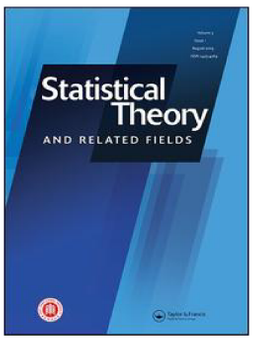

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Covariate balancing based on kernel density estimates for controlled experiments

Yiou Li , Lulu Kang & Xiao Huang

To cite this article: Yiou Li , Lulu Kang & Xiao Huang (2021): Covariate balancing based on kernel density estimates for controlled experiments, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2021.1878742

To link to this article: https://doi.org/10.1080/24754269.2021.1878742

Published online: 03 Feb 2021.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2021.1878742

Taylor & Francis
Taylor & Francis Group

Covariate balancing based on kernel density estimates for controlled experiments

Yiou Li[a], Lulu Kang [b] and Xiao Huang[b]

[a]Department of Mathematical Sciences, DePaul University, Chicago, IL, USA; [b]Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL, USA

ABSTRACT

Controlled experiments are widely used in many applications to investigate the causal relationship between input factors and experimental outcomes. A completely randomised design is usually used to randomly assign treatment levels to experimental units. When covariates of the experimental units are available, the experimental design should achieve covariate balancing among the treatment groups, such that the statistical inference of the treatment effects is not confounded with any possible effects of covariates. However, covariate imbalance often exists, because the experiment is carried out based on a single realisation of the complete randomisation. It is more likely to occur and worsen when the size of the experimental units is small or moderate. In this paper, we introduce a new covariate balancing criterion, which measures the differences between kernel density estimates of the covariates of treatment groups. To achieve covariate balance before the treatments are randomly assigned, we partition the experimental units by minimising the criterion, then randomly assign the treatment levels to the partitioned groups. Through numerical examples, we show that the proposed partition approach can improve the accuracy of the difference-in-mean estimator and outperforms the complete randomisation and rerandomisation approaches.

ARTICLE HISTORY
Received 12 August 2020
Revised 16 January 2021
Accepted 18 January 2021

KEYWORDS
Covariate balance; controlled experiment; completely randomised design; difference-in-mean estimator; kernel density estimation; rerandomisation

1. Introduction

The controlled experiment is a useful tool for investigating the causal relationship between experimental factors and responses. It has broad applications in many fields, such as science, medicine, social science, business, etc. The construction of the experimental design for a controlled experiment has two steps. First, a factorial experimental design, that is the treatment settings of the experimental factors, is specified. By *treatment settings* or *treatment levels*, we mean the combinations of the different factorial settings of all factors. There is a rich body of literature of classic and new methods on a factorial design (Wu & Hamada, 2011), and it is beyond the scope of this paper. The second step is to assign treatment settings to the available experimental units. A factorial design can involve only one factor with L treatment levels, or multiple factors with L combinations of treatment settings, which is decided by the design. In either case, there are $L-1$ treatment effects, such as main effects or interactions between multiple factors, and they are commonly estimated using the difference-in-mean estimator. Although practitioners usually use a completely randomised design for the second step, it has some limitations, as we are going to discuss next. In this paper, we focus on the second step, how to assign experimental units to treatment levels to improve the accuracy of the difference-in-mean estimator.

In many applications, the experimental units are varied with different *covariates* information. The response measurement of an experimental unit can be influenced by both the treatment setting and the covariates information of the experimental unit. For instance, in a clinical trial of a new hypoglycaemic agent, the experimental factor is the treatment that a patient receives and it has two levels, the new agent (with a fixed dosage) and placebo. The experimental units are the patients who participate in the clinical trial. The patients are partitioned into two groups. The group that receives the placebo is typically called the *control group*, and the group that receives the new agent is called the *treatment group*. The treatment effect of the new agent is then estimated by the difference of the average blood glucose level of the treatment group and that of the control group. This estimator is called *mean-difference* or *difference-in-mean estimator* (Rubin, 2005). The covariates of experimental units include the patients' age, gender, weight, and other physical and medical information. Naturally, the blood glucose level of a patient (the response) is related to the treatment, as well as the physical and medical background (covariates information) of the patient. If the distributions of covariates of the treatment and control groups are significantly different, the estimated treatment effect via the difference-in-mean estimator can be confounded with some covariate

CONTACT Lulu Kang lkang2@iit.edu Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA

effects. If the number of experimental units is large relative to L, the confounding problem can be elevated by a completely randomised design. Asymptotically, the completely randomised design results in the same distribution of covariates for all partitioned groups. In other words, it achieves *covariates balance*. However, in practice, the experimenter only conducts the experiment once or a few times, each time using a single realisation of the complete randomisation, and usually for a finite number of experimental units. As pointed out by many existing works in the causal inference literature, relying on complete randomisation can be dangerous (Bertsimas et al., 2015; Morgan & Rubin, 2012). Especially, when the number of experimental units is small or moderate, covariate imbalance among the partitioned groups under complete randomisation could be surprisingly significant, leading to inaccurate estimates and incorrect statistical inference of treatment effects (Bertsimas et al., 2015).

The problem of inaccurate estimates induced by the covariate imbalance could be addressed by two types of methods. One type of method is for observational data. The premise is that a completely randomised design is used to assign treatment levels to experimental units before data are collected. Then some adjusting methods are applied at the data analysis stage. These adjusting methods include post-stratification (McHugh & Matts, 1983; Xie & Aurisset, 2016), propensity score matching (Pearl, 2000; Rosenbaum & Rubin, 1983), Doubly-robust estimator (Funk et al., 2011), coarsened exact matching (CEM) (Blackwell et al., 2009), etc. The common theme of these methods is to apply sophisticated weighting schemes to improve the original difference-in-mean estimator and reduce the mean squared error (consisting of the bias and variance) of the estimator the imbalance of covariates. The alternative estimators also include the least-square estimator (Wu & Hamada, 2011) of the treatment effects which is based on parametric model assumption. The literature on the observational study in Causal Inference is vast and we refer readers to Imbens and Rubin (2015) and Rosenbaum (2017) for a more comprehensive review.

Another kind of method aims at achieving covariate balance before the random assignment of the treatment levels (Kallus, 2018) at the design stage and before data collection. First, the experimental units are partitioned into L groups according to a certain covariate balancing criterion. Then, the L treatment levels are randomly assigned to the L groups. Such methods include randomised block designs (Bernstein, 1927), rerandomisation (Morgan & Rubin, 2012, 2015), and the optimal partition proposed by Bertsimas et al. (2015) and Kallus (2018), etc. With a randomised block design, the experimental units are divided into subgroups called blocks such that the experimental units have similar covariates information in each block. Then, the treatment levels are randomly assigned to the experimental units within each block. Similar to the post-stratification method, blocking can not be directly applied when the covariates are continuous or mixed. Users need to choose a discrete block factor based on continuous or mixed covariates. Morgan and Rubin (2012) and Morgan and Rubin (2015) proposed a rerandomisation method in which the experimenter keeps randomising the experimental units into L groups to achieve a sufficiently small Mahalanobis distance of the covariates between the groups. In Bertsimas et al. (2015), the imbalance is measured as the sum of discrepancies in group means and variances, and the desired partition minimises this imbalance criterion. Kallus (2018) proposed a new kernel allocation to divide the experimental units into balanced groups.

In this paper, we only consider the controlled experiments with continuous covariates. We propose a new criterion to measure the covariate imbalance, and the corresponding optimal partition minimises this new criterion and achieves the best covariates balance between L treatment groups. The problem set-up and assumptions are introduced in Section 2. In Section 3, we propose the new covariate balancing criterion, which measures the differences between the kernel density estimates of the covariates of the L groups of experimental units. In Section 4, we formulate the partition problem into a quadratic integer programming and discuss the choice of the parameters in kernel density estimates and optimisation. The proposed approach is compared with complete randomisation and rerandomisation through simulation and real examples in Section 5. We conclude this paper with some discussions and potential research directions in Section 6.

2. Problem set-up

In this work, we assume that the controlled experiment has N experimental units and they are predetermined and fixed through the data collection process. We first illustrate the problem set-up using the case of $L = 2$, based on which we derive the covariates balancing criterion in Section 3. In the second part of Section 3, we extend the design criterion from $L = 2$ case to the general $L \geq 2$ case.

We assume that the response variable follows a general model

$$y_i = h(\boldsymbol{z}_i) + \alpha x_i + \epsilon_i, \quad i = 1, \ldots, N. \tag{1}$$

Here $x_i \in \{0, 1\}$ is the indicator of which of the two treatment levels the experimental unit i recieves. Conventionally, we use $x_i = 0$ to present a baseline level, or *control*, and $x_i = 1$ to present the other level, or *treatment*. We loosely use the terms control and treatment just to make the distinction between the two different levels. Assume the covariates of a experimental unit is

$\boldsymbol{z} \in \Omega \subset \mathbb{R}^d$, and $\boldsymbol{z}_i = [z_{i1}, \ldots, z_{id}]^\top$ is the observed covariates of the ith experimental unit. The function h is a square-integrable function. When $x_i = 0$, $h(\boldsymbol{z}_i)$ is the mean of the response y_i. The parameter α is the treatment effect and ϵ_i is the random noise with zero mean and constant variance σ^2. Furthermore, ϵ_i is independent of the covariates of all the experimental units, treatment assignment, and the random noise of other experimental units.

Based on (1), the sample means of the response in two groups are calculated as

$$\bar{y}_T = \frac{\sum_{i=1}^N y_i x_i}{n_T}, \quad \bar{y}_C = \frac{\sum_{i=1}^N y_i(1-x_i)}{n_C},$$

where n_T and n_C are number of experimental units in treatment group and control group, respectively. The most commonly used estimator for the treatment effect α is the difference-in-mean estimator

$$\begin{aligned}\hat{\alpha} &= \bar{y}_T - \bar{y}_C \\ &= \alpha + \int h(\boldsymbol{z}) \mathrm{d}\hat{F}_T(\boldsymbol{z}) - \int h(\boldsymbol{z})\, \mathrm{d}\hat{F}_C(\boldsymbol{z}) + \bar{\epsilon}_T - \bar{\epsilon}_C,\end{aligned} \tag{2}$$

where $\hat{F}_T$ and $\hat{F}_C$ are the empirical distributions of the covariates of the treatment group and control group, respectively, and $\bar{\epsilon}_T = \frac{\sum_{i=1}^N \epsilon_i x_i}{n_T}$ and $\bar{\epsilon}_T = \frac{\sum_{i=1}^N \epsilon_i (1-x_i)}{n_C}$ are the mean errors in the corresponding groups.

Before conducting the experiment, the randomness of $\hat{\alpha}$ comes from three sources: random noise, the partition of experimental units in the two groups (if it is done through randomisation), and the random assignment of two levels to the two partitioned groups. More importantly, the three sources of randomness are independent of each other in our framework. Accordingly, the mean of estimator $\hat{\alpha}$ is

$$\mathrm{E}(\hat{\alpha}) = \mathrm{E}_\epsilon \left\{ \mathrm{E}_{\text{Partition}} \left[\mathrm{E}_{\text{LA}} \left(\hat{\alpha} \,\middle|\, \hat{F}_1, \hat{F}_2, \epsilon \right) \middle| \epsilon \right] \right\},$$

where $\hat{F}_1$ and $\hat{F}_2$ are the empirical distribution of partitioned Group 1 and Group 2, respectively. The partition can be random or deterministic, depending on the partition method used. Define LA as the Bernoulli random variable representing the Level Assignment. Let $LA = 0$ if Group 1 is assigned as the treatment group and $LA = 1$ if Group 1 is assigned as the control group, and $\Pr(LA = 0) = \Pr(LA = 1) = \frac{1}{2}$. Obviously, if we only consider the random level assignments given the partition and random noise,

$$\begin{aligned}&\mathrm{E}_{\text{LA}} \left(\hat{\alpha} \,\middle|\, \hat{F}_1, \hat{F}_2, \epsilon \right) \\ &= \alpha + \Pr(LA = 0) \left(\int h(\boldsymbol{z}) \mathrm{d}\hat{F}_1(\boldsymbol{z}) - \int h(\boldsymbol{z}) \mathrm{d}\hat{F}_2(\boldsymbol{z}) \right) \\ &\quad + \Pr(LA = 1) \left(\int h(\boldsymbol{z}) \mathrm{d}\hat{F}_2(\boldsymbol{z}) - \int h(\boldsymbol{z}) \mathrm{d}\hat{F}_1(\boldsymbol{z}) \right) \\ &\quad + \bar{\epsilon}_T - \bar{\epsilon}_C \\ &= \alpha + \bar{\epsilon}_T - \bar{\epsilon}_C.\end{aligned}$$

Since the random noise is independent of partition, regardless of distribution of the partition,

$$\mathrm{E}(\hat{\alpha}) = \mathrm{E}_\epsilon \left\{ \mathrm{E}_{\text{Partition}} \left(\alpha + \bar{\epsilon}_T - \bar{\epsilon}_C \right) \right\} = \alpha.$$

Therefore, it does not matter what kind of partition method we use, random or deterministic, the difference-in-mean estimator is always unbiased (Kallus, 2018), as long as the Level Assignment is fairly and randomly assigned to the two partitioned groups.

Similarly, considering the three sources of randomness, the variance of $\hat{\alpha}$ is

$$\begin{aligned}\mathrm{var}(\hat{\alpha}) &= \mathrm{E}\left[\left(\int h(\boldsymbol{z}) \mathrm{d}\hat{F}_T(\boldsymbol{z}) \right.\right. \\ &\qquad \left.\left. - \int h(\boldsymbol{z})\, \mathrm{d}\hat{F}_C(\boldsymbol{z}) + \bar{\epsilon}_T - \bar{\epsilon}_C \right)^2 \right] \\ &= \mathrm{E}\left[\left(\int h(\boldsymbol{z}) \mathrm{d}\hat{F}_T(\boldsymbol{z}) - \int h(\boldsymbol{z})\, \mathrm{d}\hat{F}_C(\boldsymbol{z}) \right)^2 \right] \\ &\quad + \sigma^2 \left(\frac{1}{n_T} + \frac{1}{n_C} \right).\end{aligned} \tag{3}$$

In (3), the expectation in the last equation is with respect to the randomness in the partition of the experimental units, and the randomness of level assignment. We can use a more detailed but cumbersome notation and derive

$$\begin{aligned}&\mathrm{E}\left[\left(\int h(\boldsymbol{z}) \mathrm{d}\hat{F}_T(\boldsymbol{z}) - \int h(\boldsymbol{z}) \mathrm{d}\hat{F}_C(\boldsymbol{z}) \right)^2 \right] \\ &= \mathrm{E}_{\text{Partition}} \left\{ \mathrm{E}_{\text{LA}} \left[\left(\int h(\boldsymbol{z}) \mathrm{d}\hat{F}_T(\boldsymbol{z}) \right.\right.\right. \\ &\qquad \left.\left.\left. - \int h(\boldsymbol{z}) \mathrm{d}\hat{F}_C(\boldsymbol{z}) \right)^2 \middle| \hat{F}_1, \hat{F}_2 \right] \right\} \\ &= \mathrm{E}_{\text{Partition}} \left\{ \Pr(LA = 0) \left(\int h(\boldsymbol{z}) \mathrm{d}\hat{F}_1(\boldsymbol{z}) \right.\right. \\ &\qquad \left. - \int h(\boldsymbol{z}) \mathrm{d}\hat{F}_2(\boldsymbol{z}) \right)^2 \\ &\qquad + \Pr(LA = 1) \left(\int h(\boldsymbol{z}) \mathrm{d}\hat{F}_2(\boldsymbol{z}) \right. \\ &\qquad \left.\left. - \int h(\boldsymbol{z}) \mathrm{d}\hat{F}_1(\boldsymbol{z}) \right)^2 \right\} \\ &= \mathrm{E}_{\text{Partition}} \left[\left(\int h(\boldsymbol{z}) \mathrm{d}\hat{F}_1(\boldsymbol{z}) - \int h(\boldsymbol{z}) \mathrm{d}\hat{F}_2(\boldsymbol{z}) \right)^2 \right].\end{aligned}$$

As a result, once the partition of the experimental units is established, the variance of difference-in-mean

estimator $\text{var}(\hat{\alpha})$ in (3) is invariant to the treatment assignment of the two groups, that is,

$$\text{var}(\hat{\alpha}) = \text{E}_{\text{Partition}}\left[\left(\int h(\boldsymbol{z})\text{d}\hat{F}_1(\boldsymbol{z}) - \int h(\boldsymbol{z})\text{d}\hat{F}_2(\boldsymbol{z})\right)^2\right] + \sigma^2\left(\frac{1}{n_T}+\frac{1}{n_C}\right). \tag{4}$$

Thus, the experimenter should focus on the partition of the experimental units to reduce $\text{var}(\hat{\alpha})$. In the following section, we propose an alternative partition method that aims at regulating the variance of the difference-in-mean estimator $\hat{\alpha}$.

3. Kernel density estimation based covariate balancing criterion

In this section, we first show the derivation of the KDE-based covariate balancing criterion for the $L = 2$ following the set-up in Section 2. Then we extend the balancing criterion to the general $L \geq 2$ case.

3.1. The case of $L = 2$

Define a partition of the experimental units as $\boldsymbol{g} = [g_1, \ldots, g_N]^\top$, where $g_i = 0$ if the ith experimental unit is partitioned into Group 1 and $g_i = 1$ if the ith experimental unit is partitioned into Group 2. Note that g_i is different from the indicator variable x_i in (1). The order of partition and treatment level assignment does not matter. One can partition the experimental units into two groups first and then randomly assign treatment levels. Or, one can randomly assign treatment levels to the two groups (which are still empty) and fill the groups with experimental units afterward. Therefore, Group 1 is not necessarily the treatment or the control group, that is, $g_i = 1$ does not imply $x_i = 0$ or $x_i = 1$.

To construct a smooth approximation of the empirical distributions of the covariates in two groups, we estimate the corresponding distributions using the kernel density estimation. The kernel density estimation (KDE) is a popular technique to estimate the density function of a multivariate distribution, which is a generalisation of histogram density estimation but with improved statistical properties (Simonoff, 2012). We use this approximation for two reasons: (1) the covariates are continuous in nature, and (2) to bound $\text{var}(\hat{\alpha})$. With a sample $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$ of a $d-$dimensional random vector drawn from a distribution with density function f, the kernel density estimate is defined to be

$$\hat{f}(\boldsymbol{z}) = \frac{1}{n}|\mathsf{H}|^{-1/2}\sum_{i=1}^{n} K\left(\mathsf{H}^{-1/2}(\boldsymbol{z}-\boldsymbol{z}_i)\right), \tag{5}$$

where $K(\cdot)$ is the kernel function which is a symmetric multivariate density function, and H is the positive definite bandwidth matrix. With the smooth approximation of the empirical distributions, by (4), the variance of the difference-in-mean estimator is approximately upper bounded by

$$\begin{aligned}
\text{var}(\hat{\alpha}) &= \text{E}\left[\left(\int h(\boldsymbol{z})\text{d}\hat{F}_1(\boldsymbol{z}) - \int h(\boldsymbol{z})\text{d}\hat{F}_2(\boldsymbol{z})\right)^2\right] + \sigma^2\left(\frac{1}{n_T}+\frac{1}{n_C}\right)\\
&\approx \text{E}\left[\left(\int h(\boldsymbol{z})\hat{f}_1(\boldsymbol{z})\text{d}\boldsymbol{z} - \int h(\boldsymbol{z})\hat{f}_2(\boldsymbol{z})\text{d}\boldsymbol{z}\right)^2\right] + \sigma^2\left(\frac{1}{n_T}+\frac{1}{n_C}\right)\\
&\leq \text{E}\left[\int |h(\boldsymbol{z})|^2\text{d}\boldsymbol{z}\int\left|\hat{f}_1(\boldsymbol{z}) - \hat{f}_2(\boldsymbol{z})\right|^2\text{d}\boldsymbol{z}\right] + \sigma^2\left(\frac{1}{n_T}+\frac{1}{n_C}\right)\\
&= \|h\|_2^2\text{E}\left[\left\|\hat{f}_1 - \hat{f}_2\right\|_2^2\right] + \sigma^2\left(\frac{1}{n_T}+\frac{1}{n_C}\right),
\end{aligned} \tag{6}$$

where the inequality follows from Cauchy-Schwarz inequality, $\|\cdot\|_2^2$ denotes the $\mathcal{L}_2$-norm of a function, $\hat{F}_1$ and $\hat{F}_2$ are empirical distributions of covariates in Group 1 and 2, and $\hat{f}_1$ and $\hat{f}_2$ are the KDE of the covariates of the two groups, respectively. Here the expectation is only with respect to the partition $\boldsymbol{g}$ as explained in (4) in Section 2.

Since $\text{var}(\hat{\alpha})$ depends on the function h, we cannot directly minimise $\text{var}(\hat{\alpha})$ with respect to partition without making any assumption on h function. To make our approach robust to any assumption on h, we propose using

$$B_{\mathsf{H}}(\boldsymbol{g}) = \left\|\hat{f}_1 - \hat{f}_2\right\|_2^2 \tag{7}$$

as the covariate balancing criterion to partition the experimental units. It is the part of the upper bound that is not a constant and only depends on the partition $\boldsymbol{g}$. This criterion also appeared in Anderson et al. (1994), which proposed the same criterion as a two-sample test statistic to test whether the two samples are drawn from the same distribution. Their work further supports the idea of using $B_{\mathsf{H}}(\boldsymbol{g})$ as the partition criterion from the covariate balancing perspective. A small $B_{\mathsf{H}}(\boldsymbol{g})$ value suggests that the covariates samples in the two groups tend to be drawn from the same distribution, then the covariate information in the two groups is balanced.

To achieve a partition with small $B_{\mathsf{H}}(\boldsymbol{g})$, one can either find the optimal solution that minimises $B_{\mathsf{H}}(\boldsymbol{g})$ or rerandomise the experimental units until a sufficiently small $B_{\mathsf{H}}(\boldsymbol{g})$ is obtained. The latter way is

doable when the asymptotic distribution of the criterion is available. The asymptotic distribution of $B_{\mathsf{H}}(\boldsymbol{g})$ can be constructed using bootstrap method (Anderson et al., 1994). However, different from the simple normal asymptotic distribution of Mahalanobis distance derived in Morgan and Rubin (2012), due to the computation cost of the bootstrap method, calculating the threshold of $B_{\mathsf{H}}(\boldsymbol{g})$ is computationally expensive. As a result, we choose to construct a partition by minimising $B_{\mathsf{H}}(\boldsymbol{g})$, and we call the partition $\boldsymbol{g}^* = \text{argmin}_{\boldsymbol{g}} B_{\mathsf{H}}(\boldsymbol{g})$ the KDE-based partition. Since we use the optimisation approach, our partition scheme is actually deterministic. In other words, the partition scheme we use is to let $\boldsymbol{g} = \boldsymbol{g}^*$ with probability equal to 1. This discussion also applies to the general $L \geq 2$ case. In Section 4, we discuss in detail about how to construct the KDE-based partition that minimises $B_{\mathsf{H}}(\boldsymbol{g})$.

3.2. General case of $L \geq 2$

For the general case of $L \geq 2$, the covariate balancing criterion is generalised as follows

$$B_{\mathsf{H}}(\boldsymbol{g}) = \max_{l,s=1,\ldots,L} \left\| \hat{f}_l - \hat{f}_s \right\|_2^2. \tag{8}$$

Extending $\boldsymbol{g}$ to the case of $L>2$, the partition is defined as $\boldsymbol{g} = [g_1, \ldots, g_N]^\top$, with $g_i \in \{0, \ldots, L-1\}$, $i = 1, \ldots, N$. To show the generalised criterion in (8) is reasonable, we explain it under two scenarios: (1) the experiment involves only one factor with L treatment settings; (2) the experiment involves multiple factors and the experimental design contains L different combinations of the treatment settings of these factors. In the first scenario, the treatment effects of interest are the pairwise contrasts between any two treatment levels. The linear model in (1) is extended to

$$y_i = h(\boldsymbol{z}_i) + \sum_{l=1}^{L-1} \alpha_j x_{il} + \epsilon_i, \quad i = 1, \ldots, N.$$

Here x_{il}'s are the dummy variables corresponding to the assigned treatment level for the ith unit, and $x_{il} = 1$ if the ith unit is assigned to treatment level l, and $x_{il} = 0$ otherwise, for $l = 1, \ldots, L-1$. Using this notation, the treatment level L is set to be the baseline, and the rest of the treatment levels are compared to it. The treatment effects α_l for $l = 1, \ldots, L-1$ are estimated by the difference-in-mean estimator

$$\hat{\alpha}_l = \bar{y}_l - \bar{y}_L, \quad \text{for } l = 1, \ldots, L-1,$$

where $\bar{y}_k$ is the sample mean of the observations with treatment level k for $k = 1, \ldots, L$. Often, an experimenter is interested in the pairwise contrasts of the treatment effects $\alpha_l - \alpha_s$, $l, s = 1, \ldots, L$. Therefore, one would aim at minimising the largest variance of the corresponding estimator $\hat{\alpha}_l - \hat{\alpha}_s$, that is, $\max_{l,s=1,\ldots,L} \text{var}(\hat{\alpha}_l - \hat{\alpha}_s)$. Following the same deviation in the $L = 2$ case, $\hat{\alpha}_l - \hat{\alpha}_s$ is unbiased with variance

$$\begin{aligned}
&\text{var}(\hat{\alpha}_l - \hat{\alpha}_s) \\
&= \text{E}_{\text{Partition}} \left[\left(\int h(\boldsymbol{z}) \text{d}\hat{F}_l(\boldsymbol{z}) - \int h(\boldsymbol{z}) \text{d}\hat{F}_s(\boldsymbol{z}) \right)^2 \right] \\
&\quad + \sigma^2 \left(\frac{1}{n_l} + \frac{1}{n_s} \right).
\end{aligned}$$

Using the similar argument in (6), the largest variance is approximately upper bounded by

$$\begin{aligned}
&\max_{l,s=1,\ldots,L} \text{var}(\hat{\alpha}_l - \hat{\alpha}_s) \\
&\lessapprox \max_{l,s=1,\ldots,L} \left\{ \|h\|_2^2 \text{E} \left[\left\| \hat{f}_s - \hat{f}_l \right\|_2^2 \right] + \sigma^2 \left(\frac{1}{n_l} + \frac{1}{n_s} \right) \right\}.
\end{aligned}$$

Assuming all the L treatment groups have very similar or the same size of experimental units, one should aim at minimising $\max_{l,s=1,\ldots,L} \|\hat{f}_l - \hat{f}_s\|_2^2$. Thus, a natural covariates balancing criterion is

$$B_{\mathsf{H}}(\boldsymbol{g}) = \max_{l,s=1,\ldots,L} \left\| \hat{f}_l - \hat{f}_s \right\|_2^2.$$

When $L = 2$, the criterion reduces to (7).

In the second scenario, we explain it via a simple 2×2 full factorial design with two two-level factors, denoted by A and B. The full factorial design contains $L = 4$ treatment settings. Using the generic notation in the design of experiments literature, the four treatment settings are $(+,+)$, $(+,-)$, $(-,+)$, and $(-,-)$. The sample mean of the observations in each treatment group is $\bar{y}_l$ and the sample size is n_l for $l = 1, \ldots, L$. There are $L-1 = 3$ effects to be estimated, which are main effects of A and B (denoted as α_A and α_B) and their interaction effect (denoted by α_{AB}). For simplicity, we assume $n_1 = \ldots = n_L = n$, and the total number of experimental units is $N = 4n$. The commonly used difference-in-mean estimators for α_A, α_B, and α_{AB} are

$$\begin{aligned}
\hat{\alpha}_A &= \tfrac{1}{2} \left(\bar{y}_1 + \bar{y}_2 - \bar{y}_3 - \bar{y}_4 \right) \\
\hat{\alpha}_B &= \tfrac{1}{2} \left(\bar{y}_1 + \bar{y}_3 - \bar{y}_2 - \bar{y}_4 \right) \\
\hat{\alpha}_{AB} &= \tfrac{1}{2} \left(\bar{y}_1 - \bar{y}_2 - \bar{y}_3 + \bar{y}_4 \right).
\end{aligned}$$

Following the same derivation of (4), we obtain the variance of $\hat{\alpha}_A$

$$\begin{aligned}
\text{var}(\hat{\alpha}_A) = \text{E} \Bigg[\frac{1}{4} \Bigg(&\int h(\boldsymbol{z}) \left[\text{d}\hat{F}_1(\boldsymbol{z}) + \text{d}\hat{F}_2(\boldsymbol{z}) \right] \\
&- \int h(\boldsymbol{z}) \left[\text{d}\hat{F}_3(\boldsymbol{z}) + \text{d}\hat{F}_4(\boldsymbol{z}) \right] \Bigg)^2 \Bigg] + \frac{\sigma^2}{n}.
\end{aligned}$$

Using the argument of (6), $\mathrm{var}(\hat{\alpha}_A)$ is approximately upper bounded

$$\mathrm{var}(\hat{\alpha}_A) \lessapprox \frac{1}{4}\|h\|_2^2 \mathrm{E}\left[\left\|\hat{f}_1+\hat{f}_2-\hat{f}_3-\hat{f}_4\right\|_2^2\right]+\frac{\sigma^2}{n}.$$

In this upper bound, only $\|\hat{f}_1+\hat{f}_2-\hat{f}_3-\hat{f}_4\|_2$ depends on the partition. By triangle inequality,

$$\begin{aligned}&\left\|\hat{f}_1+\hat{f}_2-\hat{f}_3-\hat{f}_4\right\|_2\\ &\le \min\left\{\left\|\hat{f}_1-\hat{f}_3\right\|_2+\left\|\hat{f}_2-\hat{f}_4\right\|_2, \left\|\hat{f}_1-\hat{f}_4\right\|_2\right.\\ &\quad\left.+\left\|\hat{f}_2-\hat{f}_3\right\|_2\right\}.\end{aligned}$$

Similarly, the corresponding norms in the upper bounds of $\mathrm{var}(\hat{\alpha}_B)$ and $\mathrm{var}(\hat{\alpha}_{AB})$ are upper bounded by

$$\begin{aligned}&\left\|\hat{f}_1+\hat{f}_3-\hat{f}_2-\hat{f}_4\right\|_2\\ &\le \min\left\{\left\|\hat{f}_1-\hat{f}_2\right\|_2+\left\|\hat{f}_3-\hat{f}_4\right\|_2, \left\|\hat{f}_1-\hat{f}_4\right\|_2+\left\|\hat{f}_3-\hat{f}_2\right\|_2\right\},\\ &\left\|\hat{f}_1+\hat{f}_4-\hat{f}_2-\hat{f}_3\right\|_2\\ &\le \min\left\{\left\|\hat{f}_1-\hat{f}_2\right\|_2+\left\|\hat{f}_3-\hat{f}_4\right\|_2, \left\|\hat{f}_1-\hat{f}_3\right\|_2+\left\|\hat{f}_2-\hat{f}_4\right\|_2\right\}.\end{aligned}$$

Therefore, to regulate the worst-case variance of these three estimators, it makes sense to minimise the largest pairwise difference $\|\hat{f}_l-\hat{f}_s\|_2$, $l,s=1,\ldots,4$. Thus, we reach the same criterion as above

$$B_{\mathsf{H}}(\boldsymbol{g}) = \max_{l,s=1,\ldots,4}\left\|\hat{f}_l-\hat{f}_s\right\|_2^2.$$

For general full or fractional factorial design, if there are L treatment settings, there would be L sample means of the response from each treatment group. If all the $L-1$ effects (main effects, two-factor interactions, etc) are parameters of interests, then all the pairwise distance of $||\hat{f}_l-\hat{f}_s||_2$ should be as small as possible so that the covariate balancing is achieved at best across all treatment groups. If only some of the $L-1$ effects are of interest, it is still ideal to reach covariate balancing across all treatment groups because the difference-in-mean estimators are essentially linear combinations of the group sample means.

4. Construction of KDE-based partition

Constructing a KDE-based partition of the experimental units is essentially an optimisation problem.

$$\min_{\boldsymbol{g}} \quad \max_{l,s=1,\ldots,L}\left\|\hat{f}_l-\hat{f}_s\right\|_2^2 \tag{9a}$$

$$\text{s.t.} \quad 0\le g_i\le L-1, \quad i=1,\ldots,N, \tag{9b}$$

$$\sum_{i=1}^{N} \mathbb{1}_{\{g_i=j\}} = n_j, \quad j=0,\ldots,L-1, \tag{9c}$$

$$g_i \quad \text{integer}, \ i=1,\ldots,N. \tag{9d}$$

Here $\mathbb{1}_{\{\cdot\}}$ is the indicator function that $\mathbb{1}_{\{A\}}=1$ if A is true, and $\mathbb{1}_{\{A\}}=0$ otherwise, and n_j is the size of the experimental units in the $(j+1)$th group. This is an integer programming problem which can be difficult and computational to solve. In the next part, we formulate (9) into a quadratic integer programming problem for $L=2$. It can be solved efficiently using modern optimisation tools.

4.1. Optimisation

To facilitate the formulation and computation of the optimisation problem, we derive a more concrete formula of $B_{\mathsf{H}}(\boldsymbol{g})$ defined in (8) for general $L\ge 2$. For simplicity, we assume the number of experimental units is equal for all groups, so N is divisible by L. Denote the number of experimental units in the lth partitioned group as n, and $N=nL$. For any $l,s=1,\ldots,L$, we have

$$\begin{aligned}\left\|\hat{f}_l-\hat{f}_s\right\|_2^2 &= \int\left|\hat{f}_l(\boldsymbol{z})-\hat{f}_s(\boldsymbol{z})\right|^2 \mathrm{d}\boldsymbol{z}\\ &= \int\left(\frac{1}{n}|\mathsf{H}|^{-\frac{1}{2}}\sum_{g_i=l-1}K\left(\mathsf{H}^{-\frac{1}{2}}(\boldsymbol{z}-\boldsymbol{z}_i)\right)\right.\\ &\quad\left.-\frac{1}{n}|\mathsf{H}|^{-\frac{1}{2}}\sum_{g_i=s-1}K\left(\mathsf{H}^{-\frac{1}{2}}(\boldsymbol{z}-\boldsymbol{z}_k)\right)\right)^2\mathrm{d}\boldsymbol{z}\\ &= \frac{1}{n^2|\mathsf{H}|}\left\{\int\left[\sum_{g_i=l-1}K\left(\mathsf{H}^{-\frac{1}{2}}(\boldsymbol{z}-\boldsymbol{z}_i)\right)\right]^2\mathrm{d}\boldsymbol{z}\right.\\ &\quad+\int\left[\sum_{g_i=s-1}K\left(\mathsf{H}^{-\frac{1}{2}}(\boldsymbol{z}-\boldsymbol{z}_k)\right)\right]^2\mathrm{d}\boldsymbol{z}\\ &\quad-2\int\sum_{g_i=l-1}K\left(\mathsf{H}^{-\frac{1}{2}}(\boldsymbol{z}-\boldsymbol{z}_i)\right)\\ &\quad\left.\sum_{g_i=s-1}K\left(\mathsf{H}^{-\frac{1}{2}}(\boldsymbol{z}-\boldsymbol{z}_k)\right)\mathrm{d}\boldsymbol{z}\right\}\\ &= \frac{1}{n^2|\mathsf{H}|}\left[\mathrm{sum}(\mathsf{W}_{l,l})+\mathrm{sum}(\mathsf{W}_{s,s})-2\mathrm{sum}(\mathsf{W}_{l,s})\right],\end{aligned}$$

where the matrix operator $\mathrm{sum}(\mathsf{A})=\sum_{i,j}\mathsf{A}(i,j)$ is defined as the summation of all the entries of a matrix. Then, $B_{\mathsf{H}}(\boldsymbol{g})$ can be computed as

$$\begin{aligned}B_{\mathsf{H}}(\boldsymbol{g}) = \max_{\substack{l,s=1,\ldots,L\\ l\neq s}}\frac{1}{n^2|\mathsf{H}|}&\left[\mathrm{sum}(\mathsf{W}_{l,l})\right.\\ &\left.+\mathrm{sum}(\mathsf{W}_{s,s})-2\mathrm{sum}(\mathsf{W}_{l,s})\right],\end{aligned} \tag{10}$$

where the matrix W is a symmetric matrix of size $N\times N$, with elements defined as,

$$\mathsf{W}(i,i) = \int K\left(\mathsf{H}^{-1/2}(\boldsymbol{z}-\boldsymbol{z}_i)\right)^2\mathrm{d}\boldsymbol{z}, \tag{11}$$

$$\mathsf{W}(i,j) = \mathsf{W}(j,i) = \int K\left(\mathsf{H}^{-1/2}(\boldsymbol{z}-\boldsymbol{z}_i)\right) K\left(\mathsf{H}^{-1/2}(\boldsymbol{z}-\boldsymbol{z}_j)\right) \mathrm{d}\boldsymbol{z}. \tag{12}$$

It is well-known that the choice of kernel function K is not essential to the KDE (Silverman, 1986). To illustrate the partition method, we choose the commonly used multivariate Gaussian kernel $K(\boldsymbol{x}) = (2\pi)^{-d/2}\exp(-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x})$. The entries of W using the Gaussian kernel can be calculated analytically,

$$\begin{aligned}
&\mathsf{W}(i,j)\\
&= \int_{\mathbb{R}^d} K\left(\mathsf{H}^{-\frac{1}{2}}(\boldsymbol{z}-\boldsymbol{z}_i)\right) K\left(\mathsf{H}^{-\frac{1}{2}}(\boldsymbol{z}-\boldsymbol{z}_j)\right) \mathrm{d}\boldsymbol{z}\\
&= \int_{\mathbb{R}^d} (2\pi)^{-d} \exp\left(-\frac{1}{2}\left[(\boldsymbol{z}-\boldsymbol{z}_i)'\mathsf{H}^{-1}(\boldsymbol{z}-\boldsymbol{z}_i)\right.\right.\\
&\left.\left.+(\boldsymbol{z}-\boldsymbol{z}_j)'\mathsf{H}^{-1}(\boldsymbol{z}-\boldsymbol{z}_j)\right]\right) \mathrm{d}\boldsymbol{z}\\
&= (2\pi)^{-d} \int_{\mathbb{R}^p} \exp\left(-\left(\boldsymbol{z}-\frac{\boldsymbol{z}_i+\boldsymbol{z}_j}{2}\right)' \mathsf{H}^{-1}\right.\\
&\left.\left(\boldsymbol{z}-\frac{\boldsymbol{z}_i+\boldsymbol{z}_j}{2}\right) - \frac{1}{4}(\boldsymbol{z}_i-\boldsymbol{z}_j)'\mathsf{H}^{-1}(\boldsymbol{z}_i-\boldsymbol{z}_j)\right) \mathrm{d}\boldsymbol{z}\\
&= 2^{-d}\pi^{-\frac{d}{2}}|\mathsf{H}|^{\frac{1}{2}} \mathrm{e}^{\left(-\frac{1}{4}(\boldsymbol{z}_i-\boldsymbol{z}_j)'\mathsf{H}^{-1}(\boldsymbol{z}_i-\boldsymbol{z}_j)\right)}.
\end{aligned}$$

As a special case, $\mathsf{W}(i,i) = |\mathsf{H}|^{\frac{1}{2}}2^{-d}\pi^{-\frac{d}{2}}$. Note that the above calculation applies when the domain of $\boldsymbol{z}$, denoted by Ω, is unbounded, i.e. $\Omega = \mathbb{R}^d$. If Ω is a subset of $\mathbb{R}^d$, we can derive the integration in the range of Ω, and the resulting formula would involve the CDF of normal. But here we still integrate with the range of $\mathbb{R}^d$, and the approximation error is small since the value of the estimated density function should be small outside of Ω.

Given a specific partition $\boldsymbol{g}$, we partition the W matrix into $L \times L$ sub-matrices accordingly, such that each sub-matrix $\mathsf{W}_{s,r}$ corresponds to the experimental units in group r and s. That is, the entries of sub-matrix $\mathsf{W}_{r,s}$ are $\mathsf{W}(i,j)$ such that $g_i = r-1$ and $g_j = s-1$ for $r,s = 1,\ldots,L$. Notice that such a definition of the block matrices depends on the partition $\boldsymbol{g}$. Thus, the entries of the sub-matrices would change as the partition is varied. But the entries of the W for each pair of $(\boldsymbol{z}_i,\boldsymbol{z}_j)$ remain the same for all $i,j = 1,\ldots,N$. So the entries of the matrix W only need to be computed once for computing $B_{\mathsf{H}}(\boldsymbol{g})$ values for different partitions.

For $L = 2$, the objective function $B_{\mathsf{H}}(\boldsymbol{g}) = \frac{1}{n^2|\mathsf{H}|}[\mathrm{sum}(\mathsf{W}_{1,1}) + \mathrm{sum}(\mathsf{W}_{2,2}) - 2\mathrm{sum}(\mathsf{W}_{1,2})]$. Recall that by the definition of partition vector $\boldsymbol{g} = [g_1,\ldots,g_N]^\top$, $g_i = 0$ if the ith experimental unit is in Group 1, and $g_i = 1$ if the ith experimental unit is in Group 2. Then, the objective function $B_{\mathsf{H}}(\boldsymbol{g})$ could be rewritten as

$$\begin{aligned}
B_{\mathsf{H}}(\boldsymbol{g}) &= \sum_{i=1}^N\sum_{j=1}^N (2g_i-1)\mathsf{W}(i,j)(2g_j-1)\\
&= 4\left[\sum_{i=1}^N\sum_{j=1}^N g_i g_j \mathsf{W}(i,j) - \sum_{i=1}^N g_i \sum_{j=1}^N \mathsf{W}(i,j)\right]\\
&\quad + \sum_{i=1}^N\sum_{j=1}^N \mathsf{W}(i,j)\\
&= 4(\boldsymbol{g}^\top\mathsf{W}\boldsymbol{g} - \boldsymbol{g}^\top\boldsymbol{w}) + \sum_{i=1}^N\sum_{j=1}^N \mathsf{W}(i,j),
\end{aligned} \tag{13}$$

where $\boldsymbol{w} = [\sum_{j=1}^N \mathsf{W}(1,j),\ldots,\sum_{j=1}^N \mathsf{W}(N,j)]^\top$. As a result, for $L = 2$, the optimisation problem (9) is reformulated into

$$\min_{\boldsymbol{g}} \quad \boldsymbol{g}^\top\mathsf{W}\boldsymbol{g} - \boldsymbol{g}^\top\boldsymbol{w} \tag{14a}$$

$$\text{s.t.} \quad \sum_{i=1}^N g_i = N/2, \tag{14b}$$

$$g_i \quad \text{binary},\ i = 1,\ldots,N, \tag{14c}$$

This is a quadratic integer programming that can be solved efficiently by Gurobi Optimiser (Gurobi Optimization, LLC, 2020) for small- or moderately-sized experiments. For large-sized experiments or the more general case of $L \geq 3$, stochastic optimisation tools such as genetic algorithm (Miller & Goldberg, 1995) and simulated annealing (Van Laarhoven & Aarts, 1987) can be adopted to solve the optimisation. Regardless of the optimisation method, the matrix W is computed only once in the optimisation procedure, which significantly cuts down the computation.

4.2. Choice of bandwidth matrix

The accuracy of the KDE is sensitive to the choice of bandwidth matrix H (Simonoff, 2012; Wand & Jones, 1993). Many methods have been developed to construct H under various criteria (de Lima & Atuncar, 2011; Duong & Hazelton, 2005; Jones et al., 1996; Sain et al., 1994; Sheather & Jones, 1991; Wand & Jones, 1994; Zhang et al., 2006). Besides these methods, H can be chosen by some rules of thumb, including Silverman's rule of thumb (Silverman, 1986) and Scott's rule (Scott, 2015). But they may lead to a suboptimal KDE (Duong & Hazelton, 2003; Wand & Jones, 1993) due to the diagonality constraint of H. Another rule of thumb uses a full bandwidth matrix as

$$\mathsf{H} = n^{-2/(d+4)}\hat{\Sigma}, \tag{15}$$

where $\hat{\Sigma}$ is the estimated covariance matrix with sample size n, and d is the covariate dimension. It can be considered as a generalisation of Scott's rule (Härdle et al., 2012). To compromise between the computational cost and the accuracy of the KDE, we propose using the rule of thumb in (15). It is easy to compute and leads to a more accurate KDE compared to the

simple diagonal matrix. We did a series of numerical comparisons to test the impact of H on the KDE-based partition (not reported here due to the space limit). The results show that (15) leads to similar partitions compared with the more computationally demanding methods such as cross-validation (Duong & Hazelton, 2005; Sain et al., 1994) and Bayesian methods (de Lima & Atuncar, 2011; Zhang et al., 2006). Similar observations were found in the work by Anderson et al. (1994). As pointed out in Anderson et al. (1994), the criterion $B_{\mathsf{H}}(\boldsymbol{g})$ aims at measuring the discrepancy between the distributions from which the samples are drawn, but not precisely estimating those distributions. Although the bandwidth matrix H plays an important role in the estimation of distribution, it is not surprising that the KDE-based partition is robust to the choice of H.

5. Examples

Example 1. Simulation Example. We compare the performance of the proposed KDE-based partition with complete randomisation and rerandomisation (Morgan & Rubin, 2012) through a simulation example with $d = 2$ covariates. Three different types of mean function h are considered.

Model 1. Linear basis: $h(\boldsymbol{z}) = \beta_0 + \sum_{i=1}^{2} \beta_i z_i$,

$$y = \alpha x + \beta_0 + \sum_{i=1}^{2} \beta_i z_i + \epsilon.$$

Model 2. Quadratic basis: $h(\boldsymbol{z}) = \beta_0 + \sum_{i=1}^{2} \beta_i z_i + \sum_{i=1}^{2} \gamma_i z_i^2 + \theta z_1 z_2$,

$$y = \alpha x + \beta_0 + \sum_{i=1}^{2} \beta_i z_i + \sum_{i=1}^{2} \gamma_i z_i^2 + \theta z_1 z_2 + \epsilon.$$

Model 3 Sinusoidal model: $h(\boldsymbol{z}) = \beta_0 + \beta_1 \sin(\phi + \pi\gamma_1 z_1 + \pi\gamma_2 z_2)$,

$$y = \alpha x + \beta_0 + \beta_1 \sin(\phi + \pi\gamma_1 z_1 + \pi\gamma_2 z_2) + \epsilon.$$

The notation $x \in \{0, 1\}$ is the indicator of the treatment level assignment. To generate data from these models, we need to specify the values of the parameters, including the treatment effect α, and others β_i's, γ_i's, ϕ and θ. Let $\alpha = 2$ for all models. In Model 1 and Model 2, the regression coefficients β_i's, γ_i's and θ are sampled from a uniform distribution $U[-2, 2]$. In Model 3, β_0, β_1, γ_1 and γ_2 are randomly generated from $U[-1, 1]$, and ϕ is randomly generated from $U[0, 2\pi]$. The observed covariates are generated from multivariate standard normal distribution $N(\mathbf{0}, \boldsymbol{I}_2)$. All these values are fixed through the simulations. Since $\text{var}(\hat{\alpha})$ is only affected by the partition methods and is invariant to the variance of the noise σ^2, it does not matter to the comparison of different methods. Therefore in this and the next example, we set $\sigma = 0$.

To compare the proposed approach with other random methods, $m = 1000$ random partitions are generated via complete randomisation and rerandomisation approaches. For each partition, the treatment levels are randomly assigned to the two groups, and $m = 1000$ designs are generated for the two random approaches. Since the proposed KDE-partition is an optimisation-based method, the partition is deterministic. With all possible treatment-level assignments, there are only two designs. For each design obtained from the three methods, we generate the response data from the three models with the fixed parameters and covariates. Then, the estimated mean squared error $\widehat{\text{MSE}}(\hat{\alpha}) = \frac{1}{m}\sum_{i=1}^{m}(\alpha - \hat{\alpha})^2$ of the difference-in-mean estimator is calculated for increasing sample size from $N = 20$ to $N = 100$, and they are plotted in Figure 1.When the true relationship between the response y and covariate $\boldsymbol{z}$ is linear, rerandomisation outperforms the other two methods. Morgan and Rubin (2012) showed that, compared to complete randomisation, the rerandomisation with Mahalanobis distance criterion can reduce $\text{var}(\hat{\alpha})$ significantly when the mean function $h(\boldsymbol{z}) = \sum_{i=0}^{d} \beta_j z_j$ contains only the linear terms of the covariates. When a more complicated relationship such as Model 2 or 3 is considered, the KDE-based partition outperforms the complete randomisation and rerandomisation by a large margin. In practice, the true mean function h is rarely as simple as a linear function and usually contains higher-order terms of the covariates. Thus, from a practical perspective, we suggest that the KDE-based partition is a better choice.

We further explore the performance of the proposed KDE-based partition in matching the empirical distributions of the covariates in two groups. We compare the difference of the empirical distributions under complete randomisation, rerandomisation, and KDE-based partition. The discrepancy of the first and second raw moments of the two empirical distributions over the $m = 1000$ partitions are calculated and reported in Table 1. In general, rerandomisation performs the best in matching the means of the empirical distributions, which to some extent implies that it performs the best under *Model 1* (the model with main effect only). The KDE-based partition consistently outperforms the complete randomisation and is superior to the other two methods for the second moments since it matches the approximated density functions rather than just the first and second moments.

Example 2. Real Data Example. We compare the KDE-based partition with the complete randomisation and rerandomisation using a real data set. The data set is from a diabetes study (Efron et al., 2004). It contains 422 observations of $d = 10$ covariates and a univariate response. The covariates are age, sex, body mass index,

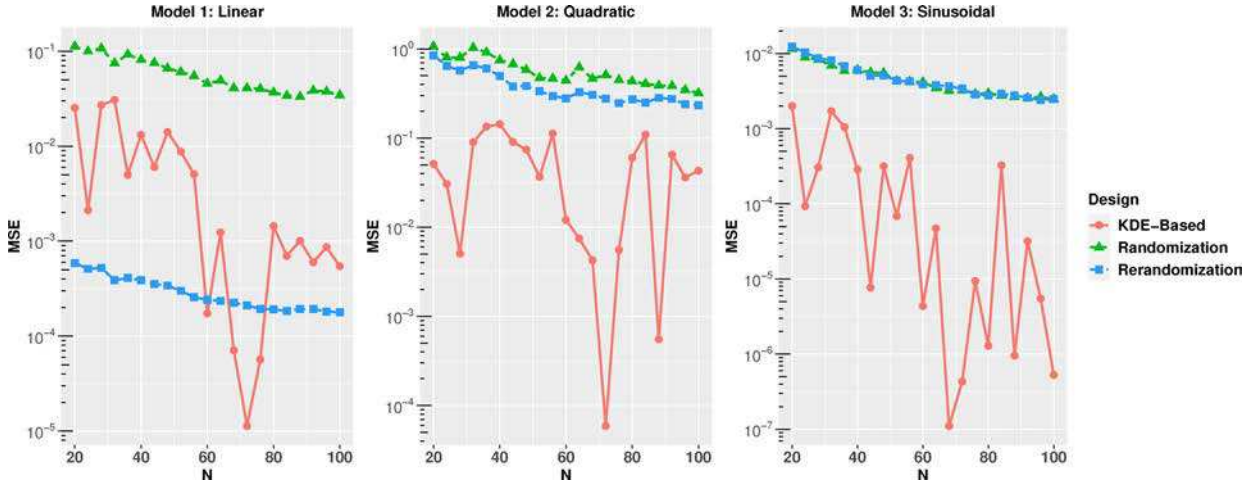

Figure 1. Comparison of the estimated mean squared error of difference-in-mean estimator using three partition methods for Example 1.

Table 1. Discrepancy of moments under different partition methods.

		Moments				
N	Method	z_1	z_2	z_1^2	z_2^2	z_1z_2
20	Random	0.356	0.298	0.605	0.292	0.217
	Re-random	0.026	0.021	0.612	0.286	0.235
	KDE-based	0.107	0.179	0.390	0.082	0.010
40	Random	0.261	0.235	0.356	0.212	0.249
	Re-random	0.019	0.017	0.366	0.233	0.258
	KDE-based	0.177	0.011	0.097	0.071	0.272
60	Random	0.204	0.182	0.278	0.181	0.189
	Re-random	0.015	0.014	0.282	0.188	0.192
	KDE-based	0.048	0.040	0.036	0.006	0.164
80	Random	0.170	0.177	0.234	0.201	0.164
	Re-random	0.013	0.013	0.238	0.210	0.163
	KDE-based	0.031	0.122	0.182	0.144	0.003
100	Random	0.167	0.150	0.238	0.193	0.161
	Re-random	0.013	0.011	0.244	0.198	0.158
	KDE-based	0.079	0.063	0.136	0.180	0.068

average blood pressure, and six blood serum measurements of the patients, and the response is a quantitative measure of disease progression. The data are only observational data and do not contain any experimental factors.

To use this data, we do not assume any functional form for $h(\boldsymbol{z})$. Instead, we assume the observed quantitative measure of disease progression is the sum $h(\boldsymbol{z}) + \epsilon$. Let the true value of the treatment effect $\alpha = 2$. Given treatment assignment x, the response data including the treatment effect is, $y = \alpha x + h(\boldsymbol{z}) + \epsilon$.

For different values of N, ranging from $N = 12$ to $N = 60$, $m = 1000$ partitions are generated using complete randomisation and rerandomisation. As explained before, for each N value, the optimal KDE-based partition is deterministic. For each of the partitions obtained from the three methods, we randomly assign treatment settings to the two partitioned groups to obtain the design, and then compute the response y accordingly. The estimated mean squared error $\widehat{\mathrm{MSE}}(\hat{\alpha}) = \frac{1}{m}\sum_{i=1}^{m}(\alpha - \hat{\alpha})^2$ of the difference-in-mean estimator is calculated for each N. They are shown in Figure 2. The KDE-based partition outperforms the other two partition methods for all sample sizes.

6. Discussion

In this paper, we introduce a KDE-based partition method for the controlled experiments. By adopting a smooth approximation of the covariate empirical distributions, we propose a new covariate balancing criterion. It measures the difference between the distributions of covariates in the partitioned groups. We use quadratic integer programming to construct the partition that minimises the covariate balancing criterion for the two-level experiments. If the number of treatment settings is more than two, other stochastic optimisation methods can be applied. The design generated via the KDE-partition can regulate the variance of the difference-in-mean estimator. Compared with the complete randomisation and rerandomisation methods, the simulation and real examples show that the proposed method leads to a more accurate difference-in-mean estimation of the treatment effect when the underlying model involves more complicated functions of the covariates. The simulation example also confirms that the covariates' distributions of the groups are better matched using the proposed method.

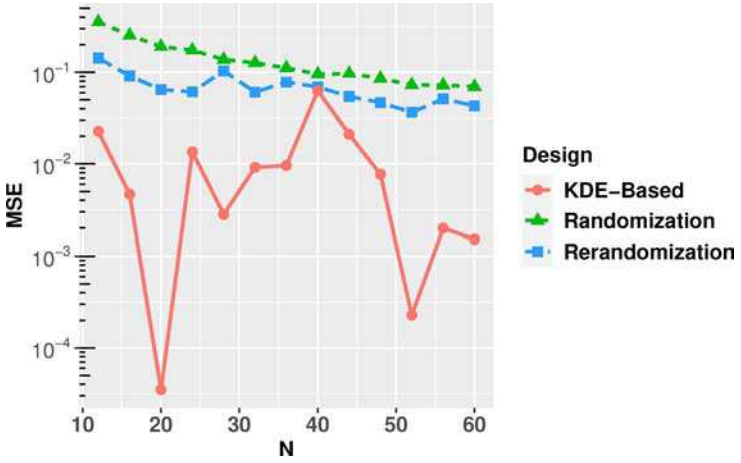

Figure 2. Comparison of the estimated mean squared error of difference-in-mean estimator using three partition methods for Example 2.

It is worth pointing out that, when the KDE-based partition is used, the classical hypothesis testing procedure for the difference-in-mean estimator is not applicable, since the partition is a deterministic solution and the random treatment assignments only provide two different designs when $L = 2$. Fortunately, a sophisticated testing procedure using bootstrap method has been established and proven to be powerful (Bertsimas et al., 2015) for the sharp null hypothesis (Rubin, 1980), H_0 : alltreatmenteffectsare0. The detailed bootstrap algorithm is in Algorithm 1.

Algorithm 1 Hypothesis testing procedure

procedure HYPOTHESIS TESTING OF TREATMENT EFFECT
  Construct KDE-based partition, randomly assign treatment levels to treatment groups, apply treatments, measure the responses y_i, $i = 1, \ldots, N$ and compute the difference-in-mean estimate $\hat{\alpha}$.
  for $t = 1, \ldots, T$ **do**
    sample $i_j^t \sim$ unif $(1, \ldots, N)$ independently for $j = 1, \ldots, N$,
    construct KDE-based partition for $\boldsymbol{z}_{i_1^t}, \ldots, \boldsymbol{z}_{i_N^t}$ and
    compute the new difference-in-mean estimate $\hat{\alpha}^t$
  end for
  the p-value of H_0 is $p = \frac{1+\sum_{t=1}^{T} \mathbb{1}_{\{|\hat{\alpha}^t| \geq |\hat{\alpha}|\}}}{1+T}$.
end procedure

One limitation of the proposed KDE-based partition is the 'curse of dimensionality'. It is well-known that KDE may perform poorly when the dimension of the covariate is large relative to the sample size. To overcome this problem, the experimenter can add a dimension reduction step prior to the partition of the experimental units. Based on the properties of the covariate data, one can choose the appropriate dimension reduction method from various choices, such as the principal component analysis (PCA) and the nonlinear variants of PCA. For instance, using PCA, the experimenter can select a small but sufficient number of the principal components and apply the KDE-based partition on the linearly transformed covariates of a much lower dimension. We also want to alert the readers with another limitation of the KDE-based partition. Similar to other optimal covariates balancing ideas, such as Bertsimas et al. (2015) and Kallus (2018), the KDE-based partition assumes all the influential covariates to the response are known to the experimenter and their data are included in the observed covariates data. Otherwise, if there are latent but important covariates, the optimal partition methods, including the proposed KDE-based partition, might lead to an estimator with large variance because it deterministically balances the experimental units based on incomplete covariate information. In this case, we recommend the randomisation or rerandomisation methods.

The proposed KDE-based partition method can be used in other scenarios beyond controlled experiments. Essentially, we have proposed a density-based partition method that minimises the differences of data between groups. It can be incorporated into any statistical tool that needs to partition data into similar groups, such as cross-validation, divide-and-conquer, etc. We hope to explore these directions in the future. In this work, we do not assume any interaction terms between the covariates and the treatment effect. However, interaction effects are likely to occur in practice. Another interesting direction is the partition of the experimental units considering the interaction terms in the model.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Division of Mathematical Sciences [grant number 1916467].

Notes on contributors

Dr Yiou Li is an Assistant Professor of the Department of Mathematical Sciences at DePaul University. She obtained her MS in Mathematical Finance and PhD in Applied Mathematics from the college of computing at Illinois Institute of Technology. Dr. Li's research focuses on various areas in Statistics, including statistical design and analysis of physical and computer experiments, statistical modeling and data analysis in economics and finance, etc. She has publications and submitted papers in statistical journals including *International Journal for Uncertainty Quantification*, *Journal of Statistical Planning and Inference*, *Canadian Journal of Statistics*, *Statistica Sinica*, etc.

Dr Lulu Kang is an Associate Professor of the Department of Applied Math at Illinois Institute of Technology (IIT). She obtained her MS in Operations Research and PhD in Industrial Engineering from the Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology. Dr Kang has worked on various areas in Statistics, including uncertainty quantification, statistical design and analysis of experiments, Bayesian computational statistics, etc. She has publications and submitted papers in top statistical journals including *Techometrics*, *SIAM/ASA Journal on Uncertainty Quantification*, *Statistica Sinica*, etc. Dr Kang is currently the associate editor for journals *SIAM/ASA Journal on Uncertainty Quantification and Technometrics*.

Dr Xiao Huang obtained his PhD degree in Applied Mathematics and a Master in Mathematical Finance from Illinois Institute of Technology. His PhD advisor was Dr Lulu Kang. He collaborated with Dr Lulu Kang and Dr Yiou Li on this work during his PhD study at Illinois Tech. After graduation from Illinois Tech in 2019, Dr Huang has worked as a Data Scientist at United Airlines.

ORCID

Lulu Kang http://orcid.org/0000-0002-6000-3436

References

Anderson, N. H., Hall, P., & Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, *50*(1), 41–54. https://doi.org/10.1006/jmva.1994.1033

Bernstein, S. (1927). Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, *97*(1), 1–59. https://doi.org/10.1007/BF01447859

Bertsimas, D., Johnson, M., & Kallus, N. (2015). The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, *63*, 868–876. https://doi.org/10.1287/opre.2015.1361

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). cem: Coarsened exact matching in Stata. *The Stata Journal*, *9*(4), 524–546. https://doi.org/10.1177/1536867X0900900402

de Lima, M. S., & G. S. Atuncar (2011). A Bayesian method to estimate the optimal bandwidth for multivariate kernel estimator. *Journal of Nonparametric Statistics*, *23*(1), 137–148. https://doi.org/10.1080/10485252.2010.485200

Duong, T., & Hazelton, M. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, *15*(1), 17–30. https://doi.org/10.1080/10485250306039

Duong, T., & Hazelton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, *32*(3), 485–506. https://doi.org/10.1111/sjos.2005.32.issue-3

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499. https://doi.org/10.1214/009053604000000067

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, *173*(7), 761–767. https://doi.org/10.1093/aje/kwq439

Gurobi Optimization, LLC (2020). Gurobi optimizer reference manual.

Härdle, W. K., Müller, M., Sperlich, S., & Werwatz, A. (2012). *Nonparametric and semiparametric models*. Springer Science & Business Media.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, *11*, 337–381.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(1), 85–112. https://doi.org/10.1111/rssb.12240

McHugh, R., & Matts, J. (1983). Post-stratification in the randomized clinical trial. *Biometrics*, *39*(1), 217–225. https://doi.org/10.2307/2530821

Miller, B. L., & Goldberg, D. E. (1995). Genetic algorithms, tournament selection, and the effects of noise. *Complex Systems*, *9*, 193–212.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *40*(2), 1263–1282. https://doi.org/10.1214/12-AOS1008

Morgan, K. L., & Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, *110*(512), 1412–1421. https://doi.org/10.1080/01621459.2015.1079528

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Rosenbaum, P. (2017). *Observation and experiment: An introduction to causal inference*. Harvard University Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, *75*, 591–593.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Sain, S. R., Baggerly, K. A., & Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association*, *89*(427), 807–817. https://doi.org/10.1080/01621459.1994.10476814

Scott, D. W. (2015). *Multivariate density estimation: Theory, practice, and visualization*. Wiley.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(3), 683–690. https://doi.org/10.1111/rssb.1991.53.issue-3

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, vol. CRC Press.

Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

Van Laarhoven, P. J., & Aarts, E. H. (1987). Simulated annealing. In *Simulated annealing: Theory and applications* (pp. 7–15). Springer.

Wand, M. P., & Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, *88*(422), 520–528. https://doi.org/10.1080/01621459.1993.10476303

Wand, M. P., & Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, *9*, 97–116.

Wu, C. J., & Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*, vol. Wiley.

Xie, H., & Aurisset, J. (2016). Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 645–654). ACM.

Zhang, X., King, M. L., & Hyndman, R. J. (2006). A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, *50*(11), 3009–3031. https://doi.org/10.1016/j.csda.2005.06.019