WARWICK
ECONOMICS

Synthetic Decomposition for Counterfactual Predictions

Nathan Canen and Kyungchul Song

July 2023 No: 1466

Warwick Economics Research Papers

ISSN 2059-4283 (online)
ISSN 0083-7350 (print)

SYNTHETIC DECOMPOSITION FOR COUNTERFACTUAL PREDICTIONS

Nathan Canen and Kyungchul Song
University of Houston, University of Warwick & NBER, and University of British Columbia

ABSTRACT. Counterfactual predictions are challenging when the policy variable goes beyond its pre-policy support. However, in many cases, information about the policy of interest is available from different ("source") regions where a similar policy has already been implemented. In this paper, we propose a novel method of using such data from source regions to predict a new policy in a target region. Instead of relying on extrapolation of a structural relationship using a parametric specification, we formulate a transferability condition and construct a synthetic outcome-policy relationship such that it is as close as possible to meeting the condition. The synthetic relationship weighs both the similarity in distributions of observables and in structural relationships. We develop a general procedure to construct asymptotic confidence intervals for counterfactual predictions and prove its asymptotic validity. We then apply our proposal to predict average teenage employment in Texas following a counterfactual increase in the minimum wage.

KEY WORDS: Counterfactual Predictions, Decomposition Analysis, Ex Ante Policy Evaluation, Synthetic Decomposition, Uniform Asymptotic Validity

JEL CLASSIFICATION: C30, C54

Date: July 11, 2023.
We thank Victor Aguirregabiria, Tim Armstrong, Xu Cheng, EunYi Chung, Wayne Gao, Yu-Chin Hsu, Hiro Kasahara, Vadim Marmer, Ismael Mourifie, Vitor Possebom, Frank Schorfheide and Paul Schrimpf, and participants in the seminars in Seoul National University, University of Calgary, University of Illinois Urbana-Champaign, University of Norte Dame, University of Pennsylvania, University of Toronto, University of Victoria, and in CIREQ Montreal Econometrics Conference, Conference on Econometrics for Modern Data Structures, and SETA 2022 for valuable comments. All errors are ours. We also thank Ratzanyel Rincón for his excellent research assistance. Song acknowledges that this research was supported by Social Sciences and Humanities Research Council of Canada.
Corresponding Address: Kyungchul Song, Vancouver School of Economics, University of British Columbia, 6000 Iona Drive, Vancouver, BC, V6T 1L4, Canada, kysong@mail.ubc.ca.

1. Introduction

Policymakers' questions are often centered around the prediction of a new policy's outcome, such as predicting the effect of a new job training program, the welfare implication of a proposed merger of firms, or the employment effect of a minimum wage increase.[1] Such questions are hard to answer because the new policy's outcome is not observed. For example, when a state in the U.S. considers increasing its minimum wage to a level never seen before within that state, this will imply that the policy is beyond its historical variations. In this situation, the researcher may consider using a parametric specification of the outcome-policy relationship and extrapolate it to a post-policy setting. However, when the new policy goes beyond the support of its historical variations, the counterfactual prediction may fail to be nonparametrically identified and the prediction inevitably relies on the particular parametrization that is chosen.

Alternatively, the researcher may use data from another region that has experienced a similar policy in the past. Transferring empirical features from one source to another has long been used in economics. In macroeconomics, it is common to calibrate a structural model by using estimates from micro-studies (see Gregory and Smith (1993) for a review). In a different context, the decomposition method in labor economics can also be viewed as following a similar idea: the researcher "transfers" information from a population in one period (before the policy) to the same population in another period (after the policy) (see Fortin, Lemieux, and Firpo (2011) for a review of this vast literature).[2] The program evaluation literature also explores various methods of transferring causal inference results from one experiment setting to another experiment or non-experiment setting (see Hotz, Imbens, and Mortimer (2005), Hartman, Grieve, Ramsahai, and Sekhon (2015), Athey, Chetty, and Imbens (2020), Gechter and Meager (2022), to name but a few). However, to the best of our knowledge, much less attention has been paid to the transfer problem when predicting counterfactual policies using structural equation models.

In this paper, we consider the problem of generating counterfactual predictions from a new policy, using data from other regions that have experienced a similar policy in the past. In order to transfer empirical features across regions in structural models, the outcome-policy relationship needs to be "transferable" from one region to another. For example, we may want to predict the average teenage employment after a minimum wage increase in Texas in the U.S.,

[1] See Heckman and Vytlacil (2005), Heckman (2010), and Wolpin (2013) for discussions on the importance and challenges of *ex ante* policy evaluations. See Hotz, Imbens, and Mortimer (2005) for related issues in a program evaluation setting.

[2] Since the population may have changed between the two periods for reasons other than the policy, we can view the decomposition method as a special form of a transfer between different populations.

and consider using data from California, assuming that its structural relationship between the teenage employment outcome and the minimum wage (after controlling for some observed characteristics) is identical to that in Texas.[3] However, the transferability between two regions can be strong in practice, especially when the market environments in the two regions exhibit salient differences.

As we show in this paper, the transferability issue can be alleviated when we have multiple "source" regions in which similar policies have been implemented in the past. In the minimum wage example with Texas as the target region, there may be several other states such as California, Oregon, and Connecticut which experienced similar minimum wage increases. However, aggregating data from these source regions is not immediately obvious. Ideally, it would be desirable to choose a source region that is most "similar" to the target region, but it is not clear which dimensions of the characteristics between the two regions would be most relevant for a given policy prediction problem.

To solve this issue, we develop a method of aggregating information from multiple source regions to generate counterfactual predictions in the target region by constructing a synthetic structural relationship from multiple source regions. First, as noted by Todd and Wolpin (2008), structural equation models often involve the policy variable in an index (called the *policy component* here) which exhibits variations at the individual level. We can classify each person in the target population into the *matched group* and *unmatched group* depending on whether the person's post-policy value of the policy component can be matched with another person's pre-policy value of the policy component in the same population. As proposed by Todd and Wolpin (2008), we can use the pre-policy data from the target population to nonparametrically identify the counterfactual predictions for the matched group even under new policies never implemented before (see Wolpin (2013) for an overview of this approach).

To generate counterfactual predictions for the unmatched group, we introduce what we call the *synthetic transferability condition* which requires that there exist a weight vector such that the weighted average of the outcome-policy relationships in the source regions coincides with that of the target region. This condition is weaker than the transferability condition with a single source region, described above, as it does not require the source regions to have the same structural relationship as the target region. Then, under a non-redundancy condition for the source regions, we can identify this weight as a minimizer of an L^2 distance between

[3] As we explain later, the transferability condition is closely related to conditional external validity or the external unconfoundedness condition in the literature of causal inference and external validity (see Hotz, Imbens, and Mortimer (2005), Hartman, Grieve, Ramsahai, and Sekhon (2015), and Athey, Chetty, and Imbens (2020)). The synthetic transferability condition has a testable implication in a spirit similar to checking the pre-treatment fit in synthetic control methods. Our proposal entails a formal test of this implication with uniform asymptotic size control.

the outcome-policy relationship in the target region and the weighted average of the outcome-policy relationships in the source regions, where the L^2 distance is restricted to the matched group in the target region. Thus, we find a weighted average of the outcome-policy relationships in the source regions that is most similar to that in the target region on the matched group. This weighted average can be viewed as a synthetic structural relationship that can be used to generate counterfactual predictions. As our proposal essentially replaces the structural relationship involved in the decomposition method with a synthetic one, we call this method a *synthetic decomposition method*.

Our method is quite general and can be applied to a wide range of counterfactual prediction settings. In particular, we consider a generic nonparametric form of an outcome-policy relationship that is nonseparable in the (potentially multi-dimensional) unobserved heterogeneity. This flexibility allows the researcher to derive a nonparametric outcome-policy relationship from a structural model that specifies peoples' incentives and choices differently across the populations. Furthermore, the type of a policy can vary, including policies that transform a certain individual-level exogenous variable (e.g., demographic-dependent tax subsidies) or an aggregate-level exogenous variable (e.g., minimum wages). The policy can be one that changes a structural parameter or a coefficient of a certain variable, or a change in the distribution of an exogenous observed variable.

We then develop inference on the counterfactual prediction from the synthetic decomposition method. In this paper, we pursue a general approach that does not require the researcher's knowledge of the details of the asymptotic properties of estimators for each source region, because such properties may vary depending on the particular model specified for each region (e.g., the specification of the structural relationship between outcomes, a policy, and observed or unobserved characteristics). More specifically, we develop an inference method for the policy predictions inspired by Rosen (2008), Moon and Schorfheide (2009), Shi and Shum (2015), Bugni, Canay, and Shi (2017), and Cox and Shi (2022). The non-standard aspect of inference in our context is that the estimated weight is chosen from a simplex, and hence, the limiting distribution of the estimated weight depends on how close the population weight is to a vertex or an edge of the simplex. While the situation is analogous to a setting with the parameter on the boundary studied by Geyer (1994) and Andrews (1999), their approach of quadratic approximation does not apply here.[4] By adapting the proposal of Mohamad, van Zwet, Cator,

[4]Asymptotic or bootstrap inference for constrained estimators has received a considerable attention in the econometrics literature. More recent examples include Kaido and Santos (2014), Kitamura and Stoye (2018), Fang and Seo (2021), Hsieh, Shi, and Shum (2022) and Li (2022) to name but a few. See those papers for more references in this literature.

and Goeman (2020) and Cox and Shi (2022) to our setting, we develop an asymptotic inference method that is valid uniformly over the behavior of the population weight. Monte Carlo simulations suggest that our procedure works well in finite samples.

We illustrate our procedure with an empirical application studying the effects of a counterfactual increase in minimum wages in Texas to US$9. (The prevailing minimum wage is the federal level of US$7.25, set in 2009.) Such increases are subject to extensive policy and academic interest, as shown by it being a central policy proposal in the 2022 Texas gubernatorial elections.[5] However, the extensive minimum wage literature in labor economics predominantly focuses on *ex post* analyses of minimum wage increases (Neumark (2019)). We implement our proposed method using Current Population Survey (CPS) data and estimate that an increase in minimum wages would decrease average (teenage) employment by 9.5-11 percentage points on a baseline of approximately 29% if minimum wages in Texas were US$9. In doing so, our synthetic comparison for Texas (i) accounts for the heterogeneous skill distributions and demand conditions across states (Flinn (2011)), (ii) does not require the researcher to choose the comparison unit (e.g., whether to focus on geographically close or distant states - see Dube, Lester, and Reich (2010) and Neumark (2019) for a discussion), (iii) accounts for the difference in causal relationships between minimum wages and employments across states (Flinn (2002)), (iv) does not rely on parametric extrapolation, which is a concern in this literature - see Flinn (2006); Gorry and Jackson (2017); Neumark (2019), for example.

Related Literature

The importance of *ex ante* policy evaluation in economics has been emphasized in the literature. See, for example, Heckman and Vytlacil (2005), Wolpin (2007) and Wolpin (2013). See also the review by Heckman (2010) and the references therein. The evaluation usually builds on an invariance condition that requires certain structural relationships to remain unchanged after the policy. While most literature on program evaluations focuses on measuring the impact of a policy, the invariance of structural relationships that underlies the measured impact is crucial for understanding the reasons for the effect of the policy and predicting the effect of a new policy.

The literature of counterfactual predictions using structural models has often been motivated by the *ex ante* policy evaluation settings in practice. Stock (1989) studied the problem of counterfactual predictions using a structural equation model, when a policy changes the distribution of an exogenous variable. A recent stream of literature studies the nonparametric identification of counterfactual predictions in structural models (see Aguirregabiria (2005), Jun and Pinkse (2020), and Gu, Russell, and Stringham (2022), to name but a few.) A consistent theme in this

[5]See `tinyurl.com/2cmv7fhz` for its presence and analysis within one of the candidate's policy platforms.

literature is that certain objects of counterfactual prediction are nonparametrically identified even when the structural model is not fully identified. Examples include Aguirregabiria and Suzuki (2014), Norets and Tang (2014), Arcidiacono and Miller (2020), Kalouptsidi, Scott, and Souza-Rodrigues (2020), Kalouptsidi, Kitamura, Lima, and Souza-Rodrigues (2020), and Canen and Song (2023). However, this literature usually considers identification using data from the target population only.

Our proposal is closely related to the decomposition method in labor economics. After the seminal papers of Oaxaca (1973) and Blinder (1973), the decomposition method has been extended and widely used in labor economics.[6] See Juhn, Murphy, and Pierce (1993) and DiNardo, Fortin, and Lemieux (1996). The causal interpretation of the decomposition method was studied by Kline (2011). See also Fortin, Lemieux, and Firpo (2011) for an extensive review of this literature. There is a growing literature of nonparametric counterfactual prediction inspired by the decomposition method. See Rothe (2010), Chernozhukov, Fernández-Val, and Melly (2013), and Hsu, Lai, and Lieli (2022). The transfer of results from a source population to a target population requires various forms of invariance conditions. (See Heckman and Vytlacil (2007) for a detailed discussion on these conditions.) See also Hsu, Lai, and Lieli (2022) for an invariance condition invoked in the problem of counterfactual predictions from one study context to another. In the literature of statistics, a similar form of an invariance condition is proposed by Bühlmann (2020).

A growing attention has been paid to the issue of external validity in field experiments, such as when the results obtained from experiments are not replicated in their scaled-up implementation. (See Allcott (2015), Bold, Kimenyi, Mwabu, Sandefur, et al. (2018), and Wang and Yang (2021) and references therein. See also Duflo (2004) and Muralidharan and Niehaus (2017) for the review of these issues and the literature.) One of the earliest approaches to detect or address the issue of external validity in program evaluations is found in Hotz, Imbens, and Mortimer (2005). They consider the problem of using past experiment results to predict its outcome for a different population. They introduce what they call the assumption of unconfounded location and show the external validity of the past experiments under this assumption. Variants of this assumption have been used in the study of external validity in the literature. Examples include Hartman, Grieve, Ramsahai, and Sekhon (2015), Athey, Chetty, and Imbens (2020), and Gui (2022). Another strand of related literature considers combining experimental and observational data (see, e.g., Athey, Chetty, and Imbens (2020) and Gechter and Meager (2022)), or performs meta analysis by aggregating experiment results across studies using a Bayesian hierarchical model (BHM) or pursuing minimax regret optimality (see,

[6] Vitor Possebom kindly let us know that there was an early appearance of a similar idea in Kitagawa (1955).

e.g., Vivalt (2020), Bandiera, Fischer, Prat, and Ytsma (2021), Ishihara and Kitagawa (2021), and Meager (2022).)

Our paper's method of constructing a counterfactual prediction is closely related to Todd and Wolpin (2006). They estimate a dynamic structural model of the labor market using pre-treatment data from a randomized experiment in Mexico and validate the model by comparing the predictions from the structural model with the randomized experiment results. Then, they use the model to generate counterfactual predictions from different policies. Instead of building up a full structural model, we follow Todd and Wolpin (2008) and focus on the nonparametric outcome-policy relationship that is relevant to the policy prediction problem. To deal with a setting where the policy variable goes out of the support, we develop a new method that uses data from multiple source populations.

Our synthetic decomposition method is inspired by the method of synthetic control which currently attracts a wide attention in applied research and the literature of econometrics (see Abadie (2021) for an overview and the related literature on the method). Both methods are similar in the sense that they aim to construct a synthetic comparison group using data from multiple populations, instead of relying on an ad-hoc comparison of various characteristics of the regions. However, the way the comparison group is constructed is fundamentally different. The synthetic control method compares the pre-treatment *outcomes* between the target population and the source populations, whereas the synthetic decomposition method compares the pre-treatment outcome-policy *relationships* between the target population and the source populations on the *matched group*. The synthetic decomposition method does not require observations over multiple time periods, but requires individual-level data for each population.

The rest of the paper proceeds as follows. In Section 2, we present our main proposal of the synthetic decomposition method and discuss conditions for the method to work. In Section 3, we provide procedures of estimation and construction of confidence intervals, assuming that we observe a random sample of data from each population. In Section 4, we present an empirical application that studies the prediction problem of average employment when the minimum wage increases in Texas. In Section 5, we conclude. In the Supplemental Note, we present general conditions for the proposed confidence intervals to be uniformly asymptotically valid, as well as formal results and proofs. The Supplemental Note also contains some more details on the Monte Carlo simulations and the empirical application.

2. Synthetic Decomposition for Counterfactual Predictions

2.1. The Target Population and Counterfactual Predictions

Suppose that there is a region where we are interested in predicting the outcome of a new policy. The outcome variable is denoted by Y_i and the observed random vector of exogenous variables by X_i. We assume that they are related as follows:

$$Y_i = g_0(\mu_0(X_i), U_i), \tag{1}$$

where U_i is an unobserved random vector, μ_0 is a map subject to a change depending on a policy, and g_0 is a map that is invariant to the policy. We call the population in the region the ***target population***.[7] In our paper, we focus on settings where we do not have long time series data, and hence, the randomness of variables arises only from their within-population cross-sectional variations. This means that, in our paper, the aggregate variables behave like nonstochastic quantities. Throughout this paper, we assume that X_i does not include any aggregate variables, and treat observed aggregate variables as "observed parameters."

A policy is a transform of the map μ_0 into μ_0^Γ. Thus, after the policy, the relation between X_i and Y_i changes as follows:

$$Y_i = g_0\left(\mu_0^\Gamma(X_i), U_i\right). \tag{2}$$

We call $\mu_0(X_i)$ and $\mu_0^\Gamma(X_i)$ the ***policy components***. The main requirement for the policy components is that they exhibit cross-sectional variations before and after the policy.

The target parameter is the predicted average outcome after the policy and is written as

$$\theta_0 = \mathbf{E}_0\left[g_0\left(\mu_0^\Gamma(X_i), U_i\right)\right], \tag{3}$$

where $\mathbf{E}_0$ denotes the expectation with respect to the target population P_0. Our framework accommodates various forms of policies. We discuss some examples of policy components below.

Example 1 (Transforming an Individual Covariate): $\mu_0(x) = x$ and $\mu_0^\Gamma(x) = f(x)$ for some function f. For example, suppose that $X_i = (X_{1,i}, X_{2,i})$ where $X_{1,i}$ represents an individual's income and $X_{2,i}$ represents other demographic characteristics. Now the policy of interest is an income subsidy by an amount, say, $\delta > 0$, for each individual with X_i in a set A. Then,

[7] It is important to note that the notion of "population" here not only refers to the joint distribution of random variables, but also depends on the causal model of how the random variables are generated. Hence, two identical joint distributions that are generated according to different causal models are treated as drawn from different populations.

we can take

$$f(x) = (x_1 + \delta, x_2)1\{x \in A\} + (x_1, x_2)1\{x \notin A\}.$$

Even if the amount δ is the same across individuals, the policy components μ_0 and μ_0^Γ generally exhibit variations at the individual level.[8]

Example 2 (Changing a Structural Parameter or an Aggregate Variable): $\mu_0(x) = q(x; v_0)$ for a parametric function $q(\cdot; v_0)$ with parameter v_0, and $\mu_0^\Gamma(x) = q(x; \tilde{v}_0)$ for a different parameter $\tilde{v}_0$. For example, the parameter can represent certain structural parameters such as parameters of the matching function in search and matching models. Alternatively, we may consider a setting where the policy affects some aggregate state variable, such as the minimum wage, the level of a sales tax or the population size through immigration policies. In such cases, we can view v_0 as the aggregate variable that the policy targets, and $\tilde{v}_0$ as its post-policy value. Note that the aggregate policy variable v_0 does not vary at the individual level, but the policy components $q(x; v_0)$ and $q(x; \tilde{v}_0)$ can.

It is convenient to introduce what we call the ***Average Response Function (ARF)*** as follows:[9]

$$m_0(\overline{\mu}, x) = \int g_0(\overline{\mu}, u) dP_{0,U|X}(u \mid x).$$

where $P_{0,U|X}$ denotes the conditional distribution of U_i given X_i in the target population (before the policy). The ARF summarizes the structural relationship between the outcome and the policy component in the model. Note that the dependence of the ARF $m_0(\overline{\mu}, x)$ on the first argument $\overline{\mu}$ is *causal* so that we can use this dependence for counterfactual analysis when we change the value of $\overline{\mu}$. However, the dependence on the second argument x is not, because in this model the causal relation between U_i and X_i is left ambiguous. The target parameter is written as follows:

$$\theta_0 = \int m_0\left(\mu_0^\Gamma(x), x\right) dP_0(x). \tag{4}$$

[8]It is important to note that this simple setting of counterfactual prediction is already different from the standard program evaluation setting. Here, the potential outcomes are given as follows:

$$Y_i(0) = g_0(\mu_0(X_i), U_i) \text{ and } Y_i(1) = g_0(\mu_0^\Gamma(X_i), U_i).$$

However, unlike the standard program evaluation setting, *everybody is treated* here. Furthermore, we focus on an *ex ante* policy evaluation where we do not observe the outcome of the policy for the target population yet. (See Heckman and Vytlacil (2007) and Wolpin (2013) for the problem of policy analysis in such a setting.)

[9]The ARF is closely related to the Local Average Response function (LARF) proposed by Altonji and Matzkin (2005). In fact, if we take $\mu_0(x) = x$, and take the derivative of the ARF with respect to the first argument and evaluate it at the same value as the second argument, the derivative becomes the LARF. The ARF is generally different from the ASF (Average Structural Function) introduced by Blundell and Powell (2003), unless X_i and U_i are independent.

From here on, we call the map $m_0\left(\mu_0^\Gamma(\cdot),\cdot\right)$ the ***post-policy ARF*** in the target population.[10] As noted by Blundell and Powell (2003) for the case of ASF (Average Structural Function), it is not always necessary to recover the reduced form g_0 from data to obtain the ARF in many settings. See Canen and Song (2023) for a similar observation in game-theoretic models.

We assume that the target population has not experienced the policy yet. The average counterfactual outcome θ_0 is not nonparametrically identified when the policy sends the policy component outside of its pre-policy support. To address this issue, we may choose a parametric specification of the map g_0 and extrapolate it beyond the support of the pre-policy data. However, since the counterfactual prediction is not nonparametrically identified, it unavoidably relies on the choice of a parametric specification. To address this challenge, we consider using information from other populations which have already implemented a similar policy. We will explain this idea later.

Let $\mathcal{X}_0$ be the support of X_i in the target population. Define

$$\mathcal{X}_0^\Gamma = \left\{x \in \mathcal{X}_0 : \mu_0^\Gamma(x) = \mu_0(\tilde{x}), \text{ for some } \tilde{x} \in \mathcal{X}_0\right\}. \tag{5}$$

Roughly speaking, the set $\mathcal{X}_0^\Gamma$ is the set of values x such that the post-policy value $\mu_0^\Gamma(x)$ matches up with the pre-policy value $\mu_0(\tilde{x})$ for some $\tilde{x} \in \mathcal{X}_0$. We follow Wolpin (2013) and call the set $\mathcal{X}_0^\Gamma$ a ***matched group***.

Example 2.1. Consider the following simple model for Y_i for the target population:

$$Y_i = \mu_0(X_i) + U_i,$$

and assume that X_i and U_i are independent with U_i having mean zero, and the policy changes X_i to $X_i + \Delta$ for some vector Δ. Then, $\mu_0^\Gamma(x) = \mu_0(x+\Delta)$, and the ARF is given by the identity map $\overline{\mu} \mapsto \overline{\mu}$, and the matched group is given by

$$\mathcal{X}_0^\Gamma = \{x \in \mathcal{X}_0 : \mu_0(x+\Delta) = \mu_0(\tilde{x}) \text{ for some } \tilde{x} \in \mathcal{X}_0\}.$$

See Figure 1 for an illustration. ■

We require the identification of the post-policy ARF for the target population only for x in the matched group.

[10] Alternatively, we might be interested in a ***post-policy Distributional Response Function (DRF)***: for some set A,

$$p_0(A; \mu_0^\Gamma(x), x) = \int 1\left\{g_0(\mu_0^\Gamma(x), u) \in A\right\} dP_{0,U|X}(u \mid x).$$

The quantity represents the conditional probability of Y_i taking values from a set A given $X_i = x$, when $\mu_0(X_i)$ is *fixed to be* $\mu_0(x)$. Once we replace $m_0(\mu_0^\Gamma(x), x)$ by $p_0(A; \mu_0^\Gamma(x), x)$, the main proposal of this paper carries over to this alternative.

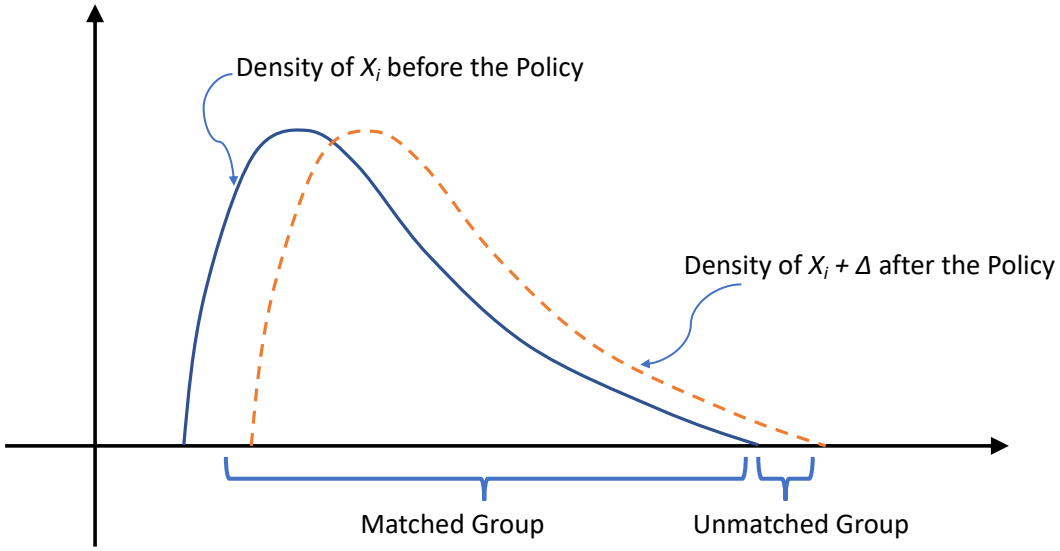


FIGURE 1. The Matched Group and Unmatched Group from the Policy Shifting the Distribution of X_i: In the analysis of the policy that changes X_i to $X_i + \Delta$, the pre-policy outcomes of people in the matched group are compared with the post-policy outcomes of other people in the same population.

Assumption 2.1 (Identification of the Post-Policy ARF on a Matched Group)**.** The set $\mathcal{X}_0^\Gamma \subset \mathcal{X}_0$ is non-empty and $m_0(\mu_0^\Gamma(x), x)$ is identified for all $x \in \mathcal{X}_0^\Gamma$.

Essentially, Assumption 2.1 requires that a portion of the target population before the policy can be used to predict outcomes after the policy. If $\mu_0(X_i) = X_i$ and X_i and U_i are independent, the ARF is immediately identified as the conditional expectation:

$$m_0(\overline{\mu}, x) = \mathbf{E}_0\left[Y_i \mid X_i = \overline{\mu}\right].$$

When X_i and U_i are not independent, one can still identify the ARF when there is an appropriate control function, as shown by Blundell and Powell (2003) for the case of ASF.

2.2. Transfer from a Source Population

Suppose that there is a region 1 which has already experienced the policy. It seems reasonable to assume that the data from this region should be useful for policy prediction in the target region in some way or another. We call the population in this region the ***source population***. Suppose that Y_i and X_i in the source region are related by the following reduced form:

$$Y_i = g_1\left(\mu_1^\Gamma(X_i), U_i\right),$$

where the source post-policy ARF, $m_1(\mu_1^\Gamma(x), x)$, is identified for each $x \in \mathcal{X}_0$. The joint distribution of (Y_i, X_i) can differ between the source and target regions, but we assume the following ***transferability condition***: for all $x \in \mathcal{X}_0$,

$$m_1\left(\mu_1^\Gamma(x), x\right) = m_0\left(\mu_0^\Gamma(x), x\right). \tag{6}$$

The condition (6) says that the average outcome-policy relationship in the source population is *transferable* to that in the target population. In this setting, we can identify θ_0 as follows:

$$\theta_0 = \int_{\mathcal{X}_0^\Gamma} m_0\left(\mu_0^\Gamma(x), x\right) dP_0(x) + \int_{\mathcal{X}_0 \setminus \mathcal{X}_0^\Gamma} m_1\left(\mu_1^\Gamma(x), x\right) dP_0(x).$$

This identification strategy can be viewed as originating from the Oaxaca-Blinder decomposition method in labor economics. To see the connection with the decomposition method, consider a setting where at time t_0, the policy is not implemented and Y_i is generated as follows:

$$Y_i = g_0\left(\mu_0(X_i), U_i\right), \text{ with } X_i \sim P_{0,X},$$

and at time $t_1 > t_0$, the policy is implemented and Y_i is generated as follows:

$$Y_i = g_1\left(\mu_1^\Gamma(X_i), U_i\right), \text{ with } X_i \sim P_{1,X}.$$

Here $P_{0,X}$ and $P_{1,X}$ denote the distribution of X_i before and after the policy. Then, the difference between the expected outcomes at times t_1 and t_0 is given by

$$\begin{aligned} \mathbf{E}_1[Y_i] - \mathbf{E}_0[Y_i] = &\int m_1\left(\mu_1^\Gamma(x), x\right)\left(dP_{1,X}(x) - dP_{0,X}(x)\right) \\ &+ \int \left(m_1\left(\mu_1^\Gamma(x), x\right) - m_0(\mu_0(x), x)\right) dP_{0,X}(x). \end{aligned} \tag{7}$$

Thus, the change in the mean of Y_i before and after the policy is decomposed into the component due to the change in the distribution of X_i and the component due to the change in the average outcome-policy relationship. The transferability condition (6) excludes the possibility that the change of the conditional average outcome given X_i between times t_0 and t_1 is due to events other than the policy. Under the transferability condition, the second term on the right hand side of (7) can be interpreted as an average causal effect of the policy in the target population.

The transferability condition (6) is related to the conditional external validity conditions used in program evaluations (see Hotz, Imbens, and Mortimer (2005), Hartman, Grieve, Ramsahai, and Sekhon (2015), Athey, Chetty, and Imbens (2020), and Gui (2022)). For example, if we take the potential outcomes

$$Y_i(0) = g_k\left(\mu_k(X_i), U_i\right) \text{ and } Y_i(1) = g_k\left(\mu_k^\Gamma(X_i), U_i\right), \quad k = 0, 1,$$

with k depending on the population the individual i belongs to, then the transferability condition (6) holds if for each x and $d = 0, 1$, the conditional distribution of $Y_i(d)$ given $X_i = x$ remains the same across the two populations. This latter condition follows from the location unconfoundedness condition of Hotz, Imbens, and Mortimer (2005).

However, the transferability condition in (6) can be strong in practice. We can relax this transferability condition when we have at least two source populations that satisfy certain conditions.

2.3. Multiple Source Populations and Synthetic Decomposition

2.3.1. **Multiple Source Populations.** We assume that there are K regions (e.g., countries, states, markets, etc.), and each region $k = 1, ..., K$ has a random vector (Y_i, X_i, U_i) with a distribution P_k, where for each $k = 1, ..., K$, we assume that Y_i is generated according to the following reduced form:

$$Y_i = g_k(\mu_k(X_i), U_i),$$

where the map g_k governs the causal relationship between the outcome variable Y_i and the exogenous variables $(\mu_k(X_i), U_i)$ in region k. Note that the causal map, g_k, varies per region, reflecting different structural relationships (e.g., laws, structural parameters, regulations, equilibria, etc.) We call P_k's the source populations. As in the target population, the map μ_k is subject to a change by a policy. Each source population k has experienced a policy that changes μ_k into μ_k^Γ. After the policy, Y_i is generated as follows:

$$Y_i = g_k\left(\mu_k^\Gamma(X_i), U_i\right).$$

Similarly as for the source population, we define the ARF:

$$m_k(\overline{\mu}, x) = \int g_k(\overline{\mu}, u) dP_{k,U|X}(u \mid x),$$

where $P_{k,U|X}$ denotes the conditional distribution of U_i given X_i in population k. We let $\mathcal{X}_k$ denote the support of X_i in the source population P_k.

We introduce the following transferability condition for the source populations. Let Δ_{K-1} denote the $(K-1)$-simplex, i.e., $\Delta_{K-1} = \{\mathbf{w} \in \mathbf{R}^K : \sum_k w_k = 1, w_k \geq 0, k = 1, ..., K\}$.

Assumption 2.2 (Synthetic Transferability)**.** There exists $\mathbf{w}^* = (w_1^*, ..., w_K^*) \in \Delta_{K-1}$ such that

$$m_0\left(\mu_0^\Gamma(x), x\right) = \sum_{k=1}^K m_k\left(\mu_k^\Gamma(x), x\right) w_k^* \tag{8}$$

for all $x \in \mathcal{X}_0$.

The synthetic transferability condition is weaker than the condition in (6) in the sense that none of the source post-policy ARFs is required to be identical to that in the target population.

The major distinction between the target population and the source population is that, unlike the target population, each source population has experienced a policy Γ_k such that $m_k(\mu_k^\Gamma(x), x)$ is identified for all $x \in \mathcal{X}_0$. We state this condition formally below.

Assumption 2.3 (Rich Support)**.** For all $k = 1, ..., K$ and all $x \in \mathcal{X}_0$, $m_k(\mu_k^\Gamma(x), x)$ is identified.

Assumption 2.3 requires that the post-policy ARF $m_k(\mu_k^\Gamma(x), x)$ is identified on the support of X_i in the target population. One can view Assumption 2.3 as an "eligibility condition" for any population to serve as a source population for the prediction problem.

Let us introduce the last condition for the source populations. It requires that these populations are not redundant in an appropriate sense. Define

$$\mathbf{m}(x) = [m_1(\mu_1^\Gamma(x), x), ..., m_K(\mu_K^\Gamma(x), x)]^\top,$$

and let

$$H = \int_{\mathcal{X}_0^\Gamma} \mathbf{m}(x)\mathbf{m}(x)^\top dP_0(x).$$

Assumption 2.4. H is invertible.

Assumption 2.4 requires that the post-policy ARFs, $m_k(\mu_k^\Gamma(\cdot), \cdot)$, $k = 1, ..., K$, be linearly independent on $\mathcal{X}_0^\Gamma$. This assumption is used to point-identify the weight $\mathbf{w}^*$ in the Synthetic Transferability assumption. This assumption can be removed with a minor modification of our proposal. Thus, in the Supplemental Note, we present a modified inference procedure which does not require the point-identification of $\mathbf{w}^*$ and thus Assumption 2.4.

2.3.2. **Synthetic Decomposition.** Here, we propose our main method of synthetic decomposition. For a given weight $\mathbf{w} = (w_1, ..., w_K)^\top \in \Delta_{K-1}$, we define

$$\theta(\mathbf{w}) = \int_{\mathcal{X}_0^\Gamma} m_0\left(\mu_0^\Gamma(x), x\right) dP_0(x) + \int_{\mathcal{X}_0 \setminus \mathcal{X}_0^\Gamma} m^{\text{syn}}(x; \mathbf{w})\, dP_0(x), \tag{9}$$

where

$$m^{\text{syn}}(x; \mathbf{w}) = \sum_{k=1}^{K} m_k(\mu_k^\Gamma(x), x) w_k.$$

The relevance of $\theta(\mathbf{w})$ to the original problem of prediction in the target population depends on the choice of the weight $\mathbf{w}$. Whenever $\mathbf{w}^*$ is the weight vector satisfying the synthetic transferability condition, we have

$$\theta_0 = \theta(\mathbf{w}^*).$$

In fact, under Assumptions 2.2-2.4, we can identify $\mathbf{w}^* = \mathbf{w}_0$, where

$$\mathbf{w}_0 = \arg\min_{\mathbf{w}\in\Delta_{K-1}} \rho^2(\mathbf{w}), \tag{10}$$

and

$$\rho^2(\mathbf{w}) = \int_{\mathcal{X}_0^\Gamma} \left(m^{\text{syn}}(x;\mathbf{w}) - m_0\left(\mu_0^\Gamma(x), x\right)\right)^2 dP_0(x). \tag{11}$$

Hence, the weight $\mathbf{w}_0$ brings the synthetic ARF $m^{\text{syn}}(x;\mathbf{w})$ as close as possible to the target ARF, $m_0(\mu_0^\Gamma(x), x)$, for $x \in \mathcal{X}_0^\Gamma$. The integral in the pseudo-distance ρ is taken only on the matched group. Therefore, both $m^{\text{syn}}(x;\mathbf{w})$ and $m_0\left(\mu_0^\Gamma(x), x\right)$ are identified on the matched group.

From this $\mathbf{w}_0$, we obtain the identification of θ_0 as follows.

Theorem 2.1. *Suppose that Assumptions 2.2-2.4 hold. Then, θ_0 is identified as $\theta(\mathbf{w}_0)$.*

When the synthetic transferability condition fails, the prediction $\theta(\mathbf{w}_0)$ is still derived from the weighted average of the outcome-policy relationships which is chosen to be *as close as possible to meeting the synthetic transferability condition*, based on their predictive performance on the pre-policy support of X_i in the target population. Naturally, those source populations with the outcome-policy relationships most similar to that of the target population on the matched group receive a highest weight by design.

To see the role of the rich support condition in Assumption 2.3, we decompose it into the following two conditions:

(a) For all $k = 1, ..., K$ and all $x \in \mathcal{X}_0^\Gamma$, $m_k(\mu_k^\Gamma(x), x)$ is identified.
(b) For all $k = 1, ..., K$ and all $x \in \mathcal{X}_0 \setminus \mathcal{X}_0^\Gamma$, $m_k(\mu_k^\Gamma(x), x)$ is identified.

Condition (a) is used to identify the objective function $\rho(\mathbf{w})$ in (10), so that we obtain $\mathbf{w}_0$. Condition (b) is used to identify $\theta(\mathbf{w})$ defined in (9).

To see the role of the invertibility condition in Assumption 2.4, first define

$$\mathbf{h} = \int_{\mathcal{X}_0^\Gamma} \mathbf{m}(x) m_0\left(\mu_0^\Gamma(x), x\right) dP_0(x).$$

Then, it is not hard to see that we can rewrite

$$\mathbf{w}_0 = \arg\min_{\mathbf{w}\in\Delta_{K-1}} \left(\mathbf{w} - H^{-1}\mathbf{h}\right)^\top H \left(\mathbf{w} - H^{-1}\mathbf{h}\right). \tag{12}$$

We can see that the solution $\mathbf{w}_0$ defined in (10) is unique.

The crucial identifying restriction for the synthetic decomposition method is that the weight $\mathbf{w}^*$ in the synthetic transferability hypothesis remains the same regardless of whether $x \in \mathcal{X}_0^\Gamma$

and $x \in \mathcal{X}_0 \setminus \mathcal{X}_0^\Gamma$. Later, we relax the synthetic transferability assumption so that $\mathbf{w}^*$ is allowed to depend on a subvector of X_i that is not part of the policy variable.[11]

2.3.3. **Synthetic Decomposition and Linear Extrapolation.** We can view the synthetic decomposition as a form of extrapolation from multiple source populations to a target population. In particular, when (i) the outcome-policy relationships follow a linear regression model and (ii) the synthetic transferability condition holds, the synthetic decomposition coincides with parametric extrapolation.

To see this, consider the simple model in Example 2.1 and suppose that the reduced forms follow the same linear regression specification,

$$Y_i = X_i^\top \beta_0 + U_i \text{ and } Y_i = X_i^\top \beta_k + U_i, \quad k = 1, ..., K,$$

where β_0 belongs to the convex hull of β_k's. This latter condition implies the synthetic transferability condition as we have

$$\sum_{k=1}^{K} w_k^* \beta_k = \beta_0, \tag{13}$$

for some weight vector $\mathbf{w}^* = (w_1^*, ..., w_K^*)$. Unsurprisingly, in this case the synthetic decomposition method chooses a weighted average of the outcome-policy relationships and yields a prediction that coincides with the one from a linear extrapolation in the target population.

2.4. Examples

We now provide two applied examples that fit our framework.

2.4.1. **Minimum Wages and Labor Supply.** Minimum wages are one of the most prevalent and widely debated policies for the labor market. When studying the effects of counterfactual raises in minimum wages on employment, rather than past raises, the literature often uses search-and-bargaining models (e.g., Flinn (2006); Ahn, Arcidiacono, and Wessels (2011); Flinn and Mullins (2015)). In Section 4, we carefully rewrite the model of Ahn, Arcidiacono, and Wessels (2011) to fit our framework. We give a brief overview here.

Let $Y_{i,j} \in \{0, 1\}$ denote the employment status of worker i after a match with firm j, X_i as worker i's observable characteristics (age, race, etc.), $\underline{W}_k$ as the prevailing minimum wage in region k that i is subject to, and $U_{i,j}$ as a match-specific unobservable (shocks, unobserved

[11] This identifying strategy is precisely the way synthetic control methods identify the counterfactual non-treatment outcome of the treated unit as a weighted average of the outcomes in the donor pool of control units. The underlying assumption here is that the weights obtained from the pre-treatment outcomes remain valid after the treatment.

types) drawn from a CDF F_k. As we explain in Section 4, the wage generation in Ahn, Arcidiacono, and Wessels (2011) can be written as:

$$W_{i,j} = \max\{\beta_k M_{i,j}, \underline{W}_k\}, \tag{14}$$

where $\beta_k \in (0,1)$ is a parameter that represents the worker i's bargaining strength in region k, M_{ij} is the match productivity drawn for worker i with firm j. We parameterize the generation of $M_{i,j}$ as follows:

$$\log M_{i,j} = X_i^\top \gamma_k + U_{i,j}.$$

The employment indicator, $Y_{i,j}$, equals one if the match surplus is higher than the wage:

$$Y_{i,j} = 1\{M_{i,j} \geq W_{i,j}\} = 1\{M_{i,j} \geq \underline{W}_k\} = 1\{X_i^\top \gamma_k + U_{i,j} \geq \log(\underline{W}_k)\} \tag{15}$$

where the first equality follows from (14) and $\beta_k \in (0,1)$ and the second one after applying logs.

Now, suppose that minimum wages in Texas increase to US$9 and we want to predict its effects on employment. Then, Texas would be the target region 0. In order to apply the synthetic decomposition method, we first set the policy component for region k as

$$\mu_k(X_i) = X_i^\top \gamma_k - \log(\underline{W}_k).$$

The counterfactual policy sets $\mu_0^\Gamma(X_i) = X_i^\top \gamma_k - \log(9)$. States that have had the policy variable $\mu_k(X_i)$ overlapping with $\mu_0^\Gamma(X_i)$ (e.g., California, Washington) would be source regions. Hence, it follows from (15) that:

$$g_k(\overline{\mu}, u) = 1\{\overline{\mu} + u \geq 0\}.$$

The ARF for state k is identified as

$$m_k(\overline{\mu}, x) = 1 - F_k(-\overline{\mu}). \tag{16}$$

The ARF is identified as the share of workers in state k whose productivity is higher than the minimum wage in state k. As we show in Section 4, the policy component parameter γ_k is identified from a semiparametric censored regression of wages:

$$W_{i,j} = \max\{\log \beta_k + X_i^\top \gamma_k + U_{i,j}, \underline{W}_k\}.$$

Then, we can identify the ARF m_k using (16) as $m_k(\overline{\mu}, x) = \mathbf{E}[Y_{i,k} \mid \mu_k(X_i) = \overline{\mu}]$. Note that we do not need to parametrize the distribution of $U_{i,j}$.

2.4.2. **Tax Policy and Immigration.** Changes to income tax rates may immediately affect tax revenue, but they may also change the composition of the population. For instance, high earners may choose to emigrate when facing higher taxes. This matters for welfare, as such

high earners are highly mobile and pay a large share of taxes.[12] (See Scheuer and Werning (2017) for a theoretical investigation and Moretti and Wilson (2017), Akcigit, Baslandze, and Stantcheva (2016); Kleven, Landais, and Saez (2013) and Kleven, Landais, Saez, and Schultz (2014) for evidence on the effects of *past* changes to tax policies, including Danish and Spanish reforms).

To evaluate the effects of a decrease in tax rates in country 0 (e.g., U.K.) on high earners' immigration, we could follow Kleven, Landais, and Saez (2013) and model this as a discrete choice problem. A high earner i's preference, $V_{i,k}$, for living in country k depends on the average tax rate τ_k on the wage W_i the individual would face, and is specified as follows:

$$V_{i,k} = \alpha \log(1-\tau_k) + \alpha \log(W_i) + Z_i^\top \beta_k + U_{i,k}. \tag{17}$$

The first two terms represent the (concave) preferences over net-of-tax wages, $Z_i'\beta_k$ captures heterogeneity of worker preferences for each country (which may depend on age, nationality, etc.), with Z_i denoting the observed characteristics of the individual i, and $U_{i,k}$ represents an idiosyncratic Extreme Value Type 1 shock which is i.i.d. across individuals and regions. (Note that for high earners, the average tax rate is approximately equal to the marginal tax rate which is the same across all the high earners.) Then, individual i's decision to live in region k is represented by a binary indicator Y_i as follows:

$$Y_i = 1\left\{V_{i,k} > \max_{j\neq k} V_{i,j}\right\}. \tag{18}$$

To apply the synthetic decomposition method, we take the policy component for the individual as: for each $k = 0, 1, ..., K$, and for each individual i in region k with $X_i = (W_i, Z_i)$,

$$\mu_k(X_i) = \alpha \log(1-\tau_k) + \alpha \log(W_i) + Z_i^\top \beta_k.$$

The target country is the U.K., and the policy of interest is lowering tax rates in the U.K. so that

$$\mu_0^\Gamma(X_i) = \alpha \log(1-\tau_0^\Gamma) + \alpha \log(W_i) + Z_i^\top \beta_0,$$

for $\tau_0^\Gamma < \tau_0$. Therefore, the ARF for country k is identified as

$$m_k(\overline{\mu}, x) = \frac{\exp(\overline{\mu})}{\exp(\overline{\mu}) + \sum_{j\neq k} \exp(\mu_j(x))}.$$

The matched group is constructed as in (5) using the definitions of $\mu_0(x)$ and $\mu_0^\Gamma(x)$ above.

[12] In 2016, the top 1% of households in the U.S. earned 16% of the total income while paying 25% of all federal taxes. However, their income, accumulated wealth and favorable immigration policies permit straightforward changes to residence status, making them very responsive to tax policy. See https://doc-research.org/2019/01/global-mobility-wealthy-push-pull-factors/ for a policy overview.

2.5. Extensions

2.5.1. **Policy Prediction with Observed Time-Varying Aggregated Variables.** In many empirical applications, we observe individuals over multiple time periods (either in panel data or in rotational cross-sectional data) and aggregate variables that affect individual outcomes. The aggregate variables may represent regional economic states and often contain a policy variable such as a change in the minimum wages or taxes. Here we show how our method applies to this case.

First, consider the generation of outcomes for populations $k = 0, 1, ..., K$ in periods $t = 1, ..., T$:

$$Y_{it} = g_k(\mu_k(X_{it}; z_{k,t}, v_{k,t}), z_{k,t}, U_{it}), \quad i \in N_k.$$

Here X_{it} denotes a vector of individual covariates and $v_{k,t}$ and $z_{k,t}$ the vectors of time-varying observed aggregate variables for population k. The policy sets the vector $v_{k,t}$ to v_0^*. The policy component $\mu_k(X_{it}; v_{k,t}, z_{k,t})$ is allowed to depend on the observed aggregate characteristics of region k. It is required to exhibit individual-level variations through X_{it}.

For each region $k = 0, 1, ..., K$, the ARF is written as

$$m_{k,t}(\overline{\mu}, x, z_{k,t}) = \int g_k(\overline{\mu}, z_{k,t}, u) dP_{k,t}(u \mid x, z_{k,t}),$$

for each $(\overline{\mu}, x)$ on the support of $(\mu_k(X_{it}; z_{k,t}, v_{k,t}), X_{it})$, where $P_{k,t}(\cdot \mid x, z)$ denotes the conditional distribution of U_{it} given $X_{it} = x$ and $z_{k,t} = z$, for $i \in N_k$. We assume that there are not many time periods in the sample, and hence, any aggregate observed variables are regarded as "observed constants."

Our main interest is in predicting the average outcome of Y_{it} for population 0, when the policy changes $v_{0,t}$ into v_0^*. Then, our target parameter is defined as

$$\theta_{0,t}(v_0^*) = \int m_{0,t}\left(\mu_0(x; z_{0,t}, v_0^*), x, z_{0,t}\right) dP_{0,t}(x),$$

where $P_{0,t}$ denotes the distribution of X_{it} in the target population. The quantity $\theta_{0,t}(v_0^*)$ represents the average outcome when the distribution of $X_{i,t}$ and the value of the aggregate variables $z_{0,t}$ are fixed at those at time t, and the policy changes the variable $v_{0,t}$ into the counterfactual one v_0^*.

Let us see how our method applies in this setting. First, we take the matched group as

$$\mathcal{X}_{0,t}^{\Gamma} = \left\{x \in \mathcal{X}_{0,t} : \mu_0(x; z_{0,t}, v_0^*) = \mu_0(\tilde{x}; z_{0,t}, v_{0,t}), \text{ for some } \tilde{x} \in \mathcal{X}_{0,t}\right\},$$

where $\mathcal{X}_{0,t}$ denotes the support of X_{it} in the target population. The synthetic transferability condition is given as follows: there exists a weight vector $\mathbf{w}^*(v_0^*)$ such that for each $x \in \mathcal{X}_{0,t}$,

we have

$$m_{0,t}\left(\mu_0(x;z_{0,t},v_0^*),x,z_{0,t}\right)=\sum_{k=1}^{K}m_{k,t}\left(\mu_{k,t}(x;z_{k,t},v_{k,t}),x,z_{k,t}\right)w_k^*(v_0^*).$$

(We make it explicit that the weight depends on the value of v_0^*.) We can identify the weights by minimizing $\rho_t(\mathbf{w})$ over $\mathbf{w}$, where

$$\rho_t^2(\mathbf{w})=\int\left(M_{0,t}(x;z_{0,t},v_0^*)-\sum_{k=1}^{K}M_{k,t}(x;z_{k,t},v_{k,t})w_k\right)^2 dP_{0,t}(x),$$

where for $k=0,1,...,K$ and $t=1,...,T$, we define

$$M_{k,t}(x;z,v)=m_{k,t}(\mu_k(x;z,v),x,z)1\{x\in\mathcal{X}_{k,t}\},$$

with $\mathcal{X}_{k,t}$ denoting the support of X_{it} for $i\in N_k$. Let $\mathbf{w}_0(v_0^*)$ be the minimizer of $\rho_t(\mathbf{w})$ over $\mathbf{w}\in\Delta_{K-1}$. Using the weight $\mathbf{w}_0(v_0^*)$, we can identify θ_0 as follows:

$$\begin{aligned}\theta_{0,t}(v_0^*)=&\int_{\mathcal{X}_{0,t}^{\Gamma}}m_{0,t}(\mu_0(x;z_{0,t},v_0^*),x,z_{0,t})dP_{0,t}(x)\\&+\sum_{k=1}^{K}\left(\int_{\mathcal{X}_{0,t}\setminus\mathcal{X}_{0,t}^{\Gamma}}m_{k,t}(\mu_k(x;z_{k,t},v_{k,t}),x,z_{k,t})dP_{0,t}(x)\right)w_{0,k}(v_0^*).\end{aligned}\tag{19}$$

The average effect of changing v_0 from v_0' to v_0^* is given by

$$\theta_{0,t}(v_0^*)-\theta_{0,t}(v_0').$$

When v_0' is chosen to be $v_{0,t}$ in the target region, we can obtain $\theta_{0,t}(v_{0,t})$ in two different ways. The first way is to obtain $\theta_{0,t}(v_{0,t})$ by replacing v_0^* with $v_{0,t}$ in (19) and the second way is to obtain

$$\theta_{0,t}(v_{0,t})=\int_{\mathcal{X}_{0,t}}m_{0,t}(\mu_0(x;z_{0,t},v_0'),x,z_{0,t})dP_{0,t}(x).$$

If the estimates of $\theta_{0,t}(v_{0,t})$ obtained through two different ways are close to each other, this suggests that the synthetic transferability condition is supported by data.

2.5.2. **Spillover of Policy Effects Across Regions.** A policy in one region can often have a spillover effect on other regions. For example, the immigration of high-skilled workers in response to a change in tax policy in a target region (as in Kleven, Landais, and Saez (2013)) would affect the number of immigrants in source regions. We show that the situation with spillover effects can be accommodated in our proposal.

We consider two types of spillover effects. The first is the spillover effect of policies from the source regions on the target region. The spillover effect is something that has already

happened and is reflected in the data at the time when the policymaker considers implementing a new policy on the target population. For instance, source countries with lower taxes have already received high-skilled immigrants from the target region. The spillover effect of source regions' policies on the target region's outcomes realizes through its impact on the exogenous variables X_i and U_i, as prescribed in the reduced form in (1). As such, the spillover effect is entirely mediated through the variations in X_i, and its presence does not alter anything in our proposal, because it is among the many sources of exogenous variations in X_i which we can use to identify the ARF. On the other hand, if the spillover effect is an additional source of endogeneity (i.e., the correlation between X_i and U_i), we need to carefully search for an identification method using instrumental variables or resorting to a control function approach (e.g., Blundell and Matzkin (2014)). Once the ARF is identified, this paper's proposal can be applied.

The second spillover effect is from the policy of the target population to other regions. This is a spillover effect that is not yet reflected in the data, and hence part of our counterfactual analysis of policy in the target population. For example, a decrease in tax rates in the target region (say, the U.K.) would induce immigration away from source regions (e.g., Spain). Our definition of the pre-policy population will then be the population that consists of people before the migration induced by the policy, and likewise the post-policy population will be the population that consists of people after the migration. Therefore, the policy effect, according to our definition, includes both the effect on the people who do not migrate as a consequence and the composition effect that arises due to the migration.

For example, suppose that the policy not only changes μ_0 into μ_0^Γ, but also alters the distribution P_0 into $P_0 \circ f^{-1}$ for some map f. The latter change corresponds to changing X_i into $f(X_i)$. Now, the post-policy prediction includes both the effects, so that we can take

$$\begin{aligned}\theta_0 &= \int_{f(\mathcal{X}_0)} m_0\left(\mu_0^\Gamma(x), x\right) d(P_0 \circ f^{-1})(x) \\ &= \int_{\mathcal{X}_0} m_0\left(\mu_0^\Gamma(f(x), f(x))\right) dP_0(x).\end{aligned}$$

Hence, by redefining the policy operator, Γ, we can study the effect of a policy that has a spillover effect through migration.[13]

2.5.3. **Covariate-Dependent Weights.** The synthetic transferability condition assumes that the weights are the same across different demographic groups, and this may be restrictive in some applications. For example, suppose that we have two source regions 1 and 2, where a

[13]However, in contrast to the previous situations, we may need to estimate the "policy" Γ as it includes its composition effect through migration from the target region.

high education group in region 1 is matched better with a high education group in the target region than region 2, whereas a low education group in region 2 is matched better with a low education group in the target region than in region 1. By allowing the weight to depend on the education indicator, we can accommodate such a situation flexibly.

Suppose that $X_i = (X_{i,1}, X_{i,2})$, where we denote the supports of $X_{i,1}$ and $X_{i,2}$ in regions $k = 0, 1, ..., K$ by $\mathcal{X}_{k,1}$ and $\mathcal{X}_{k,2}$ respectively. For each individual i who belongs to population $k = 0, 1, ..., K$, we have

$$Y_i = g_k(\mu_k(X_{i,1}), X_{i,2}, U_i), \text{ before the policy}$$
$$Y_i = g_k(\mu_k^\Gamma(X_{i,1}), X_{i,2}, U_i), \text{ after the policy.}$$

We define a generalized version of the ***synthetic ARF***: for a subvector $\tilde{x}_2$ of x_2, with $x = (x_1, x_2)$,

$$m^{\text{syn}}(x; \mathbf{w}) = \sum_{k=1}^{K} m_k\left(\mu_k^\Gamma(x_1), x_2\right) w_k(\tilde{x}_2),$$

where $w_k(\tilde{x}_2)$ is a nonnegative function such that $\sum_{k=1}^K w_k(\tilde{x}_2) = 1$, and $w_k(\tilde{x}_2)$ is the k-th entry of $\mathbf{w}(\tilde{x}_2)$. We denote by $\tilde{\mathcal{X}}_{0,2}$ the support of the corresponding subvector $\tilde{X}_{i,2}$ of $X_{i,2}$ in the target population. In contrast to the previous case, the weight given to a source region k can vary across different people in the target region depending on the value of their covariate $\tilde{x}_2$. Hence, $\mathbf{w}$ is a map from $\tilde{\mathcal{X}}_{0,2}$ to Δ_{K-1}. Similarly as above, we obtain a counterfactual prediction for policy Γ_0 for the target region 0 as $\theta(\mathbf{w})$ in (9) with this redefined $m^{\text{syn}}(x; \mathbf{w})$.

To motivate the problem of selecting the weight vector $\mathbf{w}$, we introduce a generalized version of the transferability condition.

Assumption 2.5 (Generalized Synthetic Transferability)**.** For some map $\mathbf{w}^* : \tilde{\mathcal{X}}_{0,2} \to \Delta_{K-1}$,

$$m^{\text{syn}}(x; \mathbf{w}^*) = m_0\left(\mu^\Gamma(x_1), x_2\right), \text{ for all } x = (x_1, x_2) \in \mathcal{X}_0.$$

The previous synthetic transferability condition is a special case of this condition. We now construct the optimal weight as follows. First, we can define

$$\mathbf{w}_0 = \arg \inf_{\mathbf{w}: \tilde{\mathcal{X}}_{0,2} \to \Delta_{K-1}} \rho^2(\mathbf{w}),$$

and

$$\rho^2(\mathbf{w}) = \int_{\mathcal{X}_0^\Gamma} \left(\mathbf{m}(x)^\top \mathbf{w}(\tilde{x}_2) - m_0\left(\mu_0^\Gamma(x_1), x_2\right)\right)^2 dP_0(x), \tag{20}$$

where $\mathbf{m}(x) = [m_1(\mu_1^\Gamma(x_1), x_2), ..., m_K(\mu_K^\Gamma(x_1), x_2)]^\top$. The quantity $\rho(\mathbf{w}_0)$ is smaller than that in (11), because the domain of the minimizers $\mathbf{w}$ is larger now. This means that the covariate-dependent weight will exhibit a better fit than the previous weights.

One can obtain a prediction for the target population as $\theta(\mathbf{w}_0)$ using this weight, $\mathbf{w}_0(\tilde{x}_2)$. To obtain a characterization of this generalized weight, we first define

$$H(\tilde{x}_2) = \int_{\mathcal{X}_0^\Gamma} \mathbf{m}(x)\mathbf{m}(x)^\top dP_0(x \mid \tilde{x}_2), \text{ and}$$
$$\mathbf{h}(\tilde{x}_2) = \int_{\mathcal{X}_0^\Gamma} \mathbf{m}(x) m_0(\mu_0^\Gamma(x_1), x_2) dP_0(x \mid \tilde{x}_2),$$

where $P_0(\cdot \mid \tilde{x}_2)$ denotes the conditional distribution of X_i given $\tilde{X}_{i,2} = \tilde{x}_2$ in the target population. We replace Assumption 2.4 by the following assumption.

Assumption 2.6. For each $\tilde{x}_2 \in \tilde{\mathcal{X}}_{0,2}$, $H(\tilde{x}_2)$ is invertible.

This assumption excludes the situation where $X_{i,1} = X_i$. In other words, there must be variables in X_i that are excluded from the covariate $X_{i,1}$ that the weight is allowed to depend on. Then, it is not hard to see that

$$\mathbf{w}_0(\tilde{x}_2) = \arg \inf_{\mathbf{w} \in \Delta_{K-1}} \left(\mathbf{w} - H^{-1}(\tilde{x}_2)\mathbf{h}(\tilde{x}_2)\right)^\top H(\tilde{x}_2) \left(\mathbf{w} - H^{-1}(\tilde{x}_2)\mathbf{h}(\tilde{x}_2)\right).$$

Again, under Assumption 2.6, $\mathbf{w}_0$ is uniquely determined. Hence under the generalized synthetic transferability condition, we have $\theta_0 = \theta(\mathbf{w}_0)$. This approach is not very computationally costly when $\tilde{X}_{i,2}$ is a discrete random variable with its support having only a few points.

3. Estimation and Confidence Intervals

3.1. Estimation

Let us first consider the estimation of $\mathbf{w}_0$ and $\theta(\mathbf{w}_0)$. As for the estimation of $\mathbf{w}_0$, we first make use of the characterization (12) and consider its sample counterpart. For each $k = 1, \ldots, K$, we first estimate the post-policy ARFs to obtain $\hat{m}_k\left(\hat{\mu}_k^\Gamma(x), x\right)$ using the sample from the source population P_k. We let $\hat{\mathcal{X}}_0^\Gamma$ be an estimated set of $\mathcal{X}_0^\Gamma$ using the sample from the target population and construct the sample version of H and $\mathbf{h}$ as follows:

$$\hat{H} = \frac{1}{n_0} \sum_{i \in N_0} \hat{\mathbf{m}}(X_i)\hat{\mathbf{m}}(X_i)^\top 1\left\{X_i \in \hat{\mathcal{X}}_0^\Gamma\right\}, \text{ and} \tag{21}$$
$$\hat{\mathbf{h}} = \frac{1}{n_0} \sum_{i \in N_0} \hat{\mathbf{m}}(X_i)\hat{m}_0\left(\hat{\mu}_0^\Gamma(X_i), X_i\right) 1\left\{X_i \in \hat{\mathcal{X}}_0^\Gamma\right\},$$

where $\hat{\mathbf{m}}(x) = [\hat{m}_1(\hat{\mu}_1^\Gamma(x), x), \ldots, \hat{m}_K(\hat{\mu}_K^\Gamma(x), x)]^\top$. Note that each ARF, $\hat{m}_k(\hat{\mu}_k^\Gamma(\cdot), \cdot)$, is constructed using the sample from the *source* region k, whereas in constructing $\hat{H}$ and $\hat{\mathbf{h}}$, it is

evaluated at a data point X_i of the sample from the *target* region. Using these, we obtain

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\Delta_{K-1}} \left(\mathbf{w} - \hat{H}^{-1}\hat{\mathbf{h}}\right)^\top \hat{H}\left(\mathbf{w} - \hat{H}^{-1}\hat{\mathbf{h}}\right).$$

In the Supplemental Note, we show that $\hat{\mathbf{w}}$ is $\sqrt{n_0}$-consistent for $\mathbf{w}_0$.[14]

Using the estimated weight, $\hat{\mathbf{w}}$, we obtain the prediction for the target region as follows:

$$\hat{\theta}(\hat{\mathbf{w}}) = \frac{1}{n_0}\sum_{i\in N_0} \hat{m}_0\left(\hat{\mu}_0^\Gamma(X_i), X_i\right) 1\{X_i \in \hat{\mathcal{X}}_0^\Gamma\} + \frac{1}{n_0}\sum_{i\in N_0} \hat{m}^{\text{syn}}(X_i;\hat{\mathbf{w}})\, 1\{X_i \in \mathcal{X}_0 \setminus \hat{\mathcal{X}}_0^\Gamma\}, \tag{22}$$

where

$$\hat{m}^{\text{syn}}(x;\hat{\mathbf{w}}) = \sum_{k=1}^{K} m_k\left(\hat{\mu}_k^\Gamma(x), x\right)\hat{w}_k.$$

We will show below that, under regularity conditions, the estimator $\hat{\theta}(\hat{\mathbf{w}})$ is $\sqrt{n_0}$-consistent.

3.2. Confidence Set for $\mathbf{w}_0$

Let us consider constructing confidence intervals for $\theta(\mathbf{w}_0)$. Since $\mathbf{w}_0$ can take a value arbitrarily close to the boundary of the simplex Δ_{K-1}, it turns out that

$$\sqrt{n_0}(\hat{\theta}(\hat{\mathbf{w}}) - \theta(\hat{\mathbf{w}})) \to_d \zeta, \tag{23}$$

where ζ is a complicated non-Gaussian distribution that depends on whether $\mathbf{w}_0$ is in the interior of Δ_{K-1} or on the boundary of Δ_{K-1}. And if it is on the boundary, what part of the boundary $\mathbf{w}_0$ is located in. While one might consider using a naive bootstrap where one first constructs a bootstrap counterpart of the quantity $\sqrt{n_0}(\hat{\theta}(\hat{\mathbf{w}}_0) - \theta(\hat{\mathbf{w}}))$ and uses its bootstrap distribution in place of the asymptotic distribution, this approach does not work. Such a failure of the bootstrap when the parameter is on the boundary was shown by Andrews (2000).

In this paper, we pursue an approach that does not require the researcher to find the details of the limiting distribution ζ, which may change depending on the specifications of the models and estimation methods. To construct the confidence set, we first formulate the identification of $\mathbf{w}_0$ as a solution to a constrained optimization, and using a Kuhn-Tucker condition, formulate the identification in terms of a set of equality restrictions with a nuisance parameter that is constrained to a convex cone. The problem of asymptotic inference in such a setting has been studied in the literature (see, for example, Rosen (2008), Moon and Schorfheide (2009), Kitamura and Stoye (2018)). Here, we follow the approach of Mohamad, van Zwet, Cator, and Goeman (2020) and Cox and Shi (2022) to construct a test statistic for the restrictions

[14] As we formally state later, we assume that the size of a random sample from each population is asymptotically comparable across the populations, i.e., there exists $r_k > 0$ such that $n_k/n_0 \to r_k$ as $n_0, n_k \to \infty$ for each $k = 1, ..., K$.

and invert it to form a confidence set for $\mathbf{w}_0$. This approach is simple and does not involve tuning parameters often required in the problem of testing for inequality restrictions. Later, we show that this confidence set is uniformly asymptotically valid. This approach is generally applicable whenever we have $\sqrt{n_0}$-consistent estimators of H and $\mathbf{h}$. This often follows from a wide range of estimators of the ARFs.

First, let us construct a confidence set for $\mathbf{w}_0$. For this, we form an equality restriction using the Lagrangian of the constrained optimization in (12):

$$\mathcal{L}(\mathbf{w}, \tilde{\lambda}, \boldsymbol{\lambda}) = (\mathbf{w} - H^{-1}\mathbf{h})^\top H(\mathbf{w} - H^{-1}\mathbf{h}) + \tilde{\lambda}(1 - \mathbf{w}^\top \mathbf{1}) - \boldsymbol{\lambda}^\top \mathbf{w},$$

where $\tilde{\lambda}$ and $\boldsymbol{\lambda}$ are Lagrange multipliers.

By the Kuhn-Tucker condition and the strict convexity of the objective function, the necessary and sufficient conditions for $\mathbf{w}_0 \in \Delta_{K-1}$ to be the unique minimizer of $\rho(\mathbf{w})$ are that for some $\tilde{\lambda} \in \mathbf{R}$ and $\boldsymbol{\lambda} \in \Lambda(\mathbf{w}_0)$,

$$H\mathbf{w}_0 - \mathbf{h} + \tilde{\lambda}\mathbf{1} - \boldsymbol{\lambda} = \mathbf{0}, \tag{24}$$

where $\mathbf{1}$ is the $K \times 1$ vector of ones, $\mathbf{0}$ is the $K \times 1$ vector of zeros, and

$$\Lambda(\mathbf{w}_0) = \left\{\boldsymbol{\lambda} \in \mathbf{R}^K : \boldsymbol{\lambda}^\top \mathbf{w}_0 = 0 \text{ and } \boldsymbol{\lambda} \leq \mathbf{0}\right\}. \tag{25}$$

If we concentrate out $\tilde{\lambda}$ using the restrictions $\mathbf{w}_0^\top \mathbf{1} = 1$ and $\boldsymbol{\lambda}'\mathbf{w}_0 = 0$, we obtain that

$$\mathbf{f}(\mathbf{w}_0) - \boldsymbol{\lambda} = \mathbf{0}, \text{ for some } \boldsymbol{\lambda} \in \Lambda(\mathbf{w}_0), \tag{26}$$

where $\mathbf{f}(\mathbf{w}_0) = H\mathbf{w}_0 - \mathbf{h} - \mathbf{w}_0^\top(H\mathbf{w}_0 - \mathbf{h})\mathbf{1}$. We form a test statistic that tests the restriction in (26) as follows:

$$T(\mathbf{w}_0) = n_0 \inf_{\boldsymbol{\lambda} \in \Lambda(\mathbf{w}_0)} \left(\hat{\mathbf{f}}(\mathbf{w}_0) - \boldsymbol{\lambda}\right)^\top \hat{\Omega}^{-1} \left(\hat{\mathbf{f}}(\mathbf{w}_0) - \boldsymbol{\lambda}\right), \tag{27}$$

where $\hat{\mathbf{f}}(\mathbf{w}_0)$ is the same as $\mathbf{f}(\mathbf{w}_0)$ except that H and $\mathbf{h}$ are replaced by $\hat{H}$ and $\hat{\mathbf{h}}$, and $\hat{\Omega}$ is a scale normalizer which we explain later.[15]

As for critical values, we follow the approach of Mohamad, van Zwet, Cator, and Goeman (2020) and Cox and Shi (2022). First, for each $\mathbf{w} \in \Delta_{K-1}$, we let $\hat{\boldsymbol{\lambda}}(\mathbf{w})$ be the solution $\boldsymbol{\lambda}$ in the minimization problem in (27) with $\mathbf{w}_0$ replaced by the generic $\mathbf{w} \in \Delta_{K-1}$. Then the confidence set for $\mathbf{w}_0$ is given by

$$\tilde{C}_{1-\kappa} = \{\mathbf{w} \in \Delta_{K-1} : T(\mathbf{w}) \leq \hat{c}_{1-\kappa}(\mathbf{w})\}, \tag{28}$$

[15] The method of constructing a test statistic from a constrained optimization over Lagrangian multipliers appeared in Moon and Schorfheide (2009). The main difference here is that in our case, the inequality restrictions are crucial for the point-identification of $\mathbf{w}_0$, whereas, in their case, the parameters are point-identified using only equality restrictions, and hence their use of quadratic approximation for constructing a critical value does not apply in our setting.

where $\hat{c}_{1-\kappa}(\mathbf{w})$ denotes the $1-\kappa$ percentile of the χ^2 distribution with the degree of freedom equal to the number of zero entries in $\hat{\boldsymbol{\lambda}}(\mathbf{w})$. The test $1\{T(\mathbf{w}) \leq \hat{c}_{1-\kappa}(\mathbf{w})\}$ is essentially what Cox and Shi (2022) called the CC test in their paper. The main difference is that $\mathbf{f}(\mathbf{w}_0)$ is not necessarily the expectation of a random vector in our setting. Otherwise, our setting is much simpler than Cox and Shi (2022) because the inequality restrictions (as represented by the constraint $\lambda \in \Lambda(\mathbf{w}_0)$) do not involve any unknowns.

We may be interested in checking whether data support the synthetic transferability condition in (28). The confidence set $\tilde{C}_{1-\kappa}$ for $\mathbf{w}_0$ defined in (28) can be used to test an implication from the condition. Consider testing the following implication from the synthetic transferability condition:

$$H_0: \text{ There exists } \mathbf{w} \in \Delta_{K-1} \text{ such that } m_0(\mu_0^\Gamma(x), x) = m^{\text{syn}}(x; \mathbf{w}) \text{ for all } x \in \mathcal{X}_0^\Gamma.$$

$$H_1: H_0 \text{ is false.}$$

We set κ in $\tilde{C}_{1-\alpha}$ to be the level of the test and perform the following procedure. If $\tilde{C}_{1-\alpha} = \varnothing$, we reject H_0 at level α. Otherwise, we do not reject H_0 at level α.

Now, let us discuss the bootstrap construction of $\hat{\Omega}$. First, we construct a bootstrap sample as follows. Since each population has a different distribution, we need to resample (with replacement) from each region. For each region $k = 0, 1, ..., K$, let $\{W_i^* : i \in N_k\}$ be the bootstrap sample from the sample $\{W_i : i \in N_k\}$, where $W_i = (Y_i, X_i')^\top$, $i \in N$, and

$$N = \bigcup_{k=0}^{K} N_k.$$

Then for each $k = 0, 1, ..., K$, we construct the bootstrap version of the extended post-policy ARF, $\hat{m}_k^*(\hat{\mu}_k^{\Gamma*}(\cdot), \cdot)$, using the bootstrap sample from the region k, and define

$$\hat{\mathbf{m}}^*(\cdot) = \left[\hat{m}_1^*\left(\hat{\mu}_1^{\Gamma*}(\cdot), \cdot\right), ..., \hat{m}_K^*\left(\hat{\mu}_K^{\Gamma*}(\cdot), \cdot\right)\right]^\top,$$

and let

$$\hat{H}^* = \frac{1}{n_0} \sum_{i \in N_0} \hat{\mathbf{m}}^*(X_i^*) \hat{\mathbf{m}}^*(X_i^*)^\top 1\left\{X_i^* \in \hat{\mathcal{X}}_0^{\Gamma*}\right\}, \text{ and}$$

$$\hat{\mathbf{h}}^* = \frac{1}{n_0} \sum_{i \in N_0} \hat{\mathbf{m}}^*(X_i^*) \hat{m}_0^*\left(\hat{\mu}_0^{\Gamma*}(X_i^*), X_i^*\right) 1\left\{X_i^* \in \hat{\mathcal{X}}_0^{\Gamma*}\right\}.$$

Then, we define

$$\begin{aligned}\hat{\boldsymbol{\gamma}}^* = &\sqrt{n_0}\left(\hat{H}^* - \hat{H}\right)\hat{\mathbf{w}} - \sqrt{n_0}\left(\hat{\mathbf{h}}^* - \hat{\mathbf{h}}\right) \\ &- \sqrt{n_0}\hat{\mathbf{w}}^\top\left(\hat{H}^* - \hat{H}\right)\mathbf{1} + \sqrt{n_0}\hat{\mathbf{w}}^\top\left(\hat{\mathbf{h}}^* - \hat{\mathbf{h}}\right)\mathbf{1}.\end{aligned}$$

To construct a scale normalizer $\hat{\Omega}$, we apply the truncation method of Shao (1992) as follows. For each $k = 1, ..., K$, we define

$$\hat{\tau}_k = \sqrt{n_0} \max\left\{ \left| \left[\hat{H}\hat{\mathbf{w}} - \hat{\mathbf{h}} \right]_k \right|, c_0 \right\},$$

for some constant $c_0 > 0$ such as $c_0 = 0.05$, and construct a truncated version of $\hat{\boldsymbol{\gamma}}^*$ as $\tilde{\boldsymbol{\gamma}}^* = [\tilde{\gamma}_k^*]_{k=1}^K$, where

$$\tilde{\gamma}_k^* = \begin{cases} \hat{\tau}_k, & \text{if } \gamma_k^* \geq \hat{\tau}_k, \\ \hat{\gamma}_k^*, & \text{if } -\tau_k \leq \hat{\gamma}_k^* \leq \hat{\tau}_k, \text{ and} \\ -\hat{\tau}_k, & \text{if } \gamma_k^* \leq -\hat{\tau}_k, \end{cases}$$

and $\hat{\gamma}_k^*$ denotes the k-th entry of $\hat{\boldsymbol{\gamma}}^*$. Thus, we construct $\tilde{\boldsymbol{\gamma}}^*$ for each bootstrap sample $b = 1, ..., B$. Let us denote it by $\tilde{\boldsymbol{\gamma}}_b^*$. Then we construct[16]

$$\hat{\Omega} = \frac{1}{B} \sum_{b=1}^{B} \tilde{\boldsymbol{\gamma}}_b^* \tilde{\boldsymbol{\gamma}}_b^{*\top} - \left(\frac{1}{B} \sum_{b=1}^{B} \tilde{\boldsymbol{\gamma}}_b^* \right) \left(\frac{1}{B} \sum_{b=1}^{B} \tilde{\boldsymbol{\gamma}}_b^* \right)^\top. \tag{29}$$

3.3. Confidence Intervals for $\theta(\mathbf{w}_0)$

Now, let us construct the confidence interval for $\theta(\mathbf{w}_0)$. First, we can show that

$$\frac{n_0 (\hat{\theta}(\mathbf{w}_0) - \theta(\mathbf{w}_0))^2}{\hat{\sigma}^2} \to_d \chi^2(1), \tag{30}$$

for an appropriate scale normalizer. To construct $\hat{\sigma}$, we use a bootstrap interquartile range as proposed by Chernozhukov, Fernández-Val, and Melly (2013) in a different context. More specifically, we define

$$\hat{\theta}^*(\mathbf{w}) = \frac{1}{n_0} \sum_{i \in N_0} \hat{m}_0^* \left(\hat{\mu}_0^{\Gamma *}(X_i^*), X_i^* \right) 1\left\{ X_i^* \in \hat{\mathcal{X}}_0^{\Gamma *} \right\} + \frac{1}{n_0} \sum_{i \in N_0} \hat{m}^{\text{syn}*}\left(X_i^*; \mathbf{w} \right) 1\left\{ X_i^* \notin \hat{\mathcal{X}}_0^{\Gamma *} \right\},$$

where

$$\hat{m}^{\text{syn}*}(x; \mathbf{w}) = \sum_{k=1}^{K} \hat{m}_k^* \left(\hat{\mu}_k^{\Gamma *}(x), x \right) w_k.$$

We let

$$T^* = \sqrt{n_0} \left(\hat{\theta}^*(\hat{\mathbf{w}}) - \hat{\theta}(\hat{\mathbf{w}}) \right).$$

[16] As pointed out by Hahn and Liao (2021), without using the truncation, the confidence set $\tilde{C}_{1-\kappa}$ is still asymptotically valid, although it is conservative. In our simulations, the truncation does not make any meaningful difference in the results.

We read the 0.75 quantile and 0.25 quantile of the bootstrap distribution of $\{T^* : b = 1, ..., B\}$, and denote them to be $\hat{q}_{0.75}$ and $\hat{q}_{0.25}$, respectively. Define

$$\hat{\sigma} = \frac{\hat{q}_{0.75} - \hat{q}_{0.25}}{z_{0.75} - z_{0.25}},$$

where $z_{0.75}$ and $z_{0.25}$ are the 0.75- and 0.25-quantiles of $N(0, 1)$.

Define

$$\hat{T}(\mathbf{w}, \theta) = \frac{\sqrt{n_0}(\hat{\theta}(\mathbf{w}) - \theta)}{\hat{\sigma}}.$$

We construct the $(1-\alpha)$-level confidence interval using the Bonferroni approach as follows:

$$C_{1-\alpha} = \left\{ \theta \in \Theta : \inf_{\mathbf{w} \in \tilde{C}_{1-\kappa}} \hat{T}^2(\mathbf{w}, \theta) \leq c_{1-\alpha+\kappa}(1) \right\}, \tag{31}$$

where $\kappa > 0$ is a small constant, such as $\kappa = 0.005$, and $c_{1-\alpha+\kappa}(1)$ denotes the $(1 - \alpha + \kappa)$-quantile of the $\chi^2(1)$ distribution.

3.4. Uniform Asymptotic Validity

We summarize the conditions that we use to establish the uniform asymptotic validity of the confidence set $C_{1-\alpha}$. Here, we state the conditions verbally. The formal statements and the proof are found in the Supplemental Note.

Assumption 3.1. (i) The post-policy ARFs in the target and source populations have the $4+\delta$-th finite moment uniformly over P.

(ii) The estimated post-policy ARFs and their bootstrap versions have an asymptotic linear representation uniform over P, with the influence function having the $4 + \delta$-th finite moment uniformly over P.

The moment condition is a technical condition that is often used in asymptotic inference. The asymptotic linear representation is often part of the proofs that show asymptotic normality of an estimator. Its derivation is standard in many examples.

Assumption 3.2. The matrix H and the population version of $\hat{\Omega}$ have minimum eigenvalues bounded from below uniformly over n and P.

This assumption requires that the post-policy ARFs are not redundant. As mentioned before, we can relax this assumption once we modify the procedure. Details are found in the Supplemental Note.

Assumption 3.3. For each $k = 0, 1, ..., K$, there exists a constant $r_k > 0$ such that $n_k/n_0 \to r_k$, as $n_0, n_k \to \infty$.

Under these conditions, we can show that the confidence interval $C_{1-\alpha}$ is asymptotically valid uniformly over P.

Theorem 3.1. *Suppose that Assumptions 3.1-3.3 hold. Then, for each $\alpha \in (0,1)$, the confidence interval $C_{1-\alpha}$ is asymptotically valid uniformly over P.*

The proof of the theorem is found in the Supplemental Note.

3.5. Monte Carlo Simulations

In the Supplemental Note, we present Monte Carlo simulations exploring the finite sample properties of our inference procedure.

We consider a total of eight exercises: two specifications for the ARF, two different amounts of overlap of the support of the policy variable between target and source regions (small, 50%, or large, 90%), and two sample sizes ($n_0 = 500, 1000$). In particular, the ARF specifications differ by: (i) having w_0 to be in the interior or on the boundary of the simplex, and (ii) different functional forms for the relationship between the outcome and the policy variable.

The results for the coverage probability and the average length of the confidence interval are shown in Table 3, while the results on the finite sample properties for the estimators $\hat{\mathbf{w}}, \hat{\theta}_0(\hat{\mathbf{w}})$ are shown in Table 4. Across specifications, inference for the target parameter (θ_0) is typically conservative, as seen in Table 3: their empirical coverage probabilities are usually above 95%, the nominal level for all sample sizes and specifications, with the coverage probabilities closer to 100% in most cases. Consistent with our asymptotic results, the average length of the confidence interval decreases as the sample size grows across specifications and inference approaches. The average length of the confidence interval is smaller when there is a larger overlap between the support of the policy variable in the target region and the source regions. In this case, there is more information from the target region that can be used for identification and estimation of our target parameter.

Finally, our estimators for θ_0 and $\mathbf{w}_0$ seem to perform very well pointwise: the Root Mean Square Error (RMSE) for $\hat{\theta}$ and $\hat{\mathbf{w}}$ are small. Furthermore, the average bias and variance of $\hat{\theta}(\hat{\mathbf{w}})$ across simulations are close to 0, suggesting that our estimator is close to the true values even with moderate sample sizes.

4. Empirical Application: Minimum Wages and Labor Supply

4.1. Background

Minimum wages have been among the most studied and debated policies for the labor market, spurring an immense literature in economics. The predominant paradigm in empirical work is to study their effects on employment or other outcomes by leveraging their state-level variation. This includes difference-in-difference designs with Two-Way Fixed Effects models (which Neumark (2019) summarizes as the workhorse approach), synthetic control (see Allegretto, Dube, Reich, and Zipperer (2017); Neumark and Wascher (2017) for extensive discussions), decomposition methods (DiNardo, Fortin, and Lemieux (1996)), cross-border comparisons (Dube, Lester, and Reich (2010)), among others.

While this literature can evaluate minimum wage increases that have already been implemented, they are by-and-large inappropriate to predict the effects of policies yet to occur, including increases in minimum wages beyond the support of historical variations. Indeed, even simple theoretical models predict highly nonlinear effects of minimum wages (e.g., Flinn (2006); Gorry and Jackson (2017)).[17] The synthetic decomposition method presented in this paper is able to address such policy questions.

As foreshadowed in Section 3, our empirical illustration studies a (counterfactual) increase in minimum wage in Texas beyond federally mandated levels and how it affects teenage employment. The focus on Texas, while an illustration, is of both academic and policy interest. Texas is the largest state in the U.S. with minimum wages set at the federal level (constant since 2009). Raising the minimum wages has also been a policy of the 2022 Democratic gubernatorial candidate. We illustrate our method by investigating the effects of an increase in minimum wages in Texas from US$7.25 to US$9.00, on teenage employment. We follow the structural labor economics literature in basing such predictions on an equilibrium search and matching model of labor markets (e.g., Flinn (2006); Flinn and Mullins (2015) and Ahn, Arcidiacono, and Wessels (2011), in particular). However, in contrast to such papers, we construct a synthetic comparison using other states beyond Texas where the policy has been observed (e.g., Oregon, Washington, etc.).

[17] This is best summarized by Neumark (2019) who writes in a recent review that, "even if one has a strong view of what the U.S. literature says about the employment effects of past minimum wage increases, this may provide much less guidance in projecting the consequences of much larger minimum wage increases than those studied in the prior literature. Predicting the effects of minimum wage increases of many dollars, based on research studying much smaller increases, is inherently risky for the usual statistical reasons. But the problem is potentially exacerbated because the reduced form estimates on which the prior literature is based may fail to capture changes in underlying behavior as high minimum wages affect a far greater share of workers." (p.294)

Our empirical application suits the synthetic decomposition method very well. There are two main sources of heterogeneity across regions. First, the population characteristics differ. For example, states are heterogeneous in workers' education, age and skill, among others, all of which may matter for the effects of minimum wages (Neumark (2019), and seen in the data below). More importantly, the causal structure g_0 for the source region could be very different than those for other states, g_k, even those from neighboring states. Intuitively, even if California and other states had similar characteristics to Texas, they may have very different labor market environments (e.g., state income taxation, different labor laws, etc.). In fact, Flinn (2002) argues that structural parameters are estimated to be very different across submarkets. The synthetic decomposition method respects such heterogeneity across regions. It assigns weights to those source states to form the best comparison units in terms of their causal structures.

4.2. A Two-Sided Search Model of Labor Markets with Minimum Wages

We follow Ahn, Arcidiacono, and Wessels (2011) and consider the following static model of two-sided matching between firms and workers. For each population $k = 0, 1, ..., K$, we let $\overline{N}_k$ be the total measure of the workers and $\overline{J}_k$ the total measure of the firms in the population k. Each worker-firm pair (i, j) is drawn, and then for each worker i, (R_i, K_i) is drawn, where R_i is the reservation wage of worker i and K_i the cost of searching for the worker i. The worker-firm pair is given the offer of matching with a contact rate $\lambda_k > 0$. The timing of the events for the worker-firm pair given the offer of the match proceeds as follows.

(1) The worker decides to search for a match with a firm. Once the worker decides to search, the worker pays the search cost K_i and receives an offer of match with a firm j with probability $\lambda_k > 0$. If the worker decides not to search for a firm, the worker receives zero payoff.

(2) The worker decides whether to accept the offer of the match or not. If the worker rejects the offer of the match, the worker receives a reservation wage R_i. If the worker accepts the offer, the worker-firm pair (i, j) jointly produces output $M_{i,j}$.

(3) Once the output $M_{i,j}$ is realized the firm and the worker enter a Nash bargaining to determine the wage, $W_{i,j}$, under the minimum wage constraint.

(4) After the wage $W_{i,j}$ is determined, the firm decides whether to retain the worker or not. If the firm retains the worker, the firm obtains the profit $M_{i,j} - W_{i,j}$ and the worker receives the wage $W_{i,j}$. If the firm does not retain the worker, the firm and the worker receive the zero payoff.

(5) After these events are completed, the econometrician observes a random sample of the workers, their employment status and wages, and their observed characteristics.

To close the model, we need to state the equilibrium constraints. First, it is profitable for worker i to accept the offer from the match with firm j if and only if

$$\mathbf{E}_k[1\{M_{i,j} \geq W_{i,j}\}W_{i,j} \mid R_i, K_i] \geq R_i, \tag{32}$$

where the conditional expectation $\mathbf{E}_k$ is with respect to the distribution in population k. Then, it is profitable for the worker to search for a job if and only if

$$\lambda_k \mathbf{E}_k[\max\{1\{M_{i,j} \geq W_{i,j}\}W_{i,j}, R_i\} \mid R_i, K_i] \geq K_i.$$

For the firm, it is profitable for it to retain the worker if and only if $Y_{i,j} \geq W_{i,j}$. Finally, we assume that the contact rate λ is endogenously determined as a fixed point as follows:

$$\lambda_k = \frac{\mathcal{M}_k(\lambda_k, \overline{J}_k, \overline{N}_k)}{\zeta_k(\lambda_k)\overline{N}_k},$$

where $\mathcal{M}_k(\lambda_k, \overline{J}_k, \overline{N}_k)$ denotes the matching technology, representing the total measure of matched workers, and

$$\zeta_k(\lambda_k) = P\{\lambda_k \mathbf{E}_k[\max\{1\{M_{i,j} \geq W_{i,j}\}W_{i,j}, R_i\} \mid R_i, K_i] \geq K_i\},$$

i.e., the probability of the worker deciding to search for a firm. Hence, $\zeta_k(\lambda_k)\overline{N}_k$ represents the total measure of workers searching for a match with a firm.

As for the wage determination through Nash bargaining, we follow Ahn, Arcidiacono, and Wessels (2011) and obtain the following wage generation: for $M_{i,j} \geq \underline{W}_k$,

$$W_{i,j} = \max\{\beta_k M_{i,j}, R_i, \underline{W}_k\},$$

where $\beta_k \in (0,1)$ is a parameter that represents worker i's bargaining strength. We also follow Ahn, Arcidiacono, and Wessels (2011) in simplifying the procedure by assuming that (32) is satisfied for all the workers such that $R_i \leq \underline{W}_k$. Then the wage is generated only for those workers with $R_i \leq \underline{W}_k$, and hence, the wage generation is simplified as follows: for $M_{i,j} \geq \underline{W}_k$,

$$W_{i,j} = \max\{\beta_k M_{i,j}, \underline{W}_k\}. \tag{33}$$

The employment indicator $Y_{i,j} \in \{0,1\}$ is also given as follows:

$$Y_{i,j} = 1\{M_{i,j} \geq W_{i,j}\} = 1\{M_{i,j} \geq \underline{W}_k\}, \tag{34}$$

where the last equality follows from (33) and $\beta_k \in (0,1)$.

Our counterfactual policy is to set the minimum wage to $\underline{W}_k^\Gamma$. We aim to predict the employment rate for population 0 (Texas) after the minimum wage changes to $\underline{W}_k^\Gamma$ (US$9).

To build up an empirical model, for population k, we specify the match output $M_{i,j}$ as follows:

$$\log M_{i,j} = X_i^\top \gamma_k + U_{i,j}.$$

where X_i denotes the observed characteristics of worker i, $U_{i,j}$ represents a match component that is not observed by the econometrician, and γ_k is a parameter vector. We assume that $U_{i,j}$'s are i.i.d., independent of $(X_i, W_{i,j}, \underline{W}_k)$, $i \in N_k$, and all firms j, and follow the distribution with the CDF F_k. Unlike Ahn, Arcidiacono, and Wessels (2011), we leave F_k as nonparametrically specified. Since we do not restrict $U_{i,j}$ to have mean zero, we lose no generality by assuming that the vector X_i does not include an intercept term.

It follows from this parametrization and (34) that:

$$Y_{i,j} = 1\{M_{i,j} \geq \underline{W}_k\} = 1\{X_i^\top \gamma_k + U_{i,j} \geq \log \underline{W}_k\}. \tag{35}$$

In order to check the applicability of the synthetic decomposition method, we consider the support conditions required in this setting. First, we define our policy components

$$\mu_k(X_i, \underline{W}_k) = X_i^\top \gamma_k - \log \underline{W}_k, \text{ and}$$
$$\mu_k^\Gamma(X_i, \underline{W}_k) = X_i^\top \gamma_k - \log \underline{W}_k^\Gamma.$$

We take

$$\mathcal{X}_0^\Gamma = \{x \in \mathcal{X}_0 : x^\top \gamma_0 - \log \underline{W}_0^\Gamma = \tilde{x}^\top \gamma_0 - \log \underline{W}_0, \text{ for some } \tilde{x} \in \mathcal{X}_0\}, \tag{36}$$

where we denote the support of X_i in the target population by $\mathcal{X}_0$. The set $\mathcal{X}_0^\Gamma$ represents the set of characteristics for people who have a match for comparison after the policy. The support conditions required can be summarized as follows.

(a) The support of $X_i^\top \gamma_0 - \log \underline{W}_0$ and that of $X_i^\top \gamma_0 - \log \underline{W}_0^\Gamma$ overlap in the target population (Assumption 2.1), so that the set $\mathcal{X}_0^\Gamma$ is not empty.

(b) The support of X_i in each source population $k = 1, ..., K$ contains the set $\mathcal{X}_0$.

First, note that due to the independence between $U_{i,j}$'s and X_i's, the ARF $m_k(\overline{\mu}, x)$ does not depend on the second argument, and we simply write $m_k(\overline{\mu})$. In our case, the pre-policy and post-policy ARFs take the following form:

$$m_k(\mu_k(X_i)) = \int g_k(\mu_k(X_i), u) dF_k(u), \text{ and} \tag{37}$$
$$m_k(\mu_k^\Gamma(X_i)) = \int g_k(\mu_k^\Gamma(X_i), u) dF_k(u),$$

where

$$g_k(\overline{\mu}, u) = 1\{\overline{\mu} + u \geq 0\}.$$

For each worker $i \in N_k$, we let $j(i)$ be the firm matched with this worker, and simply write $Y_i = Y_{i,j(i)}$ and $W_i = W_{i,j(i)}$. Since $(X_i, \underline{W}_k)$ and U_{ij} are independent, we have

$$m_k(\overline{\mu}) = \mathbf{E}_k\left[Y_i \mid \mu_k(X_i) = \overline{\mu}\right].$$

Then, the synthetic prediction is obtained by taking the weights w_k's which minimize the L^2-distance between

$$m_0(\mu_0^\Gamma(x)) \text{ and } \sum_{k=1}^{K} m_k(\mu_k^\Gamma(x)) w_k,$$

over the set $\mathcal{X}_0^\Gamma$ such that the intersection of the support of $X_i^\top \gamma_0 - \log \underline{W}_0$ and that of $X_i^\top \gamma_0 - \log \underline{W}_0^\Gamma$ when we restrict X_i to $\mathcal{X}_0^\Gamma$ is nonempty.

4.3. Empirical Implementation

We use the dataset from Allegretto, Dube, Reich, and Zipperer (2017) for our exercises, which is drawn from the Current Population Survey (CPS), a repeated cross-section. Following the authors, among many others, we focus on teenagers and use their individual-level employment status as the outcome, $Y_{i,j} \in \{0, 1\}$, individual-level characteristics as X_i (age, sex, marriage status, whether they are Hispanic, whether they are black or another non-white race). We further observe wages for an employed samples, $W_{i,j}$, and each state's minimum wages. Our sample is restricted from 2002 to 2014, so that it does not start during the 2001 recession (see Neumark and Wascher (2017) for a discussion).

The counterfactual sets the minimum wage in Texas (US$7.25 in 2014) to US$9 in 2014 (i.e., US$11.42 in 2022 dollars). Our parameter of interest, θ_0 is the average teenage employment in Texas in 2014 (for Texas' 2014 population) had the minimum wage been US$9. We compare this to teenage employment for those in Texas in 2014 with the prevailing minimum wage.

To make this comparison, we consider two sets of source regions. First, we use the states with the highest prevailing minimum wages within our sample, which are California, Connecticut, D.C. and Washington.[18] We note that the support conditions (36) can include more states because it is a condition on the support of the ARF and not on the policy itself. Hence, in a second exercise, we further include Florida (a large state close to Texas), Louisiana (a neighboring state) and Oregon (another state satisfying the first restriction). For illustration purposes, we

[18] Vermont also satisfies this restriction, but we drop it as its sample is too small to provide meaningful variation for estimation of Vermont-specific parameters.

TABLE 1. Summary Statistics for the Whole Sample (2002-2014)

	CA	CT	DC	FL	LA	OR	TX	WA
(Teenage) Employment	0.257	0.348	0.164	0.301	0.263	0.320	0.285	0.337
Wages (US$)	8.74	8.76	8.75	7.67	7.40	8.51	7.51	8.81
Age	17.48	17.37	17.47	17.44	17.39	17.43	17.42	17.39
Male	0.511	0.517	0.467	0.514	0.504	0.501	0.504	0.512
Married	0.014	0.005	0.003	0.010	0.007	0.015	0.021	0.010
Hispanic	0.463	0.147	0.124	0.240	0.035	0.130	0.471	0.124
Black	0.062	0.109	0.727	0.193	0.385	0.025	0.122	0.034
Share of Teenagers in Population	0.075	0.073	0.049	0.062	0.077	0.065	0.077	0.069
Average State Unemployment (%)	8.11	6.51	7.60	6.44	6.28	7.80	6.18	7.00

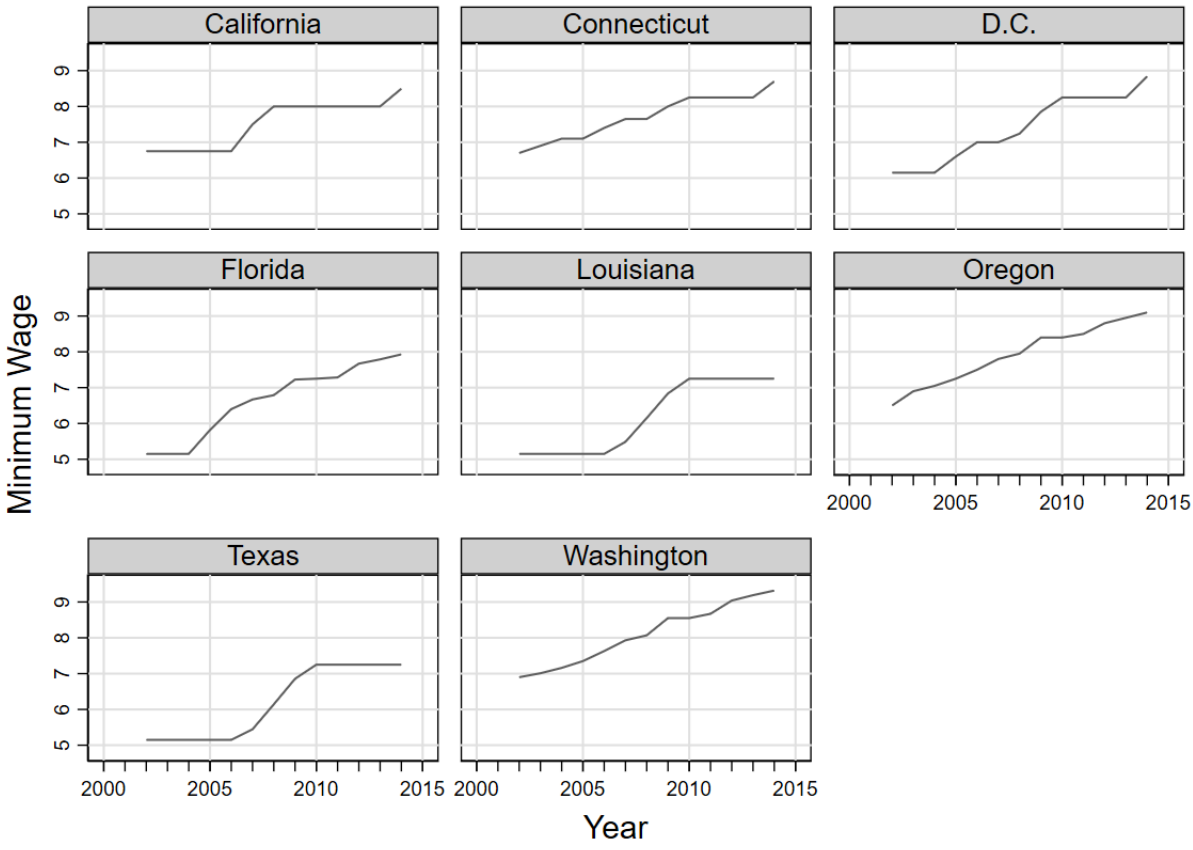
Notes: The table presents summary statistics for the variables used in the main specification. This includes the labor market outcomes (employment and wages for the employed) and observable characteristics. Note that these statistics are averaged over all years in the sample period (2002-2014), while our main comparison is to Texas in 2014.

use a 10% random sample of the data for each region. This shows the performance of our estimator with reasonably standard sample sizes.

Summary statistics are provided in Table 1, while the variation in minimum wages across all source and target regions is shown in Figure 2. In terms of demographics (e.g., the share of teenage Hispanics and African-Americans), Texas most resembles California. However, it is more similar to Florida and Louisiana in terms of average teenage employment and in wages. On the other hand, Louisiana's minimum wage policies are very similar to Texas', which may provide less information on such changes.

We estimate the ARF's for the model in two steps. First, we estimate γ_k using the pairwise differencing method of Honoré and Powell (1994). Then, we plug them into $\mu_k(X_i)$ and estimate m_k nonparametrically using a kernel regression estimation method and a cross-validated bandwidth. Details are provided in the Supplemental Note. We use $B = 200$ bootstrap draws,

FIGURE 2. The Variation in Minimum Wages Over Time Across Regions in Our Sample



Notes: The panels depict the minimum wages during the sample periods (2002-2014) in some selected states in the U.S.

set $\kappa = 0.005$ and $\alpha = 0.05$. We draw a fine grid of $\mathbf{w}$ uniformly over its simplex, using a procedure based on Rubin (1981).[19]

4.4. Results

Table 2 presents the results of the estimation. We present two specifications per exercise, which only differ in whether they accommodate aggregate variables: the share of teenagers in the state population and the average unemployment in the state.

[19]To construct each gridpoint, we first draw a vector of dimension $K-1$, where each element is drawn i.i.d. from the uniform distribution with support $[0,1]$. Then, we include 0 and 1 into that drawn vector, which is then sorted. The grid point is the vector of differences across adjacent elements of $\mathbf{w}$ (which are all nonnegative and must sum up to 1 by construction).

TABLE 2. Confidence Intervals for θ_0: Predicted Average (Teenage) Employment in Texas After a Counterfactual Minimum Wage Increase

	Increase to US$9			
θ_0	0.195 [0.144, 0.253]	0.186 [0.112, 0.265]	0.192 [0.124, 0.265]	0.186 [0.116, 0.261]
$w_0 = \begin{pmatrix} CA \\ CT \\ FL \\ DC \\ LA \\ OR \\ WA \end{pmatrix}$	$\begin{pmatrix} 0.602 \\ 0 \\ - \\ 0.236 \\ - \\ - \\ 0.162 \end{pmatrix}$	$\begin{pmatrix} 0.647 \\ 0 \\ - \\ 0.246 \\ - \\ - \\ 0.107 \end{pmatrix}$	$\begin{pmatrix} 0.367 \\ 0 \\ 0.026 \\ 0.286 \\ 0 \\ 0 \\ 0.322 \end{pmatrix}$	$\begin{pmatrix} 0.381 \\ 0 \\ 0.216 \\ 0 \\ 0 \\ 0.305 \\ 0.098 \end{pmatrix}$
Teenage Employment in Texas in 2014	0.292	0.292	0.292	0.292
Effect of Minimum Wage Increase on Employment	-9.72 p.p. (or -33.3%)	-10.6 p.p. (or -36.3%)	-10 p.p. (or -34.2%)	-10.6 p.p. (or -36.2%)
Aggregate Variables		✓		✓
More Source Regions			✓	✓

Notes: The table presents the results from synthetic decomposition for increasing minimum wages in Texas to US$9 on (teenage) employment in 2014. θ_0 represents our parameter of interest, which is the predicted average (teenage) employment after the policy, keeping the population in Texas in 2014 the same. This prediction uses information from the target region (Texas) and source regions. We present its estimates across two specifications: one using only individual-level covariates (age, sex, married, hispanic, black, other race), and another which further includes aggregate variables (teen share in the state, average unemployment rate in the state). We then use two different sets of source regions. The confidence interval for θ_0 is presented in brackets. For comparability, we also present the empirical average (pre-policy), and how the estimated θ_0 translates to changes in employment relative to the data (the baseline employment in Texas is 0.292). We also present estimates for the weights, $\mathbf{w}_0$.

Our estimates suggest that an increase in the minimum wage decrease predicted average (teenage) employment: our estimates of θ_0 and all upper bounds of their associated confidence intervals are all below the observed employment rate of 0.292. In particular, the counterfactual employment is estimated between 0.186-0.195, implying a decrease in average (teenage) employment between 9-11 percentage points. This is robust across specifications and consistent with the labor economics literature finding such negative effects (see Neumark (2019) for a

review). In terms of magnitudes, it is also very similar to those found in Flinn (2006) with a similar proportional increase in minimum wages from US$5 to US$6 – see his Figure 4.

Our synthetic comparison is predominantly based on California, D.C., Florida and Washington. This seems intuitive, as California best approximates the demographics of Texas. However, our estimates also suggest that accounting for common shocks/aggregate variables is important. Absent state-specific economic trends (here, the average teenage share in the population and the average unemployment rate), we would have estimated the effects of minimum wages on employment to be about 1 percentage point lower, thereby underestimating its negative effects. The aggregate variables also matter for the weights given to source regions: because state-level variables change the model's causal structure, as well as the characteristics of those states, there is no reason why each region would remain equally comparable to Texas with/without them. In fact, we find that California receives lower weights when including such variables. This is because its state unemployment levels are much larger than Texas's which, in turn, is more similar to Washington's.

5. Conclusion

In this paper, we propose a novel way to utilize data from other populations to generate counterfactual predictions for a target population, when we do not have enough data for the latter. We explore ways to utilize data from other populations ("source populations"), motivated by a synthetic transferability condition. This hypothesis generalizes existing invariance conditions for extrapolation of causal effects and allows us to build predictions based on a synthetic causal structure, chosen to be as close as possible to the target ARF under a certain metric. Our approach is quite general and applies to various policy settings where the researcher may have multiple source populations, regardless of how the reduced forms are originated structurally.

There are further extensions that one can explore from this research. First, it is possible that, just like in synthetic control methods, using many source populations may cause overfitting. As in synthetic control, a judicious selection of source populations based on the domain knowledge of the context of application is important in practice. We believe that a decision-theoretic guidance to help the researcher in this selection would be helpful, although to the best of our knowledge, the predominant portion of the literature focuses on a decision setting under a single population. Second, it would be useful to statistically gauge the plausibility of the synthetic transferability condition. For this, we may need to sacrifice the generality of this paper's setting and make use of further restrictions on the ARF's, such as continuity or shape constraints of the ARFs, depending on the application of focus. Finally, the current paper has assumed that the policy is known to the researcher. However, in practice, the precise form of

the policy may be unknown. The researcher may face a range of policies under consideration, or may not have precise knowledge of how the policy alters the reduced form, and may need to estimate it using additional data. This question seems relevant in practice.

References

ABADIE, A. (2021): "Using synthetic controls: Feasibility, data requirements, and methodological aspects," *Journal of Economic Literature*, 59(2), 391–425.

AGUIRREGABIRIA, V. (2005): "Nonparametric Identification of Behavioral Responses to Counterfactual Policy Intervention in Dynamic Discrete Decision Processes," *Economics Letters*, 87, 393–398.

AGUIRREGABIRIA, V., AND J. SUZUKI (2014): "Identification and Counterfactuals in Dynamic Models of Market Entry and Exit," *Quantitative Marketing and Economics*, 12(3), 267–304.

AHN, T., P. ARCIDIACONO, AND W. WESSELS (2011): "The Distributional Impacts of Minimum Wage Increases when both Labor Supply and Labor Demand are Endogenous," *Journal of Business & Economic Statistics*, 29(1), 12–23.

AKCIGIT, U., S. BASLANDZE, AND S. STANTCHEVA (2016): "Taxation and the international mobility of inventors," *American Economic Review*, 106(10), 2930–81.

ALLCOTT, H. (2015): "Site Selection Bias in Program Evaluation," *Quarterly Journal of Economics*, 130(3), 1117–1165.

ALLEGRETTO, S., A. DUBE, M. REICH, AND B. ZIPPERER (2017): "Credible Research Designs for Minimum Wage Studies: A Response to Neumark, Salas, and Wascher," *ILR Review*, 70(3), 559–592.

ALTONJI, J. G., AND R. L. MATZKIN (2005): "Cross-Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053–1102.

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

——— (2000): "Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space," *Econometrica*, pp. 399–405.

ARCIDIACONO, P., AND R. A. MILLER (2020): "Identifying Dynamic Discrete Choice Models of Short Panels," *Journal of Econometrics*, 215(2), 473–485.

ATHEY, S., R. CHETTY, AND G. IMBENS (2020): "Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes," *arXiv preprint arXiv:2006.09676*.

BANDIERA, O., G. FISCHER, A. PRAT, AND E. YTSMA (2021): "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments," *American Economic Review:*

Insights, 3, 435–454.

BLINDER, A. S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, pp. 436–455.

BLUNDELL, R., AND R. L. MATZKIN (2014): "Control Functions in Nonseparable Simultaneous Equations Models," *Quantitative Economics*, 5(2), 271–295.

BLUNDELL, R., AND J. L. POWELL (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics*, ed. by L. Dewatripont, L. Hansen, and S. Turnovsky, vol. 2, pp. 312–357. Cambridge University Press, Cambridge.

BOLD, T., M. KIMENYI, G. MWABU, J. SANDEFUR, ET AL. (2018): "Experimental Evidence on Scaling up Education Reforms in Kenya," *Journal of Public Economics*, 168, 1–20.

BUGNI, F. A., I. A. CANAY, AND X. SHI (2017): "Inference for Subvectors and Other Functions of Partially Identified Parameters in Moment Inequality Models," *Quantitative Economics*, 8, 1–38.

BÜHLMANN, P. (2020): "Invariance, Causality and Robustness," *Statistical Science*, 35, 404–426.

CANEN, N., AND K. SONG (2023): "A Decomposition Approach to Counterfactual Analysis in Game-Theoretic Models," *arXiv:2010.08868v5 [econ.EM]*.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND B. MELLY (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81(6), 2205–2268.

COX, G., AND X. SHI (2022): "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models," *Forthcoming in Review of Economic Studies*.

DINARDO, J., N. M. FORTIN, AND T. LEMIEUX (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001–1044.

DUBE, A., T. W. LESTER, AND M. REICH (2010): "Minimum Wage Effects Across State Borders: Estimates using Contiguous Counties," *The review of economics and statistics*, 92(4), 945–964.

DUFLO, E. (2004): "Scaling Up and Evaluation," *Annual World Bank Conference on Development Economics*, pp. 341–369.

FANG, Z., AND J. SEO (2021): "A Projection Framework for Testing Shape Restrictions That Form Convex Cones," *Econometrica*, 89(5), 2439–2458.

FLINN, C., AND J. MULLINS (2015): "Labor Market Search and Schooling Investment," *International Economic Review*, 56(2), 359–398.

FLINN, C. J. (2002): "Labour Market Structure and Inequality: A Comparison of Italy and the US," *The Review of Economic Studies*, 69(3), 611–645.

——— (2006): "Minimum Wage Effects on Labor Market Outcomes Under Search, Matching, and Endogenous Contact Rates," *Econometrica*, 74(4), 1013–1062.

——— (2011): *The Minimum Wage and Labor Market Outcomes*. MIT press.

FORTIN, N., T. LEMIEUX, AND S. FIRPO (2011): "Decomposition Methods in Economics," in *Handbook of Labor Economics*, vol. 4, pp. 1–102. Elsevier.

GECHTER, M., AND R. MEAGER (2022): "Combining Experimental and Observational Studies in Meta-Analysis: A Debiasing Approach," *Working Paper*.

GEYER, C. J. (1994): "On the Asymptotics of Constrained M-Estimation," *Annals of Statistics*, 22, 1993 – 2010.

GORRY, A., AND J. J. JACKSON (2017): "A Note on the Nonlinear Effect of Minimum Wage Increases," *Contemporary Economic Policy*, 35(1), 53–61.

GREGORY, A. W., AND G. W. SMITH (1993): "25 Statistical aspects of calibration in macroeconomics," .

GU, J., T. M. RUSSELL, AND T. STRINGHAM (2022): "Counterfactual Identification and Latent Space Enumeration in Discrete Outcome Models," *Working Paper*.

GUI, G. (2022): "Combining Observational and Experimental Data Using First-Stage Covariates," *arXiv preprint arXiv:2010.05117*.

HAHN, J., AND Z. LIAO (2021): "Bootstrap Standard Error Estimates and Inference," *Econometrica*, 89(4), 1963–1977.

HARTMAN, E., R. GRIEVE, R. RAMSAHAI, AND J. S. SEKHON (2015): "From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated: Combining Experimental with Observational Studies to Estimate Population Treatment Effects," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 757–778.

HECKMAN, J. J. (2010): "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 48, 356–398.

HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

HECKMAN, J. J., AND E. J. VYTLACIL (2007): "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," *Handbook of Econometrics*, 6, 4779–4874.

HONORÉ, B. E., AND J. L. POWELL (1994): "Pairwise-Difference Estimators of Censored and Truncated Regression Models," *Journal of Econometrics*, 64, 241–278.

HOTZ, J. V., G. W. IMBENS, AND J. H. MORTIMER (2005): "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations," *Journal of Econometrics*, 125, 241–270.

HSIEH, Y.-W., X. SHI, AND M. SHUM (2022): "Inference on Estimators Defined by Mathematical Programming," *Journal of Econometrics*, 226(2), 248–268.

HSU, Y.-C., T.-C. LAI, AND R. P. LIELI (2022): "Counterfactual Treatment Effects: Estimation and Inference," *Journal of Business and Economic Statistics*, 40(1), 240–255.

ISHIHARA, T., AND T. KITAGAWA (2021): "Evidence Aggregation for Treatment Choice," *arXiv:2108.06473v1*.

JUHN, C., K. M. MURPHY, AND B. PIERCE (1993): "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy*, 101(3), 410–442.

JUN, S. J., AND J. PINKSE (2020): "Counterfactual Prediction in Complete Information Games: Point Prediction Under Partial Identification," *Journal of Econometrics*, 216(2), 394–429.

KAIDO, H., AND A. SANTOS (2014): "Asymptotically Efficient Estimation of Models Defined by Convex Moment Inequalities," *Econometrica*, pp. 387–413.

KALOUPTSIDI, M., Y. KITAMURA, L. LIMA, AND E. A. SOUZA-RODRIGUES (2020): "Partial Identification and Inference for Dynamic Models and Counterfactuals," Discussion paper, National Bureau of Economic Research.

KALOUPTSIDI, M., P. T. SCOTT, AND E. SOUZA-RODRIGUES (2020): "Identification of Counterfactuals in Dynamic Discrete Choice Models," *Quantitative Economics*.

KITAGAWA, E. M. (1955): "Components of a Difference Between Two Rates," *Journal of the American Statistical Association*, 50, 1168–1194.

KITAMURA, Y., AND J. STOYE (2018): "Nonparametric Analysis of Random Utility Models," *Econometrica*, 86, 1883–1909.

KLEVEN, H. J., C. LANDAIS, AND E. SAEZ (2013): "Taxation and International Migration of Superstars: Evidence from the European Football Market," *American economic review*, 103(5), 1892–1924.

KLEVEN, H. J., C. LANDAIS, E. SAEZ, AND E. SCHULTZ (2014): "Migration and Wage Effects of Taxing Top Earners: Evidence from the Foreigners' Tax Scheme in Denmark," *The Quarterly Journal of Economics*, 129(1), 333–378.

KLINE, P. (2011): "Oaxaca-Blinder as a Reweighting Estimator," *American Economic Review: Papers and Proceedings*, 101(4), 532–537.

LI, J. (2022): "The Proximal Bootstrap for Constrained Estimation," *Working Paper*.

MEAGER, R. (2022): "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature," *Forthcoming in American Economic Review*.

MOHAMAD, D. A., E. W. VAN ZWET, E. CATOR, AND J. J. GOEMAN (2020): "Adaptive Critical Value for Constrained Likelihood Ratio Testing," *Biometrika*, 107(3), 677–688.

MOON, H. R. M., AND F. SCHORFHEIDE (2009): "Estimation with Overidentifying Inequality Moment Conditions," *Journal of Econometrics*, 153, 136–154.

MORETTI, E., AND D. J. WILSON (2017): "The Effect of State Taxes on the Geographical Location of Top Earners: Evidence from Star Scientists," *American Economic Review*, 107(7), 1858–1903.

MURALIDHARAN, K., AND P. NIEHAUS (2017): "Experimentation at Scale," *Journal of Economic Perspectives*, 31, 103–214.

NEUMARK, D. (2019): "The Econometrics and Economics of the Employment Effects of Minimum Wages: Getting from Known Unknowns to Known Knowns," *German Economic Review*, 20(3), 293–329.

NEUMARK, D., AND W. WASCHER (2017): "Reply to "Credible Research Designs for Minimum Wage Studies"," *ILR Review*, 70(3), 593–609.

NORETS, A., AND X. TANG (2014): "Semiparametric Inference in Dynamic Binary Choice Models," *Review of Economic Studies*, 81(3), 1229–1262.

OAXACA, R. (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, pp. 693–709.

ROSEN, A. M. (2008): "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146(1), 107–117.

ROTHE, C. (2010): "Nonparametric Estimation of Distributional Policy Effects," *Journal of Econometrics*, 155, 56–70.

RUBIN, D. B. (1981): "The Bayesian Bootstrap," *The Annals of Statistics*, pp. 130–134.

SCHEUER, F., AND I. WERNING (2017): "The Taxation of Superstars," *The Quarterly Journal of Economics*, 132(1), 211–270.

SHAO, J. (1992): "Bootstrap Variance Estimators with Truncation," *Statistics & probability letters*, 15(2), 95–101.

SHI, X., AND M. SHUM (2015): "Simple Two-Stage Inference for a Class of Partially Identified Models," *Econometric Theory*, 31(3), 493–520.

STOCK, J. H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84, 567–575.

TODD, P. E., AND K. I. WOLPIN (2006): "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," *American Economic Review*, pp. 1384–1417.

——— (2008): "Ex Ante Evaluation of Social Programs," *Annales d'Économie et de Statistique*, pp. 263–291.

VIVALT, E. (2020): "How Much Can We Generalize From Impact Evaluations?," *Journal of the European Economic Association*, 18(6), 3045–3089.

WANG, S., AND D. Y. YANG (2021): "Policy Experimentation in China: The Political Economy of Policy Learning," Discussion paper, National Bureau of Economic Research.

WOLPIN, K. I. (2007): "Ex ante policy Evaluation, Structural Estimation and Model Selection," *American Economic Review*, 97(2), 48–52.

——— (2013): *The Limits of Inference Without Theory*. MIT Press.

SUPPLEMENTAL NOTE TO "SYNTHETIC DECOMPOSITION FOR COUNTERFACTUAL PREDICTIONS"

Nathan Canen and Kyungchul Song
University of Houston and University of British Columbia

The supplemental note provides the proof of asymptotic validity of inference proposed in Canen and Song (2023), and some details on the Monte Carlo simulations and empirical application.

A. Uniform Asymptotic Validity

A.1. Assumptions and Results

Let us first introduce conditions that ensure uniform asymptotic validity of the confidence intervals, $C_{1-\alpha}$, defined in (31). Let $\mathcal{P}$ be the space of probability distributions that satisfy Assumptions A.1-A.5 below. From here on, we make explicit the dependence of $\mathbf{w}_0$, $\theta(\mathbf{w}_0)$, and Ω on $P \in \mathcal{P}$ by rewriting them as $\mathbf{w}_P$, $\theta_P(\mathbf{w}_P)$ and Ω_P. Similarly we write H_P and $\mathbf{h}_P$ instead of H and $\mathbf{h}$, and write $\mu_{k,P}$, $\mu^{\Gamma}_{k,P}$ and $m_{k,P}$ instead of μ_k, μ^{Γ}_k, and m_k.

The nonstandard aspect of uniform asymptotic validity in our setting comes from the fact that $\sqrt{n_0}(\hat{\mathbf{w}} - \mathbf{w}_P)$ exhibits discontinuity in its pointwise asymptotic distribution. Hence, our proof focuses on dealing with this aspect, using high level conditions for other aspects that can be handled using standard arguments.

Assumption A.1. For each $k = 0, 1, ..., K$, there exists a constant $r_k > 0$ such that $n_k/n_0 \to r_k$, as $n_k, n_0 \to \infty$.

Assumption A.1 says that the sample size from each source population is not asymptotically negligible relative to the sample size from the target population.

Assumption A.2. For each $k = 0, 1, ..., K$, there exists $\delta > 0$ such that

$$\sup_{P \in \mathcal{P}} \mathbf{E}_P \left[|m_{k,P}(\mu_{k,P}(X_i), X_i)|^{4+\delta} \right] < \infty \text{ and } \sup_{P \in \mathcal{P}} \mathbf{E}_P \left[|m_{k,P}(\mu^{\Gamma}_{k,P}(X_i), X_i)|^{4+\delta} \right] < \infty.$$

Assumption A.2 requires that the ARFs have a moment bounded uniformly over $P \in \mathcal{P}$.

Assumption A.3. There exists $\eta > 0$ such that for all $n \geq 1$,

$$\inf_{P \in \mathcal{P}} \lambda_{\min}(H_P) > \eta,$$

where $\lambda_{\min}(H_P)$ denotes the smallest eigenvalue of H_P.

Assumption A.3 requires that the matrix H_P has eigenvalues bounded away from zero uniformly over $P \in \mathcal{P}$ and over $n \geq 1$. The assumption excludes a setting where $\mathbf{w}_P$ is weakly identified. Later we will discuss how this assumption can be relaxed.

Recall the definition $W_i = (Y_i, X_i^\top)^\top$. For each $k = 0, 1, ..., K$, let us define

$$q_{k,0,P}(W_i) = m_{k,P}(\mu_{k,P}^\Gamma(X_i), X_i) 1\{X_i \in \mathcal{X}_0^\Gamma\} \text{ and}$$
$$\hat{q}_{k,0}(W_i) = \hat{m}_k(\hat{\mu}_k^\Gamma(X_i), X_i) 1\{X_i \in \hat{\mathcal{X}}_0^\Gamma\}.$$

Similarly, we define $q_{k,1,P}(W_i)$ and $\hat{q}_{k,1}(W_i)$ to be the same as $q_{k,0,P}(W_i)$ and $\hat{q}_{k,0}(W_i)$ except that $1\{X_i \in \mathcal{X}_0^\Gamma\}$ and $1\{X_i \in \hat{\mathcal{X}}_0^\Gamma\}$ are replaced by $1\{X_i \notin \mathcal{X}_0^\Gamma\}$ and $1\{X_i \notin \hat{\mathcal{X}}_0^\Gamma\}$ respectively. The following assumption requires the asymptotic linear representation of the estimated ARFs.

For any sequence of random vectors Z_{n_0} and W_{n_0} in $\mathbf{R}^d$, $n_0 \geq 1$, we denote

$$Z_{n_0} = W_{n_0} + o_{\mathcal{P}}(1),$$

if for each $\epsilon > 0$,

$$\limsup_{n_0 \to \infty} \sup_{P \in \mathcal{P}} P\left\{ \|Z_{n_0} - W_{n_0}\| > \epsilon \right\} = 0.$$

Assumption A.4. Suppose that for each $k = 0, 1, ..., K$, $\ell = 0, 1$, $\varphi_{k,\ell,P}(\cdot)$ is equal to $q_{k,\ell,P}(\cdot)$ or a constant function at one. Then, for each $j, k = 0, 1, ..., K$, $\ell = 0, 1$, as $n_0 \to \infty$,

$$\frac{1}{\sqrt{n_0}} \sum_{i \in N_0} \left(\hat{q}_{j,\ell}(W_i) - q_{j,\ell,P}(W_i) \right) \varphi_{k,\ell,P}(W_i) = \frac{1}{\sqrt{n_j}} \sum_{i \in N_j} \psi_{j,\ell,P}(W_i; \varphi_{k,\ell,P}) + o_{\mathcal{P}}(1),$$
$$\frac{1}{\sqrt{n_0}} \sum_{i \in N_0} \left(\hat{q}_{j,\ell}(W_i) - q_{j,\ell,P}(W_i) \right) \left(\hat{q}_{k,\ell}(W_i) - q_{k,\ell,P}(W_i) \right) = o_{\mathcal{P}}(1),$$

where $\psi_{j,\ell,P}(W_i; \varphi_{k,\ell,P})$ is a mean zero random variable such that for some $\delta > 0$,

$$\sup_{P \in \mathcal{P}} \mathbf{E}_P \left[|\psi_{j,\ell,P}(W_i; \varphi_{k,\ell,P})|^{4+\delta} \right] < \infty,$$

for all $k = 0, 1, ..., K$ and $\ell = 0, 1$.

Note that the estimation error in $\hat{q}_{j,\ell}(\cdot)$ comes from the sample in region j, whereas the summation is over the sample in region 0. The influence function is driven by the randomness in the estimation error $\hat{q}_{j,\ell}(\cdot) - q_{j,\ell,P}(\cdot)$. Similarly, we make the following assumption for the bootstrap version of the estimators.

Assumption A.5. Suppose that for each $k = 0, 1, ..., K$, $\ell = 0, 1$, $(\hat{\varphi}_{k,\ell}(\cdot), \varphi_{k,\ell,P}(\cdot))$ is equal to $(\hat{q}_{k,\ell}(\cdot), q_{k,\ell,P}(\cdot))$ or a pair of constant functions at one. Then, for each $j, k = 0, 1, ..., K$, $\ell = 0, 1$, the following statements hold.

(i) As $n_0 \to \infty$,

$$\frac{1}{\sqrt{n_0}} \sum_{i \in N_0} \left(\hat{q}^*_{j,\ell}(W^*_i) - \hat{q}_{j,\ell}(W^*_i) \right) \hat{\varphi}_{k,\ell}(W^*_i) = \frac{1}{\sqrt{n_j}} \sum_{i \in N_j} \hat{\psi}_{j,\ell,P}(W^*_i; \varphi_{k,\ell,P}) + o_{\mathcal{P}}(1),$$

$$\frac{1}{\sqrt{n_0}} \sum_{i \in N_0} \left(\hat{q}^*_{j,\ell}(W^*_i) - \hat{q}_{j,\ell}(W^*_i) \right) \left(\hat{q}^*_{k,\ell}(W^*_i) - \hat{q}_{k,\ell}(W^*_i) \right) = o_{\mathcal{P}}(1),$$

where

$$\hat{\psi}_{j,\ell,P}(W^*_i; \varphi_{k,\ell,P}) = \psi_{j,\ell,P}(W^*_i; \varphi_{k,\ell,P}) - \frac{1}{n_j} \sum_{i \in N_j} \psi_{j,\ell,P}(W_i; \varphi_{k,\ell,P}),$$

and $\psi_{j,\ell,P}(\cdot; \varphi_{k,\ell,P})$ is the influence function in Assumption A.4.

(ii) As $n_0 \to \infty$,

$$\frac{1}{\sqrt{n_0}} \sum_{i \in N_0} \left(\hat{q}_{j,\ell}(W^*_i) - \hat{q}_{j,\ell}(W_i) \right) \left(\hat{q}_{k,\ell}(W^*_i) - q_{k,\ell,P}(W^*_i) \right) = o_{\mathcal{P}}(1),$$

$$\frac{1}{\sqrt{n_0}} \sum_{i \in N_0} \left(\hat{q}_{j,\ell}(W^*_i) - \hat{q}_{j,\ell}(W_i) \right) \left(\hat{q}_{k,\ell}(W_i) - q_{k,\ell,P}(W_i) \right) = o_{\mathcal{P}}(1), \text{ and}$$

$$\frac{1}{\sqrt{n_0}} \sum_{i \in N_0} \left(\hat{q}^*_{j,\ell}(W^*_i) \hat{q}^*_{k,\ell}(W^*_i) - q_{j,\ell,P}(W_i) q_{k,\ell,P}(W_i) \right) = o_{\mathcal{P}}(1).$$

Define

$$\Omega_{n,P} = \frac{1}{n_0} \sum_{i \in N} \mathbf{E}_P \left[\tilde{\boldsymbol{\psi}}_{i,P} \tilde{\boldsymbol{\psi}}^\top_{i,P} \right], \tag{A.1}$$

where $\tilde{\boldsymbol{\psi}}_{i,P} = \Psi_{i,P} \mathbf{w}_P - \boldsymbol{\psi}_{i,P}$ and $\Psi_{i,P}$ and $\boldsymbol{\psi}_{i,P}$ are defined in Lemma A.7 below. Inspection of $\Omega_{n,P}$ shows that it depends on n only through n_k/n_0, $k = 1, ..., K$, and depends on the ratios continuously. Let Ω_P be the same as $\Omega_{n,P}$ with n_k/n_0 replaced by r_k, for $k = 1, ..., K$, where r_k's are positive constants in Assumption A.1. Then, it is not hard to see that from Assumption A.2,

$$\sup_{P \in \mathcal{P}} \left\| \Omega_{n,P} - \Omega_P \right\| \to 0, \tag{A.2}$$

as $n_0 \to \infty$.

The following theorem shows that the estimators $\hat{\mathbf{w}}$ and $\hat{\theta}(\hat{\mathbf{w}})$ are $\sqrt{n_0}$-consistent for $\mathbf{w}_P$ and $\theta_P(\mathbf{w}_P)$ uniformly over $P \in \mathcal{P}$.

Theorem A.1. *Suppose that Assumptions A.1-A.5 hold and that* $\inf_{P \in \mathcal{P}} \lambda_{\min}(\Omega_P) > 0$. *Then,*

$$\lim_{M \uparrow \infty} \limsup_{n_0 \to \infty} \sup_{P \in \mathcal{P}} P \left\{ \sqrt{n_0} \left\| \hat{\mathbf{w}} - \mathbf{w}_P \right\| > M \right\} = 0.$$

However, as noted earlier, depending on the sequence of probabilities in $\mathcal{P}$, $\sqrt{n_0}(\hat{\mathbf{w}} - \mathbf{w}_P)$ can be asymptotically non-normal, and so can $\sqrt{n_0}(\hat{\theta}(\hat{\mathbf{w}}) - \theta_P(\mathbf{w}_P))$ as a consequence. Nevertheless, the confidence interval $C_{1-\alpha}$ we propose in the main text turns out to be uniformly asymptotically valid as the following theorem shows.

Theorem A.2. *Suppose that Assumptions A.1-A.5 hold. Then, for each $\alpha \in (0,1)$,*

$$\liminf_{n_0 \to \infty} \inf_{P \in \mathcal{P}} P\left\{\theta_P(\mathbf{w}_P) \in C_{1-\alpha}\right\} \geq 1 - \alpha.$$

The proofs of these results are presented in the next section.

A.2. Proofs

A.2.1. **Preliminary Results.** We begin with auxiliary results on the rejection probability of a test involving the squared residual from projecting an asymptotically normal random vector onto a polyhedral cone. The main preliminary result is Lemma A.6. This is the result we use later to establish the uniform asymptotic validity of the confidence set for $\mathbf{w}_P$.

First, for any matrix $m \times K$ matrix A, we consider a polyhedral cone of the following type:

$$\Lambda(A) = \left\{\mathbf{x} \in \mathbf{R}^K : A\mathbf{x} \leq \mathbf{0}\right\}.$$

When A is replaced by

$$A(\mathbf{w}) = [I_K, \mathbf{w}, -\mathbf{w}]^\top, \tag{A.3}$$

with $\mathbf{w} \in \Delta_{K-1}$, where I_K is the K-dimensional identity matrix, we write $\Lambda(\mathbf{w})$ simply, instead of $\Lambda(A(\mathbf{w}))$. Let $\overline{K} = \{1, ..., K\}$. For each $J \subset \overline{K}$, let

$$\begin{aligned}
\Lambda_J(\mathbf{w}) &= \left\{\mathbf{x} \in \mathbf{R}^K : \mathbf{x}_J = \mathbf{0}, \mathbf{x}_{-J} \leq \mathbf{0}, \text{ and } \mathbf{w}^\top \mathbf{x} = 0\right\} \text{ and} \\
L_J(\mathbf{w}) &= \left\{\mathbf{x} \in \mathbf{R}^K : \mathbf{x}_J = \mathbf{0} \text{ and } \mathbf{w}^\top \mathbf{x} = 0\right\}.
\end{aligned}$$

(An inequality between vectors are understood as holding element-wise. We also assume that any inequality or equality that involves a vector $\mathbf{x}_J$ with $J = \varnothing$ is vacuously true.) Note that $L_J(\mathbf{w})$ is the span of $\Lambda_J(\mathbf{w})$. The relative interior of $\Lambda_J(\mathbf{w})$ (relative to $L_J(\mathbf{w})$) is given by

$$\text{ri}(\Lambda_J(\mathbf{w})) = \left\{\mathbf{x} \in \mathbf{R}^K : \mathbf{x}_J = \mathbf{0}, \mathbf{x}_{-J} < \mathbf{0}, \text{ and } \mathbf{w}^\top \mathbf{x} = 0\right\}. \tag{A.4}$$

For any vector $\mathbf{x}$, we denote $[\mathbf{x}]_j$ to mean its j-th entry. For any vector $\mathbf{x} \in \mathbf{R}^K$, we let

$$J_0(\mathbf{x}) = \{j \in \overline{K} : [\mathbf{x}]_j = 0\}.$$

Given a symmetric positive definite matrix Ω, we define the norm $\|\cdot\|_\Omega$ as $\|\mathbf{x}\|_\Omega = \sqrt{\mathbf{x}'\Omega^{-1}\mathbf{x}}$, and the projection $\Pi_\Omega(\mathbf{y} \mid \Lambda(\mathbf{w}))$ (along the norm $\|\cdot\|_\Omega$) to be the solution to the following

minimization problem:

$$\inf_{\mathbf{x}\in\Lambda(\mathbf{w})} \|\mathbf{y}-\mathbf{x}\|_\Omega^2. \tag{A.5}$$

Since Ω is positive definite and $\Lambda(\mathbf{w})$ is closed and convex, the projection $\Pi_\Omega(\mathbf{y} \mid \Lambda(\mathbf{w}))$ exists and is unique. The following lemma shows how the projection along $\|\cdot\|_\Omega$ is translated into that along $\|\cdot\|$.

Lemma A.1. *For any $J \subset \overline{K}$ and $\mathbf{x} \in \Lambda(\mathbf{w})$ with any $\mathbf{w} \in \Delta_{K-1}$, the following holds.*

(i) $\mathbf{x} \in ri(\Lambda_J(\mathbf{w}))$ if and only if $J_0(\mathbf{x}) = J$.

(ii) If $ri(\Lambda_J(\mathbf{w})) \neq \emptyset$ for some $J \subset \overline{K}$, then, $J \neq \emptyset$ and $\overline{K} \setminus J \subset J_0(\mathbf{w})$, and

$$ri(\Lambda_J(\mathbf{w})) = \left\{\mathbf{x} \in \mathbf{R}^K : \mathbf{x}_J = \mathbf{0}, \mathbf{x}_{-J} < \mathbf{0}\right\}. \tag{A.6}$$

(iii) For any $\mathbf{y} \in \mathbf{R}^K$, $J_0(\Pi_\Omega(\mathbf{y} \mid \Lambda(\mathbf{w}))) \neq \emptyset$.

Proof: (i) The result follows from the fact that $\mathrm{ri}(\Lambda_J(\mathbf{w}))$, $J \subset \overline{K}$, partition $\Lambda(\mathbf{w})$.

(ii) For the first statement, suppose to the contrary that $J = \emptyset$. Then, since $\mathbf{w} \in \Delta_{K-1}$, $\mathrm{ri}(\Lambda_J(\mathbf{w})) = \emptyset$. Hence we must have $J \neq \emptyset$. As for the second statement, suppose that $\overline{K} \setminus J \not\subset J_0(\mathbf{w})$ so that there exists $j \in \overline{K} \setminus J$, with $[\mathbf{w}]_j > 0$. Then, for such J, none of $\mathbf{x} \in \Lambda_J(\mathbf{w})$ satisfies both $\mathbf{x}_{-J} < \mathbf{0}$, and $\mathbf{w}^\top \mathbf{x} = 0$, because $\mathbf{w} \geq \mathbf{0}$, and hence, $\mathrm{ri}(\Lambda_J(\mathbf{w})) = \emptyset$. Hence the second statement holds.

Now, we turn to the third statement. Suppose that $J = \overline{K}$. Then, $\mathrm{ri}(\Lambda_J(\mathbf{w})) = \Lambda_J(\mathbf{w}) = \{\mathbf{0}\}$. Hence (A.6) follows. Suppose that $\overline{K} \setminus J \neq \emptyset$. Since $\overline{K} \setminus J \subset J_0(\mathbf{w})$ by the previous result, $\overline{K} \setminus J_0(\mathbf{w}) \subset J$. Then, for any $\mathbf{x} \in \mathbf{R}^K$ such that $\mathbf{x}_J = 0$, the condition $\mathbf{w}^\top \mathbf{x} = 0$ in (A.4) holds. Again, (A.6) follows.

(iii) Since $\Lambda(\mathbf{w})$ is closed and convex, $\Pi_\Omega(\mathbf{y} \mid \Lambda(\mathbf{w}))$ exists in $\Lambda(\mathbf{w})$ and is unique. Since $\mathrm{ri}(\Lambda_J(\mathbf{w}))$'s partition $\Lambda(\mathbf{w})$, there exists a unique $J^* \subset \overline{K}$ such that $\Pi_\Omega(\mathbf{y} \mid \Lambda(\mathbf{w})) \in \mathrm{ri}(\Lambda_{J^*}(\mathbf{w}))$. Hence, $\mathrm{ri}(\Lambda_{J^*}(\mathbf{w})) \neq \emptyset$. By the previous results (i) and (ii), we must have $J^* = J_0(\Pi_\Omega(\mathbf{y} \mid \Lambda(\mathbf{w})))$, and $J^* \neq \emptyset$. ■

Lemma A.2. *For any $m \times K$ matrix A, and any symmetric positive definite $K \times K$ matrix Ω,*

$$\Pi_\Omega(\mathbf{y} \mid \Lambda(A)) = \Omega^{1/2} \Pi_I(\Omega^{-1/2}\mathbf{y} \mid \Omega^{-1/2}\Lambda(A)).$$

Proof: Note that

$$\begin{aligned}\Pi_\Omega(\mathbf{y} \mid \Lambda(A)) &= \arg\min_{\mathbf{x}:A\mathbf{x}\leq\mathbf{0}} (\mathbf{y}-\mathbf{x})^\top \Omega^{-1}(\mathbf{y}-\mathbf{x}) \\ &= \Omega^{1/2} \arg\min_{\mathbf{x}:A\Omega^{1/2}\mathbf{x}\leq\mathbf{0}} (\Omega^{-1/2}\mathbf{y}-\mathbf{x})^\top(\Omega^{-1/2}\mathbf{y}-\mathbf{x}).\end{aligned}$$

The last term is equal to $\Omega^{1/2}\Pi_I(\Omega^{-1/2}\mathbf{y} \mid \Lambda(A\Omega^{1/2})) = \Omega^{1/2}\Pi_I(\Omega^{-1/2}\mathbf{y} \mid \Omega^{-1/2}\Lambda(A))$. ■

Lemma A.3. *Suppose that $Y \in \mathbf{R}^K$ is a random vector following $N(\mathbf{0}, \Omega)$, with a symmetric positive definite matrix Ω. Then, for any $\alpha \in (0,1)$ and $\mathbf{w} \in \Delta_{K-1}$,*

$$P\left\{\|Y - \Pi_\Omega(Y \mid \Lambda(\mathbf{w}))\|_\Omega^2 > c_{1-\alpha}(Y; \mathbf{w}, \Omega)\right\} \leq \alpha,$$

where $c_{1-\alpha}(Y; \mathbf{w}, \Omega) = G^{-1}(1-\alpha; |J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w})))|)$, and $G(\cdot; k)$ is the CDF of the χ^2-distribution with degree of freedom equal to k.

Proof: Let F_ℓ, $\ell = 1, ..., L$, be the faces of the polyhedral cone $\Lambda(\mathbf{w})$, and let $\mathrm{ri}(F_\ell)$ be the relative interior of F_ℓ. Then, by Theorem 1 of Mohamad, van Zwet, Cator, and Goeman (2020),[20] we have

$$P\left\{\|Y - \Pi_\Omega(Y \mid \Lambda(\mathbf{w}))\|_\Omega^2 > q_{1-\alpha}(Y; \mathbf{w}, \Omega)\right\} \leq \alpha,$$

where

$$q_{1-\alpha}(Y; \mathbf{w}, \Omega) = \sum_{\ell=1}^{L} 1\{\Pi_\Omega(Y \mid \Lambda(\mathbf{w})) \in \mathrm{ri}(F_\ell)\} G^{-1}(1-\alpha; K - \mathrm{rk}(P_\ell)), \tag{A.7}$$

and F_ℓ's are faces of $\Lambda(\mathbf{w})$, P_ℓ denotes the projection matrix (along $\|\cdot\|_\Omega$) onto the linear span of F_ℓ, and $\mathrm{rk}(P_\ell)$ denotes the rank of P_ℓ. It suffices to show that

$$q_{1-\alpha}(Y; \mathbf{w}, \Omega) = c_{1-\alpha}(Y; \mathbf{w}, \Omega).$$

In our case with the polyhedral cone $\Lambda(\mathbf{w})$, the faces and their relative interiors are given by $\Lambda_J(\mathbf{w})$ and $\mathrm{ri}(\Lambda_J(\mathbf{w}))$, $J \subset \overline{K}$ (see, e.g., the proof of Lemma 3.13.5 of Silvapulle and Sen (2005)). Furthermore, by Lemma A.1(ii), $\Pi_\Omega(Y \mid \Lambda(\mathbf{w})) \in \mathrm{ri}(\Lambda_J(\mathbf{w}))$ implies that $J \neq \varnothing$ and $\overline{K} \setminus J \subset J_0(\mathbf{w})$. Hence, we can rewrite $q_{1-\alpha}(Y; \mathbf{w}, \Omega)$ as

$$\begin{aligned} &\sum_{J \subset \overline{K}: J \neq \varnothing, \overline{K} \setminus J \neq \varnothing} 1\{\overline{K} \setminus J \subset J_0(\mathbf{w})\} 1\{\Pi_\Omega(Y \mid \Lambda(\mathbf{w})) \in \mathrm{ri}(\Lambda_J(\mathbf{w}))\} G^{-1}(1-\alpha; K - \mathrm{rk}(P_J)) \\ &\quad + 1\{\Pi_\Omega(Y \mid \Lambda(\mathbf{w})) \in \mathrm{ri}(\Lambda_{\overline{K}}(\mathbf{w}))\} G^{-1}\left(1-\alpha; K - \mathrm{rk}(P_{\overline{K}})\right), \end{aligned} \tag{A.8}$$

where P_J is the projection matrix onto the linear span of $\Lambda_J(\mathbf{w})$.

Since $\mathrm{ri}(\Lambda_{\overline{K}}(\mathbf{w})) = \{\mathbf{0}\}$, and $P_{\overline{K}}$ is a zero matrix, the last term in (A.8) is equal to

$$1\{\Pi_\Omega(Y \mid \Lambda(\mathbf{w})) = \mathbf{0}\} G^{-1}(1-\alpha; K).$$

We focus on the first sum in (A.8). The linear span of $\Lambda_J(\mathbf{w})$ is given by $L_J(\mathbf{w})$. However, for any nonempty J such that $\overline{K} \setminus J \subset J_0(\mathbf{w})$, the span is reduced to $\left\{\mathbf{x} \in \mathbf{R}^K : \mathbf{x}_J = \mathbf{0}\right\}$, and $\mathrm{rk}(P_J) = K - |J|$. Therefore, $K - \mathrm{rk}(P_J) = |J|$ in (A.8). As seen in the proof of Lemma A.1(iii),

[20] They apply Lemma 3.13.2 of Silvapulle and Sen (2005), p.125, which uses the orthogonal decomposition of $\mathbf{R}^K$ equipped with the inner product $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}'\mathbf{b}$. However, the lemma continues to hold with any other inner product, with the definition of projections and orthogonal complements appropriately redefined.

there exists a unique $J^* \subset \overline{K}$ such that $\Pi_\Omega(Y \mid \Lambda(\mathbf{w})) \in \mathrm{ri}(\Lambda_{J^*}(\mathbf{w}))$ and $J^* = J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w})))$, which implies that $\overline{K} \setminus J^* \subset J_0(\mathbf{w})$. Hence, from (A.8),

$$\begin{aligned} q_{1-\alpha}(Y;\mathbf{w},\Omega) &= 1\{\overline{K} \setminus J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w}))) \neq \varnothing\} G^{-1}(1-\alpha; |J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w})))|) \\ &\quad + 1\{J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w}))) = \overline{K}\} G^{-1}(1-\alpha; K) \\ &= G^{-1}(1-\alpha; |J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w})))|) = c_{1-\alpha}(Y;\mathbf{w},\Omega). \end{aligned}$$

This gives the desired result. ■

Lemma A.4. *Let $Y \in \mathbf{R}^K$ be a random vector following $N(0,\Omega)$ for a symmetric positive definite matrix Ω. Then, for any $\alpha \in (0,1)$ and $\mathbf{w} \in \Delta_{K-1}$,*

$$P\left\{\|Y - \Pi_\Omega(Y \mid \Lambda(\mathbf{w}))\|_\Omega^2 \geq c_{1-\alpha}(Y;\mathbf{w},\Omega)\right\} \leq \alpha,$$

where $c_{1-\alpha}(Y;\mathbf{w},\Omega)$ is as defined in Lemma A.3.

Proof: In light of Lemma A.3, it suffices to show that

$$P\left\{\|Y - \Pi_\Omega(Y \mid \Lambda(\mathbf{w}))\|_\Omega^2 = c_{1-\alpha}(Y;\mathbf{w},\Omega)\right\} = 0.$$

The probability on the left hand side is equal to

$$\sum_{k=1}^{K} P\left\{\|Y - \Pi_\Omega(Y \mid \Lambda(\mathbf{w}))\|_\Omega^2 = G^{-1}(1-\alpha;k) \text{ and } |J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w})))| = k\right\}.$$

Note that the summation excludes $k = 0$ by Lemma A.1(iii). It suffices to show that

$$P\{\|Y - \Pi_\Omega(Y \mid \Lambda(\mathbf{w}))\|_\Omega^2 = c\} = 0, \tag{A.9}$$

for any constant $c \geq 0$.

First, we write

$$\begin{aligned} \|Y - \Pi_\Omega(Y \mid \Lambda(\mathbf{w}))\|_\Omega^2 &= \sum_{J \subset \overline{K}} \|Y - \Pi_\Omega(Y \mid \Lambda(\mathbf{w}))\|_\Omega^2 \, 1\{\Pi_\Omega(Y \mid \Lambda(\mathbf{w})) \in \mathrm{ri}(\Lambda_J(\mathbf{w}))\} \\ &= \sum_{J \subset \overline{K}} \|Y - \Pi_\Omega(Y \mid L_J(\mathbf{w}))\|_\Omega^2 \, 1\{\Pi_\Omega(Y \mid \Lambda(\mathbf{w})) \in \mathrm{ri}(\Lambda_J(\mathbf{w}))\} \\ &= \sum_{J \subset \overline{K}} \|Y - \Pi_\Omega(Y \mid L_J(\mathbf{w}))\|_\Omega^2 \, 1\{J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w}))) = J\} \, 1\{\overline{K} \setminus J \subset J_0(\mathbf{w})\}. \end{aligned}$$

The second equality follows by Lemma 3.13.2 of Silvapulle and Sen (2005). The last equality follows by Lemma A.1. By (iii) of Lemma A.1, the last sum is equal to

$$\sum_{J \subset \overline{K}: J \neq \varnothing} \|Y - \Pi_\Omega(Y \mid L_J(\mathbf{w}))\|_\Omega^2 \, 1\{J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w}))) = J\} \, 1\{\overline{K} \setminus J \subset J_0(\mathbf{w})\}.$$

Since the events $\{J_0(\Pi_\Omega(Y \mid \Lambda(\mathbf{w}))) = J\}$, $J \subset \overline{K}$, are disjoint across $J \subset \overline{K}$, it suffices to show that $\|Y - \Pi_\Omega(Y \mid L_J(\mathbf{w}))\|_\Omega^2$ is a continuous random variable for all nonempty $J \subset \overline{K}$. We let

$$Q_J(\mathbf{w}) = [[\Omega^{1/2}]_J^\top, \Omega^{1/2}\mathbf{w}]^\top,$$

where $[\Omega^{1/2}]_J$ denotes the $J \times K$ matrix of which each row corresponds to the j-th row vector of $\Omega^{1/2}$, $j \in J$. Then, by Lemma A.2,

$$\Pi_\Omega(Y \mid L_J(\mathbf{w})) = \Omega^{1/2}\Pi_I(\Omega^{-1/2}Y \mid L_J^\Omega(\mathbf{w})),$$

where $L_J^\Omega(\mathbf{w}) = \{\mathbf{x} : Q_J(\mathbf{w})\mathbf{x} = 0\}$, which is the linear span of $\{\mathbf{x} : A(\mathbf{w})\Omega^{1/2}\mathbf{x} \le \mathbf{0}\}$, with $A(\mathbf{w})$ defined in (A.3).

Now,

$$\begin{aligned}
\|Y - \Pi_\Omega(Y \mid L_J(\mathbf{w}))\|_\Omega^2 &= \left(\Omega^{-1/2}Y - \Omega^{-1/2}\Pi_\Omega(Y \mid L_J(\mathbf{w}))\right)^\top \left(\Omega^{-1/2}Y - \Omega^{-1/2}\Pi_\Omega(Y \mid L_J(\mathbf{w}))\right) \\
&= \left(\Omega^{-1/2}Y - \Pi_I(\Omega^{-1/2}Y \mid L_J^\Omega(\mathbf{w}))\right)^\top \left(\Omega^{-1/2}Y - \Pi_I(\Omega^{-1/2}Y \mid L_J^\Omega(\mathbf{w}))\right) \\
&= \left(\Omega^{-1/2}Y\right)' M_J^\Omega(\mathbf{w}) \left(\Omega^{-1/2}Y\right),
\end{aligned}$$

where $M_J^\Omega(\mathbf{w})$ is a $K \times K$ symmetric idempotent matrix of rank equal to $K - \dim(L_J^\Omega(\mathbf{w}))$. When $J = \overline{K}$, $L_J^\Omega(\mathbf{w}) = \{\mathbf{0}\}$, and hence $\dim(L_J^\Omega(\mathbf{w})) = 0$. Suppose that J is such that $|J| = K - 1$. Then, all but one entries of $\Omega^{1/2}\mathbf{x}$ in $L_J^\Omega(\mathbf{w})$ are zero. If this nonzero entry appears in the j-th entry of $\Omega^{1/2}\mathbf{x}$, the requirement $\overline{K} \setminus J \subset J_0(\mathbf{w})$ yields that $w_j = 0$. Therefore, $\dim(L_J^\Omega(\mathbf{w})) = 1$. Similarly, if J is such that $0 < |J| < K - 1$, $\dim(L_J^\Omega(\mathbf{w})) = K - |J|$ (under the condition that $\overline{K} \setminus J \subset J_0(\mathbf{w})$). Hence, for any nonempty $J \subset \overline{K}$ such that $\overline{K} \setminus J \subset J_0(\mathbf{w})$, we have

$$K - \dim(L_J^\Omega(\mathbf{w})) = |J|.$$

This means that $\|Y - \Pi_\Omega(Y \mid L_J(\mathbf{w}))\|_\Omega^2$ follows the χ^2-distribution with degree of freedom equal to $|J|$. Since $J \neq \varnothing$, $\|Y - \Pi_\Omega(Y \mid L_J(\mathbf{w}))\|_\Omega^2$ is a continuous random variable. ■

Lemma A.5. *Suppose that Ω_n is a sequence of symmetric positive definite $K \times K$ matrices such that $\Omega_n \to \Omega_0$, as $n \to \infty$, for a symmetric positive definite matrix Ω_0. Suppose also that $\mathbf{y}_n \in \mathbf{R}^K$ and $\mathbf{w}_n \in \Delta_{K-1}$ are sequences of vectors such that $\mathbf{y}_n \to \mathbf{y}_0$, as $n \to \infty$, for some $\mathbf{y}_0 \in \mathbf{R}^K$. Then, the following statements hold.*

(i) $\lim_{n\to\infty} \left\|\Pi_{\Omega_n}(\mathbf{y}_n \mid \Lambda(\mathbf{w}_n)) - \Pi_{\Omega_0}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_n))\right\| = 0.$
(ii) $\lim_{n\to\infty} \left|\left|J_0\left(\Pi_{\Omega_n}(\mathbf{y}_n \mid \Lambda(\mathbf{w}_n))\right)\right| - \left|J_0\left(\Pi_{\Omega_0}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_n))\right)\right|\right| = 0.$

Proof: (i) Note that

$$\left\|\Pi_{\Omega_n}(\mathbf{y}_n \mid \Lambda(\mathbf{w}_n)) - \Pi_{\Omega_0}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_n))\right\| \le A_{n,1} + A_{n,2}, \tag{A.10}$$

where

$$A_{n,1} = \left\| \Pi_{\Omega_n}(\mathbf{y}_n \mid \Lambda(\mathbf{w}_n)) - \Pi_{\Omega_n}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_n)) \right\|, \text{ and}$$
$$A_{n,2} = \left\| \Pi_{\Omega_n}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_n)) - \Pi_{\Omega_0}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_n)) \right\|.$$

Since a projection map in a Hilbert space on a closed convex set is a contraction map (see, e.g. Theorem 3 of Cheney and Goldstein (1959)),

$$A_{n,1} \leq \|\mathbf{y}_n - \mathbf{y}_0\|_{\Omega_n}.$$

Since $\Omega_n \to \Omega_0$ and Ω_0 is positive definite, the above bound vanishes as $n \to \infty$.

Let us turn to $A_{n,2}$. Due to the contractive property of the projection map, and since $\mathbf{0} \in \Lambda(\mathbf{w}_n)$, and $\Omega_n \to \Omega_0$, we have

$$\begin{aligned} \left\| \Pi_{\Omega_n}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_n)) \right\|_{\Omega_n} &= \left\| \Pi_{\Omega_n}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_n)) - \Pi_{\Omega_n}(\mathbf{0} \mid \Lambda(\mathbf{w}_n)) \right\|_{\Omega_n} \\ &\leq \|\mathbf{y}_0\|_{\Omega_n} \to \|\mathbf{y}_0\|_{\Omega_0}, \end{aligned}$$

as $n \to \infty$. Hence, there exists a fixed bounded, closed, convex set $B \subset \mathbf{R}^K$ which depends only on $\mathbf{y}_0$ and Ω_0 such that for all $n \geq 1$, $\Lambda(\mathbf{w}_n) \cap B \neq \varnothing$ and

$$\inf_{\mathbf{x} \in \Lambda(\mathbf{w}_n)} \|\mathbf{y}_0 - \mathbf{x}\|_{\Omega_n}^2 = \inf_{\mathbf{x} \in \Lambda(\mathbf{w}_n) \cap B} \|\mathbf{y}_0 - \mathbf{x}\|_{\Omega_n}^2 \text{ and}$$
$$\inf_{\mathbf{x} \in \Lambda(\mathbf{w}_n)} \|\mathbf{y}_0 - \mathbf{x}\|_{\Omega_0}^2 = \inf_{\mathbf{x} \in \Lambda(\mathbf{w}_n) \cap B} \|\mathbf{y}_0 - \mathbf{x}\|_{\Omega_0}^2.$$

Furthermore, note that

$$\begin{aligned} &\left| \inf_{\mathbf{x} \in \Lambda(\mathbf{w}_n) \cap B} \|\mathbf{y}_0 - \mathbf{x}\|_{\Omega_n}^2 - \inf_{\mathbf{x} \in \Lambda(\mathbf{w}_n) \cap B} \|\mathbf{y}_0 - \mathbf{x}\|_{\Omega_0}^2 \right| \\ &\leq \left| \inf_{\mathbf{x} \in \Lambda(\mathbf{w}_n) \cap B} \left(\|\mathbf{y}_0 - \mathbf{x}\|_{\Omega_n}^2 - \|\mathbf{y}_0 - \mathbf{x}\|_{\Omega_0}^2 + \|\mathbf{y}_0 - \mathbf{x}\|_{\Omega_0}^2 - \inf_{\tilde{\mathbf{x}} \in \Lambda(\mathbf{w}_n) \cap B} \|\mathbf{y}_0 - \tilde{\mathbf{x}}\|_{\Omega_0}^2 \right) \right|. \end{aligned}$$

The last term is bounded by

$$\sup_{\mathbf{x} \in \Lambda(\mathbf{w}_n) \cap B} \left| (\mathbf{y}_0 - \mathbf{x})^\top (\Omega_n^{-1} - \Omega_0^{-1})(\mathbf{y}_0 - \mathbf{x}) \right| \to 0,$$

as $n \to \infty$. Since $\Lambda(\mathbf{w}_n) \cap B$ is a closed convex set, the projection of $\mathbf{y}_0$ onto $\Lambda(\mathbf{w}_n) \cap B$ along $\|\cdot\|_{\Omega_n}$ exists and is unique. Hence, we find that

$$\lim_{n \to \infty} A_{n,2} = 0.$$

(ii) By the result of (i), as $n \to \infty$,

$$\Pi_{\Omega_n}(\mathbf{y}_n \mid \Lambda(\mathbf{w}_n)) - \Pi_{\Omega_0}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_n)) \to 0. \tag{A.11}$$

Recall that the relative interiors $\mathrm{ri}(\Lambda_J(\mathbf{w}))$, $J \subset \overline{K}$, partition $\Lambda(\mathbf{w})$. For any subsequence of $\{n\}$, we choose a further subsequence $\{n'\}$ and $J, J' \subset \overline{K}$ such that for all n' in the subsequence, we have

$$\Pi_{\Omega_{n'}}(\mathbf{y}_{n'} \mid \Lambda(\mathbf{w}_{n'})) \in \mathrm{ri}(\Lambda_J(\mathbf{w}_{n'})), \text{ and}$$
$$\Pi_{\Omega_0}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_{n'})) \in \mathrm{ri}(\Lambda_{J'}(\mathbf{w}_{n'})).$$

Since $\mathrm{ri}(\Lambda_J(\mathbf{w}_{n'}))$ and $\mathrm{ri}(\Lambda_{J'}(\mathbf{w}_{n'}))$ are nonempty, by Lemma A.1, we have

$$\mathrm{ri}(\Lambda_J(\mathbf{w}_{n'})) = \{\mathbf{x} \in \mathbf{R}^K : \mathbf{x}_J = 0, \mathbf{x}_{-J} < 0\}, \tag{A.12}$$

and similarly with $\mathrm{ri}(\Lambda_{J'}(\mathbf{w}_{n'}))$. Hence both the relative interiors do not depend on $\mathbf{w}_{n'}$ or n'. Furthermore, from this, we have

$$J_0(\Pi_{\Omega_{n'}}(\mathbf{y}_{n'} \mid \Lambda(\mathbf{w}_{n'}))) = J \text{ and } J_0(\Pi_{\Omega_0}(\mathbf{y}_0 \mid \Lambda(\mathbf{w}_{n'}))) = J'.$$

Now, by (A.11) and (A.12), we find that from large n' on, we have

$$\Pi_{\Omega_{n'}}(\mathbf{y}_{n'} \mid \Lambda(\mathbf{w}_{n'})) \in \mathrm{ri}(\Lambda_{J'}(\mathbf{w}_{n'})).$$

Since the relative interiors $\mathrm{ri}(\Lambda_J(\mathbf{w}_{n'}))$, $J \subset \overline{K}$, partition $\Lambda(\mathbf{w}_{n'})$, we find that $J = J'$ from some large n' on. ■

Lemma A.6. *Suppose that $Y_n \in \mathbf{R}^K$, $n \geq 1$, is a sequence of random vectors, and $\mathbf{w}_n \in \Delta_{K-1}$ is a sequence of nonstochastic vectors, such that $Y_n \to_d Y$, where Y follows $N(0, \Omega_0)$ for some symmetric positive definite matrix Ω_0. Furthermore, let Ω_n be a sequence of symmetric positive definite random matrices such that $\Omega_n \to_p \Omega_0$, as $n \to \infty$.*

Then, for any $\alpha \in (0,1)$,

$$\limsup_{n\to\infty} P\left\{\left\|Y_n - \Pi_{\Omega_n}(Y_n \mid \Lambda(\mathbf{w}_n))\right\|^2_{\Omega_n} > c_{1-\alpha}(Y_n; \mathbf{w}_n, \Omega_n)\right\} \leq \alpha,$$

where $c_{1-\alpha}(Y_n; \mathbf{w}_n, \Omega_n) = G^{-1}\left(1-\alpha; |J_0(\Pi_{\Omega_n}(Y_n \mid \Lambda(\mathbf{w}_n)))|\right)$.

Proof: Due to the almost sure representation theorem (cf. Theorem 6.7 of Billingsley (1999), p.70), there is a common probability space on which we have a sequence of random vectors $\tilde{Y}_n$ and random matrices $\tilde{\Omega}_n$ such that

$$[\tilde{Y}_n^\top, \mathrm{vec}(\tilde{\Omega}_n)^\top]^\top \to_{a.s} [\tilde{Y}^\top, \mathrm{vec}(\tilde{\Omega})^\top]^\top,$$

as $n \to \infty$, where $\tilde{Y}_n$ and $\tilde{\Omega}_n$ have the same distribution as Y_n and Ω_n, and $\tilde{Y}$ and $\tilde{\Omega}_0$ have the same distribution as Y and Ω_0.

By Lemma A.5(i), we have

$$\left\|\tilde{Y}_n - \Pi_{\tilde{\Omega}_n}(\tilde{Y}_n \mid \Lambda(\mathbf{w}_n))\right\|^2_{\tilde{\Omega}_n} - \left\|\tilde{Y} - \Pi_{\tilde{\Omega}_0}(\tilde{Y} \mid \Lambda(\mathbf{w}_n))\right\|^2_{\tilde{\Omega}_0} \to_{a.s} 0,$$

as $n \to \infty$.

Let us turn to the critical values. By Lemma A.5(ii), we have

$$\lim_{n\to\infty} \left| \left| J_0\left(\Pi_{\tilde{\Omega}_n}(\tilde{Y}_n \mid \Lambda(\mathbf{w}_n))\right)\right| - \left|J_0\left(\Pi_{\tilde{\Omega}_0}(\tilde{Y} \mid \Lambda(\mathbf{w}_n))\right)\right| \right| = 0.$$

Hence,

$$\begin{aligned} c_{1-\alpha}(\tilde{Y}_n; \mathbf{w}_n, \tilde{\Omega}_n) &= G^{-1}\left(1-\alpha; |J_0(\Pi_{\tilde{\Omega}_n}(\tilde{Y}_n \mid \Lambda(\mathbf{w}_n)))|\right) \\ &= G^{-1}\left(1-\alpha; |J_0(\Pi_{\tilde{\Omega}_0}(\tilde{Y} \mid \Lambda(\mathbf{w}_n)))|\right) + o_{a.s.}(1) \equiv c_{1-\alpha}(\tilde{Y}; \mathbf{w}_n, \tilde{\Omega}_0) + o_{a.s.}(1), \end{aligned}$$

as $n \to \infty$. Thus, we find that

$$\begin{aligned} &\left\|\tilde{Y}_n - \Pi_{\tilde{\Omega}_n}(\tilde{Y}_n \mid \Lambda(\mathbf{w}_n))\right\|^2_{\tilde{\Omega}_n} - c_{1-\alpha}(\tilde{Y}_n; \mathbf{w}_n, \tilde{\Omega}_n) \\ &\quad - \left(\|\tilde{Y} - \Pi_{\tilde{\Omega}_0}(\tilde{Y} \mid \Lambda(\mathbf{w}_n))\|^2_{\tilde{\Omega}_0} - c_{1-\alpha}(\tilde{Y}; \mathbf{w}_n, \tilde{\Omega}_0)\right) \to_{a.s.} 0, \end{aligned}$$

as $n \to \infty$. Now, observe that

$$\begin{aligned} &P\left\{\left\|\tilde{Y}_n - \Pi_{\tilde{\Omega}_n}(\tilde{Y}_n \mid \Lambda(\mathbf{w}_n))\right\|^2_{\tilde{\Omega}_n} - c_{1-\alpha}(\tilde{Y}_n; \mathbf{w}_n, \tilde{\Omega}_n) > 0\right\} \\ &\le P\left\{\left\|\tilde{Y}_n - \Pi_{\tilde{\Omega}_n}(\tilde{Y}_n \mid \Lambda(\mathbf{w}_n))\right\|^2_{\tilde{\Omega}_n} - c_{1-\alpha}(\tilde{Y}_n; \mathbf{w}_n, \tilde{\Omega}_n) \ge 0\right\} \\ &= P\left\{\left\|\tilde{Y} - \Pi_{\tilde{\Omega}_0}(\tilde{Y} \mid \Lambda(\mathbf{w}_n))\right\|^2_{\tilde{\Omega}_0} - c_{1-\alpha}(\tilde{Y}; \mathbf{w}_n, \tilde{\Omega}_0) + o_{a.s.}(1) \ge 0\right\} \\ &\le P\left\{\left\|\tilde{Y} - \Pi_{\tilde{\Omega}_0}(\tilde{Y} \mid \Lambda(\mathbf{w}_n))\right\|^2_{\tilde{\Omega}_0} - c_{1-\alpha}(\tilde{Y}; \mathbf{w}_n, \tilde{\Omega}_0) \ge 0\right\} + o(1) \le \alpha + o(1), \end{aligned}$$

as $n \to \infty$. The second inequality follows by reversed Fatou's Lemma and from the fact that the map $1\{\cdot \ge 0\}$ is upper semicontinuous. The last inequality follows by Lemma A.4. ■

A.2.2. **The Proof of the Main Results.** Throughout the proofs below, we assume that Assumptions A.1-A.5 are satisfied. Define

$$\hat{\mathbf{G}}_P = \sqrt{n_0}(\hat{H} - H_P) \text{ and } \hat{\mathbf{g}}_P = \sqrt{n_0}(\hat{\mathbf{h}} - \mathbf{h}_P).$$

The following lemma gives an asymptotic linear presentation for $\hat{\mathbf{G}}_P$ and $\hat{\mathbf{g}}_P$.

Lemma A.7. *As $n_0 \to \infty$,*

$$\hat{\mathbf{G}}_P = \frac{1}{\sqrt{n_0}} \sum_{i \in N} \Psi_{i,P} + o_{\mathcal{P}}(1), \text{ and } \hat{\mathbf{g}}_P = \frac{1}{\sqrt{n_0}} \sum_{i \in N} \boldsymbol{\psi}_{i,P} + o_{\mathcal{P}}(1),$$

where $\Psi_{i,P}$ is the $K \times K$ matrix whose (j,k)-entry is given by

$$\begin{aligned} \psi_{i,P,jk} = &\sqrt{\frac{n_0}{n_j}} \psi_{j,0,P}(W_i; q_{k,0,P}) 1\{i \in N_j\} + \sqrt{\frac{n_0}{n_k}} \psi_{k,0,P}(W_i; q_{j,0,P}) 1\{i \in N_k\} \\ &+ \left\{q_{j,0,P}(W_i) q_{k,0,P}(W_i) - \mathbf{E}_P\left[q_{j,0,P}(W_i) q_{k,0,P}(W_i)\right]\right\} 1\{i \in N_0\}, \end{aligned}$$

and $\boldsymbol{\psi}_{i,P}$ *is the* $K \times 1$ *vector whose* k*-th entry is given by*

$$\begin{aligned}\psi_{i,P,k} = \sqrt{\frac{n_0}{n_k}}\psi_{k,0,P}(W_i; q_{0,0,P})1\{i \in N_k\} + \psi_{0,0,P}(W_i; q_{0,0,P})1\{i \in N_0\} \\ + \left\{q_{k,0,P}(W_i)q_{0,0,P}(W_i) - \mathbf{E}_P\left[q_{k,0,P}(W_i)q_{0,0,P}(W_i)\right]\right\}1\{i \in N_0\}.\end{aligned}$$

Proof: For $j,k = 1,...,K$, let $\hat{H}_{jk}$ be the (j,k)-th entry of $\hat{H}$ and $H_{P,jk}$ the (j,k)-th entry of H_P. As for the first statement, for each $j,k = 1,...,K$, we write

$$\begin{aligned}\sqrt{n_0}(\hat{H}_{jk} - H_{P,jk}) = \frac{1}{\sqrt{n_0}}\sum_{i\in N_0}(\hat{q}_{j,0}(W_i) - q_{j,0,P}(W_i))\hat{q}_{k,0}(W_i) \\ + \frac{1}{\sqrt{n_0}}\sum_{i\in N_0}(\hat{q}_{k,0}(W_i) - q_{k,0,P}(W_i))q_{j,0,P}(W_i) \\ + \frac{1}{\sqrt{n_0}}\sum_{i\in N_0}\left\{q_{j,0,P}(W_i)q_{k,0,P}(W_i) - \mathbf{E}_P\left[q_{j,0,P}(W_i)q_{k,0,P}(W_i)\right]\right\}.\end{aligned}$$

By Assumption A.4, we find that

$$\begin{aligned}\sqrt{n_0}(\hat{H}_{jk} - H_{P,jk}) = \sqrt{\frac{n_0}{n_j}}\sum_{i\in N_j}\psi_{j,0,P}(W_i; q_{k,0,P}) + \sqrt{\frac{n_0}{n_k}}\sum_{i\in N_k}\psi_{k,0,P}(W_i; q_{j,0,P}) \\ + \frac{1}{\sqrt{n_0}}\sum_{i\in N_0}\left\{q_{j,0,P}(W_i)q_{k,0,P}(W_i) - \mathbf{E}_P\left[q_{j,0,P}(W_i)q_{k,0,P}(W_i)\right]\right\} + o_{\mathcal{P}}(1).\end{aligned}$$

The proof for the second statement is similar and is omitted. ■

Lemma A.8. *As* $n_0 \to \infty$, $\hat{H} = H_P + o_{\mathcal{P}}(1)$ *and* $\hat{\mathbf{h}} = \mathbf{h}_P + o_{\mathcal{P}}(1)$.

Proof: Since $\sup_{P\in\mathcal{P}}\mathbf{E}_P\left[\left\|\Psi_{i,P}\right\|^2\right] < \infty$ and $\sup_{P\in\mathcal{P}}\mathbf{E}_P\left[\left\|\boldsymbol{\psi}_{i,P}\right\|^2\right] < \infty$, the result is immediate from Lemma A.7. ■

For each $\mathbf{w} \in \mathbf{R}^K$, we define

$$\begin{aligned}\hat{\mathcal{M}}(\mathbf{w}) &= \left(\mathbf{w} - \hat{H}^{-1}\hat{\mathbf{h}}\right)^\top \hat{H}\left(\mathbf{w} - \hat{H}^{-1}\hat{\mathbf{h}}\right), \text{ and} \\ \mathcal{M}_P(\mathbf{w}) &= \left(\mathbf{w} - H_P^{-1}\mathbf{h}_P\right)^\top H_P\left(\mathbf{w} - H_P^{-1}\mathbf{h}_P\right).\end{aligned}$$

Lemma A.9. *As* $n_0 \to \infty$, $\hat{\mathbf{w}} = \mathbf{w}_P + o_{\mathcal{P}}(1)$.

Proof: First, we prove the following two claims.

(i) For each $\epsilon > 0$,

$$\lim_{n_0\to\infty}\sup_{P\in\mathcal{P}} P\left\{\sup_{\mathbf{w}\in\Delta_{K-1}}|\hat{\mathcal{M}}(\mathbf{w}) - \mathcal{M}_P(\mathbf{w})| > \epsilon\right\} = 0.$$

(ii) For each $\epsilon > 0$,

$$\liminf_{n_0 \to \infty} \inf_{P \in \mathcal{P}} \inf_{\mathbf{w} \in \Delta_{K-1} \setminus B(\mathbf{w}_P : \epsilon)} \{\mathcal{M}_P(\mathbf{w}) - \mathcal{M}_P(\mathbf{w}_P)\} > 0,$$

where $B(\mathbf{w}_P; \epsilon) = \{\mathbf{w} \in \Delta_{K-1} : \|\mathbf{w} - \mathbf{w}_P\| < \epsilon\}$.

Once we have (i) and (ii), we follow the arguments in the proof of Theorem 2.1 of Newey and McFadden (1994) to complete the proof. More specifically, we invoke (ii) and take $\epsilon > 0$, $\eta_\epsilon > 0$ and n_ϵ such that for all $n \geq n_\epsilon$,

$$\inf_{P \in \mathcal{P}} \inf_{\mathbf{w} \in \Delta_{K-1} \setminus B(\mathbf{w}_P : \epsilon)} \{\mathcal{M}_P(\mathbf{w}) - \mathcal{M}_P(\mathbf{w}_P)\} > \eta_\epsilon.$$

The event of $\|\hat{\mathbf{w}} - \mathbf{w}_P\| > \epsilon$ implies $\mathcal{M}_P(\hat{\mathbf{w}}) - \mathcal{M}_P(\mathbf{w}_P) > \eta_\epsilon$, or

$$\hat{\mathcal{M}}(\mathbf{w}_P) - \mathcal{M}_P(\mathbf{w}_P) > \hat{\mathcal{M}}(\hat{\mathbf{w}}) - \mathcal{M}_P(\hat{\mathbf{w}}) + \eta_\epsilon,$$

where we use that $\hat{\mathcal{M}}(\hat{\mathbf{w}}) \leq \hat{\mathcal{M}}(\mathbf{w}_P)$. The probability of this event is bounded by

$$\sup_{P \in \mathcal{P}} P\left\{2 \sup_{\mathbf{w} \in \Delta_{K-1}} |\hat{\mathcal{M}}(\mathbf{w}) - \mathcal{M}_P(\mathbf{w})| > \eta_\epsilon\right\} \to 0,$$

as $n \to \infty$, by (i). Since the last convergence is uniform in $P \in \mathcal{P}$, we obtain the desired result of the lemma.

Let us prove (i) first. For each $\mathbf{w} \in \Delta_{K-1}$, we write

$$\hat{\mathcal{M}}(\mathbf{w}) - \mathcal{M}_P(\mathbf{w}) = \mathbf{w}^\top (\hat{H} - H_P)\mathbf{w} - 2(\hat{\mathbf{h}} - \mathbf{h}_P)^\top \mathbf{w} + \hat{\mathbf{h}}^\top \hat{H}^{-1} \hat{\mathbf{h}} - \mathbf{h}_P^\top H_P^{-1} \mathbf{h}_P.$$

The desired result of (i) follows by Lemma A.8 and Assumption A.3.

Let us turn to (ii). Note that

$$\begin{aligned} \mathcal{M}_P(\mathbf{w}) - \mathcal{M}_P(\mathbf{w}_P) &= (\mathbf{w} - \mathbf{w}_P)^\top H_P (\mathbf{w} - \mathbf{w}_P) + 2(\mathbf{w} - \mathbf{w}_P)^\top H_P (\mathbf{w}_P - H_P^{-1} \mathbf{h}_P) \\ &\geq \inf_{P \in \mathcal{P}} \lambda_{\min}(H_P) \|\mathbf{w} - \mathbf{w}_P\|^2, \end{aligned} \tag{A.13}$$

because $(\mathbf{w} - \mathbf{w}_P)^\top H_P (\mathbf{w}_P - H_P^{-1} \mathbf{h}_P) \geq 0$ for all $\mathbf{w} \in \Delta_{K-1}$ by the definition of $\mathbf{w}_P$. (See, e.g., Propositions 2.1.5 and 2.3.2 of Clarke (1990).) The desired result follows from Assumption A.3. ■

Define

$$\hat{\mathbf{G}}^* = \sqrt{n_0}(\hat{H}^* - \hat{H}) \text{ and } \hat{\mathbf{g}}^* = \sqrt{n_0}(\hat{\mathbf{h}}^* - \hat{\mathbf{h}}).$$

Lemma A.10. *As $n_0 \to \infty$,*

$$\hat{\mathbf{G}}^* = \frac{1}{\sqrt{n_0}} \sum_{i \in N} \Psi_{i,P}^* + o_{\mathcal{P}}(1) \text{ and } \hat{\mathbf{g}}_P^* = \frac{1}{\sqrt{n_0}} \sum_{i \in N} \boldsymbol{\psi}_{i,P}^* + o_{\mathcal{P}}(1),$$

where $\Psi^*_{i,P}$ *is the* $K \times K$ *matrix whose* (j,k)*-entry is given by*

$$\psi^*_{i,P,jk} = \sqrt{\frac{n_0}{n_j}}\hat{\psi}_{j,0,P}(W^*_i; q_{k,0,P})1\{i \in N_j\} + \sqrt{\frac{n_0}{n_k}}\hat{\psi}_{k,0,P}(W^*_i; q_{j,0,P})1\{i \in N_k\}$$
$$+ \left\{ q_{j,0,P}(W^*_i)q_{k,0,P}(W^*_i) - \frac{1}{n_0}\sum_{i \in N_0} q_{j,0,P}(W_i)q_{k,0,P}(W_i) \right\} 1\{i \in N_0\},$$

and $\boldsymbol{\psi}^*_{i,P}$ *is the* $K \times 1$ *vector whose* k*-th entry is given by*

$$\psi^*_{i,P,k} = \sqrt{\frac{n_0}{n_k}}\hat{\psi}_{k,0,P}(W^*_i; q_{0,0,P})1\{i \in N_k\} + \hat{\psi}_{0,0,P}(W^*_i; q_{0,0,P})1\{i \in N_0\}$$
$$+ \left\{ q_{k,0,P}(W^*_i)q_{0,0,P}(W^*_i) - \frac{1}{n_0}\sum_{i \in N_0} q_{k,0,P}(W_i)q_{0,0,P}(W_i) \right\} 1\{i \in N_0\}.$$

Proof: The proof is similar to that of Lemma A.7. Since the arguments are standard, we provide a sketch of the proof of the first statement only for brevity. Let $\hat{H}^*_{jk}$ be the (j,k)-th entry of $\hat{H}^*$. We write

$$\sqrt{n_0}(\hat{H}^*_{jk} - \hat{H}_{jk}) = A_{n,1} + A_{n,2},$$

where

$$A_{n,1} = \frac{1}{\sqrt{n_0}}\sum_{i \in N_0}(\hat{q}^*_{j,0}(W^*_i) - \hat{q}_{j,0}(W^*_i))\hat{q}^*_{k,0}(W^*_i) + \frac{1}{\sqrt{n_0}}\sum_{i \in N_0}(\hat{q}^*_{k,0}(W^*_i) - \hat{q}_{k,0}(W^*_i))\hat{q}_{j,0}(W^*_i), \text{ and}$$

$$A_{n,2} = \frac{1}{\sqrt{n_0}}\sum_{i \in N_0}\left\{ \hat{q}_{j,0}(W^*_i)\hat{q}_{k,0}(W^*_i) - \frac{1}{n_0}\sum_{i \in N_0}\hat{q}_{j,0}(W_i)\hat{q}_{k,0}(W_i) \right\}.$$

From Assumptions A.4-A.5, we can show that

$$A_{n,1} = \sqrt{\frac{n_0}{n_j}}\sum_{i \in N}\hat{\psi}_{j,0,P}(W^*_i; q_{k,0,P})1\{i \in N_j\} + \sqrt{\frac{n_0}{n_k}}\sum_{i \in N}\hat{\psi}_{k,0,P}(W^*_i; q_{j,0,P})1\{i \in N_k\} + o_{\mathcal{P}}(1).$$

By Assumption A.5(ii),

$$A_{n,2} = \frac{1}{\sqrt{n_0}}\sum_{i \in N_0}\left\{ q_{j,0,P}(W^*_i)q_{k,0,P}(W^*_i) - \frac{1}{n_0}\sum_{i \in N_0} q_{j,0,P}(W_i)q_{k,0,P}(W_i) \right\} + o_{\mathcal{P}}(1).$$

Thus, we obtain the desired result. ■

Recall the definition of $\Omega_{n,P}$ in (A.1). We construct its bootstrap version. Define

$$\tilde{\boldsymbol{\psi}}^*_{i,P} = \Psi^*_{i,P}\mathbf{w}_P - \boldsymbol{\psi}^*_{i,P},$$

where $\Psi^*_{i,P}$ and $\boldsymbol{\psi}^*_{i,P}$ are defined in Lemma A.10. We let

$$\tilde{\Omega}_{n,P} = \frac{1}{n_0}\sum_{i\in N}\mathbf{E}\left[\tilde{\boldsymbol{\psi}}^*_{i,P_n}\tilde{\boldsymbol{\psi}}^{*\top}_{i,P} \mid \mathcal{F}_n\right],$$

where $\mathcal{F}_n$ denotes the σ-field generated by $(W_i)_{i\in N}$.

Lemma A.11. *Suppose that $N_n = \{1,...,n\}$ is partitioned as*

$$N_n = \bigcup_{k=0}^{K} N_{n,j},$$

where we denote $n_{k,n} = |N_{n,j}|$, $k = 0,1,...,K$, and $n_{k,n}/n_{0,n} \to r_k$ for the constant $r_k > 0$ in Assumption A.1, and $n_{k,n}$ denotes the sample size in the region k. Then, the following statements hold for any sequence of probabilities $P_n \in \mathcal{P}$.

(i) As $n \to \infty$,

$$\sup_{t\in\mathbf{R}}\left|P_n\left\{\Omega^{-1/2}_{n,P_n}\frac{1}{\sqrt{n_{0,n}}}\sum_{i\in N_n}\tilde{\boldsymbol{\psi}}_{i,P_n} \le t\right\} - \Phi(t)\right| \to 0,$$

where Φ is the CDF of $N(0,1)$.

(ii) For any $\epsilon > 0$, as $n \to \infty$,

$$P_n\left\{\sup_{t\in\mathbf{R}}\left|P_n\left\{\tilde{\Omega}^{-1/2}_{n,P_n}\frac{1}{\sqrt{n_{0,n}}}\sum_{i\in N_n}\tilde{\boldsymbol{\psi}}^*_{i,P_n} \le t \mid \mathcal{F}_n\right\} - \Phi(t)\right| > \epsilon\right\} \to 0.$$

Proof: Both results follow from standard arguments involving the Central Limit Theorem and its bootstrap version for a sum of independent random variables. (See Chapter 3 of Shao and Tu (1995).) ■

Lemma A.12. *As $n_0 \to \infty$, $\hat{\Omega} = \Omega_{n,P} + o_{\mathcal{P}}(1)$.*

Proof: It suffices to show that as $n_0 \to \infty$,

$$\Omega_{n,P} = \tilde{\Omega}_{n,P} + o_{\mathcal{P}}(1) \text{ and } \hat{\Omega} = \tilde{\Omega}_{n,P} + o_{\mathcal{P}}(1).$$

The first statement is easy to show. For brevity, we focus on showing the second statement. We write $\hat{\Omega}_n$ instead of $\hat{\Omega}$ to make the sample size explicit. We choose a subsequence $\{n'\}$ of $\{n\}$ such that for $P_{n'} \in \mathcal{P}$,

$$\liminf_{n_0\to\infty}\sup_{P\in\mathcal{P}} P\left\{\|\hat{\Omega}_n - \tilde{\Omega}_{n,P}\| > \epsilon\right\} = \liminf_{n'\to\infty} P_{n'}\left\{\|\hat{\Omega}_{n'} - \tilde{\Omega}_{n',P}\| > \epsilon\right\}.$$

Then, there exists a further subsequence $\{n''\}$ of $\{n'\}$ such that

$$\liminf_{n'\to\infty} P_{n'}\left\{\|\hat{\Omega}_{n'} - \tilde{\Omega}_{n',P}\| > \epsilon\right\} = \lim_{n''\to\infty} P_{n''}\left\{\|\hat{\Omega}_{n''} - \tilde{\Omega}_{n'',P}\| > \epsilon\right\}.$$

Recall that $\hat{\tau}_k = \sqrt{n_0}\max\left\{\left|\left[\hat{H}\hat{\mathbf{w}} - \hat{\mathbf{h}}\right]_k\right|, c_0\right\}$. Thus, from Lemma A.7, we have

$$\lim_{M\to\infty}\liminf_{n''\to\infty} P_{n''}\{\|\hat{\boldsymbol{\tau}}\| > M\} = 0. \tag{A.14}$$

Since $\mathbf{E}\left[|\tilde{\gamma}_k^*|^{2+\delta} \mid \mathcal{F}_n\right] \leq \hat{\tau}_k^{2+\delta}$ for any $k = 1, ..., K$ and for any $\delta > 0$,

$$\lim_{M\to\infty}\liminf_{n''\to\infty} P_{n''}\left\{\mathbf{E}\left[\|\tilde{\boldsymbol{\gamma}}^*\|^{2+\delta} \mid \mathcal{F}_{n''}\right] > M\right\} = 0.$$

Hence, for each $k, \ell = 1, ..., K$, and $\epsilon > 0$,

$$\begin{aligned}
&\lim_{M\to\infty}\liminf_{n''\to\infty} P_{n''}\left\{\mathbf{E}\left[|\tilde{\gamma}_k^*\tilde{\gamma}_\ell^*|1\{|\tilde{\gamma}_k^*\tilde{\gamma}_\ell^*| > M\} \mid \mathcal{F}_{n''}\right] > \epsilon\right\} \\
&\leq \lim_{M\to\infty}\liminf_{n''\to\infty} P_{n''}\left\{\mathbf{E}\left[|\tilde{\gamma}_k^*\tilde{\gamma}_\ell^*|^{1+\delta} \mid \mathcal{F}_{n''}\right] > \epsilon M^\delta\right\} = 0.
\end{aligned}$$

Therefore, $\mathbf{E}\left[\|\tilde{\boldsymbol{\gamma}}^*\|^{2+\delta} \mid \mathcal{F}_{n''}\right]$ is asymptotically uniformly integrable uniformly over $P \in \mathcal{P}$. From Lemma A.11, we have along the subsequence $P_{n''}$,

$$\tilde{\Omega}_{n'',P_{n''}}^{-1/2}\left(\frac{1}{B}\sum_{b=1}^{B}\tilde{\boldsymbol{\gamma}}_b^*\tilde{\boldsymbol{\gamma}}_b^{*\top}\right)\tilde{\Omega}_{n'',P_{n''}}^{-1/2} = \tilde{\Omega}_{n'',P_{n''}}^{-1/2}\mathbf{E}\left[\tilde{\boldsymbol{\gamma}}_b^*\tilde{\boldsymbol{\gamma}}_b^{*\top}|\mathcal{F}_{n''}\right]\tilde{\Omega}_{n'',P_{n''}}^{-1/2} + o_P(1),$$

as $B \to \infty$ and then $n'' \to \infty$. Furthermore, from (A.14),

$$\tilde{\Omega}_{n'',P_{n''}}^{-1/2}\mathbf{E}\left[\tilde{\boldsymbol{\gamma}}_b^*\tilde{\boldsymbol{\gamma}}_b^{*\top}|\mathcal{F}_{n''}\right]\tilde{\Omega}_{n'',P_{n''}}^{-1/2} = \tilde{\Omega}_{n'',P_{n''}}^{-1/2}\mathbf{E}\left[\tilde{\boldsymbol{\psi}}_{i,P_{n''}}^*\tilde{\boldsymbol{\psi}}_{i,P_{n''}}^{*\top}|\mathcal{F}_{n''}\right]\tilde{\Omega}_{n'',P_{n''}}^{-1/2} + o_P(1).$$

From the arguments in the proof of Theorem 2.20 of van der Vaart (1998), we find that

$$\tilde{\Omega}_{n'',P_{n''}}^{-1/2}\left(\frac{1}{B}\sum_{b=1}^{B}\tilde{\boldsymbol{\gamma}}_b^*\tilde{\boldsymbol{\gamma}}_b^{*\prime}\right)\tilde{\Omega}_{n'',P_{n''}}^{-1/2} \to_P I_K,$$

as $n'' \to \infty$. ■

Lemma A.13. *For any $\kappa \in (0,1)$, we have*

$$\liminf_{n_0\to\infty}\inf_{P\in\mathcal{P}} P\left\{\mathbf{w}_P \in \tilde{C}_{1-\kappa}\right\} \geq 1-\kappa.$$

Proof: For any subsequence of $\{n_0\}$, we choose a further subsequence in Lemma A.11. Then, we apply Lemma A.6 on the subsequence, with

$$Y_n = \hat{\mathbf{G}}_{P_n}\mathbf{w}_{P_n} - \hat{\mathbf{g}}_{P_n},$$

and $Y = \Omega_P^{1/2}\mathbb{Z}$, and with Ω_n and Ω in the lemma replaced by $\hat{\Omega}_n$ and Ω_P. Then, since $Y_n \to_d Y$ and $\hat{\Omega}_n \to_P \Omega_P$ by Lemmas A.11 and A.12 and (A.2), we obtain the desired result. ■

Lemma A.14. *As $n_0 \to \infty$,*

$$\sup_{\mathbf{w}\in\Delta_{K-1}}\left|\sqrt{n_0}\left(\hat{\theta}(\mathbf{w}) - \theta_P(\mathbf{w})\right) - \sum_{k=1}^{K} w_k\frac{1}{\sqrt{n_0}}\sum_{i\in N}\psi_{k,P}^{\theta}(W_i)\right| = o_{\mathcal{P}}(1),$$

where

$$\psi^{\theta}_{k,P}(W_i) = \left(\psi_{0,0,P}(W_i;1) + q_{0,0,P}(W_i) - \mathbf{E}_P\left[q_{0,0,P}(W_i)\right]\right)1\{i \in N_0\}$$
$$+ \left(q_{k,1,P}(W_i) - \mathbf{E}_P\left[q_{k,1,P}(W_i)\right]\right)1\{i \in N_0\} + \sqrt{\frac{n_0}{n_k}}\psi_{k,1,P}(W_i;1)1\{i \in N_k\}.$$

Proof: We write

$$\sqrt{n_0}\left(\hat{\theta}(\mathbf{w}) - \theta_P(\mathbf{w})\right) = \frac{1}{\sqrt{n_0}}\sum_{i\in N_0}\left(\hat{q}_{0,0}(W_i) - q_{0,0,P}(W_i)\right) + \frac{1}{\sqrt{n_0}}\sum_{i\in N_0}\left(q_{0,0,P}(W_i) - \mathbf{E}_P\left[q_{0,0,P}(W_i)\right]\right)$$
$$+ \frac{1}{\sqrt{n_0}}\sum_{k=1}^{K} w_k \sum_{i\in N_0}\left\{\hat{q}_{k,1}(W_i) - q_{k,1,P}(W_i) + q_{k,1,P}(W_i) - \mathbf{E}_P\left[q_{k,1,P}(W_i)\right]\right\}.$$

By Assumption A.4 and by the fact that $\sum_{k=1}^{K} w_k = 1$, we find

$$\sqrt{n_0}\left(\hat{\theta}(\mathbf{w}) - \theta_P(\mathbf{w})\right) = \frac{1}{\sqrt{n_0}}\sum_{i\in N_0}\psi_{0,0,P}(W_i;1) + \frac{1}{\sqrt{n_0}}\sum_{i\in N_0}\left(q_{0,0,P}(W_i) - \mathbf{E}_P\left[q_{0,0,P}(W_i)\right]\right)$$
$$+ \sum_{k=1}^{K} w_k \frac{1}{\sqrt{n_0}}\sum_{i\in N_0}\left(q_{k,1,P}(W_i) - \mathbf{E}_P\left[q_{k,1,P}(W_i)\right]\right)$$
$$+ \sum_{k=1}^{K} w_k \sum_{i\in N_k}\sqrt{\frac{n_0}{n_k}}\left\{\psi_{k,1,P}(W_i;1) + q_{k,1,P}(W_i) - \mathbf{E}_P\left[q_{k,1,P}(W_i)\right]\right\} + o_{\mathcal{P}}(1)$$
$$= \sum_{k=1}^{K} w_k \frac{1}{\sqrt{n_0}}\sum_{i\in N}\psi^{\theta}_{k,P}(W_i) + o_{\mathcal{P}}(1).$$

We obtain the desired result. ■

Define

$$\sigma^2 = \sum_{k=1}^{K} w_k^2 \frac{1}{n_0}\sum_{i\in N}\mathbf{E}\left[\left(\psi^{\theta}_{k,P}(W_i)\right)^2\right].$$

Lemma A.15. *As* $n_0 \to \infty$,

$$\sup_{P\in\mathcal{P}}\sup_{t\in\mathbf{R}}\left|P\left\{\frac{\sqrt{n_0}\left(\hat{\theta}(\mathbf{w}_P) - \theta_P(\mathbf{w}_P)\right)}{\sigma} \le t\right\} - \Phi(t)\right| \to 0,$$

where Φ *is the CDF of* $N(0,1)$.

Proof: The result follows from Lemma A.14 and the Central Limit Theorem for independent random variables. ■

Lemma A.16. *As* $n_0 \to \infty$, $\hat{\sigma} = \sigma + o_{\mathcal{P}}(1)$.

Proof: The result follows from the uniform asymptotic normality of the bootstrap version $\sqrt{n_0}\big(\hat{\theta}^*(\hat{\mathbf{w}})-\hat{\theta}(\hat{\mathbf{w}})\big)$. The arguments are standard and omitted. ■

Lemma A.17. *(i) For any $\epsilon>0$, there exists $M>0$ such that*

$$\limsup_{n_0\to\infty}\sup_{P\in\mathcal{P}} P\left\{\sup_{\mathbf{w}\in\Delta_{K-1}}\left|\hat{\mathcal{M}}(\mathbf{w})-\mathcal{M}_P(\mathbf{w})\right|>Mn_0^{-1/2}\right\}<\epsilon.$$

(ii) For any $\epsilon>0$, there exists $M>0$ such that for any sequence $\delta_n\to 0$ as $n\to\infty$,

$$\limsup_{n_0\to\infty}\sup_{P\in\mathcal{P}} P\left\{\sup_{\mathbf{w}\in\Delta_{K-1}:\|\mathbf{w}-\mathbf{w}_P\|\le\delta_n}\left|\hat{\mathcal{M}}^{\Delta}(\mathbf{w})-\hat{\mathcal{M}}^{\Delta}(\mathbf{w}_P)\right|>M\delta_n n_0^{-1/2}\right\}<\epsilon.$$

Proof: (i) First, we write

$$\hat{\mathcal{M}}(\mathbf{w})-\mathcal{M}_P(\mathbf{w})=\mathbf{w}^\top(\hat{H}-H_P)\mathbf{w}-2(\hat{\mathbf{h}}-\mathbf{h}_P)^\top\mathbf{w}+\hat{\mathbf{h}}'\hat{H}^{-1}\hat{\mathbf{h}}-\mathbf{h}_P'H_P^{-1}\mathbf{h}_P.$$

Since the weights are from the simplex Δ_{K-1} that is a bounded set, the desired result follows from Lemma A.7 and the Central Limit Theorem.

(ii) We write

$$\begin{aligned}\hat{\mathcal{M}}^{\Delta}(\mathbf{w})-\hat{\mathcal{M}}^{\Delta}(\mathbf{w}_P)=(\mathbf{w}-\mathbf{w}_P)^\top(\hat{H}-H_P)(\mathbf{w}-\mathbf{w}_P)+2\mathbf{w}_P'(\hat{H}-H_P)(\mathbf{w}-\mathbf{w}_P)\\ -2(\hat{\mathbf{h}}-\mathbf{h}_P)^\top(\mathbf{w}-\mathbf{w}_P),\end{aligned}$$

where $\hat{\mathcal{M}}^{\Delta}(\mathbf{w})=\hat{\mathcal{M}}(\mathbf{w})-\mathcal{M}_P(\mathbf{w})$. Similarly as before, from Lemma A.10, the desired result immediately follows. ■

Lemma A.18. *Suppose that for some positive sequence $\delta_{n,1}$ such that $\lim_{n\to\infty}\delta_{n,1}=0$, we have*

$$\lim_{M\uparrow\infty}\limsup_{n_0\to\infty}\sup_{P\in\mathcal{P}} P\left\{\|\hat{\mathbf{w}}-\mathbf{w}_P\|>M\delta_{n,1}\right\}=0.$$

Then,

$$\lim_{M\uparrow\infty}\limsup_{n_0\to\infty}\sup_{P\in\mathcal{P}} P\left\{\|\hat{\mathbf{w}}-\mathbf{w}_P\|^2>Mn_0^{-1/2}\delta_{n,1}\right\}=0.$$

Proof: We take arbitrary $\epsilon>0$ and large $\overline{M}_\epsilon>0$ such that

$$\limsup_{n_0\to\infty}\sup_{P\in\mathcal{P}} P\left\{\|\hat{\mathbf{w}}-\mathbf{w}_P\|>\overline{M}_\epsilon\delta_{n,1}\right\}\le\epsilon. \tag{A.15}$$

Recall the definition $\hat{\mathcal{M}}^{\Delta}(\mathbf{w})=\hat{\mathcal{M}}(\mathbf{w})-\mathcal{M}_P(\mathbf{w})$. Since $\hat{\mathcal{M}}(\mathbf{w}_P)\ge\hat{\mathcal{M}}(\hat{\mathbf{w}})$, we have

$$\begin{aligned}\hat{\mathcal{M}}^{\Delta}(\mathbf{w}_P)-\hat{\mathcal{M}}^{\Delta}(\hat{\mathbf{w}})&\ge\mathcal{M}_P(\hat{\mathbf{w}})-\mathcal{M}_P(\mathbf{w}_P)\\ &\ge\inf_{P\in\mathcal{P}}\lambda_{\min}(H_P)\|\hat{\mathbf{w}}-\mathbf{w}_P\|^2\ge\eta\|\hat{\mathbf{w}}-\mathbf{w}_P\|^2,\end{aligned} \tag{A.16}$$

from (A.13), where $\eta > 0$ is the constant in Assumption A.3. Define the event

$$E_n(\epsilon) = \left\{ \|\hat{\mathbf{w}} - \mathbf{w}_P\| > \overline{M}_\epsilon \delta_{n,1} \right\}.$$

By Lemma A.17(ii), for any $\epsilon_1 > 0$, there exists $M > 0$ such that

$$\limsup_{n_0 \to \infty} \sup_{P \in \mathcal{P}} P\left\{ |\hat{\mathcal{M}}^{\Delta}(\mathbf{w}_P) - \hat{\mathcal{M}}^{\Delta}(\hat{\mathbf{w}})| > M n_0^{-1/2} \overline{M}_\epsilon \delta_{n,1} \right\} \cap E_n^c(\epsilon) \leq \epsilon_1.$$

Therefore, from (A.16),

$$\liminf_{n_0 \to \infty} \inf_{P \in \mathcal{P}} P\left\{ \eta \|\hat{\mathbf{w}} - \mathbf{w}_P\|^2 \leq M n_0^{-1/2} \overline{M}_\epsilon \delta_{n,1} \right\} \geq 1 - \epsilon_1 - \epsilon.$$

Since the choice of ϵ_1 and ϵ is arbitrary, the desired result follows. ■

Proof of Theorem A.1: By Lemma A.9, there exists a sequence $\delta_{n,1} \to 0$ such that

$$\lim_{M \uparrow \infty} \limsup_{n_0 \to \infty} \sup_{P \in \mathcal{P}} P\left\{ \|\hat{\mathbf{w}} - \mathbf{w}_P\| > M \delta_{n,1} \right\} = 0. \tag{A.17}$$

By Lemma A.18, we find that the above result holds for $\delta_{n,1} = n_0^{-1/4}$. Now, we use mathematical induction. Suppose that (A.17) holds with $\delta_{n,1}$ such that

$$\log(\delta_{n,1}) = \log(n_0)\left(-\frac{1}{4} - \frac{1}{8} - \ldots - \frac{1}{2^m} \right),$$

for some $m \geq 2$. Then, with this choice of $\delta_{n,1}$, we apply Lemma A.18 again to find that (A.17) holds with $\delta_{n,1}$ such that

$$\log(\delta_{n,1}) = \log(n_0)\left(-\frac{1}{4} - \frac{1}{8} - \ldots - \frac{1}{2^{m+1}} \right).$$

Hence, we find that (A.17) holds with $\delta_{n,1}$ such that

$$\log(\delta_{n,1}) = \log(n_0)\left(-\sum_{m=2}^{\infty} \frac{1}{2^m} \right) = -\frac{1}{2}\log(n_0).$$

This gives the desired result. ■

Proof of Theorem A.2 : Note that

$$\begin{aligned} P\left\{ \theta_P(\mathbf{w}_P) \notin C_{1-\alpha} \right\} &= P\left\{ \inf_{\mathbf{w} \in \tilde{C}_{1-\kappa}} \left(\frac{\sqrt{n_0}(\hat{\theta}(\mathbf{w}) - \theta_P(\mathbf{w}_P))}{\hat{\sigma}} \right)^2 > c_{1-\alpha+\kappa}(1) \right\} \\ &\leq P\left\{ \left(\frac{\sqrt{n_0}(\hat{\theta}(\mathbf{w}_P) - \theta_P(\mathbf{w}_P))}{\hat{\sigma}} \right)^2 > c_{1-\alpha+\kappa}(1) \right\} + P\left\{ \mathbf{w}_P \notin \tilde{C}_{1-\kappa} \right\}. \end{aligned}$$

The desired result follows by Lemmas A.15, A.16 and A.13. ■

A.3. When H is Not Necessarily Invertible

Let us discuss the case where H is not necessarily invertible. In this case, we show how we can still obtain uniformly valid confidence intervals for θ_0. First, we provide a modification of the method to accommodate this setting, and then present the uniform validity result.

We begin by noting that we can rewrite

$$\rho_P^2(\mathbf{w}) = \mathbf{w}^\top H_P \mathbf{w} + 2\mathbf{w}^\top \mathbf{h}_P.$$

We define

$$\mathbf{W}_P = \arg\min_{\mathbf{w}\in\Delta_{K-1}} \rho_P^2(\mathbf{w}).$$

Let us explain how we construct the confidence interval for $\theta_0(\mathbf{w}_0)$ for a fixed $\mathbf{w}_0 \in \Delta_{K-1}$. We first define

$$\hat{\theta}(\mathbf{w}) = \frac{1}{n_0}\sum_{i\in N_0} \hat{m}_0\left(\hat{\mu}_0^\Gamma(X_i), X_i\right) 1\{X_i \in \hat{\mathcal{X}}_0^\Gamma\} + \frac{1}{n_0}\sum_{i\in N_0} \hat{m}^{\text{syn}}(X_i;\mathbf{w})\, 1\{X_i \in \mathcal{X}_0 \setminus \hat{\mathcal{X}}_0^\Gamma\}, \tag{A.18}$$

where

$$\hat{m}^{\text{syn}}(x;\mathbf{w}) = \sum_{k=1}^K m_k\left(\hat{\mu}_k^\Gamma(x), x\right) w_k.$$

As in (27), we construct

$$T'(\mathbf{w}) = n_0 \inf_{\boldsymbol{\lambda}\in\Lambda(\mathbf{w})} \left(\hat{\mathbf{f}}(\mathbf{w}) - \boldsymbol{\lambda}\right)^\top \hat{\Omega}^{-1}(\mathbf{w}) \left(\hat{\mathbf{f}}(\mathbf{w}) - \boldsymbol{\lambda}\right), \tag{A.19}$$

where $\hat{\Omega}(\mathbf{w})$ is constructed as in (29) with $\mathbf{w}$ replacing $\hat{\mathbf{w}}$. Then, the confidence set for $\mathbf{w}_0$ is given by

$$\tilde{C}'_{1-\kappa} = \left\{\mathbf{w} \in \Delta_{K-1} : T'(\mathbf{w}) \le \hat{c}_{1-\kappa}(\mathbf{w})\right\}, \tag{A.20}$$

where $\hat{c}_{1-\kappa}(\mathbf{w})$ denotes the $1-\kappa$ percentile of the χ^2 distribution with degree of freedom equal to the number of zero entries in $\hat{\boldsymbol{\lambda}}(\mathbf{w})$.

We let

$$T^*(\mathbf{w}) = \sqrt{n_0}\left(\hat{\theta}^*(\mathbf{w}) - \hat{\theta}(\mathbf{w})\right).$$

We read the 0.75 quantile and 0.25 quantile of the bootstrap distribution of $\{T^*(\mathbf{w}) : b = 1, ..., B\}$, and denote them to be $\hat{q}_{0.75}(\mathbf{w})$ and $\hat{q}_{0.25}(\mathbf{w})$, respectively. Define

$$\hat{\sigma}(\mathbf{w}) = \frac{\hat{q}_{0.75}(\mathbf{w}) - \hat{q}_{0.25}(\mathbf{w})}{z_{0.75} - z_{0.25}},$$

where $z_{0.75}$ and $z_{0.25}$ are the 0.75- and 0.25-quantiles of $N(0,1)$.

Define

$$\hat{T}'(\mathbf{w},\theta) = \frac{\sqrt{n_0}(\hat{\theta}(\mathbf{w}) - \theta)}{\hat{\sigma}(\mathbf{w})}.$$

We construct the $(1-\alpha)$-level confidence interval using the Bonferroni approach as follows:

$$C'_{1-\alpha} = \left\{ \theta \in \Theta : \inf_{\mathbf{w} \in \tilde{C}_{1-\kappa}} \left(\hat{T}'(\mathbf{w},\theta)\right)^2 \leq c_{1-\alpha+\kappa}(1) \right\}, \tag{A.21}$$

where $\kappa > 0$ is a small constant, such as $\kappa = 0.005$, $c_{1-\alpha+\kappa}(1)$ denotes the $(1-\alpha+\kappa)$-quantile of the $\chi^2(1)$ distribution. By modifying the arguments in the proof of Theorem A.2, we can show that

$$\liminf_{n_0 \to \infty} \inf_{P \in \mathcal{P}} \inf_{\mathbf{w}_0 \in \mathbf{W}_P} P\left\{\theta_P(\mathbf{w}_0) \in C_{1-\alpha}\right\} \geq 1-\alpha.$$

Equipped with Lemma A.6, we can show this using standard arguments. We omit the details.

B. Further Details on Monte Carlo Simulations and Empirical Applications

B.1. Details on the Monte Carlo Simulations

In our simulations, there is an outcome Y_i which is a function of a single-dimensional policy variable X_i and an unobserved random variable U_i. We draw $U_i \sim N(0,\sigma^2)$, i.i.d. and independently from X_i. We allow for one target region (region 0), three source regions ($k = 1,2,3$), with all regions having the same sample size $n_0 \in \{500, 1000\}$.

For the policy experiment, we draw $\mu_k^\Gamma(X_i)$ i.i.d. Uniform$[0,1]$ for all source regions (i.e., the post-policy variable). However, for the target region, we assume that we only observe pre-policy X_i drawn i.i.d. Uniform$[1-s,1]$. The policy of interest is the map

$$\mu_0^\Gamma(X_i) = X_i/s - (1-s).$$

Hence, the post-policy distribution of $\mu_0^\Gamma(X_i)$ in the target region is Uniform$[0,1]$ and s measures the overlap of the support of X_i in the target region relative to the source regions. When $s = 1$, no information from source regions is necessary: the target parameter θ_0 is fully identified and estimable from the target region alone. When s is very close to 0, then the post-policy ARF for the target region is not identified for $X_i > 0$ and identification of θ_0 almost solely relies on information from source regions.

We consider two separate specifications for Average Response Functions, which we refer to as (i) "Linear" and (ii) "Non-Linear" (in X_i). They differ in specifying linear versus non-linear specifications for the ARF's, as well as in having boundary versus interior values for $\mathbf{w}^*$ (the

weights satisfying the synthetic transferability condition). This allows us to verify our inference in all of these theoretically and empirically relevant cases.

The **Linear Specification** specifies the following causal structures for the outcome, Y_i:

$$Y_i = \begin{cases} X_i + U_i, & \text{if} \quad k = 1 \\ 0.5X_i - 1 + U_i, & \text{if} \quad k = 2 \\ 0.3X_i + 1 + U_i, & \text{if} \quad k = 3 \\ 0.4X_i + U_i, & \text{if} \quad k = 0, \end{cases} \tag{B.1}$$

while the **Non-Linear Specification** is:

$$Y_i = \begin{cases} X_i + U_i, & \text{if} \quad k = 1 \\ X_i^2 - 1 + U_i, & \text{if} \quad k = 2 \\ X_i^3 - 3X_i + U_i, & \text{if} \quad k = 3 \\ 0.2X_i^3 + 0.4X_i^2 - 0.2X_i - 0.4 + U_i, & \text{if} \quad k = 0. \end{cases} \tag{B.2}$$

Our target parameter for the Linear Specification is given by $\theta_0 = 0.2$, while it is $\theta_0 = -0.317$ for the Non-Linear Specification. It follows that, for the Linear Specification, $\mathbf{w}_0 = (0, 0.5, 0.5)$, while for the Non-Linear case, $\mathbf{w}_0 = (0.4, 0.4, 0.2)$.

We estimate the ARF's for each region by Ordinary Least Squares with covariates in region k given by

$$\tilde{X}_i^k = [1, X_i, X_i^2, X_i^3]^\top.$$

Estimation of $\hat{H}$, $\hat{\mathbf{h}}$ and $\hat{\mathbf{w}}$ follows those in equations (21) above. We then estimate $\hat{\theta}(\hat{\mathbf{w}})$ as in equation (22) which, in our set-up, is given by:

$$\begin{aligned} \hat{\theta}(\hat{\mathbf{w}}) = \frac{1}{n_0} \sum_{i \in N_0} \hat{m}_0(\hat{\mu}^\Gamma(X_i)) 1\{X_i \in [1-s, 1]\} \\ + \frac{1}{n_0} \sum_{i \in N_0} \hat{m}^{\text{syn}}(X_i; \hat{\mathbf{w}}) \, 1\{X_i \in [0, 1-s]\}, \end{aligned}$$

where

$$\hat{m}^{\text{syn}}(x; \hat{\mathbf{w}}) = \sum_{k=1}^{K} \tilde{m}_k(\hat{\mu}_k^\Gamma(x)) \hat{w}_k.$$

We consider the Bonferroni-based Confidence Interval (CI), denoted $C_{1-\alpha}$ in equation (31) in the main text. Finally, we set $\alpha = 0.05$ to be the significance level, $\sigma = 0.5$, $\kappa = 0.005$, $R = 1000$ simulations and $B = 999$ bootstrap draws. The results are shown below and described in the main text.

TABLE 3. The Empirical Coverage Probability and Average Length of Confidence Intervals for θ_0 at 95% Nominal Level.

Coverage Probability

		Linear Specification	Non-Linear Specification
$n_0 = 500$	s = 0.5	0.942	1
	s = 0.9	0.991	0.987
$n_0 = 1000$	s = 0.5	0.935	0.955
	s = 0.9	0.988	0.979

Average Length of CI

		Linear Specification	Non-Linear Specification
$n_0 = 500$	s = 0.5	0.178	0.356
	s = 0.9	0.118	0.112
$n_0 = 1000$	s = 0.5	0.155	0.311
	s = 0.9	0.084	0.072

Notes: The first panel of the table reports the empirical coverage probability of the confidence interval for θ_0 computed with the two different procedures and the two different specifications described in the main text, while the second panel reports its average length. The inference procedure uses the Bonferroni-based confidence interval which computes a confidence set for $\mathbf{w}_0$ in a first-step. Simulation number is $R = 1000$, with $B = 999$ bootstrap repetitions, with $\kappa = 0.005$.

TABLE 4. Finite Sample Properties of $\hat{\mathbf{w}}$, $\hat{\theta}(\hat{\mathbf{w}})$ at 95% Nominal Level.

		Linear Specification			
		RMSE($\hat{\mathbf{w}}$)	RMSE($\hat{\theta}(\hat{\mathbf{w}})$)	Bias($\hat{\theta}(\hat{\mathbf{w}})$)	Var($\hat{\theta}(\hat{\mathbf{w}})$)
$n_0 = 500$	s = 0.5	0.414	0.044	-0.021	0.002
	s = 0.9	0.306	0.024	-0.001	0.001
$n_0 = 1000$	s = 0.5	0.356	0.032	-0.015	0.001
	s = 0.9	0.265	0.017	-0.002	0.000

		Non-Linear Specification			
		RMSE($\hat{\mathbf{w}}$)	RMSE($\hat{\theta}(\hat{\mathbf{w}})$)	Bias($\hat{\theta}(\hat{\mathbf{w}})$)	Var($\hat{\theta}(\hat{\mathbf{w}})$)
$n_0 = 500$	s = 0.5	0.355	0.051	-0.001	0.003
	s = 0.9	0.243	0.024	0.001	0.001
$n_0 = 1000$	s = 0.5	0.305	0.037	-0.002	0.001
	s = 0.9	0.207	0.017	-0.001	0.000

Notes: The table reports the Root Mean Square Error (RMSE) across simulations for our estimators $\hat{\mathbf{w}}$ and $\hat{\theta}(\hat{\mathbf{w}})$ described in our main text, as well as the bias of $\hat{\theta}(\hat{\mathbf{w}})$ and the variance of $\hat{\theta}(\hat{\mathbf{w}})$ across simulations.

B.2. Details on Empirical Application: Estimation for the ARFs

In this section, we explain the estimation of $m_k(\mu_k(X_i))$ and $m_k(\mu_k^\Gamma(X_i))$ used in our empirical application. As for estimation, we use the equilibrium wage generation in (33) so that for each

$i \in N_k$, we have

$$\log W_i = \max\{\log \beta_k + X_i^\top \gamma_k + U_{i,j}, \log \underline{W}_k\}.$$

We estimate γ_k using the pairwise difference estimation method of Honoré and Powell (1994). (Note that β_k is not identified in this semiparametric setting.) More specifically, we first define

$$s(y_1, y_2, \delta) = \begin{cases} y_1^2 - (y_2 + \delta)y_1, & \text{if } \delta \leq -y_2 \\ (y_1 - y_2 - \delta)^2, & \text{if } -y_2 < \delta < y_1 \\ (-y_2)^2 - (\delta - y_1)(-y_2), & \text{if } \delta \geq y_1. \end{cases}$$

For each $k = 0, 1, ..., K$, we let $\hat{\gamma}_k$ be the estimator obtained as a solution to the following optimization problem:

$$\min_\gamma \frac{1}{n_K(n_K - 1)} \sum_{i \in N_K} \sum_{j \in N_K : j > i} s\left(\log W_i - \log \underline{W}_k, \log W_j - \log \underline{W}_k, (X_i - X_j)^\top \gamma\right).$$

From this, we obtain $\hat{\gamma}_k$.

Then we define

$$\hat{\mu}_k(X_i) = X_i' \hat{\gamma}_k - \log \underline{W}_k \text{ and } \hat{\mu}_k^\Gamma(X_i) = X_i' \hat{\gamma}_k - \log \underline{W}_k^\Gamma,$$

and construct

$$\hat{m}_k(\overline{\mu}) = \frac{\sum_{\ell \in N_K, \ell \neq i} K_h(\overline{\mu} - \hat{\mu}_k(X_\ell)) Y_i}{\sum_{\ell \in N_K, \ell \neq i} K_h(\overline{\mu} - \hat{\mu}_k(X_\ell))},$$

where $K_h(x) = K(x/h)/h$ and K is a univariate kernel. In particular, we use a quartic kernel and choose h by cross-validation. We obtain the estimators of $m_k(\mu_k(X_i))$ and $m_k(\mu_k^\Gamma(X_i))$ as follows:

$$\hat{m}_k(\hat{\mu}_k(X_i)) \text{ and } \hat{m}_k(\hat{\mu}_k^\Gamma(X_i)).$$

Once the ARFs are estimated, we can proceed to construct a synthetic prediction after the minimum wage changes as described in the main text.

References

BILLINGSLEY, P. (1999): *Convergence of Probability Measures*. John Wiley & Sons, Inc, New York.

CANEN, N., AND K. SONG (2023): "Synthetic Decomposition for Counterfactual Predictions," *Working Paper*.

CHENEY, W., AND A. A. GOLDSTEIN (1959): "Proximity Maps for Convex Sets," *Proceedings of the American Mathematical Society*, 10, 448–450.

CLARKE, F. H. (1990): *Optimization and Nonsmooth Analysis*. SIAM, New York.

HONORÉ, B. E., AND J. L. POWELL (1994): "Pairwise-Difference Estimators of Censored and Truncated Regression Models," *Journal of Econometrics*, 64, 241–278.

MOHAMAD, D. A., E. W. VAN ZWET, E. CATOR, AND J. J. GOEMAN (2020): "Adaptive Critical Value for Constrained Likelihood Ratio Testing," *Biometrika*, 107(3), 677–688.

NEWEY, W. K., AND D. L. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, vol. IV, pp. 2111–2245. Elsevier.

SHAO, J., AND D. TU (1995): *The Jackknife and Bootstrap*. Springer-Verlag, New York.

SILVAPULLE, M. J., AND P. K. SEN (2005): *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. John Wiley & Sons, Hoboken, New Jersey.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, New York, USA.

DEPARTMENT OF ECONOMICS, UNIVERSITY OF WARWICK, COVENTRY, CV4 7AL, UNITED KINGDOM
Email address: nathan.canen@warwick.ac.uk

VANCOUVER SCHOOL OF ECONOMICS, UNIVERSITY OF BRITISH COLUMBIA, 6000 IONA DRIVE, VANCOUVER, BC, V6T 1L4, CANADA
Email address: kysong@mail.ubc.ca