Revisiting Randomization with the Cube Method

Laurent Davezies[*] Guillaume Hollard[†] Pedro Vergara Merino[‡]

Abstract

Popular randomization methods have clear limitations: stratification and pairwise matching do not allow balancing on several continuous variables; re-randomization makes computing valid confidence intervals challenging; and simple randomization may lead to substantial imbalances in covariates.

In this paper, we demonstrate how the Cube method (Deville and Tillé, 2004) can be used to overcome all these limitations. Indeed, the Cube method allows for the selection of perfectly balanced samples on any covariate (continuous or not), ensuring that balance tests are always passed. Furthermore, the Cube method permits the definition of unambiguous confidence intervals and leads to substantial gains in precision when covariates correlate with potential outcomes. Lastly, the Cube method allows for freely choosing assignment probabilities, which can vary across subgroups (e.g. to anticipate sample attrition).

Keywords: Causal inference, covariate balance, experimental design, treatment effects.

JEL Codes: C13, C21

Preliminary draft. Please do not cite or circulate

[*] CREST - ENSAE - IPParis, laurent.davezies@ensae.fr

[†] CREST - Ecole Polytechnique - IPParis, guillaume.hollard@polytechnique.edu

[‡] CREST - ENSAE -IPParis, pedro.vergaramerino@ensae.fr

We would like to thank Riccardo d'Adamo, Russell Davidson, Pauline Rossi, Felix Schleef, Sami Stouli, Xinran Li, Xavier d'Haultfoeuille and many other colleagues in workshops and conferences whose comments and suggestions helped improve this paper. In particular, we would like to thank everybody at the Bristol Econometric Study Group.

1 Introduction

Choosing a particular randomization method to create control and treatment groups is not without consequences. The principal purpose of randomization is to avoid selection bias by balancing potential outcomes across groups. But, in addition, the choice of a randomization method has an impact on the precision of the estimates, the ability to balance covariates, the possibility to use heterogeneous assignment probabilities, draw valid inferences, and computational complexity (see Athey and Imbens, 2017).

We point to three main limitations of popular randomization methods. First, naive randomization (in which the assignment probability of a unit to a given treatment arm does not depend on the assignment of others) only ensures that balancing will occur *on average*. As a result, an empiricist who checks for the balance for several covariates after naive randomization will confront high probabilities of obtaining statistically significant differences across groups. Unbalanced covariates have consequences since there is evidence of publication bias and p-hacking when observing the empirical distribution of p-values of balance tests in RCTs. For instance, Snyder and Zhuo (2018) estimate that the publication process eliminates 46% of p-values under 0.15. To avoid frequent unbalanced covariates, researchers may use methods that use information about covariates that are available before randomization takes place, like stratification, pairwise matching, and re-randomization. However, not all designs allow balancing on *all* observable covariates. A second limitation derives from these methods not using all available information, thus waiving precision gains. Last, a third limit appears when using methods like re-randomization. Indeed, the inference for treatment estimates might be challenging or computationally demanding, casting doubts on the determination of confidence intervals (Imbens, 2011; Li et al., 2018).

We here introduce the Cube method and show that it is not concerned by any of the limitations above-described. The Cube method was developed for survey sampling purposes by Deville and Tillé (2004). We here show how empiricists can use the Cube method in the context of RCTs. In particular, the Cube method allows researchers to obtain almost-perfectly-balanced covariates across treatment arms, discrete or continuous. The Cube method also increases the number of variables the empiricist can balance on. We derive exact expressions for the asymptotic variances of the sample average treatment effect (SATE) and population average treatment effect (PATE) estimators. Formal derivation also allows us to highlight the precision gains obtained. In particular, one achieves substantial precision gains when observable covariates (i.e., “baseline covariates”) correlate with the outcome of interest, a frequent situation in fieldwork. Interestingly, the Cube method allows also for freely choosing assignment probabilities, which can vary across subgroups (e.g., to anticipate sample attrition). Last, we illustrate the interest in the Cube method using data from existing RCTs.

We first review existing methods and provide a detailed account of association limitations. Section 3 introduces the potential outcome framework and covariate balancing. Section 4 presents the Cube algorithm and gives insights on how to apply it to RCTs. Section 5 gives the balancing properties of the Cube method and provides novel asymptotic expressions for the variance of average treatment effect estimators. We then specify two ways of performing inference based on asymptotic normality and the randomization mechanism. Finally, Section 6 simulates experiments using two datasets and provides insights into the gains from the Cube method.

2 State of the art and literature review

2.1 Naive randomization and balance tests

To create comparable treatment and control groups, empiricists sometimes use “naive randomization” methods, meaning they do not use any information previously available to design the assignment mechanism. The most common naive method is complete randomization. In complete randomized experiments, the empiricist chooses a fixed size for each treatment arm and randomly draws the exact number of units for each one. Moreover, every allocation probabilities to a treatment arm are identical across individuals. The analogous method to complete randomization in survey sampling is simple random sampling without replacement. Other naive methods are Bernoulli sampling and Poisson sampling. In Bernoulli sampling, the empiricist draws independently every unit with the same probability, which means that the average size of a treatment arm is known. The final allocation, however, generally deviates from this average. Poisson sampling maintains independence when drawing units but allows for different inclusion probabilities or propensity scores.

Naive randomization generates treatment and control groups that are, on average, identical. However, an empiricist allocates units only once. She can thus get groups that are different from each other. To check if randomization successfully created different groups, empiricists perform balance tests on observable covariates. Most common balance tests compare the mean of baseline covariates between control and treatment groups. The idea behind these tests is: if the control and treatment groups are similar across covariates, they are likely more comparable across potential outcomes. However, with naive randomization, *p*-values converge to a uniform distribution: there exists a 10% chance of getting significant differences at the 90%-confidence level. There always exists a high probability of obtaining statistically significant differences. These differences incite, on the one hand, the author to leave out the concerned covariates or to quit the experiment altogether. Statistically significant differences encourage, on the other hand, the editor not to publish the article in question. There is evidence of publication bias and *p*-hacking when observing the empirical distribution of *p*-values related to balance tests in RCTs. For instance, Snyder and Zhuo (2018) estimate that the publication process eliminates 46% of the p-values under 0.15. They also evidence an overrepresentation of *p*-values over 0.9. Using balance tests to check the comparability of treatment arms contributes thus to publication bias. Statisticians also argue that doing these tests can lead to poorer estimation of treatment effects (Mutz et al., 2019; Bruhn and McKenzie, 2009). Moreover, imbalances that concern covariates highly correlated to potential outcomes, such as baseline outcomes, produce less precise estimators for treatment effects.

2.2 Covariate-adaptative randomization

An alternative to checking the presence of imbalances after randomizing consists of creating a design ensuring the balance between treatment arms. We refer to the ensemble of such methods as "covariate-adaptive randomization." The most practiced covariate-adaptive methods are stratified randomization, pairwise randomization, and re-randomization. These mechanisms improve the balance between treatment and control groups but have some limitations.

2.2.1 Stratification

Stratification has a long tradition in RCTs (Fisher, 1935; Higgins et al., 2016). This method consists of using one or several baseline variables to create blocks or strata and then using complete randomization inside each stratum. A common practice in experiments is to block on gender, meaning that randomization is performed independently amongst male and female units, generating the same proportion of men and women in each treatment arm. When using dummy variables to define the strata, stratified or blocked randomization allows almost perfect balancing of the variables used to create them. We say "almost-perfectly balanced" and not "perfectly balanced" because approximations are sometimes necessary (e.g., from a population of 51 women and 49 men, it is impossible to perfectly balance gender if we want 50 units treated and 50 units in the control group).Athey and Imbens (2017) recommend balancing on small strata since this method generates substantial precision gains. However, stratification does not come without any limitations. Facing continuous covariates, such as income or grades, makes it impossible to stratify without discretizing the variable in question. This problem arises, as well, when using discrete variables with many possible values, such as age. The empiricist will obtain an almost-perfect balance for the discretized variables but not for the continuous ones. Moreover, as the number of balancing variables increases, the quantity of strata rises exponentially. With ten dummy variables, for instance, we are automatically creating 1024 possible strata, thus requiring a relatively large sample.

2.2.2 Pairing

Empiricists have also used pairwise designs for decades (Ball et al., 1973; Greevy et al., 2004; Imai et al., 2009; Basse et al., 2019). However, results about inference and asymptotic properties are very recent (Bai, 2022; Bai et al., 2022). This method consists of creating pairs of units based on one or several discrete variables or an aggregate of covariates if one of them is continuous. After this step, the empiricist randomly assigns one of the units to treatment and the other to the control group. If she uses only discrete variables to match, the empiricist will get perfectly-balanced groups across those covariates. As for stratification, however, the number of discrete variables the empiricist can use is $o(\ln(n))$. When matching on an aggregate of variables, such as the Mahalanobis distance, this aggregation worsens the balance between covariates. Any extra variable will further reduce the balance quality of previous covariates. Mechanically, pairwise randomization only allows using homogeneous treatment probabilities equal to 1/2. Cytrynbaum (2022) proposes a generalization of pairing, local randomization, that can account for treatment probabilities different than 1/2, but does not resolve the problem of balancing in many covariates. Another limitation of pairing is its computational burden, especially for large samples.

2.2.3 Re-randomization

Finally, re-randomization is another method that allows obtaining balance between covariates that has gained focus in the last decades (Morgan and Rubin, 2012; Li et al., 2018; Imbens, 2011). The main idea of re-randomization is to randomize repeatedly until having balanced groups. Some empiricists perform re-randomization without prespecifying it. This repetition affects treatment probabilities in an unknown manner, which induces invalid inference (Bruhn and McKenzie, 2009; Athey and Imbens, 2017). There are, however, several ways of performing re-randomization that allow valid inferences to some extent. A simple way consists of choosing a criterion —e.g., no imbalances for several covariates, small Mahalanobis distance across groups— and repeating randomization until we meet the criterion. If one chooses the Mahalanobis distance as the criterion, one can perform valid inference (Li et al., 2018). However, using the Mahalanobis distance as the selected criterion entails covariate aggregation. As for the other methods, this aggregation weakens the balance quality and reduces precision when including many variables. Another form of re-randomization is to perform several assignments and then select one randomly among those that satisfy a pre-defined criterion. This process allows us to obtain, mechanically, balance according to the definition of the criterion. One then can perform inference using Fisher's exact p-values. However, the null hypothesis for such tests is much more restrictive.

To perform valid inference and still have balance, Athey and Imbens (2017) recommend stratification with small blocks over re-randomization and pairing. This recommendation is similar to local randomized experiments recently proposed by Cytrynbaum (2022). We postulate that with the Cube method, an empiricist can simultaneously increase balance quality and the number of balanced baseline covariates, compared with these methods, and still draw valid inferences. An empiricist that uses the Cube method to balance baseline covariates gets rid of every significant difference in balance tests (for discrete and continuous variables), gets precision gains, and can easily use heterogeneous treatment probabilities.

3 Setup

This section presents the potential outcome framework, provides assumptions on the data-generating process, and formally defines covariate balancing.

3.1 Data Generating Process and Assignment Design

We consider the standard Neyman-Rubin framework of potential outcomes where $Y_i(0)$ is the outcome of unit i when not treated and $Y_i(1)$ is the outcome when treated. We consider X_i a vector of p covariates (including the constant for the sake of simplicity in the following).

According to the literature, we assume IIDness and the existence of second-order moments for all these variables.

Assumption 1
$(Y_i(0), Y_i(1), X_i)$ are iid across i and $\mathbb{E}\left(Y(0)^2 + Y(1)^2 + ||X||^2\right) < \infty$

The empiricist observes $(X_1, \ldots, X_n)$ for a finite sample of size n. She wants to randomly allocate these n units to treatment according to a design Π, i.e., a distribution on the set of the possible samples of treated $\{0,1\}^n$. If the design Π does not depend on the potential outcomes, it balances potential outcomes in the treatment and control groups *in average*, avoiding selection bias. The design Π could depend on $(X_1, ..., X_n)$: for instance, the treatment probability of a unit i could depend on X_i for various reasons: efficiency, cost of the treatment depending on X_i, subpopulations of particular interest,.... In the following, D_i is the dummy variable indicating if i is treated or untreated. Empiricist have to chose not only each individual selection probability $\mathbb{P}_\Pi(D_i = 1|X_1, ..., X_n)$ but the full design Π that determines $\mathbb{P}_\Pi\left(\cap_{i=1,...,n} D_i = d_i|X_1, ..., X_n\right)$ for any potential allocations $(d_i)_{i=1,...,n} \in \{0,1\}^n$. A major issue is exploiting the knowledge of $(X_1, ..., X_n)$ to define a "good" design Π to go beyond the balancing of potential outcomes in average. To study this question, let us formulate the assumption on the class of design we consider in the following.

Assumption 2 *Empiricist observes a sample $(X_i)_{i=1,...,n}$ of size n and generates a random assignment $(D_i)_{i=1,...,n}$ according to a randomization design Π such that:*

$$(D_1, ..., D_n) \perp\!\!\!\perp (Y_1(0), Y_1(1), ..., Y_n(0), Y_n(1))|X_1, ..., X_n \tag{1}$$

and for any $i = 1, ..., n$

$$\mathbb{P}_\Pi(D_i = 1|X_1, ..., X_n) = p(X_i) \in [c, 1-c], \tag{2}$$

where p is a function chosen by the empiricist and c is a positive constant.

Equation 1 means that assignment is independent of the unknown potential outcomes, conditional on the auxiliary information X. Equation 2 specifies that the assignment probability of unit i could depend on X_i but not on X_j for $j \neq i$. It also states that the propensity score $p(X_i)$ fulfills a common support condition. This restriction is usual and necessary in the literature on treatment effects estimation. In many RCTs, $p(X_i) = 1/2$ for all i, and in that case, there is (on average) the same number of treated and untreated units. But in some cases, for theoretical reasons (for instance: efficiency, population of interest) or practical reasons (e.g., budget constraints), only a smaller fraction of units could be treated ($p(X_i) < 1/2$) and/or $p(X_i)$ could be heterogeneous across i ($\mathbb{V}(p(X_i)) > 0$). In the following, we denote the propensity score $p(X_i)$ as π_i. Our proposition of design accommodates any propensity score type, offering complete flexibility to the empiricists concerning its definition.

After the experiment, the empiricist observes $Y_i = Y_i(1) \times D_i + Y_i(0) \times (1 - D_i)$. She will thus never observe both potential outcomes for the same unit.

Empiricists are generally interested in estimating the sample and population average treatment effects given by

$$\text{SATE}: \quad \theta_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0) \tag{3}$$

and

$$\text{PATE}: \quad \theta_0^* = \mathbb{E}\left[Y_i(1) - Y_i(0)\right], \tag{4}$$

respectively.[†]

In this paper, we will focus on the Horvitz-Thompson estimator (HT) and the Hajek estimator (H), which are of central interest in RCTs. The Horvitz-Thompson estimator is

$$\widehat{\theta}_{HT} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i D_i}{\pi_i} - \frac{Y_i(1-D_i)}{1-\pi_i}\right) \tag{5}$$

which is unbiased for both the SATE and the PATE and is the difference between the inverse probability weighting estimators on the treated and the control group.

Hajec estimator will also be considered:

$$\widehat{\theta}_H = \frac{1}{\sum_{i=1}^{n}\frac{D_i}{\pi_i}}\sum_{i=1}^{n}\frac{Y_i D_i}{\pi_i} - \frac{1}{\sum_{i=1}^{n}\frac{1-D_i}{1-\pi_i}}\sum_{i=1}^{n}\frac{Y_i(1-D_i)}{1-\pi_i}, \tag{6}$$

this corresponds as well to the inverse probability weighting OLS estimator

$$\widehat{\theta}_H = \arg\min_{\theta}\min_{a}\sum_{i=1}^{n} w_i\left(Y_i - a - \theta D_i\right)^2$$

for $w_i = \frac{1}{\pi_i}$ if $D_i = 1$ and $w_i = \frac{1}{1-\pi_i}$ if $D_i = 0$. Let n_T denote the number of treated units and n_C the number of control units. When π_i is constant, $\hat{\theta}_H = \frac{1}{n_T}\sum_{i:D_i=1} Y_i - \frac{1}{n_C}\sum_{i:D_i=0} Y_i$ is the difference between the average on the treated group and the control group whereas $\hat{\theta}_{HT} = \frac{1}{\mathbb{E}(n_T)}\sum_{i:D_i=1} Y_i - \frac{1}{\mathbb{E}(n_C)}\sum_{i:D_i=0} Y_i$ is a slight modification of this difference of averages.

[†] In some cases, they are interested in similar parameters for some subpopulations: $\frac{1}{\sum_{i=1}^{n}\mathbb{1}\{X_i \in \mathcal{X}\}}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))\mathbb{1}\{X_i \in \mathcal{X}\}$ or $\mathbb{E}\left[Y_i(1) - Y_i(0) | X_i \in \mathcal{X}\right]$. Estimators of these quantities are defined restricting the sample to units such that $X_i \in \mathcal{X}$ and the asymptotic properties of these estimators follows from a straightforward adaptation of what is presented in the following.

3.2 Balancing constraints

Under Assumption 2, as soon as $\mathbb{E}(D_i|(X_{i'}, Y_{i'}(0), Y_{i'}(1))_{i'=1,...,n}) = \pi_i$ and we have balancing in average for the potential outcomes:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i:D_i=1}\frac{Y_i(1)}{\pi_i}\middle|(X_{i'})_{i'=1,...,n}\right) = \frac{1}{n}\sum_{i=1}^{n} Y_i(1),$$
$$\mathbb{E}\left(\frac{1}{n}\sum_{i:D_i=0}\frac{Y_i(0)}{1-\pi_i}\middle|(X_{i'})_{i'=1,...,n}\right) = \frac{1}{n}\sum_{i=1}^{n} Y_i(0).$$

for any covariates X_j for $j = 1, ..., p$:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i:D_i=1}\frac{X_{ji}}{\pi_i}\middle|(X_{i'}, Y_{i'}(0), Y_{i'}(1))_{i'=1,...,n}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i:D_i=0}\frac{X_{ji}}{1-\pi_i}\middle|(X_{i'}Y_{i'}(0), Y_{i'}(1))_{i'=1,...,n}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} X_{ji}.$$

As explained in Sections 1 and 2, to go beyond the balancing of potential outcomes on average, empiricists can take advantage of the observation of covariates X before the experiment. A long and natural idea (Fisher, 1926) is to balance these covariates not only in average but also almost surely. Let us define more precisely a perfectly-balanced design.

Definition 1 (Perfectly-balanced Design)
A design Π is perfectly-balanced over $X = (X_1, ..., X_p)'$ if for $(D_i)_{i=1,...,n}$ sampled in Π we always have for any $j = 1, ..., p$:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ji}D_i}{\pi_i} = \frac{1}{n}\sum_{i=1}^{n} X_{ji} \tag{7}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ji}(1-D_i)}{1-\pi_i} = \frac{1}{n}\sum_{i=1}^{n} X_{ji} \tag{8}$$

Equation (7) describes equality between the covariate sample mean and the estimated mean in the treatment group, whereas equation (8) ensures perfect balance for the control group. A perfectly-balanced assignment eliminates any allocation to the treatment that does not balance perfectly the covariates between treatment and control groups. Note that when π_i is constant, conditions (7) and (8) are equivalent. But this is not the case if the π_i are heterogeneous. A common practice in experiments is to form treatment and control groups of fixed sizes, n_T, and $n_C = n - n_T$, respectively. This is equivalent to setting the constraint in (7) with $X_{ji} = \pi_i$. Indeed, for any possible allocation $(d_1, ..., d_n)$:

$$n_T = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \pi_i = \mathbb{E}(n_T) \tag{9}$$

In that case we also have: $n_C = \sum_{i=1}^n (1-d_i) = \sum_{i=1}^n (1-\pi_i) = \mathbb{E}(n_C)$, and under such assignment $\widehat{\theta}_{HT}$ defined in (5) is equal to $\widehat{\theta}_H$ defined in (6). Notice that, as in Tillé and Favre (2004), we can rewrite (7), (8), (9) as

$$\frac{1}{n}\sum_{i=1}^n \frac{Z_i D_i}{\pi_i} = \frac{1}{n}\sum_{i=1}^n Z_i \tag{10}$$

with $Z_i = (\pi_i, X_i', \frac{\pi_i}{1-\pi_i}X_i')'$. If assignment probabilities are homogeneous (i.e., $\pi_i = \pi$) and if the constant covariate is included in X, perfect balancing for the treatment group (Equation 7) implies perfect balancing on Z_i (Equation 10) but this is no more the case if the π_i are heterogeneous.

The notion of perfect balancing is closely related to the balancing tests produced by empiricists after randomization. The balancing tests check ex-post that randomization has balanced or not the treated and the treatment group on the covariates not in average only but for a particular sampling according to the design Π. A perfectly-balancing design integrates ex-ante the information contained in the observation of X to ensure balancing ex-post and improve the balancing in the potential outcomes. For these reasons, debates on the conciliation of randomization and balancing have a long history in statistical sciences (see for instance Fisher, 1926). It is worth noticing that perfect balancing is not always attainable: for instance, if $n = 101$ and $\pi_i = 1/2$. Imposing (9) implies $n_T = 50.5$, which is simply impossible. But statistical analysis ensures that balancing up to a $o_p\left(\frac{1}{\sqrt{n}}\right)$ is sufficient to take full advantage of the auxiliary information X_i, π_i. Empiricists are then reduced to find a design Π such that for $(D_i)_{i=1,\dots,n} \sim \Pi$, we have:

$$\frac{1}{n}\sum_{i=1}^n \frac{Z_i D_i}{\pi_i} = \frac{1}{n}\sum_{i=1}^n Z_i + o_p\left(\frac{1}{\sqrt{n}}\right) \text{ for } Z_i = \left(\pi_i, X_i', \frac{\pi_i}{1-\pi_i}X_i'\right)' \tag{11}$$

Various randomization strategies have been proposed and used in the literature to achieve Equation (11). Some of the oldest and widest-used strategies to do so are stratification or blocking (Fisher, 1926), rerandomization (Student, 1938; Morgan and Rubin, 2012), pairwise randomization (Imai et al., 2009; Greevy et al., 2004; Ball et al., 1973). We discussed their limitations in Section 2. In this paper, we advocate the Cube method that achieves simultaneously many desirable properties of the various usual balancing design with a large number of covariates Z that could be as large as a $O(n^{1/2-1/r})$ if covariates admit moments of order r and as large as $o(n^{1/2})$ if covariates have bounded supports. Let us briefly present the Cube method before detailing its theoretical properties, the consequence of its use on the estimators $\widehat{\theta}^H$ and $\widehat{\theta}^{HT}$ and on the inference about PATE and SATE.

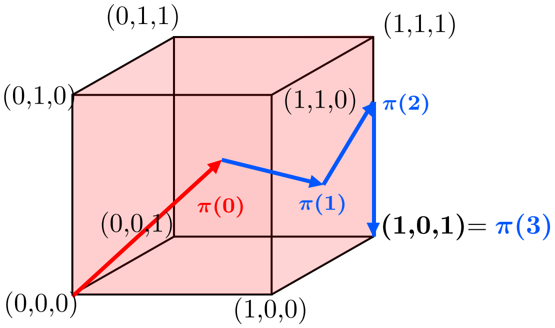
4 Balancing covariates with the Cube method

Deville and Tillé (2004) first introduced the Cube method to produce samples balanced to the population. The Cube method consists of an algorithm in two steps: the *flight* and *landing phase*s. The technique gets its name from the graphical representation of a sampling problem. Equation (10) or (11) ensure that balancing treatment and control groups in an experimental setting for some covariates is equivalent to balancing the treatment group to the entire sample. Let us consider the n-cube $C = [0,1]^n$. Each vertex of C (from 2^n possibilities) represents a possible treatment group: for instance, $(1,1,...,1)$ corresponds to the situation where all units are allocated to treatment, $(1,0,1,0,...,1,0)$ corresponds to the case where the treatment group is $\{i : i \text{ odd}\}$. A sampling design Π corresponds to how a vertex is selected. Recall that we consider a framework where empiricists impose that equation (2) holds for Π and a vector $(\pi_i)_{i=1,...,n}$.

We will first describe the Cube design without balancing constraints before moving to the more interesting case where balancing constraints (11) are considered. Whatever the set of balancing constraints, the Cube method is a discrete martingale that moves in (at most) n steps from the interior point $\boldsymbol{\pi(0)} = (\pi_i)_{i=0}^n$ to $\boldsymbol{\pi(n)} = (D_i)_{i=0}^n$ a vertex of C. At the first step, one chooses a random direction for $\boldsymbol{\pi(1)} - \boldsymbol{\pi(0)}$ and a step size such that $\boldsymbol{\pi(1)}$ belongs to a facet of C. We can then show that $\mathbb{E}[\boldsymbol{\pi(1)}|\boldsymbol{\pi(0)}] = \boldsymbol{\pi(0)}$. After this step, because $\boldsymbol{\pi}(1)$ belongs to a facet of C, one component i_0 of $\boldsymbol{\pi}(1)$ is equal to 0 or 1, selecting $D_{i_0} = \pi_{i_0}(1)$ one has thus assigned a first unit to either treatment or control group. Because a facet of a n-cube is a $(n-1)$-cube, one then repeats the process in a $(n-1)$-cube, and so on, until landing in a vertex of C. At the final step n, one will have $(D_i)_{i=1,...,n} = \boldsymbol{\pi(n)} \in \{0,1\}^n$ and $\mathbb{E}[D_i] = \pi_i$ (i.e., every unit is allocated to the treatment group with the probability specified by the empiricist). These successive steps are the *flight phase* and for the Cube method without balancing constraint, allocation $(D_i)_{i=1,...,n}$ is always determined at the end of this phase. Figure 1 illustrates graphically the method.

In Figure 1, all the vertex of the n-cube can be selected meaning that all individuals could be allocated to the control group. We now consider that empiricist wants to allocate a fixed number n_T of units to the treatment and n_c units to the control. This can be achieved with the Cube method as soon as $\sum_{i=1}^n \pi_i = n_T$. The condition that exactly n_T units are assigned to the treatment can be express as a balancing constraint. Indeed, because $n_T = \sum_i D_i$ and $\sum_i \pi_i = n_T$, the fixed size condition is equivalent to $\sum_i \frac{Z_i D_i}{\pi_i} = \sum_i Z_i$ for $Z_i = \pi_i$. Let K the

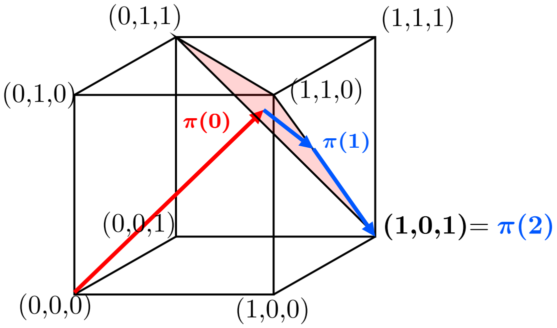
Figure 1: Cube method without balancing constraints



This figure depicts an example of the Cube algorithm with $n = 3$ when no balancing constraint is imposed. The red arrow represents the initial treatment probabilities $(\pi_i)_{i=1,\dots,n}$. Then, every blue arrow is a step of the *flight phase*. In this example, the first unit is initially assigned to the treatment group. Then, the third unit is assigned to the treatment group. Last, the second unit is assigned the control group. Therefore, the allocation – in bold – is $(1, 0, 1)$.

set of vector s in the n-cube C such that $\sum_i s_i = n_T$. K is a closed convex set and its extreme points are the vertex of C, that is the set of allocation that respects the fixed size constraints. K is contained in an affine subspace of dimension $n-1$ of direction $V := \{v : \sum_{i=1}^{n} v_i = 0\}$, we have $K = C \cap \{\boldsymbol{\pi(0)} + v : \sum_i v_i = 0\}$. The Cube method select randomly an element of V for the direction of $\boldsymbol{\pi(1)} - \boldsymbol{\pi(0)}$ and fix the step size such that $\boldsymbol{\pi(1)}$ is a border point of K. We can show $E(\boldsymbol{\pi(1)}|\boldsymbol{\pi(0)}) = \boldsymbol{\pi(0)}$. After this first step, $\boldsymbol{\pi(1)}$ belongs to a facet of C and a unit i_1 is assigned either to the treatment either to the control group. Units $i \neq i_0$ remains to assign and we have $\sum_{i:i\neq i_0} \pi_i(1) = n_T - D_{i_0}$. We can then replicate the first step after replacing n_T by $n_T - D_{i_1}$ the sample $\{1, ..., n\}$ by $\{1, ..., n\} \backslash \{i_0\}$ and to allocate a second unit and to update assignment probability as $\boldsymbol{\pi(2)}$. At step $n-1$, $\boldsymbol{\pi(n-1)}$ belong to the extreme points of K, this ends the flight phase. Because the extreme points of K are some vertices of C the assignment is achieved. Now imagine that one have 101 units to assign with equal probability to the treatment and control groups. Perfect balancing on the two groups sizes is not possible 101 is an odd integer and it is not possible to assign 50.5 units to the treatment. A popular solution is to consider $\pi_i = 50/101$ or $\pi_i = 51/101$ and to sample randomly 50 (or 51) elements among the 101 units. However, this strategy does not accommodate easily with heterogeneous probabilities of assignment and does not generalize to take into account many balancing constraints. With the cube method described above, for each step t of the flight phase we have $\sum_i \pi_i(t) = 50.5$ and the extreme points of K are no more vertices of C. In that case, at the end of the flight phase, $n-1 = 100$ units are assigned

Figure 2: Cube method with fixed sample size



This figure depicts an example of the Cube algorithm with $n = 3$ when imposing the constraint $n_T = 2$ and $\sum_{i=1}^{3} \pi_i = 2$. The red area depicts the points (s_1, s_2, s_3) in the cube satisfying the equation $\sum_{i=1}^{3} s_i = 2$. This condition is equivalent to imposing the balancing constraint $\sum_i \frac{Z_i s_i}{\pi_i} = \sum_i Z_i$ with $Z_i = \pi_i$. The red arrow represents the initial treatment probabilities. Then, every blue arrow is a step of the *flight phase*. In this example, the first unit is initially assigned to the treatment group. Then, since $n_T = 2$, only one unit among the second and third units can be assigned to the treatment group. In this case, the second blue arrow shows that, in the same step, the second unit is assigned to the control group and the third one to the treatment group. Therefore, the last allocation – in bold – is $(1, 0, 1)$.

at the end of the flight phase with $(n-1)/2 = 50$ units to the treatment and $(n-1)/2 = 50$ units to the control. The Cube method can be completed with a last phase that randomly assign to the treatment of the control the remaining unit ensuring that $n_T = 50$ or 51 and $E(n_T) = 50.5$, in that case the sizes of treatment and control groups are not exactly fixed but almost fixed (in fact as fixed as possible as soon as we respect the initial assignment probabilities $\pi_i = 1/2$). This second phase is called *landing phase*. These two phases (the flight phase and the landing phase) can be generalized to the case where the empiricist wants to impose several balancing constraints and heterogeneous probabilities of treatment.

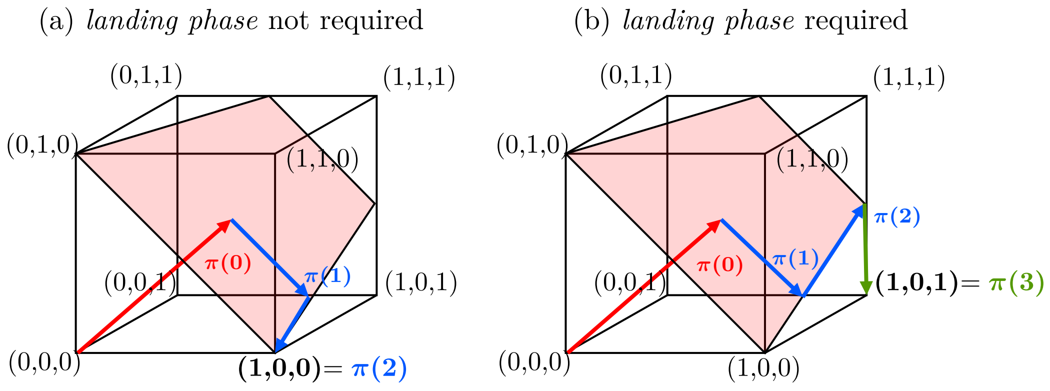
Let describe the Cube method with an arbitrary number of balancing constraints defined by q covariates Z_i. A point $\mathbf{s} \in C$ will satisfy an equation analog to (10) if

$$\sum_{i=1}^{n} \frac{Z_i s_i}{\pi_i} = \sum_{i=1}^{n} Z_i. \tag{12}$$

Let $A_i = \frac{Z_i}{\pi_i}$ and $A = (A_1, ..., A_n)$ the matrix of size $q \times n$. Then (12) is equivalent to

$$\begin{aligned} & \sum_{i=1}^{n} A_i s_i = \sum_{i=1}^{n} A_i \pi_i \\ \Leftrightarrow \quad & A\mathbf{s} = A\boldsymbol{\pi}(\mathbf{0}) \end{aligned}$$

Figure 3: CUBE method with one balancing constraint


(a) *landing phase* not required

(b) *landing phase* required

This figure depicts an example of the CUBE algorithm with $n = 3$ where we do not always get perfectly-balanced allocations. We consider the initial treatment probabilities in (2) to be $\pi_1 = \pi_2 = \pi_3 = \frac{2}{3}$. The red area depicts the points (s_1, s_2, s_3) in the cube satisfying the equation $\sum_{i=1}^{3} s_1 + s_2 - \frac{1}{2}s_3 = 1$. This is equivalent to imposing the constraint in (12) with $Z_1 = Z_2 = \frac{2}{3}$ and $Z_3 = -\frac{1}{3}$. The red arrow represents the initial treatment probabilities. Since not every vertex of the plane is a cube vertex, we cannot always satisfy the constraint. In both panels, the algorithm assigns the first unit to the treatment group (first blue arrow). The second blue arrow corresponds to the assignment of the third unit. If the algorithm assigns the third unit to the control group (panel a), it automatically assigns the second one to treatment. However, if the algorithm assigns the third unit to the treatment group (panel b), the second unit is in neither group, even if we attain a plane vertex. In the *landing phase*, the CUBE algorithm will proceed by randomly allocating the second unit. In this example, the green arrow shows that the *landing phase* allocates the second unit to the control group.

$$\Leftrightarrow \quad \mathbf{s} \in Q := \boldsymbol{\pi}(\mathbf{0}) + \ker(A).$$

$K = C \cap Q$ is, therefore, the $(n-q)$-polytope that contains all the points in C such that (12) holds. At the first step, one chooses a random direction in $v \in \ker(A)$ and we select the unique $\lambda > 0$ such that $\boldsymbol{\pi}(\mathbf{1}) := \boldsymbol{\pi}(\mathbf{0}) + \lambda v$ is on a facet of K. One can show that $\mathbb{E}[\boldsymbol{\pi}(\mathbf{1})|\boldsymbol{\pi}(\mathbf{0})] = \boldsymbol{\pi}(\mathbf{0})$. Because any facet of K is the intersection of a facet of C with Q, a component i_0 of $\boldsymbol{\pi}(\mathbf{1})$ is 0 or 1 and defining $D_{i_0} = \pi_{i_0}(1)$ one has assigned a first unit. Next, one applies a similar step for the facet of K instead of K and $\boldsymbol{\pi}(\mathbf{1})$ as a starting point instead of $\boldsymbol{\pi}(\mathbf{0})$. After $n - q$ steps, one has reached a vertex of K. This process corresponds to the *flight phase* in Deville and Tillé (2004). If this vertex of K is also a vertex of C, the *flight phase* allocates every unit, and the two groups are perfectly balanced (see Figures 2 and 3a). But in many cases, the vertex of K is not a vertex of C, and it remains at most

q units to assign during the *landing phase* (according to the wording of Deville and Tillé, 2004) (see Figure 3b).

Say that at the end of the *flight phase*, one has not assigned $r \leq q$ units and let $\boldsymbol{\pi}^* = \boldsymbol{\pi(n-q)}$ be the updated treatment probabilities at this stage. The *landing phase* of the CUBE method assigns the r missing units such that $\mathbb{E}[D_i|\boldsymbol{\pi}^*] = \boldsymbol{\pi}^*$. Grafström and Tillé (2013) describe two methods for the landing phase (these are also the options used in sampling packages): (i) Linear programming: one considers all the 2^r allocations for these units and assigns probabilities to each allocation to minimize a cost function (such as distance to K) and satisfying $\mathbb{E}[D_i|\boldsymbol{\pi}^*] = \boldsymbol{\pi}^*$. Then, using these probabilities, one randomly selects an allocation. (ii) Suppression of variables: if $r > 20$, solving a linear problem becomes computationally difficult. In that case, at the end of the flight phase, one can drop a covariate and continue with the flight phase. One can thus successively drop variables (in the order of preference) until attaining a vertex of C.

5 Statistical Properties of the Cube method

This Section shows how the Cube method allows obtaining an almost-perfect balance between the treatment and control groups and relates this balance to gains in precision for treatment effect estimators.

5.1 Balancing approximations

As explained above, designing an allocation mechanism that always produces perfectly-balanced groups is impossible. However, we here prove that the Cube method is successful, under certain conditions, in creating almost-perfectly-balanced samples.

To check balance properties after allocating individuals according to the design Π, empiricists are interested in computing the difference

$$\Delta^{\Pi}_{j,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_{ji}D_i}{\pi_i} - \frac{X_{ji}(1-D_i)}{1-\pi_i}.$$

Because $\mathbb{P}_{\Pi}(D_i = 1|(X_{i'})_{i'=1,\ldots,n}) = \pi_i$, we have $\mathbb{E}\left(\Delta^{\Pi}_{j,n}\right)$ and under weak conditions on Π, we have

$$\sqrt{n}\Delta^{\Pi}_{j,n} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}(\Delta^{\Pi}_{j,n})\right), \tag{13}$$

where $\mathbb{V}(\Delta^{\Pi}_{j,n})$ is an asymptotic variance depending on Π and the distribution of X.

For the so-called baseline balance tests, empiricists consider the t-statistic

$$t^{\Pi}_{j,n} = \sqrt{n}\frac{\Delta^{\Pi}_{j,n}}{\sqrt{\widehat{\mathbb{V}}(\Delta^{\Pi}_{j,n})}}$$

where $\widehat{\mathbb{V}}(\Delta^{\Pi}_{j,n})$ is a consistent estimator of the asymptotic variance of $\Delta^{\Pi}_{j,n}$ to test the null hypothesis of perfect balance. $t^{\Pi}_{j,n}$ is then associated to a p-value $p^{\Pi}_{j,n}$ which take values between 0 and 1. As explained in Sections 1 and 2, when creating balance tests for RCTs, p-values below 0.15 are considered problematic (Snyder and Zhuo, 2018).

Let us first consider a naive mechanism that does not use baseline information to assign units. Such situations correspond to the case where the design Π is a Poisson design, ie a design where each unit i is allocated to the treatment independently of the allocation of other units:

$$\mathbb{P}_{\Pi}\left(\bigcap_{i=1}^{n}\{D_i = d_i\}\big|(X_i)_{i=1}^{n}\right) = \prod_{i=1}^{n}\pi_i^{d_i}(1-\pi_i)^{1-d_i}.$$

A Poisson design does not balance any variable.

When $\pi_i = \frac{n_T}{n}$ for any i, another popular design is sampling without replacement of n_T treated units.

$$\mathbb{P}_{\Pi}\left(\bigcap_{i=1}^{n}\{D_i = d_i\}\big|(X_i)_{i=1}^{n}\right) = \binom{n}{n_T}^{-1}\mathbb{1}\left\{\sum_{i=1}^{n} d_i = n_T\right\}.$$

Sampling design without replacement and equal probabilityπ only balances constant variables. In that case, the sample of treated and control groups are fixed, and the design is also balanced on the constant $\sum_{i=1}^{n} D_i = n_T$, $\sum_{i=1}^{n}(1-D_i) = n - n_T$ and $\sum_{i=1}^{n}\frac{D_i}{\pi_i} = n$.

Under such assignments and Assumptions 1 and 2 and more generally for any design Π such that (13) holds with $\mathbb{V}(\Delta^{\Pi}_{j,n}) > 0$, we have $\Delta^{\Pi}_{j,n} = O_p\left(\frac{1}{\sqrt{n}}\right)$, $t^{\Pi}_{j,n} \xrightarrow{d} \mathcal{N}(0,1)$ and $p^{\Pi}_{j,n} \xrightarrow{d} \mathcal{U}(0,1)$. This means that if one randomizes naively, control and treatment groups will present imbalances with a strictly-positive probability. Moreover, for a confidence level of $100(1-\alpha)\%$, there exists always $100\alpha\%$ chance of obtaining significant differences. If an empiricist evaluates the balance of 10 independent covariates at the 85% confidence level (a level retained in the literature over rejection of the balancing assumption balancing is considered as problematic Snyder and Zhuo, 2018), there is more than 80% chance of having at least one significant difference. This magnitude questions the mere implementation of such widely used tests. Even if a multiple F-test with a confidence level of 85% mitigates this rejection rate, the null hypothesis of simultaneously balanced covariates is rejected by construction with a 15% chance.

The Cube method ensures that these tests are unnecessary since we can balance control and treatment groups in any variate $(X_j)_{j=1,...,p}$. This is achieved because $\mathbb{V}(\Delta^{\Pi}_{j,n}) = 0$ for any $j = 1, ..., p$ in (13). Performing these tests would not make sense since we never reject the null hypothesis by construction. However, one might report them if the editor worries about empiricists randomizing badly. Usual balancing strategies are stratification or pairwise matching. These methods ensure $\mathbb{V}(\Delta^{\Pi}_{j,n}) = 0$ if the covariates $(X_j)_{j=1,...,p}$ are all discrete but will always generate imbalances for continuous ones since the empiricist needs to discretize or aggregate them before randomizing.

The following proposition explains how the balancing approximations are satisfied with the Cube method. Because the number q of balancing constraints in equation (11) could be large with the Cube method, we are also explicit on how q affects balancing approximations to allow us to consider a framework where q tends to ∞.

Proposition 1 (Balancing approximations with the Cube method)
If Assumptions 1 and 2 hold, then

$$\Delta^{Cube}_{j,n} = o_p\left(\frac{q}{\sqrt{n}}\right).$$

Moreover if $\mathbb{E}\left[|X_{j1}|^r\right] < \infty$ *for* $r \geq 2$, *then* $\Delta^{Cube}_{j,n} = o_p\left(\frac{q}{n^{1-1/r}}\right)$, *if* X_{j1} *is sub-Gaussian, then* $\Delta^{Cube}_{j,n} = O_p\left(\frac{q\sqrt{\ln(n)}}{n}\right)$, *and if* X_{j1} *has a bounded support, then* $\left|\Delta^{Cube}_{j,n}\right| < \frac{Kq}{cn}$ *for* K *such that* $|X_{j1}| < K$. *As soon as* $\sqrt{n}\Delta^{Cube}_{j,n} = o_p(1)$, *we have* $t^{Cube}_{j,n} \xrightarrow{P} 0$, *and* $p^{Cube}_{j,n} \xrightarrow{P} 1$.

Proposition 1 shows that, as n grows, the Cube method ensures the balancing Equation (11) as soon as the second-order moment of X exists. Furthermore, if moment of order $r > 2$ exists for X, (11) holds as soon as $q = O\left(n^{\frac{1}{2}-\frac{1}{r}}\right)$. q can even be $o\left(\sqrt{\frac{n}{\ln(n)}}\right)$ if the covariates X are all subgaussians or $o(\sqrt{n})$ if they are bounded. This means that with probability tending to one, the p-values of the balance test tend to 1 meaning that balancing is never rejected for large n contrary to randomization under a design Π such that (11) does not hold.

5.2 Variance reduction

The balance between covariates in the control and treatment groups is also beneficial if these variables are related to the potential outcomes. In this case, using the Cube method will also reduce the variance of the HT estimator.

Assumption 3

For $d \in \{0, 1\}$,

$$Y_i(d) = \beta_d X_i + \varepsilon_i(d), \ \textit{with } \mathbb{E}[\varepsilon_i(d)|X_i] = 0$$

.

Assumption 3 states that potential outcomes are linearly related to observable covariates. However, we allow heterogeneity in treatment effects by specifying different equations for control and treatment groups.

Conjecture 1 (Poisson approximation)

For any $k \in \mathbb{N}^*$ we have with probability one:

$$\lim_{n\to\infty} \sup_{i_1,...,i_k} \left| \mathbb{E}\left(\prod_{j=1}^{k} \left(D_{i_j} - \pi_{i_j}\right) |X_1, ..., X_n \right) \right| = 0$$

This conjecture establishes that as n increases, the Cube method tends to Poisson sampling. As n goes to infinity, the dependence between the assignment of a finite number of individuals disappears. We draw this conjecture from results in Deville and Tillé (2005) and simulations that confirm it. This conjecture is unnecessary for getting results on the SATE, but it is useful when focusing on the PATE.

In order to have a benchmark for the gains in variance decline, we compare the Cube method with Poisson randomization, i.e., an unconstrained sampling with heterogeneous treatment probabilities. The results also hold for simple randomization, that is, with homogeneous treatment probabilities.

Proposition 2 (Asymptotic variance)

Let θ_0 be the SATE defined in (3), and $\widehat{\theta}$ be the HT estimator in (5). If Assumptions 1-3 hold, and if Π is a balancing sampling using the Cube method we have:

$$\mathbb{E}\left[\left(\widehat{\theta} - \theta_0\right)^2\right] = \frac{V_0}{n} + O\left(\frac{1}{n^2}\right),$$

whereas $\mathbb{E}\left[\left(\widehat{\theta} - \theta_0\right)^2\right] = \frac{V_0+\Sigma_0}{n} + O\left(\frac{1}{n^2}\right)$ *for* $V_0 = \mathbb{E}\left(\pi_i(1-\pi_i)\left(\frac{\varepsilon_i(1)}{\pi_i} + \frac{\varepsilon_i(0)}{1-\pi_i}\right)^2\right)$ *and* $\Sigma_0 = \mathbb{E}\left[\frac{1-\pi_i}{\pi_i}(X_i'\beta_1)^2\right] + \mathbb{E}\left[\frac{\pi_i}{1-\pi_i}(X_i'\beta_0)^2\right]$ *if simple randomization is used.*

Moreover, let θ_0^* the PATE in (4). If, additionally Conjecture 1 also holds, then with the Cube method:

$$\mathbb{E}\left[\left(\widehat{\theta} - \theta_0^*\right)^2\right] = \frac{V_0^*}{n} + O\left(\frac{1}{n^2}\right),$$

whereas $\mathbb{E}\left[\left(\widehat{\theta} - \theta_0^*\right)^2\right] = \frac{V_0^*+\Sigma_0^*}{n} + O\left(\frac{1}{n^2}\right)$ *for* $V_0^* = (\beta_1-\beta_0)'\mathbb{V}(X_i)(\beta_1-\beta_0) + \mathbb{E}[\frac{\varepsilon_i(1)^2}{\pi_i}] + \mathbb{E}[\frac{\varepsilon_i(0)^2}{1-\pi_i}]$ *and* $\Sigma_0^* = \mathbb{E}\left[\frac{(X_i'\beta_1)^2}{\pi_i}\right] + \mathbb{E}\left[\frac{(X_i'\beta_0)^2}{1-\pi_i}\right]$ *if simple randomization is used.*

Proposition 2 shows the gain in asymptotic variance from balancing covariates using the Cube method. The reduction is more substantial when X explains more of the potential outcomes. Estimates of the ATE are thus more precise when using the Cube method. This reduction can represent significantly lower costs when conducting an RCT. Simulations in Section 6 estimate these gains.

5.3 Inference

This section provides properties of the Cube algorithm and methods to perform inference. We elicit two main techniques of conducting inference, one based on the asymptotic properties of the HT estimator and the other based on the randomization mechanism.

5.3.1 Asymptotics-based inference

Some methods, such as re-randomization, alter the inclusion probabilities in a manner that is unclear to the empiricist (Imbens, 2011). When the criterion for selection is known and behaves in a known way, such as the Mahalanobis distance, one can perform conservative inference. However, balance is imperfect for numerous covariates. Since the Cube method assigns treatment only once, we can perform asymptotic-based inference. We here give the asymptotic properties and propose an easy way to construct less conservative confidence intervals.

Proposition 3
Let Assumptions 1-3 hold and $\pi_i = \frac{1}{2}, \forall i \in \{1, \dots, n\}$. Then, using the Cube method,

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V_0\right).$$

Moreover, if in addition of Assumptions 1-3, Conjecture 1 holds, the Cube method yields for any $\pi_i \in [c, 1-c]$, $c > 0$,

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V_0\right)$$

and

$$\sqrt{n}\left(\widehat{\theta} - \theta_0^*\right) \xrightarrow{d} \mathcal{N}\left(0, V_0^*\right).$$

To construct a confidence interval, one would like to estimate either V_0 or V_0^*. Estimating V_0/n is impossible without making assumptions on the relation between $\varepsilon_i(1)$ and $\varepsilon_i(0)$. This issue is common in RCTs. We can, nonetheless, easily construct an unbiased estimator $\widehat{V}$ for V_0^*/n. Let $\widehat{\beta}_d$ and $\widehat{\varepsilon}_i(d)$ be the estimated coefficients and residuals, respectively, of a regression of $Y_i(d)$ on X_i, for $d \in \{0, 1\}$. We then have

$$\widehat{V} = \frac{1}{n}\left[\left(\widehat{\beta}_1 - \widehat{\beta}_0\right)' \widehat{\mathbb{V}}(X_i)\left(\widehat{\beta}_1 - \widehat{\beta}_0\right) + \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\varepsilon}_i(1)^2 D_i}{\pi_i} + \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\varepsilon}_i(0)^2(1-D_i)}{1-\pi_i}\right]. \tag{14}$$

Then, we can test the weak hypothesis

$$H_0 : \theta_0^* = 0, \tag{15}$$

and construct the confidence interval based on

$$\widehat{\theta} \pm \Phi^{-1}\left(1 - \alpha/2\right)\sqrt{\widehat{V}} \tag{16}$$

In Section 6, we perform simulations that confirm the exact coverage rate of this confidence interval when n is big enough ($n \leq 200$).

5.3.2 Randomized-based inference

We here study the properties of randomization-based inference when permuting treatment status while satisfying balancing constraints. For these tests, we consider the stronger null hypothesis:

$$H_0 : (Y_i(1), X_i) \overset{d}{=} (Y_i(0), X_i). \tag{17}$$

Notice that testing this hypothesis, under Assumptions 1 and 2 is equivalent to testing $(Y_i)_{i=1}^n \perp\!\!\!\perp (D_i)_{i=1}^n | X_1, \dots X_n$ (Proof in Appendix).

To explain the test, we introduce some new notation. Let G_n be the set of all possible 2^n assignments. Then, we can define the set of assignments $G_n^{Cube} \subseteq G_n$ satisfying the constraints imposed by the Cube method. That is, with Assumptions 1 and 2,

$$G_n^{Cube} = \left\{g \in G_n : \Delta_{j,n} = o_p\left(\frac{q}{\sqrt{n}}\right) \text{ for } 1 \leq j \leq p\right\}$$

.

We note $\mathbf{P_n} = (Y_i, D_i, X_i)_{i=1}^n$ the observed values, and $\mathbf{P_n^{(g)}} = (Y_i, D_i^{(g)}, X_i)_{i=1}^n$, the new data where we have reassigned treatment according to $g \in G_n^{Cube}$. For computational facility, we can replace G_n^{Cube} by $G_n^B = \{g_1, \dots, g_B\}$, such that g_1 is the assignment really obtained and $(g_i)_{i=1}^B$ are drawn independently from a uniform distribution on G_n^{Cube}.

Then, for a given test statistic $T_n(\mathbf{P_n})$,we consider the test

$$\phi^{rand}(\mathbf{P_n}) = \mathbb{1}\left\{T_n(\mathbf{P_n}) > c_n(\mathbf{P_n}, 1-\alpha)\right\}$$

with

$$c_n(\mathbf{P_n}, 1-\alpha) = \inf\left\{t \in \mathbb{R} : \frac{1}{B}\sum_{g \in G_n^B} \mathbb{1}\{T_n(\mathbf{P_n^{(g)}}) \leq t\} \geq 1-\alpha\right\}.$$

Proposition 4
Under Assumptions 1 and 2, and the null hypothesis in (17)*,*

$$\mathbb{E}\left[\phi_n^{rand}(\mathbf{P_n})\right] \leq \alpha.$$

Proposition 4 indicates that if $T_n(\mathbf{P_n}) > c_n(\mathbf{P_n}, 1-\alpha)$, we reject the null hypothesis (17) at the α level. The proof is similar to previous results on other covariate-adaptive assignment mechanisms (Heckman et al., 2010, 2011; Lee and Shaikh, 2014; Bai et al., 2022), but it is presented for completeness. This proposition ensures that we can compute Fisher's p-values by comparing our test statistic with those produced by other assignments made by the Cube method.

6 Empirical applications

In this section, we simulate using data from Lee et al. (2021) and Gerber et al. (2020) to show the benefits of using the Cube method. We first present how the Cube method can improve balance significantly compared with naive randomization and stratification. We then explicit precision gains and their consequences on sample size reduction.

For each paper, we observe the data $\mathbf{P_n} = (Y_i, X_i, D_i)_{i=1}^n$. First, we create $\mathbf{Q_n} = (Y_i(1), Y_i(0), X_i)_{i=1}^n$. For this purpose, we consider $Y_i(d) = Y_i$ for $d = D_i$, and we use machine learning techniques to impute $Y_i(d)$ for $d \neq D_i$. First, we use treated and control units separately to train two models $f_1(X_i)$ and $f_0(X_i)$, respectively, that predict potential outcomes in each case. Then, for $D_i = d$, we compute the residuals $\hat{u}_i(d) = Y_i(d) - f_d(X_i)$, and compute variances $\sigma_{\hat{u}}(d)$. For each unit i, we consider $d \neq D_i$ and impute $Y_i(d) = f_d(X_i) + u_i(d)$, with $u_i(d) \sim \mathcal{N}(0, \sigma_{\hat{u}}(d))$. Finally, we compute $\theta_0 = \frac{1}{n}\sum_{i=1}^n Y_i D_i - Y_i(1-D_i)$, the parameter of interest.

For $k \in \{1, \ldots, K\}$, we then sample with replacement $n' \leq n$ units from $\mathbf{Q_n}$ to get $\mathbf{Q_{n'}(k)} = (Y_{i_k}(1), Y_{i_k}(0), X_{i_k})_{i_k=1}^{n'}$. We can then choose a treatment assignment mechanism to generate $\mathbf{P}_{\mathbf{n'}}(\mathbf{k}) = (Y_{i_k}, X_{i_k}, D_{i_k})_{i_k=1}^{n'}$ and compute

$$\widehat{\theta}_k = \frac{1}{n'} \sum_{i_k=1}^{n'} \frac{Y_{i_k} D_{i_k}}{\pi_{i_k}} - \frac{Y_{i_k}(1 - D_{i_k})}{1 - \pi_{i_k}}$$

and its associated p-values.

6.1 Design effect on balancing covariates

Lee et al. (2021) study the effect of mobile banking on remittances from rural-urban migrants in Bangladesh. We consider here five balancing covariates: baseline remittances, household size, age, gender, and primary education. Baseline remittances are a covariate of particular interest since they correspond to the pre-treatment outcome and are continuous, making it impossible to stratify without discretizing it. Household size and age are discrete variables with many possible values, making it hard to stratify without creating groups. Gender and primary education are dummy variables, making it easy to stratify. Gerber et al. (2020) at belief updating and voting behavior after being exposed to different polls. We here consider the beliefs about the margins of the elections as the outcome variable. We use six variables for balancing: baseline beliefs (continuous), past voting behavior (continuous), interest in politics (discrete), years of schooling (discrete), identification of Nancy Pelosi as the speaker (dummy), and gender (dummy).

For simplicity, we fix treatment probabilities to 1/2. We perform 10,000 simulations and test the balance of covariates with five different assignment mechanisms with fixed sample size: (i) Naive Randomization, (ii) Stratification on the pre-treatment outcome (discretized by quartiles), (iii) Cube Method on the baseline outcome, (iv) Stratification on all variables (baseline outcomes are discretized by quartiles, and other non-dummy variables are grouped using the median value as the cutoff), (v) and Cube Method on all the variables. We then compute for each mechanism ℓ and variable j the average squared-mean-differences (ASMD)

$$\text{ASMD}_j^{(\ell)} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{n'} \sum_{i_k=1}^{n'} \frac{X_{ji_k} D_{i_k}^{(\ell)}}{\pi_{i_k}} - \frac{X_{ji_k}(1 - D_{i_k}^{(\ell)})}{1 - \pi_{i_k}} \right)^2 .$$

Table 1 reports the ratio $\frac{\text{ASMD}_j^{\ell}}{\text{ASMD}_j^{Naive}}$ for each treatment assignment mechanism. It is easy to see the advantages of the Cube method. The Cube method outperforms stratification for both datasets and for every single variable. First, we see that as stratification cannot perform balance over continuous variables without discretizing them, the Cube method is much

Table 1: ASMD ratio

	Naive	Strata Baseline	Cube Baseline	Strata All	Cube All
	(1)	(2)	(3)	(4)	(5)
		Lee et al. (2021): $n' = 808$			
Baseline Remittances‡	1.000	0.569	0.013	0.567	0.012
Household Size†	1.000	1.032	0.984	0.592	0.009
Age†	1.000	0.968	1.001	0.581	0.007
Female	1.000	0.940	0.943	0.315	0.004
Primary Education	1.000	1.013	0.975	0.465	0.003
		Gerber et al. (2020): $n' = 6650$			
Baseline Beliefs‡	1.000	0.299	0.001	0.304	0.001
Past Votes†	1.000	1.027	1.003	0.325	0.001
Interest in Politics†	1.000	0.942	0.955	0.390	0.001
Schooling†	1.000	1.002	0.996	0.296	0.001
Identifies Pelosi	1.000	1.016	1.000	0.167	0.001
Male	1.000	1.035	1.008	0.092	0.000

This table shows the ratio $\text{ASMD}_j^{\ell} \,/\, \text{ASMD}_j^{Naive}$ for every mechanism. Column 1 is the reference mechanism: Naive randomization. Column 2 corresponds to stratified randomization on the pre-treatment outcome quartiles. Column 3 shows balance when using the Cube Method only on the pre-treatment outcome. Column 4 presents the results for stratified randomization on all covariates. Column 5 shows the AMSD ratio when using the Cube method to balance all covariates.

‡: When used in stratification (Columns 2 and 4), discretized by quartiles.

†: When used in stratification (Column 4), discretized by the median.

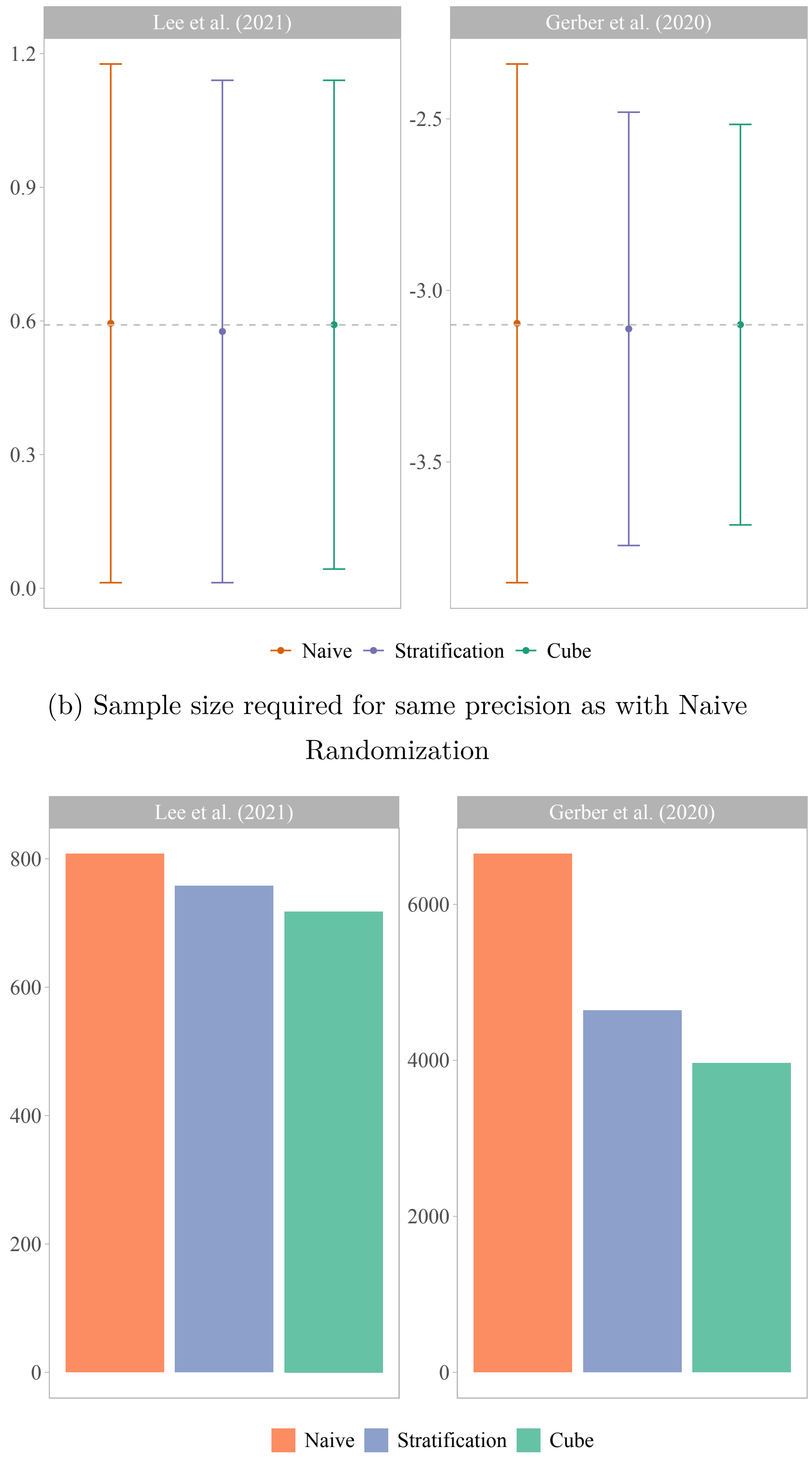
more efficient in reducing these differences. Nonetheless, the Cube method is also stronger in providing balanced dummy variables, such as gender. When balancing several covariates, approximations done with stratification are more frequent than with the Cube method, generating higher imbalances. More precisely, if we seek to balance the pre-treatment outcome only, stratification by quartiles reduces the ASMD by 43% and 77%, and the Cube method reduces the AMSD by 99% (Columns 2 and 3, respectively). If we use stratified randomization with all variables, the balance is reduced between 53% and 91% for the binary variables and only between 3% and 58% for the other ones (Column 4). It is not thus possible to anticipate the level of balance we obtain using stratification, especially if the empiricist wants to balance over multiple covariates. Finally, Column 5 shows that the Cube achieves its goal completely by reducing the AMSD by 99% or 100%, depending on the variable.

6.2 Precision gains

We now present the precision gains from using the Cube method. As explained in Section 5, gains in precision are higher when balancing covariates are correlated to potential outcomes. We hereon note $\widehat{R}_1^2$ and $\widehat{R}_0^2$, the R^2s associated with the regressions of the treated and control units, respectively. When regressing on all the covariates listed in Table 1, we get $(\widehat{R}_1^2 + \widehat{R}_0^2)/2 = 0.12$ for Lee et al. (2021) and $(\widehat{R}_1^2 + \widehat{R}_0^2)/2 = 0.41$ for Gerber et al. (2020). The expressions for asymptotic variances using Poisson randomization or the Cube method in Proposition 2 entail that for $\pi_i = 1/2$ the percentage reduction in variance can be closely estimated by $(\widehat{R}_1^2 + \widehat{R}_0^2)/2$. Precision gains are, therefore, much larger for Gerber et al. (2020). However, precision gains for Lee et al. (2021) are not negligible.

Figure 4 show the advantages of using the Cube method for both experiments. For each panel, we compare here three different methods: (i) Naive randomization with fixed-sample size, (ii) Stratified randomization on all covariates listed in Table 1, and (iii) Randomization with the Cube method on all covariates. Panel (a) shows precision in terms of 95% confidence interval sizes. We see that the Cube method gives smaller confidence intervals for both experiments. However, gains are more substantial for Gerber et al. (2020). Panel b shows benefits in sample size: how many units should participate in the experiment to obtain the same results as for Naive randomization. We see again that the Cube method is the better-performing mechanism: one would need 718 units instead of 808 for Lee et al. (2021) and 3966 instead of 6650 for Gerber et al. (2020). These gains in the number of units imply a substantial decrease in experimental costs. We see that, even if the covariates are not very explicative of potential outcomes, there are still gains from using the Cube method. Moreover, this mechanism always performs better than stratification, meaning that if one has data available before randomizing, one should consider using the Cube method.

Figure 4: Precision gains from the Cube method



(a) 95% confidence intervals

(b) Sample size required for same precision as with Naive Randomization

We now show the importance of covariate selection when using the Cube method since it is crucial in determining the precision gains. Table 2 presents the win in precision for different balancing constraints. We propose here seven mechanisms, where we always balance at least on the treatment group size: (1) No balancing covariates apart from treatment group size– i.e., naive randomization – (2) Balancing on the pre-treatment outcome and ten other covariates, (3) Balancing only on the pre-treatment outcome, (4) Balancing on the first two moments of the pre-treatment outcome, (5) Balancing on four dummies defined by the quartiles of the pre-treatment outcome – i.e., stratification –, (6) Balancing on the ten covariates, and (7) balancing on a random variable.

For each mechanism, we report the explanatory power of the balancing covariates in the form of $\widehat{R}_1^2$ and $\widehat{R}_0^2$, and precision gains in terms of percentage variance reduction for the estimator using mechanism ℓ: $\Delta\%\,\mathrm{Var}^{(\ell)} = \frac{\widehat{V}^{(\ell)} - \widehat{V}.^{Naive}}{\widehat{V}^{Naive}}$, with $\widehat{V}$ defined as in (14), percentage reduction of confidence interval size $\Delta\%\,\mathrm{CI}^{(\ell)} = \frac{\Delta\%\,\mathrm{Var}^{(\ell)}}{1+0.1\sqrt{100+\Delta\%\,\mathrm{Var}^{(\ell)}}}$ and the effective sample size – i.e., by how much could we decrease n to obtain the same precision using than naive randomization.

We indeed find that precision gains are related to the explanatory power of the covariates used for balancing. For Lee et al. (2021), we see that the best option is to balance the baseline and all the covariates. If the empiricist does so, she would reduce the variance by 11%, whereas she would only reduce it by 6% using the baseline only or 8% using all the covariates. Here, the baseline has limited explanatory power, so she should include as much auxiliary information as possible. With the simulated data from Gerber et al. (2020), gains are much higher. Notably, using the Cube method to balance the baseline outcome and ten other covariates reduces the variance by 41%. If using naive randomization, it would be necessary to increase the sample size by 68% to obtain the same precision. We see that, however, most of this power gain comes from balancing the pre-treatment outcome. Indeed, mechanisms (2) and (3), which do not balance on other covariates, give similar, though smaller, precision gains. However, using the Cube method provides advantages compared with stratification since the baseline outcome is continuous. Indeed if we balance using quartiles, wins in variance and sample size are almost 25% and 40% smaller, respectively, than when balancing using the continuous one. In this case, covariates other than the baseline outcome explain very little of observed outcome variables, so the gains when using them to balance are small. Moreover, we check that balancing on a random variable, which is thus not correlated to potential outcomes, is practically equivalent to using naive randomization.

Table 2: Balancing covariates and precision gains

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Balance on:							
Treatment group size	✓	✓	✓	✓	✓	✓	✓
Baseline outcome (continuous)		✓	✓	✓			
Other covariates		✓				✓	
Squared baseline outcome				✓			
Baseline outcome (quartiles)					✓		
Random variable							✓
Lee et al. (2021)							
Explanatory power:							
$\widehat{R}_1^2$	0.000	0.117	0.050	0.063	0.054	0.091	0.002
$\widehat{R}_0^2$	0.000	0.113	0.061	0.073	0.059	0.077	0.002
Precision gains:							
Δ% Var	0.000	-11.201	-5.968	-7.041	-5.661	-8.212	-0.648
Δ% CI size	0.000	-5.767	-3.030	-3.585	-2.872	-4.194	-0.325
Effective sample size	808	718	760	752	764	742	804
Gerber et al. (2020)							
Explanatory power:							
$\widehat{R}_1^2$	0.000	0.347	0.343	0.351	0.268	0.024	0.000
$\widehat{R}_0^2$	0.000	0.464	0.462	0.462	0.333	0.026	0.000
Precision gains:							
Δ% Var	0.000	-40.375	-40.172	-40.454	-30.023	-2.416	-0.052
Δ% CI size	0.000	-22.783	-22.652	-22.834	-16.347	-1.216	-0.026
Effective sample size	6650	3966	3980	3960	4654	6490	6648

This table shows the relation between covariate selection and precision gains. The first panel shows the set of selected variables for each column. The first column shows the benchmark case of naive randomization with a fixed sample size. The six mechanisms remaining use the Cube method to balance additional variables. For each experimental dataset, we check the explanatory level of selected variables on the potential outcomes and the gains in variance reduction, confidence interval size reduction, and sample size. The latter refers to the sample size needed to obtain the same precision as in naive randomization.

Figure 5 shows how these gains in precision change with the sample size. For this, we take $n' \in \{50, 100, 200, 500, 808, 1000, 2000, 6650\}$ and check how the coverage rate of our confidence intervals and the rejection probability change with n'. We first see that the coverage rate of the Cube method approaches 95% as n' increases. For sample sizes of 50, 100, and 200, coverage rates are more liberal for the Cube method than for naive randomization. For $n' = 200$, our confidence intervals defined in (16) are very close to the exact coverage rate. The test of the power is, furthermore, higher for the Cube method than for naive randomization, whatever the sample size. The much weaker precision in small samples undermines the better coverage rate in naive randomization. The Cube method generates estimates that are more precise, thus reducing the probability of obtaining a false positive. This difference is even present for $n' = 1000$. For Gerber et al. (2020), with the Cube algorithm, we reject the null with 98% chance, whereas there is 88% likeliness of doing so if we randomize naively.

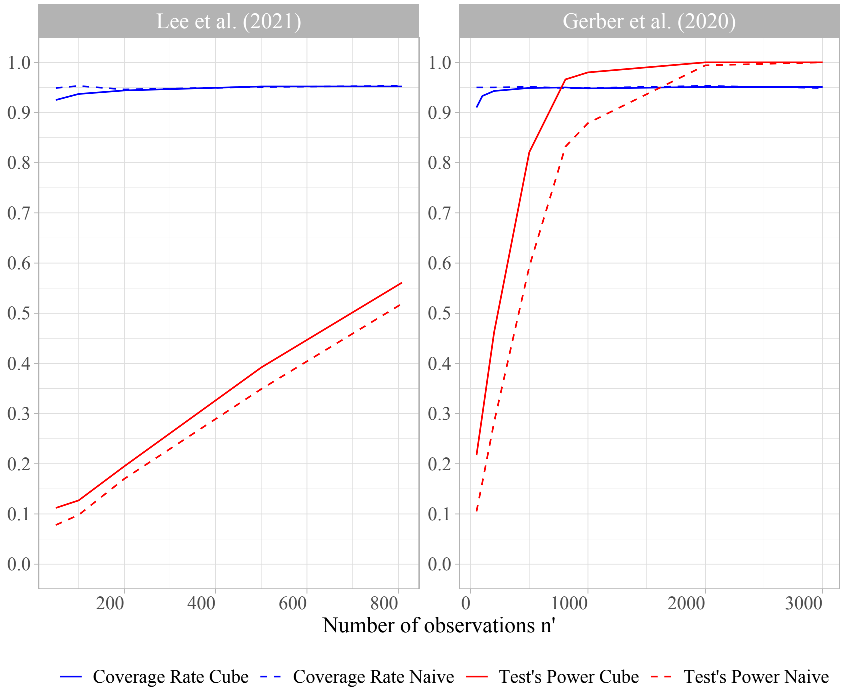


Figure 5: Sample size and Confidence Intervals

n'	50	100	200	500	808	1000	2000	6650
				Lee et al. (2021)				
Coverage rate:								
Naive	0.949	0.953	0.946	0.951	0.953	–	–	–
Cube	0.925	0.937	0.944	0.952	0.952	–	–	–
Power of the test:								
Naive	0.078	0.098	0.170	0.349	0.519	–	–	–
Cube	0.112	0.127	0.195	0.392	0.561	–	–	–
				Gerber et al. (2020)				
Coverage rate:								
Naive	0.950	0.950	0.950	0.951	0.949	0.949	0.953	0.949
Cube	0.910	0.933	0.943	0.949	0.950	0.948	0.951	0.951
Power of the test:								
Naive	0.105	0.163	0.283	0.593	0.832	0.879	0.994	1.000
Cube	0.217	0.298	0.462	0.821	0.966	0.980	1.000	1.000

7 Conclusion

The Cube method, first introduced by Deville and Tillé (2004) outperforms most common methods used in experimental settings. We here provide a set of results formalizing these gains and allowing us to implement the Cube method in the RCT context. We tackle common issues empiricists face, such as balance, inference, sample size, and precision. Our analytical results and simulation show that if an empiricist has data available beforehand, she should always use the Cube method to balance characteristics between the treatment and control groups, especially if she believes them to be rather explanatory of the outcome of interest.

Without additional costs, when compared to covariate-adaptive mechanisms, the Cube method allows to have better balance and precision. Mechanically, all differences between the treatment and control groups disappear, making balancing tests unnecessary and discarding publication bias and p-hacking they produce. This reduction generates more precise estimates and significantly reduces the sample size needed for a minimum detectable effect.

Several questions remain unanswered and should be the focus of further research on this method. Notably, there are the questions of clustered assignment, multiple treatment arms, and attrition. These questions can be treated easily by generalizing the algorithm, but some work is needed to clarify the asymptotics in these cases.

References

Aaronson, J., Burton, R., Dehling, H., Gilat, D., Hill, T., and Weiss, B. (1996). Strong Laws for L- and U- Statistics. *Research Scholars in Residence*, 348.

Athey, S. and Imbens, G. W. (2017). Chapter 3 - The Econometrics of Randomized Experimentsa. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Economic Field Experiments*, volume 1 of *Handbook of Field Experiments*, pages 73–140. North-Holland.

Bai, Y. (2022). Optimality of Matched-Pair Designs in Randomized Controlled Trials. *American Economic Review*, 112(12):3911–3940.

Bai, Y., Romano, J. P., and Shaikh, A. M. (2022). Inference in Experiments With Matched Pairs. *Journal of the American Statistical Association*, 117(540):1726–1737. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.1883437.

Ball, S., Bogatz, G. A., Rubin, D. B., and Beaton, A. E. (1973). Reading with Television: An Evaluation of the Electric Company. A Report to the Children's Television Workshop. Volumes 1 and 2. Technical Report PR-73-02, ETS Program Report.

Basse, G. W., Feller, A., and Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494.

Bruhn, M. and McKenzie, D. (2009). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, 1(4):200–232.

Chen, J. and Rao, J. N. K. (2007). Asymptotic Normality Under Two-Phase Sampling Designs. *Statistica Sinica*, 17(3):1047–1064. Publisher: Institute of Statistical Science, Academia Sinica.

Cytrynbaum, M. (2022). Designing Representative and Balanced Experiments by Local Randomization. page 101.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4):893–912.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128(2):569–591.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33:503–515. Publisher: Ministry of Agriculture and Fisheries.

Fisher, S. R. A. (1935). *The Design of Experiments*. Oliver and Boyd.

Gerber, A., Hoffman, M., Morgan, J., and Raymond, C. (2020). One in a Million: Field Experiments on Perceived Closeness of the Election and Voter Turnout. *American Economic Journal: Applied Economics*, 12(3):287–325.

Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals: Doubly balanced spatial sampling. *Environmetrics*, 24(2):120–131.

Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics (Oxford, England)*, 5(2):263–275.

Gut, A. (2013). *Probability: A Graduate Course*, volume 75 of *Springer Texts in Statistics*. Springer New York, New York, NY.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., and Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics*, 1(1):1–46. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE8.

Heckman, J. J., Pinto, R., Shaikh, A. M., and Yavitz, A. (2011). Inference with Imperfect Randomization: The Case of the Perry Preschool Program.

Higgins, M. J., Sävje, F., and Sekhon, J. S. (2016). Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences*, 113(27):7369–7376. Publisher: Proceedings of the National Academy of Sciences.

Imai, K., King, G., and Nall, C. (2009). The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*, 24(1).

Imbens, G. W. (2011). Experimental design for unit and cluster randomid trials. Technical report, Harvard University.

Lee, J. N., Morduch, J., Ravindran, S., Shonchoy, A., and Zaman, H. (2021). Poverty and Migration in the Digital Age: Experimental Evidence on Mobile Banking in Bangladesh. *American Economic Journal: Applied Economics*, 13(1):38–71.

Lee, S. and Shaikh, A. M. (2014). Multiple Testing and Heterogeneous Treatment Effects: Re-Evaluating the Effect of Progresa on School Enrollment. *Journal of Applied Econometrics*, 29(4):612–626. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.2327.

Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162. Publisher: Proceedings of the National Academy of Sciences.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2). arXiv:1207.5625 [math, stat].

Mutz, D. C., Pemantle, R., and Pham, P. (2019). The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data. *The American Statistician*, 73(1):32–42. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2017.1322143.

Snyder, C. and Zhuo, R. (2018). Sniff Tests as a Screen in the Publication Process: Throwing out the Wheat with the Chaff. Technical Report w25058, National Bureau of Economic Research, Cambridge, MA.

Student (1938). Comparison Between Balanced and Random Arrangements of Field Plots. *Biometrika*, 29(3/4):363–378. Publisher: [Oxford University Press, Biometrika Trust].

Takacs, L. (1991). A Moment Convergence Theorem. *The American Mathematical Monthly*, 98(8):742–746. Publisher: Mathematical Association of America.

Tillé, Y. and Favre, A.-C. (2004). Coordination, Combination and Extension of Balanced Samples. *Biometrika*, 91(4):913–927. Publisher: [Oxford University Press, Biometrika Trust].

A Proofs of Propositions

A.1 Proof of Balancing Approximations for the Cube Method (Proposition 1)

From Assumption 2 and Proposition 4 in Deville and Tillé (2004), we have:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ji}D_i}{\pi_i}-\frac{1}{n}\sum_{i=1}^{n}X_{ji}\right|\leq\frac{q}{n}\max_{i=1,\dots,n}\left|\frac{X_{ji}}{\pi_i}\right|\leq\frac{q\max_{i=1,\dots,n}|X_{ji}|}{cn}.$$

If Assumptions 1 holds and if moment of order r exists for X_{j1}, from Proposition 1.5 and Theorem 2.1 in Chapter 6 in Gut (2013), we have $\max_{i=1,\dots,n}|X_{ji}|=o_p(n^{1/r})$. If X_{ji} sub-Gaussian $\max_{i=1,\dots,n}|X_{ji}|=O_p\left(\sqrt{\ln(n)}\right)$ and if X_{ji} is bounded by K, $\max_{i=1,\dots,n}|X_{ji}|\leq K$.

A.2 Proof of Asymptotic Normality for the SATE (First part of Propositions 2 and 3)

We want to prove

$$\sqrt{n}\left(\widehat{\theta}_{HT}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,V_0\right)$$

and

$$\sqrt{n}\left(\widehat{\theta}_{H}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,V_0\right)$$

where $V_0=\mathbb{E}\left[\pi_i\left(1-\pi_i\right)\left(\frac{\varepsilon_i(1)}{\pi_i}+\frac{\varepsilon_i(0)}{1-\pi_i}\right)^2\right]$.

We first show that it is sufficient to prove asymptotic normality for one of the two estimators. Proposition 1 ensures that if the empiricist includes a constant in the set of covariates to balance, one has $\sum_{i=1}^{n}\frac{D_i}{\pi_i}=n+o_p\left(\frac{1}{\sqrt{n}}\right)$ and $\sum_{i=1}^{n}\frac{1-D_i}{1-\pi_i}=n+o_p\left(\frac{1}{\sqrt{n}}\right)$. Then, the Hajec estimator in (6) is given by

$$\begin{aligned}\widehat{\theta}_H&=\frac{1}{n+o_p\left(\frac{1}{\sqrt{n}}\right)}\left(\sum_{i=1}^{n}\frac{Y_iD_i}{\pi_i}-\frac{Y_i(1-D_i)}{1-\pi_i}\right)\\&=\frac{n}{n+o_p\left(\frac{1}{\sqrt{n}}\right)}\widehat{\theta}_{HT}\\&=\left(1+o_p\left(\frac{1}{\sqrt{n}}\right)\right)\widehat{\theta}_{HT}.\end{aligned}$$

Then,

$$\sqrt{n}\left(\widehat{\theta}_H-\theta_0\right)=\sqrt{n}\left(\widehat{\theta}_{HT}-\theta_0\right)+o_p\left(1\right)\left(\widehat{\theta}_{HT}-\theta_0\right).$$

By Slutsky's theorem, if $\sqrt{n}\left(\widehat{\theta}_{HT}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,V_0\right)$, then $\sqrt{n}\left(\widehat{\theta}_H-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,V_0\right)$. It is thus sufficient to prove asymptotic normality of the Horvitz-Thompson estimator, i.e.,

$$\sqrt{n}\left(\widehat{\theta}_{HT}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,V_0\right).$$

Under Assumptions 1 and 2, Proposition 1 ensures

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_{HT}-\theta_0\right)=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i-\pi_i)\left(\frac{\varepsilon_i(1)}{\pi_i}+\frac{\varepsilon_i(0)}{1-\pi_i}\right)+\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{D_iX_i'}{\pi_i}-\sum_{i=1}^{n}X_i'\right)\beta_1\\&-\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)X_i'}{1-\pi_i}-\sum_{i=1}^{n}X_i'\right)\beta_0\\=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}-\left(\frac{\varepsilon_i(1)}{\pi_i}+\frac{\varepsilon_i(0)}{1-\pi_i}\right)\pi_i+\left(\frac{\varepsilon_i(1)}{\pi_i}+\frac{\varepsilon_i(0)}{1-\pi_i}\right)D_i+o_p(1)\end{aligned}$$

Then, we have

$$\sqrt{n}\left(\widehat{\theta}_{HT}-\theta_0\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_i+g_iD_i+o_p(1)$$

with $f_i:=f\left(X_i,\varepsilon_i(0),\varepsilon_i(1)\right)=-\left(\frac{\varepsilon_i(1)}{\pi_i}+\frac{\varepsilon_i(0)}{1-\pi_i}\right)\pi_i$ and $g\left(X_i,\varepsilon_i(0),\varepsilon_i(1)\right)=\frac{\varepsilon_i(1)}{\pi_i}+\frac{\varepsilon_i(0)}{1-\pi_i}$. Slutsky's theorem ensures that we have to prove

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_i+g_iD_i\xrightarrow{d}\mathcal{N}\left(0,V_0\right).\tag{A.1}$$

By Assumption 3 $\mathbb{E}[f|X]=\mathbb{E}[g|X]=0$. Then, Conjecture 1 and Lemma 2 give that, conditional on $(X_i)_{i\geq 1}$, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_i+g_iD_i\xrightarrow{d}\mathcal{N}\left(0,V_0\right)$, with $V_0=\mathbb{E}[f_i^2+(2g_if_i+g_i^2)\pi_i]=\mathbb{E}\left[\pi_i(1-\pi_i)\left(\frac{\varepsilon_i(1)}{\pi_i}+\frac{\varepsilon_i(0)}{1-\pi_i}\right)\right]$, in the sense of Definition 2. Notice that V_0 does not depend on $(X_1)_{i\geq 1}$, so convergence in distribution is unconditional. This concludes the proof.

A.3 Proof of Asymptotic Normality for the PATE (Second part of Propositions 2 and 3)

As shown in Proof A.2, if $\sqrt{n}\left(\widehat{\theta}_{HT}-\theta_0^*\right)\xrightarrow{d}\mathcal{N}\left(0,V_0^*\right)$, we have $\sqrt{n}\left(\widehat{\theta}_H-\theta_0^*\right)\xrightarrow{d}\mathcal{N}\left(0,V_0^*\right)$, we thus restrict ourselves to proving asymptotic normality for the Horvitz-Thompson estimator, i.e,

$$\sqrt{n}\left(\widehat{\theta}_{HT}-\theta_0^*\right)\xrightarrow{d}\mathcal{N}\left(0,V_0^*\right)$$

where $V_0^* = (\beta_1 - \beta_0)'\mathbb{V}(X)(\beta_1 - \beta_0) + \mathbb{E}\left[\frac{\varepsilon_i(1)^2}{\pi_i}\right] + \mathbb{E}\left[\frac{\varepsilon_i(0)^2}{1-\pi_i}\right]$.

Let us consider $f_i \coloneqq f(X_i, \varepsilon_i(1), \varepsilon_i(0)) = -\frac{\varepsilon_i(0)}{1-\pi_i}$, $g_i \coloneqq g(X_i, \varepsilon_i(1), \varepsilon_i(0)) = \frac{\varepsilon_i(1)}{\pi_i} + \frac{\varepsilon_i(0)}{1-\pi_i}$, and $h_i \coloneqq h(X_i) = (X_i - \mathbb{E}[X_i])'(\beta_1 - \beta_0)$.
= Under Assumptions 1-2, Proposition 1 ensures

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_{HT} - \theta_0^*\right) =& \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{X_i'D_i}{\pi_i} - E[X_i]'\right)\beta_1 \\ &- \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{X_i'(1-D_i)}{1-\pi_i} - E[X_i]'\right)\beta_0 \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\varepsilon_i(1)D_i}{\pi_i} - \frac{\varepsilon_i(0)(1-D_i)}{1-\pi_i} \\ =& \frac{1}{\sqrt{n}}\sum_{i=1}^n f_i + g_iD_i + \frac{1}{\sqrt{n}}\sum_{i=1}^n h_i + o_p(1).\end{aligned}$$

Slutsky's theorem ensures that we have to prove

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n f_i + g_iD_i + \frac{1}{\sqrt{n}}\sum_{i=1}^n h_i \xrightarrow{d} \mathcal{N}\left(0, V_0^*\right). \tag{A.2}$$

By Assumption 3 $\mathbb{E}[f|X] = \mathbb{E}[g|X] = 0$. Then, Conjecture 1 and Lemma 2 give that, conditional on $(X_i)_{i\geq 1}$, $\sqrt{n}\left(\widehat{\theta}_{HT} - \theta_0^*\right) \xrightarrow{d} \mathcal{N}\left(0, V_{01}^*\right)$ with $V_{01}^* = \mathbb{E}[f_i^2 + (2g_if_i + g_i^2)\pi_i] = \mathbb{E}\left[\frac{\varepsilon_i(1)^2}{\pi_i}\right] + \mathbb{E}\left[\frac{\varepsilon_i(0)^2}{1-\pi_i}\right]$. Moreover, by the central limit theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^n h_i \xrightarrow{d} \mathcal{N}\left(0, V_{02}^*\right)$ with $V_{02}^* = (\beta_1 - \beta_0)'\mathbb{V}(X)(\beta_1 - \beta_0)$. Theorem 2 in Chen and Rao (2007) ensures that $\frac{1}{\sqrt{n}}\sum_{i=1}^n f_i + g_iD_i + \frac{1}{\sqrt{n}}\sum_{i=1}^n h_i \xrightarrow{d} \mathcal{N}\left(0, V_{01}^* + V_{02}^*\right)$. This concludes the proof.

A.4 Proof of Randomized-based Inference (Proposition 4)

For completeness, we show first, as in Bai et al. (2022), that the strong null hypothesis (17) $(Y_i(1), X_i) \overset{d}{=} (Y_i(0), X_i)$ is equivalent to stating $Y_1, \ldots, Y_n \perp\!\!\!\perp D_1, \ldots, D_n | X_1, \ldots, X_n$.
Let us consider random allocations generated by the Cube Method d and d' in the support of $D_1, \ldots, D_n | X_1, \ldots, X_n$ and any set A. Then we have,

$$\begin{aligned}&\mathbb{P}\left[(Y_1, \ldots, Y_n) \in A | (D_1, \ldots, D_n) = (d_1, \ldots, d_n), X_1, \ldots, X_n)\right] \\ =&\mathbb{P}\left[(Y_1(d_1), \ldots, Y_n(d_n)) \in A | (D_1, \ldots, D_n) = (d_1, \ldots, d_n), X_1, \ldots, X_n)\right] \\ =&\mathbb{P}\left[(Y_1(d_1), \ldots, Y_n(d_n)) \in A | (D_1, \ldots, D_n) = (d_1, \ldots, d_n), X_1, \ldots, X_n)\right] \\ =&\mathbb{P}\left[(Y_1(d_1), \ldots, Y_n(d_n)) \in A | X_1, \ldots, X_n)\right] \\ =&\mathbb{P}\left[(Y_1(d_1'), \ldots, Y_n(d_n')) \in A | X_1, \ldots, X_n)\right]\end{aligned}$$

$$=\mathbb{P}\left[(Y_1(d_1'),\ldots,Y_n(d_n'))\in A|(D_1,\ldots,D_n)=(d_1',\ldots,d_n'),X_1,\ldots,X_n)\right]$$
$$=\mathbb{P}\left[(Y_1,\ldots,Y_n)\in A|(D_1,\ldots,D_n)=(d_1',\ldots,d_n'),X_1,\ldots,X_n)\right],$$

so both hypothesis are equivalent. Then, under Assumptions 1 and 2, and the strong null hypothesis 17,

$$(Y_i,D_i,X_i)_{i\geq 1}\overset{d}{=}(Y_i,D_i^{(g)},X_i)_{i\geq 1}.$$

We thus have

$$\begin{aligned}\mathbb{E}\left[\sum_{g\in G_n^B}\phi_n^{rand}\left(\mathbf{P_n^{(g)}}\right)\right] &= \sum_{g\in G_n^B}\mathbb{E}\left[\mathbb{E}\left[\phi_n^{rand}\left(\mathbf{P_n^{(g)}}\right)|X_1,\ldots,X_n\right]\right]\\ &= \sum_{g\in G_n^B}\mathbb{E}\left[\mathbb{E}\left[\phi_n^{rand}\left(\mathbf{P_n}\right)|X_1,\ldots,X_n\right]\right]\\ &= B\mathbb{E}\left[\phi_n^{rand}\left(\mathbf{P_n}\right)\right] \qquad \text{(A.3)}\end{aligned}$$

Moreover, $c_n\left(\mathbf{P_n},1-\alpha\right)=c_n\left(\mathbf{P_n^{(g)}},1-\alpha\right)$ for any $g\in G_n^B$ ensures by definition of $c_n\left(\mathbf{P_n},1-\alpha\right)$

$$\sum_{g\in G_n^B}\phi_n^{rand}\left(\mathbf{P_n^{(g)}}\right)\leq B\alpha \qquad \text{(A.4)}$$

Combining equations (A.3) and (A.4) we get $\mathbb{E}\left[\phi_n^{rand}\left(\mathbf{P_n}\right)\right]\leq\alpha$, which concludes the proof.

B Lemmas for the Cube Method

Lemma 1 (Exchangeability)
For any permutation σ of $\{1,...,n\}$ we have:

$$(D_{\sigma(i)},\pi^*_{\sigma(i)}X_{\sigma(i)})_{i=1,\ldots,n}\overset{d}{=}(D_i,X_i)_{i=1,\ldots,n}$$

Proof:
For any value of n, the Cube algorithm ensures there exists a finite collection of independent uniform random variables $(U_1,...,U_K)$ independent of $(X_1,...,X_n)$ such that $(D_1,...,D_n,\pi_1^*,...,\pi_n^*)=f(X_1,...,X_n,U_1,...,U_K)$. Because the X are iid and independent of the U, we have:

$$(X_{\sigma(1)},...,X_{\sigma(n)},U_1,...,U_K)\overset{d}{=}(X_1,...,X_n,U_1,...,U_K).$$

The result follows.

Definition 2
$W_n \xrightarrow{d} \mathcal{N}(0,\sigma^2)$ conditional on $(X_i)_{i\geq 1}$ if and only if for any h bounded Lipschitz $\mathbb{E}\left(h(W_n)|(X_i)_{i\geq 1}\right)$ converges almost surely to $\int h(u)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{u^2}{2\sigma^2}\right)du$.

Usual criteria (Portemanteau Lemma, Levy continuity theorem, ...) to prove convergence in distribution could be adapted to prove the convergence in distribution conditional on $(X_i)_{i\geq 1}$ apply if the usual expectations and probabilities are replaced by conditional expectations and probabilities and usual convergence of sequences is replaced by almost sure convergence of random variables. More concretely, we will use the fact that if for any $k \geq 1$, $\mathbb{E}((W_n)^k|(X_i)_{i\geq 1})$ converges almost surely to the kth-raw moment of a Gaussian distribution of variance σ^2 then $W_n \xrightarrow{d} \mathcal{N}(0,\sigma^2)$ conditional on $(X_i)_{i\geq 1}$. This is an adaptation of the theorem of Takacs (1991) that states that if for any $k \geq 1$, $\mathbb{E}((W_n)^k)$ converges to the kth-raw moment of a Gaussian distribution of variance σ^2 then $W_n \xrightarrow{d} \mathcal{N}(0,\sigma^2)$. Moreover, $W_n \xrightarrow{d} \mathcal{N}(0,\sigma^2)$ conditional on $(X_i)_{i\geq 1}$ if and only if $\forall t \in \mathbb{R}$, $\mathbb{P}\left(W_n \leq t|(X_i)_{i\geq 1}\right)$ converges almost surely to $\Phi(\frac{t}{\sqrt{\sigma^2}})$ for Φ the cdf of the standard Gaussian.

Lemma 2 (Asymptotic normality)
Let f and g be two functions such that for $f_i = f(\delta_i(1),\delta_i(0),X_i)$ and $g_i = g(\delta_i(1),\delta_i(0),X_i)$ we have $\mathbb{E}(f_i^2+g_i^2) < \infty$ and $\mathbb{E}[f_i|X_i] = \mathbb{E}[g_i|X_i] = 0$.
If Assumptions 1 and 2 and Conjecture 1 hold. Then, conditional on $(X_i)_{i\geq 1}$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_i + g_i D_i \xrightarrow{d} \mathcal{N}(0,V_0) \tag{B.1}$$

with $V_0 = \mathbb{E}\left[f_1^2 + (2g_1f_1 + g_1^2)\pi_1\right]$.

Proof:

First step: $|f_i|+|g_i|$ bounded implies (B.1)
Let us assume it exists $K>0$ such that $|f_1|+|g_1| < K$ for any $k \in \mathbb{N}$. This ensures that all the moments of $f_i + g_iD_i$ exist. Let $M_{n,k} = \mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n f_i + g_iD_i\right)^k|(X_i)_{i\geq 1}\right]$.
We have

$$M_{n,k} = \mathbb{E}\left[n^{-k/2}\sum_{1\leq i_1,\ldots,i_k\leq n}\prod_{\ell=1}^{k}\left(f_{i_\ell}+g_{i_\ell}D_{i_\ell}\right)|(X_i)_{i\geq 1}\right].$$

Let us order the indices $i_1,\ldots,i_k$ as $j_1,\ldots,j_m$ for some $1 \leq m \leq k$ with each j_ℓ occurring with multiplicity a_ℓ. Let $A_{k,m} := \{a = (a_1,\ldots,a_m) \in \mathbb{N}^{*m} : \sum_{\ell=1}^m a_\ell = k\}$ and for $a \in A_{k,m}$, $c_{k,a} = \frac{k!}{\prod_{\ell=1}^m a_\ell!}$. We have:

$$M_{n,k} = \sum_{m=1}^{k} n^{-k/2}\sum_{1\leq j_1<\ldots<j_m\leq n}\sum_{a\in A_{k,m}} c_{k,a}\mathbb{E}\left[\prod_{\ell=1}^{k}\left(f_{j_\ell}+g_{j_\ell}D_{j_\ell}\right)^{a_\ell}|(X_i)_{i\geq 1}\right].$$

In order to prove the convergence of moments, we will focus on the summands

$$B_{n,k,m} = n^{-k/2} \sum_{1 \le j_1 < \ldots < j_m \le n} \sum_{a \in A_{k,m}} c_{k,a} \mathbb{E}\left[\prod_{\ell=1}^{k} (f_{j_\ell} + g_{j_\ell} D_{j_\ell})^{a_\ell} | (X_i)_{i \ge 1} \right].$$

Notice that $|B_{n,k,m}| \le n^{-k/2}\binom{n}{m} \sum_{m=1}^{k} \sum_{a \in A_{k,m}} c_{k,a} K^k = O\left(n^{m-k/2}\right)$. For $m < k/2$, we thus have $\lim_n B_{n,k,m} = 0$.

We focus now in the case $m > k/2$. For $\mathcal{K} \subseteq \{1, \ldots, m\}$, we note $\mathcal{K}^c = \{1, \ldots, m\} \setminus \mathcal{K}$. Then, the binomial theorem and the expansion $\prod_{\ell=1}^{m}(x_\ell + y_\ell) = \sum_{\mathcal{K} \subseteq \{1,\ldots,m\}} \prod_{\ell \in \mathcal{K}} x_\ell \prod_{\ell' \in \mathcal{K}^c} y_{\ell'}$ and identity $D^a = D$ for $a \ge 1$ ensure

$$\begin{aligned}
&B_{n,k,m} \\
&= n^{-k/2} \sum_{1 \le j_1 < \ldots < j_m \le n} \sum_{a \in A_{k,m}} c_{k,a} \mathbb{E}\left[\prod_{\ell=1}^{k} (f_{j_\ell} + g_{j_\ell} D_{j_\ell})^{a_\ell} \middle| (X_i)_{i \ge 1} \right] \\
&= n^{-k/2} \sum_{1 \le j_1 < \ldots < j_m \le n} \sum_{a \in A_{k,m}} c_{k,a} \mathbb{E}\left[\prod_{\ell=1}^{m} \left[f_{j_\ell}^{a_\ell} + \left(\sum_{r=1}^{a_\ell} \binom{a_\ell}{r} f_{j_\ell}^{a_\ell - r} g_{j_\ell}^{r} \right) D_{j_\ell} \right] \middle| (X_i)_{i \ge 1} \right] \\
&= n^{-k/2} \sum_{1 \le j_1 < \ldots < j_m \le n} \sum_{a \in A_{k,m}} c_{k,a} \sum_{\mathcal{K} \subseteq \{1,\ldots,m\}} \mathbb{E}\left[\prod_{\ell \in \mathcal{K}} f_{j_\ell}^{a_\ell} \prod_{\ell' \in \mathcal{K}^c} \left(\sum_{r=1}^{a_{\ell'}} \binom{a_{\ell'}}{r} f_{j_{\ell'}}^{a_{\ell'} - r} g_{j_{\ell'}}^{j} \right) \prod_{\ell'' \in \mathcal{K}^c} D_{j_{\ell''}} \middle| (X_i)_{i \ge 1} \right]
\end{aligned}$$

Then, independence of $(f_i, g_i)_{i \ge 1}$ across i and conditional independence $(f_i, g_i) \perp\!\!\!\perp D_i | (X_{i'})_{i' \ge 1}$ ensure

$$\begin{aligned}
&B_{n,k,m} \\
&= n^{-k/2} \sum_{1 \le j_1 < \ldots < j_m \le n} \sum_{a \in A_{k,m}} c_{k,a} \sum_{\mathcal{K} \subseteq \{1,\ldots,m\}} \prod_{\ell \in \mathcal{K}} \mathbb{E}\left[f_{j_\ell}^{a_\ell} | X_{j_\ell} \right] \prod_{\ell' \in \mathcal{K}^c} \left(\sum_{r=1}^{a_{\ell'}} \binom{a_{\ell'}}{r} \mathbb{E}\left[f_{j_{\ell'}}^{a_{\ell'} - r} g_{j_{\ell'}}^{r} | X_{j_{\ell'}} \right] \right) \\
&\qquad \mathbb{E}\left[\prod_{\ell'' \in \mathcal{K}^c} D_{j_{\ell''}} \middle| (X_i)_{i \ge 1} \right].
\end{aligned}$$

Because $m > k/2$, for any $a \in A_{k,m}$ there exists s such that $a_s = 1$. For any $\mathcal{K}$, if $s \in \mathcal{K}$, then $\prod_{\ell \in \mathcal{K}} \mathbb{E}\left[f_{j_\ell}^{a_\ell} | X_{j_\ell} \right] = \mathbb{E}\left[f_{j_s} | X_{j_s} \right] \prod_{\ell \in \mathcal{K} \setminus \{s\}} \mathbb{E}\left[f_{j_\ell}^{a_\ell} | X_{j_\ell} \right] = 0$, else $s \in \mathcal{K}^c$ and $\prod_{\ell' \in \mathcal{K}^c} \left(\sum_{r=1}^{a_{\ell'}} \binom{a_{\ell'}}{r} f_{j_{\ell'}}^{a_{\ell'} - r} g_{j_{\ell'}}^{j} \right) = \mathbb{E}(g_{j_s} | X_{j_s}) \prod_{\ell' \in \mathcal{K}^c \setminus \{s\}} \left(\sum_{r=1}^{a_{\ell'}} \binom{a_{\ell'}}{r} f_{j_{\ell'}}^{a_{\ell'} - r} g_{j_{\ell'}}^{j} \right) = 0$. It follows that if $m > k/2$ we have $B_{n,k,m} = 0$.

Let now consider the last case $m = k/2$. For $a \in A_{k,k/2}$ either there exists s such that $a_s = 1$ and by the previous reasoning, we have $\prod_{\ell \in \mathcal{K}} \mathbb{E}\left[f_{j_\ell}^{a_\ell} | X_{j_\ell} \right] \prod_{\ell' \in \mathcal{K}^c} \left(\sum_{r=1}^{a_{\ell'}} \binom{a_{\ell'}}{r} \mathbb{E}\left[f_{j_{\ell'}}^{a_{\ell'} - r} g_{j_{\ell'}}^{r} | X_{j_{\ell'}} \right] \right) = 0$ for any $\mathcal{K}$, either $a = (2, ..., 2)$ and it follows

$$B_{n,k,k/2} = n^{-k/2} \sum_{1 \le j_1 < \ldots < j_{k/2} \le n} \frac{k!}{2^{k/2}} \sum_{\mathcal{K} \subseteq \{1,\ldots,k/2\}} \prod_{\ell \in \mathcal{K}} \mathbb{E}\left[f_{j_\ell}^2 | X_{j_\ell} \right] \prod_{\ell' \in \mathcal{K}^c} \mathbb{E}\left[2 f_{j_{\ell'}} g_{j_{\ell'}} + g_{j_{\ell'}}^2 | X_{j_{\ell'}} \right]$$

$$\mathbb{E}\left[\prod_{\ell''\in\mathcal{K}^c} D_{j_{\ell''}}\middle|(X_i)_{i\geq 1}\right]$$

Conjecture 1 and the fact that $\max(|f_i|^2, |2f_ig_i + g_i^2|) \leq 3K^2$ ensure

$$\begin{aligned}B_{n,k,k/2} =&n^{-k/2}\sum_{1\leq j_1<...<j_{k/2}\leq n}\frac{k!}{2^{k/2}}\sum_{\mathcal{K}\subseteq\{1,...,k/2\}}\prod_{\ell\in\mathcal{K}}\mathbb{E}\left[f_{j_\ell}^2|X_{j_\ell}\right]\prod_{\ell'\in\mathcal{K}^c}\mathbb{E}\left[2f_{j_{\ell'}}g_{j_{\ell'}} + g_{j_{\ell'}}^2|X_{j_{\ell'}}\right]\prod_{\ell''\in\mathcal{K}^c}\pi_{j_{\ell''}}\\ &+ n^{-k/2}\binom{n}{k/2}\frac{k!}{2^{k/2}}2^{k/2}(3K^2)^{k/2}o(1)\\ =&n^{-k/2}\sum_{1\leq j_1<...<j_{k/2}\leq n}\frac{k!}{2^{k/2}}\sum_{\mathcal{K}\subseteq\{1,...,k/2\}}\prod_{\ell\in\mathcal{K}}\mathbb{E}\left[f_{j_\ell}^2|X_{j_\ell}\right]\prod_{\ell'\in\mathcal{K}^c}\mathbb{E}\left[\left(2f_{j_{\ell'}}g_{j_{\ell'}} + g_{j_{\ell'}}^2\right)\pi_{j_{\ell'}}|X_{j_{\ell'}}\right]\\ &+ o(1)\end{aligned}$$

Factorization formula $\sum_{\mathcal{K}\subseteq\{1,...,m\}}\prod_{\ell\in\mathcal{K}}x_\ell\prod_{\ell'\in\mathcal{K}^c}y_{\ell'} = \prod_{\ell=1}^m(x_\ell + y_\ell)$ ensures

$$\begin{aligned}B_{n,k,k/2} &= \frac{k!}{2^{k/2}}n^{-k/2}\sum_{1\leq j_1<...<j_{k/2}\leq n}\prod_{\ell=1}^{k/2}\mathbb{E}\left[f_{j_\ell}^2 + \left(2f_{j_\ell}g_{j_\ell} + g_{j_\ell}^2\right)\pi_{j_\ell}\middle|X_{j_\ell}\right] + o(1)\\ &= \frac{k!}{2^{k/2}}n^{-k/2}\binom{n}{k/2}\binom{n}{k/2}^{-1}\sum_{1\leq j_1<...<j_{k/2}\leq n}h(X_{j_1},...,X_{j_{k/2}}) + o(1)\end{aligned}$$

for $h(u_1,...,u_{k/2}) = \prod_{i=1}^{k/2}\mathbb{E}(f^2 + (2fg + g^2)\pi|X = u_i)$. Strong law of large numbers for U-statistics (Aaronson et al., 1996) ensures that $\binom{n}{k/2}^{-1}\sum_{1\leq j_1<...<j_{k/2}\leq n}h(X_{j_1},...,X_{j_{k/2}})$ converges almost surely to $\mathbb{E}(h(X_1,...,X_{k/2})) = (V_0)^{k/2}$ and $\lim_n n^{-k/2}\binom{n}{k/2} = \frac{1}{(k/2)!}$. Then, $\lim_n M_{n,k} = 0$ for k odd, and $\lim_n M_{n,k} = \frac{k!}{2^{k/2}(k/2)!}V_0^{k/2}$ for k even. By the adapted form of the theorem in Takacs (1991), if f_i and g_i, are bounded, we have that, conditional on $(X_i)_{i\geq 1}$ $\frac{1}{\sqrt{n}}\sum_{i=1}^n f_i + g_iD_i$ converges almost surely to a Gaussian of variance V_0.

Second step: $\mathbb{E}(Y(0)^2 + Y(1)^2 + ||X||^2) < \infty$ **implies** (B.1)

Assumption 1 ensures only that f_i and g_i admit moments of order 2. Then, for $M > 0$, let $f_{\leq M,i}$, $f_{>M,i}$, $g_{\leq M,i}$ and $g_{>M,i}$ the truncated variables $f_{\leq M,i} = f_i\mathbb{1}\{|f_i| \leq M\}$, $f_{>M,i} = f_i\mathbb{1}\{|f_i| > M\}$, $g_{\leq M,i} = g_i\mathbb{1}\{|g_i| \leq M\}$ and $g_{>M,i} = g_i\mathbb{1}\{|g_i| > M\}$. We define $\tilde{f}_{\leq M,i} = f_{\leq M,i} - \mathbb{E}[f_{\leq M,i}|X_i]$, $\tilde{f}_{>M,i} = f_{>M,i} - \mathbb{E}[f_{>M,i}|X_i]$, $\tilde{g}_{\leq M,i} = g_{\leq M,i} - \mathbb{E}[g_{\leq M,i}|X_i]$ and $\tilde{g}_{>M,i} = g_{>M,i} - \mathbb{E}[g_{>M,i}|X_i]$. We have:

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n(\tilde{f}_{>M,i} + \tilde{g}_{>M,i}D_i)\right|^2\middle|(X_\ell)_{\ell\geq 1}\right]\\ =&\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[(\tilde{f}_{>M,i} + \tilde{g}_{>M,i}D_i)^2|(X_\ell)_{\ell\geq 1}\right] + \frac{1}{n}\sum_{\substack{1\leq i,j\leq n\\ i\neq j}}\mathbb{E}\left[\left(\tilde{f}_{>M,i} + \tilde{g}_{>M,i}D_i\right)\left(\tilde{f}_{>M,j} + \tilde{g}_{>M,j}D_j\right)|(X_\ell)_{\ell\geq 1}\right]\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\tilde{f}^2_{>M,i}\middle|\,(X_\ell)_{\ell\geq 1}\right]+\mathbb{E}\left[\left(2\tilde{f}_{>M,i}\tilde{g}_{>M,i}+\tilde{g}^2_{>M,i}\right)\middle|\,(X_\ell)_{\ell\geq 1}\right]\mathbb{E}\left[D_i|\,(X_\ell)_{\ell\geq 1}\right]\\
+&\frac{1}{n}\sum_{\substack{1\leq i,j\leq n\\ i\neq j}}\Big(\mathbb{E}\left[\tilde{f}_{>M,i}\tilde{f}_{>M,j}|(X_\ell)_{\ell\geq 1}\right]+\mathbb{E}\left[\tilde{f}_{>M,i}\tilde{g}_{>M,j}|(X_\ell)_{\ell\geq 1}\right]\mathbb{E}\left[D_j|\,(X_\ell)_{\ell\geq 1}\right]\\
&\qquad+\mathbb{E}\left[\tilde{f}_{>M,j}\tilde{g}_{>M,i}|(X_\ell)_{\ell\geq 1}\right]\mathbb{E}\left[D_i|\,(X_\ell)_{\ell\geq 1}\right]+\mathbb{E}\left[\tilde{g}_{>M,i}\tilde{g}_{>M,j}|(X_\ell)_{\ell\geq 1}\right]\mathbb{E}\left[D_iD_j|(X_\ell)_{\ell\geq 1}\right]\Big)\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\tilde{f}^2_{>M,i}\middle|\,X_i\right]+\mathbb{E}\left[\left(2\tilde{f}_{>M,i}\tilde{g}_{>M,i}+\tilde{g}^2_{>M,i}\right)\middle|\,X_i\right]\pi_i\\
+&\frac{1}{n}\sum_{\substack{1\leq i,j\leq n\\ i\neq j}}\Big(\mathbb{E}\left[\tilde{f}_{>M,i}|X_i\right]\mathbb{E}\left[\tilde{f}_{>M,j}|X_j\right]+\mathbb{E}\left[\tilde{f}_{>M,i}|X_i\right]\mathbb{E}\left[\tilde{g}_{>M,j}|X_j\right]\pi_j\\
&\qquad+\mathbb{E}\left[\tilde{f}_{>M,j}|X_j\right]\mathbb{E}\left[\tilde{g}_{>M,i}|X_i\right]\pi_i+\mathbb{E}\left[\tilde{g}_{>M,i}|X_i\right]\mathbb{E}\left[\tilde{g}_{>M,j}|X_j\right]\mathbb{E}\left[D_iD_j|(X_\ell)_{\ell\geq 1}\right]\Big)\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\tilde{f}^2_{>M,i}+\left(2\tilde{f}_{>M,i}\tilde{g}_{>M,i}+\tilde{g}^2_{>M,i}\right)\pi_i|X_i\right]
\end{aligned}
$$

The second equality holds because $(f_1,\ldots,f_n,g_1,\ldots,g_n)\perp\!\!\!\perp(D_1,\ldots,D_n)|X_1,\ldots,X_n$. by Assumption 2. The third equality holds because $(f_i,g_i,X_i)_{i\geq 1}$ are independent across i by Assumption 1 and $\mathbb{E}\left[D_i|(X_\ell)_{\ell>1}\right]=\pi_i$ by Assumption 2. The fourth equality holds because $\mathbb{E}\left[\tilde{f}_{>M,\ell}|X_\ell\right]=\mathbb{E}\left[\tilde{g}_{>M,\ell}|X_\ell\right]=0$.

The SLLN ensures that $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\tilde{f}^2_{>M,i}+\left(2\tilde{f}_{>M,i}\tilde{g}_{>M,i}+\tilde{g}^2_{>M,i}\right)\pi_i|X_i\right]$ converges almost-surely to $\mathbb{E}\left[\tilde{f}^2_{>M,1}+\left(2\tilde{f}_{>M,1}\tilde{g}_{>M,1}+\tilde{g}^2_{>M,1}\right)\pi_1\right]$. It follows that by Cauchy-Schwarz inequality:

$$
\begin{aligned}
&\limsup_n\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{f}_{>M,i}+\tilde{g}_{>M,i}D_i)\right|\middle|(X_\ell)_{\ell\geq 1}\right]\\
&\leq\limsup_n\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{f}_{>M,i}+\tilde{g}_{>M,i}D_i)\right|^2\middle|(X_\ell)_{\ell\geq 1}\right]^{1/2}\\
&=\mathbb{E}\left[\tilde{f}^2_{>M,1}+\left(2\tilde{f}_{>M,1}\tilde{g}_{>M,1}+\tilde{g}^2_{>M,1}\right)\pi_1\right]^{1/2}
\end{aligned}
\tag{B.2}
$$

which, by dominated convergence, is arbitrarily small for a sufficiently large M.

Let h a bounded Lipschitz function of constant c_h, $V(M)=\mathbb{E}\left[\tilde{f}^2_{\leq M,1}+\left(2\tilde{f}_{>M,1}\tilde{g}_{\leq M,1}+\tilde{g}^2_{\leq M,1}\right)\pi_1\right]$, and $N\sim\mathcal{N}(0,1)$. We have by triangle, Lipschitz inequlities, and the fact that $f_i+g_iD_i=\tilde{f}_{\leq M,i}+\tilde{g}_{\leq M,i}D_i+\tilde{f}_{>M,i}+\tilde{g}_{>M,i}D_i$:

$$
\begin{aligned}
&\left|\mathbb{E}\left[h\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_i+g_iD_i\right)\middle|(X_\ell)_{\ell\geq 1}\right]-\mathbb{E}\left[h\left(V_0^{1/2}N\right)\right]\right|\\
&\leq\left|\mathbb{E}\left[h\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_i+g_iD_i\right)\middle|(X_\ell)_{\ell\geq 1}\right]-\mathbb{E}\left[h\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{f}_{\leq M,i}+\tilde{g}_{\leq M,i}D_i\right)\middle|(X_\ell)_{\ell\geq 1}\right]\right|
\end{aligned}
$$

$$
\begin{aligned}
&+\left|\mathbb{E}\left[h\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{f}_{\leq M,i}+\tilde{g}_{\leq M,i}D_i\right)\Big|(X_\ell)_{\ell\geq 1}\right]-\mathbb{E}\left[h\left(V(M)^{1/2}N\right)\right]\right|\\
&+\left|\mathbb{E}\left[h\left(V(M)^{1/2}N\right)\right]-\mathbb{E}\left[h\left(V_0^{1/2}N\right)\right]\right|\\
&\leq c_h\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{f}_{>M,i}+\tilde{g}_{>M,i}D_i\right|\Big|(X_\ell)_{\ell\geq 1}\right]\\
&+\left|\mathbb{E}\left[h\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{f}_{\leq M,i}+\tilde{g}_{\leq M,i}D_i\right)\Big|(X_\ell)_{\ell\geq 1}\right]-\mathbb{E}\left[h\left(V(M)^{1/2}N\right)\right]\right|\\
&+c_h\left|V(M)^{1/2}-V_0^{1/2}\right|\mathbb{E}(|N|).
\end{aligned}
$$

The first step of the proof and (B.2) ensure that for any value of $M>0$:

$$
\begin{aligned}
&\limsup_n\left|\mathbb{E}\left[h\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_i+g_iD_i\right)\Big|(X_\ell)_{\ell\geq 1}\right]-\mathbb{E}\left[h\left(V_0^{1/2}N\right)\right]\right|\\
&\leq c_h\left(\mathbb{E}\left[\tilde{f}^2_{>M,1}+\left(2\tilde{f}_{>M,1}\tilde{g}_{>M,1}+\tilde{g}^2_{>M,1}\right)\pi_1\right]^{1/2}+\left|V(M)^{1/2}-V_0^{1/2}\right|\right).
\end{aligned}
$$

By dominated convergence, $\lim_M V(M)=V_0$ and $\lim_M \mathbb{E}\left[\tilde{f}^2_{>M,1}+\left(2\tilde{f}_{>M,1}\tilde{g}_{>M,1}+\tilde{g}^2_{>M,1}\right)\pi_1\right]=0$. Next, considering M tending to ∞, dominated convergence ensures

$$
\limsup_n\left|\mathbb{E}\left[h\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_i+g_iD_i\right)\Big|(X_\ell)_{\ell\geq 1}\right]-\mathbb{E}\left[h\left(V_0^{1/2}N\right)\right]\right|=0.
$$

This achieves the proof.