cemfi

working paper
1905

The Role of the Propensity Score in Fixed Effect Models

Dmitry Arkhangelsky
Guido W. Imbens

March 2019

CENTRO DE ESTUDIOS
MONETARIOS Y FINANCIEROS
Casado del Alisal 5, 28014 Madrid, Spain
www.cemfi.es

CEMFI Working Paper No. 1905
March 2019

The Role of the Propensity Score in Fixed Effect Models

Abstract

We develop a new approach for estimating average treatment effects in the observational studies with unobserved group-level heterogeneity. A common approach in such settings is to use linear fixed effect specifications estimated by least squares regression. Such methods severely limit the extent of the heterogeneity between groups by making the restrictive assumption that linearly adjusting for differences between groups in average covariate values addresses all concerns with cross-group comparisons. We start by making two observations. First we note that the fixed effect method in effect adjusts only for differences between groups by adjusting for the average of covariate values and average treatment. Second, we note that weighting by the inverse of the propensity score would remove biases for comparisons between treated and control units under the fixed effect set up. We then develop three generalizations of the fixed effect approach based on these two observations. First, we suggest more general, nolinear, adjustments for the average covariate values. Second, we suggest robustifying the estimators by using propensity score weighting. Third, we motivate and develop implementations for adjustments that also adjust for group characteristics beyond the average covariate values.

JEL Codes: C14.

Keywords: Fixed effects, cross-section data, clustering, causal effects, treatment effects, unconfoundedness.

Dmitry Arkhangelsky
CEMFI
darkhangel@cemfi.es

Guido W. Imbens
University of Stanford
imbens@stanford.edu

Acknowledgement

We are grateful for comments by participants in the Harvard-MIT econometrics seminar, the SIEPR lunch at Stanford, the International Association of Applied Econometrics meeting in Montreal, Pat Kline, and Matias Cattaneo. We are also grateful to Greg Duncan for raising questions that this paper tries to answer. This research was generously supported by ONR grant N00014-17-1-2131.

1 Introduction

A common specification for regression functions in the context of data with a group structure is

$$Y_i = \alpha_{C_i} + W_i\tau + X_i^\top \beta + \varepsilon_i, \tag{1.1}$$

where the α_{C_i} are the group fixed effects (*e.g.,* Wooldridge [2010]). The coefficient on the binary treatment W_i, denoted by τ, is the object of interest. This regression function is often estimated by least squares. The use of this specification, and similar ones with two-way fixed effects and nonlinear adjustments for the covariates X_i, is widespread in empirical work, where the groups may correspond to states, cities, SMSAs, classrooms, birth cohorts, firms, or other geographic or demographic groups. The fixed effects are intended to capture unobserved differences between the groups. The motivation for including the fixed effects in the regression is that without them the least squares estimator may not have a credible causal interpretation (*e.g.,* Arellano [2003]). The main issue that we wish to address in the current paper is that the fixed effect specification can be quite restrictive, and is not naturally generalized. In particular we are interested in the case where we have a modest number of units per group, not sufficiently large to do the analysis entirely within groups, followed by averaging over the groups. With a modest number of units per group, such a flexible within-group analysis is not feasible, and we are forced to rely on comparisons of treated and control units in different groups. However, we may be concerned that simply accounting for the group differences through additive fixed effects is not sufficient to adjust for all relevant differences (*e.g.,* Altonji and Matzkin [2005], Imai and Kim [2019]). A second issue is that the fixed effect approach is focused on the average treatment effect, and does not naturally generalize to other estimands such as quantile treatment effects without changing the nature of the assumptions.

To motivate our approach, we make two observations. The first observation, following from repeated applications of omitted variable bias formulas, is that we can estimate the coefficient

on W_i in (1.1) by least squares estimation of a different regression function, namely

$$Y_i = \alpha + W_i\tau + X_i^\top\beta + \overline{W}_{C_i}\delta + \overline{X}_{C_i}^\top\gamma + \varepsilon_i, \tag{1.2}$$

where $\overline{W}_c$ and $\overline{X}_c$ are the group averages of W_i and X_i in group c. This equivalence is mentioned in Mundlak [1978] in a seminal paper on panel data, and exploited by Altonji and Mansfield [2018] to bound treatment effect variance. In the regression in (1.2) the low dimensional group averages $\overline{W}_{C_i}$ and $\overline{X}_{C_i}$ are used as control variables along with X_i instead of high-dimensional fixed effects in (1.1). This representation suggests thinking about the fundamental identification assumption underlying the estimator for τ as an unconfoundedness type assumption common in the program evaluation literature:

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, \overline{W}_{C_i}, \overline{X}_{C_i}. \tag{1.3}$$

A second observation is that if assignment to treatment is completely random, independent of both covariates and group membership, the fixed effect estimator is consistent even if the conditional expectation of the outcome is not not linear in covariates, treatment and fixed effects. More generally, the fixed effect estimator is consistent if we weight the units by the inverse of the propensity score:

$$(\hat{\alpha}_c, \hat{\tau}, \hat{\beta}) = \arg\min_{\alpha_c,\tau,\beta} \sum_{i=1}^{N} \big(Y_i - \alpha_{C_i} - W_i\tau - X_i^\top\beta\big)^2 \frac{1}{p(X_i, C_i)^{W_i}(1 - p(X_i, C_i))^{1-W_i}},$$

where $p(x, c) \equiv \text{pr}(Wi = 1|X_i = x, C_i = c)$ is the propensity score. This observation suggests using the propensity score weighting as an alternative to a fixed effect model to adjust for general group differences.

We use these two observations to motivate three distinct modifications of the fixed effect estimation strategy. These three modifications can be used individually or collectively to free up implicit and explicit restrictions in the fixed effect approach. First, the conditional independence in (1.3) suggest that we can use more flexible specifications for the regression function as a function of the control variables $(X_i, \overline{W}_{C_i}, \overline{X}_{C_i})$ and the treatment W_i. In general we could

specify the regression function as:

$$Y_i = g(W_i, X_i, \overline{W}_{C_i}, \overline{X}_{C_i}) + \varepsilon_i,$$

with a parametric or non-parametric specification for $g(\cdot)$. These specifications may include higher order moments of the control variables, or interactions with the treatment, or transformations of the linear index. Given estimates of $g(\cdot)$ we can average the difference $g(1, X_i, \overline{W}_{C_i}, \overline{X}_{C_i}) - g(0, X_i, \overline{W}_{C_i}, \overline{X}_{C_i})$ over the sample to estimate the average treatment effect.

Second, the conditional independence in (1.3) suggest that we can model the propensity score, defined as a function of X_i and the group indicator, as a function of the control variables $(X_i, \overline{W}_{C_i}, \overline{X}_{C_i})$:

$$p(X_i, C_i) = e(X_i, \overline{W}_{C_i}, \overline{X}_{C_i}) \equiv \text{pr}(W_i = 1 | X_i, \overline{W}_{C_i}, \overline{X}_{C_i}).$$

Once we have estimates of the propensity score, we can use them to develop inverse propensity score weighting estimators. In particular, an attractive approach would be to use the inverse propensity score weighting in combination with a credible specification of the regression function. For example, one could use the conventional fixed effect specification but use a weighted version to make the results more robust to misspecification of the regression function. Such double robust methods have been found in the causal inference literature on unconfoundedness to be more effective than estimators that rely solely on specifying the conditional mean of the outcome given conditioning variables and treatments.

Third, the representation in (1.2) and the associated unconfoundedness assumption in (1.3) highlight that the fixed effect approach implicitly assumes that the two averages $\overline{W}_{C_i}$ and $\overline{X}_{C_i}$ capture all the relevant differences between the groups. A natural question is whether this is so. We may wish to consider additional characteristics of the clusters beyond these two averages to improve the comparability of clusters. This is similar to how we build up the propensity score using logistic regression models with an increasingly rich set of covariates. We build a framework for doing so in the current context where the propensity score is defined as the probability of treatment given covariates and group membership. This framework suggests conditions under which the average covariate values and average treatment per group are sufficient to capture all relevant differences between groups.

In practice our recommendation is to use all three modifications: First choose what group characteristics one wishes to include in the analysis, beyond the group averages of the covariates and treatment that are included in the standard fixed effect approach. Second, specify a credible conditional mean function, possibly, but not necessarily involving fixed effects, and including these additional group characteristics. Third, estimate the propensity score and combine that with the conditional mean specification to obtain a more robust estimator for the average treatment effect.

The first contribution in the current paper is a formal characterization of the primitive assumptions that justify the unconfoundedness assumption in (1.3) as well as generalizations that include other group-level variables. These assumptions are primitive in the sense that they impose restrictions on the population distribution of units and groups, whereas (1.3) depends partly on the sampling process (*e.g.*, the properties of the group averages change with the number of units sampled per group). Our second contribution is to characterize the average treatment effects that can be identified under this assumption, which will involve some trimming along the lines of Crump et al. [2009]. In the third contribution, we develop a new estimator. We derive large sample properties of the procedures proposed here, including consistency and asymptotic normality. A major challenge is that some of the conditioning variables, $\overline{W}_{C_i}$ and $\overline{X}_{C_i}$ in (1.3) are sample averages rather than population values. To capture the relevant empirical settings, we focus on asymptotics where the number of units per group remains fixed while the number of groups increases. As a result, the within-group averages $\overline{W}_{C_i}$ and $\overline{X}_{C_i}$ are not estimating their population counterparts consistently. Nevertheless, we show that Normal distribution based confidence intervals for our proposed estimators are valid in large samples. Finally, we show how our approach can be used to estimate other estimands, such as quantile treatment effects.

2 Set Up

In this section we set up the problem and introduce the notation. Using the potential outcome set up (e.g., Imbens and Rubin [2015]), we consider a set up with a large, possibly infinite, population of units, characterized by a pair of potential outcomes $(Y_i(0), Y_i(1))$, and a K-component vector of pretreatment variables X_i. The population is partitioned into subpopulations or groups, with C_i indicating the group unit i is a member of. The number of groups in the population is large,

and so is the number of units per group.

We are interested in the average treatment effects. Ideally we might wish to estimate the population average effect,

$$\tau \equiv \mathbb{E}[Y_i(1) - Y_i(0)],$$

but this may be challenging, and we may need to settle for some other average of $Y_i(1) - Y_i(0)$, *e.g.*, the average over some subpopulation defined in terms of groups, covariates and assignments. Unit i receives treatment $W_i \in \{0, 1\}$. We first randomly sample C groups, and then draw a random sample of size N from the subpopulation defined by the sampled groups. For the sampled units we observe the quadruple (Y_i, W_i, X_i, C_i), $i = 1, \ldots, N$, where $Y_i \equiv Y_i(W_i)$ is the realized outcome, that is, the potential outcome corresponding to the treatment received, and $C_i \in \{1, \ldots, C\}$ is the group label for unit i. Also define $C_{ic} = \mathbf{1}_{C_i=c}$ as the binary group indicators, and let $N_c \equiv \sum_{i=1}^{N} C_{ic}$ be the number of sampled units in group c. For any variable Z_i, let $\overline{Z}_c \equiv \sum_{i:C_i=c} Z_i/N_c$ be the corresponding group average in group c. For each unit in the population the (partly unobserved) data tuple is given by $\{(Y_i(0), Y_i(1), W_i, X_i, U_i, C_i)\}_{i=1}^{N}$. The variable U_i is an unobserved cluster-level variable that varies only between clusters, so that it is equal to its cluster average for all units, $\overline{U}_{C_i} = U_i$ for all i.

In the settings we are interested in the number of strata or clusters in the sample, C, may be substantial, on the order of hundreds or even thousands. The dimension of X_i is modest. The number of units in the population in each cluster is large, but we observe only few units in each group, possibly as few as two or three. As a result methods that rely on accurate estimation of features of the population distribution of potential outcomes or treatments conditional on covariates within clusters may have poor properties.

The set up we consider has a large population of clusters. In the population, each cluster has a large number of units. We randomly sample a finite number of clusters and then sample a finite number of units from the subpopulation of sampled clusters. Large sample approximations to estimators are based on the number of sampled clusters increasing, with the average number of sampled units per cluster converging to a constant.

3 Identification, Estimation, and Inference

In this section we propose a new estimator for average treatment effects in the setting with grouped data. The estimator has features in common with the efficient influence function estimators from the program evaluation literature, as well as with the fixed effect estimators from the panel data literature. Unlike fixed effect estimators, it can accommodate differences in potential outcome distributions between clusters that are not additive. There are two issues involved in our approach. First, we have to be careful in defining the estimand to account for the fact that there may be few units in a cluster. In general, we can not consistently estimate the overall average causal effect, because there are likely to be clusters with no treated or no control units. To take this into account, we define a subset of units for which we estimate the average effect. Of course, this is not entirely new to our approach even in the panel data setting: implicitly standard fixed effect estimators do not estimate the average effect of the treatment if there is systematic variation in treatment effects by strata and some strata have no variation in treatments. However, by explicitly moving away from focusing on population quantities, we relax the conditions required for identification, compared to, say, those in Altonji and Matzkin [2005]. Second, a key feature in our approach is that we need to adjust for characteristics of the clusters that are not observed. Although we can estimate these features, they cannot be estimated consistently under the asymptotic sequences we consider.

3.1 Identification

This assumption describes the sampling process.

Assumption 3.1. (Balanced clustered sampling) *There is a super-population of groups, we randomly sample C clusters. We then randomly sample N units from the subpopulation of sampled clusters. Let N_c be the number of units sampled from cluster c, so that $N = \sum_{c=1}^{C} N_c$ is the total sample size.*

Our second assumption imposes restrictions on the treatment assignment process:

Assumption 3.2. (Unconfoundedness within Clusters)

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, C_i. \tag{3.1}$$

This assumption implies that we can always compare individuals with the same characteristics within the cluster. Some version of this assumption underlies most fixed effect approaches.

The second assumption imposes restrictions on the fixed effects.

Assumption 3.3. (Random effects)
There is an unobserved group-level variable $\overline{U}_{C_i}$ such that:

$$\Big(Y_i(1), Y_i(0), X_i, W_i\Big) \perp\!\!\!\perp C_i \;\Big|\; \overline{U}_{C_i} \tag{3.2}$$

This assumption, what Altonji and Matzkin [2005] (Assumption 2.3 in their paper) call exchangeability, essentially turns the problem into a random effects set up: the labels of the clusters C_i are not important, only the cluster-level characteristics $\overline{U}_{C_i}$ are. This assumption allows us to conceptualize similarity of clusters. This assumption is without essential loss of generality, as it follows in the case with infinitely sized groups from deFinetti's theorem (De Finetti [2017], Diaconis [1977]).

Since $\overline{U}_{C_i}$ is measurable with respect to cluster indicator variable, a direct implication of the previous pair of assumptions is:

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, \overline{U}_{C_i}. \tag{3.3}$$

Now we can also compare treated and control units in different clusters, as long as the clusters have the same value for $\overline{U}_{C_i}$.

For the first identification result we need some additional notation. For each cluster c define $\mathbb{P}_c$ to be the empirical distribution of (X_i, W_i) in cluster c. In the case with discrete X_i this amounts to the set of frequencies of observations in a cluster for each pair of values (W_i, X_i).[1]

Proposition 1. (Unconfoundedness with empirical measure) *Suppose Assumptions 3.1-3.3 hold. Then:*

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, N_{C_i}, \mathbb{P}_{C_i} \tag{3.4}$$

For the proofs of the results in this section see Appendix A.

[1] For the formal definition of this object including continuous X_i see Appendix A.

This result states that as long as units have the same characteristics, and they come from clusters identical in terms of $\mathbb{P}_{C_i}$, they are comparable. This is a balancing/propensity score type result in the sense that subpopulation with the same value for $(X_i, \mathbb{P}_{C_i})$ are balanced: the distribution of treatments is the same for all units within such subpopulations. See, for example, Rosenbaum and Rubin [1983].

However, the empirical relevance of this result is limited, because in most cases the dimension of the conditioning set is high. If X_i is discrete and takes on K values, with K typically large, and there are N_c units in group c, the number of possible values for $(X_i, \mathbb{P}_c)$ is $K \times (2K)^{N_c}/N_c!$. Overlap of the distributions of the conditioning variables is going to be a major problem in this case. This is the motivation for the next assumption. We put structure on the joint distribution of (W_i, X_i) within groups to reduce the dimension of the conditioning set.

Assumption 3.4. (Exponential family) *Conditional on $\overline{U}_{C_i}$ distribution of (X_i, W_i) belongs to an exponential family with a known sufficient statistic:*

$$f_{X_i,W_i|U_i}(x, w|u) = h(x, w) \exp\Big\{\eta^\top(u) S(x, w) + \eta_0(u)\Big\}, \tag{3.5}$$

with potentially unknown carrier $h(\cdot)$.

Define $\overline{S}_c \equiv (N_c, \sum_{i:C_i=c} S(X_i, W_i)/N_c$ be the sample size in cluster c and the cluster average of $S(X_i, W_i)$ for cluster c.

Comment 1: This assumption restricts the joint distribution of the treatment and covariates conditional on the cluster, $(X_i, W_i)|\overline{U}_{C_i}$ but places no restrictions on the conditional distribution of the outcome variable, $Y_i|X_i, W_i, \overline{U}_{C_i}$. □

Comment 2: If we do not restrict the dimension of $S(\cdot)$ the exponential family assumption is without essential loss of generality. To see this, note that if the distribution of X_i is discrete, one can immediately write the joint distribution of (X_i, W_i) within each cluster as an exponential family distribution with a cluster specific parameter. In addition, we can approximate any distribution arbitrarily well by a discrete distribution. □

Comment 3: One might wonder why we make any assumptions on the distribution of X_i at all and not just focus on a model for the propensity score, as is commonly done in unconfoundedness settings. The reason is key to our approach. With the number of units within the cluster not increasing with the sample size, we cannot estimate the propensity score consistently (we cannot

estimate the exponential family parameters $\eta(u)$ consistently). This situation is akin to fixed-T models in panel data, where common parameters can be identified, but individual effects are not. Modeling the joint distribution of (W_i, X_i) in the way we do we can bypass the need for consistent estimation of $\eta(u)$ and instead focus on the conditional distribution of W_i given X_i and $\overline{S}_{C_i}$. □

Lemma 1. *Suppose Assumptions 3.1–3.4 hold. Then*

$$W_i \perp\!\!\!\perp C_i \,\Big|\, X_i, \overline{S}_{C_i}. \tag{3.6}$$

Theorem 1. (Unconfoundedness with sufficient statistic) *Suppose Assumptions 3.1–3.4 hold. Then:*

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \,\Big|\, X_i, \overline{S}_{C_i}. \tag{3.7}$$

Comment 4: Theorem 1 can be viewed as essentially a direct consequence of Proposition 1, but it is substantially more operational. It reduces the potentially high-dimensional object $\mathbb{P}_{C_i}$ to a lower dimensional average $\overline{S}_{C_i}$. It is unusual in that one of the conditioning variables, $\overline{S}_{C_i}$, is not a fixed unit-level characteristic. Instead, it is a characteristic of the cluster and the sampling process. If we change the sampling process, say to sampling twice as many units per cluster, the distribution of $\overline{S}_{C_i}$ changes. Nevertheless, this conceptual difference in the nature of $\overline{S}_{C_i}$ relative to the unit-level characteristic X_i does not affect how it is used in the estimation procedures. □

Comment 5: There is another key difference between the unconfoundedness condition in Theorem 1 and in Proposition 1. With continuous covariates, the latter essentially makes it impossible to have overlap. Indeed, unless we have individuals with the same value of covariates within the cluster, the distribution of W_i given X_i and $\mathbb{P}_{C_i}$ is degenerate. It is well known that overlap is crucial in the semiparametric estimation of treatment effects and without it, the identification is possible only under functional form assumptions.□

Comment 6: The result in Theorem 1 is more useful because it allows us to control the degree of overlap as well. The higher the dimension of $S(\cdot)$ the closer we are to controlling for $\mathbb{P}_{C_i}$, and thus the smaller is the region for which we have overlap. □

Comment 7: In Altonji and Matzkin [2005] a key assumption (Assumption 2.1) requires that

there is an observed variable Z_i such that conditioning on Z_i renders the covariate of interest (the treatment in our case) exogenous. The role of this conditioning variable is in our setting played by the sufficient statistic $\overline{S}_{C_i}$. Our set up shows how this property can arise from assumptions on the joint distribution of the treatment and the other covariates, and how we can make this more plausible by expanding the set of sufficient statistics. □

In this and the next section, we l assume that $S(\cdot)$ is known, fixed and there is a known region of the covariate space where we have overlap. In Section 3.3 we discuss selecting the set of sufficient statistics. In particular, recall the definition of the propensity score:

$$e(x,s) \equiv \mathbb{E}[W_i | X_i = x, \overline{S}_{C_i} = s] \tag{3.8}$$

We are making the following assumption:

Assumption 3.5. (KNOWN OVERLAP) *We assume that there exists $\eta > 0$ and a nonempty known set $\mathbb{A}$, such that for any $(x,s) \in \mathbb{A}$ we have $\eta < e(x,s) < 1-\eta$.*

COMMENT 8: This assumption has two parts: the first part restrict $e(x,s)$ to be non-degenerate on a certain set. This is necessary if we want to identify treatment effects without relying on functional form assumptions. The second part is different: we assume that the set is known to a researcher. This is a generalization of the standard overlap assumption, where we assume that the set $\mathbb{A}$ is equal to the support of the covariate space. See Crump et al. [2009]. □

3.2 Estimation and Inference

Here we collect several inference results for the general semiparametric estimator. All proofs can be found in Appendix B.

For the further use we use following notation for the conditional mean, propensity score and residuals:

$$\begin{cases} \mu(W_i, X_i, \overline{S}_{C_i}) \equiv \mathbb{E}[Y_i | W_i, X_i, \overline{S}_{C_i}] \\ e(X_i, \overline{S}_{C_i}) \equiv \mathbb{E}[W_i | X_i, \overline{S}_{C_i}] \\ \varepsilon_i(w) \equiv Y_i(w) - \mu(w, X_i, \overline{S}_{C_i}) \end{cases} \tag{3.9}$$

Note that these expectations are defined conditional on Assumption 3.1, which determines the distribution of $\overline{S}_{C_i}$.

We will use $\hat{\mu}_i(\cdot)$ and $\hat{e}_i(\cdot)$ for generic estimators of $\mu(\cdot)$ and $e(\cdot)$. Subscript i is used to allow for cross-fitting (Chernozhukov et al. [2016]). Define $A_i \equiv \{(X_i, \overline{S}_{C_i}) \in \mathbb{A}\}$, where $\mathbb{A}$ is a (known) set with overlap in the distribution of $(X_i, \overline{S}_{C_i})$. Define true and estimated share of observations with overlap:

$$\begin{cases} \pi(\mathbb{A}) \equiv \mathbb{E}[A_i] \\ \overline{A} \equiv \frac{1}{N}\sum_{i=1}^{N} A_i \end{cases} \tag{3.10}$$

We assume the generic estimators $\hat{e}_i$ and $\hat{\mu}_i$ satisfy several high-level consistency properties. These restrictions are standard in the program evaluation literature.

Assumption 3.6. (High-level conditions) *The following conditions are satisfied for $\hat{e}_i(\cdot)$ and $\hat{\mu}_i(\cdot)$:*

$$\begin{cases} \eta < \hat{e}_i(X_i, \overline{S}_{C_i}) < 1 - \eta \ \ a.s. \\ \frac{1}{N}\sum_{i=1}^{N} A_i (e(X_i, \overline{S}_{C_i}) - \hat{e}(X_i, \overline{S}_{C_i}))^2 = o_p(1) \\ \frac{1}{N}\sum_{i=1}^{N} A_i (\mu(W_i, X_i, \overline{S}_{C_i}) - \hat{\mu}_i(W_i, X_i, \overline{S}_{C_i}))^2 = o_p(1) \\ \frac{1}{N}\sum_{i=1}^{N} A_i (e(X_i, \overline{S}_{C_i}) - \hat{e}_i(X_i, \overline{S}_{C_i}))^2 \\ \qquad \times \frac{1}{N}\sum_{i=1}^{N} A_i (\mu(W_i, X_i, \overline{S}_{C_i}) - \hat{\mu}_i(W_i, X_i, \overline{S}_{C_i}))^2 = o_p\left(\frac{1}{n}\right) \end{cases} \tag{3.11}$$

We also restrict moments of the residuals:

Assumption 3.7. (Moment conditions)

$$\begin{cases} \mathbb{E}[\varepsilon_i^2(k) | X_i, \overline{S}_{C_i}] < K \ \ a.s. \\ \mathbb{E}[\varepsilon_i^4(k)] < \infty \end{cases} \tag{3.12}$$

For arbitrary (subject to appropriate integrability conditions) functions $(\mu(\cdot), e(\cdot))$ define the

following functional:

$$\psi(y,w,x,s,\mu(\cdot),e(\cdot)) \equiv \mu(1,x,s)-\mu(0,x,s)+\left(\frac{w}{e(x,s)}-\frac{1-w}{1-e(x,s)}\right)(y-\mu(w,x,s)). \quad (3.13)$$

We focus on the following conditional estimand:[2]

$$\tilde{\tau}_{\mathbb{A}} = \frac{1}{\overline{A}}\frac{1}{N}\sum_{i=1}^{N} A_i\left(\mu(1,X_i,\overline{S}_{C_i})-\mu(0,X_i,\overline{S}_{C_i})\right) \quad (3.14)$$

Theorem 2. (Consistency) *Suppose Assumptions 3.1–3.4 and Assumption 3.6 hold. Then:*

$$\hat{\tau}_{\mathrm{dr}} \equiv \frac{1}{N\overline{A}}\sum_{i=1}^{N} A_i\psi(Y_i,W_i,X_i,\overline{S}_{C_i},\hat{\mu}(W_i,X_i,\overline{S}_{C_i}),\hat{e}(X_i,\overline{S}_{C_i})), \quad (3.15)$$

satisfies $\hat{\tau}_{\mathrm{dr}}-\tilde{\tau}_{\mathbb{A}}=o_p(1)$.

For inference results we need to use $\hat{\mu}_i$ with cross-fitting. We also need to take account of the clustering. Define

$$\rho(c,\mu(\cdot),e(\cdot)) \equiv \frac{1}{N_c}\sum_{i:C_i=c} A_i\psi(Y_i,W_i,X_i,\overline{S}_{C_i},\mu(W_i,X_i,\overline{S}_{C_i}),e(X_i,\overline{S}_{C_i})),$$

so that

$$\hat{\tau}_{\mathrm{dr}} = \frac{1}{\overline{A}}\sum_{c=1}^{C}\frac{N_c}{N}\rho(c,\hat{\mu}(\cdot),\hat{e}(\cdot)).$$

Theorem 3. (Inference for semiparametric case) *Suppose Assumptions 3.1–3.4 and Assumption 3.6 hold. Assume that $\hat{\mu}_i$ is estimated using cross-fitting with L folders. Then:*

$$\sqrt{n}(\hat{\tau}_{\mathrm{dr}}-\tilde{\tau}_{\mathbb{A}}) \xrightarrow{d} \mathcal{N}(0,\mathbb{V}), \qquad \text{where} \quad \mathbb{V}=\frac{\mathbb{E}\left[\xi_c^2\right]}{\pi^2(\mathbb{A})},$$

[2] It is straightforward to extend our inference results to a more standard target $\tau_{\mathbb{A}}$, in which case we will have a different variance.

where ξ_c is defined in the following way:

$$\xi_c \equiv \sum_{i \in c} \frac{A_i}{N_c} \left(\frac{W_i}{e(X_i, \overline{S}_{C_i})} - \frac{1 - W_i}{1 - e(X_i, \overline{S}_{C_i})} \right) (Y_i - \mu(W_i, X_i, \overline{S}_{C_i}))$$

Finally, we address the estimation of variance. For this define the following empirical version of ξ_c:

$$\hat{\xi}_c \equiv \sum_{i \in c} \frac{A_i}{N_c} \left(\left(\frac{W_i}{\hat{e}(X_i, \overline{S}_{C_i})} - \frac{1 - W_i}{1 - \hat{e}(X_i, \overline{S}_{C_i})} \right) (Y_i - \hat{\mu}(W_i, X_i, \overline{S}_{C_i})) \right) \tag{3.16}$$

The proposed variance estimator is just the variance of $\hat{\xi}_c$:

$$\hat{\mathbb{V}} := \frac{1}{\overline{A}^2} \frac{1}{C} \sum_{c=1}^{C} \left(\hat{\xi}_c - \frac{1}{C} \sum_{c'=1}^{C} \hat{\xi}_{c'} \right)^2. \tag{3.17}$$

The following proposition says that asymptotically variance of the estimated influence function is equal to the variance of the true influence function:

Proposition 2. (VARIANCE CONSISTENCY) *Suppose the assumptions of Theorem 3 hold. Then the variance estimator is consistent:*

$$\hat{\mathbb{V}} = \mathbb{V} + o_p(1). \tag{3.18}$$

3.3 Choosing the Sufficient Statistics

The suggestion to include additional group characteristics raises the question how to select these. Selecting more sufficient statistics raises concerns with overlap and the ability to adjust for these sufficient statistics adequately given the finite sample, and failure to adjust for all the relevant group characteristics may lead to biased estimators. Intuitively we would like a selection procedure to select more sufficient statistics in settings where we have a lot of units per cluster, and if the distributions vary substantially by cluster. A full treatment of this problem is an open question. However, we provide a suggestion for systematically selecting sufficient statistics in the case where we have a large set of potential sufficient statistics that includes all the relevant ones, but also some that are not relevant.

The sufficient statistics are intended to capture the differences in distributions of (X_i, W_i) between clusters. If a particular sufficient statistic is important, it should therefore be useful in predicting which cluster a unit belongs to. Hence we can cast this as a prediction or classification problem and bring to bear machine learning methods. Under the exponential family assumption, and given the sampling framework in Assumption 3.1, the conditional probability that a unit in the sample is from group c, conditional on (W_i, X_i) and conditional on the set of $\overline{U}_1, \ldots, \overline{U}_C$, has a multinomial logit form:

$$\text{pr}(C_i = c | W_i, X_i, \overline{U}_1, \cdots, \overline{U}_C) = \frac{\exp(\eta_0(\overline{U}_c) + \eta^\top(\overline{U}_c) S(X_i, W_i))}{\sum_{c'=1}^{C} \exp(\eta_0(\overline{U}_{c'}) + \eta^\top(\overline{U}_{c'}) S(X_i, W_i))}.$$

Hence the problem of selecting the sufficient statistics is similar to the problem of selecting covariates in a multinomial logistic regression model. Given a large set of potential sufficient statistics we can use standard regularization methods, such as LASSO (Tibshirani [1996]) to select a sparse set of relevant ones.

4 Extensions

In this section we discuss three extensions of the ideas introduced in this paper.

4.1 Quantile Treatment Effects

Theorem 1 states that conditional on the covariates and the sufficient statistics we have the unconfoundedness condition:

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, \overline{S}_{C_i}.$$

This implies that we can study estimation of effects other than average treatment effects. This is important in applications where we want to estimate, say, nonlinear effects controlling on cluster-level unobserved heterogeneity.

In particular, for any bounded function $f : \mathbb{R} \to \mathbb{R}$ we can estimate $\mathbb{E}[f(Y_i(w))]$ using the

following representation:

$$\mathbb{E}[f(Y_i(w)] = \mathbb{E}\left[\frac{\{W_i = w\}f(Y_i)}{e(X_i, \overline{S}_{C_i})}\right]$$

This allows us to deal with quantile treatment effects of the type introduced by Lehmann and D'Abrera [2006]. If we are interested in q-th quantile of the distribution of $Y_i(w)$ then (under appropriate continuity) we can identify it as a solution of the following problem:

$$c : \mathbb{E}\left[\frac{\{W_i = w\}\{Y_i \leq c\}}{e(X_i, \overline{S}_{C_i})}\right] = q$$

For the standard case under unconfoundedness Firpo [2007] has developed effective estimation methods that can be adapted to this case.

4.2 Panel Data

Although we focus in the current paper on a cross-section setting with clusters, as in Altonji and Mansfield [2018], the issues raised here are also relevant to proper panel or longitudinal data settings. In that literature the paper fits into a recent set of studies Abadie et al. [2010], de Chaisemartin and D'Haultfœuille [2018], Bonhomme and Manresa [2015], Imai and Kim [2019] that connects more directly with the causal (treatment effect) literature than the earlier panel data literature by allowing for general heterogeneity beyond additive effects.

Suppose we have N observations on C individuals, and T time periods, so that $N = C \times T$. We observe Y_i for all units and a binary treatment W_i. Let $T_i \in \{1, \ldots, T\}$ denote the time period observation i is from, and let $C_i \in \{1, \ldots, C\}$ denote the individual it goes with.

For any variable Z_i, define the time and individual averages:

$$\overline{Z}_{\cdot t} := \frac{1}{C}\sum_{i:T_i=t} Y_i, \qquad \overline{Z}_{c\cdot} := \frac{1}{T}\sum_{i:C_i=c} Y_i,$$

and the overall average

$$\overline{Z} := \frac{1}{N}\sum_{i=1}^{N} Z_i,$$

and the residual

$$\dot{Z}_i = Z_i - \overline{Z}_{\cdot t} - \overline{Z}_{c\cdot} + \overline{Z}$$

Let $\hat{\tau}_{\text{fe}}$ be the least squares estimator for the regression

$$Y_i = \alpha_{T_i} + \beta_{C_i} + \tau W_i + X_i^\top \gamma + \varepsilon_i \tag{4.1}$$

Compare this to the least squares regression

$$Y_i = \tau W_i + X_i^\top \gamma + \delta \overline{W}_{\cdot T_i} + \mu \overline{W}_{C_i \cdot} + \psi \overline{X}_{\cdot T_i} + \varphi \overline{X}_{C_i \cdot} + \varepsilon_i$$

The two least squares estimators for τ are numerically identical. This suggests that we can view the standard fixed-time effects approach in (4.1) as controlling for time and individual level sufficient statistics. This view opens a road to generalizing the standard estimators.

At the same time, this type of generalization is not completely satisfactory. For one, controlling for future values of X_{it} and W_{it} seems controversial. Also, it seems that the outcome information should be used to control for individual-level heterogeneity. Finally, in the panel case, the definition of treatment effects is inherently more complex, because of the dynamic structure of the problem. For these reasons, we think that the approach of this paper while insightful should be refined to make it appropriate for the panel data settings. We leave this for future research.

4.3 Beyond Exponential Families

Modelling the conditional distribution of (X_i, W_i) given U_i using exponential family is very natural for the purposes of this paper. Nevertheless, in some applications other families can be more appropriate. In particular, another operational choice is a discrete mixture. Assume that U_{C_i} can take a finite number of values $\{u_1, \ldots, u_p\}$ with probabilities $\pi_1, \ldots, \pi_p$ and the conditional distribution of X_i, W_i given U_i is given by $f(x, w|u)$. Collect all the data that we observe for cluster c in the following tuple:

$$\mathcal{D}_c \equiv ((X_{c,1}, W_{c,1}), \ldots, (X_{c,N_c}, W_{c,N_c})) \tag{4.2}$$

Marginal distribution of this object is given by the following expression:

$$f_{\mathcal{D}_c}(x_1, w_1, \ldots, x_{N_c}, w_{N_c}) = \sum_{k=1}^{p} \prod_{j=1}^{N_c} f(x_j, w_j|k)\pi_k \tag{4.3}$$

This implies that the conditional distribution of U_c given $\mathcal{D}_c$ has the following form:

$$\pi(U_c = k|x_1, w_1, \ldots, x_{N_c}, w_{N_c}) = \frac{\prod_{j=1}^{N_c} f(x_j, w_j|k)\pi_k}{\sum_{k=1}^{p} \prod_{j=1}^{N_c} f(x_j, w_j|k)\pi_k} \tag{4.4}$$

Define $S(\mathcal{D}_c) \equiv (\pi(U_c = 1|\mathcal{D}_c), \ldots, \pi(U_c = p|\mathcal{D}_c))$ and observe that as long as Assumption 3.3 holds we have the following:

$$(Y_i(1), Y_i(0)) \perp W_i|X_i, S(\mathcal{D}_c) \tag{4.5}$$

Recent results (e.g., Allman et al. [2009], Bonhomme et al. [2016]) show that $(\pi_1, \ldots, \pi_p)$ and $f(x, w|u)$ are nonparametrically identified under quite general assumptions. Using the algorithms proposed in these papers we can estimate $S(\mathcal{D}_c)$ and use it as a sufficient statistic.

This is conceptually different from the Bayesian classification that is used in unsupervised machine learning. Standard classification algorithms assign a unique value U_i to each observation (in our case, cluster). The usual way of doing this is to assign U_i that has the highest posterior probability. Here, we do not want to do this; instead, we want to find clusters that are similar in terms of the whole posterior distribution, not only its mode. If N_c is large, then this difference is not that important, because the posterior will typically concentrate on a particular value of U_i. With small N_c this is not going to happen, and the distinction is essential.

5 Conclusion

In this work, we proposed a new approach to identification and estimation in the observational studies with unobserved cluster-level heterogeneity. The identification argument is based on the combination of random effects and exponential family assumptions. We show that given this structure we can identify a specific average treatment effect even in cases where the observed number of units per cluster is small. From the operational point of view, our approach allows

researchers to utilize all the recently developed machinery from the standard observational studies. In particular, we generalize the doubly-robust estimator and prove its consistency and asymptotic normality under common high-level assumptions. We also show that the standard fixed effects estimation is a particular case of our procedure.

As a direction for future research, it will be interesting to see whether it is possible to utilize machine learning methods to learn sufficient statistics from the data. Additionally, it is essential to understand the statistical trade-off between the dimension of the sufficient statistic, cluster size and estimation rate for the propensity score. Finally, we view this work as a first step towards understanding a more challenging and arguably more practically important data design, where we observe panel data.

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105(490):493–505, 2010.

Elizabeth S Allman, Catherine Matias, John A Rhodes, et al. Identifiability of parameters in latent structure models with many observed variables. The Annals of Statistics, 37(6A): 3099–3132, 2009.

Joseph G Altonji and Richard K Mansfield. Estimating group effects using averages of observables to control for sorting on unobservables: School and neighborhood effects. American Economic Review, 108(10):2902–46, 2018.

Joseph G Altonji and Rosa L Matzkin. Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica, 73(4):1053–1102, 2005.

Manuel Arellano. Panel data econometrics. Oxford university press, 2003.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. Econometrica, 83(3):1147–1184, 2015.

Stéphane Bonhomme, Koen Jochmans, Jean-Marc Robin, et al. Estimating multivariate latent-structure models. The Annals of Statistics, 44(2):540–563, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K Newey, et al. Double machine learning for treatment and causal parameters. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 2016.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. Biometrika, 96(1):187–199, 2009.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. 2018.

Bruno De Finetti. Theory of probability: A critical introductory treatment, volume 6. John Wiley & Sons, 2017.

Persi Diaconis. Finite forms of de finetti's theorem on exchangeability. Synthese, 36(2):271–281, 1977.

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. Econometrica, 75(1):259–276, 2007.

Kosuke Imai and In Song Kim. When should we use unit fixed effects regression models for causal inference with longitudinal data? American Journal of Political Science, 2019.

Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

Erich Leo Lehmann and Howard JM D'Abrera. Nonparametrics: statistical methods based on ranks. Springer New York, 2006.

Yair Mundlak. On the pooling of time series and cross section data. Econometrica: journal of the Econometric Society, pages 69–85, 1978.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

Jeffrey M Wooldridge. Econometric analysis of cross section and panel data. MIT press, 2010.

A Identification results

First, we need to formally define $\mathbb{P}_c$. For this fix an arbitrary linear order $\succsim$ on $\mathcal{X} \times \{0,1\}$ (e.g., a lexicographic order). For any cluster c consider a tuple $A_c = \{(X_i, W_i)\}_{i \in c}$, order elements of A_c with respect to $\succsim$ and define $\mathbb{P}_c = \big((X_{(1)}, W_{(1)}), \ldots, (X_{(c)}, W_{(c)})\big) \in (\mathcal{X} \times \{0,1\})^c$. Under Assumption 3.1 this construction ensures that $\mathbb{P}_c$ is a well-defined random vector. It is clear that there is a one-to-one relationship between this vector and the empirical distribution of (X_i, W_i) within the cluster which makes the notation appropriate.

Below we will use the following definition of conditional independence. Let X, Y, Z be three random elements and A, B be the elements of the $\sigma(X)$- and $\sigma(Y)$-algebras, respectively. The $X \perp\!\!\!\perp Y | Z$ if the following holds:

$$\mathbb{E}[\{X \in A\}\{Y \in B\}|Z] = \mathbb{E}[\{X \in A\}|Z]\mathbb{E}[\{Y \in B\}|Z] \tag{1.1}$$

In the proofs below we are using A and B as generic elements of the appropriate σ-algebras, without explicitly specifying them.

We start stating several lemmas that are important for the first identification result (Proposition 1). The first lemma says that given the (X_i, W_i, U_i) other covariates cannot help in predicting $(Y_i(0), Y_i(1))$.

Lemma A1. (Statistical exclusion) *Under Assumptions 3.1, 3.3 the following is true:*

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp \{(\mathbb{P}_{C_j}, X_j, W_j)\}_{i=1}^N | X_i, W_i, U_i \tag{1.2}$$

Proof. From the repeated application of the iterated expectations and Assumptions 3.1, 3.3 we have the following:

$$\begin{aligned}
&\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}|X_i, W_i, U_i] = \\
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}|\{X_i, W_i, U_i, C_i\}_{i=1}^n]|X_i, W_i, U_i] = \\
&\mathbb{E}[\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|\{X_i, W_i, U_i, C_i\}_{i=1}^n]|X_i, W_i, U_i] = \\
&\mathbb{E}[\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, W_i, U_i]|X_i, W_i, U_i] = \\
&\qquad \mathbb{E}[\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}|X_i, W_i, U_i]\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, W_i, U_i]
\end{aligned} \tag{1.3}$$

Equality between the first and the last expression implies the independence result. □

The second lemma states that only $\mathbb{P}_{C_i}$ are useful in predicting U_i.

Lemma A2. (Statistical sufficiency) *Under Assumption 3.1 the following holds:*

$$U_i \perp\!\!\!\perp \{W_j, X_j\}_{j=1}^N | \mathbb{P}_{C_i} \tag{1.4}$$

Proof. The proof follows from the following equalities:

$$\begin{aligned}
&\mathbb{E}[\{U_i \in A\}\{\{W_j, X_j\}_{j=1}^N \in B\}|\mathbb{P}_{C_i}] = \\
&\mathbb{E}\left[\mathbb{E}[\{U_i \in A\}\{\{W_j, X_j\}_{j=1}^N \in B\}|\mathbb{P}_{C_i}, \{W_j, X_j\}_{j=1}^N, \{C_j = C_i\}_{j=1}^N]|\mathbb{P}_{C_i}\right] = \\
&\mathbb{E}\left[\{\{W_j, X_j\}_{j=1}^N \in B\}\mathbb{E}[\{U_i \in A\}|\mathbb{P}_{C_i}, \{W_j, X_j\}_{j=1}^N, \{C_j = C_i\}_{j=1}^N]|\mathbb{P}_{C_i}\right] = \\
&\mathbb{E}\left[\{\{W_j, X_j\}_{j=1}^N \in B\}\mathbb{E}[\{U_i \in A\}|\mathbb{P}_{C_i}, \{W_j, X_j\}_{j:C_j=C_i}]|\mathbb{P}_{C_i}\right] = \\
&\mathbb{E}\left[\{\{W_j, X_j\}_{j=1}^N \in B\}\mathbb{E}[\{U_i \in A\}|\mathbb{P}_{C_i}]|\mathbb{P}_{C_i}\right] = \\
&\qquad\qquad \mathbb{E}\left[\{\{W_j, X_j\}_{j=1}^N \in B\}|\mathbb{P}_{C_i}\right]\mathbb{E}\left[\{U_i \in A\}|\mathbb{P}_{C_i}\right]
\end{aligned} \tag{1.5}$$

The third equality holds by random sampling (observations in different clusters are independent), the fourth equality holds by exchangeability of data within the cluster. □

Proof of Proposition 1: We start with the following equalities:

$$\begin{aligned}
&\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, \mathbb{P}_{C_i}] = \\
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, \mathbb{P}_{C_i}, U_i]|W_i, X_i, \mathbb{P}_{C_i}] = \\
&\qquad\qquad \mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, U_i]|W_i, X_i, \mathbb{P}_{C_i}]
\end{aligned} \tag{1.6}$$

The last equality follows from Lemma A1. As a next step we have the following result:

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, U_i]|W_i, X_i, \mathbb{P}_{C_i}] = \\
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, U_i]|W_i, X_i, \mathbb{P}_{C_i}] = \\
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, U_i]|X_i, \mathbb{P}_{C_i}] = \\
&\quad \mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, \mathbb{P}_{C_i}, U_i]|X_i, \mathbb{P}_{C_i}] = \mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, \mathbb{P}_{C_i}] \quad (1.7)
\end{aligned}$$

The first equality follows directly from Assumption 3.2, the second equality follows from Lemma A2. Combining the two chains of equalities we get the following:

$$\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, \mathbb{P}_{C_i}] = \mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, \mathbb{P}_{C_i}] \quad (1.8)$$

which proves the conditional independence. □

Corollary A1. (Exclusion in exponential families) *Under the assumptions of Lemma A1 the following is true:*

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp \{(\overline{S}_{C_j}, X_j, W_j)\}_{j=1}^N | X_i, W_i, U_i \quad (1.9)$$

Proof. Because S_{C_i} is a function of $\mathbb{P}_{C_i}$ the result follows from Lemma A1. □

Lemma A3. (Sufficiency in exponential families) *Under Assumptions 3.1 and 3.4 the following holds:*

$$U_i \perp\!\!\!\perp \{W_j, X_j\}_{j=1}^N | \overline{S}_{C_i} \quad (1.10)$$

Proof. The proof is exactly the same as in Lemma A2 with S_{C_i} used instead of $\mathbb{P}_{C_i}$. The fourth equality now holds directly by the exponential family assumption. □

Proof of Theorem 1: The same as for Proposition 1, use Corollary A1 and Lemma A3 instead of Lemmas A1 and A2. □

Corollary A2. *For any function f such that $\mathbb{E}[|f(Y(k))|] < \infty$ the following is true:*

$$\mathbb{E}[f(Y_i)|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \\ \{W_i = 0\}\mathbb{E}[f(Y_i(0))|X_i, \overline{S}_{C_i}] + \{W_i = 1\}\mathbb{E}[f(Y_i(1))|X_i, \overline{S}_{C_i}] \quad (1.11)$$

Proof. The proof follows from the following equalities:

$$\mathbb{E}[f(Y_i)|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \mathbb{E}[\mathbb{E}[f(Y_i)|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N, U_i]|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \\ \mathbb{E}[\mathbb{E}[f(Y_i)|W_i, X_i, U_i]|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \mathbb{E}[f(Y_i)|W_i, X_i, \overline{S}_{C_i}] = \\ \{W_i = 0\}\mathbb{E}[f(Y_i(0))|X_i, \overline{S}_{C_i}] + \{W_i = 1\}\mathbb{E}[f(Y_i(1))|X_i, \overline{S}_{C_i}] \quad (1.12)$$

where the third equality follows from Corollary A1, the fourth from Lemma A3 and the final one from Proposition 1. □

B Inference results

Notation: We are using standard notation from the empirical processes literature adapted to our setting. For any **cluster-level** random vector X_c: $\mathbb{P}_n(X_c) \equiv \frac{1}{n}\sum_{c=1}^n X_c$ and $\mathbb{G}_n(X_c) \equiv \sqrt{n}\left(\mathbb{P}_n(X_c) - \mathbb{E}[X_c]\right)$. Define $B_i = (X_i, \overline{S}_{C_i})$ and $D_i \equiv (W_i, B_i)$.

We start with a reminder on notation:

$$\begin{cases} \mu(D_i) \equiv \mathbb{E}[Y_i|D_i] \\ e(B_i) \equiv \mathbb{E}[W_i|B_i] \\ \varepsilon(k) \equiv Y_i(k) - \mu(k, B_i) \\ \psi(y, w, x, s, \mu(\cdot), e(\cdot)) \equiv \mu(1, x, s) - \mu(0, x, s) + \left(\frac{w}{e(x,s)} - \frac{1-w}{1-e(x,s)}\right)(y - \mu(w, x, s)) \\ \rho(c, \mu(\cdot), e(\cdot)) \equiv \frac{1}{|c|}\sum_{i:C_i=c}\{A_i\}\psi(Y_i, W_i, X_i, \overline{S}_{C_i}, \mu(W_i, X_i, \overline{S}_{C_i}), e(X_i, \overline{S}_{C_i})) \\ \xi_c \equiv \sum_{i\in c}\frac{1}{N_c}\{A_i\}\left(\frac{W_i}{e(X_i, \overline{S}_{C_i})} - \frac{1-W_i}{1-e(X_i, \overline{S}_{C_i})}\right)(Y_i - \mu(W_i, X_i, \overline{S}_{C_i})) \end{cases} \quad (2.1)$$

In order to prove Theorem 2 we consider a more general case that allows for misspecification.

First we prove Lemma B4 which states that we get identification if either the propensity score or the conditional mean is potentially misspecified. Then we prove Proposition B1 which is a general consistency result under possible misspecification. Theorem 2 follows as a special case. After that we prove Theorem 3 and Proposition 2. Finally, all results below are proved assuming that $N_c = |c|$ is the same in all clusters. If this is not the case then, one can group the clusters of the same size and redo the analysis separately for each group. This approach is valid if the number of clusters of the same size grows linearly with the number of sampled clusters.

Lemma B4. *Assume that at least one of the following statements is true:*

$$\begin{cases} \tilde{\mu}(W_i, X_i, \overline{S}_{C_i}) = \mu(W_i, X_i, \overline{S}_{C_i}) \\ \tilde{e}(X_i, \overline{S}_{C_i}) = e(X_i, \overline{S}_{C_i}) \end{cases} \tag{2.2}$$

If the assumptions of Theorem 1 hold then we have the following result:

$$\mathbb{E}[\rho(c, \tilde{m}, \tilde{e})] = \mathbb{E}\left[\sum_{i \in c} \frac{1}{|c|}\{A_i\}\tau(B_i)\right] \tag{2.3}$$

where $\tau(B_i) := \mu(1, B_i) - \mu(0, B_i)$.

Proof. By construction we have the following:

$$\mathbb{E}[\rho(c, \tilde{\mu}, \tilde{e})] = \mathbb{E}\left[\sum_{i \in c} \frac{1}{|c|}\{A_i\}\left(\tilde{\mu}(1, B_i) - \tilde{\mu}(0, B_i) + \left(\frac{W_i}{\tilde{e}(B_i)} - \frac{1 - W_i}{1 - \tilde{e}(B_i)}\right)(Y_i - \tilde{\mu}(D_i))\right)\right] =$$
$$\mathbb{E}\left[\sum_{i \in c} \frac{1}{|c|}\{A_i\}(\tilde{\mu}(1, B_i) - \tilde{\mu}(0, B_i))\right] + \sum_{i \in c} \frac{1}{|c|}\mathbb{E}\left[\{A_i\}\left(\frac{W_i}{\tilde{e}(B_i)} - \frac{1 - W_i}{1 - \tilde{e}(B_i)}\right)(Y_i - \tilde{\mu}(D_i))\right] \tag{2.4}$$

For the second part we have the following (using unconfoundedness):

$$\mathbb{E}\left[\{A_i\}\left(\frac{W_i}{\tilde{e}(B_i)}-\frac{1-W_i}{1-\tilde{e}(B_i)}\right)(Y_i-\tilde{\mu}(D_i))\right]=$$
$$\mathbb{E}\left[\mathbb{E}\left[\{A_i\}\left(\frac{W_i}{\tilde{e}(B_i)}-\frac{1-W_i}{1-e(\tilde{B}_i)}\right)(Y_i-\tilde{\mu}(D_i))|B_i\right]\right]=$$
$$\mathbb{E}\left[\{A_i\}\left(\frac{e(B_i)\left(\mu(1,B_i)-\tilde{\mu}(1,B_i)\right)}{\tilde{e}(B_i)}-\frac{(1-e(B_i)\left(\mu(0,B_i)-\tilde{\mu}(0,B_i)\right)}{1-\tilde{e}(B_i)}\right)\right] \quad (2.5)$$

This implies that if either $\tilde{\mu}(D_i)=\mu(D_i)$ or $\tilde{e}(B_i)=e(B_i)$ then $\mathbb{E}[\rho(c,\tilde{m},\tilde{e})]=\mathbb{E}\left[\sum_{i\in c}\frac{1}{|c|}\{A_i\}\tau(B_i)\right]$. □

Proposition B1. (Consistency with wrong specifications) *Assume that the following conditions hold for* $(\hat{\tilde{e}},\hat{\tilde{\mu}})$*:*

$$\begin{cases}\mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\hat{\tilde{\mu}}(1,B_i)-\tilde{\mu}(1,B_i)\right)^2\right)=o_p(1)\\ \mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\hat{\tilde{e}}(B_i)-\tilde{e}(B_i)\right)^2\right)=o_p(1)\\ \eta<\tilde{e}(B_i)<1-\eta \ \ a.s.\\ \eta<\hat{\tilde{e}}(B_i)<1-\eta \ \ a.s.\\ \mathbb{E}[\tilde{\varepsilon}_i^2(k)]<\infty\end{cases} \quad (2.6)$$

where $\tilde{\varepsilon}_i(k):Y_i(k)-\tilde{\mu}(k,B_i)$*. Additionally assume that the conditions of Lemma B4 hold. Then we have the following:*

$$\mathbb{P}_n\rho(c,\tilde{\mu},\tilde{e})=\mathbb{P}_n\rho(c,\tilde{\mu},\tilde{e})+o_p(1)=\mathbb{E}[\rho(c,\tilde{\mu},\tilde{e})]+o_p(1) \quad (2.7)$$

Proof. To prove the consistency result we need to separate the functional into two parts:

$$\rho(c,\tilde{\mu},\tilde{e})=\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\tilde{\mu}(1,B_i)+\frac{W_i}{\tilde{e}(B_i)}(Y_i-\tilde{\mu}(1,B_i))\right)-$$
$$\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\tilde{\mu}(0,B_i)+\frac{1-W_i}{1-\tilde{e}(B_i)}(Y_i-\tilde{\mu}(0,B_i))\right)=\rho_1(c,\tilde{\mu},\tilde{e})-\rho_0(c,\tilde{\mu},\tilde{e}) \quad (2.8)$$

In what follows we are working only with the first part of the functional, the second can be analyzed in the exactly the same way. Define the empirical version:

$$\rho_1(c,\hat{\tilde{\mu}},\hat{\tilde{e}}) \equiv \sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\hat{\tilde{\mu}}(1,B_i)+\frac{W_i}{\hat{\tilde{e}}(B_i)}(Y_i-\hat{\tilde{\mu}}(1,B_i))\right) \tag{2.9}$$

We can decompose this expression into three parts:

$$\begin{aligned}\rho_1(c,\hat{\tilde{\mu}},\hat{\tilde{e}}) &= \sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\tilde{\mu}(1,B_i)+\frac{W_i}{\tilde{e}(B_i)}(Y_i-\tilde{\mu}(1,B_i))\right)+\\ &+\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\left(\hat{\tilde{\mu}}(1,B_i)-\tilde{\mu}(1,B_i)\right)\left(1-\frac{W_i}{\hat{\tilde{e}}(B_i)}\right)\right)+\\ &\sum_{i\in c}\frac{1}{|c|}\{A_i\}(Y_i-\tilde{\mu}(1,B_i))W_i\left(\frac{1}{\hat{\tilde{e}}(B_i)}-\frac{1}{\tilde{e}(B_i)}\right)=\rho_1(c,\tilde{\mu},\tilde{e})+R_{1c}+R_{2c}\end{aligned} \tag{2.10}$$

The result will follow once we prove two approximations:

$$\begin{cases}\mathbb{P}_n R_{1c}=o_p(1)\\ \mathbb{P}_n R_{2c}=o_p(1)\end{cases} \tag{2.11}$$

We start with the second one. Observe that we have the following:

$$\begin{aligned}|\mathbb{P}_n R_{2c}|\leq\mathbb{P}_n|R_{2c}|\leq\mathbb{P}_n\sum_{i\in c}\frac{1}{|c|}\{A_i\}|\tilde{\varepsilon}_i(1)|\left(\frac{\{A_i\}W_i}{\tilde{e}(B_i)\hat{\tilde{e}}(B_i)}\right)\{A_i\}\left|\tilde{e}(B_i)-\hat{\tilde{e}}(B_i)\right|\leq\\ \max_i\left(\frac{\{A_i\}W_i}{\tilde{e}(B_i)\hat{\tilde{e}}(B_i)}\right)\sqrt{\mathbb{P}_n\sum_{i\in c}\frac{1}{|c|}\{A_i\}\tilde{\varepsilon}_i^2(1)}\sqrt{\mathbb{P}_n\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\tilde{e}(B_i)-\hat{\tilde{e}}(B_i)\right)^2}=\\ O_p(1)\sqrt{O_p(1)}\sqrt{o_p(1)}=o_p(1)\end{aligned} \tag{2.12}$$

For the first term we have the following:

$$R_{1c} = \sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\left(\hat{\tilde{\mu}}(1, B_i) - \tilde{\mu}(1, B_i) \right) \left(1 - \frac{W_i}{\hat{\tilde{e}}(B_i)} \right) \right) =$$
$$\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\left(\hat{\tilde{\mu}}(1, B_i) - \tilde{\mu}(1, B_i) \right) \left(1 - \frac{W_i}{\tilde{e}(B_i)} \right) \right) +$$
$$\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\left(\hat{\tilde{\mu}}(1, B_i) - \tilde{\mu}(1, B_i) \right) W_i \left(\frac{\hat{\tilde{e}}(B_i) - \tilde{e}(B_i)}{\tilde{e}(B_i) \hat{\tilde{e}}(B_i)} \right) \right) = R_{11c} + R_{12c} \quad (2.13)$$

The first part can be bounded in the following way:

$$|\mathbb{P}_n R_{11c}| \leq \mathbb{P}_n |R_{11c}| \leq$$
$$\max_i \left| \frac{\{A_i\}(W_i - \tilde{e}(B_i))}{\tilde{e}(B_i)} \right| \times \sqrt{\mathbb{P}_n \left(\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\hat{\tilde{m}}(1, B_i) - \tilde{\mu}(1, B_i) \right)^2 \right)} =$$
$$O_p(1) \times o_p(1) = o_p(1) \quad (2.14)$$

The second part can be bounded in the following way:

$$|\mathbb{P}_n R_{12c}| \leq \mathbb{P}_n |R_{12c}| \leq \max_i \left(\frac{\{A_i\} W_i}{\tilde{e}(B_i) \hat{\tilde{e}}(B_i)} \right) \times$$
$$\sqrt{\mathbb{P}_n \left(\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\hat{\tilde{\mu}}(1, B_i) - \tilde{\mu}(1, B_i) \right)^2 \right)} \sqrt{\mathbb{P}_n \left(\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\hat{\tilde{e}}(B_i) - \tilde{e}(B_i) \right)^2 \right)} =$$
$$O_p(1) \times o_p(1) \, o_p(1) = o_p(1) \quad (2.15)$$

Combining all the results together we have the proof. □

Proof of Theorem 2: Observe that $\hat{e}$ and $\hat{\mu}$ satisfy the assumptions of Proposition B1 with $\tilde{\mu}$ and $\tilde{e}$ equal to m and e. As a result, combining Proposition B1 and Lemma B4 we get the

following:

$$\frac{1}{\hat{\pi}(A)}\mathbb{P}_n\rho(c,\hat{\mu},\hat{e}) = \frac{1}{\hat{\pi}(A)}\left(\mathbb{E}[\rho(c,\mu,e)] + o_p(1)\right) =$$
$$\left(\frac{1}{\pi(A)} + o_p(1)\right)\left(\mathbb{E}[\rho(c,\mu,e)] + o_p(1)\right) = \frac{1}{\pi(A)}\mathbb{E}[\rho(c,\mu,e)] + o_p(1) =$$
$$\frac{1}{\pi(A)}\mathbb{E}[\{A_i\}\tau(X_i,\overline{S}_{C_i})] + o_p(1) \quad (2.16)$$

Proof of Theorem 3: The start of the argument is the same as in proof for the consistency result. We decompose the empirical version of $\rho_1(c,\hat{m},\hat{e})$:

$$\rho_1(c,\hat{m},\tilde{e}) - \sum_{i\in c}\frac{1}{|c|}\{A_i\}\mu(1,B_i) = \sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\frac{W_i}{e(B_i)}(Y_i - \mu(1,B_i))\right) +$$
$$+\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left((\hat{\mu}(1,B_i) - \mu(1,B_i))\left(1 - \frac{W_i}{\hat{e}(B_i)}\right)\right) +$$
$$\sum_{i\in c}\frac{1}{|c|}\{A_i\}(Y_i - \mu(1,B_i))W_i\left(\frac{1}{\hat{e}(B_i)} - \frac{1}{e(B_i)}\right) = \xi_{1c} + R_{1c} + R_{2c} \quad (2.17)$$

The result will follow once we prove the following:

$$\begin{cases}\mathbb{P}_n R_{1c} = o_p\left(\frac{1}{\sqrt{n}}\right)\\ \mathbb{P}_n R_{2c} = o_p\left(\frac{1}{\sqrt{n}}\right)\end{cases} \quad (2.18)$$

In exactly the same way as before we can decompose R_{1c} into R_{11c} and R_{12c}. For R_{12c} we have the following:

$$|\mathbb{P}_n R_{12c}| \leq \mathbb{P}_n|R_{12c}| \leq \max_i\left(\frac{\{A_i\}W_i}{e(B_i)\hat{e}(B_i)}\right)\times$$
$$\sqrt{\mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\hat{\mu}(1,B_i) - \mu(1,B_i)\right)^2\right)}\sqrt{\mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\hat{e}(B_i) - e(B_i)\right)^2\right)} =$$
$$O_p(1)\times o_p\left(\frac{1}{\sqrt{n}}\right) = o_p\left(\frac{1}{\sqrt{n}}\right) \quad (2.19)$$

For R_{11c} we use the following argument:

$$\mathbb{E}\left[(\mathbb{P}_n R_{11c})^2\right] = \mathbb{E}\left[\left(\sum_{l\in L}\sum_{c:l(c)=l}\frac{1}{n}\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)\left(1-\frac{W_i}{e(B_i)}\right)\right)^2\right] \leq$$

$$|L|\sum_{l\in L}\mathbb{E}\left[\left(\sum_{c:l(c)=l}\frac{1}{n}\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)\left(1-\frac{W_i}{e(B_i)}\right)\right)^2\right] =$$

$$|L|\sum_{l\in L}\sum_{c:l(c)=l}\frac{1}{n}\mathbb{E}\left[\frac{1}{n}\left(\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)\left(1-\frac{W_i}{e(B_i)}\right)\right)^2\right] \leq$$

$$|L|\sum_{l\in L}\sum_{c:l(c)=l}\frac{1}{n}\mathbb{E}\left[\frac{1}{n}\sum_{i\in c}\frac{1}{|c|}\left(\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)\left(1-\frac{W_i}{e(B_i)}\right)\right)^2\right] =$$

$$|L|\sum_{l\in L}\sum_{c:l(c)=l}\frac{1}{n}\mathbb{E}\left[\frac{1}{n}\sum_{i\in c}\frac{1}{|c|}\left(\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)^2\left(\frac{e(B_i)(1-e(B_i)}{e^2(B_i)}\right)\right)\right] \leq$$

$$K\frac{1}{n}\mathbb{E}\left[\mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)^2\right)\right] \quad (2.20)$$

Using this we get the following:

$$\mathbb{E}\left[|\mathbb{P}_n R_{11c}|\right] \leq \sqrt{\mathbb{E}\left[(\mathbb{P}_n R_{11c})^2\right]} \leq$$

$$\frac{K}{\sqrt{n}}\mathbb{E}\left[\mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)^2\right)\right] = o\left(\frac{1}{\sqrt{n}}\right) \quad (2.21)$$

This implies (by Markov's inequality) that $\mathbb{P}_n R_{11c} = o_p\left(\frac{1}{\sqrt{n}}\right)$

$$\mathbb{E}[R_{2c}^2|\{D_i\}_{i=1}^N] \leq \sum_{i\in c}\frac{1}{|c|}\mathbb{E}[\varepsilon_i^2|D_i]\left(\frac{\{A_i\}W_i}{e^2(B_i)\hat{e}^2(B_i)}\right)\{A_i\}\left(e(B_i)-\hat{e}(B_i)\right)^2 \leq$$

$$\max_i\left(\frac{\{A_i\}\mathbb{E}[\varepsilon_i^2|D_i]W_i}{e^2(B_i)\hat{e}(B_i)}\right)\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(e(B_i)-\hat{e}(B_i)\right)^2 \quad (2.22)$$

We also have the following:

$$\mathbb{E}[R_{2c}|\{D_i\}_{i=1}^N] = 0 \quad (2.23)$$

Using these two things we get the following:

$$\mathbb{E}[(\mathbb{P}_n R_{2c})^2 \,|\{D_i\}_{i=1}^N] \leq \max_i \left(\frac{\{A_i\}\mathbb{E}[\varepsilon_i^2|D_i]W_i}{e^2(B_i)\hat{e}(B_i)}\right) \times \\ \frac{1}{n}\mathbb{P}_n \sum_{i\in c} \frac{1}{|c|}\{A_i\}\left(e(B_i)-\hat{e}(B_i)\right)^2 \leq K \times o_p\left(\frac{1}{n}\right) = o_p\left(\frac{1}{n}\right) \quad (2.24)$$

This implies that $\mathbb{E}[(\mathbb{P}_n R_{2c})^2] = o\left(\frac{1}{n}\right)$ (because $(\hat{e}-e)^2$ is bounded by 1) and thus $R_{2c} = o_p\left(\frac{1}{\sqrt{n}}\right)$.

Proof of Proposition 2: Similarly to all other proofs we can divide ξ_c into two parts ξ_{1c} and ξ_{0c}. We will analyze ξ_{1c}, analysis for ξ_{0c} is the same. We have the following decomposition:

$$\hat{\xi}_{1c} - \xi_{1c} = \sum_{i\in c} \frac{1}{|c|}\{A_i\}\left(\left(\mu(1,B_i)-\hat{\mu}(1,B_i)\right)\frac{W_i}{\hat{e}(B_i)}\right) + \\ \sum_{i\in c} \frac{1}{|c|}\{A_i\}(Y_i-\mu(1,B_i))W_i\left(\frac{1}{\hat{e}(B_i)}-\frac{1}{e(B_i)}\right) = R_{11c}+R_{12c} \quad (2.25)$$

For the first term we have the following bound:

$$\mathbb{P}_n R_{11c}^2 \leq \mathbb{P}_n \frac{1}{c}\sum_{i\in c}\{A_i\}\left(\left(\mu(1,B_i)-\hat{\mu}(1,B_i)\right)^2\frac{W_i}{\hat{e}^2(B_i)}\right) \leq \\ \left(\max_i \frac{\{A_i\}W_i}{\hat{e}^2(B_i)}\right) \times \mathbb{P}_n\left(\frac{1}{c}\sum_{i\in c}\{A_i\}\left(\mu(1,B_i)-\hat{\mu}(1,B_i)\right)^2\right) = O_p(1)o_p(1) = o_p(1) \quad (2.26)$$

For the second term we have the following bound:

$$\mathbb{P}_n R_{12c}^2 \leq \mathbb{P}_n \frac{1}{c}\sum_{i\in c}\{A_i\}\{W_i\}\varepsilon_i^2(1)\frac{(\hat{e}(B_i)-e(B_i))^2}{\hat{e}^2(B_i)e^2(B_i)} \leq \\ \sqrt{\left(\mathbb{P}_n \frac{1}{c}\sum_{i\in c}\{A_i\}\frac{\{A_i\}(\hat{e}(B_i)-e(B_i))^4}{\hat{e}^4(B_i)e^4(B_i)}\right)\left(\mathbb{P}_n\frac{1}{c}\sum_{i\in c}\{A_i\}\{W_i\}\varepsilon_i^4(1)\right)} \leq \\ K\sqrt{\left(\mathbb{P}_n \frac{1}{c}\sum_{i\in c}\{A_i\}(\hat{e}(B_i)-e(B_i))^2\right)\left(\mathbb{P}_n\frac{1}{c}\sum_{i\in c}\{A_i\}\{W_i\}\varepsilon_i^4(1)\right)} = \\ o_p(1)O_p(1) = o_p(1) \quad (2.27)$$

Putting these results together we have the following:

$$\mathbb{P}_n(\hat{\xi}_{1c}+\hat{\xi}_{2c})^2-\mathbb{P}_n(\xi_{1c}+\xi_{2c})^2=\mathbb{P}_n(\xi_{1c}+\xi_{2c}+R_{11c}+R_{12c}+R_{01c}+R_{02c})^2-\mathbb{P}_n(\xi_{1c}+\xi_{2c})^2=$$
$$\mathbb{P}_n(\xi_{1c}+\xi_{2c})(R_{11c}+R_{12c}+R_{01c}+R_{02c})+\mathbb{P}_n(R_{11c}+R_{12c}+R_{01c}+R_{02c})^2\leq$$
$$\sqrt{\mathbb{P}_n(\xi_{1c}+\xi_{2c})^2 4\mathbb{P}_n(R_{11c}^2+R_{12c}^2+R_{01c}^2+R_{02c}^2)}+\mathbb{P}_n(R_{11c}^2+R_{12c}^2+R_{01c}^2+R_{02c}^2)=$$
$$\sqrt{O_p(1)o_p(1)}+o_p(1)=o_p(1) \quad (2.28)$$

This argument also implies that $\mathbb{P}_n(\hat{\xi}_{1c})=\mathbb{P}_n(\xi_{1c})=o_p(1)$ and thus we have the final result:

$$\frac{1}{\hat{\pi}^2(A)}\left(\mathbb{P}_n(\hat{\xi}_{1c}+\hat{\xi}_{2c})^2-\left(\mathbb{P}_n(\hat{\xi}_{1c}+\hat{\xi}_{2c})\right)^2\right)-\frac{1}{\pi^2(A)}\mathbb{P}_n(\xi_{1c}+\xi_{2c})^2=$$
$$\frac{1}{\hat{\pi}^2(A)}\left(\mathbb{P}_n(\hat{\xi}_{1c}+\hat{\xi}_{2c})^2-\mathbb{P}_n(\xi_{1c}+\xi_{2c})^2\right)+\left(\frac{1}{\hat{\pi}^2(A)}-\frac{1}{\pi^2(A)}\right)\mathbb{P}_n(\xi_{1c}+\xi_{2c})^2+O_p(1)o_p(1)=$$
$$O_p(1)o_p(1)+o_p(1)O_p(1)+O_p(1)o_p(1)=o_p(1) \quad (2.29)$$