Batch effect correction with sample remeasurement in highly confounded case-control studies

Hanxuan Ye[1], Xianyang Zhang[1*], Chen Wang[2], Ellen L. Goode[2] and Jun Chen[2*]

[1]Department of Statistics, Texas A&M University, 155 Ireland Street, College Station, 77843, TX, USA.
[2]Department of Quantitative Health Sciences, Mayo Clinic, 200 First Street SW, Rochester, 55905, MN, USA.

*Corresponding author(s). E-mail(s): zhangxiany@stat.tamu.edu; chen.jun2@mayo.edu;
Contributing authors: hanxuan@tamu.edu; wang.chen@mayo.edu; egoode@mayo.edu;

Abstract

Batch effects are pervasive in biomedical studies. One approach to address the batch effects is repeatedly measuring a subset of samples in each batch. These remeasured samples are used to estimate and correct the batch effects. However, rigorous statistical methods for batch effect correction with remeasured samples are severely underdeveloped. In this study, we developed a framework for batch effect correction using remeasured samples in highly confounded case-control studies. We provided theoretical analyses of the proposed procedure, evaluated its power characteristics, and provided a power calculation tool to aid in the study design. We found that the number of samples that need to be remeasured depends strongly on the between-batch correlation. When the correlation is high, remeasuring a small subset of samples is possible to rescue most of the power.

1 Introduction

One major issue facing biological studies is that biological measurement is highly susceptible to non-biological experimental variation or "batch effects". Batch effects are pervasive in modern high-throughput omics technologies using microarrays or next-generation sequencing [1, 2]. Different experimental conditions, measurement modalities, personnel executing the experiments, and batches of reagents all contribute to batch effects. Such unwanted variation has severe statistical consequences. It could reduce statistical power by introducing extra variation or, more seriously, lead to false findings if the batch effects are confounded with the effects of interest. Although performing the biological measurement in a single batch is the most effective way to reduce batch effects, such practice may not always be possible due to various constraints such as resource availability and measuring capacity. Even if the experimental measurement is executed in a single batch, some unexpected batch effects could still arise. For example, different measuring chips, locations on the chips, DNA extraction plates, and sequencing lanes have all been found to produce batch effects in omics studies [3–5]. Therefore, addressing the batch effects in the study design and data analysis is critical to improve the statistical power, increase the robustness of the findings and reduce the developmental cost.

Over the past two decades, a number of batch effect correction methods have been developed and applied in practical data analysis. Two mainstreams for batch effect correction are location-scale (LS) matching and matrix factorization (MF). The LS methods assume the sources of the batch effects are known so that the location (e.g., mean), scale (e.g., standard deviation), or even the entire distribution are matched across batches. Methods in this category include batch mean-centering (BMC) [6], gene-wise standardization (SD) [7], ComBat [8, 9], cross-platform normalization (XPN) [10] and distance-weighted discrimination (DWD) [11]. Among these, ComBat, an empirical Bayes-based LS method, is the most widely used method due to its robustness to small batch sizes compared with earlier methods [8, 9]. In contrast, the MF-based methods do not require the sources of batch effects are known in advance. Instead, they search for directions of maximal variance associated with the batch effects and use the resulting latent factors to correct for batch effects. Methods in this category include singular value decomposition (SVD)/principal component analysis (PCA) [12, 13], surrogate variable analysis (SVA) [14], RUV [15–17] and LEAPP [18]. The SVA, RUV, and LEAPP have also been studied and expanded within a unified CATE rotation [19] framework that adjusts for the confounders in hypothesis testing.

Previous efforts for batch effects correction have been focused on estimating and correcting batch effects based on independent samples [8, 9, 14, 18, 19]. In practice, however, one intuitive approach used by investigators to address batch effects is through remeasuring a subset of samples in each batch in the hope that these remeasured samples could be used to estimate and correct the batch effects [20, 21]. Unfortunately, other than some simple approaches, statistical methods for batch correction using the remeasured samples remain

severely underdeveloped. Biostatisticians are often faced with the inability to efficiently utilize these remeasured samples in the analysis to correct for batch effects, hindering the successful completion of the proposed studies. To fill the methodological gap, this study investigates the feasibility and methodology for batch effect correction using remeasured samples in a highly confounded case-control study [22]. We specifically consider a challenging scenario, where an investigator has collected all the case samples, and she wants to compare these case samples to the control samples that have already been measured previously and a subset of which are still available for remeasurement. This scenario is quite common in clinical settings since clinical investigators usually obtain case samples more easily than control samples. For example, an investigator wants to compare her case samples to the control samples from the institutional biobank [23]. Oftentimes, the biobank samples have already been characterized in a standalone study or have been used as controls in other disease studies, resulting in a large amount of pre-existing control data that can be potentially used together with the new control data generated from remeasuring a subset of the biobank samples. Another example is the subsequent analysis in case-cohort studies [24]. One strength of case-cohort studies is that the subcohort can be used as a reference group for a variety of different case groups. The subcohort in a case-cohort study implemented early for a common disease can be used as a reference group for a series of rarer or long-latency disease. The data for the reference group already exists in study databases. For subsequent disease studies, the existing subcohort data may be re-used after a subset of the subcohort samples have been remeasured.

Obviously, if none of the control samples are remeasured, the biological effects will be completely confounded with the batch effects, and distinguishing between the biological and batch effects will be very difficult. Ideally, all the control samples need to be remeasured together with the case samples to maximize the discovery power. However, due to resource constraints and sample availability, such practice may not always be possible. Therefore, it will be of tremendous help if only remeasuring a small subset of control samples is necessary to correct the batch effects. Despite of a subject of critical importance, to our surprise, no dedicated statistical methods are available. No theoretical investigation has been performed to study the operating characteristics of batch effect correction with remeasured samples. It is unknown how many control samples need to be remeasured to recover most of the power, whether a handful of controls are sufficient to correct for batch effects, and what factors matter most in deciding the number of remeasured samples. A rigorous statistical testing method coupled with a power calculation tool is critically needed for this particular scenario. A successful tool could potentially rescue a completely confounded study and has a tremendous economic impact on the field.

In this study, we proposed a computationally efficient statistical method for batch effect correction with remeasured samples for a highly confounded case-control study. The method is based on the maximum likelihood framework, and

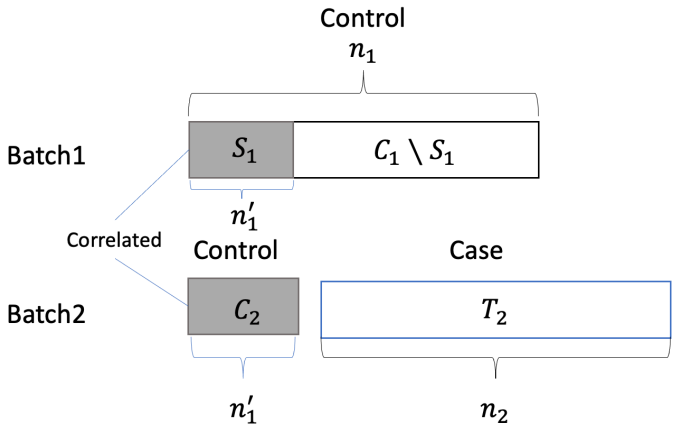


Fig. 1: **Illustration of the study design.** Here, C_1 denotes the set of n_1 control samples in Batch 1, and T_2 denotes the set of case samples in Batch 2. A subset S_1 consisting of n_1' samples from the C_1 set is remeasured. The set of these remeasured samples in Batch 2 is indicated as C_2. $C_1 \setminus S_1$ represents the unmeasured control samples.

hence the derived procedure is optimal in using the information available. We studied the theoretical properties of the procedure and proved the consistency and asymptotic normality of the resulting estimators. We investigated the power characteristics of the approach based on simulations and theoretical analysis, and identified statistical properties affecting its power. Finally, we proposed a power calculation tool to aid in the study design. A real dataset with known batch effects and a large number of remeasured samples was used to demonstrate the feasibility and efficiency of the proposed procedure.

2 Results

2.1 Problem Setup and Model

Consider that the control and case samples are measured on two different batches. We assume the linear model

$$\begin{aligned} y_i &= x_i(a_0 + a_1) + \mathbf{z}_i^\top \mathbf{b} + \epsilon_i, \\ \epsilon_i | x_i &\sim N(0, (1 - x_i)\sigma_1^2 + x_i\sigma_2^2), \end{aligned} \tag{1}$$

for $i = 1, 2, \ldots, n$, where y_i is the outcome, $x_i \in \{0, 1\}$ is the control/case group membership (0: control, 1: case), $\mathbf{z}_i$ contains measurements of other covariates including the intercept and possible covariate-batch (group) interactions, a_0 is coefficient for the true biological effect and a_1 is the coefficient for the nuisance batch effects. Since the batch effect and biological effect are indistinguishable

in this example, remeasurement of a subset of samples is thus necessary. Suppose the control and case samples are collected in the first and second batch, respectively, and a subset of control samples of size n' are remeasured in the second batch. Suppose the control and case group contain n_1 and n_2 samples, respectively. Without loss of generality, we assume that the first n_1' control samples are remeasured, where $n_1' \leq n_1$ (Figure 1). Then we have

$$\begin{aligned}
&\text{Control (batch 1):} \quad y_i = \mathbf{z}_i^\top \mathbf{b} + \epsilon_i^{(1)}, \quad i = 1, 2, \ldots, n_1, \\
&\text{Case (batch 2):} \quad y_i = a_0 + a_1 + \mathbf{z}_i^\top \mathbf{b} + \epsilon_i^{(2)}, \quad i = n_1 + 1, \ldots, n_1 + n_2 = n, \\
&\text{Control (batch 2):} \quad y_i = a_1 + \mathbf{z}_i^\top \mathbf{b} + \epsilon_i^{(2)}, \quad i = n + 1, \ldots, n + n_1',
\end{aligned}$$

where $\epsilon_i^{(1)} \sim N(0, \sigma_1^2)$ for $1 \leq i \leq n_1$, $\epsilon_i^{(2)} \sim N(0, \sigma_2^2)$ for $n_1 + 1 \leq i \leq n + n_1'$, and

$$\text{cov}(\epsilon_i^{(1)}, \epsilon_{n+i}^{(2)}) = \rho \sigma_1 \sigma_2$$

for $1 \leq i \leq n_1'$. The goal here is to develop an efficient procedure to test the null hypothesis that

$$H_0 : a_0 = 0,$$

i.e., there is no true biological effect.

We introduce some notation before describing the estimation and inference procedures. Denote by C_1 the set of control samples in the first batch and T_2 the set of case samples in the second batch. Let $S_1 = \{1, \ldots, n_1'\}$ and $C_2 = \{n+1, \ldots, n+n_1'\}$ be the subset of remeasured control samples in batch 1 and batch 2, respectively, where $|S_1| = |C_2| = n_1'$. See Figure 1 for illustration. Note that the covariates associated with the samples in S_1 and C_2 are the same. Let $\boldsymbol{\theta} = (a_0, a_1, \mathbf{b}, \rho, \sigma_1, \sigma_2)$ be the parameter vector to be estimated, and $\mathbf{N} = (n_1, n_2, n_1')$ be the vector of sample sizes. We define $\mu_{1i} = \mathbf{z}_i^\top \mathbf{b}$ for $i \in C_1$, $\mu_{2i} = a_0 + a_1 + \mathbf{z}_i^\top \mathbf{b}$ for $i \in T_2$ and $\mu_{3i} = a_1 + \mathbf{z}_i^\top \mathbf{b}$ for $i \in C_2$.

2.2 Simulation Studies

We conduct a set of simulation studies to investigate the finite sample performance of the proposed procedure in terms of estimation accuracy, type I error rate, and statistical power. Moreover, we compare our method ("ReMeasure") with three alternative procedures:

1. The location-scale matching approach ("LS", details in Supplementary Section 3).
2. Estimation and inference using only the second batch data ("Batch2").
3. Estimation and inference using the whole data set while ignoring the batch effects ("Ignore").

We study the effects of location and scale differences between the two batches, the between-batch correlations, and the number of remeasured samples. We generate the data according to the model in (1). Specifically, we set $n_1 = n_2 = 50$, and consider a univariate covariate z_i randomly drawn from

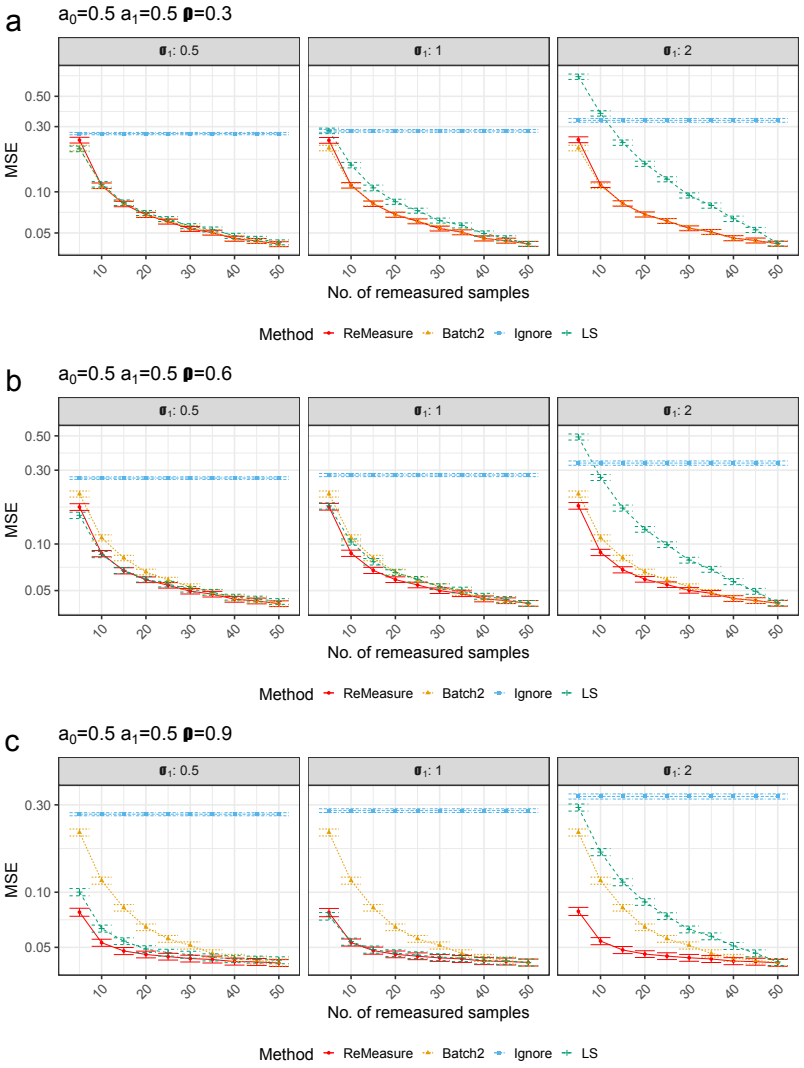


Fig. 2: **The mean square error (MSE) of a_0 estimate for different procedures when both sample sizes $n_1 = n_2 = 50$.** We vary the degrees of between-batch correlation (ρ values of 0.3, 0.6, and 0.9 for panels a, b, c, respectively) and degree of noise levels (σ_1, left to right). Both the biological effect parameter a_0 and batch location parameter a_1 are set to 0.5. For clarity, the y-axis is presented in $\log_{10}$ scale. Results are based on 1000 replications. Data are presented as mean values +/- SEM.

the standard normal distribution. We let $\mathbf{b} = -0.5$ and set $\sigma_2^2 = 1$ so that a_0 can be interpreted as the Cohen's d [25], an effect size measure for a two-sample t-test. Let ρ be the between-batch correlation for these remeasured

control samples. We investigate the batch scale parameter $\sigma_1^2 \in \{0.5^2, 1^2, 2^2\}$, the between-batch correlation $\rho \in \{0.3, 0.6, 0.9\}$, and the remeasured sample size $n_1' \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. We set the true biological effect $a_0 \in \{0, 0.5, 0.8\}$, representing no effect, moderate effect, and strong effect, according to Cohen's criterion. We found empirically (Supplementary Figure 1a) that the behavior of the proposed estimator of a_0 did not depend on the value of batch location effect a_1, so we thus set $a_1 = 0.5$ throughout the simulations.

In Figure 2 and Supplementary Table 1, we report the mean square error (MSE) of different procedures for estimating the biological effect a_0 when $a_0 = 0.5$. The MSE of a_0 estimate for other values shows the same pattern (data not shown). The method that ignores the batch effect and the remeasured samples ("Ignore") performs the worst in almost all settings. In contrast, the MSE for the other methods decreases with the number of remeasured samples but increases with σ_1^2. When the between-batch correlation ρ is small, the MSE of the method based on the second batch ("Batch2") is similar to that of the proposed method ("ReMeasure"), suggesting that the control samples in the first batch provide limited information when ρ is small. In this case, using the first batch of samples may only marginally improve the estimation efficiency. As ρ becomes larger, "Batch2" method begins to be less efficient. When the between-batch correlation is very high ($\rho = 0.9$), the control samples in the first batch help improve the estimation accuracy tremendously, and "ReMeasure" achieves a considerably smaller MSE even when the number of remeasured samples is small. The location-scale matching method ("LS"), on the other hand, has a much higher MSE than "ReMeasure" especially when the batch scale parameter for the first batch is large ($\sigma_1 = 2$). As the number of remeasured samples increases, the discrepancy decreases, indicating that a large number of remeasured samples may be needed for "LS" to work properly.

Next, we study the type I error rate and the statistical power (Figure 3). As expected, "Ignore" has the largest type I error inflation while "Batch2" controls the type I error across all settings. "LS" has severely inflated type I error when the number of remeasured samples is small, reflecting the large MSE observed. In contrast, the proposed method "ReMeasure" has much better type I error control than "LS" and it generally controls the type I error to the target level when $n_1' \geq 10$. However, when the number of remeasured samples is very small ($n_1' = 5$), "ReMeasure" has some type I error inflation. A larger between-batch correlation (ρ) reduces its inflation. The inflation is due to the use of the plug-in estimates of the variance components ($\sigma_1^2, \sigma_2^2, \rho$) in deriving the asymptotic distribution. When the number of the remeasured sample is small, the estimation of ρ is subject to large variability, and the asymptotic null distribution could deviate from the true null distribution. Indeed, if we plug in the true ρ in the test statistic instead of the estimated version ("Oracle" procedure), the type I error under $n_1' = 5$ is brought down close to the target level (Supplementary Figure 2a). In terms of statistical power, "ReMeasure" is similar to or slightly better than "Ignore" when ρ is small but is substantially

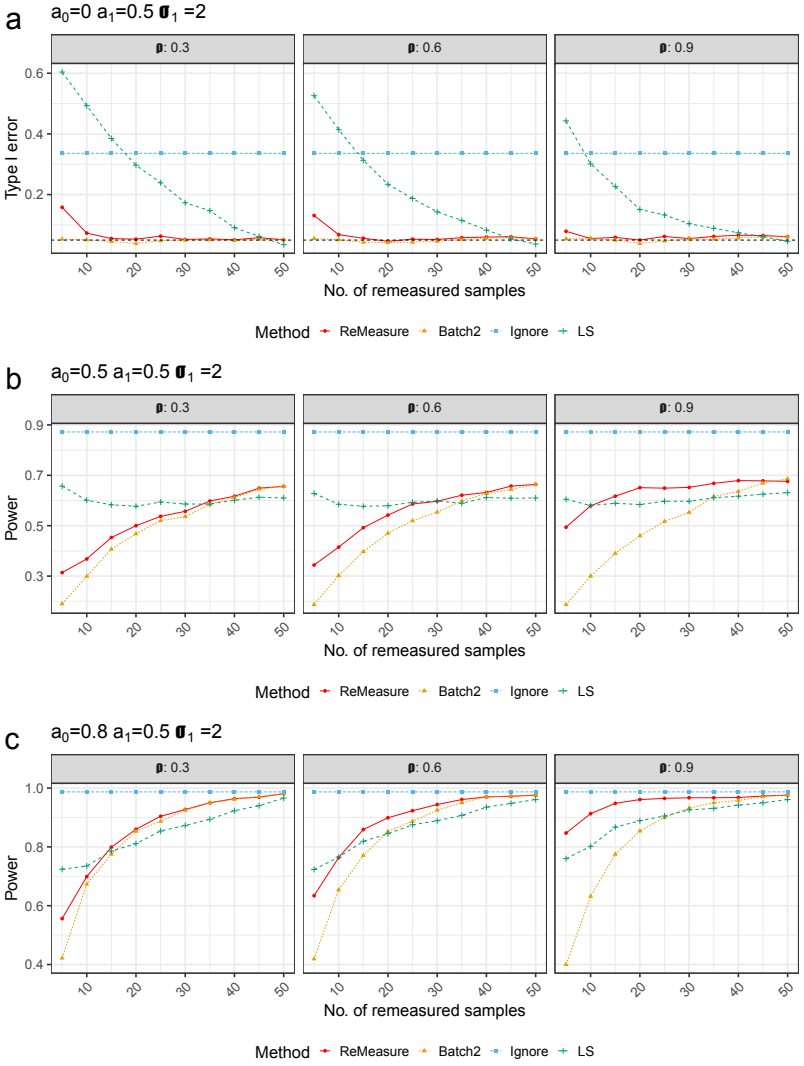


Fig. 3: **Evaluation of empirical type I error and power for different procedures in testing the biological effect $a_0 = 0$ with $n_1 = n_2 = 50$.** The true biological effects (a_0) we explored include 0 (type I error, panel a), 0.5 (power, panel b), and 0.8 (power, panel c). For each panel, from left to right, we increase the between-batch correlation (ρ) from 0.3 to 0.9. The batch location parameter a_1 is set to 0.5 and the batch scale parameter σ_1 is set to 2. The dashed line indicates the 5% nominal type I error rate used.

more powerful when ρ is large. The high power of "LS" and "Ignore" is not very meaningful since they have severe type I error inflation. We also compared

the performance under different σ_1 values, the patterns were almost identical (Supplementary Figure 1b).

To improve the type I error control under a small number of remeasured samples ($n_1' < 10$), we propose to use the bootstrap method to derive a more accurate null distribution. Supplementary Figure 3 shows that the bootstrap method could control the type I error at small n_1's across settings. However, the better type I error control is at the expense of some power and it is slightly less powerful than the asymptotic approach. When ρ is small, it may not have any advantage over the "Batch2" method. Therefore, the bootstrap method is only recommended for small n_1's when ρ is not small.

To demonstrate the robustness of the proposed method, we performed additional simulations under large sample sizes, different batch location parameters, and different error distributions. We also compared to two additional approaches: the naive approach, which fits a linear model based on all the samples adjusting the batch variable and ignoring repeated measurements, and the "LSind" approach, which is the "LS" method that uses the entire control samples to estimate the location and scale parameters. The results are summarized in Supplementary Section 4 ("Additional simulations").

2.3 Theoretical Power Analysis

In practice, one frequent question asked by an investigator is how many control samples need to be remeasured to achieve sufficient statistical power. Although the simulation-based approach can be used for power calculation, it is computationally intensive and is not amenable to large sample sizes. It also does not allow the exploration of different parameter settings flexibly. Therefore, an analytical power calculation tool is needed to aid in the study design. To achieve this end, we propose an approximate power calculator based on the asymptotic distribution. Specifically, the type I error and power can be calculated theoretically through the asymptotic normality of $\hat{a}_0$: $(\hat{a}_0 - a_0)/\mathrm{sd}(\hat{a}_0) \Rightarrow \mathcal{N}(0,1)$. The power $\Pr[|\hat{a}_0/\mathrm{sd}(\hat{a}_0)| > z_{1-\frac{\alpha}{2}}]$ for the significant level α can be calculated as

$$\Pr\left(\left|\frac{\hat{a}_0 - a_0}{\mathrm{sd}(\hat{a}_0)} + \frac{a_0}{\mathrm{sd}(\hat{a}_0)}\right| > z_{1-\frac{\alpha}{2}}\right) \approx \Pr\left(\left|Z + \frac{a_0}{\mathrm{sd}(\hat{a}_0)}\right| > z_{1-\frac{\alpha}{2}}\right), \quad (2)$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution and $Z \sim \mathcal{N}(0,1)$. In the theoretical power calculation, the oracle estimator for $\mathrm{sd}(\hat{a}_0)$ is used, where we assume that ρ, σ_1 and σ_2 are all known.

Supplementary Figure 2b provides a comparison between the theoretical power ("Theory") and the empirical power based on the asymptotic method ("ReMeasure"). The theoretical power does not deviate much from "ReMeasure" at different effect sizes. The approximation is more accurate when the number of remeasured samples is larger, and the between-batch correlation is higher. Thus, the theoretical power provides a reasonable approximation to the actual power when the proposed procedure is applied.

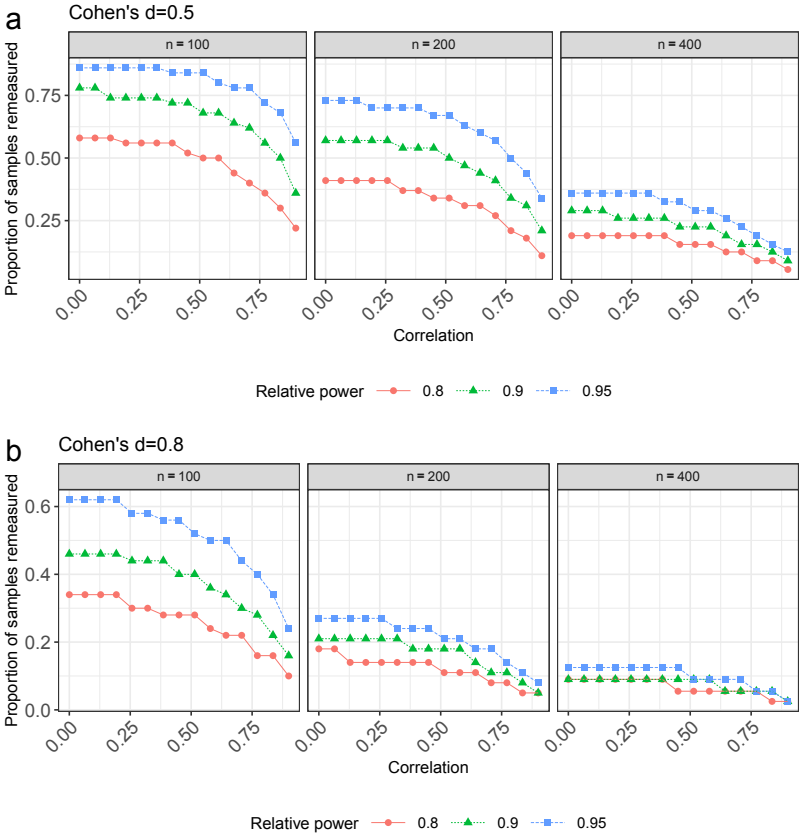


Fig. 4: **Proportion of control samples that need to be remeasured to achieve** $80\%, 90\%, 95\%$ **relative power vs. between-batch correlation** ρ **when** $n_1 = n_2 = 50, 100, 200$**.** We fix $a_1 = 0.5$ and consider settings where the effect size (Cohen's d) takes values 0.5 (panel a) and 0.8 (panel b), representing moderate and strong effects, respectively, according to Cohen's criterion. Results are derived from 500 replications.

With the theoretical power calculator, we can conduct power analysis under different parameter settings. Compared to the usual parameters used in power calculation for a two-sample t-test, such as the sample size of the control group n_1 and the case group n_2, the effect size a_0 (Cohen's d, mean difference standardized by the within-group standard deviation), significance level, and the desired power, power analysis for the proposed procedure depends on two additional parameters: the number of remeasured control samples n_1' and the between-batch correlation ρ. On the other hand, the batch location and scale parameters have little effect on power. Besides traditional power analyses such

as power vs. sample size and power vs. effect size, in our context, investigators are frequently interested in the following two types of power analysis:

- Given fixed sample sizes for the control and case group, how much power do we have at different numbers of remeasured samples?
- Given fixed sample sizes for the control and case group, how many control samples do we need to remeasure to recover, for example, 80% of the optimal power? The optimal power is defined as the power we can achieve by remeasuring all the control samples.

These questions can be easily answered by the theoretical power formula. To aid in study design, we provide an R Shiny app (https://hanxuan.shinyapps.io/PowerCalculation), which takes the user-supplied parameter values (sample size, effect size, between-batch correction, significance level) as the input and outputs the power at different numbers of remeasured samples. We provide both the absolute and relative power, where the absolute power is the statistical power in the traditional sense, i.e., the probability of rejecting the null hypothesis when the null hypothesis is false, and the relative power is the ratio of the absolute power to the optimal power defined above. Supplementary Figure 4 shows an example of power calculation for a confounded case-control study with sample remeasurement. In this example, both the case and control sample sizes are pre-fixed at 50, the expected between-batch correlation is 0.6, the effect size aimed to detect (Cohen's d) is 0.6, and the significance level used is 0.05. The Shiny app outputs a power curve at different numbers of remeasured samples, based on which we can see that 35 control samples need to be remeasured to achieve 80% absolute power (Supplementary Figure 4a) and 19 control samples need to be remeasured to achieve 80% of the optimal power (Supplementary Figure 4b).

Finally, we perform additional power analysis to gain more insights into the proposed procedure. Figure 4 shows the proportion of control samples that need to be remeasured to achieve 80%, 90%, 95% relative power at different sample sizes, effect sizes, and between-batch correlations. We can see that the larger the between-batch correlation, the smaller the number of samples that need to be remeasured to achieve desired relative power. The proportion of samples that need to be remeasured drops rapidly when the correlation is greater than 0.6.

2.4 Real Data Application

We next use a real dataset to illustrate the proposed method. The dataset came from two transcriptomics studies of ovarian cancer using different measurement platforms [26–28]. In the first study, the gene expression was profiled using Agilent micro-arrays [26, 27]. In the second study, the gene expression was profiled using RNA-Seq [28]. It is well known that different measurement platform creates strong batch effects for omics study [1]. A subset of the samples were profiled in both studies, which provides us the opportunity to evaluate the proposed method. In this analysis, we focused on high-grade

serous ovarian cancer, which is the most common type of ovarian cancer with well defined cancer subtypes [27, 29] (Agilent dataset $n = 306$, RNA-Seq dataset $n = 97$). There are 47 samples measured in both datasets. After intersecting the genes from the two platforms, we finally included 11,861 genes in the analysis. Based on these remeasured samples, we calculated the correlation of the gene expression between the two platforms. Supplementary Figure 5a shows that the distribution of the correlation coefficients has a wide range $(-0.47, 0.87)$ with a median correlation of 0.48. About 24% genes have a correlation larger than 0.6. The overall correlation is considered to be medium. To demonstrate the proposed method, we analyzed the cancer subtype variable (four subtypes: C1-MES, C2-IMM, C4-DIF, and C5-PRO) to identify subtype-specific gene signatures by comparing the expression profile of a specific subtype to that of the other subtypes. The Agilent dataset consists of 76, 77, 71, and 82 samples for C1-MES, C2-IMM, C4-DIF, and C5-PRO subtypes, respectively, while the RNA-Seq dataset consists of 25, 20, 28, and 24 samples for C1-MES, C2-IMM, C4-DIF, and C5-PRO subtypes, respectively. We artificially created two sample groups with complete confounding by letting one subtype be measured on one platform and the rest three on the other platform, mimicking a completely confounded case-control study.

We first compare the performance of "ReMeasure", "Batch2", "Ignore" and "LS" after fitting gene-wise models. We start with evaluating the type I error control of the proposed method. This is achieved by comparing the same subtypes from the Agilent and the RNA-Seq platform. To ensure sufficient statistical power, we pooled samples from all four subtypes and made the subtype composition similar between the Agilent and RNA-Seq dataset. Specifically, we compare 276 Agilent samples consisting of 69 samples in each subtype to 68 RNA-Seq samples consisting of 17 samples in each subtype, using 40 remeasured Agilent samples to correct for batch effects. Since both batches have similar subtype composition and the patient characteristics are also similar between the two batches (they are from the same Midwest population), we expect to see very few substantial differences. Indeed, based on Figure 5, we observe that "Batch2" detects about 5% "significant" genes across different numbers of remeasured samples as expected at the 5% significance cutoff. For "ReMeasure", it detects close to 5% "significant" genes when the number of remeasures samples are larger than or equal to 10, consistent with the simulation results. In contrast, "Ignore" and "LS" have made substantially more "false" discoveries, indicating that they have poor type I error control.

Next, we conduct a power study by comparing one subtype from the Agilent platform to the other three subtypes from the RNA-Seq platform, treating the RNA-Seq samples as controls and the Agilent samples as cases. RNA-Seq samples remeasured on the Agilent platform are used to correct batch effects. To objectively evaluate power, we need to know the ground truth. However, the ground truth is unknown in this case, so instead we create a list of genes that are more likely to be subtype signatures by comparing one subtype vs. others in the same Agilent dataset. Based on two-sample t-tests and

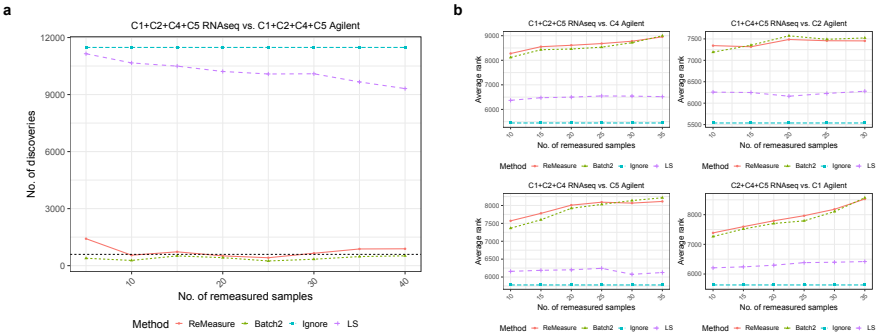


Fig. 5: **Comparison of "ReMeasure", "Batch2", "Ignore" and "LS" on the real dataset.** (a) The number of discoveries vs. the number of remeasured samples by comparing the same subtypes between the two platforms. Specifically, "C1+C2+C4+C5 RNAseq vs. C1+C2+C4+C5 Agilent" denotes comparing subsets of samples, each consisting of an identical number of samples from each subtype (C1-MES, C2-IMM, C4-DIF, C5-PRO), between the RNAseq and Agilent platforms. A two-sided z-test with a 5% significance cut-off is applied for all methods. (b) The average rank vs. the number of remeasured samples for those subtype signature genes by comparing different subtypes on the two platforms. Likewise, "C1+C2+C5 RNAseq vs. C4 Agilent" refers to comparing combined "C1-MES," "C2-IMM," and "C5-PRO" subtypes from the RNAseq platform to the "C4-DIF" subtype from the Agilent platform. The same explanation applies to other titles.

5% FDR (Benjamini-Hochberg procedure), we identified 3793, 4212, 6168, 4439 signature genes for the four subtypes, respectively. In the following, we conduct four types of comparisons: 1) C1+C2+C5 RNA-Seq vs. C4 Agilent, 2) C1+C4+C5 RNA-Seq vs. C2 Agilent, 3) C1+C2+C4 RNA-Seq vs. C5 Agilent and 4) C2+C4+C5 RNA-Seq vs. C1 Agilent. We evaluate the ability of the proposed method to retrieve those signature genes, in comparison to "Batch2", "LS" and "Ignore". If a method works, we expect that the signature genes will rank high (lower p-values) in the respective results. The average ranks of "ReMeasure" and "Batch2" are much higher than "Ignore" and "LS" (Figure 5)."ReMeasure" achieves a slightly higher rank than "Batch2", especially when the number of remeasured samples is at the lower end. However, "ReMeasure" recovers substantially more genes than "Batch2" for the four subtypes at 5% FDR (Supplementary Figure 5b), indicating that "ReMeasure" is more powerful than "Batch2" while the false positive control is similar to "Batch2".

Finally, we compare the number of discoveries for "Batch2" and "ReMeasure" on the genes with the lowest between-batch correlation (bottom quartile) and the highest between-batch correlation (top quartile). Supplementary Figures 5c and 5d reveal that our approach is more similar to "Batch2" under

weak correlation and more powerful than "Batch2" under a strong correlation, consistent with our simulation study.

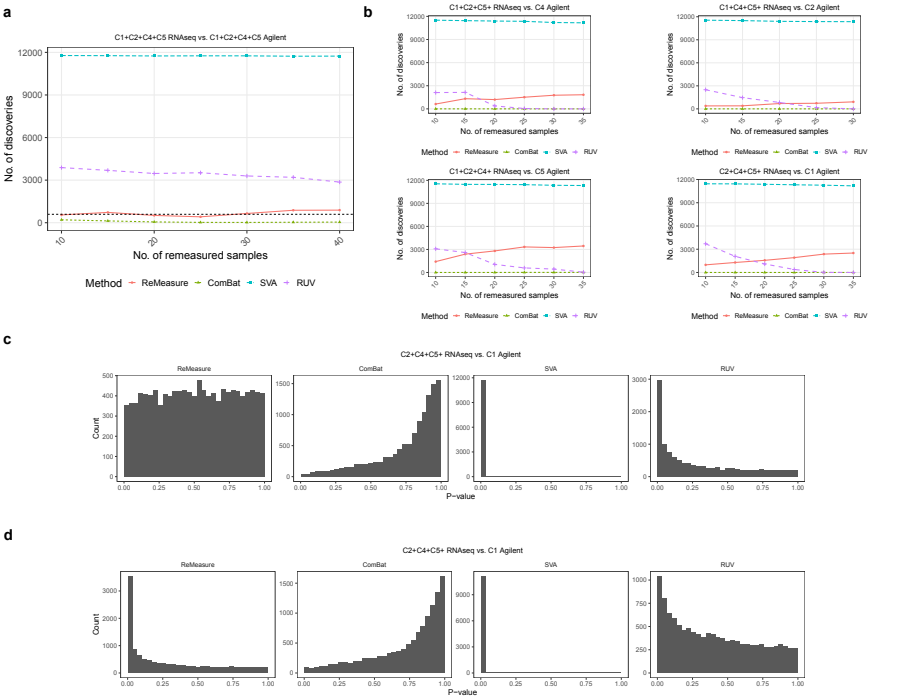


Fig. 6: **Comparison to "ComBat", "SVA" and "RUV" on the real dataset.** (a) and (b) show the number of discoveries vs. the number of remeasured samples by comparing subtypes on the two platforms. The p-values of "ComBat", "SUV" are calculated based on the F-test, while those of "RUV" and "ReMeasure" are obtained from two-sides t-test and z-test, respectively. (a) Comparing the same subtypes when the nominal type I error level is 0.05. "C1+C2+C4+C5 RNAseq vs. C1+C2+C4+C5 Agilent" denotes comparing subsets of samples, each consisting of an identical number of samples from each subtype (C1-MES, C2-IMM, C4-DIF, C5-PRO), between the RNAseq and Agilent platforms. A 5% significance cut-off is applied for all methods. (b) Comparing different subtypes with 5% FDR. "C1+C2+C5 RNAseq vs. C4 Agilent" refers to comparing combined "C1-MES," "C2-IMM," and "C5-PRO" subtypes from the RNAseq platform to the "C4-DIF" subtype from the Agilent platform. The same explanation applies to other titles. (c) and (d) show the unadjusted raw p-value distribution. (c) Comparing the same subtypes with all 40 remeasured samples included. (d) Comparing different subtypes (C2+C4+C5 RNA-Seq vs. C1 Agilent) with all 35 remeasured samples included.

We further compare our "ReMeasure" method to "ComBat" [8], "SVA" [14, 30], and "RUV" [15–17], the three most popular batch effect correction methods, on the real data set. "ComBat" directly removes the known batch effects by performing an empirical Bayesian adjustment, while "SVA" identifies and estimates the surrogate variables for unwanted variations, including batch effects and other unmeasured biological variations, with no requirement of knowing the batch a sample belongs to. "RUV" assumes a factor model that utilizes negative control genes (i.e., genes unrelated to the factor of interest) to estimate the latent factors for unwanted variations. Although in our case the batch information is known, we still run "SVA" and "RUV" to see whether they can capture the known batch effects. We used the `ComBat` and `sva` functions in the R Bioconductor `sva` package, and `naiveReplicateRUV` function in the R Bioconductor `RUVnormalize` package to run the three procedures. The remeasured samples in the second batch were included in the analysis, but their corresponding samples in the first batch were excluded to satisfy the independence assumption of both methods. For "SVA", we used the permutation method described in [31] to estimate the optimal number of surrogate variables. The resulting surrogate variables were then included in the regression model as covariates. The p-values were calculated based on the F-test, comparing the model with and without the group variable. For "ComBat", we fit the gene-wise linear regression model based on batch-corrected data. For "RUV", 364 housekeeping genes were used as negative controls, and the remeasured samples were used as the replicates, following the original paper [17].

Under the null, where we compare the gene expression of the same subtypes (C1+C2+C4+C5) between the two measurement platforms (RNA-Seq vs. Agilent), "SVA" finds a substantially higher number of significant genes than what would be predicted under the null, even with a large number of estimated surrogate variables (>24 surrogate variables for most cases, Figure 6a), indicating that the estimated surrogate variables are still not adequate to capture the full batch effects. Since "SVA" could not control the type I error properly, its high power under the alternative hypothesis is thus not meaningful (Figure 6b). On the other hand, "ComBat" is very conservative and finds very few significant genes under the null (Figure 6a). Its type I error control is at the expense of power. When we compare one subtype vs. others (Figure 6b), the power of "ComBat" is extremely low, indicating that most of the true biological signals may be removed in batch correction due to high confounding of biological and batch effects. "RUV" also has substantially increased type I error, but is less serious than "SVA". The number of detected significant genes decreases with the number of remeasured samples.

It is also interesting to compare the p-value distributions of the four methods. Under the null (C1+C2+C4+C5 RNA-Seq vs. C1+C2+C4+C5 Agilent), the p-value distribution of "ReMeasure" is close to the uniform distribution, while the p-value distributions of "ComBat", "SVA", and "RUV" deviate substantially from the uniform distribution (Figure 6c). When comparing C2+C4+C5 RNA-Seq to C1 Agilent (Figure 6d), the p-value distribution of

"ReMeasure" has the expected form for a multiple testing experiment with signals, with a spike of small p-values and a long tail of larger p-values close to the uniform distribution. In contrast, the p-value distribution of "ComBat" has a spike on the right side of the histogram due to over-adjustment, and the p-values of "SVA" concentrate on the left side of the histogram due to under-adjustment. The p-value distribution of "RUV" behaves well in this case.

We thus conclude that the existing batch adjustment methods do not work well in the severely confounded scenario, and our method can effectively leverage the remeasured samples to correct batch effects.

3 Discussion

Due to the complex technical processes involved in biological measurement, even slight variation in sample preparation and processing can cause batch effects [1]. In many cases, batch effects are not known until the data are analyzed. Batch effects are most disastrous when it is highly confounded with the variable of interest, for example, when the case and control samples are measured separately. In such scenarios, it is extremely challenging to separate the true biological effects from batch effects. Although such confounded studies could be due to a bad study design or less awareness of batch effects, they could also be due to logistics issues. For example, a clinical investigator has collected patient samples and wants to compare them to existing controls. But due to sample availability or financial constraint, the investigator may not be able to remeasure all the control samples together with the case samples. It is thus of tremendous help to the investigator if she only needs to remeasure a small subset of control samples while retaining most power.

Traditional batch effect correction methods such as "ComBat" [8], "SVA" [14, 30], and "RUV" [15–17] were mainly developed for independent samples and they have limited ability to correct batch effects in highly confounded scenario. They either removed the batch and biological effects altogether (reduced power) or retained the batch effects to a large extent (increased type I error).

Our method has several limitations. In some cases, the control samples may not be available for remeasurement, making our method not applicable. Even if they are available for remeasurement, there can still be subtle batch effects associated with difference in collection, storage, and freeze-thaw cycle [24]. Though reprocessing the samples can reduce batch effects, batch effects associated with the upstream technical variation can still persist. Our method cannot correct these residual batch effects. Furthermore, although we show our method is robust to some deviation of the Gaussian distribution, it can still perform poorly when the data are highly skewed or zero inflated. As the genomics studies move into the era of single-cell genomics, the genomics data has become even more complex with severe zero inflation [32]. Simple data transformation may not be sufficient to make the data Gaussian-like. To extend

the capability of our method to analyze such complex genomics datasets, new methodological development is needed. One potential direction is to extend our method to the generalized linear model setting, where the measurement can be modeled by more general distributions such as zero-inflated negative binomial model for zero-inflated count data [33, 34].

Our procedure is based on the maximum likelihood estimation framework, and we proved its consistency and asymptotic normality. However, when the number of remeasured samples is small (n<10), the procedure could have inflated type I error. This is a disadvantage of the proposed method since the number of samples needed to be remeasured may be small when the between-batch correlation is high. To improve the small-sample performance, we proposed a bootstrap method based on residual resampling and showed that it had a well-controlled type I error. However, when the inter-batch correlation is not high (<0.8), the bootstrap method could be less powerful than the "Batch2" method. In this case, "Batch2" is recommended. As the type I error inflation of the asymptotic procedure is mainly driven by the inaccurate estimation of the between-batch correlation when we analyze a large number of features as in omics-wide testing, it is possible to improve the estimation efficiency by pooling information from all features using empirical Bayes method [8]. We leave this as a future research direction.

4 Methods

4.1 Parameter Estimation

Under the Gaussian assumption on the errors, the log joint likelihood of the data is given by

$$\begin{aligned} &L_{\mathbf{N}}(\boldsymbol{\theta}) = -n_1' \log(\sigma_1^2) - n_1' \log(\sigma_2^2) - n_1' \log(1-\rho^2) \\ &-\frac{1}{(1-\rho^2)} \sum_{i \in S_1} \left[\left(\frac{y_i - \mu_{1i}}{\sigma_1} \right)^2 - 2\rho \left(\frac{y_i - \mu_{1i}}{\sigma_1} \right) \left(\frac{y_{n+i} - \mu_{3i}}{\sigma_2} \right) + \left(\frac{y_{n+i} - \mu_{3i}}{\sigma_2} \right)^2 \right] \\ &-(n_1 - n_1') \log(\sigma_1^2) - \sum_{i \in C_1 \setminus S_1} \left(\frac{y_i - \mu_{1i}}{\sigma_1} \right)^2 - n_2 \log(\sigma_2^2) - \sum_{i \in T_2} \left(\frac{y_i - \mu_{2i}}{\sigma_2} \right)^2. \end{aligned} \tag{3}$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ can be obtained as

$$\hat{\boldsymbol{\theta}} = (\hat{a}_0, \hat{a}_1, \hat{\mathbf{b}}, \hat{\rho}, \hat{\sigma}_1, \hat{\sigma}_2) = \arg\max_{\boldsymbol{\theta} \in \Theta} L_{\mathbf{N}}(\boldsymbol{\theta}). \tag{4}$$

The solution to (4) does not have a closed-form solution due to the correlation between the remeasured samples from the two batches. One way to find the solution is by using a generic numerical optimization algorithm such as the Newton-Raphson method or its variants, which updates the parameters via the first or second-order methods until convergence. Here we provide a more efficient algorithm (see Supplementary Section 2) that explores the specific structure of the first-order conditions associated with the objective

function. We deduce the first-order conditions by setting the partial derivative of the objective function with respect to each parameter to be zero. We then update the parameters by iteratively solving these equations. The algorithm is an order-of-magnitude faster than the generic optimization algorithm (Supplementary Figure 6).

4.2 Statistical Inference

We are mostly interested in estimating and conducting inference of the biological effect a_0. The uncertainty assessment or variance of $\hat{a}_0$ is key to hypothesis testing and power analysis. The alternate updating algorithm in Supplementary Section 2 allows us to obtain the variance estimate straightforwardly. It can be shown that

$$\hat{a}_0 = \sum_{i \in T_2} \frac{y_i - \mathbf{z}_i^\top \hat{\mathbf{b}}}{n_2} - \sum_{i \in C_2} \frac{y_i - \mathbf{z}_i^\top \hat{\mathbf{b}}}{n_1'} + \frac{\hat{\rho}\hat{\sigma}_2}{\hat{\sigma}_1} \sum_{i \in S_1} \frac{y_i - \mathbf{z}_i^\top \hat{\mathbf{b}}}{n_1'} \tag{5}$$

indicating that the MLE of a_0 can be expressed as the linear combination of response variables from different batches. The first two terms in the formula estimate the biological effect without using the first batch of samples, and the third term uses those remeasured samples in the first batch to adjust the estimate. The degree of adjustment depends on the between-batch correlation for those remeasured samples. When the correlation is low, the estimate is similar to that without using the first batch. However, when the correlation is high, the adjustment could be substantial. Based on the formula (5), we can calculate its variance accordingly. The details for the variance formula can be found in Supplementary Section 2.

Based on the large-sample theory in Supplementary Section 1, the p-value for testing $a_0 = 0$ can be computed as $2\Phi(-|\hat{a}_0|/\widehat{\mathrm{sd}}(\hat{a}_0))$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The remeasured sample size n_1' needs to be large for our large-sample theory to work. However, in practice, the remeasured sample size may be small, in which case, the estimation of the correlation parameter ρ is subject to large variability since it only depends on n_1' pairs of observations. For small n_1', the large sample theory does not provide an accurate approximation to the sampling distribution of $|\hat{a}_0|/\widehat{\mathrm{sd}}(\hat{a}_0)$. To overcome this issue, we propose to use the bootstrap method to improve the approximation accuracy of the finite sample distribution. For example, we can use the residual bootstrap. The set of residuals is obtained as

$$\begin{aligned}
&\text{Control (batch 1):} \quad \hat{\epsilon}_i^{(1)} = y_i - \mathbf{z}_i^\top \hat{\mathbf{b}}, i = 1, \ldots, n_1, \\
&\text{Case (batch 2):} \quad \hat{\epsilon}_i^{(2)} = y_i - \hat{a}_0 - \hat{a}_1 - \mathbf{z}_i^\top \hat{\mathbf{b}}, i = n_1 + 1, \ldots, n_1 + n_2 = n, \\
&\text{Control (batch 2):} \quad \hat{\epsilon}_i^{(2)} = y_i - \hat{a}_1 - \mathbf{z}_i^\top \hat{\mathbf{b}}, i = n + 1, \ldots, n + n_1'.
\end{aligned}$$

We re-sample the residuals with replacements from each group and then generate a new bootstrap sample with the fixed $\mathbf{z}_i$ but new y_i using the fitted parameters and re-sampled residuals.

Given B bootstrap samples, we can calculate $\hat{a}_0^{(b)}$ and $\widehat{\text{Var}}(\hat{a}_0^{(b)})$ for $1 \leq b \leq B$ based on each resample using Algorithm 1 and Formula (17) in Supplementary Section 2. Thereby, we obtain the bootstrap statistics $Z_b := (\hat{a}_0^{(b)} - \hat{a}_0)/\sqrt{\widehat{\text{Var}}(\hat{a}_0^{(b)})}$ for $b = 1, \ldots, B$. Given $Z = \hat{a}_0/\sqrt{\widehat{\text{Var}}(\hat{a}_0)}$, the bootstrapped p-value can be computed as $B^{-1}\sum_{b=1}^{B} \mathbf{1}\{|Z_b| > |Z|\}$.

Data Availability. Source data for Figures 2-6 is available with this manuscript. They can also be found at https://github.com/yehanxuan/BatchReMeasure-manuscript-sourcecode.

Code Availability. All the codes to reproduce the results in this paper are available at https://github.com/yehanxuan/BatchReMeasure-manuscript-sourcecode. The developed R package BatchReMeasure is available at https://github.com/yehanxuan/BatchReMeasure. The specific version used to produce the results in this manuscript is also available on Code Ocean [35].

References

[1] Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics **11**(10), 733–739 (2010)

[2] Goh, W.W.B., Wang, W., Wong, L.: Why batch effects matter in omics data, and how to avoid them. Trends in Biotechnology **35**(6), 498–507 (2017)

[3] Scherer, A.: Batch Effects and Noise in Microarray Experiments: Sources and Solutions. John Wiley & Sons, New Jersey (2009)

[4] Tom, J.A., Reeder, J., Forrest, W.F., Graham, R.R., Hunkapiller, J., Behrens, T.W., Bhangale, T.R.: Identifying and mitigating batch effects in whole genome sequencing data. BMC Bioinformatics **18**(1), 1–12 (2017)

[5] Price, E.M., Robinson, W.P.: Adjusting for batch effects in dna methylation microarray data, a lesson learned. Frontiers in Genetics **9**, 83 (2018)

[6] Sims, A.H., Smethurst, G.J., Hey, Y., Okoniewski, M.J., Pepper, S.D., Howell, A., Miller, C.J., Clarke, R.B.: The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. BMC Medical Genomics **1**(1), 1–14 (2008)

[7] Li, C., Wong, W.H.: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proceedings of the National Academy of Sciences **98**(1), 31–36 (2001)

[8] Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics **8**(1), 118–127 (2007)

[9] Zhang, Y., Parmigiani, G., Johnson, W.E.: ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics **2**(3) (2020)

[10] Shabalin, A.A., Tjelmeland, H., Fan, C., Perou, C.M., Nobel, A.B.: Merging two gene-expression studies via cross-platform normalization. Bioinformatics **24**(9), 1154–1160 (2008)

[11] Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C.M., Marron, J.S.: Adjustment of systematic microarray data biases. Bioinformatics **20**(1), 105–114 (2004)

[12] Alter, O., Brown, P.O., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences **97**(18), 10101–10106 (2000)

[13] Jolliffe, I.T.: Principal Component Analysis. Springer, New York, NY (2013)

[14] Leek, J.T., Storey, J.D.: Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics **3**(9), 161 (2007)

[15] Gagnon-Bartsch, J.A., Speed, T.P.: Using control genes to correct for unwanted variation in microarray data. Biostatistics **13**(3), 539–552 (2012)

[16] Gagnon-Bartsch, J.A., Jacob, L., Speed, T.P.: Removing unwanted variation from high dimensional data with negative controls. Berkeley: Tech Reports from Dep Stat Univ California, 1–112 (2013)

[17] Jacob, L., Gagnon-Bartsch, J.A., Speed, T.P.: Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. Biostatistics **17**(1), 16–28 (2016)

[18] Sun, Y., Zhang, N.R., Owen, A.B.: Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. The Annals of Applied Statistics **6**(4), 1664–1688 (2012)

[19] Wang, J., Zhao, Q., Hastie, T., Owen, A.B.: Confounder adjustment in multiple hypothesis testing. Annals of Statistics **45**(5), 1863 (2017)

[20] Tasaki, S., Suzuki, K., Kassai, Y., Takeshita, M., Murota, A., Kondo, Y., Ando, T., Nakayama, Y., Okuzono, Y., Takiguchi, M., *et al.*: Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. Nature Communications **9**(1), 1–12 (2018)

[21] Xia, Q., Thompson, J.A., Koestler, D.C.: Batch effect reduction of microarray data with dependent samples using an empirical bayes approach (bridge). Statistical Applications in Genetics and Molecular Biology **20**(4-6), 101–119 (2021)

[22] Zhou, L., Sue, A.C.-H., Goh, W.W.B.: Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects? Journal of Genetics and Genomics **46**(9), 433–443 (2019)

[23] Olson, J.E., Ryu, E., Hathcock, M.A., Gupta, R., Bublitz, J.T., Takahashi, P.Y., Bielinski, S.J., St Sauver, J.L., Meagher, K., Sharp, R.R., *et al.*: Characteristics and utilisation of the mayo clinic biobank, a clinic-based prospective collection in the usa: cohort profile. Bmj Open **9**(11),

032707 (2019)

[24] Rundle, A.G., Vineis, P., Ahsan, H.: Design options for molecular epidemiology research within cohort studies. Cancer Epidemiology Biomarkers & Prevention **14**(8), 1899–1907 (2005)

[25] Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Routledge, Oxfordshire, United Kingdom (2013)

[26] Wang, C., Winterhoff, B.J., Kalli, K.R., Block, M.S., Armasu, S.M., Larson, M.C., Chen, H.-W., Keeney, G.L., Hartmann, L.C., Shridhar, V., *et al.*: Expression signature distinguishing two tumour transcriptome classes associated with progression-free survival among rare histological types of epithelial ovarian cancer. British Journal of Cancer **114**(12), 1412–1420 (2016)

[27] Konecny, G.E., Wang, C., Hamidi, H., Winterhoff, B., Kalli, K.R., Dering, J., Ginther, C., Chen, H.-W., Dowdy, S., Cliby, W., et al.: Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. Journal of the National Cancer Institute **106**(10) (2014)

[28] Fridley, B.L., Dai, J., Raghavan, R., Li, Q., Winham, S.J., Hou, X., Weroha, S.J., Wang, C., Kalli, K.R., Cunningham, J.M., *et al.*: Transcriptomic characterization of endometrioid, clear cell, and high-grade serous epithelial ovarian carcinoma. Cancer epidemiology, biomarkers & prevention **27**(9), 1101–1109 (2018)

[29] Chen, G.M., Kannan, L., Geistlinger, L., Kofia, V., Safikhani, Z., Gendoo, D.M., Parmigiani, G., Birrer, M., Haibe-Kains, B., Waldron, L.: Consensus on molecular subtypes of high-grade serous ovarian carcinoma. Clinical Cancer Research **24**(20), 5037–5047 (2018)

[30] Leek, J.T., Storey, J.D.: A general framework for multiple testing dependence. Proceedings of the National Academy of Sciences **105**(48), 18718–18723 (2008)

[31] Buja, A., Eyuboglu, N.: Remarks on parallel analysis. Multivariate behavioral research **27**(4), 509–540 (1992)

[32] Stegle, O., Teichmann, S.A., Marioni, J.C.: Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics **16**(3), 133–145 (2015)

[33] Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., Ballman, K.: An omnibus test for differential distribution analysis of microbiome sequencing data. Bioinformatics **34**(4), 643–651 (2018)

[34] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., Vert, J.-P.: A general and flexible method for signal extraction from single-cell rna-seq data. Nature communications **9**(1), 284 (2018)

[35] Ye, H., Zhang, X., Chen, J.: BatchReMeasure: Batch effects correction with sample remeasurement. Code Ocean https://doi.org/10.24433/CO.4806327.v1 (2023).

[36] Takeshi, A.: Advanced Econometrics. Harvard University Press, Cambridge, Massachusetts (1985)

[37] Vaart, A.W.v.d.: Asymptotic Statistics. Cambridge University Press, Cambridge, United Kingdom (1998)

[38] Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Accelerated methods for nonconvex optimization. SIAM Journal on Optimization **28**(2), 1751–1772 (2018)

[39] Nesterov, Y., Polyak, B.T.: Cubic regularization of newton method and its global performance. Mathematical Programming **108**(1), 177–205 (2006)

Supplementary Section 1 Theoretical results

This section will show that our estimator is consistent and asymptotically normal if the sample size grows to infinity. Our theory differs from the traditional MLE theory in two aspects: (i) we do not require the likelihood function to be correctly specified as the errors are allowed to be non-Gaussian, and (ii) the data are not identically distributed in our setting as the model structure changes across batches and case/control groups, which complicates the analysis. The following set of mild assumptions is imposed for theoretical analysis.

Condition 1 *The true parameter* $\boldsymbol{\theta}_0 = (a_{00}, a_{10}, \mathbf{b}_0, \rho_0, \sigma_{10}, \sigma_{20})$ *belongs to the interior of some compact parameter space* Θ.

Condition 2 *The errors* $\epsilon_i^{(j)}$ *are independent across* i *and* j *with mean zero and finite variance, i.e.,* $\mathbb{E}[\epsilon_i^{(j)}] = 0$ *and* $\mathrm{Var}[\epsilon_i^{(j)}] = \sigma_j^2 < \infty$. *Assume that the covariates* $\mathbf{z}_i$ *are i.i.d with* $\mathbb{E}[\mathbf{z}] = \mathbf{0}$ *and* $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \boldsymbol{\Sigma}_z$, *where* $\boldsymbol{\Sigma}_z$ *is positive definite.*

Condition 3 *Suppose each batch is a non-negligible portion of the total sample, and the remeasured sample is a non-negligible portion of the batch 1 sample. Formally, we assume that* $n_1/n \to r_1$ *and* $n_1'/n_1 \to r_1'$ *as* $n_1, n_2 \to \infty$, *where* $r_1, r_1' \in (0, 1)$.

Under Conditions 1-3, the objective function normalized by the sample size converges in probability to a weighted sum of some non-stochastic functions

as $n_1', n_2 \to \infty$ by the law of large numbers,

$$\bar{L}_{\mathbf{N}}(\boldsymbol{\theta}) = n^{-1} L_{\mathbf{N}}(\boldsymbol{\theta}) \xrightarrow{p} \ell(\boldsymbol{\theta}) = \sum_{k=1}^{3} w_k \ell_k(\boldsymbol{\theta}) \tag{1}$$

with $w_1 = r_1 r_1', w_2 = r_1(1-r_1')$ and $w_3 = (1-r_1)$. Here $\ell_k(\boldsymbol{\theta})$'s are the limiting functions of the sample averages of the Gaussian log-likelihoods. The detailed forms of $\ell_k(\boldsymbol{\theta})$ can be found in Section 1.1.

Theorem 1.1 (Consistency) *Suppose Conditions 1-3 are satisfied. The estimator $\hat{\boldsymbol{\theta}}$ that maximizes the objective function $L_{\mathbf{N}}(\boldsymbol{\theta})$ is weakly consistent, namely, $\hat{\boldsymbol{\theta}}$ converges in probability to the underlying true parameter $\boldsymbol{\theta}_0 = (a_{00}, a_{10}, \mathbf{b}_0, \rho_0, \sigma_{10}, \sigma_{20})$ of Model 1 as $n_1' \to \infty, n_2 \to \infty$.*

We remark that the remeasured size n_1' has to tend to infinity to ensure the consistency of the MLE of ρ.

Theorem 1.2 (Asymptotic Normality) *Under Conditions 1-3, the estimator $\hat{\boldsymbol{\theta}}$ is asymptotically normal, i.e.,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \Longrightarrow \mathcal{N}\left(0, \mathbb{E}[\ddot{\ell}(\boldsymbol{\theta}_0)]^{-1} \mathbb{E}\left[\sum_{k=1}^{3} w_k \dot{\ell}_k(\boldsymbol{\theta}_0) \dot{\ell}_k(\boldsymbol{\theta}_0)^\top\right] \mathbb{E}[\ddot{\ell}(\boldsymbol{\theta}_0)]^{-1}\right), \tag{2}$$

as $n_1', n_2 \to \infty$, with the weights w_k's given in (1). Here $\dot{\ell}_k(\boldsymbol{\theta}_0)$ and $\ddot{\ell}(\boldsymbol{\theta}_0)$ denote the first derivative of $\ell_k(\boldsymbol{\theta})$and second derivative of $\ell(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, respectively.

Proofs of Theorem 1.1 and 1.2 are detailed below.

1.1 Proof of Theorem 1.1

The proof requires multiple steps. We first present several useful lemmas. Then, we show that the objective function converges uniformly in probability to some non-stochastic function that has a unique maximizer. The consistency is then established using Lemma 1.4.

Lemma 1.3 (Strictly concavity) *The log-likelihood of a mean-zero Gaussian distribution*

$$h_n(\boldsymbol{\Sigma}^{-1}) = -\frac{1}{2} tr(\mathbf{S}_n \boldsymbol{\Sigma}^{-1}) + \frac{1}{2} \log\det(\boldsymbol{\Sigma}^{-1}) \tag{3}$$

is strictly concave with respect to $\boldsymbol{\Sigma}^{-1}$ for some positive definite matrix $\mathbf{S}_n$. Thus $h_n(\cdot)$ has a unique global maximizer.

Proof Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Note that $-\mathrm{tr}(\mathbf{S}_n\boldsymbol{\Omega})$ is an affine function of $\boldsymbol{\Omega}$, and the log-determinant function $\log\det(\boldsymbol{\Omega})$ is strictly concave. Thus the linear combination of these two terms is strictly concave as a function of $\boldsymbol{\Omega}$. □

Lemma 1.4 (Theorem 4.1.1 of [36]) *Suppose the function* $\bar{L}_{\mathbf{N}}(\boldsymbol{\theta})$ *satisfies the following conditions:*

1. *The parameter space* Θ *is compact.*
2. $\bar{L}_{\mathbf{N}}(\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta} \in \Theta$ *almost everywhere.*
3. $\bar{L}_{\mathbf{N}}(\boldsymbol{\theta})$ *converges to a non-stochastic function* $\ell(\boldsymbol{\theta})$ *in probability uniformly over* $\boldsymbol{\theta} \in \Theta$ *and* $\ell(\boldsymbol{\theta})$ *attains a unique global maximum at* $\boldsymbol{\theta}_0$.

Then $\hat{\boldsymbol{\theta}}_{\mathbf{N}} := \arg\max_{\boldsymbol{\theta}\in\Theta} \bar{L}_{\mathbf{N}}(\boldsymbol{\theta}) \to^p \boldsymbol{\theta}_0$.

Lemma 1.5 (Uniform convergence in probability) *Let* $g(\mathbf{x}, \boldsymbol{\theta})$ *be a measurable function of* $\mathbf{x}$ *for each* $\boldsymbol{\theta}$ *in a compact space* Θ*, and a continuous function of* $\boldsymbol{\theta}$ *for each* $\mathbf{x}$. *Let* $\mathbf{x}_i$ *be a sequence of i.i.d random vectors such that* $\mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} |g(\mathbf{x}_i, \boldsymbol{\theta})|] < \infty$ *and* $\mathbb{E}[g(\mathbf{x}_i, \boldsymbol{\theta})] = 0$. *Then*

$$\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i, \boldsymbol{\theta}) \to^p 0 \quad \textit{uniformly.}$$

Proof Write $g_i(\boldsymbol{\theta}) = g(\mathbf{x}_i, \boldsymbol{\theta})$ for the ease of notation. The compact parameter space Θ has a finite non-overlapping cover $\Theta_1^K, \ldots, \Theta_K^K$ such that the distance of any two points within some Θ_i^K goes to 0 as $K \to \infty$. Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ be K-vectors such that $\boldsymbol{\theta}_i \in \Theta_i^K$. Then we have for any $\varepsilon > 0$,

$$\begin{aligned}
&\Pr[\sup_{\boldsymbol{\theta}\in\Theta} |n^{-1}\sum_{i=1}^{n} g_i(\boldsymbol{\theta})| > \varepsilon] \le \Pr[\cup_{k=1}^{K}\{\sup_{\boldsymbol{\theta}\in\Theta_k^K} |n^{-1}\sum_{i=1}^{n} g_i(\boldsymbol{\theta})| > \varepsilon\}] \\
\le &\sum_{k=1}^{K} \Pr[\sup_{\boldsymbol{\theta}\in\Theta_k^K} |n^{-1}\sum_{i=1}^{n} g_i(\boldsymbol{\theta})| > \varepsilon] \\
\le &\sum_{k=1}^{K} \Pr[|n^{-1}\sum_{i=1}^{n} g_i(\boldsymbol{\theta}_k)| > \varepsilon/2] + \sum_{k=1}^{K} \Pr[n^{-1}\sum_{i=1}^{n} \sup_{\boldsymbol{\theta}\in\Theta_k^K} |g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_k)| > \varepsilon/2].
\end{aligned}$$

Since $g_i(\boldsymbol{\theta})$ is uniformly continuous in $\boldsymbol{\theta} \in \Theta$ for every i, we have

$$\lim_{K\to\infty} \sup_{1\le k\le K} \sup_{\boldsymbol{\theta}\in\Theta_k^K} |g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_k)| = 0$$

almost surely. Meanwhile,

$$\sup_{1\le k\le K} \sup_{\boldsymbol{\theta}\in\Theta_k^K} |g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_k)| \le 2 \sup_{\boldsymbol{\theta}\in\Theta} |g_i(\boldsymbol{\theta})|. \tag{4}$$

The integrability of the right-hand side indicates that we can use the Lebesgue dominated convergence theorem to show that

$$\lim_{K\to\infty} \mathbb{E}[\sup_{1\le k\le K} \sup_{\boldsymbol{\theta}\in\Theta_k^K} |g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_k)|] = 0.$$

That is to say, there exists a finite $K = K(\varepsilon/4)$ such that

$$\mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta_k^K} |g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_k)|] < \varepsilon/4$$

for $k \in 1, \ldots, K$. Finally, the conclusion of the theorem follows from Kolmogorov's law of large numbers (KLLN). Taking $n \to \infty$, for any $k \in 1, \ldots, K$, $\sum_{i=1}^{n} g_i(\boldsymbol{\theta}_k)/n \overset{a.s.}{\to} 0$ since $\mathbb{E}[g_i(\boldsymbol{\theta}_k)] = 0$. Moreover,

$$n^{-1} \sum_{i=1}^{n} \sup_{\boldsymbol{\theta} \in \Theta_k^K} |g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_k)| \overset{a.s.}{\to} \mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta_k^K} |g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_k)|].$$

We can always choose $n > N(K, \kappa)$ for any small value $\kappa > 0$ such that $\Pr[|n^{-1} \sum_{i=1}^{n} g_i(\boldsymbol{\theta}_k)| > \varepsilon/2] < \kappa/(2K)$ and $\Pr[n^{-1} \sum_{i=1}^{n} \sup_{\boldsymbol{\theta} \in \Theta_k^K} |g_i(\boldsymbol{\theta}) - g_i(\boldsymbol{\theta}_k)| > \varepsilon/2] < \kappa/(2K)$. □

We now divide the proof of Theorem 1.1 into three major steps. We consider the compact parameter space Θ with the following form: $c_1 \le \sigma_1, \sigma_2, \sigma_3 \le c_2$, $\rho_1, \rho_2 \in [-1 + \delta, 1 - \delta]$ for some $\delta > 0$, $\mathbf{b}^\top \mathbf{b} \le c_3$, $a_0, a_1, a_3 \in [-M, M]$, where c_1 and δ are small positive constants and c_2, c_3, M are large positive constants. The true parameter $\boldsymbol{\theta}_0$ is assumed to be an interior point of Θ and the $\mathbf{z}$ is assumed to have the positive definite covariance matrix $\boldsymbol{\Sigma}_z$.

Step 1: Point-wise convergence in probability. Let $\bar{L}_{\mathbf{N}}(\boldsymbol{\theta}) = L_{\mathbf{N}}(\boldsymbol{\theta})/n$ and note that $\hat{\boldsymbol{\theta}}$ is the maximizer of $\bar{L}_{\mathbf{N}}(\boldsymbol{\theta})$. For any $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \Lambda)$, where $\boldsymbol{\alpha} = (a_0, a_1, \mathbf{b})$ and $\Lambda = (\rho_1, \sigma_1, \sigma_2)$, we have

$$\begin{aligned}
\bar{L}_{\mathbf{N}}(\boldsymbol{\theta}) =& \frac{n_1'}{n} \Big\{ -\log\det(\boldsymbol{\Sigma}) \\
& - \sum_{i \in S_1} \frac{1}{n_1'} (y_i - \mu_{1i}, y_{n+i} - \mu_{3i}) \boldsymbol{\Sigma}^{-1} (y_i - \mu_{1i}, y_{n+i} - \mu_{3i})^\top \Big\} \\
& + \frac{n_1 - n_1'}{n} \left\{ -\log(\sigma_1^2) - \sum_{i \in C_1 \setminus S_1} \frac{1}{n_1 - n_1'} \sigma_1^{-2} (y_i - \mu_{1i})^2 \right\} \\
& + \frac{n_2}{n} \left\{ -\log(\sigma_2^2) - \sum_{i \in T_2} \frac{1}{n_2} \sigma_2^{-2} (y_i - \mu_{2i})^2 \right\} \\
& \overset{p}{\to} \ell(\boldsymbol{\theta}) = r_1 r_1' \Big\{ -\log\det(\boldsymbol{\Sigma}) - \mathrm{tr}(\mathbf{V}_1^{(1)} \boldsymbol{\Sigma}^{-1}) \Big\} \\
& + r_1 (1 - r_1') \{ -\log(\sigma_1^2) - \sigma_1^{-2} V_2^{(1)} \} \\
& + (1 - r_1) \{ -\log(\sigma_2^2) - \sigma_2^{-2} V_3^{(1)} \} \\
& = w_1 \ell_1(\boldsymbol{\theta}) + w_2 \ell_2(\boldsymbol{\theta}) + w_3 \ell_3(\boldsymbol{\theta}),
\end{aligned} \tag{5}$$

as $n_1' \to \infty, n_2 \to \infty$, where $w_1 = r_1 r_1', w_2 = r_1(1 - r_1')$ and $w_3 = (1 - r_1)$. The forms of $\mathbf{V}_1^{(1)}, V_2^{(1)}$ and $V_3^{(1)}$ are given by

$$\begin{aligned}
\mathbf{V}_1^{(1)} &= \begin{bmatrix} q(\mathbf{b}_0 - \mathbf{b};\ \boldsymbol{\Sigma}_z) + \sigma_{10}^2 & q(\mathbf{b}_0 - \mathbf{b};\ \boldsymbol{\Sigma}_z) + \rho_{10}\sigma_{10}\sigma_{20} \\ q(\mathbf{b}_0 - \mathbf{b};\ \boldsymbol{\Sigma}_z) + \rho_{10}\sigma_{10}\sigma_{20} & (a_{10} - a_1)^2 + q(\mathbf{b}_0 - \mathbf{b};\ \boldsymbol{\Sigma}_z) + \sigma_{20}^2 \end{bmatrix}, \\
V_2^{(1)} &= q(\mathbf{b}_0 - \mathbf{b};\ \boldsymbol{\Sigma}_z) + \sigma_{10}^2, \\
V_3^{(1)} &= (a_{00} - a_0)^2 + (a_{10} - a_1)^2 + 2(a_{00} - a_0)(a_{10} - a_1) + q(\mathbf{b}_0 - \mathbf{b};\ \boldsymbol{\Sigma}_z) + \sigma_{20}^2,
\end{aligned}$$

respectively, where $q(\mathbf{b}_0 - \mathbf{b};\ \mathbf{\Sigma}_z)$ represents the quadratic form $(\mathbf{b}_0 - \mathbf{b})^\top \mathbf{\Sigma}_z (\mathbf{b}_0 - \mathbf{b})$.

Step 2: Uniqueness of the maximizer. The limit of the objective function $\ell(\boldsymbol{\theta})$ enjoys the following decomposition

$$\ell(\boldsymbol{\theta}) = \sum_{k=1}^{3} w_k \ell_k(\boldsymbol{\theta}) = \sum_{k=1}^{3} w_k \Big[f_k(\boldsymbol{\alpha}, \Lambda) + h_k(\Lambda) \Big], \tag{6}$$

where f_k and h_k are defined below. Consider the case of $k = 1$, we have

$$\begin{aligned}
f_1 =& -\frac{1}{(1-\rho^2)} \Big[\frac{(\mathbf{b}_0 - \mathbf{b})\mathbf{\Sigma}_z(\mathbf{b}_0 - \mathbf{b})}{\sigma_1^2} \\
& - 2\rho \frac{(\mathbf{b}_0 - \mathbf{b})^\top \mathbf{\Sigma}_z (\mathbf{b}_0 - \mathbf{b})}{\sigma_1 \sigma_2} + \frac{(a_{10} - a_1)^2 + (\mathbf{b}_0 - \mathbf{b})\mathbf{\Sigma}_z(\mathbf{b}_0 - \mathbf{b})}{\sigma_2^2} \Big] \\
=& -\big(q(\mathbf{b}_0 - \mathbf{b};\ \mathbf{\Sigma}_z) + q((0, a_{10} - a_1)^\top;\ \mathbf{I}) \big) \det(\mathbf{\Sigma}^{-1}), \\
f_2 =& -\frac{(\mathbf{b}_0 - \mathbf{b})\mathbf{\Sigma}_z(\mathbf{b}_0 - \mathbf{b})}{\sigma_1^2}, \\
f_3 =& -\frac{(\mathbf{b}_0 - \mathbf{b})\mathbf{\Sigma}_z(\mathbf{b}_0 - \mathbf{b})}{\sigma_2^2} - \frac{(a_{00} + a_{10} - a_0 - a_1)^2}{\sigma_2^2}, \\
h_1 =& -\log\det(\mathbf{\Sigma}) - \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_0), \\
h_2 =& -\log(\sigma_1^2) - \frac{\sigma_{10}^2}{\sigma_1^2} \\
h_3 =& -\log(\sigma_2^2) - \frac{\sigma_{20}^2}{\sigma_2^2}
\end{aligned}$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \qquad \mathbf{\Sigma}_0 = \begin{bmatrix} \sigma_{10}^2 & \rho_0\sigma_{10}\sigma_{20} \\ \rho_0\sigma_{10}\sigma_{20} & \sigma_{20}^2 \end{bmatrix}.$$

Observe that the f_ks are combination of quadratic forms and all of them attain the maximum value 0 only if $\mathbf{b} = \mathbf{b}_0, a_0 = a_{00}, a_1 = a_{10}$ for any positive definite $\mathbf{\Sigma}$. Regarding h_k's, the strict concavity with respect to $\mathbf{\Sigma}^{-1}$ implies that $\mathbf{\Sigma} = \mathbf{\Sigma}_0$ is the unique maximizer of h_1. Similarly, $\sigma_1 = \sigma_{10}$ ($\sigma_2 = \sigma_{20}$) is the unique maximizer for h_2 (h_3). Hence, $\boldsymbol{\theta}_0$ is the unique maximizer of $\ell(\boldsymbol{\theta})$.

Step 3: Uniform convergence in probability. Take $\mathbf{x}_i = (y_i, \mathbf{z}_i)$ to be the pair of response and covariate. Let

$$\begin{aligned}
\ell_{1,\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_{n+i}) =& -\log\det(\mathbf{\Sigma}) \\
& - (y_i - \mu_{1i}, y_{n+i} - \mu_{3i})\mathbf{\Sigma}^{-1}(y_i - \mu_{1i}, y_{n+i} - \mu_{3i})^\top
\end{aligned} \tag{7}$$

which satisfies $\mathbb{E}[\ell_{1,\boldsymbol{\theta}}] = \ell_1(\boldsymbol{\theta})$. Then

$$\begin{aligned}
&\mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} |\ell_{1,\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_{n+i})|] \\
\leq &\mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} |\log\det(\boldsymbol{\Sigma})| + \sup_{\boldsymbol{\theta}\in\Theta} (y_i - \mu_{1i}, y_{n+i} - \mu_{3i})\boldsymbol{\Sigma}^{-1}(y_i - \mu_{1i}, y_{n+i} - \mu_{3i})^\top] \\
\leq &\mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} |\log\det(\boldsymbol{\Sigma})| \\
&\quad + \sup \lambda_1(\boldsymbol{\Sigma}^{-1})\mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} (y_i - \mu_{1i}, y_{n+i} - \mu_{3i})(y_i - \mu_{1i}, y_{n+i} - \mu_{3i})^\top] \\
\overset{(i)}{\leq} &M_1 + \sup \lambda_1(\boldsymbol{\Sigma}^{-1}) \sup\{\sigma_{10}^2 + \sigma_{20}^2 + 2\|\mathbf{b}_0 - \mathbf{b}\|^2 \mathrm{tr}\boldsymbol{\Sigma}_z + (a_{10} - a_1)^2\} \\
\overset{(ii)}{\leq} &M_1 + M_2.
\end{aligned}$$

In the above derivations, (i) is due to the fact that quadratic form $\mathbf{v}^\top\boldsymbol{\Sigma}^{-1}\mathbf{v} \leq \lambda_1(\boldsymbol{\Sigma}^{-1})\|\mathbf{v}\|^2$ for any vector $\mathbf{v}$ and $\|\mathbf{z}_i^\top(\mathbf{b} - \mathbf{b}_0)\|^2 \leq \|\mathbf{z}_i\|^2\|\mathbf{b} - \mathbf{b}_0\|^2$; (ii) is because of $\boldsymbol{\theta} = (a_0, a_1, \mathbf{b}, \rho_1, \sigma_1, \sigma_2) \in [-M, M]^2 \times \mathbb{B}_p(c_3) \times [-1 + \delta, 1 - \delta] \times [c_1, c_2] \times [c_1, c_2]$. Define $g_i(\boldsymbol{\theta}) := \ell_{1,\boldsymbol{\theta}}((\mathbf{x}_i, \mathbf{x}_{n+i})) - \mathbb{E}[\ell_{1,\boldsymbol{\theta}}((\mathbf{x}_i, \mathbf{x}_{n+i}))]$. Then the uniform convergence of $n_1'^{-1}\sum_{i=1}^{n_1'} \ell_{1,\boldsymbol{\theta}}$ to $\ell_1(\boldsymbol{\theta})$ follows from Lemma 1.5. In the same spirit, we can also define

$$\begin{aligned}
\ell_{2,\boldsymbol{\theta}}(\mathbf{x}_i) &= -\log(\sigma_1^2) - (y_i - \mu_{1i})^2/\sigma_1^2, \\
\ell_{3,\boldsymbol{\theta}}(\mathbf{x}_i) &= -\log(\sigma_2^2) - (y_i - \mu_{2i})^2/\sigma_2^2.
\end{aligned} \tag{8}$$

We shall have similar conclusion regarding $\ell_{2,\boldsymbol{\theta}}(\mathbf{x}_i)$ and $\ell_{3,\boldsymbol{\theta}}(\mathbf{x}_i)$. The uniform convergence in probability of $\bar{L}_{\mathbf{N}}(\boldsymbol{\theta})$ to $\ell(\boldsymbol{\theta})$ is then verified. Finally, we employ Lemma 1.4 to establish the consistency of $\hat{\boldsymbol{\theta}}$.

1.2 Proof of Theorem 1.2

Follow the notation used in [37], we denote

$$\begin{aligned}
&\mathbb{P}_{n_1'}^{(1)}\ell_{1,\boldsymbol{\theta}} = (n_1')^{-1}\sum_{i=1}^{n_1'} \ell_{1,\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_{n+i}), && i = 1, \ldots, n_1', \\
&\mathbb{P}_{n_1 - n_1'}^{(2)}\ell_{2,\boldsymbol{\theta}} = (n_1 - n_1')^{-1}\sum_{i=1}^{n_1 - n_1'} \ell_{2,\boldsymbol{\theta}}(\mathbf{x}_i), && i = n_1' + 1, \ldots, n_1, \\
&\mathbb{P}_{n_2}^{(3)}\ell_{3,\boldsymbol{\theta}} = n_2^{-1}\sum_{i=1}^{n_2} \ell_{3,\boldsymbol{\theta}}(\mathbf{x}_i), && i = n_1, \ldots, n.
\end{aligned}$$

The objective function $\bar{L}_{\mathbf{N}}(\boldsymbol{\theta})$ can be written as

$$\bar{L}_{\mathbf{N}}(\boldsymbol{\theta}) = \frac{n_1'}{n}\mathbb{P}_{n_1'}^{(1)}\ell_{1,\boldsymbol{\theta}} + \frac{n_1 - n_1'}{n}\mathbb{P}_{n_1 - n_1'}^{(2)}\ell_{2,\boldsymbol{\theta}} + \frac{n_2}{n}\mathbb{P}_{n_2}^{(3)}\ell_{3,\boldsymbol{\theta}}.$$

Our (multivariate) Gaussian log-likelihood satisfies the Lipschitz-Hessian condition [38, 39] that

$$\|\nabla^2\ell_{1,\boldsymbol{\theta}_1}(\mathbf{x}_i,\mathbf{x}_{n+i})-\nabla^2\ell_{1,\boldsymbol{\theta}_2}(\mathbf{x}_i,\mathbf{x}_{n+i})\|_{op}\le M_1(\mathbf{x}_i,\mathbf{x}_{n+i})\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|,$$

and $\|\nabla^2\ell_{k,\boldsymbol{\theta}_1}(\mathbf{x}_i)-\nabla^2\ell_{k,\boldsymbol{\theta}_2}(\mathbf{x}_i)\|_{op}\le M_k(\mathbf{x}_i)\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|, k=2,3$ for some absolutely integrable functions $M_1(\mathbf{x}_i,\mathbf{x}_{n+i})$, $M_2(\mathbf{x}_i)$ and $M_3(\mathbf{x}_i)$ in the sense that $\mathbb{E}[M_k]<\infty, k=1,2,3$, as we will verify later in Section 1.2.1. By the first-order condition for $\hat{\boldsymbol{\theta}}$, we have

$$\begin{aligned}0=&\nabla\bar{L}_{\mathbf{N}}(\hat{\boldsymbol{\theta}})=\frac{n_1'}{n}\mathbb{P}^{(1)}_{n_1'}\nabla\ell_{1,\boldsymbol{\theta}_0}+\frac{n_1-n_1'}{n}\mathbb{P}^{(2)}_{n_1-n_1'}\nabla\ell_{2,\boldsymbol{\theta}_0}+\frac{n_2}{n}\mathbb{P}^{(3)}_{n_2}\nabla\ell_{3,\boldsymbol{\theta}_0}\\&+\Big(\frac{n_1'}{n}\mathbb{P}^{(1)}_{n_1'}\nabla^2\ell_{1,\boldsymbol{\theta}_0}+\frac{n_1-n_1'}{n}\mathbb{P}^{(2)}_{n_1-n_1'}\nabla^2\ell_{2,\boldsymbol{\theta}_0}+\frac{n_2}{n}\mathbb{P}^{(3)}_{n_2}\nabla^2\ell_{3,\boldsymbol{\theta}_0}\Big)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\&+\hat{\gamma}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0),\end{aligned}$$

where $\hat{\gamma}=\int_0^1\left(\nabla^2\bar{L}_{\mathbf{N}}((1-t)\hat{\boldsymbol{\theta}}+t\boldsymbol{\theta}_0)-\nabla^2\bar{L}_{\mathbf{N}}(\boldsymbol{\theta}_0)\right)\mathrm{d}t$ and

$$\begin{aligned}\|\hat{\gamma}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|&\le\int_0^1\|\left(\nabla^2\bar{L}_{\mathbf{N}}((1-t)\hat{\boldsymbol{\theta}}+t\boldsymbol{\theta}_0)-\nabla^2\bar{L}_{\mathbf{N}}(\boldsymbol{\theta}_0)\right)\|_{op}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|\,\mathrm{d}t\\&\le\left[\frac{n_1'}{n_1}\mathbb{P}^{(1)}_{n_1'}M_1+\frac{n_1-n_1'}{n}\mathbb{P}^{(2)}_{n_1-n_1'}M_2+\frac{n_2}{n}\mathbb{P}^{(3)}_{n_2}M_3\right]\frac{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|^2}{2}\qquad(9)\\&=O_p(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2).\end{aligned}$$

Re-arranging the terms, we have

$$\begin{aligned}&\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=\\&-\left(\frac{n_1'}{n}\mathbb{P}^{(1)}_{n_1'}\nabla^2\ell_{1,\boldsymbol{\theta}_0}+\frac{n_1-n_1'}{n}\mathbb{P}^{(2)}_{n_1-n_1'}\nabla^2\ell_{2,\boldsymbol{\theta}_0}+\frac{n_2}{n}\mathbb{P}^{(3)}_{n_2}\nabla^2\ell_{3,\boldsymbol{\theta}_0}+o_p(1)\right)^{-1}\\&\left(\frac{n_1'}{\sqrt{n}}\mathbb{P}^{(1)}_{n_1'}\nabla\ell_{1,\boldsymbol{\theta}_0}+\frac{n_1-n_1'}{\sqrt{n}}\mathbb{P}^{(2)}_{n_1-n_1'}\nabla\ell_{2,\boldsymbol{\theta}_0}+\frac{n_2}{\sqrt{n}}\mathbb{P}^{(3)}_{n_2}\nabla\ell_{3,\boldsymbol{\theta}_0}\right).\end{aligned}$$

Given the facts that $n_1/n\to w_1=r_1$, $n_1'/n_1\to w_2=r_1'$, $(n_1-n_1')/n_1\to w_3=1-r_1'$, $\mathbb{E}\dot{\ell}_{k,\boldsymbol{\theta}_0}=0$ and the sample points from case and control groups of different batches are independent, the Linderberg-Feller conditions can be verified as follows. Define

$$\begin{cases}X_{n,i}=(\dot{\ell}_{1,\boldsymbol{\theta}}(\mathbf{x}_i,\mathbf{x}_{n+i})/\sqrt{n},0,0)^\top, & i=1,\ldots,n_1',\\X_{n,i}=(0,\dot{\ell}_{2,\boldsymbol{\theta}}(\mathbf{x}_i)/\sqrt{n},0)^\top, & i=n_1'+1,\ldots,n_1,\\X_{n,i}=(0,0,\dot{\ell}_{3,\boldsymbol{\theta}}(\mathbf{x}_i)/\sqrt{n})^\top, & i=n_1,\ldots,n.\end{cases}$$

Then for any $\epsilon > 0$,

$$\sum_{i=1}^{n} \mathbb{E}\Big[\|X_{n,i}\|^2 \mathbf{1}\{\|X_{n,i}\| > \epsilon\}\Big] \\ \leq \max_k \mathbb{E}\Big[|\dot{\ell}_{k,\boldsymbol{\theta}}|^2 \mathbf{1}\{|\dot{\ell}_{k,\boldsymbol{\theta}}/\sqrt{n}| > \epsilon\}\Big].$$

Since $\mathbb{E}[|\dot{\ell}_{k,\boldsymbol{\theta}}|^2] < \infty$ for $k = 1, 2, 3$, it suffices to show that

$$\mathbf{1}\{|\dot{\ell}_{k,\boldsymbol{\theta}}/\sqrt{n}| > \epsilon\} \to^{a.s} 0,$$

which is true because $1/\sqrt{n} \to 0$. Meanwhile,

$$\sum_{i=1}^{n} \mathrm{cov}(X_{n,i}) \to \\ \begin{pmatrix} r_1 r_1' \mathbb{E}[\dot{\ell}_{1,\boldsymbol{\theta}_0} \dot{\ell}_{1,\boldsymbol{\theta}_0}^\top] & & \\ & r_1(1-r_1')\mathbb{E}[\dot{\ell}_{2,\boldsymbol{\theta}_0} \dot{\ell}_{2,\boldsymbol{\theta}_0}^\top] & \\ & & (1-r_1)\mathbb{E}[\dot{\ell}_{3,\boldsymbol{\theta}_0} \dot{\ell}_{3,\boldsymbol{\theta}_0}^\top] \end{pmatrix}. \tag{10}$$

Thus, $\sum_{i=1}^{n} X_{n,i} = (\frac{n_1'}{\sqrt{n}} \mathbb{P}_{n_1'}^{(1)} \dot{\ell}_{1,\boldsymbol{\theta}_0}, \frac{n_1 - n_1'}{\sqrt{n}} \mathbb{P}_{n_1 - n_1'}^{(2)} \dot{\ell}_{2,\boldsymbol{\theta}_0}, \frac{n_2}{\sqrt{n}} \mathbb{P}_{n_2}^{(3)} \dot{\ell}_{3,\boldsymbol{\theta}_0})^\top$ is jointly normal with the asymptotic covariance in (10). The conclusion thus follows.

1.2.1 Verification of the Lipschitz-Hessian Condition

We now verify the Lipschitz-Hessian condition under our model setting. For any positive definite symmetric matrix $\mathbf{A}$, we have $\|\mathbf{A}\|_{op} \leq \|\mathbf{A}\|_F$, where $\|\mathbf{A}\|_F$ denotes the Frobenius norm of $\mathbf{A}$. For $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q) \in \mathbb{R}^q$, we have

$$\|\nabla^2 \ell_{1,\boldsymbol{\theta}_1} - \nabla^2 \ell_{1,\boldsymbol{\theta}_2}\|_{op} \leq \|\nabla^2 \ell_{1,\boldsymbol{\theta}_1} - \nabla^2 \ell_{1,\boldsymbol{\theta}_2}\|_F \\ = \Big(\sum_{s=1}^{q} \sum_{t=1}^{q} |(\nabla^2 \ell_{1,\boldsymbol{\theta}_1} - \nabla^2 \ell_{1,\boldsymbol{\theta}_2})_{st}|^2 \Big)^{1/2},$$

and for any $s, t \in \{1, \ldots, q\}$,

$$\Big| \frac{\partial^2 \ell_{1,\boldsymbol{\theta}_1}(\mathbf{x}_i, \mathbf{x}_{n+i})}{\partial \theta_s \partial \theta_t} - \frac{\partial^2 \ell_{1,\boldsymbol{\theta}_2}(\mathbf{x}_i, \mathbf{x}_{n+i})}{\partial \theta_s \partial \theta_t} \Big| \leq M_{st}(\mathbf{x}_i, \mathbf{x}_{n+i}) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

with $M_{st}(\mathbf{x}_i, \mathbf{x}_{n+i}) := \sup_{\boldsymbol{\theta}} \|\nabla \frac{\partial^2 \ell_{1,\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_{n+i})}{\partial \theta_s \partial \theta_t}\|$ because of the smoothness of the log-likelihood and the compactness of Θ. Hence,

$$\|\nabla^2 \ell_{1,\boldsymbol{\theta}_1} - \nabla^2 \ell_{1,\boldsymbol{\theta}_2}\|_{op} \leq \left(\sum_{s=1}^{q} \sum_{t=1}^{q} M_{st}^2(\mathbf{x}_i, \mathbf{x}_{n+i}) \right)^{1/2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Using the form of $\ell_{1,\boldsymbol{\theta}}$, we have the following observations:

- Taking any derivative w.r.t the variance/covariance parameters (ρ, σ_1 and σ_2) will not change the degree of the polynomials (w.r.t to y_i's). For instance,

$$\frac{\partial^3 \ell_{1,\boldsymbol{\theta}}}{\partial \sigma_1^3} = -\frac{4}{\sigma_1^3} + \frac{1}{1-\rho^2}\Big[\frac{12(y_i - \mathbf{z}_i^\top \mathbf{b})^2}{\sigma_1^5} + \frac{12\rho(y_i - \mathbf{z}_i^\top \mathbf{b})(y_{n+i} - \mathbf{z}_i^\top \mathbf{b} - a_1)}{\sigma_1^4 \sigma_2}\Big].$$

- Any third derivative w.r.t a_0, a_1 and $\mathbf{b}$ is 0. This can be seen from

$$\frac{\partial^3 \ell_{1,\boldsymbol{\theta}}}{\partial a_0^3} = 0, \quad \frac{\partial^3 \ell_{1,\boldsymbol{\theta}}}{\partial a_1^3} = 0, \quad \frac{\partial^2 \ell_{1,\boldsymbol{\theta}}}{\partial \mathbf{b}^\top \partial \mathbf{b}} = \det(\boldsymbol{\Sigma}) \mathbf{z}_i \mathbf{z}_i^\top.$$

  Moreover, the third derivative w.r.t any component b_j in $\mathbf{b}$ is zero.

One can verify that all the third derivatives are dominated by some terms that are proportional to $\mathbf{z}_i \mathbf{z}_i^\top$, $(\epsilon_i^{(1)})^2$ and $(\epsilon_{n+i}^{(2)})^2$. In particular, we have

$$M_1(\mathbf{x}_i, \mathbf{x}_{n+i}) = A(\mathbf{1}^\top \mathbf{z}_i \mathbf{z}_i^\top \mathbf{1}) + B(\epsilon_i^{(1)})^2 + C(\epsilon_{n+i}^{(2)})^2,$$

for sufficiently large positive constants A, B and C such that

$$\|\nabla^2 \ell_{1,\boldsymbol{\theta}_1} - \nabla^2 \ell_{1,\boldsymbol{\theta}_2}\|_{op} \leq M_1(\mathbf{x}_i, \mathbf{x}_{n+i}) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

We have assumed that $\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top] = \boldsymbol{\Sigma}_z$, $\mathrm{Var}[\epsilon_i^{(1)}] = \sigma_1^2$ and $\mathrm{Var}[\epsilon_i^{(2)}] = \sigma_2^2$ in Condition 1-3, which implies that $\mathbb{E}[M_1] < \infty$. Similar arguments apply to $\ell_{2,\boldsymbol{\theta}}$ and $\ell_{3,\boldsymbol{\theta}}$.

Supplementary Section 2 Computational algorithm and statistical inference

In this section, we describe the parameters updating scheme in detail. We first introduce some notation. Let $R_1 = \sum_{i \in S_1} (y_i - \mathbf{z}_i^\top \mathbf{b})/n_1'$, $R_2 = \sum_{i \in T_2} (y_i - \mathbf{z}_i^\top \mathbf{b})/n_2$, $R_3 = \sum_{i \in S_1} (y_{n+i} - \mathbf{z}_{n+i}^\top \mathbf{b})/n_1'$. Let $W_{(S_1)} = \sum_{i \in S_1} (y_i - \mu_{1i})^2$, $W_{(C_2)} = \sum_{i \in S_1} (y_{n+i} - \mu_{3i})^2$, $W_{(S_1 \cdot C_2)} = \sum_{i \in S_1} (y_i - \mu_{1i})(y_{n+i} - \mu_{3i})$, $W_{(C_1 \setminus S_1)} = \sum_{i \in C_1 \setminus S_1} (y_i - \mu_{1i})^2$, and $W_{(T_2)} = \sum_{i \in T_2} (y_i - \mu_{2i})^2$.

Taking the first-order derivative of the objective function with respect to a_1 and a_0 separately and setting the expressions to 0, we obtain

$$\begin{aligned} &\frac{n_1'(R_3 - a_1)}{\sigma_2^2(1-\rho^2)} - \frac{\rho n_1' R_1}{\sigma_1 \sigma_2 (1-\rho^2)} + \frac{n_2(R_2 - a_0 - a_1)}{\sigma_2^2} = 0, \\ &R_2 - a_0 - a_1 = 0. \end{aligned}$$

The explicit forms of the updating rules for a_0 and a_1 are given respectively by

$$a_0 = R_2 - (R_3 - \frac{\rho\sigma_2}{\sigma_1}R_1), \tag{11}$$

and

$$a_1 = R_3 - \frac{\rho\sigma_2}{\sigma_1}R_1. \tag{12}$$

For ρ, σ_1 and σ_2, there is no closed-form updating rule. The correlation ρ is updated by finding the real positive root of the cubic equation

$$n_1'\rho(1-\rho^2) = \rho(\frac{W_{(S_1)}}{\sigma_1^2} + \frac{W_{(C_2)}}{\sigma_2^2}) - \frac{(1+\rho^2)W_{(S_1\cdot C_2)}}{\sigma_1\sigma_2}.$$

The standard deviations σ_1 and σ_2 are updated via finding the positive roots of the following two quadratic equations

$$n_1(1-\rho^2)\sigma_1^2 = W_{(S_1)} + (1-\rho^2)W_{(C_1\setminus S_1)} - \frac{\rho W_{(S_1\cdot C_2)}\sigma_1}{\sigma_2},$$
$$(n_1'+n_2)(1-\rho^2)\sigma_2^2 = W_{(C_2)} + (1-\rho^2)W_{(T_2)} - \frac{\rho W_{(S_1\cdot C_2)}\sigma_2}{\sigma_1}.$$

To describe the updating rule for $\mathbf{b}$, we need to introduce some additional symbols. Suppose the dimension of $\mathbf{b}$ is p. Write $\mathbf{Z}^{(S_1)} = \mathbf{Z}^{(C_2)} = (\mathbf{z}_1, \ldots, \mathbf{z}_{n_1'})^\top$ as the $n_1' \times p$ matrix of remeasured covariates with the rows being the covariates for each sample. Let $\mathbf{y}^{(S_1)} = (y_1, \ldots, y_{n_1'})^\top$ be the corresponding response vector of the control group in the first batch. Also, we define $\mathbf{Z}^{(T_2)}$ as the $n_2 \times p$ design matrix of the treatment group in the second batch and $\mathbf{y}^{(T_2)}$ as the corresponding response vector. Let $\mathbf{Z}^{(C_1\setminus S_1)} = (\mathbf{z}_{n_1'+1}, \ldots, \mathbf{z}_{n_1})$ be the $(n_1 - n_1') \times p$ matrix of covariates that are not remeasured and $\mathbf{y}^{(C_1\setminus S_1)}$ be the corresponding response in the first batch. Moreover, we let $\mathbf{y}^{(C_2)} = (y_{n+1}, \ldots, y_{n+n_1'})^\top$ be the vector of responses of the remeasured samples in the control group in the second batch. For any $n \times p$ matrix $\mathbf{Z}$, we define $\bar{\mathbf{Z}} = \mathbf{Z}^\top \mathbf{1}/n$ as the $p \times 1$ vector that contains the mean value for each column of $\mathbf{Z}$, and let $\mathbf{Z}_c$ be the centralized matrix by subtracting the mean vector $\bar{\mathbf{Z}}$ from each row of $\mathbf{Z}$. Similarly, we let $\mathbf{y}_c$ be the centralized vector by subtracting the mean value from each element in $\mathbf{y}$.

Algorithm 1: The alternate parameter updating for MLE.

Input: initial value $\theta^{(0)} = (a_0^{(0)}, a_1^{(0)}, \mathbf{b}^{(0)}, \rho^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)})$

Output: MLE $\hat{\boldsymbol{\theta}} = (\hat{a}_0, \hat{a}_1, \hat{\mathbf{b}}, \hat{\rho}, \hat{\sigma}_1, \hat{\sigma}_2)$

Set $k = 1$

Set $\mu_{1i}^{(0)} = \mathbf{z}_i^\top \mathbf{b}^{(0)}$, $\mu_{2i}^{(0)} = a_0^{(0)} + a_1^{(0)} + \mathbf{z}_i^\top \mathbf{b}^{(0)}$, $\mu_{3i}^{(0)} = a_1^{(0)} + \mathbf{z}_i^\top \mathbf{b}^{(0)}$.
Denote $W_{(S_1)}^{(0)} = \sum_{i\in S_1}(y_i - \mu_{1i}^{(0)})^2$, $W_{(C_2)}^{(0)} = \sum_{i\in S_1}(y_{n+i} - \mu_{3i}^{(0)})^2$,
$W_{(S_1\cdot C_2)}^{(0)} = \sum_{i\in S_1}(y_i - \mu_{1i}^{(0)})(y_{n+i} - \mu_{3i}^{(0)})$,
$W_{(C_1\setminus S_1)}^{(0)} = (y_i - \mu_{1i}^{(0)})^2, W_{(T_2)}^{(0)} = \sum_{i\in T_2}(y_i - \mu_{2i}^{(0)})^2$.

repeat

Compute $\rho^{(k)}$ by solving
$$n_1'\rho(1-\rho^2) = \rho\Big(\frac{W_{(S_1)}^{(k-1)}}{(\sigma_1^{(k-1)})^2} + \frac{W_{(C_2)}^{(k-1)}}{(\sigma_2^{(k-1)})^2}\Big) - \frac{(1+\rho^2)W_{(S_1\cdot C_2)}^{(k-1)}}{\sigma_1^{(k-1)}\sigma_2^{(k-1)}}.$$

Compute $\sigma_1^{(k)}$ by solving
$$n_1\big(1-(\rho^{(k)})^2\big)\sigma_1^2 = W_{(S_1)}^{(k-1)} + \big(1-(\rho^{(k)})^2\big)W_{(C_1\setminus S_1)}^{(k-1)} - \frac{\rho^{(k)}W_{(S_1\cdot C_2)}^{(k-1)}\sigma_1}{\sigma_2^{(k-1)}}.$$

Compute $\sigma_2^{(k)}$ by solving
$$(n_1'+n_2)\big(1-(\rho^{(k)})^2\big)\sigma_2^2 = W_{(C_2)}^{(k-1)} + \big(1-(\rho^{(k)})^2\big)W_{(T_2)}^{(k-1)} - \frac{\rho^{(k)}W_{(S_1\cdot C_2)}^{(k-1)}\sigma_2}{\sigma_1^{(k-1)}}.$$

Compute $a_1^{(k)} = R_3^{(k-1)} - \frac{\rho^{(k)}\sigma_2^{(k)}}{\sigma_1^{(k)}}R_1^{(k-1)}$, $a_0^{(k)} = R_2^{(k-1)} - a_1^{(k)}$.

Update $\mathbf{S}^{(k)}, \mathbf{t}^{(k)}$ based on (14), (15) given $\rho^{(k)}, \sigma_1^{(k)}, \sigma_2^{(k)}$.

Compute $\mathbf{b}^{(k)} = \mathbf{S}^{(k)-1}\mathbf{t}^{(k)}$

Update $R_1^{(k)}, R_2^{(k)}, R_3^{(k)}$ given $\mathbf{b}^{(k)}$

Update $W_{(S_1)}^{(k)}, W_{(C_2)}^{(k)}, W_{(S_1\cdot C_2)}^{(k)}, W_{(C_1\setminus S_1)}^{(k)}, W_{(T_2)}^{(k)}$ given $a_0^{(k)}, a_1^{(k)}, \mathbf{b}^{(k)}$

Set $k = k + 1$.

until *convergence.*

The final MLE estimator is
$\hat{\boldsymbol{\theta}} = (\hat{a}_0, \hat{a}_1, \hat{\mathbf{b}}, \hat{\rho}, \hat{\sigma}_1, \hat{\sigma}_2) = (a_0^{(k)}, a_1^{(k)}, \mathbf{b}^{(k)}, \rho^{(k)}, \sigma_1^{(k)}, \sigma_2^{(k)})$.

Using the first order condition for $\mathbf{b}$, and (11) and (12), we have

$$\begin{aligned}
&\frac{\mathbf{Z}^{(C_2)\top}(\mathbf{y}^{(S_1)} - \mathbf{Z}^{(C_2)}\mathbf{b})}{\sigma_1^2(1-\rho^2)} - \frac{\rho\mathbf{Z}^{(C_2)\top}(\mathbf{y}^{(C_2)} - \mathbf{Z}^{(C_2)}\mathbf{b} + R_3 - \frac{\rho\sigma_2}{\sigma_1}R_1)}{\sigma_1\sigma_2(1-\rho^2)} \\
&- \frac{\rho(\mathbf{Z}^{(C_2)} - \frac{\rho\sigma_2}{\sigma_1}\bar{\mathbf{Z}}^{(C_2)})^\top(\mathbf{y}^{(S_1)} - \mathbf{Z}^{(C_2)}\mathbf{b})}{\sigma_1\sigma_2(1-\rho^2)} \\
&+ \frac{(\mathbf{Z}^{(C_2)} - \frac{\rho\sigma_2}{\sigma_1}\bar{\mathbf{Z}}^{(C_2)})^\top(\mathbf{y}^{(C_2)} - \mathbf{Z}^{(C_2)}\mathbf{b} + R_3 - \frac{\rho\sigma_2}{\sigma_1}R_1)}{\sigma_2^2(1-\rho^2)} \\
&+ \frac{\mathbf{Z}^{(C_1\setminus S_1)\top}(\mathbf{y}^{(C_1\setminus S_1)} - \mathbf{Z}^{(C_1\setminus S_1)}\mathbf{b})}{\sigma_1^2} + \frac{\mathbf{Z}_c^{(T_2)\top}(\mathbf{y}_c^{(T_2)} - \mathbf{Z}_c^{(T_2)}\mathbf{b})}{\sigma_2^2} = 0.
\end{aligned} \tag{13}$$

Re-arranging the above equation by putting the terms related to $\mathbf{b}$ on the left-hand side and the rest on the right-hand side, we obtain a linear equation $\mathbf{Sb} = \mathbf{t}$, where $\mathbf{S}$ and $\mathbf{t}$ depend on $(\rho, \sigma_1, \sigma_2)$. The forms of $\mathbf{S}$ and $\mathbf{t}$ are given respectively by

$$\begin{aligned}\mathbf{S} = &\frac{\mathbf{Z}^{(C_2)\top}\mathbf{Z}^{(C_2)}}{\sigma_1^2(1-\rho^2)} - \frac{\rho\mathbf{Z}^{(C_2)\top}(\mathbf{Z}^{(C_2)} - (1-\frac{\rho\sigma_2}{\sigma_1})\bar{\mathbf{Z}}^{(C_2)})}{\sigma_1\sigma_2(1-\rho^2)} \\ &- \frac{\rho(\mathbf{Z}^{(C_2)} - (1-\frac{\rho\sigma_2}{\sigma_1})\bar{\mathbf{Z}}^{(C_2)})^\top\mathbf{Z}^{(C_2)}}{\sigma_1\sigma_2(1-\rho^2)} \\ &+ \frac{(\mathbf{Z}^{(C_2)} - (1-\frac{\rho\sigma_2}{\sigma_1})\bar{\mathbf{Z}}^{(C_2)})^\top(\mathbf{Z}^{(C_2)} - (1-\frac{\rho\sigma_2}{\sigma_1})\bar{\mathbf{Z}}^{(C_2)})}{\sigma_2(1-\rho^2)} \\ &+ \frac{\mathbf{Z}^{(C_1\setminus S_1)\top}\mathbf{Z}^{(C_1\setminus S_1)}}{\sigma_1^2} + \frac{\mathbf{Z}_c^{(T_2)\top}\mathbf{Z}_c^{(T_2)}}{\sigma_2^2},\end{aligned} \tag{14}$$

and

$$\begin{aligned}\mathbf{t} = &\frac{\mathbf{Z}^{(C_2)\top}\mathbf{y}^{S_1}}{\sigma_1^2(1-\rho^2)} - \frac{\rho\mathbf{Z}^{(C_2)\top}(\mathbf{y}^{(C_2)} - \bar{\mathbf{y}}^{(C_2)} + \frac{\rho\sigma_2}{\sigma_1}\bar{\mathbf{y}}^{(S_1)})}{\sigma_1\sigma_2(1-\rho^2)} \\ &- \frac{\rho(\mathbf{Z}^{(C_2)} - (1-\frac{\rho\sigma_2}{\sigma_1})\bar{\mathbf{Z}}^{(C_2)})^\top\mathbf{y}^{(S_1)}}{\sigma_1\sigma_2(1-\rho^2)} \\ &+ \frac{(\mathbf{Z}^{(C_2)} - (1-\frac{\rho\sigma_2}{\sigma_1})\bar{\mathbf{Z}}^{(C_2)})^\top(\mathbf{y}^{(C_2)} - \bar{\mathbf{y}}^{(C_2)} + \frac{\rho\sigma_2}{\sigma_1}\bar{\mathbf{y}}^{(S_1)})}{\sigma_2^2(1-\rho^2)} \\ &+ \frac{\mathbf{Z}^{(C_1\setminus S_1)\top}\mathbf{y}^{(C_1\setminus S_1)}}{\sigma_1^2} + \frac{\mathbf{Z}_c^{(T_2)\top}\mathbf{y}_c^{(T_2)}}{\sigma_2^2}.\end{aligned} \tag{15}$$

We summarize the iterative updating procedure in Algorithm 1.

To estimate the variance of $\hat{a}_0$, let $\tilde{\mathbf{Z}} = (1 - \hat{\rho}\hat{\sigma}_2/\hat{\sigma}_1)\mathbf{Z}^{(C_2)}$. Note that $\hat{a}_0$ can be written as

$$\begin{aligned}\hat{a}_0 =&\frac{\mathbf{1}_{n_2}^\top}{n_2}(\mathbf{y}^{(T_2)} - \mathbf{Z}^{(T_2)}\hat{\mathbf{b}}) - \hat{a}_1 \\ =&(-\frac{\mathbf{1}_{n_2}^\top}{n_2}\mathbf{Z}^{(T_2)}\mathbf{A} + \frac{\mathbf{1}_{n_1'}^\top}{n_1'}\tilde{\mathbf{Z}}\mathbf{A} + \frac{\mathbf{1}_{n_1'}^\top}{n_1'}\frac{\hat{\rho}\hat{\sigma}_2}{\hat{\sigma}_1})\mathbf{y}^{(S_1)} + (\frac{\mathbf{1}_{n_2}^\top}{n_2} - \frac{\mathbf{1}_{n_2}^\top}{n_2}\mathbf{Z}^{(T_2)}\mathbf{B} + \frac{\mathbf{1}_{n_1'}^\top}{n_1'}\tilde{\mathbf{Z}}\mathbf{B})\mathbf{y}^{(T_2)} \\ &+(-\frac{\mathbf{1}_{n_2}^\top}{n_2}\mathbf{Z}^{(T_2)}\mathbf{C} - \frac{\mathbf{1}_{n_1'}^\top}{n_1'} + \frac{\mathbf{1}_{n_1'}^\top}{n_1'}\tilde{\mathbf{Z}}\mathbf{C})\mathbf{y}^{(C_2)} + (-\frac{\mathbf{1}_{n_2}^\top}{n_2}\mathbf{Z}^{(T_2)}\mathbf{D} + \frac{\mathbf{1}_{n_1'}^\top}{n_1'}\tilde{\mathbf{Z}}\mathbf{D})\mathbf{y}^{(C_1\setminus S_1)} \\ =&\mathbf{c}_1^\top\mathbf{y}^{(S_1)} + \mathbf{c}_2^\top\mathbf{y}^{(T_2)} + \mathbf{c}_3^\top\mathbf{y}^{(C_2)} + \mathbf{c}_4^\top\mathbf{y}^{(C_1\setminus S_1)},\end{aligned} \tag{16}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and $\mathbf{D}$ are the coefficient matrices such that $\hat{\mathbf{b}} = \mathbf{A}\mathbf{y}^{(S_1)} + \mathbf{B}\mathbf{y}^{(T_2)} + \mathbf{C}\mathbf{y}^{(C_2)} + \mathbf{D}\mathbf{y}^{(C_1 \setminus S_1)}$. The explicit forms are given by

$$\mathbf{A} = \hat{\mathbf{S}}^{-1}\mathbf{Z}^{(S_1)\top}\Big(\frac{\mathbf{I}_{n_1'}}{\hat{\sigma}_1^2(1-\hat{\rho}^2)} - \frac{\hat{\rho}}{\hat{\sigma}_1\hat{\sigma}_2(1-\hat{\rho}^2)}(\mathbf{I}_{n_1'} - \frac{\mathbf{1}_{n_1'}\mathbf{1}_{n_1'}^\top}{n_1'}) - \frac{\hat{\rho}^2}{\hat{\sigma}_1^2(1-\hat{\rho}^2)}\frac{\mathbf{1}_{n_1'}\mathbf{1}_{n_1'}^\top}{n_1'}\Big),$$

$$\mathbf{B} = \hat{\mathbf{S}}^{-1}\frac{\mathbf{Z}^{(T_2)\top}}{\hat{\sigma}_2^2}(\mathbf{I}_{n_2} - \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top/n_2),$$

$$\mathbf{C} = \hat{\mathbf{S}}^{-1}\Big(\frac{1}{\hat{\sigma}_2^2(1-\hat{\rho}^2)} - \frac{\hat{\rho}}{\hat{\sigma}_1\hat{\sigma}_2(1-\hat{\rho}^2)}\Big)\mathbf{Z}^{(S_1)\top}(\mathbf{I}_{n_1'} - \mathbf{1}_{n_1'}\mathbf{1}_{n_1'}^\top/n_1'),$$

$$\mathbf{D} = \hat{\mathbf{S}}^{-1}\frac{\mathbf{Z}^{(C_1\setminus S_1)\top}}{\hat{\sigma}_1^2},$$

where $\hat{\mathbf{S}}$ is defined in the same way as $\mathbf{S}$ by replacing $(\rho, \sigma_1, \sigma_2)$ with $(\hat{\rho}, \hat{\sigma_1}, \hat{\sigma}_2)$. The variance of $\hat{a}_0$ can then be estimated by

$$\widehat{\mathrm{Var}}(\hat{a}_0) = \hat{\sigma}_1^2(\mathbf{c}_1^\top\mathbf{c}_1 + \mathbf{c}_4^\top\mathbf{c}_4) + \hat{\sigma}_2^2(\mathbf{c}_2^\top\mathbf{c}_2 + \mathbf{c}_3^\top\mathbf{c}_3) + 2\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2\mathbf{c}_1^\top\mathbf{c}_3. \tag{17}$$

Supplementary Section 3 The location-scale matching approach

The location-scale (LS) approach assumes a model for the location (mean) and scale (variance) with the batches. By standardizing the means and variances across the batches, the batch effect can then be removed. The estimation of the scale proportion is $\hat{\sigma}_2/\hat{\sigma}_1$, where $\hat{\sigma}_2$ is the standard deviation of $\mathbf{y}^{(C_2)}$, and $\hat{\sigma}_1$ is the standard deviation of $\mathbf{y}^{(S_1)}$. Then the batch-adjusted data for the first batch, $\mathbf{y}_*^{(C_1)}$, are given by

$$\mathbf{y}_*^{(C_1)} = \frac{\hat{\sigma}_2}{\hat{\sigma}_1}(\mathbf{y}^{(C_1)} - \bar{\mathbf{y}}^{(C_1)}) + \bar{\mathbf{y}}^{(C_1)} + \bar{\mathbf{y}}^{(C_2)} - \bar{\mathbf{y}}^{(S_1)}.$$

We can assume the adjusted control samples in the first batch, together with the case samples in the second batch $(\mathbf{y}_*^{(C_1)}, \mathbf{y}^{(T_2)})$ follow the model

$$\begin{aligned}
&\text{Control (batch 2):} \quad y_i = a_1 + \mathbf{z}_i^\top\mathbf{b} + \epsilon_i^{(2)}, \quad i = 1, \ldots, n_1,\\
&\text{Case (batch 2):} \quad y_i = a_0 + a_1 + \mathbf{z}_i^\top\mathbf{b} + \epsilon_i^{(2)}, \quad i = n_1+1, \ldots, n.
\end{aligned}$$

Therefore, we can use the least squares to get the parameter estimates.

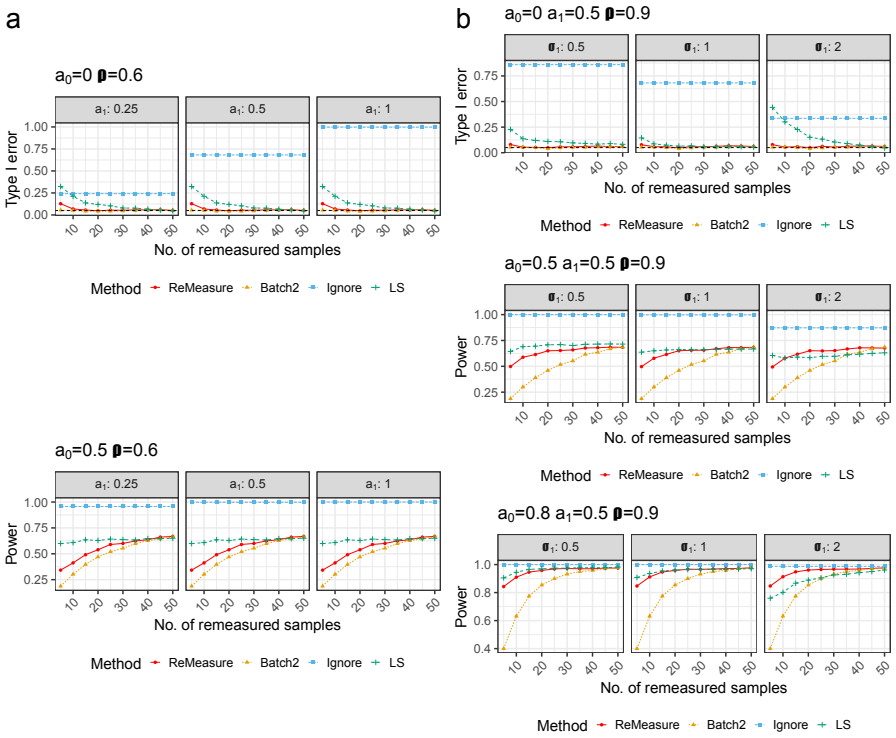
	Batch2	Ignore	LS	ReMeasure
$a_1 = 0.5$		$a_0 = 0.5$		
$\sigma_1 = 0.5, \rho = 0.3$				
$n_1' = 5$	0.209(0.010)	0.267(0.005)	**0.206(0.010)**	0.239(0.012)
$n_1' = 10$	**0.111(0.005)**	0.267(0.005)	0.114(0.005)	0.111(0.005)
$n_1' = 15$	0.083(0.004)	0.267(0.005)	0.083(0.004)	**0.082(0.004)**
$n_1' = 20$	0.068(0.003)	0.267(0.005)	0.070(0.003)	**0.068(0.003)**
$n_1' = 25$	0.061(0.002)	0.267(0.005)	0.063(0.003)	**0.061(0.002)**
$n_1' = 30$	0.055(0.002)	0.267(0.005)	0.056(0.002)	**0.054(0.002)**
$n_1' = 35$	0.051(0.002)	0.267(0.005)	0.053(0.002)	**0.050(0.002)**
$n_1' = 40$	0.046(0.002)	0.267(0.005)	0.048(0.002)	**0.045(0.002)**
$n_1' = 45$	0.044(0.002)	0.267(0.005)	0.045(0.002)	**0.044(0.002)**
$n_1' = 50$	0.042(0.002)	0.267(0.005)	0.043(0.002)	**0.041(0.002)**
$\sigma_1 = 0.5, \rho = 0.6$				
$n_1' = 5$	0.211(0.010)	0.267(0.005)	**0.153(0.007)**	0.174(0.009)
$n_1' = 10$	0.110(0.005)	0.267(0.005)	**0.086(0.004)**	0.087(0.004)
$n_1' = 15$	0.081(0.003)	0.267(0.005)	0.068(0.003)	**0.067(0.003)**
$n_1' = 20$	0.066(0.003)	0.267(0.005)	0.059(0.002)	**0.059(0.002)**
$n_1' = 25$	0.059(0.002)	0.267(0.005)	0.055(0.002)	**0.054(0.002)**
$n_1' = 30$	0.053(0.002)	0.267(0.005)	0.051(0.002)	**0.050(0.002)**
$n_1' = 35$	0.049(0.002)	0.267(0.005)	0.049(0.002)	**0.047(0.002)**
$n_1' = 40$	0.045(0.002)	0.267(0.005)	0.045(0.002)	**0.044(0.002)**
$n_1' = 45$	0.043(0.002)	0.267(0.005)	0.044(0.002)	**0.043(0.002)**
$n_1' = 50$	0.041(0.002)	0.267(0.005)	0.043(0.002)	**0.041(0.002)**
$\sigma_1 = 0.5, \rho = 0.9$				
$n_1' = 5$	0.212(0.009)	0.267(0.005)	0.100(0.005)	**0.078(0.004)**
$n_1' = 10$	0.116(0.005)	0.267(0.005)	0.063(0.003)	**0.053(0.002)**
$n_1' = 15$	0.082(0.003)	0.267(0.005)	0.054(0.002)	**0.048(0.002)**
$n_1' = 20$	0.064(0.003)	0.267(0.005)	0.049(0.002)	**0.046(0.002)**
$n_1' = 25$	0.056(0.002)	0.267(0.005)	0.047(0.002)	**0.045(0.002)**
$n_1' = 30$	0.051(0.002)	0.267(0.005)	0.046(0.002)	**0.043(0.002)**
$n_1' = 35$	0.046(0.002)	0.267(0.005)	0.044(0.002)	**0.043(0.002)**
$n_1' = 40$	0.044(0.002)	0.267(0.005)	0.043(0.002)	**0.042(0.002)**
$n_1' = 45$	0.042(0.002)	0.267(0.005)	0.043(0.002)	**0.042(0.002)**
$n_1' = 50$	0.041(0.002)	0.267(0.005)	0.043(0.002)	**0.041(0.002)**

Supplementary Table 1: **Mean square error (MSE) of a_0 for different procedures when both sample sizes $n_1 = n_2 = 50$.** We present the average MSE based on 1000 replications, with the number in the parenthesis indicating the SEM. This comparison provides insights into the performance of these estimation procedures under different parameter settings.

Supplementary Section 4 Additional simulations

4.1 The effect of the batch location parameter

To examine the influence of batch location effect a_1 on method performance, we fix the batch scale parameter $\sigma_1 = 0.5$ and the effect size $a_0 = 0.5$. Supplementary Figure 1a shows that a_1 has substantial impacts on "Ignore" while the other methods are not affected by a_1.



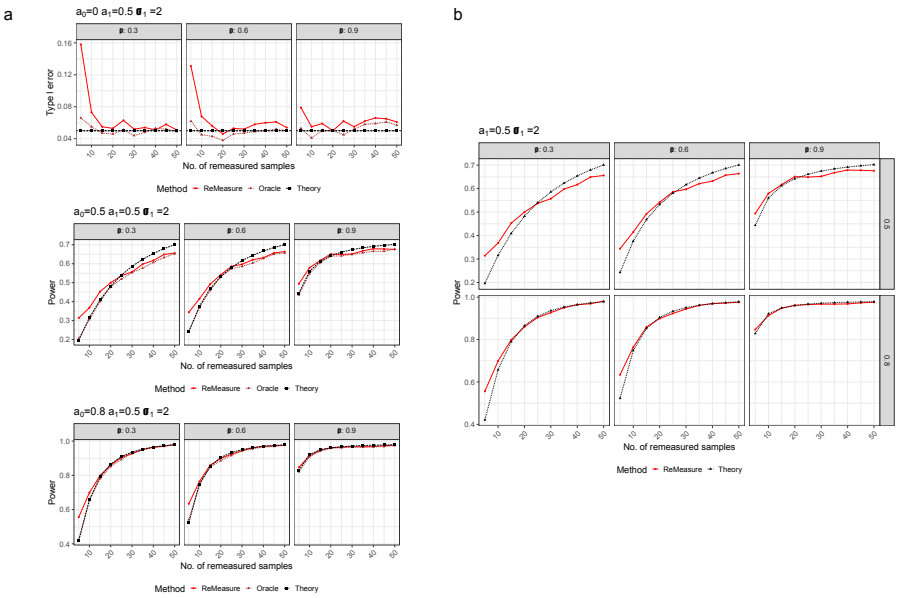
Supplementary Fig. 1: **Assessing empirical type I error and power for testing the biological effect $a_0 = 0$ for different procedures with sample sizes $n_1 = n_2 = 50$.** (a) Panels organized from left to right present the results under different values of a_1 (the batch location parameter), based on 1000 replications. The proposed estimator of a_0 does not depend on the batch location. (b) The "ReMeasure" and "Batch2" are not affected by the choices of σ_1. The dashed line indicates the nominal type I error rate used.

4.2 Performance under large sample sizes

In the case of large sample sizes (i.e., $n_1 = n_2 = 200$), we present the MSE and power curves in Supplementary Figures 7a and 7b.

4.3 Performance under non-Gaussian noises

For the proposed method to work, the error in the regression model does not have to follow the Gaussian distribution as we stated in theory. Here we consider the cases where errors follow the centered gamma distribution with the shape parameter 2 and scale parameter 1. We also consider the student t-distribution with degrees of freedom equal to 6. Supplementary Figure 8a presents the power curves under these noise distributions. The power behaviors in these three cases have similar patterns. The power is slightly higher under



Supplementary Fig. 2: **Comparison of empirical type I error and statistical power for "ReMeasure", "Oracle", and "Theory" with** $n_1 = n_2 = 50$**.** (a) In the "Oracle" scenario, we assume σ_1, σ_2, and ρ are known. The "Theory" curve is derived based on the theoretical power formula. (b) A magnified view of the power comparison plot between "ReMeasure" and "Theory," offering a clearer sight.

the non-Gaussian error at the price of a more inflated type I error compared to the Gaussian case.

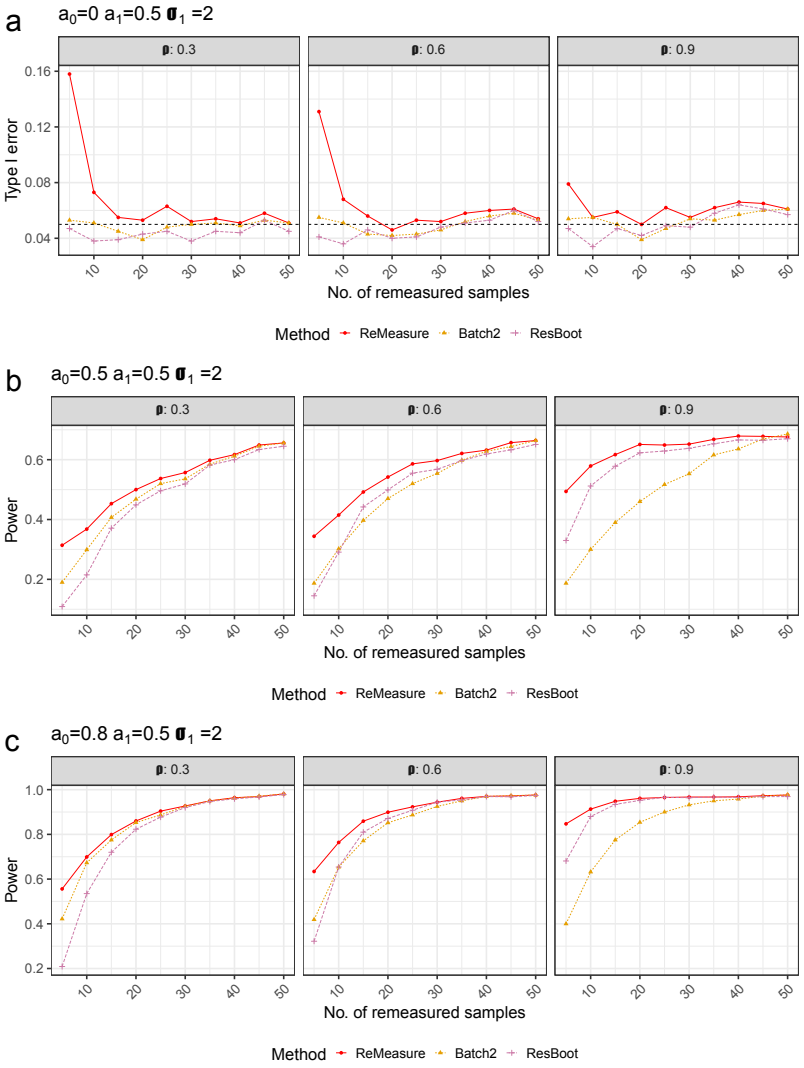
The histograms of the z-statistics in Supplementary Figure 8b are close to that of the standard normal distribution under different noise distributions, which empirically justifies the asymptotic normal approximation.

4.4 Comparison to the naive least square approach

We compare to the naive approach based on the model $Y \sim X + \text{Batch} + Z$. This approach neglects the repeated measure nature and the heterogeneity of variances, which may lead to a reduction in statistical power. Supplementary Figure 9a shows that its power is substantially lower than other competing methods on the same simulated datasets.

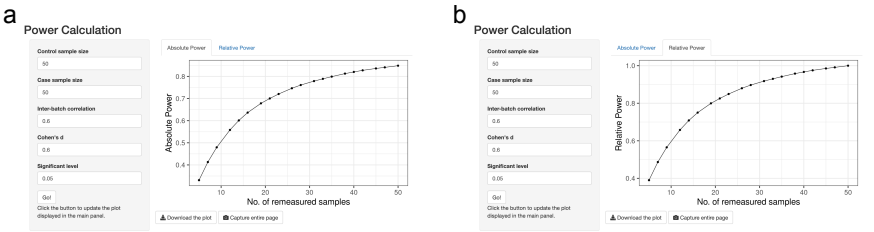
4.5 Comparison to the location-scale matching method using all the control samples

We compare to the location-scale matching method using all independent control samples in the first batch (C_1, C_2). The new approach, "LSind", has milder type I error inflation than the original "LS" based only on the controls that

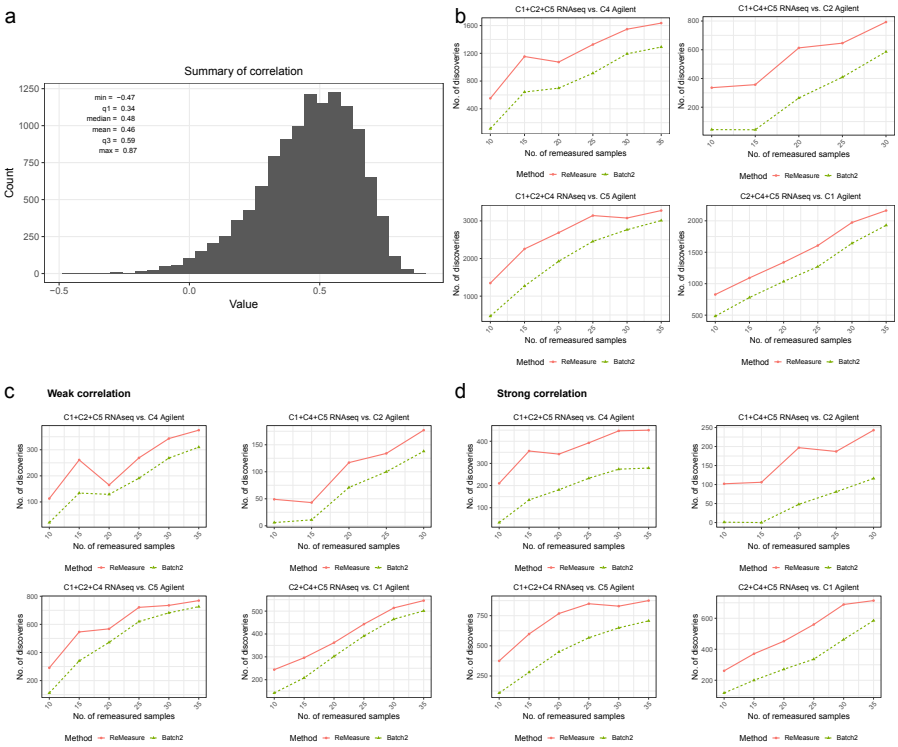


Supplementary Fig. 3: **Assessing empirical type I error and power of the residual bootstrap across parameter settings with sample sizes** $n_1 = n_2 = 50$**.** The bootstrap method can control the type I error at small remeasured sample sizes and deliver comparable power as "ReMeasure" when the remeasured sample size is large. The between-batch correlation, ρ, takes values of $0.3, 0.6$, and 0.9 from the left to the right panel.

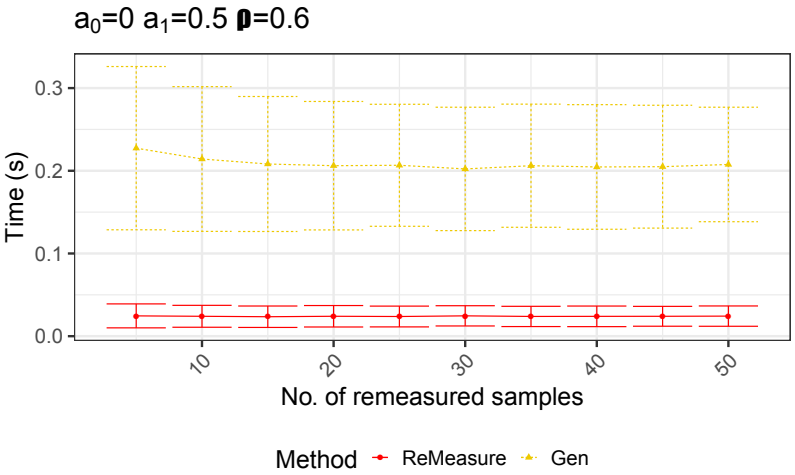
are remeasured. However, its performance deteriorates in the small-sample setting, with a substantially larger type I error above the nominal level compared to the "ReMeasure" and "Batch2" methods (Supplementary Figure 9b).



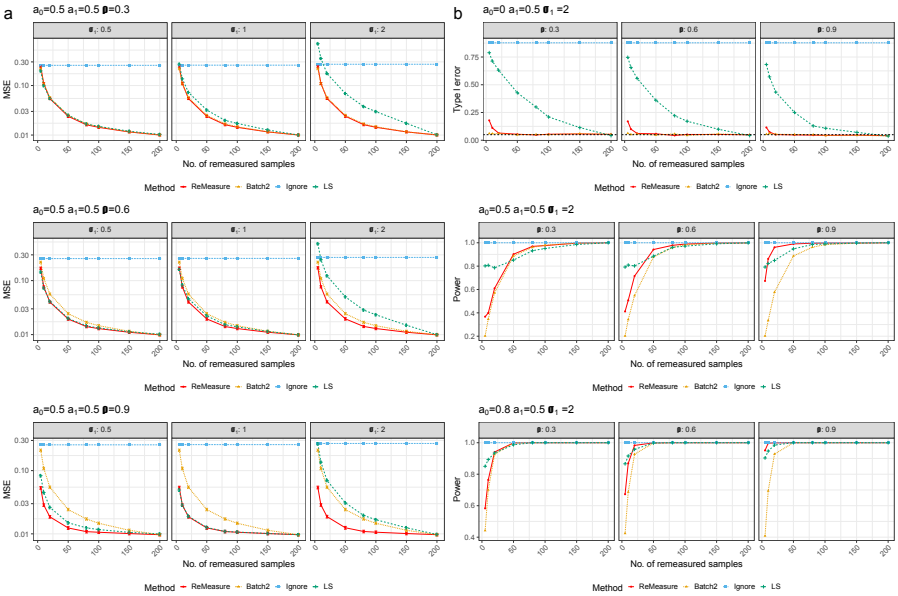
Supplementary Fig. 4: **An example of power analysis using the Shiny app for a confounded case-control study with sample measurement.** (a) The absolute power vs. No. of remeasured samples. (b) The relative power vs. No. of remeasured samples.



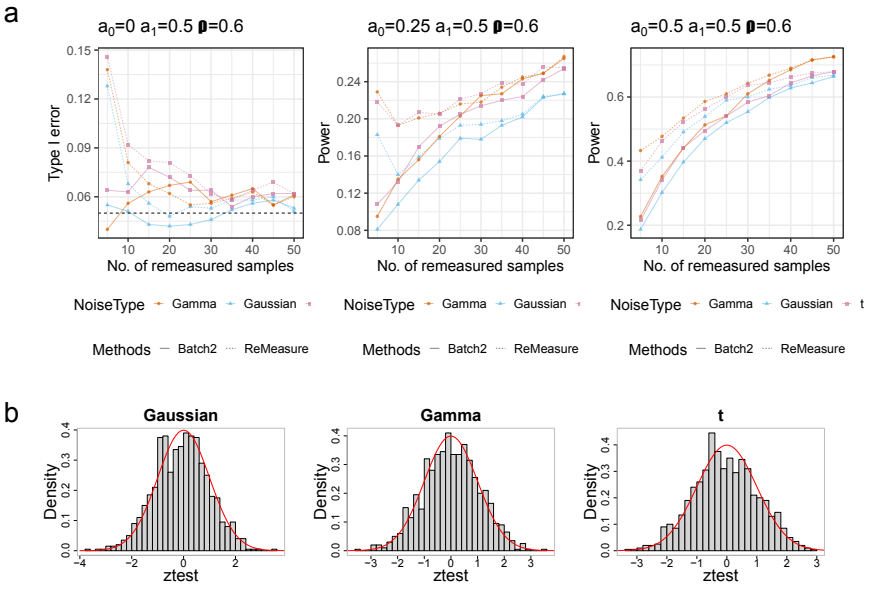
Supplementary Fig. 5: **Correlation of the gene expression between platforms and significant gene discovery in ovarian cancer dataset.** (a) Distribution of the correlation coefficients of the gene expression level between the Agilent and RNA-Seq platform estimated based on 47 common samples. (b), (c) and (d) depicts the number of discovered significant genes vs. the number of remeasured samples, comparing "ReMeasure" to "Batch2". (b) All genes are used. (c) A quarter of the genes with the lowest correlation are used. (d) A quarter of the genes with the highest correlation are used. In the title, "C1+C2+C5 RNAseq vs. C4 Agilent" refers to comparing combined "C1-MES," "C2-IMM," and "C5-PRO" subtypes from the RNAseq platform to the "C4-DIF" subtype from the Agilent platform. The same explanation applies to other titles.



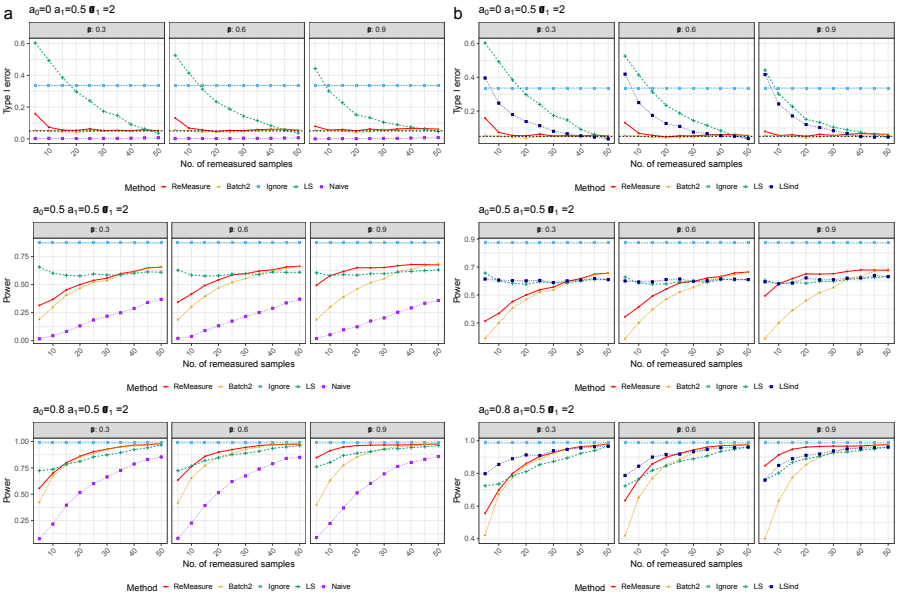
Supplementary Fig. 6: **Computational time of "ReMeasure" compared to the generic MLE optimization approach ("Gen").** "Gen" uses the *optim* function in the R *stat* package. "ReMeasure" is > 10 times faster than "Gen" ($n_1 = n_2 = 50$, and 500 replications are conducted). Data are presented as mean values +/- SD. We run all methods on the same computation platform with 2.40 GHz Intel (R) Xeon (R) E5-2680 v4 28-Core CPU.



Supplementary Fig. 7: **Mean square errors and statistical power of different procedures with sample sizes** $n_1 = n_2 = 200$**.** (a) Data are presented as mean values +/- SEM at varying noise levels (σ_1). y-axis is presented in $\log_{10}$ scale. (b) The type I error and power at different between-batch correlations (ρ). The dashed line indicates the 5% nominal type I error rate used. All results are based on 1000 replications.



Supplementary Fig. 8: **Impact of noise distributions on type I error, power, and Z-statistics.** (a) Type I error and power curves under different noise distributions: centered gamma, Gaussian, and t_6 distributions. (b) The histograms of the Z-statistics are calculated using the proposed method when $n_1' = 50$ under three different noise distributions: centered gamma, Gaussian, and t_6.



Supplementary Fig. 9: **The empirical type I error and power for testing the biological effect $a_0 = 0$ under different parameter settings.** The dashed line indicates the nominal type I error rate used. (a) "naive" is the method that fits a linear regression model based on all samples adjusting the batch variable and ignoring the repeated measurement. (b) "LSind" refers to the location-scale matching method using all control samples in C_1 and C_2.