Pairwise Valid Instruments

Zhenting Sun  
National School of Development  
Peking University  
zhentingsun@nsd.pku.edu.cn

Kaspar Wüthrich  
Department of Economics  
University of California San Diego  
kwuthrich@ucsd.edu

March 16, 2022

Abstract

Finding valid instruments is difficult. We propose Validity Set Instrumental Variable (VSIV) regression, a method for estimating treatment effects when the instruments are partially invalid. VSIV regression exploits testable implications for instrument validity to remove invalid variation in the instruments. We show that the proposed VSIV estimators are asymptotically normal under weak conditions and always remove or reduce the asymptotic bias relative to standard IV estimators.

Keywords: Invalid instruments, local average treatment effects, treatment effects, identification, instrumental variable estimation, bias reduction
1 Introduction

Instrumental variable (IV) methods based on the local average treatment effects (LATEs) framework (Imbens and Angrist, 1994; Angrist and Imbens, 1995; Angrist et al., 1996) rely on three assumptions:¹ (i) exclusion (the instrument does not have a direct effect on the outcome), (ii) random assignment (the instrument is independent of potential outcomes and treatments), and (iii) monotonicity (the instrument has a monotonic impact on treatment take-up).² In many applications, some of these assumptions are likely to be violated or at least questionable. This has motivated the derivation of testable restrictions and tests for IV validity in various settings (e.g., Balke and Pearl, 1997; Imbens and Rubin, 1997; Heckman and Vytlacil, 2005; Kitagawa, 2015; Huber and Mellace, 2015; Mourifié and Wan, 2017; Kédagni and Mourifié, 2020; Carr and Kitagawa, 2021; Sun, 2021). The main contribution of this paper is to propose a method for exploiting the information available in the testable restrictions of IV validity to remove or reduce the bias in IV estimation.

We consider a setting where the available instruments are partially invalid. Our method, which we refer to as Validity Set IV (VSIV) estimation, has two steps. First, we use testable implications of IV validity to remove invalid variation in the instruments. Second, we run an IV regression using the remaining variation in the instruments. We establish the asymptotic normality of the proposed VSIV estimators and show that they always remove or reduce the bias relative to traditional IV estimators. Thus, VSIV regression constitutes a data-driven approach for removing or reducing the bias in IV estimation as much as possible, given all the information about IV validity in the data.

Our goal is to estimate the causal effect of an endogenous treatment \( D \) on an outcome of interest \( Y \), using a potentially vector-valued discrete instrument \( Z \). In the ideal case, \( Z \) is fully valid, i.e., the LATE assumptions hold for all instrument values. However, full instrument validity is questionable in many applications, especially when there are many instruments or instrument values. To this end, we introduce the notion of pairwise valid instruments.³ Pairwise valid instruments are only valid for a subset of all pairs of instrument values, which we refer to as the validity pair set. In the first step of VSIV regression, we identify and estimate the largest validity pair set, \( \mathcal{P}_M \), using the testable restrictions for IV validity in Kitagawa (2015), Mourifié and Wan (2017), Kédagni and Mourifié (2020),

---

¹See, for example, Imbens (2014); Melly and Wüthrich (2017); Huber and Wüthrich (2018) for recent reviews and Angrist and Pischke (2008, 2014); Imbens and Rubin (2015) for textbook treatments.

²Some papers also include the instrument first stage assumption as part of the LATE assumptions. We maintain suitable first stage assumptions throughout this paper and thereby abstract from the issues arising from weak instruments.

³Pairwise validity can be viewed as a generalization of the partial monotonicity assumption of Mogstad et al. (2021). See Remark 2.1 for a discussion.
and Sun (2021). In the second step of VSIV regression, we estimate LATEs for all pairs of instrument values in the estimated validity set, \( \hat{Z}_0 \).

We study the theoretical properties of VSIV regression under two scenarios. If the estimated validity pair set, \( \hat{Z}_0 \), is consistent for the largest validity pair set \( Z_M \) in the sense that \( \Pr(\hat{Z}_0 = Z_M) \to 1 \), VSIV regression is asymptotically unbiased and normal under standard conditions. Since the estimator of the validity pair set, \( \hat{Z}_0 \), is typically constructed based on necessary (but not sufficient) conditions for IV validity, it could converge to a pseudo-true validity set \( \hat{Z}_0 \) that is larger than \( Z_M \), i.e., \( \Pr(\hat{Z}_0 = \hat{Z}_0) \to 1 \). We prove that VSIV regression always leads to a smaller asymptotic bias than standard IV methods. Taken together, our theoretical results show that, irrespective of whether the largest validity pair set can be estimated consistently or not, VSIV regression leads to asymptotically normal IV estimators with reduced bias.

VSIV regression can be applied in many different settings. In the main text, we focus on the leading case of a binary treatment. In the Appendix, we extend our results to ordered treatments and also consider unordered treatments (Heckman and Pinto, 2018). Moreover, VSIV regression is generic—it can be used in conjunction with any set of testable restrictions. For example, if additional testable restrictions beyond those in Kitagawa (2015), Mourifié and Wan (2017), Kédagni and Mourifié (2020), and Sun (2021) are available, they could be used to refine the estimator of the validity pair set \( \hat{Z}_0 \) and further reduce the bias of VSIV regression.

**Notation.** We introduce some standard notation (e.g., Sun, 2021). All random elements are defined on a probability space \((\Omega, \mathcal{A}, \mathbb{P})\). For all \( m \in \mathbb{N} \), \( B_{\mathbb{R}^m} \) is the Borel \( \sigma \)-algebra on \( \mathbb{R}^m \). The symbol \( \rightsquigarrow \) denotes weak convergence in a metric space in the Hoffmann–Jørgensen sense. For a set \( D \), the space of bounded functions on \( D \) is \( \ell^\infty(D) \), where \( \ell^\infty(D) = \{ f : D \to \mathbb{R} : \| f \|_\infty < \infty \} \) and \( \| f \|_\infty = \sup_{x \in D} |f(x)| \). For every subset \( B \subset D \), let \( 1_B \) denote the indicator function for \( B \). For a topological space \( D \), let \( C(D) \) denote the set of real-valued continuous functions on \( D \). Finally, we adopt the convention (e.g., Folland, 1999, p. 45), that

\[
0 \cdot \infty = 0. \tag{1.1}
\]
2 Binary Treatments

2.1 Setup

Consider a setting with an outcome variable $Y \in \mathbb{R}$, a treatment $D \in \mathcal{D}$, and an instrument (vector) $Z \in \mathcal{Z}$. In this section, we focus on the leading case where the treatment is binary, $D \in \mathcal{D} = \{0, 1\}$. See Appendix A for extensions to multivalued ordered and unordered treatments. The instrument is discrete, $Z \in \mathcal{Z} = \{z_1, \ldots, z_K\}$, and can be ordered or unordered. Let $Y_{dz} \in \mathbb{R}$ for $(d, z) \in \mathcal{D} \times \mathcal{Z}$ denote the potential outcomes and let $D_z$ for $z \in \mathcal{Z}$ denote the potential treatments. The following assumption generalizes the standard LATE assumptions with binary instruments to multivalued instruments.

**Assumption 2.1** LATE assumptions with binary treatments:

(i) Exclusion: For each $d \in \{0, 1\}$, $Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ almost surely (a.s.).

(ii) Random Assignment: $Z$ is jointly independent of $(Y_{0z_1}, \ldots, Y_{0z_K}, Y_{1z_1}, \ldots, Y_{1z_K})$ and $(D_{z_1}, \ldots, D_{z_K})$.

(iii) Monotonicity: For all $k = 1, \ldots, K - 1$, $D_{zk+1} \geq D_{zk}$ a.s.

Assumption 2.1 does not include a first stage assumption. Throughout this paper, we maintain suitable first stage assumptions and focus on settings where exclusion, random assignment, or monotonicity are questionable. We thereby abstract from the issues arising from weak instruments. To lighten up the exposition, we keep the first stage assumptions implicit.

Assumption 2.1 is similar to the LATE assumptions in, for example, Imbens and Angrist (1994), Angrist and Imbens (1995), Fröhlich (2007), Kitagawa (2015), and Sun (2021). It imposes exclusion, random assignment, and monotonicity with respect to all possible values of the instrument $z \in \mathcal{Z}$, which can be restrictive in applications. Therefore, we introduce the notion of *pairwise instrument validity*, which weakens the conditions in Assumption 2.1. Define the set of all possible pairs of values of $Z$ as

$$\mathcal{Z} = \{(z_1, z_2), \ldots, (z_1, z_K), \ldots, (z_K, z_1), \ldots, (z_K, z_{K-1})\}.$$ 

The number of the elements in $\mathcal{Z}$ is $K \cdot (K - 1)$. We use $Z_{(k, k')}$ to denote a pair $(z_k, z_{k'}) \in \mathcal{Z}$. 

4
**Definition 2.1** The instrument \( Z \) is pairwise valid for the treatment \( D \in \{0, 1\} \) if there is a set \( \mathcal{Z}_M = \{(z_{k_1}, z_{k_1}'), \ldots, (z_{k_M}, z_{k_M}')\} \subset \mathcal{Z} \) such that the following conditions hold for every \((z, z') \in \mathcal{Z}_M:\)

(i) Exclusion: For each \( d \in \{0, 1\} \), \( Y_{dz} = Y_{dz'} \text{ a.s.} \)

(ii) Random Assignment: \( Z \) is jointly independent of \((Y_{0z}, Y_{0z'}, Y_{1z}, Y_{1z'}, D_z, D_{z'})\).

(iii) Monotonicity: \( D_z' \geq D_z \text{ a.s.} \)

The set \( \mathcal{Z}_M \) is called a validity pair set of \( Z \). The union of all validity pair sets is the largest validity pair set, denoted by \( \bar{\mathcal{Z}}_M \).

The following lemma shows that under pairwise instrument validity, particular treatment effects can be identified.

**Lemma 2.1** Suppose that the instrument \( Z \) is pairwise valid as defined in Definition 2.1 with a known validity pair set \( \mathcal{Z}_M = \{(z_{k_1}, z_{k_1}'), \ldots, (z_{k_M}, z_{k_M}')\} \). Then the following quantity can be identified for each \((z_{k_m}, z_{k_m}') \in \mathcal{Z}_M:\)

\[
\beta_{k_m', k_m} \equiv \frac{E[Y | Z = z_{k_m}'] - E[Y | Z = z_{k_m}]}{E[D | Z = z_{k_m}'] - E[D | Z = z_{k_m}]}.
\] (2.1)

Lemma 2.1 is a direct extension of Theorem 1 of Imbens and Angrist (1994) for the case where \( Z \) is pairwise valid. We follow Imbens and Angrist (1994) and refer to \( \beta_{k_m', k_m} \) as a LATE. Lemma 2.1 shows that if a validity pair set \( \mathcal{Z}_M \) is known, we can identify every \( \beta_{k_m', k_m} \) with \((z_{k_m}, z_{k_m}') \in \mathcal{Z}_M \). In practice, however, \( \mathcal{Z}_M \) is usually unknown. In this paper, we show how to identify and estimate the largest validity pair set \( \bar{\mathcal{Z}}_M \) based on testable restrictions for IV validity, and how to use this estimate to reduce the bias in IV estimation.

We focus on the LATEs \( \beta_{k_m', k_m} \) as our objects of interest. Traditional IV estimators yield weighted averages of LATEs (e.g., Imbens and Angrist, 1994) and, thus, are strictly less informative. Moreover, we can always compute linear IV estimands based on the LATEs.

**Remark 2.1 (Relationship between Pairwise Validity and Partial Monotonicity)** The partial monotonicity condition proposed by Mogstad et al. (2021) is a special case of condition (iii) in Definition 2.1. For example, suppose \( Z = (Z_1, Z_2) \in \mathbb{R}^2 \) and each element of \( Z \) is binary. Thus, \( \mathcal{Z} = \{(0, 0), (0, 1), (1, 0), (1, 1)\} \). Suppose that Assumption PM of Mogstad et al. (2021) holds with \( D_{(0,0)} \leq D_{(1,0)} \text{ a.s. and } D_{(0,0)} \leq D_{(0,1)} \text{ a.s.} \), and that conditions (i) and (ii) of Definition 2.1 hold. Then a validity pair set is \(((0, 0), (1, 0)), ((0, 0), (0, 1))\).
2.2 Validity Set IV Estimation

The largest validity pair set \( \mathcal{Z}_M \) is typically unknown in applications. In this paper, we propose a procedure for estimating \( \mathcal{Z}_M \). That is, we seek to identify and exclude \( (z_k, z_{k'}) \notin \mathcal{Z}_M \) from \( \mathcal{Z} \), since if \( (z_k, z_{k'}) \notin \mathcal{Z}_M \), then \( \beta_{k,k} \) defined in (2.1) is not equal to a LATE in general. Suppose that there are subsets \( \mathcal{Z}_1 \subset \mathcal{Z} \) and \( \mathcal{Z}_2 \subset \mathcal{Z} \) that satisfy the testable implications in Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021), and those in Kédagni and Mourifié (2020), respectively, which we will discuss in detail in Section 3.\(^4\) Then we let \( \mathcal{Z}_0 = \mathcal{Z}_1 \cap \mathcal{Z}_2 \) so that \( \mathcal{Z}_0 \) satisfies all the above necessary conditions. We first construct separate estimators \( \hat{\mathcal{Z}}_1 \) and \( \hat{\mathcal{Z}}_2 \) for \( \mathcal{Z}_1 \) and \( \mathcal{Z}_2 \), respectively, and then construct the estimator \( \hat{\mathcal{Z}}_0 \) for \( \mathcal{Z}_0 \) as \( \hat{\mathcal{Z}}_0 = \hat{\mathcal{Z}}_1 \cap \hat{\mathcal{Z}}_2 \). We refer to the IV estimators based on \( (z_k, z_{k'}) \in \mathcal{Z}_0 \) as VSIV estimators. In the following, we assume that suitable estimators \( \hat{\mathcal{Z}}_1, \hat{\mathcal{Z}}_2, \) and \( \hat{\mathcal{Z}}_0 \) are available. We discuss the construction of such estimators in Section 3.

If \( \hat{\mathcal{Z}}_0 \) is consistent for the largest validity pair set \( \mathcal{Z}_M \) in the sense that \( \mathbb{P}(\hat{\mathcal{Z}}_0 = \mathcal{Z}_M) \to 1 \), the proposed VSIV estimators are asymptotically unbiased and normal under weak and standard regularity conditions. We consider this case in Section 2.2.1. Since \( \hat{\mathcal{Z}}_0 \) is constructed based on the necessary conditions for the pairwise IV validity, \( \mathcal{Z}_0 \) could be larger than \( \mathcal{Z}_M \). In Section 2.2.2, we show that even if \( \mathcal{Z}_0 \) is larger than \( \mathcal{Z}_M \), VSIV estimators always yield bias reductions relative to standard IV estimators.

2.2.1 VSIV Regression under Consistent Estimation of the Validity Pair Set

Suppose that the estimator, \( \hat{\mathcal{Z}}_0 \), is consistent for the largest validity pair set \( \mathcal{Z}_M \), in the sense that \( \mathbb{P}(\hat{\mathcal{Z}}_0 = \mathcal{Z}_M) \to 1 \), and we use \( \hat{\mathcal{Z}}_0 \) to construct a VSIV estimator for the LATEs.

Suppose we have a random sample \( \{Y_i, D_i, Z_i\}_{i=1}^n \). Let \( \mathcal{I}(z, \mathcal{A}, \mathcal{S}) = 1\{z \in \mathcal{A}, \mathcal{A} \in \mathcal{S}\} \) for all \( z \), all \( \mathcal{A} \in \mathcal{Z} \), and all \( \mathcal{S} \subset \mathcal{Z} \). For every random variable \( \xi_i \) and every \( \mathcal{A} \in \mathcal{Z} \), we define

\[
\mathcal{E}_n(\xi_i, \mathcal{A}) = \frac{1}{n} \sum_{i=1}^n \xi_i I\left\{Z_i \in \mathcal{A}, \mathcal{A} \in \hat{\mathcal{Z}}_0\right\} \quad \text{and} \quad \mathcal{E}(\xi_i, \mathcal{A}) = \frac{E[\xi_i 1\{Z_i \in \mathcal{A}, \mathcal{A} \in \mathcal{Z}_M\}]}{E[1\{Z_i \in \mathcal{A}, \mathcal{A} \in \mathcal{Z}_M\}]}.
\]

Given the estimated validity set \( \hat{\mathcal{Z}}_0 \), for every \( \mathcal{Z}_{(k,k')} \in \mathcal{Z} \), we run the IV regression

\[
Y_i \mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \hat{\mathcal{Z}}_0) = \gamma^0_{(k,k')} \mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \hat{\mathcal{Z}}_0) + \gamma^1_{(k,k')} D_i \mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \hat{\mathcal{Z}}_0) + \epsilon_i \mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \hat{\mathcal{Z}}_0),
\]

\[(2.2)\]

\(^4\)As discussed in Sun (2021), in general, the testable implications in Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021), and those in Kédagni and Mourifié (2020) are complementary to each other.
using \( g(Z_i)I(Z_i, Z_{(k,k')} , \widehat{Z}_0) \) as the instrument for \( D_iI(Z_i, Z_{(k,k')} , \widehat{Z}_0) \), where \( g \) is a prespecified function that maps the value of \( Z_i \) to \( \mathbb{R} \). For example, we can simply set \( g(z) = z \) for all \( z \) if \( Z_i \) is a scalar instrument. Then we obtain the VSIV estimator for each LATE as

\[
\widehat{\beta}_{1(k,k')} = \frac{E_n \left( g \left( Z_i \right) Y_i, Z_{(k,k')} \right) - E_n \left( g \left( Z_i \right), Z_{(k,k')} \right) E_n \left( Y_i, Z_{(k,k')} \right)}{E_n \left( g \left( Z_i \right) D_i, Z_{(k,k')} \right) - E_n \left( g \left( Z_i \right), Z_{(k,k')} \right) E_n \left( D_i, Z_{(k,k')} \right)} ,
\]

which is the IV estimator of \( \gamma_{1(k,k')} \) in (2.2). We define

\[
\widehat{\beta}_1 = \left( \widehat{\beta}_{1(1,2)} , \ldots , \widehat{\beta}_{1(1,K)} , \ldots , \widehat{\beta}_{1(K,1)} , \ldots , \widehat{\beta}_{1(K,K-1)} \right) ,
\]

\[
\beta_{1(k,k')} = \frac{E \left( g \left( Z_i \right) Y_i, Z_{(k,k')} \right) - E \left( g \left( Z_i \right), Z_{(k,k')} \right) E \left( Y_i, Z_{(k,k')} \right)}{E \left( g \left( Z_i \right) D_i, Z_{(k,k')} \right) - E \left( g \left( Z_i \right), Z_{(k,k')} \right) E \left( D_i, Z_{(k,k')} \right)} ,
\]

and

\[
\beta_1 = \left( \beta_{1(1,2)} , \ldots , \beta_{1(1,K)} , \ldots , \beta_{1(K,1)} , \ldots , \beta_{1(K,K-1)} \right) .
\]

Note that if \( Z_{(k,k')} \not\in \mathcal{M} \), \( \beta^1_{1(k,k')} = 0 \) by (1.1). Similarly, if \( Z_{(k,k')} \not\in \widehat{\mathcal{M}} \), \( \widehat{\beta}^1_{1(k,k')} = 0 \) by (1.1).

To establish the theoretical properties of the VSIV estimators, we impose the following standard regularity conditions.

**Assumption 2.2** \( \{(Y_i, D_i, Z_i)\}_{i=1}^n \) is an i.i.d. sample.

**Assumption 2.3** The moments \( E[Y_i], E[D_i], E[g(Z_i)], E[g(Z_i)Y_i] \), and \( E[g(Z_i)D_i] \) exist.

The next theorem establishes the asymptotic distribution of the vector of VSIV estimators \( \widehat{\beta}_1 \), obtained based on the estimator of the instrument validity set \( \widehat{Z}_0 \).

**Theorem 2.1** Suppose that the instrument \( Z \) is pairwise valid for the treatment \( D \) as defined in Definition 2.1 with the largest validity pair set \( \mathcal{M} = \{(z_{k1}, z_{k1'}), \ldots , (z_{kM}, z_{kM'})\} \), and that the estimator \( \widehat{Z}_0 \) satisfies \( \mathbb{P} (\widehat{Z}_0 = \mathcal{M}) \rightarrow 1 \). Under Assumptions 2.2 and 2.3,

\[
\sqrt{n} (\widehat{\beta}_1 - \beta_1) \xrightarrow{d} N \left( 0, \Sigma \right) ,
\]

where \( \Sigma \) is defined in (C.2) in the Appendix. In addition, \( \beta^1_{1(k,k')} = \beta_{k',k} \) as defined in (2.1) for every \( (z_k, z_{k'}) \in \mathcal{M} \).

If a validity pair set is known, we can use it as the estimator for itself and run the VSIV regression. Theorem 2.1 establishes the joint asymptotic normality of the VSIV estimator of
the LATEs. The asymptotic covariance matrix $\Sigma$ defined in the Appendix can be consistently estimated under standard conditions. Importantly, the estimation of the instrument validity pair set does not affect the asymptotic covariance matrix such that standard inference methods can be applied.

### 2.2.2 Bias Reduction using VSIV Regression

In Section 2.2.1, we show that if the estimator of the validity set is consistent, $\mathbb{P}(\hat{Z}_0 = \mathcal{Z}_M) \to 1$, VSIV estimators are consistent for LATEs under weak conditions. However, since $\mathcal{Z}_0$ is constructed based on necessary (but not necessarily sufficient) conditions for IV validity, in general we have $\mathbb{P}(\hat{Z}_0 = Z_0) \to 1$, where the pseudo-validity set $\hat{Z}_0$ could be larger than $\mathcal{Z}_M$. In this case, VSIV is not asymptotically unbiased in general. Here we show that even if $\hat{Z}_0$ is larger than $\mathcal{Z}_M$, the VSIV estimators always reduce the bias relative to standard IV estimators. Intuitively, VSIV estimators use the information in the data about IV validity to reduce the asymptotic bias as much as possible.

Since our target parameter is the vector $\beta_1$, a natural definition of the estimation bias is $\|\hat{\beta}_1 - \beta_1\|_2$ for every estimator $\hat{\beta}_1$.

**Definition 2.2** The estimation bias of an arbitrary estimator $\hat{\beta}_1$ for the true value $\beta_1$ defined in (2.5) is defined as $\|\hat{\beta}_1 - \beta_1\|_2$, where $\| \cdot \|_2$ is the $\ell^2$-norm on Euclidean spaces.

Consider an arbitrary presumed validity pair set $\mathcal{Z}_P$, which could incorporate prior information. If no prior information is available, $\mathcal{Z}_P = \mathcal{Z}$. Given $\mathcal{Z}_P$, we define $\hat{\mathcal{Z}}_0' = \hat{\mathcal{Z}}_0 \cap \mathcal{Z}_P$ and use $\hat{\mathcal{Z}}_0'$ to construct the VSIV estimators in (2.2).

The following theorem shows that the VSIV estimators based on $\hat{\mathcal{Z}}_0'$ always exhibit a smaller asymptotic bias than standard IV estimators based on $\mathcal{Z}_P$.

**Theorem 2.2** Suppose $\mathbb{P}(\hat{Z}_0 = Z_0) \to 1$ with $\mathcal{Z}_0 \supset \mathcal{Z}_M$. For every presumed validity pair set $\mathcal{Z}_P$, the asymptotic estimation bias $\text{plim}_{n \to \infty} \|\hat{\beta}_1 - \beta_1\|_2$ is always reduced by using $\hat{\mathcal{Z}}_0'$ in the regression (2.2) compared to that from using $\mathcal{Z}_P$.

As shown later in Propositions 3.1 and 3.2, the pseudo-validity pair set $\hat{\mathcal{Z}}_0$ can always be estimated consistently by $\hat{\mathcal{Z}}_0$ under mild conditions. Compared to constructing standard IV estimators based on $\mathcal{Z}_P$, Theorem 2.2 shows that the asymptotic estimation bias, $\text{plim}_{n \to \infty} \|\hat{\beta}_1 - \beta_1\|_2$, can be reduced by using VSIV estimators based on $\hat{\mathcal{Z}}_0' = \hat{\mathcal{Z}}_0 \cap \mathcal{Z}_P$.

The arguments used for establishing the asymptotic normality of the VSIV estimators in Section 2.2.1 do not rely on the consistent estimation of $\mathcal{Z}_M$. Thus, irrespective of whether
\( \mathcal{Z}_M \) can be estimated consistently, the VSIV estimators are asymptotically normal, centered at \( \beta_1 \) defined with \( \mathcal{Z}_0 \) instead of \( \mathcal{Z}_M \). However, note that \( \beta_1 \) can only be interpreted as a vector of LATEs under consistent estimation.

**Example 2.1 (Bias Reduction using VSIV Regression)** Consider a simple example where \( \mathcal{Z} = \{1, 2, 3, 4\} \) as in our application and suppose that \( \mathcal{Z}_M = \{(1, 2)\} \). In this case, by (2.4) and (1.1),

\[
\beta_1 = (\beta_{(1,2)}^1, \ldots, \beta_{(1,4)}^1, \ldots, \beta_{(4,1)}^1, \ldots, \beta_{(4,3)}^1) = (\beta_{(1,2)}^1, 0, \ldots, 0).
\]

Suppose that, by mistake, we assume \( \bar{Z} \) is valid according to Assumption 2.1 and use

\[
\mathcal{Z}_P = \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}
\]

as an estimator for \( \mathcal{Z}_M \). Then by (2.3) and (1.1),

\[
\hat{\beta}_1 = (\hat{\beta}_{(1,2)}^1, \hat{\beta}_{(1,3)}^1, \hat{\beta}_{(1,4)}^1, 0, \hat{\beta}_{(2,3)}^1, \hat{\beta}_{(2,4)}^1, 0, 0, \hat{\beta}_{(3,4)}^1, 0, 0, 0),
\]

(2.6)

where \( \hat{\beta}_{(1,3)}^1, \hat{\beta}_{(1,4)}^1, \hat{\beta}_{(2,3)}^1, \hat{\beta}_{(2,4)}^1, \) and \( \hat{\beta}_{(3,4)}^1 \) may not converge to 0 in probability. However, by definition \( \beta_{(1,3)}^1 = 0, \beta_{(1,4)}^1 = 0, \beta_{(2,3)}^1 = 0, \beta_{(2,4)}^1 = 0, \) and \( \beta_{(3,4)}^1 = 0 \). Thus, the bias \( \| \hat{\beta}_1 - \beta_1 \|_2 \) may not converge to 0 in probability. The approach proposed in this paper helps reduce this bias as much as possible. We exploit the information in the data about IV validity to obtain the estimator \( \hat{\mathcal{Z}}_0 \). Even if \( \hat{\mathcal{Z}}_0 \) converges to a set larger than \( \mathcal{Z}_M \) (because we use the necessary but not sufficient conditions for IV validity), VSIV always reduces the bias. Suppose that our estimator \( \hat{\mathcal{Z}}_0 = \mathcal{Z}_0 = \{(1, 2), (3, 4)\} \), which is larger than \( \mathcal{Z}_M \) but smaller than \( \mathcal{Z}_P \). In this case,

\[
\hat{\beta}_1 = (\hat{\beta}_{(1,2)}^1, 0, 0, 0, 0, 0, 0, 0, \hat{\beta}_{(3,4)}^1, 0, 0, 0).
\]

(2.7)

Note that \( \hat{\beta}_{(1,2)}^1 \) and \( \hat{\beta}_{(3,4)}^1 \) in (2.6) and (2.7) are the same by (2.3), because \( (1, 2), (3, 4) \in \mathcal{Z}_P \cap \mathcal{Z}_0 \). Thus, VSIV reduces the probability limit of the bias \( \| \hat{\beta}_1 - \beta_1 \|_2 \).

### 2.3 Partially Valid Instruments and Connection to Existing Results

Suppose we estimate the following canonical IV regression model,

\[
Y_i = \alpha_0 + \alpha_1 D_i + \varepsilon_i,
\]

(2.8)
using \( g(Z_i) \) as the instrument for \( D_i \). When the instrument \( Z \) is fully valid, the traditional IV estimator of \( \alpha_1 \) is
\[
\hat{\alpha}_1 = \frac{n \sum_{i=1}^{n} g(Z_i) Y_i - \sum_{i=1}^{n} g(Z_i) \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} g(Z_i) D_i - \sum_{i=1}^{n} g(Z_i) \sum_{i=1}^{n} D_i}.
\] (2.9)

The asymptotic properties of \( \hat{\alpha}_1 \) can be found in Imbens and Angrist (1994, p. 471) and Angrist and Imbens (1995, p. 436).

To connect VSIV regression to canonical IV regression with fully valid instruments, consider the following special case of pairwise IV validity.

**Definition 2.3** Suppose that the instrument \( Z \) is pairwise valid for the treatment \( D \) with the largest validity pair set \( Z_M \). If there is a validity pair set
\[
Z_M = \{(z_{k_1}, z_{k_2}), (z_{k_2}, z_{k_3}), \ldots, (z_{k_{M-1}}, z_{k_M})\}
\]
for some \( M > 0 \), then the instrument \( Z \) is called a **partially valid instrument** for the treatment \( D \). The set \( Z_M = \{z_{k_1}, \ldots, z_{k_M}\} \) is called a **validity value set** of \( Z \).

Suppose that \( Z \) is partially valid for the treatment \( D \) with a validity value set \( Z_M \), and that there is a consistent estimator \( \hat{Z}_0 \) of \( Z_M \). We then construct a VSIV estimator for \( \alpha_1 \) in (2.8) by running the IV estimation for the model
\[
Y_i 1 \left\{ Z_i \in \hat{Z}_0 \right\} = \gamma_0 1 \left\{ Z_i \in \hat{Z}_0 \right\} + \gamma_1 D_i 1 \left\{ Z_i \in \hat{Z}_0 \right\} + \epsilon_i 1 \left\{ Z_i \in \hat{Z}_0 \right\},
\] (2.10)

using \( g(Z_i) 1 \{Z_i \in \hat{Z}_0\} \) as the instrument for \( D_i 1 \{Z_i \in \hat{Z}_0\} \). We obtain the VSIV estimator for \( \alpha_1 \) in (2.8) by
\[
\hat{\theta}_1 = \frac{n_z \sum_{i=1}^{n} g(Z_i) Y_i 1 \left\{ Z_i \in \hat{Z}_0 \right\} - \sum_{i=1}^{n} g(Z_i) 1 \left\{ Z_i \in \hat{Z}_0 \right\} \sum_{i=1}^{n} Y_i 1 \left\{ Z_i \in \hat{Z}_0 \right\}}{n_z \sum_{i=1}^{n} g(Z_i) D_i 1 \left\{ Z_i \in \hat{Z}_0 \right\} - \sum_{i=1}^{n} g(Z_i) 1 \left\{ Z_i \in \hat{Z}_0 \right\} \sum_{i=1}^{n} D_i 1 \left\{ Z_i \in \hat{Z}_0 \right\}},
\] (2.11)

where \( n_z = \sum_{i=1}^{n} 1 \{Z_i \in \hat{Z}_0\} \). We can see that \( \hat{\theta}_1 \) is a generalized version of \( \hat{\alpha}_1 \) in (2.9), because when the instrument is fully valid, we can just let \( \hat{Z}_0 = Z \) and then \( \hat{\theta}_1 = \hat{\alpha}_1 \).

**Theorem 2.3** Suppose that the instrument \( Z \) is partially valid for the treatment \( D \) according to Definition 2.3 with a validity value set \( Z_M = \{z_{k_1}, \ldots, z_{k_M}\} \), and that the estimator \( \hat{Z}_0 \) for
\( Z_M \) satisfies \( \mathbb{P}(\hat{Z}_0 = Z_M) \to 1 \). Under Assumptions 2.2 and 2.3, it follows that \( \hat{\theta}_1 \overset{p}{\to} \theta_1 \), where

\[
\theta_1 = \frac{E[g(Z_i)Y_i|Z_i \in Z_M] - E[Y_i|Z_i \in Z_M]E[g(Z_i)|Z_i \in Z_M]}{E[g(Z_i)D_i|Z_i \in Z_M] - E[D_i|Z_i \in Z_M]E[g(Z_i)|Z_i \in Z_M]}
\]

Also, \( \sqrt{n}(\hat{\theta}_1 - \theta_1) \overset{d}{\to} N(0, \Sigma_1) \), where \( \Sigma_1 \) is provided in (C.17) in the Appendix. In addition, the quantity \( \theta_1 \) can be interpreted as the weighted average of \( \{\beta_{k_2,k_1}, \ldots, \beta_{k_M,k_{M-1}}\} \) defined as in (2.1). Specifically, \( \theta_1 = \sum_{m=1}^{M-1} \mu_m \beta_{km+1,k_m} \) with

\[
\mu_m = \frac{[p(z_{km+1}) - p(z_{km})] \sum_{i=m}^{M-1} \mathbb{P}(Z_i = z_{ki+1}|Z_i \in Z_M) \{g(z_{ki+1}) - E[g(Z_i)|Z_i \in Z_M]\}}{\sum_{i=1}^{M} \mathbb{P}(Z_i = z_k|Z_i \in Z_M)p(z_k)\{g(z_k) - E[g(Z_i)|Z_i \in Z_M]\}},
\]

\( p(z_k) = E[D_i|Z_i = z_k] \), and \( \sum_{m=1}^{M-1} \mu_m = 1 \).

Theorem 2.3 is an extension of Theorem 2 of Imbens and Angrist (1994) for the case where the instrument is partially but not fully valid.

To establish a connection to existing results, Theorem 2.3 assumes consistent estimation of the validity value set, \( \mathbb{P}(\hat{Z}_0 = Z_M) \to 1 \). If \( \hat{Z}_0 \) converges to a larger set than \( Z_M \), the properties of VSIV follow from the results in Section 2.2.2, because partially valid instruments are a special case of pairwise valid instruments.

### 3 Estimation of \( Z_0 \)

Here we discuss the construction of the estimators of \( Z_1 \) based on the testable implications in Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021) and the estimators of \( Z_2 \) based on the testable implications in Kédagni and Mourifié (2020). We show that under weak assumptions, these estimators are consistent in the sense that \( \mathbb{P}(\hat{Z}_1 = Z_1) \to 1 \) and \( \mathbb{P}(\hat{Z}_2 = Z_2) \to 1 \). These results imply that \( \hat{Z}_0 \) is a consistent estimator of the pseudo-true validity set \( Z_0 \), \( \mathbb{P}(\hat{Z}_0 = Z_0) \to 1 \). As a consequence, when \( Z_0 = Z_M \), the largest validity pair set can be estimated consistently.

#### 3.1 Definition and Estimation of \( Z_1 \)

The definition of \( Z_1 \) relies on the testable implications proposed in Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021). We use the notation of Sun (2021) to introduce
these testable restrictions. Define conditional probabilities
\[
P_z (B, C) = \mathbb{P} (Y \in B, D \in C | Z = z)
\]
for all Borel sets \( B, C \in \mathcal{B}_{\mathbb{R}} \) and all \( z \in Z \). With the largest validity pair set \( \mathcal{Z}_M = \{(z_{k_1}, z'_{k_1}), \ldots, (z_{k_M}, z'_{k_M})\} \), for every \( m \in \{1, \ldots, M\} \),\(^5\) it follows that
\[
P_{z_{km}} (B, \{1\}) \leq P_{z'_{km}} (B, \{1\}) \quad \text{and} \quad P_{z_{km}} (B, \{0\}) \geq P_{z'_{km}} (B, \{0\})
\]
for all \( B \in \mathcal{B}_{\mathbb{R}} \). By definition, for all \( B, C \in \mathcal{B}_{\mathbb{R}} \),
\[
\mathbb{P} (Y \in B, D \in C | Z = z) = \frac{\mathbb{P} (Y \in B, D \in C, Z = z)}{\mathbb{P} (Z = z)}.
\]
Define the function spaces
\[
\mathcal{G}_P = \left\{ (1_{\mathbb{R}^3 \times \{z_k\}}, 1_{\mathbb{R}^3 \times \{z_{k'}\}}) : k, k' \in \{1, \ldots, K\}, k \neq k' \right\},
\]
\[
\mathcal{H} = \left\{ (-1)^d \cdot 1_{B \times \{d\} \times \mathbb{R}} : B \text{ is a closed interval in } \mathbb{R}, d \in \{0, 1\} \right\}, \quad \text{and}
\]
\[
\tilde{\mathcal{H}} = \left\{ (-1)^d \cdot 1_{B \times \{d\} \times \mathbb{R}} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}, d \in \{0, 1\} \right\}.
\]
Similarly to Sun (2021), by Lemma B.7 in Kitagawa (2015), we use all closed intervals \( B \subset \mathbb{R} \) to construct \( \mathcal{H} \) instead of all Borel sets.

We denote by \( \mathcal{P} \) the set of probability measures on \( (\mathbb{R}^3, \mathcal{B}_{\mathbb{R}^3}) \). Suppose we have access to an i.i.d. sample \( \{(Y_i, D_i, Z_i)\}_{i=1}^n \) distributed according to some probability distribution \( P \) in \( \mathcal{P} \), that is, \( P(G) = \mathbb{P}((Y_i, D_i, Z_i) \in G) \) for all \( G \in \mathcal{B}_{\mathbb{R}^3} \). For every measurable function \( v \), with some abuse of notation, define
\[
P(v) = \int v \, dP.
\]
The closure of \( \mathcal{H} \) in \( L^2(P) \) is equal to \( \tilde{\mathcal{H}} \) by Lemma C.1 of Sun (2021). For every \( (h, g) \in \tilde{\mathcal{H}} \times \mathcal{G}_P \) with \( g = (g_1, g_2) \), define
\[
\phi(h, g) = \frac{P(h \cdot g_2)}{P(g_2)} - \frac{P(h \cdot g_1)}{P(g_1)}.
\]
\(^5\)The testable implications proposed by Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021) are originally for full instrument validity. We can easily obtain the testable implications for the conditions in Definition 2.1 following the proof of Kitagawa (2015).
and
\[
\sigma^2(h, g) = \Lambda(P) \cdot \left\{ \frac{P(h^2 \cdot g_2)}{P^2(g_2)} - \frac{P^2(h \cdot g_2)}{P^3(g_2)} + \frac{P(h^2 \cdot g_1)}{P^2(g_1)} - \frac{P^2(h \cdot g_1)}{P^3(g_1)} \right\}, \tag{3.3}
\]
where \( \Lambda(P) = \prod_{k=1}^{K} P(1_{\mathbb{R} \times \mathbb{R} \times \{ z_k \}}) \) and \( P^m(g_j) = [P(g_j)]^m \) for \( m \in \mathbb{N} \) and \( j \in \{1, 2\} \). We denote the sample analog of \( \phi \) as
\[
\hat{\phi}(h, g) = \frac{\hat{P}(h \cdot g_2)}{\hat{P}(g_2)} - \frac{\hat{P}(h \cdot g_1)}{\hat{P}(g_1)},
\]
where \( \hat{P} \) is the empirical probability measure corresponding to \( P \) so that for every measurable function \( v \),
\[
\hat{P}(v) = \frac{1}{n} \sum_{i=1}^{n} v(Y_i, D_i, Z_i). \tag{3.4}
\]
For every \((h, g) \in \mathcal{H} \times \mathcal{G}_P\) with \( g = (g_1, g_2) \), define the sample analog of \( \sigma^2(h, g) \) as
\[
\hat{\sigma}^2(h, g) = \frac{T_n}{n} \cdot \left\{ \frac{\hat{P}(h^2 \cdot g_2)}{\hat{P}^2(g_2)} - \frac{\hat{P}^2(h \cdot g_2)}{\hat{P}^3(g_2)} + \frac{\hat{P}(h^2 \cdot g_1)}{\hat{P}^2(g_1)} - \frac{\hat{P}^2(h \cdot g_1)}{\hat{P}^3(g_1)} \right\},
\]
where \( T_n = n \cdot \prod_{k=1}^{K} \hat{P}(1_{\mathbb{R} \times \mathbb{R} \times \{ z_k \}}) \). By (1.1), \( \hat{\sigma}^2 \) is well defined. By similar proof of Lemma 3.1 in Sun (2021), \( \sigma^2 \) and \( \hat{\sigma}^2 \) are uniformly bounded in \((h, g)\). The following lemma reformulates the testable restrictions in terms of \( \phi \). Below, we use this reformulation to define \( \mathcal{Z}_1 \) and the corresponding estimator \( \hat{\mathcal{Z}}_1 \).

**Lemma 3.1** Suppose that the instrument \( Z \) is pairwise valid for the treatment \( D \) with the largest validity pair set \( \mathcal{Z}_M = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_M}, z_{k'_M})\} \). For every \( m \in \{1, \ldots, M\} \), we have that \( \sup_{h \in \mathcal{H}} \hat{\phi}(h, g) = 0 \) with \( g = (1_{\mathbb{R} \times \mathbb{R} \times \{ z_{k_m} \}}, 1_{\mathbb{R} \times \mathbb{R} \times \{ z_{k'_m} \}}) \).

Lemma 3.1 provides a necessary condition based on Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021) for the validity pair set \( \mathcal{Z}_M \). Define
\[
\mathcal{G}_1 = \left\{ g \in \mathcal{G}_P : \sup_{h \in \mathcal{H}} \phi(h, g) = 0 \right\} \quad \text{and} \quad \hat{\mathcal{G}}_1 = \left\{ g \in \mathcal{G}_P : \sqrt{T_n} \sup_{h \in \mathcal{H}} \left| \frac{\hat{\phi}(h, g)}{\xi_0 \sqrt{\hat{\sigma}(h, g)}} \right| \leq \tau_n \right\}, \tag{3.5}
\]
where \( \tau_n \to \infty \) with \( \tau_n / \sqrt{n} \to 0 \) as \( n \to \infty \), and \( \xi_0 \) is a small positive number.\(^7\) The set \( \hat{\mathcal{G}}_1 \) is different from the contact sets defined in Beare and Shi (2019), Sun and Beare (2021), and Sun (2021) in independent contexts, because of the map \( \sup \). A further discussion about

\(^5\)Lemma 3.1 states the conditions in (3.1) in terms of \( \phi \).
\(^6\)In practice, we use \( \xi_0 = 0.001 \).
the estimation of contact sets can be found in Linton et al. (2010) and Lee et al. (2013). Define \( \mathcal{Z}_1 \) as the collection of all \((z, z')\) associated with some \( g \in G_1 \):

\[
\mathcal{Z}_1 = \{(z_k, z_{k'}) \in \mathcal{Z} : g = (1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_{k'}\}}) \in G_1\}.
\]  

(3.6)

For example, if \( K = 4 \) and \( G_1 = \{(1_{\mathbb{R} \times \mathbb{R} \times \{z_1\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_2\}}), (1_{\mathbb{R} \times \mathbb{R} \times \{z_3\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_4\}})\} \), then \( \mathcal{Z}_1 = \{(z_1, z_2), (z_3, z_4)\} \). By Lemma 3.1, \( \mathcal{Z}_1 \subset \mathcal{Z} \). We use \( \hat{G}_1 \) to construct the estimator of \( \mathcal{Z}_1 \), denoted by \( \hat{\mathcal{Z}}_1 \), which is defined as the set of all \((z, z')\) associated with some \( g \in \hat{G}_1 \):

\[
\hat{\mathcal{Z}}_1 = \{(z_k, z_{k'}) \in \mathcal{Z} : g = (1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_{k'}\}}) \in \hat{G}_1\}.
\]  

(3.7)

Note that (3.7) is the sample analog of (3.6). The following proposition establishes consistency of \( \hat{\mathcal{Z}}_1 \).

**Proposition 3.1** Under Assumptions 2.2 and 2.3, \( P(\hat{G}_1 = G_1) \to 1 \), and thus \( P(\hat{\mathcal{Z}}_1 = \mathcal{Z}_1) \to 1 \).

Proposition 3.1 is related to the contact set estimation in Sun (2021). Since by definition, \( G_1 \subset G_P \) and \( G_P \) is a finite set, we can use techniques similar to those in Sun (2021) to obtain the stronger result in Proposition 3.1, that is, \( P(\hat{G}_1 = G_1) \to 1 \).

### 3.2 Definition and Estimation of \( \mathcal{Z}_2' \)

The definition of \( \mathcal{Z}_2' \) relies on the testable implications in Kédagni and Mourifié (2020) for the exclusion restriction \( Y_{dzk_m} = Y_{dk_m} \) for \( d \in \{0, 1\} \) and the independence condition \((Y_{0zk_m}, Y_{0zk_m}', Y_{1zk_m}, Y_{1zk_m}') \perp Z)\) for every \( m \in \{1, \ldots, M\} \) with the largest validity pair set \( \mathcal{Z}_{2021}' = \{(z_{k_1}, z_{k_1'}), \ldots, (z_{k_M}, z_{k_M'})\} \). To alleviate the exposition, we define \( Y_d(z, z') \) for each \( d \in \{0, 1\} \) and every \((z, z') \in \mathcal{Z}_{2021}' \) so that \( Y_d(z, z') = Y_{dz} = Y_{dz'} \) a.s.

We consider the case where \( Y \) is continuous. Similar results can be obtained when \( Y \) is discrete. To avoid theoretical and computational complications, we introduce the following testable implications that are slightly weaker than (and implied by) those in Kédagni and Mourifié (2020) but simplify computation significantly.\(^8\) We present the general testable implications of Kédagni and Mourifié (2020) in Appendix C.2.

We start by describing the testable implications using the notation by Kédagni and Mourifié (2020). Let \( R \) denote the collection of all subsets \( C \subset \mathbb{R} \) such that \( C = (a, b] \) with

---

\(^8\)The estimation of the density functions in the testable implications of Kédagni and Mourifié (2020) may involve kernel estimation and bandwidth selection. Also, we would need the estimator to be consistent uniformly under high-level assumptions, which could cause further technical complications.
For every $Z_{(k,k')} \in \mathcal{Z}_M$, every $A \in \mathcal{R}$, each $d \in \mathcal{D}$, and each $z \in Z_{(k,k')}$,

$$
P(Y \in A, D = d|Z = z) \leq P(Y_d(z_k, z_{k'}) \in A|Z = z) = P(Y_d(z_k, z_{k'}) \in A)
$$

which implies that

$$
\max_{z \in Z_{(k,k')}} P(Y \in A, D = d|Z = z) \leq P(Y_d(z_k, z_{k'}) \in A).
$$

(3.8)

Let $\mathcal{P}$ be a prespecified finite\(^9\) collection of partitions $P_{\mathbb{R}}$ of $\mathbb{R}$ such that $P_{\mathbb{R}} = \{C_1, \ldots, C_N\}$ with $C_k \in \mathcal{R}$ for all $k$, $\cup_{k=1}^{N} C_k = \mathbb{R}$ and $C_k \cap C_l = \emptyset$ for all $k \neq l$. Then we obtain the first condition that

$$
\max_{P_k \in \mathcal{P}} \max_{d \in \mathcal{D}} \sum_{A \in P_k} \max_{z \in Z_{(k,k')}} P(Y \in A, D = d|Z = z) \leq \max_{P_k \in \mathcal{P}} \max_{d \in \mathcal{D}} \sum_{A \in P_k} P(Y_d(z_k, z_{k'}) \in A) = 1.
$$

(3.9)

Also, for all $A_0, A_1 \in \mathcal{B}_{\mathbb{R}},$

$$
P(Y_0(z_k, z_{k'}) \in A_0, Y_1(z_k, z_{k'}) \in A_1) = \min_{z \in Z_{(k,k')}} P(Y_0(z_k, z_{k'}) \in A_0, Y_1(z_k, z_{k'}) \in A_1|Z = z)
$$

$$
= \min_{z \in Z_{(k,k')}} \sum_{d \in \mathcal{D}} P(Y_0(z_k, z_{k'}) \in A_0, Y_1(z_k, z_{k'}) \in A_1, D = d|Z = z)
$$

$$
\leq \min_{z \in Z_{(k,k')}} \sum_{d \in \mathcal{D}} P(Y \in A_d, D = d|Z = z).
$$

Let $P_0^0, P_1^1 \in \mathcal{P}$. It follows that

$$
1 = \sum_{A_0 \in P_0^0} \sum_{A_1 \in P_1^1} P(Y_0(z_k, z_{k'}) \in A_0, Y_1(z_k, z_{k'}) \in A_1)
$$

$$
\leq \sum_{A_0 \in P_0^0} \sum_{A_1 \in P_1^1} \min_{z \in Z_{(k,k')}} \sum_{d \in \mathcal{D}} P(Y \in A_d, D = d|Z = z).
$$

Then we obtain the second condition that

$$
\min_{P_k \in \mathcal{P}} \sum_{A_0 \in P_0^0} \sum_{A_1 \in P_1^1} \min_{z \in Z_{(k,k')}} \sum_{d \in \mathcal{D}} P(Y \in A_d, D = d|Z = z) \geq 1.
$$

(3.10)

\(^9\)As discussed in Kédagni and Mourifié (2020, p. 666), their testable implications involve the supremum and the infimum over all partitions of $\mathbb{R}$, which may cause empirical and theoretical complications. Kédagni and Mourifié (2020) suggest that some choices of the partitions could be made in practice. We follow this idea and set $\mathcal{P}$ to be a prespecified finite set.
Next, for all $A_0, A_1 \in \mathcal{B}_\mathbb{R}$,
\[
\mathbb{P}(Y_0(z_k, z_{k'}) \in A_0) = \sum_{A_1 \in \mathcal{P}_R^1} \mathbb{P}(Y_0(z_k, z_{k'}) \in A_0, Y_1(z_k, z_{k'}) \in A_1) \leq \sum_{A_1 \in \mathcal{P}_R^1} \min_{z \in \mathcal{Z}(k, k')} \sum_{d \in \mathcal{D}} \mathbb{P}(Y \in A_d, D = d | Z = z),
\]
and
\[
\mathbb{P}(Y_1(z_k, z_{k'}) \in A_1) = \sum_{A_0 \in \mathcal{P}_R^0} \mathbb{P}(Y_0(z_k, z_{k'}) \in A_0, Y_1(z_k, z_{k'}) \in A_1) \leq \sum_{A_0 \in \mathcal{P}_R^0} \min_{z \in \mathcal{Z}(k, k')} \sum_{d \in \mathcal{D}} \mathbb{P}(Y \in A_d, D = d | Z = z),
\]
which, together with (3.8), imply
\[
\max_{P^0_\mathcal{R}, P^1_\mathcal{R} \in \mathcal{P}} \max_{d \in \mathcal{D}} \sup_{A_d \in \mathbb{R}} \sum_{z \in \mathcal{Z}(k, k')} \mathbb{P}(Y \in A_d, D = d | Z = z) - \varphi_d(A_d, \mathcal{Z}(k, k'), P^0_\mathcal{R}, P^1_\mathcal{R}) \leq 0, \tag{3.11}
\]
where
\[
\varphi_0(A_0, W, P^0_\mathcal{R}, P^1_\mathcal{R}) = \sum_{A_1 \in \mathcal{P}_R^1} \min_{z \in W} \sum_{d \in \mathcal{D}} \mathbb{P}(Y \in A_d, D = d | Z = z)
\]
and
\[
\varphi_1(A_1, W, P^0_\mathcal{R}, P^1_\mathcal{R}) = \sum_{A_0 \in \mathcal{P}_R^0} \min_{z \in W} \sum_{d \in \mathcal{D}} \mathbb{P}(Y \in A_d, D = d | Z = z)
\]
for all $W \subset \mathbb{Z}$.

We now present a reformulation of the testable implications in (3.9)–(3.11) similar to the reformulation in Section 3.1. We use this reformulation to define $\mathcal{Z}_2$ and the corresponding estimator $\hat{\mathcal{Z}}_2$. Define the function spaces
\[
\mathcal{G}_\mathcal{Z} = \{1_{\mathbb{R} \times \{z_k\}} : 1 \leq k \leq K\}, \mathcal{H}_\mathcal{D} = \{1_{\mathbb{R} \times \{d\} \times \mathcal{D}} : d \in \mathcal{D}\}, \mathcal{H}_\mathcal{B} = \{1_{\mathbb{R} \times \mathbb{R} : B \in \mathbb{R}} : B \in \mathbb{B}\},
\]
and $\mathcal{H}_\mathcal{B} = \{1_{\mathbb{B} \times \mathbb{R} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}}\}.$ \tag{3.12}

Define a map $\psi : \mathcal{H}_\mathcal{B} \times \mathcal{H}_\mathcal{D} \times \mathcal{G}_\mathcal{Z} \rightarrow \mathbb{R}$ such that
\[
\psi(h, f, g) = \frac{P(h \cdot f \cdot g)}{P(g)} \tag{3.13}
\]
for every $(h, f, g) \in \mathcal{H}_\mathcal{B} \times \mathcal{H}_\mathcal{D} \times \mathcal{G}_\mathcal{Z}$. Moreover, define a map $\mathcal{H}$ such that if $P_\mathcal{R} \in \mathcal{P}$ with
\( P_R = \{ C_1, \ldots, C_N \} \) and \( C_k \in \mathcal{R} \) for all \( k \in \{ 1, \ldots, N \} \), then

\[
\mathfrak{M}(P_R) = \{ 1_{C \times \mathbb{R} \times \mathbb{R}} : C \in P_R \}.
\] (3.14)

Let \( P(\mathcal{G}_Z) \) denote the collection of all nonempty subsets of \( \mathcal{G}_Z \). Then for every \( \mathcal{G}_S \in P(\mathcal{G}_Z) \), define

\[
\psi_1(\mathcal{G}_S) = \max_{p_k \in \mathcal{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in \mathfrak{M}(P_k)} \max_{g \in \mathcal{G}_S} \psi(h, f, g) - 1,
\]

\[
\psi_2(\mathcal{G}_S) = 1 - \min_{p_k \in \mathcal{P}, h_1 \in \mathfrak{M}(P_k)} \sum_{h_0 \in \mathfrak{M}(P_k^0)} \sum_{h_1 \in \mathfrak{M}(P_k^1)} \min_{g \in \mathcal{G}_S} \sum_{d \in \mathcal{D}} \psi(h_d, f_d, g),
\]

and

\[
\psi_3(\mathcal{G}_S) = \max_{p_k \in \mathcal{P}} \max_{a \in \mathcal{A}} \sup_{h \in \mathcal{H}_B} \left\{ \max_{g \in \mathcal{G}_S} \psi(h, f, g) - \bar{\varphi}_d(h_d, \mathcal{G}_S, p_k^0, p_k^1) \right\},
\]

where \( f_d = 1_{\mathbb{R} \times \{d\} \times \mathbb{R}} \),

\[
\bar{\varphi}_0(h_0, \mathcal{G}_S, p_k^0, p_k^1) = \sum_{h_1 \in \mathfrak{M}(P_k)} \min_{g \in \mathcal{G}_S} \sum_{d \in \mathcal{D}} \psi(h_d, f_d, g), \text{ and}
\]

\[
\bar{\varphi}_1(h_1, \mathcal{G}_S, p_k^0, p_k^1) = \sum_{h_0 \in \mathfrak{M}(P_k^0)} \min_{g \in \mathcal{G}_S} \sum_{d \in \mathcal{D}} \psi(h_d, f_d, g).
\]

For every \( Z_{(k,k')} \in \mathcal{Z}_M \), define the set \( \mathcal{G}(Z_{(k,k')}) \) as

\[
\mathcal{G}(Z_{(k,k')}) = \{ (1_{\mathbb{R} \times \mathbb{R} \times \{x_k\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{x_{k'}\}}) \}.
\]

The conditions in (3.9)–(3.11) imply that \( \psi_l(\mathcal{G}(Z_{(k,k')})) \leq 0 \) for all \( l \in \{ 1, 2, 3 \} \). Thus, we define \( \mathcal{Z}_2 \) by

\[
\mathcal{Z}_2 = \{ Z_{(k,k')} \in \mathcal{Z} : \psi_l(\mathcal{G}(Z_{(k,k')})) \leq 0, l \in \{ 1, 2, 3 \} \}.
\]

Note that \( \mathcal{Z}_M \subset \mathcal{Z}_2 \). Let \( \hat{\psi} : \mathcal{H}_B \times \mathcal{H}_D \times \mathcal{G}_Z \to \mathbb{R} \) be the sample analog of \( \psi \) such that

\[
\hat{\psi}(h, f, g) = \frac{\widehat{P}(h \cdot f \cdot g)}{\widehat{P}(g)}
\]

for every \( (h, f, g) \in \mathcal{H}_B \times \mathcal{H}_D \times \mathcal{G}_Z \), where \( \widehat{P} \) is defined as in (3.4). Let \( \hat{\psi}_l \) be the sample analog of \( \psi_l \) for \( l \in \{ 1, 2, 3 \} \), which replaces \( \psi \) in \( \hat{\psi} \) by \( \hat{\psi}_l \). We define the estimator \( \widehat{\mathcal{Z}}_2 \) for \( \mathcal{Z}_2 \) by

\[
\widehat{\mathcal{Z}}_2 = \{ Z_{(k,k')} \in \mathcal{Z} : \sqrt{T_n} \hat{\psi}_l(\mathcal{G}(Z_{(k,k')})) \leq t_n, l \in \{ 1, 2, 3 \} \},
\]

where \( t_n \to \infty \) and \( t_n/\sqrt{T_n} \to 0 \) as \( n \to \infty \).
The following proposition establishes consistency of the estimator $\hat{\mathcal{F}}_2$.

**Proposition 3.2** Under Assumptions 2.2 and 2.3, $\mathbb{P}(\hat{\mathcal{F}}_2 = \mathcal{F}_2) \to 1$.

## 4 Empirical Application

We revisit the study of Angrist and Krueger (1991) and examine the use of the classical quarter of birth (QOB) instrument for estimating the returns to schooling. As explained by Dahl et al. (2017), the validity of this instrument has been contested. For example, Bound et al. (1995) argue that the exclusion restriction (Assumption 2.1.(i)) is not plausible because of seasonal birth patterns; see also Buckles and Hungerman (2013). Moreover, the validity of the monotonicity assumption (Assumption 2.1.(iii)) is questionable due to strategic parent behavior when enrolling their children (e.g., Barua and Lang, 2016).

Here we use the proposed method to remove invalid variation in the QOB instrument. The data set is from Angrist and Krueger (1991). Following Dahl et al. (2017), the outcome $Y$ is the log weekly wage, and the binary treatment $D$ is equal to 1 if an individual has 13 or more of years of schooling and 0 otherwise. The QOB instrument $Z \in \{1, 2, 3, 4\}$ indicates the quarter in which an individual is born.

For computational simplicity, we employ the necessary restrictions in Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021) for the estimation of the validity pair set, and we assume that

$$\mathcal{F}_P = \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}.$$  

The tuning parameter $\tau_n$ is chosen from $\{1, 1.5, \ldots, 6.5\}$. To calculate the supremum in

$$\sqrt{T_n}|\sup_{h \in H} \hat{\phi} (h,g)(\xi_0 \lor \hat{\sigma}(h,g))|$$

for every $g$, we use the approach employed by Kitagawa (2015) and Sun (2021). Specifically, we compute the supremum based on the closed intervals $[a, b]$ with the realizations of $\{Y_i\}_{i=1}^n$ as endpoints, i.e., intervals $[a, b]$ where $a, b \in \{Y_i\}_{i=1}^n$ and $a \leq b$. The resulting value is equal to

$$\sqrt{T_n}|\sup_{h \in H} \hat{\phi} (h,g)(\xi_0 \lor \hat{\sigma}(h,g))|.$$

We use the same sample of 486,926 men born between 1940 and 1949 as in Dahl et al. (2017). For computational simplicity, we randomly draw a subsample of size 10,000 to estimate the validity pair set. Table 4.1 shows the estimation results. The estimated validity pair set for $\tau_n = 3.5$ is $\hat{\mathcal{F}}_0 \cap \mathcal{F}_P = \{(1, 2), (1, 3), (1, 4), (2, 3)\}$. When $\tau_n < 3.5$, $\hat{\mathcal{F}}_0 \cap \mathcal{F}_P = \emptyset$, and when $\tau_n > 3.5$, $\hat{\mathcal{F}}_0 \cap \mathcal{F}_P = \mathcal{F}_P$.

---

10The data set was downloaded from https://economics.mit.edu/faculty/angrist/data1/data/angkru1991 (last accessed February 5, 2022).
Table 4.1: Validity Pair Set Estimation

| τni | (1, 2) | (1, 3) | (1, 4) | (2, 3) | (2, 4) | (3, 4) |
|-----|--------|--------|--------|--------|--------|--------|
| 1   | 0      | 0      | 0      | 0      | 0      | 0      |
| 1.5 | 0      | 0      | 0      | 0      | 0      | 0      |
| 2   | 0      | 0      | 0      | 0      | 0      | 0      |
| 2.5 | 0      | 0      | 0      | 0      | 0      | 0      |
| 3   | 0      | 0      | 0      | 0      | 0      | 0      |
| 3.5 | 1      | 1      | 1      | 1      | 0      | 0      |
| 4   | 1      | 1      | 1      | 1      | 1      | 1      |
| 4.5 | 1      | 1      | 1      | 1      | 1      | 1      |
| 5   | 1      | 1      | 1      | 1      | 1      | 1      |
| 5.5 | 1      | 1      | 1      | 1      | 1      | 1      |
| 6   | 1      | 1      | 1      | 1      | 1      | 1      |
| 6.5 | 1      | 1      | 1      | 1      | 1      | 1      |

References

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering Metrics: The Path from Cause to Effect*. Princeton University Press.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.

Barua, R. and Lang, K. (2016). School entry, educational attainment, and quarter of birth: A cautionary tale of a local average treatment effect. *Journal of Human Capital*, 10(3):347–376.

Beare, B. K. and Shi, X. (2019). An improved bootstrap test of density ratio ordering. *Econometrics and Statistics*, 10:9–26.
Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association, 90*(430):443–450.

Buckles, K. S. and Hungerman, D. M. (2013). Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics, 95*(3):711–724.

Carr, T. and Kitagawa, T. (2021). Testing instrument validity with covariates. Working Paper arXiv:2112.08092.

Dahl, C. M., Huber, M., and Mellace, G. (2017). It’s never too late: A new look at local average treatment effects with or without defiers. Working Paper.

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics, 139*(1):35–75. Endogeneity, instruments and identification.

Heckman, J. J. and Pinto, R. (2018). Unordered monotonicity. *Econometrica, 86*(1):1–35.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica, 73*(3):669–738.

Huber, M. and Mellace, G. (2015). Testing instrument validity for LATE identification based on inequality moment constraints. *Review of Economics and Statistics, 97*(2):398–411.

Huber, M. and Wüthrich, K. (2018). Local average and quantile treatment effects under endogeneity: A review. *Journal of Econometric Methods, 8*(1).

Imbens, G. (2014). Instrumental variables: An econometrician’s perspective. Technical report, National Bureau of Economic Research.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*(2):467–475.

Imbens, G. W. and Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies, 64*(4):555–574.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
Kédagni, D. and Mourifié, I. (2020). Generalized instrumental inequalities: Testing the instrumental variable independence assumption. *Biometrika*, 107(3):661–675.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, 83(5):2043–2063.

Lee, S., Song, K., and Whang, Y.-J. (2013). Testing functional inequalities. *Journal of Econometrics*, 172(1):14–32.

Linton, O., Song, K., and Whang, Y.-J. (2010). An improved bootstrap test of stochastic dominance. *Journal of Econometrics*, 154(2):186–202.

Melly, B. and Wüthrich, K. (2017). Local quantile treatment effects. In *Handbook of Quantile Regression*, pages 145–164. Chapman and Hall/CRC.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, 111(11):3663–98.

Mourifié, I. and Wan, Y. (2017). Testing local average treatment effect assumptions. *Review of Economics and Statistics*, 99(2):305–313.

Sun, Z. (2021). Instrument validity for heterogeneous causal effects. Working Paper arXiv:2009.01995.

Sun, Z. and Beare, B. K. (2021). Improved nonparametric bootstrap tests of Lorenz dominance. *Journal of Business & Economic Statistics*, 39(1):189–199.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.
Appendix to *Pairwise Valid Instruments*  
Zhenting Sun  
Kaspar Wüthrich

**A Extension: Multivalued Ordered and Unordered Treatments**

In this section, we generalize the results in the main text to multivalued ordered and unordered treatments.

### A.1 Ordered Treatment

Suppose, in general, that the observable treatment variable $D \in \mathcal{D} = \{d_1, \ldots, d_J\}$. Without loss of generality, suppose $d_1 < \cdots < d_J$. The following assumption is a straightforward generalization of Assumption 2.1 to ordered treatments (e.g., Sun, 2021).

**Assumption A.1 IV Validity Conditions for Ordered Treatments:**

(i) **Exclusion:** For all $d \in \mathcal{D}$, $Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ a.s.

(ii) **Random Assignment:** $Z$ is jointly independent of $(Y_{d_1 z_1}, \ldots, Y_{d_1 z_K}, \ldots, Y_{d_J z_1}, \ldots, Y_{d_J z_K})$ and $(D_{z_1}, \ldots, D_{z_K})$.

(iii) **Monotonicity:** For all $k = 1, \ldots, K - 1$, $D_{z_{k+1}} \geq D_{z_k}$ a.s.

We next introduce the definition of pairwise valid instruments for ordered treatments.

**Definition A.1** An instrument $Z$ is pairwise valid for an ordered treatment $D \in \mathcal{D} = \{d_1, \ldots, d_J\}$ if there is a set $\mathcal{Z}_M = \{(z_{k_1}, z'_{k_1}), \ldots, (z_{k_M}, z'_{k_M})\}$ with $z_{k_1}, z'_{k_1}, \ldots, z_{k_M}, z'_{k_M} \in \mathcal{Z}$ such that the following conditions hold for every $(z, z') \in \mathcal{Z}_M$:

(i) **Exclusion:** For all $d \in \mathcal{D}$, $Y_{dz} = Y_{dz'}$ a.s.

(ii) **Random Assignment:** $Z$ is jointly independent of $(Y_{d_1 z}, Y_{d_1 z'}, \ldots, Y_{d_J z}, Y_{d_J z'}, D_z, D_{z'})$.

(iii) **Monotonicity:** $D_{z'} \geq D_z$ a.s.

The set $\mathcal{Z}_M$ is called a validity pair set of $Z$. The union of all validity pair sets is the largest validity pair set, denoted by $\mathcal{Z}_M$. 

1
With the exclusion condition, for every \((z, z') \in \mathcal{Z}_M\), define \(Y_d(z, z')\) such that \(Y_d(z, z') = Y_{dz} = Y_{dz'}\) a.s. for all \(d \in D\).

**Lemma A.1** Suppose that the instrument \(Z\) is pairwise valid as defined in Definition A.1 with a known validity pair set \(\mathcal{Z}_M = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_M}, z_{k'_M})\}\). Then for every \(m \in \{1, \ldots, M\}\), the following quantity can be identified:

\[
\beta_{k'_{m}, k_{m}} = \frac{E[Y|Z = z_{k'_{m}}] - E[Y|Z = z_{k_{m}}]}{E[D|Z = z_{k'_{m}}] - E[D|Z = z_{k_{m}}]},
\]

where

\[
\omega_j = \frac{\mathbb{P}(D_{z_{k'_{m}}} \geq d_j > D_{z_{k_{m}}})}{\sum_{l=2}^{J} (d_l - d_{l-1}) \mathbb{P}(D_{z_{k'_{m}}} \geq d_l > D_{z_{k_{m}}})}.
\]

Lemma A.1 is an extension of Theorem 1 of Imbens and Angrist (1994) and Theorem 1 of Angrist and Imbens (1995) for the case where \(Z\) is pairwise valid. We follow Angrist and Imbens (1995) and refer to \(\beta_{k'_{m}, k_{m}}\) as the average causal response (ACR). Lemma A.1 shows that if a validity pair set \(\mathcal{Z}_M\) is known, we can identify every \(\beta_{k'_{m}, k_{m}}\). In practice, however, \(\mathcal{Z}_M\) is usually unknown. We show how to identify the largest validity pair set \(\mathcal{Z}_M\) and use it to estimate the ACRs.

As in Section 2, we first suppose that \(\mathcal{Z}_{iM}\) can be estimated consistently by some estimator \(\widehat{\mathcal{Z}}_0\). We follow the same notation as in Section 2. With \(\widehat{\mathcal{Z}}_0\), for every \(\mathcal{Z}_{(k,k')} \in \mathcal{Z}\), we run the regression

\[
Y_i \mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \widehat{\mathcal{Z}}_0) = \gamma_{(k,k')}^0 \mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \widehat{\mathcal{Z}}_0) + \gamma_{(k,k')}^1 D_i \mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \widehat{\mathcal{Z}}_0) + \epsilon_i \mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \widehat{\mathcal{Z}}_0),
\]

(A.2)

using \(g(Z_i)\mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \widehat{\mathcal{Z}}_0)\) as the instrument for the regressor \(D_i \mathcal{I}(Z_i, \mathcal{Z}_{(k,k')}, \widehat{\mathcal{Z}}_0)\). Then we obtain the VSIIV estimator for each ACR as

\[
\hat{\beta}_{1_{(k,k')}}^1 = \frac{\mathcal{E}_n \left( g(Z_i) Y_i, \mathcal{Z}_{(k,k')} \right) - \mathcal{E}_n \left( g(Z_i), \mathcal{Z}_{(k,k')} \right) \mathcal{E}_n \left( Y_i, \mathcal{Z}_{(k,k')} \right)}{\mathcal{E}_n \left( g(Z_i) D_i, \mathcal{Z}_{(k,k')} \right) - \mathcal{E}_n \left( g(Z_i), \mathcal{Z}_{(k,k')} \right) \mathcal{E}_n \left( D_i, \mathcal{Z}_{(k,k')} \right)},
\]

(A.3)

which is the IV estimator for \(\gamma_{(k,k')}^1\) in (A.2). As in Section 2, we define

\[
\hat{\beta}_1 = \left( \hat{\beta}_{1_{(1,2)}}, \ldots, \hat{\beta}_{1_{(1,K)}}, \ldots, \hat{\beta}_{1_{(K,1)}}, \ldots, \hat{\beta}_{1_{(K,K-1)}} \right),
\]

2
\[
\beta_{(k,k')}^1 = \frac{\mathcal{E}(g(Z_i)Y_i, Z_{(k,k')}) - \mathcal{E}(g(Z_i), Z_{(k,k')})}{\mathcal{E}(g(Z_i) D_i, Z_{(k,k')}) - \mathcal{E}(g(Z_i), Z_{(k,k')})}
\]  
(A.4)

and

\[
\beta_1 = \left(\beta_{(1,2)}, \ldots, \beta_{(1,K)}, \ldots, \beta_{(K,1)}, \ldots, \beta_{(K,K-1)}\right).
\]

**Assumption A.2** \((\{Y_i, D_i, Z_i\})^n_{i=1}\) is an i.i.d. sample.

**Assumption A.3** The moments \(E[Y_i], E[D_i], E[g(Z_i)], E[g(Z_i)Y_i],\) and \(E[g(Z_i)D_i]\) exist.

**Theorem A.1** Suppose that the instrument \(Z\) is pairwise valid for the treatment \(D\) as defined in Definition A.1 with the largest validity pair set \(\mathcal{Z}_M = \{(z_{k_1}, z_{k_1'}), \ldots, (z_{k_M}, z_{k_M'})\}\), and that the estimator \(\hat{\beta}_0\) satisfies \(P(\hat{\beta}_0 = \mathcal{Z}_M) \rightarrow 1\). Under Assumptions A.2 and A.3, \(\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, \Sigma)\), where \(\Sigma\) is provided in (C.2) in the proof. In addition, \(\beta_{(k,k')}^1 = \beta_{k',k}\) as defined in (A.1) for every \((z_k, z_{k'}) \in \mathcal{Z}_M\).

The estimation of \(\mathcal{Z}_M\) is similar to that in Section 2. Suppose that there are subsets \(\mathcal{Z}_1 \subset \mathcal{Z}\) and \(\mathcal{Z}_2 \subset \mathcal{Z}\) that satisfy the testable implications in Kitagawa (2015), Mourifié and Wan (2017), and those in Kédagni and Mourifié (2020), respectively. We let \(\mathcal{Z}_0 = \mathcal{Z}_1 \cap \mathcal{Z}_2\) so that \(\mathcal{Z}_0\) satisfies all the above necessary conditions. We can first construct the estimators \(\hat{\beta}_1\) and \(\hat{\beta}_2\) for \(\mathcal{Z}_1\) and \(\mathcal{Z}_2\), respectively, and then construct the estimator \(\hat{\beta}_0\) for \(\mathcal{Z}_0\) as \(\hat{\beta}_0 = \hat{\beta}_1 \cap \hat{\beta}_2\). See Appendix C.3 for details.

Next, we generalize the results in Section 2.2.2 and show that VSIV regression always reduces the asymptotic estimation bias when the treatments are ordered. Given a presumed validity pair set \(\mathcal{Z}_p\), we apply VSIV regression based on \(\mathcal{Z}_0\) defined as in Section 2.2.2.

**Theorem A.2** Suppose \(P(\hat{\beta}_0 = \mathcal{Z}_0) \rightarrow 1\) with \(\mathcal{Z}_0 \supset \mathcal{Z}_M\). For every presumed validity pair set \(\mathcal{Z}_p\), the asymptotic estimation bias \(\text{plim}_{n \to \infty} \|\hat{\beta}_1 - \beta_1\|_2\) is always reduced by using \(\hat{\beta}_0\) in the regression (A.2) compared to that from using \(\hat{\beta}_p\).

As shown in Propositions C.1 and C.2, the pseudo-validity pair set \(\mathcal{Z}_0\) can always be estimated consistently by \(\hat{\beta}_0\) under mild conditions. Theorem A.2 shows that VSIV regression based on \(\hat{\beta}_0 \cap \mathcal{Z}_p\) always reduces the bias.

**Remark A.1** In Section 2, we provide the definition of partial IV validity for the binary treatment case. See Appendix C.4 for the extension to ordered treatments.
A.2 Unordered Treatment

A.2.1 Setup

Here, we extend our results to unordered multivalued treatments using the framework of Heckman and Pinto (2018). The treatment (choice) $D$ is discrete with support $\mathcal{D} = \{d_1, \ldots, d_J\}$, which is unordered. Heckman and Pinto (2018) consider the following monotonicity assumption.

**Assumption A.4** For all $d \in \mathcal{D}$ and all $z, z' \in \mathcal{Z}$, $1\{D_{z'} = d\} \geq 1\{D_z = d\}$ for all $\omega \in \Omega$, or $1\{D_{z'} = d\} \leq 1\{D_z = d\}$ for all $\omega \in \Omega$.

Based on Assumption A.4 we introduce the definition of the pairwise IV validity for the unordered treatment case.

**Definition A.2** An instrument $Z$ is pairwise valid for the unordered treatment $D$ if there is a set $\mathcal{Z}_M = \{(z_{k_1}, z'_{k_1}), \ldots, (z_{k_M}, z'_{k_M})\}$ with $z_{k_1}, z'_{k_1}, \ldots, z_{k_M}, z'_{k_M} \in \mathcal{Z}$ and $k_m < k'_m$ for every $m$ such that the following conditions hold for every $(z, z') \in \mathcal{Z}_M$:

(i) Exclusion: For all $d \in \mathcal{D}$, $Y_{dz} = Y_{dz'}$ a.s.

(ii) Random Assignment: $Z$ is jointly independent of $(Y_{d_1 z}, Y_{d_1 z'}, \ldots, Y_{d_J z}, Y_{d_J z'}, D_z, D_{z'})$.

(iii) Monotonicity: For all $d \in \mathcal{D}$, $1\{D_{z'} = d\} \geq 1\{D_z = d\}$ for all $\omega \in \Omega$, or $1\{D_{z'} = d\} \leq 1\{D_z = d\}$ for all $\omega \in \Omega$.

The set $\mathcal{Z}_M$ is called a validity pair set of $Z$. The union of all validity pair sets is the largest validity pair set, denoted by $\mathcal{Z}_M$.

Suppose the instrument $Z$ is pairwise valid for the treatment $D$ with the largest validity pair set $\mathcal{Z}_M = \{(z_{k_1}, z'_{k_1}), \ldots, (z_{k_M}, z'_{k_M})\}$. Define $Y_d(z, z')$ for each $d \in \mathcal{D}$ and every $(z, z') \in \mathcal{Z}_M$ such that $Y_d(z, z') = Y_{dz} = Y_{dz'}$ a.s. Following Heckman and Pinto (2018), we introduce the following notation. Define the response vector $S$ as a $K$-dimensional random vector of potential treatments with $Z$ fixed at each value of its support:

$$S = (D_{z_1}, \ldots, D_{z_K})^T.$$

The finite support of $S$ is $\mathcal{S} = \{\xi_1, \ldots, \xi_{N_S}\}$, where $N_S$ is the number of possible values of $S$. The response matrix $R$ is an array of response-types defined over $\mathcal{S}$, $R = (\xi_1, \ldots, \xi_{N_S})$.

---

11More precisely, the potential treatments should be written as functions of $\omega$, $D_z(\omega)$ and $D_{z'}(\omega)$. For simplicity of notation, we omit $\omega$ whenever there is no confusion. The inequalities can be modified to hold a.s.
For every $Z_{(k,k')} \in \mathcal{Z}$, there is a $2 \times K$ binary matrix $\mathcal{M}_{(k,k')}$ such that

$$\mathcal{M}_{(k,k')}(z_1, \ldots, z_K)^T = (z_k, z_{k'})^T.$$ 

For example, if $K = 5$ and $(k, k') = (3, 5)$, then

$$\mathcal{M}_{(3,5)} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$ 

We define a transformation $\mathcal{K}_{(k,k')}$ such that if $A$ is a $K \times L$ matrix, $\mathcal{K}_{(k,k')}A$ is the matrix that consists of all the unique columns of $\mathcal{M}_{(k,k')}A$ in the same order as in $\mathcal{M}_{(k,k')}A$. In the above example, if $A = ((x_1, \ldots, x_5)^T, (x_1, \ldots, x_5)^T, (y_1, \ldots, y_5)^T)$, then $\mathcal{K}_{(3,5)}A = ((x_3, x_5)^T, (y_3, y_5)^T)$. We write $\mathcal{K}_{(k,k')}R = (s_1, \ldots, s_{L_{(k,k')}})$, where $L_{(k,k')}$ is the column number of $\mathcal{K}_{(k,k')}R$. Let $B_{d(k,k')}$ denote a binary matrix of the same dimension as $\mathcal{K}_{(k,k')}R$, whose elements are equal to 1 if the corresponding element in $\mathcal{K}_{(k,k')}R$ is equal to $d$, and equal to 0 otherwise. We denote the element in the $m$th row and $l$th column of the matrix $B_{d(k,k')}$ by $B_{d(k,k')}(m,l)$. Finally, we use $B_{d(k,k')} = 1\{\mathcal{K}_{(k,k')}R = d\}$ to denote $B_{d(k,k')}$. 

**Lemma A.2** Suppose that the instrument $Z$ is pairwise valid for the treatment $D$ with the largest validity pair set $\mathcal{Z}_{M} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_M}, z_{k'_M})\}$. The following statements are equivalent:

(i) For every $(z_k, z_{k'}) \in \mathcal{Z}_{M}$, the binary matrix $B_{d(k,k')} = 1\{\mathcal{K}_{(k,k')}R = d\}$ is lonesum\(^{12}\) for every $d \in D$.

(ii) For every $(z_k, z_{k'}) \in \mathcal{Z}_{M}$ and all $d, d', d'' \in D$, there are no $2 \times 2$ sub-matrices of $\mathcal{K}_{(k,k')}R$ of the type

$$\begin{pmatrix} d & d' \\ d'' & d \end{pmatrix} \text{ or } \begin{pmatrix} d' & d \\ d & d'' \end{pmatrix}$$

with $d' \neq d$ and $d'' \neq d$.

(iii) For every $(z_k, z_{k'}) \in \mathcal{Z}_{M}$ and every $d \in D$, the following inequalities hold:

$$1\{D_{z'} = d\} \geq 1\{D_{z} = d\} \text{ for all } \omega \in \Omega, \text{ or } 1\{D_{z'} = d\} \leq 1\{D_{z} = d\} \text{ for all } \omega \in \Omega.$$ 

Lemma A.2 is an extension of Theorem T-3 of Heckman and Pinto (2018) for pairwise valid instruments. It provides equivalent conditions for the monotonicity condition (iii) in Definition A.2. 

\(^{12}\) A binary matrix is lonesum if it is uniquely determined by its row and column sums.” (Heckman and Pinto, 2018, p. 20).
To describe our results, following Heckman and Pinto (2018), we define some additional notation. Let $B_{d(k,k')}^+$ denote the Moore–Penrose pseudo-inverse of $B_{d(k,k')}$. Let $\kappa : \mathbb{R} \to \mathbb{R}$ be an arbitrary function of interest. Define for all $d \in \mathcal{D}$,

$$
\bar{P}_Z (d) = (\mathbb{P} (D = d | Z = z_1), \ldots, \mathbb{P} (D = d | Z = z_K))^T,
$$

$$
\bar{Q}_Z (d) = (E [\kappa (Y) \cdot 1 \{ D = d \} | Z = z_1], \ldots, E [\kappa (Y) \cdot 1 \{ D = d \} | Z = z_K])^T,
$$

$$
P_{Z(k,k')} (d) = \mathcal{M}_{(k,k')} \bar{P}_Z (d) = (\mathbb{P} (D = d | Z = z_k), \mathbb{P} (D = d | Z = z_{k'}))^T,
$$

and

$$
Q_{Z(k,k')} (d) = \mathcal{M}_{(k,k')} \bar{Q}_Z (d)
= (E [\kappa (Y) \cdot 1 \{ D = d \} | Z = z_k], E [\kappa (Y) \cdot 1 \{ D = d \} | Z = z_{k'}])^T,
$$

Moreover, we define

$$
P_{Z(k,k')} = (P_{Z(k,k')} (d_1), \ldots, P_{Z(k,k')} (d_J))^T \text{ and }
$$

$$
P_{S(k,k')} = (P (\mathcal{M}_{(k,k')} S = s_1), \ldots, P (\mathcal{M}_{(k,k')} S = s_{L(k,k')}))^T,
$$

and

$$
Q_{S(k,k')} (d)
= \left( E [\kappa (Y_{d(z_k,z_{k'})}) \cdot 1 \{ \mathcal{M}_{(k,k')} S = s_1 \}], \ldots, E \left[ \kappa (Y_{d(z_k,z_{k'})}) \cdot 1 \left\{ \mathcal{M}_{(k,k')} S = s_{L(k,k')} \right\} \right] \right)
$$

for all $d \in \mathcal{D}$. Define $\Sigma_{d(k,k')} (t)$ to be the set of response-types in which $d$ appears exactly $t$ times, that is, for every $d \in \mathcal{D}$ and every $t \in \{0,1,2\}$, define

$$
\Sigma_{d(k,k')} (t) = \left\{ s : s \text{ is some } l\text{th column of } \mathcal{K}_{(k,k')} R \text{ with } \sum_{m=1}^2 B_{d(k,k')} (m,l) = t \right\}.
$$

Let $b_{d(k,k')} (t)$ be a $L(k,k')$-dimensional binary row-vector that indicates if every column of $\mathcal{K}_{(k,k')} R$ belongs to $\Sigma_{d(k,k')} (t)$, that is, $b_{d(k,k')} (t) (l) = 1$ if $s_l \in \Sigma_{d(k,k')} (t)$, and $b_{d(k,k')} (t) (l) = 0$ otherwise for every $l \in \{1, \ldots, L(k,k')\}$, where $s_l$ is the $l$th column of $\mathcal{K}_{(k,k')} R$. In this section, we let

$$
\mathcal{F} = \{(z_1, z_2), \ldots, (z_1, z_K), \ldots, (z_{K-1}, z_K)\}.
$$

Finally, define $1(\mathcal{A}) = (1((z_1, z_2) \in \mathcal{A}), \ldots, 1((z_{K-1}, z_K) \in \mathcal{A}))$ for every $\mathcal{A} \subset \mathcal{F}$. 

6
A.2.2 VSIV Regression under Consistent Estimation of the Validity Pair Set

Here, we study the properties of VSIV regression when the validity pair set can be estimated consistently, that is, there is an estimator \( \hat{\mathcal{Z}}_0 \) such that \( \mathbb{P}(\hat{\mathcal{Z}}_0 = \mathcal{Z}_M) \to 1 \). Suppose that there are subsets \( \mathcal{Z}_1 \subset \mathcal{Z} \) and \( \mathcal{Z}_2 \subset \mathcal{Z} \) that satisfy the testable implications in Sun (2021), and those in Kédagni and Mouriﬁé (2020), respectively. Similarly to Section A.1, we let \( \mathcal{Z}_0 = \mathcal{Z}_1 \cap \mathcal{Z}_2 \) so that \( \mathcal{Z}_0 \) satisfies all the above necessary conditions. We can first construct the estimators \( \hat{\mathcal{Z}}_1 \) and \( \hat{\mathcal{Z}}_2 \) for \( \mathcal{Z}_1 \) and \( \mathcal{Z}_2 \), respectively, and then construct the estimator \( \hat{\mathcal{Z}}_0 \) for \( \mathcal{Z}_0 \) as \( \hat{\mathcal{Z}}_0 = \hat{\mathcal{Z}}_1 \cap \hat{\mathcal{Z}}_2 \). See Appendix D.2 for details. If \( \mathcal{Z}_0 = \mathcal{Z}_M \), then under mild conditions, it follows that \( \mathbb{P}(\hat{\mathcal{Z}}_0 = \mathcal{Z}_M) \to 1 \).

To state the results, define
\[
P_{DZ}(d) = (\mathbb{P}(D = d, Z = z_1), \ldots, \mathbb{P}(D = d, Z = z_K))^T, 
\]
\[
Q_{YDZ}(d) = (E[\kappa(Y) 1\{D = d, Z = z_1\}], \ldots, E[\kappa(Y) 1\{D = d, Z = z_K\}])^T, 
\]
for every \( d \in D \), and
\[
Z_P = (\mathbb{P}(Z = z_1), \ldots, \mathbb{P}(Z = z_K)),
\]
\[
W = \left(Z_P, P_{DZ}(d_1)^T, \ldots, P_{DZ}(d_J)^T, Q_{YDZ}(d_1)^T, \ldots, Q_{YDZ}(d_J)^T\right)^T.
\]
Suppose we have a random sample \( \{(Y_i, D_i, Z_i)\}_{i=1}^n \). Define the following sample analogs:
\[
\hat{\mathbb{P}}(Z = z) = \frac{1}{n} \sum_{i=1}^n 1\{Z_i = z\} \text{ for all } z,
\]
\[
\hat{\mathbb{P}}(D = d, Z = z) = \frac{1}{n} \sum_{i=1}^n 1\{D_i = d, Z_i = z\} \text{ for all } d \text{ and all } z,
\]
\[
\hat{E}[\kappa(Y) 1\{D = d, Z = z\}] = \frac{1}{n} \sum_{i=1}^n \kappa(Y_i) 1\{D_i = d, Z_i = z\} \text{ for all } d \text{ and all } z,
\]
\[
\hat{P}_{DZ}(d) = \left(\hat{\mathbb{P}}(D = d, Z = z_1), \ldots, \hat{\mathbb{P}}(D = d, Z = z_K)\right)^T \text{ for all } d,
\]
\[
\hat{Q}_{YDZ}(d) = \left(\hat{E}[\kappa(Y) 1\{D = d, Z = z_1\}], \ldots, \hat{E}[\kappa(Y) 1\{D = d, Z = z_K\}]\right)^T \text{ for all } d,
\]
\[
\hat{Z}_P = \left(\hat{\mathbb{P}}(Z = z_1), \ldots, \hat{\mathbb{P}}(Z = z_K)\right),
\]
and
\[
\hat{W} = \left(\hat{Z}_P, \hat{P}_{DZ}(d_1)^T, \ldots, \hat{P}_{DZ}(d_J)^T, \hat{Q}_{YDZ}(d_1)^T, \ldots, \hat{Q}_{YDZ}(d_J)^T\right)^T.
\]
We impose the following weak regularity conditions.

**Assumption A.5** \{\( (Y_i, D_i, Z_i) \}_{i=1}^{n} \) is an i.i.d. sample, and the moment \( E[\kappa(Y)] \) exists.

The next theorem presents the asymptotic properties of VSIV regression with unordered treatments.

**Theorem A.3** Suppose that the instrument \( Z \) is pairwise valid for the treatment \( D \) as defined in Definition A.2 with the largest validity pair set \( Z_M = \{(z_{k_1}, z_{k_2}), \ldots, (z_{k_M}, z_{k_M'})\} \), and that the estimator \( \hat{\beta} \) satisfies \( \mathbb{P}(\hat{\beta}_{0} = \beta_{M}) \to 1 \). Under Assumption A.5, the following response-type probabilities and counterfactuals are identified for every \( d \in D \), each \( t \in \{1, 2\} \), and every \((k, k')\) with \( k < k'\):

\[
\mathbb{P}\left( M_{(k,k')} S \in \Sigma_{d(k,k')} (t), (z_k, z_{k'}) \in Z_M \right) = b_{d(k,k')} (t) B_{d(k,k')}^+ P_{Z(k,k')} (d) 1 \{(z_k, z_{k'}) \in Z_M \}
\]

and

\[
E[\kappa(Y_d(z_k, z_{k'})) \mid M_{(k,k')} S \in \Sigma_{d(k,k')} (t), (z_k, z_{k'}) \in Z_M] = \frac{b_{d(k,k')} (t) B_{d(k,k')}^+ Q_{Z(k,k')} (d) 1 \{(z_k, z_{k'}) \in Z_M \}}{b_{d(k,k')} (t) B_{d(k,k')}^+ P_{Z(k,k')} (d) 1 \{(z_k, z_{k'}) \in Z_M \}}.
\]

(A.5)

In addition, we have that

\[
\sqrt{n} \left\{ \left( \bar{W}, 1(\hat{\beta}_{0}) \right) - \left( W, 1(Z_M) \right) \right\} \overset{d}{\to} (N(0, \Sigma_W), 0),
\]

where \( \Sigma_W \) is given in (D.4).

Since the probabilities \( \mathbb{P}(M_{(k,k')} S \in \Sigma_{d(k,k')} (t), (z_k, z_{k'}) \in Z_M) \) and the counterfactuals \( E[\kappa(Y_d(z_k, z_{k'})) \mid M_{(k,k')} S \in \Sigma_{d(k,k')} (t), (z_k, z_{k'}) \in Z_M] \) in (A.5) are differentiable functions of \((W, 1(Z_M))\), inferences on these quantities can be conducted based on Theorem A.3 and delta methods (e.g., Theorem 3.9.4 in van der Vaart and Wellner, 1996). As shown in Remark 7.1 in Heckman and Pinto (2018) and Theorem A.3, if \((z_k, z_{k'}) \in Z_M \) and \( \Sigma_{d(k,k')} (t) = \Sigma_{d'(k,k')} (t') \) for some \( d, d' \in D \) and some \( t, t' \in \{1, 2\} \), the mean treatment effect of \( d \) relative to \( d' \) for \( \Sigma_{d(k,k')} (t) \) can be identified, which is \( E[Y_d(z_k, z_{k'}) - Y_{d'}(z_k, z_{k'}) \mid M_{(k,k')} S \in \Sigma_{d(k,k')} (t)] \).

### A.2.3 Bias Reduction for Mean Treatment Effect

Here, we extend the results in Section 2.2.2 and show that VSIV estimation always reduces the asymptotic bias for estimating mean treatment effects with unordered treatments.
For all $d, d' \in \mathcal{D}$, all $t, t' \in \{1, 2\}$, and all $k < k'$, following Heckman and Pinto (2018), we define the mean treatment effect as

$$
\beta_{(k,k')}(d, d', t, t') = E[Y_d(z_k, z_{k'}) - Y_{d'}(z_k, z_{k'})| M_{(k,k')}(z_{k'}, z_k) S \in \Sigma_{d(k, k')}(t), (z_k, z_{k'}) \in \mathcal{Z}, \Sigma_{d(k, k')}(t) = \Sigma_{d'(k, k')(t')}].
$$

Lemma A.3 The mean treatment effect $\beta_{(k,k')}(d, d', t, t')$ can be expressed as

$$
\beta_{(k,k')}(d, d', t, t') = \frac{b_{d(k, k')}(t) B_{d(k, k')}(t) Q_{Z(k, k')}(d) 1 \{(z_k, z_{k'}) \in \mathcal{Z}, \Sigma_{d(k, k')}(t) = \Sigma_{d'(k, k')(t')}\}}{b_{d(k, k')}(t) B_{d(k, k')}(t) Q_{Z(k, k')}(d) 1 \{(z_k, z_{k'}) \in \mathcal{Z}, \Sigma_{d(k, k')}(t) = \Sigma_{d'(k, k')(t')}\}}
$$

We now define

$$
\beta_{(k,k')}(d, d') = (\beta_{(k, k')}(d, d', 1, 1), \beta_{(k, k')}(d, d', 1, 2), \beta_{(k, k')}(d, d', 2, 1), \beta_{(k, k')}(d, d', 2, 2))
$$

for all $d, d' \in \mathcal{D}$ and all $k < k'$. For all $k < k'$, we let

$$
\beta_{(k,k')} = (\beta_{(k, k')}(d_1, d_2), \ldots, \beta_{(k, k')}(d_1, d_J), \ldots, \beta_{(k, k')}(d_J, d_1), \ldots, \beta_{(k, k')}(d_J, d_{J-1})).
$$

Finally, we define

$$
\beta = (\beta_{(1, 2)}, \ldots, \beta_{(1, K)}, \ldots, \beta_{(K-1, K)}).
$$

Note that if $(z_k, z_{k'}) \notin \mathcal{Z}$, then $\beta_{(k,k')} = 0$. For the sample analogs, we define

$$
\tilde{\beta}_{(k,k')}(d, d', t, t') = \frac{b_{d(k, k')}(t) B_{d(k, k')}(t) Q_{Z(k, k')}^0(d) 1 \{(z_k, z_{k'}) \in \mathcal{Z}^0, \Sigma_{d(k, k')}(t) = \Sigma_{d'(k, k')(t')}\}}{b_{d(k, k')}(t) B_{d(k, k')}(t) Q_{Z(k, k')}^0(d) 1 \{(z_k, z_{k'}) \in \mathcal{Z}^0, \Sigma_{d(k, k')}(t) = \Sigma_{d'(k, k')(t')}\}}
$$

(A.8)
where \( P_{Z(k,k')}(d) \) and \( Q_{Z(k,k')}(d) \) can be obtained by transformations of \( \hat{W} \). We let
\[
\hat{\beta}(k,k')(d,d') = (\hat{\beta}(k,k')(d,d',1,1), \hat{\beta}(k,k')(d,d',1,2), \hat{\beta}(k,k')(d,d',2,1), \hat{\beta}(k,k')(d,d',2,2))
\]
(A.10)
for all \( d,d' \in \mathcal{D} \) and all \( k < k' \). For all \( k < k' \), we define
\[
\hat{\beta}(k',k) = (\hat{\beta}(k,k')(d_1,d_2), \ldots, \hat{\beta}(k,k')(d_1,d_K), \ldots, \hat{\beta}(k,k')(d_K,d_1), \ldots, \hat{\beta}(k,k')(d_K,d_K-1)).
\]
(A.11)
Finally, define
\[
\hat{\beta} = (\hat{\beta}(1,2), \ldots, \hat{\beta}(1,K), \ldots, \hat{\beta}(K-1,K)).
\]
(A.12)

The following theorem shows that VSIV regression always reduces the asymptotic estimation bias.

**Theorem A.4** Suppose \( \mathbb{P}(\hat{Z}_0 = Z_0) \rightarrow 1 \) with \( Z_0 \supset Z_M \). For every presumed validity pair set \( \mathcal{Z}_P \), the asymptotic bias \( \text{plim}_{n \rightarrow \infty} \| \hat{\beta} - \beta \|_2 \) is always reduced by using \( \hat{Z}_0' = \hat{Z}_0 \cap \mathcal{Z}_P \) in the estimation for (A.9) compared to that from using \( \mathcal{Z}_P \).

As shown in Propositions C.2 and D.1, the pseudo-validity pair set \( \hat{Z}_0 \) can always be estimated consistently by \( \hat{Z}_0 \) under mild conditions. Theorem A.4 shows that VSIV regression based on \( \hat{Z}_0 \cap \mathcal{Z}_P \) reduces the bias relative to standard IV regression based on \( \mathcal{Z}_P \).

**B Proofs for Section 2**

The results in Section 2 are for the special case where \( D \) is binary and follow from the general results for ordered treatments in Appendix A.1. The proofs of these general results are in Appendix C.

**C Proofs and Supplementary Results for Appendix A.1**

**C.1 Proofs for Appendix A.1**

**Proof of Lemma A.1.** The proof closely follows the strategy of that of Theorem 1 in Angrist and Imbens (1995). Let \( d_0 = 0 \) and \( Y_{d_0}(z_{km}, z_{km'}) = 0 \) for every \( m \). Let \( d_{J+1} \) be
some number such that \( d_{j+1} > d_j \). We can write
\[
Y = \sum_{k=1}^{K} 1 \{ Z = z_k \} \cdot \left\{ \sum_{j=1}^{J} 1 \{ D = d_j \} Y_{d_j z_k} \right\}.
\]

Now we have that
\[
E \left[ \frac{Y}{Z = z_{k'}_m} \right] - E \left[ \frac{Y}{Z = z_{km}} \right]
= E \left[ \sum_{j=1}^{J} Y_{d_j} (z_{km}, z_{k'_m}) \left( \frac{1 \{ D_{z_{k'_m}} \geq d_j \} - 1 \{ D_{z_{km}} \geq d_j + 1 \}}{-1 \{ D_{z_{km}} \geq d_j \} + 1 \{ D_{z_{km}} \geq d_j + 1 \}} \right) \right]
= \sum_{j=1}^{J} E \left[ (Y_{d_j} (z_{km}, z_{k'_m}) - Y_{d_{j-1}} (z_{km}, z_{k'_m})) \left( \frac{1 \{ D_{z_{k'_m}} \geq d_j \} - 1 \{ D_{z_{km}} \geq d_j \}}{-1 \{ D_{z_{km}} \geq d_j \} + 1 \{ D_{z_{km}} \geq d_j \}} \right) \right].
\]

By Definition A.1, \( 1 \{ D_{z_{k'_m}} \geq d_j \} - 1 \{ D_{z_{km}} \geq d_j \} \) \( \in \{0, 1\} \). Then we have that
\[
\sum_{j=1}^{J} E \left[ (Y_{d_j} (z_{km}, z_{k'_m}) - Y_{d_{j-1}} (z_{km}, z_{k'_m})) \left( \frac{1 \{ D_{z_{k'_m}} \geq d_j \} - 1 \{ D_{z_{km}} \geq d_j \}}{-1 \{ D_{z_{km}} \geq d_j \} + 1 \{ D_{z_{km}} \geq d_j \}} \right) \right]
= \sum_{j=1}^{J} \left\{ E \left[ (Y_{d_j} (z_{km}, z_{k'_m}) - Y_{d_{j-1}} (z_{km}, z_{k'_m})) \left( \frac{1 \{ D_{z_{k'_m}} \geq d_j \} - 1 \{ D_{z_{km}} \geq d_j \}}{-1 \{ D_{z_{km}} \geq d_j \} + 1 \{ D_{z_{km}} \geq d_j \}} \right) \right] \right\}
\cdot \mathbb{P} \left( 1 \{ D_{z_{k'_m}} \geq d_j \} - 1 \{ D_{z_{km}} \geq d_j \} = 1 \right)
= \sum_{j=1}^{J} E \left[ (Y_{d_j} (z_{km}, z_{k'_m}) - Y_{d_{j-1}} (z_{km}, z_{k'_m})) \left( D_{z_{k'_m}} \geq d_j > D_{z_{km}} \right) \right]
\cdot \mathbb{P} \left( D_{z_{k'_m}} \geq d_j > D_{z_{km}} \right).
Similarly, we have

\[
E \left[ D \mid Z = z_{k_m}' \right] - E \left[ D \mid Z = z_{k_m} \right] \\
= E \left[ \sum_{j=1}^{J} d_j \left( \mathbf{1} \{ D_{z_{k_m}'} \geq d_j \} - \mathbf{1} \{ D_{z_{k_m}} \geq d_j \} \right) \right] \\
- E \left[ \sum_{j=1}^{J} d_j \left( \mathbf{1} \{ D_{z_{k_m}'} \geq d_{j+1} \} - \mathbf{1} \{ D_{z_{k_m}} \geq d_{j+1} \} \right) \right] \\
= E \left[ \sum_{j=1}^{J} d_j \cdot \mathbf{1} \{ D_{z_{k_m}'} \geq d_j > D_{z_{k_m}} \} \right] - E \left[ \sum_{j=1}^{J} d_{j-1} \cdot \mathbf{1} \{ D_{z_{k_m}'} \geq d_j > D_{z_{k_m}} \} \right] \\
= \sum_{j=1}^{J} (d_j - d_{j-1}) P \left( D_{z_{k_m}'} \geq d_j > D_{z_{k_m}} \right).
\]

Thus, finally we have that

\[
\beta_{k_m', k_m} \equiv \frac{E \left[ Y \mid Z = z_{k_m}' \right] - E \left[ Y \mid Z = z_{k_m} \right]}{E \left[ D \mid Z = z_{k_m}' \right] - E \left[ D \mid Z = z_{k_m} \right]} \\
= \sum_{j=1}^{J} \omega_j \cdot E \left[ \left( Y_{d_j} \left( z_{k_m}, z_{k_m}' \right) - Y_{d_{j-1}} \left( z_{k_m}, z_{k_m}' \right) \right) \mid D_{z_{k_m}'} \geq d_j > D_{z_{k_m}} \right],
\]

where

\[
\omega_j = \frac{P \left( D_{z_{k_m}'} \geq d_j > D_{z_{k_m}} \right)}{\sum_{l=1}^{J} (d_l - d_{l-1}) P \left( D_{z_{k_m}'} \geq d_l > D_{z_{k_m}} \right)}.
\]

Note that by definition, \( P(D_{z_{k_m}'} \geq d_1 > D_{z_{k_m}}) = 0 \).

**Proof of Theorem A.1.** For every \( Z(k,k') \in \mathcal{Z}_M \), we define

\[
W_j \left( Z(k,k') \right) = \begin{pmatrix}
g(Z_i) Y_i I (Z_i, Z(k,k'), \mathcal{X}_M) \\
g(Z_i) Y_i I (Z_i, Z(k,k'), \mathcal{Z}_M) \\
g(Z_i) D_i I (Z_i, Z(k,k'), \mathcal{X}_M) \\
g(Z_i) D_i I (Z_i, Z(k,k'), \mathcal{Z}_M) \\
I (Z_i, Z(k,k'), \mathcal{X}_M) \\
I (Z_i, Z(k,k'), \mathcal{Z}_M)
\end{pmatrix},
\]

12
By similar arguments, we have that

\[
\hat{W}_n (Z_{(k,k')}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{g(Z_i) Y_i I(Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0)}{Y_i I(Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0)} + \frac{g(Z_i) I(Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0)}{g(Z_i) D_i I(Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0)} + \frac{g(Z_i) D_i I(Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0)}{I(Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0)} \right\},
\]

and

\[
W(Z_{(k,k')}) = E \left[ W_i (Z_{(k,k')}) \right].
\]

Also, we let

\[
\hat{W}_n = \left( \hat{W}_n^T (Z_{(1,2)}) , \ldots , \hat{W}_n^T (Z_{(1,K)}) , \ldots , \hat{W}_n^T (Z_{(K,1)}) , \ldots , \hat{W}_n^T (Z_{(K,K-1)}) \right)^T
\]

and

\[
W = \left( W^T (Z_{(1,2)}) , \ldots , W^T (Z_{(1,K)}) , \ldots , W^T (Z_{(K,1)}) , \ldots , W^T (Z_{(K,K-1)}) \right)^T.
\]

By assumption, for every small \( \varepsilon > 0 \), we have
\[
P(\sqrt{n}1\{ \hat{\mathcal{Z}}_0 \neq \mathcal{Z}_M \} > \varepsilon) \leq P(\hat{\mathcal{Z}}_0 \neq \mathcal{Z}_M) \to 0.
\]

First, we have that

\[
\sqrt{n} \left| \frac{1}{n} \sum_{i=1}^{n} I(Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0) - \frac{1}{n} \sum_{i=1}^{n} I(Z_i, Z_{(k,k')}, \mathcal{Z}_0) \right| \leq \sqrt{n}1\{ \hat{\mathcal{Z}}_0 \neq \mathcal{Z}_0 \} = o_p(1).
\]

With \( n^{-1} \sum_{i=1}^{n} |g(Z_i) Y_i| \overset{P}{\to} E(|g(Z_i) Y_i|) \) by law of large numbers under the assumptions, we have that

\[
\sqrt{n} \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i) Y_i I(Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0) - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) Y_i I(Z_i, Z_{(k,k')}, \mathcal{Z}_M) \right| \\
= \sqrt{n} \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i) Y_i \left[ I(Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0) - I(Z_i, Z_{(k,k')}, \mathcal{Z}_0) \right] \right| \\
\leq \frac{1}{n} \sum_{i=1}^{n} |g(Z_i) Y_i| \left( \sqrt{n}1\{ \hat{\mathcal{Z}}_0 \neq \mathcal{Z}_M \} \right) = o_p(1).
\]

By similar arguments, we have that

\[
\sqrt{n} \left( \hat{W}_n (Z_{(k,k')}) - W(Z_{(k,k')}) \right) = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} (W_i (Z_{(k,k')}) - W(Z_{(k,k')})) + o_p(1).
\]
Then by multivariate central limit theorem,

\[
\sqrt{n} \left( \bar{W}_n - W \right) = \sqrt{n} \begin{pmatrix}
\bar{W}_n (Z_{(1,2)}) - W (Z_{(1,2)}) \\
\vdots \\
\bar{W}_n (Z_{(K,K-1)}) - W (Z_{(K,K-1)})
\end{pmatrix}
= \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix}
W_i (Z_{(1,2)}) - W (Z_{(1,2)}) \\
\vdots \\
W_i (Z_{(K,K-1)}) - W (Z_{(K,K-1)})
\end{pmatrix} + o_p (1) \overset{d}{\to} N (0, \Sigma_P),
\]

where \( \Sigma_P = E \left[ V_P V_P^T \right] \) and

\[
V_P = \begin{pmatrix}
W_i (Z_{(1,2)}) - W (Z_{(1,2)}) \\
\vdots \\
W_i (Z_{(K,K-1)}) - W (Z_{(K,K-1)})
\end{pmatrix}.
\]

Define a function \( f : \mathbb{R}^6 \to \mathbb{R} \) by

\[
f (x) = \frac{x_1/x_6 - x_2 x_3/x_6^2}{x_4/x_6 - x_5 x_3/x_6} = \frac{x_1 x_6 - x_2 x_3}{x_4 x_6 - x_5 x_3}
\]

for every \( x \in \mathbb{R}^6 \) with \( x = (x_1, x_2, x_3, x_4, x_5)^T, x_4 x_6 - x_5 x_3 \neq 0, \) and \( x_6 \neq 0 \). We can obtain the gradient of \( f \), denoted \( f' \), by \( f' (x) = (f'_1 (x), f'_2 (x), f'_3 (x), f'_4 (x), f'_5 (x), f'_6 (x))^T \) for every \( x = (x_1, x_2, x_3, x_4, x_5, x_6)^T \), where

\[
\begin{align*}
f'_1 (x) &= \frac{x_6}{x_4 x_6 - x_5 x_3}, \\
f'_2 (x) &= \frac{-x_3}{x_4 x_6 - x_5 x_3}, \\
f'_3 (x) &= \frac{-x_2 x_4 x_6 + x_5 x_1 x_6}{(x_4 x_6 - x_5 x_3)^2}, \\
f'_4 (x) &= \frac{-x_1 x_6 - x_2 x_3}{(x_4 x_6 - x_5 x_3)^2}, \\
f'_5 (x) &= \frac{x_3 (x_1 x_6 - x_2 x_3)}{(x_4 x_6 - x_5 x_3)^2}, \quad \text{and} \quad f'_6 (x) = \frac{-x_1 x_5 x_3 + x_2 x_3 x_4}{(x_4 x_6 - x_5 x_3)^2}.
\end{align*}
\]

Let

\[
\mathcal{F} (\hat{W}_n) - \mathcal{F} (W) = \begin{pmatrix}
f (\hat{W}_n (Z_{(1,2)}) - f (W (Z_{(1,2)})) \\
\vdots \\
f (\hat{W}_n (Z_{(K,K-1)}) - f (W (Z_{(K,K-1)}))
\end{pmatrix} = \hat{\beta}_1 - \beta_1.
\]

The Jacobian matrix \( \mathcal{F}' \) of \( \mathcal{F} \) can be obtained with the derivatives of \( f \). Then by delta method, we have that

\[
\sqrt{n} \left( \hat{\beta}_1 - \beta_1 \right) \overset{d}{\to} \mathcal{F}' (W) N (0, \Sigma_P).
\]
Now we have that for every $Z_i \in \mathcal{Z}_{(k,k')}$, we have

\[
\begin{pmatrix}
E \left[ g \left( Z_i \right) Y_1 \{ Z_i \in \mathcal{Z}_{(k,k')} \} \right] & E \left[ Y_1 \{ Z_i \in \mathcal{Z}_{(k,k')} \} \right] & E \left[ g \left( Z_i \right) 1 \{ Z_i \in \mathcal{Z}_{(k,k')} \} \right] \\
\mathbb{P} \left( Z_i \in \mathcal{Z}_{(k,k')} \right) & \mathbb{P} \left( Z_i \in \mathcal{Z}_{(k,k')} \right) & \mathbb{P} \left( Z_i \in \mathcal{Z}_{(k,k')} \right)
\end{pmatrix}
\]

\[
= \sum_{l=1}^{K} \left\{ \begin{array}{l}
\mathbb{P} \left( Z_i = z_k | Z_i \in \mathcal{Z}_{(k,k')} \right) E \left[ Y_i Z_i = z_k \right] \{ g \left( z_k \right) - E \left[ g \left( Z_i \right) | Z_i \in \mathcal{Z}_{(k,k')} \right] \} \\
+ \mathbb{P} \left( Z_i = z_k' | Z_i \in \mathcal{Z}_{(k,k')} \right) E \left[ Y_i Z_i = z_k \right] \{ g \left( z_k' \right) - E \left[ g \left( Z_i \right) | Z_i \in \mathcal{Z}_{(k,k')} \right] \} \\
\end{array} \right\}
\]

By (A.1), we have

\[
E \left[ Y_i Z_i = z_k' \right] = \beta_{k',k} \left( E \left[ D_i Z_i = z_k' \right] - E \left[ D_i Z_i = z_k \right] \right) + E \left[ Y_i Z_i = z_k \right],
\]

and thus it follows that

\[
\begin{align*}
\mathbb{P} \left( Z_i = z_k | Z_i \in \mathcal{Z}_{(k,k')} \right) E \left[ Y_i Z_i = z_k \right] \{ g \left( z_k \right) - E \left[ g \left( Z_i \right) | Z_i \in \mathcal{Z}_{(k,k')} \right] \} \\
+ \mathbb{P} \left( Z_i = z_k' | Z_i \in \mathcal{Z}_{(k,k')} \right) E \left[ Y_i Z_i = z_k \right] \{ g \left( z_k' \right) - E \left[ g \left( Z_i \right) | Z_i \in \mathcal{Z}_{(k,k')} \right] \}
\end{align*}
\]

where we use the equality that

\[
\begin{align*}
\mathbb{P} \left( Z_i = z_k | Z_i \in \mathcal{Z}_{(k,k')} \right) \{ g \left( z_k \right) - E \left[ g \left( Z_i \right) | Z_i \in \mathcal{Z}_{(k,k')} \right] \} \\
+ \mathbb{P} \left( Z_i = z_k' | Z_i \in \mathcal{Z}_{(k,k')} \right) \{ g \left( z_k' \right) - E \left[ g \left( Z_i \right) | Z_i \in \mathcal{Z}_{(k,k')} \right] \} = 0. \tag{C.3}
\end{align*}
\]

Similarly, we have

\[
\begin{align*}
\frac{E \left[ g \left( Z_i \right) D_i 1 \{ Z_i \in \mathcal{Z}_{(k,k')} \} \right]}{\mathbb{P} \left( Z_i \in \mathcal{Z}_{(k,k')} \right)} - \frac{E \left[ D_i 1 \{ Z_i \in \mathcal{Z}_{(k,k')} \} \right]}{\mathbb{P} \left( Z_i \in \mathcal{Z}_{(k,k')} \right)} \frac{E \left[ g \left( Z_i \right) 1 \{ Z_i \in \mathcal{Z}_{(k,k')} \} \right]}{\mathbb{P} \left( Z_i \in \mathcal{Z}_{(k,k')} \right)}
\end{align*}
\]

\[
= \mathbb{P} \left( Z_i = z_k' | Z_i \in \mathcal{Z}_{(k,k')} \right) \{ p \left( z_k' \right) - p \left( z_k \right) \} \{ g \left( z_k' \right) - E \left[ g \left( Z_i \right) | Z_i \in \mathcal{Z}_{(k,k')} \right] \},
\]

where $p(z) = E \left[ D_i | Z_i = z \right]$ for all $z$ and we use the equality in (C.3) again. □
Proof of Theorem A.2. For every \( Z_{(k,k')} \in \mathcal{L}_M \), we define

\[
W_i \left( Z_{(k,k')} \right) = \begin{pmatrix}
g(Z_i) Y_i \mathcal{I} \left( Z_i, Z_{(k,k')}, \mathcal{Z}'_0 \right)
Y_i \mathcal{I} \left( Z_i, Z_{(k,k')}, \mathcal{Z}_0 \right)
g(Z_i) \mathcal{I} \left( Z_i, Z_{(k,k')}, \mathcal{Z}'_0 \right)
g(Z_i) D_i \mathcal{I} \left( Z_i, Z_{(k,k')}, \mathcal{Z}'_0 \right)
D_i \mathcal{I} \left( Z_i, Z_{(k,k')}, \mathcal{Z}'_0 \right)
\end{pmatrix},
\]

and

\[
\hat{W}_n \left( Z_{(k,k')} \right) = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix}
g(Z_i) Y_i \mathcal{I} \left( Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0 \right)
Y_i \mathcal{I} \left( Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0 \right)
g(Z_i) \mathcal{I} \left( Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0 \right)
g(Z_i) D_i \mathcal{I} \left( Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0 \right)
D_i \mathcal{I} \left( Z_i, Z_{(k,k')}, \hat{\mathcal{Z}}_0 \right)
\end{pmatrix},
\]

By proof similar to that of Theorem A.1, \( \hat{W}_n \left( Z_{(k,k')} \right) \xrightarrow{p} W \left( Z_{(k,k')} \right) \).

For every random variable \( \xi_i \) and every \( \mathcal{A} \in \mathcal{Z}' \), we define

\[
\mathcal{E}'_n \left( \xi_i, \mathcal{A} \right) = \frac{1}{n} \sum_{i=1}^{n} \xi_i \mathcal{I} \left( Z_i \in \mathcal{A}, \mathcal{A} \in \mathcal{Z}_P \right) \quad \text{and} \quad \mathcal{E}' \left( \xi_i, \mathcal{A} \right) = \frac{E \left[ \xi_i \mathcal{I} \left( Z_i \in \mathcal{A}, \mathcal{A} \in \mathcal{Z}_P \right) \right]}{E \left[ \mathcal{I} \left( Z_i \in \mathcal{A}, \mathcal{A} \in \mathcal{Z}_P \right) \right]}.
\]

Then we obtain the VSIV estimator using \( \mathcal{Z}_P \) for each ACR as

\[
\hat{\beta}'_{(k,k')} = \frac{\mathcal{E}'_n \left( g(Z_i) Y_i, Z_{(k,k')} \right) - \mathcal{E}'_n \left( g(Z_i), Z_{(k,k')} \right) \mathcal{E}'_n \left( Y_i, Z_{(k,k')} \right)}{\mathcal{E}'_n \left( g(Z_i) D_i, Z_{(k,k')} \right) - \mathcal{E}' \left( g(Z_i), Z_{(k,k')} \right) \mathcal{E}' \left( Y_i, Z_{(k,k')} \right)},
\]

which converges in probability to

\[
\beta'_{(k,k')} = \frac{\mathcal{E}' \left( g(Z_i) Y_i, Z_{(k,k')} \right) - \mathcal{E}' \left( g(Z_i), Z_{(k,k')} \right) \mathcal{E}' \left( Y_i, Z_{(k,k')} \right)}{\mathcal{E}' \left( g(Z_i) D_i, Z_{(k,k')} \right) - \mathcal{E}' \left( g(Z_i), Z_{(k,k')} \right) \mathcal{E}' \left( D_i, Z_{(k,k')} \right)}.
\]

We also define

\[
\mathcal{E}''_n \left( \xi_i, \mathcal{A} \right) = \frac{1}{n} \sum_{i=1}^{n} \xi_i \mathcal{I} \left( Z_i \in \mathcal{A}, \mathcal{A} \in \hat{\mathcal{Z}}_0' \right) \quad \text{and} \quad \mathcal{E}'' \left( \xi_i, \mathcal{A} \right) = \frac{E \left[ \xi_i \mathcal{I} \left( Z_i \in \mathcal{A}, \mathcal{A} \in \hat{\mathcal{Z}}_0' \right) \right]}{E \left[ \mathcal{I} \left( Z_i \in \mathcal{A}, \mathcal{A} \in \hat{\mathcal{Z}}_0' \right) \right]}.
\]
Then we obtain the VSIV estimator using \( \hat{\beta}'_{(k,k')} \) for each ACR as

\[
\hat{\beta}'_{(k,k')} = \frac{\varepsilon''(g(Z_i)Y_i, Z_{(k,k')}) - \varepsilon''(g(Z_i), Z_{(k,k')})}{\varepsilon''(g(Z_i)D_i, Z_{(k,k')}) - \varepsilon''(g(Z_i), Z_{(k,k')})} E_n(Y_i, Z_{(k,k')}),
\]

which converges in probability to

\[
\beta''_{(k,k')} = \frac{\varepsilon''(g(Z_i)Y_i, Z_{(k,k')}) - \varepsilon''(g(Z_i), Z_{(k,k')})}{\varepsilon''(g(Z_i)D_i, Z_{(k,k')}) - \varepsilon''(g(Z_i), Z_{(k,k')})} E_n(Y_i, Z_{(k,k')}).
\]

If \( Z_{(k,k')} \notin \mathcal{L}_M \) and \( Z_{(k,k')} \notin \mathcal{L}_F \), then \( \beta''_{(k,k')} = 0 \). In this case, it is possible that \( \hat{\beta}'_{(k,k')} = 0 \), because by definition \( \mathcal{L}_0 \subset \mathcal{L}_F \). Note that if \( Z_{(k,k')} \notin \mathcal{L}_0 \), then \( \beta''_{(k,k')} = \beta'_{(k,k')} \) by definition.

If \( Z_{(k,k')} \notin \mathcal{L}_M \) and \( Z_{(k,k')} \notin \mathcal{L}_F \), then \( \beta''_{(k,k')} = \beta'_{(k,k')} = 0 \). Similarly, in this case, \( \beta''_{(k,k')} = \beta'_{(k,k')} = 0 \) because \( \mathcal{L}_0 \subset \mathcal{L}_F \).

If \( Z_{(k,k')} \in \mathcal{L}_M \) and \( Z_{(k,k')} \notin \mathcal{L}_F \), then \( \beta''_{(k,k')} = \beta'_{(k,k')} = 0 \) because \( \mathcal{L}_0 \subset \mathcal{L}_F \).

If \( Z_{(k,k')} \in \mathcal{L}_M \) and \( Z_{(k,k')} \notin \mathcal{L}_F \), then \( \beta''_{(k,k')} = 0 \) because \( \mathcal{L}_0 \subset \mathcal{L}_F \).

### C.2 Testable Implications of Kédagni and Mourifié (2020)

We consider the case where \( D \in \mathcal{D} = \{d_1, \ldots, d_J \} \). The testable implications in Kédagni and Mourifié (2020) are for exclusion \( (Y_{dzkm} = Y_{dzm} \text{ for } d \in \mathcal{D}) \) and statistical independence \((Y_{dzkm}, Y_{dzkm}, \ldots, Y_{dzkm}, Y_{dzkm} \perp Z)\) for every \( m \in \{1, \ldots, M\} \) with the largest validity pair set \( \mathcal{L}_M = \{ (z_{k1}, z_{k1}'), \ldots, (z_{km}, z_{km}') \} \). Under these conditions, we can define \( Y_d(z, z') \) as in Section 3.2 for every \( d \in \mathcal{D} \) and every \( (z, z') \in \mathcal{L}_M \). Define

\[
f_{Y:D}(y, d|z) = f_{Y:D,Z}(y|d, z)P(D = d|Z = z)
\]

for every \( y \in \mathbb{R} \), every \( d \in \mathcal{D} \), and every \( z \in \mathcal{Z} \), where \( f_{Y:D,Z}(y|d, z) \) is the conditional density function of \( Y \) given \( D = d \) and \( Z = z \). For every \( Z_{(k,k')} = (z_k, z_{k'}) \in \mathcal{L}_M \), every \( A \in \mathcal{B}_\mathbb{R} \), every \( d \in \mathcal{D} \), and each \( z \in Z_{(k,k')} \),

\[
P(Y \in A, D = d|Z = z) \leq P(Y_{dz} \in A|Z = z) = P(Y_d(z_k, z_{k'}) \in A),
\]
\[ P(Y \in A, D = d | Z = z) = \frac{P(Y \in A, D = d, Z = z)}{P(Z = z)} = P(Y \in A | D = d, Z = z) P(D = d | Z = z). \]

Then, by the discussion in Section 4.1 of Kédagni and Mourifié (2020), for (almost) all \( y \),
\[ f_{Y,D}(y, d|z) = f_{Y|D,Z}(y|d, z) P(D = d | Z = z) \leq f_{Y_d(z_k, z_{k'})}(y), \]
where \( f_{Y_d(z_k, z_{k'})} \) is the density function of the potential outcome \( Y_d(z_k, z_{k'}). \) Thus, for every \( d \in D \),
\[ \max_{z \in Z(k, k')} f_{Y,D}(y, d|z) \leq f_{Y_d(z_k, z_{k'})}(y), \quad \tag{C.4} \]
and we obtain the first inequality of Kédagni and Mourifié (2020):
\[ \max_{d \in D} \int_{\mathbb{R}} \max_{z \in Z(k, k')} f_{Y,D}(y, d|z) \, dy \leq 1. \]

Also, for all \( A_1, \ldots, A_J \in \mathcal{B}_\mathbb{R} \),
\[
\begin{align*}
P(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_j}(z_k, z_{k'}) \in A_j) \\
= \min_{z \in Z(k, k')} \ P(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_j}(z_k, z_{k'}) \in A_j | Z = z) \\
= \min_{z \in Z(k, k')} \sum_{j=1}^{J} \ P(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_j}(z_k, z_{k'}) \in A_j, D = d_j | Z = z) \\
\leq \min_{z \in Z(k, k')} \sum_{j=1}^{J} \ P(Y \in A_j, D = d_j | Z = z). 
\end{align*}
\]

Let \( P^j_\mathbb{R} \) be an arbitrary partition of \( \mathbb{R} \) for \( j \in \{1, \ldots, J\} \), that is, \( P^j_\mathbb{R} = \{C^j_1, \ldots, C^j_{N_j}\} \) with \( \bigcup_{j=1}^{N_j} C^j_l = \mathbb{R} \) and \( C^j_l \cap C^j_{l'} = \emptyset \) for all \( l' \neq l \). Then
\[
1 = \sum_{A_1 \in P^1_\mathbb{R}} \cdots \sum_{A_J \in P^J_\mathbb{R}} P(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J) \\
\leq \sum_{A_1 \in P^1_\mathbb{R}} \cdots \sum_{A_J \in P^J_\mathbb{R}} \min_{z \in Z(k, k')} \sum_{j=1}^{J} P(Y \in A_j, D = d_j | Z = z). 
\]
Then we obtain the second inequality of Kédagni and Mourifié (2020):

$$\inf_{\{P^1_R, \ldots, P^J_R\}} \sum_{A_1 \in P^1_R} \cdots \sum_{A_J \in P^J_R} \min_{z \in Z_{(k,k')}} \sum_{j=1}^J \mathbb{P}(Y \in A_j, D = d_j | Z = z) \geq 1,$$

where the infimum is taken over all partitions \(\{P^1_R, \ldots, P^J_R\}\). Next, for all \(A_1, \ldots, A_J \in B_R\),

$$\mathbb{P}(Y_{d_j}(z_k, z_{k'}) \in A_j) \leq \sum_{A_1 \in P^1_R} \cdots \sum_{A_j-1 \in P^{j-1}_R} \sum_{A_{j+1} \in P^{j+1}_R} \cdots \sum_{A_J \in P^J_R} \min_{z \in Z_{(k,k')}} \sum_{\xi=1}^J \mathbb{P}(Y \in A_{\xi}, D = d_{\xi} | Z = z),$$

which, together with (C.4), implies the third inequality of Kédagni and Mourifié (2020):

$$\sup_{\{P^1_R, \ldots, P^J_R\}} \max_{j \in \{1, \ldots, J\}} \left\{ \int_{A_j} \max_{z \in Z_{(k,k')}} f_{Y,D}(y,d_{\xi}|z) \, dy - \varphi_j(A_j, Z_{(k,k')}, P^1_R, \ldots, P^J_R) \right\} \leq 0,$$

where

$$\varphi_j(A_j, W, P^1_R, \ldots, P^J_R) = \sum_{A_1 \in P^1_R} \cdots \sum_{A_{j-1} \in P^{j-1}_R} \sum_{A_{j+1} \in P^{j+1}_R} \cdots \sum_{A_J \in P^J_R} \min_{z \in W} \sum_{\xi=1}^J \int_{A_{\xi}} f(y,d_{\xi}|z) \, dy$$

for all \(W \subset Z\).

### C.3 Estimation of \(Z_0\)

#### C.3.1 Definition and Estimation of \(Z_1\)

We follow the notation of Sun (2021) to introduce the definition of \(Z_1\) and the corresponding estimator. Define conditional probabilities

$$P_z(B, C) = \mathbb{P}(Y \in B, D \in C | Z = z)$$
for all Borel sets \( B, C \in \mathcal{B}_\mathbb{R} \) and all \( z \in \mathcal{Z} \). The testable implication proposed by Sun (2021)\(^{13}\) for the conditions in Definition A.1 is that for every \( m \in \{1, \ldots, M\} \),

\[
P_{z_{km}}(B, \{d_J\}) \leq P_{z_{km}}(B, \{d_J\}) \quad \text{and} \quad P_{z_{km}}(B, \{d_1\}) \geq P_{z_{km}}(B, \{d_1\})
\]

(C.5)

for all \( B \in \mathcal{B}_\mathbb{R} \), and

\[
P_{z_{km}}(\mathbb{R}, C) \geq P_{z_{km}}(\mathbb{R}, C)
\]

(C.6)

for all \( C = (-\infty, c] \) with \( c \in \mathbb{R} \). Without loss of generality, we assume that \( d_1 = 0 \) and \( d_J = 1 \). By definition, for all \( B, C \in \mathcal{B}_\mathbb{R} \),

\[
\mathbb{P}(Y \in B, D \in C | Z = z) = \frac{\mathbb{P}(Y \in B, D \in C, Z = z)}{\mathbb{P}(Z = z)}.
\]

Next, we reformulate the testable restrictions to define \( \mathcal{X}_1 \) and its estimator. Define the following function spaces

\[
\mathcal{G}_P = \left\{ (1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_{k'}\}}) : k, k' \in \{1, \ldots, K\}, k \neq k' \right\},
\]

\[
\mathcal{H}_1 = \left\{ (-1)^d \cdot 1_{B \times \{d\} \times \mathbb{R}} : B \text{ is a closed interval in } \mathbb{R}, d \in \{0, 1\} \right\},
\]

\[
\mathcal{H}_1^\prime = \left\{ (-1)^d \cdot 1_{B \times \{d\} \times \mathbb{R}} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}, d \in \{0, 1\} \right\},
\]

\[
\mathcal{H}_2 = \{1_{\mathbb{R} \times C \times \mathbb{R}} : C = (-\infty, c], c \in \mathbb{R}\},
\]

\[
\mathcal{H}_2^\prime = \{1_{\mathbb{R} \times C \times \mathbb{R}} : C = (-\infty, c] \text{ or } C = (-\infty, c), c \in \mathbb{R}\},
\]

\[
\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2, \text{ and } \mathcal{H}^\prime = \mathcal{H}_1^\prime \cup \mathcal{H}_2^\prime.
\]

(C.7)

Let \( P, \phi, \sigma^2, \tilde{P}, \hat{\phi}, \text{ and } \tilde{\sigma}^2 \) be defined in a way similar to that in Section 2 but for all \((h, g) \in \mathcal{H} \times \mathcal{G}_P\). Also, we let \( \Lambda(P) = \prod_{k=1}^K P(1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}) \) and \( T_n = n \cdot \prod_{k=1}^K \tilde{P}(1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}) \).

By similar proof of Lemma 3.1 in Sun (2021), \( \sigma^2 \) and \( \tilde{\sigma}^2 \) are uniformly bounded in \((h, g) \in \mathcal{H} \times \mathcal{G}_P\).

The following lemma reformulates the testable restrictions in terms of \( \phi \).

**Lemma C.1** Suppose that the instrument \( Z \) is pairwise valid for the treatment \( D \) with the largest validity pair set \( \mathcal{X}_M = \{(z_{k_1}, z_{k_1'}), \ldots, (z_{k_M}, z_{k_M'})\} \). For every \( m \in \{1, \ldots, M\} \), we have that \( \sup_{h \in \mathcal{H}} \phi(h, g) = 0 \) with \( g = (1_{\mathbb{R} \times \mathbb{R} \times \{z_{km}\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_{km'}\}}) \).

**Proof of Lemma C.1.** Note that for every \( g \in \mathcal{G}_P \), we can always find some \( a \in \mathbb{R} \)

\(^{13}\) The testable implications proposed by Sun (2021) are originally for full IV validity. We can easily obtain the testable implications for the conditions in Definition A.1 following the proof of Sun (2021).
such that $\phi(h, g) = 0$ with $h = 1_{a \times \{0\} \times \mathbb{R}}$. So $\sup_{h \in \mathcal{H}} \phi(h, g) \geq 0$ for every $g \in \mathcal{G}_P$. Under assumption, for every $g = (1_{\mathbb{R} \times \mathbb{R} \times \{z_m\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_n\}})$, by Lemma 2.1 of Sun (2021), $\phi(h, g) \leq 0$ for all $h \in \mathcal{H}$. Thus, $\sup_{h \in \mathcal{H}} \phi(h, g) = 0$. ■

Lemma C.1 provides a necessary condition for $\mathcal{Z}_M$. By Lemma C.1, we define

$$
\mathcal{G}_1 = \left\{ g \in \mathcal{G}_P : \sup_{h \in \mathcal{H}} \phi(h, g) = 0 \right\} \quad \text{and} \quad \hat{\mathcal{G}}_1 = \left\{ g \in \mathcal{G}_P : \sqrt{T_n} \left( \sup_{h \in \mathcal{H}} \frac{\hat{\phi}(h, g)}{\xi_0} \right) \leq \tau_n \right\}
$$

with $\tau_n \to \infty$ and $\tau_n / \sqrt{n} \to 0$ as $n \to \infty$, where $\xi_0$ is a small positive number. We define $\mathcal{Z}_1$ as the collection of all $(z, z')$ that are associated with some $g \in \mathcal{G}_1$:

$$
\mathcal{Z}_1 = \left\{ (z_k, z'_k) \in \mathcal{Z} : g = (1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z'_k\}}) \in \mathcal{G}_1 \right\}.
$$

We use $\hat{\mathcal{G}}_1$ to construct the estimator of $\mathcal{Z}_1$, denoted by $\hat{\mathcal{Z}}_1$, which is defined as the set of all $(z, z')$ that are associated with some $g \in \hat{\mathcal{G}}_1$ in the same way $\mathcal{Z}_1$ is defined based on $\mathcal{G}_1$:

$$
\hat{\mathcal{Z}}_1 = \left\{ (z_k, z'_k) \in \mathcal{Z} : g = (1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z'_k\}}) \in \hat{\mathcal{G}}_1 \right\}.
$$

To establish consistency of $\hat{\mathcal{Z}}_1$, we state and prove an auxiliary lemma.

**Lemma C.2** Under Assumption A.2, $\hat{\phi} \to \phi$, $T_n/n \to \Lambda(P)$, and $\hat{\sigma} \to \sigma$ almost uniformly. In addition, $\sqrt{T_n}(\hat{\phi} - \phi) \to G$ for some random element $G$, and for all $(h, g) \in \mathcal{H} \times \mathcal{G}_P$ with $g = (g_1, g_2)$, the variance $\text{Var}(G(h, g)) = \sigma^2(h, g)$.

**Proof of Lemma C.2.** Note that the $\mathcal{G}_P$ defined in (C.7) is only slightly different from the $\mathcal{G}$ defined in (7) of Sun (2021). The lemma can be proved following a strategy similar to that of the proofs of Lemmas C.11 and 3.1 of Sun (2021). ■

The following proposition establishes consistency of $\hat{\mathcal{Z}}_1$.

**Proposition C.1** Under Assumptions A.2 and A.3, $P(\hat{\mathcal{G}}_1 = \mathcal{G}_1) \to 1$, and thus $P(\hat{\mathcal{Z}}_1 = \mathcal{Z}_1) \to 1$.

**Proof of Proposition C.1.** First, suppose $\mathcal{G}_1 \neq \emptyset$. Under the constructions, we have that

---

14See the definition of almost uniform convergence in van der Vaart and Wellner (1996, p. 52).
for all $\varepsilon > 0$,}
\[
\lim_{n \to \infty} \mathbb{P} \left( G_1 \setminus \hat{G}_1 \neq \emptyset \right) \\
\leq \lim_{n \to \infty} \mathbb{P} \left( \max_{g \in \tilde{G}_1} \sup_{h \in H} \frac{\tilde{\phi}(h, g) - \phi(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} > \tau_n \right) \\
\leq \lim_{n \to \infty} \mathbb{P} \left( \max_{g \in \tilde{G}_1 \setminus G_1} \sup_{h \in H} \frac{\tilde{\phi}(h, g) - \phi(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} > \tau_n \right) .
\]

By Lemma C.2, $\sqrt{T_n}(\hat{\phi} - \phi) \rightsquigarrow G$ and $\hat{\sigma} \to \sigma$ almost uniformly, which implies that $\tilde{\sigma} \rightsquigarrow \sigma$ by Lemmas 1.9.3(ii) and 1.10.2(iii) of van der Vaart and Wellner (1996). Thus by Example 1.4.7 (Slutsky’s lemma) and Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996),
\[
\max_{g \in \tilde{G}_1 \setminus G_1} \sup_{h \in H} \sqrt{T_n} \left| \frac{\tilde{\phi}(h, g) - \phi(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \right| \rightsquigarrow \max_{g \in \tilde{G}_1 \setminus G_1} \sup_{h \in H} \left| \frac{G(h, g)}{\xi_0 \vee \sigma(h, g)} \right| .
\]

Since $\tau_n \to \infty$, we have that $\lim_{n \to \infty} \mathbb{P}(G_1 \setminus \hat{G}_1 \neq \emptyset) = 0$.

If $G_1 = G_P$, then clearly $\lim_{n \to \infty} \mathbb{P}(\hat{G}_1 \setminus G_1 \neq \emptyset) = 0$. Suppose $G_1 \neq G_P$. Since $G_P$ is a finite set and $\tilde{\sigma}$ is uniformly bounded in $(h, g)$ by construction, then there is a $\delta > 0$ such that $\min_{g \in G_P \setminus G_1} \sup_{h \in H} \phi(h, g) / \xi_0 \vee \hat{\sigma}(h, g) > \delta$. Thus, we have that
\[
\lim_{n \to \infty} \mathbb{P} \left( \hat{G}_1 \setminus G_1 \neq \emptyset \right) \\
\leq \lim_{n \to \infty} \mathbb{P} \left( \max_{g \in \tilde{G}_1 \setminus G_1} \sup_{h \in H} \frac{\phi(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} > \delta, \sup_{g \in \tilde{G}_1 \setminus G_1} \sqrt{T_n} \sup_{h \in H} \frac{\tilde{\phi}(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \leq \tau_n \right) .
\]

By Lemma C.2, $\tilde{\phi} \to \phi$ almost uniformly. Thus, for every $\varepsilon > 0$, there is a measurable set $A$ with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that for sufficiently large $n$,
\[
\max_{g \in \tilde{G}_1 \setminus G_1} \sup_{h \in H} \left| \frac{\tilde{\phi}(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \right| \geq \max_{g \in \tilde{G}_1 \setminus G_1} \sup_{h \in H} \frac{\phi(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} - \delta .
\]
uniformly on \( A \). We now have that

\[
\lim_{n \to \infty} \mathbb{P}\left( \hat{G}_1 \setminus G_1 \neq \emptyset \right) \\
\leq \lim_{n \to \infty} \mathbb{P}\left( \bigcap \left\{ \max_{y \in \hat{G}_1 \setminus G_1} \sup_{\lambda \in \mathcal{H}} P_{\lambda y}^U (h, g) \bigg| \sup_{\lambda \in \mathcal{H}} \frac{\phi(h, g)}{\lambda y} \leq \tau_n \right\} \cap A ) + \mathbb{P}(A^c) \\
\leq \lim_{n \to \infty} \mathbb{P}\left( \sqrt{\frac{T_n}{n}} \frac{\delta}{\epsilon} < \max_{y \in \hat{G}_1} \frac{\phi(h, g)}{\xi_0} \right) \sup_{\lambda \in \mathcal{H}} \frac{\phi(h, g)}{\lambda y} \leq \frac{\tau_n}{\sqrt{n}} + \epsilon = \epsilon,
\]

because \( \tau_n/\sqrt{n} \to 0 \) as \( n \to \infty \). Here, \( \epsilon \) can be arbitrarily small. Thus we have that \( \mathbb{P}(\hat{G}_1 = G_1) \to 1 \), because \( \mathbb{P}(\hat{G}_1 \setminus G_1 \neq \emptyset ) \to 0 \) and \( \mathbb{P}(\hat{G}_1 \setminus G_1 \neq \emptyset ) \to 0 \).

Second, suppose \( G_1 = \emptyset \). This implies that \( \min_{y \in \Psi} \{ \sup_{\lambda \in \mathcal{H}} P_{\lambda y}^U (h, g) / (\xi_0 \vee \hat{\sigma} (h, g)) \} > \delta \)
for some \( \delta > 0 \). Since by Lemma C.2, \( \hat{\sigma} \to \hat{\sigma} \) almost uniformly, then there is a measurable
set \( A \) with \( \mathbb{P}(A) \geq 1-\epsilon \) such that for sufficiently large \( n \),

\[
\max_{y \in \hat{G}_1} \sup_{\lambda \in \mathcal{H}} \frac{\phi(h, g)}{\xi_0} \sup_{\lambda \in \mathcal{H}} \frac{\phi(h, g)}{\lambda y} \geq \max_{y \in \hat{G}_1} \sup_{\lambda \in \mathcal{H}} \frac{\phi(h, g)}{\xi_0} \sup_{\lambda \in \mathcal{H}} \frac{\phi(h, g)}{\lambda y} - \frac{\delta}{2}
\]

uniformly on \( A \). Thus we now have that

\[
\lim_{n \to \infty} \mathbb{P}\left( \hat{G}_1 \neq \emptyset \right) \leq \lim_{n \to \infty} \mathbb{P}\left( \bigcap \left\{ \max_{y \in \hat{G}_1} \sup_{\lambda \in \mathcal{H}} P_{\lambda y}^U (h, g) \bigg| \sup_{\lambda \in \mathcal{H}} \frac{\phi(h, g)}{\lambda y} \leq \tau_n \right\} \cap A ) + \mathbb{P}(A^c) \\
\leq \lim_{n \to \infty} \mathbb{P}\left( \sqrt{\frac{T_n}{n}} \frac{\delta}{\epsilon} < \max_{y \in \hat{G}_1} \frac{\phi(h, g)}{\xi_0} \sup_{\lambda \in \mathcal{H}} \frac{\phi(h, g)}{\lambda y} \leq \frac{\tau_n}{\sqrt{n}} + \epsilon = \epsilon,
\]

because \( \tau_n/\sqrt{n} \to 0 \) as \( n \to \infty \). Here, \( \epsilon \) can be arbitrarily small. Thus, \( \mathbb{P}(\hat{G}_1 = G_1) = 1 - \mathbb{P}(\hat{G}_1 \neq \emptyset ) \to 1 \).

As mentioned after Proposition 3.1, Proposition C.1 is related to the contact set estimation in Sun (2021). Since \( G_1 \subset \Psi \) and \( \Psi \) is a finite set, we can use techniques similar to those in Sun (2021) to obtain the stronger result in Proposition C.1, that is, \( \mathbb{P}(\hat{G}_1 = G_1) \to 1 \).

### C.3.2 Definition and Estimation of \( \mathcal{Y}_2 \)

The definition of \( \mathcal{Y}_2 \) relies on the testable implications in Kédagni and Mourifié (2020) for the exclusion restriction \( Y_{d_{z_k}} = Y_{d_{z_k}} \) for \( d \in \mathcal{D} \) and the independence condition \( (Y_{d_1 z_k}, Y_{d_2 z_k}, \ldots, Y_{d_j z_k}, Y_{d_j z_k}) \perp Z) \) for every \( m \in \{1, \ldots, M\} \) with the largest validity pair set \( \mathcal{Y}_N = \{(z_k, z'_{k}), \ldots, (z_{k_M}, z'_{k_M})\} \). Under these conditions, we can define
\(Y_d(z, z')\) for each \(d \in D\) and every \((z, z') \in \mathcal{X}_M\) such that \(Y_d(z, z') = Y_d = Y_{d'}\) a.s.

We consider the case where \(Y\) is continuous. Similar results can be obtained easily when \(Y\) is discrete. As in Section 3.2, to avoid theoretical and computational complications, we introduce the following testable implications that are slightly weaker than (and implied by) the original testable restrictions in Kédagni and Mourifié (2020) (see Appendix C.2).

We start by generalizing the notation in Section 3.2 to the multivalued treatments. Let \(\mathcal{R}\) denote the collection of all subsets \(C \subset \mathbb{R}\) such that \(C = (a, b]\) with \(a, b \in \mathbb{R}\) and \(a < b\). For every \(Z_{(k, k')} = (z_k, z_{k'}) \in \mathcal{X}_M\), every \(A \in \mathcal{B}_R\), every \(d \in D\), and each \(z \in Z_{(k, k')}\),

\[
P(Y \in A, D = d | Z = z) \leq P(Y_{dz} \in A | Z = z) = P(Y_d(z_k, z_{k'}) \in A),
\]

which implies that

\[
\max_{z \in Z_{(k, k')}} P(Y \in A, D = d | Z = z) \leq P(Y_d(z_k, z_{k'}) \in A). \tag{C.11}
\]

Let \(\mathcal{P}\) be a prespecified finite collection of partitions \(P_\mathcal{R}\) of \(\mathbb{R}\) such that \(P_\mathcal{R} = \{C_1, \ldots, C_N\}\) with \(C_k \in \mathcal{R}\) for all \(k\), \(\bigcup_{k=1}^N C_k = \mathbb{R}\), and \(C_k \cap C_l = \emptyset\) for all \(k \neq l\). Then we obtain the first condition:

\[
\max_{P_\mathcal{R} \in \mathcal{P}} \max_{d \in D} \sum_{A \in P_\mathcal{R}} \max_{z \in Z_{(k, k')}} P(Y \in A, D = d | Z = z) \leq \max_{P_\mathcal{R} \in \mathcal{P}} \max_{d \in D} \sum_{A \in P_\mathcal{R}} P(Y_d(z_k, z_{k'}) \in A) = 1. \tag{C.12}
\]

Also, for all \(A_1, \ldots, A_J \in \mathcal{B}_R\),

\[
P(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J) = \min_{z \in Z_{(k, k')}} P(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J | Z = z)
= \min_{z \in Z_{(k, k')}} \sum_{j=1}^J P(Y_{d_j}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_j}(z_k, z_{k'}) \in A_J, D = d_j | Z = z)
\leq \min_{z \in Z_{(k, k')}} \sum_{j=1}^J P(Y \in A_j, D = d_j | Z = z).
\]
Let $P^1_R, \ldots, P^J_R \in \mathcal{P}$. It follows that

$$1 = \sum_{A_1 \in P^1_R} \cdots \sum_{A_J \in P^J_R} \mathbb{P}(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J)$$

$$\leq \sum_{A_1 \in P^1_R} \cdots \sum_{A_J \in P^J_R} \min_{z \in Z(k,k')} \sum_{j=1}^J \mathbb{P}(Y \in A_j, D = d_j | Z = z).$$

Then we obtain the second condition:

$$\min_{P^1_R, \ldots, P^J_R \in \mathcal{P}} \sum_{A_1 \in P^1_R} \cdots \sum_{A_J \in P^J_R} \min_{z \in Z(k,k')} \sum_{j=1}^J \mathbb{P}(Y \in A_j, D = d_j | Z = z) \geq 1. \quad \text{(C.13)}$$

Next, for all $A_1, \ldots, A_J \in \mathcal{B}_R$,

$$\mathbb{P}(Y_{d_j}(z_k, z_{k'}) \in A_j)$$

$$= \sum_{A_1 \in P^1_R} \cdots \sum_{A_{j-1} \in P^{j-1}_R} \sum_{A_{j+1} \in P^{j+1}_R} \cdots \sum_{A_J \in P^J_R} \mathbb{P}(Y_1(z_k, z_{k'}) \in A_1, \ldots, Y_J(z_k, z_{k'}) \in A_J)$$

$$\leq \sum_{A_1 \in P^1_R} \cdots \sum_{A_{j-1} \in P^{j-1}_R} \sum_{A_{j+1} \in P^{j+1}_R} \cdots \sum_{A_J \in P^J_R} \min_{z \in Z(k,k')} \sum_{\xi=1}^J \mathbb{P}(Y \in A_\xi, D = d_\xi | Z = z),$$

which, together with (C.11), implies the third condition:

$$\max_{P^1_R, \ldots, P^J_R \in \mathcal{P}} \max_{j \in \{1, \ldots, J\}} \sup_{A_j \in \mathcal{B}_R} \left\{ \max_{z \in Z(k,k')} \mathbb{P}(Y \in A_j, D = d_j | Z = z) - \varphi_j \left( A_j, Z(k,k'), P^1_R, \ldots, P^J_R \right) \right\} \leq 0, \quad \text{(C.14)}$$

where

$$\varphi_j \left( A_j, W, P^1_R, \ldots, P^J_R \right)$$

$$= \sum_{A_1 \in P^1_R} \cdots \sum_{A_{j-1} \in P^{j-1}_R} \sum_{A_{j+1} \in P^{j+1}_R} \cdots \sum_{A_J \in P^J_R} \min_{z \in W} \sum_{\xi=1}^J \mathbb{P}(Y \in A_\xi, D = d_\xi | Z = z)$$

for all $W \subset Z$.

Next, we reformulate the testable implications in (C.12)–(C.14) to define $\mathcal{X}_2$ and $\hat{\mathcal{X}}_2$. 25
Define the function spaces

\[ G_Z = \{1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}} : 1 \leq k \leq K\} , \mathcal{H}_D = \{1_{\mathbb{R} \times \{d\} \times \mathbb{R}} : d \in \mathcal{D}\} , \mathcal{H}_B = \{1_{\mathbb{R} \times \mathbb{R}} : B \in \mathcal{R}\}, \]

and \( \mathcal{H}_B = \{1_{\mathbb{R} \times \mathbb{R}} : B \) is a closed, open, or half-closed interval in \( \mathbb{R} \} . \)

(C.15)

Define a map \( \psi : \mathcal{H}_B \times \mathcal{H}_D \times G_Z \rightarrow \mathbb{R} \) such that

\[ \psi(h, f, g) = \frac{P(h \cdot f \cdot g)}{P(g)} \]

for every \((h, f, g) \in \mathcal{H}_B \times \mathcal{H}_D \times G_Z\). Let \( \mathbb{H} \) and \( P(G_Z) \) be defined as in Section 3.2. Then for every \( G_S \in P(G_Z)\), define

\[ \psi_1(G_S) = \max_{P_S \in \mathcal{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in \mathbb{H}(P_S)} \max_{g \in \mathcal{G}_S} \psi(h, f, g) - 1, \]

\[ \psi_2(G_S) = 1 - \min_{P^1_S, \ldots, P^J_S \in \mathcal{P}} \sum_{h_1 \in \mathbb{H}(P^1_S)} \cdots \sum_{h_J \in \mathbb{H}(P^J_S)} \min_{g \in \mathcal{G}_S} \sum_{j=1}^J \psi(h_j, f_j, g), \]

and

\[ \psi_3(G_S) = \max_{P^1_S, \ldots, P^J_S \in \mathcal{P}} \max_{f \in \mathcal{H}_D} \sup_{h_j \in \mathbb{H}(P^j_S)} \left\{ \max_{g \in \mathcal{G}_S} \psi(h, f, g) - \tilde{\varphi}_j(h_j, G_S, P^1_S, \ldots, P^J_S) \right\}, \]

where \( f_j = 1_{\mathbb{R} \times \{d_j\} \times \mathbb{R}} \), and

\[ \tilde{\varphi}_j(h_j, G_S, P^1_S, \ldots, P^J_S) \]

\[ = \sum_{h_1 \in \mathbb{H}(P^1_S)} \cdots \sum_{h_{j-1} \in \mathbb{H}(P^{j-1}_S)} \sum_{h_{j+1} \in \mathbb{H}(P^{j+1}_S)} \cdots \sum_{h_J \in \mathbb{H}(P^J_S)} \min_{g \in \mathcal{G}_S} \sum_{\xi=1}^J \psi(h, f, g). \]

For every \( Z(k, k') \in \mathcal{Z}_M \), let \( \mathcal{G}(Z(k, k')) = \{(1_{\mathbb{R} \times \{z_k\}}, 1_{\mathbb{R} \times \{z_{k'}\}})\} \). The conditions in (C.12)–(C.14) imply that \( \psi_l(\mathcal{G}(Z(k, k'))) \leq 0 \) for all \( l \in \{1, 2, 3\} \). Thus, we define \( \mathcal{Z}_2 \) by

\[ \mathcal{Z}_2 = \{ Z(k, k') \in \mathcal{Z} : \psi_l(\mathcal{G}(Z(k, k'))) \leq 0, l \in \{1, 2, 3\} \}. \]

Let \( \hat{\psi} : \mathcal{H}_B \times \mathcal{H}_D \times G_Z \rightarrow \mathbb{R} \) be the sample analog of \( \psi \) such that

\[ \hat{\psi}(h, f, g) = \frac{\hat{P}(h \cdot f \cdot g)}{P(g)} \]

for every \((h, f, g) \in \mathcal{H}_B \times \mathcal{H}_D \times G_Z\). Let \( \hat{\psi}_l \) be the sample analog of \( \psi_l \) for \( l \in \{1, 2, 3\} \), which
replaces ψ in ψ̂ by ˆψ. We define the estimator ˆZ2 for Z2 by

\[
\hat{Z}_2 = \left\{ Z_{(k,k')} \in \mathcal{Z} : \sqrt{T_n} \hat{\psi}_l(G(Z_{(k,k')})) \leq t_n, l \in \{1, 2, 3\} \right\},
\]

where \( t_n \to \infty \) and \( t_n/\sqrt{T_n} \to 0 \) as \( n \to \infty \).

To establish consistency of ˆZ2, we state and prove some auxiliary lemmas.

**Lemma C.3** The function space \( \mathcal{H}_B \) is a VC class with VC index \( V(\mathcal{H}_B) = 3 \).

**Proof of Lemma C.3.** The proof closely follows the strategy of the proof of Lemma C.2 of Sun (2021). ■

We define

\[
\mathcal{V} = \{ h \cdot f \cdot g : h \in \bar{\mathcal{H}}_B, f \in \mathcal{H}_D, g \in \mathcal{G}_Z \} \quad \text{and} \quad \hat{\mathcal{V}} = \mathcal{V} \cup \mathcal{G}_Z.
\]

**Lemma C.4** The function space \( \hat{\mathcal{V}} \) defined in (C.16) is Donsker and pre-Gaussian uniformly in \( Q \in \mathcal{P} \), and \( \hat{\mathcal{V}} \) is Glivenko–Cantelli uniformly in \( Q \in \mathcal{P} \).

**Proof of Lemma C.4.** The proof closely follows the strategies of the proofs of Lemmas C.5 and C.6 of Sun (2021). ■

The following proposition establishes consistency of ˆZ2.

**Proposition C.2** Under Assumptions A.2 and A.3, \( P(\hat{Z}_2 = Z_2) \to 1. \)

**Proof of Proposition C.2.** Let \( C_2 \) be the set of all \( G(Z_{(k,k')}) \) with \( Z_{(k,k')} \in \mathcal{Z}_2 \) and \( \hat{C}_2 \) be the set of all \( G(Z_{(k,k')}) \) with \( Z_{(k,k')} \in \hat{Z}_2 \). First, we have that

\[
P \left( C_2 \setminus \hat{C}_2 \neq \emptyset \right) \leq P \left( \max_{G_S \in C_2 \setminus \hat{C}_2} \sqrt{T_n} \left( \hat{\psi}_1(G_S) - \psi_1(G_S) \right) > t_n \right) + P \left( \max_{G_S \in C_2 \setminus \hat{C}_2} \sqrt{T_n} \left( \hat{\psi}_2(G_S) - \psi_2(G_S) \right) > t_n \right) + P \left( \max_{G_S \in C_2 \setminus \hat{C}_2} \sqrt{T_n} \left( \hat{\psi}_3(G_S) - \psi_3(G_S) \right) > t_n \right).
\]
By Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996),

\[
\max_{G_S \in C_2} \sqrt{T_n} \left| \max_{P_k \in \mathcal{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in H(P_k)} \max_{g \in G_S} (\hat{\psi}(h, f, g) - \max_{P_k \in \mathcal{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in H(P_k)} \max_{g \in G_S} \psi(h, f, g)) \right|
\]

\[
\leq \max_{G_S \in C_2} \max_{P_k \in \mathcal{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in H(P_k)} \max_{g \in G_S} \sqrt{T_n} \left| \hat{\psi}(h, f, g) - \psi(h, f, g) \right| \xrightarrow{n \to \infty} G_1
\]

for some random element $G_1$. Then it follows that

\[
P\left( \max_{G_S \in C_2 \setminus  \hat{C}_2} \left\{ \hat{\psi}_1(G_S) - \psi_1(G_S) \right\} > t_n \right) \leq P\left( \max_{G_S \in \hat{C}_2} \sqrt{T_n} \left| \hat{\psi}_1(G_S) - \psi_1(G_S) \right| > t_n \right)
\]

\[
\to 0.
\]

Similarly, we have that

\[
P\left( \max_{G_S \in C_2 \setminus  \hat{C}_2} \left\{ \hat{\psi}_2(G_S) - \psi_2(G_S) \right\} > t_n \right) \to 0
\]

and

\[
P\left( \max_{G_S \in C_2 \setminus  \hat{C}_2} \left\{ \hat{\psi}_3(G_S) - \psi_3(G_S) \right\} > t_n \right) \to 0.
\]

Thus, $P(C_2 \setminus \hat{C}_2 \neq \emptyset) \to 0$.

Next, because $\mathcal{Z}$ is finite, we have that for some $\delta > 0$,

\[
P\left( \hat{C}_2 \setminus C_2 \neq \emptyset \right) \leq P\left( \max_{G_S \in \hat{C}_2 \setminus C_2} \psi_1(G_S) > \delta, \max_{G_S \in \hat{C}_2 \setminus C_2} \sqrt{T_n} \hat{\psi}_1(G_S) \leq t_n \right)
\]

\[
+ P\left( \max_{G_S \in \hat{C}_2 \setminus C_2} \psi_2(G_S) > \delta, \max_{G_S \in \hat{C}_2 \setminus C_2} \sqrt{T_n} \hat{\psi}_2(G_S) \leq t_n \right)
\]

\[
+ P\left( \max_{G_S \in \hat{C}_2 \setminus C_2} \psi_3(G_S) > \delta, \max_{G_S \in \hat{C}_2 \setminus C_2} \sqrt{T_n} \hat{\psi}_3(G_S) \leq t_n \right).
\]
By Lemma C.4, \( \| \hat{\psi} - \psi \|_\infty \to 0 \) a.s. Then we have that

\[
\max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \left| \hat{\psi}_1 (G_S) - \psi_1 (G_S) \right| = \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \left| \max_{P_k \in \mathcal{D}, f \in \mathcal{D}_f} \sum_{h \in H(P_k)} \max_{g \in G_S} \hat{\psi}(h, f, g) - \max_{P_k \in \mathcal{D}, f \in \mathcal{D}_f} \sum_{h \in H(P_k)} \max_{g \in G_S} \psi(h, f, g) \right|
\leq \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \max_{P_k \in \mathcal{D}, f \in \mathcal{D}_f} \sum_{h \in H(P_k)} \max_{g \in G_S} \left| \hat{\psi}(h, f, g) - \psi(h, f, g) \right| \to 0
\]
a.s. Similarly, it follows that

\[
\max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \left| \hat{\psi}_2 (G_S) - \psi_2 (G_S) \right| \to 0 \text{ and } \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \left| \hat{\psi}_3 (G_S) - \psi_3 (G_S) \right| \to 0
\]
a.s. So we have that a.s., for all large \( n \),

\[
\max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \hat{\psi}_1 (G_S) \geq \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \psi_1 (G_S) - \frac{\delta}{2}, \quad \hat{\psi}_2 (G_S) \geq \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \psi_2 (G_S) - \frac{\delta}{2},
\]
and

\[
\max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \hat{\psi}_3 (G_S) \geq \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \psi_3 (G_S) - \frac{\delta}{2}.
\]

Thus, it follows that

\[
P\left( \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \psi_1 (G_S) > \delta, \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \sqrt{T_n} \hat{\psi}_1 (G_S) \leq t_n \right) \leq P\left( \frac{\delta}{2} \leq \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \hat{\psi}_1 (G_S) \leq \frac{t_n}{\sqrt{T_n}} \right)
\to 0.
\]

Similarly,

\[
P\left( \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \psi_2 (G_S) > \delta, \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \sqrt{T_n} \hat{\psi}_2 (G_S) \leq t_n \right) \to 0,
\]
and

\[
P\left( \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \psi_3 (G_S) > \delta, \max_{g_S \in \mathcal{C}_2 \setminus \mathcal{C}_2} \sqrt{T_n} \hat{\psi}_3 (G_S) \leq t_n \right) \to 0,
\]
which implies \( P(\hat{\mathcal{C}}_2 \setminus \mathcal{C}_2 \neq \emptyset) \to 0 \). Thus,

\[
P\left( \hat{\mathcal{C}}_2 \neq \mathcal{C}_2 \right) \leq P\left( \hat{\mathcal{C}}_2 \setminus \mathcal{C}_2 \neq \emptyset \right) + P\left( \mathcal{C}_2 \setminus \hat{\mathcal{C}}_2 \neq \emptyset \right) \to 0.
\]
C.4 Partially Valid Instruments for Multivalued Ordered Treatments

Here we extend the analysis in Section 2.3 to ordered treatments. Consider the following generalized version of Definition 2.3.

**Definition C.1** Suppose the instrument $Z$ is pairwise valid for the (multivalued ordered) treatment $D$ with the largest validity pair set $Z_M$. If there is a validity pair set

$$Z_M = \{(z_{k_1}, z_{k_2}), (z_{k_2}, z_{k_3}), \ldots, (z_{k_{M-1}}, z_{k_M})\}$$

for some $M > 0$, then the instrument $Z$ is called a **partially valid instrument** for the treatment $D$. The set $Z_M = \{z_{k_1}, \ldots, z_{k_M}\}$ is called a **validity value set** of $Z$.

Suppose that we have access to a consistent estimator $\hat{Z}_0$ of the validity value set $Z_M$, that is, $P(\hat{Z}_0 = Z_M) \rightarrow 1$. Then we can use $\hat{Z}_0$ to construct a VSIV estimator, $\hat{\theta}_1$, for a weighted average of ACRs based on model (2.8), where $D$ is now a multivalued ordered treatment. The following theorem presents the asymptotic properties of the VSIV estimator, generalizing Theorem 2.3.

**Theorem C.1** Suppose that the instrument $Z$ is partially valid for the treatment $D$ as defined in Definition C.1 with a validity value set $Z_M = \{z_{k_1}, \ldots, z_{k_M}\}$, and that the estimator $\hat{Z}_0$ for $Z_M$ satisfies $P(\hat{Z}_0 = Z_M) \rightarrow 1$. Under Assumptions A.2 and A.3, it follows that $\hat{\theta}_1 \overset{p}{\rightarrow} \theta_1$, where

$$\theta_1 = \frac{E[g(Z_i)Y_i|Z_i \in Z_M] - E[Y_i|Z_i \in Z_M]E[g(Z_i)|Z_i \in Z_M]}{E[g(Z_i)D_i|Z_i \in Z_M] - E[D_i|Z_i \in Z_M]E[g(Z_i)|Z_i \in Z_M]}.$$

Also, $\sqrt{n}(\hat{\theta}_1 - \theta_1) \overset{d}{\rightarrow} N(0, \Sigma_1)$, where $\Sigma_1$ is provided in (C.17). In addition, the quantity $\theta_1$ can be interpreted as the weighted average of $\{\beta_{k_2,k_1}, \ldots, \beta_{k_M,k_{M-1}}\}$ defined in (A.1). Specifically, $\theta_1 = \sum_{m=1}^{M-1} \mu_m \beta_{k_{m+1},k_m}$ with

$$\mu_m = \frac{\left[p(z_{k_{m+1}}) - p(z_{k_m})\right] \sum_{i=m}^{M-1} P(Z_i = z_{k_{i+1}}|Z_i \in Z_M) \{g(z_{k_{i+1}}) - E[g(Z_i)|Z_i \in Z_M]\}}{\sum_{i=1}^{M} P(Z_i = z_k|Z_i \in Z_M) p(z_k) \{g(z_k) - E[g(Z_i)|Z_i \in Z_M]\}},$$

$$p(z_k) = E[D_i|Z_i = z_k],$$

and $\sum_{m=1}^{M-1} \mu_m = 1$. 

30
Proof of Theorem C.1. By the formula of the VSIV estimator in (2.11),

\[ \hat{\theta}_1 = \frac{n \sum_{i=1}^{n} g(Z_i) Y_i 1 \{ Z_i \in \tilde{Z_0} \} - \bar{Y}_{\tilde{Z_0}} \frac{1}{n} \sum_{i=1}^{n} g(Z_i) 1 \{ Z_i \in \tilde{Z_0} \} \}
\]

where

\[ \bar{Y}_{\tilde{Z_0}} = \frac{1}{n} \sum_{i=1}^{n} Y_i 1 \{ Z_i \in \tilde{Z_0} \} \text{ and } D_{\tilde{Z_0}} = \frac{1}{n} \sum_{i=1}^{n} D_i 1 \{ Z_i \in \tilde{Z_0} \}. \]

We first have

\[ \frac{1}{n} \sum_{i=1}^{n} g(Z_i) Y_i 1 \{ Z_i \in \tilde{Z_0} \} = \frac{1}{n} \sum_{i=1}^{n} g(Z_i) Y_i 1 \{ Z_i \in Z_M \} + \left[ \frac{1}{n} \sum_{i=1}^{n} g(Z_i) Y_i \left\{ 1 \left\{ Z_i \in \tilde{Z_0} \right\} - 1 \{ Z_i \in Z_M \} \right\} \right] \]

with

\[ \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i) Y_i \left\{ 1 \left\{ Z_i \in \tilde{Z_0} \right\} - 1 \{ Z_i \in Z_M \} \right\} \right| \leq \frac{1}{n} \sum_{i=1}^{n} |g(Z_i) Y_i| 1 \{ \tilde{Z_0} \neq Z_M \}. \]

Since \( n^{-1} \sum_{i=1}^{n} |g(Z_i) Y_i| \xrightarrow{p} E \|g(Z_i) Y_i\| \) and for every small \( \varepsilon > 0 \),

\[ \mathbb{P} \left( \left\{ \tilde{Z_0} \neq Z_M \right\} > \varepsilon \right) = \mathbb{P} \left( \tilde{Z_0} \neq Z_M \right) \rightarrow 0, \]

we have that

\[ \frac{1}{n} \sum_{i=1}^{n} g(Z_i) Y_i 1 \{ Z_i \in \tilde{Z_0} \} = \frac{1}{n} \sum_{i=1}^{n} g(Z_i) Y_i 1 \{ Z_i \in Z_M \} + o_p(1) \xrightarrow{p} E \left[ g(Z_i) Y_i 1 \{ Z_i \in Z_M \} \right]. \]

Recall that \( n_z = \sum_{i=1}^{n} 1 \{ Z_i \in \tilde{Z_0} \} \). Then we can show that \( n_z/n \xrightarrow{p} \mathbb{P}(Z_i \in Z_M) \) as \( n \rightarrow \infty \). Similarly, we have that \( \bar{Y}_{\tilde{Z_0}} \xrightarrow{p} E[Y_i 1 \{ Z_i \in Z_M \}], D_{\tilde{Z_0}} \xrightarrow{p} E[D_i 1 \{ Z_i \in Z_M \}], \)
\( n^{-1} \sum_{i=1}^{n} g(Z_i) 1 \{ Z_i \in \tilde{Z_0} \} \xrightarrow{p} E[g(Z_i) 1 \{ Z_i \in Z_M \}], \) and \( n^{-1} \sum_{i=1}^{n} g(Z_i) D_i 1 \{ Z_i \in \tilde{Z_0} \} \xrightarrow{p} E[g(Z_i) D_i 1 \{ Z_i \in Z_M \}] \). Thus, it follows that

\[ \hat{\theta}_1 \xrightarrow{p} \frac{E[g(Z_i) Y_i 1 \{ Z_i \in Z_M \}] - E[Y_i 1 \{ Z_i \in Z_M \}] \cdot E[g(Z_i) 1 \{ Z_i \in Z_M \}] - E[D_i 1 \{ Z_i \in Z_M \}] \cdot E[g(Z_i) 1 \{ Z_i \in Z_M \}]}{E[g(Z_i) Y_i 1 \{ Z_i \in Z_M \}] - E[g(Z_i) 1 \{ Z_i \in Z_M \}]} = \theta_1. \]

Next, we derive the asymptotic distribution of \( \sqrt{n}(\hat{\theta}_1 - \theta_1) \). Define a function \( f: \mathbb{R}^6 \rightarrow \mathbb{R} \)
By

\[ f(x) = \frac{x_1/x_6 - x_2x_3/x_6^2}{x_4/x_6 - x_5x_3/x_6^2} = \frac{x_1x_6 - x_2x_3}{x_4x_6 - x_5x_3} \]

for every \( x \in \mathbb{R}^6 \) with \( x = (x_1, x_2, x_3, x_4, x_5, x_6)^T \), \( x_6 \neq 0 \), and \( x_4x_6 - x_5x_3 \neq 0 \). We can obtain the gradient of \( f \), denoted \( f' \), by

\[
\begin{align*}
    f'_1(x) &= \frac{1/x_6}{x_4/x_6 - x_5x_3/x_6^2}, \quad f'_2(x) = \frac{-x_3/x_6^2}{x_4/x_6 - x_5x_3/x_6^2}, \quad f'_3(x) = \frac{-x_2x_4x_6 + x_5x_1x_6}{(x_4x_6 - x_5x_3)^2}, \\
    f'_4(x) &= \frac{(x_1x_6 - x_2x_3)x_6}{(x_4x_6 - x_5x_3)^2}, \quad f'_5(x) = \frac{x_3(x_1x_6 - x_2x_3)}{(x_4x_6 - x_5x_3)^2}, \quad f'_6(x) = \frac{-x_1x_5x_3 + x_2x_3x_4}{(x_4x_6 - x_5x_3)^2}.
\end{align*}
\]

Then we can rewrite

\[
\sqrt{n}(\hat{\theta}_1 - \theta_1) = \sqrt{n} \left\{ f(\bar{W}_n) - f(W) \right\},
\]

where

\[
\bar{W}_n = \left( \begin{array}{c}
    n \sum_{i=1}^n g(Z_i)Y_{i1} \left\{ Z_i \in \widehat{Z}_0 \right\} \\
    \tilde{Y}_{\widehat{Z}_0} \\
    \frac{1}{n} \sum_{i=1}^n g(Z_i) 1 \left\{ Z_i \in \widehat{Z}_0 \right\} \\
    \frac{1}{n} \sum_{i=1}^n g(Z_i)D_{i1} \left\{ Z_i \in \widehat{Z}_0 \right\} \\
    \frac{1}{n} \sum_{i=1}^n 1 \left\{ Z_i \in \widehat{Z}_0 \right\}
\end{array} \right)
\quad \text{and} \quad
W = \left( \begin{array}{c}
    E \left[ g(Z_i) Y_{i1} \left\{ Z_i \in Z_M \right\} \right] \\
    E \left[ Y_{i1} \left\{ Z_i \in Z_M \right\} \right] \\
    E \left[ g(Z_i) 1 \left\{ Z_i \in Z_M \right\} \right] \\
    E \left[ g(Z_i)D_{i1} \left\{ Z_i \in Z_M \right\} \right] \\
    E \left[ D_{i1} \left\{ Z_i \in Z_M \right\} \right] \\
    E \left[ 1 \left\{ Z_i \in Z_M \right\} \right]
\end{array} \right).
\]

For every small \( \varepsilon > 0 \), we have \( \mathbb{P}(\sqrt{n} \left\{ \widehat{Z}_0 \neq Z_M \right\} > \varepsilon) = \mathbb{P}(\widehat{Z}_0 \neq Z_M) \to 0 \). With

\[
\begin{align*}
    \sqrt{n} \left| \frac{1}{n} \sum_{i=1}^n g(Z_i) Y_{i1} \left\{ Z_i \in \widehat{Z}_0 \right\} - \frac{1}{n} \sum_{i=1}^n g(Z_i) Y_{i1} \left\{ Z_i \in Z_M \right\} \right| \\
    = \sqrt{n} \left| \frac{1}{n} \sum_{i=1}^n g(Z_i) Y_{i1} \left[ 1 \left\{ Z_i \in \widehat{Z}_0 \right\} - 1 \left\{ Z_i \in Z_M \right\} \right] \right| \\
    \leq \frac{1}{n} \sum_{i=1}^n |g(Z_i) Y_{i1}| \left( \sqrt{n} \left\{ \widehat{Z}_0 \neq Z_M \right\} \right) = o_P(1).
\end{align*}
\]
Similarly, we have that

$$
\sqrt{n} \left( \hat{W}_n - W \right) = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} \left( \begin{array}{c}
g(Z_i) Y_i 1 \{ Z_i \in Z_M \} - E[g(Z_i) Y_i 1 \{ Z_i \in Z_M \}] \\
g(Z_i) Y_i 1 \{ Z_i \in Z_M \} - E[Y_i 1 \{ Z_i \in Z_M \}] \\
g(Z_i) 1 \{ Z_i \in Z_M \} - E[g(Z_i) 1 \{ Z_i \in Z_M \}] \\
g(Z_i) D_i 1 \{ Z_i \in Z_M \} - E[g(Z_i) D_i 1 \{ Z_i \in Z_M \}] \\
D_i 1 \{ Z_i \in Z_M \} - E[D_i 1 \{ Z_i \in Z_M \}] \\
1 \{ Z_i \in Z_M \} - E[1 \{ Z_i \in Z_M \}] \\
\end{array} \right) + o_p(1) \xrightarrow{d} N(0, \Sigma).
$$

where $\Sigma = E[VV^T]$ and

$$
V = \left( \begin{array}{c}
g(Z_i) Y_i 1 \{ Z_i \in Z_M \} - E[g(Z_i) Y_i 1 \{ Z_i \in Z_M \}] \\
Y_i 1 \{ Z_i \in Z_M \} - E[Y_i 1 \{ Z_i \in Z_M \}] \\
g(Z_i) 1 \{ Z_i \in Z_M \} - E[g(Z_i) 1 \{ Z_i \in Z_M \}] \\
g(Z_i) D_i 1 \{ Z_i \in Z_M \} - E[g(Z_i) D_i 1 \{ Z_i \in Z_M \}] \\
D_i 1 \{ Z_i \in Z_M \} - E[D_i 1 \{ Z_i \in Z_M \}] \\
1 \{ Z_i \in Z_M \} - E[1 \{ Z_i \in Z_M \}] \\
\end{array} \right).
$$

By multivariate delta method, we have that

$$
\sqrt{n} (\hat{\theta}_1 - \theta_1) = \sqrt{n} \left\{ f \left( \hat{W}_n \right) - f (W) \right\} \xrightarrow{d} f'(W)^T \cdot N(0, \Sigma).
$$

(C.17)

Now we follow the strategy of Imbens and Angrist (1994) and have that

$$
\frac{E[g(Z_i) Y_i 1 \{ Z_i \in Z_M \}]}{P(Z_i \in Z_M)} - \frac{E[Y_i 1 \{ Z_i \in Z_M \}]}{P(Z_i \in Z_M)} - \frac{E[g(Z_i) 1 \{ Z_i \in Z_M \}]}{P(Z_i \in Z_M)} = \sum_{k=1}^{K} \frac{P(Z_i = z_k) E[Y_i 1 \{ Z_i \in Z_M \} | Z_i = z_k]}{P(Z_i \in Z_M)} \left\{ g(z_k) 1 \{ z_k \in Z_M \} - \frac{E[g(Z_i) 1 \{ Z_i \in Z_M \}]}{P(Z_i \in Z_M)} \right\}
$$

$$
= \sum_{m=1}^{M} \frac{P(Z_i = z_{km} | Z_i \in Z_M) E[Y_i | Z_i = z_{km}]}{P(Z_i \in Z_M)} \left\{ g(z_{km}) - E[g(Z_i) | Z_i \in Z_M] \right\}.
$$

33
Then we write
\[
\sum_{m=1}^{M} \mathbb{P}(Z_i = z_{km} | Z_i \in \mathcal{Z}_M) E[Y_i | Z_i = z_{km}] \{g(z_{km}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\}
\]
\[
= \sum_{m=1}^{M-1} \mathbb{P}(Z_i = z_{km+1} | Z_i \in \mathcal{Z}_M) E[Y_i | Z_i = z_{km+1}] \{g(z_{km+1}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\}
+ \mathbb{P}(Z_i = z_{k1} | Z_i \in \mathcal{Z}_M) E[Y_i | Z_i = z_{k1}] \{g(z_{k1}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\}.
\]  
(C.18)

By (A.1), we have
\[
E[Y_i | Z_i = z_{km+1}] = \beta_{km+1,km} (E[D_i | Z_i = z_{km+1}] - E[D_i | Z_i = z_{km}]) + E[Y_i | Z_i = z_{km}]
\]
\[
= \sum_{l=1}^{m} \beta_{l+1,kl} \left( E[D_i | Z_i = z_{kl+1}] - E[D_i | Z_i = z_{kl}] \right) + E[Y_i | Z_i = z_{k1}],
\]
and thus it follows that
\[
\sum_{m=1}^{M-1} \mathbb{P}(Z_i = z_{km+1} | Z_i \in \mathcal{Z}_M) E[Y_i | Z_i = z_{km+1}] \{g(z_{km+1}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\}
\]
\[
= \sum_{m=1}^{M-1} \mathbb{P}(Z_i = z_{km+1} | Z_i \in \mathcal{Z}_M) \left\{ \sum_{l=1}^{m} \beta_{l+1,kl} \left[ p(z_{kl+1}) - p(z_{kl}) \right] \right\}
\]
\[
\cdot \left\{ g(z_{km+1}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \right\}
\]
\[
+ \sum_{m=1}^{M-1} \mathbb{P}(Z_i = z_{km+1} | Z_i \in \mathcal{Z}_M) E[Y_i | Z_i = z_{k1}] \{g(z_{km+1}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\}.
\]

By (C.18), this implies that
\[
\sum_{m=1}^{M} \mathbb{P}(Z_i = z_{km} | Z_i \in \mathcal{Z}_M) E[Y_i | Z_i = z_{km}] \{g(z_{km}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\}
\]
\[
= \sum_{m=1}^{M} \mathbb{P}(Z_i = z_{km+1} | Z_i \in \mathcal{Z}_M) \left\{ \sum_{l=1}^{m} \beta_{l+1,kl} \left[ p(z_{kl+1}) - p(z_{kl}) \right] \right\}
\]
\[
\cdot \left\{ g(z_{km+1}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \right\},
\]  
(C.19)

where we use \( \sum_{m=1}^{M} \mathbb{P}(Z_i = z_{km} | Z_i \in \mathcal{Z}_M) \{g(z_{km}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\} = 0 \). By rewrit-
ing (C.19), we obtain

\[
\sum_{m=1}^{M-1} \mathbb{P}(Z_i = z_{m+1} | Z_i \in \mathcal{Z}_M) \left\{ \sum_{l=1}^{m} \beta_{k_{l+1},k_l} \left[ p(z_{k_{l+1}}) - p(z_{k_l}) \right] \right\} \tilde{g}(z_{m+1})
\]

\[
= \mathbb{P}(Z_i = z_{k_2} | Z_i \in \mathcal{Z}_M) \left\{ \beta_{k_2,k_1} \left[ p(z_{k_2}) - p(z_{k_1}) \right] \tilde{g}(z_{k_2}) \right\} + \cdots
\]

\[
+ \mathbb{P}(Z_i = z_{k_M} | Z_i \in \mathcal{Z}_M) \left\{ \sum_{l=1}^{M-1} \beta_{k_{l+1},k_l} \left[ p(z_{k_{l+1}}) - p(z_{k_l}) \right] \right\} \tilde{g}(z_{k_M})
\]

\[
= \sum_{m=1}^{M-1} \left\{ \beta_{k_{m+1},k_m} \left[ p(z_{k_{m+1}}) - p(z_{k_m}) \right] \sum_{l=m}^{M-1} \mathbb{P}(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M) \right\} \tilde{g}(z_{k_{m+1}}),
\]

where \( \tilde{g}(z) = g(z) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \) for all \( z \). Similarly, we have

\[
\frac{E[g(Z_i) D_{i1} \{ Z_i \in \mathcal{Z}_M \}]}{\mathbb{P}(Z_i \in \mathcal{Z}_M)} - \frac{E[D_{i1} \{ Z_i \in \mathcal{Z}_M \} E[g(Z_i) 1 \{ Z_i \in \mathcal{Z}_M \}]}{\mathbb{P}(Z_i \in \mathcal{Z}_M)}
\]

\[
= \sum_{m=1}^{M-1} \mathbb{P}(Z_i = z_{k_m} | Z_i \in \mathcal{Z}_M) p(z_{k_m}) \{ g(z_{k_m}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \}.
\]

Thus, we have \( \theta_1 = \sum_{m=1}^{M-1} \mu_m \beta_{k_{m+1},k_m} \), with

\[
\mu_m = \frac{\left[ p(z_{k_{m+1}}) - p(z_{k_m}) \right] \sum_{l=m}^{M-1} \mathbb{P}(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M) \{ g(z_{k_{l+1}}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \}}{\sum_{l=1}^{M-1} \mathbb{P}(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M) p(z_{k_l}) \{ g(z_{k_l}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \}}.
\]

Now we show that \( \sum_{m=1}^{M-1} \mu_m = 1 \). First, we have that

\[
\sum_{m=1}^{M-1} \left[p(z_{k_{m+1}}) - p(z_{k_m})\right] \sum_{l=m}^{M-1} \mathbb{P}(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M) \{ g(z_{k_{l+1}}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \}
\]

\[
= \left[p(z_{k_2}) - p(z_{k_1})\right] \sum_{l=1}^{M-1} \mathbb{P}(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M) \{ g(z_{k_{l+1}}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \} + \cdots
\]

\[
+ \left[p(z_{k_M}) - p(z_{k_{M-1}})\right] \mathbb{P}(Z_i = z_{k_M} | Z_i \in \mathcal{Z}_M) \{ g(z_{k_M}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \}
\]

\[
= \sum_{l=2}^{M} \mathbb{P}(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M) p(z_{k_l}) \{ g(z_{k_l}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \}
\]

\[
- p(z_{k_1}) \sum_{l=2}^{M} \mathbb{P}(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M) \{ g(z_{k_l}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \}
\]

\[
= \sum_{l=1}^{M} \mathbb{P}(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M) p(z_{k_l}) \{ g(z_{k_l}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M] \},
\]
where we use the equality that \( \sum_{i=1}^{M} \mathbb{P}(Z_i = z_{ki} | Z_i \in Z_M) \{ g(z_{ki}) - E[g(Z_i) | Z_i \in Z_M] \} = 0 \). This implies that \( \sum_{m=1}^{M-1} \mu_m = 1 \). □

D Proofs and Supplementary Results for Appendix A.2

D.1 Proofs for Appendix A.2

Proof of Lemma A.2. (i) \( \iff \) (ii). We closely follow the proof for “(i) \( \iff \) (ii)” in Theorem T-3 of Heckman and Pinto (2018). By Lemma L-5 of Heckman and Pinto (2018), if \( B_{d(k,k')} \) is lonesum, then no \( 2 \times 2 \) sub-matrix of \( B_{d(k,k')} \) takes the form

\[
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}
\] or \[
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}.
\]

Since \( B_{d(k,k')} = 1\{K_{k,k'} R = d\} \), (i) \( \implies \) (ii). Suppose (ii) holds. Then no \( 2 \times 2 \) sub-matrix of \( B_{d(k,k')} \) takes the form in (D.1) by the definition of \( B_{d(k,k')} \). By Lemmas L-6 and L-8 of Heckman and Pinto (2018), (i) holds.

(i) \( \implies \) (iii) \( \implies \) (ii). If for every \( d \in D \), \( B_{d(k,k')} \) is lonesum, by Lemma L-9 of Heckman and Pinto (2018),

\[
B_{d(k,k')} (1, l) \leq B_{d(k,k')} (2, l) \quad \text{for all } l, \quad \text{or} \quad B_{d(k,k')} (1, l) \geq B_{d(k,k')} (2, l) \quad \text{for all } l.
\]

Because the value of \( (D_{z_k}, D_{z_{k'}}) \) must be equal to \( (K_{k,k'} R (1, l), K_{k,k'} R (2, l)) \) for some \( l \), it follows that

\[
1 \{D_{z_k} = d\} \leq 1 \{D_{z_{k'}} = d\} \quad \text{or} \quad 1 \{D_{z_k} = d\} \geq 1 \{D_{z_{k'}} = d\}.
\]

Thus the following sub-matrices will not occur in \( K_{k,k'} R \):

\[
\begin{pmatrix}
d & d' \\
d'' & d
\end{pmatrix} \quad \text{or} \quad \begin{pmatrix}
d' & d \\
d & d''
\end{pmatrix}.
\]

Proof of Theorem A.3. The proof follows a strategy similar to that of the proof of Theorem T-6 in Heckman and Pinto (2018). We first write

\[
\mathbb{P}(M_{(k,k')} S \in \Sigma_{d(k,k')} (t)) = b_{d(k,k')} (t) P_{S(k,k')}.
\]
Also, since

\[
E \left[ \kappa (Y_d(z_k, z_{k'})) 1 \{ \mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')} (t) \} \right] \\
= E \left[ \kappa (Y_d(z_k, z_{k'})) 1 \{ \mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')} (t) \} \right] 1 \{ \mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')} (t) \} \\
= E \left[ \kappa (Y_d(z_k, z_{k'})) |\mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')} (t) \} \cdot \mathbb{P} (\mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')} (t)) \right]
\]

and

\[
E \left[ \kappa (Y_d(z_k, z_{k'})) 1 \{ \mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')} (t) \} \right] \\
= E \left[ \kappa (Y_d(z_k, z_{k'})) \sum_{t=1}^{L(k,k')} 1 \{ \mathcal{M}_{(k,k')} S = s_t \} \{ s_t \in \Sigma_{d(k,k')} (t) \} \right] = b_{d(k,k')} (t) Q_{S(k,k')} (d),
\]

we have that

\[
E \left[ \kappa (Y_d(z_k, z_{k'})) |\mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')} (t) \} \right] = \frac{b_{d(k,k')} (t) Q_{S(k,k')} (d)}{b_{d(k,k')} (t) P_{S(k,k')} (d)}.
\]

(D.3)

Now we suppose \((z_k, z_{k'}) \in \mathcal{X}_{\mathcal{M}}\). By definition, we have \(P_{Z(k,k')} (d) = B_{d(k,k')} P_{S(k,k')} \) and \(Q_{Z(k,k')} (d) = B_{d(k,k')} Q_{S(k,k')} (d)\), so by Lemma L-2 of Heckman and Pinto (2018),

\[
b_{d(k,k')} (t) P_{S(k,k')} = b_{d(k,k')} (t) \left[ B_{d(k,k')}^+ P_{Z(k,k')} (d) + (I - B_{d(k,k')}^+ B_{d(k,k')}) \lambda_P \right]
\] and

\[
b_{d(k,k')} (t) Q_{S(k,k')} (d) = b_{d(k,k')} (t) \left[ B_{d(k,k')}^+ Q_{Z(k,k')} (d) + (I - B_{d(k,k')}^+ B_{d(k,k')}) \lambda_Q \right],
\]

where \(\lambda_P\) and \(\lambda_Q\) are arbitrary real-valued vectors.

We next show that \(b_{d(k,k')} (t) [I - B_{d(k,k')}^+ B_{d(k,k')}] = 0\). First, by the proof of Lemma L-16 of Heckman and Pinto (2018) and Lemma A.2 in this paper, if \(B_{d(k,k')} (\cdot, l)\) and \(B_{d(k,k')} (\cdot, l')\) have the same sum, then these two vectors are identical. Thus, by the definition of the set \(\Sigma_{d(k,k')} (t)\), for all \(s_l, s_{l'} \in \Sigma_{d(k,k')} (t)\), \(B_{d(k,k')} (\cdot, l) = B_{d(k,k')} (\cdot, l')\). Let \(C_{d(k,k')} (t) = B_{d(k,k')} (\cdot, l)\) with \(l\) satisfying that \(s_l \in \Sigma_{d(k,k')} (t)\), where \(s_l\) is the \(l\)th column of \(\mathcal{K}(k,k') R\). Let \(C_{d(k,k')} = (C_{d(k,k')}(1), C_{d(k,k')}(2))\) be the matrix that consists of all unique nonzero vectors in \(B_{d(k,k')}\).\(^{15}\) Then clearly \(C_{d(k,k')}\) has full column rank and \(C_{d(k,k')}^T C_{d(k,k')}\) has full rank. Thus, \((C_{d(k,k')}^T C_{d(k,k')}^{-1})\) exists. Let \(D_{d(k,k')} = (b_{d(k,k')}(1), b_{d(k,k')}(2))^T\). Since by the definition of \(b_{d(k,k')} (t)\), \(b_{d(k,k')} (t) \cdot b_{d(k,k')} (t')^T = 0\) for \(t \neq t'\), \(D_{d(k,k')}\) has full row rank and \((D_{d(k,k')} D_{d(k,k')}^T)^{-1}\) exists. We then decompose \(B_{d(k,k')} = C_{d(k,k')} \cdot D_{d(k,k')}\).\(^{16}\)

Now by similar proof of Lemma L-17 of Heckman and Pinto (2018), we can show that

\(^{15}\)Without loss of generality, we assume that both \(C_{d(k,k')}(1)\) and \(C_{d(k,k')}(2)\) exist.

\(^{16}\)See Remark A.3 of Heckman and Pinto (2018).
the Moore–Penrose pseudo inverse of \( B_{d(k,k')} \) is
\[
B^+_{d(k,k')} = \left(D^T_{d(k,k')}D_{d(k,k')}ight)\left(D^T_{d(k,k')}D_{d(k,k')}ight)^{-1}\left(C^T_{d(k,k')}C_{d(k,k')}ight)^{-1}C^T_{d(k,k')}.
\]
For \( t \in \{1,2\} \), we can write \( b_{d(k,k')}(t) = e_t D_{d(k,k')} \), where \( e_t \) is a row vector in which the \( t \)th element is 1 and the other element is 0. Then we have that
\[
b_{d(k,k')}(t)\left[I - B^+_{d(k,k')}B_{d(k,k')}\right] = b_{d(k,k')}(t) - b_{d(k,k')}(t) B^+_{d(k,k')}B_{d(k,k')} = b_{d(k,k')}(t) - e_t D_{d(k,k')}D^T_{d(k,k')}\left(D^T_{d(k,k')}\left(D_{d(k,k')}\right)^{-1}\left(C^T_{d(k,k')}C_{d(k,k')}ight)^{-1}C^T_{d(k,k')}C_{d(k,k')}:D_{d(k,k')}\right) = 0.
\]
This implies that \( b_{d(k,k')}(t) P_{S(k,k')} \) and \( b_{d(k,k')}(t) Q_{S(k,k')} (d) \) can be identified by
\[
b_{d(k,k')}(t) P_{S(k,k')} = b_{d(k,k')}(t) B^+_{d(k,k')}P_{Z(k,k')} (d)
\]
and
\[
b_{d(k,k')}(t) Q_{S(k,k')} (d) = b_{d(k,k')}(t) B^+_{d(k,k')}Q_{Z(k,k')} (d).
\]
Thus, (D.2) and (D.3) show that
\[
\mathbb{P}\left(\mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')} (t)\right) = b_{d(k,k')}(t) B^+_{d(k,k')}P_{Z(k,k')} (d)
\]
and
\[
E\left[\kappa (Y_d(z_k, z_{k}')) | \mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')} (t)\right] = \frac{b_{d(k,k')}(t) B^+_{d(k,k')}Q_{Z(k,k')} (d)}{b_{d(k,k')}(t) B^T_{d(k,k')}P_{Z(k,k')} (d)}
\]
are identified. Thus, it can easily be shown that (A.5) holds by (1.1), and the quantities in (A.5) are identified. Define
\[
Z_{pi} = \{Z_i = z_1\}, \ldots, \{Z_i = z_K\},
\]
\[
P_{DZi} (d) = \{1 \{D_i = d, Z_i = z_1\}, \ldots, 1 \{D_i = d, Z_i = z_K\}\}^T \text{ for all } d,
\]
\[
Q_{YDZi} (d) = \kappa (Y_i) \{1 \{D_i = d, Z_i = z_1\}, \ldots, \kappa (Y_i) \{1 \{D_i = d, Z_i = z_K\}\}\}^T \text{ for all } d,
\]
and
\[
W_i = \left(Z_{pi}, P_{DZi} (d_1)^T, \ldots, P_{DZi} (d_J)^T, Q_{YDZi} (d_1)^T, \ldots, Q_{YDZi} (d_J)^T\right)^T.
\]
By multivariate central limit theorem, \( \sqrt{n}(\hat{W} - W) \overset{d}{\to} N (0, \Sigma_W) \), where
\[
\Sigma_W = E[(W_i - W)(W_i - W)^T], \quad (D.4)
\]
Also, we have that
\[
\frac{\sqrt{n}}{\left\langle W, 1(\mathcal{Z}_0) \right\rangle} = \sqrt{n} \left( \mathcal{W} - W, 1(\mathcal{Z}_0) \right) \xrightarrow{d} (N(0, \Sigma_W), 0).
\]

**Proof of Lemma A.3.** We first write
\[
E \left[ \left( Y_d(z_k, z_{k'}) - Y_d'(z_k, z_{k'}) \right) \cdot 1 \left\{ \mathcal{M}_{(k,k')} \in \Sigma_{d(k,k')}(t), (z_k, z_{k'}) \in \mathcal{Z}_M \right\} \right] \\
\quad \cdot 1 \left\{ \Sigma_{d(k,k')}(t) = \Sigma_{d'(k,k')}(t') \right\}
\]
\[
= E \left[ Y_d(z_k, z_{k'}) - Y_d'(z_k, z_{k'}) | \mathcal{M}_{(k,k')} \in \Sigma_{d(k,k')}(t), (z_k, z_{k'}) \in \mathcal{Z}_M, \Sigma_{d(k,k')(t)} = \Sigma_{d'(k,k')(t')} \right] \\
\quad \cdot \mathbb{P} \left( \mathcal{M}_{(k,k')} \in \Sigma_{d(k,k')}(t), (z_k, z_{k'}) \in \mathcal{Z}_M, \Sigma_{d(k,k')(t)} = \Sigma_{d'(k,k')(t')} \right).
\]

Thus, if follows that
\[
E \left[ Y_d(z_k, z_{k'}) - Y_d'(z_k, z_{k'}) | \mathcal{M}_{(k,k')} \in \Sigma_{d(k,k')(t)}, (z_k, z_{k'}) \in \mathcal{Z}_M \right] \\
\quad \cdot 1 \left\{ \Sigma_{d(k,k')(t)} = \Sigma_{d'(k,k')(t')} \right\} \cdot \mathbb{P} \left( \mathcal{M}_{(k,k')} \in \Sigma_{d(k,k')(t)}, (z_k, z_{k'}) \in \mathcal{Z}_M \right)
\]
\[
= E \left[ Y_d(z_k, z_{k'}) - Y_d'(z_k, z_{k'}) | \mathcal{M}_{(k,k')} \in \Sigma_{d(k,k')(t)}, (z_k, z_{k'}) \in \mathcal{Z}_M, \Sigma_{d(k,k')(t)} = \Sigma_{d'(k,k')(t')} \right] \\
\quad \cdot \mathbb{P} \left( \mathcal{M}_{(k,k')} \in \Sigma_{d(k,k')(t)}, (z_k, z_{k'}) \in \mathcal{Z}_M, \Sigma_{d(k,k')(t)} = \Sigma_{d'(k,k')(t')} \right).
\]

If \( \Sigma_{d(k,k')(t)} \neq \Sigma_{d'(k,k')(t')} \) or \( (z_k, z_{k'}) \notin \mathcal{Z}_M \), then the result holds by (1.1) with the mean effect \( \beta_{(k,k')}(d, d', t, t') = 0 \). If \( \Sigma_{d(k,k')(t)} = \Sigma_{d'(k,k')(t')} \) and \( (z_k, z_{k'}) \in \mathcal{Z}_M \),
\[
E \left[ Y_d(z_k, z_{k'}) - Y_d'(z_k, z_{k'}) | \mathcal{M}_{(k,k')} \in \Sigma_{d(k,k')(t)}, (z_k, z_{k'}) \in \mathcal{Z}_M, \Sigma_{d(k,k')(t)} = \Sigma_{d'(k,k')(t')} \right]
\]
\[
= E \left[ Y_d(z_k, z_{k'}) - Y_d'(z_k, z_{k'}) | \mathcal{M}_{(k,k')} \in \Sigma_{d(k,k')(t)} \right].
\]
Proof of Theorem A.4. The proof is similar to that of Theorem A.2. ■

D.2 Estimation of $\mathcal{Z}_0$

D.2.1 Definition and Estimation of $\mathcal{Z}_1$

Following Sun (2021), we provide the definitions of $\mathcal{Z}_1$ and its estimator. Suppose the instrument $Z$ is pairwise valid with $\mathcal{Z}_M = \{(z_{k1}, z_{k2}), \ldots, (z_{kN}, z_{kN})\}$. Fix $(z, z') \in \mathcal{Z}_M$. For every $d \in \mathcal{D}$, if $1 \{D_{z'} = d\} \leq 1 \{D_z = d\}$ a.s., we have that

$$\mathbb{P}(Y \in B, D = d | Z = z') = E \left[1 \{Y_d(z, z') \in B\} \times 1 \{D_{z'} = d\}\right]$$

$$\leq E \left[1 \{Y_d(z, z') \in B\} \times 1 \{D_z = d\}\right] = \mathbb{P}(Y \in B, D = d | Z = z) \quad (D.5)$$

for all Borel sets $B$. Denote $2^J$ $J$-dimensional different binary vectors by $v_1, \ldots, v_{2^J}$, where

$$v_1 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, v_2 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, v_{2^J} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Let $\mathcal{L} : \mathcal{D} \rightarrow \{1, \ldots, J\}$ map $d \in \mathcal{D}$ to $d$'s index in $\mathcal{D}$ so that if $d = d_j$, we have $\mathcal{L}(d) = j$. For every $q \in \{1, \ldots, 2^J\}$, define $f_q : \{d_1, \ldots, d_J\} \rightarrow \{1, -1\}$ by $f_q(d) = (-1)^{v_q(\mathcal{L}(d))}$. For every fixed $(z, z') \in \mathcal{Z}_M$, there is $q \in \{1, \ldots, 2^J\}$ such that

$$f_q(d) \cdot \left(\mathbb{P}(Y \in B, D = d | Z = z') - \mathbb{P}(Y \in B, D = d | Z = z)\right) \leq 0$$

for all $d \in \mathcal{D}$ and all closed intervals $B$. Then for all $q \in \{1, \ldots, 2^J\}$, define

$$H_q = \left\{f_q(d) \cdot 1_{B \times \{d\} \times \mathbb{R}} : B \text{ is a closed interval in } \mathbb{R}, d \in \mathcal{D}\right\} \quad \text{ and }$$

$$\bar{H}_q = \left\{f_q(d) \cdot 1_{B \times \{d\} \times \mathbb{R}} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}, d \in \mathcal{D}\right\}.$$

Furthermore, define the following function spaces

$$G = \left\{1_{\mathbb{R} \times \mathbb{R} \times \{z_j\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}} : j, k \in \{1, \ldots, K\}, j < k\right\}, H = \bigcup_{q=1}^{2^J} H_q, \text{ and } \bar{H} = \bigcup_{q=1}^{2^J} \bar{H}_q. \quad (D.6)$$

Let $P$, $\phi$, $\sigma^2$, $\bar{P}$, $\bar{\phi}$, and $\bar{\sigma}^2$ be defined in a way similar to that in Section 2 but for all $(h, g) \in \bar{H} \times G$. Also, we let $\Lambda(P) = \prod_{k=1}^{K} P(1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}})$ and $T_n = n \cdot \prod_{k=1}^{K} \bar{P}(1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}})$. By similar proof of Lemma 3.1 in Sun (2021), $\sigma^2$ and $\bar{\sigma}^2$ are uniformly bounded in $(h, g) \in \bar{H} \times G$. 40
The following lemma reformulates the testable restrictions in terms of \( \phi \).

**Lemma D.1** Suppose that the instrument \( Z \) is pairwise valid for the treatment \( D \) with the largest validity pair set \( \mathcal{D}^*_M = \{(z_{k_1}, z'_{k_1}), \ldots, (z_{k_M}, z'_{k_M})\} \). For every \( m \in \{1, \ldots, M\} \), we have that \( \min_{q \in \{1, \ldots, 2^j\}} \sup_{h \in H_q} \phi(h, g) = 0 \) with \( g = (1_{\mathbb{R} \times \mathbb{R} \times \{z_{k_m}\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z'_{k_m}\}}) \).

**Proof of Lemma D.1.** Since we can find \( a \in \mathbb{R} \) and \( d \in \mathcal{D} \) such that \( P\left(1_{\{a\} \times \{d\} \times \mathbb{R}}\right) = 0 \), then we have \( \sup_{h \in H_q} \phi(h, g) \geq 0 \) for every \( q \) and every \( g \in G \). So for every \( g \in G \), \( \min_{q \in \{1, \ldots, 2^j\}} \sup_{h \in H_q} \phi(h, g) \geq 0 \). Let \( h_{Bd} = 1_{B \times \{d\} \times \mathbb{R}} \) for every closed interval \( B \) and every \( d \in \mathcal{D} \). Fix \( m \in \{1, \ldots, M\} \). Under assumption, for every \( d \in \mathcal{D} \), we have

\[
\phi(h_{Bd}, g) = \frac{P(h_{Bd} \cdot g_2)}{P(g_2)} - \frac{P(h_{Bd} \cdot g_1)}{P(g_1)} \leq 0 \text{ for every closed interval } B,
\]

or

\[
\phi(-h_{Bd}, g) = -\frac{P(h_{Bd} \cdot g_2)}{P(g_2)} - \frac{P(h_{Bd} \cdot g_1)}{P(g_1)} \leq 0 \text{ for every closed interval } B,
\]

where \( g_1 = 1_{\mathbb{R} \times \mathbb{R} \times \{z_{k_m}\}}, g_2 = 1_{\mathbb{R} \times \mathbb{R} \times \{z'_{k_m}\}} \), and \( g = (g_1, g_2) \). This implies that there is \( H_q \) such that \( \sup_{h \in H_q} \phi(h, g) \leq 0 \). Thus, it follows that \( \min_{q \in \{1, \ldots, 2^j\}} \sup_{h \in H_q} \phi(h, g) = 0 \). \( \blacksquare \)

By Lemma D.1, we define

\[
G_1 = \left\{ g \in G : \min_{q \in \{1, \ldots, 2^j\}} \sup_{h \in H_q} \phi(h, g) = 0 \right\}
\]

and

\[
\widehat{G}_1 = \left\{ g \in G : \sqrt{T_n} \left| \min_{q \in \{1, \ldots, 2^j\}} \sup_{h \in H_q} \frac{\widehat{\phi}(h, g)}{\xi_0 \sigma(h, g)} \right| \leq \tau_n \right\}
\]

with \( \tau_n \to \infty \) and \( \tau_n/\sqrt{n} \to 0 \) as \( n \to \infty \), where \( \xi_0 \) is a small positive number. We define \( \mathcal{D}_1 \) as the collection of all \((z, z')\) that are associated with some \( g \in G_1 \):

\[
\mathcal{D}_1 = \{(z_k, z_{k'}) \in \mathcal{D} : g = (1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_{k'}\}}) \in G_1\}.
\]

We use \( \widehat{G}_1 \) to construct the estimator of \( \mathcal{D}_1 \), denoted by \( \widehat{\mathcal{D}}_1 \), which is defined as the set of all \((z, z')\) that are associated with some \( g \in \widehat{G}_1 \) in the same way \( \mathcal{D}_1 \) is defined based on \( G_1 \):

\[
\widehat{\mathcal{D}}_1 = \{(z_k, z_{k'}) \in \mathcal{D} : g = (1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_{k'}\}}) \in \widehat{G}_1\}.
\]

To derive the desired consistency result, we state and prove an additional auxiliary lemma.

**Lemma D.2** Under Assumption A.5, \( \widehat{\phi} \to \phi, T_n/n \to \Lambda(P), \) and \( \widehat{\sigma} \to \sigma \) almost uniformly. In addition, \( \sqrt{T_n}(\widehat{\phi} - \phi) \sim \mathcal{G} \) for some random element \( \mathcal{G} \), and for all \((h, g) \in \hat{H} \times G\) with
$g = (g_1, g_2)$, the variance $Var(G(h, g)) = \sigma^2(h, g)$.

**Proof of Lemma D.2.** Note that the spaces $\bar{H}$ and $G$ defined in (D.6) are similar to the spaces $\bar{H}$ and $\bar{G}_P$ defined in (C.7). The lemma can be proved following a strategy similar to that of the proof of Lemma C.2. ■

**Proposition D.1** Suppose the instrument $Z$ is pairwise valid for the treatment $D$ as defined in Definition A.2. Under Assumption A.5, $\mathbb{P}(\hat{G}_1 = G_1) \to 1$, and thus $\mathbb{P}(\hat{Z}_1 = Z_1) \to 1$.

**Proof of Proposition D.1.** First, suppose $G_1 \neq \emptyset$. Then we have that

$$\min_{g \in \{1, \ldots, 2^{J_1}\}} \sup_{h \in \mathcal{H}_q} \{\phi(h, g)/(\xi_0 \vee \hat{\sigma}(h, g))\} = 0$$

for all $g \in G_1$. Under the constructions, we have that

$$\lim_{n \to \infty} \mathbb{P}\left(G_1 \setminus \hat{G}_1 \neq \emptyset\right)$$

$$\leq \lim_{n \to \infty} \mathbb{P}\left(\max_{g \in G_1} \sqrt{T_n} \left| \min_{q \in \{1, \ldots, 2^{J_1}\}} \sup_{h \in \mathcal{H}_q} \frac{\hat{\phi}(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} - \min_{q \in \{1, \ldots, 2^{J_1}\}} \sup_{h \in \mathcal{H}_q} \frac{\hat{\phi}(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \right| > \tau_n\right)$$

$$= \lim_{n \to \infty} \mathbb{P}\left(\max_{g \in G_1} \sqrt{T_n} \left| - \max_{q \in \{1, \ldots, 2^{J_1}\}} \left( - \sup_{h \in \mathcal{H}_q} \frac{\hat{\phi}(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \right) + \max_{q \in \{1, \ldots, 2^{J_1}\}} \left( - \sup_{h \in \mathcal{H}_q} \frac{\hat{\phi}(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \right) \right| > \tau_n\right)$$

$$\leq \lim_{n \to \infty} \mathbb{P}\left(\max_{g \in G_1, h \in \mathcal{H}} \sqrt{T_n} \left| \frac{\hat{\phi}(h, g) - \phi(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \right| > \tau_n\right).$$

By Lemma D.2, $\sqrt{T_n}(\hat{\phi} - \phi) \rightsquigarrow G$ and $\hat{\sigma} \to \sigma$ almost uniformly, which implies that $\hat{\sigma} \rightsquigarrow \sigma$ by Lemmas 1.9.3(ii) and 1.10.2(iii) of van der Vaart and Wellner (1996). Then by Example 1.4.7 (Slutsky’s lemma) and Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996),

$$\max_{g \in G_1, h \in \mathcal{H}} \sqrt{T_n} \left| \frac{\hat{\phi}(h, g) - \phi(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \right| \rightsquigarrow \max_{g \in G_1, h \in \mathcal{H}} \left| \frac{\hat{G}(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \right|.$$

Since $\tau_n \to \infty$, we have that $\lim_{n \to \infty} \mathbb{P}(G_1 \setminus \hat{G}_1 \neq \emptyset) = 0$.

If $G_1 = G$, then clearly $\lim_{n \to \infty} \mathbb{P}(G_1 \setminus \hat{G}_1 \neq \emptyset) = 0$. Suppose now $G_1 \neq \emptyset$. Since $G$ is a finite set and $\hat{\sigma}$ is uniformly bounded, then there is a $\delta > 0$ such that

$$\min_{g \in G \setminus G_1} \min_{q \in \{1, \ldots, 2^{J_1}\}} \sup_{h \in \mathcal{H}_q} \frac{\phi(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} > \delta.$$
Thus, we have that

$$\lim_{n \to \infty} \mathbb{P}(G_1 \setminus G_1 \neq \emptyset) \leq \lim_{n \to \infty} \mathbb{P}\left( \left\{ \max_{g \in G_1 \setminus G_1} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \psi(h, q) \right| > \delta, \quad \right. \right. $$

$$\left. \left. \right. \left. \left. \max_{g \in G_1 \setminus G_1} \sqrt{T_n} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \frac{\psi(h, q)}{\xi_0 \sigma(h, g)} \right| \leq \tau_n \right\} \right)$$

By Lemma D.2, \( \hat{\phi} \to \phi \) almost uniformly. Thus, for every \( \varepsilon > 0 \), there is a measurable set \( A \) with \( \mathbb{P}(A) \geq 1 - \varepsilon \) such that for sufficiently large \( n \),

$$\max_{g \in G_1 \setminus G_1} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \frac{\hat{\psi}(h, g)}{\xi_0 \sigma(h, g)} \right| \geq \max_{g \in G_1 \setminus G_1} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \frac{\psi(h, g)}{\xi_0 \sigma(h, g)} \right| - \frac{\delta}{2}$$

uniformly on \( A \). We now have that

$$\lim_{n \to \infty} \mathbb{P}(G_1 \setminus G_1 \neq \emptyset) \leq \lim_{n \to \infty} \mathbb{P}\left( \left\{ \max_{g \in G_1 \setminus G_1} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \frac{\psi(h, g)}{\xi_0 \sigma(h, g)} \right| > \delta, \quad \right. \right. $$

$$\left. \left. \right. \left. \left. \max_{g \in G_1 \setminus G_1} \sqrt{T_n} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \frac{\psi(h, g)}{\xi_0 \sigma(h, g)} \right| \leq \tau_n \right\} \right) + \mathbb{P}(A^c)$$

$$\leq \lim_{n \to \infty} \mathbb{P}\left( \frac{T_n \delta}{n} < \frac{T_n}{n} \right. \left. \max_{g \in G_1 \setminus G_1} \sqrt{T_n} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \frac{\hat{\psi}(h, g)}{\xi_0 \sigma(h, g)} \right| \leq \frac{\tau_n}{\sqrt{n}} \right) + \varepsilon = \varepsilon,$$

because \( \frac{\tau_n}{\sqrt{n}} \to 0 \) as \( n \to \infty \). Here, \( \varepsilon \) can be arbitrarily small. Thus we have that \( \mathbb{P}(G_1 = G_1) \to 1 \), because \( \mathbb{P}(G_1 \setminus G_1 \neq \emptyset) \to 0 \) and \( \mathbb{P}(G_1 \setminus G_1 \neq \emptyset) \to 0 \).

Second, suppose \( G_1 = \emptyset \). This implies that

$$\min_{g \in G} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \frac{\phi(h, g)}{\xi_0 \sigma(h, g)} \right| > \delta$$

for some \( \delta > 0 \). Since by Lemma D.2, \( \hat{\phi} \to \phi \) almost uniformly, then there is a measurable set \( A \) with \( \mathbb{P}(A) \geq 1 - \varepsilon \) such that for sufficiently large \( n \),

$$\max_{g \in G_1} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \frac{\hat{\phi}(h, g)}{\xi_0 \sigma(h, g)} \right| \geq \max_{g \in G_1} \left| \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in H_q} \frac{\phi(h, g)}{\xi_0 \sigma(h, g)} \right| - \frac{\delta}{2}$$
uniformly on $A$. Thus we now have that
\[
\lim_{n \to \infty} P \left( G_1 \neq \emptyset \right)
\leq \lim_{n \to \infty} P \left( \max_{g \in G_1} \left\{ \min_{q \in \{1, \ldots, 2^j\}} \sup_{h \in H_q} \frac{\phi(h, g)}{\xi_0 \vee \sigma(h, g)} \right\} > \frac{\delta}{\sqrt{n}} \right) + P(A^c)
\leq \lim_{n \to \infty} P \left( \frac{\sqrt{T_n}}{n} - \frac{\delta}{2} < \max_{g \in G_1} \sqrt{T_n} \min_{q \in \{1, \ldots, 2^j\}} \sup_{h \in H_q} \frac{\hat{\phi}(h, g)}{\xi_0 \vee \hat{\sigma}(h, g)} \leq \frac{\tau_n}{\sqrt{n}} \right) + \varepsilon = \varepsilon,
\]
because $\tau_n/\sqrt{n} \to 0$ as $n \to \infty$. Here, $\varepsilon$ can be arbitrarily small. Thus, $P(\hat{G}_1 = G_1) = 1 - P(\hat{G}_1 \neq \emptyset) \to 1$.

Proposition D.1 is also related to the contact set estimation in Sun (2021). Since $G$ is a finite set, we can obtain the stronger result in Proposition D.1, that is, $P(\hat{G}_1 = G_1) \to 1$.

D.2.2 Definition and Estimation of $\mathcal{Z}_2$

The definition of $\mathcal{Z}_2$ is the same as in Appendix C.3.2 because the necessary conditions provided by Kédagni and Mourifié (2020) are for the exclusion and statistical independence conditions only. Therefore, the estimator of $\mathcal{Z}_2$ can be constructed as in Section C.3.2.

E Simulation Evidence

Here we evaluate the finite sample performance of our method in Monte Carlo simulations. We consider the case where $D \in \{0, 1\}$ and $Z \in \{0, 1, 2\}$. The presumed validity set is $\mathcal{Z}_P = \{(0, 1), (0, 2), (1, 2)\}$. For each simulation, we use 1,000 Monte Carlo iterations. To calculate the supremum in $\sqrt{T_n} \sup_{h \in H_q} \hat{\phi}(h, g)/(\xi_0 \vee \hat{\sigma}(h, g))$ for every $g$, we use the same approach as in the empirical application in Section 4.

We consider four data generating processes (DGPs), where Assumption A.1 does not fully hold. These DGPs are similar to those used in Kitagawa (2015) and Sun (2021). We consider two different sample sizes $n \in \{1500, 3000\}$. We report results for $\tau_n \in \{1, 1.5, \ldots, 6.5\}$.

For all DGPs, we specify $U \sim \text{Unif}(0, 1)$, $V \sim \text{Unif}(0, 1)$, and $Z = 2 \times 1\{U \leq 0.3\} + 1\{0.3 < U \leq 0.65\}$. For DGPs (1)–(3), we set $D_z = 1\{V \leq 0.5\}$ for $z = 0, 1, 2$, $D = \sum_{z=0}^{2} 1\{Z = z\} \times D_z$, $N_Z \sim \text{N}(0, 1)$, $N_{00} = N_Z$, and $N_{dz} = N_Z$ for $d = 0, 1$ and $z = 1, 2$. For DGP (4), we specify $N_0 \sim \text{N}(0, 1)$, $N_1 \sim \text{N}(1, 1)$, and $N_2 \sim \text{N}(2, 1)$.
(1): \( N_{10} \sim N(-0.7, 1), Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_d) \)

(2): \( N_{10} \sim N(0, 1.675^2), Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_d) \)

(3): \( N_{10} \sim N(0, 0.515^2), Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_d) \)

(4): \( D_0 = 1\{V \leq 0.6\}, D_1 = 1\{V \leq 0.1\} + 1\{V \geq 0.9\}, D_2 = D_1, D = \sum_{z=0}^{2} 1\{Z = z\} \times D_z, Y = \sum_{d=0}^{1} 1\{D = d\} \times N_d \)

The random variables \( U, V, N, Z, N_{10}, N_0, N_1, \) and \( N_2 \) are mutually independent. Note that, for all DGPs, \( \mathcal{Z}_M \cap \mathcal{Z}_P = \mathcal{Z}_1 \cap \mathcal{Z}_P = \{(1, 2)\} \). Tables E.1–E.4 show the simulation results for DGPs (1)–(4). The tables show the proportions by which each element is selected to be in \( \mathcal{Z}_1 \) in the simulations. The results show that choosing \( \tau_n \in \{3.5, 4\} \) leads to an excellent performance for \( n \in \{1500, 3000\} \). As \( n \) increases, \( \tau_n \) should be increased accordingly. Overall, the simulation results show that the proposed method performs well in identifying the validity pair set in practice.

### Table E.1: Validity Pair Set Estimation DGP (1)

| \( n \) | \( \tau_n = 1 \) | \( \tau_n = 1.5 \) | \( \tau_n = 2 \) |
| --- | --- | --- | --- |
| (0, 1) | 0.000 | 0.000 | 0.000 |
| (0, 2) | 0.000 | 0.000 | 0.000 |
| (1, 2) | 0.000 | 0.000 | 0.000 |
| \( n \) | \( \tau_n = 2.5 \) | \( \tau_n = 3 \) | \( \tau_n = 3.5 \) |
| (0, 1) | 0.000 | 0.000 | 0.000 |
| (0, 2) | 0.000 | 0.000 | 0.000 |
| (1, 2) | 0.000 | 0.000 | 0.000 |
| \( n \) | \( \tau_n = 4 \) | \( \tau_n = 4.5 \) | \( \tau_n = 5 \) |
| (0, 1) | 0.000 | 0.000 | 0.000 |
| (0, 2) | 0.000 | 0.000 | 0.000 |
| (1, 2) | 0.000 | 0.000 | 0.000 |
| \( n \) | \( \tau_n = 5.5 \) | \( \tau_n = 6 \) | \( \tau_n = 6.5 \) |
| (0, 1) | 0.000 | 0.000 | 0.000 |
| (0, 2) | 0.000 | 0.000 | 0.000 |
| (1, 2) | 0.000 | 0.000 | 0.000 |

45
Table E.2: Validity Pair Set Estimation DGP (2)

| n   | $\tau_n = 1$ | $\tau_n = 1.5$ | $\tau_n = 2$ |
|-----|---------------|-----------------|---------------|
|     | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) |
| 1500 | 0.000 0.000 0.000 | 0.000 0.000 0.000 | 0.000 0.000 0.000 |
| 3000 | 0.000 0.000 0.000 | 0.000 0.000 0.000 | 0.000 0.000 0.000 |

| n   | $\tau_n = 2.5$ | $\tau_n = 3$ | $\tau_n = 3.5$ |
|-----|-----------------|---------------|-----------------|
|     | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) |
| 1500 | 0.000 0.000 0.000 | 0.000 0.000 0.159 | 0.000 0.002 0.690 |
| 3000 | 0.000 0.000 0.000 | 0.000 0.000 0.030 | 0.000 0.000 0.552 |

| n   | $\tau_n = 4$ | $\tau_n = 4.5$ | $\tau_n = 5$ |
|-----|---------------|-----------------|---------------|
|     | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) |
| 1500 | 0.000 0.000 0.000 | 0.116 0.089 0.990 | 0.354 0.350 0.999 |
| 3000 | 0.000 0.000 0.000 | 0.000 0.000 0.983 | 0.000 0.000 0.998 |

| n   | $\tau_n = 5.5$ | $\tau_n = 6$ | $\tau_n = 6.5$ |
|-----|-----------------|---------------|-----------------|
|     | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) |
| 1500 | 0.675 0.682 1.000 | 0.891 0.904 1.000 | 0.982 0.976 1.000 |
| 3000 | 0.005 0.006 1.000 | 0.049 0.062 1.000 | 0.206 0.260 1.000 |

Table E.3: Validity Pair Set Estimation DGP (3)

| n   | $\tau_n = 1$ | $\tau_n = 1.5$ | $\tau_n = 2$ |
|-----|---------------|-----------------|---------------|
|     | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) |
| 1500 | 0.000 0.000 0.000 | 0.000 0.000 0.000 | 0.000 0.000 0.000 |
| 3000 | 0.000 0.000 0.000 | 0.000 0.000 0.000 | 0.000 0.000 0.000 |

| n   | $\tau_n = 2.5$ | $\tau_n = 3$ | $\tau_n = 3.5$ |
|-----|-----------------|---------------|-----------------|
|     | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) |
| 1500 | 0.000 0.000 0.000 | 0.000 0.000 0.159 | 0.001 0.001 0.690 |
| 3000 | 0.000 0.000 0.000 | 0.000 0.000 0.030 | 0.000 0.000 0.552 |

| n   | $\tau_n = 4$ | $\tau_n = 4.5$ | $\tau_n = 5$ |
|-----|---------------|-----------------|---------------|
|     | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) |
| 1500 | 0.009 0.005 0.939 | 0.036 0.034 0.990 | 0.089 0.121 0.999 |
| 3000 | 0.000 0.000 0.901 | 0.000 0.000 0.983 | 0.000 0.000 0.998 |

| n   | $\tau_n = 5.5$ | $\tau_n = 6$ | $\tau_n = 6.5$ |
|-----|-----------------|---------------|-----------------|
|     | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) | (0, 1) (0, 2) (1, 2) |
| 1500 | 0.202 0.249 1.000 | 0.379 0.434 1.000 | 0.592 0.658 1.000 |
| 3000 | 0.001 0.001 1.000 | 0.004 0.008 1.000 | 0.024 0.039 1.000 |
### Table E.4: Validity Pair Set Estimation DGP (4)

| \( n \) | \( \tau_n = 1 \) | \( \tau_n = 1.5 \) | \( \tau_n = 2 \) |
|---|---|---|---|
| (0, 1) | (0, 2) | (1, 2) | (0, 1) | (0, 2) | (1, 2) | (0, 1) | (0, 2) | (1, 2) |
| 1500 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 3000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| \( n \) | \( \tau_n = 2.5 \) | \( \tau_n = 3 \) | \( \tau_n = 3.5 \) |
| (0, 1) | (0, 2) | (1, 2) | (0, 1) | (0, 2) | (1, 2) | (0, 1) | (0, 2) | (1, 2) |
| 1500 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.227 | 0.000 | 0.000 | 0.761 |
| 3000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.055 | 0.000 | 0.000 | 0.645 |
| \( n \) | \( \tau_n = 4 \) | \( \tau_n = 4.5 \) | \( \tau_n = 5 \) |
| (0, 1) | (0, 2) | (1, 2) | (0, 1) | (0, 2) | (1, 2) | (0, 1) | (0, 2) | (1, 2) |
| 1500 | 0.000 | 0.000 | 0.960 | 0.000 | 0.000 | 0.995 | 0.000 | 0.000 | 1.000 |
| 3000 | 0.000 | 0.000 | 0.937 | 0.000 | 0.000 | 0.991 | 0.000 | 0.000 | 0.999 |
| \( n \) | \( \tau_n = 5.5 \) | \( \tau_n = 6 \) | \( \tau_n = 6.5 \) |
| (0, 1) | (0, 2) | (1, 2) | (0, 1) | (0, 2) | (1, 2) | (0, 1) | (0, 2) | (1, 2) |
| 1500 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 |
| 3000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 1.000 |