Minimum Integrated Distance Estimation in Simultaneous Equation Models[*]

Zhengyuan Gao[†]   Antonio F. Galvao[‡]

December 3, 2014

Abstract

This paper considers estimation and inference in semiparametric econometric models. Standard procedures estimate the model based on an independence restriction that induces a minimum distance between a joint cumulative distribution function and the product of the marginal cumulative distribution functions. This paper develops a new estimator which generalizes estimation by allowing endogeneity of the weighting measure and estimating the optimal measure nonparametrically. The optimality corresponds to the minimum of the integrated distance. To accomplish this aim we use Kantorovich's formulation of the optimal transportation problem. The minimizing distance is equivalent to the total variation distance and thus characterizes finer topological structures of the distributions. The estimation also provides greater generality by dealing with probability measures on compact metric spaces without assuming existence of densities. Asymptotic statistics of the empirical estimates have standard convergent results and are available for different statistical analyses. In addition, we provide a tractable implementation for computing the estimator in practice.

Key Words: Minimum distance, Kantorovich's duality, Kernel representation, Hilbert space

JEL Classification: C12, C13, C14

[†]Corresponding author, CORE, Universitie catholique de Louvain. E-mail: zhengyuan.gao@uclouvain.be

[‡]Department of Economics, Tippie College of Business, University of Iowa. E-mail: antonio-galvao@uiowa.edu

1   INTRODUCTION

Economists often use system of simultaneous equations to describe the relationship among economic variables. In particular, nonlinear simultaneous equation models have provided a valuable method of statistical analysis of policy variables on economic effects. This is especially true for studies where these methods help to analyze the affects of the outcome distributions of interest. Empirically, given data on the dependent and independent variables in the system, one is usually interested in estimating functions, distributions, and primitives describing the system. Identification and estimation of nonlinear structural models is often achieved by assuming that the model's latent variables are independent of the exogenous variables. Examples of such arguments include, among others, Manski (1983), Brown (1983), Roehrig (1988), Brown and Matzkin (1998), Matzkin (2003), Brown and Wegkamp (2002), Benkard and Berry (2006), Brown, Deb, and Wegkamp (2008), and Linton, Sperlich, and van Keilegom (2008).

This paper considers estimation and inference in semiparametric econometric models. We develop a new minimum distance estimator for separable models based on minimizing the distance from the independence condition, where the weighting measure is allowed to be endogenous and is estimated nonparametrically. This is an important innovation for several reasons. First, the new estimator allows general estimation without exogenous restrictions on the weighting measure. Second, the estimator is more efficient than the others available in the literature. Finally, in this general formulation, the proposed method is beneficial to practitioners since the range of models for which the methods are applicable is very broad, for instance, the framework includes an important class of nonlinear simultaneous equation models without requiring knowledge on the reduced form of the model.

We focus on nonlinear simultaneous equations models with an exogenous observed random vector X, a dependent observed random vector Y, and the model errors which are endogenous and related to the parameterized model via $\varepsilon = \rho(X, Y, \theta)$. The underlying ε is assumed to be drawn from a fixed but unknown distribution and to be stochastically independent of X. However, given the estimated parameters $\theta^* \neq \theta_0$, the model may induce $\varepsilon = \rho(X, Y, \theta^*)$ that are not necessary independent of X. The intuition on the importance and development of the new approach is as follows. Under the independence condition, the

criterion function is an integral of a distance function $\mathbf{d}(\cdot, \cdot)$ given by

$$\int \mathbf{d}(H_\theta(x, \varepsilon), P_\theta(x, \varepsilon))d\mu(x, \varepsilon), \tag{1}$$

where θ is the parameter of interest, (H_θ, P_θ) are distribution functions of (X, ε), μ is a certain probability measure. Let $\varepsilon = \rho(X, Y, \theta)$ and (X, Y) be pairs of observations. Thus, the measure μ in (1) can be expressed as $\mu(X, \rho(X, Y, \theta))$ which, in general, depends on a function of θ. We argue it is important to endogenize the measure μ because changes in θ induce changes in the other inputs. For instance, a change in θ directly induces a change in $\mu(X, \rho(X, Y, \theta))$ as well as in $P_\theta(\varepsilon)$ and $H_\theta(\varepsilon)$. A second reason to endogenize μ is practical. The weighting measure μ in the criterion function characterizes the universal differences between the distributions H_θ and P_θ via an integral. In principle, this measure could be independent of the choice of θ. However, in practice these weighted measures are unobservable as the underlying innovation $\varepsilon = \rho(X, Y, \theta_0)$ is not available. Moreover, the empirical measure on the product space of ε and x is sensitive to the choice of θ in the estimation procedure, hence, in applications, it is difficult to specify *a priori* measure to deliver the desired integrated values. Therefore, we view the endogeneity of μ thorough the interaction amongst the weighting measures, $H_\theta(\varepsilon)$ and $P_\theta(\varepsilon)$.

To achieve the goal of endogenizing the weighting measure and developing the estimator, we proceed in four steps. First, we are required to select the optimal measure. Hence, we represent $\mu(\cdot)$ as a transport measure $\gamma(\cdot)$ that pushes H_θ towards P_θ. Consider a set of joint measures $\Gamma(H_\theta, P_\theta)$. Let $\gamma(x, \varepsilon; x', \varepsilon')$ belong to this set and satisfy

$$P_\theta(x, \varepsilon) = \int d\gamma(x, \varepsilon; x', \varepsilon')dH_\theta(x', \varepsilon'), \text{ and, } H_\theta(x', \varepsilon') = \int d\gamma(x, \varepsilon; x', \varepsilon')dP_\theta(x, \varepsilon).$$

The measure $\gamma(x, \varepsilon; x', \varepsilon')$ replaces the role of μ in the problem (1) and accounts for interaction with $P_\theta(\varepsilon)$ and $H_\theta(\varepsilon)$. For two pairs of samples (X, Y), one is drawn from $P_\theta(x, \varepsilon)$ and the other is drawn from $H_\theta(x, \varepsilon)$, joint measure $\gamma(x, \varepsilon; x', \varepsilon')$ is the measure of a transport map connecting these two samples. Second, given the transport measure, we reformulate the main problem using a generalized estimation criterion function, $W(\theta)$, based on an integrated distance w.r.t. the joint probability measure γ as

$$W(\theta) := \inf_{\gamma \in \Gamma(H_\theta, P_\theta)} \left\{ \int \mathbf{d}(q, q')d\gamma(q, q') : q \sim H_\theta(x, \varepsilon) \text{ and } q' \sim P_\theta(x, \varepsilon) \right\}.$$

2

Third, to solve this problem practically, the above weighting measure is dually represented by a transition kernel that pushes one measure to another under the new metric. Fourth, from the dual representation, we establish a kernel representation theorem, such that we are able to embed the infinite dimensional criterion function into a tractable space. In doing this, the criterion function is represented in terms of a kernel-based distance. When the transition kernel induces the zero criterion under the new metric, the parameters are identified at θ_0, and a practical estimator for the parameters of interest is developed.

The contributions of this paper to the literature are as following. First, we propose a novel weighted minimum distance estimator from independence condition which employs a nonparametric estimation of the distance function. Second, we establish identification of the model in the dual representation. The dual distance characterizes finer topologies of probability measures. Third, based on the new representation, we provide both asymptotically biased (in the sense of root-n inconsistent) and unbiased statistics for estimation. In addition, we establish consistency and derive their limiting distributions. Fourth, we develop practical inference procedures. Fifth, we show that the proposed method is more efficient relative to the existing estimators. Finally, we provide tractable implementation for computing the estimator in practice.

From a technical point of view, the construction of the estimation procedure, and derivations of the statistical and limiting properties of the proposed estimator are of independent interest. The central mathematical tool used to accomplish these is the use of the optimal transportation theory and its corresponding dual problem. In particular, we make use of the Kantorovich's formulation of the optimal transportation problem. The novel integrated distance function can be interpreted as the optimal cost of transferring one mass distributed according to joint cumulative distribution function (c.d.f.) of ε and x to another mass distributed according to the product of ε and x's marginal c.d.f.s. The use of optimal weight is essential to the general formulation of integral minimum distance problems. Therefore, the main technical contribution of this paper is to introduce a new set of theorems for establishing the asymptotic results (consistency and weak convergence) for the weighted minimum distance from independence estimators for general distance functions in both primal and dual problems. The problem of approximating measures with respect to transportation distances has connections with the fields of probability theory (Talagrand, 1991), information theory (Graf and Luschgy, 2000), and optimal transport (Villani, 2009).

3

The transportation distance also appears in economics and econometrics. Current applications include rearrangements, identification, matching and quantile regression see, e.g. Galichon, Fernandez-Val, and Chernozhukov (2010), Ekeland, Galichon, and Henry (2010), Chiappori, McCann, and Nesheim (2010) and Galichon and Salanie (2010). Techniques from the transportation distance are often used for finding more tractable dual problems.

Estimation of econometric models based on the independence between the exogenous variables and the unobserved disturbance term has been explored in the literature. In a seminal work, Manski (1983) developed an approach of minimizing the distance from the independence condition for estimation. The procedure to estimate the parameters θ_0 uses a criterion function that compares the mean-square distance between a joint c.d.f. and the product of its marginal c.d.f.s. In order to derive the asymptotic properties of Manski's estimatior, Brown and Wegkamp (2002) extend the criterion function to an integrated mean-square distance criterion function. Importantly, the integral is taken with respect to a weighted measure on the product space of ε and x. Their approach provides a foundation for estimating θ_0 from general nonlinear simultaneous equation models.[1] However, it has been left unspecified the essential argument on the practical weighting measure of the product space of ε and x. The estimator studied in this paper is also related to other alternative methods. For instance, Komunjer and Santos (2010) develop a semiparametric estimator for invertible nonseparable models with scalar latent variables and an infinite dimensional component. Santos (2011) proposes an M-estimator under the assumption that the model is strictly monotonic in the scalar error term and derives its corresponding asymptotic properties.

The remaining of the paper is organized as follows. In Section 2 we present the basic formulation of the simultaneous equation models. Section 3 describes the criterion function including its primal and dual. In addition, it presents a method for representing the infinite dimensional criterion function. Section 4 discusses identification, and provides both asymptotically biased and unbiased estimators. Large sample statistical theorems are also included. Practical computation is presented in Section 5. Section 6 discusses several related estimates of the optimizing distance or divergence of probability measures. Conclusions appear in Section 7.

Notations: Throughout the paper, we use capital alphabet to denote random variable,

---

[1]As discussed in Benkard and Berry (2006), the identification condition that ε is independent of X generally does not hold for $\theta \neq \theta_0$. For an integrated criterion, such a dependence between X and ε under different θ reflects thorough the weighted measure.

i.e. X; use bold letter alphabet to denote a bundle of realizations, i.e. $\mathbf{x}$; and use the alphabet to denote the deterministic value, i.e. x; ith-realization is denoted with a subscript i.e. x_i.

In this paper, $\mathcal{H}$ always denotes a Hilbert space, i.e. a complete, norm vector space endowed with an inner product $\langle \cdot, \cdot \rangle$ giving rise to its norm via $\|x\| = \sqrt{\langle x, x \rangle}$. Let $B_{\mathcal{H}} := \{x \in \mathcal{H} : \|x\| \leq 1\}$ be the closed unit ball of $\mathcal{H}$. Every separable Hilbert space is isometrically isomorphic to the space of all square-summable sequences (ℓ_2 space). We denote $\| \cdot \|_\infty$ as the supremum norm such that $\|f\|_\infty := \sup_{x \in \mathcal{H}} |f(x)|$.

2 MODEL FRAMEWORK

2.1 MODEL SETUP AND ASSUMPTIONS

We begin by describing the model framework, notation, and main assumptions for the subsequent developments. The setup is similar to Brown and Wegkamp (2002, hereafter BW), but with important differences. We require greater generality, since we deal with distributions on compact metric spaces, and densities are not assumed to exist. To count for the general cases, we do not require existence of the reduce form.

Consider the following model

$$\varepsilon = \rho(X, Y, \theta), \tag{2}$$

where $\rho(\cdot)$ is the function describing the structural model, Y is the dependent observed random vector, X is an exogenous observed random vector, ε is the model error which is endogenous but is latent exogenous under the truth, and θ is a vector of unknown parameters.

First we define the distance used in the simultaneous equation model analyzed in this paper. We define the minimum distance from independence conditions, where the function $\mathbf{d}(\cdot, \cdot)$ is a metric on the space of joint cumulative distribution functions (c.d.f.'s) of (X, ε), where ε takes values in $\mathbb{R}^K$. Let $H_\theta(x, \varepsilon)$ be the joint c.d.f of (X, ε), and $F(x)$ and $G_\theta(\varepsilon)$ be the respective associated marginal c.d.f.'s. As a consequence of the identification assumption, which will be discussed below in detail, for any metric function $\mathbf{d}(\cdot, \cdot)$ on the space of measures,

$$\mathbf{d}(H_\theta(x, \varepsilon), F(x)G_\theta(\varepsilon)) = 0$$

if and only if X and ε are stochastically independent. Using this condition, we study ex-

tremum estimators that minimize the above distance.

We are interested in estimating the parameters θ in equation (2). Consider the following assumptions.

C1. Parameter: The true θ_0 belongs to a parameter space Θ.

C2. Observations (independent): $\mathbf{x} := (x_1, \ldots, x_n) \in \mathbb{R}^{L \times n}$ are n draws of random vector X. Any $X \in \mathcal{X}$. Let $(\Omega, \mathcal{F}_X, F)$ be a probability space, where Ω is a sample space, $\mathcal{F}_X$ is a σ-field, and F is a sample probability measure on $(\Omega, \mathcal{F}_X)$. Then $\mathbf{x} \in \{X(\omega) : \Omega \to \mathcal{X}\}$. Let $\mathcal{X}$ be a compact metric space.

C3. Observations (dependent) or model realizations: $\mathbf{y} := (y_1, \ldots, y_n) \in \mathbb{R}^{K \times n}$ are n draws of random vector Y. Any $Y \in \mathcal{Y}$. $Y := \rho^{-1}(X, \varepsilon, \theta_0) : \Upsilon \to \mathcal{Y}$ where $\Upsilon := \mathcal{X} \times \Sigma \times \Theta$. Let $\mathcal{Y}$ be a compact metric space.

C4. Observations (both independent and dependent): $\mathbf{z} := (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{(L+K) \times n}$ are n draws of joint observations $Z := (X, Y)$. Let $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.

C5. Unobservable variable: $\varepsilon = \rho(X, Y, \theta)$. For all $\theta \in \Theta$, $\rho(\cdot, \theta)$ is a mapping from $\mathcal{Z}$ into $\mathbb{R}^K$. Let $(\mathcal{Z} \times \Theta, \mathcal{B}(\mathcal{Z} \times \Theta), G_\theta)$ be a probability space, where $\mathcal{B}(\mathcal{Z} \times \Theta)$ is the Borel σ-algebra of the Cartesian product $\mathcal{Z} \times \Theta$. $G_\theta(\varepsilon)$ is a probability measure defined on $(\mathcal{Z} \times \Theta, \mathcal{B}(\mathcal{Z} \times \Theta))$.

C6. Identification condition: $\varepsilon = \rho(X, Y, \theta)$ is independent of X if and only if $\theta = \theta_0$. Let $H_\theta(x, \varepsilon)$ be the joint probability measure of (X, ε). The independent assumption between X and $\rho(X, Y, \theta_0)$ is equivalent with

$$H_\theta(x, \varepsilon) = F(x) G_\theta(\varepsilon) \quad \text{iff} \quad \theta = \theta_0$$

for any $(x, \varepsilon) \in \mathcal{X} \times \Sigma$.

C7. Structure: $\mathbb{S}$ is an ordered pair $(\rho(X, Y, \theta), H_{\theta_0}(x, \varepsilon))$. The observations $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ are generated by the structure $\mathbb{S}_0 := (\rho(X, Y, \theta_0), H_{\theta_0}(x, \varepsilon))$.

C8. Empirical probability measures: $F_n(x)$, $G_{n\theta}(\varepsilon)$ and $H_{n\theta}(x, \varepsilon)$ are the empirical measures associated with $F(x)$, $G_\theta(\varepsilon)$ and $H_\theta(x, \varepsilon)$ respectively based on the observed

data **z**. For example, $F_n(x)$ is a Borel probability measure defined on $\mathcal{X}$ as a discrete measure with a number n of points, such that

$$G_{n\theta}(\varepsilon) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\rho(z_i, \theta) \leq \varepsilon\}, \quad dG_{n\theta}(\varepsilon) := \frac{1}{n} \sum_{i=1}^{n} \delta_{\rho(z_i,\theta)}(\varepsilon) \qquad (3)$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and $\delta_{\rho(z_i,\theta)}(\cdot)$ is the Dirac function at position $\rho(z_i, \theta)$. The notation $d(\cdot)$ in this paper always denotes the differential notation.

C9. No reduce form: $\rho(X, \varepsilon, \theta)$ is non-invertible for $\theta \neq \theta_0$. It means that there is no way of estimating θ via $Y = \rho^{-1}(X, \varepsilon, \theta)$ or $\varepsilon = \rho(X, \rho^{-1}(X, \varepsilon, \theta), \theta)$. But at θ_0, $Y = f(X, \varepsilon)$ where $f \in \mathcal{F}$ is an unknown function and $\mathcal{F}$ can be embedded in a Hilbert space.

Conditions C1-C4 are standard in the econometric literature. In BW, compactness of the parameter space Θ is assumed while we stay with a general parameter space. But as a compensation, we impose the compactness for $\mathcal{Y}$, the sample space of dependent observations, in C3. The motivation of compactness in C2 and C3 is to ensure closeness and boundeness for both observable variables sample spaces $\mathcal{Z}$ in C4.[2] C5 concerns the measurability of $\mathcal{Z}$. The measurability of Z induces the measurability of $\varepsilon = \rho(z, \theta)$. Therefore, the probability measure of ε always exists. Conditions C6 and C7 impose identification and are the same as the identification conditions in Manski (1983) and BW. C6 is the independent condition for the population probability $H_\theta(x, \varepsilon)$. The joint probability $H_\theta(x, \varepsilon)$ equals the product of marginal c.d.f.s $F(x)$ and $G_\theta(\varepsilon)$ if and only if $\theta = \theta_0$. This is the key device to uniquely identify θ_0. This paper focus on estimation. But we highlight that, differently from BW, we establish identification in the dual problem. There is an extensive literature discussing identification in simultaneous equation models as that in our primal problem. We refer the reader to Benkard and Berry (2006) and the literature therein. C7 states that the pair observation (X, Y) generated by $\rho(X, Y, \theta_0)$ will give different simultaneous equation system $\rho(X, Y, \theta)$ under different value of θ. C8 defines the empirical c.d.f. and considers the weighted Dirac measure as the Randon-Nikodym derivative of the empirical c.d.f.. Similar forms of C7 and C8 can be found in Section 2 and Section 3 respectively in BW. Assumption C9 is novel and states that the simultaneous structure is not invertible. This is an important assumption allowing flexibility and generality of the proposed methods because the estimator

---

[2]If $\mathcal{Z}$ is non-compact, then a kernel function with non-compact support may not be able to identify the true parameter θ_0.

does not require knowledge on the reduced form. Thus, to attain an estimator of θ_0 by simply an inversion is not of our concern. However, with observations of Y, one can generate a nonparametric approximation for the underlying true function f. In order to ensure the approximation is feasible, the unknown true function f should come from an approximatible functional class $\mathcal{F}$. We restrict this class to reproducing kernel Hilbert space (RKHS). More discussion about RKHS will be given in Section A of the Appendix.

2.2 AN EXAMPLE

We now illustrate the model in light of the imposed conditions, especially C9. We start with a standard example satisfying our model specification (2). Consider a general separable supply and demand model

$$Q = D(Z, P; \theta_D) + \varepsilon_D$$
$$P = S(W, Q; \theta_S) + \varepsilon_S,$$

where Q is quantity, P is price, (Z, W) are characteristics, $(\varepsilon_D, \varepsilon_S)$ are shocks, (θ_D, θ_S) the parameters of interest, and the functions $D(\cdot)$ and $S(\cdot)$ are allowed to be nonlinear functions. In the representation (2) we have that $Y = (Q, P)$, $X = (Z, W)$, $\varepsilon = (\varepsilon_D, \varepsilon_S)$, $\theta = (\theta_D, \theta_S)$, and the structural model $\rho = (Q - D(\cdot), P - S(\cdot))$.

If a simultaneous equation model satisfies C1-C8 and has the specification in (2), one can establish the integrated distance function

$$\int \mathbf{d}(H_\theta(x, \varepsilon), F(x)G_\theta(\varepsilon))d\mu(x, \varepsilon), \tag{4}$$

which is continuous on Θ and θ_0 is the unique global minimum of this integrated distance function for a bounded measure $\mu(x, \varepsilon)$ of (x, ε) and a compact parameter space Θ. This result is a summary of Theorems 2 and 3 of BW.

To consider an example satisfying C9, we define a non-separable supply and demand model,

$$Q = D(Z, P, \varepsilon_D; \theta_D), \ P = S(W, Q, \varepsilon_S; \theta_S),$$

with a triangular structure for the joint distribution of P and Q such that

$$\tilde{U}_1 = \Psi(Q|Z, W), \ \tilde{U}_2 = \Psi(P|Z, W, Q),$$

8

where $\Psi(\cdot)$ is the joint distribution of the dependent variables and exogenous shifters, and $\tilde{U}$'s are constructed such that they are independent of one another as well as independent of (Z, W). The triangular construction implies that the reduced form takes the form $Q = f(Z, W, \tilde{U}_1)$, whereas in the general model the reduce form takes the form

$$Q = f(Z, W, \varepsilon_D, \varepsilon_S).$$

Thus, the triangular system cannot retrieve the true reduce form which satisfies C9.

It is important to highlight the role of condition C9 and its link with the literature. This assumption specifies the underlying relation between Y and X when $\theta = \theta_0$, which can be nonparametrically recovered. Assumption C9 excludes the possibility of reduced form relation. Models allowing for a reduced form have been addressed in Benkard and Berry (2006). Their concern is whether a derivative condition on $\rho(\cdot)$ is sufficient to derive a reduced form, and consequently identification. Differently from BW and Benkard and Berry (2006), the setup in this paper neither assume a derivative condition as in Lemma 3.1 of the later, nor assume the existence of a reduced form.

A generalization of (4) is the main concern in this paper. To estimate θ_0 via (4), BW consider a specified bounded measure μ and a compact parameter space. Importantly, both of these two conditions will be relaxed later in this paper. We will estimate an optimal measure γ (it differs from the original notation μ) and consider a general topological vector space Θ. To achieve this we first propose a new estimator. We show that the new estimator generalizes (4) and induces a dual representation. Given this dual representation setup, we show that the problem is still identifiable. The key device of our identification procedure is to select a proper integrated distance function so that $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ are from a complete separable metric space. If such an integrated distance similar to (4) one can induce a complete separable metric space for $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ on $\mathcal{Q}$, then any θ generating the associated distance value of (4) will be separated from the others. Thus θ_0, which gives a unique value of this integrated distance, is well separated from the other θ's.[3]

---

[3]This idea is based on Ascoli's theorem, see i.e. Shorack (Exercise 2.10 2000). Ascoli's theorem states that for a class of equicontinuous functions mapping from a complete separable metric space $\mathscr{P}$ to another metric space Θ, any sequence of such functions with a compact support will uniformly converge. The mapping in our setup is an estimator. The complete separable metric space is for $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$. Equicontinuouity implies that if any θ such that $d(\theta, \theta_0) \geq \delta$, then the distance between H_θ and H_{θ_0} (or P_θ and P_{θ_0}) is larger than some ϵ. This is exactly the identification condition.

3 CRITERION FUNCTION

This section establishes the grounds for constructing the minimum integrated distance estimator in the next section. First, we define the primal objective function from the independence condition. Second, we reformulate the problem with a new criterion function based on the Monge-Wasserstein distance. Third, we establish the validity of the dual problem. Finally, we establish a kernel representation theorem.

3.1 PRIMAL OBJECTIVE FUNCTION

With a bounded measure μ on $\mathcal{Z}$, under the independence assumption the criterion function for estimation of θ_0 is given by

$$M(\theta) = \int \mathbf{d}(H_\theta(x, \varepsilon), F(x)G_\theta(\varepsilon))d\mu(x, \varepsilon). \tag{5}$$

This distance function is based on mean-square distance. The criterion function is minimized at θ_0. In the literature, statistical estimation has been based on the empirical counterpart of $M(\theta)$,

$$\widehat{M}(\theta) = \int \mathbf{d}(H_{n\theta}(x, \varepsilon), F_n(x)G_{n\theta}(\varepsilon))d\mu(x, \varepsilon), \tag{6}$$

where $H_{n\theta}(x, \varepsilon)$, $F_n(x)$, and $G_{n\theta}(\varepsilon)$ are as defined in (3).

An open question for this estimation procedure is how to construct an estimator for the bounded measure μ. Actually, in practice the selection of μ will affect the estimation scheme substantially. If μ is attached to a specific measure form, then the natural question is what form of μ would be in practice. One possibility for practical implementation would be to use a grid search on μ which is shown in Brown and Wegkamp (2001). However, this procedure implies the use of a uniform empirical distribution for μ, which might not be the correct underlying distribution. Below, we show numerically that an optimal choice of μ is quite different from the uniform measure.

In (5), BW requires the integral measure $\mu(x, \varepsilon)$ to be independent of θ. Without this condition, there may exist some $\theta \neq \theta_0$ such that

$$n^{-1} \sum_{i=1}^{n} \mathbf{d}(H_{n\theta}(x_i, \varepsilon_i), F_n(x_i)G_{n\theta}(\varepsilon_i)) = 0.$$

10

A simple choice of this empirical measure is to set $\mu(x, \varepsilon) = H_{n\theta}(x, \varepsilon)$. That means the empirical summation is taken over the data (x_i, ε_i) generated by the model $\rho(X, \varepsilon, \theta)$ at θ. The summation implies that the value of θ matters for evaluating the integral. By endogenizing the role of θ in the integrated minimum distance function, we can relax the requirement of independence of μ in (5).

Therefore, the first goal of this paper is to propose a new estimator in which we endogenize the measure $\mu(x, \varepsilon)$ and then estimate the new endogenized measure and θ_0 simultaneously while making minimal assumptions regarding the measure $\mu(x, \varepsilon)$. This is an important innovation to practitioners since the estimation of $\mu(x, \varepsilon)$ affects the estimation of the parameters of interest. We will illustrate that in the numerical simulations below. In addition, we provide a general result for estimating these measures. Formally, we formulate the following inferential procedure:

(M) Let the joint probability measure $H_\theta(x, \varepsilon)$ and the mixing probability measure $P_\theta(x, \varepsilon) = F(x)G_\theta(\varepsilon)$ be Borel probability measures defined on a domain $\mathcal{Q} = \mathcal{X} \times \Sigma$. Given observations $\mathbf{x} \in \mathcal{X}$ and generating realizations $\varepsilon = (\rho(x_1, y_1, \theta), \ldots, \rho(x_n, y_n, \theta))$, we consider that $(\mathbf{x}, \varepsilon) \sim H_\theta(x, \varepsilon)$ and $(\mathbf{x}', \varepsilon') \sim P_\theta(x', \varepsilon')$ are drawn independently and identically distributed (i.i.d) from two unknown probability measures on $\mathcal{Q}$. We represent $\mu(x, \varepsilon)$ as a transport measure $\gamma(x, \varepsilon; x', \varepsilon')$ that pushes $H_\theta(x', \varepsilon')$ towards $P_\theta(x, \varepsilon)$. We show that there is an optimal γ and represent this optimal measure using kernels. By this optimal measure, we introduce a new criterion (Wasserstein distance function) and develop the associated estimator and testing procedure.

3.2 Transport Measure Representation

In this section, we reformulate the primal problem described above and represent $\mu(x, \varepsilon)$ as a transport measure. We start by reformulating the problem (5) with a new criterion $W(\theta)$. The criterion is based on the Wasserstein or Monge-Wasserstein distance. Following the definition in Dudley (p.420, 2002), we introduce the criterion:

$$W(\theta) := \inf_{\gamma \in \Gamma(H_\theta, P_\theta)} \left\{ \int \mathbf{d}(q, q') d\gamma(q, q') : q \sim H_\theta(x, \varepsilon) \text{ and } q' \sim P_\theta(x, \varepsilon) \right\}, \qquad (7)$$

where $\gamma(\cdot)$ is the transport measure, $\Gamma(H_\theta, P_\theta)$ is the set of jointly distributions on $\mathcal{Q} \times \mathcal{Q}$ with the marginals of γ given by $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$. Note that $H_\theta(x, \varepsilon)$ as a marginal of

γ is still jointly distributed for (x, ε). The set Γ refers to a product of probability measures. The function $\mathbf{d}(\cdot, \cdot)$ is a metric function for samples on $\mathcal{Q}$. The differences between the original problem (5) and the problem (7) are threefold. First, the criterion function in (7) distinguishes the joint observations $(q_1, q_2, \dots) = (\{x_1, \varepsilon_1\}, \{x_2, \varepsilon_2\}, \dots)$ and the marginal observation $(q'_1, q'_2, \dots) = (\{x_1, \varepsilon'_1\}, \{x_2, \varepsilon'_2\}, \dots)$. The observations ε in $Q := (X, \varepsilon)$ are drawn from $\varepsilon = \rho(X, Y, \theta)$ which are jointly distributed with X. While the *generic* observations ε' in $Q' := (X, \varepsilon')$ are drawn from a marginal distribution $G_\theta(\varepsilon)$ independently of X. Empirically, the number of observations in these two cases are also different. We denote $\mathbf{q}$ as $(q_1, \dots, q_n)$ and denote $\mathbf{q}'$ as $(q'_1, \dots, q'_m)$.[4] Second, the measure γ in (7) is defined as a product measure of H_θ and P_θ. Unlike the unspecified $\mu(x, \varepsilon)$ in the original problem (5), γ is restricted to the set $\Gamma(H_\theta, P_\theta)$ such that any element in this set satisfies:

$$H_{n\theta}(x_i, \varepsilon_i) = \frac{1}{m} \sum_{j=1}^{m} \gamma((x_i, \varepsilon_i), (x_j, \varepsilon'_j)), \quad P_{n\theta}(x_j, \varepsilon'_j) = \frac{1}{n} \sum_{i=1}^{n} \gamma((x_i, \varepsilon_i), (x_j, \varepsilon'_j)). \quad (8)$$

Third, the distance function in (7), $\mathbf{d}(\cdot, \cdot)$, is for samples (x_i, ε_i) and (x_j, ε'_j) not for their c.d.f.s.

Similarly to (6), the new minimum distance estimator will be based on the empirical Wasserstein distance defined as follows:

$$\widehat{W}(\theta) := \inf_{\gamma \in \Gamma(H_{n\theta}, P_{m\theta})} \left\{ \int \mathbf{d}(q, q') d\gamma(q, q') : q \sim H_{n\theta}(x, \varepsilon) \text{ and } q' \sim P_{m\theta}(x, \varepsilon) \right\}. \quad (9)$$

$$= \inf_{\gamma(q_i, q_j) \in \Gamma(H_{n\theta}, P_{m\theta})} \left\{ \frac{1}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbf{d}(q_i, q'_j) : q_i \sim H_{n\theta}(x, \varepsilon) \text{ and } q'_j \sim P_{m\theta}(x, \varepsilon) \right\}.$$

The subscript m indicates that q' are drawn from m samples. One can consider $m = n$ as a special case. The above criterion function (9) and the corresponding marginals in (8) contain $\gamma(\cdot)$. Thus, without further restrictions, $\gamma(\cdot)$ is an infeasible element with infinite dimensions. Next we provide a tractable representation of $\widehat{W}(\theta)$ in order to achieve feasible estimation. The infinite dimension issue will be solved by considering the dual representation of $\widehat{W}(\theta)$.

Remark 1. The interpretation of (7) is the cost in terms of a distance function $\mathbf{d}(\cdot, \cdot)$ of transferring a mass $Q := (X, \varepsilon)$ distributed according to $H_\theta(x, \varepsilon)$ to a mass $Q' := (X', \varepsilon')$

---

[4]More generally, we can think $\mathbf{q}' = (q'_1, \dots, q'_m)$ are from a resampling scheme of $\mathbf{q}$. Then even if $m = n$, one should realize that $\mathbf{q}' \neq \mathbf{q}$.

distributed according to $P_\theta(x, \varepsilon)$. The measure $\gamma \in \Gamma$ is the transportation schedule between Q and Q'.

Remark 2. The criterion $W(\cdot)$ comes from a general class called k-Wasserstein distance $W^k(H, P) := \left\{ \inf_{\gamma \in \Gamma(H,P)} \int \mathbf{d}(q, q')^k d\gamma(q, q') \right\}^{\frac{1}{k}}$. Let $\mathscr{P}^k(\mathcal{Q})$ denote the space of Borel probability measures on $\mathcal{Q}$ with the first k-th moments. For any sample space with finite moment, the space of measures endowed with W^k metric is a complete separable metric space (p.94, Villani, 2009). The most useful versions of k-Wasserstein distance are $k = 1, 2$. In our context, $W(\theta)$ is the 1-Wasserstein distance. Note that both measures $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ belong to $\mathscr{P}^1(\mathcal{Q})$. Thus, $W(\theta)$ induces a complete separable metric space for $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ on $\mathcal{Q}$. This complete separable metric space with compact supports imply the identification of θ_0 in our dual problem. Therefore, except the case that $H_{\theta_0}(x, \varepsilon) = P_{\theta_0}(x, \varepsilon)$, any $H_\theta(x, \varepsilon)$ will differ from $P_\theta(x, \varepsilon)$ in $W(\theta)$-distance. Thus any $W(\theta)$ will be separated from zero if $\theta \neq \theta_0$. In other words, θ_0 is well separated from the other θ's.

Remark 3. There are many possible choices of distances between probability measures, such as the Levy-Prokhorov (Brown and Matzkin, 1998), or the weak-* distance. Among them, W^k metrizes weak convergence (Theorem 6.9, Villani, 2009), that is, a sequence $\{P_n\}_{n \in \mathbb{N}}$ of measures converges weakly to P if and only if $W^k(P_n, P) \to 0$. However, as pointed out in Villani (p.98, 2009) "Wasserstein distances are rather strong,... a definite advantage over the weak-*distance". Also, it is not so difficult to combine information on convergence in Wasserstein distance with some smoothness bound, in order to get convergence in stronger distances.

Remark 4. Note that, for clarity, we adhere to the strict association of a random variable with a distribution. Thus $H_{n\theta}(x, \varepsilon)$ and $P_{m\theta}(x, \varepsilon)$ are associated with distinct random variables (q and q', respectively) representing distinct states of knowledge about the same set of variables. This view is slightly different than the usual notion of a single random variable for which we "update" our belief.

Remark 5. When the metric $\mathbf{d}(q, q')$ for distribution laws on $\mathbb{R}$ is the absolute function $|q - q'|$, $W(\cdot)$ is the Gini index, see e.g. Dudley (p.435, 2002).

3.3   The Dual Problem

While the independence condition is for c.d.f.s, the distance function in the optimization described in (7) is for samples. Thus, by transferring the focus from samples to c.d.f.s, we

need a dual problem of (7). We note that, in this paper, both he identification and estimation will be conducted in this dual problem.

The dual representation of $W(\theta)$ in (7) was introduced by Kantorovitch in order to solve of a convex linear program. For the empirical distance $\widehat{W}(\theta)$ in (9), the dual corresponds to the convex relaxation of a combinatorial problem when the densities are sums of the same number of Diracs. This relaxation extends the notion of $\Gamma(H_{n\theta}, P_{m\theta})$ to arbitrary sum of weighted Diracs, see for instance Villani (2003). The following is called Kantorovich-Rubinstein theorem. It introduces the Kantorovich's duality principle to $W(\theta)$ and $\widehat{W}(\theta)$.

Theorem 1. *Let $H_\theta, P_\theta \in \mathscr{P}^1(\mathcal{X})$. The dual problem to (7) by Kantorovich's duality principle is:*

$$W(\theta) = \|H_\theta(x, \varepsilon) - P_\theta(x, \varepsilon)\|_l^* := \sup_{\|f\|_l \leq 1} \left| \int f d\{H_\theta(x, \varepsilon) - P_\theta(x, \varepsilon)\} \right|,$$

where $\|f\|_l := \sup_{x \neq y \in \mathcal{X}} |f(q) - f(q')|/\mathbf{d}(q, q')$ is the Lipschitz semi-norm for real valued continuous f on $\mathcal{Q}$. Here $d\{\cdot\}$ is the differential notation. The optimal transportation problem for discrete measure $H_{n\theta}(x, \varepsilon)$ and $P_{m\theta}(x, \varepsilon)$ becomes

$$\widehat{W}(\theta) = \|H_{n\theta}(x, \varepsilon) - P_{m\theta}(x, \varepsilon)\|_l^* := \sup_{\|f\|_l \leq 1} \left| \int f d\{H_{n\theta}(x, \varepsilon) - P_{m\theta}(x, \varepsilon)\} \right|,$$

$$H_{n\theta} = \frac{1}{n^2} \left[\sum_{i=1}^n \sum_{j=1}^n \mathbf{1}\{x_i \leq x, \rho(z_j, \theta) \leq \varepsilon\} \right], P_{n\theta} = \frac{1}{n^2} \left[\sum_{i=1}^n \mathbf{1}\{x_i \leq x\} \right] \left[\sum_{i=1}^n \mathbf{1}\{\rho(z_i, \theta) \leq \varepsilon\} \right].$$

Proof. Please refer to Dudley (Theorem 11.8.2, 2002). Note that Dudley (Theorem 11.8.2, 2002) gives a stronger result which is for complete separable metric space. But the original result in Rubinshtein (1970) is for the compact metric space $\mathcal{Q}$. □

The integrals in Theorem 1 are with respect to empirical measures. In practice, these integrals are replaced by summations over observations. Recall that Q and Q' are drawn from different distributions. The empirical Wasserstein distance in (9) becomes

$$\widehat{W}(\theta) := \sup_{\|f\|_l \leq 1} \left| \frac{1}{n} \sum_{i=1}^n f(q_i) - \frac{1}{m} \sum_{j=1}^m f(q'_j) \right|. \tag{10}$$

where m is not necessarily the same as n.

The criterion function in (10) is similar to the total variation distance between $H_{n\theta}$ and $P_{m\theta}$. Thus, the estimation problem based on integral distance function in BW is modified to the one based on the total variation distance. The distance between $H_{n\theta}$ and $P_{m\theta}$ is evaluated by the comparison of means of $f(\cdot)$ w.r.t. $H_{n\theta}$ and $P_{n\theta}$. The way of handling two different integrals or summations separately makes the estimation more feasible. One advantage of this separation is that the empirical c.d.f. of $G_\theta(\varepsilon)$ allows for resampling scheme:

$$m^{-1} \sum_{j=1}^{m} \mathbf{1}\{\rho(x_j', y_j', \theta) \leq \varepsilon\},$$

where $\{x_j', y_j'\}_{j \leq m}$ are resamples of $\{x_i, y_i\}_{i \leq n}$. It also allows for other smoothing techniques for estimating $G_\theta(\varepsilon)$. If the sample size is small, the smoothing technique becomes important for identifying θ_0. Notice that $F_n(x)G_{n\theta}(\varepsilon) \neq H_{n\theta}(x, \varepsilon)$

$$\frac{1}{n^2}\left[\sum_{i=1}^{n} \mathbf{1}\{x_i \leq x\}\right]\left[\sum_{i=1}^{n} \mathbf{1}\{\rho(z_i, \theta) \leq \varepsilon\}\right] \neq \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{1}\{x_i \leq x, \rho(z_j, \theta) \leq \varepsilon\}\right],$$

for given (x, ε). Thus if $\theta \neq \theta_0$, we have that

$$\sum_{i=1}^{n} f(z, \varepsilon)\left[\delta_{\rho(z_i, \theta)}(\varepsilon)\right]\left[\delta_{x_i}(x)\right] \neq \sum_{i=1}^{n}\sum_{j=1}^{n} f(z, \varepsilon)\left[\delta_{\{x_i, \rho(z_j, \theta)\}}(x, \varepsilon)\right].$$

But if the sample size is small, i.e. $n = 10$, it is very likely to have an equality from the above expression even if $\theta \neq \theta_0$. Nevertheless, with re-sampling devices, one can reduce this small samples issue.

The dual distance $\|H_\theta(x, \varepsilon) - P_\theta(x, \varepsilon)\|_l^*$ is an instance of the integral probability metric which has been used in proving central limit theorems and in empirical process theory (Muller, 1997). The dual distance measures the dis-similarity between $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ on the basis of samples drawn from each of them, by finding a well behaved function which is large on the points drawn from $H_\theta(x, \varepsilon)$ and small on the points drawn from $P_\theta(x, \varepsilon)$. This is equivalent to measure the difference between the mean function values on two samples.

Theorem 1 establishes the dual problem, and hence the objective function of interest. However, there are infinite candidates f satisfying $\|f\|_l \leq 1$ in equation (10). Thus, both Kantorovich's dual $W(\theta)$ and empirical estimate criterion $\widehat{W}(\theta)$ are criterion functions over infinite dimensions. Before we establish identification and the estimator of interest we pro-

15

vide yet another representation to the problem which allows for practical implementation. Therefore, in the next section, we embed this infinite dimensional problem into a tractable form so that we can represent the criterion function.

3.4   Kernel Representation

Now we establish a kernel representation theorem, such that we are able to embed the infinite dimensional criterion function in (10) into a tractable space. In doing this, the criterion function is represented in terms of a kernel-based distance.

3.4.1   Kernel Representation Theorem

In this paper, we only consider continuous kernel because every function induced by a continuous kernel is also continuous. For a continuous kernel $\mathbf{k}$ on a compact metric space $(\mathcal{X}, d)$, we are interested in the space of all functions induced by $\mathbf{k}$ that is *dense* in $C(\mathcal{X})$ (the space of *all continuous functions*), i.e. for every function $f \in C(\mathcal{X})$ and every $\epsilon > 0$ there exists a function g induced by $\mathbf{k}$ with

$$\|f - g\|_\infty \le \epsilon.$$

Definition. If a kernel induces a dense set of $C(\mathcal{X})$, we say that the kernel is able to *represent* $C(\mathcal{X})$.

Next we provide the a kernel representation result.

Theorem 2. *(Representation Theorem) Let $(\mathcal{X}, \mathbf{d})$ be a compact metric space and $\mathbf{k}$ is a kernel on $\mathcal{X}$ with $\mathbf{k}(x, x) > 0$ for all $x \in \mathcal{X}$. Suppose that we have an injective feature map $\Phi : \mathcal{X} \to \ell_2$ of $\mathbf{k}$ with $\Phi(x) = (\Phi_n(x))_{n \in \mathbb{N}}$. If $\mathcal{A} := span\{\Phi_n : n \in \mathbb{N}\}$ is an algebra then $\mathbf{k}$ represents $C(\mathcal{X})$.*

Proof. This is a direct result of representation for reproducing kernels (Theorem 12) and Stone-Weierstrass theorem (Thereom 11).   □

In the Appendix, Section A, we state additional mathematical results on the existence of the kernel function $\mathbf{k}(q)$ for $\mathcal{F} := \{f \ : \ \|f\|_l \le 1\}$.

3.4.2 $W_{\mathcal{H}}(\theta)$-DISTANCE

The quality of empirical Wasserstein distance $\widehat{W}(\theta)$ as a statistic depends on the class $\mathcal{F} := \{f : \|f\|_l \leq 1\}$. There are two requirements: (1) $\mathcal{F}$ must be large enough so that $\widehat{W}(\theta) = 0$ if and only if $H_{n\theta}(x, \varepsilon) = P_{n\theta}(x, \varepsilon)$; (2) for the distance to be consistent, $\mathcal{F}$ needs to be restrictive enough for the empirical estimate of $\widehat{W}(\theta)$ to converge to its expectation. We will use a unit ball in the Hilbert space where all functions in the class $\mathcal{F}$ can be represented by the elements in this ball. This representation will be shown to satisfy both of the required properties.

In this paper, the class $\mathcal{F}$ is chosen as follows. Given observations $\mathbf{y} = (y_1, \ldots y_n)$ and $y_i = f(x_i, \varepsilon_i)$ for an unknown function f defined in condition C9, we consider a kernel function $\mathbf{k}(q)$ in $\mathcal{F}$. This kernel function has an inner product representation $\mathbf{k}(q, \cdot) = \langle \Phi(q), \Phi(\cdot) \rangle$. The function Φ in the representation approximates any $f \in \mathcal{F}$ such that for any $\epsilon > 0$, there exists α as the coefficient and

$$\|f(q) - \langle \alpha, \Phi(q) \rangle\|_\infty \leq \epsilon$$

with ϵ-degree of accuracy.

Now we propose a new distance $W_{\mathcal{H}}(\theta)$ that characterizes the distance between two distinct measures $H_{n\theta}(x, \varepsilon)$ and $P_{n\theta}(x, \varepsilon)$ for $\theta \neq \theta_0$ in the optimal sense. The idea is to consider $\mathcal{F} := \{f : \|f\|_l \leq 1\}$ to be embedded in a unit ball $\mathcal{B}_{\mathcal{H}}$ of a Hilbert space $\mathcal{H}$.[5] For any function $f \in \mathcal{B}_{\mathcal{H}}$ there exists a feature map $\Phi : \mathcal{Q} \to \mathcal{H}$. Then $\mathcal{H}$ is the feature space of a kernel $\mathbf{k}$ which represents all continuous bounded functions in $\mathcal{F}$. In other words, we construct a Hilbert space so that

$$\mathcal{F} := \{f : \|f\|_l \leq 1\} = \{f : \|f\|_{\mathcal{H}} \leq 1\} = \mathcal{B}_{\mathcal{H}}.$$

By Theorem 2 we know that a kernel $\mathbf{k}$ induced by the feature map $\Phi : \mathcal{Q} \to \mathcal{H}$ can represent $f \in \mathcal{B}_{\mathcal{H}}$ and hence after $f \in \mathcal{F}$. Then the same kernel will be used for defining the metric for $W(\theta)$.

Here is our construction. First, we construct a pseudo-metric on $\mathcal{B}_{\mathcal{H}}$ by using the kernel $\mathbf{k}$. Kernel $\mathbf{k}$ is induced by $\mathcal{H}$. Then we will show that the new metric characterizes the weak

---

[5]$\mathcal{F}$ is embedded in $\mathcal{B}_{\mathcal{H}}$ means that there exists a map $\Pi : \mathcal{F} \to \mathcal{B}_{\mathcal{H}}$ such that the map is injective and preserving topological structure of $\mathcal{F}$

topology of $\widehat{W}(\theta)$ and in Section 4.3 below we develop an asymptotically unbiased empirical estimate for $W(\theta)$ based on this new metric.

C10. For any $\theta \in \Theta$, the kernel $\mathbf{k}$ is square-root integrable w.r.t. both $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$. That is

$$\int_{\mathcal{Q}} \sqrt{\mathbf{k}(q, q)} dH_\theta(q) < \infty \text{ and } \int_{\mathcal{Q}} \sqrt{\mathbf{k}(q, q)} dP_\theta(q) < \infty.$$

and $\mathbf{k}(\cdot, \cdot) \in \mathcal{B}_{\mathcal{H}}$ is measurable on $\mathcal{Q} \times \mathcal{Q}$.

Let $\mathscr{P}(\mathcal{Q}) := \left\{ \mathcal{P} : \int_{\mathcal{Q}} \sqrt{\mathbf{k}(q, q)} d\mathcal{P}(q) < \infty \right\}$. Condition C10 implies $\{H_\theta(x, \varepsilon), P_\theta(x, \varepsilon)\} \subset \mathscr{P}(\mathcal{Q})$. Note that $\mathscr{P}(\mathcal{Q})$ is a subset of $\mathscr{P}^1(\mathcal{Q})$, the space of Borel probability measures on $\mathcal{Q}$ with the first moments. C10 restricts the support $\mathscr{P}^1(\mathcal{Q})$ in $W(\theta)$. However, it returns flexibility of defining pseudo-metric for $\mathscr{P}(\mathcal{Q})$ by kernels.

Theorem 3. *Assume that the simultaneous model $\rho(X, Y, \theta) = \varepsilon$ satisfies C1-C9. If $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ satisfy C10, then*

$$W_{\mathcal{H}}(\theta) := \left\| \underbrace{\int_{\mathcal{Q}} \mathbf{k}(\cdot, q) dH_\theta(q)}_{:= H_\theta \mathbf{k}} - \underbrace{\int_{\mathcal{Q}} \mathbf{k}(\cdot, q) dP_\theta(q)}_{:= P_\theta \mathbf{k}} \right\|_{\mathcal{H}}, \tag{11}$$

where $\mathcal{H}$ induces kernel $\mathbf{k} \in \mathcal{B}_{\mathcal{H}}$.

Proof. See Appendix C.1. □

Therefore, the minimum integrated distance estimator proposed in this paper will be defined below as a minimization of problem in equation (11).

As stated in Theorem 2, an algebra $\mathcal{A} := \text{span}\{\Phi_n : n \in \mathbb{N}\}$ associating with a sequence of $(\Phi_n(x))_{n \in \mathbb{N}}$ will represent any $f \in \mathcal{F} \subset C(\mathcal{Q})$. By using the inner product operation, we have

$$f = \langle \alpha, \Phi(\cdot) \rangle = \langle \langle \beta, \Phi(\cdot) \rangle, \Phi(\cdot) \rangle = \langle \alpha', \mathbf{k}(\cdot, \cdot) \rangle,$$

where α, β, and α' are some coefficients in $\mathbb{R}$. The bilinear map $\langle \alpha', \cdot \rangle : \mathcal{F} \to \mathcal{B}_{\mathcal{H}}$ is injective and preserving the topological structure of $\mathcal{F}$. Thus we can say $\mathcal{F}$ is embedded into a unit ball $\mathcal{B}_{\mathcal{H}}$ in the Hilbert space. The representation of $W_{\mathcal{H}}(\theta)$ yields a mapping from $\mathscr{P}(\mathcal{Q})$ to a Hilbert space $\mathcal{H}$:

$$\mathcal{P} \mathbf{k} : \mathcal{P} \mapsto \int_{\mathcal{Q}} \mathbf{k}(\cdot, q) d\mathcal{P}(q)$$

18

where $\mathcal{P} \in \mathscr{P}(\mathcal{Q})$ and $\mathbf{k} \in \mathcal{B}_{\mathcal{H}}$. In order to verify that $W_{\mathcal{H}}(\theta)$ is a distance function, we need the following corollary.

Corollary 1. $W_{\mathcal{H}}(\theta)$ *metrizes the weak topology on* $\mathscr{P}$.

Proof. See Appendix D.1. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

$W_{\mathcal{H}}(\theta)$ measures the dis-similarity between H_θ and P_θ, however, so far it is not clear that whether the new distance $W_{\mathcal{H}}(\theta)$ relates to $W(\theta)$. Corollary 1 states the result that all elements in $\mathscr{P}(\mathcal{Q})$ can be metrized by $W_{\mathcal{H}}(\theta)$. The following condition ensures the regularity of $W_{\mathcal{H}}(\theta)$ for all $q \in \mathcal{Q}$. With these results, $W_{\mathcal{H}}(\theta)$ is comparable with $W(\theta)$ in the probability metric space.

C11. Let $\sup_{q \in \mathcal{Q}} \mathbf{k}(q, q) \leq C$ and $\mathbf{d_k} = \|\mathbf{k}(\cdot, q) - \mathbf{k}(\cdot, q')\|_{\mathcal{H}}$. We assume that the metric space $(\mathcal{Q}, \rho)$ is separable.

Theorem 4. *For any $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ satisfy C10 and $(\mathcal{Q}, \mathbf{d_k})$ satisfies C11, then*

$$W_{\mathcal{H}}(\theta) \leq W(\theta) \leq \sqrt{W_{\mathcal{H}}^2(\theta) + 2C}$$

where the distance function in $W(\theta)$ is $\mathbf{d_k}$.

Proof. See Appendix C.2 $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Remark 6. The metric $\mathbf{d_k}(\delta_q, \delta_{q'}) = \|\mathbf{k}(\cdot, q) - \mathbf{k}(\cdot, q')\|_{\mathcal{H}}$ used in Theorem 4 is called Hilbertian metric. Given this metric, one can obtain the associated kernel $\mathbf{k}$ via

$$\mathbf{k}(q, q') = \frac{1}{2} \left[\mathbf{d_k^2}(q, z) + \mathbf{d_k^2}(q', z) - \mathbf{d_k^2}(q, q') \right]$$

for any $q, q', z \in \mathcal{Q}$. This is called three points interpolation.

To conclude this section, note that we use a general Wasserstein criterion (7) to measure the dis-similarity between these two sequences. The criterion captures infinite intrinsic connections between $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ via a transition measure γ. The measure γ in Wasserstein criterion (7) is the key mechanism to provide an optimal criterion function of the minimum distance problem over two probability measures. By Kantonovich's dual (10) and kernel representation result in (11), we embed this infinite dimensional criterion function into a

tractable space. Then the criterion function is represented in terms of a kernel-based distance. Since the new distance is equivalent to the total variation distance, the modified estimation problem maintains finer topological details.

Given these results we are in position to establish identification, propose the estimator, and derive its asymptotic properties.

4 ESTIMATION AND INFERENCE

In this section we first establish identification in the dual problem, then construct the new minimum integrated distance estimator θ^* for θ_0 and finally provide the corresponding asymptotic properties, and discuss its optimality. Practical computation of the proposed estimator will be given in Section 5.

4.1 IDENTIFICATION

In order to identify θ_0, we need to make sure that when $H_\theta \neq P_\theta$, the kernel based distance $W_{\mathcal{H}}(\theta)$ cannot attain zero. The following condition restricts the class of kernels to be strictly positive definite.

C12. The k is strictly positive definite on $\mathcal{Q}$ such that

$$\int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') d\mathcal{P}(q) d\mathcal{P}(q') > 0$$

for any $\mathcal{P} \in \mathscr{P}$ and $\mathscr{P}(\mathcal{Q}) := \left\{ \mathcal{P} : \int_{\mathcal{Q}} \sqrt{k(q, q)} d\mathcal{P}(q) < \infty \right\}$.

Example. There are many kernels satisfying the integrally strictly positive definite condition. Here we list few of them: Gaussian kernel $\exp(-\sigma \|q - q'\|_2^2)$ with $\sigma > 0$; inverse multi-quadratics $(\sigma^2 + \|q - q'\|_2^2)^{-c}$ with $c > 0$ and $\sigma > 0$, Laplacian kernel $\exp(-\sigma \|q - q'\|_1)$, with $\sigma > 0$.

Lemma 1. *Let k be a bounded kernel on a metric space $(\mathcal{Q}, \mathbf{d_k})$. Let $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ satisfy C1-C12. Then the following statement is true that $H_\theta(x, \varepsilon) \neq P_\theta(x, \varepsilon)$ but $W_{\mathcal{H}}(\theta) \neq 0$ if and only if there exists a finite non-zero signed Borel measure $\mathcal{P} \in \mathscr{P}(\mathcal{Q})$ that satisfies:*
(I) $\int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') d\mathcal{P}(q) d\mathcal{P}(q') = 0$. (II) $\mathcal{P}(\mathcal{Q}) = 0$.

Proof. See Appendix D.2. □

20

Theorem 5. *For any $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ satisfy C10 and kernel $\mathbf{k}$ satisfies C11, if $H_\theta(x, \varepsilon) \neq P_\theta(x, \varepsilon)$, then $W_\mathcal{H}(\theta)$ does not attain zero.*

Proof. Condition C12 implies that for any non-zero signed Borel measure $\mathcal{P}$, (I) in Lemma 1 is violated. Therefore, if $H_\theta(x, \varepsilon) \neq P_\theta(x, \varepsilon)$, then $W_\mathcal{H}(\theta)$ must be non-zero. $\qquad\square$

Remark 7. In fact, both $W_\mathcal{H}(\theta)$ and $W(\theta)$ are equivalent to the total variation distance

$$\sup_{q \in \mathcal{Q}} |H_\theta(q) - P_\theta(q)|$$

if the distance function $\mathbf{d}(\cdot, \cdot)$ in the primal problem (7) of $W(\theta)$ is chosen to be a Hamming metric $H(x, y) = \mathbf{1}_{x \neq y}$. This result is given in Gozlan and Leonard (Proposition 1.3, 2010).

4.2   Convergence of $P_{n\theta}$ and $H_{n\theta}$

In this paper, we only consider pre-compact space $(\mathscr{P}, \mathbf{d})$ over Hilbert spaces (complete spaces) which means that the space $(\mathscr{P}, \mathbf{d})$ is a *compact metric space*. In other words, the covering numbers of $\mathscr{P}$ are finite and distance $\mathbf{d}(\cdot, \cdot)$ metrizes $\mathscr{P}$ by a Hilbertian type metric. The covering number is to measure how big the function set $\mathscr{P}$. For a finite covering number, Glivenko-Cantelli theorem (uniform law of large number) always hold. Thus, sample counterparts of P_θ, H_θ converge. The rigorous probability theory argument is given in the Appendix B.

4.3 $\sqrt{n}$-Biased and Unbiased Estimates

Given the identification of the dual problem stated in the previous section, we are able to estimate the parameters of interest. The minimum integrated distance estimator, θ^*, is defined as follows

$$\theta^* = \arg\min_{\theta \in \Theta} \widehat{W}_\mathcal{H}(\theta) := \arg\min_{\theta \in \Theta} \left\| \int \mathbf{k}((x, \varepsilon), q) dH_{n\theta}(q) - \int \mathbf{k}((x, \varepsilon), q') dP_{n\theta}(q') \right\|_\mathcal{H}, \quad (12)$$

where $\mathbf{k}((x, \varepsilon), \cdot) = \langle \Phi(x, \varepsilon), \Phi(\cdot) \rangle$ is the kernel function on a feature space, and $\Phi(\cdot)$ is the feature map such that the representation $f = \langle \alpha, \Phi(\cdot) \rangle$ is alway valid for any $f \in \mathcal{F}$. This criterion is a restatement of the minimization problem in (11).

The estimator of the empirical counterpart of $W_\mathcal{H}(\theta)$, $\widehat{W}_\mathcal{H}(\theta)$, defined in (12), has an asymptotic bias. This is because $W_\mathcal{H}(\theta)$ is an equivalent metric as total variation distance,

as pointed out in Devroye and Gyorfi (1990), not all distributions have convergent results for empirical probability measure with respect to the total variation distance. The consequence of using $\widehat{W}_{\mathcal{H}}(\theta)$-distance is to obtain an $\sqrt{n}$ or $\sqrt{m}$-inconsistent estimator. This could bring serious issues for hypothesis testing or some other weakly convergent outcomes.

In this subsection, we first consider the regular consistency result of $\widehat{W}_{\mathcal{H}}(\theta)$. It serves a baseline criterion for the following analyses. The consistency result also shows that $\widehat{W}_{\mathcal{H}}(\theta)$ is not $\sqrt{n}$ or $\sqrt{m}$-consistent. Then, we propose an asymptotically unbiased criterion by simply taking the square of $\widehat{W}_{\mathcal{H}}(\theta)$. We show that the $\sqrt{n}$ or $\sqrt{m}$-bias terms will be eliminated.

The following theorem establishes that $\widehat{W}_{\mathcal{H}}(\theta)$ is consistent for $W_{\mathcal{H}}(\theta)$ as n or m diverges to infinity. But it is not $\sqrt{n}$ or $\sqrt{m}$-consistent.

Theorem 6. *Given Conditions C1-C12, we can assume $0 \leq \mathbf{k}(q, q') \leq C_{\mathbf{k}}$ for any q and q', then*

$$\Pr\left\{\widehat{W}_{\mathcal{H}}(\theta) - W_{\mathcal{H}}(\theta) > 2(\sqrt{C_{\mathbf{k}}/m} + \sqrt{C_{\mathbf{k}}/n}) + \epsilon\right\} \leq 2\exp\left(\frac{-\epsilon^2 mn}{2C_{\mathbf{k}}(m+n)}\right)$$

where ϵ is an arbitrarily small number.

Proof. See Appendix C.3. □

Theorem 6 shows that $\widehat{W}_{\mathcal{H}}(\theta)$ is not $\sqrt{n}$ or $\sqrt{m}$-consistent for $W_{\mathcal{H}}(\theta)$. The statistics $n^{\frac{1}{2}}\left\{\widehat{W}_{\mathcal{H}}(\theta) - W_{\mathcal{H}}(\theta)\right\}$ or $m^{\frac{1}{2}}\left\{\widehat{W}_{\mathcal{H}}(\theta) - W_{\mathcal{H}}(\theta)\right\}$ has a bias term $C_{\mathbf{k}}$ that does not vanish even if $\epsilon \to 0$ and $n, m \to \infty$. It also provides an insight that the bias $(C_{\mathbf{k}}/m)^{\frac{1}{2}} + (C_{\mathbf{k}}/n)^{\frac{1}{2}}$ is related to the kernel size.

Now we propose the square of $W_{\mathcal{H}}(\theta)$-distance as the new criterion distance and will show that this distance leads to an asymptotically unbiased estimate. As the proof in Lemma 1, the square of $W_{\mathcal{H}}(\theta)$-distance is:

$$S_{\mathcal{H}}(\theta) := [W_{\mathcal{H}}(\theta)]^2 = \|H_\theta \mathbf{k} - P_\theta \mathbf{k}\|_{\mathcal{H}}^2 = \langle H_\theta \mathbf{k} - P_\theta \mathbf{k}, \ H_\theta \mathbf{k} - P_\theta \mathbf{k} \rangle$$
$$= \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') d(H_\theta - P_\theta)(q) d(H_\theta - P_\theta)(q').$$

Given the norm $\|H_\theta \mathbf{k} - P_\theta \mathbf{k}\|_{\mathcal{H}}^2$, $S_{\mathcal{H}}(\theta)$ may be easily computed in terms of kernel functions.

This leads to our empirical estimate of the $S_{\mathcal{H}}(\theta)$:

$$\widehat{S}_{\mathcal{H}}(\theta) := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} h(q_i, q_j), \tag{13}$$

when $m = n$ in (10) and $h(\cdot, \cdot)$ is a U-statistic such that $h(q_i, q_j) := \mathbf{k}(x_i, x_j) + \mathbf{k}(\varepsilon_i, \varepsilon_j) - \mathbf{k}(x_i, \varepsilon_j) - \mathbf{k}(x_j, \varepsilon_i)$. In general, if $m \neq n$ in (10), then we have

$$\widehat{S}_{\mathcal{H}}(\theta) := \left[\frac{1}{m^2} \sum_{i \neq j}^{m} \mathbf{k}(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} \mathbf{k}(x_i, \varepsilon_j) + \frac{1}{n^2} \sum_{i \neq j}^{n} \mathbf{k}(\varepsilon_i, \varepsilon_j) \right], \tag{14}$$

The consistent statistical distance base on the following theorem, which is a straightforward application of the large deviation bound on U-statistics of the Hoeffding bound. The next theorem shows consistency of the estimator θ^*.

Theorem 7. *Given C1-C12, we can assume $0 \leq \mathbf{k}(q, q') \leq C_{\mathbf{k}}$, from which it follows $-2C_{\mathbf{k}} \leq h(q, q') \leq 2C_{\mathbf{k}}$ for any q and q'. Define θ^* as*

$$\widehat{h}(\theta^*|\Theta) = \inf_{\theta \in \Theta} \frac{1}{n^2} \sum_{i \neq j}^{n} h(q_i, q_j), \ s.t. \ q \sim H_{n\theta}, q' \sim P_{n\theta}.$$

Assume that for each $\theta \neq \theta_0$, there is an open set B_θ such that $\theta \in B_\theta$ and $h(\theta|B_\theta) > 0$

$$h(\theta^*|B_\theta) = \inf_{\theta \in B_\theta} \int h(q, q') dH_\theta(q) dP_\theta(q'), \ s.t. \ q \sim H_\theta, q' \sim P_\theta.$$

If Θ is not compact, assume further that there exists compact $\mathcal{K} \subseteq \Theta$ such that $\theta_0 \in \mathcal{K}$ and $h(\theta|\mathcal{K}^c) > 0$. Then,

$$\Pr\left\{ \limsup_{n \to \infty} |\theta^* - \theta_0| \geq \epsilon \right\} = 0. \tag{15}$$

Proof. See Appendix C.4. □

4.4 Asymptotic Distribution of the Unbiased Statistic

The $\sqrt{n}$-inconsistency of the baseline criterion $\widehat{W}_{\mathcal{H}}(\theta)$ motivates us to consider another statistic, $\widehat{S}_{\mathcal{H}}(\theta)$. Root-$n$ asymptotic bias is eliminated by taking a quadratic form of $\widehat{W}_{\mathcal{H}}(\theta)$. Then, the asymptotic unbiased estimator is valid for hypothesis testing since it induces an unbiased statistic. This unbiased statistic has a standard Gaussian limiting distribution with root-n

rate. It is a nonparametric test statistic and therefore its asymptotic properties may fit a more general framework beyond the current setting.

The following two results derive the limiting distribution of $\widehat{S}_{\mathcal{H}}(\theta^*)$, where θ^* is given in (12).

Theorem 8. *The statistic* $\widehat{S}_{\mathcal{H}}(\theta^*)$ *converges in distribution to*

$$n\widehat{S}_{\mathcal{H}}(\theta^*) \rightsquigarrow \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 1 \right]$$

where $z_l \rightsquigarrow \mathcal{N}(0,1)$ *and* z_l *are i.i.d.,* $\lambda_l = (\lambda_l^{(1)} + \lambda_l^{(2)})$ *are the solutions to the eigenvalue equations*

$$\int \tilde{\mathbf{k}}(x, \cdot)\psi_l(x)dF_{\theta_0}(x) = \lambda_l^{(1)}\psi_l(\cdot)$$

$$\int \tilde{\mathbf{k}}(\varepsilon, \cdot)\psi_l(\varepsilon)dG_{\theta_0}(\varepsilon) = \lambda_l^{(2)}\psi_l(\cdot)$$

where $\psi_l(\cdot)$ *is the eigenfunction and*

$$\tilde{\mathbf{k}}(x_i, x_j) := \mathbf{k}(x_i, x_j) - \mathbb{E}_x\mathbf{k}(x_i, x) - \mathbb{E}_x\mathbf{k}(x, x_j) + \mathbb{E}_{x,x'}\mathbf{k}(x, x')$$

is the centered kernel. In particular, $h(q_i, q_j)$ *in* (13) *is* $\tilde{\mathbf{k}}(x_i, x_j) + \tilde{\mathbf{k}}(\varepsilon_i, \varepsilon_j) - \tilde{\mathbf{k}}(x_i, \varepsilon_j) - \tilde{\mathbf{k}}(\varepsilon_i, x_j)$.

Proof. See Appendix C.5. □

Theorem 9. *Let* θ^* *be the value which gives the minimum distance of* $\widehat{S}_{\mathcal{H}}(\theta)$. *It follows that* $\widehat{S}_{\mathcal{H}}(\theta^*)$ *converges in distribution to a Gaussian random variable such that*

$$\sqrt{n}\left(\widehat{S}_{\mathcal{H}}(\theta^*) - S_{\mathcal{H}}(\theta_0)\right) \rightsquigarrow \mathcal{N}(0, \sigma_s^2),$$

where $\sigma_s^2 := 4\left(\mathbb{E}_q\left[\mathbb{E}_{q'}h(q, q')\right]^2 - [\mathbb{E}_{q,q'}h(q, q')]^2\right)$.

Proof. With Theorem 8, the proof follows directly from the result in Serfling (p. 193 Theorem B, 1980). □

4.5 Optimal Information

In this section we show the proposed estimator is more efficient than that in Brown and Wegkamp (2002) (BW). To derive an information matrix of θ^* comparable with that of BW, we need additional assumptions:

C13. The integrated minimum distance criterion function has a positive definite second derivative matrix $\mathbf{V}$ at θ_0.

C14'. Define $D_\theta(q) \equiv H_\theta(q) - P_\theta(q)$ is differentiable at θ_0 in $L_2(\mu)$ where μ is an underlying probability measure for $q = (x, \varepsilon)$. Assume that

$$\lim_{\|\theta - \theta_0\| \to 0} \int \left(\frac{D_\theta(q) - (\theta - \theta_0)^T \Delta(q)}{\|\theta - \theta_0\|} \right)^2 d\mu(q) = 0, \tag{16}$$

where Δ is the differentiable mean of $D_\theta(q)$ in $L_2(\mu)$.

Condition C13 refers to the condition A.7 in BW and it is about the second derivative matrix of the criterion function $M(\theta)$. In our setting, the integrated criterion function in C13 is $\widehat{W}_{\mathcal{H}}(\theta)$ which always has second derivative if the kernel function is smooth and second order differentiable. C14' requires the existence of differentiable mean $\Delta(q)$ in $L_2(\mu)$ for $D_\theta(q)$. The differentiable mean $\Delta(q)$ is essential to the information contained in BW's estimator. The limiting covariance matrix of $n^{-1/2}(\hat{\theta}_{BW} - \theta)$ in BW is $4\mathbf{V}^{-1}\mathbf{W}\mathbf{V}^{-1}$, where $\mathbf{V}$ is the second derivative matrix of $M(\theta)$ and $\mathbf{W}$ is

$$\mathbf{W} = \int \int \Delta(q) \Delta(q') \times$$
$$\underbrace{[F(x)F(x')G(\varepsilon_{min}) + F(x_{min})G(\varepsilon)G(\varepsilon') + H(x_{min}, \varepsilon_{min}) - 3H(q)H(q')]}_{(*)} d\mu(q) d\mu(q'),$$

and $\varepsilon_{min} := \varepsilon \wedge \varepsilon'$, $x_{min} := x \wedge x'$. The $(*)$ term in $\mathbf{W}$ is a weight for the product measure $d(\mu(q) \times \mu(q'))$.

We show that if there is a unique true μ in C14' then assumption C14' is equivalent to the minimum optimal measure γ in our setting. The idea is that $\mathbf{W}$ comes from a simplification of the following covariance term:

$$\lim_{\|\theta - \theta_0\| \to 0} \left(\int \Delta(q) D_\theta(q) d\mu(q) \right) \left(\int \Delta(q') D_\theta(q') d\mu(q') \right)^T,$$

which is equivalent (in the weak topological sense) to $\lim_{\|\theta-\theta_0\|\to 0} W_{\mathcal{H}}(\theta)$. The optimal measure corresponds to the most efficient information. We modify condition C14' in order to emphasize that endogenous effect of the measure μ in simultaneous equations.

C14. $D_\theta(q)$ has a differentiable quadratic mean (DQM) of $\Delta(q)$ at θ_0 such that

$$\lim_{\|\theta-\theta_0\|\to 0} \int \left(\frac{\sqrt{h_\theta(q)} - \sqrt{p_\theta(q)} - (\theta-\theta_0)^T \Delta_w(q) \mathbf{k}^{-1/2}(q)}{\|\theta-\theta_0\|} \right)^2 w_\theta(q) dq = 0, \qquad (17)$$

where $0 < \mathbf{k}(q) < 1$ is the kernel on $\mathcal{Q}$ such that $\mathbf{k}(q) = \langle \Phi(\cdot), \Phi(q) \rangle$ such that $Y = f(q) = \langle \alpha, \Phi(q) \rangle$ for some coefficient α and the feature map $\Phi(\cdot)$. The endogenous weight is embedded in $\mathcal{H}$ such that $w_\theta \in \mathcal{H}$ for any θ and at θ_0,

$$w_{\theta_0}^{1/2}(q) dq = d\left(\sqrt{H_{\theta_0}} - \sqrt{P_{\theta_0}}\right)(q).$$

C14 has two differences with C14'. First, we consider the DQM of $\sqrt{dD_\theta(q)}$ in C14 instead of the differentiable mean of $D_\theta(q)$. Second, DQM of $\sqrt{dD_\theta(q)}$ is associated with an endogenous weight $w_\theta(q)$. Assume that the simultaneous model $\rho(X, Y, \theta) = \varepsilon$ satisfies C1-C9. The idea of making these differences is to attain a better representation for the derivative[6]:

$$\Delta_w(q) \approx (\sqrt{dD_\theta(q)w_\theta(q)})' = \frac{1}{2\sqrt{D_\theta(q)w_\theta(q)}}(D_\theta(q)w_\theta(q))'$$

so that $(\sqrt{D_\theta(q)w_\theta(q)})'\big|_{\theta=\theta_0} \times w_{\theta_0}^{-1/2} = (\log D_\theta(q)w_\theta(q))'|_{\theta=\theta_0} = \iota_{\theta_0}(q)$ which is the information score.

The information form in a minimum Hellinger distance with DQM condition has the following result:

$$\lim_{\|\theta-\theta_0\|\to 0} \mathbf{d}_{\mathbf{H}}^2(H_\theta, FG_\theta) = \lim_{\|\theta-\theta_0\|\to 0} \frac{1}{2} \int \left(\sqrt{h_\theta(q)w_\theta(q)} - \sqrt{p_\theta(q)w_\theta(q)} \right)^2 dq$$
$$\approx \frac{1}{2}(\theta-\theta_0)^T \mathcal{I}(\theta_0)(\theta-\theta_0)^T$$

where $\mathcal{I}(\theta_0)$ is the Fisher type (or BW type) information matrix. The following theorem shows that estimator based on $W_{\mathcal{H}}(\theta)$ or $S_{\mathcal{H}}(\theta)$-distance share the same form.

---

[6] The square root of density may not exist in general. Aussming DQM condition is to induce a score involving the endogenous term w_θ that is comparable to the one in BW.

Theorem 10. *Given C1-C14, $S_{\mathcal{H}}(\theta)$ can be expressed as*

$$S_{\mathcal{H}}(\theta) = \max_{f \in \mathcal{F}} \int |h_\theta(q) - p_\theta(q)|^2 f(q) dq.$$

Let $S_{\mathcal{H}}(\theta)$ attain its minimum at $(\theta^*, \mathbf{k})$. The score of $S_{\mathcal{H}}(\theta)$ at θ_0 is

$$\int \Delta_w(q) \left(\frac{w_{\theta_0}}{\mathbf{k}} \right)^{\frac{1}{2}} d(H_{\theta_0} - P_{\theta_0})(q).$$

The covariance matrix

$$\left[\int \left(\Delta_w(q) \left(\frac{w_\theta}{w_{\theta^*}} \right)^{\frac{1}{2}} (q) \right) d(H_\theta - P_\theta)(q') \right] \left[\int \left(\Delta_w(q) \left(\frac{w_{\theta_0}}{w_{\theta^*}} \right)^{\frac{1}{2}} (q') \right) d(H_\theta - P_\theta)(q') \right]^T$$

which is optimal at $w_{\theta^*} = \mathbf{k}$.

Proof. See Appendix C.6 $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Remark. The densities $h_\theta(q)$ and $p_\theta(q)$ may not exist in the previous general inference results. However, in order to distinguish the optimal measure from the underlying measure μ in BW, we modify C14' (a condition on probability measures H_θ and P_θ) to C14 (a condition on their densities). Thus, instead of using $d\mu$, we use $w_\theta(q)dq$ to emphasize the important role of selecting the optimal measure.

5 COMPUTATION AND NUMERICAL EXPERIMENT

5.1 COMPUTATION OF THE ESTIMATOR

Theorems 6 and 7 induce the equivalence between $\widehat{W}_{\mathcal{H}}(\theta)$ and $\widehat{S}_{\mathcal{H}}(\theta)$. Thus the estimator θ^* minimizes the empirical $\widehat{W}_{\mathcal{H}}(\theta)$-distance or $\widehat{S}_{\mathcal{H}}(\theta)$, the square of $\widehat{W}_{\mathcal{H}}(\theta)$-distance. By minimizing the distance function, the statistic $\widehat{S}_{\mathcal{H}}(\theta^*)$ gives a regular consistent solution to the dual problem in (10):

$$\theta^* = \arg\min_{\theta \in \Theta} \widehat{W}(\theta) := \arg\min_{\theta \in \Theta} \sup_{\|f\|_l \leq 1} \left| \frac{1}{m} \sum_{i=1}^m f(q_i) - \frac{1}{n} \sum_{j=1}^n f(q'_j) \right|,$$

$$= \arg\min_{\theta \in \Theta} \widehat{S}_{\mathcal{H}}(\theta) := \arg\min_{\theta \in \Theta} \left[\frac{1}{m^2} \sum_{i \neq j}^m \mathbf{k}(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} \mathbf{k}(x_i, \varepsilon_j) + \frac{1}{n^2} \sum_{i \neq j}^n \mathbf{k}(\varepsilon_i, \varepsilon_j) \right],$$

s.t. $\quad (x_i, \varepsilon_i) = q_i \sim H_{n\theta}, \qquad (x_j, \varepsilon_j) = q'_j \sim P_{n\theta}.$

The simplest form of implementing this estimate is to use the grid search. One can evaluate θ at different locations and then compute $\widehat{S}_{\mathcal{H}}(\theta)$ thorough the kernel representation. This idea is feasible for one dimension case. However, due to the fact that the simultaneous equation problem in (10) has at least two variables (x_i, ε_i), grid search is computational demanding.

Based on the duality, we propose a kernel-based Kantorovitch formulation. The formulation will induce a Linear Programming (LP) problem for the dual representation of

$$C_{i,j} := [\mathbf{k}(x_i, x_j) - \mathbf{k}(x_i, \varepsilon_j)] - [\mathbf{k}(x_i, \varepsilon_j) - \mathbf{k}(\varepsilon_i, \varepsilon_j)] \geq 0.$$

Let $\gamma_{i,j}(\theta)$ be the empirical counterpart of the transportation measure $\gamma \in \Gamma(H_\theta, P_\theta)$ in (7). Then $dH_{n\theta} = \sum_{j=1}^m \gamma_{i,j}(\theta)$ and $dP_{n\theta} = \sum_{i=1}^n \gamma_{i,j}(\theta)$. The kernel-based Kantorovitch formulation of (7) is

$$\gamma_{i,j}(\theta^*) \in \arg \min_{\gamma_{i,j}(\theta) \in \Gamma(H_{n\theta}, P_{n\theta})} \sum_{i,j} \gamma_{i,j}(\theta) C_{i,j}. \tag{18}$$

Note that (18) is a standard LP problem. Every LP has a dual. The neatest way to write the dual of (18) is:

$$\max_{u_i, v_j} \sum_i u_i + \sum_j v_j \quad \text{s.t. } u_i + v_j \geq C_{i,j} \tag{19}$$

where u_i and v_j are the Lagrange multipliers of the constraints

$$dH_{n\theta}(x_i, \varepsilon_i) = \sum_{j=1}^m \gamma_{i,j}(\theta), \qquad dP_{n\theta}(x_j, \varepsilon_j) = \sum_{i=1}^n \gamma_{i,j}(\theta),$$

where the measures are sums of Diracs.

The dual problem (19) can be solved by the simplex method which is feasible for two or higher dimensional problems. It means that we could set up a simultaneous equation problem and solve it using the simplex method as with any LP problem. The duality between (19) and (18) is given by the following corollary.

Corollary 2. *If* (19) *has a solution* (u, v) *such that* $u_i + v_j = C_{i,j}$, *then this solution induces a solution* $\gamma_{i,j}(\theta^*)$ *in* (18), *and vice versa.*

Proof. See Appendix D.3. □

The optimal $\widehat{S}_{\mathcal{H}}(\theta^*)$-distance is defined as

$$\widehat{S}_{\mathcal{H}}(\theta^*) = \sum_{i,j} \gamma_{i,j}(\theta^*) C_{i,j}$$

$$\text{s.t.} \quad dH_{n\theta^*}(x_i, \varepsilon_i) = \sum_{j=1}^{m} \gamma_{i,j}(\theta^*) \text{ for any } i,$$

$$dP_{n\theta^*}(x_j, \varepsilon_j) = \sum_{i=1}^{n} \gamma_{i,j}(\theta^*) \text{ for any } j.$$

The scheme of computing θ^* is summarized below.

Step 1. Initialization.

Step 2. Generate $H_{n\theta}(x_i, \varepsilon_i)$ and $P_{n\theta}(x_j, \varepsilon_j)$ for given θ. Solve problem (18) via its dual (19). Obtain θ^*.

Step 3. Compute $\widehat{S}_{\mathcal{H}}(\theta^*)$ by the kernel representation. If $\widehat{S}_{\mathcal{H}}(\theta^*) \neq 0$, then go to Step 2, otherwise done.

Sampling for $H_{n\theta}(x_i, \varepsilon_i)$ and $P_{n\theta}(x_j, \varepsilon_j)$ in Step 2 has several options. One can draw the samples $\varepsilon(\theta)$ from $\rho(z, \theta)$ for given z and θ, then obtain the empirical c.d.f. $n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}\{x_i \leq x, \rho(z_j, \theta) \leq \varepsilon\}$ for $H_{n\theta}$ and $n^{-2} \sum_{i=1}^{n} \mathbf{1}\{x_i \leq x\} \sum_{i=1}^{n} \mathbf{1}\{\rho(z_i \theta) \leq \varepsilon\}$ for $P_{n\theta}$. On the other hand, one can resampling z in order to generate more observations for $\varepsilon(\theta)$ from $\rho(z, \theta)$. With a larger (generic) sample size of z and $\varepsilon(\theta)$, one can obtain a smoother empirical c.d.f..

Computation for $\widehat{S}_{\mathcal{H}}(\theta^*)$ in Step 3 needs a LP solver. We use Dikin's method in this solver. To outline the routine in Step 3, first we construct the kernel representation such that $\mathbf{y} = \alpha^T \Phi(\mathbf{x})$ where $\mathbf{k}(\mathbf{x}, \mathbf{x}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle$, then we use $\mathbf{k}$ to obtain $C_{i,j}$, finally we solve the LP given in (19).

5.2 NUMERICAL EXPERIMENT

The following numerical experiment illustrates more details about the computation. We generate a random vector $\mathbf{q} = (\mathbf{x}, \varepsilon)$ which is jointly distributed as $H_{n\theta}$. The product of marginal empirical distribution of $\mathbf{x}$ and ε is $P_{n\theta}$, another empirical c.d.f.. Figure 1 shows two point clouds. The squares explain the distribution of $H_{n\theta}$ and the dots explain the

distribution of $P_{n\theta_0}$ at the value $\theta = \theta_0$. When $\theta = \theta_0 = 1$, the elements $\mathbf{x}$ and ε are distributed independently as two smaller point clouds. The size of each dot (square) is proportional to its probability weight.

The specification of this computation is given as follows. The samples $q_i = (x_{1,i}, x_{2,i}, \varepsilon_{1,i,\theta}, \varepsilon_{2,i,\theta})_{i \leq 140}$ are jointly generated by a bivariate Gaussian distribution with means $(0, -0.5)$ and $(0, 0.25 + \theta)$ for $\mathbf{x}$ and ε respectively. All the variances are set to 1. The correlation between $\mathbf{x_2}$ and $\varepsilon_2(\theta)$ is $1 - \theta$ while zero for the others. When $\theta_0 = 1$, $\mathbf{x}$ and $\varepsilon(\theta_0)$ are independent since the correlation value is zero. For q'_i, $(x_{1,i}, x_{2,i})_{i \leq 150}$ and $(\varepsilon_{1,i,\theta}, \varepsilon_{2,i,\theta})_{i \leq 150}$, the samples are drawn independently from two bivariate Gaussian distributions. For $\mathbf{x}$, the mean and variance are $(0, -0.5)$ and 1. For $\varepsilon(\theta)$, the mean and variance are $(0, 0.25 + \theta)$ and 1. Obviously, these two groups of samples are equivalent under the true θ_0.

[Figure 1 about here.]

The computation of θ can be considered as pushing $\mathbf{q} \sim H_{n\theta}$ (or $\mathbf{q}' \sim P_{n\theta}$) towards a new $\mathbf{q} \sim H_{n\theta_0}$ (or $\mathbf{q}' \sim P_{n\theta_0}$). Under the independence condition, $\mathbf{q}$ is equivalently distributed as $\mathbf{q}'$ for $\theta = \theta_0$. Figure 2 gives an illustration about the transformation procedure. The initial points $\mathbf{q}' \sim P_{n\theta}$ at $\theta = 0$ are pushed towards to $\mathbf{q}' \sim P_{n\theta_0}$ at $\theta = \theta_0 = 1$. The mechanism behind this transformation relies on the transportation measure $\gamma_{i,j}(\theta)$ from the LP problem (18). By solving the LP problem, θ^* is obtained when $\gamma_{i,j}(\theta^*)$ is achieved. Figure 3 shows the matrix plot of the optimal $\gamma_{i,j}(\theta^*)$.

[Figure 2 about here.]

In the experiment, we consider the case $n = 140$, $m = 150$. So $\gamma_{i,j}(\theta)$ is contained in a 150×140 matrix. Figure 3 shows that for optimal $\gamma_{i,j}(\theta^*)$ most of the transitions betweens q_i and q'_j are never happen, namely with probability zero. The transition matrix is very sparse and has few non-zero entities. We also compute the marginal probabilities by summing up $\gamma_{i,j}(\theta^*)$ w.r.t. j and i respectively:

$$dH_{n\theta^*}(x_i, \varepsilon_i) = \sum_{j=1}^{m} \gamma_{i,j}(\theta^*), \qquad dP_{n\theta^*}(x_j, \varepsilon_j) = \sum_{i=1}^{n} \gamma_{i,j}(\theta^*).$$

The smoothed density plots in Figure 4 shows that neither $dH_{n\theta^*}(x_i, \varepsilon_i)$ nor $dP_{n\theta^*}(x_i, \varepsilon_i)$ is close to uniform distribution. Also, the pattern of the transition matrix in Figure 3 shows that

30

the optimal $\gamma_{i,j}(\theta^*)$ is not uniformly distributed. In fact, in Figure 4, the marginal densities seem have mixture patterns and have very irregular shapes. This feature is important. It means that in a simultaneous equation estimation problem:

$$\int \mathbf{d}(H_{n\theta}(x,\varepsilon), P_{n\theta}(x,\varepsilon))d\mu(x,\varepsilon),$$

the optimal choice of μ could be excluded from any regular specified distribution, i.e. uniform, normal, etc. Instead of giving a priori specification of μ, one can endogenize the effect of μ as a transportation measure. By computing the new estimator, one will obtain θ^* and an optimal measure $\gamma_{i,j}(\theta^*)$ for μ simultaneously.

[Figure 3 about here.]

[Figure 4 about here.]

Although the role of $\gamma_{i,j}(\theta^*)$ is a statistical inference mechanism, its meaning is far more than that. In economic theory, $\gamma_{i,j}(\theta^*)$ can induce an optimal transportation in the matching problem. The optimal inference is then an optimal matching mechanism. Nonparametric representation of the optimal distance is applicable for solving some general matching problems. Figure 5 shows the role of $\gamma_{i,j}(\theta^*)$ as pushing the initial points (square) towards to the posterior points (dots). The lines in Figure 5 represent the connections. A line connecting the square (x_i, ε_i) and the dot (x_j', ε_j') means that $\gamma_{i,j}(\theta^*)C_{i,j}$ for these two groups is non-zero.

[Figure 5 about here.]

6   RELATED LIKELIHOOD-BASED INFERENCE

If the model is fully specified, then the c.d.f.s $H_\theta(x,\varepsilon)$ and $P_\theta(x,\varepsilon)$ in (7) are some parametric distributions. The estimation problem becomes:

$$\min_\theta W(\theta) := \min_\theta \inf_{\gamma \in \Gamma(H_\theta, P_\theta)} \left\{ \int \mathbf{d}(q, q')d\gamma(q, q') : \ q \sim H_\theta(x,\varepsilon) \text{ and } q' \sim P_\theta(x,\varepsilon) \right\}.$$

Unlike the empirical optimal distance function $\widehat{W}(\theta)$, the integrable $W(\theta)$ can have densities for $H_\theta(x,\varepsilon)$ and $P_\theta(x,\varepsilon)$. We define $p_\theta(q') = dP_\theta$ to be the density of θ with respect to some Lebesgue measure. In parametric frameworks, we can assume that such a density always

exists. Similarly, the density of H_θ also exists and is denoted by $h_\theta(q)$. The conditional density defines the likelihood function $L(\theta; \mathbf{q}') = \Pr(\mathbf{q}'|\theta)$.

The essence of our approach is to find a map that pushes forward $H_\theta(x, \varepsilon)$ to $P_\theta(x, \varepsilon)$. Recall that q is a random variable distributed according to $H_\theta(x, \varepsilon)$ and that q' is a random variable distributed according to $P_\theta(x, \varepsilon)$. Then we seek a map f that satisfies the following specification:

$$q' = f(q), \quad \text{where } q \sim H_\theta(x, \varepsilon) \text{ and } q' \sim P_\theta(x, \varepsilon).$$

Equivalently, we seek a map f which pushes forward $H_\theta(x, \varepsilon)$ to $P_\theta(x, \varepsilon)$. The map necessarily depends on the data and the form of the likelihood function. If the independence condition is satisfied, an identity map would suffice; otherwise more complicated functions are necessary.

By using the conditional updating argument (Bayes' rule), the posterior is

$$\tilde{p}_\theta(q') = \frac{L(\theta; \mathbf{q}')p_\theta(f(q))}{\int L(\theta; \mathbf{q}')p_\theta(f(q))d\theta}$$

where $p_\theta(f(q))$ emphasizes that $q' = f(q)$. Recall that the transformed density $p_\theta(f(q))$ should be equal to $h_\theta(q)$ at $\theta = \theta_0$ such that $p_\theta(f(q)) = h_\theta(q)$. On the other hand, the optimal transportation argument says that $h_\theta(q)$ can be transformed into the probability density $p_\theta(f(q))$ by the forward push function $f(\cdot)$. Thus we can setup the following argument at $\theta = \theta_0$:

$$h(\theta) = L(\theta; \mathbf{q}')p_\theta(f(q))/c$$

where $c = \int L(\theta; \mathbf{q}')p_\theta(q')d\theta$ is a constant for any θ. This argument defines an alternate criterion for $W(\theta)$.

At $\theta = \theta_0$, $p_\theta(f(q)) = h_\theta(q)$ for any q. This condition means that the ratio $p_\theta(f(q))/h_\theta(q)$ is one or equivalently $\log p_\theta(f(q)) - \log h_\theta(q) = 0$ when $\theta = \theta_0$. Thus, we can consider the following criterion

$$\min_{\theta,f} \sum_{i=1}^{n} \left\{ \log h_\theta(q_i) - \log L(\theta; q_i') - \log p_\theta(f(q_i)) \right\}.$$

Although this criterion function has a simpler representation, it may not as tractable as $W(\theta)$ in practice due to the unknown function f. The transportation function f belongs

to a unknown functional class. Without further restriction of this class, it is infeasible to select the optimal f in the above argument. The following example is to show that with a Gaussian restriction, the push forward function f is attainable.

Example: The Linear Gaussian Case Suppose that the discrepancy between $\mathbf{q}'$ and $\mathbf{q}$ is additive: $\mathbf{q}' = A(\theta) + \mathbf{q}$. In this case, the map would thus depend on the data $\mathbf{q}'$, the forward model $A(\cdot)$, and the distribution of the observational $\mathbf{q}$.

A common further assumption is that $\mathbf{q}$ is Gaussian, zero-mean, and independent of the model parameters, i.e., $\mathbf{q} \sim \mathcal{N}(0, \Sigma_n)$, leading to the following form for the likelihood function:

$$L(\theta; \mathbf{q}') = \frac{1}{(2\pi)^{\frac{m}{2}}(\det\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\|A(\theta) - \mathbf{q}'\|_{\Sigma_n}^2\right)$$

where $\|u\|_{\Sigma} := \|\Sigma^{-\frac{1}{2}}u\|$, for any positive symmetric matrix Σ.

Let $A(\theta) = A\theta$. Because $\mathbf{q}$ is distributed as a Gaussian variable and $A\theta$ is an additive term, $\mathbf{q}'$ has the Gaussian distribution as well. Thus $p_\theta(\mathbf{q}')$ is from Gaussian family. Without loss of generality, we can set $\mathbf{q}' - \mathbf{q} \sim \mathcal{N}(0, \Sigma_M)$. Then the criterion function becomes:

$$\sum_{i=1}^{n}\left\{\log h_\theta(q_i) - \log L(\theta; q_i') - \log p_\theta(f(q_i))\right\}$$

$$\propto \left\{-\frac{1}{2}\|A\theta - \mathbf{q}'\|_{\Sigma_n}^2 + \|\theta\|_{\Sigma_M}^2\right\}$$

$$\propto \left\{-\frac{1}{2}\left(\|\theta\|_{\Sigma_n}^2 + \theta^T(A^T\Sigma_n^{-1}A + \Sigma_M^{-1})\theta - 2(\mathbf{q}')^T\Sigma_n^{-1}A\theta\right)\right\}$$

$$\propto \left\{-\frac{1}{2}\left(\|\theta\|_{\Sigma_n}^2 + \theta^T\Sigma\theta - 2\mu^T\Sigma^{-1}\theta\right)\right\}$$

where $\Sigma = (A^T\Sigma_n^{-1}A + \Sigma_M^{-1})^{-1}$ is the updated covariance matrix and $\mu = \Sigma A^T\Sigma_n^{-1}\mathbf{q}'$ is the updated mean. Then the optimal transportation mechanism is equivalent to the optimal information updating for a posterior.

7   Conclusion

In this paper we considered estimation and inference in semiparametric simultaneous equation models where the observable variables are separable of the unobserved errors. We consider estimation using a model based on the independence condition that induces a minimum

distance between a joint cumulative distribution function and the product of its marginal distribution functions. We developed a new minimum integrated distance estimator which generalizes BW approach by using Kantorovich's formulation of the optimal transportation problem. This generalization is important because it allows for estimating the weighted measure nonparametrically, and hence does not impose a priori structure on the weighting measure. Moreover, the proposed methods do not require knowledge of the reduced form of the model. The estimation also provides greater generality by dealing with probability measures on compact metric spaces without assuming existence of densities. We establish the asymptotic properties of the estimator, and show that the asymptotic statistics of empirical estimates have standard convergent results and are provided for different statistical analyses. The proposed estimator is more efficient than its competitors. In addition, we provided a tractable implementation for computing the estimator in practice.

References

Benkard, C. L., and S. Berry (2006): "On the Nonparametric Identification of Non-Linear Simultaneous Equations Models: Comment on Brown (1983) and Rhoerig (1988)," *Econometrica*, 74, 1429–1440. 1, 1, 2.1, 2.2

Brown, B. W. (1983): "The Identification Problem in Systems Nonlinear in the Variables," *Econometrica*, 51, 175–196. 1

Brown, D. J., R. Deb, and M. H. Wegkamp (2008): "Tests of Independence in Separable Econometric Models: Theory and Application," Unpublished. 1

Brown, D. J., and R. L. Matzkin (1998): "Estimation of Nonparametric Functions in Simultaneous Equation Models, with an Application to Consumer Demand," Unpublished. 1, 3

Brown, D. J., and M. H. Wegkamp (2001): "Weighted Minimum Mean-Square Distance from Independence Estimation," Cowles Foundation Discussion Paper No. 1288. 3.1

Brown, D. J., and M. H. Wegkamp (2002): "Weighted Minimum Mean-Square Distance from Independence Estimation," *Econometrica*, 70(5), 2035–2051. 1, 1, 2.1, 4.5

Chiappori, P.-A., R. J. McCann, and L. P. Nesheim (2010): "Hedonic Price Equilibria, Stable Matching and Optimal Transport: Equivalence, Topology, and Uniquness," *Economic Theory*, 42(2), 317–354. 1

Devroye, L., and L. Gyorfi (1990): "No Empirical Probability Measure can Converge in the Total Variation Sense for All Distributions," *Annals of Statistics*, 18(3), 1496–1499. 4.3

Dudley, R. M. (2002): *Real Analysis and Probability*. Cambridge University Press. 3.2, 5, 3.3, C.2, D.2

Ekeland, I., A. Galichon, and M. Henry (2010): "Optimal Transportation and the Falsifiability of Incompletely Specified Economic Models," *Economic Theory*, 42(2), 355–374. 1

GALICHON, A., I. FERNANDEZ-VAL, AND V. CHERNOZHUKOV (2010): "Quantile and Probability Curves without Crossing," *Econometrica*, 78(3), 1093–1125. 1

GALICHON, A., AND B. SALANIE (2010): "Matching with Trade-offs: Matching with Trade-offs: Revealed Preferences over Competing Characteristics," Working Paper. 1

GOZLAN, N., AND C. LEONARD (2010): "Transport Inequalities. A Survey," *Markov Processes and Related Fields*, 16, 635–736. 7

GRAF, S., AND H. LUSCHGY (2000): *Foundations of Quantization for Probability Distributions*. Springer-Verlag, New York. 1

HOEFFDING, W. (1963): "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58, 13–30. C.4

KOMUNJER, I., AND A. SANTOS (2010): "Semi-parametric Estimation of Non-separable Models: a Minimum Distance from Independence Approach," *Econometrics Journal*, 13(3), S28–S55. 1

LINTON, O., S. SPERLICH, AND I. VAN KEILEGOM (2008): "Estimation of a Semiparametric Transformation Model," *Annals of Statistics*, 36(5), 686–718. 1

MANSKI, C. (1983): "Closest Empirical Distribution Estimation," *Econometrica*, 51(2), 305–320. 1, 1, 2.1

MATZKIN, R. L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375. 1

MULLER, A. (1997): "Integral Probability Metrics and Their Generating Classes of Functions," *Advances in Applied Probability*, 29(2), 429–443. 3.3

REED, M., AND B. SIMON (1980): *Methods of Modern Mathematical Physics: Functional Analysis I*. Academic Press, New York. C.1

ROEHRIG, C. S. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56, 433–447. 1

RUBINSHTEIN, G. (1970): "Duality in Mathematical Programming and Certain Problems in Convex Analysis," *Russian Mathematical Survey*, 25(5), 171–201. 3.3

SANTOS, A. (2011): "Semiparametric Estimation of Invertible Models," Working Paper. 1

SERFLING, R. (1980): *Approximation Theorems of Mathematical Statistics*. Wiley, New York. 4.4, C.5

SHORACK, G. R. (2000): *Probability for Statisticians*. Springer-Verlag. 3

TALAGRAND, M. (1991): "A New Isoperimetric Inequality for Product Measure and the Concentration of Measure Phenomenon," in *Lecture Notes in Mathematics*, pp. 94–124. Springer-Verlag. 1

VAN DER VAART, A., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag. C.3

VILLANI, C. (2003): *Topics in Optimal Transportation*, vol. Graduates Studies in Mathematics Series. American Mathematical Society. 3.3

——— (2009): *Optimal Transport: Old and New*, Grundlehren der Mathematischen Wissenschaften. Springer. 1, 2, 3

A   Mathematical Preliminary

Definition. *(Commutative algebra)* A *commutative algebra* $\mathcal{A}$ is a vector space equipped with an additional associative $+ : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ and commutative multiplication $\cdot : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ such that

$$x \cdot (y + z) = x \cdot y + x \cdot z, \quad \lambda(x \cdot y) = (\lambda x) \cdot y$$

holds for all x, y, $z \in \mathcal{A}$ and $\lambda \in \mathbb{R}$. An important algebra is the space $C(\mathcal{X})$ of *all continuous functions* $f : \mathcal{X} \to \mathbb{R}$ on the compact metric space $(\mathcal{X}, \mathbf{d})$ endowed with the *supremum norm*

$$\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|.$$

The following approximation theorem states that certain sub-algebras of $C(\mathcal{X})$ generate the whole space. This is the starting point of representing our dual distance $\widehat{W}(\theta)$.

Theorem 11. *(Stone-Weierstrass Theorem) Let $(\mathcal{X}, d)$ be a compact metric space and $\mathcal{A} \subset C(\mathcal{X})$ be an algebra. Then $\mathcal{A}$ is dense in $C(\mathcal{X})$ if the following conditions hold.*

1. $\mathcal{A}$ does not vanish such that there is no $f(x) = 0$ for any $x \in \mathcal{X}$.

2. $\mathcal{A}$ separates points such that for all $x, y \in \mathcal{X}$ with $x \neq y$ there exists an $f \in \mathcal{A}$ with $f(x) \neq f(y)$.

Definition. *(Feature maps and Feature Spaces)* A *kernel* on $(\mathcal{X}, d)$ is a function $\mathbf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. There exists a Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ with

$$\mathbf{k}(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

for all $x, y \in \mathcal{X}$. Φ is called a *feature map* and $\mathcal{H}$ is called a *feature space* of $\mathbf{k}$. Note that both $\mathcal{H}$ and Φ are far from being unique.[7] A function $f : \mathcal{X} \to \mathbb{R}$ is induced by $\mathbf{k}$ if there exists an element $y \in \mathcal{H}$ such that $f = \langle y, \Phi(\cdot) \rangle$. The definition of this inner product is independent of Φ and $\mathcal{H}$.

Theorem 12. *(Moore-Aronszajn) Let $(\mathcal{X}, \mathbf{d})$ be a compact metric space and $\mathbf{k}$ is a kernel on $\mathcal{X}$. Suppose $\mathbf{k}$ induces a dense set of $C(\mathcal{X})$. Then for all compact and mutually disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_n \subset \mathcal{X}$, all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and any $\epsilon > 0$, there exists a function g induced by $\mathbf{k}$ with $\|g\|_\infty \leq \max_i |\alpha_i| + \epsilon$ such that*

$$\left\| g_{|\cup_{i=1}^n \mathcal{X}_i} - \sum_{i=1}^n \alpha_i \mathbf{1}\{\mathcal{X}_i\} \right\|_\infty \leq \epsilon,$$

---

[7] However, for a given kernel, there exists a canonical feature space which is called the reproducing kernel Hilbert space (RKHS).

where $g_{|\mathcal{X}}$ denotes the restriction of g to $\mathcal{X}$.

B   Probabilistic Preliminary

Definition. *(Weak topology, weak convergence) Weak topology* on a space of probability measures $\mathscr{P} := \{\mathcal{P} : \mathcal{P}$ is a probability measure$\}$ is the weakest topology such that the map $\mathcal{P} \to \int_{\mathcal{Q}} f d\mathcal{P}$ is continuous for all $\mathcal{P} \in C_b(\mathcal{Q})$ where $C_b(\mathcal{Q})$ is the *space of all bounded continuous functions.* A sequence of $\mathcal{P}_n$ is said to *converge weakly* to $\mathcal{P}$, $\mathcal{P}_n \rightsquigarrow \mathcal{P}$, if and only if $\int_{\mathcal{Q}} f d\mathcal{P}_n \to \int_{\mathcal{Q}} f d\mathcal{P}$ for every $f \in C_b(\mathcal{Q})$.

A metric $\mathbf{d}(\cdot, \cdot)$ on $\mathscr{P}$ is said to *metrize* the weak topology if the topology induced by $\mathbf{d}(\cdot, \cdot)$ coincides with the weak topology, which is defined as follows: If $\{\mathcal{P}_1, \mathcal{P}_2, \dots\} \in \mathscr{P}$ and

$$\mathcal{P}_n \rightsquigarrow \mathcal{P} \text{ is equivalent to } \mathbf{d}(\mathcal{P}_n, \mathcal{P}) \to 0 \text{ as } n \to \infty,$$

then the topology induced by $\mathbf{d}(\cdot, \cdot)$ coincides with the weak topology.

Definition. *(Compact metric space)* We denote the *closed ball* with radius r and center x by $B_{\mathbf{d}}(\mathcal{P}, r) := \{\mathcal{P}' \in \mathscr{P} : \mathbf{d}(\mathcal{P}, \mathcal{P}') \leq r\}$. The *covering numbers* of $\mathscr{P}$ are defined by

$$\mathfrak{N}((\mathscr{P}, \mathbf{d}), r) := \min\left\{n \in \mathbb{N} \cup \infty : \exists \mathcal{P}_1, \dots, \mathcal{P}_n \text{ with } \mathscr{P} \subset \cup_{i=1}^{n} B_{\mathbf{d}}(\mathcal{P}_i, r)\right\}$$

for all $r > 0$. The space $(\mathscr{P}, \mathbf{d})$ is *pre-compact* if and only if $\mathfrak{N}((\mathscr{P}, \mathbf{d}), r)$ is *finite* for all $r > 0$. If the space $(\mathscr{P}, \mathbf{d})$ is *complete*, then $\mathscr{P}$ is *compact* if and only if $\mathscr{P}$ is *pre-compact*.

Based on this argument, throughout the paper, we only consider pre-compact space $(\mathscr{P}, \mathbf{d})$ over Hilbert spaces (complete spaces). In other words, the covering numbers of $\mathscr{P}$ are finite and distance $\mathbf{d}(\cdot, \cdot)$ metrizes $\mathscr{P}$ by a Hilbertian type metric. The space $(\mathscr{P}, \mathbf{d})$ is a *compact metric space.*

C   Proof of Main Theorems

C.1   Proof of Theorem 3

Proof. Consider the family

$$\mathscr{P}(\mathcal{Q}) := \left\{\mathcal{P} : \int_{\mathcal{Q}} \sqrt{\mathbf{k}(q,q)} d\mathcal{P}(q) < \infty\right\}.$$

Let $T : \mathcal{H} \to \mathbb{R}$ be a linear functional defined as $T[h] := \int_{\mathcal{Q}} h(q) d\mathcal{P}(q)$ with

$$\|T\| = \sup_{h \in \mathcal{H}, h \neq 0} \frac{|T[h]|}{\|h\|_{\mathcal{H}}}.$$

Now we have

$$|T[h]| = \left| \int_{\mathcal{Q}} h d\mathcal{P}(q) \right| \leq \int_{\mathcal{Q}} |h| d\mathcal{P}(q) = \int_{\mathcal{Q}} | \langle h, \mathbf{k}(\cdot, q) \rangle_{\mathcal{H}} | d\mathcal{P}(q)$$

$$\leq \int_{\mathcal{Q}} \sqrt{\mathbf{k}(q, q)} \|h\|_{\mathcal{H}} d\mathcal{P}(q).$$

It implies $\|T\| < \infty$, for any $\mathcal{P} \in \mathscr{P}(\mathcal{Q})$. Then we know T is a bounded linear functional on $\mathcal{H}$. By Riesz representation theorem (Theorem II.4, Reed and Simon, 1980), for each $\mathcal{P} \in \mathbb{P}$, there exists a unique $\lambda_{\mathcal{P}} \in \mathcal{H}$ such that $T[h] = \langle h, \lambda_{\mathcal{P}} \rangle_{\mathcal{H}}$, for any $h \in \mathcal{H}$.

Let $h = \mathbf{k}(\cdot, u) \in \mathcal{B}_{\mathcal{H}}$ for some $u \in \mathcal{Q}$. Then

$$T[\mathbf{k}(\cdot, u)] = \mathcal{P}\mathbf{k} = \langle \mathbf{k}(\cdot, u), \lambda_{\mathcal{P}} \rangle_{\mathcal{H}} = \lambda_{\mathcal{P}}(u).$$

Replace $H_\theta(x, \varepsilon)$ and $P_\theta(x, \varepsilon)$ with arbitrary $\mathfrak{P} \in \mathcal{P}$ in the expression, then we have

$$\left| \int_{\mathcal{Q}} h dH_\theta(q) - \int_{\mathcal{Q}} h dP_\theta(q) \right| = \left| \langle h, \lambda_{H_\theta} \rangle_{\mathcal{H}} - \langle h, \lambda_{P_\theta} \rangle_{\mathcal{H}} \right| = \left| \langle h, \lambda_{H_\theta} - \lambda_{P_\theta} \rangle_{\mathcal{H}} \right|.$$

This implies $W_{\mathcal{H}}(\theta)$ define a norm

$$W_{\mathcal{H}}(\theta) := \sup_{\|h\|_{\mathcal{H}} \leq 1} \left| \int_{\mathcal{Q}} h dH_\theta(q) - \int_{\mathcal{Q}} h dP_\theta(q) \right|$$

$$= \sup_{\mathbf{k} \in \mathcal{B}_{\mathcal{H}}} \left| \langle \mathbf{k}(\cdot, u), \lambda_{H_\theta} \rangle_{\mathcal{H}} - \langle \mathbf{k}(\cdot, u), \lambda_{P_\theta} \rangle_{\mathcal{H}} \right|$$

$$= \sup_u |\lambda_{H_\theta}(u) - \lambda_{P_\theta}(u)| = \|\lambda_{H_\theta} - \lambda_{P_\theta}\|_{\mathcal{H}} = \|H_\theta \mathbf{k} - P_\theta \mathbf{k}\|_{\mathcal{H}}$$

The second equality uses the fact that any $\{h : \|h\|_{\mathcal{H}} \leq 1\}$ can be represented by $\mathbf{k} \in \mathcal{B}_{\mathcal{H}}$ $\qquad \square$

C.2   PROOF OF THEOREM 4

Proof. When $(\mathcal{Q}, \rho)$ is separable, we replace the general distance function $\mathbf{d}$ in (7) with $\mathbf{d_k}$, then $W(\theta)$ has the following form (p.420, Dudley, 2002)

$$W(\theta) := \inf_{\gamma \in \Gamma(H_\theta, P_\theta)} \left\{ \int \mathbf{d_k}(q, q') d\gamma(q, q') : q \sim H_\theta(x, \varepsilon) \text{ and } q' \sim P_\theta(x, \varepsilon) \right\}.$$

Condition C10 means that $(\mathcal{Q}, \rho)$ is bounded.

Lower bound: For any $\gamma \in \Gamma(H_\theta, P_\theta)$, we have

$$\left| \int_{\mathcal{Q}} f d(H_\theta - P_\theta) \right| = \left| \int_{\mathcal{Q} \times \mathcal{Q}} \left(f(q) - f(q') \right) d\gamma(q, q') \right|$$

$$\leq \int_{\mathcal{Q} \times \mathcal{Q}} \left| f(q) - f(q') \right| d\gamma(q, q')$$

$$= \int_{\mathcal{Q} \times \mathcal{Q}} \left| \langle f, \mathbf{k}(\cdot, q) - \mathbf{k}(\cdot, q') \rangle_{\mathcal{H}} \right| d\gamma(q, q')$$

$$\leq \|f\|_{\mathcal{H}} \int_{\mathcal{Q} \times \mathcal{Q}} \left\| \mathbf{k}(\cdot, q) - \mathbf{k}(\cdot, q') \right\|_{\mathcal{H}} d\gamma(q, q').$$

The first equality uses the property of marginal probability of $\gamma(q, q')$. After taking supremum over $f \in \mathcal{B}_{\mathcal{H}}$, we have

$$W_{\mathcal{H}}(\theta) = \sup_{f \in \mathcal{B}_{\mathcal{H}}} \left| \int_{\mathcal{Q}} f d(H_\theta - P_\theta) \right| \leq \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{d_k}(q, q') d\gamma(q, q')$$

for any $\gamma \in \Gamma(H_\theta, P_\theta)$. Thus after taking the infimum over γ, we have $W_{\mathcal{H}}(\theta) \leq W(\theta)$.

Upper bound: $W(\theta)$ distance is bounded as follows:

$$W(\theta) = \inf_{\gamma \in \Gamma(H_\theta, P_\theta)} \int_{\mathcal{Q} \times \mathcal{Q}} \|\mathbf{k}(\cdot, q) - \mathbf{k}(\cdot, q')\|_{\mathcal{H}} d\gamma(q, q')$$

$$\leq \int_{\mathcal{Q} \times \mathcal{Q}} \|\mathbf{k}(\cdot, q) - \mathbf{k}(\cdot, q')\|_{\mathcal{H}} dH_\theta(q) dP_\theta(q')$$

$$\overset{(1)}{\leq} \left[\int_{\mathcal{Q} \times \mathcal{Q}} \|\mathbf{k}(\cdot, q) - \mathbf{k}(\cdot, q')\|_{\mathcal{H}}^2 dH_\theta(q) dP_\theta(q') \right]^{\frac{1}{2}}$$

$$\overset{(2)}{\leq} \left[\int_{\mathcal{Q}} \mathbf{k}(q, q) d(H_\theta + P_\theta)(q) - 2 \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') H_\theta(q) dP_\theta(q') \right]^{\frac{1}{2}}$$

$$\overset{(3)}{\leq} \sqrt{2C + W_{\mathcal{H}}(\theta)}$$

$\overset{(1)}{\leq}$ follows Jensen's inequality. $\overset{(2)}{\leq}$ follows

$$\left[\int_{\mathcal{Q} \times \mathcal{Q}} \langle \mathbf{k}(\cdot, q) - \mathbf{k}(\cdot, q'), \mathbf{k}(\cdot, q) - \mathbf{k}(\cdot, q') \rangle_{\mathcal{H}} dH_\theta(q) dP_\theta(q') \right]^{\frac{1}{2}}$$

$$= \left[\int_{\mathcal{Q}} \mathbf{k}(q, q) dH_\theta(q) + \int_{\mathcal{Q}} \mathbf{k}(q', q') dP_\theta(q') - 2 \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') H_\theta(q) dP_\theta(q') \right]^{\frac{1}{2}}$$

$$\leq \left[\int_{\mathcal{Q}} \mathbf{k}(q, q) d(H_\theta + P_\theta)(q) - 2 \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') H_\theta(q) dP_\theta(q') \right]^{\frac{1}{2}}.$$

$\overset{(3)}{\leq}$ follows

$$\int_Q \mathbf{k}(q,q)d(H_\theta + P_\theta)(q) - 2\int_{Q\times Q} \mathbf{k}(q,q')dH_\theta(q)dP_\theta(q')$$

$$\leq \int_Q \mathbf{k}(q,q)d(H_\theta + P_\theta)(q) + \left[\int_{Q\times Q} \mathbf{k}(q,q')(dH_\theta(q) - dP_\theta(q'))^2\right]$$

$$\leq H_\theta\mathbf{k} + P_\theta\mathbf{k} + \left[\int_Q \mathbf{k}(q,q)d(H_\theta - P_\theta)(q)\right]^2$$

$$\leq 2C + \left[\sup_{f\in\mathcal{B}_\mathcal{H}} \int_Q fd(H_\theta - P_\theta)\right]^2 = 2C + W_\mathcal{H}(\theta)$$

where we set $C = \max\{\|H_\theta\mathbf{k}\|_\mathcal{H}, \|P_\theta\mathbf{k}\|_\mathcal{H}\}$. $\qquad\square$

C.3 PROOF OF THEOREM 6

Proof. In order to maintain simple expressions, without loss of generality the kernel functions $\mathbf{k}(q,q_i)$ is expressed by $f(q_i) = \mathbf{k}(q,q_i)$ for any q.

The upper bound of $\widehat{W}_\mathcal{H}(\theta) - W_\mathcal{H}(\theta)$ follows

$$\left|\widehat{W}_\mathcal{H}(\theta) - W_\mathcal{H}(\theta)\right| = \left|\sup_{f\in\mathcal{B}_\mathcal{H}} H_\theta f - P_\theta f - \sup_{f\in\mathcal{B}_\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^m f(q_i) - \frac{1}{n}\sum_{i=1}^n f(q_i')\right)\right|$$

$$\leq \underbrace{\sup_{f\in\mathcal{B}_\mathcal{H}}\left|H_\theta f - P_\theta f - \frac{1}{m}\sum_{i=1}^m f(q_i) + \frac{1}{n}\sum_{i=1}^n f(q_i')\right|}_{\Delta(H_\theta,P_\theta,\mathbf{q},\mathbf{q}')}.$$

We can provide an upper bound on the difference between $\Delta(H_\theta, P_\theta, Q, Q')$ and its expectation. Changing either of $(x_i, \varepsilon_i) \sim H_\theta$ or $(x_i', \varepsilon_i') \sim P_\theta$ in $\Delta(H_\theta, P_\theta, Q, Q')$ results in changes in magnitude of at most $2C_\mathbf{k}^{1/2}/m$ or $2C_\mathbf{k}^{1/2}/n$, respectively. We can then apply Theorem 13

$$\Pr\left\{(\Delta(H_\theta, P_\theta, \mathbf{q}, \mathbf{q}') - \mathbb{E}[\Delta(H_\theta, P_\theta, \mathbf{q}, \mathbf{q}')]) > \epsilon\right\} < \exp\left(-\frac{\epsilon^2 mn}{2C_\mathbf{k}(m+n)}\right) \qquad (20)$$

where the denominator in the exponent comes from

$$m\left(2C_\mathbf{k}^{\frac{1}{2}}/m\right)^2 + n\left(2C_\mathbf{k}^{\frac{1}{2}}/n\right)^2 = 4C_\mathbf{k}\left(\frac{1}{m} + \frac{1}{n}\right) = 4C_\mathbf{k}\frac{n+m}{mn}.$$

We then apply the symmetrization technique again to attain the upper bound of $\Delta(H_\theta, P_\theta, \mathbf{q}, \mathbf{q}')$. The procedure called the ghost sample approach, i.e. a second set of observations drawn from the same distribution, follows van der Vaart and Wellner (p.108, 1996). Denote $\tilde{Q} \sim H_\theta$ an i.i.d sample

of size m drawn independently of Q, similarly denote $\tilde{Q} \sim P_\theta$ in the same way.

$$\mathbb{E}[\Delta(H_\theta, P_\theta, \mathbf{q}, \mathbf{q}')]$$

$$=\mathbb{E} \sup_{f \in \mathcal{B}_\mathcal{H}} \left| H_\theta f - P_\theta f - \frac{1}{m} \sum_{i=1}^{m} f(q_i) + \frac{1}{n} \sum_{i=1}^{n} f(q_i') \right|$$

$$\stackrel{(a)}{=}\mathbb{E} \sup_{f \in \mathcal{B}_\mathcal{H}} \left| \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^{m} f(\tilde{q}_i) \right) - \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} f(\tilde{q}_i') \right) - \frac{1}{m} \sum_{i=1}^{m} f(q_i) + \frac{1}{n} \sum_{i=1}^{n} f(q_i') \right|$$

$$\stackrel{(b)}{\leq}\mathbb{E} \sup_{f \in \mathcal{B}_\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^{m} f(\tilde{q}_i) - \frac{1}{n} \sum_{i=1}^{n} f(\tilde{q}_i') - \frac{1}{m} \sum_{i=1}^{m} f(q_i) + \frac{1}{n} \sum_{i=1}^{n} f(q_i') \right|$$

$$\stackrel{(c)}{=}\mathbb{E} \sup_{f \in \mathcal{B}_\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(f(\tilde{q}_i) - f(q_i) \right) + \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(f(\tilde{q}_i') - f(q_i') \right) \right|$$

$$\stackrel{(d)}{\leq}\mathbb{E} \sup_{f \in \mathcal{B}_\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(f(\tilde{q}_i) - f(q_i) \right) \right| + \mathbb{E} \sup_{f \in \mathcal{B}_\mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(f(\tilde{q}_i') - f(q_i') \right) \right|$$

$$\stackrel{(e)}{\leq}2 \left[R_m(\mathcal{B}_\mathcal{H}, Q, H_\theta) + R_n(\mathcal{B}_\mathcal{H}, Q', P_\theta) \right] \stackrel{(f)}{\leq} 2 \left[(C_\mathbf{k}/m)^{\frac{1}{2}} + (C_\mathbf{k}/n)^{\frac{1}{2}} \right].$$

$\stackrel{(a)}{=}$ follows the ghost sample argument, $\stackrel{(b)}{\leq}$ follows Jensen's inequality, $\stackrel{(c)}{=}$ follows the symmetrization technique, $\stackrel{(d)}{\leq}$ follows the triangle inequality, $\stackrel{(e)}{\leq}$ and $\stackrel{(f)}{\leq}$ follow Theorem 14.

Substitute above result into (20), we have

$$\Pr \left\{ \Delta(H_\theta, P_\theta, \mathbf{q}, \mathbf{q}') - 2 \left[(C_\mathbf{k}/m)^{\frac{1}{2}} + (C_\mathbf{k}/n)^{\frac{1}{2}} \right] > \epsilon \right\} \leq \exp \left(-\frac{\epsilon^2 mn}{2C_\mathbf{k}(m+n)} \right).$$

Substitute this bound into the upper bound of $\left| \widehat{W}_\mathcal{H}(\theta) - W_\mathcal{H}(\theta) \right|$, then we get the result. $\qquad \square$

C.4   PROOF OF THEOREM 7

Proof. If Θ is compact, take $\mathcal{K} = \Theta$. Apply the bound in Hoeffding (p. 25 1963) to Theorem 6:

$$\Pr \left\{ \widehat{S}_\mathcal{H}(\theta) - S_\mathcal{H}(\theta) > \epsilon \right\} \leq \exp \left(\frac{-\epsilon^2 n}{8C_\mathbf{k}^2} \right).$$

Note the $S_\mathcal{H}(\theta_0) = 0$. It follows that

$$\Pr\{ \widehat{S}_\mathcal{H}(\theta^*) \geq \epsilon \} = \Pr \left\{ n^{-2} \sum_{i \neq j} h(q_i, q_j) \geq \epsilon \right\} \leq \exp \left(\frac{-\epsilon^2 n}{8C_\mathbf{k}^2} \right).$$

Note that if $\epsilon \to 0$ and $\epsilon^2 n \to \infty$, $\exp\left(-\epsilon^2 n / 8C_\mathbf{k}^2 \right) \to 0$.

If Θ is not compact, we do the following step. For fixed $\epsilon > 0$, take B_{θ_0} to be the open interval centered at θ_0 of length 2ϵ. Since $\mathcal{K} \setminus B_{\theta_0}$ is a compact set, and $\{B_\theta : \theta \in \mathcal{K} \setminus B_{\theta_0}\}$ is an open cover, we may extract a finite sub-cover, say, $B_{\theta_1}, \ldots, B_{\theta_k}$. For notational simplicity, rename $\mathcal{K}^c$ and these sets as $\Theta_1, \ldots, \Theta_m$, so that $\Theta = B_{\theta_0} \cup (\bigcup_{s=1}^{m} \Theta_s)$ and $h(\theta|\Theta_s) > 0$.

Write $c_s = h(\theta|\Theta_s)$. Then by Theorem 6, at this θ

$$\frac{1}{n^2} \sum_{i \neq j}^{n} h(q_i, q_j) \to c_s, \quad \text{with probability 1 for each } s.$$

Let A_s denote the set of data sequences for which this convergence holds and set $A = \bigcap_{s=1}^{m} A_s$. Then $\Pr(A) = 1$ and $n^{-2} \sum_{i \neq j} h(q_i, q_j) \to c_s > 0$ for all $\mathbf{q} = (q_1, q_2, \ldots,) \in A$. If "i.o." stands for "infinitely often," then we have:

$$
\begin{aligned}
\left\{\mathbf{q} : \limsup_{n \to \infty} |\theta^*(q_1, \ldots, q_n) - \theta_0| \geq \epsilon\right\} &\subseteq \bigcup_{s=1}^{m} \left\{\mathbf{q} : \theta^*(q_1, \ldots, q_n) \in \Theta_s \text{ i.o.}\right\} \\
&\subseteq \bigcup_{s=1}^{m} \left\{\mathbf{q} : \inf_{\theta \in \Theta_s} \frac{1}{n} \sum_{i \neq j} h(q_i, q_j) > 0 \text{ i.o.}\right\} \\
&\subseteq \bigcup_{s=1}^{m} \left\{\mathbf{q} : \hat{h}(\theta^*|\Theta_s) > 0 \text{ i.o.}\right\} \\
&\subseteq \bigcup_{s=1}^{m} A_s^c.
\end{aligned}
$$

Since the last set is A^c and $\Pr(A^c) = 0$, the result (15) follows. $\qquad \square$

C.5   Proof of Theorem 8

Proof. The proof follows the procedure in Serfling (Section 5.5.2, 1980). Under θ_0, by the definition of $\tilde{\mathbf{k}}(x_i, x_j)$, we have

$$
\begin{aligned}
\mathbb{E}_x \tilde{\mathbf{k}}(x, x_j) :&= \mathbb{E}_x \mathbf{k}(x, x_j) - \mathbb{E}_x \left[\mathbb{E}_{x'} \mathbf{k}(x, x')\right] - \mathbb{E}_x \mathbf{k}(x, x_j) + \mathbb{E}_{x, x'} \mathbf{k}(x, x') \\
&= -\mathbb{E}_x \left[\mathbb{E}_{x'} \mathbf{k}(x, x')\right] + \mathbb{E}_{x, x'} \mathbf{k}(x, x') = 0.
\end{aligned}
\tag{21}
$$

Step 1: To prove that the kernel $\tilde{\mathbf{k}}(x_i, x_j)$ function has the following form:

$$\tilde{\mathbf{k}}(x_i, x_j) = \sum_{l=1}^{\infty} \lambda_l^{(1)} \psi_l(x_i) \psi_l(x_j), \tag{22}$$

where $\lambda_l^{(1)}$ and ψ_l are eigenvalues and basis function of the eigenvalue equation

$$\int_Q \tilde{\mathbf{k}}(x, x')\psi_l(x)dF(x) = \lambda_l^{(2)}\psi_l(q').$$

Note that a Hilbert subspace endowed with an inner product over $\tilde{\mathbf{k}}(\cdot, \cdot)$ such that $\langle f(\cdot), \tilde{\mathbf{k}}(\cdot, x')\rangle = f(x')$ for all $f \in \mathcal{B}_\mathcal{H}$ and $\langle \tilde{\mathbf{k}}(\cdot, x), \tilde{\mathbf{k}}(\cdot, x')\rangle = \tilde{\mathbf{k}}(x, x')$. Let ψ_l be basis for $\mathcal{B}_\mathcal{H}$ and B_l be i.i.d. $\mathcal{N}(0, \sigma_l^2)$ with $\sigma_l \downarrow 0$ as $l \to \infty$. We will construct Karhunen-Loeve expansion in the following fashion. Suppose for any $x \in \mathcal{X}$,

$$f(x) = \sum_{l=1}^{\infty} B_l \psi_l(x).$$

Note that $\mathbb{E}[f(x)] = 0$ while the covariance kernel function is

$$\begin{aligned}
\tilde{\mathbf{k}}(x, x') &= \mathbb{E}_B\left[\sum_{l=1}^{\infty} B_l\psi_l(x)B_l\psi_l(x')\right] \\
&= \sum_{l=1}^{\infty} \sigma_l^2 \psi_l(x)\psi_l(x')
\end{aligned}$$

The covariance function $\tilde{\mathbf{k}}$ of zero mean Gaussian vectors completely characterizes the space $\mathcal{B}_\mathcal{H}$.

On the other hand, suppose we fix q. Then from Theorem 2, for continuous function, $\tilde{\mathbf{k}}(x, \cdot) = \sum_{i=1}^{\infty} \alpha_i(x)$ for some basis functions α_i uniformly. Now fix q' and we have

$$\tilde{\mathbf{k}}(x, x') = \sum_{i=1}^{\infty} \alpha_i(x) \sum_{j=1}^{\infty} \beta_j(x') = \sum_{i,j=1}^{\infty} \alpha_i(x)\beta_j(x')$$

Again, β_j are basis functions. For finite numbers of the basis $\alpha_1, \cdots, \alpha_L$ and $\beta_1, \cdots, \beta_L$, it can be representable as a linear combination of orthogonal basis ψ using the Gram-Schmidt orthogonalization. So we have

$$\tilde{\mathbf{k}}(x, x') = \lim_{L \to \infty} \sum_{l=1}^{L} \lambda_l^{(1)}\psi_l(x)\psi_l(x').$$

For continuous centered kernel $\tilde{\mathbf{k}}$, define linear operator $\langle \cdot, \tilde{\mathbf{k}}\rangle : \mathcal{B}_\mathcal{H} \to \mathcal{B}_\mathcal{H}$ as

$$\langle f(\cdot), \tilde{\mathbf{k}}(\cdot, x')\rangle = \int \tilde{\mathbf{k}}(x, x')f(x)\, dF(x) = f(x').$$

This is a compact self-adjoint operator[8] which yields unique countable eigenvalues λ_i and orthonor-

---

[8]An operator T is compact if the closure of T(close ball) is compact and it is self-adjoint if $\langle Tf, g\rangle = \langle f, Tg\rangle$.

mal eigenfunctions ψ_i such that[9]

$$\int \tilde{\mathbf{k}}(x, x')\psi_l(x)dF(x) = \lambda_l^{(1)}\psi_l(x'),$$

$$\int \psi_i(x)\psi_j(x)dF(x) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

with $\lambda_\infty = 0$. We will order them such that $\lambda_0 > \lambda_1 > \lambda_2 > \cdots$. Hence identifying $\lambda_i = \sigma_i^2$. Similarly, we can obtain the results for $\tilde{\mathbf{k}}(\varepsilon_i, \varepsilon_j) = \sum_{l=1}^\infty \lambda_l^{(2)}\psi_l(\varepsilon_i)\psi_l(\varepsilon_j)$.

Step 2: To prove that

$$n\widehat{S}_{\mathcal{H}}(\theta^*) := \frac{1}{n}\sum_{i \neq j}^n h(q_i, q_j) = \frac{1}{n}\sum_{i \neq j}^n \left[\tilde{\mathbf{k}}(x_i, x_j) + \tilde{\mathbf{k}}(\varepsilon_i, \varepsilon_j) - \tilde{\mathbf{k}}(x_i, \varepsilon_j) - \tilde{\mathbf{k}}(\varepsilon_i, x_j)\right]$$

$$\rightsquigarrow \sum_{l=1}^\infty \lambda_l \left[z_l^2 - 1\right].$$

Firstly, consider the limit of $n^{-1}\sum_{i \neq j}^n \tilde{\mathbf{k}}(x_i, x_j)$. With the result in (22), we have

$$\frac{1}{n}\sum_{i \neq j} \tilde{\mathbf{k}}(x_i, x_j) = \frac{1}{n}\sum_{i \neq j}\sum_{l=1}^\infty \lambda_l^{(1)}\psi_l(x_i)\psi_l(x_j)$$

$$\overset{(a)}{=} \frac{1}{n}\sum_{l=1}^\infty \lambda_l^{(1)}\left(\left(\sum_{i=1}^n \psi_l(x_i)\right)^2 - \sum_{i=1}^n \psi_l^2(x_i)\right)$$

$\overset{(a)}{=}$ comes from $(\sum_i a_i)^2 = \sum_i a_i(a_i + \sum_{i \neq j} a_j) = \sum_i a_i^2 + \sum_{i \neq j} a_i a_j$.

For any finite l, taking expectation over (23) gives

$$\lambda_l^{(1)}\mathbb{E}_x[\psi_l(x)] = \int \mathbb{E}_x\tilde{\mathbf{k}}(x', x)\psi_l(x')dF(x') = 0$$

by the result of (21). Hence $\mathbb{E}_x[\psi_l(x)] = 0$ because $\lambda_l \neq 0$ if $l \neq \infty$. Also, from (23), we know that $\text{Cov}(\psi_i(x), \psi_j(x)) = \delta_{ij}$. By the Lindeberg-Levy CLT,

$$n^{-1/2}\sum_{i=1}^n \psi_l(x_i) \rightsquigarrow \mathcal{N}(0, 1).$$

---

[9]This eigenequation is called a Fredholm equation of the first kind. Given covariance function, there is a numerical technique for estimating λ_i and ψ_i numerically.

From (23), it is also known that $\mathbb{E}_x[\psi_l^2(x)] = 1$ for any l. Then strong LLN gives

$$n^{-1}\left(\sum_{i=1}^n \psi_l^2(x_i)\right) \xrightarrow{p} 1.$$

Hence

$$\frac{1}{n}\sum_{l=1}^\infty \lambda_l^{(1)}\left(\left(\sum_{i=1}^n \psi_l(x_i)\right)^2 - \sum_{i=1}^n \psi_l^2(x_i)\right) \rightsquigarrow \sum_{l=1}^\infty \lambda_l^{(1)}\left[z_l^2 - 1\right]$$

where z_l^2 is a standard normal random variable.

Secondly, consider the limit of $n^{-1}\sum_{i\neq j}^n \tilde{\mathbf{k}}(\varepsilon_i, \varepsilon_j)$. Following the previous procedure, we can derive

$$\frac{1}{n}\sum_{i\neq j} \tilde{\mathbf{k}}(\varepsilon_i, \varepsilon_j) = \frac{1}{n}\sum_{l=1}^\infty \lambda_l^{(2)}\left(\left(\sum_{i=1}^n \psi_l(\varepsilon_i)\right)^2 - \sum_{i=1}^n \psi_l^2(\varepsilon_i)\right) \rightsquigarrow \sum_{l=1}^\infty \lambda_l^{(2)}\left[z_l^2 - 1\right].$$

Finally, consider the limit of $n^{-1}\sum_{i\neq j}^n \left[\tilde{\mathbf{k}}(x_i, \varepsilon_j) + \tilde{\mathbf{k}}(\varepsilon_i, x_j)\right]$. Because by independence condition, at θ_0, x and ε are independent. Hence $\text{Cov}(\psi_i(x), \psi_j(\varepsilon)) = 0$ for all i, j. Then we can conclude

$$\frac{1}{n}\sum_{i\neq j}^n \left[\tilde{\mathbf{k}}(x_i, \varepsilon_j) + \tilde{\mathbf{k}}(\varepsilon_i, x_j)\right] \to 0.$$

Combine these three limit forms, the result follows. $\qquad\square$

C.6 PROOF OF THEOREM 10

Proof. The proof uses some properties of Hellinger distance function. With C14, a Hellinger distance can be approximately represented as

$$\begin{aligned}
\lim_{\|\theta-\theta_0\|\to 0} \mathbf{d}_{\mathbf{H}}^2(H_\theta, FG_\theta) &= \lim_{\|\theta-\theta_0\|\to 0} \frac{1}{2}\int\left(\sqrt{h_\theta(q)w_\theta(q)} - \sqrt{p_\theta(q)w_\theta(q)}\right)^2 dq \\
&= \lim_{\|\theta-\theta_0\|\to 0} \frac{1}{2}\int\left(\sqrt{h_\theta(q)} - \sqrt{p_\theta(q)}\right)^2 w_\theta(q)dq \\
&= \lim_{\|\theta-\theta_0\|\to 0} \frac{1}{2}\int\left((\theta-\theta_0)^T\Delta_w(q)\right)\left((\theta-\theta_0)^T\Delta_w(q)\right)^T\left(\frac{w_\theta(q)}{\mathbf{k}(q)}\right)dq + o_p(1) \\
&\approx \frac{1}{2}(\theta-\theta_0)^T\int(\iota_{\theta_0}(q))d(H_\theta - P_\theta)\left[\int(\iota_{\theta_0}(q))d(H_\theta - P_\theta)\right]^T(\theta-\theta_0)^T
\end{aligned}$$

where $\iota_{\theta_0}(q) = (\log D_\theta(q)w_\theta(q))'|_{\theta=\theta_0}$.

Step 1: By definition, $(\theta^*, \mathbf{k})$ is the minimizer of $\hat{W}_{\mathcal{H}}(\theta)$ such that

$$(\theta^*, \mathbf{k}) = \arg\min_{\theta} \max_{f \in \mathcal{H}} \left| \int f(q) dH_\theta - \int f(q) dP_\theta \right|,$$

then (θ^*, w_{θ^*}) satisfies the following inequalities

$$\underbrace{\sqrt{4 \int \left(\sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right)^2 w_{\theta^*}(q) dq}}_{:= 2\sqrt{2} \mathbf{d}_{\mathbf{H}}^2(H_{\theta^*}, FG_{\theta^*})} \overset{(a)}{\geq} \sqrt{\int \left| \sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right|^2 w_{\theta^*}(q) dq}$$

$$\times \sqrt{\int \left| \sqrt{h_{\theta^*}(q)} + \sqrt{p_{\theta^*}(q)} \right|^2 w_{\theta^*}(q) dq}$$

$$\overset{(b)}{\geq} \int \left| \sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right| \left| \sqrt{h_{\theta^*}(q)} + \sqrt{p_{\theta^*}(q)} \right| w_{\theta^*}(q) dq$$

$$\overset{(c)}{=} \int \left| h_{\theta^*}(q) - p_{\theta^*}(q) \right| w_{\theta^*}(q) dq$$

$$\overset{(d)}{\geq} \left| \int h_{\theta^*}(q) w_{\theta^*}(q) dq - \int p_{\theta^*}(q) w_{\theta^*}(q) dq \right|$$

where $\overset{(a)}{\geq}$ comes from

$$4 \int \left(\sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right)^2 w_{\theta^*}(q) dq \geq \int \left| \sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right|^2 w_{\theta^*}(q) dq +$$

$$\int \left| \sqrt{h_{\theta^*}(q)} + \sqrt{p_{\theta^*}(q)} \right|^2 w_{\theta^*}(q) dq$$

$$\geq \int \left| \sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right|^2 w_{\theta^*}(q) dq \times$$

$$\int \left| \sqrt{h_{\theta^*}(q)} + \sqrt{p_{\theta^*}(q)} \right|^2 w_{\theta^*}(q) dq$$

by using the properties $\left| \sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right| < 1$ and $0 < \int \left| \sqrt{h_\theta(q)} - \sqrt{p_\theta(q)} \right|^2 w_\theta(q) dq < 1$; $\overset{(b)}{\geq}$ follows from Cauchy–Schwarz inequality

$$\sqrt{\int |a(q)|^2 w_{\theta^*}(q) dq \times \int |b(q)|^2 w_{\theta^*}(q) dq} \geq \int |a(q)| \, |b(q)|^2 w_{\theta^*}(q) dq$$

where $a(q) = \sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)}$ and $b(q) = \sqrt{h_{\theta^*}(q)} + \sqrt{p_{\theta^*}(q)}$; $\overset{(c)}{=}$ is from $|a(q)| \cdot |b(q)| = |a(q)b(q)|$; $\overset{(d)}{\geq}$ follows from Hi¿œlder's inequality.

On the other hand, we have

$$\max_{\mathcal{F}} \sqrt{S_{\mathcal{H}}(\theta^*)} = \sqrt{\int |h_{\theta^*}(q) - p_{\theta^*}(q)|^2 \, \mathbf{k}(q) dq} \overset{(a)}{\geq} \sqrt{\int |h_{\theta^*}(q) - p_{\theta^*}(q)|^2 \, w_{\theta^*}(q) dq}$$

$$\overset{(b)}{\geq} \int |h_{\theta^*}(q) - p_{\theta^*}(q)| \, w_{\theta^*}(q) dq$$

$$\overset{(c)}{=} \int \left| \sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right| w_{\theta^*}(q) dq \times$$

$$\int \left| \sqrt{h_{\theta^*}(q)} + \sqrt{p_{\theta^*}(q)} \right| w_{\theta^*}(q) dq$$

$$\overset{(d)}{\geq} \underbrace{\int \left(\sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right)^2 w_{\theta^*}(q) dq}_{:=2\mathbf{d}_{\mathbf{H}}^2(H_{\theta^*}, FG_{\theta^*})}$$

where $\overset{(a)}{\geq}$ comes from the condition C14

$$\min_{\theta} \max_{w_\theta \in \mathcal{F}} \int |h_\theta(q) - p_\theta(q)|^2 \, w_\theta(q) dq \geq \min_{\theta} \int |h_\theta(q) - p_\theta(q)|^2 \, w_\theta(q) dq$$

for any $w_\theta \in \mathcal{F}$; $\overset{(b)}{\geq}$ is from Hï¿œlder's inequality; $\overset{(c)}{=}$ is from $|a(q)| \cdot |b(q)| = |a(q)b(q)|$; $\overset{(d)}{\geq}$ is from $\left| \sqrt{h_{\theta^*}(q)} + \sqrt{p_{\theta^*}(q)} \right| \geq \left| \sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right|$ for non-negative $h_{\theta^*}(q)$ and $p_{\theta^*}(q)$.

Step 2: Information in $\mathbf{d}_{\mathbf{H}}^2(H_{\theta^*}, P_{\theta^*})$.

Consider an expansion of $2\mathbf{d}_{\mathbf{H}}^2(H_{\theta^*}, P_{\theta^*})$ around θ_0.

$$2\mathbf{d}_{\mathbf{H}}^2(H_{\theta^*}, P_{\theta^*}) = \underbrace{\int \left(\sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} - (\sqrt{h_{\theta_0}(q)} - \sqrt{p_{\theta_0}(q)}) \right)^2 w_{\theta^*}(q) dq}_{(I)}$$

$$+ \underbrace{\int \left(\sqrt{h_{\theta_0}(q)} - \sqrt{p_{\theta_0}(q)} \right)^2 w_{\theta^*}(q) dq}_{(II)}$$

$$- 2 \underbrace{\int \left(\sqrt{h_{\theta^*}(q)} - \sqrt{p_{\theta^*}(q)} \right) \left(\sqrt{h_{\theta_0}(q)} - \sqrt{p_{\theta_0}(q)} \right) w_{\theta^*}(q) dq}_{(III)}.$$

Term (I) has the following properties

$$(I) \geq \left| \int (h_{\theta^*}(q) - p_{\theta^*}(q)) \, w_{\theta^*}(q) dq \right| - \left| \int (h_{\theta_0}(q) - p_{\theta_0}(q)) \, w_{\theta_0}(q) dq \right|$$

and

$$(I) \leq \int |h_{\theta^*}(q) - p_{\theta^*}(q)| \, \mathbf{k}(q)dq + \int |h_{\theta_0}(q) - p_{\theta_0}(q)| \, \mathbf{k}(q)dq \leq \sqrt{S_{\mathcal{H}}(\theta^*)} + \sqrt{S_{\mathcal{H}}(\theta_0)},$$

and term (II) is $2\mathbf{d}_{\mathbf{H}}^2(H_{\theta_0}, P_{\theta_0})$ follows the bounds

$$\left| \int (h_{\theta_0}(q) - p_{\theta_0}(q)) \, w_{\theta_0}(q)dq \right| \leq (II) \leq \int |h_{\theta_0}(q) - p_{\theta_0}(q)| \, \mathbf{k}(q)dq \leq \sqrt{S_{\mathcal{H}}(\theta_0)},$$

and term (III) by condition C14 has the following representation

$$(III) = 2(\theta^* - \theta_0)^T \int \Delta_w(q) \mathbf{k}^{-1/2}(q)(\sqrt{h_{\theta_0}(q)} - \sqrt{p_{\theta_0}(q)})w_{\theta^*}(q)dq + o(\|\theta^* - \theta_0\|)$$

$$= (\theta^* - \theta_0)^T \int \Delta_w(q) \left(\frac{w_{\theta^*}}{w_{\theta_0}\mathbf{k}} \right) d(H_{\theta_0} - P_{\theta_0})(q) + o(\|\theta^* - \theta_0\|)$$

by using the fact that $w_{\theta_0}^{1/2}(q)dq = d\left(\sqrt{H_{\theta_0}} - \sqrt{P_{\theta_0}}\right) = (\sqrt{h_{\theta_0}(q)} - \sqrt{p_{\theta_0}(q)})^{-1}d(H_{\theta_0} - P_{\theta_0})/2$.
As $\theta^* \xrightarrow{p} \theta_0$,

$$0 = \lim_{\theta^* \xrightarrow{p} \theta_0} W_{\mathcal{H}}(\theta^*) \geq \left| \int (h_{\theta^*}(q) - p_{\theta^*}(q)) \, w_{\theta^*}(q)dq \right|$$

and $S_{\mathcal{H}}(\theta^*) \to 0$ ensure that (I) and (II) are ignorable and the representation

$$(\theta^* - \theta_0)^T \int \Delta_w(q) \left(\frac{w_{\theta_0}}{\mathbf{k}} \right)^{\frac{1}{2}} d(H_{\theta_0} - P_{\theta_0})(q)$$

in (III) gives the information score function at θ_0

$$\int \Delta_w(q) \left(\frac{w_{\theta_0}}{\mathbf{k}} \right)^{\frac{1}{2}} d(H_{\theta_0} - P_{\theta_0})(q).$$

From the result of Delta method on minimum distance estimator, see i.e. proof of Therem 3.3 (iii) in BW, it is known that

$$\theta^* = \theta_0 - 2\mathbf{V}^{-1} \int \Delta_w(q) \left(\frac{w_{\theta_0}}{\mathbf{k}} \right)^{\frac{1}{2}} d(H_{\theta_0} - P_{\theta_0})(q) + o(1)$$

where $\mathbf{V}$ is the derivative of $S_{\mathcal{H}}(\theta)$ assumed in condition C13. Then the covariance term is

$$\left[\int \left(\Delta_w(q) \left(\frac{w_{\theta_0}}{\mathbf{k}} \right)^{\frac{1}{2}}(q) \right) d(H_{\theta_0} - P_{\theta_0})(q') \right] \left[\int \left(\Delta_w(q) \left(\frac{w_{\theta_0}}{\mathbf{k}} \right)^{\frac{1}{2}}(q') \right) d(H_{\theta_0} - P_{\theta_0})(q') \right]^T.$$

Because $w_\theta \in \mathcal{F}$, $\mathbf{k} := \arg\max_{f \in \mathcal{F}} \int f d(H_\theta - P_\theta)(q)$ and $\|\mathbf{k}(q)\| \le 1$, we can see

$$\int \left(\Delta_w(q) \left(\frac{w_{\theta_0}}{\mathbf{k}} \right)^{\frac{1}{2}} \right) d(H_{\theta_0} - P_{\theta_0})(q) \le \int \left(\Delta_w(q) \left(\frac{w_{\theta_0}}{f} \right)^{\frac{1}{2}} \right) d(H_{\theta_0} - P_{\theta_0})(q)$$

for any $f \in \mathcal{F}$. Thus $(\theta^*, \mathbf{k})$ attains the optimal information. $\qquad\square$

D PROOF OF OTHER STATEMENTS

D.1 PROOF OF COROLLARY 1

Proof. By the definition of metrization, we need to show that for $\{P_1, P_2, \dots\} \subset \mathscr{P}$, $P_n \rightsquigarrow P$ if and only if $\|P_n \mathbf{k} - P\mathbf{k}\|_{\mathcal{H}} \to 0$ as $n \to \infty$.

($\Rightarrow$) Because $\mathbf{k}$ is bounded and continuous, so it is obvious that $P_n \rightsquigarrow P$ implies $\|P_n \mathbf{k} - P\mathbf{k}\|_{\mathcal{H}} \to 0$ as $n \to \infty$.

($\Leftarrow$) $\mathcal{H}$ induces a dense set in $\mathcal{B}_{\mathcal{H}} \subset C(\mathcal{Q})$. For the space of continuous bounded functions, any $\mathbf{k} \in \mathcal{F}_{\mathcal{H}}$ and any every $\epsilon > 0$, there exists a $g \in \mathcal{F}_{\mathcal{H}}$ such that $\sup |\mathbf{k} - g| \le \epsilon$. Therefore

$$\left| \int \mathbf{k} dP_n - \int \mathbf{k} dP \right| = \left| \int \mathbf{k} dP_n - \int g dP_n + \int \mathbf{k} dP - \int g dP + \int g dP_n - \int g dP \right|$$

$$\le \left| \int \mathbf{k} dP_n - \int g dP_n \right| + \left| \int \mathbf{k} dP - \int g dP \right| + \left| \int g dP_n - \int g dP \right|$$

$$\le 2\epsilon + \left| \int g dP_n - \int g dP \right| \le 2\epsilon + \|g\|_{\mathcal{H}} \|P_n \mathbf{k} - P\mathbf{k}\|_{\mathcal{H}}.$$

If $\|P_n \mathbf{k} - P\mathbf{k}\|_{\mathcal{H}} \to 0$ as $n \to \infty$ and ϵ is arbitrary, $|P_n \mathbf{k} - P\mathbf{k}| \to 0$ for any $\mathbf{k} \in \mathcal{B}_{\mathcal{H}}$ or say $P_n \rightsquigarrow P$. $\qquad\square$

D.2 PROOF OF LEMMA 1

Proof. ($\Leftarrow$) Suppose a finite non-zero signed Borel measure $\mathcal{P}$ satisfies Condition (I) and (II) in Lemma 1. By Jordan decomposition theorem (Theorem 5.6.1, Dudley, 2002), there exist unique positive measure $\mathcal{P}^+$ and $\mathcal{P}^-$ such that $\mathcal{P} = \mathcal{P}^+ - \mathcal{P}^-$ and $\mathcal{P}^+ \perp \mathcal{P}^-$. By Condition (II) in Lemma 1, we have $\mathcal{P}^+(\mathcal{Q}) = \mathcal{P}^-(\mathcal{Q}) =: \alpha$. Let $H_\theta = \alpha^{-1}\mathcal{P}^+$ and $P_\theta = \alpha^{-1}\mathcal{P}^-$. Thus, $H_\theta \ne P_\theta$. (Or one can set $P_\theta = \alpha^{-1}\mathcal{P}^+$ and $H_\theta = \alpha^{-1}\mathcal{P}^-$.) Then, by

$$W_{\mathcal{H}}^2(\theta) = \|H_\theta \mathbf{k} - P_\theta \mathbf{k}\|_{\mathcal{H}}^2 = \langle H_\theta \mathbf{k} - P_\theta \mathbf{k}, H_\theta \mathbf{k} - P_\theta(q) \rangle$$

$$= \left\langle \int_{\mathcal{Q}} \mathbf{k}(\cdot, q) dH_\theta(q), \int_{\mathcal{Q}} \mathbf{k}(\cdot, q') dH_\theta(q') \right\rangle + \left\langle \int_{\mathcal{Q}} \mathbf{k}(\cdot, q) dP_\theta(q), \int_{\mathcal{Q}} \mathbf{k}(\cdot, q') dP_\theta(q') \right\rangle$$

$$- 2 \left\langle \int_{\mathcal{Q}} \mathbf{k}(\cdot, q) dH_\theta(q), \int_{\mathcal{Q}} \mathbf{k}(\cdot, q') dP_\theta(q') \right\rangle$$

$$= \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') dH_\theta(q) dH_\theta(q') + \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') dP_\theta(q) dP_\theta(q')$$

$$- 2 \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') dH_\theta(q) dP_\theta(q')$$

$$= \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') d(H_\theta - P_\theta)(q) d(H_\theta - P_\theta)(q') = 0.$$

The last equality is set by Condition (I) in Lemma 1.

($\Rightarrow$) Suppose there exists $H_\theta \neq P_\theta$ such that $W_{\mathcal{H}}(\theta) = 0$. Let $\mathcal{P} = P_\theta - H_\theta$. Then $\mathcal{P}$ is a finite non-zero signed Borel measure that satisfies $\mathcal{P}(\mathcal{Q}) = 0$. Note that

$$W_{\mathcal{H}}^2(\theta) = \int_{\mathcal{Q} \times \mathcal{Q}} \mathbf{k}(q, q') d(H_\theta - P_\theta)(q) d(H_\theta - P_\theta)(q') = 0.$$

Thus Condition (I) in Lemma 1 satisfies. $\square$

D.3   PROOF OF COROLLARY 2

Proof. By the property of Lagrange multipliers u_i, v_j, the problem (18) can be rewritten as a minimax problem:

$$V = \min_{\gamma_{i,j}(\theta) \in \Gamma(H_{n\theta}, P_{n\theta})} \max_{u_i, v_j} \sum_{i,j} \gamma_{i,j}(\theta) C_{i,j} + \sum_i u_i - \sum_{i,j} u_i \gamma_{i,j}(\theta) + \sum_j v_j - \sum_{i,j} v_j \gamma_{i,j}(\theta)$$

$$= \min_{\gamma_{i,j}(\theta) \in \Gamma(H_{n\theta}, P_{n\theta})} \max_{u_i, v_j} \sum_i u_i + \sum_j v_j + \sum_{i,j} \gamma_{i,j}(\theta) \left(C_{i,j} - u_i - v_j \right).$$

By the saddlepoint argument, we have

$$V = \max_{u_i, v_j} \min_{\gamma_{i,j}(\theta) \in \Gamma(H_{n\theta}, P_{n\theta})} \sum_i u_i + \sum_j v_j + \sum_{i,j} \gamma_{i,j}(\theta) \left(C_{i,j} - u_i - v_j \right).$$

Consider $\gamma_{i,j}(\theta)$ as the Lagrange multipliers associated to $u_i + v_j \geq C_{i,j}$. Hence, we have the dual formulation

$$\max_{u_i, v_j} \sum_i u_i + \sum_j v_j$$

$$\text{s.t.} \quad u_i + v_j \geq C_{i,j}.$$

The result follows. $\square$

E Auxiliary Results

E.1 Large Deviation Bounds

Theorem 13. *(McDiarmid's inequality) Let $c(\cdot) : \mathcal{Q}^m \to \mathbb{R}$ where $\mathcal{Q}^m = \mathcal{Q} \times \cdots \times \mathcal{Q}$. For all $i \in \{1, \ldots, m\}$, there exists $C_i < \infty$ such that*

$$\sup_{(q_1,\ldots,q_m) \in \mathcal{Q}^m, \tilde{q} \in \mathcal{Q}} |f(q_1, \ldots, q_m) - f(q_1, \ldots, q_{i-1}, \tilde{q}, q_{i+1}, \ldots, q_m)| \leq C_i.$$

Then for all probability measures p and every $\epsilon > 0$,

$$\Pr\{(f(q_1, \ldots, q_m) - \mathbb{E}[f(q_1, \ldots, q_m)]) > \epsilon\} < \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m C_i^2}\right)$$

where the expectation is defined for $(q_1, \ldots, q_m)$.

Theorem 14. *(Rademacher symmetrization) Let $\mathcal{B}_{\mathcal{H}}$ is the unit ball in the Hilbert space $\mathcal{H}$ on a compact domain $\mathcal{Q}$. The associated kernel is bounded such that $0 \leq \mathbf{k}(q, q') \leq C_{\mathbf{k}}$. Let Q be an i.i.d. sample of size m drawn according to a probability measure $H_\theta(q)$ on $\mathcal{Q}$, and let σ_i be i.i.d random variable taking values in $\{-1, 1\}$ with equal probability. The Rademacher average*

$$R_m(\mathcal{B}_{\mathcal{H}}, Q, H_\theta) := \mathbb{E}_{H_\theta, \sigma} \sup \left|\frac{1}{m} \sum_{i=1}^m \sigma_i f(q_i)\right| \leq \sqrt{(C_{\mathbf{k}}/m)}$$

where the expectation is taken w.r.t. $Q \sim H_\theta$ and σ jointly.
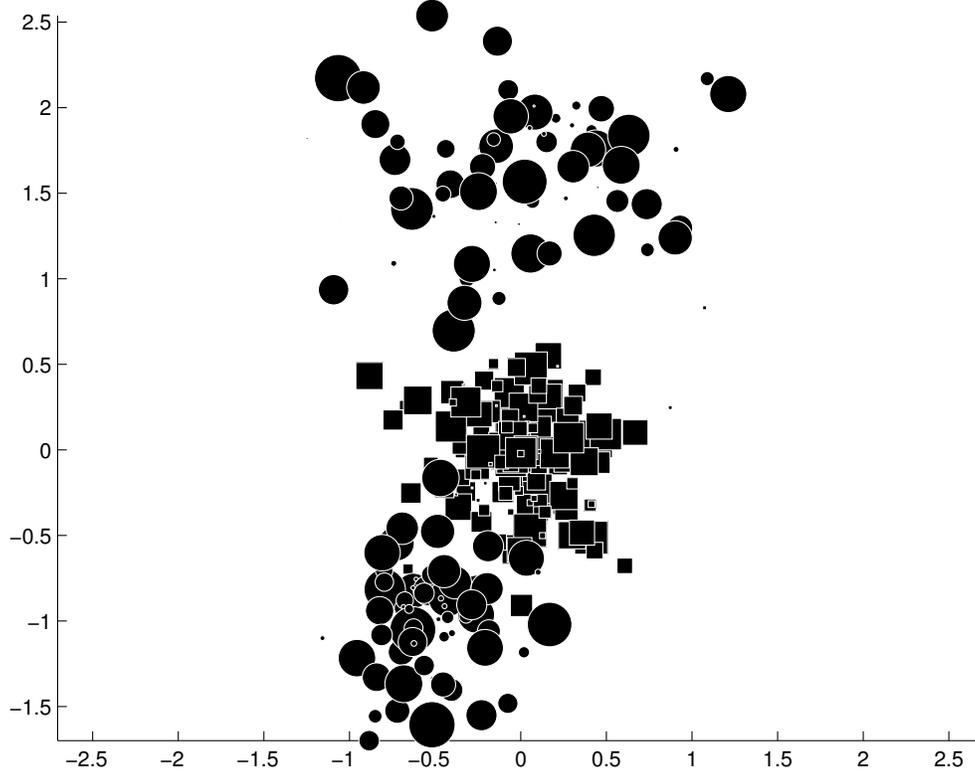
Figure 1: Point clouds of $\mathbf{q} \sim H_{n\theta}$ and $\mathbf{q}' \sim P_{n\theta_0}$.
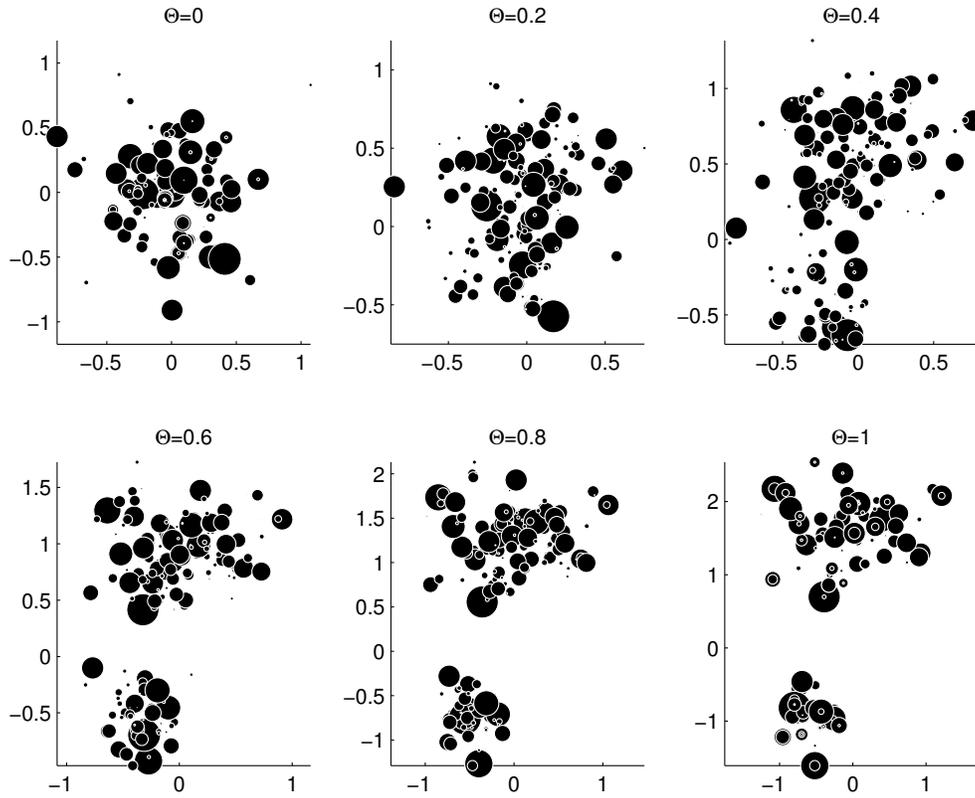
Figure 2: Point clouds of $\mathbf{q}' \sim P_{n\theta}$ are pushed towards to $\mathbf{q}' \sim P_{n\theta_0}$.
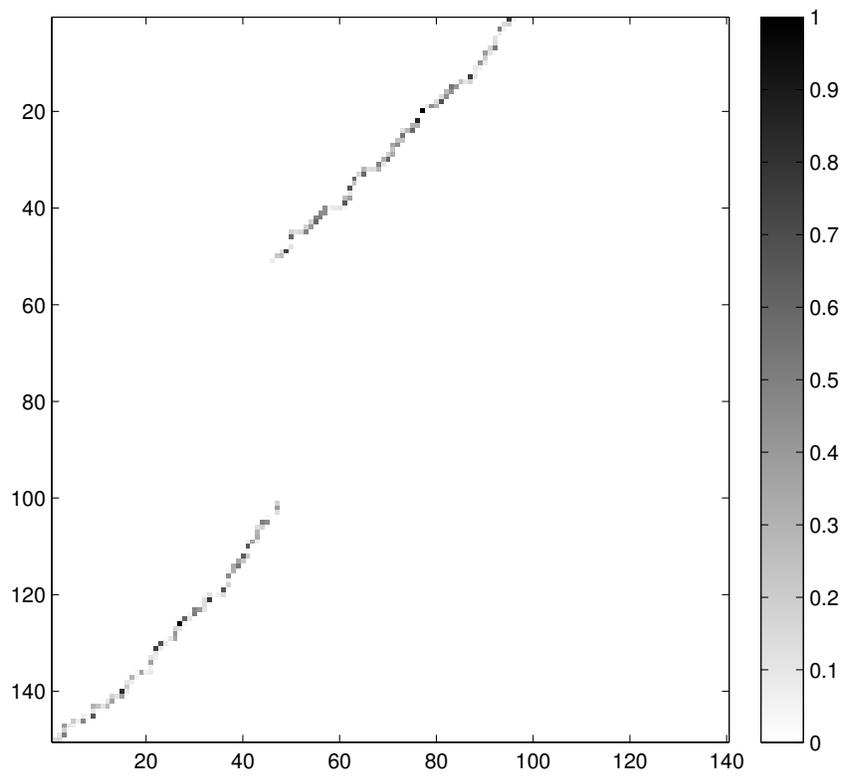
Figure 3: A matrix plot of $\gamma_{i,j}(\theta^*)$.
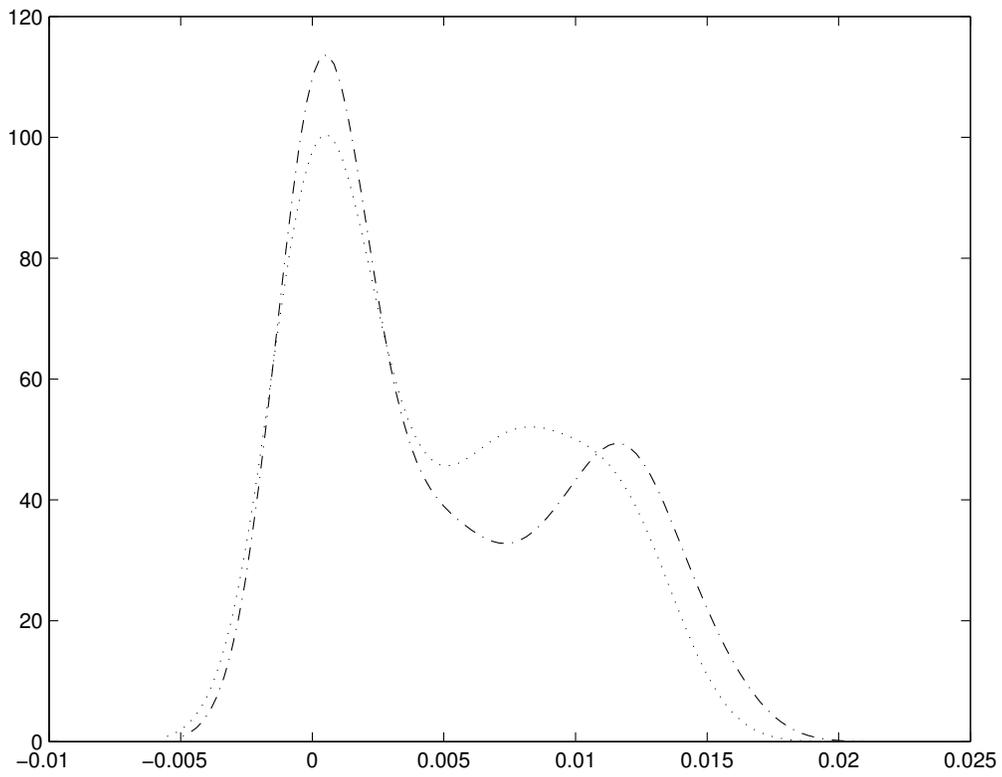
Figure 4: Density plots of $dH_{n\theta^*}$ and $dP_{n\theta^*}$.
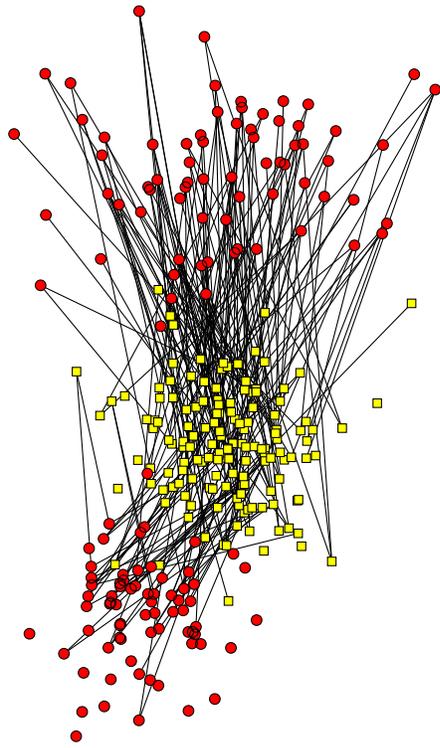
Figure 5: Measure $\gamma_{i,j}(\theta^*)$ as an optimal action.