Minimum Sliced Distance Estimation in Structural Models*

Yanqin Fan† and Hyeonseok Park ‡

This version: November 14, 2021
Current Version can be found at https://sites.google.com/view/hynskpark.

Abstract

This paper develops a simple and robust method for the estimation and inference in structural econometric models based on sliced distances. Three motivating models considered are asset pricing/state space models, aggregate demand models, and models with parameter-dependent supports. 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" such as in the asset pricing/state space models and parameter-dependent supports such as in the one-sided and two-sided models. Furthermore, our estimator is applicable 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.

Keywords: Adversarial Estimation; Generative Model; Parameter-Dependent Support; Sliced Cramer Distance; Sliced Wasserstein Distance; Stochastic Singularity.

JEL Codes: C1; C31; C32

*We thank Debopam Bhattacharya, Sergey Bobkov, Tetsuya Kaji, Ruixuan Liu, Serena Ng, and Dacheng Xiu for helpful discussions and providing important references.

†Department of Economics, University of Washington, Seattle, WA 98195, USA; email: fany88@uw.edu.

‡Department of Economics, University of Washington, Seattle, WA 98195, USA; email: parkh27@uw.edu.

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 guaranteed, 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],[1] 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 2.3 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.[2] 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 aforementioned 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 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

[1] 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.

[2] 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.

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 3.3 for a detailed discussion.

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 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 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 2.1. *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 2.1 (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 $f^*(\cdot|y_{t-1}^*; \gamma_0)$. 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 (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 $N\left(0, \tau^2 I_{J_t}\right)$ 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 $N\left(0, \sigma_\lambda^2 I_J\right)$, $N\left(0, \sigma_f^2 I_J\right)$, $N\left(0, \sigma_e^2 I_J\right)$ 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 2.3 (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 $Z_t = \left(Y_t^\top, X_t^\top\right)^\top$.

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.

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

$$\mathcal{W}_2^2(\mu, \nu) = \int_0^1 \left(F_\mu^{-1}(s) - F_\nu^{-1}(s) \right)^2 \mathrm{d}s, \tag{3.1}$$

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:

$$\mathcal{C}_2^2(\mu, \nu) = \int_{-\infty}^{\infty} (F_\mu(s) - F_\nu(s))^2 \mathrm{d}s. \tag{3.2}$$

In fact, the following result holds.

Lemma 3.1 (Theorem 2.11 in Bobkov and Ledoux [2019])**.** *Suppose μ and ν are probability measures in $\mathcal{P}_2(\mathbb{R})$. Then,*

$$\begin{aligned} \mathcal{W}_2^2(\mu, \nu) &= \int_{-\infty}^{\infty} (F_\mu(u \wedge v) - F_\nu(u \vee v))^+ \mathrm{d}u\mathrm{d}v + \int_{-\infty}^{\infty} (F_\nu(u \wedge v) - F_\mu(u \vee v))^+ \mathrm{d}u\mathrm{d}v \\ &= 2 \iint_{u \le v} \left[(F_\mu(u) - F_\nu(v))^+ + (F_\nu(u) - F_\mu(v))^+ \right] \mathrm{d}u\mathrm{d}v. \end{aligned} \tag{3.3}$$

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:

$$\mathcal{SW}(\mu, \nu) = \left[\int_{\mathbb{S}^{d-1}} \mathcal{W}^2(u^*_\# \mu, u^*_\# \nu) d\varsigma(u) \right]^{1/2}, \tag{3.4}$$

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]. 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:

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

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

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

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

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \int_{\mathcal{S}} (Q_T(s;u) - \widehat{Q}_T(s;u,\hat{\psi}_T))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \\
=\ & \inf_{\psi\in\Psi} \int_{\mathbb{S}^{d-1}} \int_{\mathcal{S}} (Q_T(s;u) - \widehat{Q}_T(s;u,\psi))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) + o_p(T^{-1}),
\end{aligned} \tag{3.5}$$

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 2.1, $\widehat{Q}_T(\cdot;u,\psi)$ 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 2.1 (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

$$Y_t = a + bY_t^*,$$

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.

For *conditional models* such as Examples 2.2 and 2.3, $\widehat{Q}_T(\cdot; u, \psi)$ is random.

Example 2.3 (Cont'd). Let $Z = \left(Y, X^\top\right)^\top$. Then the cdf of Z and $u^\top Z$ are given by

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

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

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

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

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

Example 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

$$\begin{aligned}&\int_{\mathbb{S}^{d-1}} \int (Q_T(s; u) - \widehat{Q}_T(s; u, \hat{\psi}_{T,m}))^2 w(s) \mathrm{d}s d\varsigma(u) \\ &= \inf_{\psi \in \mathcal{H}} \int_{\mathbb{S}^{d-1}} \int (Q_T(s; u) - \widehat{Q}_T(s; u, \psi))^2 w(s) \mathrm{d}s d\varsigma(u) + \nu_T, \end{aligned} \tag{3.6}$$

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

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

$$\mathcal{W}^2(u^*_{\#}\mu, u^*_{\#}\mu(\psi)) = \max_{\varphi} \left[\int \varphi(x)\, du^*_{\#}\mu(x) + \int \varphi^*(y)\, du^*_{\#}\mu(y; \psi) \right], \tag{3.7}$$

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 $u^*_{\#}\mu\left(\psi\right)$ has a continuous distribution, there exists a unique transport map T such that

$$\nabla \varphi_u\left(x;\psi\right) = x - T\left(x;u,\psi\right) = x - Q(G(s;u,\psi);u) \tag{3.8}$$

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

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

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

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

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 (3.8) 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.

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[3] exhibit consistent behavior for all examples and parameter values. We then draw QQ plots of MLE, JS, MSWD, and MSCD estimators.

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 2.1 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

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

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

$$\mathcal{W}_2^2(\mu_0,\mu_\theta) = \theta^2,\ \mathcal{SW}^2(\mu_0,\mu_\theta) = \frac{\theta^2}{2},$$

$$\mathcal{C}_2^2(\mu_0,\mu_\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(F_Y(x,z;0) - F_Y(x,z;\theta))^2\mathrm{d}z\mathrm{d}x = \frac{|\theta|}{3},$$

$$\mathcal{SC}^2(\mu_0,\mu_\theta) = \begin{cases} 0 & \text{if } \theta = 0, \\ \frac{2\left(2|\theta|^3 - 1 - (2\theta^2-1)\sqrt{\theta^2+1} - 3\theta^2\log(\tan(0.5\tan^{-1}(|\theta|)))\right)}{3\pi} & \text{otherwise.} \end{cases}$$

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 $\theta_0 \neq 0$.

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

Remark 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.

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

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

where $\mathrm{Exp}\left(\frac{1}{\theta_0}\right)$ 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

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

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 $\theta_0\ne 1/4$,

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

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,

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

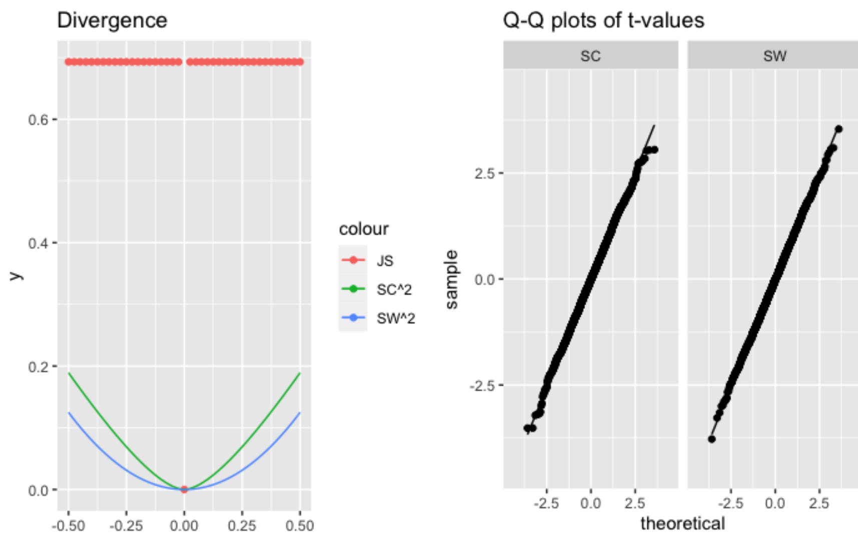

Figure 1: A comparison of several divergences for Example 1 in Arjovsky et al. [2017] with $T = 1000$

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

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

4.4 Numerical Illustration

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

Table 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.

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 with each value computed from a random sample of size 1000, see Figures 1 to 4.

Several conclusions can be drawn from Figures 1 and 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

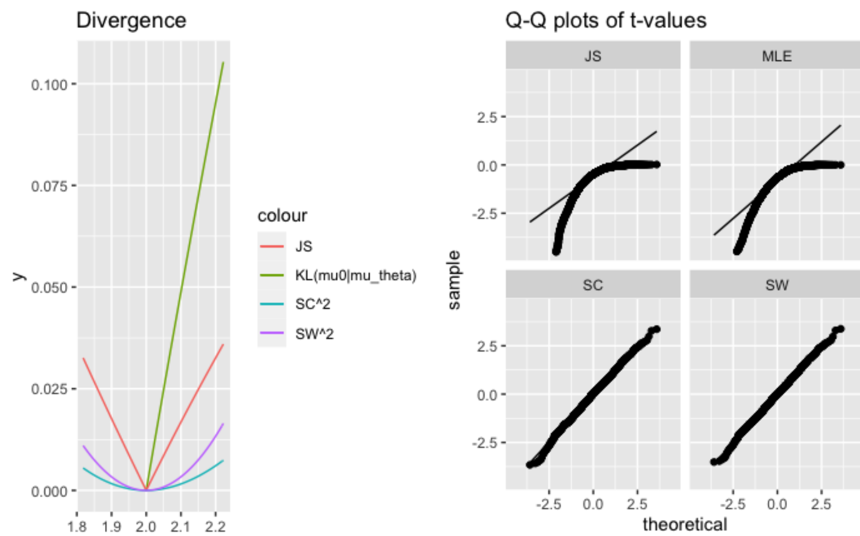

Figure 2: A comparison of several divergences when $\theta_0 = 2$ with $T = 1000$ for one-sided uniform model.

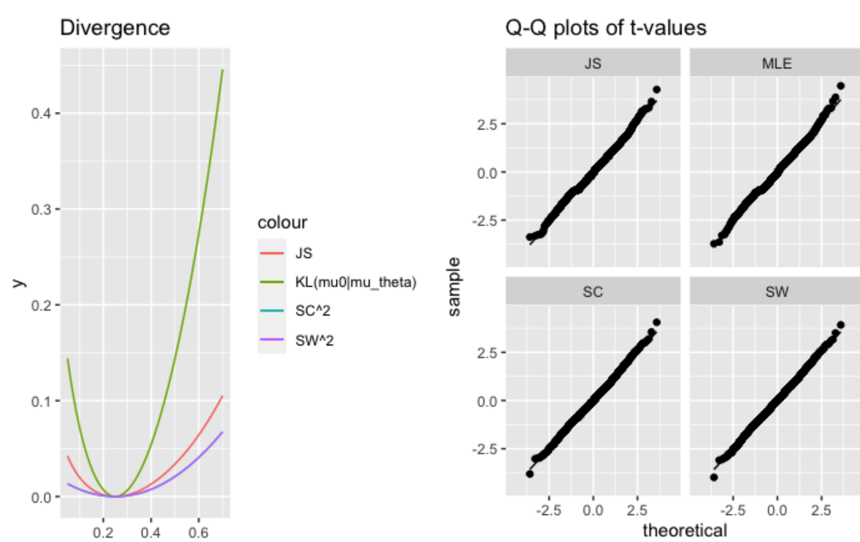

Figure 3: A comparison of several divergences when $\theta_0 = 1/4$ with $T = 1000$ for two-sided uniform model.

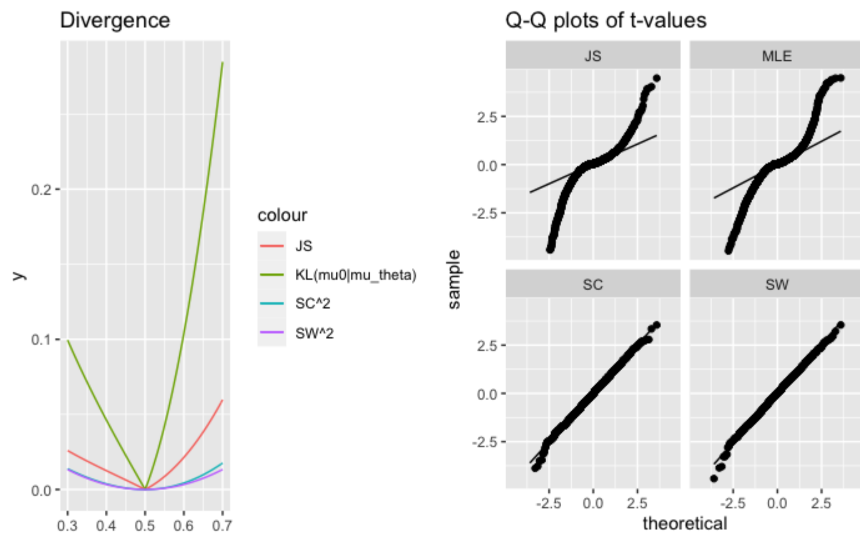

Figure 4: A comparison of several divergences when $\theta_0 = 1/2$ with $T = 1000$ for two-sided uniform model

estimator are non-normal and the distributions of MSWD and MSCD estimators are close to being normal. Indeed, we will verify in Section E 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.

5 Asymptotic Theory

In this section, we establish asymptotic theory for the general MSD estimator $\hat{\psi}_T$ defined in (3.5) 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.

5.1 Assumptions and Asymptotic Properties

Assumption 5.1. *(i)* $\int_{\mathbb{S}^{d-1}} \int_{\mathcal{S}} \left(Q_T(s;u) - Q(s;u)\right)^2 w(s) \mathrm{d}s d\varsigma(u) \xrightarrow{p} 0$*;*

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

Assumption 5.1 imposes that $Q_T(\cdot,\cdot)$ 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 5.2 below states that ψ_0 is well-separated.

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

Theorem 5.1 (Consistency of $\hat{\psi}_T$). *Suppose Assumptions 5.1 and 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 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 5.4 (i) below strengthens Assumption 5.1 (i).

Assumption 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 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 5.6.

$$B_0 := \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} D(s;u,\psi_0)D^\top(s;u,\psi_0)w(s)\mathrm{d}s d\varsigma(u) \tag{5.1}$$

is positive definite.

Theorem 5.2 (Asymptotic normality of $\hat{\psi}_T$)**.** *Suppose Assumptions 5.1 to 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 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.

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 5.4 (ii) and (iii) automatically hold. The norm-differentiability assumption in

Assumption 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 5.3 is satisfied in the one-sided and two-sided uniform models for the unweighted MSCD estimator, see Section E in the appendix for the MSWD estimator.

Example 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

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

Example 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,

$$\int_{-\infty}^{\infty}\Big(F(y;\psi)-F(y;\psi_0)-(\psi-\psi_0)\widehat{D}_T(s;\psi_0)\Big)^2 ds$$
$$=\begin{cases}-\frac{(\psi-\psi_0)^3(4\psi(\psi_0-1)+12\psi_0^2-4\psi_0+1)}{48(\psi-1)(\psi_0-1)\psi_0^2} & \text{if } \psi<\psi_0,\\ \frac{(\psi-\psi_0)^3(12\psi\psi_0+4\psi_0^2-8\psi_0+1)}{48\psi\psi_0(1-\psi_0)^2} & \text{if } \psi\geq\psi_0.\end{cases}$$

and

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

In contrast, the norm differentiability assumption for the asymptotic normality of MLE is not satisfied in the one-sided uniform model, see Remark 5.2 below.

Remark 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 $\psi\to\psi_0$,

$$\int\left[\sqrt{f(x,\psi)}-\sqrt{f(x,\psi_0)}-\frac{1}{2}(\psi-\psi_0)^\top d(x;\psi_0)\sqrt{f(x,\psi_0)}\right]^2 \mathrm{d}x=o(\|\psi-\psi_0\|^2) \tag{5.2}$$

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 (5.2). 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 (5.2) holds. When $\psi>\psi_0$,

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

However, it is a contradiction to the assumption that equation (5.2) holds.

6 Unconditional Models

For unconditional models, we can take $\widehat{Q}_T(\cdot; u, \psi) := Q(\cdot; u, \psi)$ in (3.5), 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 6.1. *(i) $\hat{\psi}_T \xrightarrow{p} \psi_0$ as $n \to \infty$ under Assumptions 5.1 (i) and 5.2; (ii) $\sqrt{T}(\hat{\psi}_T - \psi_0) \xrightarrow{d} N(0, B_0^{-1}\Omega_0 B_0^{-1})$ under Assumptions 5.1 (i), 5.2, 5.3, 5.4 (i), 5.6, and 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 5.4 (i) and 5.5 for β-mixing processes since Assumption 5.1 (i) is implied by Assumption 5.4 (i).

Definition 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 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,

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

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

6.1 MSWD Estimator

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

6.1.1 Verification of Assumption 5.4 (i)

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

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

It is known that $\{I(u^\top Z_t \le s) : u \in \mathbb{S}^{d-1}, s \in \mathbb{R}\}$ 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 C.1) to establish our first lemma below.

Lemma 6.1. *Under Condition 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*

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

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

$$\begin{aligned} y_t &= a + b y_t^*, \\ y_t^* &= \rho y_{t-1}^* + e_t,\ e_t \sim N(0,1). \end{aligned}$$

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 5.4 (i).

Lemma 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*

$$\int_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \int_0^1 \frac{w(s)}{g^2(G^{-1}(s; u, \psi_0); u, \psi_0)} \mathrm{d}s\mathrm{d}\varsigma(u) < C < \infty, \tag{6.1}$$

and

$$\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0,} \sup_{s\in supp\{w\}} \left| \frac{g(G^{-1}(s; u, \psi_0); u, \psi_0)}{g(G^{-1}(\tilde{s}(s; u); u, \psi_0); u, \psi_0)} \right| = O_p(1) \tag{6.2}$$

for $\tilde{s}(s, u, \psi_0)$ *such that* $\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s\in supp\{w\}} |\tilde{s}(s; u, \psi_0) - s| = o_p(1)$. *Then, Assumption 5.4 (i) holds under Condition 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 (6.1) and (6.2) in Lemma 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 6.1. When $w(s) = 1$ for $s \in [\delta, 1-\delta]$ for some $\delta > 0$, we can show equation (6.2) under the following assumption:

$$\begin{aligned}
&\sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s \in \mathbb{R}} (G(s; u, \psi_0)(1 - G(s; u, \psi_0))) \left| \frac{g'(s; u, \psi_0)}{g^2(s; u, \psi_0)} \right| \\
&= \sup_{u \in \mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s \in (0,1)} (s(1-s)) \left| \frac{g'(G^{-1}(s; u, \psi_0); u, \psi_0)}{g^2(G^{-1}(s; u, \psi_0); u, \psi_0)} \right| < \infty,
\end{aligned} \tag{6.3}$$

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

To show (6.3), we can use arguments in the proof of Lemma 1 in Csorgo and Revesz [1978]. For $u \in \mathbb{S}^{d-1}/\mathcal{N}_0$,

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

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

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

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

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

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

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

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

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

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

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

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

6.1.2 Verification of Assumption 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 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*

$$\int_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 \frac{w(s)}{g^2(G^{-1}(s;u,\psi_0);u,\psi_0)}\mathrm{d}u\mathrm{d}\varsigma(u) < C < \infty, \tag{6.4}$$

and

$$\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\ \sup_{s\in supp\{w\}}\left|\frac{g(G^{-1}(s;u,\psi_0);u,\psi_0)}{g(G^{-1}(\tilde{s}(s;u;\psi_0);u,\psi_0);u,\psi_0)}\right| = O(1) \tag{6.5}$$

for $\tilde{s}(s,u,\psi_0)$ such that $\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in(0,1)}|\tilde{s}(s;u,\psi_0)-s| = o(1)$. 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 (6.3) implies Equation (6.5).

Lemma 6.4. *Let's denote $\phi(\tilde{Q}):\mathbb{D}_2\to\mathbb{R}$, where*

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

Then, $\phi(\cdot)$ is Hadamard differentiable at $Q\in\mathbb{D}_2$ tangentially to $\mathbb{D}_2$ under Assumption 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 $Q_n(s;u) = Q(s,u)+t_nh_n(s,u)$ where $t_n\downarrow 0$ as $n\to\infty$, $h_n(s,u)\in\mathcal{D}_2$, $\lim_{n\to\infty}\int_{u\in\mathbb{S}^{d-1}}\int_0^1|h_n(s,u)-h(s,u)|^2w(s)ds d\varsigma(u) = 0$, and $Q, Q_n\in\mathbb{D}_2$.

Then, we have

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

□

Combining Lemmas 6.3 and 6.4, we can show that Assumption 5.5 is satisfied.

Lemma 6.5. *Suppose Assumption 5.3 and all conditions in Lemma 6.3 hold. Then, Assumption 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*

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

6.2 MSCD Estimator

Let $Q(\cdot; u, \psi) = G(\cdot; u, \psi)$. In this subsection, we will verify Assumptions 5.4 (i) and 5.5 for β-mixing process under the following condition.

Condition 6.2. *The weight function $w(s)$ is integrable. That is, $\int_{\mathbb{R}} w(s)\mathrm{d}s < \infty$.*

Lemma 6.6. *Under Conditions 6.1 and 6.2, Assumption 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 6.7. *In addition to Conditions 6.1 and 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 $\sup_{k\ge 0}\beta(k) < \infty$, we can apply Corollary 4.1 of Rio [2017] because $\int_{u\in\mathbb{S}^{d-1}}\int_{-\infty}^{\infty}\Big[I(u^\top Z_i\le s)-G(s;u,\psi_0)\Big]D(s;u,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u)$ is bounded above by a constant. ☐

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

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

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

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

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

We will verify Assumption 5.1 (ii), Assumption 5.3, Assumption 5.4 (ii) and (iii), and Assumption 5.5.

7.1 Verification of Assumptions 5.1 (ii), 5.3, 5.4 (ii) and (iii), and 5.5.

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

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

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

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

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 5.1 (ii). Furthermore, Lipchitz continuity of k is implied by that of F in the expression (7.1).

Lemma 7.1. *Suppose Condition 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 $\psi, \psi' \in \Psi$,*

$$|F(y|x, \psi) - F(y|x, \psi')| \leq M(y, x)\|\psi - \psi'\|, \tag{7.1}$$

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

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

Then, we have the following result.

Lemma 7.2. *Under Condition 6.2, Assumption 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 7.1. *$Q(\cdot; u, \psi)$ is first-order norm-differentiable at $\psi = \psi_0$. That is,*

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

for any $\tau_T \to 0$.

Under condition 7.1, it is enough to show that

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

where

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

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

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

where

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

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 5.3 and 5.4 (iii). Lipchitz continuity of k_2 is implied by that of $F(\cdot|x,\psi)$ in the expression (7.1).

Lemma 7.3. *In addition to all conditions in Lemma 7.1, suppose Condition 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 5.3 and 5.4 (iii) hold.*

Note that

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

By applying CLT, we obtain the following result.

Lemma 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 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

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

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,

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

and

$$
F(y|x,\theta,\gamma) = \begin{cases} \int_{-\infty}^{y-g(x,\theta)} f_{L,\epsilon}(\epsilon|x,\theta,\gamma)\mathrm{d}\epsilon & \text{if } y < g(x,\theta),\\ \int_{-\infty}^{0} f_{L,\epsilon}(\epsilon|x,\theta,\gamma)\mathrm{d}\epsilon + \int_0^{y-g(x,\theta)} f_{U,\epsilon}(\epsilon|x,\theta,\gamma)\mathrm{d}e & \text{if } y \ge g(x,\theta).\end{cases}
$$

7.2.1 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 $\sup_{y\neq g(x,\psi)} |\partial F(y|x,\psi)/\partial\psi|$ is uniformly bounded by a finite absolute constant. Then $F(y|x,\psi)$ is Lipchitz continuous with respect to ψ:

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

This condition is stronger than equation (7.1) 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 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*

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

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 $\psi,\psi'\in\Psi$,

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

The first two assumptions in lemma 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 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*

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

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

In addition, we assume that

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

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 $\psi,\psi'\in\Psi$,

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

Like for one-sided model, the first two assumptions in lemma 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

$$\lim_{\epsilon\downarrow0}f_{U,\epsilon}(\epsilon|x,\psi)-\lim_{\epsilon\uparrow0}f_{L,\epsilon}(\epsilon|x,\psi)>\eta>0$$

as in Chernozhukov and Hong [2004].

7.2.2 Norm Differentiability

Suppose that all assumptions in Lemmas 7.5 and 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 7.5 and 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

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

for some $\tilde{\psi}$. When the weight function is integrable,

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

Therefore, it is enough to show that

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

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 7.7. *For one-sided model, in addition to all conditions in Lemma 7.5, let us assume that $\partial f_\epsilon(\epsilon|x, \psi)/\partial \psi$ is upper semi-continuous in ϵ, ψ, and there exists $\tilde{f}_\epsilon(\epsilon|x)$ such that*

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

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 7.8. *For two-sided model, in addition to all conditions in Lemma 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 $\tilde{f}_{L,\epsilon}(\epsilon|x)$ and $\tilde{f}_{U,\epsilon}(\epsilon|x)$ such that*

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

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

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

8 Generative Models

In this section, we will investigate the asymptotic property of the SMSD estimator $\hat{\psi}_{T,m}$ described in Example 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.

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 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 $k = (k_1, \ldots, k_d)$, 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

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

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

Definition 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 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 8.1. When H_t has unbounded support, one can relax Condition 8.2 (b) by following the proof of Corollary 2.7.4 in van der Vaart and Wellner [1996].

Let

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

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 8.1. *Suppose that Conditions 8.1 and 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 8.2.(a) holds, and $\lim_{k\to\infty} k^r \beta_k = 0$ for some $r > 1$.

(ii) Condition 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 $\sum_{k\ge0} k^{-1/2} \beta_k^{(1-\zeta)\frac{r-1}{2r}} < \infty$.*
- *$\log(N_{[]}(\epsilon, \mathcal{F}_I, \|\cdot\|_H)) = O(\log|\epsilon|)$, and $\sum_{k\ge0} k^{-1} \sqrt{(\log k)^{-1} \sum_{\ell \ge k} \beta_\ell^{1-1/r}} < \infty$.*

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 C.1. For Case (ii), we have weak convergence by Theorem 1 and Example 4 (p. 405) in Doukhan et al. [1995]. □

8.1.1 The SMSWD Estimator

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

Condition 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 6.2, we have the following result.

Lemma 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

$$\sup_{\theta\in\Psi} \sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}(\theta)} \sup_{s\in supp\{w\}} \frac{g(G^{-1}(s; u, \psi); u, \psi)}{g(G^{-1}(\tilde{s}(s; u, \psi); u, \psi); u, \psi)} = O_p(1) \tag{8.1}$$

Then, Conditions in Lemma 8.1 imply

$$\sup_{\psi,s,u} m \int_0^1 (G_m^{-1}(s; u, \psi) - G^{-1}(s; u, \psi))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) = O_p(1). \tag{8.2}$$

Remark 8.2. When $w(s) = [\delta, 1-\delta]$ with $\delta > 0$, we can show equation (8.1) under the following assumptions.

$$\begin{aligned}
&\sup_{\psi} \sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}(\theta),s} (G(s; u, \psi_0)(1 - G(s; u, \psi_0))) \left| \frac{g'(s; u, \psi_0)}{g^2(s; u, \psi_0)} \right| \\
&= \sup_{\psi} \sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}(\theta),s} (s(1-s)) \left| \frac{g'(G^{-1}(s; u, \psi_0); u, \psi_0)}{g^2(G^{-1}(s; u, \psi_0); u, \psi_0)} \right| < \infty
\end{aligned} \tag{8.3}$$

by following the proof of Lemma 1 in Csorgo and Revesz [1978]. Proof is the same as the proof in Remark 6.1.

Corollary 8.1. *Suppose Assumptions 5.1 (i), 5.2, 5.3, 5.4 (i), 5.6, Condition 8.3, and conditions in Lemma 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 6.5.

Proof. From Lemma 8.2, Equation (8.2) holds. This implies Assumption 5.1 (ii) and Assumption 5.4 (ii) are satisfied.

When $T/m = o(1)$ and Conditions 8.2 and 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 (8.2) 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 $T/m = o(1)$. $\square$

8.1.2 The SMSCD Estimator

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

From Lemma 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 8.4. *All functions in $\mathcal{F}$ are Lipschitz continuous with respect to ψ and u.*

Lemma 8.3. *Let us assume that Assumptions 5.1, 5.2, 5.4, 5.6, and Conditions 6.2, 7.1, 8.4, and all assumptions in Lemma 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 6.7.

Proof. We will prove the result by showing that the assumptions in Section 5 hold.

When all conditions in Lemma 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

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

Therefore, Assumption 5.1 (i) and (ii) hold when $T/m = O(1)$, and 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

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

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 7.3. Because $D_T(s;u,\psi_0) = D(s;u,\psi_0)$, Assumption 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 6.7. Therefore, we have the desired result. □

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 8.5.

Condition 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 8.1. *In the BLP model introduced in Section 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 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 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 8.4. *Under Conditions 8.5 and 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 8.6 (a) holds, $\mathcal{F}_X$ is VC subgraph class, which impies $\mathcal{G}_X$ is P-Donsker. Under Condition 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 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 8.3.

Lemma 8.5. *Suppose Conditions 7.1, 8.5, 8.6, and 8.7 hold. Then, Assumptions 5.1 (ii) and 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 8.6. *Suppose Assumption 7.1 and Condition 6.2 hold. Then, we have*

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

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 8.5 and 8.6, we have the asymptotic distribution for $\hat{\psi}_{T,m}$ whose proof is the same as that of Lemma 8.3.

Corollary 8.2. *Suppose Assumptions 5.1 (i), 5.2, 5.4 (i), 5.6, and Conditions 6.2, 7.1, 8.5, 8.6, and 8.7 hold. When $m = T$, we have*

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

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 8.6.*

9 Numerical Results

In this section[4], we report some results on the accuracy of the asymptotic normal distribution of MSWD and MSCD estimators for 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 a term structure model. The parameter-dependent support model is an auction model.

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:

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

[4] This section is preliminary and is under revision.

Table 2: P-values when $\rho = 0.8$

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

	ρ	a_1	a_2	b_1	b_2
$T = 500$	0.0550	0.0520	0.0515	0.0465	0.0490
$T = 1500$	0.0555	0.0435	0.0435	0.0585	0.0575
$T = 2000$	0.0500	0.0515	0.0510	0.0510	0.0505
$T = 2500$	0.0505	0.0525	0.0525	0.0480	0.0490
$T = 3000$	0.0570	0.0470	0.0470	0.0500	0.0495

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

	ρ	a_1	a_2	b_1	b_2
$T = 500$	0.1065	0.1075	0.1080	0.1035	0.1045
$T = 1500$	0.0985	0.1015	0.1015	0.0970	0.0970
$T = 2000$	0.1020	0.0955	0.0955	0.1040	0.1050
$T = 2500$	0.1050	0.1000	0.0995	0.1025	0.1020
$T = 3000$	0.0950	0.1000	0.1000	0.1015	0.1015

Table 3: P-values when $\rho = 0.3$

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

	ρ	a_1	a_2	b_1	b_2
$T = 500$	0.0500	0.0495	0.0500	0.0500	0.0505
$T = 1500$	0.0525	0.0495	0.0500	0.0495	0.0485
$T = 2000$	0.0535	0.0535	0.0540	0.0505	0.0510
$T = 2500$	0.0505	0.0495	0.0495	0.0520	0.0505
$T = 3000$	0.0535	0.0465	0.0465	0.0530	0.0530

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

	ρ	a_1	a_2	b_1	b_2
$T = 500$	0.0985	0.0980	0.0975	0.0960	0.0945
$T = 1500$	0.0965	0.1015	0.1010	0.0995	0.0995
$T = 2000$	0.1020	0.1020	0.1015	0.0975	0.0975
$T = 2500$	0.0925	0.1045	0.1045	0.1040	0.1040
$T = 3000$	0.0985	0.1020	0.1015	0.0940	0.0925

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. We compute 2000 estimates of t-values for samples of size $T = 500, 1500, 2000, 2500$, and 3000. Tables 2 and 3 show that p-values are not far from the nominal sizes for all sample sizes.

9.2 Discrete-Time Vasicek Model

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 are close to empirical results in Pastorello et al. [2003]. We compute 2000 estimates of t-values from the samples of size $T = 500, 1500, 2000, 2500$, and 3000. Table 4 shows that p-values are not far from the nominal sizes for all sample sizes.

Table 4: 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

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

$$\sigma(v) = v + \frac{\int_v^\infty (1 - F_V(\xi|X,\theta,\gamma))^{m-1} d\xi}{(1 - F_V(v|X,\theta,\gamma))^{m-1}}, \tag{9.1}$$

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.

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

$$\begin{aligned}
F(y|X,\theta) &= \int_{g(X,\theta)}^{y} f(u|X,\theta)\,\mathrm{d}u \\
&= \int_{g(X,\theta)}^{y} \frac{m}{h(x,\theta)} \exp\left(-\frac{m}{h(x,\theta)}\left(u - \frac{h(x,\theta)}{m-1}\right)\right) \mathrm{d}u \text{ and } g(X,\theta) = \frac{h(x,\theta)}{m-1} \\
&= \int_0^{y-g(X,\theta)} \frac{m}{h(x,\theta)} \exp\left(-\frac{m}{h(x,\theta)}u\right) \mathrm{d}u \\
&= 1 - \exp\left(-\frac{m}{h(x,\theta)}\left(y - \frac{h(x,\theta)}{m-1}\right)\right) \text{ for } y \geq g(X,\theta).
\end{aligned}$$

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

Figure 5 shows QQ plots of t-values of the MSCD estimator and Li [2010]'s estimator. QQ plots show that both estimators are comparable.

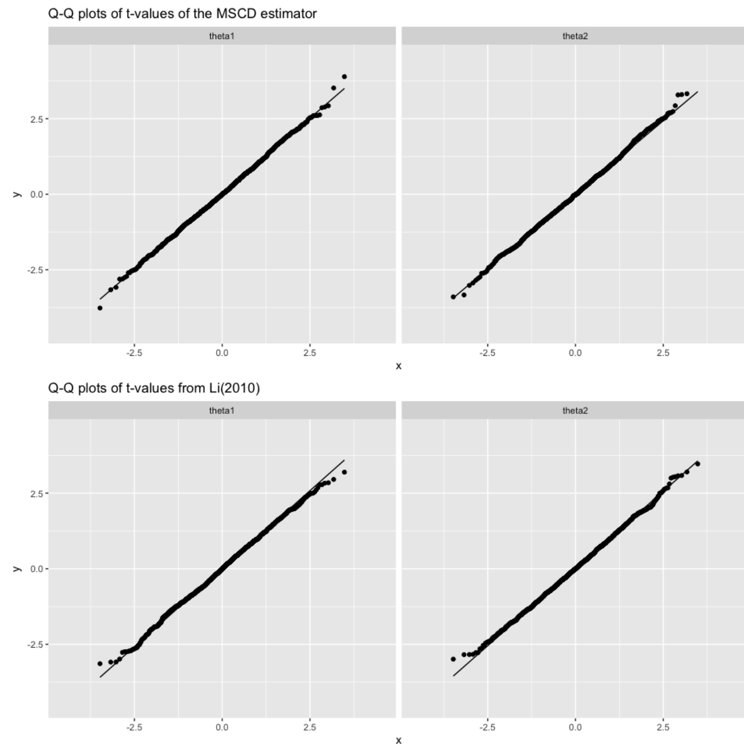

Figure 5: QQ plots of t-values of the MSCD estimator and Li [2010]'s esimator when $T = 100$

10 Concluding Remarks

Motivated by important features of structural models in economics and finance such as stochastic 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.

References

Adil Ahidar-Coutrix and Philippe Berthet. Convergence of Multivariate Quantile Surfaces. 10 2021.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savare. *Gradient Flows*. Birkhäuser-Verlag, 2008. ISBN 9783764373092. doi: 10.1007/b137080.

Sungbae An and Frank Schorfheide. Bayesian Analysis of DSGE Models. *Econometric Reviews*, 26(2-4), 4 2007. ISSN 0747-4938. doi: 10.1080/07474930701220071.

Torben G. Andersen, Luca Benzoni, and Jesper Lund. An Empirical Investigation of Continuous-Time

Equity Return Models. *The Journal of Finance*, 57(3):1239–1284, 6 2002. ISSN 00221082. doi: 10.1111/1540-6261.00460.

Donald W K Andrews. Estimation When a Parameter is on a Boundary. *Econometrica*, 67(6):1341–1383, 1999. ISSN 0012-9682. doi: 10.1111/1468-0262.00082.

M. A. Arcones and B. Yu. Central limit theorems for empirical andU-processes of stationary mixing sequences. *Journal of Theoretical Probability*, 7(1), 1 1994. ISSN 0894-9840. doi: 10.1007/BF02213360.

Martin Arjovsky and Léon Bottou. Towards Principled Methods for Training Generative Adversarial Networks. 1 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 10 2017.

David K. Backus, Silverio Foresi, and Chris Telmer. Discrete-time models of bond pricing. *NBER Working Paper*, (6736), 1998.

Eustasio Del Barrio, Evarist Giné, and Frederic Utzet. Asymptotics for L2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11 (1), 2005. doi: 10.3150/bj/1110228245.

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramer Distance as a Solution to Biased Wasserstein Gradients. 5 2017.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019. ISSN 2049-8764. doi: 10.1093/imaiai/iaz003.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile Prices in Market Equilibrium. *Econometrica*, 63(4):841, 1995. ISSN 0012-9682. doi: 10.2307/2171802.

Steven T Berry. Estimating Discrete-Choice Models of Product Differentiation. *The RAND Journal of Economics*, 25(2):242, 1994. ISSN 0741-6261. doi: 10.2307/2555829.

Eric Beutner and Laurent Bordes. Estimators Based on Data-Driven Generalized Weighted Cramér-von Mises Distances under Censoring - with Applications to Mixture Models. *Scandinavian Journal of Statistics*, 38 (1), 2011. ISSN 03036898. doi: 10.1111/j.1467-9469.2010.00708.x.

Sergey Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 261(1259):0, 2019. doi: 10.1090/memo/1259.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision*, 51(1), 1 2015. ISSN 0924-9907. doi: 10.1007/s10851-014-0506-3.

Audra J. Bowlus, Nicholas M. Kiefer, and George R. Neumann. Equilibrium search models and the transition from school to work. *International Economic Review*, 42(2), 2001. ISSN 00206598. doi: 10.1111/1468-2354.00112.

Richard C Bradley. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probability Surveys*, 2(1):107–144, 2005. doi: 10.1214/154957805100000104.

Francois-Xavier Briol, Alessandro Barp, Andrew B. Duncan, and Mark Girolami. Statistical Inference for Generative Models with Maximum Mean Discrepancy. 6 2019.

Donald J. Brown and Marten H. Wegkamp. Weighted minimum mean-square distance from independence estimation, 2002. ISSN 00129682.

Qihui Chen and Zheng Fang. Inference on functionals under first order degeneracy. *Journal of Econometrics*, 210(2):459–481, 10 2019. ISSN 0304-4076. doi: 10.1016/j.jeconom.2019.01.011.

Ren-Raw Chen and Louis Scott. Maximum Likelihood Estimation for a Multifactor Equilibrium Model of the Term Structure of Interest Rates. *The Journal of Fixed Income*, 3(3):14–31, 10 1993. ISSN 1059-8596. doi: 10.3905/jfi.1993.408090.

Mikhail Chernov. [Iterative and Recursive Estimation in Structural Nonadaptive Models]: Comment. *Journal of Business and Economic Statistics*, 21(4):485–488, 2003. ISSN 0735-0015. doi: 10.1198/073500103288619133.

Mikhail Chernov, A. Ronald Gallant, Eric Ghysels, and George Tauchen. Alternative models for stock price dynamics. *Journal of Econometrics*, 116(1-2):225–257, 9 2003. ISSN 03044076. doi: 10.1016/S0304-4076(03)00108-8.

Victor Chernozhukov and Han Hong. Likelihood Estimation and Inference in a Class of Nonregular Econometric Models. *Econometrica*, 72(5):1445–1480, 2004. doi: 10.1111/j.1468-0262.2004.00540.x.

Miklós Csörgő. *Quantile Processes with Statistical Applications*. Society for Industrial and Applied Mathematics, 1 1983. ISBN 978-0-89871-185-1. doi: 10.1137/1.9781611970289.

Miklos Csorgo and Pal Revesz. Strong Approximations of the Quantile Process. *The Annals of Statistics*, 6 (4), 7 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344261.

Miklós Csörgö and Hao Yu. Weak approximations for quantile processes of stationary sequences. *Canadian Journal of Statistics*, 24(4), 12 1996. ISSN 03195724. doi: 10.2307/3315325.

Stephen G Donald and Harry J Paarsch. Superconsistent estimation and inference in structural econometric models using extreme order statistics. *Journal of Econometrics*, 109(2):305–340, 2002. doi: 10.1016/s0304-4076(02)00116-1.

P Doukhan, P Massart, and E Rio. Invariance principles for absolutely regular empirical processes. *Annales de l'I.H.P. Probabilités et statistiques*, 31(2), 1995. ISSN 0246-0203.

Bjørn Eraker, Michael Johannes, and Nicholas Polson. The Impact of Jumps in Volatility and Returns. *The Journal of Finance*, 58(3):1269–1300, 6 2003. ISSN 00221082. doi: 10.1111/1540-6261.00566.

Amit Gandhi and Jean-François Houde. Measuring Substitution Patterns in Differentiated-Products Industries. Technical report, National Bureau of Economic Research, Cambridge, MA, 10 2019. URL http://www.nber.org/papers/w26375.pdf.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z Ghahramani, M Welling, C Cortes, N Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

T.P. Hettmansperger, I. Hueter, and J. Hüsler. Minimum distance estimators. *Journal of Statistical Planning and Inference*, 41(3), 10 1994. ISSN 03783758. doi: 10.1016/0378-3758(94)90025-6.

Keisuke Hirano and Jack R Porter. Asymptotic Efficiency in Parametric Structural Models with Parameter-Dependent Support. *Econometrica*, 71(5):1307–1338, 2003. doi: 10.1111/1468-0262.00451.

Narasimhan Jegadeesh and George G. Pennacchi. The Behavior of Interest Rates Implied by the Term Structure of Eurodollar Futures. *Journal of Money, Credit and Banking*, 28(3):426, 8 1996. ISSN 00222879. doi: 10.2307/2077890.

Renna Jiang, Puneet Manchanda, and Peter E Rossi. Bayesian analysis of random coefficient logit models using aggregate data. *Journal of Econometrics*, 149(2):136–148, 2009. ISSN 0304-4076. doi: 10.1016/j.jeconom.2008.12.010.

Tetsuya Kaji. Asymptotic Theory of L-Statistics and Integrable Empirical Processes. 10 2019.

Tetsuya Kaji, Elena Manresa, and Guillaume Pouliot. An Adversarial Approach to Structural Estimation. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3706365.

Ivana Komunjer and Serena Ng. Dynamic Identification of Dynamic Stochastic General Equilibrium Models. *Econometrica*, 79(6), 2011. ISSN 0012-9682. doi: 10.3982/ECTA8916.

Vincent N. LaRiccia. Asymptotic Properties of Weighted L^2 Quantile Distance Estimators. *The Annals of Statistics*, 10(2), 6 1982. ISSN 0090-5364. doi: 10.1214/aos/1176345803.

Vincent N. LaRiccia. Optimal weights for general L2 distance estimators. *Statistics & Probability Letters*, 2 (3):169–173, 5 1984. ISSN 01677152. doi: 10.1016/0167-7152(84)90009-9.

Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, and Xianfeng David Gu. A geometric view of optimal transportation and generative model. *Computer Aided Geometric Design*, 68, 1 2019. ISSN 01678396. doi: 10.1016/j.cagd.2018.10.005.

Tong Li. Indirect inference in structural econometric models. *Journal of Econometrics*, 157(1):120–128, 7 2010.

K L Mehra and M Sudhakara Rao. Weak Convergence of Generalized Empirical Processes Relative to d_q Under Strong Mixing. *The Annals of Probability*, 3(6), 10 1975a. ISSN 0091-1798. doi: 10.1214/aop/1176996223.

K L Mehra and M Sudhakara Rao. On Functions of Order Statistics for Mixing Processes. *The Annals of Statistics*, 3(4), 10 1975b. ISSN 0090-5364. doi: 10.1214/aos/1176343188.

Hyungsik Roger Moon, Matthew Shum, and Martin Weidner. Estimation of random coefficients logit demand models with interactive fixed effects. *Journal of Econometrics*, 206(2):613–644, 2018. ISSN 1872-6895. doi: 10.1016/j.jeconom.2018.06.016.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Şimşekli. Statistical and Topological Properties of Sliced Probability Divergences. 3 2020a.

Kimia Nadjahi, Alain Durmus, Umut Şimşekli, and Roland Badeau. Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance. 6 2020b.

Whitney K. Newey. Uniform Convergence in Probability and Stochastic Equicontinuity. *Econometrica*, 59 (4), 7 1991. ISSN 00129682. doi: 10.2307/2938179.

Ömer Öztürk and Thomas P. Hettmansperger. Generalised weighted Cramér-von Mises distance estimators. *Biometrika*, 84(2), 1997. ISSN 00063444. doi: 10.1093/biomet/84.2.283.

Harry J Paarsch. Deciding between the common and private value paradigms in empirical models of auctions. *Journal of Econometrics*, 51(1-2):191–215, 10 1992. doi: 10.1016/0304-4076(92)90035-p.

Jun Pan. The jump-risk premia implicit in options: evidence from an integrated time-series study. *Journal of Financial Economics*, 63(1):3–50, 1 2002. ISSN 0304405X. doi: 10.1016/S0304-405X(01)00088-5.

Sergio Pastorello, Valentin Patilea, and Eric Renault. Iterative and Recursive Estimation in Structural Nonadaptive Models. *Journal of Business & Economic Statistics*, 21(4):449–509, 10 2003. ISSN 0735-0015. doi: 10.1198/073500103288619124.

Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer New York, New York, NY, 1999. ISBN 978-1-4612-7190-1. doi: 10.1007/978-1-4612-1554-7.

D Pollard. The minimum distance method of testing. *Metrika*, 27(1):43–70, 1980. ISSN 0026-1335. doi: 10.1007/bf01893576.

Benedikt M. Pötscher and Ingmar R. Prucha. *Dynamic Nonlinear Econometric Models*. 1997. doi: 10.1007/978-3-662-03486-6.

Madan L. Puri and Lanh T. Tran. Empirical distribution functions and functions of order statistics for mixing random variables. *Journal of Multivariate Analysis*, 10(3):405–425, 9 1980. ISSN 0047259X. doi: 10.1016/0047-259X(80)90061-5.

Shanshan Qin and Yuehua Wu. General matching quantiles M-estimation. *Computational Statistics & Data Analysis*, 147:106941, 7 2020. ISSN 01679473. doi: 10.1016/j.csda.2020.106941.

Emmanuel Rio. *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. ISBN 978-3-662-54322-1. doi: 10.1007/978-3-662-54323-8.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-20827-5. doi: 10.1007/978-3-319-20828-2.

Bodhisattva Sen. A Gentle Introduction to Empirical Process Theory and Applications. 2021. URL http://www.stat.columbia.edu/~bodhi/Talks/Emp-Proc-Lecture-Notes.pdf.

Nikolaos Sgouropoulos, Qiwei Yao, and Claudia Yastremiz. Matching a Distribution by Matching Quantiles Estimation. *Journal of the American Statistical Association*, 110(510):742–759, 4 2015. ISSN 0162-1459. doi: 10.1080/01621459.2014.929522.

Alexander Shapiro. Statistical Inference of Stochastic Optimization Problems. In *Nonconvex Optimization and Its Applications*, pages 282–307. Springer US, 2000. doi: 10.1007/978-1-4757-3150-7{_}16.

Robert P. Sherman. Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators. *The Annals of Statistics*, 22(1), 3 1994. ISSN 0090-5364. doi: 10.1214/aos/1176325377.

Galen R Shorack and Jon A Wellner. *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, 10 2009. ISBN 0898716845. doi: 10.1137/1.9780898719017.

Alexander Torgovitsky. Minimum distance from independence estimation of nonseparable instrumental variables models. *Journal of Econometrics*, 199(1), 2017. ISSN 18726895. doi: 10.1016/j.jeconom.2017.01.009.

Sara A. van de Geer. *Empirical Processes in M-Estimation*, volume 6. Cambridge University Press, 2009. ISBN 9780521123259.

Aad W. van de Vaart. *Asymptotic Statistics*. Cambridge University Press, 10 1998. ISBN 9780521496032. doi: 10.1017/CBO9780511802256.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, New York, NY, 1996. ISBN 978-1-4757-2547-6. doi: 10.1007/978-1-4757-2545-2.

Li Xing Zhu, Kai Tai Fang, and M. Ishaq Bhatti. On estimated projection pursuit-type Crámer-Von Mises statistics. *Journal of Multivariate Analysis*, 63(1), 1997. ISSN 0047259X. doi: 10.1006/jmva.1997.1673.

A 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 $\|\cdot\|_{\mathbb{D}}$ and $\|\cdot\|_{\mathbb{K}}$ respectively. For a functional $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{K}$, we define Hadamard directional differentiability as follows.

Definition A.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 A.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 A.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 A.1 (Theorem 2.1 of Chen and Fang [2019])**.** *If Conditions A.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 A.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 A.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

$$\begin{aligned}
\alpha(\mathcal{A}, \mathcal{B}) &= \sup_{A \in \mathcal{A}, B \in\ \mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, \\
\varrho(\mathcal{A}, \mathcal{B}) &= \sup_{f \in \mathcal{L}^2(\mathcal{A}), g \in \mathcal{L}^2(\mathcal{B})} |corr(f, g)|, \\
\varphi(\mathcal{A}, \mathcal{B}) &= \sup_{A \in \mathcal{A}, B \in \mathcal{B}, \mathbb{P}(A) > 0} |\mathbb{P}(B|A) - \mathbb{P}(B)|, \\
\beta(\mathcal{A}, \mathcal{B}) &= \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|
\end{aligned}$$

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 A.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,*

$$\begin{aligned}
|cov(X, Y)| &\leq 2(2^{1/p} + 1)\alpha(\mathcal{D}, \mathcal{K})^{1/p - 1/s}(E[||X||^s])^{1/s}(E[||Y||^q])^{1/q} \\
|cov(X, Y)| &\leq 2\phi(\mathcal{D}, \mathcal{K})^{1-1/s}(E[||X||^s])^{1/s}(E[||Y||^q])^{1/q}.
\end{aligned}$$

B Proofs in Section 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.1 Proof of Theorem 5.1 (consistency)

We can show the consistency by following standard approaches in the extremum estimator. Since $\hat{\psi}_T$ satisfies $(\mathcal{M}_T(\hat{\psi}_T))^2 \le \inf_{\psi\in\Psi}(\mathcal{M}_T(\psi))^2 + T^{-1}\epsilon_T$, we have

$$\mathcal{M}_T(\hat{\psi}_T) \le \inf_{\psi\in\Psi}\mathcal{M}_T(\psi) + T^{-1/2}\epsilon_T^{1/2} = \inf_{\psi\in\Psi}\mathcal{M}_T(\psi) + o_p(1) \le \mathcal{M}(\psi_0) + o_p(1).$$

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 5.1 as follows:

$$\begin{aligned}
&\sup_{\psi\in\Psi}|\mathcal{M}_T(\psi)-\mathcal{M}(\psi)|\\
&= \sup_{\psi\in\Psi}\left|\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q_T(s;u)-\widehat{Q}_T(s;u,\psi))^2 w(s)\mathrm{d}s d\varsigma(u)\right)^{1/2} - \left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q(s;u)-Q(s;u,\psi))^2 w(s)\mathrm{d}s d\varsigma(u)\right)^{1/2}\right|\\
&\le \sup_{\psi\in\Psi}\left(\int_{\mathbb{S}^{d-1}}\int_{S}(Q_T(s;u)-Q(s;u)-\widehat{Q}_T(s;u,\psi)+Q(s;u,\psi))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}\\
&\le \widehat{SW}_T + \sup_{\psi\in\Psi}\widehat{SW}_T(\psi) \xrightarrow{p} 0.
\end{aligned}$$

Assumption 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 Proof of Theorem 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 5.1 and 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

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

for any $\tau_T \to 0$.

The objective function can be decomposed as follows:

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u)-\widehat{Q}_T(s;u,\psi))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad = \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u)-\widehat{Q}_T(s;u,\psi_0))^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) - 2(\psi-\psi_0)^\top A_T/\sqrt{T} + (\psi-\psi_0)^\top B_T(\psi-\psi_0) + R_T,
\end{aligned}$$

where

$$\begin{aligned}
A_T &= \sqrt{T}\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u)-\widehat{Q}_T(s;u,\psi_0))\widehat{D}_T(s;u,\psi_0) w(s)\mathrm{d}s\mathrm{d}\varsigma(u);\\
B_T &= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \widehat{D}_T(s;u,\psi_0)\widehat{D}_T^\top(s;u,\psi_0) w(s)\mathrm{d}s\mathrm{d}\varsigma(u);\\
R_T &= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \widehat{R}_T^2(s;u,\psi,\psi_0) w(s)\mathrm{d}s\mathrm{d}\varsigma(u) - 2\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} (Q_T(s;u)-\widehat{Q}_T(s;u,\psi_0))\widehat{R}_T(s;u,\psi,\psi_0) w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\\
&\quad + 2(\psi-\psi_0)^\top \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}} \widehat{D}_T(s;u,\psi_0)\widehat{R}_T(s;u,\psi,\psi_0) w(s)\mathrm{d}s\mathrm{d}\varsigma(u).
\end{aligned}$$

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 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 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 5.4 (iii) holds.

Step 3: We will analyze the behavior of R_T.

Here, we will show

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T}\frac{T|R_T|}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} = o_p(1).$$

Let us remind that

$$\begin{aligned}
R_T &= \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) - 2\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}(Q_T(s;u) - \widehat{Q}_T(s;u,\psi_0))\widehat{R}_T(s;u,\psi,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \\
&\quad + 2(\psi-\psi_0)^\top\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\widehat{D}_T(s;u,\psi_0)\widehat{R}_T(s;u,\psi,\psi_0)w(s)\mathrm{d}s\mathrm{d}\varsigma(u),
\end{aligned}$$

where

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

for any $\tau_T \to 0$.

Note that

$$|R_T| \leq \int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) + 2\mathcal{M}_T(\psi_0)\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}$$
$$+ 2\|\psi-\psi_0\|_\Psi\left(\overline{B}_T\right)^{1/2}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}.$$

This implies that

$$\begin{aligned}
&\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T}\frac{T|R_T|}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\\
&\quad\leq \sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T}\frac{T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\\
&\qquad+2\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T}\frac{T\mathcal{M}_T(\psi_0)\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\\
&\qquad+2\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T}\frac{T\|\psi-\psi_0\|\left(\overline{B}_T\right)^{1/2}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{n}(\psi-\psi_0)\|)^2}.
\end{aligned}$$

First, we have

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

by Assumption 5.3.

Second, we have

$$\begin{aligned}
&\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T}\frac{T\mathcal{M}_T(\psi_0)\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\\
&\quad\leq\sqrt{T}\mathcal{M}_T(\psi_0)\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T}\frac{\sqrt{T}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\\
&\quad\leq\sqrt{T}\mathcal{M}_T(\psi_0)\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T}\frac{\sqrt{T}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{1+\|\sqrt{T}(\psi-\psi_0)\|}\\
&\quad=\sqrt{T}\mathcal{M}_T(\psi_0)\left(\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T}\frac{T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\right)^{1/2}\\
&\quad=O_p(1)o_p(1)=o_p(1).
\end{aligned}$$

Here, the second last equality holds because Assumptions 5.4 implies

$$\sqrt{T}\mathcal{M}_T(\psi_0) \leq \sqrt{T}\widehat{\mathcal{SW}}_T + \sqrt{T}\widehat{\mathcal{SW}}_T(\psi_0) = O_p(1),$$

and Assumption 5.3 implies

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

Third, we have

$$\begin{aligned}
&\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \frac{T\|\psi-\psi_0\|\left(\overline{B}_T\right)^{1/2}\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\\
&= \left(\overline{B}_T\right)^{1/2}\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \frac{T\|\psi-\psi_0\|\left(\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_T(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2}\\
&= \left(\overline{B}_T\right)^{1/2}\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \frac{\sqrt{T}\|\psi-\psi_0\|\left(T\int_{\mathbb{S}^{d-1}}\int_{\mathcal{S}}\left(\widehat{R}_n(s;u,\psi,\psi_0)\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right)^{1/2}}{(1+\|\sqrt{T}(\psi-\psi_0)\|)(1+\|\sqrt{n}(\psi-\psi_0)\|)}\\
&= O_p(1)o_p(1) = o_p(1).
\end{aligned}$$

Therefore, we have

$$\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\leq\tau_T} \frac{T|R_T|}{(1+\|\sqrt{T}(\psi-\psi_0)\|)^2} = o_p(1).$$

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.

C Proofs in Section 6

C.1 Preliminary Lemmas

Lemma C.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 C.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 C.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

$$\sup_s |F(G^{-1}(s)) - s| < \sup_s |G(F^{-1}(s)) - s| \tag{C.1}$$

or

$$\sup_s |F(G^{-1}(s)) - s| > \sup_s |G(F^{-1}(s)) - s|. \tag{C.2}$$

Here, we will only prove that inequality (C.1) will result in the contradiction because we can show that the inequality (C.2) does not hold in the same way.

Suppose inequality (C.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, $s_2 \ne s_1$ because $\sup_s |F(G^{-1}(s)) - s| \ge 0$. Then, we will consider two possible cases. $s_2 < s_1$ or $s_1 < s_2$.

Case 1: We will show inequality (C.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 C.2 (2). Then, Lemma C.2 (1) implies

$$G^{-1}(s_2 + \epsilon) > F^{-1}(s_1) \Rightarrow F(G^{-1}(s_2 + \epsilon)) \ge s_1.$$

Therefore, we have

$$F(G^{-1}(s_2 + \epsilon)) - (s_2 + \epsilon) \ge s_1 - s_2 - \epsilon.$$

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 $\tilde{s}$ such that

$$\sup_s |F(G^{-1}(s)) - (s)| < |F(G^{-1}(\tilde{s})) - (\tilde{s})|.$$

This is a contradiction.

Case 2: We will show inequality (C.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 C.2 (7), we have

$$G^{-1}(s_2) = G^{-1}(G(F^{-1}(s_1))) < F^{-1}(s_1).$$

Using Lemma C.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 C.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 C.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 C.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 C.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. □

C.2 Proof of Lemma 6.1

Example 7.21 of Sen [2021] implies that

$$A = \{\{Z_t : u^\top Z_t \le s\}, u \in \mathbb{S}^{d-1}, s \in \mathbb{R}\}$$

is a VC class with VC dimension is equal to $d + 1$.

Then, we can show weak convergence of empirical process using Lemma C.1 because the indicator function is uniformly bounded.

C.3 Proof of Lemma 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

$$\begin{aligned}
&T \int_{s\in\mathbb{S}^{d-1}} \int_0^1 |G_T^{-1}(s; u, \psi_0) - G^{-1}(s; u, \psi_0)|^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u) \\
&\quad = T \int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0} \int_0^1 |G_T^{-1}(s; u, \psi_0) - G^{-1}(s; u, \psi_0)|^2 w(s) \mathrm{d}s \mathrm{d}\varsigma(u).
\end{aligned}$$

Following Proposition A.18 of Bobkov and Ledoux [2019],

$$\begin{aligned}
G_T^{-1}(s; u, \psi_0) - G^{-1}(s; u, \psi_0) &= G^{-1}(G(G_T^{-1}(s; u, \psi_0); u, \psi_0); u, \psi_0) - G(s; u, \psi_0) \\
&= \int_s^{G(G_T^{-1}(s;u,\psi_0);u,\psi_0)} \frac{1}{g(G^{-1}(z; u, \psi_0); u, \psi_0)} \mathrm{d}z \\
&\le \frac{1}{g(G^{-1}(\tilde{s}(s; u, \psi_0); u, \psi_0); u, \psi_0)} \Big(G(G_T^{-1}(s; u, \psi_0), u, \psi_0) - s \Big),
\end{aligned}$$

where $\tilde{s}(s;u,\psi_0) \in \Big(G(G_T^{-1}(s;u,\psi_0),u,\psi_0) \wedge s, G(G_T^{-1}(s;u,\psi_0),u,\psi_0) \vee s\Big)$.

This implies

$$
\begin{aligned}
&T\int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 |G_T^{-1}(s;u,\psi_0) - G^{-1}(s;u,\psi_0)|^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u) \\
&\leq T\int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 \frac{w(s)}{g^2(G^{-1}(\tilde{s};u,\psi);u,\psi_0)}|G(G_T^{-1}(s;u,\psi_0);u,\psi_0) - s|^2\mathrm{d}s\mathrm{d}\varsigma(u) \\
&\leq \left(\int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 \frac{w(s)}{g^2(G^{-1}(\tilde{s};u,\psi_0);u,\psi_0)}\mathrm{d}s\mathrm{d}\varsigma(u)\right)\{\sqrt{T}\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in(0,1)}|G(G_T^{-1}(s;u,\psi_0);u,\psi_0) - s|\}^2
\end{aligned}
$$

where $\tilde{s}(s;u,\psi_0) \in \Big(G(G_T^{-1}(s;u,\psi_0),u,\psi_0) \wedge s, G(G_T^{-1}(s;u,\psi_0),u,\psi_0) \vee s\Big)$.

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 $\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in(0,1)}|\tilde{s}(s;u,\psi_0) - s| = o_p(1)$, and we have

$$
\begin{aligned}
&\left(\int_{s\in\mathbb{S}^{d-1}/\mathcal{N}_0}\int_0^1 \frac{w(s)}{g^2(G^{-1}(\tilde{s};u,\psi_0);u,\psi_0)}\mathrm{d}s\mathrm{d}\varsigma(u)\right) \\
&\quad\leq \left(\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0}\sup_{s\in supp\{w\}}\frac{g^2(G^{-1}(s;u,\psi_0);u,\psi_0)}{g^2(G^{-1}(\tilde{s};u,\psi_0);u,\psi_0)}\right)\int_{s\in\mathbb{S}^{d-1}}\int_0^1 \frac{w(s)}{g^2(G^{-1}(s;u,\psi_0);u,\psi_0)}\mathrm{d}s\mathrm{d}\varsigma(u) \\
&\quad= O_p(1)
\end{aligned}
$$

under Conditions (6.1) and (6.2).

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

$$
\sup_s|G(G_T^{-1}(s;u,\psi_0),u,\psi_0) - s| = \sup_s\left|\frac{1}{T}\sum_{t=1}^T I(G(u^\top Z_t;u,\psi_0)\leq s) - s\right|.
$$

(c.f Equation (1.4.5) in Csörgő [1983]). Note that for each $u\in\mathbb{S}^{d-1}/\mathcal{N}_0$, we have

$$
\begin{aligned}
\sup_{s\in(0,1)}\left|\frac{1}{T}\sum_{t=1}^T I(G(u^\top Z_t;u,\psi_0)\leq s) - s\right| &= \sup_{s\in(0,1)}\left|\frac{1}{T}\sum_{t=1}^T I(u^\top Z_t\leq G^{-1}(s;u,\psi_0)) - s\right| \\
&= \sup_{s\in\mathbb{R}}|G_T(s;u,\psi_0) - G(s;u,\psi_0)|.
\end{aligned}
$$

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 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.

C.4 Proof of Lemma 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

$$\sqrt{w(s)} \frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} = \frac{1}{t_n} \frac{w(s)}{g(G^{-1}(\tilde{s}(s;u);u);u)} (G(G_n^{-1}(s;u)) - s)$$

Then,

$$\sup_s \left| \sqrt{w(s)} \frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} \right| \leq \left(\sup_s \frac{w(s)}{g(G^{-1}(\tilde{s}(s;u);u);u)} \right) \frac{1}{t_n} \sup_s \left| (G(G_n^{-1}(s;u);u) - s) \right|$$

where $\tilde{s}$ is between $G(G_n^{-1}(s;u);u)$ and s.

From Lemma C.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

$$\sup_s \left| \sqrt{w(s)} \frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} \right| \leq \left(\sup_s \frac{\sqrt{w(s)}}{g(G^{-1}(\tilde{s}(s;u);u);u)} \right) \sup_s \left| h_n(G^{-1}(s;u);u) \right|$$

It implies

$$\begin{aligned}
&\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_s \left| \sqrt{w(s)} \frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} \right| \\
&\quad \leq \left(\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_s \frac{\sqrt{w(s)}}{g(G^{-1}(\tilde{s}(s;u);u);u)} \right) \sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_s \left| h_n(G^{-1}(s;u);u) \right|.
\end{aligned}$$

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

$$\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s\in supp\{w\}} \left| \frac{g(G^{-1}(s;u,\psi_0);u,\psi_0)}{g(G^{-1}(\tilde{s}(s;u);u,\psi_0);u,\psi_0)} \right| = O(1) \tag{C.3}$$

for $\tilde{s}(s,u)$ such that $\sup_{u\in\mathbb{S}^{d-1}/\mathcal{N}_0} \sup_{s\in supp\{w\}} |\tilde{s}(s;u) - s| = o(1)$, we can find N and C such that

$$\sup_{s,u} \left| \sqrt{w(s)} \frac{G_n^{-1}(s;u) - G^{-1}(s;u)}{t_n} \right| \leq \left(\sup_{s,u} \frac{\sqrt{w(s)}}{g(G^{-1}(\tilde{s}(s;u);u);u)} \right) \sup_{s,u} \left| h_n(G^{-1}(s;u);u) \right| < C$$

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 the 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 Theorem 3.9.23 in van der Vaart and Wellner [1996].

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.

D Proofs in Section 7

D.1 Preliminary Lemma

Lemma D.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. □

D.2 Proof of Lemma 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 5.1 (ii). Since $k(X_t,X_j;\psi)$ is symmetric with respect to X_t and X_j, we have

$$\frac{1}{T^2}\sum_{t=1}^T\sum_{j=1}^T k(X_t,X_j;\psi)=\frac{1}{T^2}\sum_{t=1}^T k(X_t,X_t;\psi)+\frac{2}{T^2}\sum_{1\leq t<j\leq T}k(X_t,X_j,\psi)$$

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

$$\sup_{\psi\in\Psi}\left|\frac{2}{T^2}\sum_{1\leq t<j\leq T}k(X_t,X_j,\psi)\right|=o_p(1). \tag{D.1}$$

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 (D.1) under the Lipschitz continuity of $k(X_t,X_j,\psi)$ with respect to ψ.

Lemma D.2. *Suppose Ψ is compact and $k(x,x',\psi)$ is Lipshitz in ψ in the sense that for any $\psi,\tilde\psi\in\Psi$, we have*

$$|k(x,x',\psi)-k(x,x',\psi_0)|\leq M(x,x')\|\psi-\tilde\psi\|$$

and $\int\int M(x,\tilde x)dP_x dP_{\tilde x}<\infty$ where $P_x=P_{\tilde x}$. Then,

$$\sup_\psi\left\|\frac{2}{T^2}\sum_{1\leq t<j\leq n}k(X_t,X_j,\psi)\right\|=\sup_\psi\left\|\frac{2}{T^2}\sum_{1\leq t<j\leq T}k(X_t,X_j,\psi)-\frac{2}{T^2}\sum_{1\leq t<j\leq n}\mathbb{E}[k(X_t,X_j,\psi)]\right\|\xrightarrow{p}0.$$

Proof. See Corollary 4.1 of Newey [1991] or Lemma 4 in Appendix of Briol et al. [2019]. □

Under the Lipschitz continuity of $F(y|x,\psi)$ in ψ, $k(x,x',\psi)$ is also Lipchitz continuous in ψ by Lemma D.1. This concludes the proof.

D.3 Proof of Lemma 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 7.1.

Note that

$$\begin{aligned}
&\widehat{Q}_T(s;u,\psi) - \widehat{Q}_T(s;u,\psi_0) - (\psi-\psi_0)^\top \widehat{D}_T(s;u,\psi_0)\\
&= \left[\widehat{Q}_T(s;u,\psi) - Q(s;u,\psi)\right] - \left[\widehat{Q}_T(s;u,\psi_0) - Q(s;u,\psi_0)\right] + \left[Q(s;u,\psi) - Q(s;u,\psi_0) - (\psi-\psi_0)^\top \widehat{D}_T(s;u,\psi_0)\right].
\end{aligned}$$

Under Condition 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

$$V_{2,T}(\psi) = \left(\int_{u\in\mathbb{S}^{d-1}}\int\left((\widehat{Q}_T(s;u,\psi)-Q(s;u,\psi))-(\widehat{Q}_T(s;u,\psi_0)-Q(s;u,\psi_0))\right)^2 w(s)\mathrm{d}s\mathrm{d}\varsigma(u)\right).$$

$V_{2,T}(\psi)$ 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

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

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 D.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

$$
\mathbb{E}[k_2(X_i, X_i, \psi, \psi_0)] \le \mathbb{E}[M_2(X_i, X_i)]\|\psi - \psi_0\|
$$

and it implies that $\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_n} \mathbb{E}[k_2(X_i, X_i, \psi, \psi_0)] = o_p(1)$.

Step 3) For $t \neq j$, $\mathbb{E}[k_2(X_t, X_j, \psi, \psi_0)] = 0$ when $w(s)$ is integrable. Then,

$$
U_{2,T} := \frac{2}{T^2}\sum_{1\le t<j\le T} k_2(X_t, X_j, \psi, \psi_0)
$$

is a degenerate U-statistic.

Suppose Ψ is compact, and for any $\psi, \tilde{\psi} \in \Psi$, we have

$$
|k_2(X_t, X_j, \psi, \psi_0) - k_2(X_i, X_j, \tilde{\psi}, \psi_0)| \le M_2(X_i, X_j)\|\psi - \tilde{\psi}\|
$$

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 $F \le M(X_i, X_j)diam(\Psi)$ where $diam(\Psi) = \sup_{\theta,\tilde{\theta}\in\Psi}\|\theta - \tilde{\theta}\|$ is the diameter of Ψ Then,

$$
\int\int |k_2^2(X_i, X_j)|dF_{X_i}dF_{X_j} \le \left(\int\int M_2^2(X_i, X_j)dF_{X_i}dF_{X_j}\right)\|\psi - \psi_0\|
$$

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 $\sup_{\psi\in\Psi;\ \|\psi-\psi_0\|\le\tau_T} |TU_{2,T}| = o_p(1)$.

Therefore,

$$
\sup_{\psi\in\|\psi-\psi_0\|_2\le\tau_T} |TV_{2,T}(\psi)| = \sup_{\psi\in\|\psi-\psi_0\|_2\le\tau_T} |TV_{2,T}(\psi) - TV_{2,T}(\psi_0)| = o_p(1).
$$

E Verification of Assumptions in Section 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 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 5.3, 5.2, 5.4 (i), and 5.5, and 5.6 because all three examples are conditional models, and Assumption 5.1 (i) is implied by Assumption 5.4(i).

E.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 5.2

Assumption 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 5.3 and Assumption 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 5.6 holds as well.

Verification of Assumption 5.4 (i) and 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 $\{I(u^\top Z_t \leq s) : u \in \mathbb{S}^{d-1}, s \in \mathbb{R}\}$ 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 5.4 (i) and and 5.5 hold by Lemmas 6.2 and 6.5.

E.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 $w(s) = 1$.

E.2.1 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 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 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

$$T\int_{-\infty}^{\infty} ((F_T(s) - F(s,\theta_0)))^2 \mathrm{d}s = T\int_0^{\theta_0} ((F_T(s) - F(s,\theta_0)))^2 \mathrm{d}s \le \theta_0 \left[\sup_{s\in\mathbb{R}} (\sqrt{T}(F_T(s) - F(s,\theta_0)))\right]^2 = O_p(1).$$

Therefore, Assumption 5.4 (i) holds.

Example 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 5.3 and 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 5.5 is satisfied by Lemma 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 5.2 holds.

Assumptions 5.3 and 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 5.4 (i) and 5.5 hold by Lemmas 6.2 and 6.5.

E.2.2 Two-sided Uniform Model

In two-sided models, the CDF of Y is

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

and the quantile function is

$$F^{-1}(s;\theta_0) = \begin{cases} 4\theta_0 s & \text{if } 0 \le s \le 1/4, \\ 1 - \frac{4}{3}(1-\theta_0)(1-s) & \text{if } 1/4 \le s \le 1. \end{cases}$$

The MSCD Estimator In this case, $Q(s,\theta_0) = F(s,\theta_0)$. Let $F_T(s) = I(Z_t \le s)$. 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

$$T\int_{-\infty}^{\infty} ((F_T(s) - F(s,\theta_0)))^2 \mathrm{d}s = T\int_0^1 ((F_T(s) - F(s,\theta_0)))^2 \mathrm{d}s \le \left[\sup_{s\in\mathbb{R}}(\sqrt{T}(F_T(s) - F(s,\theta_0)))\right]^2 = O_p(1).$$

Therefore, Assumption 5.4 (i) holds.

Example 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 5.3 are satisfied.

Therefore, Assumption 5.5 by Lemma 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 5.2 holds.

Since $F^{-1}(s;\theta)$ is linear in θ, $Q(s,u,\theta)$ is norm-differentiable at $\theta = \theta_0$ with

$$D(s,\theta_0) = \begin{cases} 4s & \text{if } 0 \le s \le 1/4, \\ \frac{4}{3}(1-s) & \text{if } 1/4 \le s \le 1. \end{cases}$$

Therefore, Assumptions 5.3 and 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 5.4 (i) holds by following the proof of Lemma 6.2 and 6.5. Since $D(s,\theta_0)$ and $F^{-1}(s,\theta_0)$ are bounded by some absolute constants, we can show Assumption 5.5 by Lemma F.3.

F One-Dimensional Unconditional Models

This appendix section consists of two parts. In Section F.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 $D(s;u,\theta)$ in this appendix section.

F.1 Verification of Assumptions

We will present relevant conditions to verify Assumptions 5.4 (i) and 5.5 with $V_0 = \begin{pmatrix} \Omega_0 & 0 \\ 0 & 0 \end{pmatrix}$.

F.1.1 Verification of Assumption 5.4 (i)

When $d = 1$, there are two sets of results that we can use to verify Assumption 5.4 (i).

When $\{Z_t\}_{t=1}^T$ is a random sample, we can use results in Barrio et al. [2005].

Lemma F.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

$$\sup_{0<x<1} \frac{x(1-x)|f'(F^{-1}(x))|}{f^2(F^{-t}(x))} < \infty,$$

$$\text{either } a_F > -\infty \text{ or } \liminf_{x\to 0+} \frac{|f'(F^{-1}(x))|x}{f^2(F^{-1}(x))} > 0,$$

$$\text{either } b_F < \infty \text{ or } \liminf_{x\to 0+} \frac{|f'(F^{-1}(1-x))|x}{f^2(F^{-1}(1-x))} > 0.$$

Also, we assume that $w(s)$ is a bounded nonnegative measurable function on (0, 1) such that

$$\lim_{x\to 0+} \frac{x\int_0^x w(t)dt}{f^2(F^{-1}(x))} = 0, \ \lim_{x\to 0+} \frac{x\int_0^x w(t)dt}{f^2(F^{-1}(1-x))} = 0, \ \text{and} \ \int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))} w(t)dt < \infty.$$

Then, we have

$$\sqrt{n}(F_n^{-1}(s) - F^{-1}(s))\sqrt{w(s)} \to \frac{B(s)\sqrt{w(s)}}{\sqrt{f^2(F^{-1}(s))}} \text{ in law in } L_2(0,1)$$

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 F.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 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

$$\sigma^2(s,j) = (s \wedge t) - st + \sum_{j=2}^{\infty} [Cov(V_1 \leq s, V_j \leq t) + Cov(V_1 \leq t, V_j \leq s)] \tag{F.1}$$

where $V_j = F^{-1}(Z_j)$.

Then, we have

$$T \int_0^1 (F_T^{-1}(s) - F^{-1}(s))^2 w(s) \mathrm{d}s \xrightarrow{d} \int_0^1 \frac{B^2(s)}{f^2(F^{-1}(s))} w(s) \mathrm{d}s.$$

□

F.1.2 Verification of Assumption 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 [1975b,a], 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 F.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 F.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 F.3. *Suppose* $\{Z_t\}_{t=1}^T$ *is i.i.d. sample, and Conditions F.1 and F.2 hold. Then, Assumption 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 F.3. This is the consequence of Theorem 1 and Remark 2 in Chapter 19 of Shorack and Wellner [2009]. □

Condition F.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 F.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 F.3.*

Lemma F.4. *Conditions F.2, F.3, and F.4 hold. Then, Assumption 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]. □

F.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 $w(s) = 1$.

F.2.1 The Two-sided Uniform Model

In this subsection, we will derive the asymptotic distribution of two-sided uniform model directly using Lemma F.3. When $w(s) = 1$, the MWSD estimator θ_n has 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

$$D(s, \theta_0) = \begin{cases} 4s & \text{when } 0 \leq s \leq \frac{1}{4}, \\ \frac{4}{3}(1-s) & \text{when } \frac{1}{4} < s \leq 1. \end{cases}$$

In two-sided uniform model, Conditions F.1 and F.2 hold. This is because $Q(\cdot)$ and $D(\cdot, \theta_0)$ are uniformly bounded from below and above by constants, and $D(\cdot, \theta_0)$ is continuous on $(0, 1)$ almost everywhere with respect to Q. Therefore, by Lemma F.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 $0 \leq \theta_0 \leq 1$.

F.2.2 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

$$f_V(v, \theta)I(v \geq g_V(\theta)). \tag{F.2}$$

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

$$\sigma(v) = v + \frac{\int_v^\infty (1 - F_V(\xi))^{m-1} d\xi}{(1 - F_V(v))^{m-1}}, \tag{F.3}$$

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 (F.2) 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 F.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 F.1 and F.2 to exploit Lemma F.3 in Steps 2-1 and 2-2 below.

Note that when Lemma F.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 F.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 F.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 F.1 holds.

Step 2-2. We will verify Condition F.2. In this model, $D(s;\theta_0)$ and $Q(s)$ are continuously differentiable and strictly increasing in s. Therefore, Condition F.2 holds.

From Steps 2-1 and Step 2-2, we can use Lemma F.3. By combining all results in Steps 1 and 2, we have

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = \theta_0^2 \frac{(m-1)^2(5m^2-6m+2)}{(2m^2-2m+1)^2}.$$