sup_x \left(x^\top y - \varphi\,(x)\right).$$

It is known from Brenier theorem that for any fixed $\psi \in \Psi$, provided that $u^*_{\#}\mu\,(\psi)$ has a continuous distribution, there exists a unique transport map T such that

$$\nabla\varphi_u\,(x;\psi) = x - T\,(x;u,\psi) = x - Q(G(s;u,\psi);u) \tag{2.13}$$

where $\varphi_u\,(\cdot;\psi)$ is a Kantorovich potential and

$$\mathcal{W}^2(u^*_{\#}\mu, u^*_{\#}\mu\,(\psi)) = \int (Q(s;u) - Q(s;u,\psi))^2 \mathrm{d}s.$$

The population criterion function for the MSWD estimator without weighting can then be written as

$$\begin{aligned}
SW^2_w(\mu,\mu(\psi)) &= \int_{\mathbb{S}^{d-1}} \left[\int (Q(s;u) - Q(s;u,\psi))^2 \mathrm{d}s\right] d\varsigma(u) \\
&= \int_{\mathbb{S}^{d-1}} \mathcal{W}^2(u^*_{\#}\mu, u^*_{\#}\mu\,(\psi))d\varsigma(u) \\
&= \int_{\mathbb{S}^{d-1}} \max_{\varphi}\left[\int \varphi\,(x)\,du^*_{\#}\mu\,(x) + \int \varphi^*\,(y)\,du^*_{\#}\mu\,(y;\psi)\right] d\varsigma(u).
\end{aligned}$$

The MSWD estimator can be interpreted as a sliced Wasserstein GAN estimator:

Step 1. Given ψ, find the optimal discriminator or a Kantorovich potential $\varphi_u\left(\cdot;\psi\right)$ and compute the optimal sample objective function for the discriminator $\mathcal{W}^2(u^*_{\#}\mu, u^*_{\#}\mu\left(\psi\right))$ at each direction $u \in \mathbb{S}^{d-1}$;

Step 2. Minimize the integrated optimal sample objective function: $\int_{\mathbb{S}^{d-1}} \mathcal{W}^2(u^*_{\#}\mu, u^*_{\#}\mu\left(\psi\right))d\varsigma(u)$ with respect to ψ.

Given the expression for $\varphi_u\left(\cdot;\psi\right)$ in (2.13) or equivalently the expression for $\mathcal{W}^2(u^*_{\#}\mu, u^*_{\#}\mu\left(\psi\right))$, Step 1 is in fact redundant.

Similarly, using the dual form of the Cramer distance as an integral probability metric (see Bellemare et al. (2017)), the MSCD estimator can be computed in two steps as well.

2.4 Singular, One-sided, and Two-Sided Uniform Models

In contrast to the stochastic singular asset pricing model with $d_y > d^*$, MLE exists for the one-sided and two-sided parameter dependent support models in Example 2.3. However, the asymptotic theory is very different from the classical textbook theory and is crucially dependent on model assumptions. On the other hand, the MSWD and MSCD estimators we propose in the previous section are shown to be consistent and asymptotically normal under very mild conditions for all these models in later sections of this paper.

In this section, we use three simple examples to illustrate the different behaviors of MLE, the original oracle GAN estimator referred to as the JS estimator, and our estimators. The JS estimator is the oracle GAN estimator, which minimizes the JS divergence (See Proposition 1 of Goodfellow et al. (2014) and Example 3 in Kaji et al. (2020).):

$$\hat{\theta}_{JS} = \frac{1}{2T}\sum_{t=1}^{T}\log\left(\frac{2f(y_t;\theta_0)}{f(y_t;\theta_0)+f(y_t;\theta}\right) + \frac{1}{2}\int\log\left(\frac{2f(y;\theta)}{f(y;\theta_0)+f(y;\theta)}\right)f(y;\theta)\mathrm{d}y,$$

where $f(y_t,\theta)$ the density function of model-induced parametric distribution. Here, we consider the oracle estimator when the size of the simulated sample is infinity to minimize variation caused by the simulated sample.

We first compute the KL divergence, the JS divergence, and SW, SC distances and show that neither KL divergence nor JS divergence exhibits the same smoothness property for all three examples and all parameter values. On the contrary, the SW and SC distances[5] exhibit consistent behavior for all examples and parameter values. We then draw QQ plots of MLE, JS, MSWD, and MSCD estimators.

2.4.1 *Example 1 in Arjovsky et al. (2017)*

When μ is absolutely continuous with respect to ν,

$$KL(\mu|\nu) = \int_{-\infty}^{\infty} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) d\mu \text{ and}$$

$$JS(\mu|\nu) = \frac{1}{2}\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}(0.5\mu + 0.5\nu)}\right) \mathrm{d}\mu + \frac{1}{2}\int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}(0.5\mu + 0.5\nu)}\right) \mathrm{d}\nu,$$

where $d\mu/d\nu$ is Radon-Nikodym derivative of μ with respect to ν.

For stochastic singular models such as the asset pricing model in Example 2.2.1 when $d^* < d_y$, the support of the distribution of the observed asset prices lies within a low-dimensional manifold in a high-dimensional space and its distribution is not absolutely continuous with respect to any fixed measure. As a result, MLE is not defined, see Arjovsky and Bottou (2017) for general discussions. Example 1 in Arjovsky et al. (2017) illustrates this point nicely.

Let $Y = g_\theta(Z) = (\theta_0, Z)$, where $\theta_0 = 0$ and $Z \sim U[0, 1]$. Denote the CDF of Z by $F_U(z)$, and the CDF of Y by $F_Y(y) = F_Y(x, z; \theta_0)$. Then

$$F_Y(x, z; \theta_0) = \begin{cases} F_U(z) & \text{if } x \geq \theta_0, \\ 0 & \text{if } x < \theta_0. \end{cases}$$

Since the support of $F_Y(x, z; \theta)$ is $\{\theta\} \times [0, 1]$, $F_Y(x, z; 0)$ and $F_Y(x, z; \theta)$ have disjoint supports unless $\theta = 0$. Arjovsky et al. (2017) show that

$$KL\left(\mu_0|\mu_\theta\right) = \begin{cases} +\infty & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases} \text{ and } JS = \begin{cases} \log 2 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0. \end{cases}$$

[5]The SC distance may not have a finite second-order derivative at θ_0.

Thus, MLE is not defined. Although GAN is defined, it exhibits instable behavior, see Goodfellow et al. (2014), and JS estimator is not defined.

Straightforward calculations show that

$$\begin{aligned}
\mathcal{W}_2^2(\mu_0, \mu_\theta) &= \theta^2,\ \mathcal{SW}^2(\mu_0, \mu_\theta) = \frac{\theta^2}{2}, \\
\mathcal{C}_2^2(\mu_0, \mu_\theta) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (F_Y(x, z; 0) - F_Y(x, z; \theta))^2 \mathrm{d}z\mathrm{d}x = \frac{|\theta|}{3}, \\
\mathcal{SC}^2(\mu_0, \mu_\theta) &= \begin{cases} 0 & \text{if } \theta = 0, \\ \frac{2\left(2|\theta|^3 - 1 - (2\theta^2-1)\sqrt{\theta^2+1} - 3\theta^2 \log(\tan(0.5\tan^{-1}(|\theta|)))\right)}{3\pi} & \text{otherwise.} \end{cases}
\end{aligned}$$

In contrast to KL and JS divergences, Wasserstein distance and Cramer distance as well as their sliced versions are continuous. In addition, Wasserstein distance and the sliced WD are second-order differentiable; Sliced Cramer distance is first-order differentiable at $\theta_0 = 0$ but does not have a finite second-order derivative at $\theta_0 \neq 0$.

Remark 2.4.1. The sliced CD has better smoothness property than the Cramer distance, since the Cramer distance is not first-order differentiable at $\theta_0 = 0$. Furthermore, the sample Cramer distance is not well defined for all θ without weighting:

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (F_T(x, z) - F(x, z; \theta))^2 \mathrm{d}z\mathrm{d}x &\geq \int_{|\theta|}^{\infty}\int_{-\infty}^{\infty} (F_T(x, z) - F(x, z; \theta))^2 \mathrm{d}z\mathrm{d}x \\
&= \int_{|\theta|}^{\infty}\int_{-\infty}^{\infty} \Big(F_T(z) - F_U(z)\Big)^2 \mathrm{d}z\mathrm{d}x \\
&= \infty.
\end{aligned}$$

In the following two examples, the KL divergence is well defined and bounded so that MLE exists. However, the first-order derivative of KL divergence may or may not exist resulting in different asymptotic theory for MLE. In contrast, both SWD and SCD are twice continuously differentiable for all parameter values.

2.4.2 A One-sided Uniform Model

Let $Y \sim U[0, \theta_0]$, where $\theta_0 > 0$ is the true parameter. Its distribution function and density function are given by

$$F(y;\theta) = \frac{y}{\theta} I\left(y \leq \theta\right) \text{ and } f(y;\theta) = \frac{1}{\theta} I\left(0 \leq y \leq \theta\right) \text{ for } \theta > 0.$$

Straightforward algebra shows that

$$KL(\mu_0|\mu_\theta) = \int_0^{\theta_0} f(y;\theta_0)\log\left(\frac{f(y;\theta_0)}{f(y;\theta)}\right)dy = \begin{cases} \infty & \text{if } \theta < \theta_0, \\ \log(\theta) - \log(\theta_0) & \text{if } \theta \geq \theta_0. \end{cases}$$

Given a random sample $\{Y_t\}_{t=1}^T$ on Y, Hirano and Porter (2003) studies efficiency of the MLE and BE of θ_0. Specifically, the MLE is given by $Y_{(T)}$, the maximum order statistic of $\{Y_t\}_{t=1}^T$ and

$$T\left(Y_{(T)} - \theta_0\right) \xrightarrow{d} -\text{Exp}\left(\frac{1}{\theta_0}\right),$$

where $\text{Exp}\left(\frac{1}{\theta_0}\right)$ is the exponential distribution with rate parameter $1/\theta_0$.

Similar calculations for the JS divergence lead to

$$JS(\mu_0|\mu_\theta) = \begin{cases} \frac{1}{2}\log\left(\frac{\theta_0}{\theta+\theta_0}\right) + \frac{1}{2}\frac{\theta}{\theta_0}\log\left(\frac{\theta}{\theta+\theta_0}\right) + \log 2 & \text{if } \theta \leq \theta_0, \\ \frac{1}{2}\frac{\theta_0}{\theta}\log\left(\frac{\theta_0}{\theta+\theta_0}\right) + \frac{1}{2}\log\left(\frac{\theta}{\theta+\theta_0}\right) + \log 2 & \text{if } \theta > \theta_0. \end{cases}$$

Although $JS(\mu_0|\mu_\theta)$ is continuous at θ_0, its first order left and right derivatives at $\theta = \theta_0$ are different:

$$\lim_{\theta\uparrow\theta_0}\frac{\partial JS(\mu_0|\mu_\theta)}{\partial\theta} = -\frac{\log 2}{\theta_0} < 0 \text{ and } \lim_{\theta\downarrow\theta_0}\frac{\partial JS(\mu_0|\mu_\theta)}{\partial\theta} = \frac{\log 2}{\theta_0} > 0.$$

Straighforward calculations show that

$$\mathcal{W}_2^2(\mu_\theta, \mu_0) = \frac{1}{3}(\theta-\theta_0)^2 \text{ and } \mathcal{C}_2^2(\mu_\theta, \mu_0) = \begin{cases} \frac{(\theta-\theta_0)^2}{3\theta_0} & \text{if } \theta \leq \theta_0, \\ \frac{(\theta-\theta_0)^2}{3\theta} & \text{if } \theta > \theta_0. \end{cases}$$

In contrast to $JS(\mu_0|\mu_\theta)$, both $\mathcal{W}_2^2(\mu_\theta, \mu_0)$ and $\mathcal{C}_2^2(\mu_\theta, \mu_0)$ are twice continuously differentiable at θ_0. Although $\mathcal{C}_2^2(\mu_\theta, \mu_0)$ has different expressions for $\theta \leq \theta_0$ and $\theta_0 < \theta$, it is twice continuously differentiable at θ_0 with first and second order derivatives given by

$$\left.\frac{\partial \mathcal{C}_2^2(\mu_\theta, \mu_0)}{\partial \theta}\right|_{\theta=\theta_0} = 0 \text{ and } \left.\frac{\partial^2 \mathcal{C}_2^2(\mu_\theta, \mu_0)}{\partial \theta^2}\right|_{\theta=\theta_0} = \frac{2}{3\theta_0}.$$

2.4.3 A Two-sided Uniform Model

Suppose the support of Y is $[0, 1]$ and its density function is

$$f(y; \theta_0) = \begin{cases} \frac{1}{4\theta_0} & \text{if } 0 \le y \le \theta_0, \\ \frac{3}{4(1-\theta_0)} & \text{if } \theta_0 < y \le 1. \end{cases}$$

The density function has a jump at θ_0 with jump size $\frac{4\theta_0 - 1}{4\theta_0(1-\theta_0)}$ which is zero when $\theta_0 = 1/4$ and non-zero otherwise. The CDF of Y is given by

$$F(y; \theta) = \begin{cases} \frac{y}{4\theta} & \text{if } 0 \le y \le \theta, \\ \frac{1}{4} + \frac{3(y-\theta)}{4(1-\theta)} = 1 - \frac{3(1-y)}{4(1-\theta)} & \text{if } \theta < y \le 1 \end{cases}$$

with quantile function

$$Q(p; \theta) = \begin{cases} 4\theta p & \text{if } 0 \le p \le 1/4, \\ 1 - \frac{4}{3}(1-\theta)(1-p) & \text{if } 1/4 \le p \le 1. \end{cases}$$

The quantile function $Q(p; \theta)$ is linear in θ and has a kink at $1/4$.

Tedious but straightforward calculations yield

$$KL(\mu_0|\mu_\theta) = \begin{cases} \frac{\theta}{4\theta_0} \log\left(\frac{\theta}{\theta_0}\right) + \frac{(\theta_0-\theta)}{4\theta_0} \log\left(\frac{1-\theta}{3\theta_0}\right) + \frac{3}{4} \log\left(\frac{1-\theta}{1-\theta_0}\right) & \text{if } \theta < \theta_0, \\ \frac{1}{4} \log\left(\frac{\theta}{\theta_0}\right) + \frac{3(\theta-\theta_0)}{4(1-\theta_0)} \log\left(\frac{3\theta}{1-\theta_0}\right) + \frac{3(1-\theta)}{4(1-\theta_0)} \log\left(\frac{1-\theta}{1-\theta_0}\right) & \text{if } \theta > \theta_0. \end{cases}$$

The first-order left and right derivatives are:

$$\lim_{\theta \uparrow \theta_0} \frac{\partial KL(\mu_0|\mu_\theta)}{\partial \theta} = \frac{4\theta_0 - (\theta_0 - 1)\log\left(\frac{1-\theta_0}{3\theta_0}\right) - 1}{4(\theta_0 - 1)\theta_0} \le 0,$$

$$\lim_{\theta \downarrow \theta_0} \frac{\partial KL(\mu_0|\mu_\theta)}{\partial \theta} = \frac{4\theta_0 - 3\theta_0 \log\left(\frac{1-\theta_0}{3\theta_0}\right) - 1}{4(\theta_0 - 1)\theta_0} \ge 0.$$

The left and right derivatives imply that $KL(\mu_0|\mu_\theta)$ has a local minimum at $\theta = \theta_0$. However, it is not smooth at $\theta = \theta_0$ unless $\theta_0 = 1/4$. In the latter case,

$$\lim_{\theta\uparrow\theta_0} \frac{\partial KL(\mu_0|\mu_\theta)}{\partial\theta} = 0 = \lim_{\theta\downarrow\theta_0} \frac{\partial KL(\mu_0|\mu_\theta)}{\partial\theta}.$$

Otherwise, when $\theta_0 \neq 1/4$,

$$\lim_{\theta\uparrow\theta_0} \frac{\partial KL(\mu_0|\mu_\theta)}{\partial\theta} < 0, \;\; \lim_{\theta\downarrow\theta_0} \frac{\partial KL(\mu_0|\mu_\theta)}{\partial\theta} > 0.$$

To summarize, the KL divergence is first-order differentiable only at $\theta_0 = 1/4$; at other values for θ_0, it has left and right derivatives, but they are unequal. When $|\theta_0 - 1/4| > \eta > 0$, an application of Chernozhukov and Hong (2004) implies that MLE is asymptotically non-normally distributed. Tedious computation shows that JS divergence has the same smothness property as the KL divergence and we expect the same asymptotic properties for the JS estimator as MLE.

On the other hand,

$$\begin{aligned} \mathcal{W}_2^2(\mu_\theta, \mu_0) &= \frac{1}{3}(\theta-\theta_0)^2 \text{ and} \\ \mathcal{C}_2^2(\mu_\theta, \mu_0) &= \begin{cases} \frac{(4\theta-12\theta_0-1)(\theta-\theta_0)^2}{48(\theta-1)\theta_0} & \text{if } \theta \leq \theta_0, \\ \frac{(4\theta_0-12\theta-1)(\theta-\theta_0)^2}{48(\theta_0-1)\theta} & \text{if } \theta_0 < \theta. \end{cases} \end{aligned}$$

Both are twice continuously differentiable at $\theta_0 \in (0,1)$. For $\mathcal{C}_2^2(\mu_\theta, \mu_0)$, it has first and second order derivatives given by

$$\left.\frac{\partial \mathcal{C}_2^2(\mu_\theta, \mu_0)}{\partial\theta}\right|_{\theta=\theta_0} = 0 \text{ and } \left.\frac{\partial^2 \mathcal{C}_2^2(\mu_\theta, \mu_0)}{\partial^2\theta}\right|_{\theta=\theta_0} = \frac{(8\theta_0+1)}{24\theta_0(1-\theta_0)} > 0.$$

2.4.4 Numerical Illustration

The table below summarizes the smoothness properties of the four divergences/distances.

To illustrate the different distributions of the MLE, GAN, MSWD, and MSCD estimators, we present four figures. In each figure, the left panel plots the divergences/distances and the right panel presents the corresponding QQ plots based on 3000 estimates of t-values

Table 2.1: A Comparison of Several Divergences

Distance		KL			JS			SW			SC		
		Continuity	FD	SD	Continuity	FD	SD	Continuity	FD	SD	Continuity	FD	SD
Ex1 in Arjovsky et al. (2017)		x	x	x	x	x	x	✓	✓	✓	✓	✓	x
One-sided Uniform		✓	x	x	✓	x	x	✓	✓	✓	✓	✓	✓
Two-sided Uniform	$\theta_0 = 1/4$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	$\theta_0 \neq 1/4$	✓	x	x	✓	x	x	✓	✓	✓	✓	✓	✓

FD and SD in the table mean whether the divergence measure is first and second-order differentiable with respect to θ at $\theta = \theta_0$, respectively.

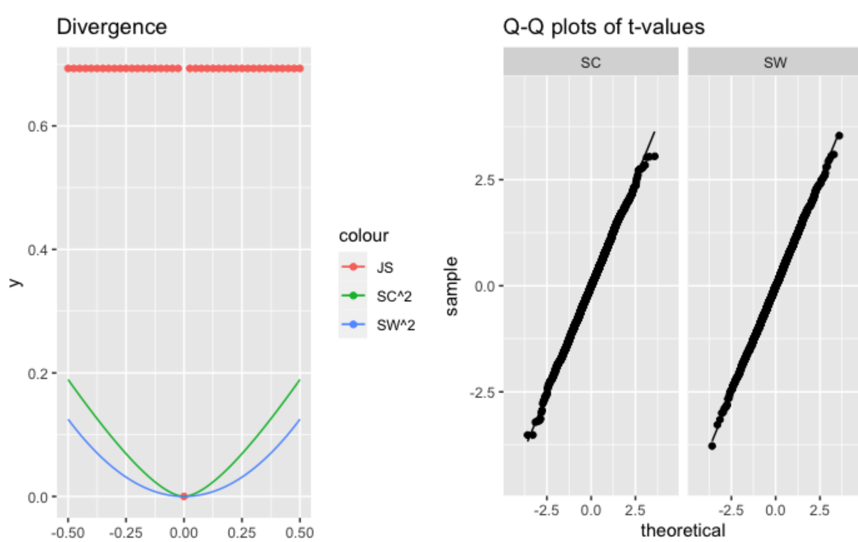

Figure 2.1: A comparison of several divergences for Example 1 in Arjovsky et al. (2017) with $T = 1000$

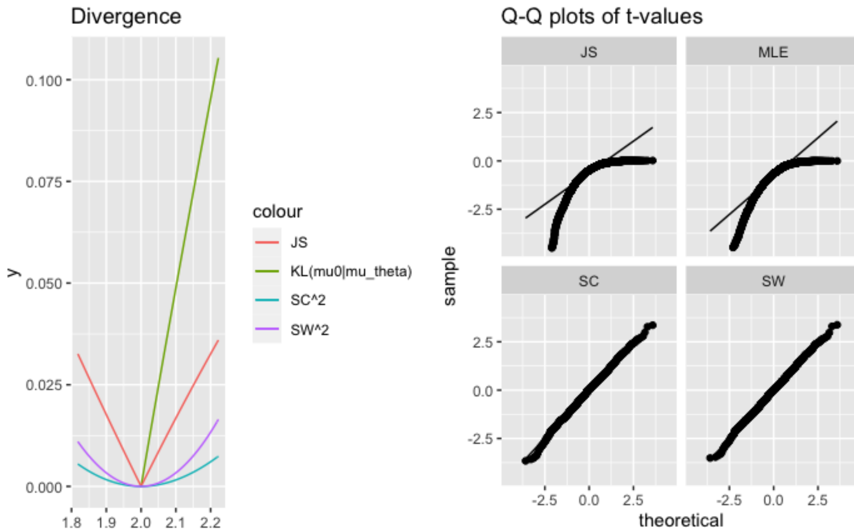

Figure 2.2: A comparison of several divergences when $\theta_0 = 2$ with $T = 1000$ for one-sided uniform model.

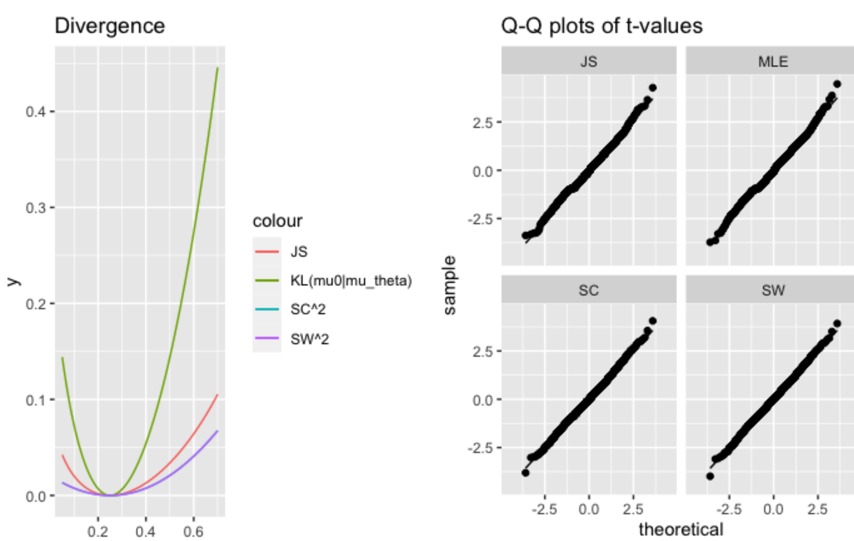

Figure 2.3: A comparison of several divergences when $\theta_0 = 1/4$ with $T = 1000$ for two-sided uniform model.

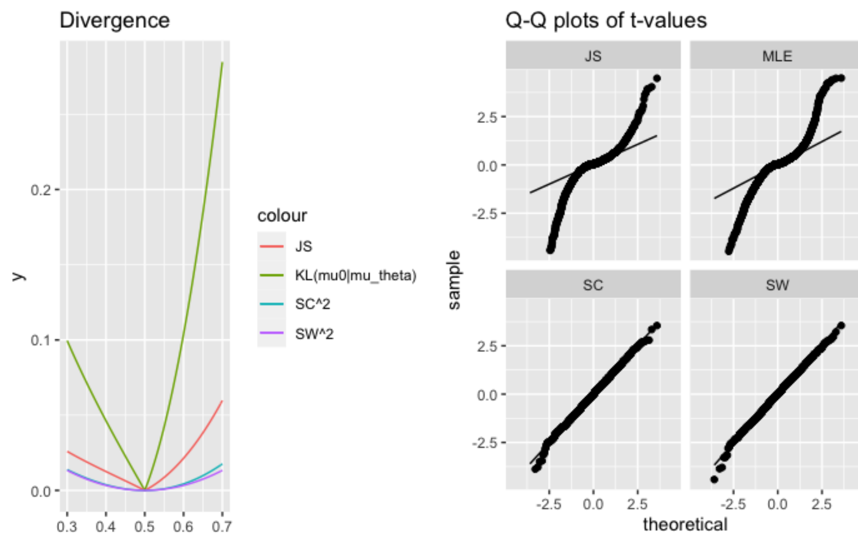

Figure 2.4: A comparison of several divergences when $\theta_0 = 1/2$ with $T = 1000$ for two-sided uniform model

with each value computed from a random sample of size 1000, see Figures 2.1 to 2.4. In the simulation, the standard deviations from the parameter estimates of Monte-Carlo simulation are used in the computation of t-values.

Several conclusions can be drawn from Figures 2.1 and 2.4. First, for the singular example in Arjovsky et al. (2017), MLE and JS estimator are not defined; both MSWD and MSCD are close to being normally distributed. Second, for the one-sided uniform model, both MLE and JS estimator are non-normally distributed. Again both MSWD and MSCD are close to being normally distributed. Third, for the two sided uniform model, when $\theta_0 = 1/4$, all four estimators are close to being normal; when θ is far from 1/4, MLE and JS estimator are non-normal and the distributions of MSWD and MSCD estimators are close to being normal. Indeed, we will verify in Section B.5 in the appendix that the assumptions for consistency and asymptotic normality of MSWD and MSCD estimators in Section 5 are satisfied in one-sided and two-sided models. For the singular model in Example 1 of Arjovsky et al. (2017), we will verify assumptions for the MSWD estimator.

2.5 *Asymptotic Theory*

In this section, we establish asymptotic theory for the general MSD estimator $\hat{\psi}_T$ defined in (2.9) under a set of high-level assumptions on $Q_T(s;u)$ and $\widehat{Q}_T(s;u,\psi)$. The high-level assumptions allow for broad classes of models and data types. They will be verified under primitive conditions for unconditional, conditional, and generative models in subsequent sections.

2.5.1 Assumptions and Asymptotic Properties

Assumption 2.5.1. *(i)*$\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(Q_T(s;u)-Q(s;u)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\xrightarrow{p}0$*;*
(ii) $\sup_{\psi\in\Psi}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{Q}_T(s;u,\psi)-Q(s;u,\psi)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\xrightarrow{p}0.$

Assumption 2.5.1 imposes that $Q_T(\cdot,\dot{\cdot})$ and $\widehat{Q}_T(\cdot,\cdot,\psi)$ converge to $Q(\cdot,\cdot)$ and $Q(\cdot,\cdot,\psi)$ in weighted L_2-norm, respectively, and the latter is uniform in $\psi\in\Psi$. Assumption 2.5.2 below states that ψ_0 is well-separated.

Assumption 2.5.2. *ψ_0 is in the interior of Ψ such that for all $\epsilon>0$,*

$$\inf_{\psi\notin B(\psi_0,\epsilon)}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q(s;u)-Q(s;u,\psi))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)>0,$$

where $B(\psi_0,\epsilon):=\{\psi\in\Psi:\|\psi-\psi_0\|\leq\epsilon\}$.

Theorem 2.5.1 (Consistency of $\hat{\psi}_T$)**.** *Suppose Assumptions 2.5.1 and 2.5.2 hold. Then $\hat{\psi}_T\xrightarrow{p}\psi_0$ as $T\to\infty$.*

To establish the asymptotic normality, additional assumptions are needed. Similar to Andrews (1999) and Pollard (1980), we impose first-order norm-differentiability on $\widehat{Q}_T$ with respect to ψ. Let

$$\widehat{R}_T(s;u,\psi,\psi_0):=\widehat{Q}_T(s;u,\psi)-\widehat{Q}_T(s;u,\psi_0)-(\psi-\psi_0)^\top\widehat{D}_T(s;u,\psi_0),$$

where $\widehat{D}_T(\cdot;\dot{;}\psi_0)$ is an $L_2(\mathcal{S}\times\mathbb{S}^{d-1},w(s)\mathrm{d}s\mathrm{d}\varsigma(u))$-measurable function.

Assumption 2.5.3. *$\widehat{Q}_T(\cdot; u, \psi)$ is first-order norm-differentiable at $\psi = \psi_0$. That is,*

$$\sup_{\psi \in \Psi;\, \|\psi - \psi_0\| \le \tau_T} \left| \frac{T \int_{\mathbb{S}^{d-1}} \int_S \left(\widehat{R}_T(s; u, \psi, \psi_0) \right)^2 w(s) ds d\varsigma(u)}{(1 + \|\sqrt{T}(\psi - \psi_0)\|)^2} \right| = o_p(1)$$

for any $\tau_T \to 0$.

Assumption 2.5.4 (i) below strengthens Assumption 2.5.1 (i).

Assumption 2.5.4. *The following conditions hold:*

(i) $T \int_{\mathbb{S}^{d-1}} \int_{\mathcal{S}} (Q_T(s; u) - Q(s; u))^2 w(s)\ \mathrm{d}s d\varsigma(u) = O_p(1)$*;*

(ii) $T \int_{\mathbb{S}^{d-1}} \int_{\mathcal{S}} (\widehat{Q}_T(s; u, \psi_0) - Q(s; u, \psi_0))^2 w(s) \mathrm{d}s d\varsigma(u) = O_p(1)$*;*

(iii) *There exists an $L_2(\mathbb{R} \times \mathbb{S}^{d-1}, w(s)\mathrm{d}s\mathrm{d}\varsigma)$-measurable function $D(\cdot; \cdot, \psi_0)$ such that*

$$\int_{\mathbb{S}^{d-1}} \int_{\mathcal{S}} \left\| \widehat{D}_T(s; u, \psi_0) - D(s; u, \psi_0) \right\|^2 w(s) \mathrm{d}s d\varsigma(u) = o_p(1).$$

Assumption 2.5.5.

$$\sqrt{T} \begin{pmatrix} \int_{\mathbb{S}^{d-1}} \int_{\mathcal{S}} (Q_T(s; u) - Q(s; u)) D(s; u.\psi_0) w(s) ds d\varsigma(u) \\ \int_{\mathbb{S}^{d-1}} \int_{\mathcal{S}} (\widehat{Q}_T(s; u, \psi_0) - Q(s; u, \psi_0)) D(s; u, \psi_0) w(s) ds d\varsigma(u) \end{pmatrix} \xrightarrow{d} N(0, V_0)$$

for some positive semidefinite matrix V_0.

Assumption 2.5.6.

$$B_0 := \int_{\mathbb{S}^{d-1}} \int_{\mathcal{S}} D(s; u, \psi_0) D^\top(s; u, \psi_0) w(s) \mathrm{d}s d\varsigma(u) \tag{2.14}$$

is positive definite.

Theorem 2.5.2 (Asymptotic normality of $\hat{\psi}_T$)**.** *Suppose Assumptions 2.5.1 to 2.5.6 hold. Then,*

$$\sqrt{n}(\hat{\psi}_T - \psi_0) \xrightarrow{d} N(0, B_0^{-1} \Omega_0 B_0^{-1}),$$

where $\Omega_0 = (e_1^\top, -e_1^\top) V_0 \begin{pmatrix} e_1 \\ -e_1 \end{pmatrix}$ in which $e_1 = (1, \ldots, 1)^\top$ is a d_ψ by 1 vector of ones.

Remark 2.5.1. Similar to Bernton et al. (2019), we can use a subsampling approach (e.g, Theorems 2.2.1. and 3.3.1 in Politis et al. (1999)) or bootstrapping for inference.

2.5.2 Norm-differentiability Assumption

We note here that when $\widehat{Q}_T(\cdot;u,\psi) = Q(\cdot;u,\psi)$ such as in unconditional models, we have $\widehat{D}_T(s;u,\psi_0) = D(s;u,\psi_0)$ so Assumptions 2.5.4 (ii) and (iii) automatically hold. The norm-differentiability assumption in Assumption 2.5.3 allows for the function $Q(\cdot;u,\psi)$ to have a finite number of kink points such as in one-sided and two-sided models, in which case one can take

$$D_T(s;u,\psi_0) = \begin{cases} \left.\frac{\partial Q(s;u,\psi)}{\partial \psi}\right|_{\psi=\psi_0} & \text{when } s \text{ is not a kink point;} \\ 0 & \text{when } s \text{ is a kink point.} \end{cases}$$

To illustrate, we show below that Assumption 2.5.3 is satisfied in the one-sided and two-sided uniform models for the unweighted MSCD estimator, see Section B.5 in the appendix for the MSWD estimator.

Example 2.5.1. *Consider the one-sided uniform model. Since*

$$F(y;\psi) = \begin{cases} 0 & \text{if } y \le 0, \\ \frac{y}{\psi} & \text{if } 0 < y < \psi, \\ 1 & \text{if } y \ge \psi, \end{cases}$$

one can take

$$\widehat{D}_T(s;\psi_0) = \begin{cases} 0 & \text{if } y \le 0, \\ -\frac{y}{\psi_0^2} & \text{if } 0 < y < \psi_0, \\ 0 & \text{if } y \ge \psi_0. \end{cases}$$

Then

$$\int_{-\infty}^{\infty} \Big(F(y;\psi) - F(y;\psi_0) - (\psi-\psi_0)\widehat{D}_T(s;\psi_0)\Big)^2 ds = \begin{cases} -\frac{(\psi-\psi_0)^3}{3\psi_0^2} & \text{if } \psi < \psi_0, \\ \frac{(\psi-\psi_0)^3}{3\psi\psi_0} & \text{if } \psi \ge \psi_0, \end{cases}$$

and

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T}\left|\frac{T\int_{\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 dsd\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\right|$$

$$\le\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T}\left|\frac{\int_{\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 dsd\varsigma(u)}{\|\psi-\psi_0\|^2}\right|=o(1).$$

Example 2.5.2. *Similarly, for the two-sided uniform model,*

$$F(y;\psi)=\begin{cases}\frac{y}{4\psi} & \text{if } 0\le y<\psi,\\ \frac{1}{4}+\frac{3(y-\psi)}{4(1-\psi)}=1-\frac{3(1-y)}{4(1-\psi)} & \text{if } \psi\le y\le 1.\end{cases}$$

Let

$$\widehat{D}_T(s;\psi_0)=\begin{cases}0 & \text{if } s\le 0,\\ -\frac{y}{4\psi_0^2} & \text{if } 0<s<\psi_0,\\ 0 & \text{if } y=\psi_0,\\ -\frac{3(1-y)}{4(1-\psi_0)^2} & \text{if } \psi_0<s<1,\\ 0 & \text{if } s\ge 1.\end{cases}$$

Then,

$$\int_{-\infty}^{\infty}\left(F(y;\psi)-F(y;\psi_0)-(\psi-\psi_0)\widehat{D}_T(s;\psi_0)\right)^2 ds$$

$$=\begin{cases}-\frac{(\psi-\psi_0)^3(4\psi(\psi_0-1)+12\psi_0^2-4\psi_0+1)}{48(\psi-1)(\psi_0-1)\psi_0^2} & \text{if } \psi<\psi_0,\\ \frac{(\psi-\psi_0)^3(12\psi\psi_0+4\psi_0^2-8\psi_0+1)}{48\psi\psi_0(1-\psi_0)^2} & \text{if } \psi\ge\psi_0.\end{cases}$$

and

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T}\left|\frac{T\int_{\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 dsd\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\right|$$

$$\le\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T}\left|\frac{\int_{\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 dsd\varsigma(u)}{\|\psi-\psi_0\|^2}\right|=o(1).$$

In contrast, the norm differentiability assumption for the asymptotic normality of MLE is not satisfied in the one-sided uniform model, see Remark 2.5.2 below.

Remark 2.5.2. (van de Vaart, 1998, Counterexample 7.9)

Suppose we have an i.i.d. real valued sample $\{Z_t\}_{t=1}^T$ with density $f(\cdot;\psi)$. According to Theorem 5.39 of van de Vaart (1998), the MLE estimator of ψ_0 follows asymptotically normal distribution when 1) there exists a measurable function $d(x;\psi_0)$ such that as $\psi \to \psi_0$,

$$\int \left[\sqrt{f(x,\psi)} - \sqrt{f(x,\psi_0)} - \frac{1}{2}(\psi-\psi_0)^\top d(x;\psi_0)\sqrt{f(x,\psi_0)}\right]^2 \mathrm{d}x = o(\|\psi-\psi_0\|^2) \quad (2.15)$$

and 2) there exists a measurable function $\dot{\ell}(\cdot)$ with $\int \dot{\ell}(x)^2 f(x;\psi_0)\mathrm{d}x < \infty$ such that for every ψ_1 and ψ_2 in a neighborhood of ψ_0,

$$|\log f(x;\psi_1) - \log f(x;\psi_2)| \leq \dot{\ell}(x)\|\psi_1-\psi_2\|.$$

Counterexample 7.9 in van de Vaart (1998) shows that the family of one-sided uniform distributions $U[0,\psi]$, $\psi > 0$, does not satisfy equation (2.15). We present this discussion in van de Vaart (1998) for completeness.

Let $f(x;\psi) = \frac{1}{\psi} I\{x \in [0,\psi]\}$ be the pdf of one-sided uniform model, and suppose there exists a measurable $d(x;\psi_0)$ such that equation (2.15) holds. When $\psi > \psi_0$,

$$\begin{aligned}
&\int \left[\sqrt{f(x,\psi)} - \sqrt{f(x,\psi_0)} - \frac{1}{2}(\psi-\psi_0)^\top d(x;\psi_0)\sqrt{f(x,\psi_0)}\right]^2 \mathrm{d}x \\
&\quad \geq \int_{f(x,\psi_0)=0} \left[\sqrt{f(x,\psi)} - \sqrt{f(x,\psi_0)} - \frac{1}{2}(\psi-\psi_0)^\top d(x;\psi_0)\sqrt{f(x,\psi_0)}\right]^2 \mathrm{d}x \\
&\quad = \int_{f(x,\psi_0)=0} f(x,\psi)\mathrm{d}x = \frac{\psi-\psi_0}{\psi} = O(\psi-\psi_0).
\end{aligned}$$

However, it is a contradiction to the assumption that equation (2.15) holds.

Remark 2.5.3. In this paper, we will not focus on the efficiency of our estimator. LaRiccia (1982, 1984) discuss the optimal weight for minimum L_2-distance estimator with $d = 1$. In detail, when $d = 1$ and $\psi \in \mathbb{R}$, LaRiccia (1982) provides an optimal weight to achieve the Cramer-Rao bound. (Refer to Remark 2 of LaRiccia (1982)). In practice, it is hard to find the optimal weight especially for the high-dimensional case, which will be future research topic.

2.6 Unconditional Models

For unconditional models, we can take $\widehat{Q}_T(\cdot; u, \psi) := Q(\cdot; u, \psi)$ in (2.9), and $\widehat{D}_T(:, u, \psi_0) := D(\cdot, u, \psi_0)$, a deterministic $L_2([0,1] \times \mathbb{S}^{d-1}, w(s)\mathrm{d}s\mathrm{d}\varsigma(u))$-measurable function. Let

$$R(s; u, \psi, \psi_0) := Q(s; u, \psi) - Q(s; u, \psi_0) - (\psi - \psi_0)^\top D(s; u, \psi_0).$$

Theorems 5.1 and 5.2 imply the following Corollary.

Corollary 2.6.1. *(i)* $\hat{\psi}_T \xrightarrow{p} \psi_0$ *as* $n \to \infty$ *under Assumptions 2.5.1 (i) and 2.5.2; (ii)* $\sqrt{T}(\hat{\psi}_T - \psi_0) \xrightarrow{d} N(0, B_0^{-1}\Omega_0 B_0^{-1})$ *under Assumptions 2.5.1 (i), 2.5.2, 2.5.3, 2.5.4 (i), 2.5.6, and 2.5.5 with* $V_0 = \begin{pmatrix} \Omega_0 & 0 \\ 0 & 0 \end{pmatrix}$. *Here,* Ω_0 *is a* $d_\psi \times d_\psi$ *positive semi-definite matrix.*

In the rest of this section, we will verify the high-level Assumptions 2.5.4 (i) and 2.5.5 for β-mixing processes since Assumption 2.5.1 (i) is implied by Assumption 2.5.4 (i).

Definition 2.6.1 (c.f., Bradley (2005))**.** *Consider the probability space* $(\Omega, \mathcal{F}, P)$. *Let* $\mathcal{A} \subset \mathcal{F}$ *and* $\mathcal{B} \subset \mathcal{F}$ *be two sigma fields. We define*

$$\beta(\mathcal{A}, \mathcal{B}) = \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|$$

where supremum is taken over all pairs of (finite) partitions $\{A_1, \ldots, A_I\}$ *and* $\{B_1, \ldots, B_J\}$ *of* Ω *such that* $A_i \in \mathcal{A}$ *for each* i *and* $B_j \in \mathcal{B}$ *for each* j.

Suppose we have a (not necessary stationary) random process $\{Z_t\}$ *for* $t \in \mathbb{Z}$, *where* $\mathbb{Z}$ *is a set of integers. Let's denote* $\mathcal{F}_J^L = \sigma(Z_t, J \le t \le L, t \in \mathbb{Z})$ *to be sigma-field generated by* $\{Z_t\}$ *where* $J \le t \le L$. *Then,* $\{Z_t\}$ *is* β*-mixing if* $\beta_k := \sup_{j \in \mathbb{Z}} \beta(\mathcal{F}_{-\infty}^j, \mathcal{F}_{j+k}^\infty) \to 0$ *as* $k \to \infty$.

Condition 2.6.1. $\{Z_t\}_{t=1}^\infty$ *is a strictly stationary* β*-mixing process with* $k^r \beta_k \to 0$ *as* $k \to \infty$ *for some* $r > 1$.

Let $F(z;\psi_0)$ denote the cdf of Z, $G(s;u,\psi_0)$ denote the cdf of $u^\top Z$, and $G_T(s;u)$ be the empirical CDF of $\{u^\top Z_t\}_{t=1}^T$, where $Z_t \sim F(z;\psi_0)$. That is,

$$G_T(s;u) = \frac{1}{T}\sum_{t=1}^{T} I(u^\top Z_t \leq s).$$

We focus on $d > 1$ in the rest of this section. Appendix B.6.1 collects results for $d = 1$.

2.6.1 MSWD Estimator

We denote $Q(\cdot;u,\psi_0)$ to be the quantile function of $u^\top Z$.

Verification of Assumption 2.5.4 (i)

For a given $u \in \mathbb{S}^{d-1}$, let

$$G(s;u,\psi) = \int 1(u^\top z \leq s)\mathrm{d}F(z;\psi), \text{ for } s \in \mathbb{R}.$$

It is known that $\{I(u^\top Z_t \leq s) : u \in \mathbb{S}^{d-1}, s \in \mathbb{R}\}$ is VC subgraph class (See Example 3.7.4c in van de Geer (2009) and Example 7.21 in Sen (2021)). We apply the weak convergence result for empirical processes indexed by VC subgraph class in Corollary 2.1 of Arcones and Yu (1994) (See Lemma B.3.1) to establish our first lemma below.

Lemma 2.6.1. *Under Condition 2.6.1, we have*

$$\sqrt{T}(G_T(\cdot;\cdot,\psi_0) - G(\cdot;\cdot,\psi_0)) \Rightarrow B(\cdot,\cdot)$$

in $\mathbb{D}_\infty$. *Here* $B(\cdot,\cdot)$ *is a Gaussian process with mean* 0 *and the covariance given by*

$$\sum_{t,j} Cov(I\{u^\top Z_t \leq s\}, I\{I\{v^\top Z_j \leq l\})$$

for $u \in \mathbb{S}^{d-1}$, $v \in \mathbb{S}^{d-1}$, $s \in \mathbb{R}$, *and* $l \in \mathbb{R}$.

For stochastic singular unconditional models, the distribution of Z_t is singular and that of $u^\top Z_t$ may be degenerate. Let

$$\mathcal{N}_0 = \{u \in \mathbb{S}^{d-1} : G(\cdot;u,\psi_0) \text{ is degenerate}\}.$$

It is nonempty when the distribution of Z_t is singular. In Example 1 in Arjovsky et al. (2017), $Y = (\theta, Z)$, where θ is a deterministic constant and $Z \sim U[0,1]$. Then $u^\top Y = u_1\theta + u_2 Z$ is degenerate when $u = (1,0)$ or $u = (-1,0)$. Thus, $\mathcal{N}_0 = \{(1,0), (-1,0)\}$.

Example 2.6.1. *Let us consider a simple asset-pricing/state space model.*

$$y_t = a + by_t^*,$$
$$y_t^* = \rho y_{t-1}^* + e_t,\ e_t \sim N(0,1).$$

where $|\rho| < 1$ *and* $d_y > d_y^* = 1$.

The unconditional distribution of y_t *is* $N(a, bb^\top/(1-\rho^2))$, *which is singular. The distribution of* $u^\top y_t$ *is degenerate when* $u^\top b = 0$. *So* $\mathcal{N}_0 = \{u \in \mathbb{S}^{d-1}, u^\top b = 0\}$.

We are now ready to verify Assumption 2.5.4 (i).

Lemma 2.6.2. *Suppose that* $G(s; u, \psi_0)$ *is absolutely continuous and differentiable with respect to* s *for all* $u \notin \mathcal{N}_0$ *and the density function* $g(s; u, \psi_0)$ *of* $G(s; u, \psi_0)$ *is bounded above with respect to* s *for each* $u \notin \mathcal{N}_0$. *Let us assume that there exists a positive constant* C *such that*

$$\int_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \int_0^1 \frac{w(s)}{g^2(G^{-1}(s; u, \psi_0); u, \psi_0)} \mathrm{d}s\mathrm{d}\varsigma(u) < C < \infty, \tag{2.16}$$

and

$$\sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0,}\ \sup_{s \in supp\{w\}} \left| \frac{g(G^{-1}(s; u, \psi_0); u, \psi_0)}{g(G^{-1}(s(s; u); u, \psi_0); u, \psi_0)} \right| = O_p(1) \tag{2.17}$$

for $s(s, u, \psi_0)$ *such that* $\sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s \in supp\{w\}} |s(s; u, \psi_0) - s| = o_p(1)$. *Then, Assumption 2.5.4 (i) holds under Condition 2.6.1.*

When the joint density $f(z; \psi)$ of Z is uniformly bounded from below by a positive absolute constant, the density $g(t; s, \psi)$ of the projected variable $u^\top Z$ is also uniformly bounded from below by a positive absolute constant. Then, equations (2.16) and (2.17) in Lemma 2.6.2 are satisfied. When $g(t; s, \psi)$ is uniformly bounded from below and above by absolute constants, Ahidar-Coutrix and Berthet (2021) provides Bahadur-Kiefer representation and uniform CLT theorem of empirical quantile process of projected variables $\{u^\top Z_t\}$ for random sample $\{Z_t\}$.

Remark 2.6.1. When $w(s) = 1$ for $s \in [\delta, 1-\delta]$ for some $\delta > 0$, we can show equation (2.17) under the following assumption:

$$\begin{aligned} &\sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s \in \mathbb{R}} (G(s; u, \psi_0)(1 - G(s; u, \psi_0))) \left| \frac{g'(s; u, \psi_0)}{g^2(s; u, \psi_0)} \right| \\ &= \sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s \in (0,1)} (s(1-s)) \left| \frac{g'(G^{-1}(s; u, \psi_0); u, \psi_0)}{g^2(G^{-1}(s; u, \psi_0); u, \psi_0)} \right| < \infty, \end{aligned} \tag{2.18}$$

where $g\prime(s; u, \psi_0) = \partial g(s; u, \psi_0)/\partial s$. When $d = 1$, this condition is the same as that in Theorem 3 of Csorgo and Revesz (1978).

To show (2.18), we can use arguments in the proof of Lemma 1 in Csorgo and Revesz (1978). For $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$,

$$\left| \frac{\partial}{\partial s} \log(g(G^{-1}(s; u, \psi_0); u, \psi_0)) \right| = \left| \frac{g'(G^{-1}(s; u, \psi_0); u, \psi_0)}{g(G^{-1}(s; u, \psi_0); u, \psi_0)} \right| \le \gamma(s(1-s))^{-1} = \gamma \frac{\partial}{\partial s} \log\left(\frac{s}{1-s} \right).$$

Then, when $s_1 > s_2$, we have for $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$,

$$\log\left(\frac{g(G^{-1}(s_1; u, \psi_0); u, \psi_0)}{g(G^{-1}(s_2; u, \psi_0); u, \psi_0)} \right) \le \gamma \log\left(\frac{s_1}{1-s_1} \right) - \gamma \log\left(\frac{s_2}{1-s_2} \right) = \gamma \log\left(\frac{s_1(1-s_2)}{s_2(1-s_1)} \right);$$

when $s_1 < s_2$ we have for $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$,

$$\log\left(\frac{g(G^{-1}(s_1; u, \psi_0); u, \psi_0)}{g(G^{-1}(s_2; u, \psi_0); u, \psi_0)} \right) \le \gamma \log\left(\frac{s_2}{1-s_2} \right) - \gamma \log\left(\frac{s_1}{1-s_1} \right) = \gamma \log\left(\frac{s_2(1-s_1)}{s_1(1-s_2)} \right).$$

Therefore, we have for $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$,

$$\left(\frac{g(G^{-1}(s_1; u, \psi_0); u, \psi_0)}{g(G^{-1}(s_2; u, \psi_0); u, \psi_0)} \right) \le \left\{ \left(\frac{s_1 \vee s_2}{s_1 \wedge s_2} \right) \left(\frac{1 - (s_1 \wedge s_2)}{1 - (s_1 \vee s_2)} \right) \right\}^\gamma.$$

This implies for $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$,

$$\left(\frac{g(G^{-1}(s; u, \psi_0); u, \psi_0)}{g(G^{-1}(\widetilde{s}(s; u); u, \psi_0); u, \psi_0)} \right) \le \left\{ \left(\frac{s \vee \widetilde{s}(s; u)}{s \wedge \widetilde{s}(s; u)} \right) \left(\frac{1 - (s \wedge \widetilde{s}(s; u))}{1 - (s \vee \widetilde{s}(s; u))} \right) \right\}^\gamma.$$

When $w(s) = 1$ for $s \in [\delta, 1-\delta]$ for some $\delta > 0$,

$$\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s\in[\delta,1-\delta]} \left(\frac{g(G^{-1}(s;u,\psi_0);u,\psi_0)}{g(G^{-1}(\widetilde{s}(s;u);u,\psi_0);u,\psi_0)} \right)$$
$$\leq \sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s\in[\delta,1-\delta]} \left\{ \left(\frac{s \vee \widetilde{s}(s;u)}{s \wedge \widetilde{s}(s;u)} \right) \left(\frac{1-(s \wedge \widetilde{s}(s;u))}{1-(s \vee \widetilde{s}(s;u))} \right) \right\}^{\gamma} = O_p(1)$$

since $\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s\in supp\{w\}} |s(s;u) - s| = o_p(1)$.

Verification of Assumption 2.5.5

Let $\mathbb{D}$ be the space of univariate distribution functions, and $\mathbb{D}_1$ be the restriction on $\mathbb{D}$ such that the domain of the distribution functions in $\mathbb{D}_1$ is the same as the support of $G(s;u)$ for each u. Let $\mathbb{D}_2 = L_2([0,1] \times \mathbb{S}^{d-1}, w(s)\mathrm{d}s\mathrm{d}\varsigma(u))$.

Lemma 2.6.3. *Assume that $G(s;u)$ is absolutely continuous and differentiable with respect to s for each $u \notin \mathcal{N}_0$, and that $g(s;u)$ is bounded above with respect to s for each $u \notin \mathcal{N}_0$. Further, suppose that there exists a positive constant C such that*

$$\int_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \int_0^1 \frac{w(s)}{g^2(G^{-1}(s;u,\psi_0);u,\psi_0)} \mathrm{d}u\mathrm{d}\varsigma(u) < C < \infty, \tag{2.19}$$

and

$$\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s\in supp\{w\}} \left| \frac{g(G^{-1}(s;u,\psi_0);u,\psi_0)}{g(G^{-1}(s(s;u;\psi_0);u,\psi_0);u,\psi_0)} \right| = O(1) \tag{2.20}$$

for $s(s,u,\psi_0)$ such that $\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s\in(0,1)} |s(s;u,\psi_0) - s| = o(1)$. Then, $\phi(G) : \mathbb{D}_1 \subset \mathbb{D} \mapsto \mathbb{D}_2$, where $\phi(G) = G^{-1}(s;u)$, is Hadamard differentiable at $G \in \mathbb{D}_1$ tangentially to $\mathbb{D}_1$, and its derivative is

$$\phi'_G(h) = \begin{cases} \frac{h(G^{-1}(s;u,\psi_0);u)}{g(G^{-1}(s;u,\psi_0);u,\psi_0)} & \text{if } u \notin \mathcal{N}_0; \\ 0 & \text{if } u \in \mathcal{N}_0. \end{cases}$$

When the weight function is $w(s) = 1$ for $s \in [\delta, 1-\delta]$ for some $\delta > 0$, Equation (2.18) implies Equation (2.20).

Lemma 2.6.4. *Let's denote $\phi(Q) : \mathbb{D}_2 \to \mathbb{R}$, where*

$$\phi(Q) = \int_{u\in\mathbb{S}^{d-1}} \int_0^1 \widetilde{Q}(s;u) D(s;u,\psi_0) w(s) \mathrm{d}s \mathrm{d}\varsigma(u)$$

Then, $\phi(\cdot)$ is Hadamard differentiable at $Q \in \mathbb{D}_2$ tangentially to $\mathbb{D}_2$ under Assumption 2.5.6, and the derivative is

$$\phi'_Q(h) = \int_{u\in\mathbb{S}^{d-1}} \int_0^1 h(s;u) D(s;u,\psi_0) w(s) \mathrm{d}s \mathrm{d}\varsigma(u)$$

Proof. Let $Q_n(s;u) = Q(s,u) + t_n h_n(s,u)$ where $t_n \downarrow 0$ as $n \to \infty$, $h_n(s,u) \in \mathcal{D}_2$, $\lim_{n\to\infty} \int_{u\in\mathbb{S}^{d-1}} \int_0^1 |h_n(s,u) - h(s,u)|^2 w(s) ds d\varsigma(u) = 0$, and $Q, Q_n \in \mathbb{D}_2$.

Then, we have

$$\begin{aligned}
&\left\| \frac{\phi(Q_n) - \phi(Q)}{t_n} - \phi'_Q(h) \right\| \\
&= \left\| \int_{u\in\mathbb{S}^{d-1}} \int_0^1 h_n(s;u) - h(s,u) D(s;u,\psi_0) w(s) \mathrm{d}s \mathrm{d}\varsigma(u) \right\| \\
&\le \left(\int_{u\in\mathbb{S}^{d-1}} \int_0^1 |h_n(s;u) - h(s,u)|^2 D(s;u,\psi_0) w(s) \mathrm{d}s \right)^{1/2} \left(\int_{u\in\mathbb{S}^{d-1}} \int_0^1 \|D(s;u,\psi_0)\|^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u) \right)^{1/2} \\
&= o(1).
\end{aligned}$$

□

Combining Lemmas 2.6.3 and 2.6.4, we can show that Assumption 2.5.5 is satisfied.

Lemma 2.6.5. *Suppose Assumption 2.5.3 and all conditions in Lemma 2.6.3 hold. Then, Assumption 2.5.5 is satisfied with $V = \begin{pmatrix} \Omega_0 & 0 \\ 0 & 0 \end{pmatrix}$ in which Ω_0 is the long-run variance of $\{K_t\}$, where*

$$K_t = \int_{u\in\mathbb{S}^{d-1}} \int_0^1 \frac{1}{g(G^{-1}(s;u,\psi_0);u,\psi_0)} \Big[I(G(u^\top Z_t;u,\psi_0) \le s) - s \Big] D(s;u,\psi_0) w(s) \mathrm{d}s \varsigma(u).$$

2.6.2 MSCD Estimator

Let $Q(\cdot;u,\psi) = G(\cdot;u,\psi)$. In this subsection, we will verify Assumptions 2.5.4 (i) and 2.5.5 for β-mixing process under the following condition.

Condition 2.6.2. *The weight function $w(s)$ is integrable. That is, $\int_{\mathbb{R}} w(s)\mathrm{d}s < \infty$.*

Lemma 2.6.6. *Under Conditions 2.6.1 and 2.6.2, Assumption 2.5.4 (i) holds.*

Proof. Note that

$$\begin{aligned}
&T\,\mathbb{E}\left[\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}(G_T(s;u,\psi_0)-G(s;u,\psi_0))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right]\\
&\quad=\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\mathbb{E}\left[\left\{\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\big(I(u^\top Z_t\le s)-G(s;u,\psi_0)\big)\right\}^2\right]w(s)\mathrm{d}s\mathrm{d}\varsigma(u),
\end{aligned}$$

and $\lim_{k\to\infty}k^r\beta_k=0$ implies $\sum_{k\ge 0}\beta_k<\infty$. When $\sum_{k\ge 0}\beta_k<\infty$, by following pages 11-12 of Rio (2017), we can show

$$\mathbb{E}\left[\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\big(I(u^\top Z_t\le s)-G(s;u,\psi_0)\big)\right\}^2\right]<C$$

for some C because $I(u^\top Z_t\le s)$ is bounded by 1. Therefore, we have the desired result. □

Note that

$$\begin{aligned}
&\sqrt{T}\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}(G_T(s;u,\psi_0)-G(s;u,\psi_0))D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\Big[I(u^\top Z_t\le s)-G(s;u,\psi_0)\Big]D(s;u,\psi_0)w(s)\mathrm{d}s\varsigma(u)
\end{aligned}$$

Then, by applying CLT, we have the desired result.

Lemma 2.6.7. *In addition to Conditions 2.6.1 and 2.6.2, let's assume $\int\|D(s,u,\psi_0)\|w(s)ds d\varsigma(u)<\infty$. Then,*

$$\begin{aligned}
&\sqrt{T}\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}(G_T(s;u,\psi_0)-G(s;u,\psi_0))D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\Big[I(u^\top Z_t\le s)-G(s;u,\psi_0)\Big]D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad\xrightarrow{d}N(0,\Omega_0),
\end{aligned}$$

where Ω_0 is the long-run variance of $\{K_t\}$:

$$K_t=\int_{u\in\mathbb{S}^{d-1}}\int\Big[I(u^\top Z_t\le s)-G(s;u,\psi_0)\Big]D(s;u,\psi_0)w(s)\mathrm{d}s\varsigma(u).$$

Proof. When $\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\|D(s,u,\psi_0)\|w(s)dsd\varsigma(u)<\infty$ and $\sup_{k\geq 0}\beta(k)<\infty$, we can apply Corollary 4.1 of Rio (2017) because $\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\Big[I(u^\top Z_i\leq s)-G(s;u,\psi_0)\Big]D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)$ is bounded above by a constant. □

2.7 MSCD Estimators for Conditional Models

For conditional models, $\widehat{Q}_T$ is random. In this section, we verify the high-level assumptions in Section 5 for MSCD estimator for random samples $\{Z_t\}_{t=1}^T$. Extensions to β-mixing processes can be done with more tedious technical derivations.

Assume that Ψ is compact. Let $Z=\left(Y,X^\top\right)^\top$, where $Y\in\mathbb{R}$ and $X\in\mathbb{R}^{d_x}$. The cdfs of Z and $u^\top Z$ are given by

$$F\left(z;\psi_0\right)=\mathbb{E}\left[F\left(y|X,\psi_0\right)I\left(X\leq x\right)\right]\ \text{ and}$$

$$\begin{aligned}G(s;u,\psi_0)&=\Pr(u_1Y+u_2^\top X\leq s)\\&=\mathbb{E}\left[\int_{-\infty}^{\infty}I(u_1y+u_2^\top X\leq s)f(y|X,\psi_0)\mathrm{d}y\right]\\&=\begin{cases}\mathbb{E}F(u_1^{-1}(s-u_2^\top X)|X,\psi_0) & \text{if } u_1>0\\ \mathbb{E}I(u_2^\top X\leq t) & \text{if } u_1=0\\ 1-\mathbb{E}F(u_1^{-1}(s-u_2^\top X)|X,\psi_0) & \text{if } u_1<0.\end{cases}\end{aligned}$$

Let $\widehat{Q}_T(\cdot;u,\psi)=\widehat{G}_T(\cdot;u,\psi)$, where

$$\begin{aligned}\widehat{G}_T(s;u,\psi)&=\frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty}I(u_1y+u_2^\top X_i\leq s)f(y|X_t,\psi)\mathrm{d}y\\&=\begin{cases}\frac{1}{T}\sum_{t=1}^{T}F(u_1^{-1}(s-u_2^\top X_t)|X_t,\psi) & \text{if } u_1>0\\ \frac{1}{T}\sum_{t=1}^{T}I(u_2^\top X_t\leq s) & \text{if } u_1=0\\ 1-\frac{1}{T}\sum_{t=1}^{T}F(u_1^{-1}(s-u_2^\top X_t)|X_t,\psi) & \text{if } u_1<0\end{cases}\end{aligned}$$

We will verify Assumption 2.5.1 (ii), Assumption 2.5.3, Assumption 2.5.4 (ii) and (iii), and Assumption 2.5.5.

2.7.1 Verification of Assumptions 2.5.1 (ii), 2.5.3, 2.5.4 (ii) and (iii), and 2.5.5.

Note that $\widehat{G}_T(s; u, \psi) = \int_{-\infty}^{\infty} I(u^\top z \leq s) dF_T(z; \psi) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[u^\top Z_t \leq s | X_t, \psi]$. We have

$$\widehat{G}_T(s; u.\psi) - G(s; u, \psi) = \frac{1}{T}\sum_{t=1}^{T}\Big(\mathbb{E}[I(u^\top Z_t \leq s)|X_t, \psi] - \mathbb{E}[I(u^\top Z_t \leq s)|\psi]\Big).$$

Hence, $\int_{u\in\mathbb{S}^{d-1}} \int_{-\infty}^{\infty} (\widehat{G}_n(t; u.\psi) - G(t; u, \psi))^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u)$ can be represented as a degenerate V-statistic of order 2. That is,

$$\int_{u\in\mathbb{S}^{d-1}} \int_{-\infty}^{\infty} (\widehat{G}_T(s; u.\psi) - G(s; u, \psi))^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u) = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{j=1}^{T} k(X_t, X_j; \psi),$$

where

$$\begin{aligned} k(X_t, X_j; \psi) = \int_{u\in\mathbb{S}^{d-1}} \int_{-\infty}^{\infty} &\big(\mathbb{E}[I(u^\top Z_t \leq s)|X_t, \psi] - \mathbb{E}[I(u^\top Z_t \leq s)|\psi]\big) \\ &\times \big(\mathbb{E}[I(u^\top Z_j \leq s)|X_j, \psi] - \mathbb{E}[I(u^\top Z_j \leq s)|\psi]\big)\, w(s) \mathrm{d}s \mathrm{d}\varsigma(u) \end{aligned}$$

is a degenerate symmetric kernel function indexed by ψ.

Under the Lipchitz continuity of $k(x, x', \psi)$ with respect to ψ for every x and x', Corollary 4.1 of Newey (1991) or Lemma 4 in appendix of Briol et al. (2019) can be used to verify Assumption 2.5.1 (ii). Furthermore, Lipchitz continuity of k is implied by that of F in the expression (2.21).

Lemma 2.7.1. *Suppose Condition 2.6.2 holds. Moreover, let us assume that $F(y|X_t, \psi)$ is Lipschitz with respect to ψ in the sense that there exists a function $M(y, x)$ such that for any $\psi, \psi' \in \Psi$,*

$$|F(y|x, \psi) - F(y|x, \psi')| \leq M(y, x)\|\psi - \psi'\|, \tag{2.21}$$

and $\int_{u\in\mathbb{S}^{d-1}, u_1\neq 0} \int_{-\infty}^{\infty} \int M^2(u_1^{-1}(s - u_2^\top x); x) \mathrm{d}F_X(x) w(s) \mathrm{d}s \mathrm{d}\varsigma(u) < \infty$, where $F_X(\cdot)$ is the CDF of X_t. Then,

$$\sup_{\psi\in\Psi} \int_{u\in\mathbb{S}^{d-1}} \int_{-\infty}^{\infty} (\widehat{G}_T(s; u.\psi) - G(s; u, \psi))^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u) = o_p(1).$$

Note that under i.i.d assumption, we have

$$
\begin{aligned}
&T\mathbb{E}\left[\int_{-\infty}^{\infty}(\widehat{G}_T(s;u.\psi_0)-G(s;u,\psi_0))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right]\\
&\quad=\int_{-\infty}^{\infty}\mathbb{E}\left[T(\widehat{G}_T(s;u,\psi_0)-G(s;u,\psi_0))^2\right]w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad=\int_{-\infty}^{\infty}\mathbb{E}\left[\Big(\mathbb{E}[I(u^\top Z_t\le s)|X_t,\psi_0]-G(s;u,\psi_0)\Big)^2\right]w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad=\int_{-\infty}^{\infty}\left\{\mathbb{E}\left[\Big(\mathbb{E}[I(u^\top Z_t\le s)|X_t,\psi_0]\Big)^2\right]-G^2(s;u,\psi_0)\right\}w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad\le\int_{-\infty}^{\infty}\left\{\mathbb{E}\left[\Big(\mathbb{E}[I(u^\top Z_t\le s)^2|X_t,\psi_0]\Big)\right]-G^2(s;u,\psi_0)\right\}w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad=\int_{-\infty}^{\infty}G(s;u,\psi_0)\Big(1-G(t;u,\psi_0)\Big)w(s)\mathrm{d}s\mathrm{d}\varsigma(u).
\end{aligned}
$$

Then, we have the following result.

Lemma 2.7.2. *Under Condition 2.6.2, Assumption 2.5.4 (ii) holds.*

Let

$$R(s;u,\psi,\psi_0)=Q(s,u,\psi)-Q(s;u,\psi_0)-(\psi-\psi_0)^\top D(s;u,\psi_0),$$

where $D(\cdot;\cdot,\psi_0)$ is a $L_2(\mathbb{R}\times\mathbb{S}^{d-1},w(s)\mathrm{d}s\mathrm{d}\varsigma)$-measurable function.

Condition 2.7.1. *$Q(\cdot;u,\psi)$ is first-order norm-differentiable at $\psi=\psi_0$. That is,*

$$\sup_{\psi\in\Psi;\,\|\psi-\psi_0\|\le\tau_T}\left|\frac{T\int_{\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\left(R(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\right|=o(1)$$

for any $\tau_T\to 0$.

Under condition 2.7.1, it is enough to show that

$$\sup_{|\psi-\psi_0|\le\tau_T}\frac{T\left(\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\left(\left[\widehat{Q}_T(s;u,\psi)-Q(s;u,\psi)\right]-\left[\widehat{Q}_T(s;u,\psi_0)-Q(s;u,\psi_0)\right]\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)}{(1+\|\psi-\psi_0\|)^2}=o_p(1),$$

where

$$\widehat{Q}_T(s;u,\psi)-Q(s;u,\psi)=\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}[I(u^\top Z_t\le s)|X_t,\psi]-G(s;u,\psi)\right).$$

Note that the integral in the numerator can be represented as a degenerate V-statistic of order 2:

$$\begin{aligned}&\left(\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\Big(\big[\widehat{Q}_T(s;u,\psi)-Q(s;u,\psi)\big]-\big[\widehat{Q}_T(s;u,\psi_0)-Q(s;u,\psi_0)\big]\Big)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)\\&=\frac{1}{T^2}\sum_{t=1}^{T}\sum_{j=1}^{T}k_2(X_t,X_j,\psi,\psi_0),\end{aligned}$$

where

$$\begin{aligned}&k_2(X_t,X_j,\psi,\psi_0)\\&=\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\Bigg\{\Big[\Big(\mathbb{E}[I(u^\top Z_t\le s)|X_t,\psi]-G(s;u,\psi)\Big)-\Big(\mathbb{E}[I(u^\top Z_t\le s)|X_t,\psi_0]-G(s;u,\psi_0)\Big)\Big]\\&\quad\times\Big[\Big(\mathbb{E}[I(u^\top Z_j\le s)|X_j,\psi]-G(s;u,\psi)\Big)-\Big(\mathbb{E}[I(u^\top Z_j\le s)|X_j,\psi_0]-G(s;u,\psi_0)\Big)\Big]\Bigg\}w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\end{aligned}$$

is a degenerate symmetric kernel.

When $k_2(x,x',\psi,\psi_0)$ is Lipschitz continuous with respect to ψ for each x, x', we can use Corollary 8 in Sherman (1994) and the proof of Lemma 4 in Appendix of Briol et al. (2019) to verify Assumptions 2.5.3 and 2.5.4 (iii). Lipchitz continuity of k_2 is implied by that of $F(\cdot|x,\psi)$ in the expression (2.21).

Lemma 2.7.3. *In addition to all conditions in Lemma 2.7.1, suppose Condition 2.7.1 holds. Then, $\widehat{Q}_T(\cdot;\cdot,\psi)$ is norm-differentiable at $\psi=\psi_0$ with $\widehat{D}_n(s;u,\psi)=D(s;u,\psi)$. This implies that Assumptions 2.5.3 and 2.5.4 (iii) hold.*

Note that

$$\begin{aligned}\sqrt{T}(\widehat{Q}_T(s;u)-Q(s;u))&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(I(u^\top Z_t\le s)-G(s;u)\big)\,,\ \text{and}\\\sqrt{T}(\widehat{Q}_T(s;u,\psi_0)-Q(s;u,\psi_0))&=\frac{1}{\sqrt{T}}\sum_{i=1}^{n}\big(\mathbb{E}[I(u^\top Z_t\le s)|X_t,\psi_0]-G(s;u,\psi_0)\big)\,.\end{aligned}$$

By applying CLT, we obtain the following result.

Lemma 2.7.4. *Suppose* $\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\|D(s,u,\psi_0)w(s)\|\mathrm{d}s\mathrm{d}\varsigma(u) < \infty$. *Then, Assumption 2.5.5 holds. That is,*

$$\begin{aligned}
&\sqrt{T}\begin{pmatrix}\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}(\widehat{Q}_T(s;u)-Q(s;u))D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}u\\ \int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}(\widehat{Q}_T(s;u,\psi_0)-Q(s;u,\psi_0))D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}u\end{pmatrix}\\
&\quad=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{pmatrix}\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\left(I(u^\top Z_t\le s)-G(s;u)\right)D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}u\\ \int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}(\mathbb{E}[I(u^\top Z_t\le s)|X_t,\psi_0]-Q(s;u;\psi_0))D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}u\end{pmatrix}\\
&\quad\xrightarrow{d} N(0,V_0),
\end{aligned}$$

where

$$V_0=\int_{u\in\mathbb{S}^{d-1}}\int_{v\in\mathbb{S}^{d-1}}\int\int\begin{pmatrix}A_{11}(t,s;u,v) & A_{12}(t,s;u,v)\\ A_{21}(t,s;u,v) & A_{22}(t,s;u,v)\end{pmatrix}\otimes D(s;u,\psi_0)D(t;u,\psi_0)^\top w(s)w(t)\mathrm{d}s\ \mathrm{d}t\mathrm{d}\varsigma(u)\mathrm{d}\varsigma(v)$$

in which $\otimes$ *is the Kronecker product, and*

$$\begin{aligned}
A_{11}(t,s;u,v)&=\mathbb{E}[I(u^\top Z\le t)I(v^\top Z\le s)]-G(t;u)G(s,v),\\
A_{22}(t,s;u,v)&=\mathbb{E}\left\{\mathbb{E}[I(u^\top Z\le t|X)]\mathbb{E}[I(v^\top Z\le s)|X]\right\}-G(t;u)G(s,v),\\
A_{12}(t,s;u,v)&=\mathbb{E}\left\{I(u^\top Z\le t)\mathbb{E}[I(v^\top Z\le s)|X]\right\}-G(t;u)G(s,v),\\
A_{21}(t,s;u,v)&=\mathbb{E}\left\{\mathbb{E}[I(u^\top Z\le t|X)]I(v^\top Z\le s)\right\}-G(t;u)G(s,v).
\end{aligned}$$

2.7.2 Example 3.3: General One-sided and Two-sided Models

We now verify conditions in the previous section for the one-sided and two-sided models studied in Chernozhukov and Hong (2004), see also Hirano and Porter (2003). Following Chernozhukov and Hong (2004), let the density function of error term $\epsilon = y - g(X,\theta)$ be $f_\epsilon(\epsilon|X,\theta,\gamma)$. In one-sided models, $f_\epsilon(\epsilon|X,\theta,\gamma)=0$ for $\epsilon<0$ and

$$F(y|x,\theta,\gamma)=\begin{cases}\int_0^{y-g(x,\theta)}f_\epsilon(\epsilon|x,\theta,\gamma)\mathrm{d}\epsilon & \text{if } y\ge g(x,\theta),\\ 0 & \text{otherwise.}\end{cases}$$

In two-sided models,

$$f_\epsilon(\epsilon|X,\theta,\gamma) = \begin{cases} f_{L,\epsilon}(\epsilon|X,\theta,\gamma) & \text{if } \epsilon < 0, \\ f_{U,,\epsilon}(\epsilon|X,\theta,\gamma) & \text{if } \epsilon \geq 0 \end{cases}$$

and

$$F(y|x,\theta,\gamma) = \begin{cases} \int_{-\infty}^{y-g(x,\theta)} f_{L,\epsilon}(\epsilon|x,\theta,\gamma)\mathrm{d}\epsilon & \text{if } y < g(x,\theta), \\ \int_{-\infty}^{0} f_{L,\epsilon}(\epsilon|x,\theta,\gamma)\mathrm{d}\epsilon + \int_{0}^{y-g(x,\theta)} f_{U,\epsilon}(\epsilon|x,\theta,\gamma)\mathrm{d}e & \text{if } y \geq g(x,\theta). \end{cases}$$

Lipchitz Continuity of $F(y|x,\cdot)$

Suppose $F(y|x,\psi)$ is absolutely continuous with respect to ψ for each y and x, and $\sup_{y\neq g(x,\psi)} \left|\frac{\partial F(y|x,\psi)}{\partial \psi}\right|$ is uniformly bounded by a finite absolute constant. Then $F(y|x,\psi)$ is Lipchitz continuous with respect to ψ:

$$|F(y|x,\psi) - F(y|x,\psi')| \leq C\|\psi - \psi'\|.$$

This condition is stronger than equation (2.21) in the sense that a Lipchitz constant does not depend on (y,x).

One-sided Model Note that

$$\frac{\partial F(y|x,\psi)}{\partial \psi} = \begin{cases} 0 & \text{if } y < g(x,\theta) \\ -\frac{\partial g(x,\psi)}{\partial \psi} f_\epsilon(y - g(X,\psi)|x,\psi), + \int_0^{y-g(x,\psi)} \frac{\partial f_\epsilon(\epsilon|x,\psi)}{\partial \psi}\mathrm{d}\epsilon & \text{if } y > g(x,\theta). \end{cases}$$

Here, $\psi = (\theta^\top, \gamma^\top)^\top$, and $\partial g(x,\psi)/\partial\psi := [\partial g(x,\theta)/\partial\theta^\top, 0]^\top$ because $g(x,\theta)$ does not contain γ.

Lemma 2.7.5. *For one-sided model, let's assume that $f_\epsilon(\epsilon|x,\psi)$ is upper semi-continuous in (ϵ,ψ) for each x, and there exists $\bar{f}_\epsilon(\epsilon|x)$ such that*

$$\left\|\frac{f_\epsilon(\epsilon|x,\psi)}{\partial\psi}\right\| \leq \bar{f}_\epsilon(\epsilon|x) \textit{ for all } \epsilon, x, \psi, \textit{ and } \int_0^{y-g(x,\psi)} \bar{f}_\epsilon(\epsilon|x)\mathrm{d}\epsilon \textit{ is finite for each } x, \psi, \textit{ and } y.$$

In addition, we assume that

$$\sup_{x,\psi}\left\|\frac{\partial g(x,\psi)}{\partial\psi}\right\|<\infty,\quad \sup_{\epsilon.x,\psi}|f_\epsilon(\epsilon|x,\psi)|<\infty,\ \textit{and}\ \sup_{y,x,\psi}\left|\int_0^{y-g(x,\psi)}\frac{\partial f_\epsilon(\epsilon|y,\psi)}{\partial\psi}\mathrm{d}\epsilon\right|<\infty.$$

Then, $F(y|x,\psi)$ is Lipchitz with respect to ψ uniformly in y and x. That is, there exists a finite absolute constant C such that for any $\psi,\psi'\in\Psi$,

$$F(y|x,\psi)-F(y|x,\psi')\leq C\|\psi-\psi'\|.$$

The first two assumptions in lemma 2.7.5 ensure validity of interchanging the limit operation and integration. The third and fourth assumptions are included in Condition C2 in Chernozhukov and Hong (2004). The fifth condition is slightly stronger than their assumption that $\frac{\partial f_\epsilon(\epsilon|y,\psi)}{\partial\psi}$ is uniformly bounded. But we do not require that $f_\epsilon(0|x,\psi)>\eta>0$ as in Chernozhukov and Hong (2004).

Two-sided Model When $y<g(x,\psi)$,

$$\frac{\partial F(y|X,\psi)}{\partial\psi}=-\frac{\partial g(x,\psi)}{\partial\psi}f_{L,\epsilon}(y-g(x,\psi)|x,\psi)+\int_{-\infty}^{y-g(x,\psi)}\frac{\partial f_{L,\epsilon}(\epsilon|x,\psi)}{\partial\psi}\mathrm{d}\epsilon;$$

when $y>g(x,\psi)$,

$$\frac{\partial F(y|X,\psi)}{\partial\psi}=-\frac{\partial g(X,\psi)}{\partial\psi}f_{U,\epsilon}(y-g(X,\psi)|X,\psi)+\int_{-\infty}^{0}\frac{\partial f_{L,\epsilon}(\epsilon|X,\psi)}{\partial\psi}\mathrm{d}\epsilon+\int_0^{y-g(x,\psi)}\frac{\partial f_{U,\epsilon}(\epsilon|x,\psi)}{\partial\psi}\mathrm{d}\epsilon.$$

Lemma 2.7.6. *For two-sided model, let's assume that $f_{L,\epsilon}(\epsilon|x.\psi)$ and $f_{U,\epsilon}(\epsilon|x,\psi)$ are continuous in (ϵ,ψ) for each x, and there exist $\bar f_{L,\epsilon}(y|x)$ and $\bar f_{L,\epsilon}(y|x)$ such that*

$$\left\|\frac{f_{L,\epsilon}(y|x,\psi)}{\partial\psi}\right\|\leq\bar f_{L,\epsilon}(y|x),\quad\left\|\frac{f_{U,\epsilon}(\epsilon|x,\psi)}{\partial\psi}\right\|\leq\bar f_{U,\epsilon}(\epsilon|x),$$

$$\sup_{x,\psi}\int_{-\infty}^{0}\bar f_{L,\epsilon}(\epsilon|x)\mathrm{d}\epsilon<\infty,\ \textit{and}\ \int_0^{y-g(x,\psi)}\bar f_{U,\epsilon}(\epsilon|x)\mathrm{d}\epsilon<\infty\ \textit{for each}\ y,x,\textit{and}\ \psi.$$

In addition, we assume that

$$\sup_{x,\psi}\left\|\frac{\partial g(x,\psi)}{\partial\psi}\right\|<\infty,\quad\sup_{\epsilon,x,\psi}|f_{L,\epsilon}(\epsilon|x,\psi)|<\infty,\quad\sup_{\epsilon,x,\psi}|f_{U,\epsilon}(\epsilon|x,\psi)|<\infty,$$

$$\sup_{y<g(x,\psi),x,\psi}\left\|\int_{-\infty}^{y-g(x,\psi)}\frac{\partial f_{L,\epsilon}(\epsilon|x,\psi)}{\partial\psi}\mathrm{d}\epsilon\right\|<\infty,\quad\sup_{y\geq g(x,\psi),x,\psi}\left\|\int_0^{y-g(x,\psi)}\frac{\partial f_{U,\epsilon}(\epsilon|x,\psi)}{\partial\psi}\mathrm{d}\epsilon\right\|<\infty.$$

Then, $F(y|x,\psi)$ is Lipchitz continuous with respect to ψ uniformly in y, x. That is, there exists an absolute positive constant C such that for any $\psi, \psi' \in \Psi$,

$$F(y|x,\psi) - F(y|x,\psi') \leq C\|\psi - \psi'\|.$$

Like for one-sided model, the first two assumptions in lemma 2.7.6 ensure validity of interchanging the limit operation and integration. The third and fourth assumptions hold under Condition C2 in Chernozhukov and Hong (2004). The fifth and sixth conditions are slightly stronger than their assumption that $\frac{\partial f_{L,\epsilon}(\epsilon|x,\psi)}{\partial \psi}, \frac{\partial f_{U,\epsilon}(\epsilon|x,\psi)}{\partial \psi}$ are uniformly bounded. But we do not require that

$$\lim_{\epsilon \downarrow 0} f_{U,\epsilon}(\epsilon|x,\psi) - \lim_{\epsilon \uparrow 0} f_{L,\epsilon}(\epsilon|x,\psi) > \eta > 0$$

as in Chernozhukov and Hong (2004).

Norm Differentiability

Suppose that all assumptions in Lemmas 2.7.5 and 2.7.6 hold. Then,

$$\frac{\partial}{\partial \psi}\mathbb{E}[F(u_1^{-1}(s - u_2^\top X)|X,\psi)] = \mathbb{E}\left[\frac{\partial}{\partial \psi}F(u_1^{-1}(s - u_2^\top X)|X,\psi)\right].$$

For simplicity of notation, let

$$\begin{aligned}
G(u,s,\psi) &= \mathbb{E}[F(u_1^{-1}(s - u_2^\top X_t)|X_t,\psi)], \\
G_t(u,s,\psi) &= F(u_1^{-1}(s - u_2^\top X_t)|X_t,\psi), \\
D_t(u,s,\psi) &= \frac{\partial}{\partial \psi}F(u_1^{-1}(s - u_2^\top X_t)|X_t,\psi), \\
D(u,s,\psi) &= \mathbb{E}\left[\frac{\partial}{\partial \psi}F(u_1^{-1}(s - u_2^\top X_t)|X_t,\psi)\right], \\
R(s,u,\psi) &= G(u,s,\psi) - G(u,s,\psi_0) - (\psi - \psi_0)^\top D(u,s,\psi), \\
R_t(s,u,\psi) &= G_t(u,s,\psi) - G_t(u,s,\psi_0) - (\psi - \psi_0)^\top D_t(u,s,\psi).
\end{aligned}$$

Here, we set $D_t(s,u,\psi) = 0$ at kink point s. Note that $R(s,u,\psi) = \mathbb{E}[R_t(s,u,\psi)]$.

We will show that for any $\tau_T \to 0$,

$$\sup_{\|\psi-\psi_0\|\le\tau_T} \frac{T\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}|R(s,u,\psi)|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|^2)} = o(1)$$

by proving that

$$\sup_{\|\psi-\psi_0\|\le\tau_T} \frac{T\mathbb{E}\left[\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}|R_t(s,u,\psi)|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right]}{(1+\|\sqrt{T}(\psi-\psi_0)\|^2)} = o(1)$$

for any $\tau_T \to 0$.

First, we assume that $g(x,\psi)$ is Lipshitz continuous at ψ_0 uniformly in x :

$$|g(x,\psi)-g(x,\psi')| \le C_g\|\psi-\psi'\|$$

and all assumptions in Lemmas 2.7.5 and 2.7.6 so $|D_t(s,u,\psi)|$ is uniformly bounded by an absolute constant, and for any $\psi,\psi'\in\Psi$,

$$|F(y|x,\psi)-F(y|x,\psi')| \le C\|\psi-\psi'\|.$$

When $u^{-1}(s-u_2^\top X_i) \in A_T := [g(x,\psi)\wedge g(x,\psi_0) - C_g\tau_T, g(x,\psi)\vee g(x,\psi_0) + C_g\tau_T]$,

$$|R_t(s,u,\psi)| \le C\|\psi-\psi'\|.$$

for some absolute constant C. This implies that when $\|\psi-\psi_0\|\le\tau_T$,

$$\begin{aligned}&\int_{u\in\mathbb{S}^{d-1}}\int_{u^{-1}(s-u_2^\top X_i)\in A_T}|R_t(s,u,\psi)|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\&\le C\|\psi-\psi_0\|^2|u_1|\left[(|g(X_i,\psi)-g(X_i,\psi_0)|+2C\tau_T\right] = O(\tau_T^3).\end{aligned}$$

When $\|\psi-\psi_0\|\le\tau_T$ and $u^{-1}(s-u_2^\top X_t)\notin A_T$, by Taylor theorem, we have

$$\begin{aligned}&F(u^{-1}(s-u_2^\top X_t)|x,\psi) - F(u^{-1}(s-u_2^\top X_t)|x,\psi_0) - (\psi-\psi)^\top\left(\frac{\partial F(u^{-1}(s-u_2^\top X_t)|x,\psi_0)}{\partial\psi}\right)\\&\quad= (\psi-\psi_0)^\top\frac{\partial^2 F(u^{-1}(s-u_2^\top X_t)|X_t,\psi)}{\partial\psi\partial\psi'}(\psi-\psi_0)\end{aligned}$$

for some ψ. When the weight function is integrable,

$$\int_{u\in\mathbb{S}^{d-1}}\int_{u^{-1}(s-u_2^\top X_t)\notin A_T}|R_t(s,u,\psi)|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \le \sup_{y,x,\psi}\left\|\frac{\partial^2 F(u_1^{-1}(s-u_2^\top x)|x,\psi)}{\partial\psi\partial\psi'}\right\|^2 \|\psi-\psi_0\|^4.$$

Therefore, it is enough to show that

$$\sup_{y,x,\psi}\left\|\frac{\partial^2 F(y|x,\psi)}{\partial\psi\partial\psi'}\right\| < \infty.$$

In the one-sided model, when $y > g(x,\psi)$,

$$\begin{aligned}\frac{\partial^2 F(y|x,\psi)}{\partial\psi\partial\psi'} &= -\frac{\partial^2 g(x,\psi)}{\partial\psi\partial\psi'}f_\epsilon(y-g(x,\psi)|x,\psi)\\ &\quad+\frac{\partial g(x,\psi)}{\partial\psi}\left[\frac{\partial g(x,\psi)}{\partial\psi'}f'_\epsilon(y-g(x,\psi)|x,\psi)+\frac{\partial f_\epsilon(y-g(x,\psi)|x,\psi)}{\partial\psi'}\right]\\ &\quad-\frac{\partial g(x,\psi)}{\partial\psi}\frac{f_\epsilon(y-g(x,\psi)|x,\psi)}{\partial\psi'}+\int_0^{y-g(x,\psi)}\frac{\partial^2 f_\epsilon(\epsilon,x,\psi)}{\partial\psi\partial\psi'}\mathrm{d}\epsilon.\end{aligned}$$

Lemma 2.7.7. *For one-sided model, in addition to all conditions in Lemma 2.7.5, let us assume that $\partial f_\epsilon(\epsilon|x,\psi)/\partial\psi$ is upper semi-continuous in ϵ,ψ, and there exists $f_\epsilon(\epsilon|x)$ such that*

$$\left\|\frac{\partial f_\epsilon(\epsilon|X,\psi)}{\partial\psi\partial\psi'}\right\| \le f(y|x),\ \int_0^{y-g(x,\theta)} f_\epsilon(\epsilon|x)\mathrm{d}\epsilon < \infty \textit{ for all } y,x,\psi,$$

and

$$\sup_{x,\psi}\left\|\frac{\partial^2 g(x,\psi)}{\partial\psi\partial\psi'}\right\|,\ \sup_{\epsilon,x,\psi}\left\|\frac{\partial f_\epsilon(\epsilon|x,\psi)}{\partial\psi}\right\|,\ \sup_{\epsilon,x,\psi}\left\|\frac{\partial f_\epsilon(\epsilon|x,\psi)}{\partial\epsilon}\right\|,\ \sup_{y,x,\psi}\left\|\int_0^{y-g(x,\psi)}\frac{\partial^2 f_\epsilon(\epsilon|x,\psi)}{\partial\psi\partial\psi'}\mathrm{d}\epsilon\right\|$$

are bounded above by finite constants. Then, $\widehat{Q}_T(\cdot;u,\psi)$ is norm-differentiable at $\psi=\psi_0$ with $D(s;u,\psi)=\mathbb{E}\left[\frac{\partial F(u_1^{-1}(s-u_2^\top X_i)|X_i,\psi)}{\partial\psi}\right]$.

In the two-sided model, when $y < g(x,\psi)$,

$$\begin{aligned}\frac{\partial^2 F(y|x,\psi)}{\partial\psi\partial\psi'} &= -\frac{\partial^2 g(x,\psi)}{\partial\psi\partial\psi'}f_{L,\epsilon}(y-g(x,\psi)|x,\psi)\\ &\quad+\frac{\partial g(x,\psi)}{\partial\psi}\left[\frac{\partial g(x,\psi)}{\partial\psi'}f'_{L,\epsilon}(y-g(x,\psi)|x,\psi)+\frac{\partial f_{L,\epsilon}(y-g(x,\psi)|x,\psi)}{\partial\psi'}\right]\\ &\quad-\frac{\partial g(x,\psi)}{\partial\psi}\frac{f_{L,\epsilon}(y-g(x,\psi)|x,\psi)}{\partial\psi'}+\int_{-\infty}^{y-g(x,\psi)}\frac{\partial^2 f_{L,\epsilon}(\epsilon,x,\psi)}{\partial\psi\partial\psi'}\mathrm{d}\epsilon.\end{aligned}$$

When $y > g(x,\psi)$,

$$
\begin{aligned}
\frac{\partial^2 F(y|x,\psi)}{\partial\psi\partial\psi'} = &-\frac{\partial^2 g(x,\psi)}{\partial\psi\partial\psi'} f_{U,\epsilon}(y-g(x,\psi)|x,\psi) \\
&+ \frac{\partial g(x,\psi)}{\partial\psi}\left[\frac{\partial g(x,\psi)}{\partial\psi'} f'_{U,\epsilon}(y-g(x,\psi)|x,\psi) + \frac{\partial f_{U,\epsilon}(y-g(x,\psi)|x,\psi)}{\partial\psi'}\right] \\
&- \frac{\partial g(x,\psi)}{\partial\psi}\frac{f_{U,\epsilon}(y-g(x,\psi)|x,\psi)}{\partial\psi'} + \int_{-\infty}^{y-g(x,\psi)} \frac{\partial^2 f_{U,\epsilon}(\epsilon,x,\psi)}{\partial\psi\partial\psi'}\mathrm{d}\epsilon \\
&+ \int_{-\infty}^{y-g(x,\psi)} \frac{\partial^2 f_{L,\epsilon}(\epsilon,x,\psi)}{\partial\psi\partial\psi'}\mathrm{d}\epsilon.
\end{aligned}
$$

Lemma 2.7.8. *For two-sided model, in addition to all conditions in Lemma 2.7.6, let us assume that $\partial f_{L,\epsilon}(\epsilon|x,\psi)/\partial\psi$ and $\partial f_{U,\epsilon}(y|x,\psi)/\partial\psi$ are continuous in y,ψ, and there exist $f_{L,\epsilon}(\epsilon|x)$ and $f_{U,\epsilon}(\epsilon|x)$ such that*

$$\left\|\frac{\partial f_{L,\epsilon}(\epsilon|x,\psi)}{\partial\psi\partial\psi'}\right\| \le f_{L,\epsilon}(\epsilon|x), \ \int_{-\infty}^{0} f_{L,\epsilon}(\epsilon|x)\mathrm{d}\epsilon < \infty \text{ for all } x,$$

$$\left\|\frac{\partial f_{U,\epsilon}(\epsilon|x,\psi)}{\partial\psi\partial\psi'}\right\| \le f_{U,\epsilon}(\epsilon|x), \ \int_{0}^{y-g(x,\psi)} f_{U}(\epsilon|x)\mathrm{d}\epsilon < \infty \text{ for all } y,x,\psi,$$

and

$$\sup_{x,\psi}\left\|\frac{\partial^2 g(x,\psi)}{\partial\psi\partial\psi'}\right\|, \sup_{\epsilon,x_i,\psi}\left\|\frac{\partial f_{L,\epsilon}(\epsilon|x,\psi)}{\partial\psi}\right\|, \sup_{\epsilon,x_i,\psi}\left\|\frac{\partial f_{U,\epsilon}(\epsilon|x,\psi)}{\partial\psi}\right\|, \sup_{\epsilon,x_i,\psi}\left\|\frac{\partial f_{L,\epsilon}(\epsilon|x,\psi)}{\partial\epsilon}\right\|, \sup_{\epsilon,x_i,\psi}\left\|\frac{\partial f_{U,\epsilon}(\epsilon|x,\psi)}{\partial\epsilon}\right\|,$$
$$\sup_{y<g(x,\beta),x,\psi}\left\|\int_{-\infty}^{y-g(x,\psi)} \frac{\partial^2 f_{L,\epsilon}(\epsilon|x_i,\psi)}{\partial\psi\partial\psi'}\mathrm{d}\epsilon\right\|, \sup_{y,x,\psi}\left\|\int_{0}^{y-g(x.\psi)} \frac{\partial^2 f_{U,\epsilon}(\epsilon|x,\psi)}{\partial\psi\partial\psi'}\mathrm{d}\epsilon\right\|$$

are bounded above by finite constants. Then, $\widehat{Q}_T(\cdot;u,\psi)$ is norm-differentiable at $\psi=\psi_0$ with $D(s;u,\psi) = \mathbb{E}\left[\frac{\partial F(u_1^{-1}(s-u_2^\top X_t)|X_t,\psi)}{\partial\psi}\right]$.

2.8 Generative Models

In this section, we will investigate the asymptotic property of the SMSD estimator $\hat{\psi}_{T,m}$ described in Example 2.3.1. For simplicity, we consider the case where $K=1$ so we have $\widehat{Q}_T(\cdot;u,\psi) = \widehat{Q}_m(\cdot;u,\psi)$, where $\widehat{Q}_m$ is computed from a simulated sample $\{\widetilde{Z}_t^\psi\}_{t=1}^m$ with size m from the generative model with parameter ψ. Here, m is a function of T. As long as K is fixed, the conclusions carry over.

2.8.1 Unconditional Generative Model

Following GAN literature (c.f Arjovsky et al. (2017) or Kaji et al. (2020)), we consider the case where Z_t is generated from some latent variable H_t whose distribution is known.

Condition 2.8.1 (Unconditional Generative Model). *$Z_t = L(H_t, \psi_0)$ for some known function L and some non-degenerate random variable H_t with a known distribution $F_H(\cdot)$.*

In the unconditional generative model, we generate $\{\widetilde{H}_t\}_{t=1}^m$ from $F_H(\cdot)$ once, and construct $\{\widetilde{Z}_t^\psi\}_{t=1}^m$, where $\widetilde{Z}_t^\psi = L(\widetilde{H}_t, \psi)$ for each $\psi \in \Psi$.

For each $u \in \mathbb{S}^{d-1}$, we denote $G_m(s; u, \psi)$ to be the empirical distribution of $\{u^\top \widetilde{Z}_t^\psi\}_{t=1}^m$:

$$G_m(s; u, \psi) = \frac{1}{m}\sum_{t=1}^m I(u^\top \widetilde{Z}_t^\psi \le s) = \frac{1}{m}\sum_{t=1}^m I(u^\top L(\widetilde{H}_t, \psi) \le s).$$

The asymptotic theory of the SMSD estimator $\widehat{\psi}_{T,m}$ is based on the empirical process

$$\mathcal{G} := \{\sqrt{m}(G_m(s; u, \psi) - G(s; u, \psi)) : s \in \mathbb{R}, u \in \mathbb{S}^{d-1}, \psi \in \Psi\}.$$

To derive the theoretical results, we will put restrictions on the following class of functions:

$$\mathcal{F} = \{u^\top L(\cdot, \psi) - s : u \in \mathbb{S}^{d-1}, s \in \mathbb{R}, \psi \in \Psi\}.$$

Following van der Vaart and Wellner (1996), we define the Hölder class. For any vector $k = (k_1, \dots, k_d)$, let

$$D^{\bar{k}} = \frac{\partial^k}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}},$$

where $\bar{k} = \sum_{j=1}^d k_j$. For a function $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$, let

$$\|f\|_\alpha = \max_{\bar{k} \le \underline{\alpha}} \sup_{x \in \mathcal{X}} |D^k f(x)| + \max_{\bar{k} = \underline{\alpha}} \sup_{x \ne y, x, y \in \mathcal{X}} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}},$$

where $\underline{\alpha}$ is the greatest integer smaller than α.

Definition 2.8.1 (Hölder Class). *Let $C_M^\alpha(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_\alpha \le M < \infty$.*

Similar to Brown and Wegkamp (2002) and Torgovitsky (2017), we impose some sufficient assumptions on $\mathcal{F}$ to ensure weak convergence of $\mathcal{G}$.

Condition 2.8.2. *Suppose the class of functions $\mathcal{F}$ satisfies either (a) or (b) below.*

(a) It is a subset of a finite dimensional vector space

(b) The support of H_t, $\mathcal{H}$, is bounded, and the density of H_t is uniformly bounded. Moreover, $\mathcal{F}$ is a subset of $\mathcal{C}_M^{\alpha}(\mathcal{H})$ with $\alpha > \dim(H_t)$.

Remark 2.8.1. When H_t has unbounded support, one can relax Condition 2.8.2 (b) by following the proof of Corollary 2.7.4 in van der Vaart and Wellner (1996).

Let

$$\mathcal{F}_I := \{I(u^\top L(\cdot, \psi) \leq s) : u \in \mathbb{S}^{d-1}, s \in \mathbb{R}, \psi \in \Psi\}$$

and $N_{[]}(\epsilon, \mathcal{F}_I, \|\cdot\|_H)$ be the ϵ-bracketing number of $\mathcal{G}$, where $\|f\|_H = \left(\int (f(h; u, s, \psi))^2 \mathrm{d}F_H(h)\right)^{1/2}$ for $f \in \mathcal{F}_I$. (We refer to Definition 2.1.6 in van der Vaart and Wellner (1996) for the definition of ϵ-bracketing number.)

Lemma 2.8.1. *Suppose that Conditions 2.8.1 and 2.8.2 hold. In addition, assume that $\{\widetilde{H}_t\}$ is a stationary and ergodic process with the β-mixing coefficient β_k, and one of the following conditions holds.*

(i) Condition 2.8.2.(a) holds, and $\lim_{k\to\infty} k^r \beta_k = 0$ for some $r > 1$.

(ii) Condition 2.8.2.(b) holds, and one of the following holds.

- $\beta_k = O(k^{-b})$ *with* $b > r/(r-1)$, *and* $\int_0^1 \epsilon^{-r/(b(r-1))}\Big(\log(N_{[]}(\epsilon, \mathcal{F}_I, \|\cdot\|_H))\Big)^{1/2} \mathrm{d}\epsilon < \infty$.
- $\log(N_{[]}(\epsilon, \mathcal{F}_I, \|\cdot\|_H)) = O(\epsilon^{-2\zeta})$ *with* $\zeta \in (0,1)$, *and* $\sum_{k\geq 0} k^{-1/2} \beta_k^{(1-\zeta)\frac{r-1}{2r}} < \infty$.
- $\log(N_{[]}(\epsilon, \mathcal{F}_I, \|\cdot\|_H)) = O(\log|\epsilon|)$, *and* $\sum_{k\geq 0} k^{-1} \sqrt{(\log k)^{-1} \sum_{\ell\geq k} \beta_\ell^{1-1/r}} < \infty$.

Then, $\mathcal{G}$ is P-Donsker so that $m^{1/2} \sup_{u,s,\psi} |G_m(s; u, \psi) - G(s, u, \psi)| = O_p(1)$.

Proof. For Case (i) in the lemma, we have weak convergence by Lemma B.3.1. For Case (ii), we have weak convergence by Theorem 1 and Example 4 (p. 405) in Doukhan et al. (1995). $\square$

The SMSWD Estimator

For the SMSWD estimator, $\widehat{Q}_T(\cdot; u, \psi) = G_m^{-1}(\cdot; u, \psi)$.

Condition 2.8.3. *$Q(\cdot; u, \psi) = G^{-1}(\cdot, u, \psi)$ is norm-differentiable at $\psi = \psi_0$.*

Using the same argument as in the proof of Lemma 2.6.2, we have the following result.

Lemma 2.8.2. *Suppose $G(t; u, \psi)$ is absolute continuous and differentiable with respect to t for every u, ψ except for $(u, \psi) \in \mathcal{N}(\theta) := \{u \in \mathbb{S}^{d-1}; G(t; u, \psi)$ is degenerate$\}$. Let us assume that there is an absolute positive constant C such that*

$$\sup_{\psi \in \Psi} \int_{u \in \mathbb{S}^{d-1}/\mathcal{N}(\theta)} \int_0^1 \frac{w(s)}{g^2(G^{-1}(s; u, \psi); u, \psi)} \mathrm{d}s\mathrm{d}\varsigma(u) < C,$$

and

$$\sup_{\theta \in \Psi} \sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}(\theta)} \sup_{s \in supp\{w\}} \frac{g(G^{-1}(s; u, \psi); u, \psi)}{g(G^{-1}(s(s; u, \psi); u, \psi); u, \psi)} = O_p(1) \tag{2.22}$$

Then, Conditions in Lemma 2.8.1 imply

$$\sup_{\psi, s, u} m \int_0^1 (G_m^{-1}(s; u, \psi) - G^{-1}(s; u, \psi))^2 w(s) \mathrm{d}s\mathrm{d}\varsigma(u) = O_p(1). \tag{2.23}$$

Remark 2.8.2. When $w(s) = [\delta, 1 - \delta]$ with $\delta > 0$, we can show equation (2.22) under the following assumptions.

$$\begin{aligned}
&\sup_{\psi} \sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}(\theta), s} (G(s; u, \psi_0)(1 - G(s; u, \psi_0))) \left| \frac{g'(s; u, \psi_0)}{g^2(s; u, \psi_0)} \right| \\
&\quad = \sup_{\psi} \sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}(\theta), s} (s(1 - s)) \left| \frac{g'(G^{-1}(s; u, \psi_0); u, \psi_0)}{g^2(G^{-1}(s; u, \psi_0); u, \psi_0)} \right| < \infty
\end{aligned} \tag{2.24}$$

by following the proof of Lemma 1 in Csorgo and Revesz (1978). Proof is the same as the proof in Remark 2.6.1.

Corollary 2.8.1. *Suppose Assumptions 2.5.1 (i), 2.5.2, 2.5.3, 2.5.4 (i), 2.5.6, Condition 2.8.3, and conditions in Lemma 2.8.2 hold. When $T/m = o(1)$, we have*

$$\sqrt{T}(\hat{\psi}_{T,m} - \psi_0) \xrightarrow{d} N(0, B_0^{-1}\Omega_0 B_0^{-1}),$$

where Ω_0 is the asymptotic variance in Lemma 2.6.5.

Proof. From Lemma 2.8.2, Equation (2.23) holds. This implies Assumption 2.5.1 (ii) and Assumption 2.5.4 (ii) are satisfied.

When $T/m = o(1)$ and Conditions 2.23 and 2.8.3 hold, $Q_m(\cdot; u, \psi)$ is norm-differentiable at $\psi = \psi_0$ with $\widehat{D}_n(s; u, \psi_0) = D(s; u, \psi_0)$ because

$$T \int_{u\in\mathbb{S}^{d-1}} \int_0^1 (\widehat{Q}_m(s; u, \psi) - Q(s; u, \psi) - \widehat{Q}_m(s; u, \psi_0) - Q(s; u, \psi) + Q(s; u, \psi_0))^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u) = o(1).$$

When $T/m = o(1)$, $\int_{u\in\mathbb{R}^{d-1}} \int_0^1 \|D(s; u, \psi_0)\|^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u) < \infty$, and Equation (2.23) holds, we have

$$\sqrt{T} \int_{u\in\mathbb{S}^{s-1}} \int_0^1 (\widehat{Q}_T(s; u, \psi_0) - Q(s; u; \psi_0)) D(s; u, \psi_0) w(s) \mathrm{d}s \mathrm{d}\varsigma(u) = o_p(1).$$

Therefore, we have the desired result when $T/m = o(1)$. □

The SMSCD Estimator

We set $Q_T(s; u, \psi) = G_m(s; u, \psi)$ for the SMSCD estimator.

From Lemma 2.8.1, $\mathcal{G}$ is P-Donsker. For stochastic equicontinuity of $\mathcal{G}$, we impose Lipschitz condition on functions in $\mathcal{F}$ by following Brown and Wegkamp (2002).

Condition 2.8.4. *All functions in $\mathcal{F}$ are Lipschitz continuous with respect to ψ and u.*

Lemma 2.8.3. *Let us assume that Assumptions 2.5.1, 2.5.2, 2.5.4, 2.5.6, and Conditions 2.6.2, 2.7.1, 2.8.4, and all assumptions in Lemma 2.8.1 hold. Then, we have the following results.*

(a) *When $T/m = o(1)$, we have $\sqrt{n}(\hat{\psi}_{T,m} - \psi_0) \Rightarrow N(0, B_0^{-1}\Omega_0 B_0^{-1})$;*

(b) When $T = m$, we have $\sqrt{n}(\hat{\psi}_{T,m} - \psi_0) \Rightarrow N(0, 2B_0^{-1}\Omega_0 B_0^{-1})$,

where Ω_0 is the asymptotic variance in Corollary 2.6.7.

Proof. We will prove the result by showing that the assumptions in Section 2.5 hold.

When all conditions in Lemma 2.8.1 hold, we have $\{\sqrt{m}(G_m(s;u,\psi) - G(s;u,\psi)); s \in \mathbb{R}, u \in \mathbb{S}^{d-1}, \psi \in \Psi\}$ is P-Donsker with uniform norm. Then, when the weight function is integrable, we have

$$\sup_{\psi\in\Psi} m \int_{u\in\mathbb{S}^{d-1}} \int_{-\infty}^{\infty} (G_m(s;u,\psi) - G(s;u,\psi))^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u)$$
$$\leq \sup_{\psi\in\Psi, u\in\mathbb{S}^{d-1}, s\in\mathbb{R}} \left(\sqrt{m}(G_m(s;u,\psi) - G(s;u,\psi))\right)^2 = O_p(1).$$

Therefore, Assumption 2.5.1 (i) and (ii) hold when $T/m = O(1)$, and 2.5.4 (i) and (ii) hold.

Also, $\mathcal{G}$ satisfies stochastic equicontinuity by following the proof of Lemma 3 in Brown and Wegkamp (2002). Then, we have

$$m \int_{u\in\mathbb{S}^{d-1}} \int_{-\infty}^{\infty} (G_m(s;u,\psi) - G_m(s;u,\psi_0) - G(s;u,\psi) + G(s;u,\psi_0))^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u)$$
$$\leq \left[\sup_{u\in\mathbb{S}^{d-1}, t\in\mathbb{R}} \sqrt{m}(G_m(s;u,\psi) - G(s;u,\psi)) - \sqrt{m}(G_m(s;u,\psi_0) - G(s;u,\psi_0))\right]^2 = o(\|(\psi - \psi_0)\|).$$

Therefore, when $T/m = O(1)$, $Q_T(\cdot;u,\psi)$ is norm-differentiable at $\psi = \psi_0$ with $D_T(s;u,\psi_0) = D(s;u,\psi_0)$ by following the proof of Lemma 2.7.3. Because $D_T(s;u,\psi_0) = D(s;u,\psi_0)$, Assumption 2.5.4 (iii) hold.

Because $\{Z_i\}$ and $\{\widetilde{Z}_t^{\psi_0}\}$ are independent, and have the same distribution, we have

$$\begin{pmatrix} \sqrt{T}\int_{u\in\mathbb{S}^{s-1}}\int_0^1 (G_T(s;u) - G(s;u)) D(s;u,\psi_0) w(s) \mathrm{d}s \mathrm{d}\varsigma(u) \\ \sqrt{m}\int_{u\in\mathbb{S}^{s-1}}\int_0^1 (G_m(s;u,\psi_0) - G(s;u;\psi_0)) D(s;u,\psi_0) w(s) \mathrm{d}s \mathrm{d}\varsigma(u) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \Omega_0 & 0 \\ 0 & \Omega_0 \end{pmatrix}\right)$$

where Ω_0 is the variance in Lemma 2.6.7. Therefore, we have the desired result. □

2.8.2 Conditional Generative Model

Similar to the previous section, we consider the case where Y_t is generated from X_t and some latent variable H_t with a known distribution as described in Condition 2.8.5.

Condition 2.8.5 (Conditional Generative Model)**.** $Y_t = L(H_t, X_t, \psi_0)$ *for some known function L and some random variable H_t with a known distribution $F_H(\cdot)$.*

Example 2.8.1. *In the BLP model introduced in Section 2.2.2,*

$$s_t = \sigma(\delta_t, x_t^{(2)}; \lambda_v) = \sigma(x_t^\top \beta + \xi_t, x_t^{(2)}; \lambda_v).$$

where $\xi_t \sim G(\cdot|x_t, \lambda_\xi)$. When ξ_t is a conditional generative model (i.e, $\xi_t = L_1(H_t, x_t, \lambda_\xi)$ for some function L_1),

$$s_t = \sigma(x_t^\top \beta + L_1(H_t, x_t, \lambda_\xi), x_t^{(2)}; \lambda_v) = L(H_t, x_t, \beta, \lambda_\xi, \lambda_v)$$

for some function L.

In the conditional generative model, we generate $\{\widetilde{H}_t\}_{t=1}^m$ from $F_H(\cdot)$ once, and construct $\{\widetilde{Y}_t^\psi\}_{t=1}^m$ where $\widetilde{Y}_t^\psi = L(\widetilde{H}_t, X_t, \psi)$ for each ψ in the estimation. Let's denote $\widetilde{Z}_t^\psi = (\widetilde{Y}_t^\psi, X_t^\top)^\top$.

In this section, we will consider a SMSCD estimator when $\{Z_t\}_{t=1}^T$ is a random sample, and Ψ is compact in line with Section 2.7. Then, $\widehat{Q}_T(\cdot; u, \psi) = G_T(\cdot; u, \psi)$, where

$$G_T(s; u, \psi) := \frac{1}{T}\sum_{t=1}^T I(u_1 \widetilde{Z}_t^\psi \leq s) = \frac{1}{T}\sum_{t=1}^T I(u_1 L(H_t, X_t, \psi) + u_2^\top X_t \leq s),$$

where $u = (u_1, u_2^\top)^\top$ with $u_1 \in \mathbb{R}$.

To show weak convergence of $\{\sqrt{T}(G_T(s; u, \psi) - G(s; u, \psi)); s \in \mathbb{R}, u \in \mathbb{S}^{d-1}, \psi \in \Psi\}$, we put restrictions on the class of functions:

$$\mathcal{F}_X := \{u_1 L(H_t, X_t, \psi) + u_2^\top X_t - s : \psi \in \Psi, (u_1, u_2^\top)^\top \in \mathbb{S}^{d-1}, s \in \mathbb{R}\}$$

Condition 2.8.6. *Suppose the class of functions $\mathcal{F}_X$ satisfies either (a) or (b) below.*

(a) It is a subset of a finite dimensional vector space

(b) The support of $(H_t^\top, X_t^\top)^\top$, $\mathcal{HX}$, is bounded and the joint density of (H_t, X_t) is uniformly bounded. Moreover, $\mathcal{F}$ is a subset of $\mathcal{C}_M^\alpha(\mathcal{HX})$ with $\alpha > \dim(H_t) + \dim(X_t)$.

Lemma 2.8.4. *Under Conditions 2.8.5 and 2.8.6,*

$$\mathcal{G}_X := \{\sqrt{T}(G_T(s;u,\psi) - G(s;u,\psi)); s \in \mathbb{R}, u \in \mathbb{S}^{d-1}, \psi \in \Psi\}$$

is P-Donsker.

Proof. When Condition 2.8.6 (a) holds, $\mathcal{F}_X$ is VC subgraph class, which impies $\mathcal{G}_X$ is P-Donsker. Under Condition 2.8.6 (b), we have same conclusion by Lemma 2 in Brown and Wegkamp (2002). (We also refer Corollary 2.7.3 of van der Vaart and Wellner (1996) for ϵ-Bracketing number for Holder class.) □

For stochastic equicontinuity of $\mathcal{G}_X$, we assume Lipschitz condition similar to Brown and Wegkamp (2002).

Condition 2.8.7. *Functions in $\mathcal{F}_X$ are Lipschitz with respect to (ψ, u, s).*

We can show the following lemma using the same arguments in Lemma 2.8.3.

Lemma 2.8.5. *Suppose Conditions 2.7.1, 2.8.5, 2.8.6, and 2.8.7 hold. Then, Assumptions 2.5.1 (ii) and 2.5.4 (ii) are satisfied, and $Q_T(\cdot; u, \psi)$ is norm-differentiable at $\psi = \psi_0$ with $D_T(s; u, \psi_0) = D(s; u, \psi_0)$.*

By applying CLT, we obtain the following result.

Lemma 2.8.6. *Suppose Assumption 2.7.1 and Condition 2.6.2 hold. Then, we have*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\int_{u\in\mathbb{R}^{d-1}}\int_{-\infty}^{\infty}\begin{pmatrix} I(u^\top Z_t \le s) - G(s;u) \\ I(u^\top \widetilde{Z}_t^{\psi_0} \le s) - G(s;u,\psi_0)\end{pmatrix} D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \xrightarrow{d} N(0, V_0)$$

where

$$V_0 = \int_{u\in\mathbb{S}^{d-1}}\int_{v\in\mathbb{S}^{d-1}}\int\int\begin{pmatrix} A_{11}(t,s;u,v) & A_{12}(t,s;u,v) \\ A_{21}(t,s;u,v) & A_{22}(t,s;u,v)\end{pmatrix} \otimes D(s;u,\psi_0)D(t;u,\psi_0)^\top w(s)w(t)\mathrm{d}s\ \mathrm{d}t\mathrm{d}\varsigma(u)\mathrm{d}\varsigma(v)$$

in which $\otimes$ is the Kronecker product, and

$$
\begin{aligned}
A_{11}(t,s;u,v) &= \mathbb{E}[I(u^\top Z \le t)I(v^\top Z \le s)] - G(t;u)G(s,v),\\
A_{22}(t,s;u,v) &= \mathbb{E}\left\{I(u^\top \widetilde{Z}^{\psi_0} \le t)I(v^\top \widetilde{Z} \le s)\right\} - G(t;u)G(s,v),\\
A_{12}(t,s;u,v) &= \mathbb{E}\left\{I(u^\top Z \le t)I(v^\top \widetilde{Z}^{\psi_0} \le s)\right\} - G(t;u)G(s,v),\\
A_{21}(t,s;u,v) &= \mathbb{E}\left\{I(u^\top \widetilde{Z}^{\psi_0} \le t)I(v^\top Z \le s)\right\} - G(t;u)G(s,v)
\end{aligned}
$$

Proof. The proof is skipped because it is trivial. □

Using Lemmas 2.8.5 and 2.8.6, we have the asymptotic distribution for $\hat{\psi}_{T,m}$ whose proof is the same as that of Lemma 2.8.3.

Corollary 2.8.2. *Suppose Assumptions 2.5.1 (i), 2.5.2, 2.5.4 (i), 2.5.6, and Conditions 2.6.2, 2.7.1, 2.8.5, 2.8.6, and 2.8.7 hold. When $m = T$, we have*

$$\sqrt{T}(\hat{\psi}_{T,m} - \psi_0) \xrightarrow{d} N(0, B_0^{-1}\Omega_0 B_0^{-1}).$$

where $\Omega_0 = (e_1^\top, -e_1^\top)V_0 \begin{pmatrix} e_1 \\ -e_1 \end{pmatrix}$ in which V_0 is the asymptotic variance in Lemma 2.8.6.

2.9 Numerical Results

In this section, we report some preliminary results on the accuracy of the asymptotic normal distribution of MSCD and MSCD estimators based on the three models: two singular stochastic models and a parameter-dependent support model. The first singular model is a simple state-space model, and the second is the term structure model. The parameter-dependent support model is an auction model.

2.9.1 A Singular State-Space Model

We consider a simple state-space model:

$$Y_t = a + bY_t^*, \quad Y_t = \rho Y_{t-1}^* + \epsilon_t, \quad \epsilon_t \sim N(0,1),$$

where $|\rho| < 1$, and $a = (a_1, a_2)^\top$ and $b = (b_1, b_2)^\top$ are 2 by 1 vectors. For identification, we assume that $b_1 > 0$. We set $Z_t = (Y_t^\top, Y_{t-1}^\top)^\top$ and estimate $\psi = (\rho, a_1, a_2, b_1, b_2)$ using the MSWD estimator. For each $u \in \mathbb{S}^3$, the projected variable $u^\top Z_t$ follows a normal distribution:

$$u^\top Z_t \sim N\left(u^\top \begin{pmatrix} a \\ a \end{pmatrix}, \frac{1}{1-\rho^2} u^\top \left[\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \otimes bb^\top\right] u\right),$$

where $\otimes$ is the Kronecker product.

We set $a = (0.1, 0.3)^\top$ and $b = (1, 0.5)^\top$, and consider two values of ρ: $\rho = 0.8$ and 0.3. The weight function is set to $w(s) = I(s \in [0.1, 0.9])$. We compute 2000 estimates of t-values for samples of size $T = 500, 1500, 2000, 2500,$ and 3000. Note that when we evaluate t-values, the standard deviation from the parameter estimates of Monte-Carlo simulation is used for all simulations in Section 2.9. Table 2.2 shows that p-values are not far from the nominal sizes for all sample sizes.

2.9.2 Discrete-Time Vasicek Model [6]

Version 1:

We consider discrete-time Vasicek model in Example 2.1 (Cont'd). We set $Z_t = (Y_t^\top, Y_{t-1}^\top)^\top$ and estimate $\psi = (c, \rho, \sigma, \lambda)$ using the MSWD estimator. We set $c = 0.07$, $\rho = 0.95$, $\sigma = 0.02$, and $\lambda = -0.005$. These values except for ρ are close to empirical results in Pastorello et al. (2003). The weight function is set to $w(s) = I(s \in [0.1, 0.9])$. We compute 2000 estimates of t-values from the samples of size $T = 500, 1500, 2000, 2500,$ and 3000. Table 2.3 shows that p-values are not far from the nominal sizes for all sample sizes.

Version 2:

We consider discrete-time Vasicek model in Example 2.1 (Cont'd). We set $Z_t = (Y_t^\top, Y_{t-1}^\top)^\top$ and estimate $\psi = (c, \rho, \sigma, \lambda)$ using the MSWD estimator. We set $c = 0.004428$, $\rho = 0.976$,

[6] Here, we minimize the log of $n \times SW$ since SW is too small around the true value.

Table 2.2: The sizes for a singular state-space model

$\rho = 0.3$

	5%					10%				
	ρ	a_1	a_2	b_1	b_2	rho	a_1	a_2	b_1	b_2
$T = 500$	0.0500	0.0495	0.0500	0.0500	0.0505	0.0985	0.0980	0.0975	0.0960	0.0945
$T = 1500$	0.0525	0.0495	0.0500	0.0495	0.0485	0.0965	0.1015	0.1010	0.0995	0.0995
$T = 2000$	0.0535	0.0535	0.0540	0.0505	0.0510	0.1020	0.1020	0.1015	0.0975	0.0975
$T = 2500$	0.0505	0.0495	0.0495	0.0520	0.0505	0.0925	0.1045	0.1045	0.1040	0.1040
$T = 3000$	0.0535	0.0465	0.0465	0.0530	0.0530	0.0985	0.1020	0.1015	0.0940	0.0925

$\rho = 0.8$

	5%					10%				
	ρ	a_1	a_2	b_1	b_2	ρ	a_1	a_2	b_1	b_2
$T = 500$	0.0550	0.0520	0.0515	0.0465	0.0490	0.1065	0.1075	0.1080	0.1035	0.1045
$T = 1500$	0.0555	0.0435	0.0435	0.0585	0.0575	0.0985	0.1015	0.1015	0.0970	0.0970
$T = 2000$	0.0500	0.0515	0.0510	0.0510	0.0505	0.1020	0.0955	0.0955	0.1040	0.1050
$T = 2500$	0.0505	0.0525	0.0525	0.0480	0.0490	0.1050	0.1000	0.0995	0.1025	0.1020
$T = 3000$	0.0570	0.0470	0.0470	0.0500	0.0495	0.0950	0.1000	0.1000	0.1015	0.1015

Table 2.3: P-values when $(c, \rho, \sigma, \lambda) = (0.07, 0.95, 0.02, -0.005)$ in the discrete-time Vasicek Model

(a) $\Pr(|t| > 1.96)$

	c	ρ	σ	λ
$T = 500$	0.0520	0.0580	0.0515	0.0605
$T = 1500$	0.0450	0.0575	0.0480	0.0555
$T = 2000$	0.0470	0.0590	0.0565	0.0495
$T = 2500$	0.0505	0.0590	0.0555	0.0510
$T = 3000$	0.0515	0.0520	0.0495	0.0510

(b) $\Pr(|t| > 1.645)$

	c	ρ	σ	λ
$T = 500$	0.0980	0.1120	0.0970	0.1095
$T = 1500$	0.0960	0.1055	0.0975	0.1015
$T = 2000$	0.0960	0.1155	0.1090	0.1050
$T = 2500$	0.0995	0.1080	0.1030	0.0990
$T = 3000$	0.0995	0.1005	0.1020	0.1065

$\sigma = 0.000556$, and $\lambda = -0.0824$. These values come from Backus et al. (1998), and match the empirical data in McCulloch and Kwon (1993). The weight function is set to $w(s) = I(s \in [0.1, 0.9])$. We compute 2000 estimates of t-values from the samples of size $T = 500, 1500, 2000, 2500$, and 3000. Table 2.4 shows that p-values are not far from the nominal sizes when the sample size is large.

Version 3:

We consider discrete-time Vasicek model in Example 2.1 (Cont'd). We set $Z_t = (Y_t^\top, Y_{t-1}^\top)^\top$ and estimate $\psi = (c, \rho, \sigma, \lambda)$ using the MSWD estimator. We set $c = 0.004428$, $\rho = 0.95$, $\sigma = 0.000556$, and $\lambda = -0.0824$. These values come from Backus et al. (1998) except for ρ. The weight function is set to $w(s) = I(s \in [0.1, 0.9])$. We compute 2000 estimates of t-values from the samples of size $T = 500, 1500, 2000, 2500$, and 3000. Table 2.5 shows that p-values are not far from the nominal sizes for all sample sizes.

Table 2.4: P-values when $(c, \rho, \sigma, \lambda) = (0.004428, 0.976, 0.000556, -0.0824)$ in the discrete-time Vasicek Model

(a) $\Pr(|t| > 1.96)$

	c	ρ	σ	λ
$T = 500$	0.0490	0.0925	0.0515	0.0540
$T = 1500$	0.0530	0.0655	0.0495	0.0570
$T = 2000$	0.0520	0.0615	0.0475	0.0575
$T = 2500$	0.0510	0.0615	0.0485	0.0460
$T = 3000$	0.0535	0.0610	0.0485	0.0510

(b) $\Pr(|t| > 1.645)$

	c	ρ	σ	λ
$T = 500$	0.0915	0.1265	0.1140	0.0970
$T = 1500$	0.1015	0.1140	0.1015	0.1070
$T = 2000$	0.0995	0.1115	0.1035	0.1055
$T = 2500$	0.1020	0.1075	0.1075	0.1035
$T = 3000$	0.0995	0.1075	0.0995	0.0975

2.9.3 Auction Model

We consider an econometric model of an independent private value procurement auction formulated in Paarsch (1992) and Donald and Paarsch (2002). In first-price procurement auction model, there is only one buyer. Sellers will provide their bids, and the lowest one will be the winning bid.

Let Y be the winning bid for the auction considered and X denote observable auction characteristics. Suppose the bidder's private value V follows a conditional distribution given X of the form

$$f_V(v|X, \theta, \gamma) I(v \geq g_V(X, \theta)).$$

Assuming a Bayes-Nash Equilibrium solution concept, the equilibrium biddning function satisfies

$$\sigma(v) = v + \frac{\int_v^\infty (1 - F_V(\xi|X, \theta, \gamma))^{m-1} d\xi}{(1 - F_V(v|X, \theta, \gamma))^{m-1}}, \tag{2.25}$$

where m is the number of bidders in the auction and $F_V(v|X, \theta, \gamma)$ denote the conditional cdf of V given X, θ, and γ.

Table 2.5: P-values when $(c, \rho, \sigma, \lambda) = (0.004428, 0.95, 0.000556, -0.0824)$ in the discrete-time Vasicek Model

(a) $\Pr(|t| > 1.96)$

	c	ρ	σ	λ
$T = 500$	0.0465	0.0630	0.0600	0.0490
$T = 1500$	0.0535	0.0510	0.0445	0.0515
$T = 2000$	0.0540	0.0500	0.0490	0.0530
$T = 2500$	0.0515	0.0565	0.0550	0.0490
$T = 3000$	0.0490	0.0570	0.0495	0.0490

(b) $\Pr(|t| > 1.645)$

	c	ρ	σ	λ
$T = 500$	0.0950	0.1070	0.1080	0.1000
$T = 1500$	0.1015	0.0980	0.0985	0.1090
$T = 2000$	0.0995	0.1010	0.0920	0.1010
$T = 2500$	0.0970	0.1015	0.1100	0.1010
$T = 3000$	0.1015	0.0975	0.1015	0.0990

Suppose V given X follows an exponential distribution with

$$f_V(v|X, \theta, \gamma) = \frac{1}{h(x, \theta)} \exp\left(-\frac{v}{h(x, \theta)}\right) \text{ and } g_V(X, \theta) = 0,$$

where $E(V|X, \theta) = h(X, \theta)$, the winning bid distribution given X is given by

$$f(y|X, \theta, \gamma) I(y \geq g(X, \theta))$$

with

$$f(y|X, \theta, \gamma) = \frac{m}{h(x, \theta)} \exp\left(-\frac{m}{h(x, \theta)}\left(y - \frac{h(x, \theta)}{m-1}\right)\right) \text{ and } g(X, \theta) = \frac{h(x, \theta)}{m-1}.$$

Hirano and Porter (2003) study MLE and BE for this model. Li (2010) proposes an indirect inference approach for the first-price sealed-bid auction where there are many buyers and one seller with $h(x, \theta) = \exp((1, x')\theta)$. The linear regression model with unit exponential error in Chernozhukov and Hong (2004) is the same as this model with $m = 1$ and $h(x, \theta) = x'\theta$.

The conditional cumulative distribution function $F(y|X,\theta)$ is given by

$$
\begin{aligned}
F(y|X,\theta) &= \int_{g(X,\theta)}^{y} f\left(u|X,\theta\right) \mathrm{d}u \\
&= \int_{g(X,\theta)}^{y} \frac{m}{h\left(x,\theta\right)} \exp\left(-\frac{m}{h\left(x,\theta\right)}\left(u-\frac{h\left(x,\theta\right)}{m-1}\right)\right) \mathrm{d}u \text{ and } g\left(X,\theta\right) = \frac{h\left(x,\theta\right)}{m-1} \\
&= \int_{0}^{y-g(X,\theta)} \frac{m}{h\left(x,\theta\right)} \exp\left(-\frac{m}{h\left(x,\theta\right)}u\right) \mathrm{d}u \\
&= 1-\exp\left(-\frac{m}{h(x,\theta)}\left(y-\frac{h(x,\theta)}{m-1}\right)\right) \text{ for } y \geq g(X,\theta).
\end{aligned}
$$

In this simulation, we set $h(x,\theta) = \exp(\theta_1 + \theta_2 x)$ folloing Li (2010), and estimate θ using the MSCD estimator.[7] We set $T = 100$, $\theta = (2.5, 0.5)$, and $m = 6$, and we conduct Monte-Carlo simulation 1000 times.

Figure 2.5 shows QQ plots of t-values of the MSCD estimator and Li (2010)'s estimator. QQ plots show that both estimators are comparable.

2.10 Concluding Remarks

Motivated by important features of structural models in economics and finance such as stochastc singularity, intractable likelihood functions, and parameter dependent supports, this paper has proposed a simple and robust method for estimation based on minimizing sliced distances between empirical and model-induced measures of the distribution. We have developed a unified asymptotic theory under high level assumptions and verified them for the motivating examples in this paper. Important issues remain to be addressed. They include applications to specific models such as DSGE models and aggregate demand models. The complexity of such models may require more sophisticated computational algorithms than used in the numerical section of the current paper. Theoretically this paper has focused on correctly specified models and future work should extend the results in this paper to possibly misspecified models and investigate rigorously robustness properties of the minimum sliced distance estimators.

[7] Here, we use unweighted version of MSCD estimator.

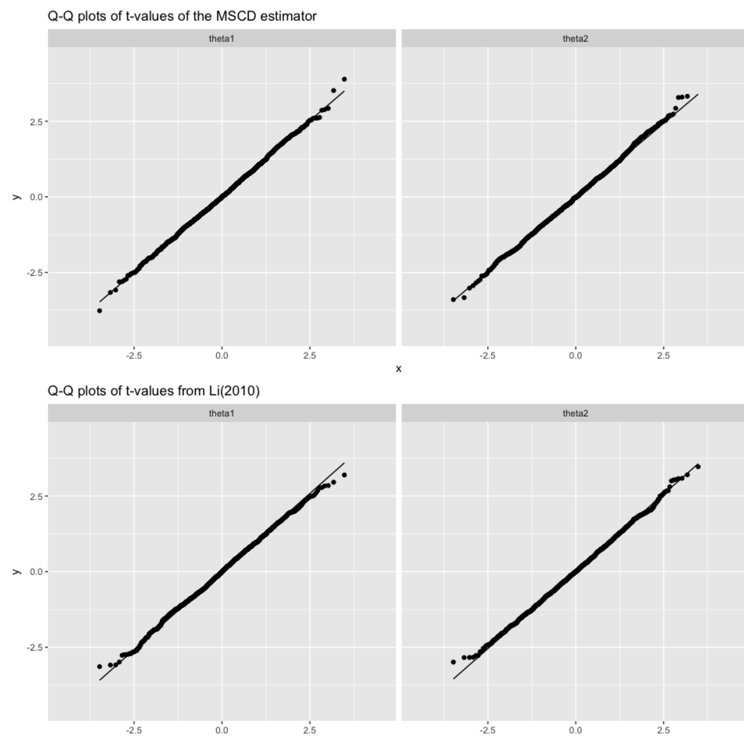

Figure 2.5: QQ plots of t-values of the MSCD estimator and Li (2010)'s esimator when $T = 100$

BIBLIOGRAPHY

AHIDAR-COUTRIX, A. AND P. BERTHET (2021): "Convergence of Multivariate Quantile Surfaces," .

AMBROSIO, L., N. GIGLI, AND G. SAVARE (2008): *Gradient Flows*, Birkhäuser-Verlag.

AN, S. AND F. SCHORFHEIDE (2007): "Bayesian Analysis of DSGE Models," *Econometric Reviews*, 26.

ANDERSEN, T. G., L. BENZONI, AND J. LUND (2002): "An Empirical Investigation of Continuous-Time Equity Return Models," *The Journal of Finance*, 57, 1239–1284.

ANDREWS, D. W. K. (1999): "Estimation When a Parameter is on a Boundary," *Econometrica*, 67, 1341–1383.

ARCONES, M. A. AND B. YU (1994): "Central limit theorems for empirical andU-processes of stationary mixing sequences," *Journal of Theoretical Probability*, 7.

ARJOVSKY, M. AND L. BOTTOU (2017): "Towards Principled Methods for Training Generative Adversarial Networks," .

ARJOVSKY, M., S. CHINTALA, AND L. BOTTOU (2017): "Wasserstein Generative Adversarial Networks," in *Proceedings of the 34th International Conference on Machine Learning*, ed. by D. Precup and Y. W. Teh, PMLR, vol. 70 of *Proceedings of Machine Learning Research*, 214–223.

BACKUS, D. K., S. FORESI, AND C. TELMER (1998): "Discrete-time models of bond pricing," *NBER Working Paper*.

Barrio, E. D., E. Giné, and F. Utzet (2005): "Asymptotics for L2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances," *Bernoulli*, 11.

Basu, S. and G. Michailidis (2015): "Regularized estimation in sparse high-dimensional time series models," *The Annals of Statistics*, 43, 1535–1567.

Beare, B. K. (2010): "Copulas and Temporal Dependence," *Econometrica*, 78, 395–410.

Bellemare, M. G., I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos (2017): "The Cramer Distance as a Solution to Biased Wasserstein Gradients," .

Bernton, E., P. E. Jacob, M. Gerber, and C. P. Robert (2019): "On parameter estimation with the Wasserstein distance," *Information and Inference: A Journal of the IMA*, 8, 657–676.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841.

Berry, S. T. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 25, 242.

Beutner, E. and L. Bordes (2011): "Estimators Based on Data-Driven Generalized Weighted Cramér-von Mises Distances under Censoring - with Applications to Mixture Models," *Scandinavian Journal of Statistics*, 38.

Bobkov, S. and M. Ledoux (2019): "One-dimensional empirical measures, order statistics, and Kantorovich transport distances," *Memoirs of the American Mathematical Society*, 261, 0.

Bonneel, N., J. Rabin, G. Peyré, and H. Pfister (2015): "Sliced and Radon Wasserstein Barycenters of Measures," *Journal of Mathematical Imaging and Vision*, 51.

Bowlus, A. J., N. M. Kiefer, and G. R. Neumann (2001): "Equilibrium search models and the transition from school to work," *International Economic Review*, 42.

Bradley, R. C. (1983): "Approximation theorems for strongly mixing random variables." *The Michigan Mathematical Journal*, 30, 69–81.

——— (2005): "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," *Probability Surveys*, 2, 107–144.

Briol, F.-X., A. Barp, A. B. Duncan, and M. Girolami (2019): "Statistical Inference for Generative Models with Maximum Mean Discrepancy," .

Brown, D. J. and M. H. Wegkamp (2002): "Weighted minimum mean-square distance from independence estimation," .

Chen, Q. and Z. Fang (2019): "Inference on functionals under first order degeneracy," *Journal of Econometrics*, 210, 459–481.

Chen, R.-R. and L. Scott (1993): "Maximum Likelihood Estimation for a Multifactor Equilibrium Model of the Term Structure of Interest Rates," *The Journal of Fixed Income*, 3, 14–31.

Chen, X. and Y. Fan (2006a): "Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification," *Journal of Econometrics*, 135, 125–154.

——— (2006b): "Estimation of copula-based semiparametric time series models," *Journal of Econometrics*, 130, 307–335.

Chen, X., Z. Huang, and Y. Yi (2018): "On Estimation of Multivariate Semiparametric GARCH Filtered Copula Models," China Center for Economic Research Working Paper Series E2018025. http://img.bimba.pku.edu.cn/resources/file/13/2018/12/31/201812311606304.pdf.

Chen, X., R. Koenker, and Z. Xiao (2009a): "Copula-based nonlinear quantile autoregression," *Econometrics Journal*, 12, S50–S67.

Chen, X., W. B. Wu, and Y. Yi (2009b): "Efficient estimation of copula-based semiparametric Markov models," *The Annals of Statistics*, 37, 4214–4253.

Chernov, M. (2003): "[Iterative and Recursive Estimation in Structural Nonadaptive Models]: Comment," *Journal of Business and Economic Statistics*, 21, 485–488.

Chernov, M., A. Ronald Gallant, E. Ghysels, and G. Tauchen (2003): "Alternative models for stock price dynamics," *Journal of Econometrics*, 116, 225–257.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "INFERENCE ON COUNTERFACTUAL DISTRIBUTIONS," *Econometrica*, 81, 2205–2268.

Chernozhukov, V. and H. Hong (2004): "Likelihood Estimation and Inference in a Class of Nonregular Econometric Models," *Econometrica*, 72, 1445–1480.

Csörgő, M. (1983): *Quantile Processes with Statistical Applications*, Society for Industrial and Applied Mathematics.

Csorgo, M. and P. Revesz (1978): "Strong Approximations of the Quantile Process," *The Annals of Statistics*, 6.

Csörgö, M. and H. Yu (1996): "Weak approximations for quantile processes of stationary sequences," *Canadian Journal of Statistics*, 24.

Dehling, H., D. Vogel, M. Wendler, and D. Wied (2017): "TESTING for CHANGES in KENDALL'S TAU," *Econometric Theory*, 33, 1352–1386.

Dehling, H. and M. Wendler (2010): "Central limit theorem and the bootstrap for U-statistics of strongly mixing data," *Journal of Multivariate Analysis*, 101, 126–137.

DONALD, S. G. AND H. J. PAARSCH (2002): "Superconsistent estimation and inference in structural econometric models using extreme order statistics," *Journal of Econometrics*, 109, 305–340.

DOUKHAN, P. (1994): *Mixing: Properties and Examples*, no. 85 in Lecture Notes in Statistics, Springer-Verlag New York.

DOUKHAN, P., P. MASSART, AND E. RIO (1995): "Invariance principles for absolutely regular empirical processes," *Annales de l'I.H.P. Probabilités et statistiques*, 31.

EKSTRÖM, M. (2014): "A general central limit theorem for strong mixing sequences," *Statistics and Probability Letters*, 94, 236–238.

ERAKER, B., M. JOHANNES, AND N. POLSON (2003): "The Impact of Jumps in Volatility and Returns," *The Journal of Finance*, 58, 1269–1300.

FAN, J., F. HAN, H. LIU, AND B. VICKERS (2016): "Robust inference of risks of large portfolios," *Journal of Econometrics*, 194, 298–308.

FAN, Y. AND A. J. PATTON (2014): "Copulas in Econometrics," *Annual Review of Economics*, 6, 179–200.

FILZMOSER, P., H. FRITZ, AND K. KALCHER (2021): *pcaPP: Robust PCA by Projection Pursuit*, r package version 1.9-74.

GANDHI, A. AND J.-F. HOUDE (2019): "Measuring Substitution Patterns in Differentiated-Products Industries," Tech. rep., Cambridge, MA.

GÓMEZ, V. (2016): *Multivariate Time Series With Linear State Space Structure*, Springer International Publishing.

GONÇALVES, S. AND H. WHITE (2002): "the Bootstrap of the Mean for Dependent Heterogeneous Arrays," *Econometric Theory*, 18, 1367–1384.

GOODFELLOW, I., J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO (2014): "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Curran Associates, Inc., vol. 27.

HAN, F. AND Y. LI (2019): "Moment Bounds for Large Autocovariance Matrices Under Dependence," *Journal of Theoretical Probability*, 1–48.

HAN, F., H. LU, AND H. LIU (2015): "A direct estimation of high dimensional stationary vector autoregressions," *Journal of Machine Learning Research*, 16, 3115–3150.

HAN, F. AND W. B. WU (2019): "Probability inequalities for high dimensional time series under a triangular array framework," *arXiv preprint arXiv:1907.06577 [math.ST]*, 1–21.

HECQ, A., L. MARGARITELLA, AND S. SMEEKES (2019): "Granger Causality Testing in High-Dimensional VARs: a Post-Double-Selection Procedure," *arXiv preprint arXiv:1902.10991*.

HETTMANSPERGER, T., I. HUETER, AND J. HÜSLER (1994): "Minimum distance estimators," *Journal of Statistical Planning and Inference*, 41.

HIRANO, K. AND J. R. PORTER (2003): "Asymptotic Efficiency in Parametric Structural Models with Parameter-Dependent Support," *Econometrica*, 71, 1307–1338.

HORN, R. A. AND C. R. JOHNSON (2012): *Matrix analysis*, Cambridge University Press, second ed.

HSING, T. AND W. B. WU (2004): "On weighted U-statistics for stationary processes," *The Annals of Probability*, 32, 1600–1631.

JEGADEESH, N. AND G. G. PENNACCHI (1996): "The Behavior of Interest Rates Implied by the Term Structure of Eurodollar Futures," *Journal of Money, Credit and Banking*, 28, 426.

Jiang, R., P. Manchanda, and P. E. Rossi (2009): "Bayesian analysis of random coefficient logit models using aggregate data," *Journal of Econometrics*, 149, 136–148.

Kaji, T. (2019): "Asymptotic Theory of L-Statistics and Integrable Empirical Processes," .

Kaji, T., E. Manresa, and G. Pouliot (2020): "An Adversarial Approach to Structural Estimation," *SSRN Electronic Journal*.

Kock, A. B. and L. Callot (2015): "Oracle inequalities for high dimensional vector autoregressions," *Journal of Econometrics*, 186, 325–344.

Kolmogorov, A. N. and Y. A. Rozanov (1960): "On Strong Mixing Conditions for Stationary Gaussian Processes," *Theory of Probability & Its Applications*, 5, 204–208.

Kolouri, S., P. E. Pope, C. E. Martin, and G. K. Rohde (2018): "Sliced-Wasserstein Autoencoder: An Embarrassingly Simple Generative Model," .

Komunjer, I. and S. Ng (2011): "Dynamic Identification of Dynamic Stochastic General Equilibrium Models," *Econometrica*, 79.

Lahiri, S. N. (2003): *Resampling methods for dependent data*, Springer-Verlag New York.

LaRiccia, V. N. (1982): "Asymptotic Properties of Weighted L^2 Quantile Distance Estimators," *The Annals of Statistics*, 10.

——— (1984): "Optimal weights for general L2 distance estimators," *Statistics & Probability Letters*, 2, 169–173.

Lee, T. H. and X. Long (2009): "Copula-based multivariate GARCH model with uncorrelated dependent errors," *Journal of Econometrics*, 150, 207–218.

Lei, N., K. Su, L. Cui, S.-T. Yau, and X. D. Gu (2019): "A geometric view of optimal transportation and generative model," *Computer Aided Geometric Design*, 68.

Li, T. (2010): "Indirect inference in structural econometric models," *Journal of Econometrics*, 157, 120–128.

Liu, H., F. Han, M. Yuan, J. Lafferty, and L. Wasserman (2012): "High-dimensional semiparametric Gaussian copula graphical models," *The Annals of Statistics*, 40, 2293–2326.

Liu, H., J. Lafferty, and L. Wasserman (2009): "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs," *Journal of Machine Learning Research*, 10, 2295–2328.

Lütkepohl, H. (2005): *New introduction to multiple time series analysis*, Springer-Verlag Berlin Heidelberg.

McCulloch, J. H. and H.-C. Kwon (1993): "U.S. Term Structure Data, 1947-1991," .

Mehra, K. L. and M. S. Rao (1975a): "On Functions of Order Statistics for Mixing Processes," *The Annals of Statistics*, 3.

——— (1975b): "Weak Convergence of Generalized Empirical Processes Relative to d_q Under Strong Mixing," *The Annals of Probability*, 3.

Merlevède, F., M. Peligrad, and E. Rio (2009): "Bernstein inequality and moderate deviations under strong mixing conditions," in *High Dimensional Probability V: The Luminy Volume*, ed. by C. Houdre, V. Koltchinskii, D. M. Mason, and M. Peligrad, Beachwood, Ohio, USA: Institute of Mathematical Statistics, vol. 5 of *Institute of Mathematical Statistics (IMS) Collections*, 273–292.

Min, A. and C. Czado (2014): "SCOMDY models based on pair-copula constructions with application to exchange rates," *Computational Statistics and Data Analysis*, 76, 523–535.

Moon, H. R., M. Shum, and M. Weidner (2018): "Estimation of random coefficients logit demand models with interactive fixed effects," *Journal of Econometrics*, 206, 613–644.

Nadjahi, K., A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Şimşekli (2020a): "Statistical and Topological Properties of Sliced Probability Divergences," .

Nadjahi, K., A. Durmus, U. Şimşekli, and R. Badeau (2020b): "Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance," .

Negahban, S. and M. J. Wainwright (2011): "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, 39, 1069–1097.

Newey, W. K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59.

Neykov, M., Y. Ning, J. S. Liu, and H. Liu (2018a): "A Unified Theory of Confidence Regions and Testing for High-Dimensional Estimating Equations," *Statistical Science*, 33, 427–443.

——— (2018b): "Supplement to "A Unified theory of Confidence Regions and Testing for High-Dimensional Estimating Equations"," *Statistical Science*, 33.

Öztürk, Ö. and T. P. Hettmansperger (1997): "Generalised weighted Cramér-von Mises distance estimators," *Biometrika*, 84.

Paarsch, H. J. (1992): "Deciding between the common and private value paradigms in empirical models of auctions," *Journal of Econometrics*, 51, 191–215.

Pan, J. (2002): "The jump-risk premia implicit in options: evidence from an integrated time-series study," *Journal of Financial Economics*, 63, 3–50.

Pastorello, S., V. Patilea, and E. Renault (2003): "Iterative and Recursive Estimation in Structural Nonadaptive Models," *Journal of Business & Economic Statistics*, 21, 449–509.

Patton, A. (2013): *Chapter 16 - Copula Methods for Forecasting Multivariate Time Series*, vol. 2 of *Handbook of Economic Forecasting*, Elsevier.

Patton, A. J. (2012): "A review of copula models for economic time series," *Journal of Multivariate Analysis*, 110, 4–18.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, New York, NY: Springer New York.

Pollard, D. (1980): "The minimum distance method of testing," *Metrika*, 27, 43–70.

Pötscher, B. M. and I. R. Prucha (1997): *Dynamic Nonlinear Econometric Models: Asymptotic Theory*, Springer-Verlag Berlin Heidelberg.

Puri, M. L. and L. T. Tran (1980): "Empirical distribution functions and functions of order statistics for mixing random variables," *Journal of Multivariate Analysis*, 10, 405–425.

Qin, S. and Y. Wu (2020): "General matching quantiles M-estimation," *Computational Statistics & Data Analysis*, 147, 106941.

Qiu, H., S. Xu, F. Han, H. Liu, and B. Caffo (2015): "Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes," *32nd International Conference on Machine Learning, ICML 2015*, 3, 1843–1851.

Rémillard, B., N. Papageorgiou, and F. Soustra (2012): "Copula-based semiparametric models for multivariate time series," *Journal of Multivariate Analysis*, 110, 30–42.

Rio, E. (2017): *Asymptotic Theory of Weakly Dependent Random Processes*, vol. 80, Berlin, Heidelberg: Springer Berlin Heidelberg.

Santambrogio, F. (2015): *Optimal Transport for Applied Mathematicians*, vol. 87 of *Progress in Nonlinear Differential Equations and Their Applications*, Cham: Springer International Publishing.

Sen, B. (2021): "A Gentle Introduction to Empirical Process Theory and Applications," .

Sgouropoulos, N., Q. Yao, and C. Yastremiz (2015): "Matching a Distribution by Matching Quantiles Estimation," *Journal of the American Statistical Association*, 110, 742–759.

Shapiro, A. (2000): "Statistical Inference of Stochastic Optimization Problems," in *Nonconvex Optimization and Its Applications*, Springer US, 282–307.

Shen, Y., F. Han, and D. Witten (2019): "Tail behavior of dependent V-statistics and its applications," *arXiv preprint arXiv:1902.02761 [math.ST]*.

Sherman, R. P. (1994): "Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators," *The Annals of Statistics*, 22.

Shorack, G. R. and J. A. Wellner (2009): *Empirical Processes with Applications to Statistics*, Society for Industrial and Applied Mathematics.

Torgovitsky, A. (2017): "Minimum distance from independence estimation of nonseparable instrumental variables models," *Journal of Econometrics*, 199.

van de Geer, S. A. (2009): *Empirical Processes in M-Estimation*, vol. 6, Cambridge University Press.

van de Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer Series in Statistics, New York, NY: Springer New York.

Wang, T. and Y. Xia (2015): "Whittle Likelihood Estimation of Nonlinear Autoregressive Models With Moving Average Residuals," *Journal of the American Statistical Association*, 110, 1083–1099.

Xue, L. and H. Zou (2012): "Regularized rank-based estimation of high-dimensional nonparanormal graphical models," *The Annals of Statistics*, 40, 2541–2571.

Yoshihara, K.-i. (1976): "Limiting behavior of U-statistics for stationary, absolutely regular processes," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35, 237–252.

Zhu, L. X., K. T. Fang, and M. I. Bhatti (1997): "On estimated projection pursuit-type Crámer-Von Mises statistics," *Journal of Multivariate Analysis*, 63.

Appendix A

ESTIMATION AND INFERENCE IN HIGH-DIMENSIONAL SEMIPARAMETRIC GAUSSIAN COPULA VECTOR AUTOREGRESSION

A.1 Proofs

In the sequel, for any random variable $X \in \mathbb{R}$ and any $p \in (0, \infty]$, we will use $\|X\|_p$ to denote its L_p norm. Let $\mathbb{N}$ denote the set of all positive integers.

A.1.1 Proofs in Section 1.2

Notation

Some additional notations are introduced below. For a random variable $\boldsymbol{X}$, we denote $\widetilde{\boldsymbol{X}}$ as an independent copy of $\boldsymbol{X}$. Then, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$, we define the following Hoeffding decomposition components,

- $f(\boldsymbol{x}, \boldsymbol{y}) = \text{sign}(x_1 - y_1)\text{sign}(x_2 - y_2)$, $g(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y}) - \mathbb{E}[f(\boldsymbol{X}, \widetilde{\boldsymbol{X}})]$;
- $g_1(\boldsymbol{x}) = \mathbb{E}[g(\boldsymbol{x}, \boldsymbol{X})]$, $g_2(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y}) - \mathbb{E}[f(\boldsymbol{x}, \boldsymbol{X})] - \mathbb{E}[f(\boldsymbol{X}, \boldsymbol{y})] + \mathbb{E}[f(\boldsymbol{X}, \widetilde{\boldsymbol{X}})]$.

For a random process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ following VAR(1) model (1.1), we define $\boldsymbol{U}_i = (\boldsymbol{Z}_i^\top, \boldsymbol{Z}_{i+1}^\top)^\top \in \mathbb{R}^{2d}$. Let $U_{i,k}$ denote the k-th element of the random vector $\boldsymbol{U}_i$. Based on the observation $\{\boldsymbol{U}_t\}_{t=1}^{n-1}$, we define the following V-statistics,

$$\widehat{\mathbf{V}} = \frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1}\text{sign}(\boldsymbol{U}_i - \boldsymbol{U}_j)\text{sign}(\boldsymbol{U}_i - \boldsymbol{U}_j)^\top; \mathbf{V} = \int\int \text{sign}(\boldsymbol{U}_i - \boldsymbol{U}_j)\text{sign}(\boldsymbol{U}_i - \boldsymbol{U}_j)^\top dF_{\boldsymbol{U}_i} dF_{\boldsymbol{U}_j},$$

where $F_{\boldsymbol{U}_i}$ is the cumulative distribution function of $\boldsymbol{U}_i$. Then, we can represent the (k,m)-th element $\widehat{T}_{km}$ of $\widehat{\mathbf{T}}$ and the (k,m)-th element $\widehat{V}_{km}$ of $\widehat{\mathbf{V}}$ as

$$\widehat{T}_{km} = \frac{2}{(n-1)(n-2)}\sum_{i<j}\text{sign}(U_{i,k}-U_{j,k})\text{sign}(U_{i,m}-U_{j,m}) = \frac{2}{(n-1)(n-2)}\sum_{i<j} f\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix}, \begin{pmatrix}U_{j,k}\\ U_{j,m}\end{pmatrix}\right),$$

$$\widehat{V}_{km} = \frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1}\text{sign}(U_{i,k}-U_{j,k})\text{sign}(U_{i,m}-U_{j,m}) = \frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1} f\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix}, \begin{pmatrix}U_{j,k}\\ U_{j,m}\end{pmatrix}\right).$$

Proof of Theorem 1.2.1

The proof technique is based on the proofs of Theorem 4 and Proposition 5 in Shen et al. (2019). Because $\mathbf{T} = \mathbf{V}$, we have the following relationship when $n \geq 3$,

$$\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\max} \leq \frac{n-1}{n-2}\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\max} + \frac{1}{n-2}\|\mathbf{V}\|_{\max} \leq 2\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\max} + \frac{1}{n-2}.$$

This implies that, when $n \geq 3$, $\mathbb{P}\left(\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\max} \geq 2z + \frac{1}{n-2}\right) \leq \mathbb{P}\left(\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\max} \geq z\right)$.

Similar to the proofs of Theorem 4 and Proposition 5 in Shen et al. (2019), we can conduct Hoeffding decomposition for the (k,m)-element $\widehat{V}_{km} - V_{km}$ of $\widehat{\mathbf{V}} - \mathbf{V}$ as follows,

$$\widehat{V}_{km} - V_{km} = \frac{2}{n-1}\sum_{i=1}^{n-1} g_1\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix}\right) + \frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1} g_2\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix}, \begin{pmatrix}U_{j,k}\\ U_{j,m}\end{pmatrix}\right).$$

Then, we have

$$\begin{aligned}
&\mathbb{P}(|\widehat{V}_{km} - V_{km}| \geq z)\\
&\leq \mathbb{P}\left(\left|\frac{1}{n-1}\sum_{i=1}^{n-1} g_1\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix}\right)\right| \geq \frac{z}{4}\right) + \mathbb{P}\left(\left|\frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1} g_2\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix}, \begin{pmatrix}U_{j,k}\\ U_{j,m}\end{pmatrix}\right)\right| \geq \frac{z}{2}\right).
\end{aligned}$$

It implies that

$$\begin{aligned}
&\mathbb{P}\left(\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\max} \geq z\right) \leq \sum_{k,m=1}^{2d}\mathbb{P}\left(|\widehat{V}_{km} - V_{km}| \geq z\right)\\
&\leq \sum_{k,m=1}^{2d}\mathbb{P}\left(\left|\frac{1}{n-1}\sum_{i=1}^{n-1} g_1\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix}\right)\right| \geq \frac{z}{4}\right) + \sum_{k,m=1}^{2d}\mathbb{P}\left(\left|\frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1} g_2\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix}, \begin{pmatrix}U_{j,k}\\ U_{j,m}\end{pmatrix}\right)\right| \geq \frac{z}{2}\right).
\end{aligned} \tag{A.1}$$

The proof consists of the following five steps. Step 1 will analyze the first part of the right-hand side of inequality (A.1) and Steps 2 to 5 will analyze the second part of the right-hand side of inequality (A.1). The exact form of z in inequality (A.1) will be provided later after analyzing each term.

Step 1. Because $\{\boldsymbol{U}_i = (\boldsymbol{Z}_i^\top, \boldsymbol{Z}_{i+1}^\top)^\top\}_{t\in\mathbb{Z}}$ is geometrically α-mixing, and $g_1(\boldsymbol{U}_{i,\{k,m\}})$ has mean zero with $\sup_i \|g_1(B_{i,\{k,m\}})\|_\infty \leq 2$, we have the following relationship using Lemma A.1.2 in the supplementary material, that is, for $c_1 > 0$ and all sufficiently large n,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n-1}\sum_{i=1}^{n-1} g_1\left(\begin{pmatrix} U_{i,k} \\ U_{i,m}\end{pmatrix}\right)\right| \geq c_1\sqrt{\frac{\log(ed)}{n-1}}\right) \\
&\leq 2\exp\left(-\frac{C_3(n-1)c_1^2\log(ed)}{4(n-1)+2c_1(n-1)\sqrt{\frac{\log(ed)}{n-1}}(\log(n-1))(\log(\log(n-1)))}\right) \\
&= 2\exp\left(-\frac{C_3 c_1^2\log(ed)}{4+2c_1\sqrt{\frac{\log(ed)(\log^2(n-1))(\log^2(\log(n-1)))}{n-1}}}\right) \leq 2\exp\left(-\widetilde{K}_0\frac{c_1^2}{1+c_1}\log(ed)\right),
\end{aligned}$$

where the constant C_3 only depends on (κ_1, κ_2) and $\widetilde{K}_0$ only depends on $(K_0, \kappa_1, \kappa_2)$. Here, κ_1 and κ_2 are the α-mixing coefficients in Proposition 1.2.2 and K_0 is the absolute constant defined in Assumption E. The last inequality holds because of Assumption E. Therefore, for $c_1 \geq 1$ and all sufficiently large n,

$$\sum_{k,m=1}^{2d}\mathbb{P}\left(\left|\frac{1}{n-1}\sum_{i=1}^{n-1} g_1\left(\begin{pmatrix} U_{i,k} \\ U_{i,m}\end{pmatrix}\right)\right| \geq c_1\sqrt{\frac{\log(ed)}{n-1}}\right) \leq 8d^2\exp\left(-\widetilde{K}_0\frac{c_1^2}{1+c_1}\log(ed)\right).$$

Step 2. Because $g_2(\boldsymbol{U}_{i,\{k,m\}}, \boldsymbol{U}_{j,\{k,m\}})$ is degenerate and centered, using the proof in Proposition 1 and Corollary 7(a) in Shen et al. (2019), we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1} g_2\left(\begin{pmatrix} U_{i,k} \\ U_{i,m}\end{pmatrix}, \begin{pmatrix} U_{j,k} \\ U_{j,m}\end{pmatrix}\right)\right| \geq x + C_4 t'\right) \\
&\leq 2\exp\left(-\frac{C_5(n-1)y}{A_2^{1/2}+y^{1/2}M_2^{1/2}}\right) + (n-1)^2\left(\sum_{\ell=1}^{2} J_\ell\right) D\widetilde{M}_2 + (n-1)\left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right],
\end{aligned}$$

where C_4 is an absolute positive constant; and C_5 only depends on (κ_1, κ_2); and all other parameters are defined in Shen et al. (2019).

According to the discussion on page 5, Table 1, and the proof of Proposition 5 in Shen et al. (2019), we have

$$D \coloneqq \text{ the uniform upper bound of density functions of } U_{i,k} - U_{j,k},$$
$$A_2^{1/2} = F(t)\left\{\sigma^2 + \frac{(\log(n-1))^4}{n-1}\right\}, \quad M_2^{1/2} = F(t)^{1/2}(\log(n-1))^2, \quad J_\ell = 3,$$
$$\sigma^2 = \frac{64\kappa_1^{\delta/(2+\delta)}}{1-\exp(-\kappa_2\delta/(2+\delta))}, \quad y = x - \frac{F(t)}{n-1}, \quad t' = C_6\left(t + \sum_{i=0}^{1} v_i + F(t)\sum_{i=0}^{1} v_i\right),$$

where δ and C_6 are absolute positive constants, $F(t) \asymp \log^2(\widetilde{M}_1/\widetilde{M}_2) + \log^2\log(1/t)$ for a sufficiently large value of $\widetilde{M}_1$ and sufficiently small values of $\widetilde{M}_2$ and t, and v_0 and v_1 satisfy

$$v_0^2 \leq 2\left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right] + 12\widetilde{M}_2 D,$$
$$v_1^2 \leq \left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right] + 12\widetilde{M}_2 D.$$

Note that $F(t) \geq 0$ by construction. Then, we have

$$\sum_{k,m=1}^{2d} \mathbb{P}\left(\left|\frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1} g_2\left(\begin{pmatrix}U_{i,k}\\U_{i,m}\end{pmatrix}, \begin{pmatrix}U_{j,k}\\U_{j,m}\end{pmatrix}\right)\right| \geq y + \frac{F(t)}{n-1} + C_4 t'\right)$$
$$\leq 8d^2\exp\left(-\frac{C_5(n-1)y}{A_2^{1/2} + y^{1/2}M_2^{1/2}}\right) + 24(n-1)^2 d^2 D\widetilde{M}_2 + 4(n-1)d^2\left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right]. \tag{A.2}$$

Let $y = c_2\sqrt{\frac{\log(ed)}{n-1}}$ for some positive absolute constant c_2, and $t = \frac{1}{2}\sqrt{\frac{\log(ed)}{n-1}}$. Note that $t = o(1)$ under Assumption E. Then, the remaining steps will be as follows. Step 3 will bound the second and third parts of the right-hand side of inequality (A.2). Step 4 will bound the first part of the right-hand side of inequality (A.2). Step 5 will upper bound t', for all sufficiently large n and all sufficiently small t.

Step 3. Step 3 will bound

$$24(n-1)^2 d^2 D\widetilde{M}_2 + 4(n-1)d^2\left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right].$$

Using the Markov inequality and the fact that $U_{i,k} - U_{j,k}$ follows normal distribution, we have

$$\begin{aligned}
&24(n-1)^2 d^2 D\widetilde{M}_2 + 4(n-1)d^2\left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right] \\
\leq &24(n-1)^2 d^2 D\widetilde{M}_2 + 4(n-1)d^2\left[\frac{\mathbb{E}[|U_{1,k}|]}{\widetilde{M}_1} + \frac{\mathbb{E}[|U_{1,m}|]}{\widetilde{M}_1}\right] \\
\leq &24(n-1)^2 d^2 D\widetilde{M}_2 + 4(n-1)d^2\left[\frac{\mathbb{E}[|U_{1,k}|^2]^{1/2}}{\widetilde{M}_1} + \frac{\mathbb{E}[|U_{1,m}|^2]^{1/2}}{\widetilde{M}_1}\right] \\
\leq &24((n-1)d)^2\frac{1}{\sqrt{2\pi\underline{\sigma}^2}}\widetilde{M}_2 + 8((n-1)d)^2\frac{\sqrt{\bar{\sigma}^2}}{\widetilde{M}_1},
\end{aligned}$$

where $\underline{\sigma}^2$ is the uniform lower bound of the variance of $U_{i,k} - U_{j,k}$, and $\bar{\sigma}^2$ is the uniform upper bound of the variance of $U_{i,k}$. Because Lemma A.1.3 in the supplementary material implies $0 < c_{\mathbf{\Omega}} \leq \lambda_{\min}(\mathbf{\Omega}) \leq \lambda_{\max}(\mathbf{\Omega}) \leq C_{\mathbf{\Omega}} < \infty$, where $c_{\mathbf{\Omega}}$ and $C_{\mathbf{\Omega}}$ only depend on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$, $\underline{\sigma}^2$ and $\bar{\sigma}^2$ are lower and upper bounded by some positive constants which only depend on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$. Therefore,

$$24(n-1)^2 d^2 D\widetilde{M}_2 + 4(n-1)d^2\left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right] \leq \frac{1}{2}K_1((n-1)d)^2\left[\widetilde{M}_2 + \frac{1}{\widetilde{M}_1}\right],$$

where K_1 only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$. Then, by choosing $\widetilde{M}_2 = \frac{1}{\widetilde{M}_1} = ((n-1)d)^{-4}$, we have

$$24(n-1)^2 d^2 D\widetilde{M}_2 + 4(n-1)d^2\left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right] \leq K_1((n-1)d)^{-2}.$$

Step 4. Step 4 will bound the first part of the right-hand side of inequality (A.2) when $y = c_2\sqrt{\frac{\log(ed)}{n-1}}$ for some positive constant c_2, $t = \frac{1}{2}\sqrt{\frac{\log(ed)}{n-1}}$, $\widetilde{M}_2 = \frac{1}{\widetilde{M}_1} = ((n-1)d)^{-4}$, and n is sufficiently large so that t is sufficiently small under Assumption E.

Step 4-1. We will show that $\frac{F(t)}{n-1} = o(t)$ when $t = m_1\sqrt{\frac{\log(ed)}{n-1}}$ and $\widetilde{M}_2 = \frac{1}{\widetilde{M}_1} = ((n-1)d)^{-m_2}$ for some positive numbers m_1 and m_2. Note that we set $m_1 = \frac{1}{2}$ and $m_2 = 4$ in Step

3. Under Assumption E, for all sufficiently large n and sufficiently small t, we have

$$
\begin{aligned}
\frac{F(t)}{t(n-1)} &\asymp \frac{4m_2^2\log^2((n-1)d)) + \log^2\left(\log\left(m_1\sqrt{\frac{n-1}{\log(ed)}}\right)\right)}{m_1\sqrt{(n-1)\log(ed)}} \\
&= \frac{4m_2^2\log^2((n-1)d)}{m_1\sqrt{(n-1)\log(ed)}} + \frac{\log^2\left(\log\left(m_1\sqrt{\frac{n-1}{\log(ed)}}\right)\right)}{m_1\sqrt{(n-1)\log(ed)}} \\
&\le \frac{8m_2^2\log^2(n-1)}{m_1\sqrt{(n-1)\log(ed)}} + \frac{8m_2^2\log^2(d)}{m_1\sqrt{(n-1)\log(ed)}} + \frac{\log^2\left(\log\left(m_1\sqrt{\frac{n-1}{\log(ed)}}\right)\right)}{m_1\sqrt{\frac{n-1}{\log(ed)}}}\frac{1}{\log(ed)} \\
&\le \frac{8m_2^2\log^2(n-1)}{m_1\sqrt{(n-1)\log(ed)}} + \left[\frac{8m_2^2}{m_1}\sqrt{\frac{\log^5(ed)}{(n-1)}} + \frac{\log^2\left(\log\left(m_1\sqrt{\frac{n-1}{\log(ed)}}\right)\right)}{m_1\sqrt{\frac{n-1}{\log(ed)}}}\right]\frac{1}{\log(ed)} \\
&\le 2C_7\frac{\log^2(n-1)}{\sqrt{(n-1)\log(ed)}} + K_2\frac{1}{\log(ed)} = o(1),
\end{aligned}
$$

where C_7 is an absolute positive constant, and K_2 is a positive constant that only depend on K_0. Therefore, we have $\frac{F(t)}{n-1} = o(t)$.

Step 4-2. In this step, we will bound the first part of the right-hand side of inequality (A.2) with $y = c_2\sqrt{\frac{\log(ed)}{n-1}}$ for some positive constant c_2; and $t = \frac{1}{2}\sqrt{\frac{\log(ed)}{n-1}}$ when n is sufficiently large and $\frac{\log ed}{n-1}$ is sufficiently small, which is true under Assumption E. For the first part, we have

$$
\exp\left(-\frac{C_5(n-1)y}{A_2^{1/2} + y^{1/2}M_2^{1/2}}\right) = \exp\left(-\frac{C_5c_2\sqrt{n-1}\sqrt{\log(ed)}}{A_2^{1/2} + c_2^{1/2}\left(\frac{\log(ed)}{n-1}\right)^{1/4}M_2^{1/2}}\right) = \exp\left(-\frac{C_5c_2\log(ed)}{\left(\frac{\log(ed)}{n-1}\right)^{1/2}A_2^{1/2} + c_2^{1/2}\left(\frac{\log(ed)}{n-1}\right)^{3/4}M_2^{1/2}}\right)
$$

Then, Steps 4-2-1 and 4-2-2 combined will show $\left(\frac{\log(ed)}{n-1}\right)^{1/2}A_2^{1/2} = O(1)$ and $\left(\frac{\log(ed)}{n-1}\right)^{3/4}M_2^{1/2} = O(1)$.

Step 4-2-1. This step shows that $\left(\frac{\log(ed)}{n-1}\right)^{1/2}A_2^{1/2} = O(1)$.

Under Assumption E, for all sufficiently large n and sufficiently small $t = \frac{1}{2}\sqrt{\frac{\log(ed)}{n-1}}$, we have

$$
\begin{aligned}
(\log(ed)/n)^{1/2}A_2^{1/2} \asymp tF(t) &\asymp t\log^2(\widetilde{M}_1/\widetilde{M}_2) + t\log^2(\log(1/t)) \\
&= 64t\log^2((n-1)d) + t\log^2(\log(1/t)) \\
&= 32\frac{\sqrt{\log(ed)}\log^2((n-1)d)}{\sqrt{n-1}} + \frac{\log^2(\log(1/t))}{1/t} \\
&\leq 64\frac{\sqrt{\log(ed)}\log^2((n-1))}{\sqrt{n-1}} + 64\frac{\sqrt{\log(ed)}\log^2(d)}{\sqrt{n-1}} + \frac{\log^2(\log(1/t))}{1/t} \\
&\leq K_{3A},
\end{aligned}
$$

where K_{3A} only depends on K_0. Therefore, $\left(\frac{\log(ed)}{n-1}\right)^{1/2} A_2^{1/2} = O(1)$.

Step 4-2-2. This step shows that $\left(\frac{\log(ed)}{n-1}\right)^{3/4} M_2^{1/2} = O(1)$. Under Assumption E, for all sufficiently large n and sufficiently small $t \asymp \sqrt{\frac{\log(ed)}{n-1}}$, we have

$$
\left(\left(\frac{\log(ed)}{n-1}\right)^{3/4} M_2^{1/2}\right)^2 \asymp t^3F(t)\log^4(n) = tF(t)(t^2\log^4(n)) \leq K_{3A}\overline{K}_0,
$$

where $\overline{K}_0$ only depends on K_0. Therefore, for all sufficiently large n and sufficiently small t, we have $\left(\frac{\log(ed)}{n-1}\right)^{3/4} M_2^{1/2} \leq K_{3M}$, where $K_{3M} = \sqrt{K_{3A}\overline{K}_0}$ only depends on K_0.

According to Steps 4-2-1 and 4-2-2, we have $\left(\frac{\log(ed)}{n-1}\right)^{1/2} A_2^{1/2} \leq K_{3A}$ and $\left(\frac{\log(ed)}{n-1}\right)^{3/4} M_2^{1/2} \leq K_{3M}$ for all sufficiently large n and sufficiently small $t \asymp \sqrt{\frac{\log(ed)}{n-1}}$. Therefore, Steps 4-2-1 and 4-2-2 combined show that, for all sufficiently large n, sufficiently small $t \asymp \sqrt{\frac{\log(ed)}{n-1}}$, and $y = c_2\sqrt{\frac{\log(ed)}{n-1}}$ where $c_2 > 0$, we have

$$
8d^2\exp\left(-\frac{C_5(n-1)y}{A_2^{1/2} + y^{1/2}M_2^{1/2}}\right) \leq 8d^2\exp\left(-\frac{K_3c_2\log(ed)}{1 + c_2^{1/2}}\right),
$$

where K_3 only depends on $(K_0, \kappa_1, \kappa_2)$.

Step 5. Step 5 will upper bound $t' = C_6\left(t + \sum_{i=0}^1 v_i + F(t)\sum_{i=0}^1 v_i\right)$.

Following the proof of Proposition 5 in Shen et al. (2019), we have

$$
\begin{aligned}
v_0^2 &\leq 2\left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right] + 12\widetilde{M}_2D, \\
v_1^2 &\leq \left[\mathbb{P}(|U_{1,k}| \geq \widetilde{M}_1) + \mathbb{P}(|U_{1,m}| \geq \widetilde{M}_1)\right] + 12\widetilde{M}_2D.
\end{aligned}
$$

Step 3 has shown that

$$\begin{aligned}\{(n-1)d\}^2(v_0+v_1)^2 &\le 2\{(n-1)d\}^2(v_1^2+v_2^2)\\ &\le 8\{(n-1)d\}^2\left[\left[\mathbb{P}(|U_{1,k}|\ge\widetilde{M}_1)+\mathbb{P}(|U_{1,m}|\ge\widetilde{M}_1)\right]+6\widetilde{M}_2D\right]\le 4K_1\{(n-1)d\}^{-2}.\end{aligned}$$

Therefore, $\sum_{i=0}^1 v_i = \{(n-1)d\}^{-1}\{(n-1)d\}\sum_{i=0}^1 v_i \le 2\sqrt{K_1}\{(n-1)d\}^{-2}$. Also, for all sufficiently large n and the sufficiently small $t=\frac{1}{2}\sqrt{\frac{\log(ed)}{n-1}}$,

$$\begin{aligned}F(t)\sum_{i=0}^1 v_i = \frac{1}{d}\frac{F(t)}{t(n-1)}t\{(n-1)d\}\sum_{i=0}^1 v_i &\le \frac{1}{d}\left[2C_7\frac{\log^2(n-1)}{\sqrt{(n-1)\log(ed)}}+K_2\frac{1}{\log(ed)}\right]2\sqrt{K_1}\{(n-1)d\}^{-1}t\\ &= 2C_7\sqrt{K_1}\frac{\log^2(n-1)}{(n-1)^2d^2}+\frac{\sqrt{K_1}K_2}{d^2(n-1)\sqrt{(n-1)\log(ed)}}.\end{aligned}$$

Therefore, for all sufficiently large n and sufficiently small $t=\frac{1}{2}\sqrt{\frac{\log(ed)}{n-1}}$,

$$\begin{aligned}t' &= C_6\left(t+\sum_{i=0}^1 v_i+F(t)\sum_{i=0}^1 v_i\right)\\ &\le \frac{C_6}{2}\sqrt{\frac{\log(ed)}{n-1}}+\frac{2C_6\sqrt{K_1}}{(n-1)^2d^2}+\frac{2C_6C_7\sqrt{K_1}\log^2(n-1)}{(n-1)^2d^2}+\frac{C_6\sqrt{K_1}K_2}{d^2(n-1)\sqrt{(n-1)\log(ed)}}.\end{aligned}$$

From Steps 3 to 5, when $y=c_2\sqrt{\frac{\log(ed)}{n-1}}$ for some positive constant c_2 and $t=\frac{1}{4\sqrt{K_0}}\sqrt{\frac{\log(ed)}{n-1}}$, for all sufficiently large n and sufficiently small t,

$$\begin{aligned}&\sum_{k,m=1}^{2d}\mathbb{P}\left(\left|\frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1}g_2\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix},\begin{pmatrix}U_{j,k}\\ U_{j,m}\end{pmatrix}\right)\right|\ge c_2\sqrt{\frac{\log(ed)}{n-1}}+C_7\frac{\log^2(n-1)}{(n-1)}+\frac{K_2}{2\sqrt{(n-1)\log(ed)}}\right.\\ &\qquad\left.+\frac{C_4C_6}{2}\sqrt{\frac{\log(ed)}{n-1}}+\frac{2C_4C_6\sqrt{K_1}}{(n-1)^2d^2}+\frac{2C_4C_6C_7\sqrt{K_1}\log^2(n-1)}{(n-1)^2d^2}+\frac{C_4C_6\sqrt{K_1}K_2}{d^2(n-1)\sqrt{(n-1)\log(ed)}}\right)\\ &\le \sum_{k,m=1}^{2d}\mathbb{P}\left(\left|\frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1}g_2\left(\begin{pmatrix}U_{i,k}\\ U_{i,m}\end{pmatrix},\begin{pmatrix}U_{j,k}\\ U_{j,m}\end{pmatrix}\right)\right|\ge c_2\sqrt{\frac{\log(ed)}{n-1}}+\frac{F(t)}{n-1}+C_4t'\right)\\ &\le 8d^2\exp\left(-K_3\frac{c_2}{1+c_2^{1/2}}\log(ed)\right)+K_1\{(n-1)d\}^{-2}.\end{aligned}$$

For some positive constant Q to be specified later, let's set

$$z = Q\sqrt{\frac{\log(ed)}{n-1}} + \frac{C\log^2(n-1)}{(n-1)} + \frac{K_2}{2\sqrt{(n-1)\log(ed)}} + \frac{C}{2}\sqrt{\frac{\log(ed)}{n-1}} + \frac{2C\sqrt{K_1}}{(n-1)^2d^2} + 2C^2\sqrt{K_1}\frac{\log^2(n-1)}{(n-1)^2d^2} + \frac{C\sqrt{K_1}K_2}{d^2(n-1)\sqrt{(n-1)\log(ed)}},$$

where $C := \max\{C_4C_6, C_7\}$ is an absolute positive constant. Then, for all sufficiently large n and sufficiently small $t \asymp \sqrt{\frac{\log(ed)}{n-1}}$,

$$\begin{aligned}
&\mathbb{P}\left(\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\max} \geq z\right) \leq \sum_{k,m=1}^{2d} \mathbb{P}\left(\|\widehat{V}_{km} - V_{km}\| \geq z\right) \\
&\leq \sum_{k,m=1}^{2d} \mathbb{P}\left(\left|\frac{1}{n-1}\sum_{i=1}^{n-1} g_1\left(\begin{pmatrix} U_{i,k} \\ U_{i,m}\end{pmatrix}\right)\right| \geq \frac{z}{4}\right) + \sum_{k,m=1}^{2d} \mathbb{P}\left(\left|\frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1} g_2\left(\begin{pmatrix} U_{i,k} \\ U_{i,m}\end{pmatrix}, \begin{pmatrix} U_{j,k} \\ U_{j,m}\end{pmatrix}\right)\right| \geq \frac{z}{2}\right) \\
&\leq \sum_{k,m=1}^{2d} \mathbb{P}\left(\left|\frac{1}{n-1}\sum_{i=1}^{n-1} g_1\left(\begin{pmatrix} U_{i,k} \\ U_{i,m}\end{pmatrix}\right)\right| \geq \frac{Q}{4}\sqrt{\frac{\log(ed)}{n-1}}\right) + \sum_{k,m=1}^{2d} \mathbb{P}\left(\left|\frac{1}{(n-1)^2}\sum_{i,j=1}^{n-1} g_2\left(\begin{pmatrix} U_{i,k} \\ U_{i,m}\end{pmatrix}, \begin{pmatrix} U_{j,k} \\ U_{j,m}\end{pmatrix}\right)\right| \geq \frac{z}{2}\right) \\
&\leq 8d^2\exp\left(-\widetilde{K}_0\frac{(Q/4)^2}{1+(Q/4)}\log(ed)\right) + 8d^2\exp\left(-K_3\frac{(Q/2)}{1+(Q/2)^{1/2}}\log(ed)\right) + K_1\{(n-1)d\}^{-2}.
\end{aligned}$$

Therefore, by choosing sufficiently large Q, we have $\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\max} = O_{\mathbb{P}}(\sqrt{\log(ed)/n})$. By picking the value of Q large enough so that $\widetilde{K}_0\frac{(Q/4)^2}{1+(Q/4)} \geq 4$ and $K_3\frac{(Q/2)}{1+(Q/2)^{1/2}} \geq 4$, we have

$$\mathbb{P}\left(\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\max} \geq z\right) \leq 16e^{-4}d^{-2} + K_1\{(n-1)d\}^{-2}.$$

Lastly, noticing the following relationship

$$\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\max} \leq \frac{n-1}{n-2}\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\max} + \frac{1}{n-2}\|\mathbf{V}\|_{\max} \leq \frac{n-1}{n-2}\|\widehat{\mathbf{V}} - \mathbf{V}\|_{\max} + \frac{1}{n-2},$$

the proof is finished.

Proof of Theorem 1.2.2

The proof of Theorem 1.2.2 immediately follows from the lemma below, which is in the proof of Theorem 5.3 in Qiu et al. (2015).

Lemma A.1.1 (Master theorem)**.** *Suppose $A \in \mathcal{M}(s, M)$. Further suppose $\|\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\max} \leq \epsilon_1$ and $\|\widehat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\|_{\max} \leq \epsilon_2$ with probability no less than $1 - \epsilon_0$. Let λ in equation (1.6) be $\lambda \geq \epsilon_1 M + \epsilon_2$. Then with probability no less than $1 - \epsilon_0$, it holds that*

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\max} \leq 2\|\mathbf{\Sigma}_0^{-1}\|_\infty[\epsilon_1 M + \epsilon_2] = 2\|\mathbf{\Sigma}^{-1}\|_\infty \lambda \quad \text{and}$$
$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_\infty \leq 4s\|\mathbf{\Sigma}_0^{-1}\|_\infty[\epsilon_1 M + \epsilon_2] = 4s\|\mathbf{\Sigma}^{-1}\|_\infty \lambda.$$

Technical lemmas used in Section 1.2

Lemma A.1.2 (Theorem 1 in Merlevède et al. (2009))**.** *Let $\{X_t\}_{t\in\mathbb{Z}}$ be a sequence of centered real-valued random variables. Suppose that the sequence satisfies that $\alpha(m) \leq \kappa_1 \exp(-\kappa_2 m)$ for all $m \geq 1$, and that there exists a positive M such that $\sup_{i\geq 1}\|X_i\|_\infty \leq M$. Then there exist positive constants C_1 and C_2 depending only on (κ_1, κ_2) such that for all $n \geq 4$ and t satisfying that for any $0 < t < \{C_1 M(\log n)(\log\log n)\}^{-1}$, we have*

$$\log \mathbb{E}\left(\exp\left(t\sum_{i=1}^n X_i\right)\right) \leq \frac{C_2 t^2 n M^2}{1 - C_1 t M(\log n)(\log\log n)}.$$

In terms of probability, there exists a constant C_3 only depending on (κ_1, κ_2) such that for all $n \geq 4$ and $x \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq x\right) \leq \exp\left(-\frac{C_3 x^2}{nM^2 + Mx(\log n)(\log\log n)}\right). \tag{A.3}$$

Lemma A.1.3. *Suppose that the latent process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ follows VAR(1) model (1.1) with Assumption M(ii) and (iii). Let $\mathbf{\Omega}_j := \mathrm{Var}\left((\boldsymbol{Z}_t^\top, \ldots, \boldsymbol{Z}_{t+j}^\top)^\top\right)$ for any $j \in \{0, 1, 2, \cdots\}$. Then, $\mathbf{\Omega}_j$ is positive definite and $0 < c_{\mathbf{\Omega}} < \lambda_{\min}(\mathbf{\Omega}_j) \leq \lambda_{\max}(\mathbf{\Omega}_j) < C_{\mathbf{\Omega}} < \infty$ for some constants $c_{\mathbf{\Omega}}$ and $C_{\mathbf{\Omega}}$ that only depend on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$.*

Proof. We will prove the lemma in two steps: first for the case that $j = 0$ and then the case that $j > 0$.

Step 1. When $j=0$, $\mathbf{\Omega}_0 = \mathbf{\Sigma}_0 = \sum_{k=0}^{\infty} \mathbf{A}^k \mathbf{\Sigma}_{\boldsymbol{E}} (\mathbf{A}^\top)^k$. Then,

$$\lambda_{\min}(\mathbf{\Omega}_j) = \lambda_{\min}\left(\mathbf{\Sigma}_{\boldsymbol{E}} + \sum_{k=1}^{\infty} \mathbf{A}^k \mathbf{\Sigma}_{\boldsymbol{E}} (\mathbf{A}^\top)^k\right) \geq \lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{E}}) \geq c_{\boldsymbol{E}} > 0, \text{ and}$$

$$\lambda_{\max}(\mathbf{\Omega}_j) \leq \sum_{k=0}^{\infty} \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{E}}) \|(\mathbf{A}^\top)^k\|_2^2 \leq \frac{\lambda_{\max}(\mathbf{\Sigma}_E)}{1-\|\mathbf{A}\|_2^2} \leq \frac{\lambda_{\max}(\mathbf{\Sigma}_E)}{1-C_{\mathbf{A}}^2}.$$

Step 2. We investigate the case when $j>0$. We can represent $\mathbf{\Omega}_j$ as follows (cf. Section 1.4 in Gómez (2016)).

$$\mathbf{\Omega}_j = \mathrm{Var}\left(\begin{pmatrix} \mathbf{I}_d & \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} \\ \mathbf{A} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \cdots \\ \vdots & \ddots & \ddots & \ddots \\ \mathbf{A}^j & \cdots & \mathbf{A} & \mathbf{I}_d \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_t \\ \boldsymbol{E}_{t+1} \\ \vdots \\ \boldsymbol{E}_{t+j} \end{pmatrix}\right) = \mathbf{L} \begin{pmatrix} \mathbf{\Omega}_0 & \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \mathbf{\Sigma}_{\boldsymbol{E}} & \mathbf{0}_{d\times d} & \cdots \\ \vdots & \ddots & \ddots & \ddots \\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{\Sigma}_{\boldsymbol{E}} \end{pmatrix} \mathbf{L}^\top,$$

with

$$\mathbf{L} = \begin{pmatrix} \mathbf{I}_d & \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} \\ \mathbf{A} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \cdots \\ \vdots & \ddots & \ddots & \ddots \\ \mathbf{A}^j & \cdots & \mathbf{A} & \mathbf{I}_d \end{pmatrix} \text{ and } \mathbf{L}^{-1} = \begin{pmatrix} \mathbf{I}_d & \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} \\ -\mathbf{A} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \ddots \\ \mathbf{0}_{d\times d} & \ddots & \ddots & \ddots \\ \vdots & \mathbf{0}_{d\times d} & -\mathbf{A} & \mathbf{I}_d \end{pmatrix}.$$

We have $\|\mathbf{L}\|_2 = \sum_{k=0}^{j} \|\mathbf{A}^k\|_2 \leq \frac{1}{1-\|\mathbf{A}\|_2}$ and $\|\mathbf{L}^{-1}\|_2 \leq 1 + \|\mathbf{A}\|_2$. Therefore,

$$\lambda_{\max}(\mathbf{\Omega}_j) \leq \lambda_{\max}(\mathbf{\Omega}_0)\|\mathbf{L}^\top\|_2^2 \leq \frac{\lambda_{\max}(\mathbf{\Omega}_0)}{(1-\|\mathbf{A}\|_2)^2}, \text{ and}$$

$$\frac{1}{\lambda_{\min}(\mathbf{\Omega}_j)} = \lambda_{\max}(\mathbf{\Omega}_j^{-1}) \leq \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{E}}^{-1})\|\mathbf{L}^{-1}\|_2^2 \leq \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{E}}^{-1})(1+\|\mathbf{A}\|_2)^2.$$

Combining Steps 1 and 2 concludes. □

A.1.2 Proofs in Section 1.3

Notation

In addition to the notations introduced in the main text and Section A.1.1, more are defined as follows. For a generic vector $\boldsymbol{v} \in \mathbb{R}^d$, let $S(\boldsymbol{v}) := \boldsymbol{w}^\top \widehat{\boldsymbol{h}}(\boldsymbol{v}) = \boldsymbol{w}^\top(\widehat{\mathbf{\Sigma}}_0 \boldsymbol{v} - \widehat{\mathbf{\Sigma}}_{1,*m})$. We denote $\boldsymbol{\beta}(c) := (c, \boldsymbol{\gamma}^\top)^\top$ when we use c in the position of θ. Similarly, we denote $\widehat{\boldsymbol{\beta}}(c) := (c, \widehat{\boldsymbol{\gamma}}^\top)^\top$.

For Kendall's tau $\mathbf{T}$, estimator $\widehat{\mathbf{T}}$, and bootstrap version $\widehat{\mathbf{T}}^*$, which are defined in subsections 1.2.2 and 1.3.2, we denote

- $\mathbf{T}_0 = \mathbf{T}_{[d],[d]}$, $\mathbf{T}_1 = \mathbf{T}_{[d],d+[d]}$, $\widehat{\mathbf{T}}_0 = \widehat{\mathbf{T}}_{[d],[d]}$, $\widehat{\mathbf{T}}_1 = \widehat{\mathbf{T}}_{[d],d+[d]}$, $\widehat{\mathbf{T}}_0^* = \widehat{\mathbf{T}}_{[d],[d]}^*$, and $\widehat{\mathbf{T}}_1^* = \widehat{\mathbf{T}}_{[d],d+[d]}^*$;

- $\tau_{0,jk}$, $\widehat{\tau}_{0,jk}$, and $\widehat{\tau}_{0,jk}^*$ are the (j,k)-th element of $\mathbf{T}_0$, $\widehat{\mathbf{T}}_0$, and $\widehat{\mathbf{T}}_0^*$, respectively;

- $\tau_{1,jk}$, $\widehat{\tau}_{1,jk}$, and $\widehat{\tau}_{1,jk}^*$ are the (j,k)-th element of $\mathbf{T}_1$, $\widehat{\mathbf{T}}_1$, and $\widehat{\mathbf{T}}_1^*$, respectively.

Then, $\boldsymbol{\Sigma}_0 = \sin\left(\frac{\pi}{2}\mathbf{T}_0\right)$, $\boldsymbol{\Sigma}_1 = \sin\left(\frac{\pi}{2}\mathbf{T}_1\right)$, $\widehat{\boldsymbol{\Sigma}}_0 = \sin\left(\frac{\pi}{2}\widehat{\mathbf{T}}_0\right)$, $\widehat{\boldsymbol{\Sigma}}_1 = \sin\left(\frac{\pi}{2}\widehat{\mathbf{T}}_1\right)$, $\widehat{\boldsymbol{\Sigma}}_0^* = \sin\left(\frac{\pi}{2}\widehat{\mathbf{T}}_0^*\right)$, and $\widehat{\boldsymbol{\Sigma}}_1^* = \sin\left(\frac{\pi}{2}\widehat{\mathbf{T}}_1^*\right)$. We define bootstrap version V-statistic $\widehat{\mathbf{V}}^*$ as

$$\widehat{\mathbf{V}}^* := \frac{1}{(b\ell)^2}\sum_{i,j=1}^{b\ell}\operatorname{sign}(\boldsymbol{Y}_i^* - \boldsymbol{Y}_j^*)\operatorname{sign}(\boldsymbol{Y}_i^* - \boldsymbol{Y}_j^*)^\top = \frac{1}{(b\ell)^2}\sum_{i,j=1}^{b\ell}\operatorname{sign}(\boldsymbol{U}_i^* - \boldsymbol{U}_j^*)\operatorname{sign}(\boldsymbol{U}_i^* - \boldsymbol{U}_j^*)^\top,$$

where $\boldsymbol{Y}_t^*$ is the bootstrap sample from $\boldsymbol{Y}_t = (\boldsymbol{X}_t^\top, \boldsymbol{X}_{t+1}^\top)^\top$ and $\boldsymbol{U}_t^*$ is the latent variable corresponding to $\boldsymbol{Y}_t^*$ via $\boldsymbol{Y}_t^* = \boldsymbol{f}(\boldsymbol{U}_t^*)$.

For $\mathbf{V}$, its estimator $\widehat{\mathbf{V}}$, and bootstrap estimator $\widehat{\mathbf{V}}^*$, we denote

- $\mathbf{V}_0 = \mathbf{V}_{[d],[d]}$, $\mathbf{V}_1 = \mathbf{V}_{[d],d+[d]}$, $\widehat{\mathbf{V}}_0 = \widehat{\mathbf{V}}_{[d],[d]}$, $\widehat{\mathbf{V}}_1 = \widehat{\mathbf{V}}_{[d],d+[d]}$, $\widehat{\mathbf{V}}_0^* = \widehat{\mathbf{V}}_{[d],[d]}^*$, and $\widehat{\mathbf{V}}_1^* = \widehat{\mathbf{V}}_{[d],d+[d]}^*$;

- $V_{0,jk}$, $\widehat{V}_{0,jk}$, and $\widehat{V}_{0,jk}^*$ are the (j,k)-th element of $\mathbf{V}_0$, $\widehat{\mathbf{V}}_0$, and $\widehat{\mathbf{V}}_0^*$, respectively;

- $V_{1,jk}$, $\widehat{V}_{1,jk}$, and $\widehat{V}_{1,jk}^*$ are the (j,k)-th element of $\mathbf{V}_1$, $\widehat{\mathbf{V}}_1$, and $\widehat{\mathbf{V}}_1^*$, respectively.

Then, we have the following relationships:

$$\widehat{V}_{0,jk} - V_{0,jk} = \frac{1}{(n-1)^2}\sum_{t,t'=1}^{n-1} g\left(\begin{pmatrix} Z_{t,j} \\ Z_{t,k}\end{pmatrix}, \begin{pmatrix} Z_{t',j} \\ Z_{t',k}\end{pmatrix}\right),$$

$$\widehat{V}_{1,jk} - V_{1,jk} = \frac{1}{(n-1)^2}\sum_{t,t'=1}^{n-1} g\left(\begin{pmatrix} Z_{t,j} \\ Z_{t+1,k}\end{pmatrix}, \begin{pmatrix} Z_{t',j} \\ Z_{t'+1,k}\end{pmatrix}\right),$$

$$\widehat{V}^*_{0,jk} - V_{0,jk} = \frac{1}{(b\ell)^2}\sum_{t,t'=1}^{b\ell} g\left(\begin{pmatrix} Z^*_{t,j} \\ Z^*_{t,k}\end{pmatrix}, \begin{pmatrix} Z^*_{t',j} \\ Z^*_{t',k}\end{pmatrix}\right), \text{ and}$$

$$\widehat{V}^*_{1,jk} - V_{1,jk} = \frac{1}{(b\ell)^2}\sum_{t,t'=1}^{b\ell} g\left(\begin{pmatrix} Z^*_{t,j} \\ Z^*_{t+1,k}\end{pmatrix}, \begin{pmatrix} Z^*_{t',j} \\ Z^*_{t'+1,k}\end{pmatrix}\right),$$

where $(X^*_{t,j}, X^*_{t,k})^\top$ and $(X^*_{t,j}, X^*_{t+1,k})^\top$ are bootstrap sample points from $\boldsymbol{Y}_t = (\boldsymbol{X}_t^\top, \boldsymbol{X}_{t+1}^\top)^\top$ and $(Z^*_{t,j}, Z^*_{t,k})^\top$ and $(Z^*_{t,j}, Z^*_{t+1,k})^\top$ are bootstrap sample points corresponding to $(X^*_{t,j}, X^*_{t,k})^\top$ and $(X^*_{t,j}, X^*_{t+1,k})^\top$, respectively, via $\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{Z}_t)$. Note that $\max_{j,k}|\widehat{V}_{0,jk} - V_{0,jk}| = O_{\mathbb{P}}\left(\sqrt{\frac{\log(ed)}{n}}\right)$ and $\max_{j,k}|\widehat{V}_{1,jk} - V_{1,jk}| = O_{\mathbb{P}}\left(\sqrt{\frac{\log(ed)}{n}}\right)$ by the proof of Theorem 1.2.1.

Verification of Assumptions 1-5 in Neykov et al. (2018a)

In this subsection, we will verify Assumptions 1 to 5 in Neykov et al. (2018a). Assumptions 1, 2, 4, and 5 have been verified in Appendix J in Neykov et al. (2018b). We will verify Assumption 3 by following the proofs of Lemma H.4 in Neykov et al. (2018b), Lemmas 3.3, 3.6 to 3.8 in Dehling and Wendler (2010), Lemma 3 in Yoshihara (1976), and Theorem 3.1 in Fan et al. (2016).

Lemma A.1.4 (Assumption 3 in Neykov et al. (2018a))**.** *Suppose all assumptions in Theorem 1.3.1 and Theorem 1.3.2 hold. Then, Assumption 3 in Neykov et al. (2018a) holds, that is,*

$$\frac{\sqrt{n-1}}{\sigma_n} S(\boldsymbol{\beta}) \xrightarrow{d} N(0,1).$$

Proof. In this part, we will follow the proofs of Lemma H.4 in Neykov et al. (2018b), Lemmas 3.6 and 3.8 in Dehling and Wendler (2010), and Theorem 3.1 in Fan et al. (2016).

Following the proof of Lemma H.4 in Neykov et al. (2018b), we have

$$
\begin{aligned}
\sqrt{n-1}S(\boldsymbol{\beta}) &= \sqrt{n-1}\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}) \\
&= \sqrt{n-1}\boldsymbol{w}^\top((\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0)\boldsymbol{\beta} - (\widehat{\boldsymbol{\Sigma}}_{1,*m} - \boldsymbol{\Sigma}_{1,*m})) \\
&= \sqrt{n-1}\boldsymbol{w}^\top\left\{\left(\sin\left(\frac{\pi}{2}\widehat{\mathbf{T}}_0\right) - \sin\left(\frac{\pi}{2}\mathbf{T}_0\right)\right)\boldsymbol{\beta} - \left(\sin\left(\frac{\pi}{2}\widehat{\mathbf{T}}_{1,*m}\right) - \sin\left(\frac{\pi}{2}\mathbf{T}_{1,*m}\right)\right)\right\} \\
&= \sqrt{n-1}\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k\left(\sin\left(\frac{\pi}{2}\widehat{\tau}_{0,jk}\right) - \sin\left(\frac{\pi}{2}\tau_{0,jk}\right)\right) + \sum_{j\in S_{\boldsymbol{w}}} w_j\left(\sin\left(\frac{\pi}{2}\widehat{\tau}_{1,jm}\right) - \sin\left(\frac{\pi}{2}\tau_{1,jm}\right)\right)\right\} \\
&= \underbrace{\sqrt{n-1}\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k\cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\frac{\pi}{2}(\widehat{\tau}_{0,jk} - \tau_{0,jk}) - \sum_{j\in S_{\boldsymbol{w}}} w_j\cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\frac{\pi}{2}(\widehat{\tau}_{1,jm} - \tau_{1,jm})\right\}}_{K_1} \\
&\quad \underbrace{- \sqrt{n-1}\left\{\frac{1}{2}\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k\sin\left(\frac{\pi}{2}\tau_{0,jk}\right)\left(\frac{\pi}{2}(\widehat{\tau}_{0,jk} - \tau_{0,jk})\right)^2 - \frac{1}{2}\sum_{j\in S_{\boldsymbol{w}}} w_j\sin\left(\frac{\pi}{2}\tau_{1,jm}\right)\left(\frac{\pi}{2}(\widehat{\tau}_{1,jm} - \tau_{1,jm})\right)^2\right\}}_{K_2},
\end{aligned}
$$

where $\widetilde{\tau}_{0,jk}$ is between $\widehat{\tau}_{0,jk}$ and $\tau_{0,jk}$, and $\widetilde{\tau}_{1,jm}$ is between $\widehat{\tau}_{1,jm}$ and $\tau_{1,jm}$. The second equality holds because $\boldsymbol{\Sigma}_0\boldsymbol{\beta} - \boldsymbol{\Sigma}_{1,*m} = \mathbf{0}$. Steps 1 and 2 show $K_2 = o_{\mathbb{P}}(1)$. We also have $K_1/\nu_n \xrightarrow{d} N(0,1)$ for an appropriate value of ν_n, which will be presented in Step 2-2.

Step 1. We will show $K_2 = o_{\mathbb{P}}(1)$ by proving $\mathbb{E}[K_2^2] = o(1)$ using Lemma A.1.10 in the supplementary material, which follows the proof of Lemma 3.6 in Dehling and Wendler (2010).

Because $(a+b)^2 \le 2(a^2+b^2)$, it is sufficient to show that

$$
(n-1)\mathbb{E}\left[\left(\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k\sin\left(\frac{\pi}{2}\tau_{0,jk}\right)(\widehat{\tau}_{0,jk} - \tau_{0,jk})^2\right)^2\right] = o(1) \text{ and}
$$

$$
(n-1)\mathbb{E}\left[\left(\sum_{j\in S_{\boldsymbol{w}}} w_j\sin\left(\frac{\pi}{2}\tau_{1,jm}\right)(\widehat{\tau}_{1,jm} - \tau_{1,jm})^2\right)^2\right] = o(1).
$$

For the first one, we have

$$
\begin{aligned}
&\left|(n-1)\mathbb{E}\left[\left(\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \sin\left(\frac{\pi}{2}\tau_{0,jk}\right)(\widehat{\tau}_{0,jk}-\tau_{0,jk})^2\right)^2\right]\right| \\
&= \left|(n-1)\mathbb{E}\left[\sum_{j\in S_{\boldsymbol{w}},k\in S}\sum_{j'\in S_{\boldsymbol{w}},k'\in S} w_j\beta_k w_{j'}\beta_{k'} \sin\left(\frac{\pi}{2}\tau_{0,jk}\right)\sin\left(\frac{\pi}{2}\tau_{0,j'k'}\right)(\widehat{\tau}_{0,jk}-\tau_{0,jk})^2(\widehat{\tau}_{0,j'k'}-\tau_{0,j'k'})^2\right]\right| \\
&\le (n-1)\mathbb{E}\left[\sum_{j\in S_{\boldsymbol{w}},k\in S}\sum_{j'\in S_{\boldsymbol{w}},k'\in S}|w_j||\beta_k||w_{j'}||\beta_{k'}|(\widehat{\tau}_{0,jk}-\tau_{0,jk})^2(\widehat{\tau}_{0,j'k'}-\tau_{0,j'k'})^2\right] \\
&\le \sum_{j\in S_{\boldsymbol{w}},k\in S}\sum_{j'\in S_{\boldsymbol{w}},k'\in S}|w_j||\beta_k||w_{j'}||\beta_{k'}|(n-1)\mathbb{E}[(\widehat{\tau}_{0,jk}-\tau_{0,jk})^2(\widehat{\tau}_{0,j'k'}-\tau_{0,j'k'})^2] \\
&\le \frac{1}{2}\sum_{j\in S_{\boldsymbol{w}},k\in S}\sum_{j'\in S_{\boldsymbol{w}},k'\in S}|w_j||\beta_k||w_{j'}||\beta_{k'}|(n-1)\left\{\mathbb{E}[(\widehat{\tau}_{0,jk}-\tau_{0,jk})^4]+\mathbb{E}[(\widehat{\tau}_{0,j'k'}-\tau_{0,j'k'})^4]\right\} \\
&= \|\boldsymbol{w}\|_1\|\boldsymbol{\beta}\|_1\sum_{j\in S_{\boldsymbol{w}},k\in S}|w_j||\beta_k|(n-1)\mathbb{E}[(\widehat{\tau}_{0,jk}-\tau_{0,jk})^4],
\end{aligned}
$$

where the last inequality comes from $|ab|\le (a^2+b^2)/2$. Because $(a+b)^4\le 8a^4+8a^4$, we have

$$
\begin{aligned}
&\mathbb{E}[(\widehat{\tau}_{0,jk}-\tau_{0,jk})^4] \\
=&\mathbb{E}\left[\left\{\frac{2}{n-1}\sum_{t=1}^{n-1} g_1(\boldsymbol{Z}_{t,\{j,k\}})+\frac{2}{(n-1)(n-2)}\sum_{t<t'} g_2(\boldsymbol{Z}_{t,\{j,k\}},\boldsymbol{Z}_{t',\{j,k\}})\right\}^4\right] \\
\le&\frac{128}{(n-1)^4}\mathbb{E}\left[\left(\sum_{t=1}^{n-1} g_1(\boldsymbol{Z}_{t,\{j,k\}})\right)^4\right]+\frac{128}{(n-1)^4(n-2)^4}\mathbb{E}\left[\left(\sum_{t<t'} g_2(\boldsymbol{Z}_{t,\{j,k\}},\boldsymbol{Z}_{t',\{j,k\}})\right)^4\right] \\
\le&\frac{128}{(n-1)^4}\sum_{i_1,i_2,i_3,i_4=1}^{n-1}\left|\mathbb{E}[g_1(\boldsymbol{Z}_{i_1,\{j,k\}})g_1(\boldsymbol{Z}_{i_2,\{j,k\}})g_1(\boldsymbol{Z}_{i_3,\{j,k\}})g_1(\boldsymbol{Z}_{i_4,\{j,k\}})]\right| \\
&\quad+\frac{8}{(n-1)^4(n-2)^4}\sum_{i_1,\dots,i_8=1}^{n-1}\left|\mathbb{E}[g_2(\boldsymbol{Z}_{i_1,\{j,k\}},\boldsymbol{Z}_{i_2,\{j,k\}})g_2(\boldsymbol{Z}_{i_3,\{j,k\}},\boldsymbol{Z}_{i_4,\{j,k\}})g_2(\boldsymbol{Z}_{i_5,\{j,k\}},\boldsymbol{Z}_{i_6,\{j,k\}})g_2(\boldsymbol{Z}_{i_7,\{j,k\}},\boldsymbol{Z}_{i_8,\{j,k\}})]\right|.
\end{aligned}
$$

Because $\{\boldsymbol{U}_t\}_{t\in\mathbb{Z}}$ is geometrically α-mixing, Lemma A.1.10 in the supplementary material implies $\mathbb{E}[(\widehat{\tau}_{0,jk}-\tau_{0,jk})^4]\le C(n-1)^{-2}$, where C is a constant which only depends on ($C_{\mathbf{A}}$, $c_{\boldsymbol{E}}$, $C_{\boldsymbol{E}}$, κ_1, κ_2). Therefore, using information that $\|\boldsymbol{\beta}\|_1\le M=O(1)$ and $\|\boldsymbol{w}\|_1\le\|\boldsymbol{\Sigma}_0^{-1}\|_1=\|\boldsymbol{\Sigma}_0^{-1}\|_\infty=O(1)$, one has

$$
\left|(n-1)\mathbb{E}\left[\left(\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \sin\left(\frac{\pi}{2}\tau_{0,jk}\right)(\widehat{\tau}_{0,jk}-\tau_{0,jk})^2\right)^2\right]\right|\le C\|\boldsymbol{w}\|_1^2\|\boldsymbol{\beta}\|_1^2 O(n^{-1})=o(1).
$$

Similarly, we have $(n-1)\mathbb{E}\left[\left(\sum_{j\in S_{\boldsymbol{w}}} w_j \sin\left(\frac{\pi}{2}\tau_{1,jm}\right)(\widehat{\tau}_{1,jm}-\tau_{1,jm})^2\right)^2\right] = o(1)$. In conclusion, we have $\mathbb{E}[K_2^2] = o(1)$, $\mathrm{Var}(K_2) = o(1)$, and $K_2 = o_{\mathbb{P}}(1)$.

Step 2. We will investigate the asymptotic distribution of K_1.

Using Hoeffding decomposition of $\widehat{\tau}_{0,jm} - \tau_{0,jm}$ and $\widehat{\tau}_{1,jm} - \tau_{1,jm}$, we can decompose K_1 as

$$K_1 = \sqrt{n-1}\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\frac{\pi}{2}(\widehat{\tau}_{0,jk}-\tau_{0,jk}) - \sum_{j\in S_w} w_j\cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\frac{\pi}{2}(\widehat{\tau}_{1,jm}-\tau_{1,jm})\right\}$$

$$= K_{11} + K_{12}, \text{ where}$$

$$K_{11} = \sqrt{n-1}\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k\cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\frac{\pi}{n-1}\sum_{t=1}^{n-1} g_1\left(\begin{pmatrix} Z_{t,j}\\ Z_{t,k}\end{pmatrix}\right) - \sum_{j\in S_{\boldsymbol{w}}} w_j\cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\frac{\pi}{n-1}\sum_{t=1}^{n-1} g_1\left(\begin{pmatrix} Z_{t,j}\\ Z_{t+1,m}\end{pmatrix}\right)\right\},$$

$$\begin{aligned} K_{12} = \sqrt{n-1}\Bigg\{&\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k\cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\frac{\pi}{(n-1)(n-2)}\sum_{t<t'} g_2\left(\begin{pmatrix} Z_{t,j}\\ Z_{t,k}\end{pmatrix},\begin{pmatrix} Z_{t',j}\\ Z_{t',k}\end{pmatrix}\right)\\ &-\sum_{j\in S_{\boldsymbol{w}}} w_j\cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\frac{\pi}{(n-1)(n-2)}\sum_{t<t'} g_2\left(\begin{pmatrix} Z_{t,j}\\ Z_{t+1,m}\end{pmatrix},\begin{pmatrix} Z_{t',j}\\ Z_{t'+1,m}\end{pmatrix}\right)\Bigg\}.\end{aligned}$$

Steps 2-1 and 2-2 combined show that $K_{12} = o_{\mathbb{P}}(1)$ and $K_{11}/\nu_n \xrightarrow{d} N(0,1)$, where $\nu_n^2 = \mathbb{E}[K_{11}^2]$.

Step 2-1. In this part, we will show $K_{12} = o_{\mathbb{P}}(1)$ by proving $\mathbb{E}[K_{12}^2] = o(1)$. Similar to Step 1, Lemma A.1.10 in the supplementary material implies

$$(n-1)\mathbb{E}\left[\left(\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k\cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\frac{\pi}{(n-1)(n-2)}\sum_{t<t'} g_2\left(\begin{pmatrix} Z_{t,j}\\ Z_{t,k}\end{pmatrix},\begin{pmatrix} Z_{t',j}\\ Z_{t',k}\end{pmatrix}\right)\right)^2\right] \leq C\|\boldsymbol{w}\|_1^2\|\boldsymbol{\beta}\|_1^2 O(n^{-}$$

and

$$(n-1)\mathbb{E}\left[\left(\sum_{j\in S_{\boldsymbol{w}}} w_j\cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\frac{\pi}{(n-1)(n-2)}\sum_{t<t'} g_2\left(\begin{pmatrix} Z_{t,j}\\ Z_{t+1,m}\end{pmatrix},\begin{pmatrix} Z_{t',j}\\ Z_{t'+1,m}\end{pmatrix}\right)\right)^2\right] \leq C'\|\boldsymbol{w}\|_1^2 O(n^{-1}),$$

where C and C' are some constants which only depend on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, \kappa_1, \kappa_2)$. Therefore, we have $\mathbb{E}[K_{12}^2] = o(1)$, $\mathrm{Var}(K_{12}) = o(1)$, and $K_{12} = o_{\mathbb{P}}(1)$.

Step 2-2. We show $K_{11}/\nu_n \xrightarrow{d} N(0,1)$, where $\nu_n^2 = \mathbb{E}[K_{11}^2]$. First, we can rewrite K_{11} as

$$K_{11} = \frac{1}{\sqrt{n-1}} \sum_{t=1}^{n-1} \left[\underbrace{\sum_{j \in S_{\boldsymbol{w}}, k \in S} w_j \beta_k \pi \cos\left(\frac{\pi}{2}\tau_{0,jk}\right) g_1\left(\begin{pmatrix} Z_{t,j} \\ Z_{t,k} \end{pmatrix}\right) - \sum_{j \in S_{\boldsymbol{w}}} w_j \pi \cos\left(\frac{\pi}{2}\tau_{1,jm}\right) g_1\left(\begin{pmatrix} Z_{t,j} \\ Z_{t+1,m} \end{pmatrix}\right)}_{G_{n1}(\boldsymbol{U}_t)} \right].$$

Then, $\{G_{n1}(\boldsymbol{U}_t)\}_{t\in\mathbb{Z}}$ is also geometrically α-mixing because $G_{n1}(\boldsymbol{U}_t)$ is the function of $\boldsymbol{U}_t = (\boldsymbol{Z}_t^\top, \boldsymbol{Z}_{t+1}^\top)^\top$ and $\boldsymbol{U}_t$ is geometrically α-mixing. Also, $G_{n1}(\boldsymbol{U}_t)$ is bounded above by some absolute positive constant under our assumption because $|G_{n1}(\boldsymbol{U}_t)| \le 2\pi\|\boldsymbol{w}\|_1(\|\boldsymbol{\beta}\|_1 + 1) \le 2\pi\|\boldsymbol{\Sigma}_0^{-1}\|_1(M+1) = O(1)$. It implies that $\nu_n^2 < C$ by some absolute positive constant C. Also, similar to the proof of Theorem 3.1 in Fan et al. (2016), we have

$$|\sigma_n^2 - \mathbb{E}[K_{11}^2]| \le \mathrm{Var}(K_{12}+K_2) + \sqrt{\mathbb{E}[K_{11}^2]}\sqrt{\mathrm{Var}(K_{12}+K_2)} = o(1) + O(1)o(1), \tag{A.4}$$

which implies $\sigma_n^2 - \mathbb{E}[K_{11}^2] = o(1)$. Assumptions in Theorem 1.3.2 imply that we $\sigma_n^2 = \mathrm{Var}(K_{11} + K_{12} + K_2) > C > 0$ by some absolute constant C. This implies that there exist absolute positive constants c and C, and a positive constant N which only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, \kappa_1, \kappa_2)$ such that we have $c < \nu_n^2 = \mathbb{E}[K_{11}^2] < C$ for all $n \ge N$. Therefore, we can use triangular array CLT and get the following result (see Theorem 10.2 in Pötscher and Prucha (1997), and also Ekström (2014)), $K_{11}/\nu_n \xrightarrow{d} N(0,1)$ where $\nu_n^2 = \mathbb{E}[K_{11}^2]$. Because Steps 1 and 2-1 imply $\mathrm{Var}(K_{12}) = o(1)$, $\mathrm{Var}(K_2) = o(1)$, $K_{12} = o_{\mathbb{P}}(1)$, and $K_2 = o_{\mathbb{P}}(1)$, we have the desired result. $\square$

Proofs of Theorems 1.3.1 and 1.3.2

Given Lemma A.1.4 in the appendix of the main paper, Theorem 1.3.1 now holds by checking all conditions in Theorem 1 in Neykov et al. (2018a), and Theorem 1.3.2 now holds by Theorem 2 in Neykov et al. (2018a).

Proof of Theorem 1.3.3

Note that under assumptions in Theorem 1.3.3, Theorem 1.2.2 implies $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = O_{\mathbb{P}}(s\sqrt{\log(ed)/n})$, and the proof of Theorem 1.3.1 implies $\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_1 = O_{\mathbb{P}}(s_{\boldsymbol{w}}\sqrt{\log(ed)/n})$ by verifying Assumption 2 in Neykov et al. (2018a). Given this information, we will prove Theorem 1.3.3 in two steps. Step 1 will show $(b\ell)\mathrm{Var}^*(\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)) - \sigma_n^2 = o_{\mathbb{P}}(1)$. Step 2 will show $\widehat{\sigma}_n^{2*} - (b\ell)\mathrm{Var}^*(\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)) = o_{\mathbb{P}}(1)$.

Step 1. In this step, we will show $(b\ell)\mathrm{Var}^*(\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)) - \sigma_n^2 = o_{\mathbb{P}}(1)$. Let $S^*(\boldsymbol{\beta}) = \boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)$. Then

$$
\begin{aligned}
&\sqrt{b\ell}S^*(\boldsymbol{\beta})\\
&= \underbrace{\sqrt{b\ell}\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\frac{\pi}{2}(\widehat{\tau}_{0,jk}^* - \tau_{0,jk}) - \sum_{j\in S_{\boldsymbol{w}}} w_j \cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\frac{\pi}{2}(\widehat{\tau}_{1,jm}^* - \tau_{1,jm})\right\}}_{K_1^*}\\
&\quad \underbrace{-\frac{\sqrt{b\ell}}{2}\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \sin\left(\frac{\pi}{2}\widetilde{\tau}_{jk}^*\right)\left[\frac{\pi}{2}(\widehat{\tau}_{0,jk}^* - \tau_{0,jk})\right]^2 - \sum_{j\in S_{\boldsymbol{w}}} w_j \sin\left(\frac{\pi}{2}\widetilde{\tau}_{1,jm}^*\right)\left[\frac{\pi}{2}(\widehat{\tau}_{1,jm}^* - \tau_{1,jm})\right]^2\right\}}_{K_2^*},
\end{aligned}
$$

where $\widetilde{\tau}_{0,jk}^*$ is between $\widehat{\tau}_{0,jk}^*$ and $\tau_{0,jk}$, and $\widetilde{\tau}_{1,jm}^*$ is between $\widehat{\tau}_{1,jm}^*$ and $\tau_{1,jm}$. We have

$$
\begin{aligned}
K_1^* &= \sqrt{b\ell}\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\frac{\pi}{2}(\widehat{\tau}_{0,jk}^* - \tau_{0,jk}) - \sum_{j\in S_{\boldsymbol{w}}} w_j \cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\frac{\pi}{2}(\widehat{\tau}_{1,jm}^* - \tau_{1,jm})\right\}\\
&= \underbrace{\frac{1}{\sqrt{b\ell}}\sum_{t=1}^n\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\pi g_1\left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t,k}^*\end{pmatrix}\right) - \sum_{j\in S_{\boldsymbol{w}}} w_j \cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\pi g_1\left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t+1,m}^*\end{pmatrix}\right)\right\}}_{K_{11}^*}\\
&\quad + K_{12}^*,
\end{aligned}
$$

where

$$
\begin{aligned}
K_{12}^* := &\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\pi\frac{\sqrt{b\ell}}{(b\ell)(b\ell-1)}\sum_{t<t'} g_2\left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t,k}^*\end{pmatrix}, \begin{pmatrix} Z_{t',j}^* \\ Z_{t',k}^*\end{pmatrix}\right)\\
&- \sum_{j\in S_{\boldsymbol{w}}} w_j \cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\pi\frac{\sqrt{b\ell}}{(b\ell)(b\ell-1)}\sum_{t<t'} g_2\left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t+1,m}^*\end{pmatrix}, \begin{pmatrix} Z_{t',j}^* \\ Z_{t'+1,m}^*\end{pmatrix}\right).
\end{aligned}
$$

Steps 1-1 to 1-3 below combined will show $\mathbb{E}^*[(K_2^*)^2] = o_{\mathbb{P}}(1)$, $\mathrm{Var}^*(K_{12}^*) = o_{\mathbb{P}}(1)$, and $\mathrm{Var}^*(K_{11}^*) - \mathbb{E}[K_{11}^2] = o_{\mathbb{P}}(1)$ by following the proofs of Theorem 2.1 in Dehling and Wendler (2010), Theorem 2.1 in Gonçalves and White (2002), and Theorem 3.1 in Lahiri (2003). Note that Steps 1-1 to 1-3 imply $\mathrm{Var}^*(K_{11}^*) - (b\ell)\mathrm{Var}^*(\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)) = o_{\mathbb{P}}(1)$ because, similar to the proof of Theorem 3.2 in Fan et al. (2016), we have

$$|\mathrm{Var}^*(K_{11}^*) - (b\ell)\mathrm{Var}^*(\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*))| \leq \mathrm{Var}^*(K_{12}^* + K_2^*) + \sqrt{\mathrm{Var}^*(K_{11}^*)}\sqrt{\mathrm{Var}^*(K_{12}^* + K_2^*)} = o_{\mathbb{P}}(1). \tag{A.5}$$

Step 1-1. We show that $\mathbb{E}^*[(K_2^*)^2] = o_{\mathbb{P}}(1)$. Note that

$$K_2^* = \frac{\sqrt{b\ell}}{2}\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \sin\left(\frac{\pi}{2}\tau_{0,jk}^*\right)\left[\frac{\pi}{2}(\widehat{\tau}_{0,jk}^* - \tau_{0,jk})\right]^2 - \sum_{j\in S_{\boldsymbol{w}}} w_j \sin\left(\frac{\pi}{2}\tau_{1,jm}^*\right)\left[\frac{\pi}{2}(\widehat{\tau}_{1,jm}^* - \tau_{1,jm})\right]^2\right\}.$$

Then, similar to the proof of Lemma A.1.4, it is sufficient to bound $(b\ell)\mathbb{E}[\mathbb{E}^*[(\widehat{\tau}_{0,jk}^* - \tau_{0,jk})^4]$ and $(b\ell)\mathbb{E}[\mathbb{E}^*[(\widehat{\tau}_{1,jm}^* - \tau_{1,jm})^4]]$ uniformly. Because $(a+b)^4 \leq 8a^4 + 8a^4$, we have

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}^*[(\widehat{\tau}_{0,jk}^* - \tau_{0,jk})^4]] \\
=&\mathbb{E}\left[\mathbb{E}^*\left[\left\{\frac{2}{b\ell}\sum_{t=1}^{b\ell} g_1\left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t,k}^*\end{pmatrix}\right) + \frac{2}{b\ell(b\ell-1)}\sum_{t<t'} g_2\left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t,k}^*\end{pmatrix}, \begin{pmatrix} Z_{t',j}^* \\ Z_{t',k}^*\end{pmatrix}\right)\right\}^4\right]\right] \\
\leq&\frac{128}{(b\ell)^4}\mathbb{E}\left[\mathbb{E}^*\left[\left\{\sum_{t=1}^{b\ell} g_1\left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t,k}^*\end{pmatrix}\right)\right\}^4\right]\right] + \frac{128}{(b\ell)^4(b\ell-1)^4}\mathbb{E}\left[\mathbb{E}^*\left[\left\{\sum_{t<t'} g_2\left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t,k}^*\end{pmatrix}, \begin{pmatrix} Z_{t',j}^* \\ Z_{t',k}^*\end{pmatrix}\right)\right\}^4\right]\right] \\
\leq&\frac{128}{(b\ell)^4}\sum_{i_1,i_2,i_3,i_4=1}^{n-1}\left|\mathbb{E}\left[g_1\left(\begin{pmatrix} Z_{i_1,j} \\ Z_{i_1,k}\end{pmatrix}\right) g_1\left(\begin{pmatrix} Z_{i_2,j} \\ Z_{i_2,k}\end{pmatrix}\right) g_1\left(\begin{pmatrix} Z_{i_3,j} \\ Z_{i_3,k}\end{pmatrix}\right) g_1\left(\begin{pmatrix} Z_{i_4,j} \\ Z_{i_4,k}\end{pmatrix}\right)\right]\right| \\
&+\frac{8}{(b\ell)^4(b\ell-1)^4}\sum_{i_1,\ldots,i_8=1}^{n-1}\left|\mathbb{E}\left[g_2\left(\begin{pmatrix} Z_{i_1,j} \\ Z_{i_1,k}\end{pmatrix}, \begin{pmatrix} Z_{i_2,j} \\ Z_{i_2,k}\end{pmatrix}\right) g_2\left(\begin{pmatrix} Z_{i_3,j} \\ Z_{i_3,k}\end{pmatrix}, \begin{pmatrix} Z_{i_4,j} \\ Z_{i_4,k}\end{pmatrix}\right) g_2\left(\begin{pmatrix} Z_{i_5,j} \\ Z_{i_5,k}\end{pmatrix}, \begin{pmatrix} Z_{i_6,j} \\ Z_{i_6,k}\end{pmatrix}\right) g_2\left(\begin{pmatrix} Z_{i_7,j} \\ Z_{i_7,k}\end{pmatrix}, \begin{pmatrix} Z_{i_8,j} \\ Z_{i_8,k}\end{pmatrix}\right)\right]\right| \\
&\leq C(n-1)^{-2},
\end{aligned}$$

where C is some constant which only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, \kappa_1, \kappa_2)$. The last two inequalities come from the proofs of Lemmas 3.7 and 3.8 in Dehling and Wendler (2010) and Lemmas A.1.8 to A.1.10 in the supplementary material. Similarly, one can show that $\mathbb{E}[\mathbb{E}^*[(\widehat{\tau}_{1,jk}^* - \tau_{1,jk})^4]] \leq C'(n-1)^{-2}$, where C' is some constant which only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, \kappa_1, \kappa_2)$. Then, similar to the proof of Lemma A.1.4, we have $\mathbb{E}[\mathbb{E}^*[(K_2^*)^2]] \leq C\|\boldsymbol{w}\|_1^2(\|\boldsymbol{\beta}\|_1^2 + 1)O(n^{-1}) = o(1)$, where C is some constant which only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, \kappa_1, \kappa_2)$. Therefore, we have $\mathbb{E}^*[(K_2^*)^2] = o_{\mathbb{P}}(1)$.

Step 1-2. We show that $\mathrm{Var}^*(K_{12}^*) = o_{\mathbb{P}}(1)$ by proving $\mathbb{E}[\mathbb{E}^*[(K_{12}^*)^2]] = o(1)$. For simplicity, let's denote

$$U^*_{0,jk}(g_2) = \frac{2}{(b\ell)(b\ell-1)} \sum_{t<t'} g_2\left(\begin{pmatrix} Z^*_{t,j} \\ Z^*_{t,k} \end{pmatrix}, \begin{pmatrix} Z^*_{t',j} \\ Z^*_{t',k} \end{pmatrix}\right) \text{ and}$$

$$U^*_{1,jk}(g_2) = \frac{2}{(b\ell)(b\ell-1)} \sum_{t<t'} g_2\left(\begin{pmatrix} Z^*_{t,j} \\ Z^*_{t+1,k} \end{pmatrix}, \begin{pmatrix} Z^*_{t',j} \\ Z^*_{t'+1,k} \end{pmatrix}\right).$$

Then, we can rewrite K^*_{12} as

$$K^*_{12} = \frac{\pi}{2} \sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\sqrt{b\ell}U^*_{0,jk}(g_2) - \frac{\pi}{2}\sum_{j\in S_{\boldsymbol{w}}} w_j \cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\sqrt{b\ell}U^*_{1,jk}(g_2).$$

Because $(a+b)^2 \le 2(a^2+b^2)$, it is sufficient to show

$$\mathbb{E}\left[\mathbb{E}^*\left[\left(\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\sqrt{b\ell}U^*_{0,jk}(g_2)\right)^2\right]\right] = o(1) \text{ and}$$

$$\mathbb{E}\left[\mathbb{E}^*\left[\left(\sum_{j\in S_{\boldsymbol{w}}} w_j \cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\sqrt{b\ell}U^*_{1,jm}(g_2)\right)^2\right]\right] = o(1).$$

From proofs of Lemmas 3.6 and 3.7 in Dehling and Wendler (2010) and Lemma A.1.8 in the supplementary material, we have

$$(b\ell)\mathbb{E}[\mathbb{E}^*[(U^*_{0,jk}(g_2))^2]] \le \frac{C}{(b\ell)(b\ell-1)^2}\sum_{i_1,i_2,i_3,i_4}\left|\mathbb{E}\left[g_2\left(\begin{pmatrix} Z_{i_1,j} \\ Z_{i_1,k} \end{pmatrix}, \begin{pmatrix} Z_{i_2,j} \\ Z_{i_2,k} \end{pmatrix}\right) g_2\left(\begin{pmatrix} Z_{i_3,j} \\ Z_{i_3,k} \end{pmatrix}, \begin{pmatrix} Z_{i_4,j} \\ Z_{i_4,k} \end{pmatrix}\right)\right]\right|$$
$$\le \frac{C'}{(b\ell)(b\ell-1)^2} n^2 \sum_{m=1}^{n} m\alpha_{\boldsymbol{U}}(m)^{2/5} \le C''(n-1)^{-1},$$

where C is an absolute constant, C' only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$, and C'' only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, \kappa_1, \kappa_2)$. Therefore, by doing the same calculation as in Lemma A.1.4, we have

$$\mathbb{E}\left[\mathbb{E}^*\left[\left(\sum_{j\in S_{\boldsymbol{w}},k\in S} w_j\beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right)\sqrt{b\ell}U^*_{0,jk}(g_2)\right)^2\right]\right] \le 2C''\|\boldsymbol{w}\|_1^2\|\boldsymbol{\beta}\|_1^2 O(n^{-1}) = o(1).$$

Using the same approach, we have $\mathbb{E}\left[\mathbb{E}^*\left[\left(\sum_{j\in S_{\boldsymbol{w}}} w_j \cos\left(\frac{\pi}{2}\tau_{1,jm}\right)\sqrt{b\ell}U^*_{1,jm}(g_2)\right)^2\right]\right] = o(1)$. Therefore, we have $\mathbb{E}[\mathbb{E}^*[(K^*_{12})^2]] = o(1)$ and it implies that $\mathrm{Var}^*(K^*_{12}) \le \mathbb{E}^*(K^*_{12}) = o_{\mathbb{P}}(1)$.

Step 1-3. We will show $\mathrm{Var}^*(K_{11}^*) - \mathbb{E}^*[K_{11}^2] = o_{\mathbb{P}}(1)$. Define

$$G_{n1}(\boldsymbol{U}_t) = \left\{ \sum_{j \in S_{\boldsymbol{w}}, k \in S} w_j \beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right) \pi g_1 \left(\begin{pmatrix} Z_{t,j} \\ Z_{t,k} \end{pmatrix} \right) - \sum_{j \in S_{\boldsymbol{w}}} w_j \cos\left(\frac{\pi}{2}\tau_{1,jm}\right) \pi g_1 \left(\begin{pmatrix} Z_{t,j} \\ Z_{t+1,m} \end{pmatrix} \right) \right\},$$

$$G_{n1}^*(\boldsymbol{U}_t^*) = \left\{ \sum_{j \in S_{\boldsymbol{w}}, k \in S} w_j \beta_k \cos\left(\frac{\pi}{2}\tau_{0,jk}\right) \pi g_1 \left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t,k}^* \end{pmatrix} \right) - \sum_{j \in S_{\boldsymbol{w}}} w_j \cos\left(\frac{\pi}{2}\tau_{1,jm}\right) \pi g_1 \left(\begin{pmatrix} Z_{t,j}^* \\ Z_{t+1,m}^* \end{pmatrix} \right) \right\},$$

where $\boldsymbol{U}_t^* = \boldsymbol{f}^{-1}(\boldsymbol{Y}_t^*)$. Then, we can rewrite K_{11} and K_{11}^* as $K_{11} = \frac{1}{\sqrt{n-1}} \sum_{t=1}^{n-1} G_{n1}(\boldsymbol{U}_t)$ and $K_{11}^* = \frac{1}{\sqrt{b\ell}} \sum_{t=1}^{b\ell} G_{n1}^*(\boldsymbol{U}_t^*)$. Note that $\{G_{n1}(\boldsymbol{U}_t)\}_{t \in \mathbb{Z}}$ is a triangular array with geometrically α-mixing process because $\{\boldsymbol{U}_t\}_{t \in \mathbb{Z}}$ is a geometrically α-mixing process and $G_{n1}(\cdot)$ is continuous non-random function. Then, we can show the consistency of bootstrap version of variance following the proofs of Theorem 2.1 in Gonçalves and White (2002) and Theorem 3.1 in Lahiri (2003).

Let $\hat{\sigma}_{n,\mathrm{MBB}}^{2*}$ be the bootstrap variance of K_{11}^* when $G_{n1}^*(U_t^*)$ is constructed from the moving block bootstrap (MBB) sample, not the circular block bootstrap (CBB) sample. Let $\hat{\sigma}_{n,\mathrm{CBB}}^{2*} := \mathrm{Var}^*(K_{11}^*)$ be the bootstrap variance of K_{11}^* when $G_{n1}^*(U_t^*)$ is constructed from the circular block bootstrap (CBB) sample. We refer to Chapter 3 of Lahiri (2003) for definitions of MBB and CBB. Then, we finish Step 1-3 by doing the following procedure.

Step 1-3-1. We show that $\widehat{\sigma}_{n,\mathrm{MBB}}^{2*} - \widehat{\sigma}_{n,\mathrm{CBB}}^{2*} = o_{\mathbb{P}}(1)$ by following the proof of Theorem 3.1 in Lahiri (2003) even though we are dealing with the triangular array because $G_{n1}(\cdot)$ is bounded above by some absolute constant under the assumptions that $M = O(1)$ and $\|\boldsymbol{\Sigma}_0^{-1}\|_\infty = O(1)$.

Step 1-3-2. Because an α-mixing process is the special case of an NED process, we can show $\widehat{\sigma}_{n,\mathrm{MBB}}^{2*} - \mathbb{E}[K_{11}^2] = o_{\mathbb{P}}(1)$ by following the proof of Theorem 2.1 in Gonçalves and White (2002).

We skip the proofs of Steps 1-3-1 and 1-3-2 because it can be easily verified by following the proofs of Theorem 3.1 in Lahiri (2003) and Theorem 2.1 in Gonçalves and White (2002) step by step. In conclusion, the proof of Lemma A.1.4 and Steps 1-1 to 1-3 imply that $(b\ell)\mathrm{Var}^*(\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^* \boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)) - \sigma_n^2 = o_{\mathbb{P}}(1)$.

Step 2. We will show that $\widehat{\sigma}_n^{2*} - (b\ell)\mathrm{Var}^*(\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)) = o_{\mathbb{P}}(1)$. Note that

$$\sqrt{b\ell}\widehat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\widehat{\boldsymbol{\beta}}(\theta) - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*) - \sqrt{b\ell}\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*) = \sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*) + \sqrt{b\ell}\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Sigma}}_0^*(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}).$$

Steps 2-1 and 2-2 below combined will show that

$$\mathbb{E}^*[|\sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)|^2] = o_{\mathbb{P}}(1) \text{ and } \mathbb{E}^*[|\sqrt{b\ell}\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Sigma}}_0^*(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta})|^2] = o_{\mathbb{P}}(1).$$

Note that Steps 2-1 and 2-2 imply $\widehat{\sigma}_n^{2*} - (b\ell)\mathrm{Var}^*(\boldsymbol{w}^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)) = o_{\mathbb{P}}(1)$ by the similar reasoning as in Equation (A.5) in Step 1.

Step 2-1. We show that $\mathbb{E}^*[|\sqrt{b\ell}(\widehat{\boldsymbol{w}}^\top - \boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*)|^2] = o_{\mathbb{P}}(1)$.

Because $\boldsymbol{\Sigma}_0\boldsymbol{\beta} - \boldsymbol{\Sigma}_{1,*m} = \mathbf{0}$, we have

$$\sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}_0^*\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}^*) = \sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top((\widehat{\boldsymbol{\Sigma}}_0^* - \boldsymbol{\Sigma}_0)\boldsymbol{\beta} - (\widehat{\boldsymbol{\Sigma}}_{1,*m}^* - \boldsymbol{\Sigma}_{1,*m})).$$

Therefore, it is sufficient to show that $\mathbb{E}^*[|\sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}_0^* - \boldsymbol{\Sigma}_0)\boldsymbol{\beta}|^2] = o_{\mathbb{P}}(1)$ and $\mathbb{E}^*[|\sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}_{1,*m}^* - \boldsymbol{\Sigma}_{1,*m})|^2] = o_{\mathbb{P}}(1)$. Because $|\sin(x) - \sin(y)| \leq |x - y|$ for any $x, y \in \mathbb{R}$, and

$$|\widehat{\tau}_{0,jk}^* - \tau_{0,jk}| \leq \frac{b\ell}{(b\ell - 1)}|V_{0,jk}^* - V_{0,jk}| + \frac{1}{(b\ell - 1)}|V_{0,jk}| \leq \frac{b\ell}{(b\ell - 1)}|V_{0,jk}^* - V_{0,jk}| + \frac{1}{(b\ell - 1)},$$

we have

$$\begin{aligned}
&|\sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}_0^* - \boldsymbol{\Sigma}_0)\boldsymbol{\beta}| \\
=&\left|\sqrt{b\ell}\sum_{j\in S_{\boldsymbol{w}}, k\in S}(\widehat{w}_j - w_j)\left(\sin\left(\frac{\pi}{2}\widehat{\tau}_{0,jk}^*\right) - \sin\left(\frac{\pi}{2}\tau_{0,jk}\right)\right)\beta_k\right| \\
\leq&\frac{\pi}{2}\sqrt{b\ell}\sum_{j\in S_{\boldsymbol{w}}, k\in S}|\widehat{w}_j - w_j||\beta_w||\widehat{\tau}_{0,jk}^* - \tau_{0,jk}| \\
\leq&\frac{\pi}{2}\sqrt{b\ell}\frac{b\ell}{(b\ell - 1)}\sum_{j\in S_{\boldsymbol{w}}, k\in S}|\widehat{w}_j - w_j||\beta_k||V_{0,jk}^* - V_{0,jk}| + \frac{\pi}{2}\sqrt{b\ell}\frac{1}{(b\ell - 1)}\sum_{j\in S_{\boldsymbol{w}}, k\in S}|\widehat{w}_j - w_j||\beta_k| \\
\leq&\frac{\pi}{2}\sqrt{b\ell}\frac{b\ell}{(b\ell - 1)}\sum_{j\in S_{\boldsymbol{w}}, k\in S}|\widehat{w}_j - w_j||\beta_k||V_{0,jk}^* - V_{0,jk}| + \frac{\pi}{2}\sqrt{b\ell}\frac{1}{(b\ell - 1)}\|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_1\|\boldsymbol{\beta}\|_1.
\end{aligned}$$

This implies

$$
\begin{aligned}
&\mathbb{E}^*[|\sqrt{b\ell}(\widehat{\boldsymbol{w}}-\boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}_0^*-\boldsymbol{\Sigma}_0)\boldsymbol{\beta}|^2]\\
\leq&\frac{\pi^2}{2}(b\ell)\left(\frac{b\ell}{(b\ell-1)}\right)^2\mathbb{E}^*\left[\left(\sum_{j\in S_{\boldsymbol{w}},k\in S}|\widehat{\boldsymbol{w}}_j-\boldsymbol{w}_j||\boldsymbol{\beta}_k||V_{0,jk}^*-V_{0,jk}|\right)^2\right]+\frac{\pi^2}{2}(b\ell)\frac{1}{(b\ell-1)^2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\|_1^2\|\boldsymbol{\beta}\|_1^2.
\end{aligned}
$$

First, we have $(b\ell)\frac{1}{(b\ell-1)^2}\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\|_1^2\|\boldsymbol{\beta}\|_1^2=\|\boldsymbol{\beta}\|_1^2O_{\mathbb{P}}\left(\frac{s_{\boldsymbol{w}}^2\log(ed)}{n^2}\right)=o_{\mathbb{P}}(1)$. Second, by doing a similar calculation as in Step 1 in Lemma A.1.4, one has

$$
\begin{aligned}
&(b\ell)\left(\frac{b\ell}{(b\ell-1)}\right)^2\mathbb{E}^*\left[\left(\sum_{j\in S_{\boldsymbol{w}},k\in S}|\widehat{\boldsymbol{w}}_j-\boldsymbol{w}_j||\boldsymbol{\beta}_k||V_{0,jk}^*-V_{0,jk}|\right)^2\right]\\
\leq&(b\ell)\left(\frac{b\ell}{(b\ell-1)}\right)^2\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\|_1\|\boldsymbol{\beta}\|_1\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S}|\widehat{\boldsymbol{w}}_j-\boldsymbol{w}_j||\boldsymbol{\beta}_k|\mathbb{E}^*[(V_{0,jk}^*-V_{0,jk})^2]\right\}.
\end{aligned}
$$

Note that Lemma A.1.11 in the supplementary material implies

$$
\left|\mathbb{E}^*[(V_{0,jk}^*-V_{0,jk})^2]-\frac{b\ell(b\ell-\ell)(b\ell-2\ell)(b\ell-3\ell)}{(b\ell)^4}(\widehat{V}_{0,jk}-V_{0,jk})^2\right|\leq\frac{50\ell}{b\ell}.
$$

Therefore, by combining the above three equations, we have

$$
\begin{aligned}
&\mathbb{E}^*[|\sqrt{b\ell}(\widehat{\boldsymbol{w}}-\boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}_0^*-\boldsymbol{\Sigma}_0)\boldsymbol{\beta}|^2]\\
\leq&(b\ell)\left(\frac{b\ell}{(b\ell-1)}\right)^2\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\|_1\|\boldsymbol{\beta}\|_1\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S}|\widehat{\boldsymbol{w}}_j-\boldsymbol{w}_j||\boldsymbol{\beta}_k|\mathbb{E}^*[(V_{0,jk}^*-V_{0,jk})^2]\right\}+o_{\mathbb{P}}(1)\\
\leq&(b\ell)\left(\frac{b\ell}{(b\ell-1)}\right)^2\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\|_1\|\boldsymbol{\beta}\|_1\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S}|\widehat{\boldsymbol{w}}_j-\boldsymbol{w}_j||\boldsymbol{\beta}_k|\frac{b(b-1)(b-2)(b-3)\ell^4}{(b\ell)^4}(\widehat{V}_{0,jk}-V_{0,jk})^2\right\}\\
&+(b\ell)\left(\frac{b\ell}{(b\ell-1)}\right)^2\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\|_1\|\boldsymbol{\beta}\|_1\left\{\sum_{j\in S_{\boldsymbol{w}},k\in S}|\widehat{\boldsymbol{w}}_j-\boldsymbol{w}_j||\boldsymbol{\beta}_k|50\frac{\ell}{b\ell}\right\}+o_{\mathbb{P}}(1)\\
\leq&\frac{b(b-1)(b-2)(b-3)\ell^4}{(b\ell)^4}\left(\frac{b\ell}{(b\ell-1)}\right)^2(b\ell)\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\|_1^2\|\boldsymbol{\beta}\|_1^2(\max_{j,k}\{|\widehat{V}_{0,jk}-V_{0,jk}|\})^2\\
&+50\left(\frac{b\ell}{(b\ell-1)}\right)^2\ell\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\|_1^2\|\boldsymbol{\beta}\|_1^2+o_{\mathbb{P}}(1)\\
=&O(n)O_{\mathbb{P}}\left(\frac{s_{\boldsymbol{w}}^2\log(ed)}{n}\right)\|\boldsymbol{\beta}\|_1^2O_{\mathbb{P}}\left(\frac{\log(ed)}{n}\right)+O(\sqrt{n})O_{\mathbb{P}}\left(\frac{s_{\boldsymbol{w}}^2\log(ed)}{n}\right)\|\boldsymbol{\beta}\|_1^2+o_{\mathbb{P}}(1)\\
=&\|\boldsymbol{\beta}\|_1^2O_{\mathbb{P}}\left(\frac{s_{\boldsymbol{w}}^2(\log(ed))^2}{n}\right)+\|\boldsymbol{\beta}\|_1^2O_{\mathbb{P}}\left(\frac{s_{\boldsymbol{w}}^2\log(ed)}{\sqrt{n}}\right)+o_{\mathbb{P}}(1)=o_{\mathbb{P}}(1),
\end{aligned}
$$

because $b\ell \asymp n$ and $\ell^{-1}+\ell^2/n = o(1)$ imply $b(b-1)(b-2)(b-3)\ell^4/(b\ell)^4 = O(1)$, $\ell/\sqrt{n} = o(1)$, and $\left(\frac{b\ell}{(b\ell-1)}\right)^2 = O(1)$. Similarly, we can show $\mathbb{E}^*[(\sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}^*_{1,*m} - \boldsymbol{\Sigma}_{1,*m}))^2] = o_{\mathbb{P}}(1)$. Therefore, we have $\mathbb{E}^*[|\sqrt{b\ell}(\widehat{\boldsymbol{w}}^\top - \boldsymbol{w})^\top(\widehat{\boldsymbol{\Sigma}}^*_0\boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}^*_{1,*m})|^2] = o_{\mathbb{P}}(1)$.

Step 2-2. We show $\mathbb{E}^*[|\sqrt{b\ell}\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Sigma}}^*_0(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta})|^2] = o_{\mathbb{P}}(1)$.

Because $\boldsymbol{w}^\top\boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}) = 0$, we have

$$\begin{aligned}\sqrt{b\ell}\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Sigma}}^*_0(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}) &= \sqrt{b\ell}(\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{\Sigma}}^*_0 - \boldsymbol{w}^\top\boldsymbol{\Sigma}_0)(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}) \\ &= \sqrt{b\ell}\widehat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Sigma}}^*_0 - \boldsymbol{\Sigma}_0)(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}) + \sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top\boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}).\end{aligned}$$

Then, Steps 2-2-1 and 2-2-2 below combined will show $\mathbb{E}^*[(\sqrt{b\ell}\widehat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Sigma}}^*_0 - \boldsymbol{\Sigma}_0)(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}))^2] = o_{\mathbb{P}}(1)$ and $E^*[(\sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top\boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}))^2] = o_{\mathbb{P}}(1)$.

Step 2-2-1. We show that $\mathbb{E}^*[(\sqrt{b\ell}\widehat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Sigma}}^*_0 - \boldsymbol{\Sigma}_0)(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}))^2] = o_{\mathbb{P}}(1)$. We can do the same calculation as in Step 2-1 and obtain

$$\begin{aligned}&\mathbb{E}^*[(\sqrt{b\ell}\widehat{\boldsymbol{w}}^\top(\widehat{\boldsymbol{\Sigma}}^*_0 - \boldsymbol{\Sigma}_0)(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}))^2] \\ \le&\frac{b(b-1)(b-2)(b-3)\ell^4}{(b\ell)^4}\left(\frac{b\ell}{(b\ell-1)}\right)^2(b\ell)\|\widehat{\boldsymbol{w}}\|_1^2\|\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}\|_1^2(\max_{j,k}\{|\widehat{V}_{0,jk} - V_{0,jk}|\})^2 \\ &+ 50\left(\frac{b\ell}{(b\ell-1)}\right)^2\ell\|\widehat{\boldsymbol{w}}\|_1^2\|\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}\|_1^2 + o_{\mathbb{P}}(1) \\ =&O(1)O(n)\left(\|\boldsymbol{w}\|_1^2 + O_{\mathbb{P}}\left(\frac{s_{\boldsymbol{w}}^2\log(ed)}{n}\right)\right)O_{\mathbb{P}}\left(\frac{s^2\log(ed)}{n}\right)O_{\mathbb{P}}\left(\frac{\log(ed)}{n}\right) \\ &+ O(1)O(\sqrt{n})\left(\|\boldsymbol{w}\|_1^2 + O_{\mathbb{P}}\left(\frac{s_{\boldsymbol{w}}^2\log(ed)}{n}\right)\right)O_{\mathbb{P}}\left(\frac{s^2\log(ed)}{n}\right) + o_{\mathbb{P}}(1) \\ =&o_{\mathbb{P}}(1).\end{aligned}$$

Step 2-2-2. We show that $\mathbb{E}^*[(\sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top\boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}))^2] = o_{\mathbb{P}}(1)$. This holds because $|\mathbb{E}^*[(\sqrt{b\ell}(\widehat{\boldsymbol{w}} - \boldsymbol{w})^\top\boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}))^2]| \le (b\ell)\|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_1^2\|(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta})\|_1^2 = O_{\mathbb{P}}\left(\frac{s_{\boldsymbol{w}}^2 s^2(\log(ed))^2}{n}\right) = o_{\mathbb{P}}(1)$.

By combining Steps 1 and 2, we have the desired result.

Technical Lemmas used in Section 1.3

The following lemmas apply Dehling and Wendler (2010) to our case. In the following we use $\mathbb{1}_A$ as a shorthand of $\mathbb{1}(x \in A)$.

Lemma A.1.5 (Theorem 2 in Bradley (1983)). *Suppose $\boldsymbol{R}$ and $\boldsymbol{W}$ are random variables taking their values in Borel spaces S_1 and S_2 respectively, and suppose U is a uniform-$[0,1]$ random variable independent of $(\boldsymbol{R}, \boldsymbol{W})$. Suppose N is a positive integer and $H = \{H_1, H_2, \ldots, H_N\}$ is a measurable partition of S_2. Then, there exists an S_2-valued random variable $\breve{\boldsymbol{W}} = f(\boldsymbol{R}, \boldsymbol{W}, U)$, where f is a measurable function from $S_1 \times S_2 \times [0,1]$ into S_2, such that*

1. *$\breve{\boldsymbol{W}}$ is independent of $\boldsymbol{R}$;*
2. *the probability distributions of $\breve{\boldsymbol{W}}$ and $\boldsymbol{W}$ on S_2 are identical;*
3. *$\mathbb{P}(\breve{\boldsymbol{W}}$ and $\boldsymbol{W}$ are not elements of the same $H_i \in H) \leq (8N)^{1/2}\alpha(\sigma(\boldsymbol{R}), \sigma(\boldsymbol{W}))$.*

The following is Bradley's lemma with $S_2 = \mathbb{R}^2$ by following the proof of Theorem 3 in Bradley (1983), which presents Bradley's lemma when $S_2 = \mathbb{R}$.

Lemma A.1.6. *Suppose $\boldsymbol{R}$ and $\boldsymbol{W}$ are random variables and $\boldsymbol{W} = (W_1, W_2)^\top \in \mathbb{R}^2$. Let's denote $M = \max\{\|W_1\|_\gamma, \|W_2\|_\gamma\}$. Then there exists a random variable $\breve{\boldsymbol{W}}$ such that*

1. *$\breve{\boldsymbol{W}}$ is independent of $\boldsymbol{R}$;*
2. *the probability distributions of $\breve{\boldsymbol{W}}$ and $\boldsymbol{W}$ on $\mathbb{R}^2$ are identical;*
3. *For any $0 < q \leq M$, $\mathbb{P}(\|\boldsymbol{W} - \breve{\boldsymbol{W}}\|_\infty \geq q) \leq 50\left(\frac{M}{q}\right)^{\frac{\gamma}{\gamma+1}} \alpha(\sigma(\boldsymbol{R}), \sigma(\boldsymbol{W}))^{\frac{\gamma}{\gamma+1}}$.*

Proof. Following the proof of Theorem 3 in Bradley (1983), let $\alpha = \alpha(\sigma(\boldsymbol{R}), \sigma(\boldsymbol{W}))$, $M = \max\{\|W_1\|_\gamma, \|W_2\|_\gamma\}$, and $x = [(1/\alpha)(M/q)^\gamma, \max\{\gamma, 1\}]^{1/(\gamma+1)}$. Note that $x \geq 1$ by construction. Further let m be such that $x \leq m \leq 2x$. We construct the rectangular H_{ij} for $-m \leq i, j \leq m$ and $H_{m+1,m+1}$ as follows: $H_{ij} = [qi - q/2, qi + q/2) \times [qj - q/2, qj + q/2)$ for $-m \leq i, j \leq m$, and $H_{m+1,m+1} = \left(\bigcup_{-m \leq i,j \leq m} H_{ij}\right)^c$. Then, we can set H as the union of $H_{m+1,m+1}$ and $\bigcup_{-m \leq i,j \leq m} H_{ij}$. From Lemma A.1.5 , we have that the random variable $\breve{\boldsymbol{W}} = (\breve{W}_1, \breve{W}_2)^\top$ is independent of $\boldsymbol{R}$, has the same distribution as $\boldsymbol{W}$, and

$$\mathbb{P}(\breve{\boldsymbol{W}} \text{ and } \boldsymbol{W} \text{ are not elements of the same } H_{ij} \in H) \leq 4((2m+1)^2 + 1)^{1/2}\alpha.$$

Therefore,

$$
\begin{aligned}
\mathbb{P}(\|\boldsymbol{W}-\breve{\boldsymbol{W}}\|_\infty \geq q) &\leq 4((2m+1)^2+1)^{1/2}\alpha + P(|W_1| \geq mq + q/2) + \mathbb{P}(|W_2| \geq mq+q/2) \\
&\quad + \mathbb{P}(|\breve{W}_1| \geq mq+q/2) + \mathbb{P}(|\breve{W}_2| \geq mq+q/2) \\
&= 4((2m+1)^2+1)^{1/2}\alpha + 2\mathbb{P}(|W_1| \geq mq+q/2) + 2\mathbb{P}(|W_2| \geq mq+q/2) \\
&\leq 24x\alpha + 2\mathbb{P}(|W_1| \geq xq) + 2\mathbb{P}(|W_2| \geq xq) \\
&\leq 24x\alpha + 2\left(\frac{\|W_1\|_\gamma}{q}\right)^\gamma x^{-\gamma} + 2\left(\frac{\|W_2\|_\gamma}{q}\right)^\gamma x^{-\gamma} \\
&\leq 24x\alpha + 4\left(\frac{M}{q}\right)^\gamma x^{-\gamma} \leq C_\gamma \left(\frac{M}{q}\right)^{\frac{\gamma}{\gamma+1}} \alpha^{\frac{\gamma}{\gamma+1}} \leq 50\left(\frac{M}{q}\right)^{\frac{\gamma}{\gamma+1}} \alpha^{\frac{\gamma}{\gamma+1}},
\end{aligned}
$$

where $C_\gamma := 2 + 24[\max\{\gamma, 1\}]^{1/(\gamma+1)}$. The last inequality comes from the last sentence of the proof of Theorem 3 in Bradley (1983). □

Lemma A.1.7 ($\mathbb{P}$-Lipschitz continuity of $f(\cdot,\cdot)$)**.** *Suppose all assumptions in Proposition 1.2.2 hold. Then, there exists a constant L only depending on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$ such that*

$$
\mathbb{E}[|f(\boldsymbol{W},\boldsymbol{R}) - f(\boldsymbol{W}',\boldsymbol{R})|\mathbb{1}_{\|\boldsymbol{W}-\boldsymbol{W}'\|_\infty \leq \epsilon}] \leq L\epsilon
$$

for every $\epsilon > 0$, every pair $(\boldsymbol{W}^\top, \boldsymbol{R}^\top)^\top$ with the common distribution $\mathbb{P}_{\boldsymbol{U}_{1,\{j,k\}},\boldsymbol{U}_{t,\{j,k\}}}$ for a $t \in \mathbb{N}$ or the product measure $\mathbb{P}_{\boldsymbol{U}_{1,jk}} \times \mathbb{P}_{\boldsymbol{U}_{1,jk}}$, and $(\boldsymbol{W}'^\top, \boldsymbol{R}^\top)^\top$ also with one of these common distributions. Here, $\boldsymbol{W}$, $\boldsymbol{W}'$, and $\boldsymbol{R}$ are in $\mathbb{R}^2$.

Proof. Because $|\mathrm{sign}(\cdot)| \leq 1$, we have

$$
|f(\boldsymbol{W},\boldsymbol{R}) - f(\boldsymbol{W}',\boldsymbol{R})| \leq |\mathrm{sign}(W_1 - R_1) - \mathrm{sign}(W_1' - R_1)| + |\mathrm{sign}(W_2 - R_2) - \mathrm{sign}(W_2' - R_2)|.
$$

We then have

$$
\begin{aligned}
&\mathbb{E}[|f(\boldsymbol{W},\boldsymbol{R}) - f(\boldsymbol{W}',\boldsymbol{R})|\mathbb{1}_{\|\boldsymbol{W}-\boldsymbol{W}'\|_\infty \leq \epsilon}] \\
\leq&\mathbb{E}[|\mathrm{sign}(W_1 - R_1) - \mathrm{sign}(W_1' - R_1)|\mathbb{1}_{\|\boldsymbol{W}-\boldsymbol{W}'\|_\infty \leq \epsilon}] + \mathbb{E}[|\mathrm{sign}(W_2 - R_2) - \mathrm{sign}(W_2' - R_2)|\mathbb{1}_{\|\boldsymbol{W}-\boldsymbol{W}'\|_\infty \leq \epsilon}] \\
\leq&\mathbb{E}[|\mathrm{sign}(W_1 - R_1) - \mathrm{sign}(W_1' - R_1)|\mathbb{1}_{|W_1 - W_1'| \leq \epsilon}] + \mathbb{E}[|\mathrm{sign}(W_2 - R_2) - \mathrm{sign}(W_2' - R_2)|\mathbb{1}_{|W_2 - R_2'| \leq \epsilon}] \\
\leq&2\mathbb{P}(|W_1 - R_1| \leq \epsilon) + 2\mathbb{P}(|W_2 - R_2| \leq \epsilon).
\end{aligned}
$$

The last inequality holds because $|x - x'| \leq \epsilon$ and $|x - y| > \epsilon$ imply $|\text{sign}(x - y) - \text{sign}(x' - y)| = 0$ for any real numbers x, x', and y . Under the assumptions in Proposition 1.2.2, $\mathbb{P}(|W_1 - R_1| \leq \epsilon) \leq D_1\epsilon$ for some absolute $D_1 > 0$. This is because of the following reasons:

- Note that the density of $W_1 - R_1$ is the density of $U_{1,j} - U_{t,j}$. When the distribution of $(\boldsymbol{W}^\top, \boldsymbol{R}^\top)^\top$ is from $\mathbb{P}_{\boldsymbol{U}_{1,\{j,k\}}} \times \mathbb{P}_{\boldsymbol{U}_{1,\{j,k\}}}$, W_1 and R_1 are independent of each other, so that the density function of $W_1 - R_1$ is bounded above under the assumptions. Therefore, let's consider the case when $(\boldsymbol{W}^\top, \boldsymbol{R}^\top)^\top$ is from $\mathbb{P}_{\boldsymbol{U}_{1,\{j,k\}}, \boldsymbol{U}_{t,\{j,k\}}}$;

- Because $\{\boldsymbol{Z}_t\}_{t=1}^n$ are jointly Gaussian, the maximum of the values of the density functions of $Z_{1,j} - Z_{t,j}$ is determined by the variance of $Z_{1,j} - Z_{t,j}$;

- Under the assumptions in Proposition 1.2.2, we can show that the variance of $Z_{1,j} - Z_{t,j}$ is uniformly bounded below using Lemma A.1.3. This implies that the density functions are bounded above by a constant D_1 which only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$;

- Therefore, $\mathbb{P}(|W_1 - R_1| \leq \epsilon) \leq D_1\epsilon$.

By the same logic, we have $\mathbb{P}(|W_2 - R_2| \leq \epsilon) \leq D_1\epsilon$. Therefore,

$$\mathbb{E}[|f(\boldsymbol{W}, \boldsymbol{R}) - f(\boldsymbol{W}', \boldsymbol{R})| \mathbb{1}_{\|\boldsymbol{W} - \boldsymbol{W}'\|_\infty \leq \epsilon}] \leq 2\mathbb{P}(|W_1 - R_1| \leq \epsilon) + 2\mathbb{P}(|W_2 - R_2| \leq \epsilon) \leq 4D_1\epsilon = L\epsilon,$$

where $L := 4D_1$. That is, $f(\cdot, \cdot)$ is $\mathbb{P}$ -Lipschitz-continuous with constant $L = 4D_1$. □

Note A.1.1. The proof of Lemma 3.3 in Dehling and Wendler (2010) implies that $g_1(\cdot)$ is $\mathbb{P}$-Lipschitz continuous with constant L and $g_2(\cdot, \cdot)$ is also $\mathbb{P}$-Lipschitz-continuous with constant $2L$.

In the following, we will use $\boldsymbol{W}_t$ to represent $(U_{t,j}, U_{t,k})^\top$, and $\alpha_{\boldsymbol{U}}(\cdot)$ as the α-mixing coefficient of the process $\{\boldsymbol{U}_t = (\boldsymbol{Z}_t^\top, \boldsymbol{Z}_{t+1}^\top)^\top\}_{t \in \mathbb{Z}}$.

The second lemma is to check if Lemma 3.3 in Dehling and Wendler (2010) holds.

Lemma A.1.8. *Suppose all assumptions in Proposition 1.2.2 hold. Then, for some constants C_8 and C_9 only depending on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$, we have*

$$|\mathbb{E}[g_1(\boldsymbol{W}_{i_1})g_1(\boldsymbol{W}_{i_2})g_1(\boldsymbol{W}_{i_3})g_1(\boldsymbol{W}_{i_4})]| \leq C_8\alpha_{\boldsymbol{U}}(m)^{2/5}, \text{ and } \mathbb{E}[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2})g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4})]| \leq C_9\alpha_{\boldsymbol{U}}(m)^{2/5},$$

where $m := \max\{a_2 - a_1, a_4 - a_3\}$ and $a_1, \ldots, a_4$ are the increasingly rearranged numbers of $i_1, \ldots, i_4$.

Proof. The proof follows the proof of Lemma 3.3 in Dehling and Wendler (2010) using Lemma A.1.6. In this proof, we will only look at the case where $i_1 < i_2 \leq i_3 < i_4$ with $m_1 = i_2 - i_1 = \max\{i_2 - i_1, i_4 - i_3\}$. Other cases could be handled similarly.

By Lemma A.1.6, there exists a random vector $\breve{\boldsymbol{W}}_{i_1} \in \mathbb{R}^2$ such that (a) $\breve{\boldsymbol{W}}_{i_1}$ has the same distribution as $\boldsymbol{W}_{i_1}$, (b) $\breve{\boldsymbol{W}}_{i_1}$ is independent of $\boldsymbol{W}_{i_2} \ldots \boldsymbol{W}_{i_4}$, and (c) when $0 \leq \epsilon \leq 1$, we have $\mathbb{P}(\|\boldsymbol{W}_{i_1} - \breve{\boldsymbol{W}}_{i_1}\|_\infty \geq \epsilon) \leq 50\epsilon^{-2/3}\alpha_{\boldsymbol{U}}(m)^{2/3}$. Then, $\mathbb{E}[g_1(\breve{\boldsymbol{W}}_{i_1})g_1(\boldsymbol{W}_{i_2})g_1(\boldsymbol{W}_{i_3})g_1(\boldsymbol{W}_{i_4})] = 0$. Because $|g_1(\cdot)| \leq 2$ and Lemma A.1.7 implies $g_1(\cdot)$ is $\mathbb{P}$-Lipschitz continuous with L, we have

$$\begin{aligned}
&|\mathbb{E}[g_1(\boldsymbol{W}_{i_1})g_1(\boldsymbol{W}_{i_2})g_1(\boldsymbol{W}_{i_3})g_1(\boldsymbol{W}_{i_4})]| \\
=&\left|\mathbb{E}\left[[g_1(\boldsymbol{W}_{i_1}) - g_1(\breve{\boldsymbol{W}}_{i_1})]g_1(\boldsymbol{W}_{i_2})g_1(\boldsymbol{W}_{i_3})g_1(\boldsymbol{W}_{i_4})\right]\right| \\
\leq&8\mathbb{E}[|g_1(\boldsymbol{W}_{i_1}) - g_1(\breve{\boldsymbol{W}}_{i_1})|] \\
=&8\mathbb{E}[|g_1(\boldsymbol{W}_{i_1}) - g_1(\breve{\boldsymbol{W}}_{i_1})|\mathbb{1}_{\|\boldsymbol{W}_{i_1} - \breve{\boldsymbol{W}}_{i_1}\|_\infty \geq \epsilon}] + 8\mathbb{E}[|g_1(\boldsymbol{W}_{i_1}) - g_1(\breve{\boldsymbol{W}}_{i_1})|\mathbb{1}_{\|\boldsymbol{W}_{i_1} - \breve{\boldsymbol{W}}_{i_1}\|_\infty \leq \epsilon}] \\
\leq&32\mathbb{P}(\|\boldsymbol{W}_{i_1} - \breve{\boldsymbol{W}}_{i_1}\|_\infty \geq \epsilon) + 8L\epsilon \leq 1600\epsilon^{-2/3}\alpha_{\boldsymbol{U}}(m)^{2/3} + 8\widetilde{L}\epsilon,
\end{aligned}$$

where $\widetilde{L} := \max\{L, 1\}$. Picking $\epsilon = \widetilde{L}^{-3/5}\alpha_{\boldsymbol{U}}(m)^{2/5}$, we have

$$|\mathbb{E}[g_1(\boldsymbol{W}_{i_1})g_1(\boldsymbol{W}_{i_2})g_1(\boldsymbol{W}_{i_3})g_1(\boldsymbol{W}_{i_4})]| \leq C_8\widetilde{L}^{2/5}\alpha_{\boldsymbol{U}}(m)^{2/5},$$

where C_8 is some absolute constant. Using the same approach, one could obtain

$$|\mathbb{E}[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2})g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4})]| \leq C_9\widetilde{L}^{2/5}\alpha_{\boldsymbol{U}}(m)^{2/5},$$

where C_9 is some absolute constant. These conclude the proof. ∎

Lemma A.1.9. *Suppose all assumptions in Proposition 1.2.2 hold. Suppose that among* $i_1, i_2, \ldots, i_8$*, at most two of them are identical. Then, for some constant* C *which only depends on* $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$*, we have* $|\mathbb{E}[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2})g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4})g_2(\boldsymbol{W}_{i_5}, \boldsymbol{W}_{i_6})g_2(\boldsymbol{W}_{i_7}, \boldsymbol{W}_{i_8})]| \leq C\alpha_{\boldsymbol{U}}(m)^{2/5}$*, where* $m := \max\{a_2 - a_1, a_8 - a_7\}$ *and* $a_1, \ldots, a_8$ *are the increasingly rearranged numbers of* $i_1, \ldots, i_8$*.*

Proof. The proof is the same as Lemma A.1.8, and Lemma 3.3 in Dehling and Wendler (2010).

Note that $m \geq 1$ because we assume that at most two of i's are identical.

Without loss of generality, we can assume that $a_1 = i_1$; $a_8 = i_8$; $i_1 \leq i_2$; $i_3 \leq i_4$; $i_5 \leq i_6$; and $i_7 \leq i_8$ because $g_2(\boldsymbol{W}_1, \boldsymbol{W}_2) = g_2(\boldsymbol{W}_2, \boldsymbol{W}_1)$ for any generic vectors $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$.

Suppose $m = a_2 - a_1 \geq a_8 - a_7$. From Lemma A.1.6, we have $\breve{\boldsymbol{W}}_{i_1} \in \mathbb{R}^2$ such that (a) $\breve{\boldsymbol{W}}_{i_1}$ has the same distribution as $\boldsymbol{W}_{i_1}$; (b) $\breve{\boldsymbol{W}}_{i_1}$ is independent of $\boldsymbol{W}_{i_2} \ldots \boldsymbol{W}_{i_8}$; and (c) when $0 \leq \epsilon \leq 1$, $\mathbb{P}(\|\boldsymbol{W}_{i_1} - \breve{\boldsymbol{W}}_{i_1}\|_\infty \geq \epsilon) \leq 50\epsilon^{-2/3}\alpha_{\boldsymbol{U}}(m)^{2/3}$. Then, we have $\mathbb{E}[g_2(\breve{\boldsymbol{W}}_{i_1}, \boldsymbol{W}_{i_2})g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4})g_2(\boldsymbol{W}_{i_5}, \boldsymbol{W}_{i_6})g_2(\boldsymbol{W}_{i_7}, \boldsymbol{W}_{i_8})] = 0$. Because $|g_2(\cdot)| \leq 4$, and Lemma A.1.7 implies that $g_2(\cdot)$ is $\mathbb{P}$ -Lipschitz continuous with $2L$, we have

$$
\begin{aligned}
&|\mathbb{E}[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2})g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4})g_2(\boldsymbol{W}_{i_5}, \boldsymbol{W}_{i_6})g_2(\boldsymbol{W}_{i_7}, \boldsymbol{W}_{i_8})]| \\
=&\big|\mathbb{E}\big[[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2}) - g_2(\breve{\boldsymbol{W}}_{i_1}, \boldsymbol{W}_{i_2})]g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4})g_2(\boldsymbol{W}_{i_5}, \boldsymbol{W}_{i_6})g_2(\boldsymbol{W}_{i_7}, \boldsymbol{W}_{i_8})\big]\big| \\
\leq&64\mathbb{E}[|g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2}) - g_2(\breve{\boldsymbol{W}}_{i_1}, \boldsymbol{W}_{i_2})|] \\
=&64\mathbb{E}[|g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2}) - g_2(\breve{\boldsymbol{W}}_{i_1}, \boldsymbol{W}_{i_2})|\mathbb{1}_{\|\boldsymbol{W}_{i_1} - \breve{\boldsymbol{W}}_{i_1}\|_\infty \geq \epsilon}] + 64\mathbb{E}[|g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2}) - g_2(\breve{\boldsymbol{W}}_{i_1}, \boldsymbol{W}_{i_2})|\mathbb{1}_{\|\boldsymbol{W}_{i_1} - \breve{\boldsymbol{W}}_{i_1}\|_\infty \leq \epsilon}] \\
\leq&512\mathbb{P}(\|\boldsymbol{W}_{i_1} - \breve{\boldsymbol{W}}_{i_1}\|_\infty \geq \epsilon) + 128L\epsilon \leq 25600\epsilon^{-2/3}\alpha_{\boldsymbol{U}}(m)^{2/3} + 128\widetilde{L}\epsilon,
\end{aligned}
$$

where $\widetilde{L} := \max\{L, 1\}$. When we pick $\epsilon = \widetilde{L}^{-3/5}\alpha_{\boldsymbol{U}}(m)^{2/5}$, we have

$$
|\mathbb{E}[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2})g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4})g_2(\boldsymbol{W}_{i_5}, \boldsymbol{W}_{i_6})g_2(\boldsymbol{W}_{i_7}, \boldsymbol{W}_{i_8})]| \leq 12928\widetilde{L}^{2/5}\alpha_{\boldsymbol{u}}(m)^{2/5}.
$$

This completes the proof. □

Lemma A.1.10. *Suppose all assumptions in Proposition 1.2.2 hold. Let's denote*

$$
U(g_1) = \frac{2}{n}\sum_{i=1}^{n} g_1(\boldsymbol{W}_i) \ \textit{and}\ U(g_2) = \frac{2}{n(n-1)}\sum_{i<j} g_2(\boldsymbol{W}_i, \boldsymbol{W}_j).
$$

Then, for some constants C, C', and C'' which only depend on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, \kappa_1, \kappa_2)$, we have $|\mathbb{E}[U^2(g_1)]| \leq Cn^{-2}$, $|\mathbb{E}[U^2(g_2)]| \leq C'n^{-2}$, *and*

$$|\mathbb{E}[U^4(g_2)]| = \left| \frac{16}{n^4(n-1)^4} \sum_{i_1<i_2} \sum_{i_3<i_4} \sum_{i_5<i_6} \sum_{i_7<i_8} E[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2}) g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4}) g_2(\boldsymbol{W}_{i_5}, \boldsymbol{W}_{i_6}) g_2(\boldsymbol{W}_{i_7}, \boldsymbol{W}_{i_8})] \right|$$
$$\leq C''n^{-2}.$$

Proof. We only have to present the proof for $|\mathbb{E}[U^4(g_2)]| \leq C''n^{-2}$ because we can prove $|\mathbb{E}[U^2(g_1)]| \leq Cn^{-2}$ and $|\mathbb{E}[U^2(g_2)]| \leq C'n^{-2}$ by following the proof of Lemma 3.6 in Dehling and Wendler (2010). The proof is similar to Lemma 3 in Yoshihara (1976). Note that

$$|\mathbb{E}[U^2(g_2)]| \leq \frac{16}{n^4(n-1)^4} \sum_{1\leq i_1,\ldots,i_8 \leq n} |\mathbb{E}[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2}) g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4}) g_2(\boldsymbol{W}_{i_5}, \boldsymbol{W}_{i_6}) g_2(\boldsymbol{W}_{i_7}, \boldsymbol{W}_{i_8})]|.$$

Case 1. Suppose more than two indices among $i_1, \ldots, i_8$ are identical. Then, we have at most $(8!/(3!5!)) \times n^6$ possible cases for $i_1, \ldots, i_8$. Here $k!$ means the factorial of a number $k \in \mathbb{N}$. Therefore, for some constant C_1'' which only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$,

$$\frac{16}{n^4(n-1)^4} \sum_{\substack{0\leq i_1,\ldots,i_8\leq n,\\ \text{at least three of } i\text{'s are identical}}} |\mathbb{E}[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2}) g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4}) g_2(\boldsymbol{W}_{i_5}, \boldsymbol{W}_{i_6}) g_2(\boldsymbol{W}_{i_7}, \boldsymbol{W}_{i_8})]| \leq C_1''n^{-2},$$

because $|g_2(\cdot,\cdot)| \leq 4$.

Case 2. Suppose at most two indices are identical. Let $a_1 \leq a_2 \leq \ldots \leq a_8$ be ordered indices of $i_1, \ldots, i_8$ and let's denote $b_1 := a_2 - a_1$, $b_3 := a_4 - a_3$, $b_5 := a_6 - a_5$, and $b_7 := a_8 - a_7$. Let's set $m = \max\{b_1, b_7\}$. Note that $m \geq 1$ by construction. For a fixed m, we have at most $2 \times n^6 m$ number of combinations for the values of $a_1, \ldots, a_8$. This implies that for a fixed m, we have at most $2 \times 8! \times n^6 m$ number of combinations for the values of $i_1, \ldots i_8$. Therefore, using Lemma A.1.9, we have

$$\frac{16}{n^4(n-1)^4} \sum_{\substack{0\leq i_1,\ldots,i_8\leq n,\\ \text{at most two of } i\text{'s are identical}}} |\mathbb{E}[g_2(\boldsymbol{W}_{i_1}, \boldsymbol{W}_{i_2}) g_2(\boldsymbol{W}_{i_3}, \boldsymbol{W}_{i_4}) g_2(\boldsymbol{W}_{i_5}, \boldsymbol{W}_{i_6}) g_2(\boldsymbol{W}_{i_7}, \boldsymbol{W}_{i_8})]|$$
$$\leq \frac{16}{n^4(n-1)^4} C_2'' \sum_{m=1}^{n} n^6 m \alpha_{\boldsymbol{U}}(m)^{2/5} \leq C_2'' n^{-2},$$

for some constants C_2'' and C_2'' where C_2'' only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}})$, and C_2'' only depends on $(C_{\mathbf{A}}, c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, \kappa_1, \kappa_2)$. Therefore, Cases 1 and 2 combined imply the desired result. □

Lemma A.1.11. *The CBB bootstrap V-statistics $V^*_{0,jk}$ and $V^*_{1,jk}$, and V-statistics $V_{0,jk}$ and $V_{1,jk}$ using $\{\mathbf{Y}_t\}_{t=1}^{n-1}$ have the following relationships:*

$$\left|\mathbb{E}^*[(V^*_{0,jk} - V_{0,jk})^2] - \frac{b\ell(b\ell-\ell)(b\ell-2\ell)(b\ell-3\ell)}{(b\ell)^4}(\widehat{V}_{0,jk} - V_{0,jk})^2\right| \leq \frac{50\ell}{b\ell},$$

$$\left|\mathbb{E}^*[(V^*_{1,jk} - V_{1,jk})^2] - \frac{b\ell(b\ell-\ell)(b\ell-2\ell)(b\ell-3\ell)}{(b\ell)^4}(\widehat{V}_{1,jk} - V_{1,jk})^2\right| \leq \frac{50\ell}{b\ell},$$

where $\mathbb{E}^*(\cdot)$ stands for the expectation operator under $\mathbb{P}^*$.

Proof. Because the proof is identical for $V^*_{0,jk}$ and $V^*_{1,jk}$, we will examine the case of $V^*_{0,jk}$ only. One can decompose $\mathbb{E}^*[(V^*_{0,jk} - V_{0,jk})^2]$ as

$$\begin{aligned}
&\mathbb{E}^*[(V^*_{0,jk} - V_{0,jk})^2] \\
=&\frac{1}{(b\ell)^4}\sum_{1\leq i_1,i_2,i_3,i_4\leq b\ell}\mathbb{E}^*\left[g\left(\begin{pmatrix}Z^*_{i_1,j}\\ Z^*_{i_1,k}\end{pmatrix},\begin{pmatrix}Z^*_{i_2,j}\\ Z^*_{i_2,k}\end{pmatrix}\right)g\left(\begin{pmatrix}Z^*_{i_3,j}\\ Z^*_{i_3,k}\end{pmatrix},\begin{pmatrix}Z^*_{i_4,j}\\ Z^*_{i_4,k}\end{pmatrix}\right)\right] \\
=&\frac{1}{(b\ell)^4}\sum_{\substack{1\leq i_1,i_2,i_3,i_4\leq b\ell,\\ \text{all } i\text{'s are in different blocks}}}\mathbb{E}^*\left[g\left(\begin{pmatrix}Z^*_{i_1,j}\\ Z^*_{i_1,k}\end{pmatrix},\begin{pmatrix}Z^*_{i_2,j}\\ Z^*_{i_2,k}\end{pmatrix}\right)g\left(\begin{pmatrix}Z^*_{i_3,j}\\ Z^*_{i_3,k}\end{pmatrix},\begin{pmatrix}X^*_{i_4,j}\\ X^*_{i_4,k}\end{pmatrix}\right)\right] \\
&+\frac{1}{(b\ell)^4}\sum_{\substack{1\leq i_1,i_2,i_3,i_4\leq b\ell,\\ \text{at least two of } i\text{'s is in the same block}}}\mathbb{E}^*\left[g\left(\begin{pmatrix}Z^*_{i_1,j}\\ Z^*_{i_1,k}\end{pmatrix},\begin{pmatrix}Z^*_{i_2,j}\\ Z^*_{i_2,k}\end{pmatrix}\right)g\left(\begin{pmatrix}Z^*_{i_3,j}\\ Z^*_{i_3,k}\end{pmatrix},\begin{pmatrix}Z^*_{i_4,j}\\ Z^*_{i_4,k}\end{pmatrix}\right)\right] \\
=&\frac{b(b-1)(b-2)(b-3)\ell^4}{(b\ell)^4}(\widehat{V}_{0,jk} - V_{0,jk})^2 \\
&+\frac{1}{(b\ell)^4}\sum_{\substack{1\leq i_1,i_2,i_3,i_4\leq b\ell,\\ \text{at least two of } i\text{'s is in the same block}}}\mathbb{E}^*\left[g\left(\begin{pmatrix}Z^*_{i_1,j}\\ Z^*_{i_1,k}\end{pmatrix},\begin{pmatrix}Z^*_{i_2,j}\\ Z^*_{i_2,k}\end{pmatrix}\right)g\left(\begin{pmatrix}Z^*_{i_3,j}\\ Z^*_{i_3,k}\end{pmatrix},\begin{pmatrix}Z^*_{i_4,j}\\ Z^*_{i_4,k}\end{pmatrix}\right)\right].
\end{aligned}$$

The last equality holds because there are $b(b-1)(b-2)(b-3)\ell^4$ possibilities that i_1, i_2, i_3, and i_4 are in different blocks; and when i_1, i_2, i_3, and i_4 are in different blocks, similar to the proof of Lemma 3.7 in Dehling and Wendler (2010), we have

$$\begin{aligned}
&\mathbb{E}^*\left[g\left(\begin{pmatrix}Z^*_{i_1,j}\\ Z^*_{i_1,k}\end{pmatrix},\begin{pmatrix}Z^*_{i_2,j}\\ Z^*_{i_2,k}\end{pmatrix}\right)g\left(\begin{pmatrix}Z^*_{i_3,j}\\ Z^*_{i_3,k}\end{pmatrix},\begin{pmatrix}Z^*_{i_4,j}\\ Z^*_{i_4,k}\end{pmatrix}\right)\right] \\
=&\frac{1}{(n-1)^4}\sum_{1\leq i_1,i_2,i_3,i_4\leq n-1}\left[g\left(\begin{pmatrix}Z_{i_1,j}\\ Z_{i_1,k}\end{pmatrix},\begin{pmatrix}Z_{i_2,j}\\ Z_{i_2,k}\end{pmatrix}\right)g\left(\begin{pmatrix}Z_{i_3,j}\\ Z_{i_3,k}\end{pmatrix},\begin{pmatrix}Z_{i_4,j}\\ Z_{i_4,k}\end{pmatrix}\right)\right] = (\widehat{V}_{0,jk} - V_{0,jk})^2,
\end{aligned}$$

because $\begin{pmatrix} Z^*_{i_1,j} \\ Z^*_{i_1,k} \end{pmatrix}$, $\begin{pmatrix} Z^*_{i_2,j} \\ Z^*_{i_2,k} \end{pmatrix}$, $\begin{pmatrix} Z^*_{i_3,j} \\ Z^*_{i_3,k} \end{pmatrix}$, and $\begin{pmatrix} Z^*_{i_4,j} \\ Z^*_{i_4,k} \end{pmatrix}$ are independent under $\mathbb{P}^*$. Therefore, we have

$$\begin{aligned}
&\left| \mathbb{E}^*[(V^*_{0,jk} - V_{0,jk})^2] - \frac{b(b-1)(b-2)(b-3)\ell^4}{(b\ell)^4}(\widehat{V}_{0,jk} - V_{0,jk})^2 \right| \\
=&\left| \frac{1}{(b\ell)^4} \sum_{\substack{1 \le i_1, i_2, i_3, i_4 \le b\ell, \\ \text{at least one of } i\text{'s in the same block}}} \mathbb{E}^* \left[g\left(\begin{pmatrix} Z^*_{i_1,j} \\ Z^*_{i_1,k} \end{pmatrix}, \begin{pmatrix} Z^*_{i_2,j} \\ Z^*_{i_2,k} \end{pmatrix} \right) g\left(\begin{pmatrix} Z^*_{i_3,j} \\ Z^*_{i_3,k} \end{pmatrix}, \begin{pmatrix} Z^*_{i_4,j} \\ Z^*_{i_4,k} \end{pmatrix} \right) \right] \right| \\
\le& 4\frac{(b\ell)^4 - b(b-1)(b-2)(b-3)\ell^4}{(b\ell)^4} = 4\frac{6b^3\ell^4 - 11b^2\ell^4 + 6b\ell^4}{(b\ell)^4} \le \frac{50b^3\ell^4}{(b\ell)^4} = \frac{50\ell}{b\ell},
\end{aligned}$$

because $b \in \mathbb{N}$ and $|g(\cdot)| \le 2$. □

A.2 Estimation and Inference in Latent VAR(p) Model

Let $\boldsymbol{Z}_t \in \mathbb{R}^d$ follow the VAR(p) process below for a finite and fixed p:

$$\boldsymbol{Z}_t = \mathbf{A}_1 \boldsymbol{Z}_{t-1} + \mathbf{A}_2 \boldsymbol{Z}_{t-2} + \ldots + \mathbf{A}_p \boldsymbol{Z}_{t-p} + \boldsymbol{E}_t, \quad \boldsymbol{E}_t \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{E}}), \tag{A.6}$$

where $\mathbf{A}_1, ..., \mathbf{A}_p$ are unknown transition matrices.

In this section, we assume that Assumption M-p below and Assumption S in Section 1.2.1 are satisfied. We extend the estimation and inference methods developed in the previous sections to the VAR(p) process for a finite and fixed p. To save space, we present our rank-based estimators of $\mathbf{A}_1, ..., \mathbf{A}_p$ and tests for Granger causality without providing the corresponding rate results for the estimators. For proofs, see Section A.2.3 in the appendix.

Assumption M-p: i) $\mathrm{diag}(\boldsymbol{\Sigma}_0) = \mathbf{I}_d$, i.e., $\boldsymbol{\Sigma}_0$ is a correlation matrix; ii) $\{\boldsymbol{E}_t\}_{t\in\mathbb{Z}} \overset{\text{i.i.d.}}{\sim} N(0, \boldsymbol{\Sigma}_{\boldsymbol{E}})$, where $0 < c_{\boldsymbol{E}} < \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{E}}) \le \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{E}}) < C_{\boldsymbol{E}} < \infty$ for some absolute constants $c_{\boldsymbol{E}}$ and $C_{\boldsymbol{E}}$; iii) $\|\mathbf{A}_i\|_2 \le a_i < 1$ by an absolute constant a_i for all $i = 1, \ldots, p$, and $\sum_{i=1}^p a_i < 1$; and iv) the matrix $\mathbf{A} := (\mathbf{A}_p, \ldots, \mathbf{A}_1) \in \mathbb{R}^{d \times pd}$ satisfies:

$$\mathbf{A} \in \mathcal{M}(s, M) \text{ with } \mathcal{M}(s, M) := \left\{ \mathbf{M} \in \mathbb{R}^{d \times pd} : \max_{1 \le j \le d} \sum_{k=1}^{pd} \mathbb{1}(M_{jk} \neq 0) \le s, \|\mathbf{M}\|_\infty \le M \right\}, \tag{A.7}$$

where s is a positive integer which may depend on d and p, and M is a positive constant which may also depend on d and p.

Assumption M-p reduces to Assumption M in Section 1.2.1 when $p = 1$. For a Gaussian VAR(p) process $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$, it is well-known that the sequence $\{Z_{t,k}\}_{t\in\mathbb{Z}}$ *Granger causes* $\{Z_{t,j}\}_{t\in\mathbb{Z}}$ if and only if there exists $i \in [p]$ such that the (j,k)-th element of $\mathbf{A}_i$, $A_{i,jk}$, is non-zero (cf. Corollary 2.2.1 in Lütkepohl (2005)). The following proposition holds for the observable process $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$.

Proposition A.2.1. *Suppose $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ follows VAR(p) model (A.6) with finite and fixed p. Under Assumptions M-p(ii), M-p(iii), and S(i), we obtain that the sequence $\{X_{t,k}\}_{t\in\mathbb{Z}}$ does not Granger cause $\{X_{t,j}\}_{t\in\mathbb{Z}}$ if and only if $A_{i,jk} = 0$ for all $i \in [p]$.*

We establish α-mixing property of $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ stated below in Theorem A.2.1.

Theorem A.2.1. *Suppose $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ follows VAR(p) model (A.6) with finite and fixed p. Under Assumptions M-p and S, $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is geometrically α-mixing with mixing coefficient satisfying*

$$\alpha(n; \{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}) = \alpha(n; \{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}) \leq \gamma_1 \exp(-\gamma_2 n), \text{ for all } n \geq 1.$$

Here γ_1 and γ_2 are positive constants that only depend on $c_{\boldsymbol{E}}, C_{\boldsymbol{E}}$, $a_1, \ldots, a_p$, and p.

A.2.1 Estimators of $\mathbf{A}_1, ..., \mathbf{A}_p$

Let $\boldsymbol{\Omega} = \mathbb{E}\left[\mathfrak{U}_t \mathfrak{U}_t^\top\right]$, where $\mathfrak{U}_t := (\boldsymbol{Z}_t^\top, \ldots, \boldsymbol{Z}_{t+p}^\top)^\top \in \mathbb{R}^{(p+1)d}$. We set $\boldsymbol{\Sigma}_0 := \boldsymbol{\Omega}_{[pd],[pd]}$ and $\boldsymbol{\Sigma}_1 := \boldsymbol{\Omega}_{[pd],pd+[d]}$. Then, we estimate $\mathbf{A}_1, \ldots, \mathbf{A}_p$ via the following two steps.

Step 1. Estimate $\boldsymbol{\Omega}$ by $\widehat{\boldsymbol{\Omega}} = \sin\left(0.5\pi \widehat{\mathfrak{T}}\right)$, where

$$\widehat{\mathfrak{T}} = \frac{2}{(n-p)(n-p-1)} \sum_{i=1}^{n-p} \sum_{j=i+1}^{n-p} \left[\sin(\mathfrak{Y}_i - \mathfrak{Y}_j) \sin(\mathfrak{Y}_i - \mathfrak{Y}_j)^\top\right],$$

in which $\mathfrak{Y}_t := (\boldsymbol{X}_t^\top, \ldots, \boldsymbol{X}_{t+p}^\top)^\top \in \mathbb{R}^{(p+1)d}$. Let $\widehat{\boldsymbol{\Sigma}}_0 := \widehat{\boldsymbol{\Omega}}_{[pd],[pd]}$ and $\widehat{\boldsymbol{\Sigma}}_1 := \widehat{\boldsymbol{\Omega}}_{[pd],pd+[d]}$.

Step 2. Estimate the stacked matrix $\mathbf{A} = (\mathbf{A}_p, \dots, \mathbf{A}_1) \in \mathbb{R}^{d \times pd}$ by

$$\widehat{\mathbf{A}} := \underset{\mathbf{M} \in \mathbb{R}^{d \times pd}}{\operatorname{argmin}} \sum_{j=1}^{d} \sum_{k=1}^{pd} |M_{jk}| \text{ such that } \|\widehat{\boldsymbol{\Sigma}}_0 \mathbf{M}^\top - \widehat{\boldsymbol{\Sigma}}_1\|_{\max} \leq \lambda,$$

where λ is a tuning parameter.

A.2.2 Tests for Granger Causality

We present two tests for Granger non-causality of an individual series $\{X_{t,k}\}_{t \in \mathbb{Z}}$ on another individual series $\{X_{t,m}\}_{t \in \mathbb{Z}}$ under the latent VAR(p) model (A.6) for $m \neq k$, $m \in [d]$, and $k \in [d]$. From Proposition A.2.1, it is sufficient to consider

$$H_0 : A_{i,mk} = 0 \text{ for all } i \in [p] \text{ against } H_1 : A_{i,mk} \neq 0 \text{ for some } i \in [p]. \tag{A.8}$$

Following Section 3, we first construct a de-biased estimator of $(A_{1,mk}, \dots, A_{p,mk})^\top \in \mathbb{R}^p$, then provide a consistent estimator of the asymptotic variance of the de-biased estimator under the null hypothesis via CBB, and finally develops two tests.

De-biased Estimation

For notational compactness, we let

$$G := \{k, k+d, k+2d, \dots, k+(p-1)d\}, \boldsymbol{e}_j := (\underbrace{0, \dots, 0}_{j-1}, 1, 0, \dots, 0)^\top \in \mathbb{R}^{pd},$$

$$\boldsymbol{\beta} := (\mathbf{A}_{p,m*}, \dots, \mathbf{A}_{1,m*})^\top \in \mathbb{R}^{pd}, \text{ and } \boldsymbol{\theta} := \boldsymbol{\beta}_G = (A_{p,mk}, \dots, A_{1,mk})^\top \in \mathbb{R}^p.$$

For any generic vectors $\boldsymbol{v} = (v_1, \dots, v_{pd})^\top \in \mathbb{R}^{pd}$ and $\boldsymbol{x} = (x_1, \dots, x_p)^\top \in \mathbb{R}^p$, we define $\boldsymbol{v}(\boldsymbol{x}) \in \mathbb{R}^{pd}$ as follows,

$$[\boldsymbol{v}(x)]_j = \begin{cases} x_{(j-k)/d+1} & \text{when } j \in G, \\ v_j & \text{otherwise.} \end{cases}$$

That is, $\boldsymbol{v}(\boldsymbol{x})$ is a vector such that the values of $\boldsymbol{v}$ indexed by G are replaced by $\boldsymbol{x}$, keeping the remaining parts intact.

The de-biased estimator $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is constructed via the following three steps.

Step 1. Estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{v} \in \mathbb{R}^{pd}}{\operatorname{argmin}} \|\boldsymbol{v}\|_1 \text{ such that } \|\widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{v} - \widehat{\boldsymbol{\Sigma}}_{1,*m}\|_\infty \le \lambda,$$

where λ is a tuning parameter.

Step 2. Estimate $\mathbf{W} := [\boldsymbol{\Sigma}_0^{-1}]_{*G} \in \mathbb{R}^{pd \times p}$ by

$$\hat{\mathbf{W}} := \underset{\mathbf{M} \in \mathbb{R}^{pd \times p}}{\operatorname{argmin}} \sum_{i=1}^{pd} \sum_{j=1}^{p} |M_{ij}| \text{ such that } \|\widehat{\boldsymbol{\Sigma}}_0 \mathbf{M} - [\mathbf{I}_{pd}]_{*G}\|_{\max} \le \lambda',$$

where λ' is another tuning parameter. Equivalently, $\hat{\mathbf{W}} = (\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_p) \in \mathbb{R}^{pd \times p}$, where

$$\hat{\boldsymbol{w}}_j := \underset{\boldsymbol{v} \in \mathbb{R}^{pd}}{\operatorname{argmin}} \|\boldsymbol{v}\|_1 \text{ such that } \|\hat{\boldsymbol{\Sigma}}_0 \boldsymbol{v} - \mathbf{e}_{k+(j-1)d}\|_\infty \le \lambda' \text{ for each } j \in [p].$$

Step 3. Estimate $\boldsymbol{\theta}$ by solving the estimating equation $\widehat{\boldsymbol{S}}(\hat{\boldsymbol{\beta}}(\boldsymbol{x})) = \mathbf{0}_{p \times 1}$, where $\boldsymbol{x} \in \mathbb{R}^p$ and

$$\widehat{\boldsymbol{S}}(\boldsymbol{v}) = \widehat{\mathbf{W}}^\top (\widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{v} - \widehat{\boldsymbol{\Sigma}}_{1,*m}).$$

The solution $\widetilde{\boldsymbol{\theta}}$ to the above equation has the following closed form:

$$\widetilde{\boldsymbol{\theta}} = -\left(\widehat{\mathbf{W}}^\top \widehat{\boldsymbol{\Sigma}}_{0,*G}\right)^{-1} \hat{\mathbf{W}}^\top \left(\widehat{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\beta}}(\mathbf{0}_{p \times 1}) - \widehat{\boldsymbol{\Sigma}}_{1,*m}\right) = \hat{\boldsymbol{\theta}} - \left(\widehat{\mathbf{W}}^\top \widehat{\boldsymbol{\Sigma}}_{0,*G}\right)^{-1} \hat{\mathbf{W}}^\top \left(\widehat{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Sigma}}_{1,*m}\right),$$

where $\widehat{\boldsymbol{\theta}} := \widehat{\boldsymbol{\beta}}_G = (\widehat{A}_{p,mk}, \ldots, \widehat{A}_{1,mk})^\top \in \mathbb{R}^p$.

Bootstrap Estimation of the Asymptotic Variance

Similar to Theorem 1.3.2, we obtain that under conditions in Theorem A.2.2 below, the de-biased estimator $\widetilde{\boldsymbol{\theta}}$ is asymptotically normally distributed with the asymptotic variance given by

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},n} := (n-p)\mathrm{Var}\left\{\mathbf{W}^\top \left(\widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{\beta} - \widehat{\boldsymbol{\Sigma}}_{1,*m}\right)\right\}. \tag{A.9}$$

We provide the following estimator of the asymptotic variance $\boldsymbol{\Sigma}_{\boldsymbol{\theta},n}$ in equation (A.9) under $H_0 : \boldsymbol{\theta} = \mathbf{0}_{p \times 1}$.

Step 1. Construct $\{\mathfrak{Y}_t\}_{t=1}^{n-p}$, and draw CBB samples $\{\mathfrak{Y}_t^*\}_{t=1}^{b\ell}$ from $\{\mathfrak{Y}_t\}_{t=1}^{n-p}$, where $b = [(n-p)/\ell]$.

Step 2. Construct $\widehat{\boldsymbol{\Omega}}^* := \sin\left(0.5\pi\widehat{\mathfrak{T}}\right)$, where

$$\widehat{\mathfrak{T}} = \frac{2}{(b\ell)(b\ell-1)} \sum_{i=1}^{b\ell} \sum_{j=i+1}^{b\ell} \left[\sin(\mathfrak{Y}_i^* - \mathfrak{Y}_j^*) \sin(\mathfrak{Y}_i^* - \mathfrak{Y}_j^*)^\top\right].$$

Then, we set $\widehat{\boldsymbol{\Sigma}}_0^* := \widehat{\boldsymbol{\Omega}}^*_{[pd],[pd]}$, and $\widehat{\boldsymbol{\Sigma}}_1^* := \widehat{\boldsymbol{\Omega}}^*_{[pd],pd+[d]}$.

Step 3. Calculate the bootstrap variance as

$$\widehat{\boldsymbol{\Sigma}}^*_{\boldsymbol{\theta},n} := (b\ell)\mathrm{Var}^*\left\{\hat{\mathbf{W}}^\top\left(\widehat{\boldsymbol{\Sigma}}_0^*\hat{\boldsymbol{\beta}}(\mathbf{0}_{p\times 1}) - \widehat{\boldsymbol{\Sigma}}^*_{1,*m}\right)\right\}.$$

Tests for Granger non-causality

Define the following test statistics,

$$\begin{aligned} \mathcal{W}_n &:= (n-p)\boldsymbol{\theta}^\top\left(\widehat{\boldsymbol{\Sigma}}^*_{\boldsymbol{\theta},n}\right)^{-1}\boldsymbol{\theta} \text{ and} \\ \mathcal{W}_n^{\mathrm{adj}} &:= (n-p)\left(\widehat{\mathbf{W}}^\top\widehat{\boldsymbol{\Sigma}}_{0,*G}\boldsymbol{\theta}\right)^\top\left(\widehat{\boldsymbol{\Sigma}}^*_{\boldsymbol{\theta},n}\right)^{-1}\left(\widehat{\mathbf{W}}^\top\widehat{\Sigma}_{0,*G}\boldsymbol{\theta}\right). \end{aligned}$$

Based on $\mathcal{W}_n$ and $\mathcal{W}_n^{\mathrm{adj}}$, we define the following two tests,

$$\mathbb{T}_{n,\alpha} = \mathbb{1}\left\{\mathcal{W}_n > \chi^2_{p,1-\alpha}\right\} \text{ and } \mathbb{T}^{\mathrm{adj}}_{n,\alpha} = \mathbb{1}\left\{\mathcal{W}_n^{\mathrm{adj}} > \chi^2_{p,1-\alpha}\right\},$$

where $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$ quantile of the chi-square distribution with the degree of the freedom p.

Theorem A.2.2. *For the latent VAR(p) model (A.6) with finite and fixed p, suppose that Assumptions M-p, S, and E hold. We assume that* $\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\theta},n}) \geq C_{\boldsymbol{\theta}} > 0$ *by some absolute constant* $C_{\boldsymbol{\theta}}$*. In addition, we assume that* $\ell \to \infty$, $\ell^2/n = o(1)$,

$$M = O(1), \ \|\boldsymbol{\Sigma}_0^{-1}\|_\infty = O(1), \ \textit{and} \ \max\{s^2_{\mathbf{W}}, s^2\}\frac{\log(ed)}{\sqrt{n}} = o(1),$$

where $s_{\mathbf{W}} := \max_{1\le k\le p}\sum_{j=1}^{pd}\mathbb{1}(W_{jk}\neq 0)$ *is a positive integer. Let* $\lambda \asymp \sqrt{\log(ed)/n}$*,* $\lambda' \asymp \sqrt{\log(ed)/n}$*,* $\lambda \ge C\sqrt{\log(ed)/n}$*, and* $\lambda' \ge C'\sqrt{\log(ed)/n}$ *for some sufficiently large* C *and* C'*. We then have*

$$\mathbb{P}(\mathbb{T}_{n,\alpha}=1|H_0)=\alpha+o(1) \text{ and } \mathbb{P}(\mathbb{T}^{\mathrm{adj}}_{n,\alpha}=1|H_0)=\alpha+o(1).$$

A.2.3 Proofs in Appendix A.2

Notation

For any square matrix $\mathbf{M}$, let $\rho(\mathbf{M})$ denote its spectral radius. Let

$$\bar{\mathbf{e}}_p := \begin{pmatrix}\mathbf{I}_d\\ \mathbf{0}_{(p-1)d\times d}\end{pmatrix},\ \underline{\mathbf{e}}_p := \begin{pmatrix}\mathbf{0}_{(p-1)d\times d}\\ \mathbf{I}_d\end{pmatrix},\ \mathbf{C}_p := \begin{pmatrix}\mathbf{A}_p\\ \mathbf{A}_{p-1}\\ \vdots\\ \mathbf{A}_1\end{pmatrix},\ \mathbf{Q}_p := \begin{pmatrix}-\mathbf{A}_{p-1} & \cdots & -\mathbf{A}_2 & -\mathbf{A}_1 & \mathbf{I}_d\\ -\mathbf{A}_{p-2} & \cdots & -\mathbf{A}_1 & \mathbf{I}_d & \mathbf{0}_{d\times d}\\ \iddots & \iddots & \iddots & \iddots & \iddots\\ -\mathbf{A}_1 & \mathbf{I}_d & \mathbf{0}_{d\times d} & \cdots & \cdots\\ \mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d}\end{pmatrix}$$

with $\bar{\mathbf{e}}_1=\underline{\mathbf{e}}_1=\mathbf{I}_d$ and $\mathbf{Q}_1 := \mathbf{I}_d$. We set $\boldsymbol{\mathfrak{Z}}_t \equiv (\boldsymbol{Z}_t^\top, \boldsymbol{Z}_{t+1}^\top,\ldots,\boldsymbol{Z}_{t+p-1}^\top)^\top \in \mathbb{R}^{pd}$, $\boldsymbol{\mathfrak{E}}_t = (\mathbf{0}_{d\times(p-1)d}, \boldsymbol{E}_{t+p-1}^\top)^\top \in \mathbb{R}^{pd}$,

$$\boldsymbol{\mathfrak{A}}_p = \begin{pmatrix}\mathbf{0}_{d\times d} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d}\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{I}_d & \mathbf{0}_{d\times d}\\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{I}_d\\ \mathbf{A}_p & \cdots & \mathbf{A}_3 & \mathbf{A}_2 & \mathbf{A}_1\end{pmatrix} \in \mathbb{R}^{pd\times pd},\ \text{and } \bar{\boldsymbol{\mathfrak{A}}}_p := \begin{pmatrix}0 & 1 & 0 & \cdots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ 0 & \cdots & 0 & 1 & 0\\ 0 & \cdots & 0 & 0 & 1\\ a_p & \cdots & a_3 & a_2 & a_1\end{pmatrix} \in \mathbb{R}^{p\times p}.$$

For a pd by pd matrix $\boldsymbol{\mathfrak{A}}_p$, we partition $\boldsymbol{\mathfrak{A}}_p^k$ by p^2 blocks where each block is a d by d matrix. We denote by $[\boldsymbol{\mathfrak{A}}_p^k]_{jm}$ the (j,m)-th block of $\boldsymbol{\mathfrak{A}}_p^k$. Let $\overline{\boldsymbol{\mathfrak{A}}_p^k}$ denote a p by p matrix where the (j,m)-th element $a_{k,jm}$ of $\overline{\boldsymbol{\mathfrak{A}}_p^k}$ is defined as $a_{k,jm} = \|[\boldsymbol{\mathfrak{A}}_p^k]_{jm}\|_2$. Then, we can write

$\mathfrak{A}_p^k$ and $\overline{\mathfrak{A}_p^k}$ as

$$\mathfrak{A}_p^k = \begin{pmatrix} [\mathfrak{A}_p^k]_{11} & \cdots & [\mathfrak{A}_p^k]_{1p} \\ \vdots & \ddots & \vdots \\ [\mathfrak{A}_p^k]_{p1} & \cdots & [\mathfrak{A}_p^k]_{pp} \end{pmatrix} \text{ and } \overline{\mathfrak{A}_p^k} = \begin{pmatrix} a_{k,11} & \cdots & a_{k,1p} \\ \vdots & \ddots & \vdots \\ a_{k,p1} & \cdots & a_{k,pp} \end{pmatrix}.$$

We have $\|\mathfrak{A}_p^k\|_2 \le \|\overline{\mathfrak{A}_p^k}\|_2 \le \|\overline{\mathfrak{A}_p}^k\|_2$ because for any vectors $\boldsymbol{v} = (\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top, \ldots, \boldsymbol{v}_p^\top)^\top \in \mathbb{R}^{pd}$ and $\bar{\boldsymbol{v}} = (\|\boldsymbol{v}_1\|_2, \ldots, \|\boldsymbol{v}_p\|_2)^\top \in \mathbb{R}^p$ where $\{\boldsymbol{v}_j\}_{j=1}^p$ is a $d \times 1$ vector, we have

$$\|\mathfrak{A}_p^k \boldsymbol{v}\|_2 \le \|\overline{\mathfrak{A}_p^k}\bar{\boldsymbol{v}}\|_2 \le \|\overline{\mathfrak{A}_p^{k-1}}\overline{\mathfrak{A}_p}\bar{\boldsymbol{v}}\|_2 \le \ldots \le \|\overline{\mathfrak{A}_p}^k \bar{\boldsymbol{v}}\|_2, \text{ and } \|\boldsymbol{v}\|_2 = \|\bar{\boldsymbol{v}}\|_2.$$

Proof of Proposition A.2.1

For random vector $\boldsymbol{X}_t \in \mathbb{R}^d$, let $\boldsymbol{X}_{t,\setminus k}$ denote the sub-vector of $\boldsymbol{X}_t$ which removes $X_{t,k}$ from $\boldsymbol{X}_t$.

Note that $\boldsymbol{f} = \{f_1, \ldots, f_d\}$ consists of univariate strictly increasing functions. This implies that, for any measurable set B,

$$\mathbb{P}(X_{t+1,j} \in B | \{\boldsymbol{X}_s\}_{s \le t}) = \mathbb{P}(X_{t+1,j} \in B | \{\boldsymbol{X}_{s,\setminus k}\}_{s \le t})$$

is identical to

$$\mathbb{P}(Z_{t+1,j} \in f_j^{-1}(B) | \{\boldsymbol{Z}_s\}_{s \le t}) = \mathbb{P}(Z_{t+1,j} \in f_j^{-1}(B) | \{\boldsymbol{Z}_{s,\setminus k}\}_{s \le t}).$$

It is well known that the above condition for $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ holds if and only if $A_{i,jk} = 0$ for all $i \in [p]$ since $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is a gaussian VAR process (cf. Corollary 2.2.1 in Lütkepohl (2005)). Therefore, we have the desired result.

Proof of Theorem A.2.1

When $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ follows VAR(p) model (A.6), we know that $\{\mathfrak{Z}_t\}_{t\in\mathbb{Z}}$ follows VAR(1) model: $\mathfrak{Z}_t = \mathfrak{A}_p \mathfrak{Z}_{t-1} + \mathfrak{E}_t$ with $\boldsymbol{\Sigma}_0 = \mathrm{Var}(\mathfrak{Z}_t)$. Under Assumption M-p, the spectral radius of $\bar{\mathfrak{A}}_p$, $\rho(\bar{\mathfrak{A}}_p)$, satisfies: $\rho(\bar{\mathfrak{A}}_p) \le C_{\bar{\mathfrak{A}}_p} < 1$ for some absolute constant $C_{\bar{\mathfrak{A}}_p}$ by Lemma 4.10 in Han and Li (2019).

By following the proof of Theorem 3.1 in Han and Wu (2019), we know the ϱ-mixing coefficient

$$\varrho\left(\sigma(\{\mathfrak{Z}_t\}_{t\le 0}),\sigma(\{\mathfrak{Z}_t\}_{t\ge k})\right)\begin{cases}=1 & \text{when } k<p,\\ \le\sqrt{(\lambda_{\max}(\boldsymbol{\Sigma}_0)/\lambda_{\min}(\boldsymbol{\Sigma}_0))}\|\mathfrak{A}_p^k\|_2 & \text{when } k\ge p.\end{cases}\tag{A.10}$$

We refer to Bradley (2005) for the definition and the basic properties of ϱ-mixing and α-mixing coefficients. In inequality (A.10), $\lambda_{\max}(\boldsymbol{\Sigma}_0)/\lambda_{\min}(\boldsymbol{\Sigma}_0)$ is upper bounded by a positive constant that only depends on $(c_{\boldsymbol{E}},C_{\boldsymbol{E}},a_1,\ldots,a_p,p)$ by the following lemma, whose proof is presented in Section A.2.3.

Lemma A.2.1. *Suppose $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ follows VAR(p) model (A.6) with finite and fixed p. Under Assumptions M-p and S, the following results hold: For the covariance matrix $\boldsymbol{\Omega}_j := \mathrm{Var}((\boldsymbol{Z}_t^\top,\ldots,\boldsymbol{Z}_{t+j}^\top)^\top)$ with $j\in\{0,1,2,\cdots\}$, we have $0<c_p\le\lambda_{\min}(\mathrm{Var}(\boldsymbol{\Omega}_j))\le\lambda_{\max}(\mathrm{Var}(\boldsymbol{\Omega}_j))\le C_p<\infty$ for some constants c_p and C_p that only depend on $(c_{\boldsymbol{E}},C_{\boldsymbol{E}},a_1,\ldots,a_p,p)$.*

Because p is finite and fixed, we only have to consider the case that $k\ge p$. Then,

$$\varrho\left(\sigma(\{\mathfrak{Z}_t\}_{t\le 0}),\sigma(\{\mathfrak{Z}_t\}_{t\ge k})\right)\le\sqrt{\frac{\lambda_{\max}(\boldsymbol{\Sigma}_0)}{\lambda_{\min}(\boldsymbol{\Sigma}_0)}}\|\mathfrak{A}^k\|_2\le\sqrt{\frac{\lambda_{\max}(\boldsymbol{\Sigma}_0)}{\lambda_{\min}(\boldsymbol{\Sigma}_0)}}\|\overline{\mathfrak{A}_p}^k\|_2$$

because $\|\mathfrak{A}_p^k\|_2\le\|\overline{\mathfrak{A}_p^k}\|_2\le\|\overline{\mathfrak{A}_p}^k\|_2$. When p is finite and fixed, Gelfand's formula implies that there exists a K_ϵ such that $\|\overline{\mathfrak{A}_p}^k\|_2<(\rho(\overline{\mathfrak{A}_p})+\epsilon)^k$ for all $k\ge K_\epsilon$, by choosing some $\epsilon>0$ such that $\rho(\overline{\mathfrak{A}_p})+\epsilon<C<1$ in which C is a positive constant that only depends on $(a_1,\ldots,a_p)$ (cf. Corollary 5.6.14 in Horn and Johnson (2012)). Then, when $k>\bar{K}=\max\{K_\epsilon,p\}$, we have

$$\varrho\left(\sigma(\{\mathfrak{Z}_t\}_{t\le 0}),\sigma(\{\mathfrak{Z}_t\}_{t\ge k})\right)\le\sqrt{\frac{\lambda_{\max}(\boldsymbol{\Sigma}_0)}{\lambda_{\min}(\boldsymbol{\Sigma}_0)}}\|\overline{\mathfrak{A}_p}^k\|_2<\sqrt{\frac{\lambda_{\max}(\boldsymbol{\Sigma}_0)}{\lambda_{\min}(\boldsymbol{\Sigma}_0)}}(\rho(\overline{\mathfrak{A}_p})+\epsilon)^k.$$

Therefore, $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is geometrically α-mixing with $\alpha(k;\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}})\le\gamma_1\exp(-\gamma_2 k)$, where γ_1 and γ_2 are positive constants that only depend on $(c_{\boldsymbol{E}},C_{\boldsymbol{E}},a_1,\ldots,a_p,p)$. This completes the proof.

Proof of Theorem A.2.2

Because the proof of this theorem is similar to VAR(1) case, we will just sketch the proof. First, we replace Lemma A.1.3 and Theorem 1.2.2 with Lemma A.2.1 and Theorem A.2.1, respectively. Then, we have the conclusion that is similar to Section 1.2 because $\{\boldsymbol{\mathfrak{Z}}_t\}_{t\in\mathbb{Z}}$ follows a VAR(1) model when $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ follows VAR(p) model (A.6).

Regarding the inference, our estimator $\widetilde{\boldsymbol{\theta}}$ and the related equation $\widehat{\boldsymbol{S}}(\boldsymbol{v}) = \widehat{\mathbf{W}}^\top(\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{v} - \widehat{\boldsymbol{\Sigma}}_1)$ are p-dimensional vectors, not scalar values. Therefore, we need appropriate norms and additional inequalities in order to prove the theorem. Regarding this issue, we have the following properties.

(i) Because p is finite and fixed, the choice of norm is not important. In particular, for a square matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, we have $\|\mathbf{M}\|_{\max} \le \|\mathbf{M}\|_2 \le p\|\mathbf{M}\|_{\max}$.

(ii) For any random variables $\boldsymbol{R}_t \in \mathbb{R}^p$ and $\boldsymbol{F}_t \in \mathbb{R}^p$, Cauchy-Schwartz theorem implies

$$\|\mathrm{Cov}(\boldsymbol{R}_t, \boldsymbol{F}_t)\|_{\max} \le \sqrt{\|\mathrm{Var}(\boldsymbol{R}_t)\|_{\max}\|\mathrm{Var}(\boldsymbol{F}_t)\|_{\max}} \text{ and } \|\mathrm{Var}(\boldsymbol{R}_t)\|_{\max} = \max_{j\in p}|\mathrm{Var}(\boldsymbol{R}_{t,j})|.$$

(iii) For positive semi-definite matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{p,p}$, we have the following result (cf. Wielandt-Hoffman inequality or Corollary 7.3.5 in Horn and Johnson (2012)):

$$\sum_{j=1}^{p}|\sigma_i(\mathbf{M}) - \sigma_i(\mathbf{N})|^2 \le \|\mathbf{M}-\mathbf{N}\|_F^2 \le p\|\mathbf{M}-\mathbf{N}\|_2^2,$$

where $\|\mathbf{M}-\mathbf{N}\|_F$ is the Frobenius norm of $\mathbf{M}-\mathbf{N}$.

These properties imply that we can prove the theorem by following the proofs in VAR(1) case under assumption in Theorem A.2.2. First, properties (i) and (ii) imply that under the max norm, we can apply lemmas and proofs in VAR(1) case to VAR(p) case. Second, property (iii) helps to present the line of reasoning which is similar to the logic below equation (A.4) in the appendix of the main paper. Therefore, we will skip the proof.

Proof of Lemma A.2.1

We will prove this lemma as follows. First, we will analyze the case where $j = p-1$ in Case 1. Then, we will consider the case where $j \geq p$ in Case 2. We only need to consider these two cases to prove this lemma because Cauchy interlacing theorem implies that this lemma also holds when $0 \leq j < p-1$ if it is true when $j \geq p-1$. (cf. Theorem 4.3.17 in Horn and Johnson (2012))

Case 1. We analyze the case where $j = p-1$.

Because $\boldsymbol{\Sigma}_0 = \sum_{k=0}^{\infty} \mathfrak{A}_p^k \boldsymbol{\Sigma}_{\mathfrak{E}} (\mathfrak{A}_p^\top)^k = \sum_{k=0}^{\infty} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)\ \boldsymbol{\Sigma}_{\boldsymbol{E}} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)^\top$, we have

$$\lambda_{\max}(\boldsymbol{\Sigma}_0) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{E}}) \sum_{k=0}^{\infty} \|\mathfrak{A}_p^k \underline{\mathbf{e}}_p\|_2^2 \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{E}}) \sum_{k=0}^{\infty} \|\bar{\mathfrak{A}}_p{}^k\|_2^2 \leq C < \infty,$$

where C is a positive constant which only depends on $(C_{\boldsymbol{E}}, a_1, \ldots, a_p, p)$. The second last inequality holds because of Assumption M-p(ii) and (iii). From Weyl's inequality, we have

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Sigma}_0) &= \lambda_{\min}\left(\sum_{k=0}^{p-1} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p) \boldsymbol{\Sigma}_{\boldsymbol{E}} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)^\top + \sum_{k=p}^{\infty} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p) \boldsymbol{\Sigma}_{\boldsymbol{E}} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)^\top\right) \\ &\geq \lambda_{\min}\left(\sum_{k=0}^{p-1} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p) \boldsymbol{\Sigma}_{\boldsymbol{E}} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)^\top\right) = \frac{1}{\lambda_{\max}\left(\left(\sum_{k=0}^{p-1} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p) \boldsymbol{\Sigma}_{\boldsymbol{E}} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)^\top\right)^{-1}\right)},\end{aligned}$$

if $\left(\sum_{k=0}^{p-1} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p) \boldsymbol{\Sigma}_{\boldsymbol{E}} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)^\top\right)^{-1}$ exists. Note that $\sum_{k=0}^{p-1} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)\ \boldsymbol{\Sigma}_{\boldsymbol{E}} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)^\top = \mathbf{G}_p (\mathbf{I}_p \otimes \boldsymbol{\Sigma}_{\boldsymbol{E}}) \mathbf{G}_p^\top$, where $\mathbf{G}_p = \begin{pmatrix} \mathbf{I}_{pd} \underline{\mathbf{e}}_p & \mathfrak{A}_p \underline{\mathbf{e}}_p & \cdots & \mathfrak{A}_p^{p-1} \underline{\mathbf{e}}_p \end{pmatrix}$, and $\otimes$ is the Kronecker product. The eigenvalues of $(\mathbf{I}_p \otimes\ \boldsymbol{\Sigma}_{\boldsymbol{E}})$ are lower bounded by some positive constant which only depends on $c_{\boldsymbol{E}}$. Lemma A.2.2 below implies $\mathbf{G}_p^{-1} = \mathbf{Q}_p$, and $\|\mathbf{G}_p^{-1}\|_2 \leq 1 + \sum_{k=1}^{p-1} \|\mathbf{A}_k\|_2 \leq 1 + \sum_{k=1}^{p-1} a_k < \infty$. Therefore, $\lambda_{\max}\left(\left(\sum_{k=0}^{p-1} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p) \boldsymbol{\Sigma}_{\boldsymbol{E}} (\mathfrak{A}_p^k \underline{\mathbf{e}}_p)^\top\right)^{-1}\right)$ is upper bounded by some positive constant that only depends on $(c_{\boldsymbol{E}}, a_1, \ldots, a_p)$, and we have $\lambda_{\min}(\boldsymbol{\Sigma}_0) \geq c > 0$ by some constant c that only depends on $(c_{\boldsymbol{E}}, a_1, \ldots, a_p)$.

Case 2. We analyze the case where $j \geq p$.

From $\mathfrak{Z}_{t+k} = \mathfrak{E}_{t+k} + \cdots + \mathfrak{A}_p^{k-1}\mathfrak{E}_{t+1} + \mathfrak{A}_p^k \mathfrak{Z}_t$ for $k \in \mathbb{N}$, we have

$$\begin{pmatrix} \boldsymbol{Z}_t \\ \boldsymbol{Z}_{t+1} \\ \vdots \\ \boldsymbol{Z}_{t+p-1} \\ \boldsymbol{Z}_{t+p} \\ \boldsymbol{Z}_{t+p+1} \\ \vdots \\ \boldsymbol{Z}_{t+j} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \dots & \cdots & \cdots & \cdots & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \dots & \cdots & \cdots & \mathbf{0}_{d\times d} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \ddots & \ddots & \mathbf{0}_{d\times d} \\ [\mathfrak{A}_p]_{p1} & [\mathfrak{A}_p]_{p2} & \cdots & [\mathfrak{A}_p]_{pp} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \ddots & \ddots \\ [\mathfrak{A}_p^2]_{p1} & [\mathfrak{A}_p^2]_{p2} & \cdots & [\mathfrak{A}_p^2]_{pp} & [\mathfrak{A}_p]_{pp} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \ddots \\ \vdots & \cdots & \cdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ [\mathfrak{A}_p^{j-p+1}]_{p1} & \cdots & [\mathfrak{A}_p^{j-p+1}]_{12} & [\mathfrak{A}_p^{j-p+1}]_{pp} & \cdots & [\mathfrak{A}_p^2]_{pp} & [\mathfrak{A}_p^1]_{pp} & \mathbf{I}_d \end{pmatrix}}_{\mathbf{L}} \begin{pmatrix} \boldsymbol{Z}_t \\ \boldsymbol{Z}_{t+1} \\ \vdots \\ \boldsymbol{Z}_{t+p-1} \\ \boldsymbol{E}_{t+p} \\ \boldsymbol{E}_{t+p+1} \\ \vdots \\ \boldsymbol{E}_j \end{pmatrix}.$$

From $\boldsymbol{Z}_t - \mathbf{A}_1\boldsymbol{Z}_{t-1} - \cdots - \mathbf{A}_p\boldsymbol{Z}_{t-p} = \boldsymbol{E}_t$, we have

$$\begin{pmatrix} \boldsymbol{Z}_t \\ \boldsymbol{Z}_{t+1} \\ \vdots \\ \boldsymbol{Z}_{t+p-1} \\ \boldsymbol{E}_{t+p} \\ \boldsymbol{E}_{t+p+1} \\ \vdots \\ \boldsymbol{E}_j \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{I}_d & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \cdots & \cdots & \cdots & \cdots & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & \mathbf{I}_d & \mathbf{0}_{d\times d} & \ddots & \ddots & \ddots \\ -\mathbf{A}_p & -\mathbf{A}_{p-1} & \dots & -\mathbf{A}_1 & \mathbf{I}_d & \mathbf{0}_{d\times d} & \ddots & \ddots \\ \mathbf{0}_{d\times d} & -\mathbf{A}_p & -\mathbf{A}_{p-1} & \cdots & -\mathbf{A}_1 & \mathbf{I}_d & \mathbf{0}_{d\times d} & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \mathbf{0}_{d\times d} & \cdots & \mathbf{0}_{d\times d} & -\mathbf{A}_p & -\mathbf{A}_{p-1} & \cdots & -\mathbf{A}_1 & \mathbf{I}_d \end{pmatrix}}_{\mathbf{L}^{-1}} \begin{pmatrix} \boldsymbol{Z}_t \\ \boldsymbol{Z}_{t+1} \\ \vdots \\ \boldsymbol{Z}_{t+p-1} \\ \boldsymbol{Z}_{t+p} \\ \boldsymbol{Z}_{t+p+1} \\ \vdots \\ \boldsymbol{Z}_{t+j} \end{pmatrix}.$$

Then, it suffices to show that $\|\mathbf{L}\|_2$ and $\|\mathbf{L}^{-1}\|_2$ are bounded from above by some absolute positive constants because Case 1 implies that the eigenvalues of $\mathrm{Var}((\boldsymbol{Z}_t^\top, \dots, \boldsymbol{Z}_{t+p-1}^\top, \boldsymbol{E}_{t+p}^\top, \dots, \boldsymbol{E}_{t+j}^\top)^\top)$ are lower and upper bounded by some positive constants that only depend on $(c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, a_1, \dots, a_p, p)$. We have $\|\mathbf{L}^{-1}\|_2 \le 1 + \sum_{k=1}^p \|\mathbf{A}_k\|_2 \le 1 + a_1 + \dots + a_p < \infty$, and

$$\|\mathbf{L}\|_2 \le 1 + \sum_{k=1}^{j-p+1} \|\mathfrak{A}_p^k\|_2 + (p-1) \sum_{k=1}^{j-p+1} \|\mathfrak{A}_p^k\|_2 \le 1 + p \sum_{k=1}^{\infty} \|\overline{\mathfrak{A}_p}^k\|_2 \le C_{\mathbf{L}} < \infty,$$

where $C_{\mathbf{L}}$ is a positive constant that only depends on $(a_1, \ldots, a_p, p)$. Therefore, $0 < c \leq \lambda_{\min}(\boldsymbol{\Omega}_j) \leq \lambda_{\max}(\boldsymbol{\Omega}_j) \leq C < \infty$ by constants c and C which only depend on $(c_{\boldsymbol{E}}, C_{\boldsymbol{E}}, a_1, \ldots, a_p, p)$.

Technical lemmas used in Appendix A.2

Lemma A.2.2. *Let* $\mathbf{G}_p = \begin{pmatrix} \underline{\mathbf{e}}_p & \mathfrak{A}_p \underline{\mathbf{e}}_p & \cdots & \mathfrak{A}_p^{p-1} \underline{\mathbf{e}}_p \end{pmatrix} \in \mathbb{R}^{pd \times pd}$*. For any* $p \in \mathbb{N}$*, the inverse* $\mathbf{G}_p^{-1}$ *exists and* $\mathbf{G}_p^{-1} = \mathbf{Q}_p$*, where* $\mathbf{Q}_p$ *is defined in Section A.2.3.*

Proof. Throughout the proof, it is noted that we can represent $\mathfrak{A}_p$ and $\mathbf{Q}_p$ recursively as follows,

$$\mathfrak{A}_p = \begin{pmatrix} \mathbf{0}_{d \times d} & \bar{\mathbf{e}}_{p-1}^{\top} \\ \underline{\mathbf{e}}_{p-1} \mathbf{A}_p & \mathfrak{A}_{p-1} \end{pmatrix} \text{ and } \mathbf{Q}_p = \begin{pmatrix} -\mathbf{C}_{p-1} & \mathbf{Q}_{p-1} \\ \mathbf{I}_d & \mathbf{0}_{d \times (p-1)d} \end{pmatrix} \text{ with } \mathfrak{A}_1 = \mathbf{A}_1 \text{ and } \mathbf{Q}_1 = \mathbf{I}_d.$$

With this information, we will prove this lemma with two steps. In Step 1, we will show that $\mathbf{G}_p$ can be written recursively, and we have $\mathbf{Q}_{p-1} \mathfrak{A}_{p-1}^{p-1} \underline{\mathbf{e}}_{p-1} = \mathbf{C}_{p-1}$. In Step 2, using the result in Step 1 and the information that $\mathbf{Q}_p$ can be represented recursively, we will prove this lemma using mathematical induction.

Step 1. We will show that $\mathbf{G}_p$ can be written as $\mathbf{G}_p = \begin{pmatrix} \mathbf{0}_{d \times (p-1)d} & \mathbf{I}_d \\ \mathbf{G}_{p-1} & \mathfrak{A}_{p-1}^{p-1} \underline{\mathbf{e}}_{p-1} \end{pmatrix}$ for any $p \geq 2$, and we have $\mathbf{Q}_{p-1} \mathfrak{A}_{p-1}^{p-1} \underline{\mathbf{e}}_{p-1} = \mathbf{C}_{p-1}$.

For simplicity of notation, we partition $\mathfrak{A}_p^{k-1} \underline{\mathbf{e}}_p \in \mathbb{R}^{pd \times d}$ with $k \in [p]$ to two blocks so that $\mathfrak{A}_p^{k-1} \underline{\mathbf{e}}_p = \begin{pmatrix} \mathbf{D}_{p,k} \\ \mathbf{F}_{p,k} \end{pmatrix}$, where $\mathbf{D}_{p,k}$ is a d by d matrix and $\mathbf{F}_{p,k}$ is a $(p-1)d$ by d matrix. Then, we have

$$\mathbf{G}_p = \begin{pmatrix} \underline{\mathbf{e}}_p & \mathfrak{A}_p \underline{\mathbf{e}}_p & \cdots & \mathfrak{A}_p^{p-1} \underline{\mathbf{e}}_p \end{pmatrix} = \begin{pmatrix} \mathbf{D}_{p,1} & \cdots & \mathbf{D}_{p,p-1} & \mathbf{D}_{p,p} \\ \mathbf{F}_{p,1} & \cdots & \mathbf{F}_{p,p-1} & \mathbf{F}_{p,p} \end{pmatrix}.$$

In order to finish Step 1, it is sufficient to show that (1) $\mathbf{D}_{p,k} = \mathbf{0}_{d \times d}$ and $\mathbf{F}_{p,k} = \mathfrak{A}_{p-1}^{k-1} \underline{\mathbf{e}}_{p-1}$ for $k \in [p-1]$; and (2) $\mathbf{D}_{p,p} = \mathbf{I}_d$, $\mathbf{F}_{p,p} = \mathfrak{A}_{p-1}^{p-1} \underline{\mathbf{e}}_{p-1}$, and $\mathbf{Q}_{p-1} \mathbf{F}_{p,p} = \mathbf{C}_{p-1}$.

Note that we have $\mathbf{D}_{p,k} = \bar{\mathbf{e}}_{p-1}^\top \mathbf{F}_{p,k-1}$ and $\mathbf{F}_{p,k} = \underline{\mathbf{e}}_{p-1}\mathbf{A}_p\mathbf{D}_{p,k-1} + \mathfrak{A}_{p-1}\mathbf{F}_{p,k-1}$ with $\mathbf{D}_{p,1} = \mathbf{0}_{d\times d}$ and $\mathbf{F}_{p,1} = \underline{\mathbf{e}}_{p-1}$. This is because

$$\mathfrak{A}_p^{k-1}\underline{\mathbf{e}}_p = \mathfrak{A}_p\mathfrak{A}_p^{k-2}\underline{\mathbf{e}}_p = \begin{pmatrix} \mathbf{0}_{d\times d} & \bar{\mathbf{e}}_{p-1}^\top \\ \underline{\mathbf{e}}_{p-1}\mathbf{A}_p & \mathfrak{A}_{p-1} \end{pmatrix} \begin{pmatrix} \mathbf{D}_{p,k-1} \\ \mathbf{F}_{p,k-1} \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{e}}_{p-1}^\top \mathbf{F}_{p,k-1} \\ \underline{\mathbf{e}}_{p-1}\mathbf{A}_p\mathbf{D}_{p,k-1} + \mathfrak{A}_{p-1}\mathbf{F}_{p,k-1} \end{pmatrix}.$$

With this information, it is possible to show that (1) and (2) are true for any $p \geq 2$ using mathematical induction as follows.

Base Step. When $p = 2$, we have $\mathbf{D}_{2,1} = \mathbf{0}_{d\times d}$, $\mathbf{F}_{2,1} = \underline{\mathbf{e}}_1 = \mathfrak{A}_1^0\underline{\mathbf{e}}_1$,

$$\mathbf{D}_{2,2} = \bar{\mathbf{e}}_1^\top\mathbf{F}_{2,1} = \bar{\mathbf{e}}_1^\top\underline{\mathbf{e}}_1 = \mathbf{I}_d, \mathbf{F}_{2,2} = \underline{\mathbf{e}}_1\mathbf{A}_2\mathbf{D}_{2,1} + \mathfrak{A}_1\mathbf{F}_{2,1} = \mathfrak{A}_1\underline{\mathbf{e}}_1, \text{ and } \mathbf{Q}_1\mathbf{F}_{2,2} = \mathbf{I}_d\mathfrak{A}_1\underline{\mathbf{e}}_1 = \mathbf{A}_1 = \mathbf{C}_1.$$

Therefore, (1) and (2) are true when $p = 2$.

Induction Step. Suppose (1) and (2) hold when $p = m$. Then, the following three steps will show that (1) and (2) are true for $p = m + 1$.

(i) We have $\mathbf{D}_{m+1,1} = \mathbf{0}_{d\times d}$ and $\mathbf{F}_{m+1,1} = \underline{\mathbf{e}}_m = \mathfrak{A}_m^0\underline{\mathbf{e}}_m$.

(ii) Suppose we have that $\mathbf{D}_{m+1,j} = \mathbf{0}_{d\times d}$ and $\mathbf{F}_{m+1,j} = \mathfrak{A}_m^{j-1}\underline{\mathbf{e}}_m$ for $j \in [m-1]$. Then, we have

$$\mathbf{D}_{m+1,j+1} = \bar{\mathbf{e}}_m^\top\mathbf{F}_{m+1,j} = \bar{\mathbf{e}}_m^\top\mathfrak{A}_m^{j-1}\underline{\mathbf{e}}_m = \begin{pmatrix} \mathbf{I}_{d\times d} & \mathbf{0}_{d\times(m-1)d} \end{pmatrix} \begin{pmatrix} \mathbf{D}_{m,j} \\ \mathbf{F}_{m,j} \end{pmatrix} = \mathbf{D}_{m,j} = \mathbf{0}_{d\times d} \text{ and}$$

$$\mathbf{F}_{m+1,j+1} = \underline{\mathbf{e}}_m\mathbf{A}_{m+1}\mathbf{D}_{m+1,j} + \mathfrak{A}_m\mathbf{F}_{m+1,j} = \mathfrak{A}_m\mathfrak{A}_m^{j-1}\underline{\mathbf{e}}_m = \mathfrak{A}_m^j\underline{\mathbf{e}}_m.$$

The results of (i) and (ii) combined imply that $\mathbf{D}_{m+1,k} = \mathbf{0}_{d\times d}$ and $\mathbf{F}_{m+1,k} = \mathfrak{A}_m^{k-1}\underline{\mathbf{e}}_m$ for $k \in [m]$.

(iii) From (i),(ii), and the assumption in the induction step, we have

$$\mathbf{D}_{m+1,m+1} = \bar{\mathbf{e}}_m^\top\mathbf{F}_{m+1,m} = \bar{\mathbf{e}}_m^\top\mathfrak{A}_m^{m-1}\underline{\mathbf{e}}_m = \begin{pmatrix} \mathbf{I}_{d\times d} & \mathbf{0}_{d\times(m-1)d} \end{pmatrix} \begin{pmatrix} \mathbf{D}_{m,m} \\ \mathbf{F}_{m,m} \end{pmatrix} = \mathbf{D}_{m,m} = \mathbf{I}_d,$$

$$\mathbf{F}_{m+1,m+1} = \underline{\mathbf{e}}_m\mathbf{A}_{m+1}\mathbf{D}_{m+1,m} + \mathfrak{A}_m\mathbf{F}_{m+1,m} = \mathfrak{A}_m\mathfrak{A}_m^{m-1}\underline{\mathbf{e}}_m = \mathfrak{A}_m^m\underline{\mathbf{e}}_m, \text{ and}$$

$$\mathbf{Q}_m\mathbf{F}_{m+1,m+1} = (\mathbf{Q}_m\mathfrak{A}_m)(\mathfrak{A}_m^{m-1}\underline{\mathbf{e}}_m) = \begin{pmatrix} \mathbf{A}_m & \mathbf{0}_{d\times(m-1)d} \\ \mathbf{0}_{(m-1)d\times d} & \mathbf{Q}_{m-1} \end{pmatrix} \begin{pmatrix} \mathbf{D}_{m,m} \\ \mathbf{F}_{m,m} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_m \\ \mathbf{C}_{m-1} \end{pmatrix} = \mathbf{C}_m.$$

The results of (i), (ii), and (iii) combined imply that (1) and (2) are also true for $p = m + 1$. In conclusion, by mathematical induction, we have the desired result.

Step 2. We prove $\mathbf{G}_p^{-1} = \mathbf{Q}_p$ using mathematical induction as follows.

Base Step. When $p = 1$, we have $\mathbf{G}_p^{-1} = \mathbf{Q}_p$ because $\mathbf{G}_1 = \mathbf{I}_d$ and $\mathbf{Q}_1 = \mathbf{I}_d$.

Induction Step. Suppose $\mathbf{G}_p^{-1} = \mathbf{Q}_p$ is true for $p = m$, where $m \geq 1$. Using the result in Step 1, we have

$$\mathbf{Q}_{m+1}\mathbf{G}_{m+1} = \begin{pmatrix} -\mathbf{C}_m & \mathbf{Q}_m \\ \mathbf{I}_d & \mathbf{0}_{d\times md} \end{pmatrix} \begin{pmatrix} \mathbf{0}_{d\times md} & \mathbf{I}_d \\ \mathbf{G}_m & \mathbf{F}_{m+1,m+1} \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_m\mathbf{G}_m & \mathbf{Q}_m\mathbf{F}_{m+1,m+1} - \mathbf{C}_m \\ \mathbf{0}_{d\times md} & \mathbf{I}_d \end{pmatrix} = \mathbf{I}_{(m+1)d}.$$

The last equality holds because Step 1 implies $\mathbf{Q}_m\mathbf{F}_{m+1,m+1} = \mathbf{C}_m$. This shows that $\mathbf{G}_p^{-1} = \mathbf{Q}_p$ is true for $p = m + 1$.

Therefore, by mathematical induction, we have the desired result. $\square$

Appendix B

MINIMUM SLICED DISTANCE ESTIMATION IN STRUCTURAL MODELS

B.1 A Brief Review of Hadamard Differentiability and Mixing Conditions

First, we will review the definition of Hadamard differentiability and related results (c.f, Appendix A of Chen and Fang (2019), and Shapiro (2000).)

Let $\mathbb{D}$ and $\mathbb{K}$ be normed spaces equipped with norms $\|\|_{\mathbb{D}}$ and $\|\|_{\mathbb{K}}$ respectively. For a functional $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{K}$, we define Hadamard directional differentiability as follows.

Definition B.1.1 (First-order Hadamard directional differentiability)**.** *We say that the functional ϕ is Hadamard differentiable at $G \in \mathbb{D}_\phi$ tangentially to a set $\mathbb{D}_0 \subset \mathbb{D}$, if there is a continuous map $\phi'_G : \mathbb{D} \to \mathbb{K}$ such that:*

$$\left\| \frac{\phi(G + t_n h_n) - \phi(G)}{t_n} - \phi_G(h) \right\|_{\mathbb{K}} \to 0 \text{ as } n \to \infty$$

for any sequences $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbb{R}_+$ such that $t_n \downarrow 0$, $h_n \to h \in \mathbb{D}_0$ as $n \to \infty$; and $G + t_n h_n \in \mathbb{D}_\phi$.

Definition B.1.2 (Second-order Hadamard directional differentiability)**.** *Suppose ϕ is Hadamard directionally differentiable tangetly to a set $\mathbb{D}_0 \subset \mathbb{D}$ such that the derivative $\phi'_G : \mathbb{D}_0 \to \mathbb{K}$ is well defined on $\mathbb{D}$. We say that ϕ is second-order Hadamard directionally differentiable at $G \in \mathbb{D}_\phi$ tangently to $\mathbb{D}_0$ if there is a map $\phi''_G : \mathbb{D}_0 \to \mathbb{K}$ such that*

$$\left\| \frac{\phi(G + t_n h_n) - \phi(G) - t_n \phi_G(h_n)}{t_n^2} - \phi''_G(h) \right\|_{\mathbb{K}} \to 0 \text{ as } n \to \infty$$

for any sequences $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbb{R}_+$ such that $t_n \downarrow 0$, $h_n \to h \in \mathbb{D}_0$ as $n \to \infty$; and $G + t_n h_n \in \mathbb{D}_\phi$.

Condition B.1.1 (Assumptions 2.1 and 2.2 in Chen and Fang (2019)).

1 *(i)* $\mathbb{D}$ *and* $\mathbb{K}$ *are normed space with norms* $||\cdot||_{\mathbb{D}}$ *and* $||\cdot||_{\mathbb{K}}$*, respectively. (ii)* $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E}$ *is second-order Hadamard directionally differentiable at* $\theta_0 \in \mathbb{D}_\phi$ *tangetly to* $\mathbb{D}_0 \in \mathbb{D}$*; (iii)* $\phi'_{\theta_0}(h) = 0$ *for all* $h \in \mathbb{D}_0$*.*

2 *(i) There is* $\hat{\theta}_n$ *such that* $r_n(\hat{\theta}_n - \theta) \xrightarrow{L} \mathbb{G}$ *in* $\mathbb{D}$ *for some* $r_n \uparrow \infty$*; (ii)* $\mathbb{G}$ *is tight and its support is in* $\mathbb{D}_0$*. (iii)* $\mathbb{D}_0$ *is closed under vector addition, (i.e.,* $h_1 + h_2 \in \mathbb{D}_0$ *whenever* $h_1, h_2 \in \mathbb{D}_0$*).*

Lemma B.1.1 (Theorem 2.1 of Chen and Fang (2019)). *If Conditions B.1.1 (1.i), (1.ii), (2.i), and (2.ii) hold, then*

$$r_n^2(\phi(\hat{\theta}_n) - \hat{\theta} - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)) = \phi''_{\theta_0}(r_n(\hat{\theta}_n - \theta)) + o_p(1)$$

Therefore, we have

$$r_n^2(\phi(\hat{\theta}_n) - \hat{\theta} - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)) \xrightarrow{L} \phi''_{\theta_0}(\mathbb{G}).$$

In addition, if Condition B.1.1 (1.iii) holds, then we have

$$r_n^2(\phi(\hat{\theta}_n) - \hat{\theta}) \xrightarrow{L} \phi''_{\theta_0}(\mathbb{G}).$$

Here, we will review some useful inequalities related to mixing coefficients.

Definition B.1.3 (c.f., Bradley (2005)). *Let us consider the probability space* $(\Omega, \mathcal{F}, P)$*. For any sigma-field* $\mathcal{A} \subset \mathcal{F}$*, we denote* $\mathcal{L}^2(\mathcal{A})$ *to be the space of square-integrable,* $\mathcal{A}$ *-measurable random variables.*

Let $\mathcal{A}$ and $\mathcal{B} \subset \mathcal{F}$ be two sigma fields We define

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \ \mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

$$\varrho(\mathcal{A}, \mathcal{B}) = \sup_{f \in \mathcal{L}^2(\mathcal{A}), g \in \mathcal{L}^2(\mathcal{B})} |corr(f, g)|,$$

$$\varphi(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}, \mathbb{P}(A) > 0} |\mathbb{P}(B|A) - \mathbb{P}(B)|,$$

$$\beta(\mathcal{A}, \mathcal{B}) = \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|$$

where supremum is taken over all pairs of (finite) partitions $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_J\}$ of Ω such that $A_i \in \mathcal{A}$ for each i and $B_j \in \mathcal{B}$ for each j.

Suppose we have X_t for $t \in \mathbb{Z}$, where $\mathbb{Z}$ is a set of integers. Let's denote $\mathcal{F}_J^L = \sigma(X_k, J \leq k \leq L, k \in \mathbb{Z})$ to be sigma-field generated by random variable X_t where $J \leq t \leq L$. Then, the random variable X_t is β -mixing if $\beta_k := \sup_{j \in \mathbb{Z}} \beta(\mathcal{F}_{-\infty}^j,\ \mathcal{F}_{j+k}^\infty) \to 0$ as $n \to \infty$.

For any two σ-fields $\mathcal{A}$ and $\mathcal{B}$, we have

$$2\alpha(\mathcal{A}, \mathcal{B}) \leq \beta(\mathcal{A}, \mathcal{B}) \leq \varphi(\mathcal{A}, \mathcal{B}) \text{ and } 4\alpha(\mathcal{A}, \mathcal{B}) \leq \varrho(\mathcal{A}, \mathcal{B}).$$

(see Bradley (2005).)

Lemma B.1.2 (Lemma I1 in Pötscher and Prucha (1997))**.** *Let X be a $\mathcal{D}$-measurable random variable, and Y be a $\mathcal{K}$-measurable random variable such that for $1 \leq p \leq s \leq \infty$ and $1/p + 1/q = 1$ we have $(\mathbb{E}[\|X\|^p])^{1/p} < \infty$ and $(\mathbb{E}[\|Y\|^q])^{1/q} < \infty$. Then,*

$$|cov(X, Y)| \leq 2(2^{1/p} + 1)\alpha(\mathcal{D}, \mathcal{K})^{1/p - 1/s}(E[||X||^s])^{1/s}(E[||Y||^q])^{1/q}$$

$$|cov(X, Y)| \leq 2\phi(\mathcal{D}, \mathcal{K})^{1-1/s}(E[||X||^s])^{1/s}(E[||Y||^q])^{1/q}.$$

B.2 *Proofs in Section 2.5*

For simplicity of notation, we denote

$$\begin{aligned}
\mathcal{M}_T(\psi) &= \left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q_T(s;u) - \widehat{Q}_T(s;u,\psi))^2 w(s)\mathrm{d}s d\varsigma(u)\right)^{1/2},\\
\mathcal{M}(\psi) &= \left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q(s;u) - Q(s;u,\psi))^2 w(s)\mathrm{d}s d\varsigma(u)\right)^{1/2},\\
\widehat{\mathcal{SW}}_T &= \left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q_T(s;u) - Q(s;u))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2},\\
\widehat{\mathcal{SW}}_T(\psi) &= \left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(\widehat{Q}_T(s,u,\psi) - Q(s,u,\psi))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2},\\
\overline{B}_T &= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\|\widehat{D}_T(s;u,\psi_0)\|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u),\\
\overline{B}_0 &= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\|D(s;u,\psi_0)\|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u).
\end{aligned}$$

B.2.1 *Proof of Theorem 2.5.1 (consistency)*

We can show the consistency by following standard approaches in the extremum estimator. Since $\hat{\psi}_T$ satisfies $(\mathcal{M}_T(\hat{\psi}_T))^2 \leq \inf_{\psi\in\Psi}(\mathcal{M}_T(\psi))^2 + T^{-1}\epsilon_T$, we have

$$\mathcal{M}_T(\hat{\psi}_T) \leq \inf_{\psi\in\Psi}\mathcal{M}_T(\psi) + T^{-1/2}\epsilon_T^{1/2} = \inf_{\psi\in\Psi}\mathcal{M}_T(\psi) + o_p(1) \leq \mathcal{M}(\psi_0) + o_p(1).$$

By the triangle inequality of sliced Wasserstein distance, and Minkowski inequality (c.f., Nadjahi et al. (2020a)), we can show that $\sup_{\psi\in\Psi}|\mathcal{M}_T(\psi) - \mathcal{M}(\psi)|$ under Assumption 2.5.1

as follows:

$$
\begin{aligned}
&\sup_{\psi\in\Psi}|\mathcal{M}_T(\psi)-\mathcal{M}(\psi)| \\
&=\sup_{\psi\in\Psi}\left|\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q_T(s;u)-\widehat{Q}_T(s;u,\psi))^2w(s)\mathrm{d}sd\varsigma(u)\right)^{1/2}\right. \\
&\qquad\qquad \left.-\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q(s;u)-Q(s;u,\psi))^2w(s)\mathrm{d}sd\varsigma(u)\right)^{1/2}\right| \\
&\le \sup_{\psi\in\Psi}\left(\int_{\mathbb{S}^{d-1}}\int_{S}(Q_T(s;u)-Q(s;u)-\widehat{Q}_T(s;u,\psi)+Q(s;u,\psi))^2w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2} \\
&\le \widehat{SW}_T+\sup_{\psi\in\Psi}\widehat{SW}_T(\psi)\xrightarrow{p}0.
\end{aligned}
$$

Assumption 2.5.2 implies that for each $\epsilon>0$, there exists $\delta>0$ such that for any $\psi\notin B(\psi_0,\epsilon)\Rightarrow \mathcal{M}(\psi)-\mathcal{M}(\psi_0)\ge\delta>0$.

Using these results, we have

$$
\begin{aligned}
\Pr(\hat{\psi}_T\notin B(\psi_0,\epsilon)) &\le \Pr(\mathcal{M}(\hat{\psi}_T)-\mathcal{M}(\psi_0)\ge\delta) \\
&=\Pr(\mathcal{M}(\hat{\psi}_T)-\mathcal{M}_T(\hat{\psi}_T)+\quad \mathcal{M}_T(\hat{\psi}_T)-\mathcal{M}(\psi_0)\ge\delta) \\
&\le \Pr(\mathcal{M}(\hat{\psi}_T)-\mathcal{M}_T(\hat{\psi}_T)+\mathcal{M}_T(\psi_0)+o_p(1)-\mathcal{M}(\psi_0)\ge\delta) \\
&\le \Pr(2\sup_{\psi\in\Psi}|\mathcal{M}_T(\psi)-\mathcal{M}(\psi)|+o_p(1)\ge\delta) \\
&\to 0.
\end{aligned}
$$

Therefore, $\hat{\psi}_T\xrightarrow{p}\psi_0$.

B.2.2 Proof of Theorem 2.5.2 (asymptotic distribution)

Here, we will verify Assumptions 1 to 6 in Andrews (1999). When ψ_0 is in the interior of Ψ, Assumptions 5 and 6 in Andrews (1999) hold. Therefore, it is enough to prove Assumptions 1 to 4 in Andrews (1999).

First, Assumption 1 holds in Andrews (1999) are satisfied under Assumptions 2.5.1 and 2.5.2 by Theorem 4.1 in Andrews (1999).

Let us remind that

$$\widehat{R}_T(s;u,\psi,\psi_0) := \widehat{Q}_T(s;u,\psi) - \widehat{Q}_T(s;u,\psi_0) - (\psi-\psi_0)^\top \widehat{D}_T(s;u,\psi_0)$$

with

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \left| \frac{T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)dsd\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} \right| = o_p(1)$$

for any $\tau_T \to 0$.

The objective function can be decomposed as follows:

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u) - \widehat{Q}_T(s;u,\psi))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u) - \widehat{Q}_T(s;u,\psi_0))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) - 2(\psi-\psi_0)^\top A_T/\sqrt{T} + (\psi-\psi_0)^\top B_T(\psi-\psi_0) + R_T,
\end{aligned}$$

where

$$\begin{aligned}
A_T &= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u) - \widehat{Q}_T(s;u,\psi_0))\widehat{D}_T(s;u,\psi_0) w(s)\mathrm{d}s\mathrm{d}\varsigma(u);\\
B_T &= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \widehat{D}_T(s;u,\psi_0)\widehat{D}_T^\top(s;u,\psi_0) w(s)\mathrm{d}s d\varsigma(u);\\
R_T &= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \widehat{R}_T^2(s;u,\psi,\psi_0) w(s)\mathrm{d}s\mathrm{d}\varsigma(u) - 2\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u) - \widehat{Q}_T(s;u,\psi_0))\widehat{R}_T(s;u,\psi,\psi_0) w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad + 2(\psi-\psi_0)^\top \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \widehat{D}_T(s;u,\psi_0)\widehat{R}_T(s;u,\psi,\psi_0) w(s)\mathrm{d}s\mathrm{d}\varsigma(u).
\end{aligned}$$

Then, it is enough to show

$$A_T \xrightarrow{d} N(0,V_0),\ B_T \xrightarrow{p} B_0,\ \text{and} \sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{T|R_T|}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} \xrightarrow{p} 0$$

for any $\tau_T \to 0$ as $T \to \infty$.

This is because the first two conditions satisfy Assumption 3 in Andrews (1999), and the third condition implies Assumption 2* in Andrews (1999). Assumptions 1, 2*, and 3 Andrews (1999) imply Assumption 4 in Andrews (1999) by Theorem 1 in Andrews (1999).

In Steps 1 to 3 below, we will analyze the behaviors of A_T, B_T, and R_T, respectively.

Step 1: We will analyze the behavior of A_T.

Note that $Q(\cdot,\cdot) = Q(\cdot,\cdot,\psi_0)$.

$$\begin{aligned}
A_T &= \sqrt{T}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q_T(s;u) - \widehat{Q}_T(s;u,\psi_0))\widehat{D}_T(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&= \sqrt{T}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q_T(s;u) - Q(s;u))\widehat{D}_T(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad + \sqrt{T}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q(s;u,\psi_0) - \widehat{Q}_T(s;u,\psi_0))\widehat{D}_T(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)
\end{aligned}$$

Under Assumption 2.5.4, we have

$$\sqrt{T}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q_T(s;u) - Q(s;u))(\widehat{D}_T(s;u,\psi_0) - D(s;u,\psi_0))w(s)\mathrm{d}s\mathrm{d}\varsigma(u) = o_p(1)$$
$$\sqrt{T}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q(s;u,\psi_0) - \widehat{Q}_T(s;u,\psi_0))(\widehat{D}_T(s;u,\psi_0) - D(s;u,\psi_0))w(s)\mathrm{d}s\mathrm{d}\varsigma(u) = o_p(1).$$

Therefore, we have

$$\begin{aligned}
A_T &= \sqrt{T}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q_T(s;u) - Q(s;u))D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad + \sqrt{T}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q(s;u,\psi_0) - \widehat{Q}_T(s;u,\psi_0))D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u) + o_p(1),
\end{aligned}$$

which is asymptotically normal under Assumption 2.5.5.

Step 2: We will analyze the behavior of B_T.

Note that

$$B_T = \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\widehat{D}_T(s;u,\psi_0)\widehat{D}_T^\top(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)$$

Let us denote

$$B_0 = \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}D(s;u,\psi_0)D^\top(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u).$$

Then, we have

$$\begin{aligned}
B_T - B_0 &= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\widehat{D}_T(s;u,\psi_0)\widehat{D}_T^\top(s;u,\psi_0)w(s)\mathrm{d}s d\varsigma(u) - \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}D(s;u,\psi_0)D^\top(s;u,\psi_0)w(s)\mathrm{d}s d\varsigma(u)\\
&= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(\widehat{D}_T(s;u,\psi_0) - D(s;u,\psi_0))\widehat{D}_T^\top(s;u,\psi_0)w(s)\mathrm{d}s d\varsigma(u)\\
&\quad + \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}D(s;u,\psi_0)(\widehat{D}_T(s;u,\psi_0) - D(s;u,\psi_0))^\top w(s)\mathrm{d}s\mathrm{d}\varsigma(u),
\end{aligned}$$

which is $o_p(1)$ when Assumption 2.5.4 (iii) holds.

Step 3: We will analyze the behavior of R_T.

Here, we will show

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \frac{T|R_T|}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} = o_p(1).$$

Let us remind that

$$\begin{aligned} R_T = &\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \\ &- 2\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u)-\widehat{Q}_T(s;u,\psi_0))\widehat{R}_T(s;u,\psi,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \\ &+ 2(\psi-\psi_0)^\top \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \widehat{D}_T(s;u,\psi_0)\widehat{R}_T(s;u,\psi,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u), \end{aligned}$$

where

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \left| \frac{T\int_{\mathbb{S}^{d-1}}\int \left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)ds d\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} \right| = o_p(1)$$

for any $\tau_T \to 0$.

Note that

$$\begin{aligned} |R_T| \leq &\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) + 2\mathcal{M}_T(\psi_0)\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2} \\ &+ 2\|\psi-\psi_0\|_\Psi \left(\overline{B}_T\right)^{1/2}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}. \end{aligned}$$

This implies that

$$\begin{aligned} &\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \frac{T|R_T|}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} \\ &\leq \sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \frac{T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} \\ &\quad + 2\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \frac{T\mathcal{M}_T(\psi_0)\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} \\ &\quad + 2\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \frac{T\|\psi-\psi_0\|\left(\overline{B}_T\right)^{1/2}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{n}(\psi-\psi_0)\|)^2}. \end{aligned}$$

First, we have

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_*\|\le\tau_T} \frac{T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_*)\|)^2} = o_p(1)$$

by Assumption 2.5.3.

Second, we have

$$\begin{aligned}
&\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{T\mathcal{M}_T(\psi_0)\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\\
&\le \sqrt{T}\mathcal{M}_T(\psi_0)\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{\sqrt{T}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\\
&\le \sqrt{T}\mathcal{M}_T(\psi_0)\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{\sqrt{T}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{1+\|\sqrt{T}(\psi-\psi_0)\|}\\
&= \sqrt{T}\mathcal{M}_T(\psi_0)\left(\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\right)^{1/2}\\
&= O_p(1)o_p(1) = o_p(1).
\end{aligned}$$

Here, the second last equality holds because Assumptions 2.5.4 implies

$$\sqrt{T}\mathcal{M}_T(\psi_0) \le \sqrt{T}\widehat{\mathcal{SW}}_T + \sqrt{T}\widehat{\mathcal{SW}}_T(\psi_0) = O_p(1),$$

and Assumption 2.5.3 implies

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} = o_p(1).$$

Third, we have

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{T\|\psi-\psi_0\|\left(\overline{B}_T\right)^{1/2}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}$$

$$= \left(\overline{B}_T\right)^{1/2} \sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{T\|\psi-\psi_0\|\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}$$

$$= \left(\overline{B}_T\right)^{1/2} \sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{\sqrt{T}\|\psi-\psi_0\|\left(T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_n(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)(1+\|\sqrt{n}(\psi-\psi_0)\|)}$$

$$= O_p(1)o_p(1) = o_p(1).$$

Therefore, we have

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \frac{T|R_T|}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} = o_p(1).$$

Therefore, Assumptions 1 to 6 in Andrews (1999) hold, and we can apply the Theorem 3 in Andrews (1999). This implies

$$\sqrt{T}(\hat{\psi}_T-\psi_0) = B_T^{-1}A_T + o_p(1) \xrightarrow{d} B_0^{-1}N(0,\Omega_0),$$

where $\Omega_0 = (e_1^\top, -e_1^\top)V_0\begin{pmatrix} e_1 \\ -e_1 \end{pmatrix}$ in which $e_1 = (1,\ldots,1)^\top$ is a d_ψ by 1 vector of ones.

B.3 Proofs in Section 2.6

B.3.1 Preliminary Lemmas

Lemma B.3.1 (Corollary 2.1 of Arcones and Yu (1994))**.** *Suppose that $\mathcal{F}$ is a measurable uniformly bounded VC subgraph class of functions. If the β-mixing coefficient of the stationary sequence satisfies $\lim_{k\to 0} k^r\beta_k = 0$ for some $r > 1$, then there is a Gaussian process $[(G(f)) < \infty]_{f\in\mathcal{F}}$ which has a version with uniformly bounded and uniformly continuous path* *with respect to the* $\|\cdot\|_2$*-norm such that*

$$\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(f(X_t)-Pf)\right\}_{f\in\mathcal{F}} \xrightarrow{w} \{G(f)\}_{f\in\mathcal{F}} \text{ in } \ell_\infty(\mathcal{F}).$$

Lemma B.3.2 (Lemma A.3 in Bobkov and Ledoux (2019))**.** *Given a cumulative distribution* F*, the followings hold for all* $0<t<s<1$ *and* $x\in\mathbb{R}$*.*

(1) $F^{-1}(t)\le x$ *if and only if* $F(x)\ge t$*.*

(2) $F^{-1}(t)>x$ *if and only if* $F(x)<t$*.*

(3) $F^{-1}(t)\le x<F^{-1}(s)$ *if and only if* $t\le F(x)<s$*.*

(4) $F(F^{-1}(t))\ge t$ *with equality if and only if* $t=F(y)$ *for some* $y\in\mathbb{R}$*. In particular,*

(5) $F(F^{-1}(t))=t$ *if* F *is continuous.*

(6) $F^{-1}(F(x))\le x$ *if and only if* $x\ge F^{-1}(0)$*. Moreover,*

(7) $F^{-1}(F(x))<x$ *if and only if* $F(y)=F(x)$ *for some* $y<x$ *(*$x>F^{-1}(0)$*).*

(8) F^{-1} *is strictly increasing on* $(0,1)$ *if and only if* F *is continuous.*

Lemma B.3.3. *Let* $F(\cdot)$ *and* $G(\cdot)$ *are univariate distribution functions. When at least one of them is continuous, we have*

$$\sup_s |F(G^{-1}(s))-s| = \sup_s |G(F^{-1}(s))-s|.$$

Proof. Suppose the statement is not true. Then, we have

$$\sup_s |F(G^{-1}(s))-s| < \sup_s |G(F^{-1}(s))-s| \tag{B.1}$$

or

$$\sup_s |F(G^{-1}(s))-s| > \sup_s |G(F^{-1}(s))-s|. \tag{B.2}$$

Here, we will only prove that inequality (B.1) will result in the contradiction because we can show that the inequality (B.2) does not hold in the same way.

Suppose inequality (B.1) is true. Then, there exists $0 < s_1 < 1$ such that

$$\sup_s |F(G^{-1}(s)) - s| < |G(F^{-1}(s_1)) - s_1|.$$

Let us denote $s_2 = G(F^{-1}(s_1))$. Then, $s_2 \neq s_1$ because $\sup_s |F(G^{-1}(s)) - s| \geq 0$. Then, we will consider two possible cases. $s_2 < s_1$ or $s_1 < s_2$.

Case 1: We will show inequality (B.1) does not hold when $s_2 < s_1$. Note that $s_2 = G(F^{-1}(s_1)) < 1$ in this case because $s_2 < s_1 < 1$.

For a sufficiently small $\epsilon > 0$, we have

$$s_2 + \epsilon > G(F^{-1}(s_1)) \Leftrightarrow G^{-1}(s_2 + \epsilon) > F^{-1}(s_1)$$

by Lemma B.3.2 (2). Then, Lemma B.3.2 (1) implies

$$G^{-1}(s_2 + \epsilon) > F^{-1}(s_1) \Rightarrow F(G^{-1}(s_2 + \epsilon)) \geq s_1.$$

Therefore, we have

$$F(G^{-1}(s_2 + \epsilon)) - (s_2 + \epsilon) \geq s_1 - s_2 - \epsilon.$$

When we pick sufficiently small $\epsilon > 0$, we have

$$F(G^{-1}(s_2 + \epsilon)) - (s_2 + \epsilon) > s_1 - s_2 - \epsilon > \sup_s |F(G^{-1}(s)) - (s)|.$$

Therefore, we have s such that

$$\sup_s |F(G^{-1}(s)) - (s)| < |F(G^{-1}(s)) - (s)|.$$

This is a contradiction.

Case 2: We will show inequality (B.1) does not hold when $s_2 > s_1$. Note that $s_2 > 0$ and $F^{-1}(s_1) > G^{-1}(0)$ because $s_2 > s_1 > 0$. The proof depends on whether there exists y such that $G(y) = s_2$ for $y < F^{-1}(s_1)$ or not.

Case 2-1: Suppose there exists y such that $G(y) = s_2$ for $y < F^{-1}(s_1)$.

Then, by Lemma B.3.2 (7), we have

$$G^{-1}(s_2) = G^{-1}(G(F^{-1}(s_1))) < F^{-1}(s_1).$$

Using Lemma B.3.2 (2), we have

$$G^{-1}(s_2) < F^{-1}(s_1) \Rightarrow F(G^{-1}(s_2)) < s_1.$$

Therefore,

$$\sup_s |F(G^{-1}(s_2)) - s_2| < s_2 - s_1 < |F(G^{-1}(s_2)) - s_2|.$$

But, this is a contradiction.

Case 2-2: Suppose there is no y such that $G(y) = s_2$ for $y < F^{-1}(s_1)$.

Then, by Lemma B.3.2 (6) and (7), we have

$$G^{-1}(s_2 - \epsilon) \leq G^{-1}(G(F^{-1}(s_1))) = F^{-1}(s_1).$$

for sufficiently small $\epsilon > 0$.

Case 2-2-1: Suppose there exists y such that $s_1 = F(y)$. When F is continuous, $F^{-1}(s)$ is strictly increasing, and we can find y such that $F(y) = s_1$.

Then, we have for sufficiently small $\epsilon > 0$,

$$F(G^{-1}(s_2 - \epsilon)) \leq F(F^{-1}(s_1)) = s_1$$

by Lemma B.3.2 (4).

From this, for sufficiently small $\epsilon > 0$, we have

$$F(G^{-1}(s_2 - \epsilon)) - (s_2 - \epsilon) \leq s_1 - (s_2 - \epsilon) < 0,$$

and

$$|F(G^{-1}(s_2 - \epsilon) - (s_2 - \epsilon)| \leq (s_2 - \epsilon) - s_1.$$

Therefore, for a sufficiently small $\epsilon > 0$, we have

$$\sup_s |F(G^{-1}(s)) - s| < |F(G^{-1}(s_2 - \epsilon) - (s_2 - \epsilon)|.$$

But, this is a contradiction.

Case 2-2-2: Suppose there is no y such that $s_1 = F(y)$.

In this case, $G(\cdot)$ is continuous because we assume that at least one of $F(\cdot)$ and $G(\cdot)$ is continuous. When $G(\cdot)$ is continuous, then $G^{-1}(s)$ is strictly increasing function by B.3.2 (8). Then, we have

$$G^{-1}(s_2 - \epsilon) < G^{-1}(s_2) = F^{-1}(s_1).$$

for sufficiently small $\epsilon > 0$. Then, by Lemma B.3.2 (2), we have

$$G^{-1}(s_2 - \epsilon) < F^{-1}(s_1) \Leftrightarrow F(G^{-1}(s_2 - \epsilon)) < s_1.$$

From this, for sufficiently small $\epsilon > 0$, we have

$$F(G^{-1}(s_2 - \epsilon)) - (s_2 - \epsilon) < s_1 - (s_2 - \epsilon) < 0,$$

and

$$|F(G^{-1}(s_2 - \epsilon)) - (s_2 - \epsilon)| > (s_2 - \epsilon) - s_1.$$

Therefore, for a sufficiently small $\epsilon > 0$, we have

$$\sup_s |F(G^{-1}(s)) - s| < |F(G^{-1}(s_2 - \epsilon) - (s_2 - \epsilon)|.$$

However, this is a contradiction. □

B.3.2 Proof of Lemma 2.6.1

Example 7.21 of Sen (2021) implies that

$$A = \{\{Z_t : u^\top Z_t \leq s\}, u \in \mathbb{S}^{d-1}, s \in \mathbb{R}\}$$

is a VC class with VC dimension is equal to $d + 1$.

Then, we can show weak convergence of empirical process using Lemma B.3.1 because the indicator function is uniformly bounded.

B.3.3 *Proof of Lemma 2.6.2*

For $u \in \mathcal{N}_0 := \{u \in \mathbb{S}^{d-1}; G(s;u,\psi_0) \text{ is degenerate}\}$, $G_T(s;u,\psi_0) = G(s;u,\psi_0)$. Then, we have

$$\begin{aligned} &T\int_{s\in\mathbb{S}^{d-1}}\int_0^1 |G_T^{-1}(s;u,\psi_0) - G^{-1}(s;u,\psi_0)|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \\ &\qquad = T\int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 |G_T^{-1}(s;u,\psi_0) - G^{-1}(s;u,\psi_0)|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u). \end{aligned}$$

Following Proposition A.18 of Bobkov and Ledoux (2019),

$$\begin{aligned} G_T^{-1}(s;u,\psi_0) - G^{-1}(s;u,\psi_0) &= G^{-1}(G(G_T^{-1}(s;u,\psi_0);u,\psi_0);u,\psi_0) - G(s;u,\psi_0) \\ &= \int_s^{G(G_T^{-1}(s;u,\psi_0);u,\psi_0)} \frac{1}{g(G^{-1}(z;u,\psi_0);u,\psi_0)}\mathrm{d}z \\ &\leq \frac{1}{g(G^{-1}(s(s;u,\psi_0);u,\psi_0);u,\psi_0)}\Big(G(G_T^{-1}(s;u,\psi_0),u,\psi_0) - s\Big), \end{aligned}$$

where $s(s;u,\psi_0) \in \Big(G(G_T^{-1}(s;u,\psi_0),u,\psi_0) \wedge s, G(G_T^{-1}(s;u,\psi_0),u,\psi_0) \vee s\Big)$.

This implies

$$\begin{aligned} &T\int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 |G_T^{-1}(s;u,\psi_0) - G^{-1}(s;u,\psi_0)|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \\ &\leq T\int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 \frac{w(s)}{g^2(G^{-1}(s;u,\psi);u,\psi_0)}|G(G_T^{-1}(s;u,\psi_0);u,\psi_0) - s|^2\mathrm{d}s\mathrm{d}\varsigma(u) \\ &\leq \left(\int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 \frac{w(s)}{g^2(G^{-1}(s;u,\psi_0);u,\psi_0)}\mathrm{d}s\mathrm{d}\varsigma(u)\right)\{\sqrt{T}\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in(0,1)}|G(G_T^{-1}(s;u,\psi_0);u,\psi_0) - s|\}^2 \end{aligned}$$

where $s(s;u,\psi_0) \in \Big(G(G_T^{-1}(s;u,\psi_0),u,\psi_0) \wedge s, G(G_T^{-1}(s;u,\psi_0),u,\psi_0) \vee s\Big)$.

Then, it is enough to show $\sqrt{T}\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in(0,1)}|G(G_T^{-1}(s;u,\psi_0);u,\psi_0) - s| = O_p(1)$. This is because it implies $\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in(0,1)}|s(s;u,\psi_0) - s| = o_p(1)$, and we have

$$\begin{aligned} &\left(\int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 \frac{w(s)}{g^2(G^{-1}(s;u,\psi_0);u,\psi_0)}\mathrm{d}s\mathrm{d}\varsigma(u)\right) \\ &\quad\leq \left(\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in supp\{w\}}\frac{g^2(G^{-1}(s;u,\psi_0);u,\psi_0)}{g^2(G^{-1}(s;u,\psi_0);u,\psi_0)}\right)\int_{s\in\mathbb{S}^{d-1}}\int_0^1 \frac{w(s)}{g^2(G^{-1}(s;u,\psi_0);u,\psi_0)}\mathrm{d}s\mathrm{d}\varsigma(u) \\ &\quad= O_p(1) \end{aligned}$$

under Conditions (2.16) and (2.17).

For each $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$, $G(G_T^{-1}(s;u,\psi_0),u,\psi_0)$ is the quantile function of $G(u^\top Z_t,u,\psi_0)$. It implies

$$\sup_s |G(G_T^{-1}(s;u,\psi_0),u,\psi_0)-s| = \sup_s \left|\frac{1}{T}\sum_{t=1}^{T} I(G(u^\top Z_t;u,\psi_0)\le s)-s\right|.$$

(c.f Equation (1.4.5) in Csörgő (1983)). Note that for each $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$, we have

$$\begin{aligned}\sup_{s\in(0,1)}\left|\frac{1}{T}\sum_{t=1}^{T} I(G(u^\top Z_t;u,\psi_0)\le s)-s\right| &= \sup_{s\in(0,1)}\left|\frac{1}{T}\sum_{t=1}^{T} I(u^\top Z_t\le G^{-1}(s;u,\psi_0))-s\right| \\ &= \sup_{s\in\mathbb{R}} |G_T(s;u,\psi_0)-G(s;u,\psi_0)|.\end{aligned}$$

Therefore,

$$\sqrt{T}\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in(0,1)}|G(G_T^{-1}(s;u,\psi_0),u,\psi_0)-s| = \sqrt{T}\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in\mathbb{R}}|G_T(s;u,\psi_0)-G(s;u,psi_0)|.$$

When all assumptions in Lemma 2.6.1 are satisfied,

$$\sqrt{T}\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in\mathbb{R}}|G_T(s;u,\psi_0)-G(s;u,psi_0)| = O_p(1).$$

This concludes the proof.

B.3.4 Proof of Lemma 2.6.3

Let $\mathbb{D}$ be the space of univariate distribution functions, and $\mathbb{D}_1$ be the restriction on $\mathbb{D}$ such that the domain of the cdf functions in $\mathbb{D}_1$ is the same as the support of $G(s;u)$ for each u.

Let $G \in \mathbb{D}_1$ and $G_n = G(t;u)+t_n h_n(t;u) \in \mathbb{D}_1$ such that $t_n \downarrow 0$, and $\sup_{t;u}|h_n - h| = o(1)$, $\sup_{t,u} h(t,u) < C$. Let's denote $\mathcal{N}_0 = \{u; u^\top Z \text{ is degenerate}\}$.

Because we are interested in L2 convergence, and $\mathcal{N}_0$ is degenerate, it is enough to think

$$\int_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1\left[\frac{G_n^{-1}(s;u)-G^{-1}(s;u)}{t_n}-\frac{1}{g(G^{-1}(s;u);u)}h(G^{-1}(s;u);u)\right]^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) = o(1)$$

We will show this using the dominated convergence theorem. In Step 1, We will show that there exists N and C such that

$$\sup_u \int \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)\mathrm{d}s < C$$

for all $n > N$.

In step 2, we will show

$$\int \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)\mathrm{d}s = o(1)$$

for each $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$

Step 1: We will show that there exists N and C such that

$$\sup_{u \in \mathbb{S}^{d=1}} \int \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)ds < C$$

for all $n > N$.

First, from the model assumption, we have

$$\int \left[\frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)ds = \int \frac{w(s)}{g^2(G^{-1}(s;u);u)} \mathrm{d}s \sup_{s,u} |h(G^{-1}(s,u),u)| < C$$

Second, for almost all u, we have

$$\sqrt{w(s)}\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} = \frac{1}{t_n}\frac{w(s)}{g(G^{-1}(s(s;u);u);u)}(G(G_n^{-1}(s;u)) - s)$$

Then,

$$\sup_s \left|\sqrt{w(s)}\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n}\right| \leq \left(\sup_s \frac{w(s)}{g(G^{-1}(s(s;u);u);u)}\right) \frac{1}{t_n} \sup_s \left|(G(G_n^{-1}(s;u);u) - s)\right|$$

where s is between $G(G_n^{-1}(s;u);u)$ and s.

From Lemma B.3.3, we have

$$\begin{aligned}
\frac{1}{t_n} \sup_s \left|(G(G_n^{-1}(s;u)) - s)\right| &= \frac{1}{t_n} \sup_s \left|(G_n(G^{-1}(s;u)) - s)\right| \\
&= \frac{1}{t_n} \sup_s \left|(G(G^{-1}(s;u);u) + t_n h_n(G^{-1}(s;u);u) - s)\right| \\
&= \sup_s \left|h_n(G^{-1}(s;u);u)\right|
\end{aligned}$$

Therefore, we have

$$\sup_s \left| \sqrt{w(s)} \frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} \right| \leq \left(\sup_s \frac{\sqrt{w(s)}}{g(G^{-1}(s(s;u);u);u)} \right) \sup_s \left| h_n(G^{-1}(s;u);u) \right|$$

It implies

$$\begin{aligned} &\sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_s \left| \sqrt{w(s)} \frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} \right| \\ &\quad \leq \left(\sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_s \frac{\sqrt{w(s)}}{g(G^{-1}(s(s;u);u);u)} \right) \sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_s \left| h_n(G^{-1}(s;u);u) \right|. \end{aligned}$$

Note that we have the above inequality for all u because we have $h_n(s;u) = 0$ for $u \in \mathcal{N}_0$ when the domain of $h_n(t;u)$ is the same as the support of $G(t;u)$ for each u, and we can set $g(\cdot, u) = \infty$ when $u \in \mathcal{N}_0$.

Because

$$\sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s \in supp\{w\}} \left| \frac{g(G^{-1}(s;u,\psi_0);u,\psi_0)}{g(G^{-1}(s(s;u);u,\psi_0);u,\psi_0)} \right| = O(1) \tag{B.3}$$

for $s(s,u)$ such that $\sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s \in supp\{w\}} |s(s;u) - s| = o(1)$, we can find N and C such that

$$\sup_{s,u} \left| \sqrt{w(s)} \frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} \right| \leq \left(\sup_{s,u} \frac{\sqrt{w(s)}}{g(G^{-1}(s(s;u);u);u)} \right) \sup_{s,u} \left| h_n(G^{-1}(s;u);u) \right| < C$$

for all $n > N$.

Therefore, we have N and C such that

$$\sup_{s,u} \int \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u) \right]^2 w(s) ds < C$$

for all $n > N$.

Step 2: We will show that for each $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$,

$$\int \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u) \right]^2 w(s) ds = o(1)$$

with dominated convergence theorem.

Then, similar to the logic in Kaji (2019), we can show this as follows.

$$\begin{aligned}
&\int \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)ds \\
&\quad = \int_0^\delta \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)ds \\
&\qquad + \int_\delta^{1-\delta} \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)ds \\
&\qquad + \int_{1-\delta}^1 \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)ds \\
&\quad \leq 2\delta \int \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)ds \\
&\qquad + \sup_{s\in[\delta,1-\delta]} \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 .
\end{aligned}$$

First,

$$2\delta \int \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)ds \leq 2\delta C$$

for $n > N$. So, we can choose sufficiently small δ in order to make this small.

Second,

$$\sup_{s\in[\delta,1-\delta]} \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 = o(1)$$

from van de Vaart and Wellner (1996, Theorem 3.9.23).

Therefore, for each $\epsilon > 0$, there exists a positive constant N such that

$$\int \left[\frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} - \frac{1}{g(G^{-1}(s;u);u)} h(G^{-1}(s;u);u)\right]^2 w(s)ds < \epsilon$$

for $n \geq N$.

Therefore, using Steps 1 and 2 and the dominated convergence theorem, we will show the desired result.

B.4 Proofs in Section 2.7

B.4.1 Preliminary Lemma

Lemma B.4.1. *Suppose $F(y|x,\psi)$ is Lipschitz with respect to ψ in the sense that there exists a function $M(y,x)$ for any $\psi,\psi' \in \Psi$,*

$$|F(y|x,\psi) - F(y|x,\psi')| \leq M(y,x)\|\psi - \psi'\|,$$

and $\int_{u\in\mathbb{S}^{d-1},u_1\neq 0}\int_{-\infty}^{\infty}\int M^2(u_1^{-1}(s-u_2^\top x);x)\mathrm{d}F_X(x)w(s)\mathrm{d}s\mathrm{d}\varsigma(u) < \infty$, where $F_X(\cdot)$ is the CDF of X_t. Then, $k(x,x',\psi)$ and $k_2(x,x',\psi,\psi_0)$ are Lipschitz in ψ.

Proof. Note that

$$\mathbb{E}[I(u^\top Z_t \leq s)|X_t,\psi] = \begin{cases} F(u_1^{-1}(s-u_2^\top X_t)|X_t,\psi) & \text{if } u_1 > 0 \\ I(u_2^\top X_t \leq s) & \text{if } u_2 < 0 \\ 1 - F(u_1^{-1}(s-u_2^\top X_t)|X_t,\psi) & \text{if } u_1 > 0 \end{cases}$$

Then,

$$\begin{aligned}
&|k(X_t,X_j,\psi) - k(X_t,X_j,\psi')| \\
&\quad\leq 2\int_{u\in\mathbb{S}^{d-1},u_1\neq 0}\int_{-\infty}^{\infty}\Big|\big(F(u_1^{-1}(s-u_2^\top X_t)|X_t,\psi) - \mathbb{E}[F(u_1^{-1}(s-u_2^\top X_t)|X_t,\psi)]\big) \\
&\qquad\qquad - \big(F(u_1^{-1}(s-u_2^\top X_t)|X_t,\psi) - \mathbb{E}[F(u_1^{-1}(s-u_2^\top X_t)|X_t,\psi)]\big)\Big| w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \\
&\qquad + 2\int_{u\in\mathbb{S}^{d-1},u_1\neq 0}\int_{-\infty}^{\infty}\Big|\big(F(u_1^{-1}(s-u_2^\top X_j)|X_j,\psi) - \mathbb{E}[F(u_1^{-1}(s-u_2^\top X_j)|X_j,\psi)]\big) \\
&\qquad\qquad - \big(F(u_1^{-1}(s-u_2^\top X_j)|X_j,\psi) - \mathbb{E}[F(u_1^{-1}(s-u_2^\top X_j)|X_j,\psi)]\big)\Big| w(s)\mathrm{d}s\mathrm{d}\varsigma(u).
\end{aligned}$$

Since we have

$$|F(u_1^{-1}(s-u_2^\top x)|x,\psi) - F(u_1^{-1}(s-u_2^\top x)|x,\psi')| \leq M(u_1^{-1}(s-u_2^\top x);x)\|\psi - \psi'\|,$$

where $\int_{u\in\mathbb{S}^{d-1},u_1\neq 0}\int_{-\infty}^{\infty}\int M^2(u_1^{-1}(s-u_2^\top x);x)\mathrm{d}F_X(x)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)<\infty$, $k(x,x',\psi)$ is Lipchitz continuous with respect to ψ. With a similar calculation, we can show that $k_2(x,x',\psi,\psi_0)$ is Lipchitz continuous with respect to ψ when $F(y|x,\psi)$ is Lipchitz continuous as well. $\square$

B.4.2 Proof of Lemma 2.7.1

We will investigate the properties of $\left\{\sup_\psi \frac{1}{T^2}\sum_{t=1}^{n}\sum_{j=1}^{T}k(X_t,X_j;\psi)\right\}$ to verify Assumption 2.5.1 (ii). Since $k(X_t,X_j;\psi)$ is symmetric with respect to X_t and X_j, we have

$$\frac{1}{T^2}\sum_{t=1}^{T}\sum_{j=1}^{T}k(X_t,X_j;\psi)=\frac{1}{T^2}\sum_{t=1}^{T}k(X_t,X_t;\psi)+\frac{2}{T^2}\sum_{1\le t<j\le T}k(X_t,X_j,\psi)$$

When $w(s)$ is integrable, $k(X_t,X_t,\psi)$ is uniformly bounded by an absolute positive constant, which implies $\frac{1}{T^2}\sum_{t=1}^{T}k(X_t,X_t;\psi)=o(1)$. Therefore, it is enough to show

$$\sup_{\psi\in\Psi}\left|\frac{2}{T^2}\sum_{1\le t<j\le T}k(X_t,X_j,\psi)\right|=o_p(1). \tag{B.4}$$

Since $w(s)$ is integrable, we can interchange the expectation and integral when we evaluate $\mathbb{E}[k(X_i,X_j;\psi)]$, and

$$\mathbb{E}[k(X_t,X_j;\psi)]=\begin{cases}\mathbb{E}[k(X_t,X_t;\psi)] & \text{if } t=j,\\ 0 & \text{if } t\neq j.\end{cases}$$

Then, we can verify Equation (B.4) under the Lipschitz continuity of $k(X_t,X_j,\psi)$ with respect to ψ.

Lemma B.4.2. *Suppose Ψ is compact and $k(X_t,X_j,\psi)$ is Lipshitz in ψ in the sense that for any $\psi,\psi\in\Psi$, we have*

$$|k(X_t,X_j,\psi)-k(X_t,X_j,\psi_0)|\le M(X_t,X_j)\|\psi-\psi\|$$

and $\int\int M(x,x)dP_x dP_x<\infty$ where $P_x=P_x$. Then,

$$\sup_{\psi}\left\|\frac{2}{T^2}\sum_{1\le t<j\le n}k(X_t,X_j,\psi)\right\|=\sup_{\psi}\left\|\frac{2}{T^2}\sum_{1\le t<j\le T}k(X_t,X_j,\psi)-\frac{2}{T^2}\sum_{1\le t<j\le n}\mathbb{E}[k(X_t,X_j,\psi)]\right\|\xrightarrow{p}0.$$

Proof. See Corollary 4.1 of Newey (1991) or Lemma 4 in Appendix of Briol et al. (2019). □

Under the Lipschitz continuity of $F(\cdot|x,\theta)$ in ψ, $k(\cdot,\cdot;\psi)$ is also Lipchitz continuous in ψ by Lemma B.4.1. This concludes the proof.

B.4.3 Proof of Lemma 2.7.3

Let us remind that

$$\widehat{R}_T(s;u,\psi,\psi_0) := \widehat{Q}_T(s;u,\psi) - \widehat{Q}_T(s;u,\psi_0) - (\psi-\psi_0)^\top \widehat{D}_T(s;u,\psi_0).$$

We would like to show

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} \left| \frac{T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\right| = o_p(1)$$

for any $\tau_T \to 0$.

Here, we will show it with $\widehat{D}_T(s;u,\psi_0) = D(s;u,\psi_0)$ which are defined in Condition 2.7.1.

Note that

$$\begin{aligned}
&\widehat{Q}_T(s;u,\psi) - \widehat{Q}_T(s;u,\psi_0) - (\psi-\psi_0)^\top \widehat{D}_T(s;u,\psi_0)\\
&= \left[\widehat{Q}_T(s;u,\psi) - Q(s;u,\psi)\right] - \left[\widehat{Q}_T(s;u,\psi_0) - Q(s;u,\psi_0)\right]\\
&\quad + \left[Q(s;u,\psi) - Q(s;u,\psi_0) - (\psi-\psi_0)^\top \widehat{D}_T(s;u,\psi_0)\right].
\end{aligned}$$

Under Condition 2.7.1, it is enough to show

$$\sup_{|\psi-\psi_0|\le\tau_T} \frac{T\left(\int_{u\in\mathbb{S}^{d-1}}\int\left(\left[\widehat{Q}_T(s;u,\psi)-Q(s;u,\psi)\right]-\left[\widehat{Q}_T(s;u,\psi_0)-Q(s;u,\psi_0)\right]\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} = o_p(1),$$

where

$$\widehat{Q}_T(s;u,\psi) - Q(s;u,\psi) = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}[I(u^\top Z_t\le s)|X_t,\psi] - G(s;u,\psi)\right).$$

Because $\frac{1}{1+\|\sqrt{T}(\psi-\psi_0)\|^2} \le 1$, it is enough to show

$$\sup_{|\psi-\psi_0|\le\tau_T} T\left(\int_{u\in\mathbb{S}^{d-1}}\int\left((\widehat{Q}_T(s;u,\psi)-Q(s;u,\psi)) - (\widehat{Q}_T(s;u,\psi_0)-Q(s;u,\psi_0))\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right) = o_p(1).$$

Let us denote

$$V_{2,T}(\psi) = \left(\int_{u \in \mathbb{S}^{d-1}} \int \Big((\widehat{Q}_T(s;u,\psi) - Q(s;u,\psi)) - (\widehat{Q}_T(s;u,\psi_0) - Q(s;u,\psi_0)) \Big)^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u) \right).$$

$V_{2,T}(\psi)$ is a V-statistic. That is,

$$V_{2,T}(\psi) = \frac{1}{T^2} \sum_{t=1}^{T} \sum_{j=1}^{T} k_2(X_t, X_j, \psi, \psi_0),$$

where

$$\begin{aligned}
&k_2(X_t, X_j, \psi, \psi_0) \\
&= \int_{u \in \mathbb{S}^{d-1}} \int \Bigg\{ \Bigg[\Big(\mathbb{E}[I(u^\top Z_t \le s) | X_t, \psi] - G(s;u,\psi) \Big) - \Big(\mathbb{E}[I(u^\top Z_t \le s) | X_t, \psi_0] - G(s;u,\psi_0) \Big) \Bigg] \\
&\quad \times \Bigg[\Big(\mathbb{E}[I(u^\top Z_j \le s) | X_j, \psi] - G(s;u,\psi) \Big) - \Big(\mathbb{E}[I(u^\top Z_j \le s) | X_j, \psi_0] - G(s;u,\psi_0) \Big) \Bigg] \Bigg\} w(s) \mathrm{d}s \mathrm{d}\varsigma(u)
\end{aligned}$$

Note that $k_2(X_t, X_j, \psi, \psi_0)$ is symmetric kernel, and $V_{2,T}(\psi_0) = 0$ since $k_2(X_t, X_j, \psi_0, \psi_0) = 0$.

Then, we need to handle $nV_{2,T}(\psi)$. It can be decomposed as follows.

$$\begin{aligned}
TV_{2,T}(\psi) &= \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{T} k_2(X_t, X_j; \psi, \psi_0) \\
&= \frac{1}{T} \sum_{t=1}^{T} k_2(X_t, X_t; \psi, \psi_0) + \frac{2}{T} \sum_{1 \le t < j \le T} k_2(X_t, X_j, \psi, \psi_0) \\
&= \frac{1}{T} \sum_{t=1}^{T} \Big(k_2(X_t, X_t; \psi, \psi_0) - \mathbb{E}[k_2(X_t, X_t; \psi, \psi_0)] \Big) + \mathbb{E}[k_2(X_t, X_t; \psi, \psi_0)] \\
&\quad + \frac{2}{T} \sum_{1 \le t < j \le T} k_2(X_t, X_j, \psi, \psi_0).
\end{aligned}$$

We will show that each term in the above expressions is $o_p(1)$ as $\tau_T \to 0$. Note that $k_2(\cdot, \cdot, \psi, \psi_0)$ is Lipchitz continuous in ψ by Lemma B.4.1.

Step 1) Because Ψ is compact and $k_2(X_t, X_t, \psi, \psi_0)$ is Lipschitz continuous with respect to ψ for each X_t, by ULLN, we have

$$\sup_{\psi \in \Psi} \left| \frac{1}{T} \sum_{i=1}^{T} \Big(k_2(X_t, X_t; \psi) - \mathbb{E}[k_2(X_t, X_t; \psi, \psi_0)] \Big) \right| \xrightarrow{p} 0.$$

Step 2) Because $k_2(X_i, X_i, \psi_0, \psi_0) = 0$, and $k_2(X_i, X_i, \psi, \psi_0)$ is Lipschitz continuous for each X_i, we have

$$\mathbb{E}[k_2(X_i, X_i, \psi, \psi_0)] \leq \mathbb{E}[M_2(X_i, X_i)]\|\psi - \psi_0\|$$

and it implies that $\sup_{\psi \in \Psi;\, \|\psi - \psi_0\| \leq \tau_n} \mathbb{E}[k_2(X_i, X_i, \psi, \psi_0)] = o_p(1)$.

Step 3) For $t \neq j$, $\mathbb{E}[k_2(X_t, X_j, \psi, \psi_0)] = 0$ when $w(s)$ is integrable. Then,

$$U_{2,T} := \frac{2}{T^2} \sum_{1 \leq t < j \leq T} k_2(X_t, X_j, \psi, \psi_0)$$

is a degenerate U-statistic.

Suppose Ψ is compact, and for any $\psi, \psi \in \Psi$, we have

$$|k_2(X_t, X_j, \psi, \psi_0) - k_2(X_i, X_j, \psi, \psi_0)| \leq M_2(X_i, X_j)\|\psi - \psi\|$$

with $\iint M_2^2(x, y) dF_{X_t}(x) dF_{X_j}(y) < \infty$ and $\int M_2^2(x, x) dF_{X_t}(x) < \infty$. Then, by the proof of Lemma 4 in the appendix of Briol et al. (2019), we have $k_2(X_i, X_i, \psi, \psi_0)$ is Euclidean with envelope $F \leq M(X_i, X_j) diam(\Psi)$ where $diam(\Psi) = \sup_{\theta, \theta \in \Psi} \|\theta - \theta\|$ is the diameter of Ψ Then,

$$\int\int |k_2^2(X_i, X_j)| dF_{X_i} dF_{X_j} \leq \left(\int\int M_2^2(X_i, X_j) dF_{X_i} dF_{X_j} \right) \|\psi - \psi_0\|$$

So, $\int\int |k_2^2(X_i, X_j)| dF_{X_i} dF_{X_j} \to 0$ as $|\psi - \psi_0| \to 0$.

By Corollary 8 in Sherman (1994), we have $\sup_{\psi \in \Psi;\, \|\psi - \psi_0\| \leq \tau_T} |TU_{2,T}| = o_p(1)$.

Therefore,

$$\sup_{\psi \in \|\psi - \psi_0\|_2 \leq \tau_T} |TV_{2,T}(\psi)| = \sup_{\psi \in \|\psi - \psi_0\|_2 \leq \tau_T} |TV_{2,T}(\psi) - TV_{2,T}(\psi_0)| = o_p(1).$$

B.5 Verification of Assumptions in Section 2.5 for Singular, One-sided, and Two-sided Uniform Models

In this subsection, we will verify that the assumptions for consistency and asymptotic normality of MSWD and MSCD estimators in Section 2.5 are satisfied in one-sided and two-sided models. For singular model in Example 1 of Arjovsky et al. (2017), we will deal with the MSWD estimator only.

We will verify Assumptions 2.5.3, 2.5.2, 2.5.4 (i), and 2.5.5, and 2.5.6 because all three examples are conditional models, and Assumption 2.5.1 (i) is implied by Assumption 2.5.4(i).

B.5.1 Singular Model in Example of Arjovsky et al. (2017)

In Example 1 in Arjovsky et al. (2017), $Y = (\theta, Z)$, where θ is a deterministic constant and $Z \sim U[0,1]$. Consider the MSWD estimator of θ. In this case, $Q(\cdot; u, \theta)$ is quantile function of the projected variable $u^\top Z$, which is

$$Q(s; u, \theta) = \begin{cases} u_1\theta + u_2 s & \text{if } u_2 \geq 0, \\ u_1\theta + u_2(1-s) & \text{if } u_2 < 0. \end{cases}$$

Verification of Assumption 2.5.2

Assumption 2.5.2 holds because

$$\int_{u\in\mathbb{S}^{d-1}} \int_0^1 (Q(s; u, \theta_0) - Q(s; u, \theta))^2 \mathrm{d}s\mathrm{d}\varsigma(u) = \frac{1}{2}(\theta - \theta_0)^2.$$

Verification of Assumption 2.5.3 and Assumption 2.5.6

$Q(s; u, \theta)$ is norm-differentiable at $\theta = \theta_0$ with $D(s; u, \theta) = u_1$ because

$$Q(s; u, \theta) - Q(s; u, \theta_0) = u_1(\theta - \theta_0), \text{ and } \int_{-\infty}^{\infty} \int_0^1 D(s; u, \theta)^2 \mathrm{d}s\mathrm{d}\varsigma(u) = \frac{1}{2}.$$

It implies that Assumption 2.5.6 holds as well.

Verification of Assumption 2.5.4 (i) and 2.5.5

Let $G(s; u, \psi_0)$ be the distribution function of the projected variable $u^\top Z_t$, and

$$G_T(s; u, \psi_0) = \frac{1}{T} \sum_{t=1}^{T} I(u^\top Z_t \leq s).$$

Since $\{Z_t\}_{t=1}^n$ i.i.d, and $\{I(u^\top Z_t \le s) : u \in \mathbb{S}^{d-1}, s \in \mathbb{R}\}$ is VC subgraph class,

$$\{\sqrt{T}(G_T(s; u, \psi_0) - G(s; u, \psi_0)) : s \in \mathbb{R}, u \in \mathbb{S}^{d-1}\}$$

are P-Donsker. Also, when $u_2 \neq 0$,

$$\frac{1}{g(Q^{-1}(s; u, \psi_0); u, \psi_0)} = \frac{\partial}{\partial s} Q(s; u, \theta) = |u_2| < \infty.$$

Therefore, Assumptions 2.5.4 (i) and and 2.5.5 hold by Lemmas 2.6.2 and 2.6.5.

B.5.2 One-sided and Two-sided Uniform Models

Since we deal with unconditional models for univariate variable Y_t, we can take $\widehat{Q}(s; u, \theta) := Q(s; \theta)$, and $\widehat{D}(\cdot, u, \theta_0) = D(\cdot, \theta_0)$ where $D(\cdot, \theta_0)$ is a deterministic $L_2(\mathcal{S}, w(s)\mathrm{d}s)$-measurable function. $Q(s; \theta)$ will be the parametric distribution function of Y_t for MSCD estimators, and the parametric quantile function for MSWD estimators. In one-sided and two-sided models, we will consider the case where $w(s) = 1$.

One-sided Uniform Model

In one-sided uniform model, The CDF of Y_t is

$$F(s, \theta_0) = \begin{cases} 0 & \text{if } s < 0, \\ s/\theta_0 & \text{if } 0 \le s \le \theta_0, \\ 1 & \text{if} s > \theta_0, \end{cases}$$

and

$$F^{-1}(s, \theta_0) = \theta_0 s \text{ for } 0 < s < 1.$$

We deal with real-valued random variable, and do not require projection. We can take $\widehat{Q}_T(\cdot, u, \psi) = Q(\cdot, \psi)$.

The MSCD Estimator In this case, $Q(s,\theta_0)=F(s,\theta_0)$. Let $F_T(s)=I(Z_t\le s)$. Section 2.4.2 shows

$$\mathcal{C}_2^2(\mu_\theta,\mu_0)=\int_{-\infty}^{\infty}(F(s;\theta)-F(s;\theta_0))^2\mathrm{d}s=\begin{cases}\frac{(\theta-\theta_0)^2}{3\theta_0} & \text{if } \theta\le\theta_0,\\ \frac{(\theta-\theta_0)^2}{3\theta} & \text{if } \theta>\theta_0.\end{cases}$$

Since

$$\frac{\partial\mathcal{C}_2^2(\mu_\theta,\mu_0)}{\partial\theta}=\begin{cases}\frac{2(\theta-\theta_0)}{3\theta_0}<0 & \text{if } \theta<\theta_0,\\ 0 & \text{if } \theta=\theta_0,\\ \frac{\theta^2-\theta_0^2}{3\theta^2}>0 & \text{if } \theta>\theta_0.\end{cases}$$

Assumption 2.5.2 holds when $\theta_0>0$.

Because $\{Y_t\}_{t=1}^T$ is a random sample, $\{\sqrt{T}(F_T(s)-F(s,\theta_0)):s\in\mathbb{R}\}$ is P-Donsker, and

$$\begin{aligned}T\int_{-\infty}^{\infty}((F_T(s)-F(s,\theta_0)))^2\mathrm{d}s&=T\int_0^{\theta_0}((F_T(s)-F(s,\theta_0)))^2\mathrm{d}s\\&\le\theta_0\left[\sup_{s\in\mathbb{R}}(\sqrt{T}(F_T(s)-F(s,\theta_0)))\right]^2=O_p(1).\end{aligned}$$

Therefore, Assumption 2.5.4 (i) holds.

Example 2.5.1 shows that $F(\cdot,\theta)$ is norm-differentiable at $\psi=\psi_0$, and $D(s;\psi_0)$ is $L_2(\mathbb{R},\mathrm{d}s)$-measurable, and

$$\int_{-\infty}^{\infty}D^2(s;\psi_0)\mathrm{d}s=\int_0^{\theta_0}\frac{s^2}{\theta_0^4}\mathrm{d}s=\frac{1}{3\theta_0}<\infty.$$

Therefore, Assumptions 2.5.3 and 2.5.6 hold.

Since $\{\sqrt{T}(F_T(s)-F(s,\theta_0)):s\in\mathbb{R}\}$ is P-Donsker and $\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}D^2(s;u,\psi_0)\mathrm{d}s\mathrm{d}\varsigma(u)<\infty$, Assumption 2.5.5 is satisfied by Lemma 2.6.7.

The MSWD Estimator In this case, $Q(s,\theta_0)=F^{-1}(s,\theta_0)$, and $Q_T(s)=F_T^{-1}(s)$, where $F_T(s)=\frac{1}{T}\sum_{t=1}^T I(Z_t\le s)$.

Since

$$\int_0^1((F^{-1}(s,\theta_0)-F^{-1}(s,\theta)))^2\mathrm{d}s=(\theta-\theta_0)^2,$$

Assumption 2.5.2 holds.

Assumptions 2.5.3 and 2.5.6 hold with $D(s, u, \psi_0) = s$ because $F^{-1}(s, \theta) - F^{-1}(s, \theta_0) = s(\theta - \theta_0)$.

Since $\{\sqrt{T}(F_T(s) - F(s, \theta_0)) : s \in \mathbb{R}\}$ is P-Donsker, and $\frac{\partial F^{-1}(s, \psi_0)}{\partial s} = \theta_0 < \infty$, Assumptions 2.5.4 (i) and 2.5.5 hold by Lemmas 2.6.2 and 2.6.5.

Two-sided Uniform Model

In two-sided models, the CDF of Y is

$$F(y; \theta_0) = \begin{cases} \frac{y}{4\theta_0} & \text{if } 0 \leq y < \theta_0 \\ \frac{1}{4} + \frac{3(y-\theta_0)}{4(1-\theta_0)} = 1 - \frac{3(1-y)}{4(1-\theta_0)} & \text{if } \theta \leq y \leq 1. \end{cases}$$

and the quantile function is

$$F^{-1}(s; \theta_0) = \begin{cases} 4\theta_0 s & \text{if } 0 \leq s \leq 1/4, \\ 1 - \frac{4}{3}(1-\theta_0)(1-s) & \text{if } 1/4 \leq s \leq 1. \end{cases}$$

The MSCD Estimator In this case, $Q(s, \theta_0) = F(s, \theta_0)$. Let $F_T(s) = I(Z_t \leq s)$. Because $\{Z_t\}_{t=1}^T$ is a random sample, $\{\sqrt{T}(F_T(s) - F(s, \theta_0)) : s \in \mathbb{R}\}$ is P-Donsker, and

$$\begin{aligned} T \int_{-\infty}^{\infty} ((F_T(s) - F(s, \theta_0)))^2 \mathrm{d}s &= T \int_0^1 ((F_T(s) - F(s, \theta_0)))^2 \mathrm{d}s \\ &\leq \left[\sup_{s \in \mathbb{R}} (\sqrt{T}(F_T(s) - F(s, \theta_0))) \right]^2 = O_p(1). \end{aligned}$$

Therefore, Assumption 2.5.4 (i) holds.

Example 2.5.2 shows that $Q(s; \psi)$ is norm-differentiable at $\psi = \psi_0$, and $D(s; \psi_0)$ is $L_2(\mathbb{R} \times \mathbb{S}^{d-1}, \mathrm{d}s\mathrm{d}\varsigma(u))$-measurable. Therefore, Assumption 2.5.3 are satisfied.

Therefore, Assumption 2.5.5 by Lemma 2.6.7.

The MSWD Estimator In MSWD estimator, $Q(s, \theta_0) = F^{-1}(s, \theta_0)$.

Since

$$\int_0^1 (F^{-1}(s;\theta) - F^{-1}(s;\theta_0))^2 \mathrm{d}s = \frac{1}{3}(\theta - \theta_0)^2,$$

Assumption 2.5.2 holds.

Since $F^{-1}(s;\theta)$ is linear in θ, $Q(s,u,\theta)$ is norm-differentiable at $\theta = \theta_0$ with

$$D(s,\theta_0) = \begin{cases} 4s & \text{if } 0 \leq s \leq 1/4, \\ \frac{4}{3}(1-s) & \text{if } 1/4 \leq s \leq 1. \end{cases}$$

Therefore, Assumptions 2.5.3 and 2.5.6 hold.

Since $\{\sqrt{T}(F_T(s) - F(s,\theta_0)) : s \in \mathbb{R}\}$ is P-Donsker, and $\frac{\partial F^{-1}(s,\theta_0)}{\partial s}$ is uniformly bounded above by an absolute positive constant, we can show that Assumption 2.5.4 (i) holds by following the proof of Lemma 2.6.2 and 2.6.5. Since $D(s,\theta_0)$ and $F^{-1}(s,\theta_0)$ are bounded by some absolute constants, we can show Assumption 2.5.5 by Lemma B.6.3.

B.6 One-Dimensional Unconditional Models

This appendix section consists of two parts. In Section B.6.1 below, We will present relevant conditions to verify Assumptions in Section 5. In subsequent sections, we will present some examples of one-dimensional unconditional models.

Because we deal with one-dimensional models, we will use $Q(s;\theta)$ and $D(s;\theta)$ instead of using $Q(s;u,\theta)$ and $D(s;\theta)$ in this appendix section.

B.6.1 Verification of Assumptions

We will present relevant conditions to verify Assumptions 2.5.4 (i) and 2.5.5 with

$$V_0 = \begin{pmatrix} \Omega_0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Verification of Assumption 2.5.4 (i)

When $d = 1$, there are two sets of results that we can use to verify Assumption 2.5.4 (i).

When $\{Z_t\}_{t=1}^T$ is a random sample, we can use results in Barrio et al. (2005).

Lemma B.6.1 (Theorem 4.6 (i) in Barrio et al. (2005)). *Suppose $d = 1$, and we have a random sample $\{Z_t\}_{t=1}^T$ with the cdf $F(z)$. Let B be the Brownian bridge on $(0, 1)$.*

If F is twice differentiable on its open support (a_F, b_F) with $f(x) = F'(x) > 0$, and

$$\sup_{0<x<1} \frac{x(1-x)|f'(F^{-1}(x))|}{f^2(F^{-t}(x))} < \infty,$$

$$\textit{either } a_F > -\infty \textit{ or } \liminf_{x\to 0+} \frac{|f'(F^{-1}(x))|x}{f^2(F^{-1}(x))} > 0,$$

$$\textit{either } b_F < \infty \textit{ or } \liminf_{x\to 0+} \frac{|f'(F^{-1}(1-x))|x}{f^2(F^{-1}(1-x))} > 0.$$

Also, we assume that $w(s)$ is a bounded nonnegative measurable function on (0, 1) such that

$$\lim_{x\to 0+} \frac{x\int_0^x w(t)dt}{f^2(F^{-1}(x))} = 0, \ \lim_{x\to 0+} \frac{x\int_0^x w(t)dt}{f^2(F^{-1}(1-x))} = 0, \ \textit{and} \int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))} w(t)dt < \infty.$$

Then, we have

$$\sqrt{n}(F_n^{-1}(s) - F^{-1}(s))\sqrt{w(s)} \to \frac{B(s)\sqrt{w(s)}}{\sqrt{f^2(F^{-1}(s))}} \textit{ in law in } L_2(0,1)$$

In particular,

$$T\int_0^1 (F_T^{-1}(s) - F^{-1}(s))^2 w(s)\mathrm{d}s \xrightarrow{d} \int_0^1 \frac{B^2(s)}{f^2(F^{-1}(s))} w(s)\mathrm{d}s.$$

When $\{Z_t\}_{t=1}^T$ is a stationary data, we can apply Corollary 3.3 in Csörgö and Yu (1996).

Lemma B.6.2. *Let $\{Z_t\}$ be a stationary sequence of real-valued random variable with common continuous distribution function $F(z)$. Assume that F is twice differentiable on its open support (a_F, b_F) with $f(x) = F'(x) > 0$, and*

$$\sup_{0<x<1} \frac{x(1-x)|f'(F^{-1}(x))|}{f^2(F^{-t}(x))} := \gamma < \infty.$$

We also assume that

- $\min\{A_F, B_F\} > 0$ *where* $A_F := \lim_{x\downarrow a_F} f(x) < \infty$ *and* $B_F := \lim_{x\downarrow b_F} f(x) < \infty$, *and*

- *If $A_F = 0$ ($B_F = 0$), then f is non-decreasing (non-increasing) on an interval to the right of a (to the left of b).*

In addition, we assume that the weight function $w(s)$ is a bounded nonnegative measurable function on (0, 1) such that

$$\int_0^1 \frac{w(s)}{f^2(F^{-1}(s))} ds < \infty.$$

Then, Assumption 2.5.4 (i) holds if one of the following holds.

- *$\{Z_t\}$ is a stationary α-mixing process with α-mixing coefficient α_k satisfying*

$$\alpha_k = O(k^{-\theta-\epsilon}) \text{ for some } \theta \geq \max\{1+\sqrt{2}, 2\gamma - 1\} \text{ and } \epsilon > 0.$$

- *$\{Z_t\}$ is a stationary ϱ-mixing process with α-mixing coefficient ρ_k satisfying $\sum_{k=1}^{\infty} \varrho(2^k) < \infty$.*

Proof. Under the conditions on F and mixing cofficient, by Corollary 3.3 in Csörgö and Yu (1996) we have

$$\sqrt{T} f(F^{-1}(s))(F_T^{-1}(s) - F^{-1}(s)) \Rightarrow \mathcal{B}(\cdot) \text{ in } D[0,1] \text{ with Skorokhod topology,}$$

where $\{\mathcal{B}(s)\}$ is zero-mean Gaussian process with $\mathcal{B}(0) = 1$ and $\mathcal{B}(1) = 0$ and covariance function

$$\sigma^2(s,j) = (s \wedge t) - st + \sum_{j=2}^{\infty} [Cov(V_1 \leq s, V_j \leq t) + Cov(V_1 \leq t, V_j \leq s)] \tag{B.5}$$

where $V_j = F^{-1}(Z_j)$.

Then, we have

$$T \int_0^1 (F_T^{-1}(s) - F^{-1}(s))^2 w(s) \mathrm{d}s \xrightarrow{d} \int_0^1 \frac{B^2(s)}{f^2(F^{-1}(s))} w(s) \mathrm{d}s.$$

□

Verification of Assumption 2.5.5

When $d = 1$, this assumption can be shown by using the asymptotic distribution of L-statistics, see, e.g., Shorack and Wellner (2009). Mehra and Rao (1975a,b), and Puri and Tran (1980)

For some constants b_1, b_2, and positive constants M, ϵ, we define

$$B(s) = Mt^{-b_1}(1-t)^{-b_2}, \text{ and } K(s) = Mt^{-1+b_1+\epsilon}(1-t)^{-1+b_2+\epsilon}.$$

Condition B.6.1 (Bounded growth). *There exist some constants b_1, b_2, and positive constants M, ϵ with $\epsilon > \frac{1}{2}$ satisfying the following conditions.*

(i) $|D_i(s, \psi_0)w(s)| \leq B(s)$ *for all i and* $0 < s < 1$ *with* $\max(b_1, b_2) < 1$, *where D_i is the i-th element of $D(s, \psi_0)$;*

(ii) $|F^{-1}(s)| \leq K(s)$ *for* $0 < s < 1$.

Condition B.6.2 (Smoothness). *$D_i(s, \psi_*)w(s)$ is continuous except for a set of points s such that its measure with respect to Q is zero.*

Lemma B.6.3. *Suppose $\{Z_t\}_{t=1}^T$ is i.i.d. sample, and Conditions B.6.1 and B.6.2 hold. Then, Assumption 2.5.5 holds with* $V_0 = \begin{pmatrix} \Omega_0 & 0 \\ 0 & 0 \end{pmatrix}$, *where*

$$\Omega_0 = \int_0^1 \int_0^1 [s \wedge t - st] D(s, \psi_0) D(s, \psi_0)^\top w(s) w(t) dQ(s) dQ(t).$$

Proof of Lemma B.6.3. This is the consequence of Theorem 1 and Remark 2 in Chapter 19 of Shorack and Wellner (2009). □

Condition B.6.3. *$\{Z_t\}_{t=1}^T$ is strong mixing process with* $\sum_{k=1}^{\infty} k^2 \alpha^{\delta}(k) < \infty$ *for some* $0 < \delta < 1$.

Condition B.6.4. *There exist some constants b_1, b_2, and positive constants M, ϵ with $\epsilon > \frac{1}{2}$ satisfying the following conditions.*

(i) $|D_i(s,\psi_0)w(s)| \leq B(s)$ *for all* i *and* $0 < s < 1$ *with* $\max(b_1, b_2) < 1$*, where* D_i *is the* i*-th element of* $D(s,\psi_0)$*;*

(ii) $|F^{-1}(s)| \leq K(s)^{1/2+\delta/2(b_1+1)} \times (1-t)^{1/2+\delta/2(b_2+1)}$ *for* $0 < s < 1$*;*

(iii) $\int_0^1 [s(1-s)]^{(1-\delta)/2}\overline{B}(s)\mathrm{d}|Q| < \infty$ *where* $\overline{B}(s) = s^{-b_1(1+\delta/2)}(1-s)^{-b_2(1+\delta/2)}$.

Here δ *is a constant in Condition B.6.3.*

Lemma B.6.4. *Conditions B.6.2, B.6.3, and B.6.4 hold. Then, Assumption 2.5.5 holds with* $V_0 = \begin{pmatrix} \Omega_0 & 0 \\ 0 & 0 \end{pmatrix}$ *where*

$$\Omega_0 = \int_0^1 \int_0^1 \sigma^2(s,t) D(s,\psi_0) D(s,\psi_0)^\top w(s) w(t) dQ(s) dQ(t)$$

in which

$$\sigma^2(s,j) = (s \wedge t) + \sum_{j=1}^{\infty} [\Pr(V_1 \leq s, V_j \leq t) - st] + \sum_{j=2}^{\infty} [\Pr(V_1 \leq t, V_j \leq s) - st]$$

where $V_j = F(Z_j)$ *follows* $U[0,1]$*.*

Proof. Proof is skipped because the proof is the same as Theorem 5.2 of Puri and Tran (1980) with replacing J_n by J in Puri and Tran (1980). □

B.6.2 Examples

In this subsection, we will deal with the MSWD estimator of two-sided uniform models, and independent private procurement auction. For simplicity, we will consider the case where data is a random sample, and $w(s) = 1$.

The Two-sided Uniform Model

In this subsection, we will derive the asymptotic distribution of two-sided uniform model directly using Lemma B.6.3. When $w(s) = 1$, the MWSD estimator θ_n have the following form.

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = 3\sqrt{T}\int_0^1 (Q_T(s) - Q(s))D(s, \theta_0)ds,$$

where $Q(\cdot)$ and $Q_T(\cdot)$ are parametric and empirical quantile functions,respectively, and

$$D(s, \theta_0) = \begin{cases} 4s & \text{when } 0 \le s \le \frac{1}{4}, \\ \frac{4}{3}(1-s) & \text{when } \frac{1}{4} < s \le 1. \end{cases}$$

In two-sided uniform model, Condition B.6.1 and B.6.2 hold. This is because $Q(\cdot)$ and $D(\cdot, \theta_0)$ are uniformly bounded below and above by constants, and $D(\cdot, \theta_0)$ is continuous on $(0, 1)$ almost everywhere with respect to Q. Therefore, by Lemma B.6.3,

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = 3\sqrt{T}\int_0^1 (Q_T(s) - Q(s))D(s, \theta_0)ds \xrightarrow{d} N(0, \sigma_0^2),$$

where

$$\begin{aligned} \sigma_0^2 &= \int_0^1 \int_0^1 [s \wedge t - st]D(s, \theta_0)D(t, \theta_0)dQ(s)dQ(t) \\ &= \frac{19}{720}\theta_0 + \frac{1}{24}\theta_0(1-\theta_0) + \frac{3}{80}(1-\theta_0)^2. \end{aligned}$$

Note that σ_0^2 is positive when $0 \le \theta_0 \le 1$.

Independent Private Value Procurement Auction

Consider an econometric model of an independent private value procurement auction formulated in Paarsch (1992) and Donald and Paarsch (2002). Let Y be the winning bid for the auction considered. Suppose the bidder's private value V follows a distribution of the form

$$f_V(v, \theta)I(v \ge g_V(\theta)). \tag{B.6}$$

Assuming a Bayes-Nash Equilibrium solution concept, the equilibrium biding function satisfies

$$\sigma(v) = v + \frac{\int_v^\infty (1 - F_V(\xi))^{m-1} d\xi}{(1 - F_V(v))^{m-1}}, \tag{B.7}$$

where m is the number of bidders in the auction and $F_V(v)$ denote the cdf of V.

Suppose V follows an exponential distribution with

$$f_V(v;\theta) = \frac{1}{h(\theta)} \exp\left(-\frac{v}{h(\theta)}\right) \text{ and } g_V(\theta) = 0,$$

where $E(V) = h(\theta)$. The winning bid distribution is given by (B.6) with

$$f(y;\theta) = \frac{m}{h(\theta)} \exp\left(-\frac{m}{h(\theta)}\left(y - \frac{h(\theta)}{m-1}\right)\right) \text{ and } g(\theta) = \frac{h(\theta)}{m-1}.$$

The cumulative distribution function $F(y;\theta)$ is given by

$$F(y;\theta) = 1 - \exp\left(-\frac{m}{h(\theta)}\left(y - \frac{h(\theta)}{m-1}\right)\right) \text{ for } y \geq g(\theta) = \frac{h(\theta)}{m-1}.$$

Inverting $F(\cdot;\theta)$, we obtain the conditional quantile function $Q(s;\theta)$ given by

$$Q(s;\theta) = h(\theta)\left[\frac{1}{m-1} - \frac{1}{m}\log(1-s)\right].$$

Without loss of generality, suppose $h(\theta) = \theta$. Then, the MSWD estimator is

$$\hat{\theta}_T = \left[\int_0^1 \left(\frac{1}{m-1} - \log(1-s)\right)^2 \mathrm{d}s\right]^{-1} \int_0^1 \left[\frac{1}{m-1} - \log(1-s)\right] Q_T(s)\mathrm{d}s$$

and

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = \left[\int_0^1 D^2(s;\theta_0)\mathrm{d}s\right]^{-1} \sqrt{T}\int_0^1 (Q_T(s) - Q(s))D(s;\theta_0)\mathrm{d}s,$$

where $Q(s) = Q(s;\theta_0)$ and

$$D(s;\theta_0) = \frac{1}{m-1} - \log(1-s).$$

Then, we will show the positive definiteness of hessian term $\int_0^1 D^2(s;\theta_0)\mathrm{d}s$, and asymptotic normality of $\sqrt{T}\int_0^1 (Q_T(s) - Q(s))D(s;\theta_0)\mathrm{d}s$ using Lemma B.6.3 in Steps 1 and 2 below.

Step 1. We can show $\int_0^1 D^2(s;\theta_0)\mathrm{d}s$ is positive definite and finite. Straightforward calculation shows

$$\int_0^1 \left(\frac{1}{m-1} - \log(1-x)\right)^2 ds = \frac{1}{(m-1)^2} + \frac{2}{m-1} + 2 < \infty.$$

Step 2. We will verify Conditions B.6.1 and B.6.2 to exploit Lemma B.6.3 in Steps 2-1 and 2-2 below.

Note that when Lemma B.6.3 holds, we have

$$\sqrt{T}\int_0^1 (Q_T(s) - Q(s))D(s;\theta_0)\mathrm{d}s \xrightarrow{d} N(0,\sigma^2),$$

where

$$\sigma^2 = \int_0^1\int_0^1 [\min\{s,t\} - st]D(s)D(t)dQ(s)dQ(t) = \theta_0^2\left[5 + \frac{4m-3}{(m-1)^2}\right].$$

Steps 2-1. We will verify Conditions B.6.1.

Since $Q(s;\theta_0) = \theta_0 D(s,\theta_0)$, we will show that we can choose sufficiently small positive constant c and sufficiently large positive constant C such that

$$\frac{|D(s,\theta_0)|}{s^{-c}(1-s)^{-c}} < C \text{ for all } 0 < s < 1.$$

If there exists sufficiently small and positive constant $c < 1/4$, and sufficiently large C satisfying above condition, we can find $\epsilon > 1/2$, b_1, and b_2 with $\max\{b_1, b_2\} < 1$ in Condition B.6.1: $b_1 = b_2 = c$, and $\epsilon = 1 - 2c$.

Note that when $c > 0$,

$$s^c(1-s)^c|Q(u)| \le s^c(1-s)^c\left|\frac{1}{m-1} + \log(1-s)\right| \le C_1 + \frac{|\log(1-s)|}{(1-s)^{-c}}$$

for some positive constant C which only depends on m, and

$$\lim_{s\to 0+}\frac{|\log(1-s)|}{(1-s)^{-c}} = 0, \text{ and } \lim_{s\to 1+}\frac{|\log(1-s)|}{(1-s)^{-c}} = 0 \text{ for any } c > 0.$$

This implies that for any $c > 0$, we can choose sufficiently large C_c such that we can pick sufficiently large constant C_c satisfying

$$s^c(1-s)^c|Q(s)| \le C_1 + \frac{|\log(1-s)|}{(1-s)^{-c}} \le C_c \text{ for } 0 < s < 1.$$

Therefore, Condition B.6.1 holds.

Step 2-2. We will verify Condition B.6.2. In this model, $D(s;\theta_0)$ and $Q(s)$ are continuously differentiable and strictly increasing in s. Therefore, Condition B.6.2 holds.

From Steps 2-1 and Step 2-2, we can use Lemma B.6.3. By combining all results in Steps 1 and 2, we have

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N\left(0, \mathit{\Omega}\right),$$

where

$$\mathit{\Omega} = \theta_0^2 \frac{(m-1)^2(5m^2-6m+2)}{(2m^2-2m+1)^2}.$$