The Role of the Propensity Score in Fixed Effect Models*

Dmitry Arkhangelsky † Guido W. Imbens‡

February 2022

Abstract

We develop a new approach for estimating average treatment effects in observational studies with unobserved group-level heterogeneity. In such settings, a common empirical strategy is to use linear fixed-effect specifications estimated by least squares. We make two observations about this approach. First, the fixed-effect regression controls for differences between groups only by adjusting linearly for the average values of covariates and treatments. Second, weighting by the inverse of the propensity score would also remove biases for comparisons between treated and control units under the fixed effect setup. We develop three generalizations of the fixed effect approach based on these observations. First, we suggest using flexible adjustments for the average covariate values. Second, we propose robustifying the estimators by using inverse propensity score weighting. Third, we consider adjustments for group characteristics beyond average covariate values. In practice, we recommend researchers use all three generalizations.

Keywords: fixed effects, cross-section data, clustering, causal effects, treatment effects, unconoundedness.

additional references: Rosenbaum [1984], Pesaran [2006]

*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, Matias Cattaneo, and Manuel Arellano. 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.

†Assistant Professor, CEMFI, darkhangel@cemfi.es.

‡Professor of Economics, Graduate School of Business, Stanford University, SIEPR, and NBER, imbens@stanford.edu.
1 Introduction

A common specification for regression functions when estimating causal effects with grouped or clustered data is a fixed effect specification:

\[ Y_i = \alpha_{C_i} + W_i \tau + X_i^\top \beta + \varepsilon_i, \]  

(1.1)

where \( C_i \in \{1, \ldots, C\} \) indicates which cluster a unit belongs to, and the \( \alpha_c \) are the cluster fixed effects (see Wooldridge [2010], Angrist and Pischke [2008] for textbook discussions). The regression function is typically estimated by least squares. The coefficient \( \tau \), interpreted as the causal effect of the treatment or causal variable \( W_i \) (binary through most of this paper), is the object of interest. The fixed effects \( \alpha_c \) are intended to capture unobserved differences between the clusters. The motivation for including them in the regression is that their presence improves the credibility of a causal interpretation of the least squares estimator (e.g., Arellano [2003], Angrist and Pischke [2008]). Inference for \( \tau \) is often based on asymptotic approximation with the growing number of clusters and fixed number of units per cluster. This specification is popular in empirical work where the clusters may correspond to states, cities, SMSAs, classrooms, birth cohorts, firms, or other geographic or demographic groups.

The popularity of the fixed effect specification may be due partly because it appears to account for differences between clusters in a flexible way. However, the specification in (1.1) embodies multiple assumptions. It is not immediately clear which ones are critical and which ones are not, or how the latter can be relaxed. In the current paper, we unpack these assumptions in a novel way allowing the researcher to make more informed specification choices. We focus on three issues. First, the specification in (1.1) assumes a constant treatment effect. In practice, there is likely heterogeneity in treatment effects. This is important for the interpretation of the fixed effect estimator. We show that, in general, the average effect of the treatment is not point-identified. We then characterize the estimand corresponding to the fixed effect estimator under general treatment-effect heterogeneity and demonstrate that it estimates, under some conditions, a weighted average treatment effect. The weights depend in a somewhat unusual way on the sampling scheme. Second, the typical assumptions motivating the specification in (1.1) always validate within-cluster comparisons of treated and control units with the same covariates. However, these assumptions implicitly validate some, but not all, cross-cluster comparisons of treated and control units with the same covariates. We clarify precisely which cross-cluster
comparisons are validated by the fixed effect specification and how the researcher can control directly which cross-cluster comparisons have causal interpretations. This is critical in grouped data settings because often, there are too few units within clusters to make inferences relying solely on within-cluster comparisons sufficiently precise. Third, the fixed effect specification embodies functional form restrictions. We show how to relax these functional form assumptions.

We focus on formal results for the case with a modest number of units per group. In particular, we discuss settings where the group size is not large enough to carry out a two-stage procedure where we first estimate the effects entirely within clusters by flexibly adjusting for covariates, followed by averaging over the clusters. In other words, we need to rely on comparisons of treated and control units in different clusters, and there is concern that accounting for the cluster differences solely through additive fixed effects is not sufficient to adjust for all relevant differences (e.g., Altonji and Matzkin [2005], Imai and Kim [2019]).

2 Unconfoundedness within Clusters, Two Observations and A Simple Example

In this section, we give intuition for the main results and insights of the paper through an informal discussion in some special cases. In Section 3 we present a more formal discussion of the general case. We start in this section by making two observations about the fixed effect estimator that suggest a way to relax the functional form assumptions. Next, we discuss two versions of a simple example with two units per cluster. This serves to make some points about the interpretation of the estimand and the validation of particular cross-cluster comparisons.

2.1 Unconfoundedness within Clusters

A key assumption underlying most cluster analyses is that assignment is unconfounded within the clusters. Suppose each unit $i$ is characterized by a pair of potential outcomes, $(Y_i(0), Y_i(1))$, and is assigned to a binary treatment $W_i$. For each unit we also observe a cluster $C_i$ that this unit belongs to. In the absence of individual-level covariates, we can express cluster unconfoundedness in the following way:

$$W_i \perp \perp (Y_i(0), Y_i(1)) \mid C_i.$$  \hspace{1cm} (2.1)
This assumption validates comparisons between treated and control units within clusters. In addition we make the common assumption that units within a cluster are exchangeable.

We focus here on the case where for all clusters in the sample we observe the same number of units per cluster. Then cluster unconfoundedness implies the following conditional independence restriction:

\[ W_i \perp (Y_i(0), Y_i(1)) \mid W_c, \]

where \( W_c \) is the share of treated units in cluster \( c \). This unconfoundedness condition implies that we can combine clusters with the same value for \( W_c \). This does not immediately have a lot of empirical content. To see this, consider the special case with two units per cluster. This implies there are three values for \( W_c \), namely 0, 1/2, and 1. If \( W_c = 0 \) or \( W_c = 1 \), the combined set of clusters has only control units or only treated units, so there is no basis to estimate treatment effects. If we look at a set of clusters with \( W_c = 1/2 \), comparing treated and control units gives us an average of within-cluster comparisons of treated and control units, which were validated already by (2.1).

Despite the lack of empirical content, there are two critical insights in going from (2.1) to (2.2). Both insights are about operationalizing the notion that some clusters are more similar than others, which is absent in (2.1). Because the conditioning in (2.1) is on an unordered discrete characteristic, there is no distance metric to make the case that cluster \( c \) is closer to cluster \( c' \) or to cluster \( c'' \). In contrast, (2.2) allows for the establishment of such a metric. The first insight is directly about smoothing. Based on (2.2) we now do have a meaningful distance metric: a cluster \( c \) with \( W_c = 0.10 \) is closer to cluster \( c' \) with \( W_{c'} = 0.11 \) than it is to cluster \( c'' \) with \( W_{c''} = 0.50 \). The second insight is even more central to the current paper. It allows us to find a middle ground between conditioning on the full set of cluster indicators and not conditioning on the clustering indicators at all. This insight depends on the presence of covariates. In that case, there is an extension of the equivalence result between (2.1) to (2.2) where conditioning on the cluster indicator \( C_i \) and the covariates \( X_i \) is equivalent to conditioning on \( X_i \) and a set of cluster-level averages. Specifically, in the single binary covariate case cluster unconfoundedness

\[ W_i \perp (Y_i(0), Y_i(1)) \mid X_i, C_i. \]
is (given independence of the units within a cluster), equivalent to

\[ W_i \perp \perp \left( Y_i(0), Y_i(1) \right) \mid X_i, \overline{W}_C, \overline{X}_C, \overline{WX}_C. \] (2.4)

The advantage of (2.4) is that it allows us to consider stronger conditions where conditioning on only a subset of \((X_i, \overline{W}_C, \overline{X}_C, \overline{WX}_C)\) is sufficient to remove the dependence between \(W_i\) and \((Y_i(0), Y_i(1))\). For example by assuming that adjustment for \((X_i, \overline{W}_C, \overline{X}_C)\) is sufficient:

\[ W_i \perp \perp \left( Y_i(0), Y_i(1) \right) \mid X_i, \overline{W}_C, \overline{X}_C, \] (2.5)

or even

\[ W_i \perp \perp \left( Y_i(0), Y_i(1) \right) \mid X_i, \overline{W}_C. \] (2.6)

We discuss in Section ?? whether, and when, it is enough to adjust for some subset of these averages. The point here is that moving from (2.3) to (2.4) allows us at the very least to consider such intermediate cases.

### 2.2 Two Observations

Next we make two observations regarding the fixed effect specification. The first observation is that we can estimate \(\tau\), the coefficient of interest in (1.1), by least squares estimation of an ostensibly different regression function, namely

\[ Y_i = \alpha + W_i \tau + X_i^\top \beta + \overline{W}_C \delta + \overline{X}_C \gamma + \varepsilon_i, \] (2.7)

where \(\overline{W}_c\) and \(\overline{X}_c\) are the cluster averages of \(W_i\) and \(X_i\) in cluster \(c\). This numerical equivalence follows from repeated applications of textbook omitted variable bias formulas. This observation is not new to us. It was first mentioned in Mundlak [1978] in a seminal paper on panel data. The representation in (2.7) suggests viewing the fixed effect regression as adjusting for differences between clusters by adjusting linearly for cluster differences in the cluster averages \(\overline{W}_c\) and \(\overline{X}_c\). That in turn suggest thinking about relaxing the assumption underlying the estimator for \(\tau\) to a functional-form-free unconfoundedness-type assumption common in the program evaluation literature (Rosenbaum and Rubin [1983], Imbens and Wooldridge [2009], Abadie and Cattaneo [2018]). Instead of assuming that the potential outcomes depend linearly on \((X_i, \overline{W}_C, \overline{X}_C)\), one
could assume that nonparametrically conditioning on them removes the dependence between $W_i$ and $(Y_i(0), Y_i(1))$:

$$W_i \perp \perp (Y_i(0), Y_i(1)) \mid X_i, W_{Ci}, \bar{X}_{Ci}. \quad (2.8)$$

The second observation, also inspired by the modern causal inference literature, is that if assignment to treatment is completely random, independent of both covariates and cluster membership, the fixed effect estimator is consistent for the population average treatment effect even if the conditional expectation of the outcome is not not linear in covariates, treatment, and fixed effects. More generally, the difference in outcomes by treatment status is consistent for the average treatment effect as long as we weight the units by the inverse of the propensity score:

$$(\hat{\alpha}, \hat{\tau}) = \arg \min_{\alpha, \tau} \frac{1}{N} \sum_{i=1}^{N} (Y_i - \alpha - W_i \tau)^2 \frac{1}{p(X_i, C_i)^{W_i} (1 - p(X_i, C_i))^{1-W_i}}, \quad (2.9)$$

where $p(x, c) \equiv \text{pr}(W_i = 1|X_i = x, C_i = c)$ is the propensity score. This observation suggests using inverse propensity score weighting as an alternative to outcome modelling to adjust for general group differences. This observation is not straightforward to exploit because consistently estimating the propensity function as a function of the cluster indicator is impossible under asymptotics with a fixed number of units per cluster.

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 functional form restrictions in the fixed effect approach embodied in (1.1). First, the conditional independence in (2.8) suggest that we can use more flexible specifications for the regression function as a function of the control variables $(X_i, W_{Ci}, \bar{X}_{Ci})$ and the treatment $W_i$. In general we can specify the regression function as:

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

with a parametric or non-parametric specification for $g(\cdot)$ that generalizes the linear additive form in (1.1). 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 $\hat{g}(1, X_i, \bar{W}_{Ci}, \bar{X}_{Ci}) - \hat{g}(0, X_i, \bar{W}_{Ci}, \bar{X}_{Ci})$ over the sample to estimate the average treatment effect.
Second, the conditional independence in (2.8) suggest that we can model the propensity score as a function of the control variables \((X_i, W_{C_i}, X_{C_i})\):

\[
e(X_i, W_{C_i}, X_{C_i}) \equiv \text{pr}(W_i = 1|X_i, W_{C_i}, X_{C_i}),
\]

using this instead of \(p(X_i, C_i)\) in (2.9). 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 argued in the recent 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 (e.g., Robins and Rotnitzky [1995], Chernozhukov et al. [2017]).

Third, and this is perhaps the most important conceptual insight in the paper, the representation in (2.7) and the associated unconfoundedness assumption in (2.8) highlight that the fixed effect specification implicitly assumes that the two averages \(W_c\) and \(X_c\) capture all the relevant differences between the clusters. In other words, we can compare treated and control units in different clusters, as long as the clusters have the same values for \(W_c\) and \(X_c\). A natural question is whether these two averages do in fact capture all the relevant differences between clusters. 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 (Belloni et al. [2014], Imbens and Rubin [2015]).

In practice our recommendation is to use all three modifications: First choose what cluster characteristics one wishes to include in the analysis, beyond the cluster averages of the covariates and treatment that are included in the standard fixed effect approach. Second, specify a credible conditional mean function including these additional cluster characteristics. Third, estimate the propensity score as a function of individuals and cluster characteristics and combine that with the conditional mean specification to obtain a more robust estimator for the average treatment effect.
2.3 An Example

Next, we discuss two versions of an example, one version with and one without covariates. This example illustrates two themes that will come up repeatedly in our discussion. The first point we make in the no-covariate case is that with heterogeneous treatment effects, one needs to be careful about the definition or interpretation of the estimand. More specifically, under the standard assumptions, the fixed effect estimator is not estimating the population average effect. Instead, it estimates a weighted average treatment effect, with weights depending on the number of units sampled per cluster. The second point we make is that with a single binary covariate the additive specification validates a very particular set of cross-cluster comparisons.

In both versions of the example, we have a population with a large number of clusters, all with the same large number of units. The sampling process has two stages. First, we randomly sample \( C \) clusters from the population of clusters. Second, we sample \( N_c \) units from cluster \( c \) for all the sampled clusters. In addition to outcomes and treatment indicators, for each unit we observe a binary covariate \( X_i \). We focus on the case \( N_c = 2 \), where we sample exactly two units from each cluster. Let \( C_i \in \{1, \ldots, C\} \) be the cluster that unit \( i \) belongs to. Because the number of units in the sample for each sampled cluster is two, \( W_c \), the fraction of treated units in cluster \( c \), can take on three values, \( W_c \in \{0, 1/2, 1\} \). For \( s \in \{0, 1/2, 1\} \), let \( M_s \) denote the number of clusters in the sample with \( W_c = s \). Finally, let \( \tau_c \) be the population average treatment effect within cluster \( c \).

2.3.1 A Simple Example Without Covariates

First, suppose there are no covariates. In this case, the fixed effect estimator has a straightforward form. The estimator reduces to averaging the difference between the treated unit and the control unit over the \( M_{1/2} \) clusters with exactly one treated and one control unit:

\[
\hat{\tau} = \frac{1}{M_{1/2}} \sum_{c: W_c = 1/2} \hat{\tau}_c, \quad \text{where} \quad \hat{\tau}_c = \sum_{i: C_i = c} W_i Y_i - \sum_{i: C_i = c} (1 - W_i) Y_i.
\]

Note that this estimator does not depend on outcomes for clusters with either only treated units \( W_c = 1 \), or clusters with only control units, \( W_c = 0 \). For clusters with \( W_c = 1/2 \), \( \hat{\tau}_c \) is the difference between a treated and a control outcome from the same cluster, so it is a natural, and in fact the only natural, estimator for the average effect within that cluster given that there are
only two units from that cluster in the sample.

The first question is what \( \hat{\tau} \) is estimating in settings with heterogeneous treatment effects under unconfoundedness. The unconfoundedness assumption implies that \( \hat{\tau}_c \) is unbiased for \( \tau_c \) for clusters with \( \overline{W}_c = 1/2 \). The population average treatment effect is \( \tau = \mathbb{E}[\tau_{C_i}] \). Define for \( s \in \{0, 1/2, 1\} \) the weighted average effect

\[
\tau(s) = \mathbb{E}[\tau_{C_i}|\overline{W}_{C_i} = s].
\] (2.10)

This is a critical, though somewhat unusual, object. \( \tau(s) \) is a conditional average treatment effect, but the conditioning depends on the sampling scheme and the assignment mechanism. The overall average effect can be expressed in terms of the \( \tau(s) \):

\[
\tau = \mathbb{E}[\tau_{C_i}] = \text{pr}(\overline{W}_{C_i} = 0)\mathbb{E}[\tau_{C_i}|\overline{W}_{C_i} = 0] + \text{pr}(\overline{W}_{C_i} = 1/2)\mathbb{E}[\tau_{C_i}|\overline{W}_{C_i} = 1/2]
\]

\[
+ \text{pr}(\overline{W}_{C_i} = 1)\mathbb{E}[\tau_{C_i}|\overline{W}_{C_i} = 1]
\]

\[
= \text{pr}(\overline{W}_{C_i} = 0)\tau(0) + \text{pr}(\overline{W}_{C_i} = 1/2)\tau(1/2) + \text{pr}(\overline{W}_{C_i} = 1)\tau(1).
\]

The expected value of \( \hat{\tau} \) conditional on \( \overline{W}_{C_i} \) is \( \tau(1/2) \), and if \( \tau(s) \) varies by \( s \), this means that \( \tau(1/2) \) differs from \( \tau \), and thus that \( \hat{\tau} \) estimates something different from \( \tau \).

To further illustrate what \( \hat{\tau} \) is estimating, let us define the propensity score,

\[ U_{C_i} \equiv \text{pr}(W_i = 1|C_i). \]

Let us assume that it has the following distribution over clusters,

\[ \text{pr}(U_c = u) = 1/3, \ u \in \{0, 1/2, 1\}. \]

In addition, we assume

\[ \mathbb{E}[Y_i(1) - Y_i(0)|U_{C_i} = u] = u^2. \]

In this setup, the overall average treatment effect is \( \tau = \mathbb{E}[Y_i(1) - Y_i(0)] = E[U_{C_i}^2] = 5/12. \)

However, \( \text{pr}(U_{C_i} = 1/2|\overline{W}_{C_i} = 1/2) = 1 \), so that

\[ \tau(1/2) = 1/4. \]

Conditioning on clusters with one treated and one control observation shifts the distribution of
from its marginal distribution to a distribution that puts all the weight on values of \( u \) that make the observation of both a treated and control unit possible. It is easy to see that the interpretation of the fixed effect estimator changes if we change the sampling scheme, e.g., if we sample \( N_c = 3 \) units for each cluster, or if the assignment mechanism changes.

### 2.3.2 An Example With Covariates

Next we generalize the example to allow for the presence of a single binary covariate. Define for each cluster \( c \) and each covariate value \( x \) the average treatment effect:

\[
\tau_c(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x, C_i = c].
\] (2.11)

Like \( \tau_c \) in the previous subsection, this quantity is well defined for all \((x, c)\) such that \( x \) is in the support of \( X_i \) in cluster \( c \), irrespective of whether the number of units in a cluster is finite or infinite.

For any statistic \( S : \mathbb{X} \times \{0, 1\} \rightarrow \mathbb{R}^d \), let \( \overline{S}_c \) denote the average of \( S(X_i, W_i) \) over the units in the sample in cluster \( c \). Define the conditional expectation of \( \tau_{C_i}(X_i) \) where the expectation is over all units with \( X_i = x \) and over all clusters with \( S_{C_i} = s \):

\[
\tau(s, x) \equiv \mathbb{E} [\tau_{C_i(x)} | \overline{S}_{C_i} = s, X_i = x] = \mathbb{E} [Y_i(1) - Y_i(0) | \overline{S}_{C_i} = s, X_i = x]
\] (2.12)

Like \( \tau(s) \) in the previous subsection, this is a somewhat unusual object because it is partly defined in terms of the sampling scheme, and not a fixed population characteristic.

We assume cluster unconfoundedness,

\[
W_i \perp \perp (Y_i(0), Y_i(1)) \mid X_i, C_i.
\] (2.13)

or, in this binary covariate setting, its equivalent, given exchangeability,

\[
W_i \perp \perp (Y_i(0), Y_i(1)) \mid X_i, \overline{W}_{C_i}, \overline{X}_{C_i}, \overline{WX}_{C_i}.
\] (2.14)

This assumption validates within-cluster comparisons of treated and control units with the same value for the covariate. In particular, we can only compare units for clusters \( c \) such that there is variation in the treatment (\( \overline{W}_c \in \{(0, 1), (1, 0)\} \)), where \( \overline{W}_c \) be the pair of treatment values \((W_i, W_j)\) for the two units \( i \) and \( j \) in cluster \( c \) and there is no variation in the covariate
\((\mathbf{X}_c \in \{(0,0),(1,1)\})\), where \(\mathbf{X}_c\) be the pair of treatment values \((X_i, X_j)\) for the two units \(i\) and \(j\) in cluster \(c\). Denote the set of such clusters by 
\[
\mathcal{B} \equiv \left\{ c = 1, \ldots, C \mid \mathbf{W}_c \in \{(0,1),(1,0)\}, \mathbf{X}_c \in \{(0,0),(1,1)\}\right\},
\]
and consider the average treatment effect for these clusters:
\[
\tau_{\mathcal{B}} \equiv \mathbb{E}[\tau_{C_i}(X_i)|C_i \in \mathcal{B}]. \tag{2.15}
\]
We can identify this average using only within-cluster variation, in a setting with a fixed number of units per cluster. Because \(\mathcal{B}\) depends on the realizations of the treatment indicators, \(\tau_{\mathcal{B}}\) is a random quantity. We cannot identify more general effects, e.g., the overall average treatment effect, without additional assumptions.

Now consider weakening the cluster unconfoundedness condition from (2.14) to
\[
W_i \perp \perp \left( Y_i(0), Y_i(1) \right) \mid X_i, W_{C_i}, X_{C_i}. \tag{2.16}
\]
Thus, (2.16) weakens the cluster unconfoundedness in (2.14) to require only conditioning on \(W_{C_i}\) and \(X_{C_i}\), encapsulating a key feature of the fixed effect specification that \(W_{C_i}\) and \(X_{C_i}\) fully capture all relevant differences between clusters, and that \(WX_{C_i}\) is not needed for this. As we show below, (2.16) allows for some particular cross-cluster comparisons where within-cluster comparisons are not feasible. We leave the justification of this assumption to next section, and focus now on its implications. Restriction (2.16) implies that we can compare treated and control units as long as they have the same values of \(X_i\) and \(S_i \equiv (W_{C_i}, X_{C_i})\). By construction \(S_i\) takes 9 possible values, but only 3 of them correspond to clusters with variation in treatment status: \(S_i \in \{(\frac{1}{2},0), (\frac{1}{2},1), (\frac{1}{2},1)\}\). For other values of \(S_i\) we either have only control or only treated units and thus condition (2.16) is not useful (and neither is (2.13)).

The three remaining values of \(S_i\) lead to conceptually different comparisons. Values \(S_i \in \{ (\frac{1}{2},0), (\frac{1}{2},1) \} \) correspond to clusters where both units are identical in terms of \(X_i\) but have different treatment statuses. For these groups (2.16) justifies within-cluster comparisons which were already validated by the more restrictive assumption (2.13). This leaves on remaining value, \(S_i = (\frac{1}{2}, \frac{1}{2})\). If \(S_i = (\frac{1}{2}, \frac{1}{2})\) then by definition the two units in the cluster have different values of \(X_i\), and within-cluster comparisons have no causal interpretation: the pair of units in the same cluster are either \((W_i, X_i) = (0,0)\) and \((W_j, X_j) = (1,1)\), or the pair are \((W_i, X_i) = (1,0)\)
and \((W_j, X_j) = (0, 1)\). In both cases the pairs cannot directly be compared. In this case, 
(2.16) justifies between-cluster comparisons that are not justified solely based on (2.13). In particular, (2.16) justifies combining clusters where \(W_i = X_i\) with those where \(W_i = 1 - X_i\). It is useful to consider this comparison in more detail. Suppose that cluster \(c_1\) has units with \((W_i, X_i) \in \{(0, 0), (1, 1)\}\), and cluster \(c_2\) has units with \((W_i, X_i) \in \{(0, 1), (1, 0)\}\). We cannot directly compare a treated unit in cluster \(c_1\) with a control unit in cluster \(c_1\). But we can compare a treated unit in cluster \(c_1\) (with \((W_i, X_i) = (1, 1)\)) with a control unit in cluster \(c_2\) (with \((W_i, X_i) = (0, 1)\)) because they satisfy the two conditions that (i) they have the same value of \(X_i\) (namely \(X_i = 1\)), and (ii) they belong to clusters with the same value for \(\overline{W}_c\) and \(\overline{X}_c\) (namely \(\overline{W}_c = 1/2\) and \(\overline{X}_c = 1/2\)).

We show later that the linear additive fixed effect version uses this comparison, but in addition uses others that are not validated by (2.16) but instead rely intrinsically on additional assumptions.

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 Set Up and Estimands

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,

\[\tau \equiv E[Y_i(1) - Y_i(0)],\]

but this may be challenging, and we may need to settle for some other average of \(Y_i(1) - Y_i(0)\). This is a subtle issue, because we do not want to define the set we average over to be defined in terms of the outcomes. Formally, we define it formally in terms of groups, covariates and assignments below. 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, \(C_i \in \{1, \ldots, C\}\) is the group label for unit \(i\), and \(X_i \in \mathbb{X}\) is an exogenous covariate. Also define \(C_{ic} = 1_{C_i = c}\) 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 \(\bar{Z}_c \equiv \sum_{i:C_i = c} Z_i / N_c\) 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, \(\bar{U}_{C_i} = U_i\) for all \(i\).

In the settings we are interested in the number of strata or clusters in the sample 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 co-
variates 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.

To be more specific about the estimands we focus on, let $S : \mathbb{X} \times \{0, 1\} \mapsto S \subset \mathbb{R}^d$ be a generic function, with $\overline{S}_c$ the sample average of $S(X_i, W_i)$ over the units in cluster $c$. Now consider for a particular pair $(x, s)$ the expected treatment effect,

$$
\tau(x, s) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x, \overline{S}_{C_i} = s].
$$

This is a somewhat unusual object because it is partly defined in terms of the sample, but such definitions are common in the program evaluation literature. See for similar sample-defined average treatment effects the discussion in Imbens [2004]. To make the discussion more specific, suppose there are no covariates, and $\overline{S}_c = \overline{W}_c$, the fraction treated in cluster $c$. In that case $\tau(s)$ is the average treatment effect, averaged over all clusters with the fraction treated equal to $s$. Because we focus on the setting with a modest number of units observed per cluster, we cannot condition on the population value of this fraction, and we need to be more careful in defining the estimands.

To illustrate this, suppose that for all sampled clusters we observe only two units. Then $\overline{S}_c$ takes on only three values, 0, 1/2, 1. Then clearly we cannot estimate $\tau(0$ and $\tau(1)$ in the absence of restrictions on the relation between the assignment probabilities and the potential outcomes: the clusters with $\overline{S}_c$ may be quite different from the clusters with $\overline{S}_c = 1/2$. We therefore restrict the estimands we focus on further. Specifically we postulate the existence of a known set $A \subset \mathbb{X} \times S$ such that for all $(X_i, \overline{S}_{C_i}) \in A$ the probability of being treated is between $\eta$ and $1 - \eta$ for some $\eta > 0$. Then we focus on the averages of the $\tau(X_i, \overline{S}_{C_i})$ of the following class:

$$
\tau_A = \frac{1}{\sum_{i=1}^N \sum_{(X_i, \overline{S}_{C_i}) \in A} 1(X_i, \overline{S}_{C_i})} \sum_{i=1}^N 1(X_i, \overline{S}_{C_i}) \tau(1, X_i, \overline{S}_{C_i})
$$

(3.1)
In the example with two units per cluster and no covariates, the estimand would be equal to the average treatment effect for units in clusters with one treated and one control unit observed. This is similar in spirit to panel data settings where we often can only estimate parameters for units with changes in some of the covariates over time.

In the next two subsections we discuss the choices of \( S(x, w) \) and the set \( A \) that characterize the estimands.

### 3.2 Assumptions and Preliminary Results

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)**

\[
W_i \perp \perp \left( Y_i(0), Y_i(1) \right) \mid X_i, C_i. \tag{3.2}
\]

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:

\[
\left( Y_i(1), Y_i(0), X_i, W_i \right) \perp \perp C_i \mid \overline{U}_{C_i}. \tag{3.3}
\]

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 $U_{Ci}$ is measurable with respect to cluster indicator variable, a direct implication of the
previous pair of assumptions is:

$$W_i \perp \perp (Y_i(0), Y_i(1)) \mid X_i, U_{Ci},$$

(3.4)

Now we can also compare treated and control units in different clusters, as long as the clusters
have the same value for $U_{Ci}$.

For the first identification result we need some additional notation. For each cluster $c$ define
$P_c$ 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 $(W_i, X_i)$.

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

$$W_i \perp \perp (Y_i(0), Y_i(1)) \mid X_i, N_{Ci}, P_{Ci}$$

(3.5)

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

This result states that as long as units have the same characteristics, and they come from
clusters identical in terms of $P_{Ci}$, they are comparable. This is a balancing/propensity score
type result in the sense that subpopulation with the same value for $(X_i, P_{Ci})$ 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 $(X_i, P_c)$ is $K \times (2^K)^{N_c}/N_c!$.
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 $U_{Ci}$ distribution of $(X_i, W_i)$ belongs

\footnote{For the formal definition of this object including continuous $X_i$ see Appendix A.}
to an exponential family with a known sufficient statistic:

\[ f_{X_i,W_i|U_i}(x,w|u) = h(x,w) \exp\left\{ \eta^\top(u)S(x,w) + \eta_0(u) \right\}, \quad (3.6) \]

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, \( (X_i, W_i)|U_{C_i} \) but places no restrictions on the conditional distribution of the outcome variable, \( Y_i|X_i, W_i, U_{C_i} \). □

**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 \( \overline{S}_{C_i} \). □

**Proposition 2.** Suppose Assumptions 3.1–3.4 hold. Then

\[ W_i \perp \perp C_i \bigg| X_i, \overline{S}_{C_i}. \quad (3.7) \]

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

\[ W_i \perp \perp \left( Y_i(0), Y_i(1) \right) \bigg| X_i, \overline{S}_{C_i}. \quad (3.8) \]
Comment 4: Lemma 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 $P_{C_i}$ 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 Lemma 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 $P_{C_i}$ 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 Lemma 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 $P_{C_i}$, 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 assume that $S(\cdot)$ is known, fixed and there is a known region of the covariate space where we have overlap. In Section 3.4 we discuss selecting the set of sufficient statistics. In particular, recall the definition of the propensity score:

$$e(x, s) \equiv E[W_i|X_i = x, \overline{S}_{C_i} = s]$$  \hspace{1cm} (3.9)

We are making the following assumption:

**Assumption 3.5. (Known overlap)** We assume that there exists $\eta > 0$ and a nonempty known set $A$, such that for any $(x, s) \in A$ 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 $A$ is equal to the support of the covariate space. See Crump et al. [2009]. □

3.3 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:

\[
\begin{align*}
\mu(W_i, X_i, S_{C_i}) &\equiv \mathbb{E}[Y_i|W_i, X_i, S_{C_i}] \\
e(X_i, S_{C_i}) &\equiv \mathbb{E}[W_i|X_i, S_{C_i}] \\
\varepsilon_i(w) &\equiv Y_i(w) - \mu(w, X_i, S_{C_i})
\end{align*}
\] (3.10)

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

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 true and estimated share of observations with overlap:

\[
\begin{align*}
\pi(A) &\equiv \mathbb{E}[A_i] \\
\overline{A} &\equiv \frac{1}{N} \sum_{i=1}^{N} A_i
\end{align*}
\] (3.11)

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 $\hat{\mu}_i(\cdot)$...
and \( \hat{\mu}_i(\cdot) \):

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

(3.12)

We also restrict moments of the residuals:

**Assumption 3.7. (Moment conditions)**

\[
\begin{align*}
\mathbb{E}[\epsilon_i^2(k)|X_i, \overline{S}_{C_i}] < K \ a.s. \\
\mathbb{E}[\epsilon_i^4(k)] < \infty
\end{align*}
\]

(3.13)

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

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

(3.14)

**Theorem 1. (Consistency)** Suppose Assumptions 3.1–3.4 and Assumption 3.6 hold. Then:

\[
\hat{\tau}_{dr} \equiv \frac{1}{NA} \sum_{i=1}^{N} A_i \psi(Y_i, W_i, X_i, \overline{S}_{C_i}, \hat{\mu}(W_i, X_i, \overline{S}_{C_i}), \hat{e}(X_i, \overline{S}_{C_i})),
\]

(3.15)

satisfies \( \hat{\tau}_{dr} - \tilde{\tau}_A = o_p(1) \).

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

\[
\hat{\tau}_{dr} = \frac{1}{A} \sum_{c=1}^{C} \frac{N_c}{N} \rho(c, \hat{\mu}(\cdot), \hat{e}(\cdot)).
\]
Theorem 2. (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:

$$\sqrt{n}(\hat{\tau}_{dr} - \tilde{\tau}_A) \overset{d}{\to} \mathcal{N}(0, \mathbb{V}),$$

where $\mathbb{V} = \frac{\mathbb{E}[\xi_c^2]}{\pi^2(A)}$,

where $\xi_c$ is defined in the following way:

$$\xi_c \equiv \sum_{i \in c} A_i \left( \frac{W_i}{\hat{e}(X_i, \bar{S}_c)} - \frac{1 - W_i}{1 - \hat{e}(X_i, \bar{S}_c)} \right) (Y_i - \hat{\mu}(W_i, X_i, \bar{S}_c))$$

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

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

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

$$\hat{\mathbb{V}} := \frac{1}{A} C \sum_{c=1}^C \left( \hat{\xi}_c - \frac{1}{C} \sum_{c'=1}^C \hat{\xi}_{c'} \right)^2.$$ (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 3. (Variance consistency) Suppose the assumptions of Theorem 2 hold. Then the variance estimator is consistent:

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

3.4 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:

\[
\Pr(C_i = c | W_i, X_i, \overline{U}_1, \ldots, \overline{U}_C) = \frac{\exp(\eta_0(\overline{U}_c) + \eta^\top(\overline{U}_c) S(X_i, W_i))}{\sum_{c'=1}^C \exp(\eta_0(\overline{U}_{c'}) + \eta^\top(\overline{U}_{c'}) S(X_i, W_i))}.
\]

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

\[
W_i \perp \perp \left( Y_i(0), Y_i(1) \right) \mid X_i, \overline{S}_C.
\]

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’Haultfoeuille [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:

$$\overline{Z}_t := \frac{1}{C} \sum_{t:T_i = t} Y_i, \quad \overline{Z}_c := \frac{1}{T} \sum_{t:C_i = c} Y_i,$$

and the overall average

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

and the residual

$$\hat{Z}_i = Z_i - \overline{Z}_t - \overline{Z}_c + \overline{Z}$$
Let $\hat{\tau}_{fe}$ be the least squares estimator for the regression
\[ Y_i = \alpha_{Ti} + \beta_{Ci} + \tau W_i + X_i^\top \gamma + \varepsilon_i \] (4.1)

Compare this to the least squares regression
\[ Y_i = \tau W_i + X_i^\top \gamma + \delta \bar{W}_{iT} + \mu \bar{W}_{Ci} + \psi \bar{X}_{iT} + \varphi \bar{X}_{Ci} + \varepsilon_i \]

The two least squares estimators for $\tau$ 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 $U_{Ci}$ 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:

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

Marginal distribution of this object is given by the following expression:
\[ f_{\mathcal{D}_c}(x_1, w_1, \ldots, x_{N_c}, w_{N_c}) = \sum_{k=1}^{p} \prod_{j=1}^{N_c} f(x_j, w_j | k) \pi_k \] (4.3)
This implies that the conditional distribution of $U_c$ given $D_c$ has the following form:

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

(4.4)

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

$$W_i \perp \perp (Y_i(0), Y_i(1)) \mid X_i, S(D_c_i)$$

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 $S(D_c)$ 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 $U_{C_i}$ to each observation (in our case, cluster). The usual way of doing this is to assign $U_{C_i}$ 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 $U_{C_i}$. With small $N_c$ this is not going to happen, and the distinction is essential.
4.4 Empirical Illustration
5 Related Literature

Our identification results are related to those of Altonji and Matzkin [2005] (AM). Indeed, one way of interpreting their assumptions is to say that there exist a statistic $S_i$ that depends on individual characteristics as well as group level variables, and which proxies for an unobserved group-level characteristic $U_i$ such that

$$W_i \perp \perp (Y_i(0), Y_i(1)) \mid S_i$$

holds. AM discuss that a candidate for such statistic is an exchangeable function of the data, but emphasise that exchangeability alone is not sufficient for identification. One can interpret our identification results as a foundation for AM, because they explicitly show how $S_i$ can be constructed under additional assumptions. Also, because we work in a more specific setting with binary treatment and exogenous covariates we can get a precise characterisation of the identified parameter.

Our approach is also related to random effects models studied in panel data literature (see Arellano and Bonhomme [2011] for a recent review). For example, Arellano and Bonhomme [2016] explicitly model the conditional distributions of unobservables given the individual characteristics and achieve identification using results from nonparametric nonlinear deconvolution literature (Hu and Schennach [2008]). There are two important differences between our approach and theirs. First, we do not put restriction on the dimension of the unobserved group-level characteristics, while such restriction are key in Arellano and Bonhomme [2016]. Second, we do not use outcome data to deal with unobserved heterogeneity. In this sense we follow a design-based tradition used in causal inference literature (Rosenbaum [2002], Imbens and Rubin [2015], Abadie et al. [2020]).
6 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 and Matias D Cattaneo. Econometric methods for program evaluation. Annual Review of Economics, 10:465–503, 2018.

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.

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. Econometrica, 88(1):265–296, 2020.

Elizabeth S Allman, Catherine Matias, and John A Rhodes. 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.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist’s companion*. Princeton University Press, 2008.

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

Manuel Arellano and Stéphane Bonhomme. Nonlinear panel data analysis. 2011.

Manuel Arellano and Stéphane Bonhomme. Nonlinear panel data estimation via quantile regressions. *The Econometrics Journal*, 19(3):C61–C94, 2016.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2): 608–650, 2014.

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

Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. 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.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, pages 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.

Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

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*, 63(2):467–490, 2019.

Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, pages 1–29, 2004.

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

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

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.

M Hashem Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012, 2006.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.
Paul R Rosenbaum. Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574, 1984.

Paul R Rosenbaum. *Observational Studies*. Springer, 2002.

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.

J.M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press. MIT Press, 2010. ISBN 9780262232586.
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 $\mathbb{P}_c = \{(X(1), W(1)), \ldots, (X(c), W(c))\} \in (\mathcal{X} \times \{0, 1\})^c$. 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 $\sigma$-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:

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

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

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

Equality between the first and the last expression implies the independence result. \qed
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:

\[ U_i \perp \{W_j, X_j\}_{j=1}^N | \mathbb{P}_{C_i} \]  \hspace{1cm} (1.4)

**Proof.** The proof follows from the following equalities:

\[ \mathbb{E}[\{U_i \in A\} | \{W_j, X_j\}_{j=1}^N \in B] | \mathbb{P}_{C_i} = \]
\[ \mathbb{E}[\{U_i \in A\} | \{W_j, X_j\}_{j=1}^N \in B, \{W_j, X_j\}_{j=1}^N, \{C_j = C_i\}_{j=1}^N | \mathbb{P}_{C_i} = \]
\[ \mathbb{E}[\{W_j, X_j\}_{j=1}^N \in B, \{U_i \in A\} | \mathbb{P}_{C_i}, \{W_j, X_j\}_{j=1}^N, \{C_j = C_i\}_{j=1}^N | \mathbb{P}_{C_i} = \]
\[ \mathbb{E}[\{W_j, X_j\}_{j=1}^N \in B, \{U_i \in A\} | \mathbb{P}_{C_i}, \{W_j, X_j\}_{j=1}^N, \{C_j = C_i\}_{j=1}^N | \mathbb{P}_{C_i} = \]
\[ \mathbb{E}[\{W_j, X_j\}_{j=1}^N \in B | \mathbb{P}_{C_i}] \mathbb{E}[\{U_i \in A\} | \mathbb{P}_{C_i}] \]  \hspace{1cm} (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:

\[ \mathbb{E}[(Y_i(1), Y_i(0)) \in A] | W_i, X_i, \mathbb{P}_{C_i} = \]
\[ \mathbb{E}[\mathbb{E}[(Y_i(1), Y_i(0)) \in A] | W_i, X_i, \mathbb{P}_{C_i}, U_i | W_i, X_i, \mathbb{P}_{C_i} = \]
\[ \mathbb{E}[\mathbb{E}[(Y_i(1), Y_i(0)) \in A] | W_i, X_i, U_i | W_i, X_i, \mathbb{P}_{C_i}] \]  \hspace{1cm} (1.6)

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

\[ \mathbb{E}[\mathbb{E}[(Y_i(1), Y_i(0)) \in A] | W_i, X_i, U_i | W_i, X_i, \mathbb{P}_{C_i}] = \]
\[ \mathbb{E}[\mathbb{E}[(Y_i(1), Y_i(0)) \in A] | X_i, U_i | W_i, X_i, \mathbb{P}_{C_i} = \]
\[ \mathbb{E}[\mathbb{E}[(Y_i(1), Y_i(0)) \in A] | X_i, U_i | X_i, \mathbb{P}_{C_i}] = \]
\[ \mathbb{E}[\mathbb{E}[(Y_i(1), Y_i(0)) \in A] | X_i, \mathbb{P}_{C_i}, U_i | X_i, \mathbb{P}_{C_i} = \mathbb{E}[\{Y_i(1), Y_i(0)\} \in A] | X_i, \mathbb{P}_{C_i}] \]  \hspace{1cm} (1.7)

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:
\[
E\left[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, \mathbb{P}_{C_i}\right] = E\left[\{(Y_i(1), Y_i(0)) \in A\}|X_i, \mathbb{P}_{C_i}\right]
\] (1.8)
which proves the conditional independence. □

**Corollary A1.** **(Exclusion in exponential families)** Under the assumptions of Lemma A1 the following is true:
\[
(Y_i(1), Y_i(0)) \perp \perp \{(S_{C_j}, X_j, W_j)\}_{j=1}^N |X_i, W_i, U_i
\] (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:
\[
U_i \perp \perp \{W_j, X_j\}_{j=1}^N |\overline{S}_{C_i}
\] (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 ??:** 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 $E[|f(Y(k))|] < \infty$ the following is true:
\[
E[f(Y_i)|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \\
\{W_i = 0\} E[f(Y_i(0))|X_i, \overline{S}_{C_i}] + \{W_i = 1\} E[f(Y_i(1))|X_i, \overline{S}_{C_i}]
\] (1.11)

**Proof.** The proof follows from the following equalities:
\[
E[f(Y_i)|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = E[E[f(Y_i)|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N, U_i]|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \\
E[E[f(Y_i)|W_i, X_i, U_i]|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = E[f(Y_i)|W_i, X_i, \overline{S}_{C_i}] = \\
\{W_i = 0\} E[f(Y_i(0))|X_i, \overline{S}_{C_i}] + \{W_i = 1\} E[f(Y_i(1))|X_i, \overline{S}_{C_i}]
\] (1.12)
where the third equality follows from Corollary A1, the fourth from Lemma A3 and the final one from Proposition ??.

34
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 \( \mathbb{G}_n(X_c) \equiv \sqrt{n} (\mathbb{P}_n(X_c) - \mathbb{E}[X_c]) \). Define \( B_i = (X_i, \overline{S}_{C_i}) \) and \( D_i \equiv (W_i, B_i) \).

We start with a reminder on notation:

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

In order to prove Theorem 1 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 1 follows as a special case. After that we prove Theorem 2 and Proposition 3. 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:

\[
\begin{align*}
\bar{\mu}(W_i, X_i, \overline{S}_{C_i}) & = \mu(W_i, X_i, \overline{S}_{C_i}) \\
\bar{\varepsilon}(X_i, \overline{S}_{C_i}) & = e(X_i, \overline{S}_{C_i})
\end{align*}
\]

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

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

where \( \tau(B_i) := \mu(1, B_i) - \mu(0, B_i) \).
Proof. By construction we have the following:

\[
\begin{align*}
\mathbb{E}[\rho(c, \hat{\mu}, \hat{\epsilon})] &= \mathbb{E} \left[ \sum_{i \in c} \frac{1}{|c|} \{A_i\} \left( \hat{\mu}(1, B_i) - \hat{\mu}(0, B_i) + \left( \frac{W_i}{\hat{\epsilon}(B_i)} - \frac{1 - W_i}{1 - \hat{\epsilon}(B_i)} \right) (Y_i - \hat{\mu}(D_i)) \right) \right] \\
&= \mathbb{E} \left[ \sum_{i \in c} \frac{1}{|c|} \{A_i\} (\hat{\mu}(1, B_i) - \hat{\mu}(0, B_i)) \right] + \sum_{i \in c} \frac{1}{|c|} \mathbb{E} \left[ \{A_i\} \left( \frac{W_i}{\hat{\epsilon}(B_i)} - \frac{1 - W_i}{1 - \hat{\epsilon}(B_i)} \right) (Y_i - \hat{\mu}(D_i)) \right]
\end{align*}
\]

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

\[
\begin{align*}
\mathbb{E} \left[ \{A_i\} \left( \frac{W_i}{\hat{\epsilon}(B_i)} - \frac{1 - W_i}{1 - \hat{\epsilon}(B_i)} \right) (Y_i - \hat{\mu}(D_i)) \right] &= \\
\mathbb{E} \left[ \mathbb{E} \left[ \{A_i\} \left( \frac{W_i}{\hat{\epsilon}(B_i)} - \frac{1 - W_i}{1 - \hat{\epsilon}(B_i)} \right) (Y_i - \hat{\mu}(D_i)) | B_i \right] \right] &= \\
\mathbb{E} \left[ \{A_i\} \left( \frac{e(B_i) (\mu(1, B_i) - \hat{\mu}(1, B_i))}{\hat{\epsilon}(B_i)} - \frac{(1 - e(B_i) (\mu(0, B_i) - \hat{\mu}(0, B_i)))}{1 - \hat{\epsilon}(B_i)} \right) \right]
\end{align*}
\]

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

**Proposition B1. (Consistency with Wrong Specifications)** Assume that the following conditions hold for \((\hat{\epsilon}, \hat{\mu})\):

\[
\begin{align*}
\mathbb{P}_n \left( \sum_{i \in c} \frac{1}{|c|} \{A_i\} \left( \hat{\mu}(1, B_i) - \hat{\mu}(1, B_i) \right)^2 \right) &= o_p(1) \\
\mathbb{P}_n \left( \sum_{i \in c} \frac{1}{|c|} \{A_i\} \left( \hat{\epsilon}(B_i) - \hat{\epsilon}(B_i) \right)^2 \right) &= o_p(1) \\
\eta < \hat{\epsilon}(B_i) < 1 - \eta \text{ a.s.} \\
\eta < \hat{\epsilon}(B_i) < 1 - \eta \text{ a.s.} \\
\mathbb{E}[\hat{\epsilon}_i^2(k)] < \infty
\end{align*}
\]

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

\[
\mathbb{P}_n \rho(c, \hat{\mu}, \hat{\epsilon}) = \mathbb{P}_n \rho(c, \hat{\mu}, \hat{\epsilon}) + o_p(1) = \mathbb{E}[\rho(c, \hat{\mu}, \hat{\epsilon})] + o_p(1)
\]
Proof. To prove the consistency result we need to separate the functional into two parts:

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

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:

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

We can decompose this expression into three parts:

\[
\rho_1(c, \hat{\mu}, \hat{e}) = \sum_{i \in c} \frac{1}{|c|} \left\{ A_i \right\} \left( \hat{\mu}(1, B_i) + \frac{W_i}{\hat{e}(B_i)} (Y_i - \hat{\mu}(1, B_i)) \right) + \sum_{i \in c} \frac{1}{|c|} \left\{ A_i \right\} \left( \left( \hat{\mu}(1, B_i) - \tilde{\mu}(1, B_i) \right) \left( 1 - \frac{W_i}{\hat{e}(B_i)} \right) \right) + \sum_{i \in c} \frac{1}{|c|} \left\{ A_i \right\} (Y_i - \tilde{\mu}(1, B_i)) W_i \left( \frac{1}{\hat{e}(B_i)} - \frac{1}{\hat{\mu}(B_i)} \right) = \rho_1(c, \tilde{\mu}, \tilde{e}) + R_{1c} + R_{2c} \quad (2.10)
\]

The result will follow once we prove two approximations:

\[
\begin{cases}
\mathbb{P}_n R_{1c} = o_p(1) \\
\mathbb{P}_n R_{2c} = o_p(1)
\end{cases} \quad (2.11)
\]

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

\[
|\mathbb{P}_n R_{2c}| \leq \mathbb{P}_n |R_{2c}| \leq \mathbb{P}_n \sum_{i \in c} \frac{1}{|c|} \left\{ A_i \right\} |\hat{\epsilon}_i(1)| \left( \frac{\{A_i\} W_i}{\hat{e}(B_i) \hat{\epsilon}(B_i)} \right) \{A_i\} \left| \hat{\epsilon}(B_i) - \hat{\epsilon}(B_i) \right| \right| \leq \\
\max_i \left( \frac{\{A_i\} W_i}{\hat{e}(B_i) \hat{\epsilon}(B_i)} \right) \sqrt{\mathbb{P}_n \sum_{i \in c} \frac{1}{|c|} \{A_i\} \hat{\epsilon}_i^2(1)} \left( \sqrt{\mathbb{P}_n \sum_{i \in c} \frac{1}{|c|} \{A_i\} \left( \hat{\epsilon}(B_i) - \hat{\epsilon}(B_i) \right)^2 = O_p(1) \sqrt{O_p(1)} \sqrt{o_p(1)} = o_p(1) \quad (2.12)
\right.
\]
For the first term we have the following:

\[ R_{1c} = \sum_{i \in c} \frac{1}{|c|} \left\{ A_i \right\} \left( \left( \hat{\mu}(1, B_i) - \bar{\mu}(1, B_i) \right) \left( 1 - \frac{W_i}{\bar{e}(B_i)} \right) \right) = \]

\[ \sum_{i \in c} \frac{1}{|c|} \left\{ A_i \right\} \left( \left( \hat{\mu}(1, B_i) - \bar{\mu}(1, B_i) \right) \left( 1 - \frac{W_i}{\bar{e}(B_i)} \right) \right) + \]

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

The first part can be bounded in the following way:

\[ |\mathbb{P}_n R_{11c}| \leq \mathbb{P}_n |R_{11c}| \leq \]

\[ \max_i \left| \left\{ A_i \right\} \left( W_i - \bar{e}(B_i) \right) \right| \times \mathbb{P}_n \left( \sum_{i \in c} \frac{1}{|c|} \left\{ A_i \right\} \left( \hat{\mu}(1, B_i) - \bar{\mu}(1, B_i) \right)^2 \right) = \]

\[ O_p(1) \times o_p(1) = o_p(1) \quad (2.14) \]

The second part can be bounded in the following way:

\[ |\mathbb{P}_n R_{12c}| \leq \mathbb{P}_n |R_{12c}| \leq \max_i \left( \frac{\left\{ A_i \right\} W_i}{\hat{e}(B_i)\hat{e}(B_i)} \right) \times \]

\[ \mathbb{P}_n \left( \sum_{i \in c} \frac{1}{|c|} \left\{ A_i \right\} \left( \hat{\mu}(1, B_i) - \bar{\mu}(1, B_i) \right)^2 \right) \sqrt{ \mathbb{P}_n \left( \sum_{i \in c} \frac{1}{|c|} \left\{ A_i \right\} \left( \hat{\mu}(1, B_i) - \bar{\mu}(1, B_i) \right)^2 \right) } = \]

\[ O_p(1) \times o_p(1) o_p(1) = o_p(1) \quad (2.15) \]

Combining all the results together we have the proof.  

\[ \square \]

**Proof of Theorem 1**: Observe that \( \hat{e} \) and \( \hat{\mu} \) satisfy the assumptions of Proposition B1 with \( \bar{\mu} \) and \( \bar{e} \) equal to \( m \) and \( e \). As a result, combining Proposition B1 and Lemma B4 we get the following:

\[ \frac{1}{\pi(A)} \mathbb{P}_n \rho(c, \hat{\mu}, \hat{e}) = \frac{1}{\pi(A)} \left( \mathbb{E}[\rho(c, \mu, e)] + o_p(1) \right) = \]

\[ \left( \frac{1}{\pi(A)} + o_p(1) \right) \left( \mathbb{E}[\rho(c, \mu, e)] + o_p(1) \right) = \frac{1}{\pi(A)} \mathbb{E}[\rho(c, \mu, e)] + o_p(1) = \]

\[ \frac{1}{\pi(A)} \mathbb{E}[\left\{ A_i \right\} \tau(X_i, \mathcal{F}_{C_i})] + o_p(1) \quad (2.16) \]
Proof of Theorem 2: The start of the argument is the same as in proof for the consistency result. We decompose the empirical version of $\rho(c, \hat{m}, \hat{e})$:

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

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:

$$
\mathbb{P}_n R_{12c} \leq \mathbb{P}_n |R_{12c}| \leq \max_i \left( \frac{\{A_i\} W_i}{e(B_i) \hat{e}(B_i)} \right) \times \\
\sqrt{\mathbb{P}_n \left( \sum_{i \in c} \frac{1}{|c|} \{A_i\} (\hat{\mu}(1, B_i) - \mu(1, B_i))^2 \right)} \sqrt{\mathbb{P}_n \left( \sum_{i \in c} \frac{1}{|c|} \{A_i\} (\hat{e}(B_i) - e(B_i))^2 \right)} = \\
O_p(1) \times o_p \left( \frac{1}{\sqrt{n}} \right) = o_p \left( \frac{1}{\sqrt{n}} \right) \quad (2.19)
$$
For $R_{11c}$ we use the following argument:

$$
E \left[ (P_n R_{11c})^2 \right] = E \left[ \left( \sum_{l \in L} \frac{1}{n} \sum_{i \in c} \frac{1}{|c|} (\hat{\mu}_{-l(c)}(1, B_i) - \mu(1, B_i)) \left( 1 - \frac{W_i}{\epsilon(B_i)} \right) \right)^2 \right] \\
|L| \sum_{l \in L} \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i \in c} \frac{1}{|c|} (\hat{\mu}_{-l(c)}(1, B_i) - \mu(1, B_i)) \left( 1 - \frac{W_i}{\epsilon(B_i)} \right) \right)^2 \right] = \leq \leq \leq \\
|L| \sum_{l \in L} \frac{1}{n} \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i \in c} \frac{1}{|c|} (\hat{\mu}_{-l(c)}(1, B_i) - \mu(1, B_i)) \left( 1 - \frac{W_i}{\epsilon(B_i)} \right) \right)^2 \right] \\
|L| \sum_{l \in L} \frac{1}{n} \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i \in c} \frac{1}{|c|} (\hat{\mu}_{-l(c)}(1, B_i) - \mu(1, B_i)) \left( 1 - \frac{W_i}{\epsilon(B_i)} \right) \right)^2 \right] \leq \\
\frac{K}{n} \mathbb{E} \left[ P_n \left( \sum_{i \in c} \frac{1}{|c|} (\hat{\mu}_{-l(c)}(1, B_i) - \mu(1, B_i))^2 \right) \right] \leq (2.20) \\
\mathbb{E} \left[ \left| P_n R_{11c} \right| \right] \leq \sqrt{E \left[ (P_n R_{11c})^2 \right]} \leq \frac{K}{\sqrt{n}} \mathbb{E} \left[ P_n \left( \sum_{i \in c} \frac{1}{|c|} (\hat{\mu}_{-l(c)}(1, B_i) - \mu(1, B_i))^2 \right) \right] = o \left( \frac{1}{\sqrt{n}} \right) (2.21) \\
\text{This implies (by Markov’s inequality) that } P_n R_{11c} = o \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{\epsilon}(B_i)} \right) \{A_i\} (e(B_i) - \hat{e}(B_i))^2 \leq \\
\max_i \left( \frac{\{A_i\} \mathbb{E}[\varepsilon_i^2 | D_i] W_i}{e^2(B_i) \hat{\epsilon}(B_i)} \right) \sum_{i \in c} \frac{1}{|c|} \{A_i\} (e(B_i) - \hat{e}(B_i))^2 (2.22) \\
\text{We also have the following:} \\
\mathbb{E}[R_{2c}|\{D_i\}_{i=1}^N] = 0 (2.23)
Similarly to all other proofs we can divide

Proof of Proposition 3:

Using these two things we get the following:

\[ \left( \begin{array}{c} \mathbb{E}\left[ \left( \sum_{i=1}^{N} A_i \right) \mathbb{E}\left[ \varepsilon_i^2 | D_i \right] W_i \right] \right) \times \right. \\
\frac{1}{n} \sum_{i \in \mathcal{C}} \left( \begin{array}{c} \frac{1}{c} \right) \left( \begin{array}{c} A_i \right) \left( e(B_i) - \hat{e}(B_i) \right)^2 \leq K \times o_p \left( \frac{1}{n} \right) = o_p \left( \frac{1}{n} \right) \] (2.24)

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

**Proof of Proposition 3:** Similarly to all other proofs we can divide \( \xi_c \) into two parts \( \xi_{1c} \) and \( \xi_{0c} \). We will analyze \( \xi_{1c} \), analysis for \( \xi_{0c} \) is the same. We have the following decomposition:

\[ \hat{\xi}_{1c} - \xi_{1c} = \sum_{i \in \mathcal{C}} \frac{1}{|c|} \{ A_i \} \left( (\mu(1, B_i) - \hat{\mu}(1, B_i)) \frac{W_i}{\hat{e}(B_i)} \right) + \]

\[ \sum_{i \in \mathcal{C}} \frac{1}{|c|} \{ A_i \} (Y_i - \mu(1, B_i)) W_i \left( \frac{1}{\hat{e}(B_i)} - \frac{1}{e(B_i)} \right) = R_{11c} + R_{12c} \] (2.25)

For the first term we have the following bound:

\[ \mathbb{P}_n R_{11c} \leq \mathbb{P}_n \left( \frac{1}{c} \sum_{i \in \mathcal{C}} \{ A_i \} \left( (\mu(1, B_i) - \hat{\mu}(1, B_i))^2 \frac{W_i}{\hat{e}^2(B_i)} \right) \right) \leq \]

\[ \left( \max_{i} \frac{\{ A_i \} W_i}{\hat{e}^2(B_i)} \right) \times \mathbb{P}_n \left( \frac{1}{c} \sum_{i \in \mathcal{C}} \{ A_i \} (\mu(1, B_i) - \hat{\mu}(1, B_i))^2 \right) = O_p(1) o_p(1) = o_p(1) \] (2.26)

For the second term we have the following bound:

\[ \mathbb{P}_n R_{12c} \leq \mathbb{P}_n \left( \frac{1}{c} \sum_{i \in \mathcal{C}} \{ A_i \} \{ W_i \} \varepsilon_i^2(1) \frac{(\hat{e}(B_i) - e(B_i))^2}{\hat{e}^2(B_i)e^2(B_i)} \right) \leq \]

\[ \sqrt{\left( \mathbb{P}_n \left( \frac{1}{c} \sum_{i \in \mathcal{C}} \{ A_i \} \frac{(\hat{e}(B_i) - e(B_i))^4}{\hat{e}^4(B_i)e^4(B_i)} \right) \right) \left( \mathbb{P}_n \left( \frac{1}{c} \sum_{i \in \mathcal{C}} \{ A_i \} \{ W_i \} \varepsilon_i^4(1) \right) \right)} \leq \]

\[ K \sqrt{\left( \mathbb{P}_n \left( \frac{1}{c} \sum_{i \in \mathcal{C}} \{ A_i \} (\hat{e}(B_i) - e(B_i))^2 \right) \right) \left( \mathbb{P}_n \left( \frac{1}{c} \sum_{i \in \mathcal{C}} \{ A_i \} \{ W_i \} \varepsilon_i^4(1) \right) \right)} = \]

\[ o_p(1) O_p(1) = o_p(1) \] (2.27)
Putting these results together we have the following:

\[
\mathbb{P}_n(\hat{\xi}_1 c + \hat{\xi}_2 c)^2 - \mathbb{P}_n(\xi_1 c + \xi_2 c)^2 = \mathbb{P}_n(\xi_1 + \xi_2 + R_{11c} + R_{12c} + R_{01c} + R_{02c})^2 - \mathbb{P}_n(\xi_1 + \xi_2 c)^2 = \\
\mathbb{P}_n(\xi_1 + \xi_2 c)(R_{11c} + R_{12c} + R_{01c} + R_{02c}) + \mathbb{P}_n(\xi_1 + \xi_2 c + R_{11c} + R_{12c} + R_{01c} + R_{02c})^2 \leq \\
\sqrt{\mathbb{P}_n(\xi_1 + \xi_2 c)^2 4\mathbb{P}_n(R_{11c}^2 + R_{12c}^2 + R_{01c}^2 + R_{02c}^2) + \mathbb{P}_n(R_{11c}^2 + R_{12c}^2 + R_{01c}^2 + R_{02c}^2)} = \\
\sqrt{O_p(1) + o_p(1)} = o_p(1) \quad (2.28)
\]

This argument also implies that \(\mathbb{P}_n(\hat{\xi}_1 c) = \mathbb{P}_n(\xi_1 c) = o_p(1)\) and thus we have the final result:

\[
\frac{1}{\pi^2(A)} \left( \mathbb{P}_n(\hat{\xi}_1 c + \hat{\xi}_2 c)^2 - \left( \mathbb{P}_n(\hat{\xi}_1 c + \hat{\xi}_2 c) \right)^2 \right) - \frac{1}{\pi^2(A)} \mathbb{P}_n(\xi_1 + \xi_2 c)^2 = \\
\frac{1}{\pi^2(A)} \left( \mathbb{P}_n(\hat{\xi}_1 c + \hat{\xi}_2 c)^2 - \mathbb{P}_n(\xi_1 + \xi_2 c)^2 \right) + \left( \frac{1}{\pi^2(A)} - \frac{1}{\pi^2(A)} \right) \mathbb{P}_n(\xi_1 + \xi_2 c)^2 + O_p(1) = \\
O_p(1) + o_p(1) + O_p(1) + o_p(1) = o_p(1) \quad (2.29)
\]